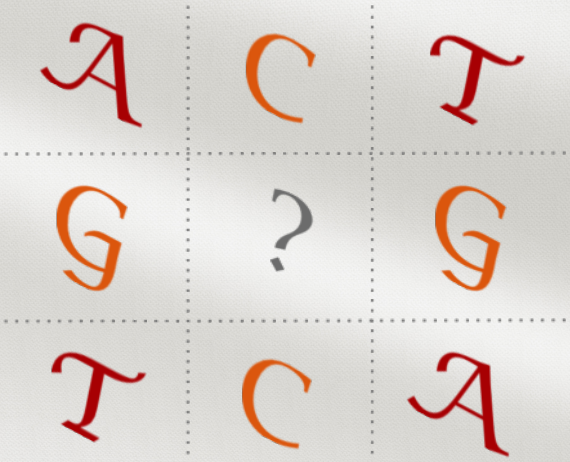
GENOMIC IMPERFECTIONS

Ramesh Hariharan



Strand Life Sciences

In Nature's playground
Characters jog around
Playing hide-and-seek
In all innocent fun
But every so often
A fight breaks out...

PUBLISHED BY STRAND LIFE SCIENCES
WWW.STRANDLS.COM
Pre-Publication Edition, May 2015

Preface

The *Genome*, our parents' gift to us, comprises a staggering 6 billion characters. These characters make us resemble our parents. They also do many other things. And therein lies a fundamental question. What do these 6 billion characters tell us about ourselves? Can they predict the future trajectory of our health? Can they predict if we will grow up to be good sportsmen or scientists or orators?

The answer is far less dramatic than we might like. For instance, identical twins are born with supposedly identical, or at least near-identical, genomes. Yet, one twin sometimes develops autism or schizophrenia or diabetes or cancer while the other doesn't. So genomic characters are not the sole determiners of our fate. They act in concert with various environmental factors. And most frustratingly, in concert with sheer random chance. These extra factors become increasingly vocal as we age, making it difficult for us to predict the future.

Yet, there are a few genomic characters which drive home their agenda almost single-handedly! The stories in this book are real stories about such characters. Their impact is sometimes subtle, as in the opening story on my own red-green color blindness. At other times, their impact is dramatic: siblings whose organs are out of place, a boy whose blood can't carry enough oxygen, a family whose members lose vision in their 40s, a baby with cancer of the eye, a boy who is highly susceptible to cancer, and a middle-aged patient actually battling cancer. And occasionally, their impact is downright tragic: brothers who pass away mysteriously, early in life, and siblings whose hearts fail in their 30s.

Fishing out the relevant genomic characters from among 6 billion candidates in each of these stories is a challenge. This makes each story a detective quest. A quest that attempts to connect the world of clinical practice, built upon centuries of careful observation of external form, to the world of molecular biology, with its deep internal secrets. The path between these two worlds passes through the world of computer algorithms, which distills large amounts of data down to its essence. The interplay between these three worlds is fascinating, as you will hopefully see in the stories that follow.

The stories in this book do touch upon the often grave impact these genomic characters have on people's lives, but only lightly so. The greater focus of this book is on the characters themselves. Regardless, the gravity of their impact is always in the backdrop. Any attempt at simplification and lightness for lucidity is not intended to disrespect this gravity in any way.

These genomic characters come in many varieties. Accordingly, each story deals with a different type of character that needs a different bag of tricks to uncover. And each type of character has a different way of making its presence felt. The purpose of this book is to bring these genomic characters to the attention of a wider audience. The science around these characters will hopefully fascinate. And their ubiquity and stark yet stealthy impact will hopefully inspire many to devote their energies towards taming these characters.

Acknowledgements

This interplay between the worlds of clinical practice, molecular biology and computer algorithms, while fascinating, necessarily pushes us into zones well outside our individual expertise. These stories therefore reflect the combined efforts of several individuals from all three worlds, anchored by our team at *Strand Life Sciences*.

Dr. Meenakshi Bhat, Clinical Geneticist and Professor at the Center for Human Genetics, Bangalore, provided us the first, and most inspiring, introduction to the clinical world. The team at Strand cut its teeth working with Dr. Bhat. Indeed, Chapter 4 describes one of the cases from our collaboration with Dr. Bhat.

Dr. Rajani Battu, Ophthalmologist and Retina Specialist at Narayana Nethralaya, Bangalore,worked with us extensively. The workshops she organized provided a most fascinating introduction to the world of retinal dystrophies. Chapter 2 describes one of our first cases in collaboration with Dr. Battu.

Dr. Subhash Chandra, Cardiologist at Vikram Hospital, Bangalore, and a Fellow of the Royal College of Physicians, gave us our introduction to cardiology. Chapter 3 describes one of our cases in collaboration with Dr. Chandra.

Prof. Stephen Kingsmore, Professor of Pediatrics, University of Missouri-Kansas City School of Medicine, and a Fellow of the Royal College of Pathologists, was a serendipitous connection. Chapter 5 is based on his patient(s). This is the only story here for which genome sequencing data was generated outside the laboratories at Strand. A collaboration with *Illumina*,

the company whose genome sequencing hardware enabled all our stories, gave us access to this data. This chapter recounts our analysis of this data and the discovery of a new gene. Prof. Kingsmore's group discovered this gene independently from the same data and is in the process of publishing this discovery along with several collaborators. Prof. Kingsmore is one of the pioneers of genomic diagnosis in pediatric medicine, so it was an honor for us to compare notes on this case.

Dr. Sunil Bhat, Consultant in Pediatric Hematology, Oncology and Bone Marrow Transplantation at Narayana Hrudayalaya, Bangalore, referred the patient described in Chapter 6.

Dr. Ashwin Mallipatna, Pediatric Ophthalmologist and Retinoblastoma Specialist at Narayana Nethralaya, Bangalore, gave us a fascinating introduction to Retinoblastoma. He also worked with us on optimizing our measurement techniques. Chapter 7 describes one of the cases we worked together on.

Dr. Annie Hasan, Senior Consultant in Genetics, Kamineni Hospitals, Hyderabad, has worked with us on several cases. A few of these have been eye-openers; an extra trick or two was needed to obtain the right diagnoses in these cases. Chapter 8 describes one such case.

Dr. Amit Verma, Consultant in Molecular Oncology and Cancer Genetics, Max Healthcare, New Delhi, has worked with us on several cancer cases. Chapter 9 describes one such case.

My gratitude to all the above clinicians for working with us and for providing consent for inclusion of these stories in this book.

Erica Ramos, *Elliot Margulies*, *Sean Humphray* and *John Peden*, all at *Illumina*, were instrumental in giving us access to the data in Chapter 5 and helped in drawing some of the conclusions.

A huge team at *Strand Life Sciences* provided the foundation for these stories. A few names are listed below. My gratitude to all others at Strand who have not been explicitly listed; this book stands on your shoulders. The stories presented here are but a few of the numerous cases that we've worked on together, touching many lives in the process.

Vaijayanti Gupta, *Binay Panda*, *Preveen Ramamoorthy* and *Satish Sankaran* led the molecular biology laboratories in which genomes of

most of the patients in these stories were deciphered.

Vamsi Veeramachaneni, *Shanmukh Katragadda*, *Bhupender Singh*, *Rohit Gupta*, *Radhakrishna Bettadapura* and the teams they led designed computer algorithms to distill the vast amounts of data generated by our molecular biology laboratories.

Arunabha Ghosh, *Anand Janakiraman*, *Nimisha Gupta*, *Sujaya Srinivasan* and the teams they led built the software interfaces that brought all relevant scientific literature to our fingertips, thus enabling our hunt for the genomic characters in these stories.

Shuba Krishna, *Smita Agarwal*, *Ashraf Mannan*, *Urvashi Bahadur*, *Rupali Gadkari*, *Jemima Jacob*, *Aparna Ganapathy* and the teams they led were the detectives who worked on these and several other cases, often scouring through reams of information to fish out one offending character among 6 billion.

Jyoti Bajpai, *Prasanna Shirol*, *Payal Manek* and *Khaleel Ahmed*, interfaced with several of the clinicians involved in these stories. *Prasanna Shirol*'s personal battle with one such genomic character and his untiring activism in spreading awareness and helping others is one of the inspirations behind this book.

Achintya Das and *Nilesh Tiwari* helped with digging deep into the impact of the offending characters in some of these stories.

Finally, *Kas Subramanian*, *Thiru Reddy* and *Vijay Chandru* provided overall cover for all of this, financially, as well as through their relationships with clinicians and hospitals.

Several people read and commented on early versions of these stories. *Ananya Ramgopal*, an 11th grade student, was my benchmark reader. Her comments on the first several stories helped calibrate the presentation for readability. *Sowmya Raghavan* was invariably the first to read each chapter. *Rajani Battu* and *Manoj Varma*'s detailed comments provided great encouragement. *Maya Malpani*, *Sandeep Tyagi*, *Rajesh Sundaresan*, *Swaroop Aradhya*, *Rebecca Manohar* and *Nitin Deshpande* all provided helpful comments. The presentation of these stories will surely become more accessible and enjoyable as all their feedback is incorporated in future editions.

Typesetting these stories in a presentable way required help from two notable quarters. First, the Legrand Orange LaTeX book style by Mathias Legrand[1] served as an excellent starting point for the internals of this book. And second, PSTricks, by Yuri Robbers and Annemarie Skjold,[2] served an a great starting point for designing the cover.

My gratitude to all the people listed above. In particular, special thanks to the patients and their families for allowing their stories to be represented here.

Much of this writing was done at home, and did take away from valuable family time. My gratitude to my family for bearing simultaneously with my physical presence and my mental absence. Hopefully, that is about to change!

Finally, my thanks to you, the reader, for taking this book in hand, even if for a quick browse and not an extensive read. Much needs to be done to quell the impact of these genomic characters on patients in general. Hopefully, these stories will inspire you to devote your energies to this cause.

Contents

Chapter 1

Three Colors or Two?

I had just qualified in my college entrance examinations after finishing high school. The final hurdle was a medical test; a mere formality, usually, for someone without an abnormal medical condition. The test progressed uneventfully until someone flashed a series of cards and asked me to call out the number written on each card, as in the example below.

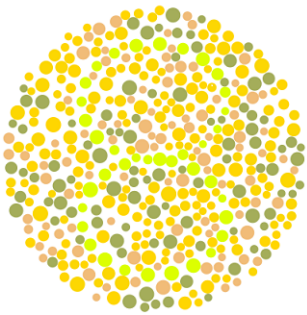


Figure 1.1: An example Ishihara card. Can you spot the hidden number? Or any number at all?

The candidate who preceded me breezed through his turn without batting an eyelid. My turn came. I saw dots of different colors and different sizes. But these dots did not a number form! Not to my eyes. I struggled to find a pattern, but in vain. I muttered out unsure guesses.

Partial color blindness said my report. There was a moment of concern, but allayed soon enough because this only disqualified me from working in mines, something I didn't aspire to do anyway. What remained, of course, was the mystery of this strange disability which had remained unnoticed for the the first 17 years of my life. What was it, and how did it come about?

Years later, I was giving a talk to an audience of about a 100 people. I projected one of these card pictures on the screen and asked anyone who had trouble spotting the number to raise their hand. 5 people raised their hands. All were males! It was good to know that I was not alone. But why were all 5 who raised their hands males?

Pinning it Down

The card shown above is an example of the Ishihara test,[3] named after its inventor Shinobu Ishihara, a University of Tokyo Professor. The card has dots of various colors (and various sizes, to confound matters further; but sizes are unimportant for our discussion). The hidden number or letter comprises dots of a particular color (call these *number dots*). All other dots (call these *noise dots*) have various other colors and are thrown randomly around the hidden number. The color contrast between noise dots and number dots is chosen with some care. For most people, this contrast is sufficient for number dots to stand out strongly, allowing the brain to connect these individual dots together into a number shape. For some people, however, this contrast is just not sufficient. All these people see is a mess of dots. They just cannot spot the number!

The following simple experiment helps concretize this phenomenon a little further. To simplify matters, let us work with only 2 colors. Number dots are set in Green. Noise dots also start as pure Green, at which point these are completely indistinguishable from the number dots. We then generate a series of Ishihara cards by mixing in increasing amounts of Red into the noise dots. At what point does the added Red tip the contrast over?

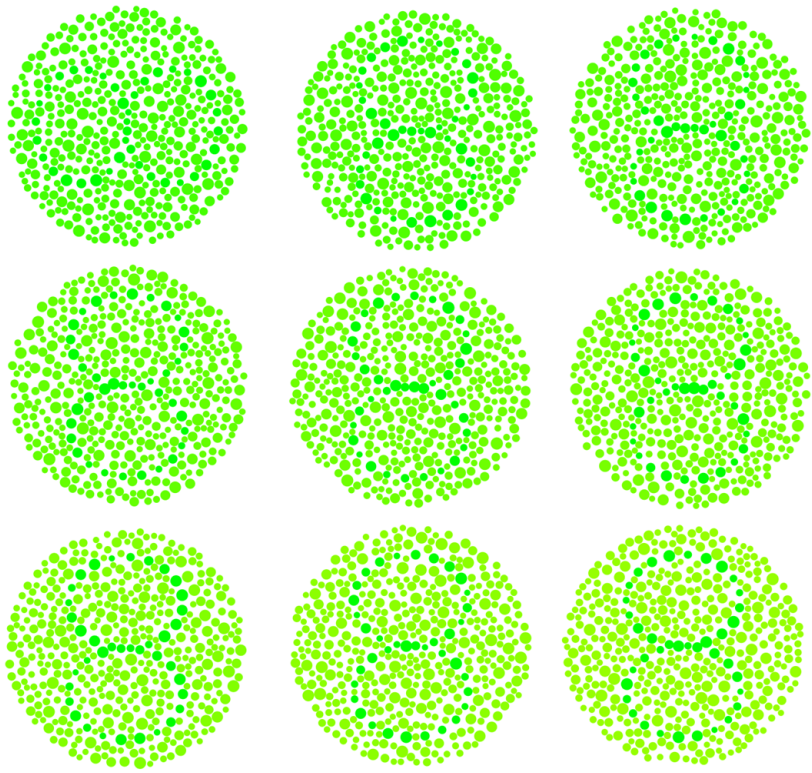


Figure 1.2: A Series of Ishihara Cards with the same hidden number, but with increasing amounts of Red mixed in (from 80 units at the top left to 160 units at the bottom right); at what point can you clearly spot the number?

This is easily tested on a computer. Colors on a computer screen can be generated by specifying a mix of Red, Green and Blue. For each color, the amount mixed is specified by a number between 0 and 255. Higher the number, more is the amount mixed in. Number dots are assigned pure Green (i.e., 255 units of Green). Noise dots are assigned pure Green,

initially. We then mix increasing amounts of Red into the number dots, in steps of 10 (i.e., we start with 0 units of Red, then 10 units, then 20 units, and so on). At each step, several Ishihara cards are generated by randomly choosing among several number and letter shapes, to guard against the brain getting used to a particular pattern. These cards are shown in quick succession to guard against the brain fishing out patterns laboriously. The viewer is asked to call out the hidden pattern in each case. Laptop screens are not suitable for this exercise because colors change quite dramatically when viewed from even slightly different angles. A tablet screen works well though.

I called on a few volunteers to participate in the above exercise. A summary of their performance is shown in Table 1.

Red Mix	**Volunteer Performance**	**My Performance**
80	< 30	< 10
90	$30 - 60$	< 10
100	$80 - 100$	< 10
110/120	100	< 10
130	100	40
140	100	90
150	100	$98 - 100$
160	100	100

Table 1.1: Performance: Mine and Others'.

When the amount of Red mixed was less than or equal to 80 units, most patterns were hard to identify for most volunteers. However, 110 units of added Red provided ample contrast for most people to spot the hidden patterns correctly. My eyes were far weaker though. At 110 units of added Red, I could barely spot any of the patterns. I could spot patterns

effortlessly only when 160 units of Red were added. So I needed 50 extra units of Red to perceive this clear contrast. Why?

Was this only an issue of confusion between Red and Green. Would Blue be any different? In the Blue version of the above experiment, we again start with pure Green number dots, and mix in noise dots with increasing amounts of Blue (instead of Red). What happens now? Normal volunteers were able to spot patterns when 110 units of Blue were mixed. My eyes reached 90% success at 110, and 100% at 120 units of Blue. Only a marginal difference.

In summary, my ability to distinguish between Blue and Green was comparable to most. But my ability to distinguish between Red and Green was significantly poorer than most. What was the cause?

Sensing Color

To start, recall a classical experiment by Isaac Newton. Newton used a prism to split a beam of sunlight into a spectrum of rainbow colors. One of these colors was Yellow. He screened out all the other colors, letting only Yellow through, and tried to split this Yellow beam again using a prism. The beam just wouldn't split! It stayed Yellow. He then tried another way of generating Yellow. He screened off all other colors, letting only Red and Green through this time. He then combined these Red and Green beams. The resulting beam appeared Yellow to the eye, no different from the original Yellow beam above. But when he attempted to split this Yellow beam using a prism, it gladly obliged and split into Red and Green. Thus, Newton showed that what the eye perceives as Yellow is not a single type of light. Many different combinations result in the same perceptual effect.

How does the eye actually detect these colors, and how does it perceive different combinations in the same way? We know now that each color of light has a distinct wavelength. The range of wavelengths we can perceive stretches from one end of a rainbow to the other. As you move from the Violet end of the rainbow to the Red end, the wavelength changes continuously from about 400nm to 650nm (nm stands for *nanometer*, 1 billionth of a meter). Does the eye have different sensors for each

wavelength? That would be just too many sensors. But then, if the eye has just a few sensors, say Red (650nm), Green (510nm), and Blue (475nm) sensors, it can certainly detect Red, Green and Blue; but how does it detect all the other wavelengths? If a combination of Red and Green light falls on the eye, both the Red and the Green sensors will detect their respective colors, and maybe the fact that both sensors detect something tricks the brain into seeing Yellow. But then how does the eye detect the pure unsplittable Yellow (570nm) which Newton generated?

The trick is as follows. The eye indeed has only 3 types of color sensors. However, each type of sensor is not sensitive to just one sharp wavelength. Rather, it responds to a range of wavelengths. The picture below can help visualize this better.

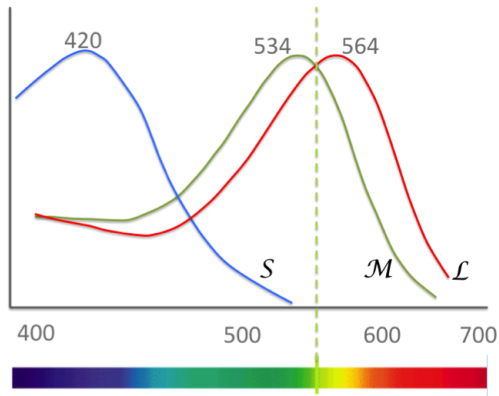


Figure 1.3: The Short ($\mathscr{S}$), Medium ($\mathscr{M}$), and Long ($\mathscr{L}$) wavelength sensor response intensities at various wavelengths.

For instance, the Green sensor might detect not just pure Green, it would also respond to Yellow, albeit more weakly that it would for Green. Likewise, the Red sensor would respond most strongly to Red but somewhat more weakly to Yellow. When the eye sees a Red-Green light combination, both sensors respond strongly, and the brain interprets this as Yellow. When the eye sees pure unsplittable Yellow, again both sensors

respond, albeit a bit more weakly; regardless, they both respond with sufficient intensity for the brain to still see Yellow. Since the three sensors now are sensitive to a range of wavelengths rather than a specific color wavelength, it is better to refer to them as *Short*, *Medium* and *Long* wavelength sensors, instead of Blue, Green and Red, respectively.

The Short wavelength sensor ($\mathscr{S}$) responds the most at 420nm (roughly Violet), the Medium wavelength sensor ($\mathscr{M}$) at 534nm (roughly Green), and the Long wavelength sensor ($\mathscr{L}$) at 564nm (roughly Orange).[4] However, these sensors do respond, though to lesser extents, even at other wavelengths. Curiously, the $\mathscr{M}$ and $\mathscr{L}$ sensors have their peaks really close together; the peak for the $\mathscr{S}$ sensor appears far away in comparison. This would mean that Green and Red is likely to be confused a lot more than Green and Blue. No wonder my eyes showed more confusion between Green and Red than between Green and Blue.

The fact that the $\mathscr{M}$ and the $\mathscr{L}$ sensors have peaks nearby is puzzling at first sight: how then do we distinguish between Green and Red? Both sensors will react at roughly the same intensity for both Green and Red wavelengths (and all wavelengths in between, including Yellow, Orange etc), providing little discrimination between these colors. It must be the case that the brain is very sensitive to the small difference in the responses between the $\mathscr{M}$ and $\mathscr{L}$ sensors. Wavelengths to the right of the dotted line in Fig. 1.3 elicit a greater response from the $\mathscr{L}$ sensor than the $\mathscr{M}$ sensor, i.e. $\mathscr{L} - \mathscr{M} > 0$, and are perceived broadly as Red. And wavelengths to the left of the dotted line in Fig. 1.3 elicit a greater response from the $\mathscr{M}$ sensor than the $\mathscr{L}$ sensor, i.e. $\mathscr{L} - \mathscr{M} < 0$, and are perceived broadly as Green.

Next, how do we distinguish between Orange and Red? Both wavelengths are to the right of the dotted line in Fig. 1.3, and the $\mathscr{L} - \mathscr{M}$ difference is almost the same. So this discrimination may not possible using just the $\mathscr{L} - \mathscr{M}$ difference. Possibly, the brain uses both $\mathscr{L} - \mathscr{M}$ and $\mathscr{L} + \mathscr{M}$ to provide this discrimination. Large values of $\mathscr{L} + \mathscr{M}$ appear as Orange, smaller values as Red.

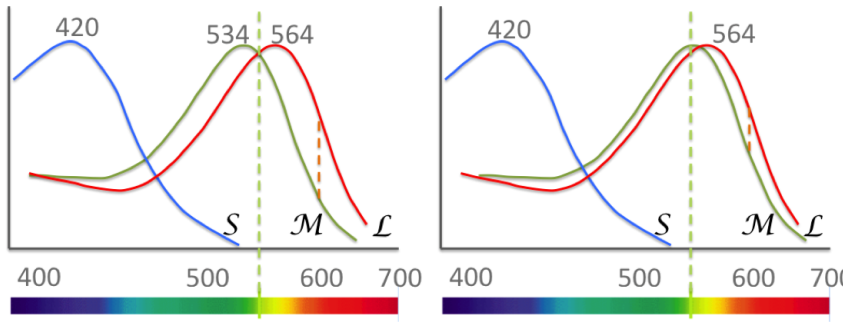


Figure 1.4: The $\mathscr{M}$ sensor response shifted slightly to higher wavelengths; $\mathscr{L}-\mathscr{M}$, as shown by the dotted line at 600nm, is reduced compared to normal.

What happens when either the $\mathscr{L}$ or the $\mathscr{M}$ sensors is slightly defective so the two corresponding peaks move even closer together than their usual gap of about 30nm (as in Fig. 1.4). The $\mathscr{L}-\mathscr{M}$ value becomes smaller than it usually is. This is indicated by the dotted line for the Red wavelength (600nm); this line is shorter in the defective case, i.e., the figure on the right, compared to the normal on the left. As a result, to elicit the same response with these defective sensors, the intensity of the 600nm Red light needed is much larger. At an extreme, if the $\mathscr{L}$ and $\mathscr{M}$ sensors peaks became identical, then $\mathscr{L}-\mathscr{M}$ will always be 0, i.e., all colors in the Green-Red range will be perceived broadly as Yellow.

Could this explain why my eyes needed 50 more units of Red compared to others to spot the correct answers in the Ishihara experiment above? Was the gap between my $\mathscr{L}$ and $\mathscr{M}$ sensors indeed smaller than normal? If so then what made this gap smaller?

Color Sensors in the Eye

Light sensors reside in the cells of the *Retina*, a screen at the back of the eye. Light from objects travels through the lens of our eye and is then focused on the retina. Sensors in the retinal cells detect this light. Nerves

then carry the detected signal to the brain.

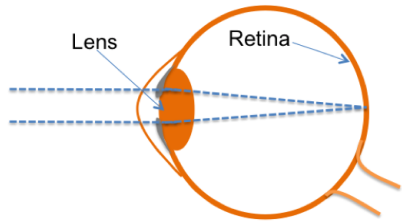


Figure 1.5: A Cartoon showing Light Focused on the Retina by the Lens

There are many different types of cells in the retina, and each type carries a different type of sensor. Much of the surface of the retina comprises cells called *Rods*; the sensors in these cells are light-sensitive but not color-sensitive. The center of the retina is populated with cells called *Cones* (Fig.1.6). These carry color sensors. Since cones appear more in the center of the retina than the periphery, we can see color nuances better where our eyes focus rather than at the periphery of our field of vision. And fewer cones means that our ability to see colors under low-light conditions (i.e., at night or in the dark) is reduced. But a large number of rods allow us to nevertheless see at night.

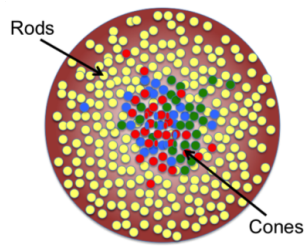


Figure 1.6: Retina showing Rods and Cones.

If you were to zoom into one of these rods or cones, you would see a

picture like this. Each such cell is enclosed by a sheet or membrane. Inside the cell are various *organelles*: these are to a cell, what organs like heart, brain, kidney etc are to the body. A key organelle is the nucleus.

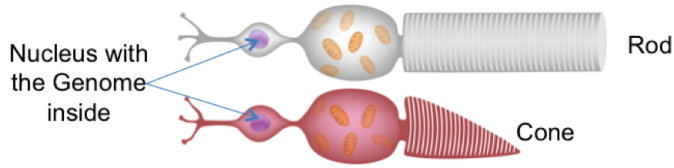


Figure 1.7: A Rod/Cone Cell Cartoon; Rods are cylindrical as shown, Cones taper off to the right.

Rod and cone cells have different behaviors because they manufacture and use different light-sensitive sensor molecules to detect light. Inside the nucleus of every cell is the *genome*. The recipes for the manufacture of these sensor molecules are written in this genome.

The Genome

The genome is a collection of books (to use a metaphor), an entire copy of which is carried by every cell in our body. Unlike English books which comprises the 26 characters of the English alphabet, these books comprise just 4 characters: A, C, G and T. There are about 6 billion such characters over all the books put together. Of course, the genome is tiny in terms of physical size, so it is not visible to the eye or even to standard microscopes. Here is how the pages of one of these books might look like, if only one could see them.

Each book in the genome is called a *chromosome*. There are 46 such chromosomes. 44 of these actually occur as 22 pairs, each pair carrying 2 copies of the same chromosome. These are shown in Fig.1.9, much like books on a bookshelf. One copy in each pair is inherited from the mother, and the other from the father. Chromosome 23 is called the *sex* chromosome, and comes in two flavors: X and Y. Females have two copies of the X chromosome, one copy coming from each parent. Males have one

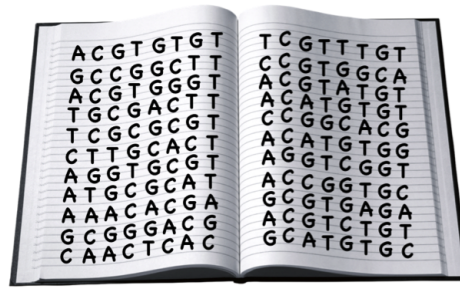


Figure 1.8: One of the Genome Books.

X chromosome and one Y chromosome; the X comes from the mother and Y comes from the father. The gender of a child is therefore determined by which sex chromosome the father passes along to the child: an X results in a girl, while a Y results in a boy.

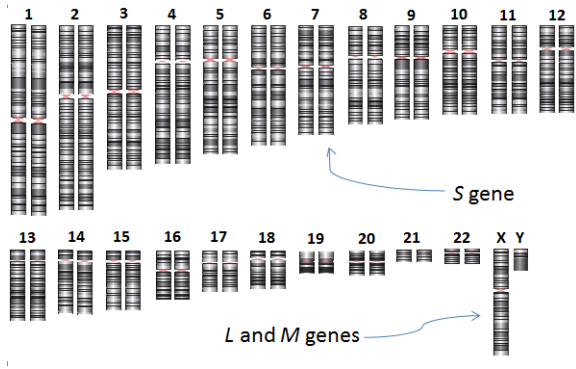


Figure 1.9: 23 pairs of chromosomes comprising the genome for a male; females have two X chromosomes instead of one X and one Y. We haven't yet met the genes labeled here but will do so soon.

And *Genes* are special stretches of text embedded within these chromosomes.

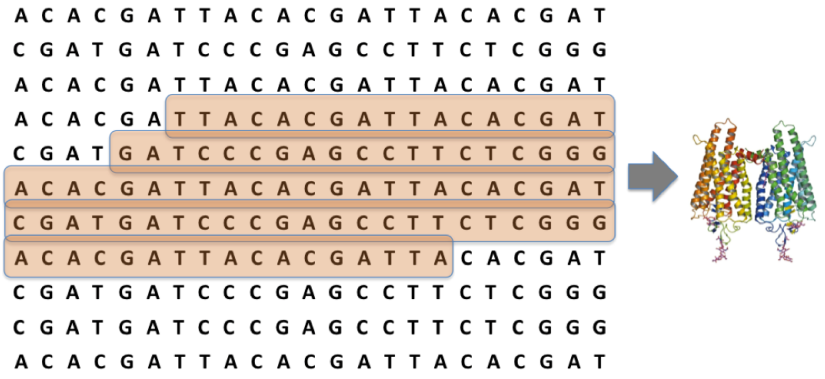


Figure 1.10: A Stretch of Text comprising a Gene, and the Protein created using this Gene's Recipe

What makes a gene special is that it carries the recipe for creation of a useful molecule, called a *protein*. In contrast, all the other text in our genome typically does not carry such recipes. There are about 21,000 genes in our genome. Each carries recipes for its own distinctive set of proteins. And each also carries a name. Indeed, we well meet several genes by name in these chapters. Which of these plays a role in sensing color?

Color Sensor Recipes

Even though every cell in our body contains the genome in its entirety, different types of cells choose to manufacture different amounts of proteins for the various genes. For instance, cells in the heart interpret recipes from certain genes to create proteins that sustain heartbeat, while cells in the blood interpret recipes from certain other genes to create proteins that help transport oxygen. The light-sensor molecules in the rod and cone cells in our retina are also proteins, and the recipes for the manufacturing of these proteins are described by some specific genes. Which genes are these?

The sensor protein in Rod cells is called *Rhodopsin* and the corresponding gene is called the *RHO* gene. Cones are not all the same though;

they come in 3 distinct types.

Recall, just a little earlier, we talked about 3 distinct types of color sensors in the eye. The Short wavelength sensor ($\mathscr{S}$) responds the most at 420nm (roughly Violet), the Medium wavelength sensor ($\mathscr{M}$) at 534nm (roughly Green), and the Long wavelength sensor ($\mathscr{L}$) at 564nm (roughly Orange) (Fig. 1.3). Indeed, there are 3 distinct types of cone cells, one type for each of the above sensors. The $\mathscr{S}$ cones manufacture the $\mathscr{S}$ sensor from the recipe described in a gene called *OPN1SW* (or just the $\mathscr{S}$ gene for short). The $\mathscr{M}$ cones manufacture the $\mathscr{M}$ sensor from the recipe described in a gene called *OPN1MW* (or just the $\mathscr{M}$ gene). And the $\mathscr{L}$ cones manufacture the $\mathscr{L}$ sensor from the recipe described in a gene called *OPN1LW* (or just the $\mathscr{L}$ gene).

The $\mathscr{S}$ gene is present on chromosome 7 in the genome. In contrast, the $\mathscr{L}$ and $\mathscr{M}$ genes are both located on the X chromosome. Since women have two copies of chromosome X while men have only one, this will make for some interesting gender variations in color perception, as we shall see later. For the moment, let us take a deeper look at how the recipes in the $\mathscr{S}$, $\mathscr{L}$ and $\mathscr{M}$ genes differ.

Exons and Introns

Even though a gene is a contiguous stretch of genomic text, the recipe for creation of the corresponding protein is not encoded all through the gene. The recipe is in fact written out in multiple contiguous sub-stretches of text within the gene, called *exons*, which are separated by intervening text sub-stretches called *introns*. Only exons carry the recipe; introns don't. Our cells know how to skip over the introns when reading and interpreting this recipe. Here is how we would represent this for the $\mathscr{S}$ gene.

Figure 1.11: Exons (cyan) and Introns (horizontal lines) in the $\mathscr{S}$ genes

The recipe encoded by the $\mathscr{S}$ gene is written out in 5 exons, with

the text in the 4 intervening introns being skipped when this recipe is interpreted. Here is a corresponding picture for the $\mathscr{L}$ and $\mathscr{M}$ genes.

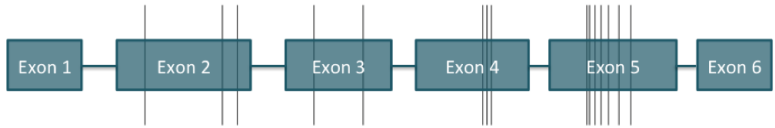


Figure 1.12: Exons and Introns in the $\mathscr{L}$ and $\mathscr{M}$ genes. The dark vertical lines indicate 14 places where the two genes differ. The differences in exon 5 account for more than 21nm of the 30nm difference between the two sensors.

There is just one picture above. But we have two genes: $\mathscr{L}$ and $\mathscr{M}$. It so turns out that these two genes are uncannily similar. And hence we are able to show both genes in the same picture. The text comprising the two genes is almost identical. Almost, that is; there are 14 characters at which the two genes differ, and their locations are marked in Fig. 1.12 by the black vertical line segments. The two resulting proteins (the $\mathscr{L}$ sensor and the $\mathscr{M}$ sensor, respectively) are therefore very similar as well. And that is why the corresponding curves in Fig. 1.3 are very close to each other.

The 14 places where $\mathscr{L}$ and $\mathscr{M}$ differ account for the 30nm difference between the maximum response wavelengths for the $\mathscr{L}$ and $\mathscr{M}$ sensors proteins (534nm for $\mathscr{M}$ vs 564nm for $\mathscr{L}$ in Fig. 1.3). Of these, *2 of the differences in exon 5 account for 21nm*,[5] and one of the differences in exon 3 accounts for another 5nm. The remainder of the gap is spread over the remaining 12 differences. In contrast, the $\mathscr{S}$ gene is very different, with many more differences from either the $\mathscr{L}$ or the $\mathscr{M}$ gene. Thus, confusion between the $\mathscr{L}$ and $\mathscr{M}$ sensors is more likely than between the $\mathscr{S}$ sensor and one of the other two sensors. Could some changes in my genome have reduced the gap between my $\mathscr{L}$ and $\mathscr{M}$ sensors?

My Recipes and Yours

The picture painted above of the $\mathscr{S}$, $\mathscr{L}$ and $\mathscr{M}$ genes is a canonical picture. In reality, there are variations from person to person. Some individuals may

have fewer than 14 differences between their $\mathscr{L}$ and $\mathscr{M}$ genes, and some may have more. Roughly speaking, one person's genome differs from another's at one in every 1000 characters. So of the 6 billion characters in our genome book, any two of us will differ in about 5-6 million. Could one of these differences be the cause of the reduced gap between my $\mathscr{L}$ and $\mathscr{M}$ sensors?

For instance, I know I have a difference at a particular character in exon 3 in my $\mathscr{M}$ gene; what is an A in most people here becomes a G for me. It is very close to a key character that causes a 5nm difference between $\mathscr{L}$ and $\mathscr{M}$ sensors. Could this be the change that causes my color perception problems?

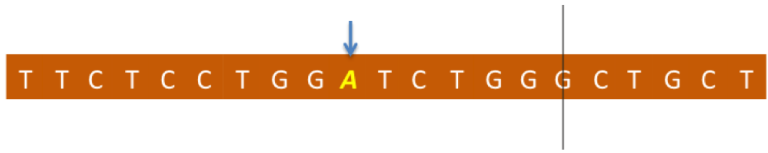


Figure 1.13: Zoomed-in view into Exon 3 of the $\mathscr{M}$ Gene. The yellow A in most people becomes a G in me. The dark vertical line is one of those shown in Fig. 1.12.

Unlikely, because this is a very *common variant*, i.e., more than 30% of the X chromosomes in all the people on earth appear to carry a G, while the remaining 70% carry an A. If this were indeed the character responsible for the phenomenon described in the introduction of this chapter, then many more than 5 in a 100 people would have failed to recognize the number in the Ishihara card I showed in my talk. There must be another, more severe, cause. Let's look further.

Inheritance and Recombination

We inherit our chromosomes from our parents. Remember most chromosomes come in pairs. We inherit one copy of each chromosome from each of our parents. Of course, the X and Y chromosomes are exceptions. So consider a male with a single X chromosome. This X chromosome comes

from his mother. The mother herself has two copies of the X chromosome. These copies need not be identical, they have small differences, as described above. Which of these two copies, if any, does she give her son?

The answer, is neither! What a mother actually gives to her son is a mosaic of the two chromosome copies. Alternate pieces from the two chromosome copies are stitched together into a new chromosome (Fig. 1.14). This process is called *Recombination*. The horizontal dotted lines show recombination breakpoints at which switches happen between the two copies.

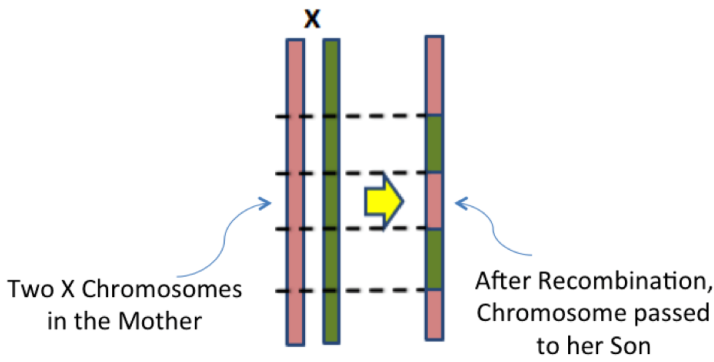


Figure 1.14: The Mother's two X chromosomes together create a Recombined Mosaic Chromosome.

Can recombination happen between two different chromosomes, say between chromosome X and chromosome 1? The answer is rarely, because recombination requires the two participating chromosomes be very similar to each other. Therefore, typically, recombination takes place between two copies of the same chromosome.

How might recombination affect the $\mathscr{L}$ and $\mathscr{M}$ genes? Both genes are on chromosome X. Further, they are actually next to each other on chromosome X. Even more curiously, there are actually two copies of the $\mathscr{M}$ gene right next to each other. Only the first of these copies, the one next to the $\mathscr{L}$ gene, is functional, i.e., the recipe given by this copy is

interpreted for construction of the $\mathscr{M}$ sensor. The other copy is not used for recipe interpretation. Now imagine what happens when a recombination breakpoint cuts through these genes as in the picture below (Fig. 1.15).

Which $\mathscr{L}$ and $\mathscr{M}$ genes are passed to her son by this mother in this scenario? The $\mathscr{L}$ gene passed to her son is derived by concatenating the first 3 exons from one copy with the last 3 exons from the other copy. The $\mathscr{M}$ genes are both derived from this latter copy. This, by itself, is unremarkable.

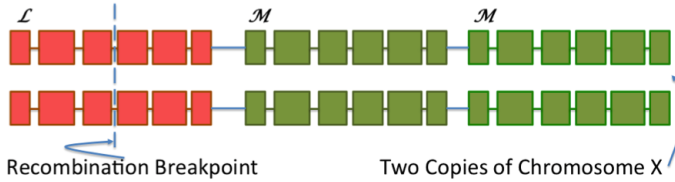


Figure 1.15: Two Copies of Chromosome X in the Mother, zoomed into the region with the $\mathscr{L}$ and $\mathscr{M}$ Genes

A more interesting scenario arises when the breakpoint does not cut through the same location in the two chromosomal copies, as in Fig. 1.16. Here, the breakpoint cuts between exons 3 and 4 in the $\mathscr{L}$ gene in one copy of chromosome X, and between exons 3 and 4 of the $\mathscr{M}$ gene in the other copy.

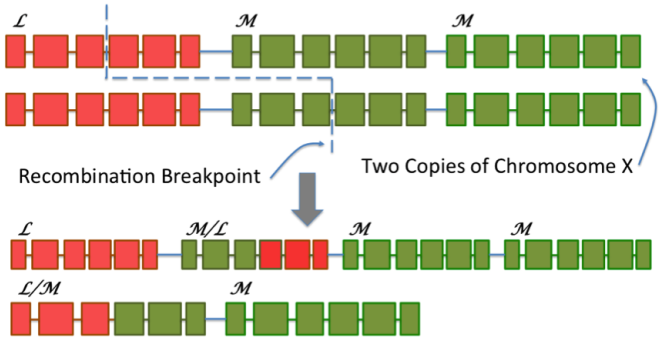


Figure 1.16: An Recombination Scenario resulting in Hybrid Genes and Extra/Missing Genes

In this case, the two resulting chromosome copies after recombination are shown above as well. One chromosome has an $\mathscr{L}$ gene followed by a hybrid $\mathscr{M}/\mathscr{L}$ gene followed by two $\mathscr{M}$ genes; so 4 genes in all. The second has a hybrid $\mathscr{L}/\mathscr{M}$ gene followed by a single $\mathscr{M}$ gene; so just 2 genes in all. So we now have two rather curious combinations of the $\mathscr{L}$ and $\mathscr{M}$ genes; one of these is passed on by the mother to her son.

Wait a minute! Can a recombination breakpoint really cut across the two chromosomal copies at two different locations? Would that not wreak havoc, with too many or too few genes in the resulting chromosome that is passed to the child? Indeed, it turns out that a breakpoint line will typically cut through the same location in both chromosome copies, as in Fig. 1.15. This is because the sequence of characters in the two chromosome copies around the recombination breakpoints have to be similar for recombination to work; and this usually happens only if the breakpoint line cuts through the same location in both copies. However, remember that the $\mathscr{L}$ and $\mathscr{M}$ genes are very similar, with only 14 differences. So in this special case, this condition is satisfied even when the recombination breakpoint line appears as in Fig. 1.16.[4] In a way, the cellular system is tricked into believing that it is recombining as per Fig. 1.15, but it is actually recombining as per Fig. 1.16!

The mother now passes one of these two curious chromosome X copies to her son. Her son might in turn pass this copy on to his daughter, and so on. Over time, this curious new chromosome X may find its way into several individuals. What impact might this have?

Recall only the $\mathscr{L}$ gene and the first $\mathscr{M}$ gene are used for recipe interpretation. In one case, the $\mathscr{L}$ gene is intact but the $\mathscr{M}$ gene is replaced by a $\mathscr{M}/\mathscr{L}$ hybrid. In the other case, the $\mathscr{L}$ gene is replaced by a $\mathscr{L}/\mathscr{M}$ hybrid but the $\mathscr{M}$ gene is intact. In either case, note that exon 5 is common to both the hybrid gene and the intact gene. Remember that differences in exon 5 between the $\mathscr{M}$ and $\mathscr{L}$ genes account for 21nm of the 30nm difference in wavelength sensitivity between the $\mathscr{M}$ and $\mathscr{L}$ sensors[5] (see Fig. 1.12). It follows that the wavelength sensitivity gap between the hybrid and intact gene is now not more than 30-21=9nm. This could lead to the situation in Fig. 1.4, possibly explaining my color sensitivity woes. Did I indeed have one of these curious chromosomes with a hybrid gene on it?

A Peek at the Recipes

The only way to identify whether I have pure $\mathscr{L}$ and $\mathscr{M}$ genes or a hybrid gene is to peek into my genome and see what it holds. The genome of course is quite tiny, sitting deep inside the nucleus in every cell in my body. The naked eye can't see it, neither can a microscope. Cleverer methods are needed.

There are tiny naturally occurring molecules whose daily job is to read the genome, interpret the recipes coded in various genes, make proteins from these recipes, and make entire copies of the genome as cells divide. Scientists have figured out how these very tiny molecules can be manipulated in clever ways so they actually provide a read-out of the genome on to a computer disk. This is called *sequencing* of the genome. And it works as follows. We start with a little saliva collected in the test tube (or blood, or any other body tissue, for that matter). This saliva contains some cells from which my genome is extracted via a chemical process and subject to sequencing using a process called *Next Generation Sequencing* or *NGS*.

Unfortunately, the NGS process today is not powerful enough to read

each of my chromosomes end to end. Instead, these chromosomes are chopped into small pieces (or *reads*) of about a 100 characters each, often randomly. Tiny genome copying molecules are then let loose on these reads. These molecules start making copies of these reads, one character at a time. With each character they copy, they are tricked into emitting a visual cue; the color of this cue depends upon the character copied. These color cues are captured by a camera and then processed by a computer to extract the character sequence for each of the reads. The contents of each read are now available, but the order of reads in the chromosome has been completely lost!

How do we proceed next? Tricks and techniques to recover the order of reads do exist but are quite onerous and expensive. Certainly not suitable for sequencing of individuals on a personal basis to aid in diagnosis of health conditions. So we've managed to tear our genomic book into shreds of paper and know what is written on each piece of paper, but we still need to put the pieces together.

Putting Reads Together

To our rescue now comes the landmark Human Genome project. Several groups of scientists launched a monumental effort on putting these pieces together around the turn of the millennium. They used huge amounts of computing power as well as follow-up experiments to accomplish this task, culminating in the publication of a (close to) finished human genome sequence in 2004.[6] This was not the genome of a single individual; rather, it was the pooled genome of a few unnamed, and presumably healthy, individuals. We call this genome, the *reference genome*. Of what use is this reference sequence is our quest?

It so turns out that the genomes of any two individuals are very similar to each other; they differ in only one in approximately a thousand characters. This means that the reference genome and my genome are very similar. Which, in turn, means that if we get a read sequence, e.g., AGGTTCTG, from my genome, then this sequence will be present somewhere in the reference genome as well, either in identical or in very slightly altered form. We do need to allow for slight alterations though because

the reference genome is not identical to my genome. Fig. 1.17 shows an example. A computer can be used to quickly identify where each read from my genome appears in the reference genome, in identical of slightly altered form. This procedure is called *read alignment*. Once all the reads are anchored to their respective places in the reference genome, we effectively have the order of reads with us. This procedure is fully automated and can be done quite quickly on powerful computers.

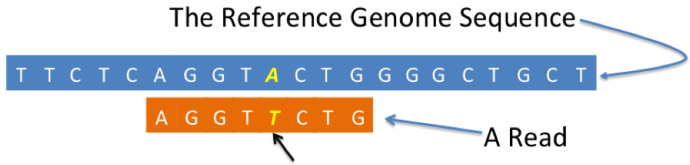


Figure 1.17: A Read Anchored to the right place in the Reference Genome

Once each read from my genome is anchored to the right place in the reference genome, differences between my genome and the reference genome become apparent right away. For instance, see highlight in yellow in Fig. 1.17: my read shows a T while the reference genome shows an A. Does that mean that I differ from the reference genome at this location? Yes, but only if we can handle measurement error.

Most measurement methods carry small amounts of error; the NGS process above is no exception. Typically one in a few hundred characters might be read wrong. Since this error rate is quite low and since the read alignment process does allow for slight alterations to the read, an erroneous read will still continue to be anchored to its rightful place in the reference genome. So that is not a problem. The problem, as illustrated in Fig. 1.17, arises when the read sequence and the reference sequence disagree in some characters. This may happen because my genome is truly different from the reference genome, or it may happen because there was an error in generating the read. It is important to identify which of these is the case. A true difference between the reference genome and my genome might be medically relevant. An error on the other hand would not have any medical implications. How do we differentiate between these two cases?

The trick lies in sequencing not one copy of my genome but many copies of my genome simultaneously. Each of my cells carries the genome; we take several cells together, chop up the genomes in all of these cells into reads, obtain the sequence of each read, and finally anchor each read to the where it might have come from in the reference sequence using the read alignment process. Now, as shown in Fig. 1.18, many different reads overlap any given character in the reference genome.

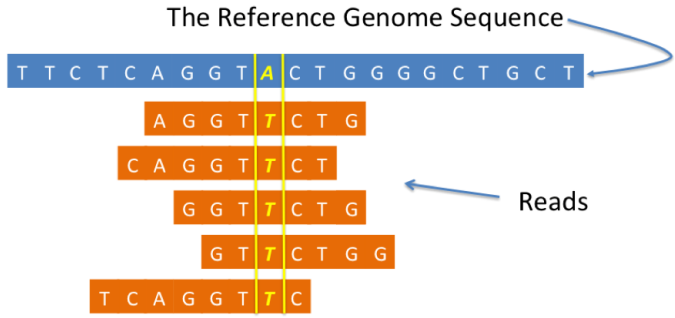


Figure 1.18: A Pile-up of Reads anchored to their respective right places in the Reference Genome

If several reads overlap a character, as in the picture above, and most indicate a T while the reference has an A, then my genome must have a T at this location. For, errors are unlikely to occur in more than a few of these reads simultaneously. What if roughly half the reads indicate a T while the rest indicate an A? This means that one of my chromosome copies (recall, most chromosomes come in pairs) has a T and the other has an A. And if very few reads indicate a T and most indicate an A, then both of my chromosome copies have an A and the T's are probably on account of error. So just by analyzing read pile-ups at each character in the reference, differences between my genome and the reference genome can be readily derived. Further, for each difference, we can identify whether this difference arises in one or both of my chromosome copies.

So that's how the NGS procedure works. Sequencing the entire genome with this procedure costs about $5000 at the time of this writing, and is

called *Whole Genome Sequencing*. This cost can be reduced substantially by chemically pulling out just the exons of the various genes from the genome, and sequencing just these. Recall from Fig. 1.12 that these exons carry recipes for manufacturing proteins and hence are the most important of the genomic regions. Sequencing just the exons of all the genes is called *Whole Exome Sequencing*; this costs about $800 at the time of this writing, under a fifth of Whole Genome Sequencing.

What did whole exome sequencing on my genome tell me about my $\mathscr{L}$ and $\mathscr{M}$ genes? Did I indeed have an $\mathscr{M}/\mathscr{L}$ hybrid gene?

Unraveling the Mystery

Take a look at Fig. 1.16 which shows the various possibilities of $\mathscr{L}$ and $\mathscr{M}$ genes that the X chromosome in my genome might carry. Which of these is actually true for my X chromosome? Let us try to decipher this from the reads that whole exome sequencing gives us. As an example, take exon 5 of the $\mathscr{L}$ and $\mathscr{M}$ genes.

If my X chromosome has the usual configuration (shown at the top of Fig. 1.16), then I should have one exon 5 copy from the $\mathscr{L}$ gene and two exon 5 copies from the $\mathscr{M}$ gene. Since exon 5 has the exact same length in both genes, the process of taking several copies of my genome and chopping it into pieces should yield twice as many reads from exon 5 of $\mathscr{M}$ than exon 5 of $\mathscr{L}$. The read alignment process will then anchor twice as many reads to exon 5 of $\mathscr{M}$ than to exon 5 of $\mathscr{L}$ in the reference genome (of course, we include reads anchored to exon 5 of both $\mathscr{M}$ copies in this count). In other words, 33.33% of reads anchoring to exon 5 will anchor to exon 5 of $\mathscr{L}$ and the remaining 66.67% to exon 5 of $\mathscr{M}$.

What if my X chromosome has an unusual configuration, shown at the bottom of Fig. 1.16? In the first of these scenarios, there are 2 copies of exon 5 from $\mathscr{L}$ and 2 copies of exon 5 from $\mathscr{M}$; by analogous calculations, 50% of reads anchoring to exon 5 should anchor to exon 5 of $\mathscr{L}$ and the remaining 50% to exon 5 of $\mathscr{M}$ in the reference genome. In the second of these scenarios, there are 0 copies of exon 5 from $\mathscr{L}$ and 2 copies of exon 5 from $\mathscr{M}$; so 100% of reads anchoring to exon 5 should anchor to exon 5 of $\mathscr{M}$ in this case.

Now that we know what to expect in each of the above cases, let us do the actual count of reads from my genome that anchor to exon 5 of the $\mathscr{L}$ gene and compare this with the count of reads that anchor to exon 5 of the $\mathscr{M}$ gene. The numbers are: 53% $\mathscr{L}$, and 47% $\mathscr{M}$.

We can do this exercise for the other exons as well. Because exon 1 and exon 6 are identical in both $\mathscr{L}$ and $\mathscr{M}$ (see Fig. 1.12), we have no way of getting comparative counts for these. But counts for all the other exons based on reads from my genome are given in the picture below.

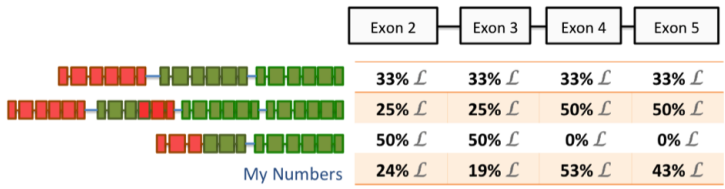


Figure 1.19: Read percentages for the $\mathscr{L}$ gene in the typical case, in the abnormal recombination cases of Fig. 1.16, and in my whole exome sequencing data.

And what do these counts indicate? Of all the configurations shown in the picture, my numbers closely match those for the second row of the table, which corresponds to a configuration comprising an $\mathscr{L}$ gene, a hybrid $\mathscr{M}/\mathscr{L}$ gene, and two non-functional $\mathscr{M}$'s. Not an exact match, but that is to be expected (after all, if you toss a fair coin 10 times, you don't get exactly 5 heads). Voila! This suggests that my functional $\mathscr{M}$ gene has been replaced with an hybrid $\mathscr{M}/\mathscr{L}$ gene. And so my $\mathscr{M}$ sensor behaves much more like my $\mathscr{L}$ sensor than it normally should, the usual 30nm gap between the two reducing to below 9nm. And this leads to markedly reduced ability in distinguishing between Red and Green. My $\mathscr{S}$ sensor is normal though, so my ability to distinguish between Blue and Green remains normal.

Wrapping Up

27 years elapsed between when I first discovered this issue to when I could finally diagnose and understand its cause. I now proudly wear the *Deuteranomaly* label, which means my $\mathscr{M}$ sensor behaves much more like my $\mathscr{L}$ sensor than it normally should.

Things could have been worse had I inherited an X chromosome obtained via even more abnormal recombination, as shown in Fig. 1.20. My functional $\mathscr{M}$ gene would have been completely replaced by an $\mathscr{L}$ gene, or vice versa, making the $\mathscr{M}$ and $\mathscr{L}$ sensors identical (recall, exon 1 is identical in the two genes, so that difference doesn't matter). I would then have only two sensors instead of three and the ability to differentiate between Red and Green would be completely gone!

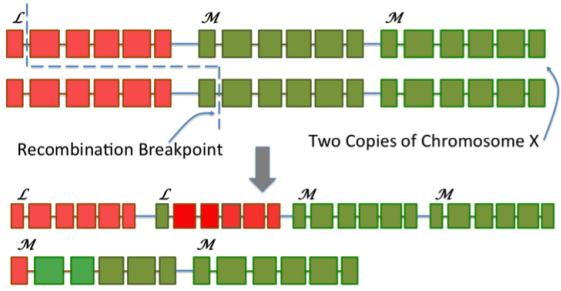


Figure 1.20: Another interesting Recombination Scenario leading to just Two Sensors instead of Three.

The famous scientist John Dalton noticed his own inability to distinguish between colors in the Red-Green spectrum almost two centuries ago. It was only in 1995 that sequencing of his $\mathscr{L}$ and $\mathscr{M}$ genes from preserved eye remains[7] confirmed that he lacked the $\mathscr{M}$ gene and had only the $\mathscr{L}$ gene. This was a much more rigorous assessment done by sequencing only the relevant $\mathscr{M}$ and $\mathscr{L}$ genes in a much more detailed way than what I have presented here. However, the elegance of whole exome sequencing is that it gives us information about almost all our genes at modest cost in one shot. This information can be drawn upon as and when needed to

diagnose varied health phenomena.

Most deuteranomalous individuals are male. Why? The $\mathscr{L}$ and $\mathscr{M}$ genes appear on the X chromosome and females have two copies of this chromosome, one inherited from each parent. Copies of the $\mathscr{L}$ gene or the functioning $\mathscr{M}$ gene on both the chromosomes will need to be mutated for Deuteranomaly to arise in a female. Males, by virtue of having just a single X chromosome, are much more susceptible. The similarity between the $\mathscr{L}$ and $\mathscr{M}$ genes, their nearness on the X chromosome, and the relative distinctiveness of the $\mathscr{S}$ gene on Chromosome 7, makes confusion between $\mathscr{L}$ and $\mathscr{M}$ much more likely than confusion between $\mathscr{S}$ and $\mathscr{L}$, or $\mathscr{S}$ and $\mathscr{M}$. Therefore, confusion between Red and Green is much more prevalent than confusion between Blue and Green.

My children can both read Ishihara cards effortlessly. When I struggle on these cards, they think I am playing dumb. They just can't believe that is the way I am made. If I were a child, I can well imagine parents and teachers pushing me to practice harder and harder on these cards; little would they know that no amount of practice helps. Genomic characters are like that sometimes: discreet, yet overwhelming.

One could argue that the $\mathscr{M}/\mathscr{L}$ hybrid gene is fairly innocuous. But that is not the case with several other changes in the genome. We will see examples later of genetic mutations that are far more serious: mutations that could debilitate or even kill. And examples of how mankind can navigate around or confront these mutations.

Chapter 2

The Picture Gets Blurred

$\mathscr{X}$ had normal vision late into her 30s. Then the picture began to blur. When reading, characters began to lose clarity. When in conversation with someone, the other person's face started to blur. This blurriness worsened over the years. A blank patch in the center of the vision field became a permanent fixture by the time she was in her 40s. Faces became hard to recognize. Reading was barely manageable with the aid of a magnifying glass. Navigation around the house as well as outdoors was still possible, though more labored. Trips to the doctor were of no avail towards diagnosing this phenomenon, let alone curing it. I am imagining this is how a face would appear to her eyes (Fig. 2.1).

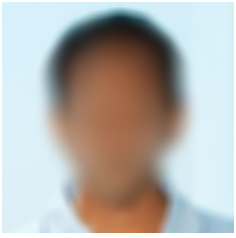


Figure 2.1: Blurring of central vision.

$\mathscr{X}$ was not alone in her ordeal. She had 7 siblings: 4 brothers and 3 sisters; her extended family is shown in Fig. 2.2. One brother ($\mathscr{B}$) and one sister ($\mathscr{S}$) shared this affliction. The other siblings remained unaffected. $\mathscr{B}$ had an active career in the police somewhat stunted by this impediment.

Several attempts at seeking mainstream and alternative medical help were made, all in vain. Now in his 70s, he could not read at all or watch TV, could barely recognize faces, and could walk on the road only with great difficulty. Curiously, $\mathscr{X}$'s mother ($\mathscr{M}$) had lived a long life with perfect vision. $\mathscr{X}$'s father ($\mathscr{F}$) had passed away in his early 30s; whether he would have suffered loss of vision if he had lived longer will never be known.

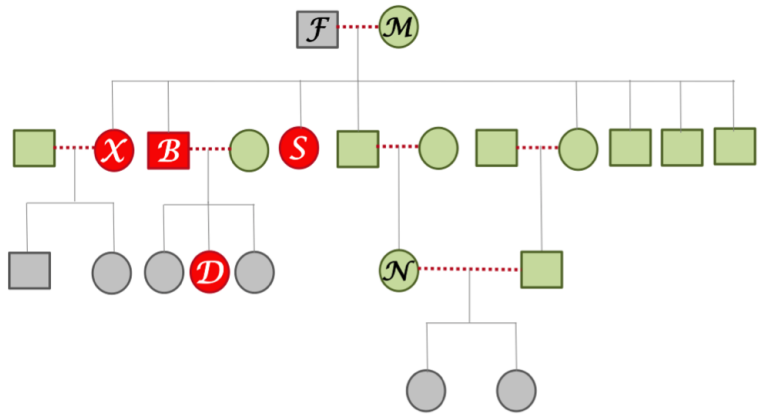


Figure 2.2: $\mathscr{X}$'s Extended Family. Circles are females, squares are males. Dotted lines connect spouses. Children of each couple are drawn below that couple. Red indicates individuals with vision loss. Green denotes healthy individuals. Gray indicates individuals for whom risk of vision loss is yet unknown.

$\mathscr{X}$, $\mathscr{B}$ and $\mathscr{S}$ had all learned to live fruitful lives in spite of this diminishing faculty. The frustration of several unanswered questions remained though. Why was this disorder so common in the family? Whom did it target and whom did it choose to spare? And why did it set in only in the third or fourth decade of life? Would it rear its ugly head again in the next generation? $\mathscr{X}$ and her siblings had several children between them. $\mathscr{B}$'s 3 daughters ranged in age from early to the late 40s. One daughter $\mathscr{D}$ had started showing signs of central vision loss; the other two daughters

were yet unaffected. $\mathscr{X}$'s children were a daughter in the early 40s and a son in the late 30s, both currently unaffected. But would they remain so, or was it just a matter of time? Was there hope of a cure around the horizon?

Inheritance

What is it that went around in this family, taking vision away from some, and leaving the rest untouched? Could an offending character in the *genome* be the cause? The genome is a book with three billion characters of text that drives every cell in our body. No two individuals have the same genome, unless they are identical twins; there are always small differences. Affected individuals might have a particular offending character (call it $\star$) which is absent from the others. This offending character could be inherited by children from their parents. And every individual carrying this character would then show visual deterioration. Let us examine this inheritance more closely.

Will a parent who has the $\star$ character in their genome pass this character on to each of their children? From Chapter 1, we know that each of us has two complete copies of the genome. One copy is inherited from our mother and one from our father. These two copies are very similar to each other, but they are not identical. Differences occur at 1 in a 1000 characters, roughly speaking. The offending character $\star$ may be present in only one of these copies, or it may be present in both copies. From Chapter 1 again, we know that a parent does not pass on one of their two copies to the child directly. Rather, the parent creates a mosaic of the two copies by taking alternating sections from each copy. This mosaic is what the child inherits. One mosaic from the mother and one from the father. If both copies in the parent have $\star$, then the mosaic that a child inherits from this parent will necessarily have $\star$. On the other hand, if only one of the copies in the parent has $\star$, then the mosaic that a child inherits from this parent will have $\star$ with only a 50% chance (Fig. 2.3).

We do not know if $\mathscr{X}$'s father (call him $\mathscr{F}$) would have experienced vision loss had he lived longer. But suppose for a moment that he would. And suppose for a moment that the source of his vision loss was a single offending character $\star$ in one copy of $\mathscr{F}$'s genome. This is often called

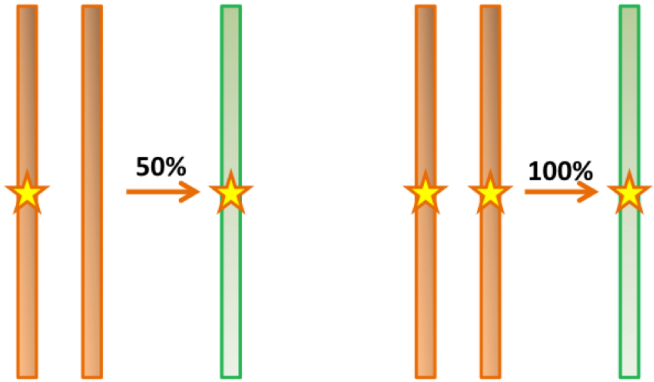


Figure 2.3: Left: The Dominant Scenario. A defective character in one copy of the genome suffices to cause disease in a parent; there is a 50% chance that a child inherits this character. Right: The Recessive Scenario. Defective characters are needed in both copies of the genome to cause disease in a parent; a child surely inherits one of these defective characters.

the *Dominant* scenario (Fig. 2.3), where one faulty copy of the genome suffices to cause disease. Since each child has a 50% chance of inheriting $\star$, roughly half his children would be expected to have the problem (roughly but not necessarily exactly half, because tossing a coin 10 times gives roughly but not necessarily exactly 5 heads). This is broadly consistent with what actually happened: 3 out of $\mathscr{F}$'s 8 children were affected. How about $\mathscr{F}$'s grandchildren from these affected children? Each of these grandchildren would have a 50% chance of inheriting $\star$. So one would expect roughly half to show symptoms by the time they reach the 40s. $\mathscr{S}$, $\mathscr{X}$'s affected sister, had 3 children, all well into their 50s, and none showed any signs of visual deterioration. $\mathscr{B}$,$\mathscr{X}$'s affected brother, had 3 children, all in their 40s; only one of them showed signs of being affected. $\mathscr{X}$ herself had two children, one in the early 40s and one in the late 30s. Neither showed any signs of being affected yet. For some of these younger grandchildren, it was probably too early to say. Nevertheless, the number of affected grandchildren appears to be fewer than expected. So maybe

our assumption of the dominant scenario, i.e., a single offending character in one genome copy responsible for the problem, is not quite correct. Is there another scenario?

Sometimes, offending characters may be needed in both genomic copies. It could be the same character in both genomic copies. Or it could be two distinct characters in the two copies. This is often called the *Recessive* scenario (Fig. 2.3). Consider $\mathscr{X}$. In this new scenario, each of her genomic copies has an offending character. She inherited one of these copies from her mother $\mathscr{M}$ and one from her father $\mathscr{F}$. So both $\mathscr{M}$ and $\mathscr{F}$ have offending characters in at least one of their respective genomic copies. $\mathscr{M}$ remained unaffected through her 80s, so it is likely that only one and not both of her genomic copies carried the offending character. The other copy was *normal*. In other words, she was a *carrier* of the offending character, without being affected by the offending character herself. Less is known of $\mathscr{F}$ since he died early, but it is quite possible that he too was a carrier (for a different offending character, possibly) without being affected directly. A child of $\mathscr{M}$ and $\mathscr{F}$ would be affected if they inherited both the offending character from $\mathscr{M}$ (a 50% chance) as well as the offending character from $\mathscr{F}$ (also a 50% chance). The odds of both events happening together is 25%. So roughly 2 out of the 8 children of $\mathscr{M}$ and $\mathscr{F}$ ought to be affected. This is not far from the actual number: 3 out of 8 were affected.

