On Matters Of Nature and Science

L A Liggett

2019-04-29

Contents

Chapter 1

Introduction

Here is an unexceptional glimpse of the universe as it was during my moment in time.

Chapter 2

Genetics and Genomics

2.1 Introduction

A haplotype block is a set of closely linked alleles or markers on a chromosome that tend to be inherited together over evolutionary time.

Across Eukaryotes, the frequency of recombination is inversely proportional to overall genome size. The result is that yeast have a recombination rate a few orders of magnitude higher than that of humans (She and Jarosz, 2018).

2.1.1 Subpoint

This is some sub info

2.1.2 Second subpoint

A haplotype block is a set of closely linked alleles or markers on a chromosome that tend to be inherited together over evolutionary time.

Across Eukaryotes, the frequency of recombination is inversely proportional to overall genome size. The result is that yeast have a recombination rate a few orders of magnitude higher than that of humans (She and Jarosz, 2018).

2.2 DNA Replication

When the origin of replication(s) is removed from bacteria or eukaryotes, growth and division is restricted or entirely eliminated, but in some strains of archaea like H volcanii, deletion of the origin of replication accelerates cell growth rates. It turns out that this archaea can use a process that is similar to homologous recombination to create a replication fork and replicate its chromosome (Hawkins et al., 2013).

2.3 Mutation Rate

2.3.1 Observed Mutation Rates

Using whole-genome sequencing or next-gen sequencing to determine mutation rates by base, it appears that C>T mutations at CpG sites mutate at a frequency of 10-7 changes/cell division, and all other sites are within the range of 10^-8 - 10^-9 base changes per division (Arnheim and Calabrese, 2009, 2016; Campbell and Eichler, 2013; Ségurel et al., 2014; **?**; **?**).

It appears that humans have the highest germline mutation rate of all analyzed species (Lynch, 2016). It also appears from trio sequencing that both human and chimp mutation rates are in the range of 1.0-1.25x10^-8 mutations per site per generation (**?**).

While germline mutation rates are similar in chimps and humans, there are important differences, one example being that CpG mutations are more prevalent in the chimp germline (**?**).

Possibly supporting the somatic theory of aging, increases and reductions in DNA damage rates may accelerate and decelerate, respectively, some aspects of aging (**??**).

2.3.2 Mutation Rate Evolution

Providing an example of how human mutation rates can differ by geographic origination, Europeans compared to African/Asian populations have a 1.6 increased mutation rate of a TCC->TTC transitions (Harris, 2015). This change in DNA replication fidelity appears to have happened 40-80kya when Europeans and Asians diverged and illustrates that DNA repair rates have not remained entirely stable throughout human evolution.

Just as in humans show a number of different mutational likelihoods, the great apes apes also show evolution in the rates of different triplet-context mutation rates (**?**). It would be interesting to investigate how these different rates have constrained evolution, if at all.

2.4 Mutation Hotspots

There are a number of sporadic mutation hotspots associated with disease incidence, like achondroplasia which has a sporadic incidence rate of 4.5 x 10-5 per generation (Arnheim and Calabrese, 2016; Waller et al., 2008). This disease originates from a single mutation in the FGFR3 gene at a mutation rate ~450 times higher than what would ordinarily be expected at a CpG site (Bellus et al., 1995; Rousseau et al., 1994; Shiang et al., 1994).

2.5 Mutation Detection

Pyrophosphorolysis-activated polymerization is a mutation detection method that can detect a single mutant molecule of DNA within 25,000 genomes (Liu and Sommer, 2004; Qin et al., 2007).

2.6 Genetic Modifications

Caffeine was cloned to allow for caffeine-deficient coffee and teas without the decaffeination process (Kato et al., 2000).

When a DNA-associating protein from tardigrades was cloned into mammalian cells, they became about 40% more tolerant to radiation (Hashimoto et al., 2016).

2.7 Sequencing Methods

Using a new sequencing method called sci-RNA-seq, the transcriptome of every cell of 762 cells in C Elegans was sequenced to yield single-cell sequencing results and transcriptome profiling of every cell in the body. The way this is done is by methanol fixing nuclei and then incorporating a UMI when converting to cDNA, then mixing cells again and incorporating another UMI when synthesizing the other strand (Cao et al., 2017).

It appears that DAPI does not increase sequencing error rates by Illumina sequencing (Leung et al., 2016).

