Chapter 1: Themes and challenges in mathematics of cancer.

Nigel J. Burroughs

January 18, 2017

1 Introduction: What is cancer

Cancer is a disease caused by mutation during normal growth generating cells that escape the cellular control processes that normally keep growth in check. The human genome comprises 3 billion paired bases (the 4 letters, A, T, G, C of DNA), which needs to be duplicated at each cell division. Duplication is not 100% accurate and has a small error rate for inserting the wrong base of about 1 error in 10^{10} replications per base in mammals []; so on average there are 0.3 errors per duplication, errors that will normally not have consequences. This small error rate when passed through the germline (to one's children/descendants) allows organisms to evolve, adapting to new environments. However, it also means that detrimental events can occur, such as generation of 'rogue' cells in the body that can grow unchecked. With a human body being composed of 3×10^{13} cells, with tissues being turned over continuously, there is significant scope for developing cells which escape normal control processes over the course of a lifetime.

Cancer is a Darwinian process where mutation and selection play a major role in determining the progress of the disease - mutations that give a cell an advantage will be selected for, i.e. given sufficient time they would numerically dominate. This is reasonably well understood for colorectal cancer, [1], where the sequence of mutations that give increasing growth advantages is known, Figure 1. The mutations that give the growth advantage are called *driver mutations*, and the genes they occur in the *driver genes*, since multiple sites in those genes may confer a growth advantage (mutation site is not necessarily unique). Further, as the cancerous mass grows, resources may become limiting, such as oxygen and sugars; hence cells will then compete and the less fit clones die. The tumour environment is thus not static, selection pressure altering through the lifespan of a cancer. In particular this could drive spatial heterogeneity, whilst angiogenesis (development of the blood vasculature for the tumour) and metastasis (secondary tumours are present at distant sites in the body) tend to be significant events in cancer progression with deteriorating prognosis.

There are 3 key observations that pertain to this Darwinian perspective:

Cancer tends to take time to develop. Cancers tend to have a number of key mutations (driver mutations) that cause the disease, these mutations giving the clone a growth advantage, Figure 1. This has lead to the idea that cancer is a multistage process. Early work analysing age of incidence, Figure 2, concluded that 6 mutations are required. This can be derived by treating mutations as independent, giving a Gamma distribution (a sum of iid exponential distributions is Gamma), or arguing that mutations are rare so acquiring n mutations by time t has probability $\propto t^{n-1}$. Recent sequencing studies have allowed driver genes to be detected across a range of cancers suggesting that the number of mutations needed varies with the type of cancer, Figure 3, [2].

Driver mutations hit the same 12 pathways across different cancers. Recent evidence suggests that the catalogue of driver genes is beginning to saturate [2]; i.e. despite more cancers being sequenced the same driver genes apeear to be important. This indicates that escape of normal growth control processes can only be acheived in a set number of ways. There are 125 driver genes[1] that fall into 12 pathways, [2], that in turn can be classified into 3 cellular process:

[1]Different detection methods, and thresholds (to control false positives) will give different numbers, but robust methods should all give similar results.