Continuing with the recessive scenario, how about $\mathscr{F}$'s grandchildren from the affected children $\mathscr{X}$, $\mathscr{B}$ and $\mathscr{S}$? For instance, $\mathscr{B}$'s children would stand a 25% chance if their mother (i.e., $\mathscr{B}$'s spouse) was a carrier as well. If not, then this would drop down to 0%. Given that one of B's children was indeed affected, their mother must have been a carrier. And given the mother is a carrier, roughly 25% of the children would be affected. The actual figure was 1 in 3, not wide off the mark. None of $\mathscr{S}$'s children were affected, even though they were all into their 50s. So maybe $\mathscr{S}$'s spouse just wasn't a carrier. $\mathscr{X}$'s children were in their late 30s and early 40s and yet unaffected. Maybe too early to call, or maybe $\mathscr{X}$'s spouse was just not a carrier either. It is hard to say which of the two is true, unless we investigate further. So begins our hunt for the offending character or characters, keeping the above possibilities in mind.

Where Do We Look?

Our window into $\mathscr{X}$'s genome is *Whole Exome Sequencing*. Remember, we introduced this in Chapter 1. But here is a quick summary again.

There are about 21,000 *genes* in our genome. Each gene is a relatively small stretch of text embedded within the three billion long genome sequence. The text for a particular gene describes a recipe for the creation of an associated molecule called a *protein*. This protein molecule then plays a specific role in keeping our cells and our bodies healthy. Some genes actually carry recipes for multiple proteins. It is likely that the offending character in $\mathscr{X}$ and others in her family appears in one of these 21,000 genes. There is a vast ocean of text in the genome outside of these genes as well. This text itself does not contain any recipes. But it may control the rate at which recipes for various genes are interpreted by cells to generate the corresponding proteins.

The recipe for a particular gene is not coded in a single contiguous stretch of text in the genome. Rather, it is written out in multiple contiguous stretches of text, called *exons*, separated by intervening text stretches called *introns*. Only exons carry the recipe; introns don't. There are roughly 200,000 exons, considering exons from all 21,000 genes. The genomic text stretches in these 200,000 exons together comprise the *exome*, just over 1% of the entire genome. Whole Exome Sequencing followed by a lot of data crunching on a computer gets us $\mathscr{X}$'s exome. In particular, it identifies every character in the exome where $\mathscr{X}$ differs from a supposedly healthy genome sequence. These characters are called *variants*. The only problem: there are tens of thousands of these variants! How do we identify which of these is the culprit?

We start by narrowing down genes of importance. Research over the last 30 years as captured in hundreds of thousands of publications tells us that not all 21,000 genes are equally important. Only around 4,000 are known to cause severe disease (specifically, diseases of the form that are not related to old age). This list will of course grow with time. For now, these genes serve as a good starting point for our hunt. We can restrict the range for our hunt even further, by considering only those genes that cause malfunction in the eye. Or even further, to genes which cause malfunction in the *retina*, the screen at the back of the eye where the eye lens projects

the image which nerves carry to the brain. A computational and manual scan of published literature yields this list of genes. We are down to about 100 genes now from our starting list of 21,000 genes!

The shortlist of genes we have spans a number of different retinal malfunctions or *dystrophies*, as diseases one is born with are called. Some of these like *Leber Congenital Amaurosis* cause severe visual impairment beginning in early infancy, but are relatively rare, at 1 in a 100,000 births. Other dystrophies which cause progressive loss of vision over several years are more common. The most common among these is a disease called *Retinitis Pigmentosa*, with an incidence of 1 in 4000.

Retinitis Pigmentosa occurs because of loss of *sensor* cells in the retina, cells which sense light. There are two types of sensor cells: *rods* and *cones*. Cones are more abundant in the center of the retina, in a region called the *macula*. Rods are more abundant outside the macula. Cones sense color but function only when the light intensity is substantial. Rods do not sense color, but are very sensitive and enable night vision. In Retinitis Pigmentosa, rods are lost first, leading to tunnel vision of the form seen in Fig. 2.4, and cone loss follows later. More than 70 genes are known, defects in which cause Retinitis Pigmentosa.

Figure 2.4: Tunnel Vision in Retinitis Pigmentosa.

Recall $\mathscr{X}'s$ vision was blurred more at the center, i.e., in the macula. There are other dystrophies where cones are lost first, or cones and rods are lost simultaneously. In such cases, central vision is often affected first, and peripheral vision later. *Stargardt* disease is one such, where degeneration of cells in the macula often starts in adolescence. *Best Vitelliform Macular Dystrophy* is yet another, with onset in childhood. There is also a type of

macular degeneration which sets in with age (called *Age-related Macular Degeneration* or *AMD*), typically in the 60's. In contrast to all these, $\mathscr{X}'s$ vision loss started, at least perceptibly so, only in the 30s. This does not fit directly into any of the above patterns. So we need to identify the variant(s) in our haystack of a 100 genes to diagnose $\mathscr{X}$'s condition.

The Genetic Code

We have a 100 genes to look through, each with potentially many variants, quite a large haystack! Not all these variants have equal impact on their respective genes though. Variants which disrupt a gene's recipe are more likely to be problematic. This recipe is written out in a code, called the *genetic code*.

First, imagine stringing the exons of a gene together, excising out the intervening introns. Take the resulting sequence of characters and divide these into blocks of 3 characters each. Each character triple specifies what is called an *amino acid*. So a gene whose recipe comprises 300 characters codes for a sequence of 100 amino acids. These amino acids are strung together to form the corresponding protein. The entire genome sequence has only 4 distinct characters, A, C, G and T (much like English has 26 distinct characters). On the other hand, there are 20 distinct amino acids. Which character triples code for which amino acids is shown in Fig. 2.5.

For instance, the triplet *CTT* codes for the amino acid *Leucine* (short form *L*). And triplet *CCA* codes for amino acid *Proline* (short form *P*). So does *CCT*, i.e., more than one triplet could code for the same amino acid. Also notice that some triplets, specifically *TGA*, *TAA* and *TAG* code for *Stop* (short form *X*); this is not an amino acid, rather an instruction to terminate the recipe execution process, declaring that the protein has now been fully created. The process of generating the protein from the recipe coded by the gene simply takes each triplet, creates the corresponding amino acid, and strings these amino acids together, stopping when a *Stop* triplet is seen.

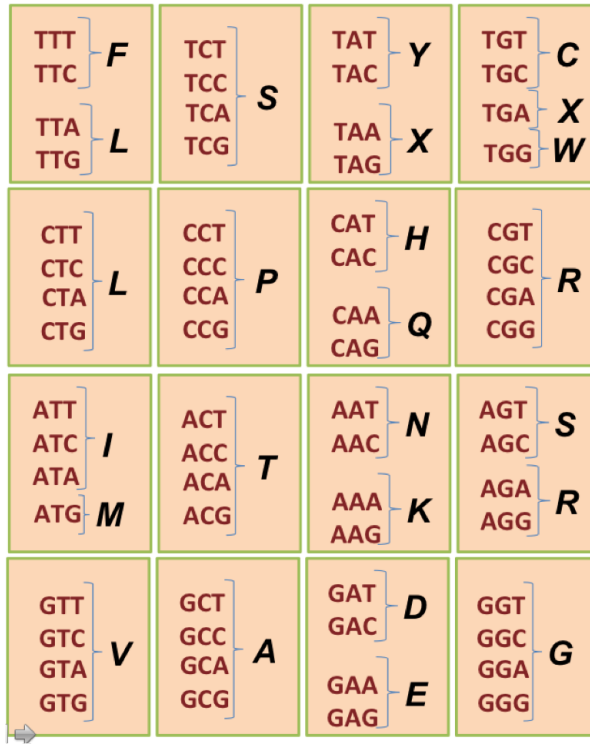


Figure 2.5: The Genetic Code. Amino acids are in bold italics.

Synonymous, Missense, Nonsense

Of course, variants in the genome will impact the above process. For instance, imagine that a variant character in your genome causes the triplet *CCT* to become *CTT*. This change of character from *C* to *T* causes the corresponding amino acid to switch from *Proline* to *Leucine*. The amino acid *Proline* is much smaller in size than *Leucine*, and therefore the protein created by replacing *Proline* by *Leucine* might have a very different shape, and consequently, very different properties. Such variants are called *Non-Synonymous* variants or *Missense* variants, because they lead to a change in amino acid. In contrast, imagine a variant which changes triplet *CCT* to

CCC. Both triplets code for *Leucine*, so no change in the protein. This is called a *Synonymous* variant. Next, imagine a variant which changes triplet *AGA* to *TGA*. *AGA* stands for the amino acid *Arginine* (short form *R*). And *TGA* codes for *Stop* (short form *X*). The recipe execution is prematurely halted in its tracks! What results is a partial protein, often quite incapable of performing its mandated duties. Such variants are called *Nonsense* variants.

Nonsense variants are often quite drastic in their impact. Synonymous variants most often are silent, nothing changes for the protein, so there are typically no adverse effects. Missense variants are a mixed bag though, very hard to assess. Depending on what amino acid changes to what, and where in the protein this change happens, the impact on the protein varies from completely benign to highly damaging. Moreover, these are not the only types of variants that lie within exons of genes. There are others which we will see in subsequent chapters. But for now, as a first pass, we take our 100 genes and see which of these have missense or nonsense variants.

Some variants are present in many people. Since the type of loss of vision we see in $\mathscr{X}$ and her siblings is fairly rare, it is also likely that the variants causing this condition are rare, i.e., they are not present in most people. So we can focus our attention not on all missense and nonsense variants, but *rare* missense and nonsense variants. How do we identify which of the variants in $\mathscr{X}$ are rare variants? For this, we sequence the genomes of many people, particularly those without a clear genetic disease, and create a database of all their variants. Variants which are found rarely in this database, or not found at all, qualify as rare variants.

So we look for rare missense and nonsense variants in $\mathscr{X}$'s exome, and our list of 100 reduces dramatically to under 10. But wait, we argued earlier that a recessive scenario is most likely, that there are offending characters on both copies of the genome. Each of us has two copies of the entire genome, one copy inherited from each of our parents. So there are two copies of each gene. It is likely that the culprit gene has variants in both copies. It could be the same variant in both copies. Or it could be two different variants in the two copies. The former case of the same rare variant being present in both gene copies is particularly common when

the parents of the affected person are related (*consanguineous* is the term). For instance, uncle-niece marriages are common in some communities in southern India. Children born of such parents are at increased risk for carrying the same rare variant in both copies. We will return to this theme a little later in this story. The latter case of two distinct variants in the two gene copies is more likely when consanguinity is not involved. How many genes have rare missense or nonsense variants in both copies of $\mathscr{X}$'s exome? Just one!

The Culprit Gene

The only surviving gene goes by the name *ABCA4*. The first three letters of the English alphabet are coincidental. Here, they actually stand for *ATP-Binding Cassette*. We will see why, shortly. *ABCA4* is located on *Chromosome* 1 (remember, the genome is divided into chapters called chromosomes).

$\mathscr{X}$ has two distinct rare variants, one in each of the two gene copies of *ABCA4*. $\mathscr{X}$'s affected siblings, $\mathscr{B}$ and $\mathscr{S}$, also have both variants. Ditto for $\mathscr{B}$'s affected daughter $\mathscr{D}$. The presence of two rare variants in one gene suggests that $\mathscr{X}$'s disease follows a recessive scenario. But are these variants really the cause of $\mathscr{X}$'s problem? And if so, why? For that, we need to understand how we see.

How we See

Imagine your eyes focused on an object. Light from that object falls on your eye. Then what happens?

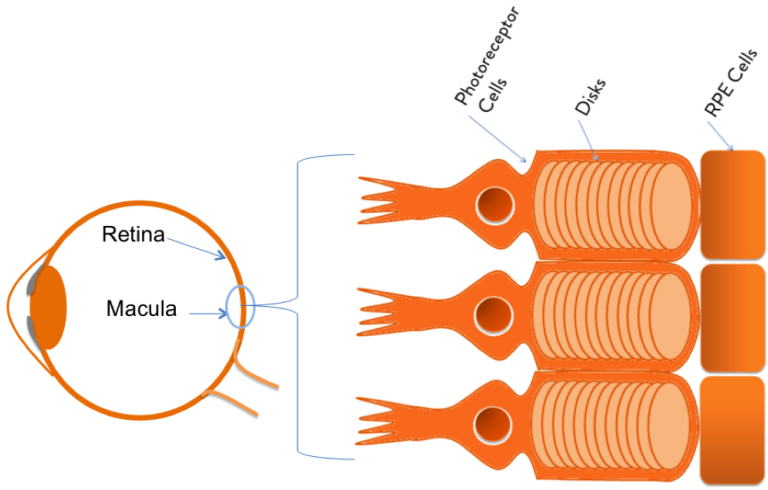


Figure 2.6: Disks in Photoreceptor Cells.

The lens in the eye then bends this light and focusses it on to a screen much like a projector projects presentations on to a screen on the wall. This screen at the back of the eye is called the *Retina* (Fig. 2.6). The retina has special sensor cells called *photoreceptors*, i.e., those cells which receive and detect light. These photoreceptors then convert this light into electrical signals. These electrical signals are then taken to the brain by neurons. The brain then interprets these electrical signals. And that is how we *see*.

The conversion of light signals into electrical signals in the photoreceptor cells is of course the key step in this process. Let's look deeper into that. Photoreceptor cells in the retina are elongated in shape and are shown in Fig. 2.6. They are supported on a base formed by another type of cell, called *RPE*, or *Retinal Pigment Epithelium* cell. Inside each photoreceptor cell is a collection of *disks* These disks are constantly refreshed, i.e., older discs are discarded and newer disks form to take their place. Discarded disks are *eaten up* by the *RPE* cells, thereby cleaning up unwanted debris and creating space for new disks.

Each disk has a hollow interior *lumen* enclosed by a outer *membrane* (Fig. 2.7). On the membranes are embedded the actual light sensor molecules: *Opsins*. Remember, we met them in Chapter 1. Attached to each such sensor molecule is a tag molecule which we'll call *cis-R*. When light hits cis-R, its shape changes into what is called *trans-R*. Both cis-R and trans-R have the exact same chemical composition, just different shapes, because the bonds are bent differently. This change in shape in response to light is the critical event. It sets off a cascade of further events which eventually culminate in a electric signal. Before we get to the cascade, where does cis-R come from, to begin with? From the *Vitamin A* in our diet. Hence the importance of vitamin A for good vision.

This cascade may be hard to remember, but here it is in any case, just to give you a flavor.

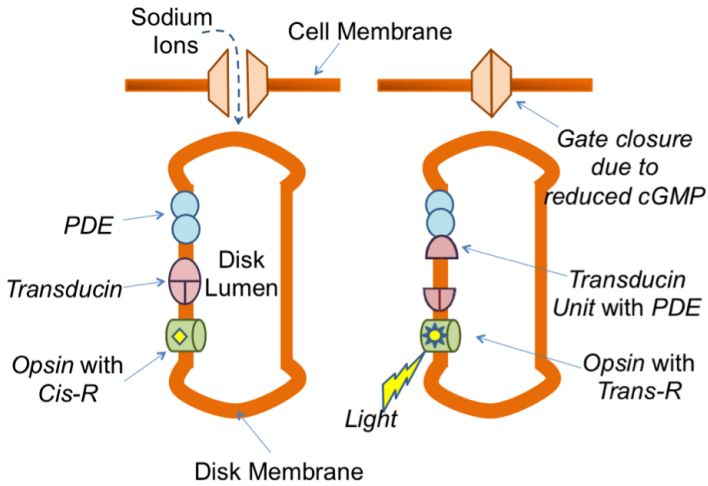


Figure 2.7: Light Response: The picture on the left shows a single disk. The picture of the right shows changes that happen when light falls.

First, trans-R detaches from the opsin sensor. The opsin sensor molecule readjusts its own shape in response. This readjusted shape attracts another molecule called *Transducin*. Transducin has multiple parts

or units. On interaction with opsin, one of these units falls off. This unit then switches on another molecule called *PDE* which in turn breaks down yet another molecule called *cGMP*. cGMP controls a gate on the cell membrane that lets electric current into the cell from outside in the form of positively charged sodium atoms (or *ions*). Breakdown of cGMP causes this gate to close. Electric current can no longer enter the cell through this gate. This switching off of electric current creates an electric signal that is then transmitted to the brain by nerves.

This is how light signals are detected and carried to the brain for interpretation, via clever cascades of molecular interactions. Our interest here, from $\mathscr{X}$'s perspective, is actually not this process itself, but some collateral damage that this process causes.

Collateral Damage

What does the trans-R that detaches from opsin go? Trans-R is toxic to cells; so it cannot be allowed to accumulate. Some of this trans-R falls out of the disk and the rest gets trapped inside the lumen, the hollow interior of the disk. The part that falls out is cleaned up and recycled back into cis-R, needed for detecting further impulses of light. This is done by proteins derived from the *RDH* genes. The part that gets trapped inside the lumen is more problematic. For, there are no *RDH* proteins in the lumen to mop up trans-R. So trans-R accumulates inside disks.

As more light falls on the retina, more trans-R accumulates. As disks age and are eaten up by the surrounding RPE cells, trans-R and its derivatives then accumulate in these cells. Some of these molecules are degraded by these cells, but some molecules (one specifically called *A2E*[8,9]) stay undegraded. A single RPE cell is estimated to eat up around 3 billion disks over 70 years.[10] The resulting accumulated A2E is quite substantial in amount and is clearly visible as a yellow fluorescent pigment using special cameras that can photograph the retina. This accumulation eventually causes these RPE cells to die. As RPE cells die, photoreceptor cells supported by these RPE cells also die. As photoreceptor cells die, vision is lost.

Of course, most of us do not lose our vision. So there must be a gene

here that comes to our rescue. Which gene is that?

ABCA4 to the Rescue

Even though there are two copies of the *ABCA4* gene in every cell in our body, not all cells actually interpret this gene's recipe to manufacture the corresponding protein. Only cells in the retina do (Fig 2.6). Like all other genes, the *ABCA4* gene encodes the recipe for a distinctive protein. Once the cell executes this recipe, a chain of 2273 amino acids is created. This chain then automatically folds into a distinctive 3-dimensional shape. This is the *ABCA4* protein. This protein then moves to and parks itself on the disk membrane (Fig. 2.8).

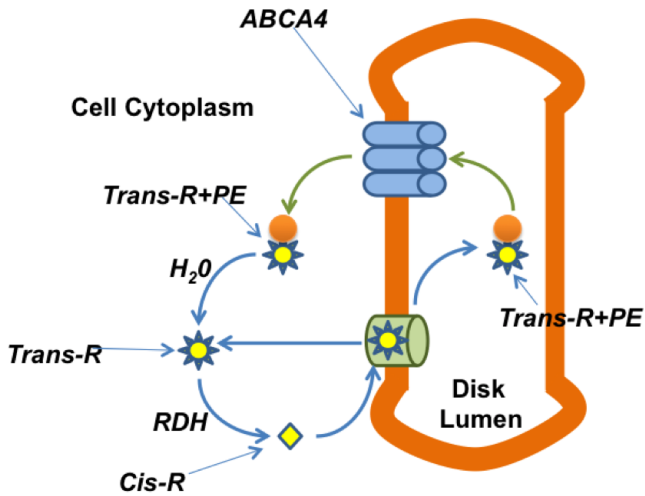


Figure 2.8: *trans-R* Recycling and *ABCA4*'s Role

From its resting place on the disk membrane, *ABCA4*, the janitor, cleans up the disk innards by pumping the trapped trans-R out of the disk. The trans-R trapped inside the disk is in a temporarily safer form called trans-R^{PE}. Once *ABCA4* pumps trans-R^{PE} out, it gets recycled back to

cis-R as shown in the above picture. This ensures that trans-R does not accumulate within disks. There are many fewer *ABCA4* protein molecules than opsin molecules, so one *ABCA4* protein molecule cleans up trans-R from many many opsin molecules. And so does *ABCA4* keep our house tidy and our vision going.

What happens when genomic variants in *ABCA4* compromise its janitorial abilities? Trans-R then starts to accumulate inside disks. Over several years of accumulation, photoreceptor cells die and vision is lost. In $\mathscr{X}$'s case, 30-40 years of accumulation was required before consequent loss of vision. Of course, the accumulation period needed for loss of vision would depend upon how badly compromised *ABCA4* was. Some genomic variants could compromise *ABCA4* more than others. Do $\mathscr{X}$'s genomic variants indeed compromise *ABCA4*'s ability to pump out trans-R? To assess this, we need to understand how *ABCA4* actually works.

How *ABCA4* Works

Here is how an *ABCA4* protein molecule looks like, parked on the disk membrane (Fig. 2.9). Some parts of the *ABCA4* protein appear inside the disk, some outside, and the rest sit on the disk membrane itself. With this placement, the *ABCA4* protein is set to move trans-R across the disk membrane as follows.

First, of course, *ABCA4* needs energy to pump trans-R out. Energy is derived from food. But the energy so derived has to be collected and packaged into batteries so it can be moved around and drawn upon in appropriate amounts, on demand. A molecule called *ATP* or *Adenosine Tri-Phosphate* serves this purpose. As its name indicates, it comprises 3 phosphates; removal of one of these releases just the right amount of energy to fuel many activities in the cell. After this removal, ATP becomes *ADP*; a separate mechanism then recycles this *ADP* and uses energy from food to convert it back to ATP.

Coming back to *ABCA4*, the process of moving trans-R^{PE} across the disk membrane starts with trans-R^{PE} attaching itself to the *ABCA4* protein inside the disk. This triggers changes in the *ABCA4* structure. In turn, this allows ATP to attach to the NBD2 part of *ABCA4*,[11] shown in Fig.

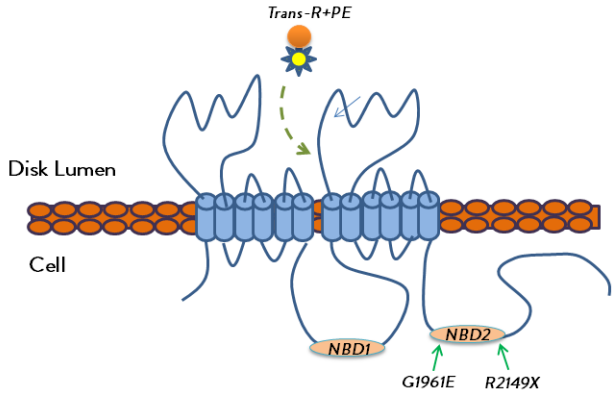


Figure 2.9: *ABCA4* in the Disk Membrane.

2.9 outside the disk. Phosphate removal from ATP then yields a burst of energy which twists the *ABCA4* protein so the NDB1 and NBD2 parts are brought together and gaps are exposed in the portions embedded in the membrane, allowing trans-R^{PE} to move across the membrane. So does *ABCA4* literally shepherd trans-R^{PE} out of the disk.

Do $\mathscr{X}$'s genomic variants indeed compromise this orchestration? If so, how?

$\mathscr{X}$'s *ABCA4* Variants

$\mathscr{X}$ and her affected siblings each have two genomic variants in the *ABCA4* gene. One variant occurs in the one copy of the gene and the other variant occurs in the other copy.

Figure 2.10: *R* changes to *X*: Nonsense

The first variant is called *R2149X*, a code-name of sorts, but quite straightforward to decipher. The *ABCA4* protein comprises 2273 amino acids. The 2149th amino acid is usually *Arginine* (abbreviated to *R*). See Fig. 2.5, `CGA` is one of the three-character combinations which code for *Arginine*. If the `C` in `CGA` somehow changes to a `T`, then `CGA` becomes `TGA`, which changes the amino acid to *Stop* (abbreviated to *X*) (Fig. 2.10). So, in $\mathscr{X}$, amino acid *R* at position 2149 is replaced by an *X*. So this is a nonsense variant which terminates the protein prematurely at the 2149th amino acid. The protein just stops dead in its tracks, mid-way!

And where does the truncation happen? Right within the NBD2 part, where an ATP molecule attaches twisting *ABCA4* so NDB1 and NBD2 come together, thus squeezing trans-R^{PE} out across the membrane. With only a partial NBD2 domain, it becomes much harder or even impossible for the ATP molecule to attach (in fact, this truncated protein may not even be able to make their way to the disk membrane).[12] With this handicap, *ABCA4* is just unable to move trans-R^{PE} across the disk membrane.

R2149X is very rare; it has been reported (in scientific peer-reviewed literature) in just a handful of individuals with macular degeneration.[13] In each of these individuals, *R2149X* appears on only one of the two *ABCA4* gene copies, never on both. The other gene copy carries another genomic variant in each case. This is indeed the case for $\mathscr{X}$ and her affected siblings as well. *R2149X* appears on one of the two *ABCA4* gene copies and the other gene copy carries a genomic variant code-named *G1961E* (Fig. 2.11).

Figure 2.11: *G* changes to *E*: Missense

Of the *ABCA4* protein's 2273 amino acids, amino acid 1961 is usually a *Glycine* (abbreviated to *G*). In $\mathscr{X}$ and her affected siblings, this *Glycine* is replaced by another amino acid *Glutamic Acid* (abbreviated to *E*). So this is a missense variant, i.e., one amino acid has been changed to another. The *G1961E* variant is relatively more common, 2 in 1000 people have an *E* instead of a *G* for amino acid 1961, in one of their two *ABCA4* gene copies. Do all these people suffer from some form of blindness? No. Because, the second *ABCA4* gene copy is still intact and capable of cleaning up trans-R^{PE} by itself. The problem arises in only those who have a compromised second copy, like $\mathscr{X}$. This happens in only 1 in 10,000 individuals.

Like *R2149X*, *G1961E* is also located within the *NBD2* domain of the *ABCA4* protein. Since *G1961E* is relatively much more common, it has been well studied. Scientists have shown *G1961E* reduces the ability of *ABCA4* to attach to ATP.[14] This ability reduces even further when trans-R^{PE} attaches to the *ABCA4* protein inside the disk. Unlike *R2149X* which completely annuls *ABCA*'s ability to clean up toxic trans-R^{PE}, *G1961E* does not annul this ability completely, rather it just reduces this ability. Nevertheless, for $\mathscr{X}$ and her affected siblings $\mathscr{B}$ and $\mathscr{S}$, these two genomic variants together reduce this ability to such an extent that toxic trans-R^{PE} accumulates over three decades leading to loss of vision.

What this means

$\mathscr{X}$, her affected siblings $\mathscr{B}$ and $\mathscr{S}$, as well as $\mathscr{B}$'s affected daughter $\mathscr{D}$, all have both the *G1961E* and the *R2149X* variants. Together, these two variants compromise both *ABCA4* gene copies. Based on this, an

ophthalmologist would classify their condition as Stargardt disease, named after Karl Stargardt, an ophthalmologist in Berlin, who first characterized this disease, in 1909. Stargardt disease is generally considered to initiate in the second decade of life (indeed, that is why it is also called *juveline macular dystrophy*). However, this is not a uniform rule. Indeed, the age of onset can vary from case to case, depending upon many factors, including the actual variants in the *ABCA4* gene. In $\mathscr{X}$'s family, the relatively milder nature of the *G1961E* variant pushes out the age of onset to the fourth decade of life. In contrast, two copies of the *R2149X* variant would likely have lead to much earlier onset.

There is good news for $\mathscr{X}$'s children though. Both children (one in the late 30s and one in early 40s) have only the *G1961E* variant, i.e., neither has the *R2149X* variant. This means that one copy of the *ABCA4* gene is intact and fully functional in both, and the other copy is partly but not fully compromised. So neither is at risk for developing the same symptoms as their mother, a fact that can be asserted with confidence, simply by studying their genomic sequence.

$\mathscr{B}$'s two unaffected daughters (keeping aside $\mathscr{D}$) are both in their 40s, and have no symptoms thus far. One of these daughters $\mathscr{E}$ has an interesting combination: two copies of the *G1961E* variant, one on each copy of the *ABCA4* gene. Remember, *G1961E* does not annul the ability to transport toxic trans-R^{PE} across the disk membrane completely, rather it just reduces this ability. This begs the following question: can two copies of the *G1961E* variant, one on each copy of the *ABCA4* gene, be causative of vision loss? If so, then at what age of onset and with what severity? Observations on patients with two copies of the *G1961E* variant show that such patients usually have a rather mild form of the disease, with widely varying ages of onset reported, ranging from 19 to 64.[15] So $\mathscr{E}$ may yet see some impact of this variant, but likely in milder form.

Two copies of the *R2149X* variant would be far more serious in comparison. Fortunately, no member of this family has this combination. $\mathscr{X}$'s niece $\mathscr{N}$ from an unaffected sibling has one copy of the *R2149X* variant though. Since the other copy of the *ABCA4* gene does not have a significant variant, $\mathscr{N}$ remains a carrier and is herself not affected. $\mathscr{N}$'s children are still in their late teens and mid-twenties. Could they be susceptible to

Stargardt disease even if neither of their parents is?

$\mathscr{N}$ has two copies of the *ABCA4* gene and each child inherits one of these two copies. So each child has a 50% chance of inheriting *R2149X*. Even if a child inherits *R2149X* from $\mathscr{N}$, (s)he will not be affected by Stargardt disease unless (s)he inherits another problematic variant from the father. Since the *G1961E* variant is found in just 2 in a 1000 people and since *R2149X* is even rarer, it is unlikely, in general, that the father would carry any of these variants. However, not so unlikely in this case! Because $\mathscr{N}$ and her husband form a consanguineous couple, i.e., they are genetically related first-cousins (Fig. 2.2). Since $\mathscr{X}$'s extended family has many instances of *G1961E* and *R2149X*, $\mathscr{N}$'s husband's chance of carrying one or more of *R2149X* or *G1961E* is far higher than 2 in a 1000; in fact, as high as 1 in 3, or about 33.3%. So each of $\mathscr{N}$'s children has a 16.6% chance of inheriting one of these problematic variants from their father, and a 50% chance of inheriting *R2149X* from the mother, yielding a net 8.3% chance of susceptibility to Stargardt disease. Not large, but not that small either!

$\mathscr{N}$'s children have not been tested for variants in the genome, neither has $\mathscr{N}$'s husband. The dangling sword (the 8.3% chance) could easily be removed by checking their *ABCA4* gene for variants. With more than a 90% chance, this test will set that dangling sword to rest, for good. But, the test could come out positive as well, albeit with under 10% chance, subjecting the individual to added psychological burden of impending vision loss. Should $\mathscr{N}$'s children indeed get tested? What can they do to ward off or push out the disease, in the even the test comes out positive?

Stargardt disease is caused by accumulation of trans-R^{PE} inside the disk. trans-R^{PE} is generated each time light causes a change from cis-R to trans-R. Reduced exposure to light could reduce rate of accumulation of trans-R^{PE}, thus postponing disease onset. Experiments done in mice kept in total darkness indeed show lack of accumulation of toxic material.[8] So avoidance of direct exposure to sunlight (via sunglasses, for instance) is potentially useful. Conventional wisdom also prescribes the use of antioxidants like Vitamin A for good vision. But cis-R, trans-R and A2E are all derivatives of vitamin A and increased vitamin A consumption could lead to greater toxic accumulation in Stargardt patients. So avoiding

Vitamin A supplementation is potentially useful as well. These measures could slow down disease progression but are unlikely to stop the disease in its tracks.

Are there stronger measures that could help $\mathscr{N}$'s children prevent vision loss (if they were indeed susceptible, that is)? Or restore vision loss in $\mathscr{X}$ and her affected siblings? Unfortunately, not yet. However, help might be around the corner, though it may be several years before this help becomes routinely available.

Gene Therapy

If one is born with defective copies of the *ABCA4* gene, can one not correct for this by providing good copies of this gene artificially? *Gene Therapy* intends to do exactly this.[16] There are many challenges though.

First, externally administered copies of *ABCA4* need to reach their intended destination (the photoreceptor cells in the retina) unharmed. This is not easily achieved in general. Packaging these gene copies into oral pills would be convenient but will just not work. The digestive system will degrade the gene well before it can reach the bloodstream. Even if it reaches the bloodstream unharmed, it will be diffused all over the body. It would then be difficult to control the amounts that reach specific destination cells. And there may be side-effects on account of the gene reaching cells other than the intended recipients. Fortunately, the retina is easily accessible and gene copies can be directly deposited on specific cells in the retina via an injection. Reaching a more deeply buried and less localized organ, e.g., the lung, would be much harder.

Second, photoreceptor cells should be able to interpret the recipes coded by these injected gene copies to create *ABCA4* protein copies. Photoreceptor cells should also be able to direct these protein copies to the disk membrane. For this to happen, the externally injected gene copies have to be packaged in a certain form that enables these copies to enter the cell and be amenable for recipe interpretation. The cleverness of gene therapy lies is its use of specially engineered viruses for this job. Certain viruses have naturally evolved to attach themselves to certain types of human cells and let their genomic sequences loose into the cell. Some even go further and

integrate their genomic sequences into the human genome sequence inside this cell! Human cells can no longer distinguish between human and the viral genes, and interpret all of these recipes, creating the corresponding proteins in the process. The *HIV* virus that causes *AIDS* is an example of such a virus. Of course, such viruses also cause undesirable disease. So special versions of these viruses have been engineered to package good copies of *ABCA4* and carry these into the photoreceptor cell without the threat of disease.

Third, if a virus incorporates its own genomic sequences into the human genome sequences, there is the danger of the human genome sequence itself getting modified in a potentially dangerous way. This could happen if the viral sequence is inserted in the middle of an important human gene. This gene's recipe is then disrupted by the inserted viral sequence and it now codes for a different protein, which may be unable to perform its intended function. To get around this issue, special *non-integrating* viruses are used. These viruses do push their genomic sequences into human cells. However, they do not integrate these sequences into the human genome. Instead, viral genome sequences are kept as separate entities (called *episomes*) inside the human cells, still available for recipe interpretation but without disrupting the human genome sequence.

Fourth, is the problem of dividing cells. Many cells in the human body are constantly dividing. When a cell divides, it breaks into two new cells. Episomes in the dividing cell then get split among the two daughter cells, reducing the number of episomes in any one cell. Over many rounds of division, few or no episomes will be left per cell. So a single viral injection may not suffice to provide good copies of *ABCA4* to all the relevant cells; repeated viral injections may be needed. Fortunately, cells in the retina are not actively dividing cells, so a single viral injection suffices to provide sustained supply of a good *ABCA4* gene copies.

Fifth, is the issue of side-effects. The human immune system often identifies and destroys foreign viruses, causing inflammation and scars in the process. This is one of the key stumbling blocks of gene therapy. The retina is however blessed in this regard; it can tolerate viral injection without a strong immune response.

So retinal diseases are quite well-suited for gene therapy and much

progress has been made in the last few years. The process of testing a therapy and readying it for routine medical use is quite slow and elaborate though. New therapies are tested first in animals. Trials in mice with severely compromised *ABCA4* genes have indeed shown that a single injection of healthy *ABCA4* packaged in a specifically designed virus (called *Lentivirus*) reduces accumulation of A2E in RPE cells down to levels comparable to those in healthy mice, even 1 year after the injection.[17] These promising results have justified launch of human trials, conducted under the code name *StarGen*.[18] Human clinical trials proceed in phases. Phase I checks whether the therapy is safe and without adverse effects, and if so, at what dosage. Effectiveness of the therapy is studied only in Phase II, on a small number of patients. If Phase I shows no adverse affects and Phase II indicates that the therapy is effective then Phase III tests the therapy in a larger pool of patients. Phase I/II for *StarGen* started in 2011 and is expected to complete in 2015. If all goes well, *StarGen* may be ready for medical use a few years after 2015.

StarGen is probably too late to be of use to $\mathscr{X}$ and her affected siblings. gene therapy can only help live cells get around malfunctioning *ABCA4*. It cannot resurrect dead cells. $\mathscr{X}$ and her siblings have already suffered substantial death of photoreceptor cells. These cannot be brought back by *StarGen*. However, it may help arrest further death of cells in $\mathscr{B}$'s affected daughter $\mathscr{D}$, still in her late 40s. She has the option of enrolling in the *StarGen* trial, which is indeed recruiting patients, as of mid-2014. $\mathscr{N}$'s children, with a 8.3% chance of Stargardt disease 10 years from now, would do well to carefully track the health of their eyes via regular check-ups so any vision loss can be detected early. Sequencing the *ABCA4* gene at the earliest sign of disease will confirm that the disease is due to *ABCA4* variants. Hopefully, *StarGen* would be well established and ready to help if and when this happens.

But what if *StarGen*'s clinical trial fails to deliver? Are there alternatives on the horizon?

Cell Therapy

A Phase I/II clinical trial is on for a completely different form of therapy.[19,20] This form of therapy does not seek to replace bad copies of the *ABCA4* gene with good copies. Rather, it seeks to replace dying or dead RPE cells with fresh, healthy RPE cells, and is aptly called *Cell Therapy*. If the RPE cells in the Retina are functional and healthy, then the photoreceptor cells supported by these RPE cells will also remain healthy, and normal vision will continue. Cell therapy requires a supply of fresh, healthy RPE cells. Where does this come from? Finding a donor and extracting RPE cells from the eyes of that donor is cumbersome if not infeasible. *Embryonic Stem Cells* serve as an alternative and plentiful source. These stem cells are easily obtained from early-stage embryos derived by combining sperm and egg cells from donors in a test tube. Left in the embryo, these cells will eventually *differentiate* into different cell types: some cells will mature into heart cells, others in brain cells etc. But if these cells are taken out of the early-stage embryo, they continue to stay in their undifferentiated state. Then, by clever programming, i.e., manipulating certain combinations of genes, these undifferentiated cells are forced to differentiate into RPE cells. These *RPE* cells are then injected into the Retina. Ongoing trials will determine whether this therapy is safe and effective, but preliminary experiences on 2 patients appear encouraging.

Are there any further alternatives as well?

Nonsense Suppression

One of $\mathscr{X}$'s *ABCA4* variants was a nonsense variant, causing premature truncation of the *ABCA4* recipe. In its own way, this may turn out to be a stroke of luck for this family. For, drugs which allow the cellular machinery to continue interpreting a gene recipe past a premature truncation (to some extent) are under active exploration.

Note the beauty of this approach. In theory, it is applicable to any gene and any disease, as long as the disease is caused by a nonsense variant in the gene. Of course, clinical trials have to be performed to prove the practicality of this approach. And clinical trials usually focus on specific diseases and genes. For instance, the *DMD* gene causing

Duchenne Muscular Dystrophy in young boys is one focus. Boys with this disease progressively lose muscle strength and end up paralyzed at a very young age. The *CFTR* gene causing *Cystic Fibrosis* has been another focus. This diseases causes the accumulation of sticky mucous in the airways, causing breathing difficulties, and eventually leading to lung failure. Trials in both cases are ongoing and do seem to show positive results, at least in some individuals.[21] Unfortunately, Stargardt disease is not one of those for which trials have been conducted. But the promise of this approach holds nevertheless.

Wrapping Up

$\mathscr{X}$, $\mathscr{B}$ and $\mathscr{S}$ have been living with increasing levels of vision loss for the last 30-40 years. Yet, a definitive diagnosis has become possible only in the last 10 years. Possible, but not necessarily affordable and easily accessible. The latter only in the last couple of years. The cause of their affliction in now understood in substantial detail. And it reminds us of the inevitable: that if unwanted debris, whether on our streets or in the cells in our eyes, is not cleaned up and disposed properly, it will take its toll eventually.

Corrective intervention by supplementing with good copies of *ABCA4* or with healthy RPE cells or via nonsense suppression will hopefully provide a preventive strategy for this problem soon. Though too late for $\mathscr{X}$, $\mathscr{B}$ and $\mathscr{S}$, $\mathscr{D}$ may be able to arrest further disease progression with these methods. $\mathscr{N}$'s children do have time on their side and would do well to keep abreast of these developments. Unfortunately, the rarity of Stargardt disease makes such efforts economically unviable or uninteresting for most pharmaceutical companies; non-profit foundations and governments have to play active roles in funding these trials and speeding them up.

$\mathscr{X}$'s children can breathe easy though, with the definitive knowledge that they are not at risk. In the meantime, *G1961E* and *R2149X* will continue to survive and propagate unnoticed in carrier individuals in this family. Their presence will be felt again when they cross paths (either with each other, or with other problematic variants). Hopefully, by the time this happens, therapies would have entered routine medical practice.

Chapter 3

The Rhythm Goes Awry

$\mathscr{X}$, a young woman in her mid twenties, suffered frequent bouts of breathlessness. Her twin sister, $\mathscr{Y}$, had the same issue. Tests to monitor the heart were duly performed. These indicated heart failure, i.e., the inability of the heart to pump sufficient blood to meet the needs of the body.

The tests also indicated an abnormally fast heart rhythm. Well-spaced, periodic heartbeats allow the heart chambers to fill with blood before the heart contracts to pump this blood out to the body. A faster, chaotic rhythm makes the heart contract even before its chambers fill up with blood. Little blood is pumped out then, causing unconsciousness suddenly. Death could follow in minutes unless emergency medical attention can keep blood and oxygen flowing and an electric shock can be administered to the chest to restore normal heart rhythm.

Promptly, an *ICD* (implantable cardiac defibrillator) device was implanted in both twins. This device is implanted in the chest with electrodes that connect to the heart chambers. It detects any irregular hear rhythms and offsets these with small electrical shock pulses in an effort to restore normal rhythm. The issue of heart failure remained though. The heart was just not pumping out blood in sufficient quantities, so the situation remained life-threatening. Indeed, $\mathscr{Y}$ passed away due to sudden cardiac arrest in her early 30s.

Sudden cardiac death was not new to us when we took on this case. One of our first hires at Strand was $\mathscr{M}$, a very talented scientist in his mid-twenties. He had a frail build but appeared healthy otherwise. Three years of productive and often intense work exposed no signs of trouble

lurking underneath. Until increased irregularity in his heart rhythms and failure of existing medication to control this irregularity forced him to take a few weeks off. He never returned, succumbing to sudden cardiac arrest.

$\mathscr{M}$, $\mathscr{X}$ and $\mathscr{Y}$ all show(ed) warning signs of irregular heart rhythm. But warning signs are not always present. The onset can be sudden and dramatic sometimes, as in the case of footballer Miklós Fehér, a Hungarian footballer playing for Portuguese club Benfica. On 25 January 2004, 24 year old Fehér came on as substitute in a match against the club Vitória de Guimarães, and assisted with a goal. At the fag end of the game, in injury time, he collapsed to the ground. Emergency medical attention did not help. The cause of death was soon confirmed to be sudden onset of arrhythmia.

Apparently $\mathscr{X}$'s father, $\mathscr{F}$, had also passed away in his 40s. The exact cause of death was not known. His family members remembered it simply as a "heart attack". $\mathscr{X}$ has a sister $\mathscr{S}$ and a brother $\mathscr{B}$. Both had no symptoms whatsoever. To be doubly sure, they were tested for heart electrical activity, structure, and blood flow. And these tests revealed the unexpected. $\mathscr{B}$'s heart appeared completely normal. On the other hand, $\mathscr{S}$'s heart showed the same dangerous heartbeat rhythm as $\mathscr{X}$, even though she had not felt breathless or shown any other symptoms yet. An ICD was implanted in her as a precaution.

What was the cause of abnormal heart rhythms and heart failure in this family? Could $\mathscr{B}$'s currently normal heart rhythms turn abnormal over time? Were children of $\mathscr{S}$ and $\mathscr{B}$ at risk?

Where Do We Look?

With genome sequencing, we have a way forward now to answer these questions. Sisters $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{S}$ and their father $\mathscr{F}$, all very likely share a genomic character which causes these abnormal heart rhythms. Can we identify this offending character by sequencing $\mathscr{X}$'s and $\mathscr{S}$'s genomes?

As in Chapter 1 and Chapter 2, our window into the sisters' genomes is *Whole Exome Sequencing*. Remember, there are about 21,000 *genes* in our genome. Each gene is a stretch of text in the genome comprising exons and intervening introns (see Fig. 1.12). Only exons carry recipes for

creation of protein molecules, introns don't. There are roughly 200,000 exons over all 21,000 genes. The genomic text stretches in these 200,000 exons together comprise what is called the *exome*. The exome measures just over 1% of the entire genome. Whole exome sequencing followed by a lot of data crunching on a computer gets us $\mathscr{X}$'s and $\mathscr{S}$'s exomes. In particular, it identifies every character in the exome where $\mathscr{X}$ and $\mathscr{S}$ differ from a supposedly healthy genome sequence. These characters are called *variants*. The main problem again: there are tens of thousands of these variants! How do we identify which of these is the culprit?

As in Chapter 2, we start by narrowing down genes of importance. We do this by making a list of genes which impact the development and functioning of the heart, as reported in scientific literature. This literature is vast, comprising research over the last 3 decades. And, of course, it is not complete; more genes will be identified over time. For now, these genes serve as a good starting point for our hunt. A computer-assisted scan of published literature gets us this shortlist. How do the 200 or so genes in this shortlist impact cardiac rhythms? And which of these genes is the cause of $\mathscr{X}$ and $\mathscr{S}$'s condition? The quest for the answer takes us through a quick tour of the heart.

The Heartbeat

The heart beats day-in and day-out, steadily and tirelessly. Where does this heartbeat come from? It turns out that the heart can beat by itself even when it is cut off from the brain. Which means the heartbeat comes from within the heart itself; signals from the brain only make this beat go faster or slower. A built-in *clock* in the heart supplies this heartbeat. This clock comprises special cells with *gates* on their surface which open and close systematically under the actions of at least 12 genes, called *channel* genes. This systematic opening and closing, in a certain sequence, causes the voltage difference between the outside and inside of the clock cells to follow a certain periodic pattern, shown in Fig. 3.1. That is the genesis of the heartbeat.

This electrical signal now needs to travel from the clock cells to the muscle cells, which contract in response to this electrical signal. It is this

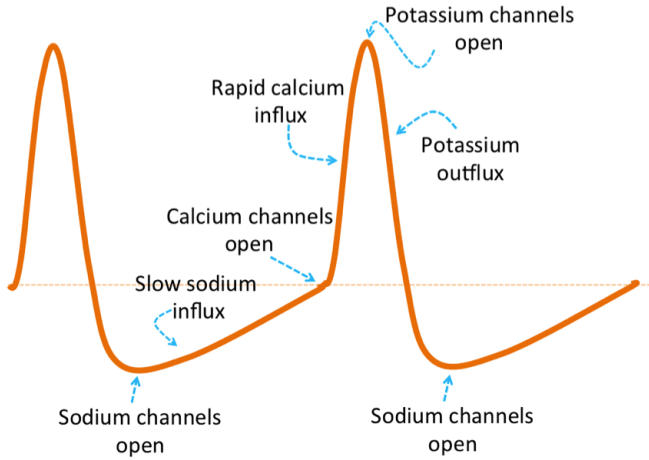


Figure 3.1: The Electrical Signal Generated by the Built-in Clock.

contraction that enables the heart to perform its main function: pumping of blood to the rest of the body. How does the electrical signal generated by the clock cells reach these muscle cells? Are there wires in the heart that carry this signal?

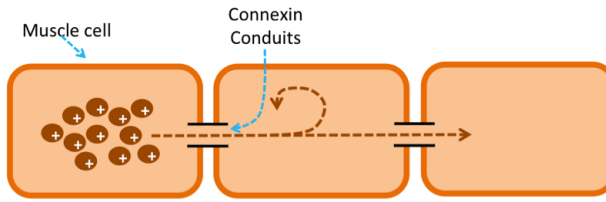


Figure 3.2: Conduits between Cells Transmit the HeartBeat.

Of course not. But cells in the heart are packed in a special way with each cell having special conduits to its neighbors. These conduits are made from recipes coded by genes called *connexins*. They allow charged

ions from one cell to move freely to the neighboring cell (see Fig. 3.2). When the voltage inside one cell increases, ions flow from that cell to its neighbors, thus raising the voltage inside the neighboring cells as well. The electrical signal thus spreads from one cell to its neighbors, and so on. Some cells have conduits made from faster connexins; these behave like fast moving highways. Yet others have conduits made from slower connexins; these behave like slower-moving city roads. Why does the heart need both fast-moving highways and slow-moving city roads?

The answer lies in the careful orchestration required between different parts of the heart for effective pumping of blood. The heart has 4 chambers, the right and left *atria*, and the right and left *ventricles* (see Fig. 3.3). These 4 chambers together pump blood in two loops: the first loop takes blood to the lungs and back for oxygenation; the second loop takes this oxygenated blood to the rest of the body and back. Blood returning from the body enters the heart via the two atria: oxygen-rich blood from the lungs enters via the left atrium and oxygen-poor blood from the rest of the body enters via the right atrium. The atria then contract to push the blood into the ventricles. Contraction of the right atrium pushes blood into the right ventricle; contraction of the left atrium pushes blood into the left ventricle. When the atria contract, the ventricles must relax, so they accept blood from the atria without resistance. Then the ventricles contract. Contraction of the right ventricle pumps blood into the artery which takes oxygen-poor blood to the lungs for oxygenation. Contraction of the left ventricle pumps oxygen-rich blood out to the rest of the body.

Note the sequence above, atria contract first and then the ventricles contract. Then, when the ventricles contract, they must do so with force and rapidly. How is this sequence orchestrated with every heartbeat?

The clock cells are located in a tiny clump in the right atrium. The signal from these cells must rapidly spread throughout both atria, so they can contract quickly and simultaneously. The signal must wait before reaching the ventricles though, so the ventricles relax when the atria contract. Once the atria have contracted, the signal must rapidly spread throughout the ventricles so all muscle cells in the ventricles contract together. This careful orchestration requires that the signal from the clock spread at different speeds through different parts of the heart. This requires

connexins of different kinds, some faster, some slower.

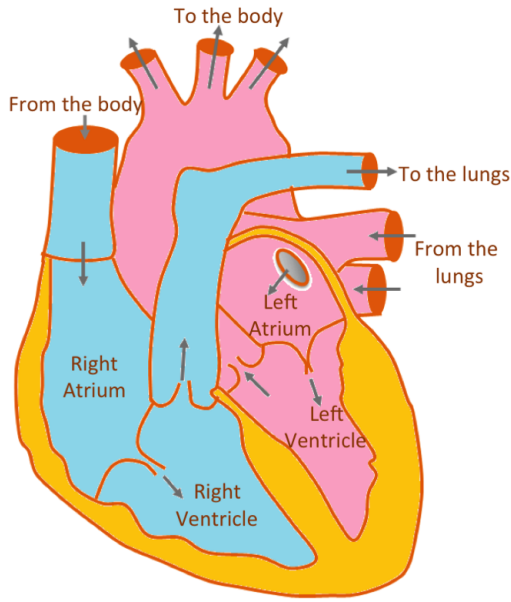


Figure 3.3: Blood Flow in the Heart

Of course, faulty channel genes and connexin genes upset this careful orchestration leading to uncoordinated contraction and arrhythmia. Roughly 330,000 people die of sudden cardiac arrest every year in the United States. Around 5% of these patients have no structural problems in or around the heart. The cause of death in these cases remains unexplained even after comprehensive forensic investigations and autopsy.

The office of the Chief Medical Examiner, New York City, studied 274 such deaths,[22] half of whom were infants and most of the other half were people between the ages of 19 and 58, i.e., not old enough for the heart to fail due to age. They checked each of these individuals for variants in 6 channel genes (two channels which control sodium ions, two which control potassium ions, and one which controls calcium ions). Problematic genomic variants were indeed present in roughly 16% of these individuals.

Even though an autopsy and forensic examination could not uncover cause of death for these cases, genetic testing revealed that the underlying disease was a *Channelopathy*, i.e., defective electrical signal on account of a faulty channel gene. About half of these cases involved a single sodium channel gene, *SCN5A*.

It is possible that several of these individuals would have shown no symptoms or abnormal patterns in standard heart tests.[23] Genetic testing would have identified some of these individuals, who could then potentially be treated, e.g., by having an ICD device implanted, as for $\mathscr{X}$. This study of course looked for variants in only 6 of close to 25 known channel-related genes,[23] altered recipes in which cause heart disease; screening more genes may have identified the problem in more cases.

Could genomic variants in the channel genes or the connexin genes be the cause of $\mathscr{X}$'s condition? Their heartbeat was indeed abnormally fast. But problems in heartbeat generation or transmission are pure electrical problems. In $\mathscr{X}$'s case, the heart showed additional structural changes, which suggest that $\mathscr{X}$'s problem was not purely electrical. Faster heartbeats could also result, for instance, due to problems in the heart muscle. Maybe that was a more likely cause?

The Muscle

The heart has special muscle cells. The heartbeat generated by the clock cells eventually reaches every one of these muscle cells. These cells contract in response. This rhythmic contraction and relaxation is what the heart is most associated with: its pumping action. How does heartbeat lead to this rhythmic contraction and relaxation? Nature has designed a clever spring-like structure called a *sarcomere* for this purpose.

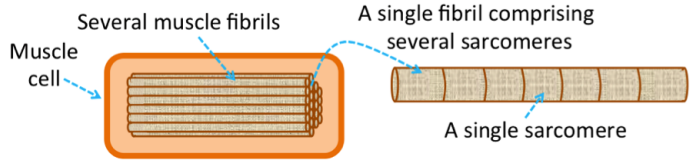


Figure 3.4: Bundles of Sarcomeres in a Muscle Cell.

Muscle cells contain bundles of fibrils, each fibril comprising several sarcomeres connected end-to-end in a chain, as shown in Fig. 3.4. A single sarcomere is shown in Fig. 3.5.

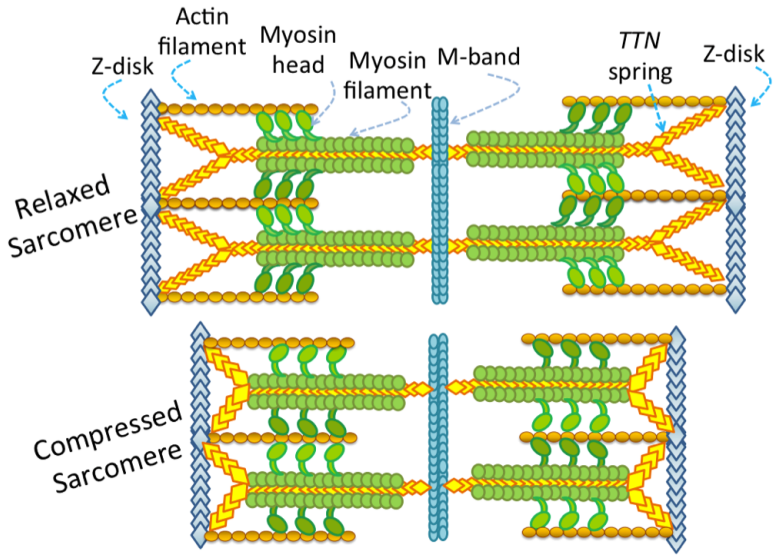


Figure 3.5: Sarcomere Contraction.

As you can see, it is sandwiched between the blue lines on the outside, with a cyan center line dividing the sarcomere into symmetric left and right halves. In each half, there are ropes, or *filaments*, marked *actin* and *myosin*. Myosin filaments have projecting heads. In response to the heartbeat signal,

these heads attach to the actin filament and pull the filament towards the center line. In turn, the actin filaments pull both ends of the sarcomere towards the center, leading to contraction of the sarcomere. Myosin heads do of course need energy to pull the actin filaments; this energy is derived from the *ATP* battery which packages energy derived from the food we eat (sounds familiar? we saw this in Chapter 2).

Each component of the sarcomere is derived from recipes encoded in one or more genes. The myosin filament itself comprises proteins from several different genes. The actin filament is derived from a single gene. And did you notice that the myosin and actin filaments themselves are not elastic, they do not compress in length when the sarcomere contracts and the outer ends move closer to each other. But something must compress. What is the spring that holds the sarcomere together? That's the protein derived from the *TTN* gene, the largest gene in the body. It connects the outer ends to the center line, folding and unfolding as the sarcomere compresses and relaxes, as shown in yellow in Fig. 3.5.

And how does the sarcomere detect the arrival of the heartbeat? Attached to the actin filaments are proteins from several genes. These proteins prevent the myosin heads from attaching and pulling the actin filaments. When the heartbeat arrives in the form of calcium ions, these proteins detach from the actin filament; myosin heads are now able to pull the actin filament, thus causing sarcomere compression. Sensitivity to this electrical signal is controlled by other genes, most notably the *MYBPC3* gene.

Sarcomeres are complex structures. And complexity comes with its chinks. Indeed, faults in several of these sarcomere genes result in defective muscle cells which in turn leads to abnormal heart structure and rhythms. The most common of these structural abnormalities is called *Hypertrophic Cardiomyopathy* (*HCM*, in short). The walls of the left ventricle are abnormally thicker and stiffer in individuals with HCM (see Fig. 3.6). This reduces the volume of blood that the left ventricle can carry and pump (remember that the left ventricle pumps out blood to most of the body). Physical exertion can then bring about chaotic heart rhythms, leading potentially to death, as seen in athletes like Fehér. HCM is not uncommon, found in as many as 1 in 500 young adults.[24] And it is the most common cause of heart-related sudden death in people under 30 years of age.

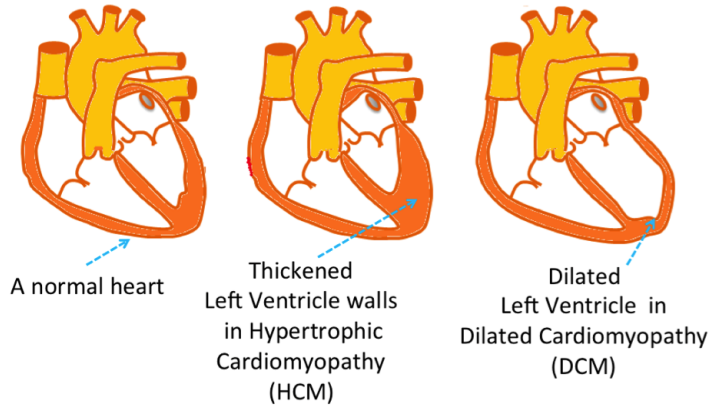


Figure 3.6: A Hypertrophic Heart and a Dilated Heart.

Another condition called *Dilated Cardiomyopathy* or *DCM* is less common, 1 in 2500 individuals. Unlike HCM, the heart appears dilated in DCM, i.e., the left ventricle wall is thinner and stretched out; the ventricle itself looks larger than normal. The muscle strength of the ventricle is not adequate to pump out most of the blood present in the ventricle with every heartbeat.

Variants in genes of the sarcomere are the typical causes of HCM and DCM. Could any of these genes be the cause of $\mathscr{X}$'s condition? Their condition did not resemble DCM or HCM directly. There appeared to be abnormal fissures in the muscle of the left ventricle, a rarer phenomenon. What else could have gone wrong?

The Glue

When sarcomeres within a muscle cell contract, what happens to the cell itself? Does the whole cell contract in synchrony with the sarcomeres? If so, does it get pulled away from its neighboring cells, which themselves are also contracting? Does each muscle cell then throb separately? How then do the various muscle cells combine their individual contractions into a forceful combined contraction of the atria or the ventricles?

To provide a strong supporting structure for forceful, combined contraction, nature has invented many glueing mechanisms using recipes encoded in several genes (Fig. 3.7). A muscle cell is riveted to its neighboring cells by rivets called *desmosomes*, made of proteins from 5 genes. Sarcomeres within a muscle cell are connected to these desmosomes and to various parts of the cell by a network of filaments (remember, ropes). Some filaments are thicker (made with *desmin* proteins), others are thinner (made with *actin* proteins). And several other proteins anchor these filaments to a cushioning material called *collagen* that fills the space outside the muscle cells. All this glue then forces the entire collection to contract together as one unified whole, rather than as individual parts, providing enough force to pump blood to all parts of the body.

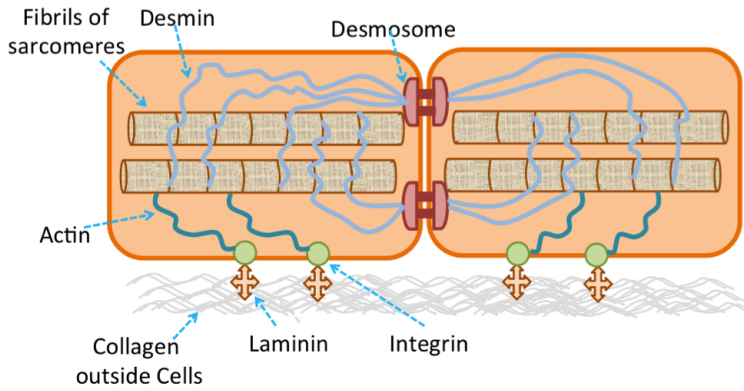


Figure 3.7: Glueing Mechanisms

Just as a building subject to chronic high vibration levels develops cracks unless the pillars are strong, so does the heart structure suffer damage with years of relentless contraction and relaxation, unless the glue is strong. Variants in genes sometimes compromise this glue. For example, variants in the genes which contribute to the rivets (or desmosomes) cause *Arrhythmogenic Right Ventricular Dysplasia (ARVD)*, where heart muscle gets progressively damaged and replaced by fat and fibrous tissue, typically in the right ventricle. This fat and fibrous tissue interferes with the

transmission of heartbeat, causing arrhythmias, and possibly death. More males seem to have ARVD than females.

Could any of these genes involved in glueing mechanisms be the cause of $\mathscr{X}$'s condition? $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{S}$ were all females. But that only makes them slightly less susceptible than males to ARVD. More importantly, structural damage appeared to be restricted to the left ventricle in their case, not quite matching the classic ARVD right ventricle description.

Altered Recipes: Dominant or Recessive?

Did the source of $\mathscr{X}$'s problem lie in the heartbeat (unlikely), the muscle (maybe), or in the glue (maybe)? We cannot make this diagnosis without diving into their genomes. An interesting observation in this regard: unlike what we saw with *ABCA4* in Chapter 2, many variants in genes involved in the heartbeat, the muscles, and the glue, behave in a *dominant* fashion. In other words, a problematic variant present in one of the two such gene copies suffices to cause disease. Problematic variants in both copies of a gene of course lead to more severe disease. This is odd, isn't it? When one copy of *ABCA4* has a problematic variant and is unable to clear up toxic material, the other good copy takes over and does the job. Why does the same thing not happen here?

An analogy might help answer this question. If a building needs 100 pillars for strength, and half of those pillars are faulty, then the building is compromised. The good pillars alone cannot rescue this building. Unless of course the architect decided to put in at least 200 pillars, allowing for half of these to go bad. Another analogy: if a 100 identical strings vibrating in unison produce a certain pleasant note and half of these strings go bad, then the result will be cacophonous, even though 50 of these strings are good. Likewise, a mix of good and bad protein resulting from variants in one gene copy could upset synchrony in the heart. Here is one of the cases from our practice with illustrates this.

A man $\mathscr{N}$ in his 20s was diagnosed with DCM. His left ventricle was enlarged with walls that were thinner then normal. So his heart could pump out only 20-25% of the blood in the left ventricle with every heartbeat. He was advised an ICD implant, much like $\mathscr{X}$. Sequencing his genome

indicated that he carried a variant in one copy of his *TTN* gene (remember the spring in the sarcomere, Fig. 3.5). This resulted in his gene recipe being truncated prematurely after 32,946 amino acids instead of the usual 34,350 amino acids. Only proteins obtained from one of the two gene copies had this problem. Proteins obtained from the other gene copy were normal full-length. There is some evidence to suggest that $\mathscr{N}$'s sarcomeres indeed contained a mix of these truncated and full-length protein copies.[25,26] The full-length protein is able to connect properly to the center line of the sarcomere but the truncated protein is not. A mix of these two protein types then leads to imbalanced springiness of the sarcomere. This, in turn, leads to dilated muscle and DCM. $\mathscr{N}$ was unlucky on this count by just a whisker; had the truncation been 500 or so amino acids further down the recipe, this truncated protein would have had better connection with the center line leading to nearly normal heart muscle.[25] Such truncations in one gene copy would have no impact then; truncations would be needed in both copies to compromise heart function.

In the dominant scenario, suppose a parent has a problematic variant in one gene copy while the other gene copy is normal. Since a child inherits one of the two gene copies of each parent, there is a 50% chance of inheriting this problematic copy. In other words, we expect half of $\mathscr{N}$'s children also to carry the problematic variant (of course, not exactly, but just roughly half). For $\mathscr{N}$, a heartwarming act of fate ensured that neither of his two young children, an infant and a toddler, had inherited this variant. $\mathscr{X}$'s family was less fortunate. It appears as if three of $\mathscr{F}$'s four children, including $\mathscr{X}$, have indeed inherited the problematic gene copy from their father. Which gene is this?

Finding the Culprit Gene

We go back to our shortlist of 200 or so genes, altered recipes in which are known to impact the heart. As in Chapter 2, we look for variants which either truncate the gene recipe prematurely, or modify it so one amino acid is replaced by another (*missense* variants, as these are called). These are the most common recipe alterations. There are a few others we will meet in due course which we also look for but do not describe here. We

also look for variants which are present in both $\mathscr{X}$ and $\mathscr{S}$, given both sisters are affected. Of these, we keep only *rare* variants, i.e., variants that are not commonly found in many people (because such common variants are unlikely to be causative of $\mathscr{X}$'s condition, which is a rare one). But, unlike in Chapter 2, we look for just one gene copy carrying the variant, rather than insisting that both gene copies carry such variants. With this, our shortlist reduces dramatically to the low single digits.