One group came up with a method that is essentially identical to mine in which they use barcoded probes to detect leukemia but they tracked the mutation manually and ignored background (Wong et al. 2015)

2.8 Diagnostics

In Li and Snyder Cell 2018, the EHR from hospitals is used to integrate with a machine learning algorithm trained on aneurysm detection. Patients are then whole genome sequenced, and the genome sequencing plus the lifestyle of the individual on EHR is then used to predict if the person has an aneurysm. They were able to achieve pretty robust detection results that could then be used in a prediction setting in the clinic.

2.9 Detecting Common Diseases

Much of the following information comes from this review: (**?**).

Linkage disequilibrium studies were designed to detect Mendelian diseases GWAS designed back in 1996 to detect non-mendelian multigenic traits that have much less penetrant effects The promise that GWAS could risk stratify people for diseases has been challenging because most diseases seem to be driven by an extremely large number of variants with small effects that will likely require extremely large sample sizes There exists a problem of missing heritability, and it was often believed that common SNPs only held part of the puzzle, and more rare variants accounted for a great deal of heritability, but this does not yet seem to be the case, and SNPs seem to have a much greater effect size Another problem with GWAS is it is haplotype specific in that it can implicate a stretch of DNA inherited from one parent, but is blind to the individual effect sizes of each of the individual variants A challenge raised by Jonathan Pritchard is that gene regulatory networks are so interconnected that variants in one gene may actually cause changes in other genes and are therefore only peripherally relevant to a phenotype One continued promise of the utility of GWAS to identify the causal genetics behind diseases is that most of the strongest GWAS associations came from small studies of european populations that identified mutations of large effect sizes. By expanding studies to populations, especially those like african populations that have less linkage disequilibrium many more variants of large effect sizes could be identified and used to tease out relationships of smaller effect sizes in other populations. Methods are also improving fo linking regulatory elements to the genes they regulate like (Gasperini et al. 2018; Gasperini et al. 2019). Linking regulatory elements to their corresponding genes can be quite helpful, because this information can be incorporated into GWAS calculations to refine causal linkage probabilities. Polygenic risk scores have often been used to predict phenotypic variance in plants and animals, and have yet to really be applied to human genomics (Khera et al. 2018). Training of PRSs seem to not require fine-mapping, and their use has been aided by the UK Biobank (Bycroft et al. 2018).

2.10 Detecting Rare Diseases

There are some 7k Mendelian monogenic disorders that impact about 0.5% of live births, but contribute to about 70% of pediatric hospital admissions An important surprise has been that de-novo mutations account for a substantial amount of intellectual disabilities and autism, where as many as 30-60% of ASD is caused

by de-novo mutations Currently as many as half of acutely ill inpatient infants can be diagnosed from WGS. There are currently 59 genes designated by the American College of Medical Genetics as being sufficiently clinically actionable as to warrant sequencing and reporting in patients (Kalia et al. 2017).

2.11 Tissue Evolution

By sequencing 7,664 tumors spanning 29 different cancer types, it appears that unlike species evolution, the force of positive selection in developing tumors outweighs that of negative selection as evidenced by the loss of less than 1 coding nucleotide substitution per tumor (**?**). The number of mutations per cancer varied from 1 per thyroid and testicle cancer to over 10 per endometrial and colorectal cancers. This information helps to answer how many mutations are needed to effectively create cancers and how this can vary with across tissue types. A number of groups have tried to answer these questions in the past by mathematically estimating the number of rate limiting steps required in the process of oncogenesis (**??**). There are two important problems with this approach, first that not all driver mutations need to be rate-limiting (**?**), and not all rate limiting steps in oncogenesis need to be driver mutations (**?**). It has also been problematic to sequence tumors and count the number of high frequency mutations in oncogenic genes, but this has the added challenges of distinguishing passenger from driver mutations and is limited to current lists of oncogenes. Lists of genes involved in cancer have become increasingly detailed, but are still limited (**??**). The absence of negative selection in cancer may well indicate how dispensable the majority of genes are for somatic cells.

In their paper, Martincorena and Campbell show that the dN/dS ratio for somatic tissues and cancer tissue is 1 or greater showing that the effect of negative selection is minimal. In contrast, the dN/dS for germline species evolution is less than 0.5 showing a much greater effect of negative selection. Surprisingly, the non-synonymous mutations showed a dN/dS ration of 1 whether they existed in haploid or diploid regions suggesting the cells were not simply tolerating the mutations by having two copies.