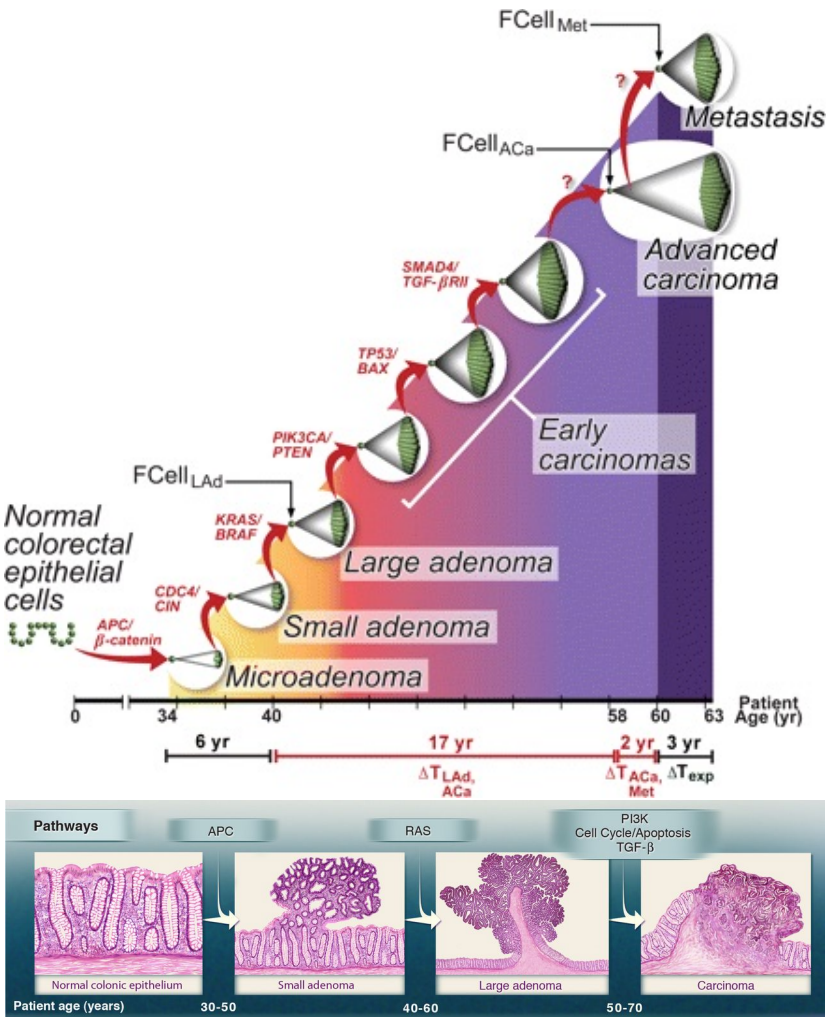

Figure 1: Evolution in colorectal tumourigenesis. **A.** Schematic for the evolution of a lethal cancer over a 63 year timescale. Each cell-filled cone represents one or more clonal expansions, cone gradient representing clone growth rate. Founder cells are indicated (FCell). Driver genes that are mutated for each transition are shown (red). Estimates for the times required for the evolution of each type are shown. The last two clonal expansions, denoted by question marks, that are not associated with any known genetic alterations. Reproduced from [1]. **B.** Progression, with associated mutated pathway, shown as cross sections of typical tissue structure. Reproduced from [2].

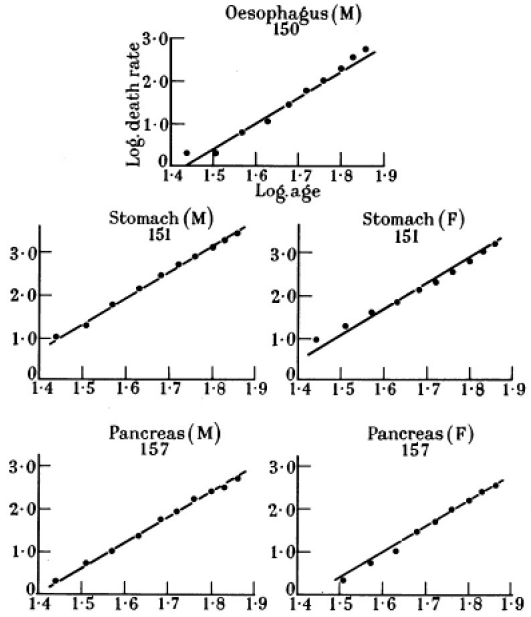

Figure 2: Death rate per million from various cancers against age group (log-log plot). M male, F female. Gradients are around 6. Reproduced from [3]; see this article for additional cancers.