We wade through this shortlist. There is the *OBSCN* gene, which plays a role in the construction of sarcomeres. There are two missense variants in this gene. Both variants are very rare. We look to see if any other patient whom we have tested earlier has any of these variants. One other patient indeed has one of these variants. But this patient has an eye disease with no cardiac complications. The other variant has not been seen in any of our previous patients. But this variant doesn't seem to be at an important location in the gene. How do we know this? We will postpone this description to Chapter 4. For now, the *OBSCN* gene does not appear to be a promising candidate. We move on to the *TTN* gene, which we saw was the cause of DCM in $\mathscr{N}$ above. There are a large number of missense variants in this gene. Not surprising; every person whose genome we have sequenced has many such variants. Remember, this is the largest gene in the human body so a lot of variants are expected in this gene even by sheer chance. Most of these missense variants are not problematic, though we cannot be sure. No variants which truncate the recipe prematurely though, of the type we saw in $\mathscr{N}$. So we keep this aside for now. Then there is the *NOS3* gene, altered recipes in which predispose one to hardening of blood vessels; but this was not the case for $\mathscr{X}$. So variants in this gene are not the likely cause.

Then there is a missense variant in the *LAMA4* gene which looks intriguing. This variant has never been seen earlier in anyone else, to the best of our knowledge. That makes it tough to evaluate whether this variant is indeed the cause of $\mathscr{X}$'s problem. Let us catalog whatever minimal information we know about *LAMA4*. Only two other variants in this gene causing heart disease have ever been published in scientific literature, both in individuals with DCM.[27] One of these is missense, the other is a premature recipe truncation. In both cases, the variant appeared in only

one gene copy and not in both gene copies. Our variant here is a different missense variant also present in only one gene copy. As part of the laminin molecule shown in Fig. 3.7, the *LAMA4* gene anchors the cushioning material collagen outside the muscle cell to the filament network inside the cell, thus contributing to the glue that ensures combined contraction of muscle cells. Is this glue weakened by our missense variant? Hard to say.

This is the challenge of genome sequencing. There are tens of thousands of variants to be assessed. Imagine finding the needle in this haystack in the absence of all-encompassing scientific knowledge! We keep the *TTN* and *LAMA4* variants aside for the moment, making a mental note to come back to them and dig deeper if we cannot find a more promising candidate. We have only one candidate left, which we dig into next, with bated breath.

Missing Characters

The variant in our last gene candidate is unlike any seen in previous chapters. It is not that a character in a healthy genome is replaced by another character in $\mathscr{X}$ and $\mathscr{S}$'s genomes. Rather, a couple of characters present in a healthy genome are completely missing from $\mathscr{X}$ and $\mathscr{S}$'s genomes!

How do we know that these characters are missing? Let us go back to how genome sequencing works (you may remember this from Chapter 1). The genome is 3 billion characters long. When $\mathscr{X}$'s genome is sequenced, what we get in our hands are little snippets of the genome called *reads*, each only a few hundred characters long. This is much like taking a book and tearing it to pieces. We have to put these pieces together to get the book back. This is not easy in itself. So we use the human *reference genome*, the entire genome sequence of a *healthy* individual, as a guide. Since any two human genomes are very similar to each other, $\mathscr{X}$'s genome is very similar to this reference genome. So if we get a read, say , AGGTCCTG, from $\mathscr{X}$'s genome, then this sequence will be present somewhere in the reference genome as well, either in identical or in very slightly altered form. What could this slightly altered form be? It could be AGGCCCTG for instance. Here, a T appears in the read in place of the underlined C

in the reference. Alternatively, it could also be AGGTTCCTG. Here, an extra underlined T appears in the reference. It could even be AGG_CCTG. Here, there is a T missing completely from the reference, indicated by an empty underlined slot. So, we have to hunt for each read in the reference genome sequence, allowing for modifications, additions and removals of characters. To find a stretch of, say, 25 missing characters in $\mathscr{X}$'s genome, our hunt will need to allow for 25 additional characters in the reference genome. Of course, we use computer algorithms to do this. How does the computer know which additional characters it must allow for? Well, it doesn't. It just has to try all possibilities! Would that not take forever, that too with tens of millions of reads to process? With clever algorithms and heuristics, this can be made to run quite fast.

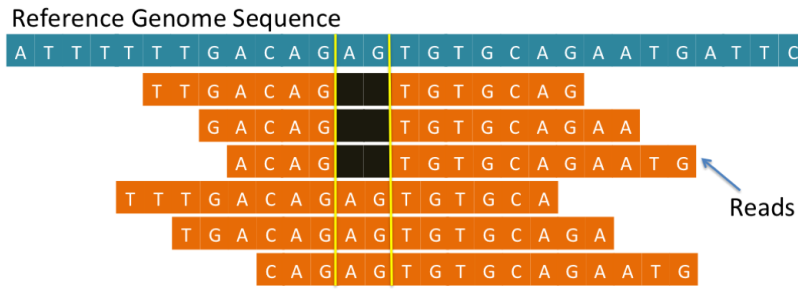


Figure 3.8: A Missing AG in $\mathscr{X}$'s Genome.

Once we do that, we can draw a picture like the one in Fig. 3.8 where each read is placed at its corresponding location in the reference genome, with added, removed or modified characters marked. One look at this picture shows that $\mathscr{X}$ has the characters AG missing from half the reads, i.e., from one copy of her genome!

Out of Frame

Both $\mathscr{X}$ and $\mathscr{S}$ have the characters AG missing from a gene called *DSP*. How does this variant impact *DSP*'s recipe? Remember how this recipe is described in the *genetic code*, described in Chapter 2? First, imagine

stringing the exons of this gene together, excising out the intervening introns. Take the resulting sequence of characters and divide these into blocks (or *frames*) of 3 characters each. Each character triplet specifies an *amino acid*. The *DSP* gene's recipe has 2871 such triplets, which together specify a protein with 2871 amino acids. What happens to these amino acids when the gene loses two consecutive characters? Fig. 3.9 illustrates this.

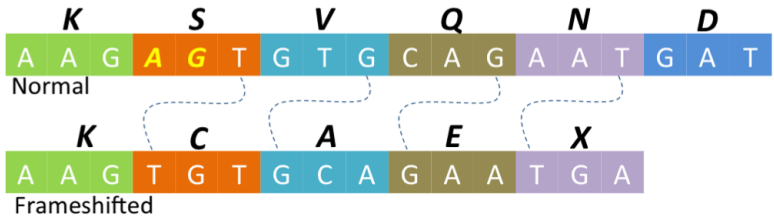


Figure 3.9: The Missing AG causes a Frameshift

Compare the two scenarios above. The first is the normal gene. The second is the gene with the characters AG (marked in yellow italics) removed, as in $\mathscr{X}$ and $\mathscr{S}$. The frames in the latter scenario are shifted compared to frames in former scenario, as indicated by the dotted lines (and hence this is called a *frameshift*). Consequently, the amino acids generated are completely different, starting from amino acid 711 (the amino acids for each frame are indicated in bold above the respective frame, based on Fig. 2.5). The recipe has changed completely!

Then something even more dramatic happens. After amino acid 713, the recipe encounters a TGA triplet, which represents a *Stop* instruction marked by an *X*. That is the end of the recipe. The protein now has only 713 amino acids as opposed to 2871, a fraction of its normal self!

$\mathscr{X}$ and $\mathscr{S}$ have two characters missing from the genome. What if they had three missing characters? Say, triplet AGT was missing in its entirety? Then the amino acid Serine (S) would no longer be present in the protein. All amino acids to the right of this Serine would be intact though. And unless this Serine was particularly important, $\mathscr{X}$ and $\mathscr{S}$ would not have felt the impact of this change at all. Problematic frameshifts happen

when the length of the missing segment is not a multiple of 3. Usually, a frameshifted recipe encounters a *Stop* instruction very soon and terminates prematurely, as it did for $\mathscr{X}$ and $\mathscr{S}$. So was this premature truncation of *DSP* gene indeed the cause of their problem?

The *DSP* Gene and *Desmosomes*

What role does *DSP* play? Remember, desmosomes rivet together neighboring heart muscle cells (Fig. 3.7). And desmin filaments connect sarcomeres to the desmosomes and to other parts of the cell. Together, these provide the glue needed for muscle cells to contract in unison. Fig. 3.10 shows a desmosome.

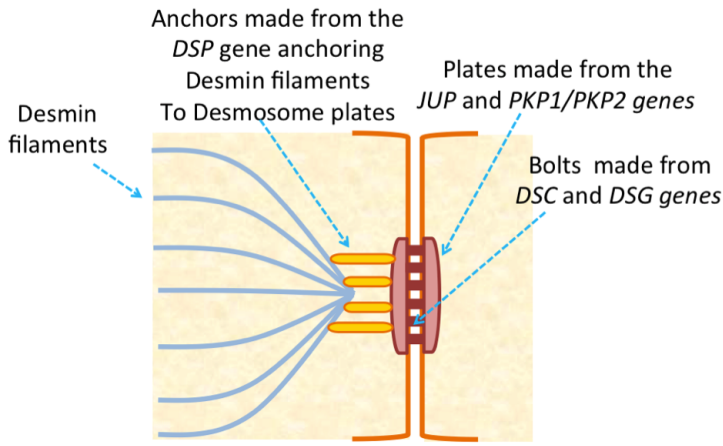


Figure 3.10: *DSP* and a Desmosome riveting cells together.

A desmosome comprises two plates, one inside each of the cells. The two plates are connected together by bolts. The plates and the bolts are both made from many proteins. *DSP* actually anchors desmin filaments to these desmosome plates. So *DSP* is an essential ingredient of the glue. If this glue is weakened, then the ability of cells to withstand mechanical stresses may be affected, resulting in increased wear-and-tear over time.

Does the truncated recipe that $\mathscr{X}$ carries in one copy of *DSP* indeed compromise this glue?

No Nonsense

The severely truncated *DSP* in $\mathscr{X}$ and $\mathscr{S}$ is about a quarter of it normal length. Chances are that it will not serve as a strong anchor between the desmin filaments and the desmosome plates. So it would not be wise for cells to use this truncated molecule. Indeed, cells have mechanisms by which they can recognize and destroy proteins generated from truncated recipes (also called *nonsense* recipes). But that is magical! How would a cell know that a recipe has been truncated? After all, every recipe has a *Stop* at its end. Imagine a cell as it interprets the recipe for a particular gene. It encounters a *Stop*. Is this the normal intended *Stop*, or is it an abnormal variant *Stop*? How would the cell differentiate between these two scenarios?

Here's how. Remember, the *DSP* gene recipe is encoded in the exons of the *DSP* gene. Exons are separated by intervening introns. Introns do not carry the recipe, only exons do. A gene could have many exons (see Fig. 3.11). To the left of the first exon of this gene is another stretch of text, called *UTR* (short for untranslated region, i.e., no recipe interpretation happens in this region). Similarly, to the right of the last exon for this gene lies another stretch of text, also called *UTR*. Recipe interpretation now happens in two steps.

In the first step, the entire stretch of the genome from the beginning for the first UTR to the end of the last UTR is copied. This copy contains the two UTRs, all the exons and all the intervening introns. Special-purpose molecules then excise out the introns so the exons now appear contiguously. These molecules are diligent though, so they leave markers at the ends of each exon to let everyone know where the exon boundaries were before they removed the introns. This template shown in Fig. 3.11.

In the second step, this template is scanned from left to right. Recipe interpretation begins when the *Start* triplet *ATG* is seen. Then every successive triplet is converted to its corresponding amino acid, based on the code in Fig. 2.5. As soon as an exon boundary is crossed in the process,

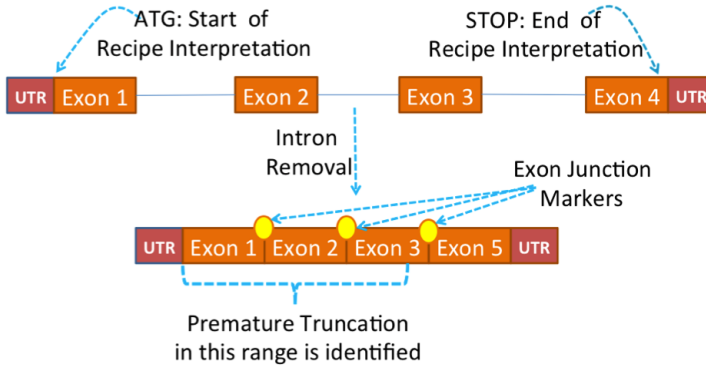


Figure 3.11: Spotting Premature Truncations: Nonsense-mediated Decay.

the marker at that boundary is removed. Recipe interpretation stops when a *Stop* triplet is encountered. So far so good. But how does the cell know whether this is the normal intended *Stop*, or whether it is an abnormal variant?

A simple check is performed now. If no exon boundary markers are left when the *Stop* is encountered, then all is well. On the other hand, if an exon boundary marker is still there when a *Stop* is encountered, then this *Stop* is likely to be a premature; the entire template is then destroyed so no recipe interpretation can happen from it. This is called *Nonsense-mediated Decay*.

Of course, not all premature truncations can be identified this way. For instance, truncations within the last exon are clearly spared. In reality, truncations close to the boundary between the last and the penultimate exon, i.e., within about 50 characters, are also spared. The average length of an exon is only around 150 characters; this means that only truncations in the last 200 or so characters in the gene recipe are spared. Other premature truncations are identified and destroyed.

However, as with several processes in biology, this process is not perfect. There are rare exceptions, i.e., truncated proteins that escape nonsense-mediated decay. We saw such an example of a *TTN* truncation

in the context of $\mathscr{N}$'s case above. So this leaves two scenarios for $\mathscr{X}$'s premature *DSP* truncation. Chances are that $\mathscr{X}$'s truncated *DSP* protein is destroyed by nonsense-mediated decay. Then all the protein copies that remain are good. But the total number of copies is half of what it normally is. On the other hand, there is a chance that the truncated *DSP* protein escapes nonsense-mediated decay. Then the total number of *DSP* protein copies remains unchanged. But half of these copies are truncated and cannot provide strong anchor. Which of these situations poses the greater danger?

Here is a simplistic argument to show that the latter situation poses the greater danger. Suppose 100 good *DSP* protein copies are needing to provide strong enough glue. And a normal individual actually generates (say) x copies, where x is 100 or more (any leftover copies may be disposed off). In an individual with a *DSP* truncation, there are $x/2$ good copies and $x/2$ truncated copies. If the truncated copies are destroyed by nonsense-mediated decay then problems set in only if $x < 200$. But if the truncated copies are not destroyed by nonsense-mediated decay then problems set in even $x \geq 200$. Why? For instance, suppose x were 400. Then there would be 200 good copies and 200 truncated copies. These would compete against each other for the 100 slots available. These 100 slots would then be filled in roughly equal proportion. So half the slots would get truncated copies, compromising the glue considerably. So the best case is that all truncated *DSP* protein copies are destroyed leaving only good protein copies, but at half the normal amount. Could this best case itself be not good enough for $\mathscr{X}$'s heart?

Of Mice..

Maybe half the normal amount of *DSP* protein is indeed sufficient and it is business as usual. Or maybe cells sense that only half the amount is available; they then kick recipe interpretation into overdrive to create more protein from the good gene copy. Or, maybe other genes kick in to do what *DSP* ought to do. All valid possibilities. Or, maybe half the normal amount of *DSP* protein just does not provide enough desmosomal riveting between cells to resist wear-and-tear, placing $\mathscr{X}$ at risk for sudden cardiac

death. Which of these is the case? Answers to this question emerge from somewhat unexpected quarters: from experiments on other animals!

Mice can be created in a laboratory with a specific gene *knocked-out*, i.e., their genomes can be so engineered that the recipe for this specific gene is disrupted substantially and no protein can be created based on this recipe. When both copies of *DSP* were knocked-out, mice died while they were still tiny embryos.[28] So the complete lack of *DSP* is a non-starter. Of course, desmin filaments could not connect to desmosome plates in these mice at all. But, more surprisingly, the number of desmosomes itself was far fewer than normal. Mice can also be engineered so that the *DSP* knock-out effect happens only in the heart cells; in all other cells, *DSP* continues as always. Such mice did survive as embryos. But they died within a couple of weeks of birth.[29] Heart muscle cells were clearly disorganized in these mice without the structure provided by *DSP*. All this when both copies of *DSP* are knocked-out.

What happens when only one *DSP* copy is knocked-out, mimicking $\mathscr{X}$'s condition? These mice were born normally. So maybe one copy of *DSP* suffices. But, wait! 20% of these mice died within 6 months of birth![29] The typical lifespan of a mouse is about 2 years, so 20% of these mice lived less than a quarter of that span. There were fewer connexins, leading to electrical signal being transmitted slower than normal through the muscle cells.[30] In spite of this, there were no structural changes in the heart and no abnormal rhythms visible, initially, when the mice were 2 months old. Apparently, the rate of spread of electrical signal did not decrease enough to be bothersome under normal conditions. However, it did decrease enough to be bothersome when stressed by exercise, leading to arrhythmia. Over time, structural and rhythmic changes become more overt. The right and left ventricle chambers became enlarged. The left ventricle walls became thinner. As a combined effect, the ability to pump out all the blood in the ventricular chamber with each heartbeat was poorer. Heartbeat rates were abnormally fast. Not very different from what was observed in $\mathscr{X}$ and $\mathscr{S}$.

There was more in store when hearts of mice with only one copy of *DSP* were observed under a microscope. Muscle cells in these mice were not organized in a normal way. There was an excess of fat among these

muscle cells.[29] There was also evidence of abnormal levels of cell death. And an excess of fibrous tissue, similar to what one sees in scars which form when cells die. These scars possibly interrupt the spread of electrical signal from cell to cell, thus leading to arrhythmia.

So there is convincing evidence that just half the normal amount of *DSP* protein is insufficient for a strong glue effect. However, this is evidence in mice. Does that translate to humans? After all, humans are not mice.

And Men..

Mice and humans do look very different to the untrained eye. Yet, we are close(r) relatives in a genomic sense. For several genes, our genomic text stretches actually resemble theirs! Several genes whose knock-outs cause early death in mice also cause disease in humans. So, it is reasonable to surmise that premature truncation of one copy of the *DSP* gene is indeed the source of $\mathscr{X}$'s problems. Substantiating evidence in humans can add to our conviction here. So we search scientific literature for patients with heart complications and with one copy of *DSP* prematurely truncated. A number of such cases appear in literature (for instance,[31] itself reports 7 such individuals;[32] reports 2 more such individuals).

But there is also a twist! Take one example: a 49 year old woman with serious heart complications (abnormal fast heartbeat and thinning of ventricular walls) has a variant that truncates one copy of *DSP* after 1269 amino acids.[33] Her brother and son too have the same variant. Their complications are milder, almost borderline though. So far so good. Then the woman's 85 year old father too has the variant; but he shows no signs of any expected heart complications! There are other such examples as well. For instance, a repository with exomes of 6502 presumed healthy individuals has two individuals with premature truncations in *DSP* at amino acids 1394 and 2334 respectively; both truncations are in only one gene copy. To summarize, it appears as if not everyone with one truncated copy of *DSP* suffers from what we see in $\mathscr{X}$. Are we on the wrong track completely? Maybe this *DSP* variant is not the cause of $\mathscr{X}$'s problems at all.

Let's do a simple calculation to take stock. Among these 6502 individuals mentioned above, only 2 have a premature *DSP* truncation. We also know that these 6502 individuals are all unrelated. So among unrelated healthy individuals, premature *DSP* truncations appear with a frequency of roughly 1 in 3000. Now take figures published by two studies on patients with heart disease. The first study[31] screened 100 families with ARVD for variants in the five genes whose recipes code for components of the desmosome; they found 7 premature *DSP* truncations. The second study[32] screened 121 DCM patients for variants in 46 genes and found 2 individuals with premature *DSP* truncations. So among unrelated individuals with either ARVD or DCM, premature *DSP* truncations appear with a frequency of roughly 9 of 221. There are a few nuances here, but in broad strokes, it appears that premature truncations in *DSP* are 100 times more likely to be present in individuals with ARVD or DCM than in healthy individuals!

A 100-fold greater likelihood along with behavior in mice described above makes a convincing case for the premature truncation in one copy of *DSP* being the cause of $\mathscr{X}$'s and $\mathscr{S}$'s condition. But, no doubt, the effect of premature truncation in one copy is not fully consistent. Not fully *penetrant* is the right term, i.e., not everyone with the variant gets the disease. There is not enough data to estimate what fraction of individuals with a single truncated copy of *DSP* experience disease symptoms. However, there are consolidated estimates over all known problematic single gene copy variants in all genes associated with desmosomes; about 35% of the individuals with such variants do show signs of disease by the time they are 40, and 60% show signs of disease in their lifetime.

Why do some individuals with a single copy truncation show heart disease while others don't? Possibly, some truncations escape nonsense-mediated decay while the others don't. And those which do lead to more severe disease. There is at least one example of a single copy *DSP* truncation at amino acid 586 that escapes nonsense-mediated decay found in 12 patients with heart disease, all from the same family;[34] $\mathscr{X}$'s truncation at amino acid 713, is not far away. Maybe proteins created from this truncated recipe too escape nonsense-mediated decay, at least to some extent, and compromise the glue.

Even in the above family, there appear to be four individuals who likely

have the same *DSP* truncation at amino acid 586 as the 12 patients, but whose hearts function normally.[34] What causes this variation within the same family? Maybe half the normal amount of *DSP* protein pushes the heart close to the edge in its ability to withstand mechanical stresses, but does not quite tip it over fully. Other unknown factors that vary from person to person, might tip it over. One such factor could be additional genomic variants in other heart-related genes. There are indeed reports of single copy variants in two distinct desmosome genes increasing penetrance 5-fold.[31] Remember the *TTN* and *LAMA4* gene variants we kept aside earlier, with a note to get back if the need so arises. Could one or more of these compound the effect of the premature *DSP* truncation? We just don't know.

Wrapping Up

When an illness of mysterious cause runs in a family, it leaves a sense of acute helplessness, a sense of not knowing whose turn it will be next and what they did wrong to be subject to this fate. Against this backdrop, a diagnosis is a relief. $\mathscr{X}$ and her family, through their difficult times, now have the consolation of knowing that premature truncation of the *DSP* gene is the likely cause of their affliction.

A *DSP* variant would conventionally lead to a diagnosis of *ARVD* or *Arrhythmogenic Right Ventricle Dysplasia*, characterized by abnormalities of the right ventricle primarily. Right ventricle muscle is progressively replaced with fat and fibrous tissue in *ARVD*. The muscle wall thins and the ventricle cavity becomes larger. This leads to arrhythmia and possible sudden death. The left ventricle is also affected but only as disease progresses. But $\mathscr{X}$'s condition was primarily left ventricular! Isn't that unexpected?

Interestingly, there is at least one reference in literature to a truncating variant in *DSP* which affects the left ventricle primarily.[34] And $\mathscr{X}$'s truncated *DSP* is 713 amino acids long, somewhat similar to the truncation after 586 amino acids reported in that case. The left ventricle is more muscular than the right but experiences much greater mechanical stress; so we could speculate that serious weakening of the glue is most exposed

in the left ventricle, while more moderate weakening would first appear on the right and then progress to the left. Thus, the right diagnosis for $\mathscr{X}$'s condition would be *ALVD*, or *Arrhythmogenic Left Ventricle Dysplasia*.[34]

X has an an ICD implanted to counter arrhythmia. She has another implanted device (a *LVAD* or *Left Ventricle Assist Device*) assist her left ventricle in the task of pumping blood to the rest of the body. She awaits a heart transplant. That is the only cure known. Knowledge of the *DSP* variant does not offer her a cure today. Gene therapy of the type we saw in Chapter 2 was feasible in the eye, but not yet an option for the heart.

$\mathscr{X}$'s sister $\mathscr{S}$ shows abnormal heart rhythms as well and has an ICD implanted too. She does not have heart failure, i.e., her left ventricle is not heavily compromised, so she has no overt symptoms. These could develop over time, so she needs to be monitored closely. It is known that stress-inducing exercise advances the age at which symptoms manifest in carriers of desmosomal variants, so $\mathscr{S}$ may need to restrict exercise.[35]

$\mathscr{X}$'s brother $\mathscr{B}$, and $\mathscr{S}$'s young children show no abnormalities today. We do not know yet if they carry the *DSP* variant. If they indeed do and if behavior in mice is any indication, they will be at increased risk for cardiac arrhythmia (when stressed) even before abnormal structural or heartbeat changes set in, so monitoring them regularly will be important.[30] On the other hand, if they do not carry the *DSP* variant, then all cause for concern goes away. So they have the option of getting tested for this variant via a simple and inexpensive test.

Our fundamental inability to repair damaged hearts to cure people like $\mathscr{X}$ remains though. Is there reason to hope that we might be able to do this some day? Indeed, we have a source of inspiration: unlike human hearts, zebrafish hearts self-repair after an injury. Research into genes which enable this recovery might enable us to replicate this system in humans some day.[36] So, some day, we just might be able to repair hearts with *DSP* variants, if we could learn from this little fish.

Chapter 4

The Mystery of the Eyes

$\mathscr{M}$ and $\mathscr{F}$ were healthy young parents, both in their twenties. Their second child $\mathscr{X}$, a boy, had an uneventful birth. However, repeated vomiting after birth triggered some concerns. Investigations revealed thickening of muscles at the junction of the stomach and the small intestine. These thicker muscles blocked the regular passage of food leading to forceful vomiting. By itself, this was not an uncommon condition; 3 in a 1000 babies are so affected. A surgical procedure was performed to cut through and relax these thickened muscles, thus opening the blocked passage. Vomiting stopped as a result. But investigations revealed additional complications.

$\mathscr{X}$ had *hypertension* or high blood pressure. Blood pressure is the pressure that blood exerts on the blood vessel walls. Increased pressure causes damage to these walls and to other organs. The root cause of this high blood pressure remained hard to pinpoint. The kidneys control blood pressure via some enzymes and problems in the kidney could upset this mechanism leading to hypertension; however, a kidney scan on $\mathscr{X}$ showed no abnormalities. Defects in heart structure could also cause hypertension; again, a scan of $\mathscr{X}$'s heart showed no abnormalities. Defects in hormones produced by various glands could also cause hypertension; yet again, blood tests to check for these showed no abnormalities. Usually, hypertension develops over the years. Its presence in babies is rarer, but again, not so uncommon, about 1-2%. $\mathscr{X}$'s hypertension was of a even rarer form, with high pressure present in both the blood vessels that carry blood from the heart to most of the body (*systemic hypertension*), as well as the blood vessels that carry blood from the heart to the lung (*pulmonary*

hypertension). Since unchecked hypertension could cause organ damage, $\mathscr{X}$ was treated with blood pressure medication.

$\mathscr{X}$'s complications did not end there. At 10 months of age, $\mathscr{X}$ got a lung infection (a *pneumonia* attack). He recovered from this. And then, things took a turn for the worse. When he was 17 months old, he developed breathing difficulties and passed away suddenly. This was, of course, a tragic fate for $\mathscr{M}$ and $\mathscr{F}$. Even more so that their first child $\mathscr{Y}$, also a boy, has met an identical fate, passing away due to breathing difficulties when he was 13 months old.

The loss of their two children, no doubt, left them distraught. Their childrens' symptoms did not fit a textbook description, making diagnosis difficult even for the best of geneticists. This inability of medicine to explain why their children were at the receiving end when so many other children around them were born normally and lived healthy lives left them helpless. There was no record of any such early deaths in the family. They were still young enough to have more children. But would they be able to bear this emotional trauma yet again?

$\mathscr{X}$ and $\mathscr{Y}$ both had another strikingly unusual feature which we haven't talked about yet; can you spot this feature in in Fig. 4.1?

Figure 4.1: Prominent Eyes in $\mathscr{X}$ and $\mathscr{Y}$

See the very prominent, bulging eyes. Aren't they unusually prominent? Unusual, yes, but not necessarily alarming by themselves. But, in the context of the other abnormalities, could these hold a clue to the root cause?

Where Do We Look?

Genome sequencing provides a way forward for $\mathscr{M}$ and $\mathscr{F}$. Since both boys seemed to suffer from similar complications, a shared genomic character is the likely root cause of their complications. Unfortunately, $\mathscr{Y}$ had passed away earlier and we had no access to his genome. But we did have access to the genomes of $\mathscr{M}$, $\mathscr{F}$, and $\mathscr{X}$. Not having $\mathscr{Y}$'s genome makes the job harder but not impossible. How do we identify this offending character by sequencing just the genomes of $\mathscr{M}$, $\mathscr{F}$, and $\mathscr{X}$?

As in previous chapters, our window into the genomes of $\mathscr{M}$, $\mathscr{F}$, and $\mathscr{X}$ is *Whole Exome Sequencing*. Remember, there are about 21,000 *genes* in our genome. Each gene is a stretch of text in the genome comprising exons and intervening introns (see Fig. 1.12). Only exons carry recipes for creation of protein molecules, introns don't. There are roughly 200,000 exons over all 21,000 genes. The genomic text stretches in these 200,000 exons together comprise what is called the *exome*. The exome measures just over 1% of the entire genome. Whole exome sequencing followed by a lot of data crunching on a computer gets us the exomes of $\mathscr{M}$, $\mathscr{F}$, and $\mathscr{X}$. In particular, it identifies every character in the exome where $\mathscr{M}$, $\mathscr{F}$, and $\mathscr{X}$ differ from a supposedly healthy genome sequence. These characters are called *variants*. The main problem again: there are tens of thousands of these variants! How do we identify which of these is the culprit?

Since both $\mathscr{X}$ and $\mathscr{Y}$ suffered similar fates, it is likely that they both inherited the same offending variant from their parents. We need to be a bit careful in jumping to this conclusion though, because not all genomic characters are inherited from parents. Each of us do have a few *de novo* variants; these appear in a person's genome when they are conceived rather than being inherited from that person's parents. De novo variants are relatively rare, each of us may have few tens of these in our genome. The chances of de novo variants appearing independently in both $\mathscr{X}$ and in $\mathscr{Y}$ and causing similar effects would be very very low. So chances are that they both inherited the same offending variant from their parents. If $\mathscr{X}$ and $\mathscr{Y}$ both inherited this offending variant from their parents, how is it that the parents remained healthy? Why did this variant manifest itself only in the boys but not in their parents, even though the parents too have

this character?

Recall from Fig.1.9 that the genome is divided into chapters called *chromosomes*. There are 46 such chromosomes. 44 of these actually occur as 22 pairs, each pair carrying 2 copies of the same chromosome. One copy in each pair is inherited from the mother, and the other from the father. Chromosome 23 is called the *sex* chromosome, and comes in two flavors: X and Y. Females have two copies of the X chromosome, one copy coming from each parent. Males have one X chromosome and one Y chromosome; the X comes from the mother and Y comes from the father.

Suppose the offending variant was present on a chromosome other than the sex chromosomes. Then, it is possible that each parent had that variant in just one chromosome copy, but not in both. Since a child inherits one of the two chromosome copies from each parent, there would be a 50% chance of inheriting the offending variant from the mother, and a 50% chance of inheriting the offending variant from the father. This gives a combined 25% chance that the child gets two copies of the offending variant. And maybe two copies of the variant are needed for it to cause abnormalities, while a single copy would be both harmless and silent. This would be one explanation for abnormalities in both boys, while the parents remain healthy. The chances of this happening in both boys, each at 25%, is 1 in 16, or 6.25%. Small, but not improbable.

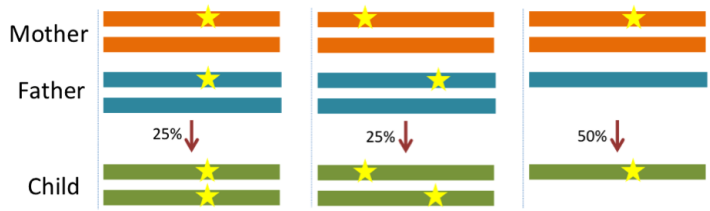


Figure 4.2: Various possible Inheritance Scenarios; shown are chromosomes for the mother, father and single male child. The last scenario involves the X chromosome

What other scenarios are possible? It could be that there are two

distinct variants, one in $\mathscr{M}$, and the other in $\mathscr{F}$. Each variant occurs in only one chromosome copy in the respective individual. Both variants appear within the same gene though (remember this is exactly what happened in Chapter 2). Again, each boy inherits both of these offending characters. The first character would render one copy of the gene incapable in some way, and the second character would render the other copy incapable. Between them, both copies of the gene would be rendered incapable, leading to manifestation of abnormalities. Each parent would have at least one good copy of the gene, and hence would remain healthy. The chances of this happening in both boys, each at 25%, is again 1 in 16, or 6.25%.

There is one more scenario. What if the offending variant is on the X chromosome? Both boys have only one copy each of the X chromosome, and they get this copy from their mother $\mathscr{M}$. $\mathscr{F}$ has no role here. $\mathscr{M}$ herself has two copies of the X chromosome. She may have the offending variant in only one of her X chromosome copies. And both boys inherit this offending variant. The chances of this happening in both boys, each at 50%, is 1 in 4, or 25%. Not that small at all! Of all the above possibilities, this one has the highest odds.

We do not know which of these scenarios holds for our family. So we keep all options open and look for variants which satisfy any of the above scenarios. Note that these are all *recessive* scenarios, i.e., one good copy of the gene suffices for good health; contrast with Chapter 3 where our quest was focussed on the *dominant* scenario). But which genes do we focus on?

In Chapter 2 and Chapter 3, the impact of the gene of interest was contained to one organ (the eye and the heart, respectively). The impact here is broader. The are changes in facial structure (prominent, bulging eyes), possibly in the blood vessels (hypertension), possibly in the development of the lungs (respiratory difficulties), and possibly in the stomach/intestine (thickening of the stomach muscle). The object of our hunt seems to have its tentacles spread far and wide. So we consider all genes, variants in which are known to directly cause abnormalities in one or more of these organ systems. This list currently comprises 500 or so genes. Still too many genes to sift through.

As in previous chapters, we look for variants which either truncate

the gene recipe prematurely, or modify it so one amino acid is replaced by another (*missense* variants, as these are called). These are the most common recipe alterations. There are a few others we will meet in due course which we also look for but do not describe here. Of these, we keep only *rare* variants, i.e., variants that are not commonly found in many people (because such common variants are unlikely to be causative of $\mathscr{X}$'s condition, a very rare condition). It so happens that $\mathscr{F}$ and $\mathscr{M}$ are related (or *consanguineous*); they are actually uncle and niece, respectively! If they were unrelated, then chances of the same rare variant appearing in both $\mathscr{F}$ and $\mathscr{M}$ would be very low. But consanguinity increases the chances of the same rare variant being present in both individuals.

We instruct our computer to wade through the tens of thousands of variants in $\mathscr{X}$, looking for variants which satisfy all the above constraints. And the computer obliges by narrowing our hunt down to a handful. Quite a dramatic reduction! Primarily because we are considering a recessive scenario and because we are able to compare variants in $\mathscr{X}$ with those in the parents. If we had been considering the dominant scenario instead, we would be looking for variants affecting just one copy of a gene rather than both copies; this would have yielded many more candidates. And if we did not have access to the genomes of $\mathscr{X}$'s parents, we would not be able to eliminate several variants which were present in both gene copies in one of the parents as well.

As we pore over our handful of candidates, we eliminate all but just two for various technical reasons of no great importance. And that leaves us with just two genes. Meet *Fibulin 5* or *FBLN5*, and *Filamin A* or *FLNA*. What are the genes and which, if any, of these presents sufficient evidence for prosecution in a court of law (figuratively speaking, of course)?

FBLN5 and False Starts

Scientific literature does mention variants in the *FBLN5* gene causing pulmonary hypertension, repeated respiratory infections, and death due to respiratory failure in the first two years of life.[37] These were all features which $\mathscr{X}$ shared, making us sit up and take note. But the geneticist who had treated $\mathscr{X}$ would have none of it. Just the mere mention of the *FBLN5*

gene met with categorical disapproval: ”certainly not FBLN5, these boys had no *Cutis Laxa*”, she said. Indeed, cutis laxa (which translates to loose skin) is the hallmark of variants in *FBLN5*. Typically, our skin is elastic; you pinch it, and it pulls right back in place. This elasticity is lost in Cutis laxa; the skin is loose, like saggy cloth. But both $\mathscr{X}$ and $\mathscr{Y}$ had normal, elastic skin. Why is it that $\mathscr{X}$'s *FBLN5* variant did not lead to Cutis laxa?

Remember Fig. 2.5? The recipe in the *FBLN5* gene in coded in a series of character triplets. Each triplet codes for a particular amino acid. When the recipe carried by this gene is interpreted, each triplet is read in sequence, the corresponding amino acid is taken, and the resulting amino acids are strung together to create the object of this recipe, namely the *FBLN5* protein. The first triplet, also called the *Start*, is represented by the triplet ATG. Recipe interpretation begins at this triplet, and continues until a *Stop* triplet is reached (which is one of TGA, TAA, or TAG). In $\mathscr{X}$, the start triplet ATG became a GTG, in both gene copies, as in Fig. 4.3.

Figure 4.3: The *Start* Triplet Changes in *FBLN5*

Both parents had this change in only one gene copy each. Neither parent would be affected because recipe interpretation would proceed normally from the other copy. It is well known that one good gene copy of *FBLN5* is sufficient for healthy behavior, unless the other copy becomes defective in a specific way that interferes with the functioning of this good copy. What impact would this have on $\mathscr{X}$? Recipe interpretation may not know where to begin anymore. Could this lead to complete inability to create the *FBLN5* protein in $\mathscr{X}$? If that were the case, then $\mathscr{X}$ should have had loose skin, which he clearly didn't. How does one explain this mystery?

A small amount of recipe interpretation might start from this GTG;.[38]

but likely small enough to still cause loose skin. More likely, recipe interpretation picks up from an alternative start point, if one exists further along the recipe. Could the very next ATG in this recipe serve as this start point? Not really! This ATG must have a certain pattern of characters in its neighborhood for it to be considered a likely start point. This pattern of characters is called the *Kozak* sequence and is illustrated in the picture below.

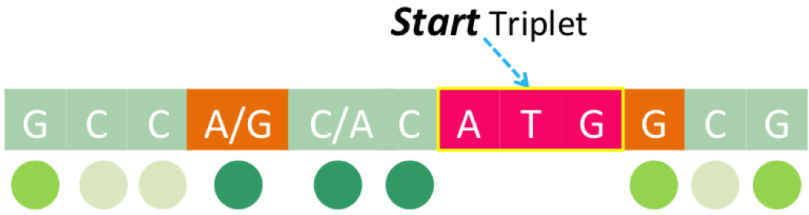


Figure 4.4: The Kozak Pattern: The start triplet ATG is shown in red. The orange characters denote important locations.

The Kozak pattern is not an absolute rule though, i.e., the start ATG triplet of every gene has a neighborhood that looks approximately, but not exactly, like the Kozak sequence. Of these, a match at the orange locations is more important. In particular, the first orange location, specified to be either A or G, has a dominating effect. A match at this location allows some lack of conformity at the second orange location. The more the neighborhood of the ATG triplet matches the Kozak sequence, the greater the chance of recipe interpretation starting from this triplet. Of course, there could be several ATG triplets all vying to be the start points for recipe interpretation; the ones with stronger Kozak matches stand a better chance in this competition.

For *FBLN5*, there is another ATG triplet that appears 168 characters down into the recipe from the original start ATG triplet (which has become GTG in $\mathscr{X}$). The neighborhood of this downstream triple shows modest conformity to the Kozak pattern. The colored dots underneath in Fig. 4.4 indicate the strength of match for each character in the Kozak pattern;

darker the color, greater the match strength. As you can see, the most important orange location and a few others have very good match strengths. Other locations have modest to poor match strengths. Several methods to quantify the goodness of this match have been invented by researchers. According to one of these,[39] this match is better than the matches that start triplets of 25% of the genes have. And only 17% of ATG triplets known to NOT be valid start triplets have better matches. So there is a modest possibility that recipe interpretation picks up from the ATG triplet 168 characters into the recipe, yielding a protein that is shorter by $168/3 = 56$ amino acids. What impact could the loss of these 56 amino acids have in $\mathscr{X}$? Could it explain X's condition and simultaneously also explain the lack of loose skin?

Fig. 4.5 shows the new start at amino acid 57 and the stretch of the *FBLN5* protein where variants known to cause Cutis laxa is various patients lie. The number of patients known with Cutis laxa on account of variants in *FBLN5* is still small, so more variants in other parts of the protein may be discovered over time. But, for now, it appears as if all variants causing Cutis laxa lie deeper into the protein, in parts which would be well-preserved even when the first 56 amino acids are lost in $\mathscr{X}$. So, maybe, recipe interpretation does occur from this new start point in $\mathscr{X}$ and the parts of the protein responsible for skin elasticity are preserved as a result; therefore, no Cutis laxa in $\mathscr{X}$. So far so good. But could the loss of the first 56 amino acids somehow cause hypertension and respiratory distress in $\mathscr{X}$, while steering clear of Cutis laxa?

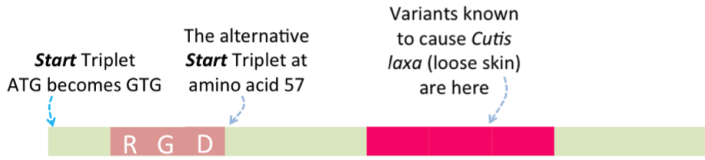


Figure 4.5: The *FBLN5* protein

A brief digression first. The elastic nature of skin comes from elastic fibers comprising several proteins. These fibers are found in the space

outside the cells in the skin. They are found in several other organs as well, e.g., blood vessels and the lung, and maintain the elasticity of these organs. *FBLN5* is not a part of these elastic fibers per se, but it plays a role in organizing these fibers. We know this because mice in which *FBLN5* has been knocked out (remember, mice can be created with specific genes removed from their genomes completely; we saw this in Chapter 3) have disorganized elastic fibers leading to inelastic skin (Cutis laxa), and complications in the lung and in the blood vessels.[40] Experiments have shown how *FBLN5* might potentially help organize these elastic fibers. The red parts of *FBLN5* attach to these fibers while a part marked by the three amino acids *R-G-D* attaches to proteins called *integrins* which are connected to the cells themselves.[40] *FBLN5* thus helps anchor these elastic fibers to cells, as shown in Fig. 4.6.

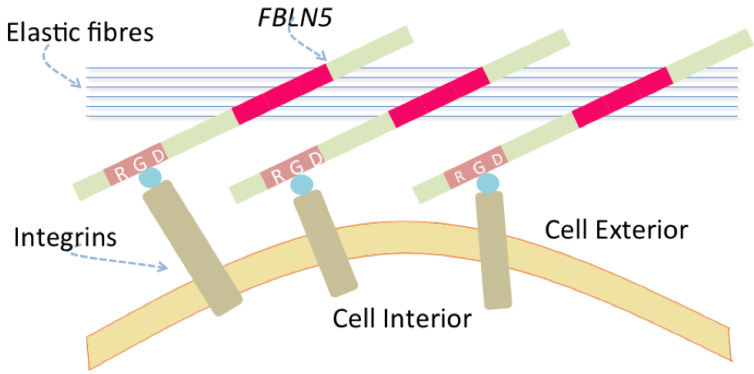


Figure 4.6: Elastic Fibers and their organization by *FBLN5*

Coming back to point: could the loss of the first 56 amino acids somehow cause hypertension and respiratory distress in $\mathscr{X}$? The red parts are, of course, unaffected. The R-G-D part that attaches to integrins comprises amino acids 54 (R, short for *Arginine*), 55 (G, short for *Glycine*) and 56 (D, short for *Aspartic acid*). These, of course, are lost. The loss of these R-G-D amino acids no longer allows elastic fibers to be anchored to cells. It would be reasonable to hypothesize that this reduces elasticity.

For blood vessels, it might mean less elastic, more stiffer walls, leading to increased blood pressure. For the lungs, it might mean that the walls of the air sacs lose elasticity making it difficult for them to expand and contract fluently in synchrony with the breathing. For the skin, this would mean Cutis laxa, but maybe, just maybe, the retention of the red parts ensures that the reduction of elasticity in the skin is not significant enough to cause Cutis laxa. Could the loss of R-G-D lead to reduction of elasticity in the blood vessels and in the lungs, but not in the skin?

One way, and possibly the only way, to answer this question definitively is by experiment. With bated breath, we search for experiments which will confirm the preferential reduction in elasticity in the blood vessels and in the lungs, relative to the skin, due to removal of the R-G-D amino acids. Our quest leads to a relevant experiment, where researchers created mice with the R-G-D replaced instead by R-G-E in both gene copies (i.e., the *Aspartic Acid* amino acid represented by character D was replaced by *Glutamic acid*, represented by the character E). This is not quite the same as removing the R-G-D, but it mimics that effect because R-G-E, unlike R-G-D, cannot attach to integrins and anchor elastic fibers to cells. Do these mice show reduction of elasticity in the blood vessels and in the lungs, but not in the skin?

The answer is disappointing. These R-G-E mice show no abnormalities, not in the lungs, not in the blood vessels and not in the skin.[41] We don't know why; possibly, other genes take over where *FBLN5* fails, a patent reminder of the redundancy that protects us. The saving grace: we might have an explanation of why $\mathscr{X}$ showed no Cutis laxa in spite of the loss of the start triplet. This probably lead to loss of the first 56 amino acids. And if behavior in mice held in humans as well, this loss would have no perceptible impact. So $\mathscr{X}$ would not have Cutis laxa, and neither would he have hypertension or lung abnormalities on this account. So we are at a dead end with *FBLN5*. But dead ends in detective stories are not uncommon. No reason for despair though, we still have one lead to pursue.

FLNA, the Master Orchestrator

As we survey literature on our last lead, the *FLNA* gene, what first emerges is a picture of awe. Unlike genes in previous chapters which played specialized roles in the eye or in the heart, this gene seems to have its tentacles spread widely. It appears to play a role in the formation of the brain, the heart, the blood vessels, the lungs, the bones, and even the intestines! How does it contribute to the formation of all these organs?

Imagine a single cell dividing and growing into a complete organism. Many different organs have to be created. Each organ has a specific shape and structure. For instance, blood vessels are long and tubular, the lungs are like balloons, the heart is a four-chambered thick-walled structure, etc. As cells divide, more cells become available for the creation of these varied structures. Then these cells have to organize themselves into these varied structures. Imagine you were given a collection of bricks and asked to make a house. You would have to move bricks around to create various parts of the house. And you would have cut various bricks into various different shapes depending upon the geometrical structure of the house you are building. Similarly, cells have to be moved around. And they have to be morphed into various shapes. And what moves them around and morphs them into shape? Proteins created from recipes described in some genes.

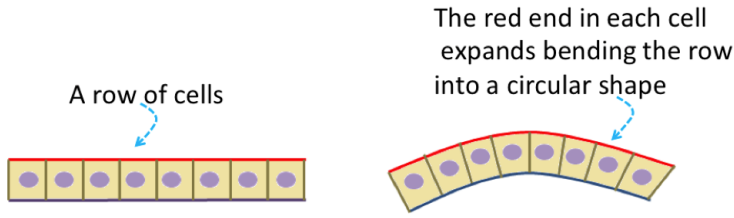


Figure 4.7: Tube formation from a sheet via cell shape changes

Wait a minute! Does an individual cell have a definite shape? Aren't cells simply like squishy round grapes? Not really; cells do have definite shapes. And not just simple fixed shapes, but shapes that they can change in

deliberate ways. This ability of a cell to change its shape allows collections of cells to then morph themselves into specific structures. For instance, imagine a line of cells (Fig. 4.7) which then need to be sculpted into a circular form. And assume these cells are glued together firmly. To accomplish a collective circular structure, each cell then modifies its shape so the end marked in red expands and the opposite end contracts. As a net result, the entire collection twists into a circular form. Similarly, this ability to change shape deliberately also enables a cell to move, as follows. A cell forms a protuberance that stretches forward in the direction of motion at the leading edge. Simultaneously, it retracts the opposite end at the trailing edge, much like an earthworm.

What gives shape to a cell? A structural framework formed by *actin* filaments (or ropes), which we met in Chapter 3. This framework is called the *actin cytoskeleton*, a skeleton, but of the cell. In Chapter 3, we also met a network of *desmin* filaments; these are thicker than actin filaments, but found only in specific organs which encounter greater mechanical stress, like the heart. Actin filaments are found in all cells. The filaments have to be organized in structurally sound ways for the cell to retain a specific shape. This organization is performed by several proteins. Some proteins help parallel filaments connect together into bundles, for greater strength. Yet other proteins help form connections between criss-crossing filaments, creating a 3 dimensional network of filaments (Fig. 4.8). By adjusting these filaments in different parts of the cell in different ways, cells can now change their own shapes. For instance, in Fig. 4.7, the filaments at the blue end are tightened while those at the red end are loosened, causing the row of cells to bend into a circular form.

What role does *FLNA* play is this context? The recipe in the *FLNA* gene yields a protein with a special structure, tailor-made for the purpose of connecting criss-crossing actin filaments (Fig. 4.8). Imagine a network of ropes in 3 dimensions and imagine a person holding a pair of criss-crossing ropes firmly, one in each hand; and imagine many such people, each holding a different pair of ropes. Together, these people keep the 3 dimensional network coherent. The *FLNA* protein does the same for the filaments in the actin cytoskeleton.

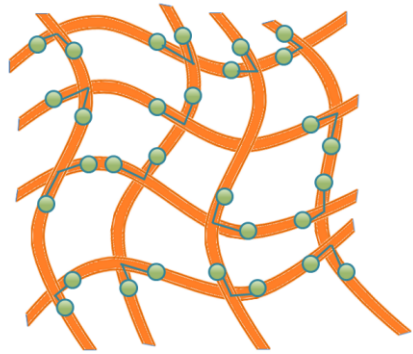


Figure 4.8: *FLNA* connecting criss-crossing actin filaments

The *FLNA* protein has 4 parts: the *Actin-binding Domain*, *Rod 1*, *Rod 2* and the *Pairing Domain* (Fig. 4.9). These parts are connected together not by rigid connectors but by flexible hinges. The actin-binding domain is where it attaches to an actin filament. In addition, a latter section of rod 1 also attaches to the same actin filament. A single *FLNA* protein does not operate individually; instead, two such proteins come together and join at their respective pairing domains to form a single combined structure. These two proteins attach to distinct criss-crossing filaments, typically holding these two filaments at close to 90 degrees to each other. The flexible hinges keep the network structure intact even when mechanical forces acting on the cell cause shear forcing this angle to change. What role does rod 2 play in all of this?

Organ formation requires cells to change shape and move. But they also require cells to modulate the amount of proteins generated by recipe interpretation from the various genes. For instance, to enable a cell to move, the protein which serves as glue between adjacent cells needs to be destroyed. So mechanical changes and chemical changes go hand in hand. This is where rod 2 of *FLNA* plays a key role. It has a special structure which allows it to sense mechanical forces acting on the cell. In turn, it initiates various chemical signals by interacting with a very large number of proteins, spanning a range of very diverse functions. Through this

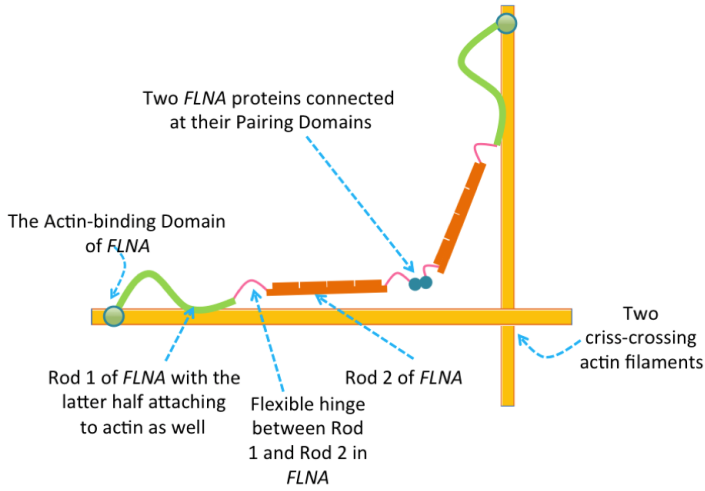


Figure 4.9: *FLNA* structure

sensing, *FLNA* converts mechanical signals to a large number of chemical signals.[42] So *FLNA* appears to be a master orchestrator of sorts.

Altered Recipes in *FLNA*

The best understanding of the far-reaching impact of this master orchestrator comes from studies on patients in whom the *FLNA* recipe is altered. And as we listen to the stories of these patients, the picture turns from one of awe to one of horror. If there were ever a gene with an extensive criminal record, this might be the one.

This impact of *FLNA* on patients is gender-dependent; the impact on females is morbid, but the impact on males is just brutally gruesome. Why this discrimination? This is a consequence of *FLNA*'s location in the genome; it appears on the X chromosome. Females have two copies of this chromosome while males have just one copy. So females have two copies of the *FLNA* gene while males have only one. Should the only copy of this gene in a male be somehow rendered dysfunctional, the

impact is lethal. Indeed, literature is rife with instances where male babies die while still in the womb or shortly thereafter, with various horrifying malformations. Females can tolerate one copy becoming dysfunctional because the other copy can hold the fort. But the gene is so potent that even one copy becoming dysfunctional sometimes causes severe issues in females, though immediate fatalities are rare.

One set of variants in *FLNA* causes what is called *Periventricular Heterotopia*.[43] Ventricles here are 4 fluid-filled cavities at the center of the brain (different from ventricles in the heart we met in Chapter 3). When the brain is forming early in life, neurons first form in the region around these ventricles; they then move outwards to form the exterior surface of the brain. If this movement does not happen correctly, then neurons clump together near these ventricles, hence the name Periventricular Heterotopia (periventricular means near the ventricles, and heterotopia means appearance at an unexpected position). The impact of this mislocation is typically mild or even silent, detectable only on an MRI scan (*Magnetic Resonance Imaging*). That is until the person in question starts having seizures (episodes of vigorous shaking), which start in the teens or the twenties. Mislocation of neurons does not affect mental ability typically, though some individuals do show mild mental disability. While mislocation of neurons is the typical theme, there are sometimes added abnormalities as well. In some individuals, heart valves and blood vessels close to the heart may be malformed. In others, the intestines may be affected leading to constipation. Some individuals might also have problems with blood clotting. Quite a range!

One curious fact about variants which cause Periventricular Heterotopia: several of these are nonsense and frameshift variants. Remember these are variants which truncate the gene pre-maturely and render it non-functional. A few missense variants (where one amino acid is replaced by another) also cause Periventricular Heterotopia. In general, a missense variant could go either way; it could reduce the protein's ability to function, or it could enhance its ability to function; either could be the case. Which way do the various missense variants that cause Periventricular Heterotopia go? The presence of many truncating variants causing Periventricular Heterotopia allows us to guess the answer to this question. Chances are

that these missense variants also reduce the function of the *FLNA* protein.

Most variants which reduce the gene's ability to function do not allow males to survive. So the above symptoms show primarily in living females. There are a few exceptions: some missense variants which cause only a minor reduction in the gene's function; these do allow males to survive, but with the above symptoms. Even in females, the impact may be quite severe, particularly on the respiratory apparatus. For instance,[44] reports 4 girls with progress respiratory failure in the first 4 months after birth. They also had abnormal cross-flow from the aorta (the blood vessel that carries oxygen-rich blood from the heart to the rest of the body) to the pulmonary artery (the blood vessel that carries oxygen-poor blood from the heart to lungs), and pulmonary hypertension. All 4 girls needed lung transplantation for survival. This is by no means a common procedure, and probably out of reach for most.

There is an additional set of missense variants in *FLNA*. The abnormalities found in individuals with these variants have their primary impact not in misplaced neurons, but on facial structure and on the bones. These abnormalities come in various shades and are accordingly classified into a number of complicated disease names: *Otopalatodigital syndrome type I (OPDI)*, *Otopalatodigital syndrome type II (OPDII)*, *Frontometaphyseal dysplasia (FMD)*, *Melnick-Needles syndrome (MNS)* and *Terminal osseous dysplasia (TOD)*. Bones are often not fully developed, leading to bowing of the long bones of the legs and insufficient strength in the chest structure. Malformation of fingers and toes are common. And facial structure shows uncommon patterns. For instance, the bony ridge of the forehead, just above the eyebrows, may be unusually prominent (if you look at pictures of chimpanzees and gorillas, you will find this ridge is very prominent; humans have in general lost this prominence). And the eyes may be further apart than usual (though only a trained eye may be able to spot this; indeed, as we realized, there is a whole branch of medicine dedicated to study of facial structure; doctors trained in this area can spot subtle abnormalities, or dysmorphologies, as they are called). Deafness, and defects in the heart, brain, intestine and kidneys are additionally present in some cases.

In contrast to variants which cause *Periventricular Heterotopia*, there is evidence to suggest that these variants actually enhance the gene's ability

to function.[45] This is another testament to how critically poised *FLNA*'s role in the development of various parts of the body is; both increased as well as decreased potency leads to problems! The impact of these variants on male and female lives is quite variable though. Males born with *OPD1* and *FMD* do survive with relatively milder symptoms. On the other hand, males with *OPD2* and *MNS* sometimes die before birth. And those who are born often die early in life, due to respiratory failure; rarely do they survive beyond the second year, except in rare cases with intensive treatment.[46] Note *death due to respiratory failure* appears to be a recurring theme in *FLNA*'s record book. Of course, conditions are typically not life-threatening for females. However, other symptoms show a wide range of expressivity. Some females appear completely normal (sub-clinical is the term, there are no apparent abnormalities, though detailed tests might unearth subtle internal issues). Others show as much severity as males would.

And that leads us to the burning issue: is *FLNA* the object of our hunt?

Is it *FLNA*?

The *FLNA* variant that we found in $\mathscr{X}$ happens to be a missense variant. The sole copy of *FLNA* in $\mathscr{X}$ has this variant. $\mathscr{X}$'s mother $\mathscr{M}$ too has the variant, but in exactly one of her two copies. $\mathscr{X}$'s father does not have this variant. We don't know whether $\mathscr{Y}$, $\mathscr{X}$'s brother had this variant; he had passed away earlier and no remains which carried his genome were preserved. Is this *FLNA* variant indeed the the root cause of this family's troubles? If so, then the symptoms shown by these two boys should fit somewhat into one of the typical scenarios we just discussed. Is that indeed the case?

If the missense variant we found reduced the gene's ability to function, then misplaced neurons should be a typical occurrence. Both boys has passed away well before *FLNA* entered the picture as a candidate. So neither was checked for these misplaced neurons. $\mathscr{M}$ also had the variant in one copy, which meant she should also be susceptible to misplaced neurons. This would typically result in epilepsy (episodes of vigorous shaking) starting in the teens or twenties. $\mathscr{M}$ was in her twenties, but had

no such episodes. She was still in her early twenties, so it is possible that these episodes might yet begin. An MRI scan would reveal if there were misplaced neurons in brain, even if these hadn't made their presence felt in the form of seizures. Performing an MRI on $\mathscr{M}$ was an option. This had not been done for practical reasons: an MRI is expensive, given our family came from a not too well-to-do background and from an eco-system where insurance was not the norm. But no one in the family showed any heart abnormalities or severe constipation or problems in blood clotting. So it seems likely that the missense variant we have found does not fit the bill as far as variants which reduce *FLNA*'s ability to function go.

Figure 4.10: Typical Facial Structure in Melnick-Needles syndrome. Picture reproduced with permission.[47]

Could this variant fit the bill for variants which enhance *FLNA*'s ability to function? The hallmark effects here are on facial and skeletal structure. The boys in our case did have some very distinctive facial features. Do these fit any of the profiles typical for *FLNA* variants? Let us remind ourselves with a picture: Fig. 4.10 shows a girl with *MNS*. What in this picture strikes you as distinctive?

A geneticist trained at spotting facial dysmorphology will not miss the prominent eyes, the full cheeks, and the small lower jaw. These are hallmark *MNS* features. The boys in this story (Fig. 4.1) too had prominent

eyes. Very prominent actually, almost popping out of their sockets. Fairly full cheeks. And small lower jaws. Would this suggest that the boys in our case had *MNS*?

The geneticist treating this family was not convinced. The classic facial features of *MNS* listed above are typically found in females. Females do show prominent eyes, full cheeks and small lower jaws. While many live full lives, some females die of respiratory failure, typically in the second or third decade of life. Frequent respiratory infections occur. Pulmonary hypertension (high blood pressure in the blood vessels linking the heart and the lung) is also known to occur.[48] $\mathscr{X}$ and $\mathscr{Y}$ clearly died of respiratory failure. $\mathscr{X}$ did have one reported incident of pneumonia when he was 10 months old. And both showed pulmonary hypertension. The one catch: $\mathscr{X}$ and $\mathscr{Y}$ were boys. And boys with *MNS* were more likely to pass away while still in the womb, or soon after birth, and suffer serious skeletal deformities.

Indeed, most references to *MNS* in scientific literature are to affected females. Male children born to these affected females show severe skeletal and facial malformations and die of respiratory failure very early. So living males are rare and hence not studied very much. However, scientific literature points out some striking dual behavior for males. Apparently, male *MNS* patients can be classified into two very different groups. The first group comprises males born to affected mothers; these show severe disease and die very early. The second, rarer, group comprises males whose mothers are unaffected and showed no signs of disease; these males show a milder form of the disease, similar to what affected females would show.[49] Maybe the boys in our family fall into the latter group.

What might cause this dual behavior? Maybe the specific variant within the *FLNA* gene determines which of the two groups one would fall in. Some variants could lead to moderate effects in females but very severe effects in males. Other variants could lead to minimal or no effect in females and moderate-to-severe effects in males. What can we say about the specific variant we found in $\mathscr{X}$ and $\mathscr{M}$?

The *FLNA* Variant: In a Fertile Desert

The genomic variant that $\mathscr{X}$ and his mother $\mathscr{M}$ have in the *FLNA* gene is called *D1970N*. *D* is short for *Aspartic Acid*, one of the 20 amino acids that triplets in gene recipes code for (Fig.2.5). *N* is short for *Asparagine*, another amino acid; as you might guess, it gets its name because it was first isolated from asparagus juice. The 1970th amino acid generated from the *FLNA* gene recipe is altered to an *N* from a *D*, hence the name *D1970N*. Of course, this is a missense variant. In $\mathscr{X}$, it is present in the sole copy of *FLNA*. In his mother $\mathscr{M}$, it in present in one of her two *FLNA* copies. We know that this variant has never been seen earlier in anyone else. Or at least no one has previously reported that they saw this variant in another individual. So we know nothing about this variant directly. We have to make indirect inferences from other knowledge that is available.

Now, most missense variants, where one amino acid is replaced by another, are quite harmless (or at least, more subtle in their effects). After all, each of us has tens of thousands of these variants in our genes. But a few can be dramatically problematic. Is *D1970N* really one of these problematic variants?

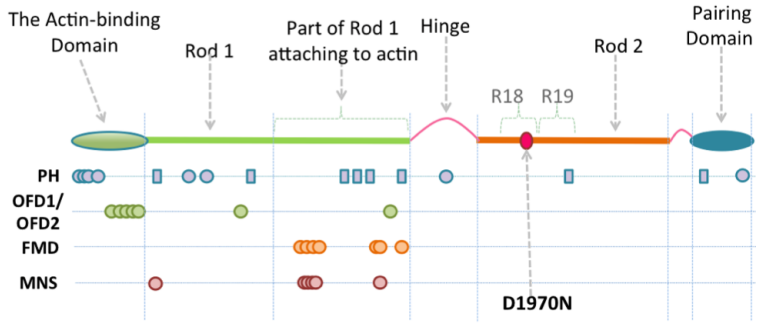


Figure 4.11: Disease-causing Variants in *FLNA* found in Patients as reported in Scientific Literature. Round dots are missense variants, rectangular dots are premature truncations. All variants causing the same disease appear on the same horizontal line. PH stands for Periventricular Heterotopia, MNS for Melnick-Needles Syndrome.

Fig. 4.11 may be helpful to place this variant in the context of problematic variants found in various patients with one of the diseases we talked about a bit earlier. What does this picture scream out loud? Square dots, representing premature gene truncations, cause Periventricular Heterotopia exclusively. In contrast, round dots representing missense variants appear in all diseases. And they cluster in certain regions. Most are in the actin-binding domain or in the section of rod 1 that binds to actin filaments. These variants are problematic because they modify the way *FLNA* attaches to these actin filaments, either increasing or decreasing the strength of this attachment. There is one round dot in the pairing domain at the very end. This is problematic because it affects the pairing of two *FLNA* proteins to create the combined structure shown in Fig. 4.9. There are a few variants in the section of rod 1 that does not bind to actin filaments, and there is one in the hinge region. But rod 2 is a desert as far as problematic missense variants go. And our variant *D1970N* sits bang in the middle of rod 2. We are clearly in new territory here!

Would it be right to conclude that *D1970N* is really harmless because practically no missense variants in rod 2 have been found so far, in spite of substantial research? Not really. Genomics research has been picking up pace over the last 20 years but our ability to study a large number of patients effortlessly and inexpensively is just a few years old at the moment. Hence our knowledge of the variant landscape is still very partial. For instance, until 2010, only 4 variants in FLNA were known that cause *MNS*. All 4 were tightly clustered at one place in the *FLNA* gene as shown, within the section of rod 1 that binds to actin filaments. This might suggest that *MNS* (remember, Melnick Needles syndrome) is caused by variants only in this region. But, in 2010, another missense variant causing *MNS* was discovered much earlier in the gene, outside the section that binds to actin filaments;[47] this expanded the possible range for variants in *FLNA* causing *MNS*. However, even to date, the number of *MNS* variants known still remains very small. No doubt more variants will emerge over time. possibly in different parts of the *FLNA* gene. For the moment, our variant in this case remains in a desert, with little help that be derived from known variants to establish its culpability.