Using sequencing of blood to find spontaneous mutations and go back and use population biology to calculate the number of HSCs, it was found that there were between 50k-200k HSCs contributing to hematopoiesis at any given time (**?**).

2.12 Mutational Processes in Cancer

Most cancers carry between 1,000 and 20,000 somatic point mutations and a few hundred insertions, deletions, and rearrangements (**?**). Leukemias and pediatric brain cancers typically contain the lowest numbers of mutations while those tissues exposed to mutagens like lung and skin cancers tend to have the highest numbers (**????**). It is interesting to note that from the Cancer Gene Census database, it appears that only three genes; TP53, PIK3CA, and BRAF are mutated in 10% or more patients (**?**).

Carcinogens or particular mutations often cause identifiable mutational patterns. BRCA1 and BRCA2 mutations commonly associated with breast, ovarian, and pancreatic cancers are associated with particular substitution patterns, small indels, and large chromosome duplications (**?**).

While mutations in TP53 and NOTCH1 can cause an imbalance in the division symmetry of stem cells in intestinal crypts by increasing the ratio of proliferation to differentiation(**??**), mutations like APC and KRAS can increase the rates of crypt fission to allow multiple crypts to become clonal (**??**).

2.13 Mutagenesis in Cancer

2.13.1 Mutation Rates

Substitution rates in B/T lymphocytes appears to be on the order of 2 to 10 mutations per diploid genome per cell division, a rate which may be about 10 times higher than that occurring in germ cells (**?**).

2.13.2 Mutation Load

Sun exposed skin cells can carry thousands of point mutations, and about 25-30% of these cells have acquired at least one driver mutation, and yet, while positive selection is evident as the clones expand, clone sizes are relatively similar across individuals suggesting that some mechanism is constraining the expansion of driver clones (**?**).

The progression of Barrett's esophagus to esophageal adenocarcinoma shows no significant increase in mutation load. This may mean that Barrett's esophagus is an advanced precancerous lesion, but some other step in the tissue is required for it to transition into such a stage (**?**).

2.14 Genetic Manipulation

2.14.1 Cloning

A macaque was the first primate to be cloned by SCNT (Liu et al., 2018).

2.14.2 Gene Therapy

X-linked severe combined immunodeficiency (SCID-X1) (commonly referred to as bubble-boy disease) is caused by mutations within the IL2RG gene which impairs the receptor for IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21. The result of these mutations is that T and NK cells fail to develop. This disease is often treated with bone marrow transplant which involves irradiation. Instead of such a treatment, low-dose busulfan has been used in combination with WT IL2TG DNA carrying lentivirus to give HSCs a functional copy of the receptor. This treatment seems to allow T and NK cells to develop with minimal harm to the patient (**?**).

2.14.3 Genetic Rescue

Inbreeding of a population can result in an accumulation of deleterious mutations. The accompanying loss in fitness can be mitigated through outbreeding, as it can increase heterozygosity and thereby mask recessive deleterious alleles (**???**). This effect depends on the principle of overdominance or heterozygous vigour, which is a condition where a heterozygous offspring has a higher fitness than either of the parents.

Using a modeling approach, it appears that genetic rescue of a species will result in a substantial loss of the original genes of an organism. According to the models, the degree of genetic replacement of additive genes will proportionally match the genetic rescue, while the replacement of recessive genes may not need to occur as much as they will be masked by the foreign alleles (**?**). One caveat to this, is that even if a large fraction of the genome in the inbred organism is recessive, it is likely that a substantial portion of it must be lost to restore fitness, as it is more likely that most of the recessive mutations contribute to the reduction in fitness, each with a small fitness effect, rather than only a handful of the mutations contributing with a large fitness effect.

Chapter 3

Neuroscience

3.1 Alzheimer's Disease

It seems that the bacteria Porphyromonas gingivalis may be partially responsible for exacerbating symptoms in people with AD. Observation of the bacteria in the brains of people with AD showed evidence that the bacteria secreted proteases called gingipains that increased A production and were also neurotoxic. Inhibition of the proteases reduced neuroinflammation and rescued neurons in the hippocampus (Dominy 2019).

Tied both to sleep and AD, it appears that chronic sleep deprivation increases the formation of A plaques, perhaps suggesting that sleep is protective against the formation of the plaques that eventually lead to AD (**?**).