So what information do we have about this desert? Remember, *FLNA*

converts mechanical signals to electrical signals by sensing mechanical strain on the cell and interacting with a large number of other proteins to convert this strain into chemical cues.[42] *FLNA* thus plays the role of a master orchestrator of sorts. And most if not all these interactions happen with rod 2. So, while rod 2 is a desert as far as variants reported in scientific literature go, it is eminently fertile when it comes to being in the scene of action. Could a variant in this fertile desert impact the ability of *FLNA* to convert mechanical signals to chemical signals?

From what we know about the ability of *FLNA* to interact with various molecules,[42] much of the action appears to happen in the part of *FLNA* marked *R19*, in Fig. 4.11, or further to the right. *D1970N* is located to the left of *R19*, in a region called *R18*. Then how could it have any influence on these interactions? It turns out that the sequence of amino acids generated from the recipe encoded in the *FLNA* gene folds up in a rather interesting way (Fig. 4.12).

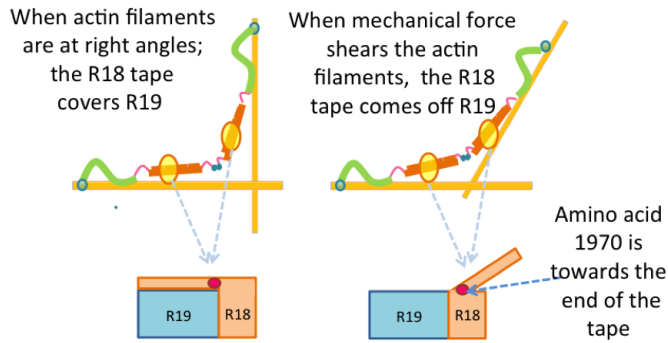


Figure 4.12: Opening and closing of the *R18* tape on *R19*.

A part of *R18* stretches out like a piece of tape and closes off the important parts of *R19* involved in these interactions. In this configuration, *R19* cannot perform its usual functions. When the cell changes its shape, the actin cytoskeleton network is strained as in (Fig. 4.12). This strain forces *FLNA* to stretch out. This stretch forces the tape to peel off *R19*. *R19*

can now perform its usual functions. So mechanical changes like change in cell shape force the tape on and off, thus switching off and on the sensing capabilities of *R19*. This is how *FLNA* is able to take mechanical signals and convert these to chemical signals. And *D1970N* appears right at the edge of this all-important tape. Could this change from *D* to *N* somehow change the dynamics of the tape opening and closing?

Unfortunately, just not enough seems to be known about this process to answer this question. In fact, *D* and *N* are actually very similar in chemical composition. As shown in Fig. 4.13, the only difference is that OH in *D* is replaced by NH_2 in *N*. So isn't this change from *D* to *N* quite minor, and unlikely to affect anything?

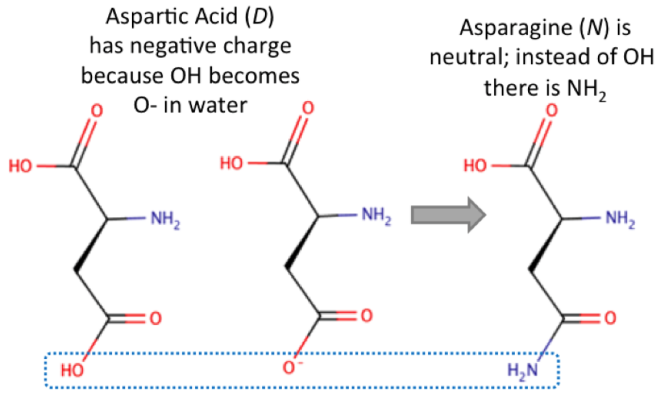


Figure 4.13: Aspartic Acid and Asparagine

There is indeed one important consequence of the *D* to *N* change in the watery environment within a cell. In this environment, *D* tends to lose a proton (i.e., a H^+ ion) and thus becomes negatively charged. *N*, on the other hand, does not lose or gain any ions, and therefore remains uncharged. Due to this -ve charge, *D* can form bonds (with other positively charged amino acids) that are stronger than what *N* can. *FLNA*'s ability to interact with several proteins depends on its ability to form chemical bonds with these proteins of appropriate strengths. Possibly, the *D1970N*

change weakens some of these bonds. There are examples known where an *N* to *D* change leads to the formation of stronger bonds and serious alteration in protein function. Ours could be a similar scenario. A vote for culpability, yes. But can we say something more definitive about *D1970N*? Or at least add more votes?

A Peek far into the Past

We know that *D1970N* has never been seen earlier in any other individual. If this variant had indeed been found in an adult male, we could have concluded that this variant is not the cause of early death in the boys in our family. Short of that, if it had been found in many females, we might have again reached the same conclusion. Why? After all, the presence of this variant in one of the two copies of *FLNA* in a female, as in $\mathscr{M}$, does no harm. The reason is the following. If it were indeed present in many females, then roughly half of the children from these females would be males, and half of these males would have the *D1970N* variant and would die early. So the presence of the variant in many females would signal a high prevalence of early male death on account of *D1970N*, which would have made news. However, this variants has just not been seen in anyone at all! And that does add one more vote in favor of its culpability. But have we looked hard enough to see if this variant is indeed found in other people?

The total number of individuals whose genomes have been sequenced today probably runs shy of a 100,000. And much of this data is probably private, leaving under a 10,000 individual genomes publicly accessible. And none of these have *D1970N*. Would one of the remaining 90,000 genomes have this variant? Or for that matter, would the genomes of one of the almost 6 billion individuals on earth have this variant? Unfortunately, there isn't a way to find out. We just have to wait for publicly accessible genomics databases to grow, which they will quite dramatically over the next several years. What can we do in the interim? There is a trick. A clever one at that! It does not substitute for the inability to look at the genomes of all individuals alive today; however, in an indirect way, it actually allows us to look well beyond!

Each time a genome is passed from parents to children, a few changes happen to the genome. Over generations, this number adds up substantially, so genomes of individuals appear more and more different as time goes by. By comparing the genome sequences of, say, 10,000 individuals alive today, we can identify how much these differ. And then knowing the rate at which differences appear on average from generation to generation, we can work backwards and calculate how many generation ago did our common ancestors live. This calculation yields an estimate of about 200,000 years. So it appears that we 6 billion individuals on earth today were all descended from a relatively small group of people who lived 200,000 years ago. But aren't we digressing?

Hold on for a bit. We can go further than humans here and do this exercise with human genomes and chimpanzee genomes. We identify the genome sequences of a few chimpanzees and compare these with our various human genomes, identify the number of differences, and again work backwards to calculate when our common ancestors lived. Of course, this should be earlier than 200,000 years ago. A lot earlier, it turns out. This number is about 6 million years! So by studying humans alone, we get a glimpse into only 200,000 years of evolutionary history. By studying chimpanzee genomes, we can get a glimpse of 6 million years of evolution! But why just chimps, we could go even further back in time. We do the same exercise with elephants. And that yields a common ancestor that lived about a 100 million years ago! Humans, chimps and elephants are all mammals. With the duck-billed platypus: half mammal and half egg-laying reptile (*metroneme* is the term for such creatures): 200 million years! With snakes, clearly reptiles: 300 million years! With amphibious frogs, who live on both land and in water: About 400 million years ago! And with fish, which live fully in water: About 450 million years ago!

So by reading the genomes of fishes, frogs, reptiles, metronemes, and various mammals, we get a glimpse into 450 million years of evolutionary history. Over the 450 million years, starting with the ancient fishes, genomes have passed on from parents to children, with each such event bringing about a few changes to the genome. Over several several generations, these changes have accumulated to yield one life-form from another. Charles Darwin's famous theory suggests that these change occur

randomly in the genome, and not in some calculated way with the intent of creating a particular new life-form. Yet, some of these changes stay and others are weeded out over time. Genomic characters that improve the *fitness* of a species tend to stay (these are *selected for*). Genomic characters that compromise the fitness of a species tend to get weeded out (these are *selected against*). Genomics characters which do not impact fitness at all could go either way, depending on where they lie in the genome (these undergo what is called *random drift*). Fitness here means the ability to survive against adverse environmental conditions, find mates, and have offspring. More offspring means greater propagation of genomic characters. And genomic characters which provide a parent the ability to have more offspring get propagated more widely, and hence are selected for.

What has this 450 million years of evolution done to the *FLNA* gene? Believe it or not, fishes, frogs, reptiles, metronemes and mammals all have an *FLNA* gene. What's more, the recipe encoded in this gene is strikingly similar in all of these very diverse life-forms. Surprising? Well, all of these organisms have a challenge in common: they all have skeletons, hearts, brains and blood vessels that have to be formed starting from a single cell. And this organ formation needs cells to change shape and move. And *FLNA* is a key player here. So no wonder its recipe is strikingly similar across these life-forms. Let us zoom into amino acid 1970, the site of the *D1970N*, and explore this similarity further.

Fig. 4.14 shows the *FLNA* amino acid sequences from various life-forms stacked using a computer algorithm so we can compare characters across life-forms. Remember, there are 20 amino acids, and each is denoted by a single letter (Fig. 2.5). For instance, *R* stands for *Arginine* and *V* for *Valine*. Since the *FLNA* amino acid sequence is quite long, 2647 amino acids, this picture shows only a few amino acids around amino acid 1970.

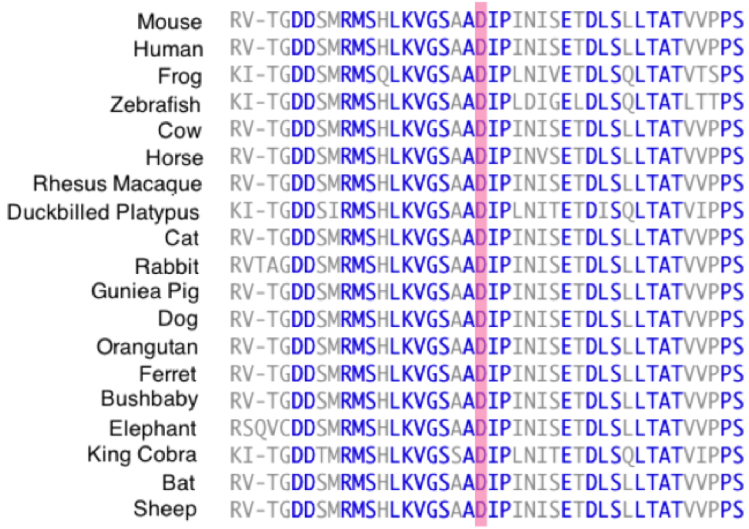


Figure 4.14: A comparison of *FLNA* sequences for various vertebrates (life-forms with backbones) around amino acid 1970 (in the pink column). Blue columns have the exact same amino acids in all organisms; gray columns have at least two distinct amino acids.

The stacking of sequences above has to be done carefully. For instance, elephants and rabbits have an extra character in the third column (Q and T, respectively). Since none of the other organisms have this extra character, we must introduce an extra - character in the sequences for all the other organisms and shift all the other characters to the right. If we didn't do that, the characters for elephants and rabbits would be staggered relative to the other sequences, and we wouldn't be able to really check whether a particular character was common to all the organisms or not. With all that done, what stares us in the face? Amino acid D is common to all the organisms shown in the figure. In other words, fishes, frogs, reptiles, metronemes, mammals all have a D at 1970. The D at 1970 has been *conserved* for 450 million years of evolution!

Of course, when we say fishes, we mean just a few fishes whose genomes have been studied. Certainly, genomes have not been studied

for tens or hundreds of thousands of fishes. So the fish *FLNA* recipe we're talking about is just a randomly drawn representative; it is possible that other fishes have other amino acids at 1970. But if randomly drawn individuals from so many different life-forms all show a *D* at 1970, then one of the following must be true. Either, other characters at 1970 appear only in a small minority of individuals in every species, too few for us to encounter. Or, a change from *D* causes a reduction in fitness and is therefore selected against. In the latter case, *D1970N* is likely to be problematic, one more vote for *D1970N*. Of course, we cannot rule out the former scenario with certainty.

So let us dig a bit more into the past. It so happens that *FLNA* is a member of a family of related genes called *Filamins*. There are 3 genes in this family, *FLNA*, *FLNB* and *FLNC*. The recipes in all these genes are similar to each other, not 100% but 64%. So, the guess is that all these 3 genes evolved from a common ancestor gene.[50] Remember that genomes are modified as they are passed from parents to children. Most such changes are just replacements of one character with another. On rare occasions, these changes can be more complicated. For instance, one section of the genome may get copied twice. Such events can create multiple gene copies from a single ancestor gene. This is possibly how a single ancestor gene lead to the formation of 3 distinct, but similar genes, *FLNA*, *FLNB* and *FLNC*. Once formed, these 3 genes evolve independently. Each will experience random changes that could then be selected for or against or undergo random drift. So over time, differences appear between these genes.

Figure 4.15: A comparison of *FLNA*, *FLNB* and *FLNC* around amino acid 1970. Blue columns have the exact same amino acids in all three genes; gray columns have at least two distinct amino acids.

Again, we can calculate when the ancestor gene split into these multiple genes based on the rate of appearance of these differences: about 500-550 million years ago.[50] So by comparing *FLNA*, *FLNB* and *FLNC* amino acid sequences, we can push our window of observation into the past from 450 million years to 500-550 million years. Fig. 4.15 shows this comparison, focused around amino acid 1970. And the *D* at 1970 is still conserved, over 500-550 million years of evolution! And $\mathscr{M}$'s and $\mathscr{X}$'s genomes have modified this *D* to an *N*; an event which we have been unable to observe in the last 500-550 million years of evolution! Another vote for *D1970N*.

Taking Stock

It's time to take stock and bring all the evidence we have against *FLNA* to book. Is this evidence against *D1970N* convincing?

The facial characteristics of $\mathscr{X}$ and $\mathscr{Y}$ were our first clue. Most visibly, prominent eyes popping out of their sockets. Also, full cheeks and a small chin. These are all typical of Melnick-Needles syndrome (*MNS*), one of the various diseases caused by problematic variants in *FLNA*. Both $\mathscr{X}$ and $\mathscr{Y}$ died of respiratory insufficiency. This is also a common finding for patients with *FLNA* variants, including Melnick-Needles syndrome. Pulmonary hypertension, i.e., high blood pressure in the artery carrying blood from the heart to the lungs for oxygenation, was present in both $\mathscr{X}$ and $\mathscr{Y}$. While not a common occurrence, this too has been reported in patients with *FLNA* variants, including Melnick-Needles syndrome. But Melnick-Needles syndrome is typically found in females; males from affected mothers usually die before or shortly after birth with severe skeletal malformations. A bit puzzling at first sight. But scientific literature does reports rarer instances of males with unaffected mothers having milder manifestations. $\mathscr{X}$ and $\mathscr{Y}$ would fit this description.

But here is a catch! Abnormalities of the skeleton are typical for Melnick-Needles syndrome: bowed bones, a small rib cage, and ribbon-like ribs, suggesting that the chest was not fully formed. Sometimes, these malformations are subtle, visible only when X-Rays are performed. $\mathscr{X}$ and $\mathscr{Y}$ did not suffer from any clearly visible skeletal abnormalities. We

do not know if X-Rays were performed on them to investigate any subtle abnormalities. Given both died of respiratory insufficiency, it is likely that a thorough investigation would have revealed some underdevelopment of the chest area. However, no such investigations were mentioned in their medical records. Certainly, the doctors treating $\mathscr{X}$ and $\mathscr{Y}$ did not have any indication that an *FLNA* variant might be the culprit, so they may not have looked for skeletal abnormalities in the ribs. And the other symptoms that $\mathscr{X}$ and $\mathscr{Y}$ showed probably did not warrant a thorough skeletal investigation. And $\mathscr{X}$ and $\mathscr{Y}$ were long gone by them time our genomic investigation started.

Given this lingering element of doubt, a more careful look at the *FLNA* variant *D1970N* was warranted. For which we travelled back 500-550 million years in time. And we were unable to spot anything but a *D* at location 1970. If indeed someone did have a different character at 1970, he, she or it was likely in minority. So there is a good chance that this *D* plays an important role and has been selected for over this long evolutionary period. But what role could it be? We do not have the knowledge today to answer this question. We know that *D1970N* lies in the *R18* region of *FLNA*, right at the edge of the tape that opens and closes to control the ability of *R19* to interact with a number of other proteins. Could *D1970N* impact this process? Or some other process involving *R18*? We just don't know.

We do know that *FLNA* lies on the X chromosome. And there is a 25% chance that both male children would inherit *D1970N* from a mother who has this variant in one of her two gene copies. For a gene on another chromosome (chromosomes 1-22, that is), this number would be only 6.25%. So a gene on the X chromosome is a good candidate from this perspective.

If only we had access to $\mathscr{Y}$'s genome (remember, $\mathscr{Y}$ was $\mathscr{X}$'s deceased brother), we could have checked for *D1970N* in $\mathscr{Y}$. A positive finding would have been another vote for *D1970N*. A negative finding would have ruled out *D1970N*. But $\mathscr{Y}$'s genome was out of reach. We could check if any other males in the extended family had *D1970N*. Unfortunately, $\mathscr{M}$'s parents had passed away much earlier. Had they been alive, we could have checked them for this variant. Since $\mathscr{M}$ had this

variant in one copy of her X chromosome, she could have got it in one of three ways. She could have inherited it from her father. Or from her mother. Or it could be de novo, i.e., it crept in when she was conceived. If she had inherited it from her father, then we could rule out *D1970N* as the cause of this family's problems. Why? Because her father would then have *D1970N* in his sole copy of *FLNA*, similar to $\mathscr{X}$. And, unlike $\mathscr{X}$, he clearly lived long enough to have children. Unfortunately, $\mathscr{M}$'s father was no more. No remains of his (hair, etc) which carried his genome were left behind. How could we then get access to his genome?

Through his children, of course. $\mathscr{M}$ and her sister $\mathscr{S}$ (note, both females) were his only children. So we checked $\mathscr{S}$. No sign of *D1970N* in her; both her *FLNA* copies had a *D*. One of these two copies was the sole copy in her father. So $\mathscr{M}$'s father did not carry *D1970N* either. How about any other males in the family. The only other male surviving in the immediate family was $\mathscr{M}$'s husband, and also her uncle, $\mathscr{F}$. Remember, this was a consanguineous couple: $\mathscr{F}$ was $\mathscr{M}$'s mother's brother. He, of course, did not have *D1970N*. So *D1970N* was not present in any male in the family for 2 generations. This would be another vote for *D1970N*.

There is one key symptom in $\mathscr{X}$ which we haven't addressed at all: systemic hypertension. This means high blood pressure in the arteries carrying oxygenated blood from the heart to the rest of the body (and not the lungs). There are a few patients with *FLNA* variants who have pulmonary hypertension, i.e., high blood pressure in the artery carrying blood from the heart to the lungs. But systemic hypertension has not yet been seen in patients with *FLNA* variants. Hypertension could occur due to defects in the heart. But that was ruled out in $\mathscr{X}$'s case; the heart was normally formed. What could explain this?

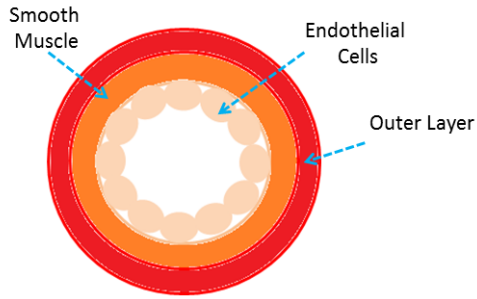


Figure 4.16: Blood Vessel Cross-section

The presence of both systemic and pulmonary hypertension suggests that the cause of hypertension lies in the blood vessels themselves. Blood vessels are made of many different types of cells (Fig. 4.16). The innermost layer comprises *endothelial cells* that release *nitric oxide* which smooth muscle cells then react to by contracting or dilating. This contraction and dilation of the vessels controls the pressure exerted by blood flow in the vessels. Elastic arteries that dilate and contract easily keep this pressure in the normal range. Stiffer arteries that cannot dilate and contract so easily face higher amounts of pressure. Would abnormalities of contraction and dilation be an implausible consequence for an *FLNA* variant?

Quite the contrary! We know that *FLNA* plays a very important role in the formation of these blood vessels. By its usual action in modulating cell movement and cell shape, *FLNA* plays a key role in the formation of the various layers of a blood vessel, the endothelial layer in particular. We know this because mice bred without the *FLNA* gene show defects in how these endothelial cells are connected to each other; there is widespread hemorrhage in these mice (or breakage of blood vessels).[51] It is certainly possible that a specific *FLNA* variant like *D1970N* could lead to a sort of blood vessel structure where the defects were milder; these blood vessels would not hemorrhage but their ability to contract and dilate as usual may be impacted, leading to hypertension. It is possible that systemic hypertension has not been seen in patients with *FLNA* variants so far, because neither *D1970N* nor other missense variants in the vicinity have

been found in these patients.

So we have against *FLNA*'s *D1970N* a dossier comprising plausible explanations. Strong enough that multiple experts when presented this evidence agree *D1970* is, at the very least, highly suspicious. No doubt, a very strong candidate. But, this isn't conclusive scientific proof yet. That would require experiments in which *D1970N* is induced into some living system and its consequences studied. That could take months, if not years, given appropriate funds. Generating funding itself would be yet another time-consuming exercise. And time is what we didn't have, as we will see next.

Will *FLNA* strike again?

Before we knew it, $\mathscr{M}$ and $\mathscr{F}$ decided to move on from their deceased children, and have their next child. Of course, the last thing they wanted was a repeat of the traumatic experience they went through with $\mathscr{X}$ and $\mathscr{Y}$. Would they have to wait till their child was a year old or more before they would know what was in store? Or at least until the child was born and medical investigations conducted to establish whether there was a serious issue or not? Or could we help them cut short this agonizing period, with each moment spent in fear of potential impending doom?

For several years, medical science has had the ability to extract the baby's genome prenatally, i.e., while it is still inside the mother's womb. How does one get to the baby's genome? Typically, to get a person's genome, one takes a sample of saliva or blood, or some other collection of cells in the body. Even hair or nail will do. But how do you get any of these for a tiny foetus deep inside the mother. It turns out that there is a way to get to the baby's genome without getting inside the baby.

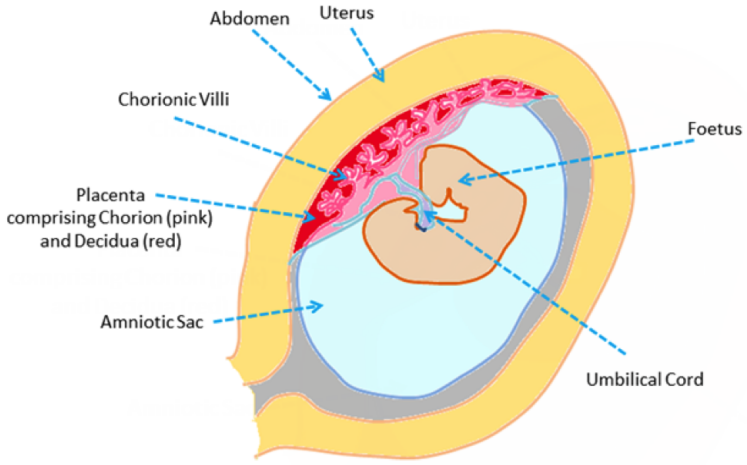


Figure 4.17: Chorionic Villi in the Placenta.

The umbilical cord from the baby is attached to an organ called the placenta. It is through the placenta that the baby gets its nutrients and disposes off waste. A part of the placenta called the chorion comes from the baby and carries its genome. The other part of the placenta called the decidua comes from the mother. To cause the least disturbance to the baby, a little tissue sample is taken from tiny protuberances on the chorion called chorionic villi (Fig. 4.17). This is done by a needle inserted into the abdomen under the watchful guidance of an ultrasound probe. The sampling procedure (aptly called *CVS* or chorionic villus sampling) can be performed when the baby is just 10 weeks old. It has to be done carefully though, otherwise the sample extracted could contain a mixture of the mother's and the baby's cells, and genomic analysis could be let astray.

CVS does pose a small risk to the pregnancy. Hence it is performed only in higher risk pregnancies. For instance, when there is a family history of a genetic disorder, or when ultrasound screens indicate something is wrong. For instance, sometimes an ultrasound will indicate *nuchal translucency*, i.e., that the baby has too much fluid collecting around the neck. This is often indicative of an abnormality called *Down's Syndrome*.

Instead of two copies of chromosome 21, there are three copies of this chromosome in this syndrome! This leads to growth delays and moderate mental disability. If nuchal transparency is seen in an ultrasound at the end of 12 weeks, then typically a CVS sample is taken and a genetic test is performed to check for 3 copies of chromosome 21. And if the test is positive, the family can choose to consider an abortion or they can choose to continue. This gives the expecting family a choice of terminating early, or equally well, setting their expectations in accordance with a differently-abled child. In recent times, the need to take a CVS sample for this purpose is being replaced by a more modern and less riskier method: taking a bit of the mother's blood. Mother's blood has free floating genomic fragments, 5-10% of which come from the baby and the remaining from the mother. These fragments are all mixed up, so we don't know which fragment came from whom. In spite of this, it is possible to identify whether the baby has three copies of chromosome 21 or two copies, just by sequencing all these fragments collectively! However, this method is not quite as developed when it comes to identifying whether the baby has a specific missense variant, like *D1970N*.

Getting back to our story, $\mathscr{M}$ was in the 12th week of her pregnancy now. Clearly, there was a strong family history of disease, which would justify a CVS. Added to that, an ultrasound scan in the 12th week showed nuchal translucency and a depressed nasal bridge. Both of these were indicative of Down's syndrome. A CVS was promptly performed. The number of copies of chromosome 21 was then checked. There were exactly two copies in the baby; so all was well so far. The question that remained on everyone's mind: did this baby has *D1970N* and would it encounter the same fate as its deceased brothers?

What if the baby turned out to be male with the *D1970N* variant in it sole copy of *FLNA*? The family would have to be explained possible future consequences. That there was a good chance of this baby following the same trajectory as its ill-fated brothers. Like most families who can only fathom so much of the detail here, this family will most likely turn to their clinician for her recommendation, and follow that blindly. The clinician would have a difficult decision to make. Let the pregnancy continue, and subject this family to the constant threat of another tragic event down the

line. They were looking up to her for help, and she finally did have a tool to help them now. Could she ignore this tool and see them plunge into sorrow again? On the other hand, could she recommend early termination of pregnancy with complete conviction? One was terminating a life after all! This was her first experience with the use of large scale genome sequencing to diagnose what appeared to be a hard-to-diagnose non-standard disease. Not only was this her first experience, it was also early days for the entire medical community. Barring a few thought leader clinicians and clinical researchers, the use of large scale genome sequencing had not yet entered into regular medical practice. In all likelihood, this was the first case of its kind in India. How would a drastic decision such as early termination of pregnancy be viewed in this light? Since the culpability of *D1970N* had not been proven conclusively, what in the off-chance that it really wasn't the culprit. What if an autopsy on the aborted foetus showed that all was well and the foetus was aborted for no reason? There are also strict laws in India against revealing the gender of the baby before it is born. How could she take this decision without revealing the gender? All tough questions as she geared up to take a decision.

Wrapping Up

For several diseases, skilled geneticists can actually pinpoint the underlying defective gene without so much as a peek into the genome. They do this simply by matching symptoms to a classic textbook description of the disease. This does need a very sharp, trained eye though. However, new diseases which don't quite fit into a geneticist's textbook are not uncommon. Pinning down the gene then becomes difficult. And in the absence of a genic diagnosis, families such as $\mathscr{M}$'s struggle with the tragic prospect of giving birth and raising a child that they might soon lose. Against this backdrop, the ability to look deep into the genome and identify the likely cause of disease provides a tool to such families, empowering them with a choice very early in their pregnancy. The identification of *D1970N* provides such a tool to $\mathscr{M}$ and her family.

Several different pieces of evidence had to be brought together to identify *D1970N* as the likely candidate, including some clever thought

experiments that took us back 500 million years ago in time. But, unlike Chapters 2 and 3, where we could pin down the gene candidates conclusively, we did not have conclusive proof for *D1970N*. Indeed, conclusive proof for missense variants which have never been observed in anyone else, or for that matter in any other organism with a backbone, for the last 500 million years, requires expensive and time-consuming scientific experiments. When researchers proactively perform these experiments, some very difficult cases get solved conclusively, as we will see in Chapter 5. But decision making in medical practice cannot wait in general. $\mathscr{M}$'s clinician had to make the best decision possible, taking all evidence available and all ethical and legal concerns into account. And she had a tough decision at hand!

The ultrasound is the one other tool she had at her disposal. The pregnancy had reached its 15th week. The ultrasound still showed a depressed nasal bridge, indicating potential lack of normalcy. A decision had to be made soon. The genetic results for *D1970N* arrived in the 16th week. And they showed that the baby had two copies of *FLNA*, one with a *D* and another with an *N*, just like its mother $\mathscr{M}$. By the 17th week, signs of a depressed nasal bridge on the ultrasound also cleared up. There was every indication to suggest that this would be a healthy baby. So the pregnancy continued. A little girl was born uneventfully several months later. The parents were happy to note that her eyes were normal and not popping out conspicuously as in her ill-fated brothers. A year old now, she appears normal on all counts. And blissfully unaware of the history she had possibly created: as the first child in India whose birth was supported by with an army of biologists and computer scientists, who scoured every inch of her family's genetic sequence in a bid to ensure that she would live a healthy life.

Chapter 5

Not Quite a Mirror Image

The liver is usually on the right side of the body, and the stomach and spleen are usually on the left (Fig. 5.1). $\mathscr{X}$, on the other hand, was born with his liver to his right and his stomach and spleen to his left!

That wasn't all. The *aorta*, the main blood vessel carrying oxygenated blood from the heart to the body, usually runs down along the left of the backbone. The *vena cava*, the main blood vessel carrying de-oxygenated blood from the rest of the body to the heart, usually runs up along the right of the backbone. $\mathscr{X}$'s aorta ran down along the right of his backbone and his vena cava ran up along the left of his backbone, as in the picture below. His anatomy appeared to be a mirror image of its usual self.

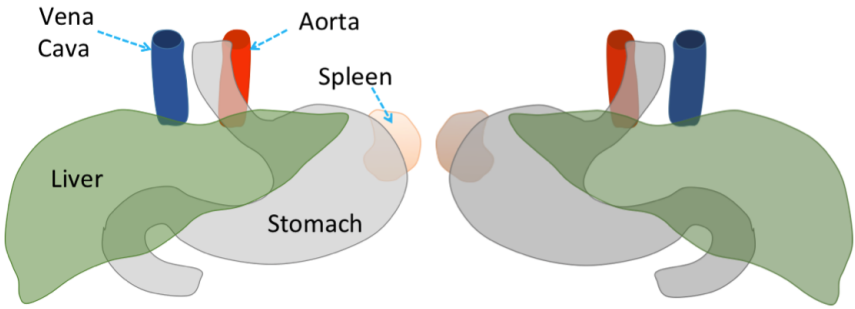


Figure 5.1: Mirror Image Flipping of Liver, Spleen and Stomach.

Not a complete mirror image, though; some of the organs did appear

in their usual places. Organs like the eyes and the kidneys come in pairs so mirror image switches on these may not matter. Organs like the urinary bladder are solitary and largely left-right symmetric; mirror image switches on these also may not matter. That leaves one very important solitary organ which is not left-right symmetric, namely, the heart.

If the heart were also a mirror image of its usual self, this novel anatomy may not cause any problems; after all, all organs were flipped in a consistent way. However, if some organs were mirror-imaged but the heart wasn't (a phenomenon called *heterotaxy*), then nature would face challenges in connecting these organs together. $\mathscr{X}$'s heart was not flipped around. And, indeed, the wiring of his blood vessels (which were indeed flipped) into his heart had gone completely awry, as shown below.

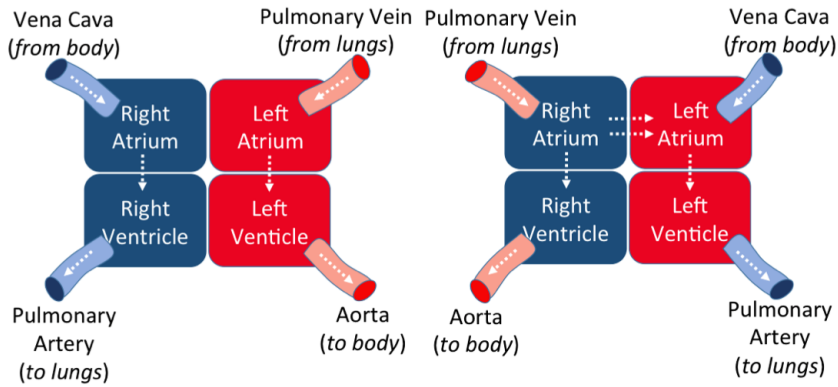


Figure 5.2: Wiring of Blood Vessels into Heart. The normal picture on the left and $\mathscr{X}$'s picture on the right.

Remember the flow of blood through the heart from Chapter 3? The *pulmonary vein* brings in oxygen-rich blood from the lungs to the left atrium. The left atrium pumps this blood to the left ventricle. The left ventricle in turn pumps this blood into the aorta, which then takes it to the rest of the body. So the left atrium and ventricle together provide oxygen-rich blood to the rest of the body. Similarly, the vena cava brings oxygen-poor blood from the rest of the body to the right atrium. The

right atrium pumps this blood to the right ventricle. The right ventricle in turn pumps this blood into the *pulmonary artery* which then takes it to the lungs for oxygenation. So the right atrium and ventricle together take oxygen-poor blood to the lungs for oxygenation.

In $\mathscr{X}$'s heart, the pulmonary artery and the aorta were transposed. The pulmonary artery was connected to the left ventricle instead of the right. And the aorta to the right ventricle instead of the left. The pulmonary vein now entered the right atrium instead of the left. This meant that the right ventricle was now responsible for providing blood to the rest of the body (the lungs excluded) instead of the left. This by itself was less serious, at least in the short term. In the long term, the right ventricle would be under excessive stress because its walls are not as muscularly designed as those of the left ventricle to be able to pump blood to the whole body over a prolonged period.

But there were additional issues, as is often the case in heterotaxy. The wall separating the left and the right atria had a hole through which blood leaked from the right atrium to the left. This meant that oxygen-rich blood from the lungs would enter the right atrium and then get split between the right ventricle and the left atrium. From the left atrium, this blood would go to the left ventricle and thence back to the lung for oxygenation via the pulmonary artery. So some of the oxygen-rich blood coming from the lung to the heart would go back to the lung in a short circuit, without delivering oxygen to the rest of the body. And only some of this oxygen-rich blood would flow to the rest of body, depriving several parts of the body of much-needed oxygen, causing parts of the body to turn blue.

$\mathscr{X}$'s brother $\mathscr{Y}$ was 6 now. He too had been born with similar, though not identical, alterations in anatomy. In both cases, surgery on the heart to fix some of these defects was necessary for the children to survive. Both parents and other siblings of $\mathscr{X}$ and $\mathscr{Y}$ were normal, conforming to the standard architectural plan. From where, then, did $\mathscr{X}$ and $\mathscr{Y}$ acquire this curious architectural plan?

Where Do We Look?

Unlike in previous chapters, our window into the genomes of $\mathscr{X}$, $\mathscr{Y}$, and their parents was not *Whole Exome Sequencing*. As the name indicates, the whole exome sequence of an individual comprises the text stretches in the 200,000 exons spanning 21,000 genes. Remember, gene recipes are encoded in the exons (see Fig. 1.12). These exons comprise only just about 1% of whole genome. The remaining 99% are introns, which separate exons, and *inter-genic regions*, which lie completely outside all the genes (see Fig. 1.12). So the vast expanse of introns and inter-genic regions comprise the overwhelming majority of the genome. For $\mathscr{X}$ and his family, the whole genome was sequenced; introns, inter-genic regions and all! And this difference will be critical in our quest, as we will see soon.

Whole Genome Sequencing gets us much more of the genome. At added cost, of course. Added cost, that is often unnecessary, given an overwhelming majority of variants causing disease are present in the exons. However, whole genome sequencing can be a life saver in some situations. While more expensive, it actually takes less time. Because, separating the exons out from the rest of the genome takes a few days longer than simply going ahead and sequencing the whole genome. For instance, Stephen Kingsmore and team at Children's Mercy Hospital, Kansas City, pioneered the practice of rapid sequencing for diagnosing children with genetic defects:[52] they sequence the whole genome in 24 hours, and deliver a diagnosis after analyzing the resulting variants in another 24 hours! $\mathscr{X}$ and his family comprised one such case. A particularly difficult case though, as we shall see.

Whole genome sequencing generates an enormous amount of data. But data crunching on a computer can still be performed in a few hours with clever algorithms and systems. It yields about 4 million genomic variants per person, i.e., characters in the genome which differ from a supposedly healthy genome sequence. How do we identify which of these 4 million is the culprit here?

Figure 5.3: Various Possible Inheritance Scenarios; shown are chromosomes for the mother, father and single male child. The last scenario involves the X chromosome.

Our experience from Chapter 4 reminds us of the possible scenarios in this case (see Fig. 5.3). Both $\mathscr{X}$ and $\mathscr{Y}$ were boys. So was this yet another case of a variant lurking in one copy of the mother's X chromosome? Remember, boys have only one X chromosome copy, inherited from their mothers. $\mathscr{X}$ would have a 50% chance of inheriting this variant. Both $\mathscr{X}$ and $\mathscr{Y}$ would have inherited this variant with a 25% chance; not that small! Or was it one of the other scenarios in Fig. 5.3? Maybe both mother and father carry problematic variants in one copy each of some other gene. In this scenario, there is a 25% chance that $\mathscr{X}$ inherited the problematic variant from his father in one gene copy, and the problematic variant from his mother in the second gene copy, possibly crippling both copies of that gene. The chances of this happening in both boys, each at 25%, is 1 in 16, or 6.25%. Which of these scenarios was responsible for the curious architectural plan in $\mathscr{X}$ and $\mathscr{Y}$? And which gene was responsible?

The first shot, of course, is at genes well known to cause shades of mirror image anatomy. Scouring through literature yields a handful of these genes. As in previous chapters, we look for variants in these genes which either truncate the gene recipe prematurely, or modify it so one amino acid is replaced by another (*missense* variants, as these are called). Of course, we focus only on *rare* variants, i.e., variants that are not commonly found in many people. We do so because such common variants are unlikely to be causative of $\mathscr{X}$'s curious anatomical rearrangement, a very rare condition. And we insist that these variants be found in both $\mathscr{X}$

and $\mathscr{Y}$ and that they satisfy one of the scenarios above (Fig. 5.3). The computer takes moment, while we hold our breath. And then it returns a blank! None of the genes known to cause shades of mirror image anatomy seem to apply to our case.

The next shot is at genes which are known to cause some sort of disorder in development of any part of the body but serious enough to manifest at birth or in early childhood. These include genes like *FLNA* which we met in the last chapter. Such genes number a few thousand, of the 21,000 or so total genes. Again we look for variants in these genes as we did for the handful of genes above. And yet another blank! Here we have a family waiting for a diagnosis, and the state-of-the-art offers us no help at all. There is little alternative but to foray bravely into virgin territory now.

We move on to the remaining 17,000 or so genes, none of which are known to cause even the remotest of architectural changes, let alone changes of the type seen in $\mathscr{X}$. We instruct our computer to wade through the millions of variants in $\mathscr{X}$ and $\mathscr{Y}$, looking for variants with the properties described above. The computer takes a little longer this time. Yet another blank? Thankfully, not this time. One gene appears on the screen. A gene called *BCL9L*. What is *BCL9L*? And how could it be instrumental in placing the liver on the left, and the stomach and the spleen on the right?

How does Nature know Left from Right?

Why is the stomach on the left in almost every one of us? Note that nature is not picking randomly between right and left; had it done so, several of us would have the stomach on the left and the remaining several on the right. Nature somehow knows what is left from what is right? And it gets this right almost every time. How does nature know what is left and what is right?

We could begin by asking: why are our bodies shaped the way they are, and not as perfectly symmetric round balls? Because symmetry-breaking events abound in nature. Here's one example. We all begin life when a single sperm cell from our father fuses with an egg cell from our mother. The picture looks like that in Fig. 5.4 because the sperm cell is

much smaller than the egg cell. After the sperm and egg cells combine, the distribution of various molecules inside the combined cell becomes asymmetric; a different mix of molecules appears at the point where the sperm combines with the egg as compared to the antipodal point on the other side. In frogs, the sperm entry point eventually develops into the back end and the antipodal point develops into the belly end.[53] Differences in the mix of molecules at the two ends, over many many cell division steps, causes the back to be shaped differently from the belly

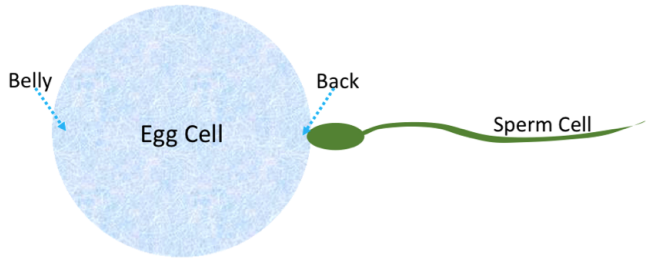


Figure 5.4: Sperm Attachment Point breaks Back-Belly Symmetry.

In humans and mice, the sperm attachment point is less important and symmetries are broken via more complicated events. But, nevertheless, symmetries do get broken. We live in a 3 dimensional world and so our bodies have 3 axes: the head-toe axis, the back-belly axis and the left-right axis. These 3 axes are at 90 degrees to each other. Our bodies are not symmetric along the head-toe axis: the head looks very different from the toe. The same is true of the back-belly axis: our back-side looks very different from our belly-side. The left-right axis is more subtle, though. Externally, our left side looks largely the same as our right, barring subtle differences. The placement of internal organs is highly asymmetric though.

Usually symmetry breaking along the back-belly axis and the head-tail axes happens first. Only then is symmetry broken along the left-right axis. And therein lies the mystery. We may not care if nature picks up the sperm entry point as the back-end and antipodal point as the belly-end, or vice versa. Whichever option it picks, one end will have the back, the other

the belly, and we'll look much the same. Once nature makes that pick, we may not even care if it interchanges the head and tail ends; a moment of thought will tell you that once you fix your back-belly axis, whether the head goes at one end of the head-tail axis or the other doesn't matter; we'll look much the same either way, once we rotate our bodies so the head comes at the top. But once the back-belly and the head-tail axes are fixed, nature has to more careful with the left-right axis. If it interchanges left and right then the liver will go on the left and the stomach on the right, very different from the usual placement. And no amount of rotation of the body will get the liver to the right, and still keep the head on top and the belly at the front. Hence the mystery: how does nature know left from right, once the back-belly and the head-tail axes are fixed?

Imagine an embryo and suppose various symmetry breaking events have blessed this embryo with distinct head and tail ends, and distinct back and belly ends. Much of what we know about embryos is from studies in other animals, mice being closest to humans.[54] So imagine a picture like the one in Fig. 5.5 showing a schematic outline of a 7.5 day old mouse embryo. It appears to be left-right symmetric. In shape, of course. But also in the distribution of various molecules, at least from what we know.

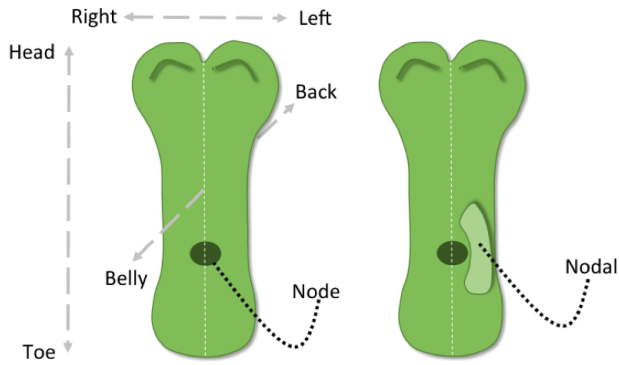


Figure 5.5: Mouse Embryos: 7.5 day old and 8.5 day old, respectively.

At some point, there is a symmetry breaking event. As a consequence,

a 8.5 day old mouse embryo is left-right asymmetric: certain molecules become more preponderant on the left than on the right. An example of such a molecule is the protein obtained by recipe interpretation of the *Nodal* gene. Then onwards, this asymmetry cascades, with *Nodal* instigating several other proteins differentially one the left. Eventually this causes some organs to form on the left (e.g., the stomach) and others on the right (e.g., the liver). The question: what is this symmetry breaking event? How does this event know what is left and what is right? Does this information come from within the embryo itself, from the mother, or from the world more broadly?

That it comes from the embryo itself was shown by studies which introduced specific variants into mice genomes.[55] These variants interfered with the usual process of breaking left-right symmetry; so symmetry was no longer broken consistently in the same direction in all individuals. Half the individuals had symmetry broken in the usual direction (so the liver appeared on the right in these) and the other half had symmetry broken in the opposite direction (so the liver appeared on the left in these). Most importantly, the direction of symmetry breaking was independent of which side the mother's liver was placed; in other words, some children had symmetry broken in one direction and other children of the same mother had symmetry broken in the opposite direction. So clearly, the symmetry breaking process was doing something independent of the mother and independent of the world at large. How does the embryo know within itself what is left and what is right?

The answer is fascinating, and clever.[54] A clump of cells call the *node* in Fig. 5.5 is the scene of action. Each cell on the belly side of the node carries a single *cilium*, a hair-like projection (Fig. 5.6). These cilia are equipped with motors (made from proteins derived by recipe interpretation from several genes) that enable them to move in circular motion.

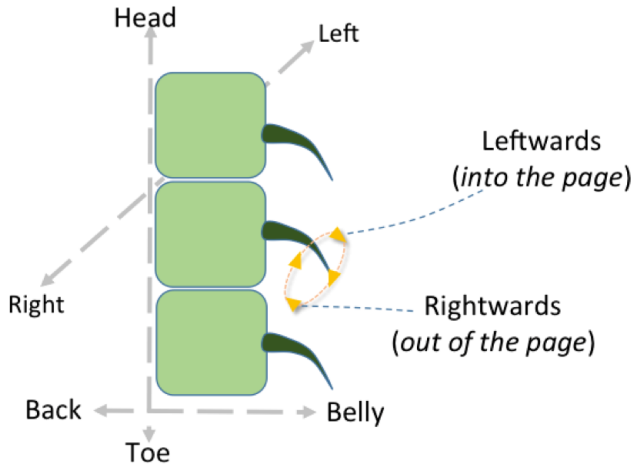


Figure 5.6: Clockwise Movement of Cilia.

Recall, the head-toe and back-belly orientations have already been established. With that, the left and right directions get defined. The circular motion of the cilia is clockwise, going into the page at the head-end and emerging out of the page at the toe-end. Note that all the cilia move in the same clockwise circular motion; in particular, it isn't the case that half the cells have clockwise-rotating cilia and the other half have counter-clockwise-rotating cilia. This clockwise circular motion causes some fluid around these cells to move. So far, nothing breaks left-right symmetry.

Now, here's the magic: the collective circular motion of all the cilia pushes this fluid to the left. Specifically to the left! Regardless of whether the embryo and mother are facing north, south, east or west, and sleeping or standing or walking! And there goes left-right symmetry.

Not so fast. Why does the circular motion of the cilia push fluid to the left? This circular motion has a leftward stroke at the top and a rightward stroke at the bottom, as shown in Fig. 5.6, and more clearly in Fig. 5.7. The former should push fluid to the left and the latter should push it to the right; these forces in opposite directions should cancel out, yielding net

zero flow of fluid.

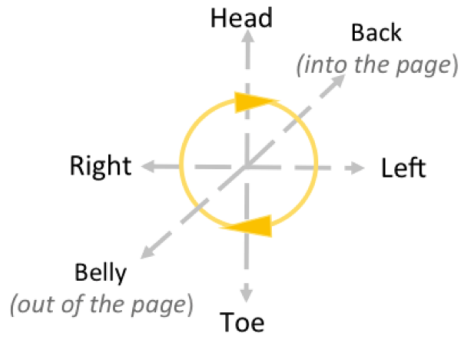


Figure 5.7: Clockwise Movement of Cilia Viewed from the Belly Side.

A small yet critical geometrical artefact helps circumvent this cancellation: each cilium, when it sticks out, points not straight ahead, but toe-wards at an angle. This, by itself, preserves left-right symmetry. Watch the circular trajectory of each cilium carefully now. This toe-wards angle causes the leftward stroke of the circular trajectory to occur further away from the cell surface than the rightwards stroke. All, still preserving left-right symmetry. Now, fluid movement is freer further from the surface. Closer to the surface, fluid experiences more friction and therefore is harder to move. So while the circular trajectory is left-right symmetric, differences in friction cause the leftwards stroke to move fluid more easily that the rightwards stroke. As a net effect, fluid gets pushed to the left.

Why do we believe this story? Because a specific genomic variant that cripples the the motor protein instrumental for the circular movement of cilia has been introduced in some mice. And lo and behold, such embryos break symmetry randomly, some in favor of left, others in favor of right, and yet others partially in both directions (which means that the *Nodal* protein mentioned above is found on both sides). So some of these embryos result in the normal placement of all organs, others result in the mirror image placement of all organs, and yet others show heterotaxy, with some organs being mirror-imaged and some not (as in our boys $\mathscr{X}$ and

$\mathscr{Y}$).[55] Even more convincingly, when the fluid is moved by artificial means in embryos with the above genomic variant to the left, symmetry is broken consistently towards the left. And when the fluid is moved by artificial means to the right, symmetry is broken consistently towards the right![56]

So fluid flow driven by cilia is the earliest known left-right symmetry breaker (though active research is on to identify other symmetry breakers[57]). Embryos somehow sense the direction of fluid flow generated by the cilia. Recipe interpretation from the *NODAL*[1] gene is then triggered in the region into which fluid flows: almost always the left side of the embryo. *NODAL*, in turn, sets off a cascade of recipe interpretation from several other genes, accentuating the asymmetry, eventually culminating in asymmetric organ formation. When genomic variants interfere with this process in one of several ways, the body plan is altered: either as a complete mirror image, or as heterotaxy (some but not all organs mirror-imaged). Such genomic variants are known in several genes. Unfortunately, *BCL9L* is not one of these. Could it really be the culprit here? For it to be so, it must create problems either in cilia formation, or in enabling cilia to rotate, or in performing recipe interpretation on the *NODAL* gene, or on other genes in its consequent cascade.

BCL9L: Is it the Culprit?

As far back as 1910, Thomas Hunt Morgan who studied fruit flies (flies that hover if you leave a basket of fruit uncovered) noticed an unusual fly: with white eyes, instead of the usual brilliant red. His pioneering studies exploring the inheritance of these white eyes established a long tradition of such genetic studies in fruit flies. Since then, scientists have bred fruit flies with various odd traits: orange eyes, legs coming out of their heads, curly wings etc. In several cases, they've traced these traits back to the causative gene in the fruit fly genome.

One such trait is winglessness. Researchers observed some wingless flies in the 1970's and eventually traced the cause back to a gene which they named *wingless*, for obvious reasons. Further research on the *wingless*

[1] Using capital letters now, as we start to shift from mice to humans; that is the convention

gene showed that it sets off a cascade of events in the forming embryo, now called *Wnt Signaling*. This cascade plays a crucial role in symmetry breaking along the head-toe axis as well and the back-belly axis.[58,53] In that process, it also plays a role in the formation of the cilia whose clockwise motion creates left-right asymmetry, as described above.[59] We will see this cascade again in a later chapter, in a very tragic circumstance. But, for now, what does this cascade have to do with *BCL9L*?

BCL9L appears somewhere in this cascade,[60] making us hopeful. But precious little is known about its exact role, just not enough to make any assessment of its role in left-right symmetry. Hoping for some clarifying insight, we turn towards the specific genomics variants that $\mathscr{X}$ and $\mathscr{Y}$ have. $\mathscr{X}$ has inherited one variant from his mother and one from his father. His mother's variant affects only one of her gene copies, and likewise for his father's variant. But $\mathscr{X}$ and *Y* have variants in both gene copies. Are these variants indeed problematic variants? Or are they innocent passers-by who just happened to be at the scene of crime?

Unfortunately, both variants are missense variants: one amino acid has changed to another in each variant. Does this change of amino acids render the gene incapable in some way? As we saw in Chapter 4, it is not easy to make this assessment. Nonsense variants and frameshifts of the type we saw in Chapter 2 and Chapter 3 are easier, because they render the gene incapable, more predictably.

One of these variants is a change of amino acid *Alanine* for *Valine* at location 185. It has been seen previously in only one healthy individual among several thousand whose genomic variants are archived in various databases (this individual has the variant in only one gene copy; the other copy of the gene is most likely in good shape, so he stays healthy). So rare enough: a possible vote for this variant. Remember our time-travel thought experiment from Chapter 4? We compare the *BCL9L* protein in several organisms to identify how long in evolutionary history this *Alanine* has been around at location 185. The longer it has been around, the greater the chance that this *Alanine* is key to the functioning of *BCL9L*, and that the change to *Valine* would be problematic. Mammals do show an *Alanine* here. But birds, frogs and fishes don't. So not that much of a vote. The other variant is a change of amino acid *Glycine* to *Aspartic Acid* at location

701. Again, birds, frogs and fishes don't have a *Glycine* here. And neither do some mammals like dogs and cats. So not much of a vote at all.

But *BCL9L* was the only candidate to come out of our analysis. As Sherlock Holmes might have said: *Once the impossible has been eliminated, whatever remains, however improbable, must be the truth*! So *BCL9L* it must be, we just have to verify this by experiment. We have to introduce these variants in a mouse or another animal and see what happens to the placement of the various organs, however time consuming that might be. So we brace for the long haul.

Mice to the Rescue

While we were holding, and quite unbeknownst to us, scientist Cecilia Lo's group at the University of Pittsburgh had embarked on an ambitious study to uncover more genes responsible for heart malformation. An example would be the unusual wiring of blood vessels into $\mathscr{X}$'s heart. The approach used in this study was even more unusual. Cecilia Lo's group first created random variants in mice genomes by treating mice with a chemical called *ethylnitrosourea* (or *ENU* for short). ENU creates a variant every 1000 or so genomic characters in the sperm cells of male mice. These mice would then pass on these variants to their offspring. The hope was that some of these variants would indeed cause heart malformation (most would of course not). These malformation-causing variants would then be isolated by sequencing the genomes of those mice which did show some malformation. The genes carrying these variants would be candidates to be studied further.

Imagine that ENU does succeed in creating a variant V causing heart malformation in a gene G. Chances are that V is introduced just in one copy of G. Most heart malformations require that both copies of G be rendered dysfunctional by variants. Therefore, for heart malformations to manifest, we need mice which have the variant V in both copies of gene G. How then do we obtain such mice, starting with mice which have V in one copy of G?

For centuries, breeders of all kinds (plant breeders, dog breeders etc) have had an answer to this challenge (Fig. 5.8). They would mate an

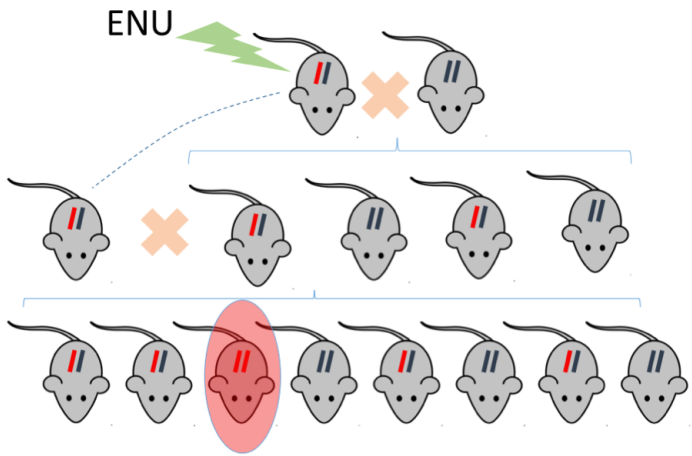


Figure 5.8: Crossing ENU-Treated Mice with Normal Mice to obtain Mice with Two Copies of the Variant.

ENU-treated male mouse carrying variant V in one gene copy (call this mouse $\mathscr{A}$) with standard female mice. The resulting offspring will inherit one copy of G from the ENU-treated father and one copy from the normal mother. Depending on which copy it inherits from its father, it will have zero or one copy of the variant V. Not yet what we want. But we can repeat this trick now. The female among these offspring are mated again with $\mathscr{A}$. Some of these females will indeed have one copy of V. Their further offspring with $\mathscr{A}$ now stand a chance of inheriting one copy of V from each parent (the actual odds are shown in Fig. 5.8). Done enough times, this will give us several mice, each with two copies of V. Of course, the chances of ENU creating a variant V causing heart malformation are not that high in the first place. So the above process has to be repeated with many many mice. Quite a mammoth exercise!

How does one know whether a mouse indeed has heart malformation? By scanning its heart using an ultrasound device. Cecilia Lo's group scanned an impressive 14,000 mice for heart malformations! And once a mouse with such malformation was found upon scanning, whole exome

sequencing was used, as in previous chapters, to isolate the causative variant. The act of fishing this variant out from several would have, no doubt, been challenging, as it was in previous chapters. But, unlike in previous chapters where just a few individuals were available in each case, this endeavor would have benefited from the availability of several mice, all of which had the same causative variant. With this marathon effort, Cecilia Lo's group found several new genes causing heart malformations. Not unexpectedly, many of these genes were involved in the formation and functioning of cilia. Unfortunately, for us, *BCL9L* was not on this list.

What Else is Possible?

Sherlock Holmes words could still ring true: *Once the impossible has been eliminated, whatever remains, however improbable, must be the truth*! *BCL9L* could well be it; it may just be that random variant generation by ENU happens not to induce the right variant in *BCL9L* and therefore it does not appear in the above list. But there is catch in Holmes's claim. Unless we have considered every possibility, eliminating the impossible is not sufficient. Have we really considered every possibility?

Recall our hunt for variants in $\mathscr{X}$ and $\mathscr{Y}$'s genomes. We considered only rare variants which either truncate the gene recipe prematurely (*nonsense* and *frameshift* variants), or modify it so one amino acid is replaced by another (*missense* variants). Should we consider variants that are not rare, i.e., found in many normal people? No, unlikely that such variants could cause $\mathscr{X}$'s rare and curious condition. Should we consider variants other than nonsense, frameshift and missense variants? We did in fact consider most relevant variant types though we haven't mentioned these here, for simplicity. What other possibility could we have missed? We insisted that our variants of interest be found in both $\mathscr{X}$ and $\mathscr{Y}$ and that they satisfy one of the scenarios in Fig. 5.3. The first two cases in Fig. 5.3 both have something in common: the child has variants in both copies of a gene while the unaffected parents have variants in exactly one copy each. Did we miss out some possibilities here?

Should we look for variants present only in $\mathscr{X}$, or in $\mathscr{Y}$, but not necessarily in both? No, both had disorderly organ placement, so both

should have the same causative variant(s). Should we look for variants present in both copies of a gene in $\mathscr{X}$ and $\mathscr{Y}$ but present in one *or zero* copies in each parent (Fig. 5.9)? Note the *zero* copies. Maybe we should consider this scenario as well. Wait a minute: if a child has variants in both gene copies, and he or she inherits one copy from each parent, should the parent not have variants in at least one gene copy? Do we really need to consider the case of *zero* copies in a parent?

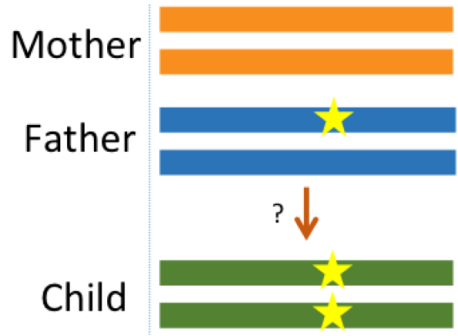


Figure 5.9: What we may have Missed.

We indeed do, as we will see, but for reasons that are not at all obvious. We will come to the reasons in due course. Yet again, we instruct our computer to wade through millions of variants and find any genes we might have missed earlier, because they had variants in both copies in the children, but variants in one or *zero* copies in each parent. We wait, hoping for something new that would change the face of this investigation. The computer produces it's new list. Much the same as the previous one. But with one notable addition. Right at the top is our new candidate gene: *XYZ99* (*you might guess that this is not a real gene name; the real name has been hidden here, pending its publication in a scientific journal*).

Unfortunately, even less is known about *XYZ99* than about *BCL9L*. No reports of any variants causing disease in humans. Little or no knowledge on what the gene does. Practically nothing useful for us. Except for one remarkable fact: hot off the press, the list, which Cecilia Lo's group

generated from their marathon effort inducing variants in mice, has *XYZ99* on it!

XYZ99: Zero Copies in the Mother?

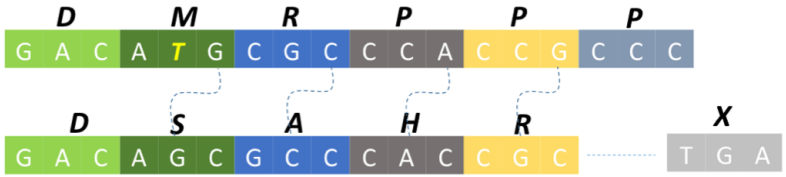


Figure 5.10: The Missing T in Yellow in a Typical Genome (top). This T deleted in $\mathscr{X}$ and $\mathscr{Y}$ causing a Frameshift in the Gene Recipe (bot). Amino acids are indicated in bold italics.

In both $\mathscr{X}$ and $\mathscr{Y}$, a single character T in the *XYZ99* gene is deleted, as shown in Fig. 5.10. This causes the triplet frame to shift (remember the triplet coding table from Fig. 2.5). The sequence of amino acids changes as a consequence, as shown. And as in Chapter 3, a *Stop* triplet comes soon enough halting the gene's recipe midway in its tracks, at amino acid 175 instead of the usual 569. For both $\mathscr{X}$ and $\mathscr{Y}$, this happens in both copies of *XYZ99*. For their father, this happens in only one copy. But, for their mother, it does not happen at all; both copies are fine!

If the mother doesn't have this variant at all, then how did both children acquire this variant in both gene copies? It is very unlikely that both children acquired the variant *de Novo*: de Novo means that the variant was introduced at the time of their conception and not inherited from their mother. Each of us has only few tens of such de Novo variants in the entire 3 billion characters of our genomes. What are the odds then that this precise variant originates in this manner, and that too in both children? Very very bleak. What else could be a possible explanation? When no further possibility suggests itself, Holmes' claim provides a prescription, the right one in this case: *either we are wrong about the mother, or we are wrong about the two children*!

How did we conclude that the children had the *XYZ99* variant in both copies but the mother in neither copy? For this, we need to go back to how genome sequencing works. When $\mathscr{X}$'s genome is sequenced, what we get in our hands are little snippets of his genome called *reads*, each only about a hundred characters long. Much like taking a book and tearing it to pieces. We have to put these pieces together to get the book back.

For this, we use the human *reference genome*, the entire genome sequence of a *healthy* individual, as a guide. Since any two human genomes are very similar to each other, $\mathscr{X}$'s genome is very similar to this reference genome. So if we get a read, say, AGGTCCTG, from $\mathscr{X}$'s genome, then this read sequence will be present somewhere in the reference genome as well, either in identical or in very slightly altered form. We have to allow for slight alterations because no two genomes are the same; on average, 1 in a 1000 characters differs between any two individuals. So a read from $\mathscr{X}$'s genome may not be present in identical form in the reference genome. This slightly altered form could change one character for another, or drop a few characters, or include a few additional characters.

So, we hunt for each read in the reference genome sequence, allowing for modifications, additions and removals of characters. When we find a hit for a read, we place that read at that location, and we get the picture in Fig. 5.11.

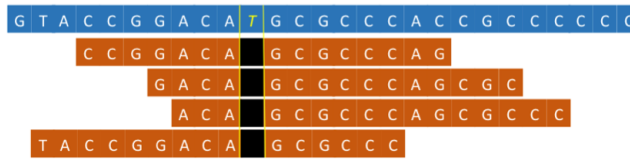


Figure 5.11: Reads in $\mathscr{X}$ and $\mathscr{Y}$ with the Alterations in Black, juxtaposed against the Reference Genome (in Blue).

As this picture shows, every read in $\mathscr{X}$ is missing a T. That missing T has been replaced by an additional blank, so the remaining characters are in register with those in the reference. This blank signifies that the character T in the reference genome is actually missing from $\mathscr{X}$ completely (i.e., in

both copies). Ditto for $\mathscr{Y}$.

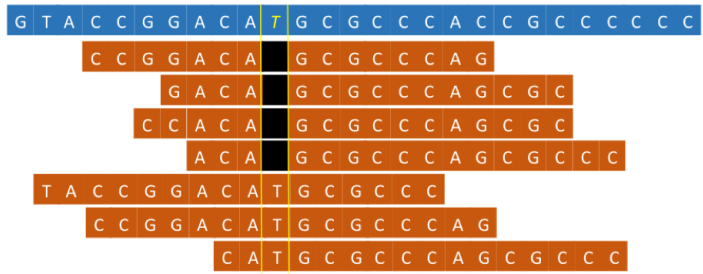


Figure 5.12: Reads in the Father.

For their father, roughly half the reads have this additional blank (Fig. 5.12). Since one expects roughly half the reads to come from one gene copy and half from the other, the father is missing this character in only one gene copy. And the mother? Her reads are shown in Fig. 5.13. All these reads have the T in them, so they do not require the additional blank. In other words, the mother appears to not have this variant at all. Both gene copies have the T. So our conclusion appears solid. Then what are we missing?