]

3.2 Sleep

Sleep deprivation results in the global phosphorylation of the brain proteome, an effect which is reversed by sleep. A mouse mutant for SIK3 causes mice to sleep more, and may play an important role in the cause of sleep desire (Wang et al. 2018).

Chapter 4

Aging

4.1 Somatic Mutation Theory

In their study covering 7,664 tumors, Martincorena and Campbell illustrate an almost complete absence of negative selection against somatic mutations (**?**). This is lack of negative selection is quite different from the force at the population level where it acts quite strongly on the germline. This result may be an important finding for the somatic theory of aging (**?**), as it may be an indication that point mutations are largely of negligible effect size within somatic cells. As such, if point mutations do play a role in aging, it seems that this may be limited to those that are neutral or advantageous to a cell.

Chapter 5

Cell Competition

5.1 Introduction

One of the first examples of cell competition was the discovery of Minutes heterozygous cells (M+/-) in 1975 (Morata Dev Bio 1975 and Gogna Ann Rev Gen 2015). This along with Myc and Wnt mediated cell competition provide a good understanding of what it takes for certain cells to survive instead of their peers, but recent deep sequencing has revealed that alone, the Minutes mechanism may not always be sufficient to understand cell competition.

5.2 Trophic Theory

Phenotypically, Drosophila with M(+/-) showed a reduced growth rate, but eventually reached full size with no noticeable defects. If M+/- were induced early during development, they were subsequently lost, and only retained into adulthood if they were induced late in development (Simpson Dev Bio 1979). Moreover, the number of WT or M+/- that were recovered at any given point were as would be predicted by just the difference in growth rates alone. 25 years later, this process of elimination was found to be apoptosis dependent, and driven by reduced Decapentaplegic (Dpp) pathway activation in M+/- cells (Moreno Nature 2002) (The Dpp morphogen is ortholog of the vertebrate Bone Morphogenetic Protein (BMP) and is necessary for patterning of the 15 imaginal discs in Drosophila). These experiments demonstrated how the Minutes gene switches between a 'winner' and a 'loser' phenotype. In the resultant 'ligand capture model', WT cells can outcompete Minutes cells because they can bind up the prosurvival trophic factor Dpp, while depriving the Minutes cells of this signal, leading to their apoptosis.

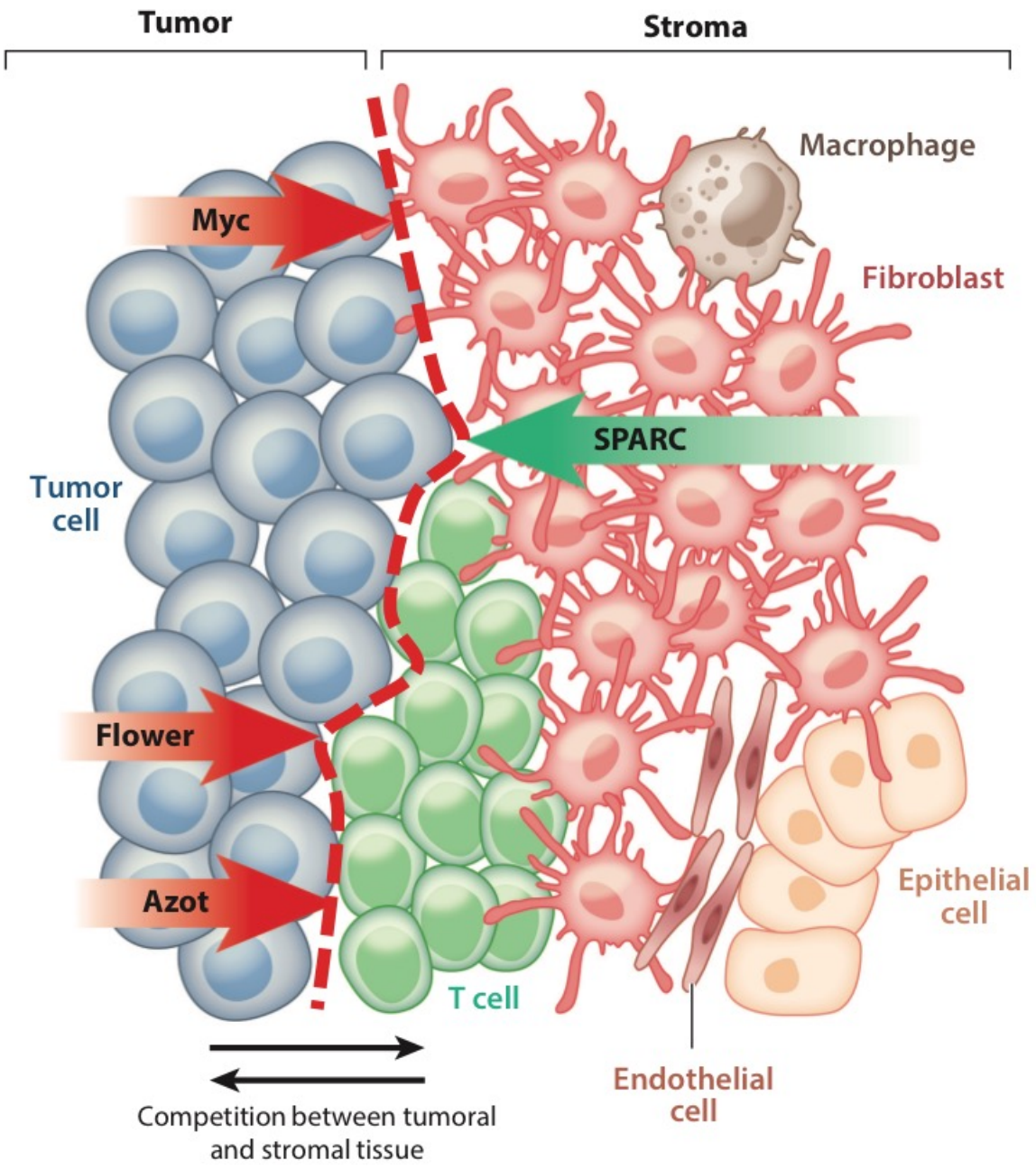
This idea of supercompetitors has been demonstrated with the proto-oncogene Myc in Drosophila. When Myc is overexpressed in developing Drosophila the Myc overexpressing cells grow rapidly and cause the elimination of surrounding WT cells through induction of apoptosis, as seen in the outcompetition of the Minutes cells (de la Cova C Cell 2004 and Moreno Cell 2008). The idea of supercompetitors is that the particular mutation carried by a cell confers on it some significant advantage over its peers, yet when Myc is downregulated, cells are at a disadvantage and are lost (Johnston Cell 1999). It appears that population levels of Myc regulate the fitness of a cell such that a cell must have more Myc than its peers to have an advantage, and the number of copies of Myc a cell has over its peers determines the magnitude of its advantage (Moreno Cell 2004).

People have continued to explore the 'trophic theory' of cell competition, finding other supporting evidence in the Wingless (Wg, Wnt ortholog) pathway. Just as Minute cells and Myc overexpressing cells do, Wg overexpressing cells in Drosophila induce JNK mediated apoptosis (Vincent Dev Cell 2011) in peers as a means of outcompeting them (Giraldez Development 2003).

5.3 Modern Theory

Secreted protein acidic and rich in cysteine (SPARC, osteonectin) is often expressed by loser cells (Portela Drosophila 2010), and as a secreted protein that modulates cell-cell and cell-ECM interactions, it is induced during morphogenesis, injury remodeling, and development (Clark J Cell Biochem 2008) and during cancer progression (Bradshaw Int J Biochem Cell Bio 2012). SPARC is important for cell competition because it raises the threshold for caspase activation in loser cells, thereby protecting cells of otherwise reduced fitness (Portela Drosophila 2010). SPARC can be a biomarker of certain types of cancer (Chlenski Semin Cell Dev Bio 2010) and may play a significant role in competition regulation during oncogenesis (Yamada Med Mol Morph 2015). SPARC is often associated with a number of cancers including breast cancer (Witkiewicz Cancer Bio Ther 2010, Bergamaschi J Path 2008), melanoma (Fenouille Pigment Cell Melanoma Res 2011, Massi Human Path 1999), osteosarcoma (Dalla-Torre BMC Cancer 2006), glioblastoma (Rich Cancer Res 2005), and bladder cancer (Yamanaka J Urol 2001). SPARC expression in surrounding stromal cells is indicative of a better prognosis in NSCLC (Koukourakis Cancer Res 2003), colon cancer (Lian J Mol Cell Card 2003), and pancreatic adenocarcinoma (Mantoni Cancer Bio Ther 2008).

```
knitr::include_graphics(rep("images/04-1.jpg", 1))
```




```
knitr::opts_chunk$set(comment=NA, fig.width=1, fig.height=1)
```