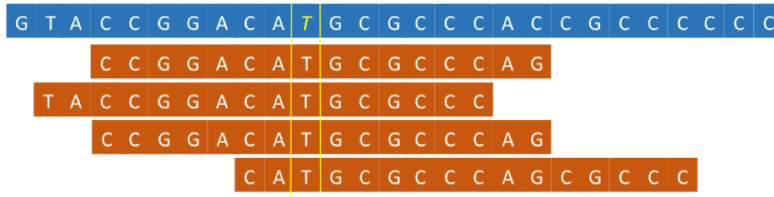


Figure 5.13: Reads in the Mother.

The sharp eye will notice one curious fact. The number of reads in each of $\mathscr{X}$, $\mathscr{Y}$ and their mother is roughly half of what it is in the father. Why should that be so? Could this suggest that $\mathscr{X}$, $\mathscr{Y}$ and their mother were missing one complete copy of *XYZ99*? Or at least a large chunk of

XYZ99 around the T in question?

This scenario is illustrated in Fig. 5.14, which shows the two copies of *XYZ99* in each of the individuals.

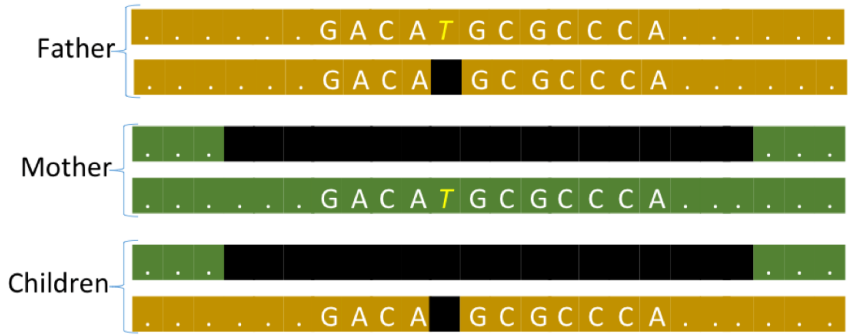


Figure 5.14: The Two Copies of *XYZ99* in each Individual, with Deletions shown in Black.

In this new scenario, the missing T we discussed above applies to only one gene copy in both $\mathscr{X}$ and $\mathscr{Y}$. Ditto for their father. What happens to the other gene copy? In the father, the other gene copy is intact. But, most importantly, this other copy has a large missing chunk in both $\mathscr{X}$ and $\mathscr{Y}$. This large missing chunk has been inherited from their mother, who also has this chunk missing in one of her gene copies. The other gene copy in the mother is intact. Could this new scenario have tricked us into believing that $\mathscr{X}$ and $\mathscr{Y}$ are missing the T from both their gene copies?

Let us see. The father then has two almost complete copies of *XYZ99* and each copy would be read, say, x times by the sequencing process, providing $2x$ reads in Fig. 5.12. But the mother and both children have a large chunk missing in one copy; there would be nothing to read from in that copy. The other copy would be read x times (roughly speaking) yielding half the number of reads as in the father. All the reads we see in Fig. 5.11 and Fig. 5.13 will come from this other copy. In $\mathscr{X}$ and $\mathscr{Y}$, this other copy is missing the T, hence all the reads will show a missing T, tricking us into believing that both copies are missing the T, when, in fact, only one of the copies is! In their mother, this other copy has the T intact,

hence none of the reads will show a missing T.

So, if our hypothesis of a missing chunk in the mother and in the children is correct, $\mathscr{X}$ and $\mathscr{Y}$ have the missing T is just one copy, not in both copies as we mistakenly thought earlier. The missing T is in the copy they inherited from their father. What they inherited from their mother had a big chunk missing from the *XYZ99* gene. That would explain it! If only we could pinpoint this large missing chunk.

Sequencing the Missing Chunk

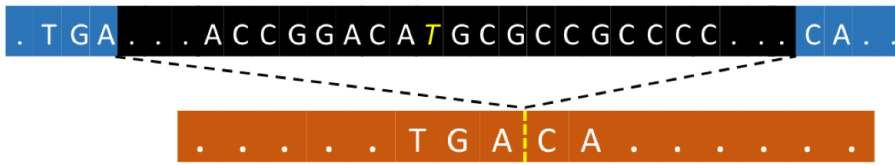


Figure 5.15: The Reference Genome (top) with the Chunk Missing from the Mother in Black. A Read derived from the Mother's Genome (bottom) straddling the Missing Chunk.

Imagine the mother's genome, with a large chunk of the *XYZ99* gene missing in one copy. Most of us have this chunk and therefore so does the reference genome. But the mother doesn't have this chunk, that too in only one of her two gene copies. This chunk is indicated in the reference genome in Fig. 5.15. Genome sequencing of the mother will then generate reads straddling this missing chunk. Each such read has a prefix derived from the left of the chunk and a suffix derived from the right of the chunk. The yellow dotted line demarcating these two parts is purely imaginary; the read does not come with this line. That, in fact is the problem. We have no way of knowing which reads have this yellow imaginary line, from the billions or so reads that genome sequencing gives us.

What happens when we hunt for one such read in the reference sequence, in its original form or in slightly altered form? Remember we had to allow an additional blank in the reads in Fig. 5.11 to allow for the

single missing T in these reads. A single additional blank would qualify for slightly altered form. Here, the mother is missing a much larger chunk. So we would have to insert a large number of additional blanks in the read; that would certainly not qualify for slightly altered form. Therefore, chances are that our hunt for these reads in the reference genome would not have been successful. It is typical for few tens of millions of reads to not find their anchor point in the reference genome, for various reasons. These reads would have been parked aside and ignored thereafter. Hidden among these tens of millions of reads might just be the handful we need: the handful that straddle the large missing chunk. How could we could find these few needles in such a massive haystack?

We have to take each read from this haystack and then launch a hunt for this read in the reference genome. This time, slight alteration is not sufficient; we have to allow for substantial alteration, i.e., many more additional blanks. We don't even know how many. And we have no idea where the yellow dotted line lies: towards the beginning, or towards the middle, or towards the end? The read itself is, say, a 100 characters long. Too long for us to try every possibility exhaustively. We need a faster method.

Read Hunting with Few Alterations

Let us start with the hunt for a read in the reference genome allowing only a few alterations. Say 3 alterations. To this end, consider one read.

The reference genome is huge, 3 billion characters long. It takes way too long to check for a match of this read at every one of these 3 billion places. Fortunately, there are clever ways to quickly shortlist a small handful of candidate places in the reference genome where a match with slight alteration might possibly occur.[61] At each of these places, we need to check whether the read at hand does indeed match with slight alteration. So consider a particular place in the reference genome (Fig. 5.16). How do we identify the minimum number of alterations needed to make the read match at this place?

Two of many ways to alter the read are shown in Fig. 5.16. In the first way, one character is modified and 3 blanks are added to the read,

yielding a total of 4 alterations. In the second way, 1 blank is added and one character is removed, yielding a total of 2 alterations. There are many many such ways. And each requires a different number of alterations.

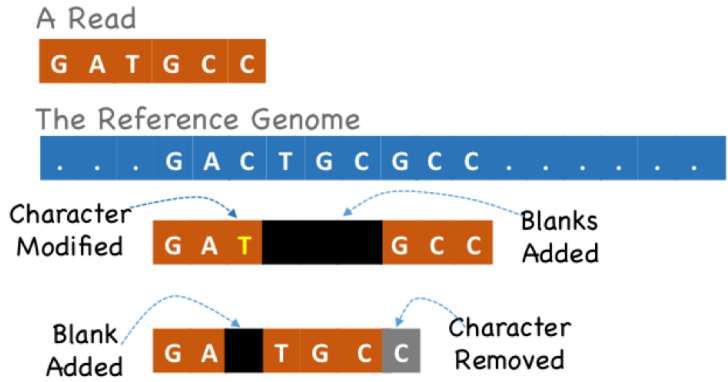


Figure 5.16: Two Ways to Alter a Read.

Our goal is to identify the one with the least number of alterations. And we have to do this without trying every possible way explicitly, because there are just too many ways. Here's how.

A picture, they say, is worth a thousand words. In that spirit, Fig. 5.17 concisely captures every one of the myriad ways to make alterations. It shows a grid of points with edges which go rightwards, downwards or diagonally downwards. Note that rightwards and downwards edges are shown dotted, while diagonal edges are sometimes dotted and sometimes solid. Can you guess when they are solid? When the reference genome character in that column and the read character in that row are identical. So what has this grid got to do with read alterations? It turns out that every distinct way of altering the read corresponds to a path in this grid from the pink dot at the top-left to the pink solid line at the bottom.

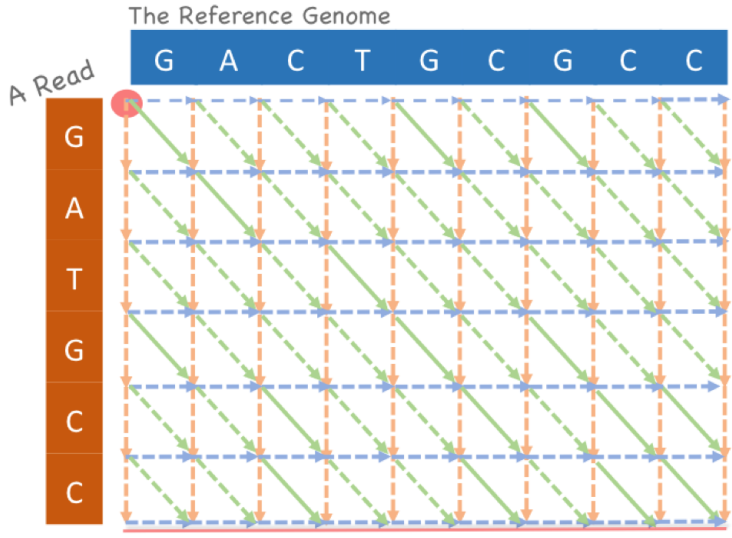


Figure 5.17: All Possible Alterations in a Picture.

For example, consider the two ways to alter the read shown in Fig. 5.16. For each of these two ways, the corresponding paths are shown in Fig. 5.18. The rule: we go right when we add a blank, we go down when we remove a character, and we go diagonally otherwise.

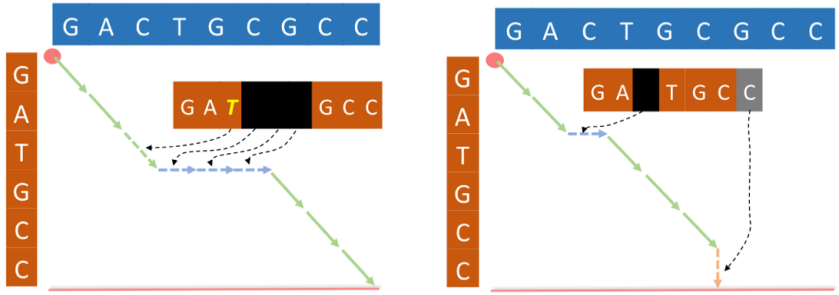


Figure 5.18: Paths Corresponding to the Two Ways to Alter from Fig. 5.16.

Just as there are many many ways to alter the read, there are also many many such paths. The number of alterations corresponding to any such path is just the number of dotted lines encountered on that path. This is because an alteration is needed only if you add a blank (go right), or if you remove a character (go down), or if the read character is not the same as the corresponding reference genome character (go diagonally on a dotted line); in all three scenarios, you cross dotted lines. The goal is to minimize the number of alterations by taking solid lines as far as possible. So our task boils down to finding the path with the fewest dotted lines from the pink dot at the top-left to the pink solid line at the bottom. And we have to find this path without going through every such path explicitly.

The trick lies in calculating the path with the fewest dotted lines from the pink dot to every one of the grid-points in Fig. 5.17, and not just to the pink solid line at the bottom. Isn't that more work than is needed? Only at first sight. It will actually be less work on the whole, as will see. We will consider the grid-points in a certain order, shown in Fig. 5.19. First, we consider all grid-points in the diagonal marked 1. Then all grid-points in the diagonal marked 2, then 3, and so on.

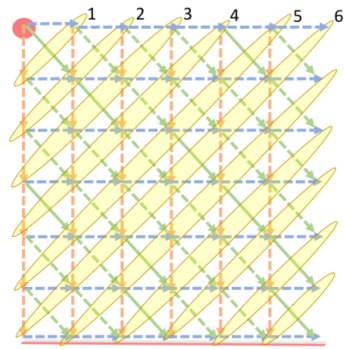


Figure 5.19: Processing Diagonals in Order.

What do we do for each grid-point? Consider one such grid-point E, and suppose it is on diagonal 4 (Fig. 5.20).

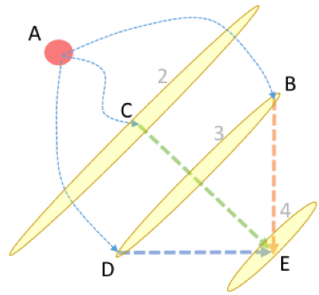


Figure 5.20: Processing One Grid-Point.

Any path from A to E must go through one of the points B, C or D. C is on diagonal 2, and B, D are on diagonal 3; so we already know the paths with the fewest dotted lines from A to each of these points. Next, we need to consider just 3 possibilities for the path with fewest dotted lines from from A to E:

- Take the path with the fewest dotted lines from A to B and then go one further step downward along a dotted line.
- Take the path with the fewest dotted lines from A to D and then go one further step rightward along a dotted line.
- Take the path with the fewest dotted lines from A to C and then go one further step along the diagonal line, which may be solid or dotted.

By taking the path with the fewest dotted lines among these 3 possibilities, we have our answer for grid-point E in just 3 steps. We can do this for every grid-point in diagonal 4. Once done, we proceed to diagonal 5, and so on, until all grid-points are done. Once all grid-points are done, we have our answer.

How many grid-points are there in all? Since we allow at most 3 alterations, we can go rightward or downward at most 3 times; we have to stick to diagonal moves the rest of the time. Which means we need to consider only grid-points in the band shown in Fig. 5.21: up to 7 grid-points on each row, 3 to the right and 3 to the left of the diagonal.

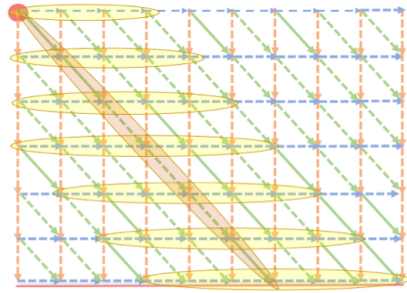


Figure 5.21: Relevant Grid-Points.

For a read of length 100, this band has just $100 * (3 + 3 + 1) = 700$ grid-points. At 3 steps per point, 2100 steps is all we need to identify the minimum number of alterations needed to make the read match at this place in the reference genome. In contrast, if we have to consider all possible ways of making up to 3 alterations, the number of steps would be well above 100,000! Quite some savings, particularly since we have tens of millions of reads to consider.

Read Hunting with Many Alterations

To identify the missing chunk in $\mathscr{X}$'s and $\mathscr{Y}$'s mother, we need to allow for many more alterations. We don't even know how many, since we do not know the size of the missing chunk. If the chunk were a million long, we would need to allow a million alterations. Which would mean about 200 million steps in the procedure above for this read! Multiply that by the tens of millions of reads we need to sift through. This is not going to be quick.

How do we discover our missing chunk quicker than that? Take a look again at Fig. 5.15. Each read which straddles the missing chunk in the mother has two parts: the part to the left of the chunk, and a part to the right of the chunk. What if we broke up the read into these two parts; a prefix and a suffix? Each part, by itself, would match with just a few alterations. The prefix would match where the chunk starts, and the suffix

would match where it ends. So we just split each read into two parts and hunt for each part separately, allowing slight alterations in each case, as we did earlier. As simple as that! And we repeat this for every one of the tens of millions of reads in our haystack.

There is a small catch though. There are 99 different ways to split a read comprising 100 characters into two parts: you could split after the first character, or after the second, and so on. How do we find the right split? One can almost imagine Dr. House from the American medical drama, House, barking at us: *try ALL of them*. Too laborious? It turns out that there is a clever way to do this implicitly, so one discovers the right split as one goes along rather than follow Dr. House's instructions to the letter. We'll spare the details, but in summary, the tens of millions of reads in our haystack are set rolling on this procedure. Entrusted to a computer, of course. Once that finishes, which it does quite quickly, we scan the reference genome for two locations: the first, where prefixes of several reads stack up, and the second, where suffixes of several reads stack up. And we look to see if these two stacks sandwich any part of the *XYZ99* gene. And what do we find?

The Missing Chunk, Finally

Staring at us finally is the answer to our riddle of $\mathscr{X}$ and $\mathscr{Y}$'s curious architectural plan: a sharp stack of read prefixes and another of read suffixes sandwiching the first three exons of *XYZ99* in the mother (Fig. 5.22). The missing chunk, obvious from this picture, starts about 2000 characters before the start of *XYZ99*. This is no-man's land, no genes or exons here. So only whole genomes sequencing would have helped us identify the start of this chunk; just sequencing the exons may not have been sufficient.

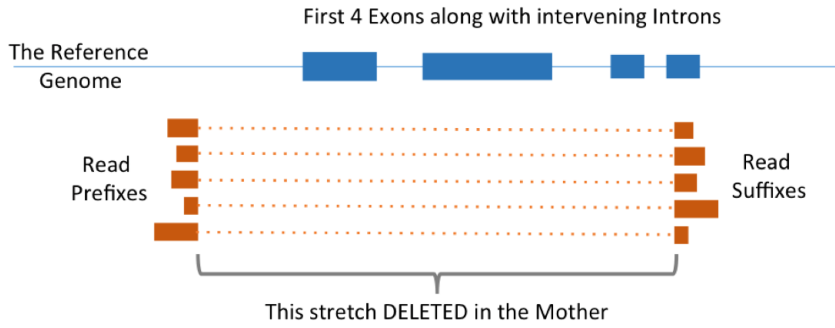


Figure 5.22: The Missing Chunk in *XYZ99*.

The missing chunk ends in the middle of the fourth exon of *XYZ99*. The chunk has a total length of 5925 characters. These 5925 characters are all missing from one copy of the gene in the mother. In all likelihood, this copy of the gene is rendered dysfunctional by the missing chunk. In spite of this, the mother's organs are all in the right place, rescued by the other normal gene copy.

We repeat the above procedure on both $\mathscr{X}$ and $\mathscr{Y}$ to verify that both indeed inherited the abnormal gene copy from their mother, the missing chunk and all. As expected, the missing chunk appears yet again. Both children have inherited the missing chunk from their mother, making the corresponding copy of their *XYZ99* gene dysfunctional. And the missing T which they inherited from their father causes the recipe in the other gene copy to be truncated prematurely. Taken together, $\mathscr{X}$ and $\mathscr{Y}$ have no functioning copies of *XYZ99* at all! And that is the likely cause of the odd placement of their organs.

Wrapping Up

$\mathscr{X}$'s and $\mathscr{Y}$'s curious case has lead to the identification of a new gene *XYZ99* for heterotaxy (some but not all organs mirror-imaged) and heart malformation. Variants in *XYZ99* had never been seen hitherto in any

human patient and little is known about its functions. Yet, we have strong evidence that *XYZ99* is indeed the cause of $\mathscr{X}$'s and $\mathscr{Y}$'s condition, thanks to the marathon efforts of Cecilia Lo and her group in screening thousands of mice to identify one with an *XYZ99* variant carrying similar heterotaxy and heart malformation features as $\mathscr{X}$ and $\mathscr{Y}$. The initial *XYZ99* variant they found was a missense variant: a change from amino acid *Tryptophan* to *Leucine* at amino acid 177. Subsequently, they also created mice with both copies of *XYZ99* knocked-out (as they effectively are in $\mathscr{X}$ and $\mathscr{Y}$) and verified that these mice too showed heterotaxy and heart malformation, as expected. So *XYZ99* is indeed the cause of $\mathscr{X}$'s and $\mathscr{Y}$'s curious condition.

Having said that, we don't know how *XYZ99* perturbs the usual process of left-right symmetry breaking, resulting in this odd placement of organs. Experiments seem to suggest that recipe interpretation from *XYZ99* is apparent only in a 10.5 day old mouse embryo.[62] Since moving cilia break left-right symmetry between day 7.5 and 8.5, *XYZ99* very likely plays its role in the subsequent cascade of events and not in the formation or functioning of the cilia. Only further experiments will uncover its exact role.

$\mathscr{X}$ and $\mathscr{Y}$ have both needed heart surgery for survival. Their parents now know the exact reason why two of their children were born with unusual anatomical features requiring surgery for survival, while they themselves, and their other children, had the standard anatomical plan. This information could be invaluable should they choose to have another child, as we saw in Chapter 4.

It did take a while to get to the bottom of this case. The answer lay buried all along under huge mounds of data. Fishing this needle out of this gargantuan haystack required a most unlikely combination: very careful sleuthing, some clever computer algorithms, and a marathon effort on screening mice!

Chapter 6

The Blood Can't Carry

$\mathscr{X}$, a two year old girl, had an unusually pale complexion. She appeared malnourished, was often short of breath, and seemed far less energetic than other infants of her age. She fell sick frequently. Her liver and spleen seemed enlarged. A blood test which observes a drop of her blood under a microscope was performed. It clearly showed that $\mathscr{X}$'s red blood cells were paler and more abnormally shaped than normal. $\mathscr{X}$'s doctor had no trouble diagnosing her condition: Beta-Thalassemia Major.

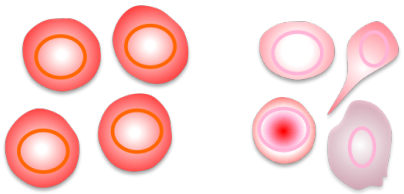


Figure 6.1: Normal Red Blood Cells and Thalassemic Red Blood Cells

The redness of red blood cells comes from a molecule called hemoglobin. The paleness of these cells in $\mathscr{X}$ indicated that she had abnormally low levels of functional hemoglobin. And hemoglobin is instrumental in carrying oxygen from the lungs to various parts of the body. So the ability of $\mathscr{X}$'s blood to carry and deliver much-needed oxygen was severely compromised. Left untreated, this could be fatal.

The immediate corrective course of action was a blood transfusion: blood from a willing donor was transfused into $\mathscr{X}$. This provided her with

a supply of red blood cells equipped with functional hemoglobin. But red blood cells in the body live on an average for just 120 days, after which they are destroyed. $\mathscr{X}$'s body would need to generate new red blood cells to compensate for these dying cells. But her body could only create red blood cells with defective hemoglobin. Which meant that $\mathscr{X}$ would need repeated blood transfusions, once a month or so. Indeed $\mathscr{X}$, now three and a half, had undergone several such transfusions.

What quirk in $\mathscr{X}$'s genome rendered her hemoglobin incapable of carrying its usual load of oxygen? Would she be able to have repeated transfusions, life-long? What lay in store for her?

The Oxygen Carrier

The blood is our lifeline, carrying oxygen from our lungs and delivering it to the rest of the body. Some amount of oxygen from the lungs simply dissolves in the blood, much as air dissolves in water. However, this amount is simply insufficient to meet the needs of the body. The blood needs an oxygen sponge to attract, carry, and then deliver much larger amounts of oxygen. Hemoglobin is indeed this sponge.

The recipe for the creation of this hemoglobin is written in the genome. Not in a single gene, but over two different genes, called *HBA* and *HBB*. The recipe described in each gene is interpreted separately and then the resulting products combine together to make hemoglobin. The recipe in the *HBA* gene is interpreted not once but twice to yield two copies of a resulting molecule. Similarly, the recipe in *HBB* gene is also interpreted twice to yield two copies of another molecule. These four resulting molecules, two of each kind, are then packaged together into a single entity, shown in Fig. 6.2. This entity is called *globin*. Not yet hemoglobin though, that needs one more step.

Globin, by itself, is not capable of carrying oxygen. It needs an additional attachment called *heme* for this purpose. There are four such attachments, one wedged deep inside each of the four globin sub-parts. However, unlike these globin sub-parts, the recipe for heme is not directly coded in any gene. Rather, the body makes heme from raw materials which we obtain via food. Together, four hemes and the four globin sub-parts

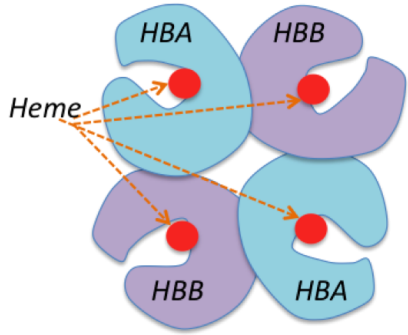


Figure 6.2: Hemoglobin.

make a single molecule of hemoglobin. And each red blood cell in our blood contains around 250 million hemoblobin molecules. And each of us has approximately 25 trillion red blood cells. That is a lot of hemoglobin!

At the center of each heme molecule is an *Iron* atom. This is where oxygen attaches itself. In the lungs, where lot of oxygen is available, only so much dissolves in the blood; the rest, a much larger amount, is soaked up by the iron in heme. This attachment of oxygen to heme is what gives oxygenated blood its brilliant red color. The red blood cells then travel through the bloodstream to other parts of the body. The small amount of oxygen dissolved in blood gets used up quickly in these parts. The resulting drop of oxygen concentration serves to squeeze the sponge, forcing oxygen to dissociate from heme and making it available to the various tissues.

An intricate structural quirk in hemoglobin enables it to soak up and deliver large amounts of oxygen very quickly. Imagine a hemoglobin molecule, with no oxygen attached to any of its four heme attachments. (Fig. 6.3).

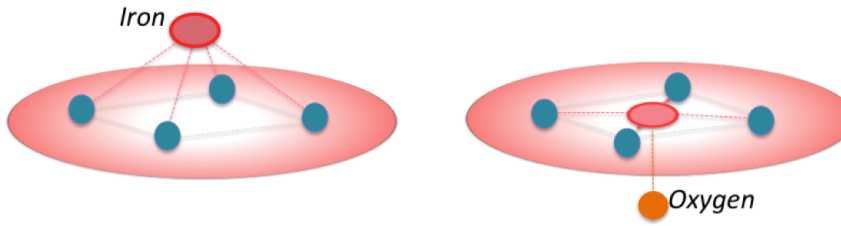


Figure 6.3: Heme: without, and with, attached Oxygen.

Each heme molecule has a pyramidal shape around the iron at the center. In this form, hemoglobin shows relatively low interest in oxygen. When oxygen does attach to the iron in any of the four heme molecules, the structure of that heme flattens out. This flattening has a cascading effect, forcing structure adjustments in all four globin sub-parts and enabling oxygen to attach to the other three hemes much more easily. So attachment of oxygen to one heme makes it easier for oxygen to attach to the other three hemes. This allows hemoglobin to soak up oxygen quickly in the lungs. Likewise, detachment of oxygen from one heme makes it easier for oxygen attached to the other three hemes to detach. This lets hemoglobin deliver oxygen quickly to other parts of the body.

Such intricate structural orchestration is not without its accidental quirks. Carbon monoxide attaches to the iron in heme even better than oxygen. In fact, 250 times better. Thankfully, the amount of oxygen in our atmosphere is a million times that of carbon monoxide! If, for whatever reason, the amount of oxygen were to reduce to only 250 times that of carbon monoxide, 50% of our hemoglobin would be busy carrying carbon monoxide instead of oxygen, starving our cells of necessary oxygen, fatally so! Hopefully, there is no immediate threat of carbon monoxide levels reaching anywhere near such dangerous levels. The greater threat, at least for the moment, lies in various altered gene recipes that prevent the oxygen sponge from functioning as intended, as seemed to be the case with $\mathscr{X}$.

Sticky Hemoglobin

Altered gene recipes compromise hemoglobin in various ways. Sometimes, by reducing the snugness with which the heme attachment fits into globin, thus spoiling the intricate structural orchestration that is required for soaking-in and dispensing oxygen. And sometimes, in other, rather surprising ways.

An example is a single character A to T flip in the *HBB* gene. This single character flip results in amino acid *Valine* at position 6 instead of the usual *Glutamic Acid* (see Fig. 2.5). Fortunately, the snugness of heme inside globin remains unaffected by this change. But something far more unexpected happens: hemoglobin molecules, which are tightly packed in red blood cells, start sticking to each other at amino acid 6 and forming chains. These chains are quite rigid, so they force deformation of red blood cells from their usual round shape into a sickle-shape.

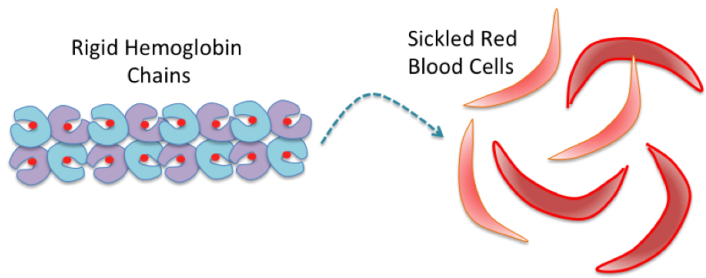


Figure 6.4: Hemoglobin Molecules Sticking Together Causing Sickling.

So what? The diameter of a red blood cell is about 7 microns (a micron is a millionth part of a meter). Many arteries and veins are much wider, allowing easy passage for these cells. But arteries and veins are major highways. Leading off these highways is an elaborate network of much smaller blood vessels called *capillaries* which take blood to every corner of the body. Many of these capillaries have diameters which could even be smaller than that of a red blood cell. Red blood cells have to literally squeeze through these capillaries!

The shape and flexibility of red blood cells is what allows them to squeeze through capillaries smoothly. But when forcefully deformed into a more rigid sickle-shape, this no longer applies. These red blood cells can no longer squeeze through easily. This leads to blockage, increased wear-and-tear, and consequent death of red blood cells. The lifetime of a red blood cell is markedly reduced from 120 days down to about 20 days. The result is a shortage of red blood cells, and therefore a shortage of hemoglobin that can deliver enough oxygen. And blockage of capillaries leads to pain and sometimes organ damage. As a result, lifespan is markedly reduced. For obvious reasons, this disease is called *sickle cell anemia*.

In certain tropical regions of Africa and India, more than 10% of us carry the aberrant *Valine* in one copy of our *HBB* genes. Fortunately, only those who carry this variant in both copies of *HBB* are affected seriously; still several hundreds of thousands of new patients each year, worldwide. $\mathscr{X}$ was not one of these patients though. Her red blood cells were not sickle-shaped. Chances are that she had a different type of recipe alteration.

Deficient Hemoglobin

Hemoglobin is a joint product of two genes, so recipe alterations in either gene could be problematic. Oxygen delivery is, of course, very critical, so nature has provided an extra level of protection. Each of us has not just two, but four copies of *HBA*, along with the usual two copies of *HBB*.

Figure 6.5: Copies of *HBA* and *HBB* Genes

Remember, both *HBA* and *HBB* contribute equally (two sub-units each) to hemoglobin. Yet we have more copies of *HBA* than *HBB* in our genome. Our cells do a counter-balancing act by running recipe interpretation on

HBA at half the rate of *HBB*, so equal amounts of protein are generated from both genes. Altered recipes sometimes upset this fine balancing act.

For instance, altered recipes in the *HBA* gene could lead to loss of some of the copies of this gene. This, of course, leads to a shortage of good *HBA* protein. Fortunately, this shortage has only mild impact if one or two copies are lost. Luck runs out when three copies are lost. The body can generate only small amounts of regular hemoglobin from the only remaining good copy. Oxygen delivery is then compromised substantially. The consequent symptoms vary from mild to requiring occasional blood transfusions. Of course, when all four copies of *HBA* are lost, the impact is drastic: death before or shortly after birth.

Similarly, if altered recipes in the *HBB* gene lead to loss of one copy of *HBB*, the impact is very mild. Things are more serious when both copies are lost. Then there is no *HBB* protein to create regular hemoglobin with. Some oxygen is still delivered, either dissolved in the blood, or carried by hemoglobin-like molecules created from combinations of *HBA* and genes other than *HBB*. Not enough, though. This leads to a serious condition. More serious even than the loss of 3 copies of *HBA*. Periodic blood transfusions which provide good red blood cells from donors become essential.

The loss of both copies of *HBB* is not as immediately lethal as the loss of all four copies of *HBA*. Because a small amount of hemoglobin can be created by combining *HBA* with other genes, but no hemoglobin can be created without *HBA*. However, as one would expect, loss of all four copies of *HBA* is a very unlikely event. Even the loss of three copies of *HBA* has low odds. But the loss of one or two copies of *HBB* is far more common. Studies on children in Mumbai and New Delhi report that, on average, 1 in 25 individuals is a carrier of a problematic variant in one copy of *HBB*.[63] Roughly one in each classroom! Most of these individuals have problematic variants in only one gene copy and are unlikely to be affected in any serious way. However, at one child per classroom carrying a problematic variant, it is no surprise that several children are indeed born with problematic variants in both *HBB* copies. $\mathscr{X}$ was likely to be one such child. How was her *HBB* recipe altered?

The Search for the Genomic Variant

Whole Genome Sequencing and *Whole Exome Sequencing*, our windows to peek into genomes in previous chapters, are probably an overkill to answer this question. These methods consider all 21,000 or so genes in our genome, getting us the text stretches in all exons of all these genes. These are powerful tools that cast a wide net. But the clinician treating $\mathscr{X}$ has a very specific guess on the gene that caused her problem, namely, the *HBB* gene. Could we get the sequence of this gene alone without necessarily going through the very elaborate process of casting a wider net?

The *HBB* gene makes things particularly convenient in this regard. It is all of 1605 characters long, spread over three exons and two introns. Very small as genes go; the average gene is much longer about 10 times as much.

Figure 6.6: The *HBB* Gene: Three Exons and Two Intervening Introns.

It is easy to just sequence a short gene directly instead of casting a wider net. So we sequence this gene and identify all variants, i.e., characters where $\mathscr{X}$ differs from a supposedly healthy genome sequence. Of course, our focus is on the three exons of *HBB*, since the gene recipe is encoded in the exons; remember, the introns are skipped in the process of recipe interpretation. There must be a problematic variant here in these exons, vindicating $\mathscr{X}$'s clinician's guess. At first sight, there isn't one. We check and recheck. But no variant here!

Splicing in Recipe Interpretation

Where was the elusive variant responsible for $\mathscr{X}$'s condition hiding? It wasn't in the three exons of *HBB*. Where else could it be? Could it be in the introns? Or was $\mathscr{X}$'s doctor mistaken about the *HBB* gene altogether? If it were indeed in the introns, then how could it alter the gene recipe?

Remember again, recipes are written in the exons of a gene. The intervening introns do not carry this recipe. These introns could be very long, a few thousand text characters on average, but going up to tens of thousands of text characters in some cases.[64] The recipe interpretation machinery knows how to skip these introns (or, to use the right terminology, how to *splice out* these introns). It does so by looking for exon boundary markers. The end of an exon is marked by two characters, invariably GT. The beginning of an exon is also marked by two characters, invariably AG. Using these markers, both introns of *HBB* are first spliced out by the recipe interpretation machinery. Subsequently, characters in the exons are partitioned into frames of 3 and then the table in Fig. 2.5 applied to convert each frame to an amino acid.

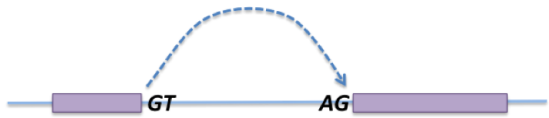


Figure 6.7: Recipe Interpretation Jumps from Exon End Mark to Exon Start Mark.

Of course, if genomic variants change either the GT or the AG to something else, these splices go haywire. For instance, if the GT becomes, say GG, then the splice fails to initiate, as shown below.

Figure 6.8: Recipe Interpretation Continues Into the Intron when GT becomes GG.

Recipe interpretation then continues into the intron rather than jumping across the intron to the next exon. This intronic recipe can be quite different from the recipe described by the next exon located on the other shore of this

intron. Which means that the protein generated is now quite different. And, very likely, as was the case in Chapter 3, this intronic recipe encounters a *Stop* instruction resulting a severely truncated recipe.

A similar situation would arise if the exon start mark AG became, say AT. Recipe interpretation launches its jump at the GT but does not land where it should; chances are it lands somewhere else, maybe much further away, possibly at the end of the next intron, splicing out both the intron as well as the next exon completely. This again generates a very different, and possibly truncated protein.

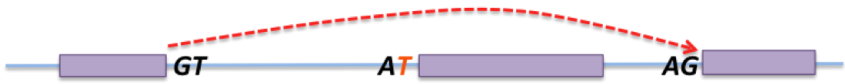


Figure 6.9: Recipe Interpretation Skips Both Introns and an Exon when AG becomes AT.

So changes at the GT or AG pairs flanking each exon have a drastic impact on the recipe. Maybe this was the source of $\mathscr{X}$'s problem? So we look for variants in $\mathscr{X}$'s genome flanking each of the 3 exons of *HBB*? Again, we draw a blank!

Long Hops and Short Steps

Where could $\mathscr{X}$'s variant be hiding? Understanding the splicing process better might yield a clue. So consider, for instance, the entire sequence of characters in the second intron of the two *HBB* introns. Not $\mathscr{X}$'s sequence, that'll come later. Just the sequence in the average person. As expected, it begins with a GT marking the end of the previous exon. And it ends with an AT marking the beginning of the next exon.

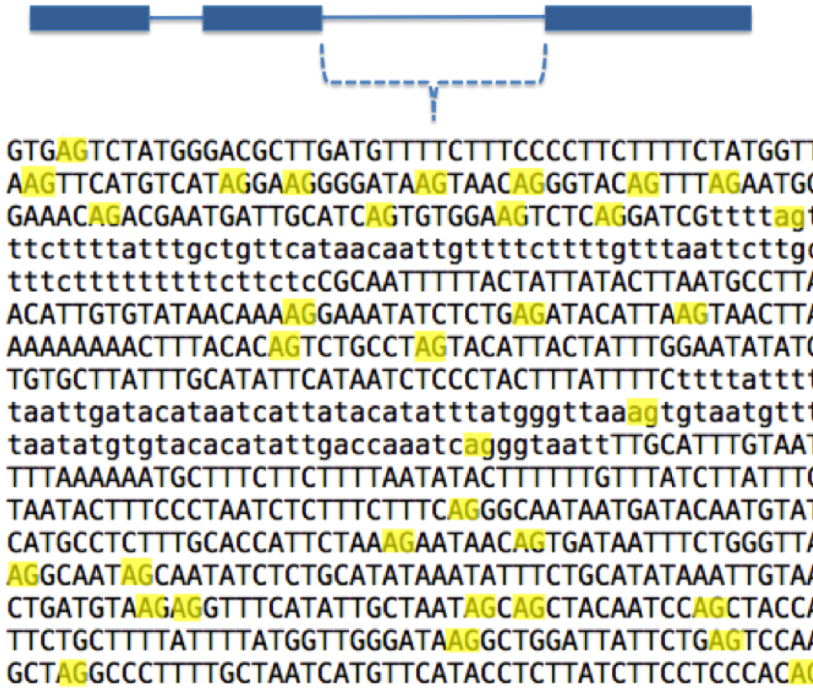


Figure 6.10: 34 AGs in the Second Intron of *HBB*.

A question pops up immediately. There are as many as 34 AG's inside this intron. But the last of these AG's, the one at the very end, is special. When the splicing process launches a jump at the GT, it picks this AG instead of 33 other AG's for its landing. What guides the splicing process to this special AG?

An intermediate character called the *Branch Point* does the honors. This point breaks the long jump into two parts: a long hop, and a step.[65] The long hop lands at the branch point, close to the eventual destination. The step bridges the remaining gap between this branch point and the special AG at the end.

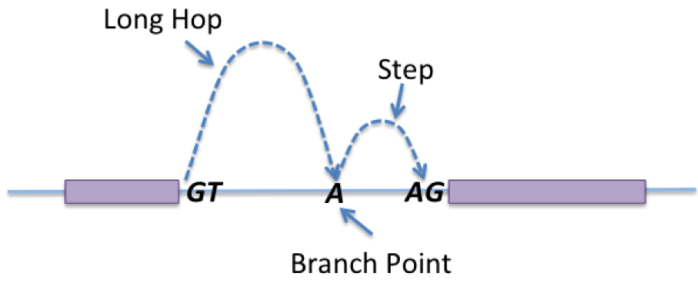


Figure 6.11: The Long Hop and the Step.

That begs the next question. Where in the vast intronic wilderness does the branch point lie? A few facts about branch point help us in this search. First, we know that the character at the branch point is most often an A.[66] There are well over a 100 A's to choose from in Fig. 6.10. But only those whose surrounding characters match a certain pattern, shown below, are candidates for the branch point.

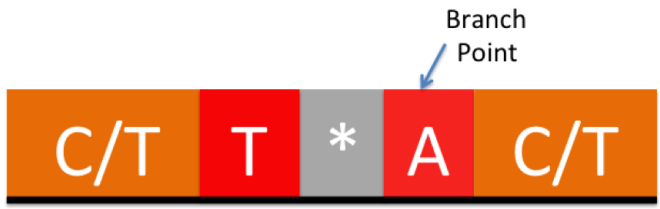


Figure 6.12: The Branch Point Pattern.

The pattern above indicates that the character two steps to the left of this A is almost always a T. The character in between the two could be anything (hence the *). And the two sandwiching locations show a preference for C's and T's. This preference is not overwhelming: occasionally A's or G's might appear at these locations; hence they are depicted in orange instead of red in Fig. 6.12. But even considering this preference to be

absolute, there are still 27 such candidates in Fig. 6.10. Which of these is the right branch point?

Some more facts lead us further in this search. A bit to the right of the branch point must lie a contiguous stretch of characters comprised predominantly of C's and T's. The occasional A or G could appear here but C's and T's are predominant. This stretch is typically about 20 characters long,[66] with the first few characters more likely to be T's than C's.

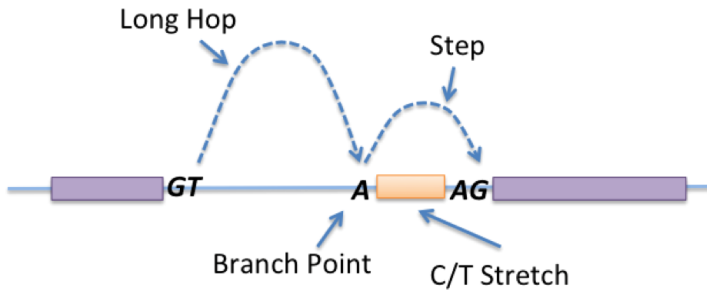


Figure 6.13: The C/T Stretch Between the Branch Point and the AG.

A careful look at Fig. 6.10 shows that only 7 of 27 branch point candidates have such an accompanying stretch. So our hunt for the branch point is down to 7 candidates now. A final property of branch points now leads us to the likely answer. The branch point is typically within 40 characters of the last AG shown in Fig. 6.10.[67] Only the very last of our 7 candidates fits this bill. And that is our branch point in all likelihood!

Here is how the picture looks around this branch point. Notice the CTAAT pattern around this branch point matches the pattern in Fig. 6.12. And the stretch shown in blue comprises primarily C's and T's, as required.

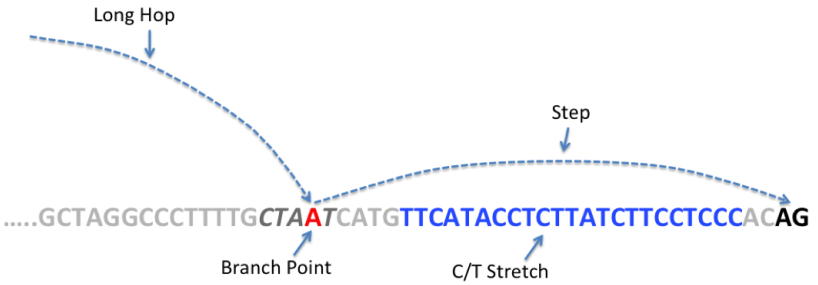


Figure 6.14: The Likely Branch Point in Intron 2 of *HBB*.

Maybe a variant at this branch point was the source of $\mathscr{X}$'s problem? Such a variant could confuse the long hop causing it to land at one of the 6 previous candidate landing spots. A step from there would end up at the wrong AG: not the AG at the end of Fig. 6.10 but some previous AG. The consequence would be a change in the *HBB* recipe, leading possibly to a defective protein. So we look for variants in $\mathscr{X}$'s genome at the breakpoints in both the introns of *HBB*? And we draw a blank, yet again!

Where Hides the Variant?

Where else could a variant be hiding? It wasn't in the exons. It wasn't at the GT or the AG markers flanking exon ends. It wasn't at the likely branch point. The sequence around this branch point was CTAAT, consistent with the branch point pattern (Fig. 6.12); no variants here either. What remained?

There was an innocuous-looking variant buried within the C/T stretch though. What is usually a T had become a G in $\mathscr{X}$, giving the C/T stretch an extra unwanted character. Only 3 of the 23 characters in this stretch are anything but C or T in the average person. But $\mathscr{X}$ has an extra G in this stretch, taking this number up to 4. Here is how this extra G appears, right in the middle of the C/T stretch.

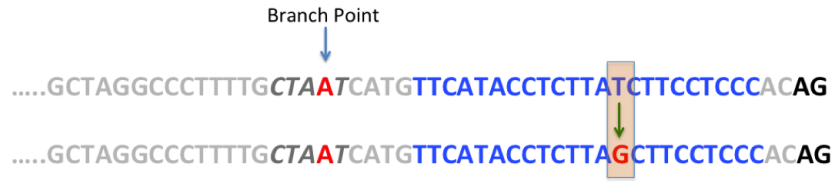


Figure 6.15: A Variant in the C/T Tract

So 3 A/G's and 20 C/T's in the average person. 4 A/G's and 19 C/T's in $\mathscr{X}$. Could this weaken the C/T stretch substantially in $\mathscr{X}$? Possibly not, thanks to the CTAAT pattern around the branch point; it matches the pattern in Fig. 6.12 perfectly, and a perfect match here can compensate for a few violations in the C/T stretch.[68]

What else could this variant do? Let us walk through the trail of events again. Introns in the *HBB* gene are spliced out during recipe interpretation. Splicing of an intron, say intron 2 of *HBB* involves a long jump from the GT at the end of exon 2 to the AG at the beginning of exon 3. This long jump occurs in two parts. A long hop first finds the branch point A, modulated by the pattern of characters around this branch point, and the pattern of characters in the C/T stretch. Then a step takes the splicing process from the branch point to the next AG to the right of the branch point. Presumably, the AG right at the end of intron 2?

Wait a minute! Which is the next AG to the right of the branch point? Usually, this is the AG at the very end in Fig. 6.15. But not for $\mathscr{X}$. By a stroke of bad luck, the T$\rightarrow$ G variant in $\mathscr{X}$ happens to be next to an A. And this creates a new AG closer to the branch point. The long hop may well have found the branch point correctly, only for it to be deluded at the very next step by this new AG. This AG entices the step from the branch point towards itself and away from the intended AG, which lies 11 characters further to the right. The outcome: recipe interpretation resumes 11 characters too soon!

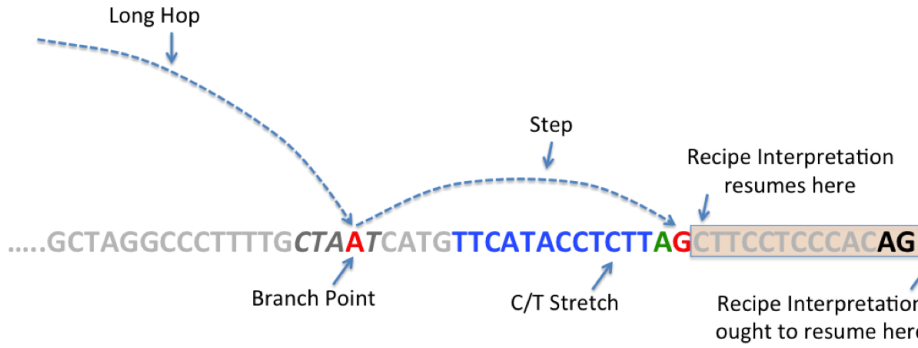


Figure 6.16: The Long Jump Lands Elsewhere.

The recipe interpretation process then partitions subsequent characters into frames of 3 characters each, converting each frame to an amino acid (remember the table in Fig. 2.5). Since it resumes 11 characters too soon in $\mathscr{X}$, it creates a few more amino acids in $\mathscr{X}$ than it would do in the average person. But that is not all. By a second stroke of bad luck for $\mathscr{X}$, 11 is not a multiple of 3. So not only does recipe interpretation create a few extra amino acids in $\mathscr{X}$, it also interprets the subsequent characters with a shifted set of frames, as shown below, leading to a very different sequence of amino acids.

Figure 6.17: And the Recipe Changes Dramatically!

Thus, a seemingly innocuous T→ G variant buried inside an intron results in $\mathscr{X}$'s *HBB* recipe yielding a sequence of amino acids whose latter third is quite different from what it would be in the average person. But is it different enough to undermine the oxygen-carrying capability of *HBB*?

Does Heme Fit Snugly?

Remember, the heme attachment to the *HBB* protein (Fig. 6.3). In the average person, this attachment fits snugly into the *HBB* protein. But $\mathscr{X}$'s *HBB* recipe is different from that of the average person. Does it still afford a snug fit for heme?

Imagine the average healthy person. And imagine the chain of amino acids strung together by recipe interpretation on the *HBB* gene in this person. This chain does not stay as such, rather, it quickly folds itself into a compact shape, as the cartoon below indicates.

Figure 6.18: A Cartoon view of a Generic Amino acid Sequence: stretched out on the left and folded on the right. This is just indicative and not quite the actual folded shape for the *HBB* protein. That shape follows shortly.

The folded shape is not a random shape but a highly repeatable one. Which means that the same shape is produced every time recipe interpretation runs on the *HBB* gene. This shape is written in the gene recipe, so to say, and it has special capabilities. Alarm bells should go off now, for a different recipe, such as the one is $\mathscr{X}$, could result in a different shape.

But before we get to $\mathscr{X}$, here is what makes this shape special.

Visualize the chain of amino acids strung together by recipe interpretation on the *HBB* gene as a ribbon; in particular, a ribbon made of metal that usually holds its shape, but can also be bent. Much like a vine would, some stretches of this ribbon twist themselves into spring-like helices. These helices behave like stiff rods that don't bend. The portions in-between these helices do bend though, causing these stiff helices to orient themselves so they wall off a central pocket from all sides. This pocket is tailor-made for the heme attachment to fit in, as shown in the picture below.

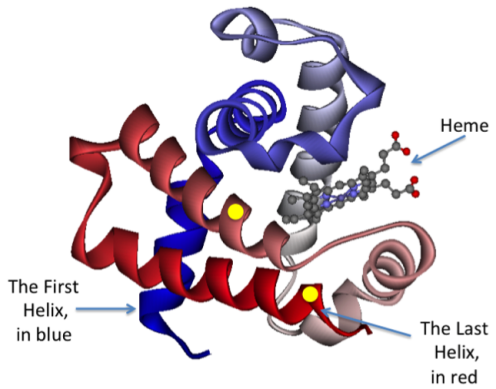


Figure 6.19: The Folded *HBB* Protein Structure. The helices in this picture are colored so the first helix is dark blue and the last helix is dark red. The colors of the intermediate helixes slowly transition from blue to red.

It might appear from this picture that the heme attachment is free to slide in and out of this pocket. That appearance is deceptive. Forces of attraction between the heme molecule and several amino-acids in the *HBB* structure hold the heme in place, providing the required snugness. Two of these amino acids, one in each of the last two helices, are amino acids 107 and 142, both marked with yellow dots in the picture above.

The picture for $\mathscr{X}$ might be a different one. The shape into which her *HBB* protein folds is written in her distinctive *HBB* recipe, which

is different from the average person's recipe. The first 106 amino acids remain identical in the two recipes. But amino acid 107 onwards, the recipe is dramatically different. How do we tell if amino acids 107 and 142 (or any others for that matter) will continue to provide anchor to heme in this new recipe? For this, we have to decipher the folded shape that results from this new recipe. Unfortunately, there is no simple way to do this.

Expensive experiments will give us the correct answer but are impractical to do for every patient. Instead, we simulate the folding process on a computer. Folding happens because the various amino acids (and other surrounding molecules) exert forces on each other. These forces cause the atoms in these molecules to move from their current locations to new locations. In their new locations, they exert a new set of forces on each other. These new forces, in turn, cause the atoms to move to yet newer locations. And so on, repeatedly. The whole process takes just microseconds inside a living cell. But simulating even a microsecond on a single computer can take weeks or even months! Some tricks do get us an approximate folded view of the last two helices (which carry the yellow dots) relatively quickly though. And here is how they look.

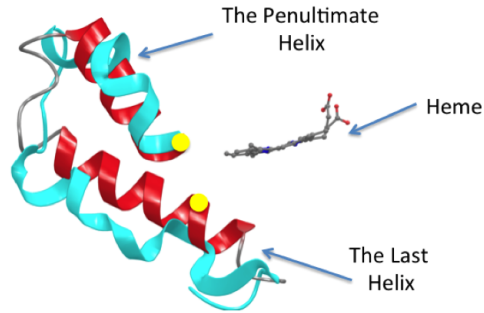


Figure 6.20: The Last Two Helices: The Average Person's in Red Juxtaposed with $\mathscr{X}$'s in Cyan.

Contrast the tight, spring-like regularity of the stiff, red helices to the rather floppy and disordered appearance of $\mathscr{X}$'s structure, shown in

cyan. These stiff helices have started to come loose in $\mathscr{X}$. The pocket is probably much bigger in $\mathscr{X}$ than it should be. And heme is unlikely to fit snugly into this pocket anymore.

Had the T→ G variant appeared in $\mathscr{X}$ in only one of her two *HBB* gene copies, then half of the protein copies generated in $\mathscr{X}$ would have been normal. However, $\mathscr{X}$ had this variant in both gene copies. So chances are that none of $\mathscr{X}$'s protein copies would be able to provide strong anchor to heme attachments.

Nailing Down the Variant

From all accounts, it appears that the T→ G variant is the cause of $\mathscr{X}$'s problems. It is likely that this variant deludes the splicing machinery into resuming recipe interpretation too prematurely, thus resulting in a new and different recipe. It is also likely that this new recipe disrupts the structure of globin, compromising its ability to anchor heme and carry oxygen. Could scientific literature offer some more evidence for the culpability of this variant, in the absence of hard experimental proof?

Studies on European and American patients seem to have no mention of $\mathscr{X}$'s T→ G variant. Neither do studies on north Indian patients. Then there is one study[69] which screened several Thalassemia patients in the since-bifurcated south Indian state of Andhra Pradesh for genomic variants in the *HBB* gene. It found 6 variants, each of which was present in multiple patients. The T→ G variant found in $\mathscr{X}$ was not one of these. However, this study also found 3 additional variants which were very rare, i.e., found in only 1 patient each. Two of these variants were in the exons of the *HBB* gene. The third variant, found in just 1 patient, was exactly $\mathscr{X}$'s T→ G variant!

Both copies of the *HBB* gene in this patient had this T→ G variant. And there were no other variants of significance in the *HBB* gene in this patient. A more recent study,[70] this time in the neighboring south Indian state of Karnataka, found this variant in 2 thalassemia patients. And yet another study,[71] an older one, found this variant in a patient of Indian origin. So 4 patients in all with this variant, in addition to $\mathscr{X}$. This has to be the one!

$\mathscr{X}$ too was from south India. Maybe this variant had arisen somewhere in this region and had then passed down the generations. Still very rare, it usually appeared in only one copy of the *HBB* gene in carriers of the disease. The unfortunate confluence of two defective copies, while very very rare, did rear its head from time to time. $\mathscr{X}$'s genome was one such confluence point. What could $\mathscr{X}$ do keep the ill-effects of this variant at bay?

Hemoglobin from Donors

$\mathscr{X}$'s genome could only make defective hemoglobin, unable to carry and deliver oxygen. Providing good hemoglobin to $\mathscr{X}$ from outside via a blood transfusion was the only immediate way to keep $\mathscr{X}$ healthy. Blood transfusions are done quite routinely nowadays, and healthy and willing donors with a matching blood group are not hard to find. But red blood cells have short life spans. $\mathscr{X}$'s doctor might suggest removal of her spleen, an organ which accelerates the destruction of donor blood cells. Even then, $\mathscr{X}$ would need blood transfusions every month or so. Such repeated transfusions would no doubt be cumbersome for $\mathscr{X}$. They would also place her at an increased risk for blood-borne infections. Most importantly, though, they would expose her to a greater long-term risk.

Remember, each hemoglobin molecule has four heme attachments, and each heme carries an iron atom. Our body controls the total iron level very tightly. Repeated infusions of iron-containing hemoglobin defy this control. Excessive iron accumulates in the body as a result. These high levels of iron lead to organ damage, particularly to the heart and the liver. $\mathscr{X}$ could face ongoing complications on this account. There are treatments for managing iron levels which could mitigate some of these complications. Even with these, $\mathscr{X}$ would be fortunate to live for a few decades.

$\mathscr{X}$ was still young and her family would appreciate the reprieve provided by repeated blood transfusions. But the specter of on-going complications and impending organ failure will no doubt weight heavily on their minds. Was there an alternative?

Replacing the Blood Genome

Here is a anecdote, which the internet attributes to Abirami Chidambaram of the Alaska State Scientific Crime Detection Laboratory. It pertains to a case of sexual assault. Semen traces were collected from the scene of crime. The genome of the criminal was extracted from this semen. It was then matched against a genomic database of various criminals, constructed from their blood samples.

For purposes of this matching, it is too expensive, and a huge overkill, to compare the entire genome; it suffices to consider only a few carefully chosen genomic locations. These so called *repeat locations* carry sets of repeating genomic characters. For instance, ACGACTACGACTACGACTACGACT.. and so on, with the unit ACGACT repeating several times. These repeat locations are chosen so that the number of repeats of the repeating unit is very likely to be different between any two individuals. For instance, you might have 5 repeats of ACGACT, I might have 8. By considering a few different repeat locations, a person can be identified uniquely. Coming back to the point, when the genome of the criminal in this case was matched against known criminals in the police database, there was a match. A match with a person locked-up in prison!

This was baffling. Did the prisoner quietly escape, commit the crime, and then return himself to jail? Unlikely. Indeed, further detective spadework unearthed yet another match. Now, two people matching is a highly unlikely event, unless they are identical twins. But these two were not identical twins. They were brothers! Brothers do share more of their genomic sequences than two unrelated people. But even brothers have genomes that differ enough so both cannot match. What could explain this mystery?

Another genomic test was performed to compare the genomes of the two brothers with that of the semen sample from the scene of crime. This time, DNA was taken by a cheek swab instead of a blood sample. A cheek swab uses cotton to scrape off cheek cells from the inside of the cheek; the genome is then extracted from these cells. And this time, only the genome of the second suspect matched that of the semen sample. The genome of the prisoner did not! And this nailed the prisoner's brother conclusively, absolving the prisoner in the process. But the mystery remained yet to be

explained.

Why did blood samples from both brothers, but cheek swabs from only one, match? Light was shed only with the realization that the prisoner had undergone a *bone marrow transplant* previously. The donor for this transplant was none other than his brother! This made the prisoner's blood genome identical to his brother's. No wonder both matched the sample from the crime scene. But the cheek, and indeed the rest of the body, carried different genomes in the two brothers. The real culprit, the prisoner's brother, could be uniquely identified only by his cheek swab and not by his blood.

In summary, a bone marrow transplant completely replaces the genome found in the blood cells of a person. All other cells continue to have the genome the person was born with. As the name indicates, the bone marrow is a factory located in the interior of our bones. This factory comprises of so called *stem cells*. Such cells can divide to generate new daughter cells, thus assuring a self-contained life-long supply of stem cells. Through a controlled process of change, some of these stem cells then become red blood cells. A red blood cell has a short lifespan and cannot divide further into multiple daughter cells. So only the stem cells in the bone marrow factory can produce new red blood cells. A bone marrow transplant seeks to replace these stem cells by alternative stem cells from a healthy donor.

If $\mathscr{X}$ were to get a bone marrow transplant, then the very first step would be to kill the stem cells in his bone marrow using drugs. Then bone marrow stem cells would be extracted from a donor's hip bones. These would then be fused into $\mathscr{X}$. These stem cells would now divide inside $\mathscr{X}$ and some will convert into red blood cells. The hemoglobin in these red blood cells would now be manufactured based on the recipe in the donor's genome and not based on the defective recipe in $\mathscr{X}$'s genome. So, with this one-time procedure, $\mathscr{X}$ would have a lifelong supply of good hemoglobin! Of course, there are risks, as always.

Our immune system (which is primarily driven by white blood cells made in the bone marrow) has this uncanny ability to differentiate between our own cells on the one hand, and foreign cells that come from another person on the other hand. And foreign cells are typically attacked and destroyed. Accordingly, $\mathscr{X}$'s immune system could recognize the donor's

cells as foreign and could proceed to attack and destroy these cells. To prevent this from happening, $\mathscr{X}$'s immune system is suppressed by killing the stem cells in her bone marrow before the transplant.

The converse issue is more problematic though. White blood cells created by the donor's stem cells might turn upon $\mathscr{X}$'s cells, thinking that these are foreign! How would these cells know what is foreign and what is not? It turns out that another set of genes called the *HLA* genes plays a key role in distinguishing self from foreign. Every individual has several variants in these genes. If the variants in the donor match those in $\mathscr{X}$, then the chances of the donor's cells turning upon $\mathscr{X}$'s cells is much reduced. Indeed, a bone-marrow transplant offers $\mathscr{X}$ a greater than 80% probability of complete cure provided the donor's *HLA* variants closely match $\mathscr{X}$'s variants.[72] No further need for blood transfusions and no progressive complications. Unfortunately, there is a 3% risk that the procedure will fail badly[73] even when the donor has matched *HLA* variants, leading to death. Was this the only option for $\mathscr{X}$?

Editing the Blood Genome

Wouldn't it be wonderful if we could somehow edit $\mathscr{X}$'s blood genome and restore the problematic G in her *HBB* gene back to a T (remember we are dealing with a T$\rightarrow$ G variant)? So instead of killing $\mathscr{X}$'s bone marrow stem cells and replacing them with someone else's stem cells, what if we could extract some bone marrow stem cells from $\mathscr{X}$, kill the remaining bone marrow cells, edit the genomes in the extracted cells, and inject these back into $\mathscr{X}$. As these stem cells divide inside $\mathscr{X}$, the new daughter cells will continue to have a T instead of the G that $\mathscr{X}$ was born with. And as some of these cells convert into red blood cells, recipe interpretation on the *HBB* gene will now result in healthy hemoglobin, capable to carrying and delivering the required oxygen to $\mathscr{X}$'s body. And since the rest of the genome stays as before, $\mathscr{X}$ wouldn't need to contend with any immune system mismatches.

So how do we take a few cells and edit one chosen character among billions of characters? Editing files on a computer is easy. Editing a printed book, less so. Editing a real, physical molecule inside a live cell? Even

less so. A major hurdle in editing a genome inside a live cell is to locate the point where the edit needs to be made. Specifically, the point which carries the T→ G variant above. The editing mechanism has to locate this point among 3 billion genomic characters. Fortunately, nature has provided us with tools to perform this search accurately. And these tools came to our notice from rather unexpected quarters.

In the last decade or so, scientists have observed an interesting phenomenon in the genomes of bacteria attacked by a virus. The attacked bacterium has its own distinctive genome sequence, as does the attacking virus. Interestingly, after the attack, stretches of characters from the viral genome are inserted into the bacterial genome. The bacterium has thus managed to save the identity of the virus into its own genome. Going further, it also has a mechanism to use this identity (i.e., these stretches of viral genome characters saved into its own genome) to ward off similar virus attacks in the future.[74,75] This mechanism is driven by a special gene called *CAS9*. Proteins derived from *CAS9* can take these stretches, match them against the genomes of the invading viruses, and clip these viral genomes at precise points where matches is found. Clipping a viral genome at these places invariably leads to destruction of the virus, thus protecting the bacterium from the attack.

Note the remarkable feature of *CAS9* above. If you give it a template (a specific stretch of roughly 20 genomic characters), then it will fish out those locations in the entire genome where this template matches. This works for the human genome as well. Even though the human genome is huge, 3 billion characters long, *CAS9* will accurately locate places where this template matches. It will then clip the genome at these match points. If the template is chosen carefully, the result will be clipping of the genome only near the point which needs to be edited. For $\mathscr{X}$, this point is the location of the T→ G variant. In fact, using two distinct templates,[76] one can get two clips, one on either side of this point, as shown in this picture.

Figure 6.21: Genome Editing: The blue stretches are two distinct templates. *CAS9* locates these in the genome and clips the genome three characters into each template.

What next? Genome clipping is actually a common event that occurs routinely in nature. To protect the genome from the impact of this clipping, nature has provided cells with repair mechanisms that help rejoin these clipped portions. One of these repair mechanisms can be fooled into replacing the genomic stretch of characters in-between the two clip points with a new chosen set of characters. By carefully choosing these characters, the aberrant G can be edited back to T!

Since the very recent realization that genomes can be edited using this so called *CRISPR-CAS9* method,[75] a massive scientific effort is on to make this approach robust. There are several promising signs already. For instance, scientists have successfully edited the genome in liver cells of live, adult mice, consequently curing these mice of a disease called tyrosinemia.[77] There remains some danger though that *CAS9* might clip the genome at unintended places, leading to some unexpected consequences. Or that the repair mechanisms that rejoin clipped portions might not quite do their job as expected. So this approach is not yet ready for medical use. However, there is every hope that further improvements will allow us to edit genomes of patients like $\mathscr{X}$ sooner rather than later.

Wrapping Up

The vast expanse of the genome is sparsely dotted by exons carrying gene recipes. These recipes amount to a paltry 1-2% of the entire genome. The remainder of the genome comprises introns and inter-genic regions (i.e.,

genomic regions that lie in-between genes), which do not carry recipes directly. It would appear, therefore, that genomic variants in these regions do not impact gene recipes at all. But $\mathscr{X}$'s story tells us otherwise.

Genomic characters in the introns guide the recipe interpretation process as it jumps from one exon to the next. And variants here could well confuse these jumps. And a jump gone awry could well yield dramatically different recipes, as seems to be the case for $\mathscr{X}$. The *HBB* gene is notorious for jumps gone awry on account of such intronic characters. For instance, 20% of Beta-Thalassemia cases in China are caused by a variant sitting deep inside the second intron of *HBB*.[78] This variant causes the splicing jump to land prematurely in the middle of this intron; recipe interpretation resume from this point and continues for a while before jumping again to the end of the intron, where it should have landed in the first place. Of course, the recipe is modified dramatically as a result.

Meanwhile, $\mathscr{X}$, who has had numerous transfusions, waits for a suitable donor for a bone marrow transplant. If a suitable donor is found, there is a good chance that the transplant will successfully correct $\mathscr{X}$'s hemoglobin without introducing any other complications. But there is a small risk of things going the wrong way.

Hopefully, the ability to edit genomes will become a reality in medical practice in the not too distant future, removing the need for donors and associated complications. $\mathscr{X}$'s doctor did not need to know which variant $\mathscr{X}$ had in her *HBB* gene to diagnose her Thalassemia. But precise knowledge of this variant would be needed for editing $\mathscr{X}$'s genome. That knowledge is now available.

Chapter 7

The Ominous Reflection

Two year old $\mathscr{X}$ was bothered by redness and irritation in his eyes. His parents also noticed a squint: it seemed as if his eyes were looking in different directions. A little whiteness in the center of the eye was visible as well. So $\mathscr{X}$ was shown to a doctor who examined the insides of his eyes. The doctor first dilated, or enlarged, $\mathscr{X}$'s pupils with some eye-drops (a pupil is the little hole at the front through which light enters the eye). He then shone light through the pupil into $\mathscr{X}$'s eyes. This light reflected back from $\mathscr{X}$'s retina. The doctor expected this reflection to be a reddish-orange circle around a tiny white dot. Here is what he saw instead.

Figure 7.1: The Expected Reddish-Orange Reflection on the Left,[79] and the Reflection from $\mathscr{X}$'s Eyes, on the Right.[80]

The large white spot instead of the reddish-orange one was unexpected. Surely, there was a problem in $\mathscr{X}$'s retina. Leaky blood vessels in the retina were a possible cause: a serious one at that, for it results in blindness. But there was another, far more serious, possibility. Indeed, further

investigation into $\mathscr{X}$'s retina indicated that the large white spot came from a tumor in $\mathscr{X}$'s eye; a potentially life-threatening cancerous growth, called a *Retinoblastoma*.

We usually think of cancer as a disease of aging. Indeed, it appears most often in the 60s and 70s. Cancer in infancy is rare. Retinoblastoma, even more so; just around 350 new cases are diagnosed each year in the United States, and around 1500 in India. $\mathscr{X}$ was one of these unfortunate few. Some cells in his retina had turned rogue, defying the social contract that binds them in discipline with their neighboring cells. These bad citizens were dividing uncontrollably, destroying the organized cellular arrangements needed for proper vision. If left untreated, blindness was imminent. But that was not all.

Cancerous cells often leave the organs they belong to and wander off to other parts of the body, carrying their destructive effect with them. When this happens to organs vital for life, say the brain or the bone marrow, life itself gets threatened. Indeed, the large white spot was a threat to $\mathscr{X}$'s life itself. What caused this life-threatening condition in $\mathscr{X}$?

Heeding the Warning

The white spot is ominous on the one hand. It can be a life-saver though by providing early warning. A leading newspaper in Bangalore carried this heart-warming story on May 12th, 2012,[81] as reported to us by Dr. Ashwin Mallipatna, a specialist in Retinoblastoma, practising at Narayana Nethralaya, Bangalore.

A couple with a young daughter in the suburbs of Bangalore noticed the white spot in their daughter's eye when she was 6 months old. The spot was still very small, so no alarms rang at that moment. On the contrary, it was interpreted as good omen!

A few months later, they took their daughter to a photo studio for her to be photographed for an occasion. Typically, the pupil of the eye shows as a red dot when the face is photographed. This happens because light from the camera flash enters the eye through the pupil, reaches the retina, and is then reflected back from the blood vessels there; reflection from blood leads to this reddish color. Cameras have sophisticated methods to

correct for this redness so you may not always spot it on a photograph.

However, when a Retinoblastoma grows on the retina, the reflection no longer appears as a red dot; in place of the red dot, is the dreaded larger white spot. This white spot is much more prominent on a photograph that it is to the direct naked eye. Indeed, the daughter's photograph showed a very prominent white spot.

Fortunately for the family, the photographer did not think this was a good omen. He had actually lost his two-year old niece a few years ago to Retinoblastoma, and was therefore well aware of the ominous nature of this white spot. Accordingly, he suggested that the daughter be shown to specialist. Just in time, it turned out; appropriate treatment saved her eyes and her life. Was $\mathscr{X}$ so lucky as well? We'll see. But a question first: what causes tumors in the eyes of $\mathscr{X}$ and other little infants like him?

Inheritance or Ill Luck?

Was the Retinoblastoma in $\mathscr{X}$ on account of inheritance from his parents? Or was it due to some other factor acquired during or after conception? Was it a controllable factor or was it just plain bad luck? Some clever deduction from simple patient observations yielded the first insights on this question in the early seventies.

Doctors had long noticed that some patients with Retinoblastoma presented with tumors in both eyes (the bilateral type), and yet others presented with tumors in only one eye (the unilateral type). These two types appeared to have different behaviors. The bilateral type appeared clearly hereditary, because children born of such patients (those fortunate enough to survive their tumors and have children of their own eventually) showed a 50% of chance of developing Retinoblastoma. Variants in some yet unknown gene $\mathscr{G}$ were therefore the likely cause.

Remember, each of us has two copies of each of our 20,000 or so genes. Genes on the X and Y chromosomes are key exceptions to this rule, but these genes are not relevant to our story here. So we have two copies of $\mathscr{G}$, one copy inherited from each of our parents. And as parents, we pass on one of our two copies of $\mathscr{G}$ to each of our children. The 50% inheritance figure for bilateral cases suggested that a variant in one of these

two copies was sufficient to cause the disease; a parent with the variant in one copy would pass on that defective copy 50% of the time. And children inheriting this defective copy would develop Retinoblastoma. But there was a catch.

A few infants with bilateral Retinoblastoma (1%-10%)[82] showed a different trend. The disease did indeed appear to be hereditary in these infants because there was a close relative (a sibling, perhaps) who also had the disease. But, strangely, neither parent showed any sign of the disease! One of these parents must have had the same problematic variant in $\mathscr{G}$ as their child. Yet, they did not develop the disease. Why? Maybe a variant in one copy of $\mathscr{G}$, while instrumental in causing the disease, was not sufficient by itself. There was at least one, and maybe many, additional causes working in tandem with the inherited variant in $\mathscr{G}$ to cause disease. These causes appeared to be involved in most but not all individuals who had a variant in one copy of $\mathscr{G}$.

The unilateral cases further announced the need to identify these additional causes. For these cases appeared to be not hereditary at all, by and large (because children of parents with unilateral Retinoblastoma rarely seemed to develop the disease). The question then arose: what were these additional causes?

This was all, of course, well before sequencing the genome was possible, so scientists had to answer this question with much simpler forms of data then available. To this end, Alfred Knudson at the M.D. Anderson Hospital in Texas dug deeper into the records of 48 patients at his hospital with Retinoblastoma in 1971. 23 of these had bilateral tumors and 25 had unilateral tumors.[82]

Knudson's Bilateral Cases: The Single Hit Roulette

Starting with the 23 bilateral cases, Knudson examined the ages at which they had been diagnosed. Retinoblastomas are usually diagnosed within within the first two years of life, so the age of diagnosis is measured in months rather than years. Take a moment to stare at these numbers.

2,3,3,3,4,4,5,6,7,8,9,11,12,12,13,15,18,18,22,24,30,44,60

Do you see a pattern here? If not, then read on! Also note the average age of diagnosis: 14 months.

What do these numbers say about any additional causes of disease in these children? Are these causes already present at the time of conception? If so, one would expect tumors to start forming at roughly the same age in all these children. Of course, never exactly the same age, for there is always natural variation. So some would start a little earlier, some a little later. These tumors would then grow at roughly the same rate in all these children until they become discernible, leading eventually to the point when they are diagnosed. Diagnosis happens, therefore, at roughly the same age in all these children, with some natural variation. Which means that most of these numbers should be around the average mark of 14, with some leaders appearing earlier and roughly an equal number of stragglers following later. But the numbers above show no such pattern!

Instead, our numbers appear tightly clustered at the beginning (many 3s and 4s) and very sparse towards the end (after 24, they skip all the way to 30, and then even further to 44, and thence all the way to 60). Our assumption that the additional causes were present at the time of conception itself appears incorrect, or so the numbers say! There must be another explanation.

Maybe the additional causes are acquired after conception. Which means the following. Each of us starts life as a single cell. This cell divides many many times to yield a multitude of cells; a hundred trillion cells actually! The genomes in each of these daughter cells is supposedly a copy of the genome in the first cell which started it all. Not a perfectly exact copy though! The copying process is remarkably accurate but not perfect, so it does make mistakes once in a while. The net result is that the genome in any particular cell in the body may carry variants which we did not inherit from our parents. Most of these variants, with the possible exceptions of those in our testes (for males) and ovaries (for females) are not be passed on to our children. These variants live and die with us. These acquired variants are called *somatic* variants. When a cell carrying a somatic variant divides, the daughter cells also inherit this variant. A particular somatic variant thus finds its way into many cells. Coming back to Knudson's bilateral cases, each of these children probably had a variant

in the gene $\mathscr{G}$ which they had inherited from their parents. Additionally, did they have one or more somatic variants, which some of the cells in the retina had acquired later? And was Retinoblastoma in these children caused by a combination of these two variants? What answers do the numbers above give us?

We need a little thought experiment to get to the answer. Let's start with the following assumption: somatic mutations arise in a random manner, much like a disoriented but persistent dart-player taking repeated shots at this dart-board. The dart-board is special, in that it has, not one, but millions of bulls-eyes, just as the retina comprises millions of cells. Each bulls-eye represents one of these cells.

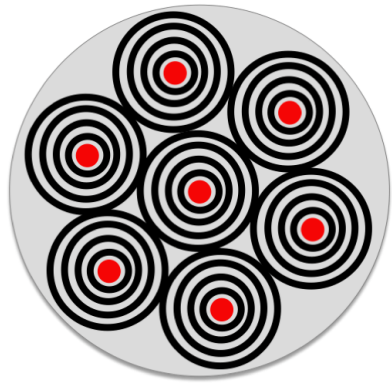


Figure 7.2: The Dart-board with Many Bulls-Eyes.

The dart-player has to hit any one of these bulls-eyes. Hitting a particular bulls-eye is akin to the corresponding cell acquiring a somatic variant. The dart-player is disoriented and has poor control on his throws. So his chances of hitting any specific one of these bulls-eyes on any given attempt are miniscule. But, since there are millions of bulls-eyes, the chances that he hits some bulls-eye, no matter which, are higher. Since retinal cells are constantly dividing in the early months of life leading to the formation of more cells, the dart-player gets repeated attempts at his task. Then his chances of success, i.e., hitting one of the many bulls-eyes in one of these

several repeated attempts, are even higher.

Next, imagine many such dart-players, each aiming at his or her respective board. Each dart-board is a metaphor for a distinct patient. As time elapses, more and more of these dart-throwers succeed, i.e., more and more of these patients acquire an additional somatic variant leading to tumor formation. At what rate does this happen? At a steady, uniform rate? Which would mean that a certain number of players succeed in the first hundred shots, and a roughly equal number succeed in the next hundred shots, and so on. Or at an accelerating rate? Which would mean that fewer players succeed in the first hundred shots than in the next hundred shots. Or at a decelerating rate? Which would mean that more players succeed in the first hundred shots than in the next hundred shots. Which of these is the true scenario?

Suppose each dart-player has a 50% chance of success in hitting some bulls-eye after, say, a hundred shots. Then, after each player has taken a hundred shots, roughly 50% of the players will succeed, and the remainder not. The players who succeed step off. The remaining (roughly) 50% continue to take further shots. What happens after another hundred shots? Will all these remaining players succeed? Not really. A little thought will show that roughly only 50% of these remaining players will succeed in this second round, and the remainder not. In other words, roughly 25% of all players will succeed in the second round of hundred shots. So 50%+25%=75% of all players are done after two hundred shots. What happens in the third round of hundred shots? Roughly 50% of the yet unsuccessful players will succeed in this round. So, after three hundred shots: 50%+25%+12.5%=87.5%! And so on.

Note the pattern: a lot initially, a little less next, even lesser next, and so on. We could plot this as a curve. A curve that indicates the fraction of all players who have succeeded, as time progresses. It is easy to calculate this curve precisely. One could do this calculation mathematically, or simply by simulation on a computer. As expected, the curve rises steeply in the beginning, because many players succeed in a given time interval, say one month. In the next month, fewer succeed, so the curve flattens out a bit. In the further next month, even fewer succeed, so the curve flattens out a bit more. And so on. The curve keeps flattening out more and more

over time. If indeed a single somatic variant, acting on top of the inherited variant, was the cause of disease in Knudson's bilateral cases, the ages of diagnosis above should also show a similar pattern. Do they indeed? The picture below gives us the answer.

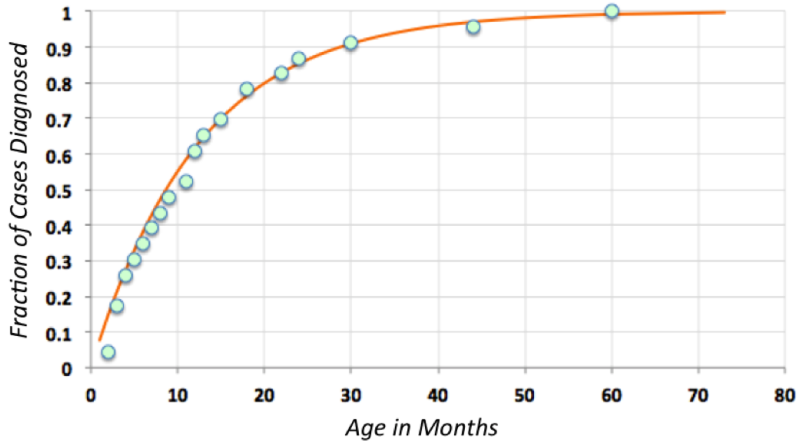


Figure 7.3: The Diagnosis Ages for Bilateral Cases.

The green dots, indicating the ages of diagnosis, lie bang on the above curve, calculated with each dart-player's success rate taken as about 7.7% per month. Roughly, half the cases are diagnosed in about 9 months, as predicted by this curve, another quarter are predicted in the next 9 months, and so on. Knudson thus concluded that an inherited variant followed by a single acquired somatic variant was necessary to cause Retinoblastoma in these bilateral cases. The first hereditary *hit* alone was insufficient, a second somatic *hit* was required. These two *hits*, together, were sufficient.

Note the element of randomness and luck in the second hit though. The dart-player has a 95% chance of success with repeated attempts over 36 months, and 99% chance of success over 60 months. So a child inheriting the first hit was very likely to acquire a second hit and develop

Retinoblastoma in the first few years of their life. But there was a slim chance that the dart-player would fail altogether and then there would be no second hit; so the occasional child with an inherited first hit could escape disease altogether. Of course, given enough time, the dart-players success rate will approach 100%. Thankfully, active division of cells in the retina stops within a few years of birth, throwing the dart-player out of commission. So children who escape the second hit in the first few years of life usually escape Retinoblastoma forever.

Knudson's Unilateral Cases: The Two Hit Roulette

Knudson still had the the unilateral cases to investigate before firmly concluding on his two-hit theory. Observations made over the years had suggested that these cases were not hereditary, by and large (because children of parents with unilateral Retinoblastoma rarely seemed to develop the disease). Where then did the first hit result from, if there was no inherited hit? Could both hits be random somatic hits? The answer is hidden in the ages of onset for 25 of Knudson's unilateral patients[82] given below, if only you can spot it!

5,8,10,15,19,21,22,24,27,28,29,29,31,32,33,34,36,36,38,46,47,48,50,52,73

Note the average age of diagnosis now: 31 months. Significantly longer than the 14 month average for bilateral cases. Tumors in unilateral patients seem to be setting in much later. And the pattern itself appears to be different. Remember how the bilateral cases were: steep at the beginning (many cases diagnosed early) and getting flatter and flatter as time progresses (fewer and fewer cases diagnosed as time elapses). Instead the unilateral cases appear to be slightly flat at the beginning with only 3 cases in the first 14 months (5, 8, 10). Then it gets a little steep in the middle with as many as 6 cases in 5 months 27-32 (27,28,29,29,31,32). At the end, it again goes flat with just 3 cases in the 23 months 50-73 (50, 52, 73). Is this the profile one would expect with two random somatic hits?

Think of the dart-board analogy again, but with a difference. As before, each dart-player takes repeated shots at a dart-board which carries millions

of bulls-eyes. However, hitting just any one of these several bulls-eyes once no longer qualifies as success. A player now succeeds when they hit the same bulls-eye twice. Note: hit the same bulls-eye twice. Not just hit two distinct bulls-eyes. Remember each bulls-eye is a metaphor for a distinct cell; both hits must occur in the same cell; two variants in two different cells will tip neither cell over into dividing uncontrollably. How quickly do the dart-players succeed in this modified challenge?

Certainly at a slower rate, for the challenge is harder. In initial attempts, some bulls-eyes on the dart-board may be hit once each but none would be hit twice. After a while, the chances of hitting one of these already-hit bulls eyes again will become substantial. So success rates will see slow take-off and then a more rapid rise. This will be followed by the same flattening-off we saw in the bilateral case. This profile can also be derived mathematically, or simulated on a computer. How well did the unilateral cases fit this two-hit profile? See below.

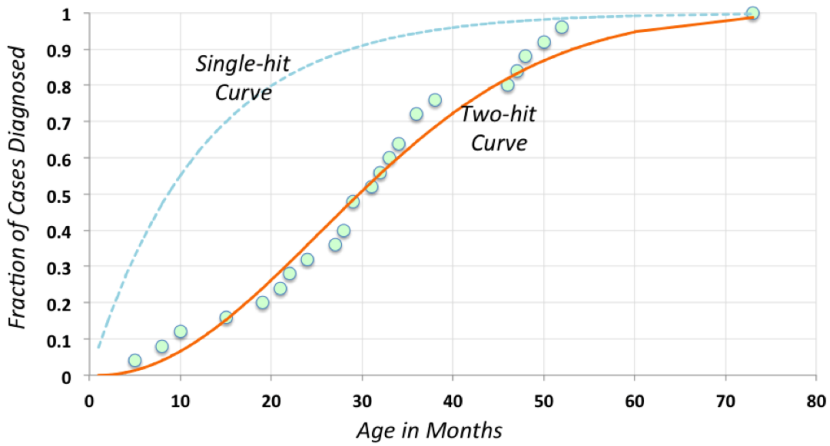


Figure 7.4: The Two-Hit Curve and Diagnosis Ages for Unilateral Cases.

Again, the dots representing diagnosis ages of the unilateral patients match the two-hit curve very nicely. In contrast, the single-hit curve that we used in the bilateral case bears no resemblance at all to the green dots.

It climbs far too aggressively while unilateral cases are diagnosed much more slowly. Reconfirmation that two hits, i.e., two variants, in the same cell, together cause Retinoblastoma!

A somatic variant hit has only a bleak chance of hitting any specific cell. But the number of cells in the retina is very large, in the millions. That makes it likely that one or more of these cells will pick up a somatic hit in the first five years of life. In those who have already inherited another hit from their parents, there is now a cell with two hits, and that cell sparks off the cancer. Those who haven't inherited such a hit from their parents are safe for the moment, until another somatic hit happens in the same cell that incurred the first somatic hit. The chances of this happening in the first five years of life are relatively remote though, just 1 in 40,000 or so. So 1 in 40,000 children is unfortunate enough to develop Retinoblastoma on account of two somatic hits.

$\mathscr{X}$'s case was bilateral; he had tumors in both eyes. He seemed to have inherited one problematic variant from his parents. In addition, he had suffered at least two subsequent somatic hits, one in each eye. Which gene did these hits happen in? And why did these variants send his retinal cells into a flurry of uncontrolled division? The answer takes us through a tour of the cell cycle.

The Cycle of a Cell

Each of us carries a 100 trillion cells. These are all derived by repeated division starting with a single cell. This act of a cell division has been perfected by nature over a billion years. Well, we should be careful and say nearly perfected, for $\mathscr{X}$'s travails do indeed stem from errors in this cell division process.

Before a cell divides into two, there is much to be done in preparation. First, the cell grows larger and accumulates material enough for both its daughter cells. Once there is enough, the cell then makes a complete copy of its entire genome. It then grows a bit more. And then comes the time for the cell to split itself and all its assets, including its duplicated genome, into two, creating two daughter cells in the process. These daughter cells then undergo the same cycle, in turn. And this cycle goes on and on.

Each of phases of the cell cycle: the first growth phase, the copy phase, the second growth phase, and the split phase, are elaborate orchestrations carried out by their own distinctive sets of genes. But someone has to signal the end of a phase and the beginning of a new phase. How does this happen?

Take the transition from the first growth phase to the copy phase for example. Recipe interpretation is initiated on a number of special genes capable of copying the genome during this transition. Protein derived from family of genes called the *E2F* genes perform this task. These proteins attach themselves to locations in the genome where a TTTCCCGC or similar sequence occurs. This attachment serves as the switching-on signal for the recipe interpretation machinery to start recipe interpretation on genes located close-by; these happen to be genes instrumental in the copy phase.

Figure 7.5: *E2F* Attaching to the Genome to Switch-On Recipe Interpretation of Genes Instrumental in the Copy Phase.

And this leads to the next question. What keeps the *E2F* proteins in check during the first growth phase, so they do not switch on these genes prematurely? This task is left to another very important gene. The protein obtained from this key gene attaches itself to the *E2F* proteins and holds them back from the genome. This holding pattern continues until the cell has grown sufficiently. Sensors in the cell detect that the cell has grown sufficiently, for instance, when it starts touching its neighboring cells. A cascade of events is then triggered. This cascade eventually causes our key protein to detach from *E2F*. The *E2F* proteins are now free to initiate the copy phase by attaching to the genome.

How does our key protein switch from holding on to letting go of

the *E2F* proteins? Remember how this key protein is created. Recipe interpretation on our key gene creates a chain of amino acids (Fig. 2.5). This chain then folds into a distinctive three-dimensional shape. This shape changes a bit when some of these amino acids are marked with additional markers (actually, phosphate groups). Our key protein usually has only a single amino acid marked; the resulting shape allows it to attach to the *E2F* proteins.[83] When the growth sensor detects that sufficient growth has happened, a cascade of events causes many of the amino acids in this key protein to be are marked. This hyper-marking forces a change in shape and our key protein then detaches from *E2F* as a consequence.

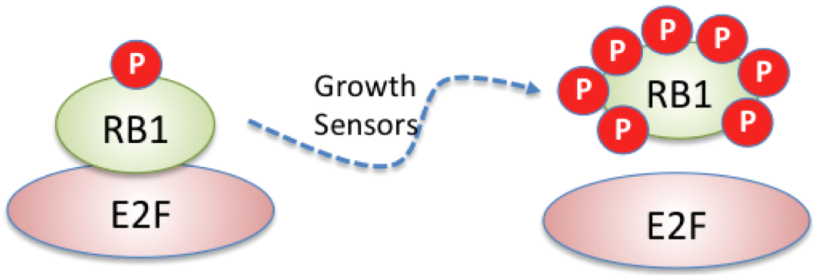


Figure 7.6: Our Key Protein Switching from One Marking State to Another.

Which is the key gene referred above? A gene which was found in the 1980s to be frequently aberrant in Retinoblastoma patients and eventually confirmed as the causative gene behind Retinoblastoma. Not surprisingly, it was named *RB*. $\mathscr{X}$'s travails too likely stemmed from this gene. Could it be that two hits on this gene in $\mathscr{X}$, one inherited and another somatic, had together incapacitated this gene? Were the *E2F* proteins then free to initiate the copying phase prematurely, unhindered by *RB*? And could this have accelerated the cell division process, leading to too much cell division, and as a consequence, tumor formation?

A Peek at *RB*

We need to peek into $\mathscr{X}$'s genome to identify any hits he might have on the *RB* gene. *Whole Genome Sequencing* and *Whole Exome Sequencing*, our windows to peek into genomes in previous chapters, are probably an overkill for this task. These methods cast a wide net, considering all 20,000 or so genes in our genome. The *RB* gene is specifically where we need to look. But unlike the *HBB* gene from the previous chapter which was all of 1605 characters long, the *RB* gene is huge: approximately 200,000 characters long! There are 27 exons (remember gene recipes are specified in the exons) separated by 26 intervening introns (remember, these introns do not carry recipes, they simply separate out the exons). The 27 exons together add up to just 2783 characters of the 200,000 or so characters in by the gene. All the other characters are sitting in the introns.

Now, sequencing a short gene in its entirety is easy. Sequencing a large gene such as *RB* in its entirety, much more cumbersome and expensive. Since the recipe of the gene is focussed in just 2783 exonic characters, why not focus on the exons alone? Indeed, that is what whole exome sequencing does. It picks out the exons alone with some clever means, but of all 20,000 genes, and sequences just these. We could instead pick out just the exons of the *RB* gene and sequence these alone. This would be far less expensive. It turns out though that we can spread the net to include maybe a hundred genes, pick up exons from all of these, and sequence all these exons, at minimal extra cost. So we create a *panel* of genes, all related to disorders of the eye, and sequence all the exons in all these genes.

We extract $\mathscr{X}$'s genome from a few drops of his saliva and sequence the genome using this gene panel. This gives us the text stretches in all the exons of all these genes. Including, of course, the *RB* gene. We then turn eagerly to this gene looking for problematic variants.

Remember the different types of variants we encountered in previous chapters. There are missense variants, where the usual amino acid in the gene recipe is replaced by another, possibly more problematic one (Fig. 2.5). But we find none for $\mathscr{X}$. Then there are nonsense variants, where the recipe change is even more dramatic; the recipe stops dead in its tracks prematurely. Again, we find no such variant in $\mathscr{X}$. Then there are

frameshift variants, where the grouping of the recipe into triplets itself is jittered due to either extra characters or missing characters, whose count is not a multiple of three (Fig. 3.9). No such variants either! How about variants which might lead the recipe interpretation process astray as it makes splicing jumps from one exon to the next? Variants of the kind we saw in Chapter 6 leading to the blood disorder Thalassemia. None of these either! Remember that particular variant was in an intron. But here we have sequenced only the exons for $\mathscr{X}$. Aren't we missing any variants that might be in the introns of *RB*?

Fortunately, we have actually sequenced a bit more than just the exons. Our attempt to extract out the exons of the genes in our gene panel usually drags along with it bits of the introns as well. Each exon drags along 50-100 characters of the neighboring introns on both sides. And we have sequenced these characters as well. But, again, no variants here either! Where is this elusive variant hiding?

Maybe our elusive variant is hiding deeper inside the intronic ocean, far away from the shores of any of the 27 exons of *RB*. Our sequencing hasn't plumbed these depths, for dredging this intronic ocean is considerably more expensive. So it is certainly a possibility, but a rare one. A variant hiding deep inside an intron is less likely to lead splicing jumps astray; those which do so are typically not too far from the exon shores. Is there a more likely scenario that we are missing?

Gene Panels and Missing Chunks

Could there be a large chunk missing from the *RB* gene in $\mathscr{X}$? Something along the lines of the picture below, where the entire shaded region comprising exon 2 and substantial parts of the two sandwiching introns are missing in $\mathscr{X}$. This region is of course present in the reference genome (the genome sequence of an average, healthy individual), and therefore in most of us, but is missing from the $\mathscr{X}$'s genome. How would we check if there is indeed such a chunk?

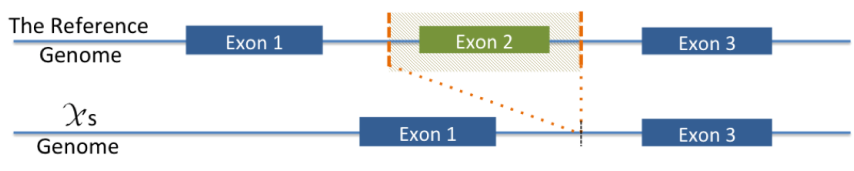


Figure 7.7: A Hypothetical Missing Chunk in $\mathcal{X}$.

Remember how we found the missing chunk in the genomes of two children who had their organs out of place, in Chapter 5. Many many copies of the genome were taken and shredded into small pieces, or reads. The sequencing process gave us the character sequence for each read. We took each read in turn and searched for it in the reference genome. This search was successful in most cases, particularly if we modified a character or two, or if we added a few extra characters or dropped a few characters. The search did fail for some reads though, hidden amongst which were reads which straddled the missing chunk. We split such reads into two parts as shown in the picture below.

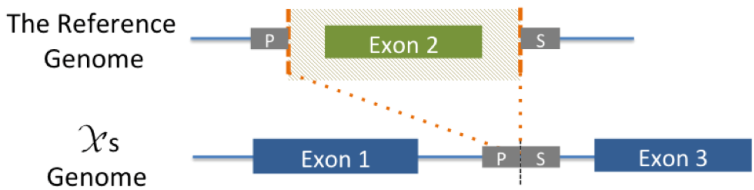


Figure 7.8: Finding the Missing Chunk by Splitting Reads.

We then searched for each part separately in the reference genome. The two parts, even though adjacent in $\mathcal{X}$'s genome, would find their respective matches far apart in the reference genome: one on either side of exon 2. And this would tell us that the stretch in between these matches was missing from $\mathcal{X}$'s genome.

Couldn't we use this method to discover if any big chunks of the

genomic recipe were missing from the *RB* gene in $\mathscr{X}$? We could, but for one catch. A rather serious catch, unfortunately! We have sequenced only the exons of *RB*. We have additionally nibbled a bit into the introns. Only a bit, not too much. So if the missing chunk has its boundaries deep into the introns, we would get no reads at all in $\mathscr{X}$ which straddle this chunk. And without any such reads, this method is a non-starter!

What had rescued us in Chapter 5 was whole genome sequencing; we had reads from all across the genome, from the exons and from the introns. Here, with our more economical gene panels, we have reads only from the exons. How could we determine if the first hit on the *RB* gene in $\mathscr{X}$ was on account of a missing chunk, without having to spend 5-10 times more to do whole genome sequencing?

The Search for the Missing Chunk

Our gene panel has roughly a 100 genes, together comprising roughly 2000 exons. Normally,$\mathscr{X}$ would have two copies of each of these exons. However, if the first hit on his *RB* gene was on account of a large missing chunk spanning an exon, $\mathscr{X}$ would be left with only one copy of that exon. He will continue to have two copies of all the remaining exons though. So *one* copy of the missing exon, versus *two* copies of all remaining exons. Can we exploit this key distinction to identify the missing exon, if there is one?

Remember, we took each read we got from sequencing $\mathscr{X}$'s genome and searched for this read in the reference genome. A read that came from exon 1 of the *RB* gene in $\mathscr{X}$'s genome would likely find its match in exon 1 of the same gene in the reference genome. Likewise, for reads from the other exons. We can then count how many reads find a match on each of the 27 exons of *RB*. Wouldn't the missing exon now stand out because it would get roughly half the number of reads as each of the other exons?

As always, there is a catch. For one, all exons do not have the same length. Some are longer, some shorter. Naturally, longer exons will have get more reads and shorter exons will get fewer reads. In this lop-sided playing field, the missing exon will not quite stand out, particularly if it were to be on the longer side. So we even out this playing field by dividing

the count on each exon by the length of that exon: call this the *coverage* of an exon. Here is the picture showing the coverages for each of the 27 exons of $\mathscr{X}$'s *RB* gene. Can we now identify the missing exon, if one exists? The hope of course is that the missing exon stands out on account on much smaller coverage. Does it?

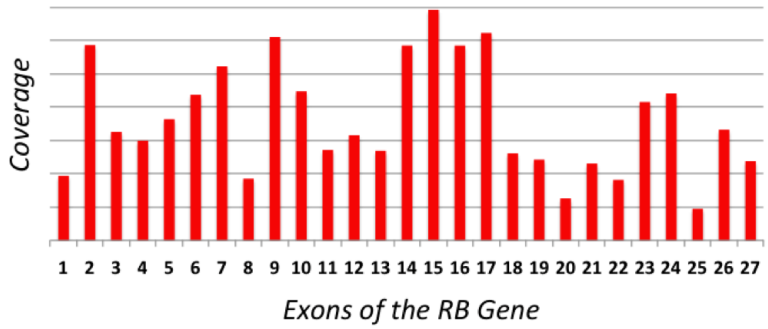


Figure 7.9: Coverages on the Various Exons of *RB*.

Well, exon 25 has low coverage. But so does exon 20. Could the missing chunk in $\mathscr{X}$ encompass both these exons? Unlikely, for exons 23 and 24 in between have good coverage; they don't appear to be missing at all. Disappointingly, there is no neat split of exons with poor coverage versus exons with good coverage. Some exons have poor coverage, others have coverage 7 times as much, and there is a continuum in-between! The missing exon, if there is one, hasn't announced itself in this picture. Maybe there is no missing chunk after all! Or maybe we made an incorrect assumption.

The First Hit

We assumed that all exons present in both *RB* gene copies in $\mathscr{X}$ will get roughly the same coverage. And that an exon present in only one of the two copies of *RB* will get roughly half of this coverage. But all exons are not the same, even if they occur in both gene copies. They carry distinct

sequences of characters. The process of extracting exons out from the entire genome for sequencing might provide preferential treatment to an exon depending upon this sequence of characters. Comparing the coverage on one exon with that on another would be like comparing apples and oranges then. Is there a way to just compare apples against apples, and oranges against oranges?

What if we compared exon coverages from $\mathscr{X}$'s genome with the corresponding exon coverages from another person's genome. A normal person, that is. Specifically, one who has never had Retinoblastoma, and therefore carries two copies of all the *RB* gene exons. Even better, we could average exon coverages over 50 such normal persons, so freak events for any one person do not upset our calculations. So we take the first exon's coverage from $\mathscr{X}$'s genome and compare it to the first exon's coverage from this pool of healthy people. We repeat this for each of the 27 exons of *RB*. Now, we are comparing apples to apples, oranges to oranges. For most exons, we expect the coverages to be comparable across the two individuals. However, if there is a missing exon, then the coverage in $\mathscr{X}$ will be roughly half of that in the other person. And that exon will stand out quite clearly: a ratio of roughly 1/2 versus a ratio of roughly 1. Did it?

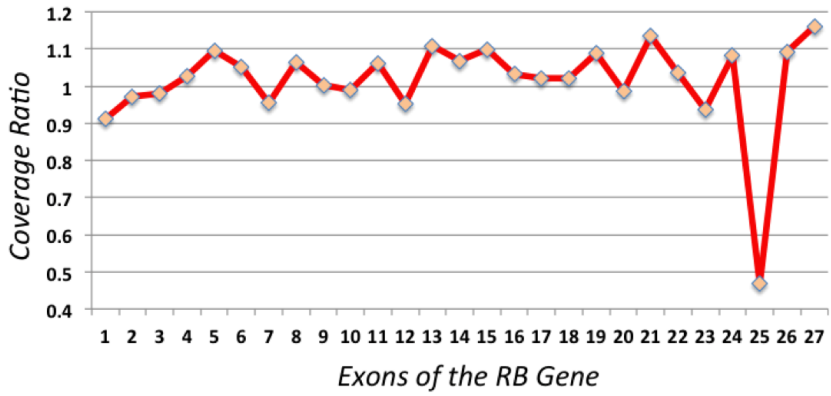


Figure 7.10: The First Hit Stares at Us.

Here, staring at us, mockingly, is the first hit on the *RB* gene in $\mathscr{X}$: exon number 25, dipping down to the 0.5 mark, leaving behind all the other exons around the 1 mark. Surely, there is a missing chunk around exon 25 in one of $\mathscr{X}$'s two *RB* gene copies! The boundaries of this missing chunk are probably deep inside the introns. We can't get to these boundaries because we have not sequenced so deep into the introns. But we know for sure that exon 25 is part of this missing chunk. This was the first hit on the *RB* gene that $\mathscr{X}$ was born with.

The Second Hit

Remember, $\mathscr{X}$ had bilateral tumors. So at least one cell in the retina of his left eye and one cell in the retina of his right eye had each incurred a second hit in the *RB* gene. As Knudson's data indicated, these second hits in $\mathscr{X}$ were not inherited; rather they were acquired after conception via a random process. What were these hits?

Only the first inherited hit in visible in the genome extracted from cells in $\mathscr{X}$'s saliva. This hit is present in every cell in $\mathscr{X}$'s body, so it is present in the cells in his saliva as well. In contrast, the second hits are present only in a few cells in $\mathscr{X}$'s retina. These cells, possibly along with others, comprise the tumors in $\mathscr{X}$'s eyes. To identify these second hits, we will need to extract the genome from $\mathscr{X}$'s tumor cells and then sequence this tumor genome.

Extracting tumor cells from $\mathscr{X}$'s tumors requires surgery, which is far more complicated than just asking $\mathscr{X}$ to spit out some saliva. Surgery of this type to identify imperfections in the tumor genome is an important part of our story, but in a later chapter. For $\mathscr{X}$, we will have to guess what happened without necessarily peeking into his tumor genome. And a very plausible guess is that the second hit in $\mathscr{X}$ is the same as the first one, but on the other copy of the *RB* gene![84]

Really, you might ask. What are the odds of that? How does the second hit manage to mimic the first one? The answer lies, possibly, in *recombination*, the phenomenon we saw in Chapter 1, albeit in a very different context.

Recombination Again

Remember, we have two copies each of our chromosomes (if you ignore the $\mathscr{X}$ and $\mathscr{Y}$ chromosomes). Which of these two copies does a parent pass on to his or her child? The answer is neither! What is passed on is a mosaic of the two copies (Fig. 1.14). This mosaic is created during *meiosis*, when cells in our testes or ovaries divide to create sperm and egg, respectively, capable of forming a new individual. Cells in our retina and other parts of the body divide using a different process, called *mitosis*. The goal of this process is to create new cells within an individual, rather than create cells capable of becoming a whole new individual. A similar mosaic formation occurs in mitosis, albeit much more infrequently. In fact, about a 100 times more infrequently.[85] Yet not infrequently enough to spare individuals like $\mathscr{X}$ who already had a first hit. Here is what might have happened to $\mathscr{X}$.

Remember, when cells divide, they go through a growth phase following which a copy phase copies the entire genome. Now focus on chromosome 13, the chapter of the genome containing the *RB* gene, and imagine one of the cells in $\mathscr{X}$'s retina as it divides.

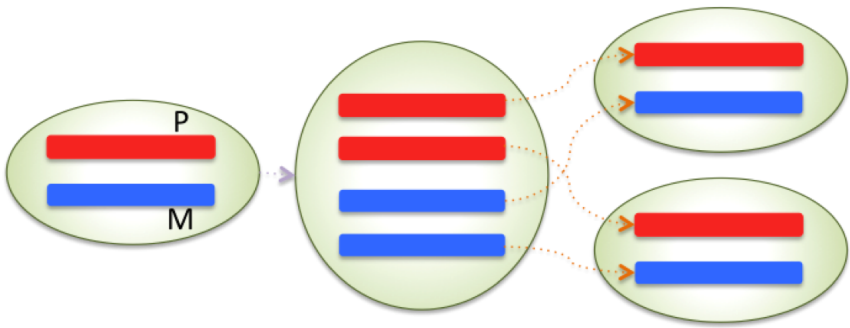


Figure 7.11: Genome Replication when a Cell Divides.

This cell has two copies of chromosome 13, one inherited from each

of $\mathscr{X}$'s parents. Accordingly, we can call these copies, *P* and *M*. The character sequences in these two copies are slightly different. Indeed, only one of these, *P*, say, carries the first hit. The *M* copy is perfectly fine. In the copy phase, both copies are copied yet again, yielding 2 copies of *P* and 2 copies of *M*.

A little after the copy phase, a split phase splits the cell into two new daughter cells. Usually, each daughter cell gets one *P* copy and one *M* copy. The good *M* copy protects each of the daughter cells from the inadequacies of first-hit stricken *P* copy. So, usually, all is well, in spite of the first hit, in both daughter cells.

Occasionally, a freak, rare recombination event spoils this harmony. This freak event usually occurs during or after the copy phase and before the split happens.[86] One of the *P* copies exchanges segments with one of the *M* copies, as shown below. This *M* copy now carries a segment of *P*. And the *P* copy, in turn, carries a segment of *M*.

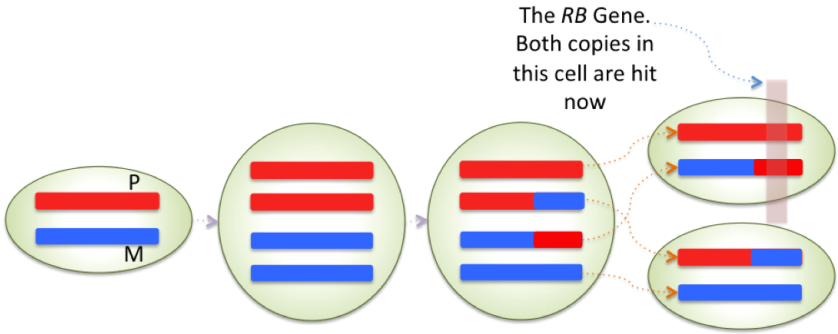


Figure 7.12: Mitotic Recombination Leading to the Second Hit.

Then, when the cell splits into two, one of the daughter cells gets a copy of *P*, along with this tainted copy of *M*: the one with a bit of *P* in it. If the *RB* gene happens to be located within the exchanged segment, the consequence is most unfortunate. One daughter cell now has the first hit in both copies of the *RB* gene. In other words, a second hit mimicking the

first!

Some cell in $\mathscr{X}$'s retina lost exon 25 from both its *RB* gene copies possibly due to such an event. A event rarer than one in a million maybe. But with a million cells in the retina, the odds of this happening? Not that low! In fact, high enough for it to happen in cells from both eyes. Then what? How did the loss of both copies of exon 25 of the *RB* gene in these cells lead to tumor formation?

Password to Nucleus Lost?

The *RB* gene has 27 exons. The 26th and 27th exons are quite short, so exon 25 is very close to the end. What happens to the *RB* gene recipe when this exon is no longer present?

When the exons of *RB* are strung together and partitioned into frames of 3 characters each (remember Fig. 2.5), a new frame begins at the start of exon 25 as shown in the picture below. However, the last frame in exon 25 spills over by one character to exon 26. With the deletion of exon 25, this character then starts a new frame in exon 26. Unfortunately, this triplet happens to be a *Stop* triplet, signaling the end of the recipe. And the *RB* gene recipe thus comes to an abrupt halt after exon 24 itself!

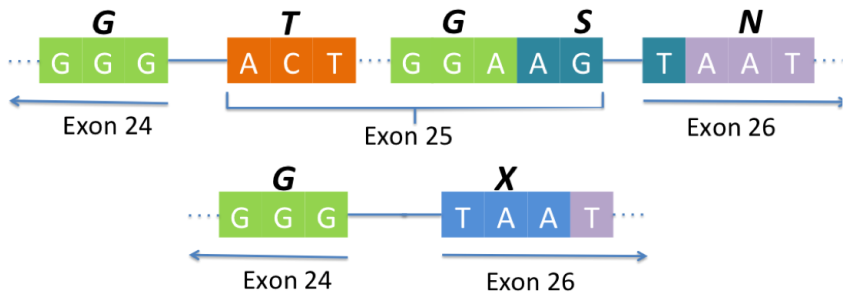


Figure 7.13: Amino Acid Sequence with and without Exon 25. The *Stop* triplet, indicated by *X* brings about an abrupt end.

Remember, proteins created from such prematurely truncated recipes

are often destroyed by cells (remember Fig. 3.11, a mechanism to not have useless debris lying around). If that is indeed the case, then the first and second hits together will leave $\mathscr{X}$ with no *RB* proteins at all. The picture is complicated though by the fact that exon 26 is the penultimate exon, and a short one at that, at 50 characters. Premature truncations within 50 characters of the end of the penultimate exon sometimes do escape destruction. So it is possible that $\mathscr{X}$ had the normal amount of *RB* protein, just that all of these protein copies were truncated at the end of exon 24. Could these truncated copies be the cause of tumors in $\mathscr{X}$?

The genome, being the precious treasure-house of gene recipes, requires careful protection inside the cell. This protection is provided by a spherical compartment in the cell we know as the *nucleus*. The nucleus is covered by a membrane (a sort of sheet) that ensures that the genome stays inside and that other molecules outside do not easily get in. However, the recipe interpretation process requires that the membrane allow some traffic through itself, as follows.

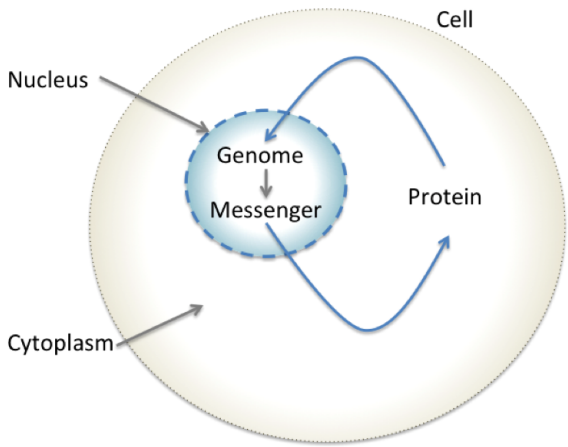


Figure 7.14: Recipe Interpretation: Out and Back into the Nucleus through Pores.

Recipe interpretation starts inside the nucleus. The exons and the introns of the *RB* gene in the genome are read and copied first. Splicing jumps then remove the introns from this copy, yielding a modified copy comprising only exonic characters. This modified copy is called the *messenger*. Messenger, because it carries the recipe from within the nucleus to the outside, the region of the cell called the *cytoplasm*. In the cytoplasm, this messenger is read again, and successive triplets of characters are converted to corresponding amino acids using the code in Fig. 2.5. The resulting chain of amino acids then folds into a characteristic shape to yield the *RB* protein. This protein must then find its way back into the nucleus from the cytoplasm.[87]

Blocking its march to the nucleus is the membrane of the nucleus, whose explicit goal is to watch out for gatecrashers. Little pores in this membrane do allow some traffic; however, these pores are not large enough for the *RB* protein to get through. Other helper proteins are needed to force this cargo through these pores. These helper proteins are very selective though, so not all cargo is equally eligible. The cargo must carry a password, i.e., a specific sequence of amino acids, for it to be eligible. Typically, the *RB* gene recipe includes this password: a sequence of characters located right in the middle of exon 25![88]

With exon 25 lost in the *RB* gene in $\mathscr{X}$, the exclusive password to the nucleus is lost. Some *RB* protein still manages to sneak in. But nevertheless, the net result: deletion of exon 25 in both copies of *RB* results in lesser amounts of *RB* protein in the nucleus and more in the cytoplasm[88] in $\mathscr{X}$'s cells: a signature typical of many cancers![87] How exactly does this lead to cancer?

Programmed Suicide

Remember, cells are programmed to divide. The *RB* protein, in its singly-marked state inside the nucleus (Fig. 7.6), puts a brake on this program. It holds the cell in its growth phase, disallowing it from progressing to the copy phase. When the time is right, multiple-marking of *RB* causes this brake to come off, letting the cell slip out into the copy phase. Thereafter, there is no going back; the cell must head inevitably towards division.

In some of $\mathscr{X}$'s cells, specifically those with two hits, there is less *RB* in the nucleus to hold back the copy phase. These cells then lose their patience, rushing prematurely into the copy phase, and therefore towards cell division. Was this loss of patience and rush to divide the cause of $\mathscr{X}$'s cancer?

At the heart of this question lies a mystery? Why do two hits on the *RB* gene within the same cell lead to cancer, only in the retina, and only in the first few years of life? The *RB* gene does its job of stemming the mad rush to divide in many other cells in $\mathscr{X}$'s body. Why then do cancers occur only in the retina? We know only bits and pieces of the answer today. But in essence, nature poses further hurdles that prevent such impatient cells from marching on in the quest for unhindered division.

One such hurdle is is best exemplified by comparing a duck's paddle feet to our fingers and toes. You will surely notice the key difference: webbing. Ducks' feet have fingers which are connected together. Human fingers and toes lack this webbing and can therefore move more independently of each other.

Figure 7.15: Webbing in Ducks and No Webbing in Humans.

We may never realize this, but our fingers and toes do indeed start off with this webbing. Somewhere along the way, it is lost; well before birth, of course. 6-8 weeks from conception, when we are still tiny foetuses, the cells which form this webbing just die! The cells in the fingers and toes are then left to grow and divide to create full-sized digits.

How do cells in the webbing know they have to die? And why to they voluntarily die while their neighbors continue to live and thrive? All the cells together secrete some substances which then spread out over the entire area spanned by the digits and the webbing. The concentration of these substances though is not uniform; some regions build up higher concentrations and others get lower concentrations. Cells which are placed in regions with, say, higher concentrations, detect this and initiate several changes, including recipe interpretation on specific genes, culminating in very organized suicide. This *programmed death*, i.e., the unselfish ability of cells to undergo organized suicide in response to various triggers, is nature's way to counterbalance excessive cell division. And nature maintains this balance using several double-edged swords. The protagonist of our plot, the *RB* gene, is one such!

Just as the *RB* protein puts a brake on the cell division, so does it put a brake on programmed cell death. With its two hits in $\mathscr{X}$ reducing the amount of *RB* in the nucleus, the brakes on cell division are removed. But so are the brakes on programmed death. So $\mathscr{X}$'s two hits do not lead to cancer. Not yet! Instead, cells with these two hits typically die.[89] And that takes our mystery one step deeper, for $\mathscr{X}$ definitely had tumors. Not just one, but two, one in each eye. How did these come about?

Divide, Die, Differentiate?

In addition to division and death, there is a third piece to the cellular puzzle. Different cells specialize in different functions. Heart muscle cells specialize in contraction, retinal photoreceptor cells specialize in detecting light, and red blood cells specialize in carrying oxygen. Each type of specialization is obtained by running a different program, which means turning on and off recipe interpretation on different sets of genes. And there are many many different specializations.

Coming to the retina: it too comprises cells of not one but many specializations. Broadly, there are three layers of cells in the retina. The first layer comprises cells that detect light and convert these to electrical signals: these are the rods and cones we saw in Chapter 1. The middle layer comprises many cell types which transform signals generated by

the rods and cones to extract additional features, like contrast. From this middle layer, the signal flows into the third layer. This layer comprises neural cells which transmit the signal to the brain. The cells in these layers are laid out in a specific, careful architecture, as shown below.

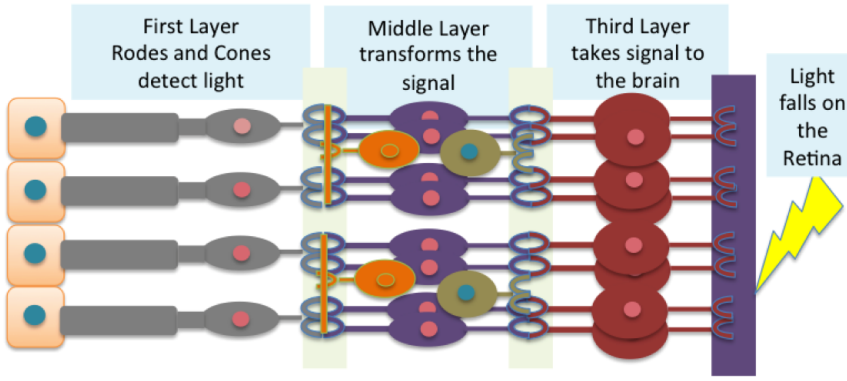


Figure 7.16: Cell Types in the Retina.

Of course, this variety of cells, laid out so carefully, is obtained by repeated cell division starting from a single cell. As division progresses, a different program takes hold: the differentiation program. It drives different cells towards different specializations by turning on or off recipe interpretation on different sets of genes. Some of these cells are driven towards rods and cones by this differentiation program, others are driven towards cells in the middle layer, yet others towards cells in the third layer. This happens over multiple generations of cell division. And not all these cells survive.

Much like the webbing between our fingers, nature creates sophisticated architecture by first generating more cells than it needs and then inducing programmed suicide in carefully chosen cells. The cells which survive, now fully specialized, enter a sort of cellular nirvana, an escape from the cycle of repeated division. Brakes are put on the cell division program in these cells, so they no longer divide. They are then frozen as

they have to hold the architecture needed for the retina to function.

And thus is our retina formed by a careful interplay between the division, differentiation and death programs. Just as the *RB* protein in its single-marked form keeps the brakes on division, letting go when it is multiply-marked, the *RB* protein without any marks on it at all controls parts of the differentiation program.[83] What happens to the interplay between the division, differentiation and death programs in $\mathscr{X}$'s retinal cells, when some of these receive two hits on this master orchestrater gene?

$\mathscr{X}$'s Cancer

Scientists have studied what happens to the various cell types in the retina when the *RB* protein is completely removed. Many cell types still continue differentiating and enter their cellular nirvana, escaping from the cell division cycle. The death program consumes the others. The rare cell though hangs in balance. In such cells, the death program is overwhelmed by the differentiation and the division programs. The differentiation program proceeds to an extent, but is in turn, overwhelmed by the division program. Only very recently have these cells been pinned down:[90] these are partly differentiated versions of cone cells, the same cells we met in Chapter 1 responsible for color sensing.

Only in these pre-cone cells is the combination of genes turned on or off by recipe interpretation capable of upsetting the fine balance between division, differentiation and death, making cell division predominant. With no *RB* protein at all, there are no brakes on cell division, so it proceeds at reckless speeds. Cancer follows! Other cell types, in the retina or in the rest of the body, provide a more hostile environment for cell division, so tumors only form starting from these pre-cone cells.

We can only guess, but it is likely that something similar happened in $\mathscr{X}$. Some pre-cone cell in his retina lost exon 25 from both gene copies due to a second hit. This forced lesser amounts of *RB* protein in the nucleus of this cell and its daughter cells, mimicking the situation above. The balance between division, differentiation and death was then won by division. And this happened not just in one eye, but both eyes, leading to bilateral tumors. Had the second hit just waited for a few years giving time

for these pre-cone cells to differentiate enough and attain their cellular nirvana, $\mathscr{X}$ could have escaped cancer. But the odds of the second hit ensured that this was not to be.

Wrapping Up

Cancer happens when nature's intricate balance between cell division, cell death and cell differentiation, goes awry. It just takes one abnormal cell for this balance to be disturbed. That cell then divides to generate more abnormal cells. Like a snowball that gathers more and more snow as it rolls down the slope, these cells become more and more abnormal as they divide repeatedly. This uncontrolled division destroys organized cellular arrangements required for proper functioning. When abnormality levels cross a certain further threshold, these cells break away and wander off to other parts of the body, carrying the same destructive effect to those parts, eventually bringing down the whole system. Unless things change dramatically, one in 4 or 5 persons alive today will eventually die of cancer.

On a comforting note, it is not easy for a normal cell to turn cancerous. Several hills have to be climbed by the normal cell to this end, and that requires several hits on the genome. As early as 1953, Carl Nordling studied the incidence of cancer with age (shown in the picture below) and observed a certain pattern of increase. For instance, from age 20 to 40, the risk goes up 5-fold, and from age 40 to 60, the risk goes up 10-fold. A more careful analysis of this curve suggested that the incidence of cancer increases as the 5th-6th power of age. This would suggest that 5-6 hits, or genomic variants, might be needed for several cancers.[91] Of course, this is a very coarse figure. The actual number can vary depending upon which genes and which cell types these hits occur in and what types of hits these are.

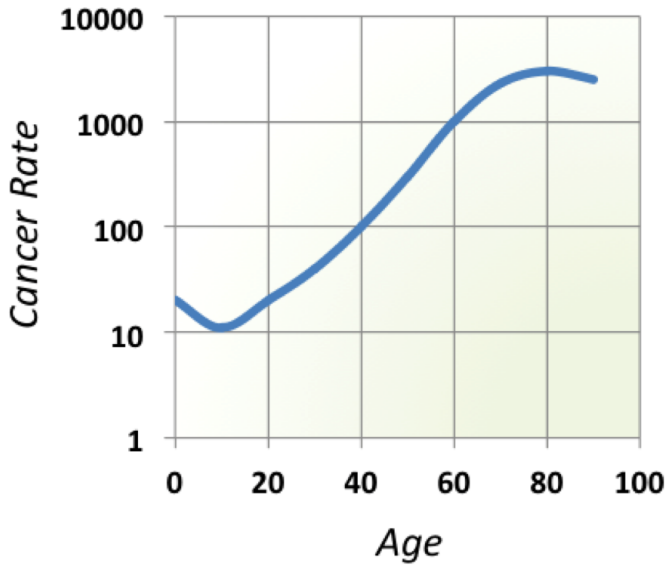


Figure 7.17: Cancer Risk with Age, from.[92]

Knudson's simple but clever analysis from handful of observations suggested that at least one cancer, namely Retinoblastoma, is caused by just 2 genomic hits in a gene which we now call *RB*. In $\mathscr{X}$'s case, the first hit on this gene was on account of a missing chunk comprising the whole of exon 25. We managed to tease out this chunk at a fraction of the cost we incurred for teasing out a similar chunk in Chapter 5. The second hit was a somatic variant, not inherited, but acquired by some of $\mathscr{X}$'s cells. There is a good chance that this second hit mimicked the first, possibly due to a random recombination event. A type of event we can only estimate chances of but cannot predict with certainty.

Fortunately, $\mathscr{X}$'s condition was diagnosed early enough to save his life, before the cancerous cells spread from his retina to other parts of his body. One of his eyes, the one in which the tumor had progressed further, had to be removed though. Cancerous cells in the other eye were killed by

treatment, while retaining vision. $\mathscr{X}$ now goes to a regular school and can do much of what other children can.

Neither of $\mathscr{X}$'s parents has Retinoblastoma. However, his father, without even realizing it, had what is called a *Retinoma*, a tumor that had started to form and then regressed. Maybe, the odds of the random second hit, or some randomness in the balance between division, differentiation and death had just worked in his favor. Chances are $\mathscr{X}$ inherited his first hit from his father.

If $\mathscr{X}$'s parents were ever to have another child then they would know what to do now. They could test whether this child, when still inside the mother's womb, carries the first hit, the removal of exon 25, or not. Even if it did, there wouldn't be cause for despair, for much of the development of the retina happens at the fag end of the pregnancy period. Simply by delivering this baby a few weeks pre-term and starting treatment promptly,[93] any tumors in the retina can be caught early and halted in their tracks, saving both sight and life!

Chapter 8

Repair out of Repair

8 year old $\mathscr{X}$ appeared much more fatigued than other children his age. He had also recently developed several tiny purple spots on his skin, caused by bleeding in small blood vessels lying just below. His 6 year old sister too had started to show similar symptoms. A blood test on $\mathscr{X}$ showed markedly reduced counts of various blood cells. After eliminating several possible causes of these phenomena, $\mathscr{X}$'s pediatrician performed what is called a *chromosome breakage* test.

Remember chromosomes (Fig. 1.9); the 46 books which together comprise the genomic bookshelf. A chromosome breakage test takes a few cells from $\mathscr{X}$'s blood and induces these to grow and divide in a dish in the laboratory. These cells are then subjected to specific drugs. The impact of these drugs on the chromosomes is studied by viewing them under a microscope. Normally, each of the 46 chromosomes would appear as in the picture below.

Figure 8.1: A Normal Chromosome and Its Copy.

This picture shows just one of the 46 chromosomes. Two entire copies

have been made of this chromosome as part of the genome copying process which takes place when cells divide. The two copies are connected at the center. Eventually, the copies will separate and make their way into distinct daughter cells. 46 such chromosome pairs would comprise the normal picture.

Some of $\mathscr{X}$'s chromosome pairs indeed matched the expected picture above. Not all though. Some appeared to be quite odd, as if multiple pairs had fused together, as shown below!

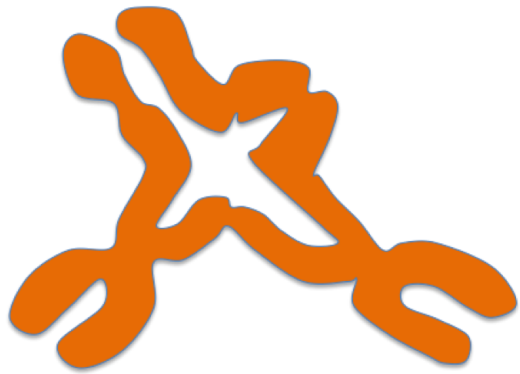


Figure 8.2: Multiple Chromosome Pairs Fused in $\mathscr{X}$.

As the test name indicates, the drugs used in the test cause certain abnormalities in the chromosomes. Cells from normal individuals usually fix these abnormalities. In contrast, $\mathscr{X}$'s cells seemed to have jumbled up some of these chromosome pairs in the process of correcting these abnormalities. A Swiss pediatrician, Guido Fanconi, was the first to identify this condition, as early as in 1927. $\mathscr{X}$'s condition now carries his name: Fanconi's anemia.

Anemia, or the lack of sufficiently many good red blood cells, was the cause of $\mathscr{X}$'s fatigue. There weren't enough red blood cells to carry oxygen in sufficient quantities to his tissues. And the purple clots under

his skin were caused by bleeding due to insufficient numbers of platelets in his blood. Platelets help blood clot; too few platelets and the blood refuses to clot. Both red blood cells and platelets (along with white blood cells) are manufactured in the bone marrow, a blood-cell making factory located inside our bones. $\mathscr{X}$'s symptoms indicated that his bone marrow was failing to generate these cells in sufficient quantities. Further progression of this failure could be life-threatening. For instance, $\mathscr{X}$ could be at risk of bleeding to death on account of a slight bruise, or suffering from the onslaught of repeated infections, or suffering organ failure on account of decreased oxygen supply to the various organs.

There are therapies available that could stimulate $\mathscr{X}$'s bone marrow to produce more red blood cells and platelets. In the medium term though, $\mathscr{X}$ might need blood transfusions, or even a bone marrow transplant, much like the girl with Thalassemia we met in Chapter 6. A successful transplant would give $\mathscr{X}$ a reprieve. But time could present further challenges. There was a 75% chance that one of $\mathscr{X}$'s cells would spin out of control and turn cancerous by age 45.[94] In comparison, the average person's risk by age 45 is under 5%. $\mathscr{X}$'s risk was 15 times as much.

What in $\mathscr{X}$'s genome caused his bone marrow to fail and his risk of cancer to shoot up dramatically? And how was that related to his chromosomes appearing jumbled up under the microscope?

The Cancer Hill

For the average person, the risk of cancer grows with age. The juggernaut of life inevitably leads each of us up this cancer hill. And women appear to climb this hill earlier, as the picture below indicates.

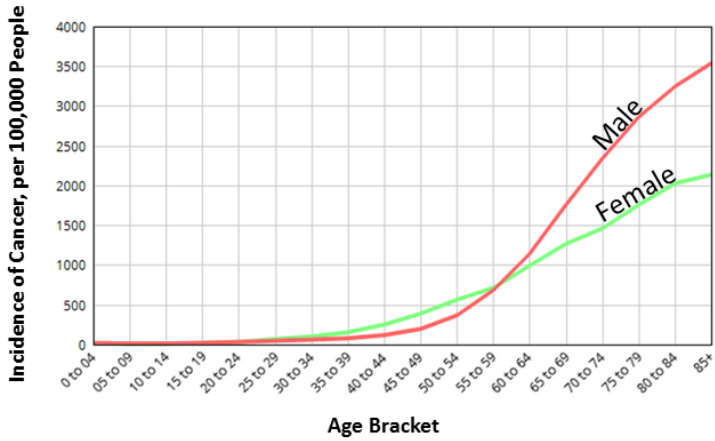


Figure 8.3: The Cancer Hill: Increasing Frequency of Cancer with Age, for Males and Females. This is based on data from Cancer Research, UK, 2009-2011.

Men are late starters, relatively speaking, but more than catch up with women in their 60s and 70s. However, between the ages of 30 to 60, the risk for women runs higher. And much of this increased risk is on account of breast cancer. Statistics today tell us that 1 in 8 or 9 women will get breast cancer in their lifetime. Most such women will not have a strong family history of breast cancer; the disease seemingly comes out of nowhere in these women. But there are exceptions.

As early as in the 1940s, doctors had observed that 16% of women with breast cancer did have a strong family history of breast cancer.[95] As more and more patients were observed and the cancer risk in immediate relatives of breast cancer patients was carefully quantified,[96] the hypothesis of familial risk grew stronger. The risk for a woman was higher if an immediate relative had breast cancer. And it was much higher when the immediate relative got cancer at an earlier age rather than a later one.[96] For instance, imagine a woman with a sister and mother affected by breast cancer. If the sister was diagnosed at age 50, then the woman's risk was 5.9 times that of the average person. But if the sister was diagnosed at age

30, then the woman's risk was as much as 15.1 times that of the average person! What did the woman and her sister share in common? Maybe a home, upbringing, neighborhood, environment, family, friends? But also characters in the genome. Which of these was the common cancer link between the sisters?