Flower is a putative transmembrane protein that has three isoforms, two are losers and one is a winner in Drosophila growing disc epithelium (Rhiner Drosophila Dev Cell 2010). RNAi against the losers is sufficient to make a cell a winner and vice versa for flower expressing cells in Drosophila epithelium. When DMBA-TPA is used to promote papilloma formation in mice, loser flower isotypes are expressed by surrounding stromal cells, while winner flower isotypes are expressed by the papilloma, suggesting that this mechanism is conserved from flies to vertebrates (Petrova Dis Models Mech 2012).

Azot is an intracellular protein that integrates the information from flower and SPARC both from the same cell and from surrounding cells and decide if its cell should apoptose (Petrova Commun Integr Biol 2011).

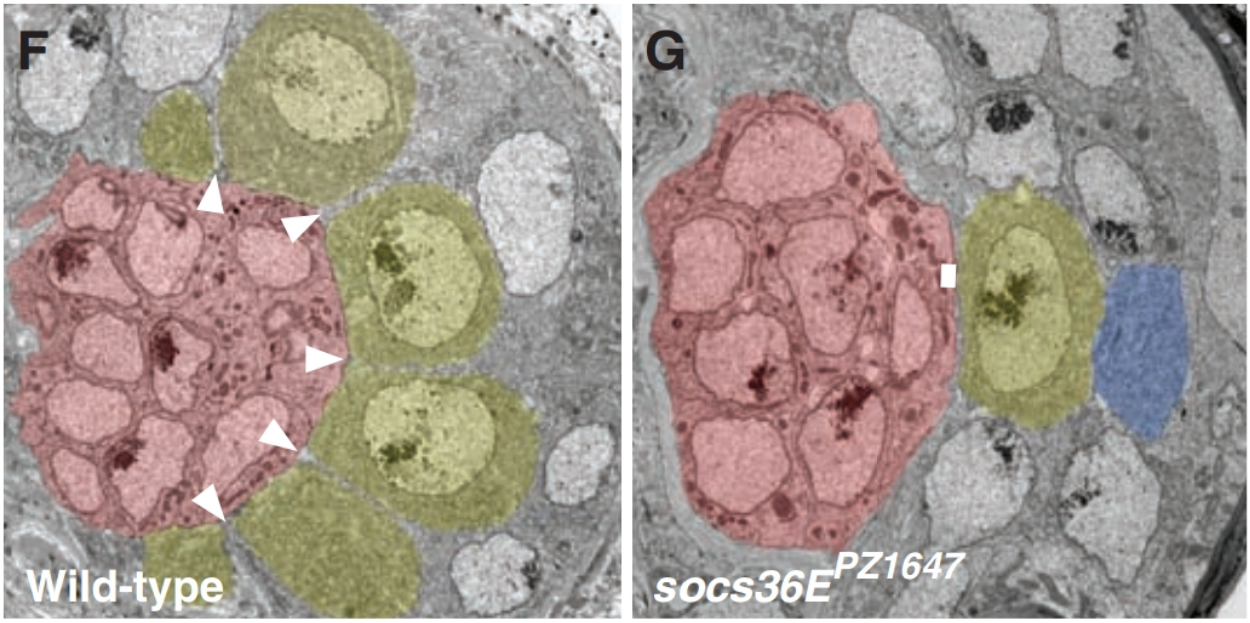
If cells have low expression levels of loser flower isoform or high SPARC, or surrounding cells have higher expression levels of loser flower, azot is not transcribed and the cell will survive (Merino Cell 2015).

It might be possible that while tumor cells are growing and expanding, they express fitness impacting genes like flower, Azot, and Myc, which gives them a competitive advantage over surrounding stromal cells.