A 17-year marathon effort was conducted by noted researcher Marie-Claire King and her colleagues at the University of Berkeley from 1974 to 1990 to answer this question. In that era, comprehensively scouring the genomes of families with multiple breast cancer instances was impossible. Even getting a few glimpses into their genomes was painstakingly difficult. No encouragement came from the fact that conventional wisdom of the day doubted the presence of genes in the genome predisposing individuals to cancer; particularly, cancer that set in after a few decades of life. Nevertheless, this marathon effort eventually bore fruit: incontrovertible evidence that variants in a specific gene on chromosome 17 increased the risk of breast cancer substantially.[97] The gene itself was identified only a few years later, in 1994. It is, arguably, the gene with the greatest presence in the public arena among all the 20,000 or so genes in our genome. A gene that will tie, albeit indirectly, to $\mathscr{X}$'s travails.

The *BRCA* Genes

This gene was appropriately christened the *BRCA1* gene (pronounced *braca one*, short for breast cancer one, early onset). Shortly after its discovery, another risk susceptibility gene was identified by Michael Stratton and his group at Cancer Research, UK, in 1995, and named along expected lines: *BRCA2*.[98] Around 60% percent of women who inherit a problematic *BRCA1* variant will develop breast cancer by age 70. Likewise, for around 45% of women who inherit a harmful *BRCA2* variant. And this happens often with an earlier age of onset. Compare this with the average woman, who has only a 12% lifetime chance.

So significant is this increased risk that women carrying problematic *BRCA1* or *BRCA2* variants are offered the rather drastic choice of risk-reducing surgery: preemptive removal of their breast tissue, followed by reconstruction. Of the many women have exercised this option, actress

Angelina Jolie's case was the most-publicized.

In a open letter in the New York Times in 2013[99] that created much awareness about the *BRCA1* genes, Jolie stated that her mother had fought cancer for almost a decade before she succumbed to it at age 56. Jolie had a problematic *BRCA1* variant and decided take the drastic step of risk-reducing surgery to preempt that tragic fate. As a result, her risk of breast cancer reduced from the 85% number down to about 5%, as estimated by her doctors. She continued to carry a 55% risk of developing ovarian cancer, for problematic *BRCA1* variants predispose to both breast cancer and ovarian cancer (and a few other related cancers). Women with *BRCA1* and *BRCA2* variants are indeed offered the option of surgically removing their ovaries after expediting plans to have children, if required. A little later, Jolie announced that she had indeed undergone elective surgery to remove her ovaries and fallopian tubes as well.[100]

Given these choices, testing of women for *BRCA1/2* gene mutations became a topic of much business interest from the late 1990s onwards. This, in turn, led to one of the most intensely controversial patenting questions of all time: can genes be patented?

Patents for the *BRCA1* and *BRCA2* genes were granted to a company called Myriad Genetics and its academic partners in 1997 and 1998, respectively. This provided Myriad Genetics the exclusive right to test women for problematic *BRCA1/2* mutations, blocking all other companies from doing so, resulting in the testing price being held high, at above $3000. Arguments and counter arguments were made for the validity of this patent over the years. Genes were, after all, natural entities on the one had, which cannot be patented. On the other hand, Myriad was running their test not on the gene itself, but on a copy of the gene, a modified entity which they argued, was a man-made invention. What extent of modification qualifies for patentability then became the moot-point. That debate still goes on. Meanwhile, the costs of sequencing just these two genes dropped well below just a few hundred dollars, adding to much public opinion against these patents. Several disputes are on-going in courts, though the wind seems to be blowing in favor of the courts setting aside Myriad's patent, paving the way for broader access and lower costs.

But what do the *BRCA1/2* genes have to do with $\mathscr{X}$'s condition? These

genes have alternate names: *FANCS* and *FANCD1*, respectively. In both cases, the prefix *FANC* stands for Fanconi's Anemia. The same Fanconi's Anemia that $\mathscr{X}$ was afflicted with!

Genomic Assaults

FANCS and *FANCD1* are not the only genes with the *FANC* prefix. There are several others. Variants in all these genes are known to cause Fanconi's Anemia. And all these genes have a common role: to protect the genome from the impact of various destructive onslaughts that it is constantly subjected to. Some of these onslaughts occur occasionally: for instance, radiation from X-Rays. But there are others that are more regular; every minute, in fact!

As our cells generate energy from food to fuel their daily activities, they also generate some highly reactive molecules containing oxygen (called *ROS*, or reactive oxygen species), as a by-product. These ROS alter our genomes in subtle ways by reacting with various genomic characters. Genomic characters are, after all, not just plain characters; they are molecules, as shown in this picture, which can react with ROS and get modified as a result.

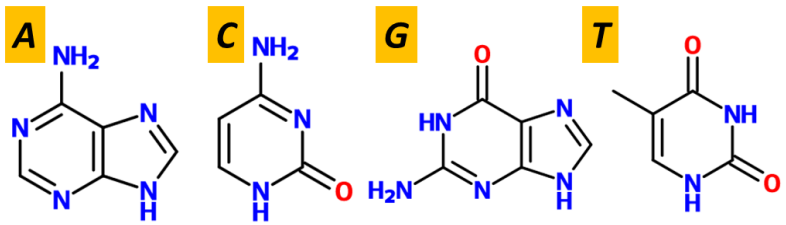


Figure 8.4: Molecules corresponding to the 4 Genomic Characters. Note that C and T looks somewhat similar.

For instance, reaction with ROS can sometimes convert a G to a form called OH8-G, due to the addition of an extra Hydrogen and Oxygen atom, as shown below. This unusual G has some strange properties, to understand which we need to appreciate the *double-stranded* nature of the genome.

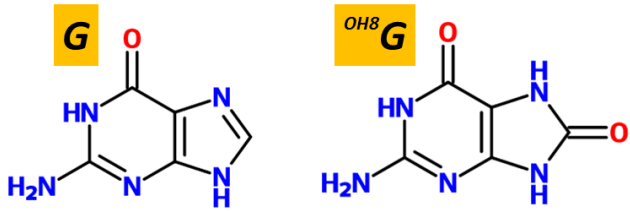


Figure 8.5: The Bona-Fide G, and OH8-G with its Extra Oxygen and Hydrogen.

So far, we have portrayed the genome rather simplistically as just a sequence of characters A, C, G, and T. Each copy of the genome actually comprises, not one, but two sequences, or *strands* as they are called. And the two strands are complementary to each other. For instance, if there is an A on one strand, there is a T on the other. And if there is a G on one strand, there is a C on the other.

Figure 8.6: The Two Strands of the Genome.

These A-T pairs and C-G pairs are held together by chemical bonds. The A-T pair has two bonds. The G-C pair comes in stronger at 3 bonds. Together, these bonds hold the two strands together. Most often, we just talk about characters on only one of these two strands. This is convenient and we don't lose any information in the process, for the other strand is perfectly complementary. But occasionally, it is useful to bring in the other strand.

For instance, while G pairs only with a C on the other strand, the unusual OH8-G can actually pair with an A, thus breaking the standard rules! This capability for abnormal pairing has unintended consequences, when the genome is copied. Suppose we start with a proper G on one strand and the expected C on the complementary strand (call this the C-G pair). Now, say, reactive oxygen containing molecules convert this G to OH8-G.

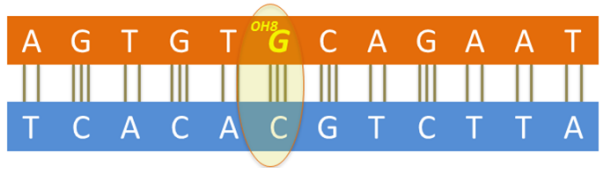


Figure 8.7: ROS Convert a G to OH8-G.

Next, suppose the cell starts to divide. In the process, an entire copy of the genome is made, so both daughter cells can get a copy each. This copying process first strips apart the two strands by breaking the bonds which hold complementary characters together, much like unzipping a zipper. Once unzipped, a brand new complementary strand is created for each of the two original strands, as in the picture below.

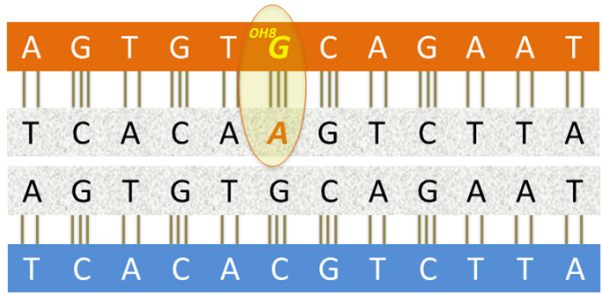


Figure 8.8: How the Genome is Copied.

This yields two copies of the genome, one for each daughter cell. Note, the first copy in the picture above is strange: the aberrant OH8-G is paired with an A instead of a C! What happens next when the daughter cell inheriting this copy divides? One of the resulting granddaughter cells will now have a T-A pair where there should have been a C-G pair. And all her descendants will carry this T-A pair, blissfully unaware that life began with the C-G pair which was altered somewhere in history due to damage caused by ROS.

There are several other character changes which ROS could cause. For example, you might notice above that the molecule for character C looks somewhat similar to that for character T (Fig. 8.4). Indeed, ROS can convert the occasional C to a form close to T; when the genome is then copied, that character actually does become a T. Occasionally, ROS can cause a G to be simply removed from the genome. When the genome is then copied, it will have one character less.

And so do characters in the genome mutate as cells divide, under the impact of ROS. These mutations accumulate with age. When enough of these accumulate in a single cell, things could start going wrong, as we'll see.

Note the pedestrian nature of these genomic changes. They are the by-products of routine everyday life. Our cells do have protective mechanisms: some of these rogue ROS molecules are scavenged by proteins created from other genes in our genome; others are stopped in their tracks by anti-oxidant vitamins in our diet. Yet, enough escape to create 10,000-100,000 genomic hits per day per cell comprising tens of different types of changes.[101] Adding fuel to fire are additional insults: cigarette smoke, asbestos fibers, diesel exhaust, fine particles in the air,[102] ultraviolet light,[103] and various other pollutants. All of these add to this baseline risk substantially.

$\mathscr{X}$'s genome was no doubt exposed to all these onslaughts. But so was everyone else's genome! What then made $\mathscr{X}$ different?

Repairing the Damage

What happens when the disk in your computer develops a mechanical fault? The disk carries all the programs that your computer runs. If it develops a problem, say a scratch or a mechanical failure, then these programs can no longer be read correctly; the computer then fails to function normally. So would parts of $\mathscr{X}$'s system, if the genome in certain cells in his body were to develop scratches and faults. Unless, of course, the genome can repair itself. Which it can, to good measure, in most people!

For instance, suppose a G becomes OH8-G on reaction with ROS. Our cells will promptly detect this change. Several proteins will then be set upon the task of correcting this error. The erroneous characters, and sometimes a few surrounding characters, are snipped off by these proteins. That leaves a gap in the genome. Other proteins then come in to fill this gap. But how do they know the right characters to fill into this gap? This is where the double-stranded structure of the genome helps, as in the picture below.

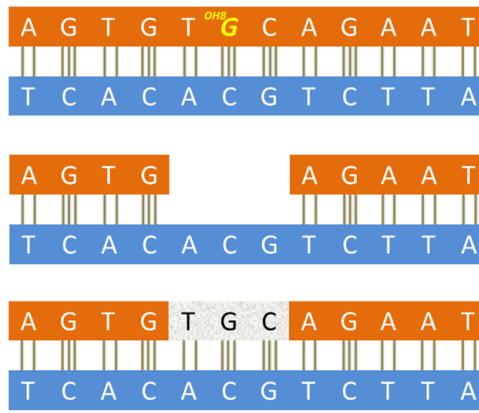


Figure 8.9: Repairing a Defective Character by Snipping-Out and Filling-In Using the Other Strand.

As long as the gap left by snipping out characters is on only one of the two strands, the other strand informs the right characters to be filled in.

Remember, if one strand has a C, the other has a G, and if one has an A, the other has a T. The gap is then filled in using the other intact strand as a template to copy from, as illustrated in the picture above.

Of course, if opposing characters on both strands were to be damaged, the task of repair would be tougher! Possibly, what made $\mathscr{X}$ different from everyone else was the ability to handle some of these tougher repair scenarios. How would such situations arise?

Cross-Links and Genomic Surgeries

Usually, a genomic character in one of the two strands, say an A, is bonded with the complementary T on the other strand. This bond is just strong enough to keep the two strands together and is easily broken when the strands need to be unzipped for the genome to be copied. Occasionally though, a much stronger link can form between a character on one strand, say a G, and not quite the complementary C but some other nearby character on the other strand, say a T. These two characters are linked together by other problematic molecules.

Figure 8.10: Cross-Links in the Genome.

This *cross-link* is so strong that it prevents the two strands from being unzipped easily during genome copying. The whole genome copying process then stalls. Which means that the cell carrying such cross-links cannot divide to create daughter cells anymore.

Fixing these cross-links is harder and no wonder the repair procedure is an elaborate and challenging one.[104] Let's start with the situation where the two strands have been unzipped and copies created on both sides of the

cross-link as shown in this picture. The cross-link prevents the unzipping and copying from completing.

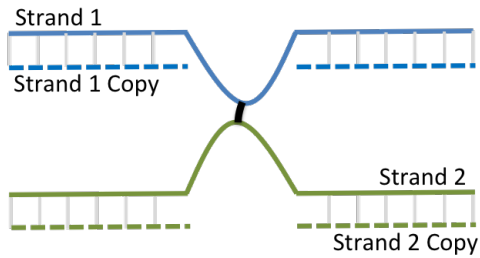


Figure 8.11: The Cross-Link Stalls Genome Copying.

As before, special proteins now detect that the cross-link has stalled the copying process. They recruit an army of other proteins, each of which performs one or more of the various steps of this elaborate process. First, several characters around the cross-link are snipped off from one of the two strands, thus unhooking the stretch with these characters. This stretch completely detaches from its strand, but remains attached to the other strand because of the cross-link, as in the picture below.

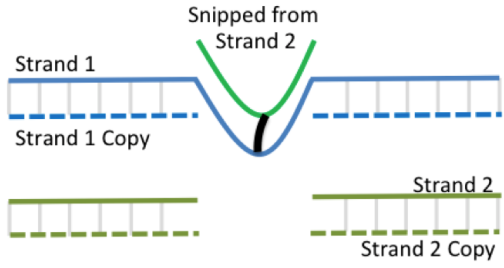


Figure 8.12: A Stretch Containing the Cross-Link is Unhooked from Strand 2.

Strand 1's copy is now ready to be completed. Since strand 1 itself is still intact, this is easily done; each base in the strand is just replaced by a complementary base in the newly created copy, as shown below. The snipped-off portion is eventually dropped as well, using a similar process of snipping-off and filling-in.

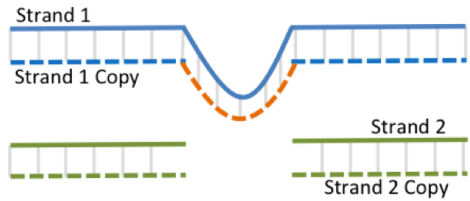


Figure 8.13: The Copy of Strand 1 is Now Complete.

Copying strand 2 poses a greater challenge: what does one fill into the missing region? The only place where this information is available is in strand 1 and its copy, which is now complete. But strand 1 and its copy together constitute a separate, very large molecule. And the information required to fill in the missing region in strand 2 is buried somewhere deep inside this molecule and needs to be fished out. Indeed, our cells have evolved clever ways to fish out this information. Rather strangely, this process begins with the removal of some more portions from strand 2 and its copy, as shown in this picture.

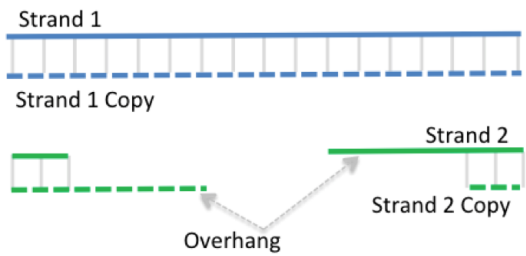


Figure 8.14: Some More Portions of Strand 2 and its Copy Removed.

Momentarily, the removal of good stretches from strand 2 and its copy adds to already existing damage. However that is part of a large game-plan towards repair. Both strand 2 and its copy now have overhanging portions in which the characters are not bonded to their complementary characters. These overhang portions, assisted by some helper proteins, probe strand 1 and its copy for information on the missing section of strand 2, as in the picture below.

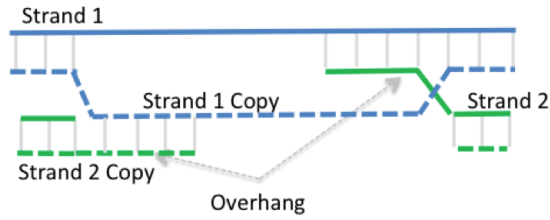


Figure 8.15: The Overhangs in Strand 2 and its Copy perform a Homology Search to Locate Complementary Stretches in Strand 1 and its Copy, respectively.

For instance, strand 2's overhang probes strand 1 for a complementary stretch (which means that for every A/T/G/C in the overhang there is a T/A/C/G, respectively, in strand 1). Miraculously, this so called *homology search* is accomplished quite accurately by our cells. Once such a stretch is found, the overhang causes unzipping of strand 1 and its copy, as shown in the picture above, allowing the overhang to bond to its complementary stretch. And likewise, the overhang in strand 2's copy does the same to a complementary stretch in strand 1's copy. The stage is now set for filling in the missing piece, as shown in the next picture.

The gap in strand 2 is filled in using the information provided by strand 1. To this end, strand 2's overhang is simply extended with characters that are complementary to strand 1. Likewise, the overhang in strand 2's copy is extended with characters that are complementary to strand 1's copy. Once both extensions are done, we have two full copies of the genome.

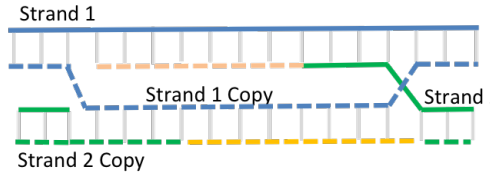


Figure 8.16: The Overhangs Now Extended Using the Complementary Strands.

One final task still remains: that of disentangling the two copies, for they are rather strangely entangled at the moment. There are different ways in which this surgery is performed, but one common way ends up cutting the various strands in a way that yields the final result shown in this picture.

Figure 8.17: Both Strands Fully Copied Now.

So many surgical steps, just to fix a little cross-link. It would have been far simpler had nature evolved a direct way to break and dissolve these cross-links. But all we have is this more convoluted multi-step surgical route. This surgery has to be carefully shepherded by proteins obtained from a number of genes. Over the years, several individuals have been observed with problematic variants in these genes. In these individuals, the above surgery doesn't go quite per plan. Something fails, possibly in the homology search step and the subsequent step of filling in the missing stretch using the other copy.

When a drop of blood is taken from such individuals and the genome subjected to artificially high levels of cross-linking using certain drugs, the outcome can be surprising: some chromosomes mistakenly stitched

together, sometimes in pairs, sometimes even triplets: the disorderly aftermath of several botched surgeries. Indeed, $\mathscr{X}$'s chromosomes too presented such a chaotic picture (Fig. 8.2). Clearly, one or more of the genes responsible for the above careful genomic surgery was malfunctioning in $\mathscr{X}$, leading to Fanconi's Anemia. There are several such genes known. Like the *BRCA1* and *BRCA2* genes, they all carry aliases with begin with the prefix *FANC*, short for Fanconi's Anemia.

The *FANC* Genes

The list of *FANC* genes is long and neatly labeled in alphabetical order: *FANCA*, *FANCB*, *FANCC* and so on; 17 and counting at the moment. Problematic variants in each of these genes have been found in patients with Fanconi Anemia. For simplicity, let us drop the *FANC* prefix which is common to all of these and refer to these genes just as *A*, *B*, *C* and so on. Most letters are used only once, but the occasional letter may be taken by two genes, for instance, *D1* and *D2*. Together, these genes, along with a several others, work in tandem to carry out their mission: detecting certain types of damage to the genome and orchestrating the careful surgery required to fix this damage.[105]

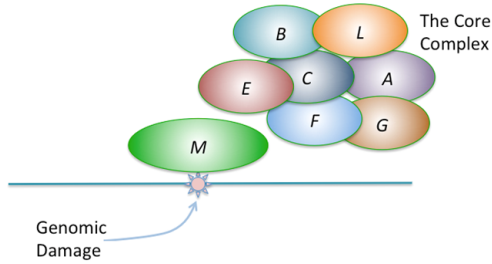


Figure 8.18: *M* Senses the Damage and Recruits the Core Complex.

This mission of repair is initiated when the *FANCM* protein (just *M* for short) senses that genome copying has been stalled by a cross-link.[106] It

promptly signals for help, recruiting several of its brethren *A*, *B*, *C*, *E*, *F*, *G* and *L* into the mission. These proteins band together to form a combined structure, called the *core complex*.

This core complex then primes its emissaries, the *I* and *D2* proteins, by marking them with a special mark. This mark is actually a small protein called *Ubiquitin*. This mark is delivered specifically by the *L* protein in the core complex.

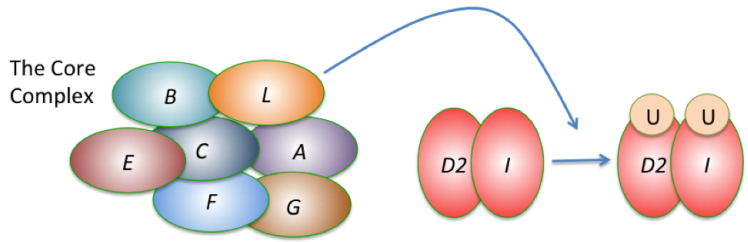


Figure 8.19: The *D2* and *I* Proteins, Primed by a Ubiquitin Mark put by the *L* Protein.

Once marked so, the *D2* and *I* proteins locate near sites which require attention in the genome. Here, they set about their task of recruiting various proteins required for the multi-step genomic surgery shown in the earlier pictures. These including proteins that snip the genome, proteins that fill-in missing pieces, and proteins that help perform the homology search and subsequent steps required for filling in the missing piece shown in Fig. 8.12. Several of these proteins are shown in the picture below (note the *BRCA1* and *BRCA2* proteins also appear in this list[107]). Detailed roles of these proteins are still being studied. We do know, for instance, that *O* helps recruit and manage proteins which attach to the overhangs in Fig. 8.14 and help in the homology search.[104]

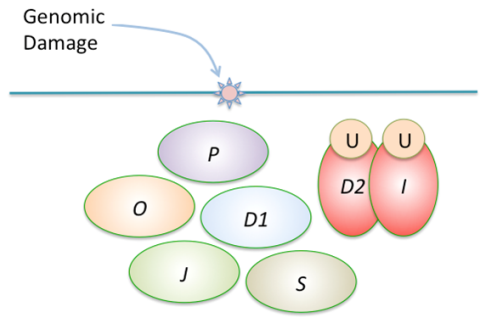


Figure 8.20: The Other *FANC* proteins, which Coodinate the Repair Process.

Once the surgery is completed and the damage repaired, the ubiquitin markers are removed off the *I* and *D2* proteins, indicating that they are no longer busy. They then wait for the core complex to set them busy again, in response to the next assault on the genome, and so on.

And so do these *FANC* proteins keep up their vigil, ensuring the genome is protected. In $\mathscr{X}$ though, one or more of these ever-vigil proteins had relaxed its guard, thus leaving the genome unattended and exposed to accumulation of uncorrected damage. Which gene's recipe did that protein come from?

The Hunt for the Variant

The cause of $\mathscr{X}$'s challenges was likely a problematic variant in one of the 17 known *FANC* genes. More *FANC* genes causing conditions like those of $\mathscr{X}$ are still being discovered; nevertheless, as a first measure, we focus on just these 17 genes.

As in previous chapters, it is economical to use a focused gene panel instead of sequencing all exons of all 20,000 or so known genes. Our focused gene panel comprises only around a 100 genes, problematic variants in which are well-known to increase the risk of cancer in an individual. The 17 *FANC* genes are included in this list. Our panel sequences all the

exons of all these genes. In addition it nibbles a bit into the introns, but not much.

Remember again, recipes are written in the exons of a gene. The intervening introns do not carry this recipe. So we skip the introns, for most part; the only exceptions are around 100 characters past the boundaries of each exon into the introns, where variants that affect jumps from exon to exon during recipe interpretation often lie. Our hope, of course, is that the cause of $\mathscr{X}$'s condition is a problematic variant in one of the exons or neighboring intronic stretches of one of the *FANC* genes.

Note that both of $\mathscr{X}$'s parents are healthy; they do not have any of the problems that $\mathscr{X}$ has. $\mathscr{X}$'s Fanconi's Anemia thus appears to be a *recessive* disease, of the type we saw in Chapter 2.1. Both copies of the relevant gene must then be made dysfunctional by problematic variants; just rendering a single gene copy dysfunctional does not result in disease. Each parent probably has a problematic variant in only one gene copy, leaving the other copy healthy and functional and keeping the disease at bay. An unfortunate confluence of both these problematic variants in $\mathscr{X}$ probably made both gene copies dysfunctional, paving the way for disease. So we look specifically for problematic variants present in both gene copies in $\mathscr{X}$. It could be the same variant present in both gene copies, or two distinct variants present in one copy each.

One *FANC* gene present an exception to this rule, though. The *B* gene is actually present on the X chromosome, of which $\mathscr{X}$ has only one copy (remember $\mathscr{X}$ was a boy, and males have only copy of chromosome X). As a special case for this gene, we look for variants present in its sole copy in $\mathscr{X}$.

Our search, as usual, hits roadblocks at the very outset. At first sight, there appear to be no variants of the types we are seeking in $\mathscr{X}$'s genome. The only variant of note, if any, is an innocuous-looking variant, a G to A character change in the *FANCL* gene. Innocuous-looking because it causes no change in the recipe. Remember Fig. 2.5; the recipe is obtained by grouping characters into triplets; each triplet is then converted to an amino acid. The G to A character change at hand causes the triplet to change from AAG to AAA. But no change in the amino acid! Both triplets are *synonymous*; they indicate the very same amino acid, namely *Lysine* (short

form *K*). Effectively, the recipe of the *FANCL* gene does not change at all. How can this variant then cause $\mathscr{X}$'s rather extreme condition?

Figure 8.21: A Synonymous Variant in Exon 13 of the *FANCL* Gene. Above the triplets are Amino Acid symbols.

Aside this variant, there appears to be little of note. None of the usual characters in the *FANC* genes seem to have been done away with in $\mathscr{X}$. No extra characters that shouldn't be there are present. No large chunks appear to have been disposed off or put in. No variant that might truncate any of the *FANC* gene recipes prematurely. No variants that cause a recipe change from one amino acid to another. No variants in the overhangs into the introns that our sequencing process has nibbled into that might mislead recipe interpretation jumps from exon to exon (remember, we saw such a variant in Chapter 6). In summary, nothing of significance that might help us explain the cause of $\mathscr{X}$'s singular condition.

Accordingly, we send a note to $\mathscr{X}$'s doctor stating our inability to identify the offending character in $\mathscr{X}$'s genome. Pop comes back the challenge. *How can that be? The chromosome breakage test of $\mathscr{X}$ was crystal clear, putting the diagnosis of Fanconi's anemia well beyond doubt. There must be a variant responsible for this in $\mathscr{X}$'s genome. Please look harder!*

The Edge of the Exonic Cliff

Hectic team meetings follow in the face of this renewed challenge. The only variant we have to work with is the G to A change in the *FANCL* gene. This variant causes no change to the gene recipe though, at least at first sight, so there is little evidence to its culpability. Then someone notices a curious fact.

As the picture above shows, this variant appears at the very end of exon 13, after which the genome slips off into the intron that follows. Perched precariously at this exonic edge, might this variant rather insidiously throw trans-intronic jumps off-balance during recipe interpretation?

Remember, recipes are described only in the exons. When interpreting the recipe written in these exons, intervening introns are skipped over. One cannot but marvel at this process which launches from the end of an exon and lands precisely at the beginning of the next exon, leaping across thousands if not tens of thousands of characters in the intronic ocean. How does the recipe interpretation process know where an intron begins and where it ends?

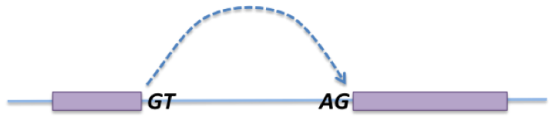


Figure 8.22: Recipe Interpretation Jumps from Exon End Mark to Exon Start Mark.

The above picture from Chapter 6 should refresh your memory. There are markers at the beginning and end of each intron which serve as guides in this process. At the beginning of most introns are the characters GT, and at the end, the character AG. These characters inform the recipe interpretation process when to take-off and where to land. Of course, these are not the only guides, for there are GT's and AG's galore even in the middle of most introns. As we saw in Chapter 6, the landing of the trans-intronic jump is guided by additional characters: the branch point and the C/T stretch (Fig. 6.13), both close of the end of the intron. But how about the take-off? What additional characters guide the take-off? Does the last character of an exon have a say in this?

Such questions are answered by studying hundreds of thousands of exons and introns of all genes together, and identifying which characters are common at various positions around the exon-intron boundary. The picture below depicts this information.

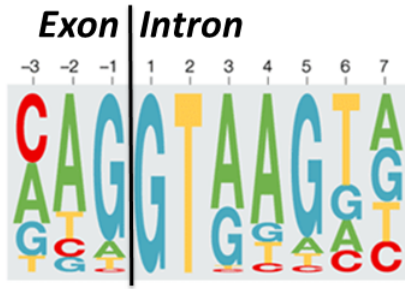


Figure 8.23: Character Commonness Around and Exon-Intron Boundary, from.[108]

As this picture shows, most introns have G and T, respectively, as their first two characters, as indicated by the full-sized G and T at positions 1 and 2. This overwhelming commonness suggests a strong role strong role for these two characters in guiding the trans-intronic jump. Of course, there are many GTs in the middle of an intron as well, so other characters in this picture must provide some guidance as well. The smaller sizes of these other characters in the picture indicates that they are less preserved across hundreds of thousands of exons and introns. So the exon-intron boundary does not appear as clearly demarcated to our eyes, as it is to the cell. Nevertheless, there are some patterns of note.

For instance, the third character in an intron (position 3) is most commonly an A; a G is only slightly less likely, while C or T are very very rare; hence the large A, slightly smaller G, and practically invisible C/T. On the other hand, the seventh character (position 7) is equally likely to be an A, G, T or C; hence all four characters are drawn at equal size. So neither of these positions show great domination by a single character. In fact, after positions 1 and 2, where G and T are completely dominant, respectively, positions -1 and 5 offer the greatest dominance by a single character, albeit far from complete.

We are most interested in position -1. This represents the last character in an exon and is most likely to be a G, with other characters occurring in relative minority. Most exons thus have a G as their last character.

This would suggest that a G here plays an important role in guiding the trans-intronic jump.[109] The G to A variant in the *FANCL* gene in $\mathscr{X}$'s genome is precisely at this position: the last character in exon 13. Could this alteration from G to A in $\mathscr{X}$ have misled the jump into not taking-off?

Exon Extended, Contracted, or Skipped?

Let's assume for a moment that this character change at the last exon position in $\mathscr{X}$'s genome does mislead the jump. What happens next during recipe interpretation? Does the process simply continue reading and interpreting into the intron all the way to the next exon rather than jumping directly across the intron to the next exon? That is indeed a possibility. But there is another, less obvious, possibility as well:[109] the jump to the next exon still happens, but launches from a different launch point instead of its usual launch pad at the end of the exon. A glance into the recipe interpretation process suggests why this is a possibility.

First, the exons and introns are all read and copied into a new molecule. From this new molecule, all the introns are removed. That leaves the exons only, ready and ripe for recipe interpretation. Removing the introns requires that cells know where introns begin and end. These boundaries are indicated by the character pattern depicted in Fig. 8.23. Several places in the genome might match this pattern, some to a greater extent than others. For instance, the character sequence CAGGTAAGTA matches the pattern in the picture quite strongly (each character is the most likely character at that position). But CAGGT*GG*GTA also matches the pattern reasonably, albeit not as strongly, with the most likely characters AA at positions 3 and 4 replaced by GG. So several places in the genome might stake their claim to mark an exon-intron boundary, some to a greater extent and some to a lesser extent. Among these competing choices, the cell chooses only a few to launch its trans-intronic jumps. How does it pick its choices?

The rules which determine the winners of this competition are still a subject of research. One such rule is, of course, that places which show greater closeness to the pattern in Fig. 8.23 have a better shot at winning.[110] There are other rules as well,[111] but none that is fully predictive of the winner. Nevertheless, a variant, such as the G to A variant in $\mathscr{X}$, might

suddenly alter the results of this competition.[112] One of the erstwhile winners might now present a poorer match to the pattern in Fig. 8.23, and an erstwhile loser might now emerge the winner. This new winner might well provide an alternative launch point. Sometimes, this launch point is a a little further down into the intron, extending the exon a bit and adding characters to the gene's recipe. Sometimes, this launch point appears earlier in the exon, effectively removing some characters from the gene's recipe. And sometimes, something even more surprising happens.

Sometimes, the launch point moves all the way to the end of the previous exon! In other words, from the end of the previous exon, a jump is launched that skips an intron, an exon, and another intron, landing at the beginning of the next exon; a super long jump at that. It is hard to predict when this happens, but we know the following. The various introns are not all removed at once, rather they are removed in some order, which is not necessarily left to right. Variants near the beginning of introns that are removed earlier in the process tend to result in this super long jump, in contrast to variants at the beginning of introns removed earlier in the process.[113]

Which of the above scenarios is the case for the G to A variant in the last character of exon 13 of the *FANCL* gene in $\mathscr{X}$? Does it result in more characters from the intron being added to the recipe? Or, some characters from the exon being deleted from the recipe? Or, the whole of the exon being deleted from the recipe? Or, does it make absolutely no change to the recipe at all, as was our first impression? How do we answer these questions?

Remember that the process of recipe interpretation reads all the exons and introns and copies them into a new molecule. From this new molecule, all the introns are removed. If we could read the sequence of characters at this stage, we will know exactly which exons remain in this molecule and which introns are removed. And that will give us our answer. Fortunately for us, someone had already performed this experiment that too, very recently![114] And the answer is in the picture below.

Figure 8.24: Exon 13 Completely Skipped Due to the G to A Variant Shown in Yellow at the End of Exon 13.

The entire exon 13 is missing from this molecule; the recipe interpretation process jumps straight from exon 12 to exon 14, removing both exon 13 and the two intermediate introns because of the G to A variant at the end of exon 13. What appears to be an innocuous variant at first glance actually has a major, insidious effect! Does it explain $\mathscr{X}$'s condition though?

Botched Genomic Surgery

Like a single bad apple bringing down the whole pack, a single character alteration in in $\mathscr{X}$ knocks all 72 characters of exon 13 out from the *FANCL* gene recipe. When grouped into triplets (remember Fig. 2.5), these characters comprise 24 amino acids (numbers 341-364). Since the G to A variant appears in both copies of the *FANCL* gene in $\mathscr{X}$, these 24 amino acids are completely missing from the *FANCL* protein in $\mathscr{X}$. How important were these missing amino acids?

Remember *FANCL*'s role in fixing cross-links in the genome (Fig. 8.19). *FANCL* is recruited when cross-links in the genome are detected. In turn, it commands its emissaries, *FANCD2* and *FANCI*, to go recruit other proteins that can perform the various genomic surgeries required to fix the cross-link (Fig. 8.20). This command is issued by the addition of a ubiquitin mark to *FANCD2* and *FANCI* by *FANCL*. By deliberately altering specific amino acids in *FANCL* and studying the consequences, the hand that adds these marks has been pinned down to lie in amino acids 307-359. Specifically, when amino acids 341 or 359 are altered, *FANCL* loses its marking ability.[115,116] And both these critical amino acids were missing from $\mathscr{X}$'s recipe!

With both amino acids 341 and 359 lost in both copies, *FANCL* completely loses its ability to mark *FANCD2* and *FANCI*, in all likelihood. The

command to go and fetch proteins that can perform corrective genomic surgery is then no longer issued in its usual manner. And what is the consequence?

Remember the missing piece in strand 2 which is generated when fixing the cross-link? It is shown again in Fig. 8.17 below. Instead of filling this piece by reading off strand 1, a different, more erroneous, surgical step is employed now. The two fragments are just joined together and the missing piece is simply lost - a dangerous situation in instances where the missing piece carries some critical parts of the recipe. Sometimes, this joining process goes even more horribly wrong, joining one fragment to another fragment resulting from another cross-link somewhere else in the genome.[117] Sometimes, even a fragment on another chromosome! The net result: fused chromosomes of the type shown in Fig. 8.2.

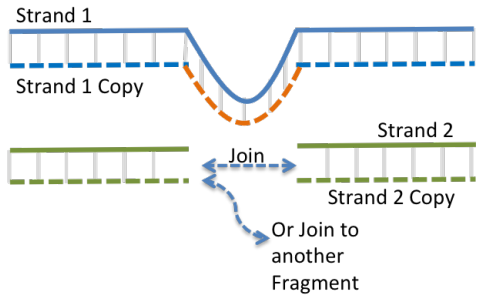


Figure 8.25: Faulty Repair in $\mathscr{X}$. The dotted lines with arrows indicate that the two genomic pieces are just joined together.

This inability to fill-in missing genomic pieces accurately causes genomic errors to accumulate in $\mathscr{X}$'s cells at a faster rate than in the average person. In response, cells detect these high levels of genomic damage and slow down in their cycle of division. This allows them time for repair before the genome is copied. However, this damage is irreparable, more often for $\mathscr{X}$ than for the average person. To protect from the ill-effects of such damaged genomes, nature provides a culling mechanism: cells with irreparable damage simply choose to eliminate themselves by programmed

suicide rather than continuing to multiply with defective genomes.[118]

The result: cells in $\mathscr{X}$ have much higher rate of death due to programmed suicide. The consequence is an under-supply of healthy cells. This is most apparent in the bone marrow, the factory located in the interior of our bones where so called *stem cells* divide continually to generate new blood cells, replacing old blood cells which are turned over once in several days. In $\mathscr{X}$, many stems cells choose to commit suicide rather than divide because of genomic damage. This paucity of cells culminates in bone marrow failure, the inability to generated blood cells of all types in sufficient quantities. All on account of a single character in the genome, which we dismissed at first glance as triflingly innocuous!

Wrapping Up

An individual begins life as a single cell, which then divides into trillions of cells. Some of these cells continue to divide throughout our lives, replacing other cells that die. Over these 10,000 trillion or so cell divisions we experience in our lifetime, the genome doesn't stay unchanged. Various routine exposures of life (food, sunlight etc) damage the genome on an ongoing basis. An army of repair genes ensures this damage is repaired as it occurs using complex genomic surgery. This repair process is fairly accurate. Nevertheless, over 10,000 trillion or so cell divisions, errors are bound to creep in. These cause the genome in one cell to become different from the genome in another.

Once the genome in a cell changes, all daughter cells arising from the division of this cell inherit these changes. Over time, our body becomes a genomic mosaic, with different cells carrying slightly different versions of the genome. Good repair machinery in our cells slows down the rate at which this variety builds up. Regardless, as the juggernaut of age rolls on, the mosaic gets more and more varied. Eventually, some cell or the other accumulates sufficiently many changes so its genome becomes unstable. Unstable means that when this cell divides, many more changes than usual occur in the genome; and in turn, when these daughter cells divide, even more changes occur; acceleration that is reminiscent of a snowball that picks up more and more snow as it rolls down the hill.

Our cells do detect this instability and induce programmed suicide so bad apples are removed. But what if the very genes which detect instability and cause programmed suicide are mutated by snowballing genomic change? As you can guess, these bad apples then do not commit programmed suicide. They continue to grow and divide. Nature poses further challenges to these cells. Typically, a cell is limited in the number of times it can divide; after about 50 or generations, a cell cannot divide anymore. Again, genomic instability can lead to accumulation of variations which enable escape past this hurdle, and several other such hurdles. So a few cells which escape all these hurdles simply continue to divide.

Among these, those which can divide faster generate more daughter cells and therefore hog a greater share of the available resources (oxygen, nutrients), enabling them to grow and divide faster than other cells. The snowball effect again; the faster a cell divides, the more wherewithal it's descendants have to divide even faster. Eventually, this results in a collection of cells which break several rules of organized functioning, leading to cancer. The odds with age: as in the cancer hill picture in Fig. 8.3.

Of course, if the ability to repair is compromised, then this picture is accelerated and cancer sets in earlier. Indeed, women with problematic variants in just one copy of the *BRCA1* or *BRCA2* genes (also called *FANCS* and *FANCD1* as in Fig. 8.20) have a 70-80% and a 50-60% chance, respectively, of getting breast cancer in their lifetime.[119] However, such cancer usually sets in only after a few decades of life. Presumably, the one working gene copy holds the fort for a while, until some erroneous genomic surgery or related event leads to loss of this good copy (which happens often, but not always[119]), at which point the genomic snowball starts accelerating.

Unfortunately, $\mathscr{X}$ had both copies of a key repair gene disabled by a subtle variant whose impact on the gene's recipe was most insidious. Each of his parents carried the variant in one copy, so they were completely unaffected and would never have realized what was in store for their son. Genomic instability in $\mathscr{X}$'s bone marrow cells had set-in in his infancy itself as a consequence. The protective measure of programmed suicide then resulted in an acute shortage of bone marrow cells. $\mathscr{X}$

would need a bone marrow transplant to cure this condition. If he were to successfully negotiate this hurdle, then the next challenge will be an increased likelihood of cancer as he grows older. Hopefully, time is on $\mathscr{X}$'s side, and tremendous advancements in detecting and treating cancer will be made in the interim, so the world is ready to manage $\mathscr{X}$'s cancer when it sets in. We will get a glimpse of where the world is headed on this front, in the next chapter. And how the very cross-links that bother $\mathscr{X}$ are also important weapons in our arsenal in the fight against cancer.

Chapter 9

Moves and Countermoves

Persistent pain in the abdomen prompted $\mathscr{X}$ to see a doctor when he was 37. An ultrasound scan identified excessive fluid in the abdomen as the cause of this pain. Why this fluid had accumulated was explained only when a CT scan was performed.

A CT scan takes a series of X-ray images from different angles to generate various cross-sections of internal body parts. In $\mathscr{X}$, this scan showed several nodules growing abnormally around the lungs. There was also an abnormal growth in the abdomen.

To identify this growth precisely, doctors performed a biopsy: a procedure to obtain a tiny bit of tissue from one of these clumps of growth. They made a small cut in $\mathscr{X}$'s chest-wall and inserted a scope that enabled them to look into his chest. They then guided a tool to the growth to pick a little piece of the tissue from around the lungs. This tissue was stained for observation under a microscope. Abnormal cells often appear distinctively different from normal cells, at least to a highly trained eye. And the shape and structure of these cells can indicate where they came from.

Looking at the cells from $\mathscr{X}$'s biopsy under a microscope, doctors concluded that the abnormal growth comprised cells from his *mesothelieum*, a membrane that covers and protects the lungs and various organs in the abdomen. $\mathscr{X}$ was accordingly diagnosed with *mesothelioma*, a *cancer* of cells from the mesothelium. Here is a rough impression of how the cancerous growths might look.

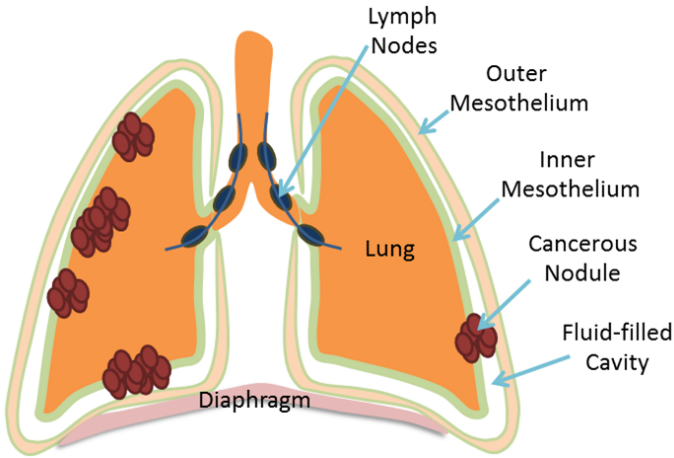


Figure 9.1: Abnormal Growth in the Lung Mesothelium.

You might wonder how our internal organs are cushioned from impact as we jump up and down, dive, fall, bang into each other, etc. The mesothelium provides this cushion. The lung mesothelium is composed of two thin, protective linings, as shown in the picture above. The gap between these linings is filled with a few teaspoons of lubricating fluid. This fluid allows the lungs to glide smoothly when they expand and contract.

Some cells in $\mathscr{X}$'s mesothelium had lost their usual discipline and were dividing uncontrollably. These cells were also growing haphazardly, no longer respecting the organized architecture of their normal neighbors. Some of these cells had in fact broken off from their neighbors, traveled through the the lubricating fluid, and planted themselves at other locations, continuing to grow and divide in their new homes. As a result, $\mathscr{X}$ showed several nodules around the lung. Some nodules were even invading into the lung.

A similar pair of linings, called the abdominal mesothelium, cushions the abdominal organs, as shown in the picture below. The gap between the inner and outer linings is larger here. But, as before, it is filled with cushioning fluid. Presumably, some the cancerous cells from the lung

mesothelium had broken off and lodged themselves in the abdomen and were growing here as well. These cancerous cells were secreting fluid into the fluid-filled cavity, so excessive fluid had accumulated in the gap between the two mesothelium linings. This was the cause of $\mathscr{X}$'s abdominal pain.

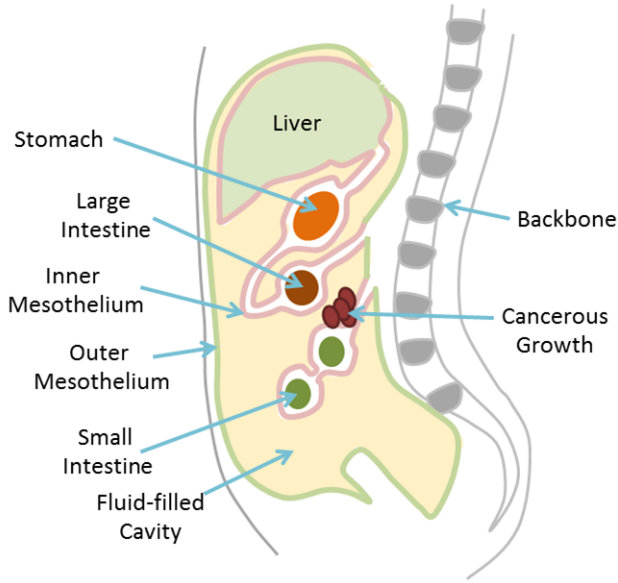


Figure 9.2: Abnormal Growth in the Abdominal Mesothelium: A Sideways Profile.

In due course, excess fluid could build up in the lung as well, offering resistance to lung expansion and impairing breathing. Even worse, if left to itself, the cancer would continue to divide and expand, compromising organ function, and possibly proving fatal. Indeed, statistics indicate that only 40% of the mesothelioma patients survive 1 year past diagnosis. And only 10% survive 5 years past diagnosis.

$\mathscr{X}$ was relatively young at 38. There was no family history of the disease in $\mathscr{X}$'s family, or anything else to suggest an inherited cause.

Then what led cells in $\mathscr{X}$'s mesothelium to lose their usual discipline and start dividing uncontrollably? And what tools could modern science and medicine offer to $\mathscr{X}$ to fight this cancer.

Asbestos?

The word *asbestos* has its origins in a Greek word which means "inextinguishable". For good reason - sheets and clothes made from this naturally-occurring mineral have long been known to be heat and fire resistant. Roman legend mentions an emperor who would throw his asbestos-derived tablecloth into the fire at the end of a meal and marvel as it emerged unscathed! Over the years, this "miracle material" became more and more ubiquitous. Asbestos materials were used for heat and fire insulation in all types of construction material: pipes, cement, plaster, roofing sheets and tiles. Soon enough, our buildings contained substantial amounts of asbestos inside.

As early as 2000 years ago, Pliny the Elder, the well-known Roman philosopher, noted that slaves working in asbestos mines suffered from sickness of the lungs and died at a relatively early age. This observation was reaffirmed again in the latter half of the 20th century. The aftermath of the second world war saw massive emphasis on rebuilding and construction, employing workers in large numbers. These workers experienced prolonged exposure to asbestos. Over the next few decades, several of them developed mesothelioma.

We now know that asbestos used in construction material leaks tiny microscopic fibers which we can inhale.[120] These fibers eventually make their way into the mesothelial linings of the lung and lodge themselves there. Here they cause production of reactive molecules containing oxygen (called *ROS*, or reactive oxygen species), which react with genomic characters and cause various forms of damage to the genome, as we saw in Chapter 8. The genomic repair mechanisms in our cells succeed in repairing several of these errors. Not all though; some errors do leak through nevertheless. As cells with these damaged characters divide, a daughter cell inherits a genome that is slightly different from the one in its parent. Accumulation of these changes over a few decades results, with a not

too small likelihood, in a cell carrying a potent combination of character changes in its genome. This combination enables the cell to escape the natural forces of programmed suicide that it is normally subjected to. As a consequence, the cell starts dividing uncontrollably. One study estimated that 1 out of 17 British men born in the 1940s and employed in carpentry for more than 10 years before the age of 30 would develop mesothelioma.

The need to control the use of asbestos became increasingly obvious in the latter half of the 20th century. In a series of legislations in the 1970s and 1980s, the Environmental Protection Agency in the United States progressively banned several uses of asbestos. A final ruling was made in 1989 to ban almost all uses of asbestos. However, this ruling was appealed by the asbestos industry and overturned in a landmark judgement in 1991. So several but not all uses of asbestos continue to be banned in the US. The European Union completely banned all uses of asbestos in 1999. In contrast, the use of asbestos in India and China not only continues unhindered, but is growing rapidly, exposing large numbers of people to its deadly fibers.[121] It is hard to say, but given the ubiquity of asbestos in India, possibly, the genesis of $\mathscr{X}$'s mesothelioma lay in some early over-exposure to asbestos fibers.

Virus?

Can cancer actually be transmitted from one person to another? Like common cold, which spreads through the air? Or less easily, like AIDS, which requires contact of body fluids? Fortunately, the answer seems to be no: most cancers do not appear to be directly infectious. However, observations made as early as 1842 suggested some communicability for certain cancers: for instance, nuns in Verona seemed to develop cervical cancer much more rarely than married women.[122] Experiments performed in the early part of the 21st century then slowly started unraveling the infectious component of cancer.

In 1911, a farmer showed a hen with a large tumor (a *sarcoma*) in her breast to Peyton Rous, a pathologist at the Rockefeller Institute in New York. Rous's curiosity led him to explore whether the tumor could be transplanted to other chicken. To his surprise, it indeed could, though only

to closely related chicken! To identify what had *carried* the cancer from one chicken to another, Rous filtered out cells and bacteria and again tried the transplant. Rather surprisingly, this transplant too caused tumors in the recipient. Rous hypothesized that a virus might be carrying the cancer from one chicken to another.[123]

Rous was the not to first to reach this conclusion though: in 1908, Danish veterinarians Wilhelm Ellerman and Olaf Bang showed that leukemia, a type of blood cancer, could be induced in healthy chicken, by taking blood from a chicken with leukemia, filtering out cells and bacteria in a similar way, and then injecting the residue into healthy chicken. Of course, microscopes of the day were not powerful enough to precisely identify these viruses were (electron microscopes, which made viruses visible, were invented a few decades later).

We now know the virus in Rous' experiment; it is aptly named *RSV* or the *Rous Sarcoma Virus*. Most viruses destroy the cells that they infect. Not all though. Some, like *RSV*, alter the cell so its starts dividing more rapidly. Daughter cells resulting from this division too divide rapidly even in the absence of the virus. So the transformation from healthy to cancerous is committed to memory in some way, most likely in the genome. Indeed, we know now that the virus succeeds in incorporating its own genes into the genomes of the cells in its host!

A specific viral gene, called *v-src*, has now been identified as the gene responsible for inducing cancer.[124] A chicken cell can no longer distinguish between genes of its own and genes which the virus introduced into its genome; it therefore interprets recipes from both types of genes. The protein created by recipe interpretation from the v-src oncogene triggers uncontrolled cell division. This is no accident, rather a cold deliberate act: for, the virus has no other use for this gene save inducing its host chick cell to divide so its hunting ground can grow!

Interestingly, chick genomes have their very own version of this gene, called *c-src*[125] (v- for viral, c-actually for cell, but you could as well think c- for chicken here). The c-src and v-src genes carry similar recipes. Then why does the c-src gene not result in cancerous transformation like the v-src does? Because of some notable character differences in the two recipes. The c-src recipe provides for a switch which helps the cell modulate protein

activity, thus switching on the protein only when needed.[126] In contrast, the v-src recipe carries no such switch, and so it creates a protein that is always active and cannot be controlled!

The story above on the Rous Sarcoma Virus involves chicken. What happens to us humans? Fortunately, the Rous Sarcoma Virus does not affect humans. The first virus to cause cancerous transformation in humans was discovered in 1964 and is now named Epstein-Barr Virus after 2 of its 3 discoverers: Anthony Epstein and Yvonne Barr (the third was Bert Achong). Just a handful of other such viruses have since been discovered. And 10-15% of all human cancers are now understood to have a viral cause.[127]

Of these, the greatest contribution comes from the Human Papilloma Virus, which accounts for an overwhelming fraction of the cervical cancer cases in the world. Remember nuns in Verona were observed to have a much lower incidence of cervical cancer than married women; this is because HPV virus spreads primarily through sexual contact. The virus goes beyond cervical cancer though; a fraction of cancers of the head and neck are also caused by HPV. Genes provided by this virus waylay two important human genes. First, the *RB* gene which we saw in Chapter 7 as the cause cancer in the eye of infants. And second, a gene called *TP53*, arguably the most vocal gene in cancer; we will see why, shortly. Both genes are key players in keeping cell division in check.

In the early 1960s, a series of studies brought a new virus to the fore in a most unexpected setting.[128] Vaccinations for polio were then made using Rhesus monkey kidney cells. Some of these vaccines were found to harbor a new virus, called SV40, whose natural host is the Rhesus monkey. This virus usually kills the cells that it infects in the Rhesus. Surprisingly, hamster (a type of rodent) cells did not die when inoculated; rather they started developed tumors! Millions of people were inadvertently exposed to SV40 via the polio vaccine between 1955 and 1963, in the US and in Europe. This prompted fear of increased cancer rates among the vaccine recipients. Fortunately, several studies have now shown that cancer rates in this population are not significantly different from those in the general population.

Coming back to $\mathscr{X}$'s mesothelioma. In 1993, suspicion of SV40 as

a cause for mesothelioma was spurred by the finding that Syrian hamsters (a species of hamster originating in Syria) when infected with SV40 developed mesothelioma.[129] Adding to this suspicion, studies on human patients found traces of the virus in a significant fraction of mesothelioma patients.[130] Not all studies have replicated this phenomenon though. And mere presence of SV40 isn't proof that it is indeed the cause of cancerous transformation; that proof remains elusive. Human cells in a laboratory dish or test tube do turn cancerous though under the influence of SV40.[131] So, again, hard to determine, but it is possible that the SV40 virus played a role in the genesis of $\mathscr{X}$'s mesothelioma.

Or Plain Random Chance?

Other than environmental exposure to say asbestos, or a virus, where else could the cause of $\mathscr{X}$'s mesothelioma lie? A very interesting observation, crystallized quite recently, provides some insights. This observation was driven by the following question. Cells in the inner surface of the small intestine are exposed as much to chemicals in our environment via food (pesticides, for example) as are cells in the inner surface of the large intestine. Yet, tumors in the large intestine are much more likely than tumors in the small intestine. Why so?

We begin life as a single cell. This cell divides repeatedly to generate trillions of cells. Along the way, these cells differentiate, or specialize, into various tissues: heart, brain, nerve, blood, lung, kidney etc. They do this by turning collections of genes on and off. At some point, the number of cells reaches a steady level. Beyond this point, further cell divisions are needed primarily to compensate for dying cells. Many fully specialized cells cannot divide anymore. So most tissues have a pool of unspecialized cells, or *stem cells* as they are called, which divide further to compensate for dying cells.

The point now is that the number of stem cells varies widely between tissues. And so does the rate at which they divide. The large intestine has more stem cells than the small intestine. And cells in the large intestine divide faster. Over a lifetime, the large intestine sees roughly 4 times as many cell divisions as the small intestine.[132] Could this explain the higher

incidence of cancer in the large intestine when compared to the small intestine? The following picture answers this question, using data from 31 different tissues.[132]

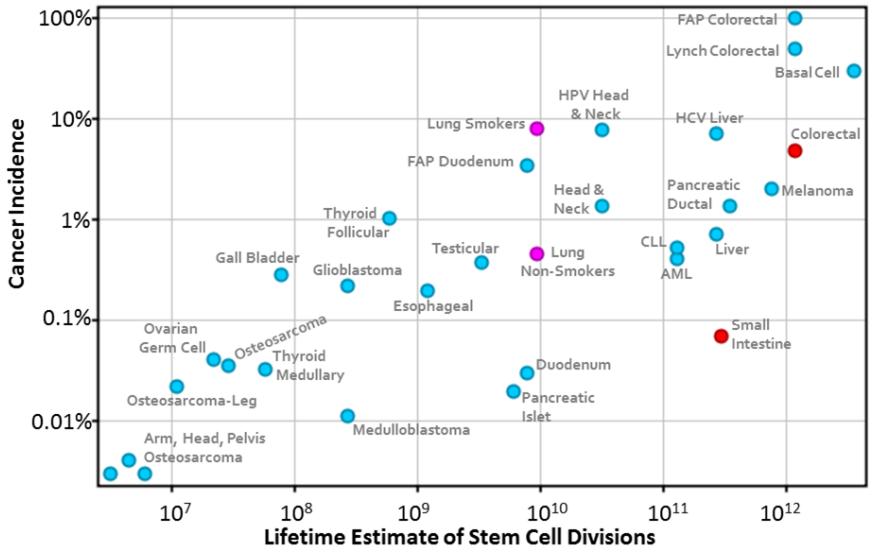


Figure 9.3: Cancer Risk in Different Tissues of the Body vs. The Number of Stem Cell Divisions in that Tissue, from.[132]

As the spread of points in this picture indicates, tissues with more cells divisions appear to have a higher risk of turning cancerous. For example, the lung (marked in pink) sees about 10 billion stem cell divisions, while the large intestine (marked as colorectal, in red) sees about 1000 billion. The lifetime risk of cancer arising from mucous secreting cells of the lung is roughly 0.45%. The corresponding figure for the large intestine is roughly 4.8%, about 10-fold higher.[132]

The correlation between stem cell division rates and cancer incidence is not perfect though, the points are spread out a bit rather than being on a tight straight line. Careful math shows that the stem cell division rate can *explain* not all, but roughly 2/3rds, of the variation in cancer incidence

between tissues. Interpreted in a way, this implies that a substantial part of our risk for cancer might just arise from the way we are made. Our cells have to divide billions to trillions of times over our lifespans. Each division has to copy the entire genome. And the copying process is not perfect, it does make mistakes, in some random ways. Over many many cell divisions, these random errors could accumulate, leading to cancer.

No doubt environmental influences (asbestos, viruses etc) add to the risk. For instance, look at the two pink dots in the above picture labeled Lung (Smokers) and Lung (Non-Smokers) (both labeled in pink). Both refer to the same tissue, so the number of stem cell divisions is the same for both the dots. Yet, the risk of lung cancer for a smoker is 18 times higher. Interestingly, the number is 23 times for males and 13 times for females, averaging to 18 on the whole.[132] Smoking leads to widespread genomic damage, so errors accumulate at a much faster rate with each cycle of cell division. Regardless, it appears as if there is a substantial baseline risk of cancer based purely on random errors that accumulate in the genomes of our cells over the several trillions of cell divisions in our body over our lifetimes. Whether this will happen to any particular individual is hard to predict. Maybe the roll of the dice in this case was simply not in $\mathscr{X}$'s favor.

Whichever way, whether it was asbestos or a virus or just plain random chance or a combination of all of these, the cancer had made it first move by announcing itself. It was now for $\mathscr{X}$ and his caregivers to make their first moves.

Preparatory Moves

The National Comprehensive Cancer Network (NCCN), an alliance of 26 of the world's leading cancer centers, provides guidelines on treating and managing cancer. Each type of cancer has its own dedicated set of guidelines. These guidelines are widely used by treating clinicians. Of course, there are guidelines specifically for lung mesothelioma as well, of the type that $\mathscr{X}$ had.[133] And $\mathscr{X}$'s doctors made their first few moves following these. The first step was to determine how localized or spread out $\mathscr{X}$'s cancer was.

$\mathscr{X}$'s cancer arose from the cells in the mesothelium cushioning his lungs. The initial tumor would have been localized at one spot in the mesothelium. Some cells had since broken away and spawned new tumors at other locations, so there were many tumor nodules around his lungs. Worse, some tumor cells had also moved to his *lymph nodes*.

The lymph system is a network of vessels within which circulates a colorless fluid called *lymph*. The junctions in this network are called lymph nodes (Fig. 9). As blood circulates around the body, fluid leaks out from the blood vessels into the spaces between the cells, providing nutrients to these cells. This fluid then collects waste, microbes, and damaged cells, and drains back into the blood vessels via the lymph network. The occasional vagrant tumor cell may break away from its parent tumor and also drain into the lymph system. Inside the lymph network, it could travel to the lymph nodes, and potentially to distant organs. Of course, a similar phenomenon could happen through the blood circulation instead of the lymph circulation.

Indeed, tumor cells were present in the lymph nodes around $\mathscr{X}$'s lungs. And some tumor cells had traveled all the way to the mesothelium of the abdomen. $\mathscr{X}$'s cancer has spread quite a bit before it had announced itself.

If a tumor is localized, then removing it surgically is as close to a cure as one can get. Even then, the expected survival of a mesothelioma patient after surgery is only a few years.[134] For, removing all visible traces of the tumor surgically is not the same as removing *all* traces of the tumor; invariably, a few cancer cells are left behind. These can grow back into full-fledged tumors in due course. So doctors use radiation focussed sharply on areas around the tumor site to destroy any such microscopic traces. Radiation causes large scale genomic damage in these cells, forcing the cells to die.

$\mathscr{X}$ was young and healthy enough to tolerate surgery and radiation. Removal of the lung mesothelium while sparing the lungs would be one form of surgery. Removal of an entire lung would be another. Unfortunately for $\mathscr{X}$, his cancer had spread widely, ruling out the option of surgery and radiation. Other approaches were needed to stop his cancer cells from continuing to divide.

The Opening Move

Could a drug be used to kill $\mathscr{X}$'s tumor cells? Or at least stop these cells from dividing? A drug that the bloodstream could carry to every remote cell in the body and that would preferentially kill cancer cells while leaving normal cells undisturbed? Answering this question requires answering a more fundamental question: what makes a cancer cell different from a normal cell?

Cancer cells divide much faster than normal cells do. So cancer cells can be preferentially killed by a drug that interferes with the cell division process. For instance, an entire copy of the genome is made during cell division. This process requires the core building blocks of the genome, the characters A, C, G and T, in sufficient quantities (Fig. 8.4). These 4 molecules are manufactured inside our cells. Some drugs which interfere with this manufacturing process, making these building blocks scarce and slowing down cell division. Indeed, such a drug, called Pemetrexed, is recommended by the NCCN as part of its guidelines for treating mesothelioma.[133]

There are other ways to interfere with the cell division process. One of these goes back to the boy with Fanconi's Anemia in the previous chapter, whose ability to resolve cross-links was highly compromised. Remember cross-links in the genome from Fig. 8.10? Those unauthorized connections between genomic characters that effectively blocked cell division and made genomic repair genes work hard to overcome the block.

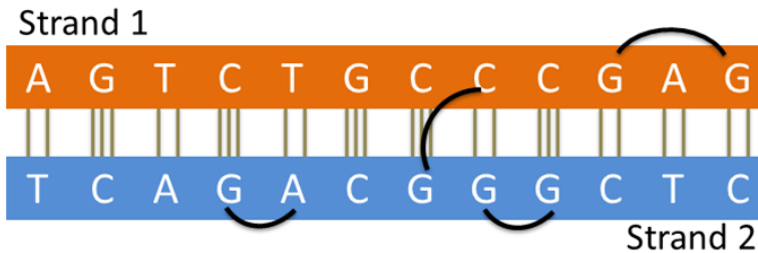


Figure 9.4: Cross-Links Caused by Cisplatin.

Some drugs, like Cisplatin and Carboplatin, can indeed introduce cross-links in large numbers in the genome. As their name indicates, they are compounds of platinum. Remember, platinum is among the most expensive precious metals on earth, ahead of gold. A combination of the drugs Pemetrexed and Cisplatin is used as the first line of treatment for mesothelioma.[135] This was the combination doctors prescribed for $\mathscr{X}$ as their first move.

Both drugs seek to stall cell division. Cancer cells divide fast and have the least time available for repairing these cross-links, and are therefore affected the most. But other normal cells in our body divide too, albeit not as fast. Unfortunately, these drugs are blunt instruments; unaware of the differences between normal and cancerous cells. In their ignorance, they interfere with normal cell division as well to an extent. And this causes painful side-effects. Nausea and vomiting result from interference in the gastro-intestinal lining. Fewer blood cells are manufactured in the bone marrow (as was the case with the boy with Fanconi's Anemia in the previous chapter). Unfortunately, sharper attacks, which can discriminate between cancer cells and normal cells, are few and far between. And none are recommended by the National Comprehensive Cancer Network as standard for mesothelioma. So the brunt of these side-effects has to be borne, in the interest of suppressing the cancer, a far greater threat.

Note a point of irony here. The boy with Fanconi's Anemia had much reduced ability to resolve cross-links. Not cross-links artificially induced by drugs, but simply those that occurred in the normal course of life. And that put him at very high risk for cancer over 3 decades or so. And here was $\mathscr{X}$, a cancer patient, on whom current standard treatment forced a barrage of cross-links at a much higher than normal rate, in an attempt to cure the cancer!