5.4 Localization

In the Drosophila gonad, both germline stem cells (GSCs) and somatic stem cells (cyst progenitor cells, CPCs) share a niche created by stromal cells that make up the hub and both share a requirement for JAK-STAT signaling to maintain their stemness. This system provides an understanding for how a single niche with multiple cell types is not overrun by the one that cycles the most rapidly.

```
knitr::include_graphics(rep("images/04-2.jpg", 1))
```




```
knitr::opts_chunk$set(comment=NA, fig.width=1, fig.height=1)
```

In the Drosophila testis, suppressor of cytokine signaling (SOCS36E) which is a JAK-STAT antagonist is expressed at high levels in the hub and at low levels in the CPCs. The JAK-STAT signaling in part drives the expression of position specific PS-integrin which is used by both the GSCs and the CPCs to bind and localize to the hub. The image to the right shows in F how typically the GSCs (yellow) have a large surface area integrin mediated connection with the hub, while CPCs (gray) only make small integrin mediated connections. As shown in G, when SOCS36E is inhibited in the CPCs, JAK-STAT signaling is downregulated resulting in an upregulation of PS-integrin which then allows the CPCs to bind more readily to the hub. This increased surface area of association between the hub and the CPCs allows the CPCs to outcompete the GSCs for binding spots, and eventually results in loss of GPCs (Issigonis Science 2009). Aberrant JAK-STAT signaling which includes loss of SOCS expression in mammals(Bowman Oncogene 2000, Pontier J Cell Sci 2009, Leeman Expert Opin Biol Ther 2006), and may play a role in stem cell misregulation driven cancers (Reya Nature 2001).

Similar to mammals, germline stem cell transplants in Drosophila are more efficient if the original GSCs are first depleted, supporting a space-filling model of GSC-hub binding (Bhattacharya Eur J Imm 2008, Oatley

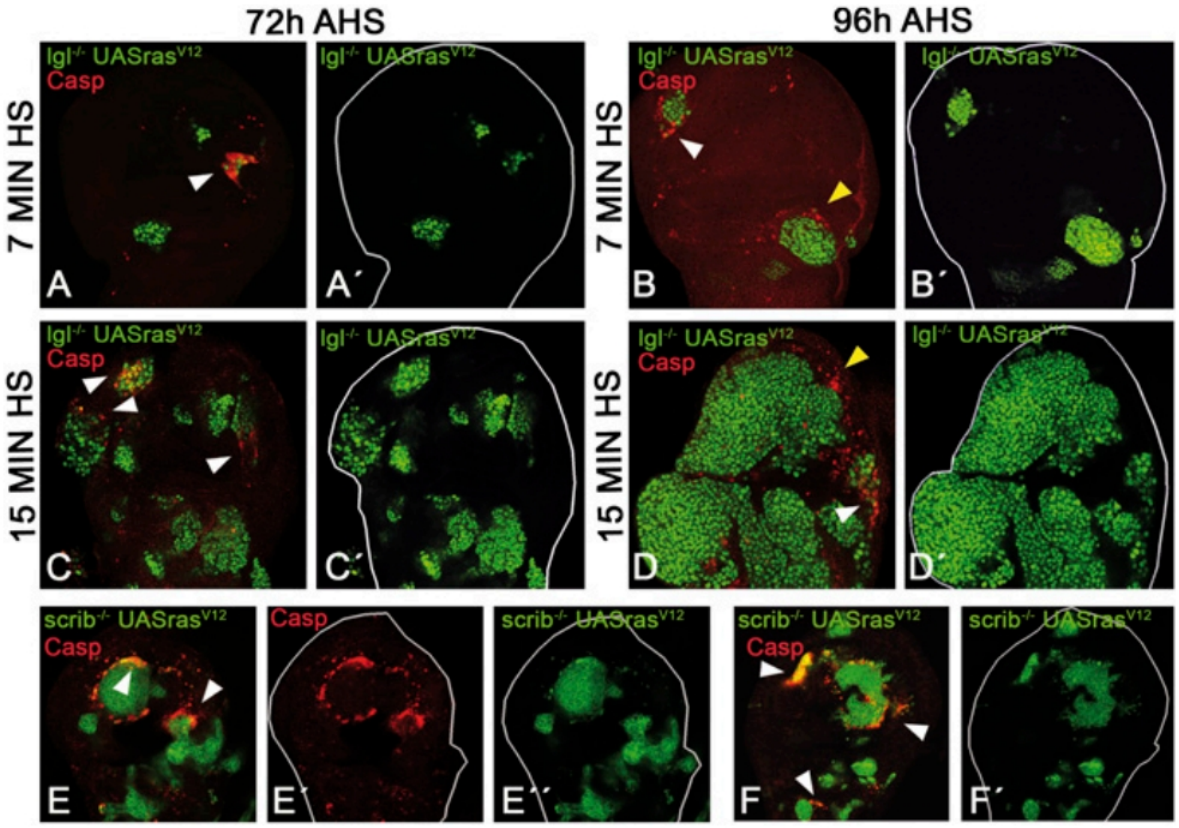
Ann Rev Cell Dev Bio 2008). Upon GSC loss, hub binding space is freed, which drives dedifferentiation of somatic stem cells that then fill the empty area adjacent to other GSCs (Sheng Cell Stem Cell 2009).