Of course, the goal of inducing cross-links in large numbers was to challenge the genomic repair machinery beyond its capability. Cells would then detect the genome is going unrepaired and invoke programmed suicide and die. So, in the short term, introducing cross-links would kill dividing cells. In the long term, the risk of cancer would increase. For $\mathscr{X}$, the battle was clearly focused on the short-term, for the average life expectancy of a mesothelioma patient is only a few years. The mission at hand was to kill

cancer cells even at the expense of an increased long-term risk for cancer.

Indeed, several of $\mathscr{X}$'s cancer cells died under the onslaught of Pemetrexed and Cisplatin. A CT scan showed that some tumor nodules had become smaller and others had vanished! And $\mathscr{X}$'s abdomen pain reduced substantially. The battle had been won!

But the war wasn't over yet. Some cancer cells survived and continued to experiment with various genomic combinations, via genomic surgeries performed during the genome repair process. The occasional cell then evolved a notorious combination, possibly just by chance. It managed to shut its doors to Cisplatin, which no longer could make its way into the cell and create cross-links as easily as it normally could.[136] It also managed to suppress the innate urge for programmed death when things go wrong in the genome. This cell become resistant to Cisplatin, and continued to divide recklessly. A year and half later, as more such cells accumulated, the cancer was back.

More Blunt Attacks

$\mathscr{X}$'s cancer cells were now resistant to the Pemetrexed/Cisplatin combination. So his doctors needed to switch strategy. Their next move, in accordance with the guidelines provided by the NCCN,[133] was to replace Cisplatin by its cousin, a drug called Carboplatin.[137]

Carboplatin, like Cisplatin, works by inducing cross-links in the genome, albeit less intensively and with far fewer side-effects.[138] The attack by Pemetrexed and Carboplatin pushed $\mathscr{X}$'s cancer to the back-foot once again. The tumors started to shrink and the symptoms abated.

The cancer again lay low for a while, figuring out its next move. But move it did. After about 9 months, it was back. The tumors were growing once again. And there was further growth in the abdomen.

It was now close to 3 years since $\mathscr{X}$ had been diagnosed. Already longer than the typical survival period from diagnosis for a mesothelioma patient. Possibly, $\mathscr{X}$'s relatively young age of 38 at diagnosis had helped. $\mathscr{X}$'s doctors now needed to make their next move to keep $\mathscr{X}$ alive and relatively healthy. They added a drug called Gemcitabine to the Pemetrexed/Carboplatin combination, again in accordance with treatment

guidelines provided by the NCCN.[139]

Gemcitabine has its own clever way of stalling cell division.[140] Remember, cell division first requires an entire copy of the genome to be made. This is done by assembling the copy one character at a time. Of course, this requires an ample supply of genomic characters (remember Fig. 8.4, each character is actually a molecule). Gemcitabine depletes this pool. It then posits as a Trojan, offering itself (more precisely, a modification of itself) as a replacement character. The genome copying process falls for this Trojan trap: it uses this replacement character in the copy being assembled. Only after it has added this and then one more character to the copy does it realize that this replacement character was a Trojan that would allow no further addition of characters. The genome copying process stalls, again prompting the cell to undergo programmed suicide.

Fortunately, the cancer in the lung responded to this new regimen. Unfortunately, the cancer in the abdomen remained stubborn, and continued to grow. Presumably, these cells had worked out a genomic combination that defused all the drugs tried so far. Close to 4 years from diagnosis, $\mathscr{X}$'s doctors were running out of options. The cancer was threatening to move towards its endgame.

The only other drug which finds mention in the treatment guidelines provided by the NCCN, the official standard of care, was a drug called Vinorelbine.[133] This drug too blocks cell division in its own distinctive way. Unlike the earlier drugs mentioned above, Vinorelbine does not obstruct genome copying itself. Instead, it obstructs the creation and function of a scaffold structure which supports the copying of the genome.

So $\mathscr{X}$'s doctors tried combinations of Pemetrexed, Carboplatin and Vinorelbine. Some caused serious side-effects and had to be pulled back. None seemed to unsettle the cancer, which, by now, had honed its defenses, and continued its unrelenting march forward.

The NCCN standard of care guidelines had nothing more to offer $\mathscr{X}$. All the blunt attacks, or *cytotoxic chemotherapies* as they are called, no longer challenged the cancer. $\mathscr{X}$'s doctors now needed to think out of the box to make their next move. Could understanding the cancer genome help?

The Cancer Genome

The ancestors of $\mathscr{X}$'s cancer cells were once normal. Then a genomic event arose in an ancestor: perhaps a new character, perhaps a missing one, perhaps not one but several characters. Perhaps, this event was introduced in the process of correcting damage to the genome. This cell then started dividing faster than usual. Some of its descendants succumbed under the onslaught of the various drugs that $\mathscr{X}$'s doctors threw at them. However, some managed to dodge these drugs by evolving key genomic changes. And like a snowball, these descendant cells evolved further genomic changes as time progressed, accompanied by more and more sinister characteristics.

These changes, or *somatic variants* as they are called, were not inherited by $\mathscr{X}$ from his parents. Rather, these were acquired after birth. Which variants were these? And how could these inform the next move that $\mathscr{X}$'s doctors needed to make?

To identify these somatic variants, we first need a few cancer cells from $\mathscr{X}$. Not as easy as drawing a bit of blood or saliva from a person, but nevertheless doable by a routine biopsy procedure. Usually, this procedure yields a mix of cancer cells and normal cells. Cancer cells then have to be carefully separated, often by a trained eye which can tease apart a cluster of cancer cells in an ocean of normal cells. This separation is far from perfect, but good enough to proceed.

Next, we sequence the genomes in these cancer cells and the normal cells separately, much as in previous chapters. We then identify somatic variants simply by spotting differences between the cancer genome and the normal genome. This exercise has been done for several thousands of tumors today. And an interesting picture of the cancer genome has emerged.

Of the 20,000 or so genes in the human genome, a typical cancer genome has somatic variants which cause recipe changes in relatively few genes: ranging from few tens to a couple of hundred.[141] As expected, lung cancer cells from smokers occupy the upper end of this spectrum. Likewise for skin cancer cells.[141] Clearly, exposure to stimuli that damage the genome (cigarette smoke in one case, and ultraviolet rays in the other) increases the rate of appearance of somatic variants quite substantially. In

any case, modified recipes in few tens to a couple of hundred genes appear to convert a normal, well-disciplined cell to a rogue, cancer cell.

These recipe changes include all types of changes that we have seen in previous chapters. For instance, character changes where the usual amino acid in the gene recipe is replaced by another, possibly more problematic one (Fig. 2.5). And character changes which cause a gene recipe to stop dead in its tracks prematurely (Fig. 2.10). And character changes where the grouping of the recipe into triplets itself is jittered due to either extra characters or missing characters, whose count is not a multiple of three (Fig. 3.9). And character changes where many copies of a gene are created, possibly causing it to function in overdrive. And character changes where entire copies of a gene are deleted!

Of course, the set of genes with recipe changes varies from one tumor specimen to another. In that sense, one person's cancer is different from that of another. Cancer is therefore, not one disease, but many diseases!

Fortunately, not all genes carrying recipe-altering somatic variations in any given tumor specimen are instrumental in converting a normal cell to a cancer cell. The genes which are instrumental are called *driver genes*. By cataloging genes in which recipe-altering somatic variations are found repeatedly over many many tumor specimens, we now have a list of roughly a 140 driver genes.[141] This list is not expected to grow dramatically, for new driver genes are appearing rarely as more and more tumors are being studied.

So, for the most part, these 140 or so genes carry the burden of converting a normal cell to a cancer cell. And any given tumor specimen has only a handful of recipe-altering somatic variants in these driver genes. Sometimes, just 1 or 2, sometimes as many as 8, so called *driver variants*. One person's cancer is still different from another's, but the variety appears more contained now. Even among the driver genes, a few genes seem to dominate the landscape as shown below in statistics gathered over 300 or so tumors that we have sequenced at Strand.

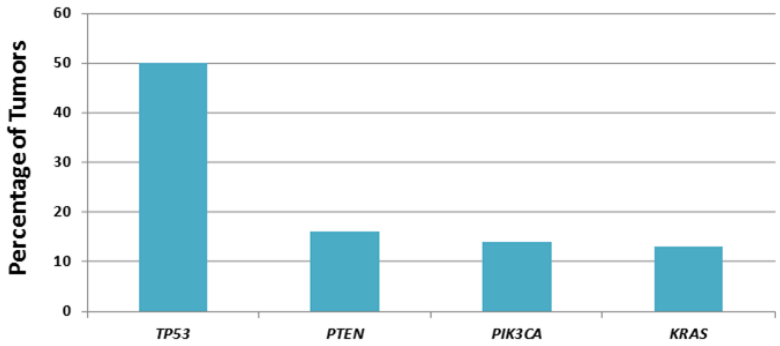


Figure 9.5: Gene Most Commonly Mutated in Cancers.

Note the gene that stands out most prominently in the above picture. 50% of the tumors carry a recipe-altering somatic variant in the *TP53* gene. 50%! Remember this gene? We saw it briefly earlier: the Human Papilloma Virus, the main cause of cervical cancer, inactivates *TP53*, thus pushing the infected cells towards cancer. Something similar happens due to somatic variants in about 50% of the tumors. There are an estimated million new cases of cancer in India every year. A like number in the US. And half of these are likely to have recipe-altering somatic mutations in *TP53*! This gene is clearly a key sentinel on the path from normalcy to cancer. Hence the name *TP* for *tumor protein* (the number 53 denotes its weight).

Getting back to the other genes in the above picture, more than 90% of the tumors carry somatic variants in one or more of these 4 genes: *TP53*, *PTEN*, *PIK3CA* and *KRAS*. These players appear again and again and again!

Which were the genes driving $\mathscr{X}$'s cancer? To find the answer, we take some cancer cells and some normal cells from $\mathscr{X}$. We then sequence all exons of carefully chosen genes in both cases. Ideally, we would have liked to sequence the exons of all 140 or so driver genes. However, not all of these driver genes can help $\mathscr{X}$ doctors in planning their next move. So we sequence all exons of 48 carefully chosen driver genes. Finally, we identify the differences between the cancer genome and the normal

genome. And we find 3 genes with key recipe-altering variants among these 48 genes.

Not surprisingly, two of these genes also appear in the picture above: *TP53* and *PTEN*. The third gene is also quite commonly found to carry somatic variants in cancer genomes; a gene called *EGFR*. Could this knowledge help $\mathscr{X}$'s doctors determine the next move against his cancer, which was growing increasingly stubborn?

The Guardian Lowers its Guard

TP53 carries the sobriquet *guardian of the genome*.[142] When cells detect genomic damage, they activate *TP53*, which then puts a brake on the cell division process. This gives time for genomic repair to happen before the genome is copied and the brake is released. And if the genome is considered beyond repair, *TP53* initiates programmed suicide, ensuring that cells with damaged genomes do not continue to replicate. This provides protection from the dangerous effects of genomic damage. Similarly, when certain genes which trigger cell division go on an overdrive, *TP53* provides a similar protective umbrella by inducing programmed suicide.

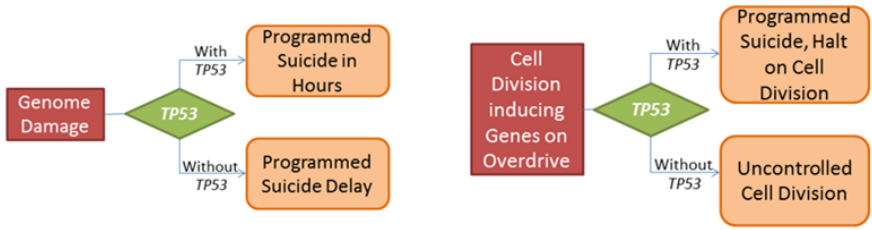


Figure 9.6: *TP53* Triggers Programmed Suicide, thus Protecting from Cancer.[142]

Given how important this watchdog role is, it is no wonder that as many as 50% of the cancers carry a recipe-altering character change in *TP53*. Curiously, this recipe-altering character change usually occurs in just one of the two *TP53* copies. The other copy is perfectly fine.

Given how absolutely critical this watchdog is, one would have expected nature to ensure that the good copy would cover for the bad one. Indeed, nature has built this sort of redundancy for many genes. For instance, the genes we encountered in Chapter 2 (the family with progressive loss of central vision), Chapter 5 (the two boys with their organs out of place), and Chapter 7 (the boy with cancer in the eye) are all protected by this redundancy; both copies of the gene have to be altered for things to go wrong. In fact, nature has built this redundancy for the *TP53* gene too. If one copy of the gene is completely lost for some reason, then it is still business as usual because our cells can detect this and generate twice the amount of protein from the remaining copy. However, nature has let a key structural lacuna slip through in case one copy of the gene is not lost, but simply modified!

The *TP53* gene recipe translates to a protein with 393 amino acids (remember how gene recipes code for proteins as shown in Fig. 2.5). This protein has several parts, two of which are shown prominently in blue and brown, respectively, in the picture below. The brown section is capable of attaching to the genome. The blue section is used to attach to other *TP53* protein instances. Four such proteins join their blue portions together, yielding a quartet, that works in concert.

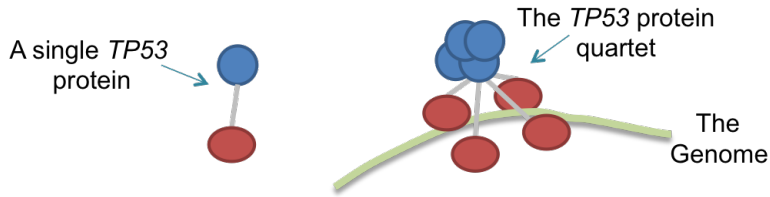


Figure 9.7: *TP53* Quartet Binds to the Genome Turning Near-by Genes On or Off.

When this quartet attaches to a specific place in the genome, a nearby gene is turned on (or off); which means, more (or less) recipe interpretation happens on that gene. Thus, by carefully attaching to specific places in the genome, this quartet is able to turn on or off specific genes needed

to accomplish the intended agenda, namely stalling cell division, and inducing programmed death, if needed.

Which locations in the genome can this quartet attach to? When the four constituents of the quartet behave independently, then only certain locations in the genome qualify.[143] And the genes that are turned on or off as a result stall cell division in response to genome damage but do not induce programmed suicide!

To turn on or off genes which induce programmed suicide, the quartet must attach to certain other locations in the genome. This happens when all four members of a quartet interact via bonds.[143] These bonds between oppositely charged amino acids in the brown portions in Fig. 9.7 provide a better grip on the genome. As a consequence, the quartet attaches more strongly to locations in the genome where individual constituents would have only attached weakly. This cooperation between the four members of the quartet enables *TP53* to induce programmed suicide in a cell in which genes that promote cell division have gone on an overdrive. And therein lies the catch!

Imagine now that one copy of *TP53* has a problematic character modification, and the other copy is fine. So 50% of the *TP53* protein created will come from the bad copy. When quartets are created now, they will have a mix of good and bad constituents; more than 2/3rds of the quartets will have at least two bad constituents. Since all the members of a quartet interact and function together, the presence of 2 bad components in a quartet is enough to weaken its attachment to the relevant genomic locations.[144] Induction of programmed suicide no longer happens as it should. So cells induced to divide on account of certain genes going on overdrive continue to divide instead of dying, leading to cancer.

Coming back to $\mathscr{X}$, two recipe-altering single character modifications in *TP53* are apparent when we sequence his cancer genome. At least some of his cells have the 272nd amino acid, usually a *Valine*, modified to a *Methionine* (remember the amino acid code from Fig. 2.5?). And at least some of his cells have the 143rd amino acid, again usually a *Valine*, modified to a *Methionine*. $\mathscr{X}$'s cells had neither of these variants at birth, so these were clearly acquired as the cancer evolved. Both variants are in the brown region of the protein that attaches to the genome (Fig

9.7). These variants are among the less potent variants known in *TP53*;[145] nevertheless, it appears that they would cause some reduction in *TP53*'s ability to induce programmed suicide. $\mathscr{X}$'s cancer cells had a weaker than usual hurdle opposing their unrelenting march. And $\mathscr{X}$'s doctors had a greater adversary to fight against.

Was there a drug that could somehow neutralize bad *TP53* protein copies, leaving the good ones to do their job? Unfortunately, the combined efforts of several pharmaceutical companies and academic researchers hasn't produced such a drug, though research in on-going on several candidates.[144] So $\mathscr{X}$'s doctors needed to look further to plan their next move.

A Sensor on Overdrive

The development and functioning of our bodies requires that our cell be told when to divide and when not to. For instance, the skull, the face, the eyes, the teeth etc, are all sculpted by carefully coordinating which cells divide and at what point of time. Who issues these signals and how do cells respond to these signals?

A historic observation in 1950 by Rita Levi-Montalcini provided early answers to this question. Levi-Montalcini implanted a mouse tumor into developing chick embryos. She was curious how the chick cells would react to the presence of these foreign cancer cells. And she noticed a particularly dramatic outgrowth of chick nerve fibers towards the implanted tumor cells. These nerve cells seemed to be dividing fast in response to something that the tumor cells were doing. How could the tumor cells induce the division of these nerve cells some distance away? It appeared as if the tumor cells were releasing a substance into the chick bloodstream; and nerve cells some distance away were then responding to this released substance by growing and dividing.

Today, we know that cells can release substances called *growth factors*. These are typically proteins obtained by recipe interpretation on certain genes. Other cells sense these growth factors and respond by growing and dividing. These responding cells could be near-by in some cases, and quite far away in others. How do these cells sense the presence of these growth

factors?

The boundary of a cell has sensor proteins whose task it is to detect the presence of growth factors outside the cell. These sensor proteins are called *growth factor receptors*. In its usual position on the cell boundary, some parts of the receptor appears outside the cell and other parts appear inside. When a growth factor attaches to the outer part, something changes in the inner part as well: either the shape of the protein or some markings on the amino acids (remember the marks in Fig. 7.6). These changes fire a cascade of events inside the cell. This cascade induces the cell to divide eventually.

Note that not all cells may have sensor proteins for a particular growth factor. Only those that do can respond. For instance, only certain nerve cells responded to the growth factor secreted by the implanted tumor in Levi-Montalcini's experiment.

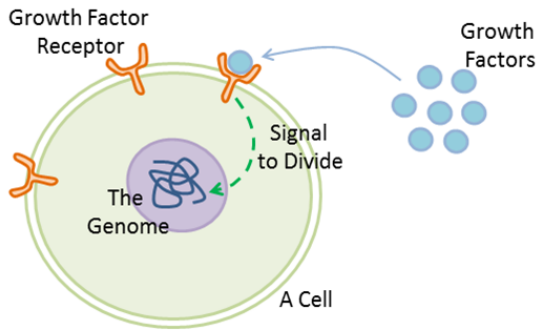


Figure 9.8: A Growth Factor Induces Cell Division via its Receptor.

Aha, you might say now! What if some cells run amok secreting growth factors in much larger than usual amounts? Would responding cells not lose their usual restraint and divide recklessly. Indeed they would. To an extent, cells control the impact of excess growth factor by controlling which cells generate growth factor receptors, and in what amounts. A receptor molecule is typically generated by recipe interpretation on one or more genes. By controlling the extent of recipe interpretation, the

number of receptor molecules can be controlled. By keeping this number in check, cell division is maintained at a sustainable rate. But what if a cell starts performing recipe interpretation on overdrive to generate many more receptors than needed? That too would be a problem.

And what if the receptor gene itself becomes trigger-happy due to a recipe-altering character modification? Very curiously, the sensor protein it generates starts firing signals inside the cell, all by itself, even when no growth factor is attached to it! This indeed appeared to be the case for $\mathscr{X}$, with the receptor in question being *EGFR*, short for *epidermal growth factor receptor*.

EGFR senses a few different growth factors. When a growth factor attaches to *EGFR*, two molecules of *EGFR* come together to form a pair. Each instance of *EGFR* in this pair has a section called the *kinase domain*. The kinase domain of one instance can add markers (actually phosphate groups, marked with a P in the picture below) to the other instance. This marking is analogous to turning on a switch, which is otherwise off. Turning on this switch fires the downstream cascade of events inside the cell that induces the cell to divide.

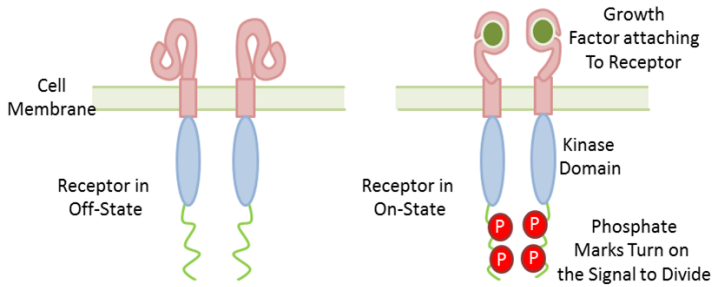


Figure 9.9: A Pair of *EGFR* Sensors Mark Each Other when a Growth Factor Attaches.

In $\mathscr{X}$, a single character change in some of his cancer cells caused the 725th amino acid in the kinase domain of *EGFR*, usually a *Threonine*, to become a *Methionine*. Remember, the kinase domain is responsible for adding the markers shown in Fig. 9.9. This recipe change causes the

kinase domain to become hyperactive, marking the amino acids shown in Fig. 9.9 to an extent even when no growth factor is attached to *EGFR*.[146] The downstream cascade inducing cell division now fires on its own, even when there is no growth factor attached!

This excessive firing by *EGFR* was one of the likely drivers of $\mathscr{X}$'s cancer. *TP53* would usually serve as a policeman by inducing programmed suicide, thus protecting cells from excessive cell division on count of *EGFR* and other similar genes going on overdrive.[142] But the recipe change in *TP53* we saw earlier had lowered this guard at least a bit. The combination then provided $\mathscr{X}$'s cancer an unhindered path to progress. What move could $\mathscr{X}$'s doctors make to control this progress by suppressing *EGFR*'s autonomous, unprovoked firing?

A Sharper Attack

By now, $\mathscr{X}$'s cancer had gathered much momentum, having evolved mechanisms to sidestep attacks with Cisplatin, Carboplatin etc. $\mathscr{X}$'s doctors had to attack where the cancer still hadn't built up its defenses by stopping *EGFR* from firing indiscriminately. A drug called *Erlotinib* is approved by the US Food and Drug Administration (FDA) as anti-cancer therapy, does exactly this.

Erlotinib blocks the kinase domain of *EGFR* from easily marking the amino acids shown in Fig. 9.9. With the marking process becoming more sluggish, the downstream cascade fires less often thus blocking cells from dividing on account of misfiring of *EGFR*. So this could potentially rein in the cancer. Actual verification would, of course, require clinical trials. And several such trials have been conducted.

Trials on patients with lung cancer who had developed resistance to cytotoxic chemotherapies like Cisplatin and Carboplatin have shown that about 7-9% of these patients do respond to Erlotinib.[147] The patients in these trials were chosen more broadly and probably comprised a mix of those whose cancers were driven by *EGFR* and those cancers weren't. One would expect that patients with recipe-altering variants in the kinase domain of *EGFR* would respond even better. Indeed, the response rate for these patients were much higher, at 27%.[148] 27% would appear as a small

number for the uninitiated, but for those who have seen cancer from near, every bit counts!

Based on the above trials, the US FDA approved the use of Erlotinib for patients with lung cancer who had developed resistance to cytotoxic chemotherapy in 2004. More recently, in 2013, the US FDA approved the use of Erlotinib as the very first line of therapy for lung cancer patients with a few specific recipe-altering variants in the kinase domain of *EGFR*.

$\mathscr{X}$'s cancer, a mesothelioma which arises not in the lungs but in the linings surrounding the lung, wasn't quite lung cancer of the type for which Erlotinib was approved by the US FDA. And therein lies an interesting dilemma: could drugs proven to work in one type of cancer work for another? Strictly speaking, a lung cell and a mesothelial cell turn different sets of genes on and off. So they could have different responses to the same therapy. Traditional wisdom therefore doesn't easily lend itself to transferring therapies for one cell type to another. And the use of Erlotinib in mesothelioma is far from proven, with some trials showing little effect[149] and others still on-going.[150]

The other side of the argument is no less compelling, even if hard to prove. $\mathscr{X}$'s cancer cells were indeed driven by a kinase domain variant in *EGFR*, roughly of the type Erlotinib is effective against. And given $\mathscr{X}$ had run out of all standard options, $\mathscr{X}$'s doctors had to try an experimental option. Erlotinib was the natural option to try. Indeed, on a hunch, $\mathscr{X}$'s doctors had made their move and put $\mathscr{X}$ on Erlotinib even before $\mathscr{X}$'s cancer genome was fully elucidated.

$\mathscr{X}$'s cancer responded well to this new move. The tumors starting reducing or at least stopped growing and his symptoms stayed in control without many side-effects. The cancer appeared to be have been reined in.

But, yet again, this time about 9 months later, the cancer figured out a way to get around and start growing again. Yet again, $\mathscr{X}$'s doctors needed to understand how the cancer had circumvented Erlotinib, and decide their next move accordingly. Could the third gene which carried a recipe-altering variant in $\mathscr{X}$'s cancer cells provide further options?

Another Brake Removed?

Remember when a particular growth factor attaches to *EFGR* on the cell boundary, it sets off a cascade of events inside the cell which induces the cell to divide. In fact, this sets off not one, but several, cascades. One such cascade is shown in the picture below, in much simplified form.

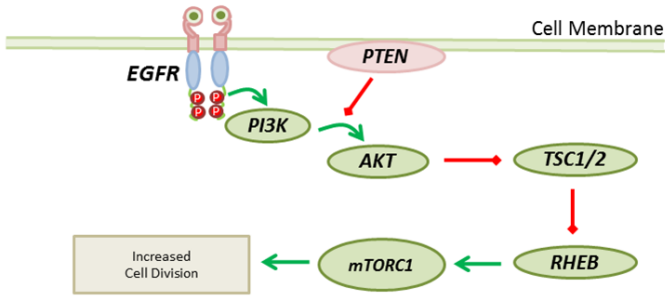


Figure 9.10: A Signal Cascade Initiated by *EGFR*. Green arrows indicate activation, red indicate brakes.

Each step in the cascade above involves one protein influencing another. This influence is either stepping on the brake to stop a particular function, or stepping on a gas pedal to activate that function. In the picture above, a red line indicates a brake, while a green line indicates a gas pedal. This is, of course, a simplified picture, for, the true picture has many more proteins, and many feedback loops, which together maintain tight control on the amounts of each of these proteins.

In this simplified picture, the cascade begins by activating *PI3K* once *EGFR* in switched on. This then activates *AKT*. Activated *AKT* works through a few intermediates to activate *mTORC1*.

Really, you might ask! Aren't there two brakes (red lines) on the path from *AKT* to *mTORC1*? So shouldn't an activated *AKT* put the brake on *mTORC1* rather than activating *mTORC1*? The answer: note there are two brakes on that path; and a brake applied on a brake is tantamount to stepping on the gas pedal!

So yes, activated *AKT* in turn activates *mTORC1*. Hyperactive *mTORC1*

does several things. It promotes the manufacture of proteins so there are ample protein resources to support a larger number of cells. It stops parts of the cell that are damaged from getting degraded, allowing such damage to accumulate. It also turns on recipe interpretation on certain genes which promote cell division. Together, these effects push the cell towards a cancerous state.[151]

Cells have an additional level of control on this cascade. Another protein applies a brake on the activation of *AKT* by *PI3K*. So, while a hyperactive *EGFR* leads to hyperactive *mTORC1*, this hyperactivity is somewhat limited by the protein in question: a protein obtained from a gene called *PTEN*. The very gene in which a recipe-altering somatic variant was found when we sequenced the genome in $\mathscr{X}$'s cancer cells. And a gene whose recipe is observed to be altered very often in cancerous cells (Fig. 9.5).

In $\mathscr{X}$, *EGFR* was firing on overdrive activating *AKT* and *mTORC1* in the process. And $\mathscr{X}$'s doctors had attempted to correct this via Erlotinib. But what if the recipe-alterations in *PTEN* were to remove the brake on the activation of *AKT* by *PI3K*. Wouldn't this cascade continue to fire even with Erlotinib?

Our study of $\mathscr{X}$'s genome showed at least some of his cancer cells carried single character recipe-alterations in *PTEN*. There are two such variants: one in amino acid 233 and another in amino acid 247. Both these alterations are exceedingly rare. One has been seen only in a single cancer patient before. The other is completely new and has never been reported before. Much further experimentation will be needed to establish if these indeed compromise the function of *PTEN* to the point where *AKT* is more freely activated by *PI3K*.

Of course, $\mathscr{X}$'s doctors could not wait that long to make the next move. The cancer was progressing ahead and demanded a more immediate answer. Research on some other variants located roughly in the same region of *PTEN* as these variants in $\mathscr{X}$ has shown that that 44% of the variants appear to indeed compromise the function of *PTEN*.[152] So there was some chance that *PTEN* was indeed compromised in some of $\mathscr{X}$'s cancer cells. Given that $\mathscr{X}$ was running out of options, making this assumption was possibly a risk worth taking. Even so, could $\mathscr{X}$'s doctors

use this information to make their next move?

The Next Move

A drug called *Everolimus* was being tried as a therapy for bladder cancer as part of a clinical trial on 37 patients. The trial itself failed because the drug did not produce an effect that was significant enough. There was one exception though: one patient showed a dramatic response. This lead to a natural question: what was special about this one patient? What made Everolimus work in this patient but fail in many others?

Sequencing the cancer genome of this patient gave the answer: a recipe alteration in the *TSC1* gene.[153] A frameshift alteration of the type we saw in Chapter 3, which caused the recipe to truncate prematurely. If you look at the cascade of events shown in Fig. 9.10, you will spot *TSC1* with a red line to *RHEB*), which in turn has a green line to *mTORC1*. So *TSC1* keeps a brake on *RHEB*, which would otherwise press on the gas pedal pushing *mTORC1* to fire. So effectively, *TSC1* keeps a brake on *mTORC1*. And with *TSC1* gone or at least reduced, the foot was partially or fully off the brake, and *mTORC1* was firing more freely than usual in the patient in question. And, Everolimus, the drug being trialed, put the brake back on *mTORC1*, as it was intended to do!

In a nutshell, the patient in question responded because Everolimus corrected the error that drove the cancer in the first place. Other patients responded less favorably because the gene variants that drove their cancers were probably different.

Coming back to $\mathscr{X}$, if you follow the the cascade of brakes and gas pedals between *PTEN* and *mTORC1* in Fig. 9.10, you will be sure to notice that the net effect of *PTEN* is to put a brake on *mTORC1*. The variants in *PTEN* in $\mathscr{X}$'s cancer cells could have possibly taken this brake off. And *Everolimus* could potentially put that brake back. For $\mathscr{X}$'s doctors, this was a natural candidate for the next move.

Studies showed that the side effects of combining *Everolimus* with *Erlotinib* were manageable.[154] Accordingly, and after appropriate permissions for trying experimental therapy, $\mathscr{X}$, was put on that combination. *Erlotinib* would stem *EGFR* from firing indiscriminately, and *Everolimus*

would apply an additional brake on *mTORC1*.

$\mathscr{X}$ had cycled through several therapies earlier. Each had pushed the cancer back and bought $\mathscr{X}$ time. But his cancer was now resistant to all these therapies. Would the combination of *Everolimus* with *Erlotinib* give $\mathscr{X}$ a fresh new lease of life?

Wrapping Up

The genome in our cells is constantly under attack: from various chemicals we inhale or ingest, from sunlight and other forms of radiation, and from viruses which infect us. Add to these genome-copying errors as our genome is copied trillions of times to create new cells. Damage to the genome and genome-copying errors cause somatic variants: variants which we do not inherit from our parents. The accumulation of sufficiently many such variants in important regions of the genome, in a single normal cell, then transforms that cell to a cancer cell.

This evolution often takes years. Typically, half or more of the somatic recipe-altering variants which appear in a cancer cell occur before the cancer begins.[141] These are not driver variants, so they just ride along without tipping the cell over into a cancerous state. Then the first driver variant occurs, initiating the formation of a slow-growing tumor. Cells now divide slightly faster, often only slightly faster, than the rate at which they die. Over many years, this slight imbalance is sufficient to produce a large pool of cells. One of these then receives a second driver variant, which confers an even faster division rate to this cell and its descendants. This snowballing effect eventually results in a few cells accumulating several driver variants, leading to a fast growing, malignant tumor.

By the time such tumors announce themselves, they are often too diffused to eliminate completely by surgery and radiation. The only possible approach then is via drugs that can travel to all parts of the body via the blood stream. Some such drugs, like Cisplatin and Carboplatin, have blunt effects, broadly killing all dividing cells. Others are more specific, e.g., Erlotinib which has an impact only on cells which have substantial amount of *EGFR* protein. Sequencing the cancer genome often suggests such drug candidates, providing doctors additional options to fight the cancer.

But cancer is a dynamic adversary, reacting and evolving with every such drug. Several cancer cells do die with each attack providing a period of time where the illness is eliminated or at least controlled. Sometimes, this can bring an happy end to the ordeal. More often, a few cancer cells remain and evolve further somatic variants which make them resistant. The disease then comes back, with renewed vigor. A different drug has to be tried then.

Indeed, several such moves had been tried by $\mathscr{X}$'s doctors, with each such move buying a fresh lease of life, albeit temporarily. Each time the cancer had made its countermove and bounced back. And one by one, possible options for treatment had been eliminated, leaving $\mathscr{X}$'s doctors struggling for new options. A peek into the genomes of $\mathscr{X}$'s cancer cells suggested first Erlotinib, and now the Erlotinib-Everolimus combination.

The Erlotinib-Everolimus pushed $\mathscr{X}$'s cancer to the backfoot once again, giving him yet another fresh lease of life. A fresh lease of modestly good quality life, for his symptoms and side-effects were not too debilitating.

Of course, a lasting response would have been a miracle. That was not to be. The response lasted for 6 months or so. The formidable adversary that it was, the cancer had once again found a way around.

Just 6 months, you might ask, was this worth all the trouble? Remember, mesothelioma is aggressive and most patients survive only a year past diagnosis. In contrast, $\mathscr{X}$ was still alive 5 years past diagnosis. Each move had bought him several months to a year of life. In particular, Erlotinib and then the Erlotinib-Everolimus combination, both suggested by sequencing his cancer genome, had added almost a year and a half of modestly good quality life! A small, albeit solemn, victory for modern medical science.

Imagine, in the future, that doctors have an extended repertoire of drugs to combat cancer. With each countermove the cancer makes, a new move pushes the cancer back to the wall, forcing it to work hard to dodge the bullet. Imagine further, that doctors have the ability to repeatedly check the cancer genome to identify what somatic variants the cancer has evolved as it responds to each move, helping guide doctors on making the next move. Could this game of moves and countermoves with cancer be extended much longer? Even indefinitely?

Indeed, believes Harold Varmus, ex-director of the US National Cancer Institute and noted cancer researcher:[155] *I think we can turn most if not all cancers into chronic, controllable, non-lethal disorders, and sometimes cure them, even after they have widely metastasized, if we make the necessary investments in development and testing of new therapies*. On that note, and with the hope that the stories in this book will inspire its readers to devote their energies to the quest laid out by Varmus, we bring this book to an end.

REFERENCES

[1] Mathias Legrand. “The Legrand Orange Book”. In: *http://www.latextemplates.com/template/the-legrand-orange-book* ().

[2] Y. Robbers and A. Skjold. “Creating Book Covers using PSTricks”. In: *https://tug.org/pracjourn/2007-1/robbers/robbers.pdf* ().

[3] S. Ishihara. “Tests for colour-blindness”. In: *Handaya, Tokyo* (1917).

[4] S. S. Deeb. “The molecular basis of variation in human color vision”. In: *Clinical Genetics* 67 (2005), pp. 369–377.

[5] B. C. Verrelli and S. A. Tishkoff. “Signatures of Selection and Gene Conversion Associated with Human Color Vision Variation”. In: *American Journal of Human Genetics* 75(3) (2004), pp. 363–375.

[6] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. In: *Nature* (2004), pp. 931–945.

[7] D. M. Hunt et al. “The chemistry of John Dalton’s color blindness”. In: *Science* 267 (1995), pp. 964–988.

[8] J. R. Sparrow, K. Nakanishi, and C. A. Parish. “The lipofuscin fluorophore A2E mediates blue light-induced damage to retinal pigmented epithelial cells”. In: *Investigative Ophthalmology and Visual Science* 7 (2000), pp. 1981–9.

[9] J. R. Sparrow et al. “A2E, a byproduct of the visual cycle”. In: *Vision Research* 28 (2003), pp. 2983–2990.

[10] S. Schmitz-Valckenberg et al. “Fundus autofluorescence imaging: review and perspectives”. In: *Retina* 3 (2008), pp. 385–409.

[11] J. Ahn et al. “Functional interaction between the two halves of the photoreceptor-specific ATP binding cassette protein ABCR (ABCA4). Evidence for a non-exchangeable ADP in the first nucleotide binding domain”. In: *Journal of Biological Chemistry* 278.41 (2003), pp. 39600–39608.

[12] M. Zhong, L. L. Molday, and S. R. Molday. “Role of the C Terminus of the Photoreceptor ABCA4 Transporter in Protein Folding, Function, and Retinal Degenerative Diseases”. In: *Journal of Biological Chemistry* 284 (2009), pp. 3640–3649.

[13] R. Battu et al. “Identification of Novel Mutations in ABCA4 Gene: Clinical and Genetic Analysis of Indian Patients with Stargardt Disease”. In: *Biomedical Research International* (2015).

[14] H. Sun, P. M. Smallwood, and J. Nathans. “Biochemical defects in ABCR protein variants associated with human retinopathies”. In: *Nature Genetics* 26 (2000), pp. 242–246.

[15] T. R. Burke et al. “Retinal Phenotypes in Patients Homozygous for the G1961E Mutation in the ABCA4 Gene”. In: *Investigative Ophthalmology and Visual Science* 53.8 (2012), pp. 4458–4467.

[16] D. M. Lipinski, M. Thake, and R. E. MacLaren. “Clinical applications of retinal gene therapy”. In: *Progress in Retinal and Eye Research* (2013), pp. 22–47.

[17] J. Kong et al. “Correction of the disease phenotype in the mouse model of Stargardt disease by lentiviral gene therapy”. In: *Gene Therapy* 19 (2008), pp. 1311–1320.

[18] Oxford Biomedica. “Phase I/IIa Study of StarGen in Patients With Stargardt Macular Degeneration”. In: *http://clinicaltrials.gov/ct2/show/study/NCT01367444?term=stargen&rank=2* ().

[19] S. D. Schwartz et al. “Embryonic stem cell trials for macular degeneration: a preliminary report”. In: *The Lancet* 9817 (2012), pp. 713–720.

[20] Advanced Cell Technology. “Sub-retinal Transplantation of hESC Derived RPE(MA09-hRPE)Cells in Patients With Stargardt’s Macular Dystrophy”. In: *http://clinicaltrials.gov/ct2/show/NCT01345006* ().

[21] S.W. Peltz et al. “Ataluren as an agent for therapeutic nonsense suppression”. In: *Annual Review of Medicine* (2013), pp. 407–425.

[22] D. Wanga et al. “Cardiac channelopathy testing in 274 ethnically diverse sudden unexplained deaths”. In: *Forensic Science International* (2014), pp. 90–99.

[23] C. Napolitano et al. “Sudden Cardiac Death and Genetic Ion Channelopathies: Long QT, Brugada, Short QT, Catecholaminergic Polymorphic Ventricular Tachycardia, and Idiopathic Ventricular Fibrillation”. In: *Circulation* (2012), pp. 2027–2034.

[24] B. J. Maron et al. “Prevalence of Hypertrophic Cardiomyopathy in a General Population of Young Adults: Echocardiographic Analysis of 4111 Subjects in the CARDIA Study”. In: *Circulation* (1995), pp. 785–789.

[25] D. S. Herman et al. “Truncations of Titin Causing Dilated Cardiomyopathy”. In: *New England Journal of Medicine* 7 (2012), pp. 619–628.

[26] V. Carmignac V et al. “C-terminal titin deletions cause a novel early-onset myopathy with fatal cardiomyopathy.” In: *Annals of Neurology* 5 (2012), p. 728.

[27] R. Knoll et al. “Laminin-alpha4 and integrin-linked kinase mutations cause human cardiomyopathy via simultaneous defects in cardiomyocytes and endothelial cells”. In: *Circulation* 5 (2007), pp. 515–525.

[28] G. I. Gallicano et al. “Desmoplakin is required early in development for assembly of desmosomes and cytoskeletal linkage.” In: *Journal of Cellular Biology* 7 (1998), pp. 2009–2022.

[29] E. Garcia-Gras et al. “Suppression of canonical Wnt/beta-catenin signaling by nuclear plakoglobin recapitulates phenotype of arrhythmogenic right ventricular cardiomyopathy”. In: *Journal of Clinical Investigation* 7 (2006), pp. 2012–2021.

[30] J. Gomes et al. “Electrophysiological abnormalities precede overt structural changes in arrhythmogenic right ventricular cardiomyopathy due to mutations in desmoplakin-A combined murine and human study”. In: *European Heart Journal* 15 (2012), pp. 1942–1953.

[31] G. Quarta et al. “Familial evaluation in arrhythmogenic right ventricular cardiomyopathy: impact of genetics and revised task force criteria”. In: *Circulation* 23 (2011), pp. 2701–2709.

[32] T.J. Pugh et al. “The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing”. In: *Genetics in Medicine* (2014).

[33] T. B. Rasmussen et al. “Protein expression studies of desmoplakin mutations in cardiomyopathy patients reveal different molecular disease mechanisms”. In: *Clinical Genetics* 1 (2013), pp. 20–30.

[34] M. Norman et al. “Novel Mutation in Desmoplakin Causes Arrhythmogenic Left Ventricular Cardiomyopathy”. In: *Circulation* (2005), pp. 636–642.

[35] C. A. James et al. “Exercise increases age-related penetrance and arrhythmic risk in arrhythmogenic right ventricular dysplasia (cardiomyopathy) associated desmosomal mutation carriers”. In: *Journal of the American College of Cardiolgy* 14 (2013), pp. 1290–1297.

[36] C. L. Lien et al. “Heart repair and regeneration: recent insights from zebrafish studies”. In: *Wound Repair and Regeneration* 5 (2012), pp. 638–646.

[37] B. Loeys et al. “Homozygosity for a missense mutation in fibulin-5 results in a severe form of cutis laxa”. In: *Human Molecular Genetics* 18 (2002), pp. 2113–2118.

[38] D. S. Peabody. “Translation Initiation at Non-AUG Triplets in Mammalian Cells”. In: *Journal of Biological Chemistry* 9 (1989), pp. 5031–5035.

[39] H. Liu and L. Wong. “Data mining tools for biological sequences”. In: *Journal of Bioinformatics and Computational Biology* 1 (2003), pp. 139–167.

[40] T. Nakamura et al. “Fibulin-5/DANCE is essential for elastogenesis in vivo”. In: *Nature* 6868 (2002), pp. 171–175.

[41] M. Budatha et al. “Extracellular matrix proteases contribute to progression of pelvic organ prolapse in mice and humans”. In: *Journal of Clinical Investigation* 5 (2011), pp. 2048–2059.

[42] Z. Razinia et al. “Filamins in mechanosensing and signaling”. In: *Annual Review of Biophysics* (2012), pp. 227–246.

[43] E. Parrini et al. “Periventricular heterotopia: phenotypic heterogeneity and correlation with Filamin A mutations”. In: *Brain* (2006), pp. 1892–1906.

[44] S. Singh et al. “Case series of 4 infants with severe infantile respiratory failure associated with Filamin A mutation leading to lung transplantation”. In: *American Thoracic Society International Conference Abstracts: A106. INTERESTING PEDIATRIC CASES* (2013).

[45] S.P. Robertson. “Filamin A:phenotypic diversity”. In: *Current Opinion in Genetics and Development* 3 (2005), pp. 301–307.

[46] A. Verloes et al. “Fronto-otopalatodigital osteodysplasia: clinical evidence for a single entity encompassing Melnick-Needles syndrome, otopalatodigital syndrome types 1 and 2, and frontometaphyseal dysplasia”. In: *Americal Journal of Medical Genetics* 5 (2000), pp. 407–422.

[47] C. Foley et al. “Expansion of the Spectrum of FLNA Mutations Associated with Melnick-Needles Syndrome”. In: *Molecular Syndromology* 3 (2010), pp. 121–128.

[48] K.L. Jones, M.C. Jones, and M. Del Campo. *Smith's Recognizable Patterns of Human Malformation*. 2013, p. 762.

[49] H.H Santos et al. "Mutational analysis of two boys with the severe perinatally lethal Melnick-Needles syndrome". In: *American Journal of Medical Genetics A* 3 (2010), pp. 726–731.

[50] B.A. Kesner et al. "Isoform Divergence of the Filamin Family of Proteins". In: *Molecular Biology and Evolution* 2 (2009), pp. 283–295.

[51] Y. Feng et al. "*Filamin A (FLNA)* is required for cell–cell contact in vascular development and cardiac morphogenesis". In: *Proceedings of the National Academy of Sciences* 52 (2006), pp. 19836–19841.

[52] C.J. Saunders et al. "Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units". In: *Science Translational Medicine* (2012).

[53] R. T. Moon. "Xenopus Embryo: Beta-Catenin and Dorsal–Ventral Axis Formation". In: *eLS. John Wiley and Sons Ltd, Chichester* (2005).

[54] S. Yoshiba and H. Hamada. "Roles of cilia, fluid flow, and Ca2+ signaling in breaking of left–right symmetry". In: *Trends in Genetics* 1 (2013), pp. 10–17.

[55] W.M. Layton Jr. "Random determination of a developmental process". In: *The Journal of Heredity* (1976), pp. 336–338.

[56] C.J. Tabin and K.J. Vogan. "A two-cilia model for vertebrate left-right axis specification". In: *Genes and Develeopment* (2003), pp. 1–6.

[57] S. Sauer and A. Klar. "Left-right symmetry breaking in mice by left-right dynein may occur via a biased chromatid segregation mechanism, without directly involving the Nodal gene". In: *Frontiers in Oncology* (2012).

[58] T.P. Yamaguchi. "Heads or tails: Wnts and anterior-posterior patterning". In: *Current Biology* (2001), pp. 713–724.

[59] A. Caron, X. Xu, and X. Lin. “Wnt/Beta-catenin signaling directly regulates Foxj1 expression and ciliogenesis in zebrafish Kupffer’s vesicle”. In: *Development* (2012), pp. 514–524.

[60] J. Deka et al. “Bcl9/Bcl9l are critical for Wnt-mediated regulation of stem cell traits in colon epithelium and adenocarcinomas”. In: *Cancer Research* (2010), pp. 6619–6628.

[61] R. Cole and R. Hariharan. “Approximate String Matching: A Simpler Faster Algorithm”. In: *SIAM Journal on Computing* (2002), pp. 1761–1783.

[62] K. Ahokas et al. “Matrix metalloproteinase-21 is expressed epithelially during development and in cancer and is up-regulated by transforming growth factor-beta1 in keratinocytes”. In: *Lab Investigation* (2003), pp. 1887–1899.

[63] N. Madan et al. “Frequency of beta-thalassemia trait and other hemoglobinopathies in northern and western India”. In: *Indian Journal Human Genetics* 1 (2010).

[64] X. Hong, D.G. Scofield, and M. Lynch. “Intron size, abundance, and distribution within untranslated regions of genes”. In: *Molecular Biology and Evolution* 12 (2006), pp. 2392–2404.

[65] S. Clancy. “RNA Splicing: Introns, Exons and Spliceosome”. In: *Nature Education* 1 (2008), p. 31.

[66] K. Gao et al. “Human branch point consensus sequence is yUnAy”. In: *Nucleic Acids Research* 7 (2008), pp. 2257–2267.

[67] A. Corvelo et al. “Genome-wide association between branch point properties and alternative splicing”. In: *Public Library of Science, Computational Biology* 11 (2010).

[68] Z-M. Zheng et al. “Optimization of a Weak 3-prime Splice Site Counteracts the Function of a Bovine Papillomavirus Type 1 Exonic Splicing Suppressor In Vitro and In Vivo”. In: *Journal of Virology* 13 (2000), pp. 5902–5910.

[69] M.D. Bashyam et al. “Molecular genetic analyses of beta-thalassemia in South India reveals rare mutations in the beta-globin gene”. In: *Journal of Human Genetics* (2004), pp. 408–413.

[70] M.D. Bashyam, A.K. Chaudhary, and V. Bhat. “The IVS-II-837 (T>G) appears to be a relatively common ’rare’ beta-globin gene mutation in beta-thalassemia patients in Karnataka State, South India”. In: *Hemoglobin* 5 (2012), pp. 497–503.

[71] N.Y. Varawalla, J.M. Old, and D.J. Weatherall. “Rare beta-thalassaemia mutations in Asian Indians”. In: *British Journal of Haemotology* 4 (1991), pp. 640–644.

[72] J. Gaziev and G. Lucarelli. “Hematopoietic Stem Cell Transplantation for Thalassemia”. In: *Current Stem Cell Research and Therapy* 2 (2011), pp. 162–169.

[73] M. Sabloff et al. “HLA-matched sibling bone marrow transplantation for beta-thalassemia major”. In: *Blood* 5 (2011), pp. 1745–1750.

[74] R. Barrangou et al. “CRISPR provides acquired resistance against viruses in prokaryotes”. In: *Science* 5819 (2007), pp. 1709–1712.

[75] M. Jinek M et al. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. In: *Science* 6096 (2012), pp. 816–821.

[76] B. Shen et al. “Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects”. In: *Nature Methods* (2014), pp. 399–402.

[77] H. Yin et al. “Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype”. In: *Nature Biotechnology* (2014), pp. 551–553.

[78] J. Lewis et al. “A Common Human Beta Globin Splicing Mutation Modeled in Mice”. In: *Blood* 6 (1998), pp. 2152–2156.

[79] J. Morley Smith. “Red Reflex Image, in Public Domain”. In: *http://upload.wikimedia.org/wikipedia/commons/d/d8/Rb_whiteeye.PNG* (2008).

[80] Ashwin Mallipatna. “White Reflex Image, Private Communication, Copyright: Narayana Nethralaya”. In: (2015).

[81] Hetal Vyas. “Photographer’s awareness saved girl’s vision”. In: *http://timesofindia.indiatimes.com/city/bengaluru/Photographers-awareness-saved-girls-vision/articleshow/13160469.cms* (2012).

[82] A. Knudson. “Mutation and Cancer: Statistical Study of Retinoblastoma”. In: *Proceedings of the National Academy of Sciences* 4 (1971), pp. 820–823.

[83] A.M. Narasimha et al. “Cyclin D activates the Rb tumor suppressor by mono-phosphorylation”. In: *ELife* (2014).

[84] C.R. de Andrade et al. “A molecular study of first and second RB1 mutational hits in retinoblastoma patients”. In: *Cancer Genetics and Cytogenetics* (2006), pp. 43–46.

[85] W. Chen and S. Jinks-Robertson. “The Role of the Mismatch Repair Machinery in Regulating Mitotic and Meiotic Recombination Between Diverged Sequences in Yeast”. In: *Genetics* (1999), pp. 1299–1313.

[86] M.C. LaFave and J. Sekelsky. “Mitotic Recombination: Why? When? How? Where?” In: *Public Library of Science, Genetics* 3 (2009).

[87] W. Jiao et al. “Aberrant nucleocytoplasmic localization of the retinoblastoma tumor suppressor protein in human cancer correlates with moderate/poor tumor differentiation”. In: *Oncogene* (2008), pp. 3156–3164.

[88] E. Zacksenhaus et al. “A bipartite nuclear localization signal in the retinoblastoma gene product and its importance for biological activity”. In: *Molecular and Cellular Biology* 4 (1993), pp. 4588–4599.

[89] B.L. Gallie et al. “Developmental Basis of Retinal-specific Induction of Cancer by RB Mutation”. In: *Cancer Research* (1999), pp. 1731–1735.

[90] X.L. Xu et al. “Rb suppresses human cone-precursor-derived retinoblastoma tumours”. In: *Nature* (2014), pp. 385–388.

[91] C.O. Nordling. “A New Theory on the Cancer-inducing Mechanism”. In: *British Journal of Cancer* 1 (1953), pp. 68–72.

[92] J.P. de Magalhaes. “How ageing processes influence cancer”. In: *Nature Reviews Cancer* 5 (2013), pp. 357–365.

[93] L. Zografos A. Balmer and F. Munier. “Diagnosis and current management of retinoblastoma”. In: *Nature Reviews Cancer* (2006), pp. 5341–5349.

[94] B.P. Alter et al. “Cancer in Fanconi anemia”. In: *Blood* 5 (2003).

[95] D.W. Smithers. “Family histories of 459 patients with cancer of the breast”. In: *British Journal of Cancer* 2 (1948), pp. 163–167.

[96] E.B. Claus, N.J. Risch, and W.D. Thompson. “Age at onset as an indicator of familial risk of breast cancer”. In: *American Journal of Epidemiology* 6 (1990), pp. 961–972.

[97] J.M. Hall et al. “Linkage of early-onset familial breast cancer to chromosome 17q21”. In: *Science* 4988 (1990), pp. 1684–1689.

[98] R. Wooster et al. “Identification of the breast cancer susceptibility gene BRCA2”. In: *Nature* (1995), pp. 789–792.

[99] A. Jolie. “My Medical Choice”. In: *http://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html* (2013).

[100] A. Jolie Pitt. “Angelina Jolie Pitt: Diary of a Surgery”. In: *http://www.nytimes.com/2015/03/24/opinion/angelina-jolie-pitt-diary-of-a-surgery.html* (2015).

[101] D.A. Kreutzer and J.M. Essigmann. “Oxidized, deaminated cytosines are a source of C->T transitions in vivo”. In: *Proceedings of the National Academy of Sciences* 7 (1998), pp. 3578–3582.

[102] A. Valavanidis, T. Vlachogianni, and K. Fiotakis. “Tobacco Smoke: Involvement of Reactive Oxygen Species and Stable Free Radicals in Mechanisms of Oxidative Damage, Carcinogenesis and Synergistic Effects with Other Respirable Particles”. In: *International Journal of Environmental Research and Public Health* 2 (2009), pp. 445–462.

[103] D.E. Heck A.M. Vetrano, T.M. Mariano, and J.D. Laskin. “UVB Light Stimulates Production of Reactive Oxygen Species”. In: *The Journal of Biological Chemistry* 25 (2003), pp. 22432–22436.

[104] Y. Huang and L. Li. “DNA crosslinking damage and cancer - a tale of friend and foe”. In: *Translational Cancer Research* 3 (2013), pp. 144–154.

[105] G.L. Moldovan and A.D. D’Andrea. “How the fanconi anemia pathway guards the genome”. In: *Annual Reviews in Genetics* (2009), pp. 223–249.

[106] M. Huang et al. “Human MutS and FANCM complexes function as redundant DNA damage sensors in the Fanconi Anemia pathway”. In: *DNA Repair* 12 (2011), pp. 1203–1212.

[107] A.D. D’Andrea. “BRCA1: A Missing Link in the Fanconi Anemia/BRCA Pathway”. In: *Cancer Discovery* 4 (2013), pp. 376–378.

[108] G. Ast. “How did alternative splicing evolve”. In: *Nature Reviews Genetics* (2004), pp. 773–782.

[109] E. Buratti et al. “Aberrant 5’ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization”. In: *Nucleic Acids Research* 13 (2007), pp. 4250–4263.

[110] X. Roca, R. Sachidanandam, and A.R. Krainer. “Intrinsic differences between authentic and cryptic 5’ splice sites”. In: *Nucleic Acids Research* 21 (2003), pp. 6321–6333.

[111] M.J. Hicks et al. “Intrinsic differences between authentic and cryptic 5’ splice sites”. In: *Molecular and Cellular Biology* 8 (2010), pp. 1878–1886.

[112] E. Buratti1 et al. “Intrinsic differences between authentic and cryptic 5’ splice sites”. In: *Nucleic Acids Research* 13 (2007), pp. 4250–4263.

[113] C. Attanasio, A. David, and M. Neerman-Arbez. “Outcome of donor splice site mutations accounting for congenital afibrinogenemia reflects order of intron removal in the fibrinogen alpha gene (FGA)”. In: *Blood* 5 (2003), pp. 1851–1856.

[114] S.C. Chandrasekharappa et al. “Massively parallel sequencing, aCGH, and RNA-Seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia”. In: *Blood* 22 (2013), pp. 138–148.

[115] Y.J. Machida et al. “UBE2T is the E2 in the Fanconi anemia pathway and undergoes negative autoregulation”. In: *Molecular Cell* (2006), pp. 589–596.

[116] A.F. Alpi et al. “Mechanistic insight into site-restricted monoubiquitination of FANCD2 by Ube2t, FANCL, and FANCI”. In: *Molecular Cell* (2008), pp. 767–777.

[117] A.J. Deans and S.C. West. “DNA interstrand crosslink repair and cancer”. In: *Nature Reviews* (2011), pp. 467–480.

[118] R. Ceccaldi et al. “Bone Marrow Failure in Fanconi Anemia Is Triggered by an Exacerbated p53-p21 DNA Damage Response that Impairs Hematopoietic Stem and Progenitor Cells”. In: *Cell Stem Cell* 1 (2012), pp. 36–49.

[119] R. Roy, J. Chun, and S.N. Powell. “BRCA1 and BRCA2: different roles in a common pathway of genome protection”. In: *Nature Reviews* (2012), pp. 68–78.

[120] R.L. Attanoos. “Malignant Mesothelioma: Asbestos Exposure”. In: *Occupational Cancers* (2014), pp. 273–284.

[121] M. Krishnan and S.G. Ray. “Banned in 52 countries, asbestos is India’s next big killer”. In: *http://archive.tehelka.com/story_main46.asp?filename=Cr070810banned.asp* (2010).

[122] R.A. Weiss and P.K. Vogt. “100 years of Rous sarcoma virus”. In: *Journal of Experimental Medicine* 208.12 (2011), pp. 2351–2355.

[123] P. Rous. “A sarcoma of the fowl transmissible by an agent separable from the tumor cells”. In: *Journal of Experimental Medicine* 13 (1911), pp. 397–411.

[124] A.P. Czernilofsky et al. “Nucleotide sequence of an avian sarcoma virus oncogene (src) and proposed amino acid sequence for gene product”. In: *Nature* 287.5779 (1980), pp. 198–203.

[125] D. Stehelin et al. “Detection and enumeration of transformation-defective strains of avian sarcoma virus with molecular hybridization”. In: *Virology* 76.2 (1977), pp. 675–684.

[126] J.A. Cooper et al. “Tyr527 is phosphorylated in pp60c-src: implications for regulation”. In: *Science* 231.4744 (1986), pp. 1431–1434.

[127] D.M. Parkin. “The global health burden of infection-associated cancers in the year 2002”. In: *International Journal of Cancer* 118.12 (2006), pp. 3030–3044.

[128] J.S. Butel and J.A. Lednicky. “Cell and molecular biology of simian virus 40: implications for human infections and disease.” In: *Journal of the National Cancer Institute* 91.2 (1999), pp. 119–134.

[129] A.L. Cleaver et al. “Long-term exposure of mesothelial cells to SV40 and asbestos leads to malignant transformation and chemotherapy resistance”. In: *Carcinogenesis* 35.2 (2014), pp. 407–414.

[130] M. Carbone. “Simian virus 40 and human tumors: It is time to study mechanisms.” In: *Journal of Cellular Biochemistry* 76.2 (1999), pp. 189–193.

[131] L. Zhang et al. “Tissue Tropism of SV40 Transformation of Human Cells: Role of the Viral Regulatory Region and of Cellular Oncogenes.” In: *Genes and Cancer* 1.10 (2010), pp. 1008–1020.

[132] C. Tomasetti and B. Vogelstein. “Variation in cancer risk among tissues can be explained by the number of stem cell divisions”. In: *Science* 347.6217 (2015), pp. 78–81.

[133] National Comprehensive Cancer Network. “Malignant Pleural Mesothelioma”. In: *http://www.nccn.org/patients/guidelines/mpm/* (2014).

[134] S. Papaspyros and S. Papaspyros. “Variation in cancer risk among tissues can be explained by the number of stem cell divisions”. In: *International Scholarly Research Notes, Surgery* 2014 (2014).

[135] N.J. Vogelzang et al. “Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma”. In: *Journal of Clinical Oncology* 21.14 (2014), pp. 2636–2644.

[136] D-W. Shen et al. “Cisplatin Resistance: A Cellular Self-Defense Mechanism Resulting from Multiple Epigenetic and Genetic Changes”. In: *Pharmacological Reviews* 64.3 (2012), pp. 706–721.

[137] B. Castagneto et al. “Phase II study of pemetrexed in combination with carboplatin in patients with malignant pleural mesothelioma (MPM)”. In: *Annals of Oncology* 19.2 (2008), pp. 370–373.

[138] W. DeNeve et al. “Discrepancy between cytotoxicity and DNA interstrand crosslinking of carboplatin and cisplatin in vivo”. In: *Investigational New Drugs* 8.1 (1990), pp. 17–24.

[139] A. K. Nowak et al. “A multicentre phase II study of cisplatin and gemcitabine for malignant mesothelioma”. In: *British Journal of Cancer* 87.5 (2002), pp. 491–496.

[140] E. Mini et al. “Cellular pharmacology of gemcitabine”. In: *Annals of Oncology* 17.5 (2006), pp. 7–12.

[141] B. Vogelstein et al. “Cancer Genome Landscapes”. In: *Science* 339.6127 (2013), pp. 1546–1558.

[142] A. Efeyan and M. Serrano. “p53: Guardian of the Genome and Policeman of the Oncogenes”. In: *Cell Cycle* 6.9 (2007), pp. 1006–1010.

[143] K. Schlereth et al. “Characterization of the p53 Cistrome–DNA Binding Cooperativity Dissects p53’s Tumor Suppressor Functions”. In: *Public Library of Science, Genetics* 9.8 (2013).

[144] A.C. Joerger and A.R. Fersht. “The Tumor Suppressor p53: From Structures to Drug Discovery”. In: *Cold Spring Harbor Perspectives in Biology* 2.6 (2010).

[145] L.R. Dearth et al. “Inactive full-length p53 mutants lacking dominant wild-type p53 inhibition highlight loss of heterozygosity as an important aspect of p53 status in human cancers”. In: *Carcinogenesis* 28.2 (2007), pp. 289–298.

[146] U. ManChon et al. “Prediction and Prioritization of Rare Oncogenic Mutations in the Cancer Kinome Using Novel Features and Multiple Classifiers”. In: *Public Library of Science, Computational Biology* 10.4 (2014).

[147] F.A. Shepherd et al. “Erlotinib in previously treated non-small-cell lung cancer”. In: *New England Journal of Medicine* 353.2 (2005), pp. 123–132.

[148] C-Q Zhu et al. “Role of KRAS and EGFR As Biomarkers of Response to Erlotinib in National Cancer Institute of Canada Clinical Trials Group Study BR.21”. In: *Journal of Clinical Oncology* 26.26 (2008), pp. 4268–4275.

[149] L.L. Garland et al. “Phase II Study of Erlotinib in Patients With Malignant Pleural Mesothelioma: A Southwest Oncology Group Study”. In: *Journal of Clinical Oncology* 25.17 (2007), pp. 2406–2413.

[150] “Erlotinib in Treating Patients With Malignant Mesothelioma of the Lung”. In: *https://clinicaltrials.gov/ct2/show/NCT00039182* (2013).

[151] S. Mori et al. “The mTOR Pathway Controls Cell Proliferation by Regulating the FoxO3a Transcription Factor via SGK1 Kinase”. In: *Public Library of Science One* 9.2 (2014).

[152] S-Y. Han et al. “Functional Evaluation of PTEN Missense Mutations Using in Vitro Phosphoinositide Phosphatase Assay”. In: *Cancer Research* 60.12 (2000), pp. 3147–3151.

[153] G. Iyer et al. “Genome Sequencing Identifies a Basis for Everolimus Sensitivity”. In: *Science* 338.6104 (2012), p. 221.

[154] V.A. Papadimitrakopoulou et al. “Everolimus and erlotinib as second- or third-line therapy in patients with advanced non-small-cell lung cancer”. In: *Journal of Thoracic Oncology* 7.10 (2010), pp. 1594–1601.

[155] Harold Varmus. “Can cancer be stopped?” In: *http://www.nydailynews.com/opinion/harold-varmus-cancer-stopped-article-1.2181020* (2015).

Our Genome comprises a staggering 6 billion characters! These characters impact our lives, often in subtle ways, and occasionally in more dramatic and serious ways. The stories in this book are all real sagas of such genomic characters and their consequences: brothers who pass away mysteriously early in life, a boy whose blood can't carry enough oxygen, siblings whose organs appear out of place, a family whose members lose vision in their 30s, siblings whose hearts fail in their 30s, a baby with cancer of the eye, a middle-aged patient battling cancer, etc. Each story proves to be a detective quest that connects the world of medical practice with that of molecular biology, traversing the world of computer algorithms along the way.

Ramesh Hariharan is the Chief Technology Officer at Strand Life Sciences and an Adjunct Professor at the Indian Institute of Science. Ramesh holds a bachelor's degree in Computer Science from the Indian Institute of Technology, New Delhi, and a Ph.D. in Computer Science from the Courant Institute, New York University. Ramesh served as faculty at the Indian Institute of Science for several years, researching computer algorithms. Fascinated by the use of computation to understand biological systems, Ramesh and 3 colleagues founded Strand Life Sciences in 2000. At Strand, Ramesh works on making genome sequencing efficient and ubiquitously accessible to patients and doctors. Do send your feedback on this book to him at ramesh@strandls.com.

Strand Life Sciences
http://www.strandls.com