In the skin, collagen 17 (COL17A1), a component of the hemidesmosome, fluctuates in response to genomic and oxidative stress induced proteolysis. The stem cells that maintain high levels of COL17A1 are able to maintain contact with the niche and symmetrically divide, allowing them to outcompete neighbors that express low levels of COL17A1 as they will divide asymmetrically. With age it seems that all skin stem cells begin to lose their COL17A1 expression and eventually start to delaminate, and contributing to aging. Forced maintenance of COL17A1 however, seems to somewhat rescue skin aging (**?**).

5.5 Microenvironment

Lethal Giant Larvae (lgl) and scribble (scrib) are tumor suppressor genes in Drosophila that play necessary roles in establishing cell polarity and asymmetric cell divisions (Knoblich Cell 2008). Larvae that are constitutively mutant for lgl develop diffuse, ultimately lethal tumors, but clones of mutant cells surrounded by WT cells do not produce tumors (Froldi BMC Biol 2010, Igaki Dev Cell 2009).

```
knitr::opts_chunk$set(comment=NA, fig.width=1, fig.height=1)
knitr::include_graphics(rep("images/04-3.jpg", 1))
```



It appears that scrib mutant cells undergo JNK-driven apoptosis when surrounded by WT cells (Brumby EMBO 2003). This resembles minutes and Dpp cell competition (Moreno Nature 2002, Morata Dev Bio 1975). When lgl- cells are created using a heat shock method in Drosophila in order to affect only a small percentage of cells, the cells would still apoptose, and do so preferentially at the border of clones indicating a short-range apoptosis inducing mechanism originating from WT cells. Interestingly, when larvae were

heat shocked for different amounts of time, there was an exponential increase in surviving non-apoptosing lgl cells (figure) (Menéndez et al. 2010). The theory for why this occurs is that lgl- clones need to form a protective niche to survive, and merging clones can facilitate this. This may solve the riddle of how all oncogenic mutations could be present in a given tissue and not give rise to a cancer, it may be that a critical mass of oncogenically initiated cells is necessary before they can avoid WT cell outcompetition.

Chapter 6

Bioinformatics

6.1 Genome stuff

There are many instances of 'NNN' repeats within the human genome, because there are parts that are very challenging to sequence, including telomere and centromere regions. Lowercase letters in the genome specify things like repetitive sequences and introns. k-mers are fragments of DNA in the genome that are of a length k, so for the string ATGCA, the 2-mers are AT, TG, GC, and CA, and the 3-mers are ATG, TGC, and GCA k-mers can be useful for error correction because sequencing errors can bias towards producing k-mers, but they can also be used in alignment or in genome classification No free lunch theorem

6.2 Software

Google has designed a new variant caller called DeepVariant that uses deep neural networks to learn how best to call variants. This appears to provide a significant improvement to variant calling accuracy (Poplin et al., 2016).

A group put together well defined pipelines for common bioinfomatics analysis like RNASeq into bioconductor that can be easily used and then cited as a method of succinctly referring to the analysis process (http://www.nature.com/authors/policies/license.html).

A group came up with a deep machine learning algorithm that can be used to predict patient data and genomics patterns (Ching et al., 2017).

A group created a machine-learning algorithm for variant calling that seems to significantly outperform many common variant calling software like somatic sniper and mutect, though they do not compare against freebayes. Might be worth taking a look at (Wood et al., 2018)

Selene is a deep learning python library designed using PyTorch that is designed to ask questions using genomic and sequencing data (**?**). They also mention in their paper a number of other deep learning python libraries specifically designed to analyze genomics data, that may be worth checking out: DragoNN, pysster(**?**), and Kipoi(**?**).

Chapter 7

Cancer

7.1 Introduction

7.1.1 Incidence Rates

Time and age plays a major role in the occurrence of many cancers, where typically the risk of suffering most cancers is about 2% by the age of 40 in humans, while the risk increases to about 50% by the age of 80 (**?**).

```
knitr::include_graphics(rep("images/04-4.jpg", 1))
```