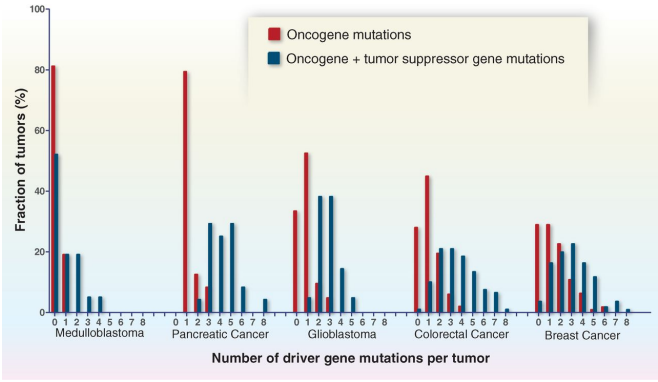

Figure 3: Distribution of driver gene mutations in five tumor types. Driver genes are separated into oncogenes, genes that promote cancer and are thus activated in cancer, and tumour suppressor genes (TGS), genes that are deleted or their activity reduced in cancers. Reproduced from [2], where source data can be found.

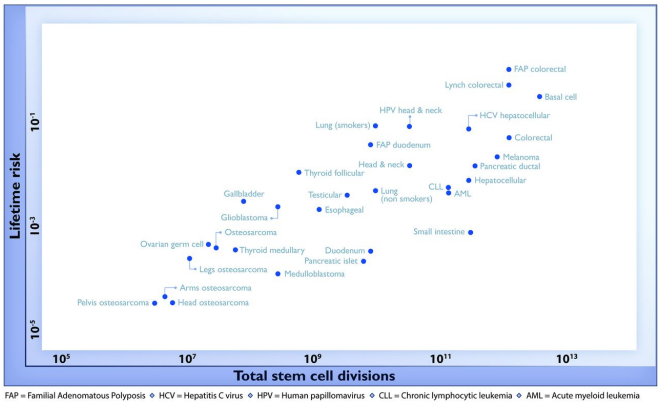

Figure 4: The relationship between the number of stem cell divisions in the lifetime of a given tissue and the lifetime risk of cancer in that tissue. Reproduced from [4]

- **Cell fate**. Most cells in the body are not capable of continual replication. This distinguishes stem cells, which under appropriate conditions will replicate indefinitely, and thus are capable of self-renewal, from the majority of cells that are either fully, or partially differentiated, i.e. on the route towards specialisation of cellular function. The latter cells have limited replication cabalities and eventually die. This replication/differentiation balance is clearly important for cancer; many of the driver mutations abrogate this balance, favouring cell division. Cell fate is also governed by epigenetic modifications, modifications that change the packing and access of the genome; some cancers have mutations in the epigenetic modification apparatus.

- **Cell survival**. Normal cells have quite strigent environment requirements for growth, and in absence of appropriate survival compounds they trigger cell death (apoptosis). Given that the tumour environment will be highly competitive and lacks the normal tissue structure, cancer cells that manage to proliferate under harsh conditions (e.g. limited glucose, growth factors) will have a selective advantage.

- **Genome maintenance**. The genome is normally stable from one cell generation to the next, acheived through multiple checks on genome integrity. Cancer cells often lose these processes, and thus tolerate the accumulation of errors in the genome. This in turn can accelerate mutation; the genome repair processes are thus also often mutated in cancers.

An additional 13 genes can be added that are not point mutated (as the 125 driver genes above) but are frequently amplified in translocations, amplifications or lost in large scale delations, [2], giving 138 'driver' genes. There are thus a large number of targets (genes, sites in the gene sequence where mutation alters function) that could affect the cell growth rate and provide a selective advantage.

Cancer risk is determined by the tissue replication rate. [4]. Risk of cancer is a combination of environmental factors (e.g. smoking), inherited genetic variation (5-10% of cancers have a heritable component) and chance, through accumulation of random mutations. The latter suggests that the number of divisions is a determinant. Since most cells are partially or fully differentiated, these are unlikely to give rise to tumours (having limited replication capacity), stem cells (those with self-renewing capacity) are more likely to give rise to tumours. Cancer risk in a tissue in fact correlates with the replication rate of stem cells in that tissue, Figure 4. This is likely a key factor in the vast differences in the accumulation of mutations across different cancers, Figure 5.

Environmental factors are however dominant in certain types of cancers; smoking is a significant risk factor increasing the mutation burden in a variety of tissues [6]. Thus, cancer is a heterogeneous grouping, with common mechanisms of mutation, selection and escape from growth control processes, but significant differences across cancer type.

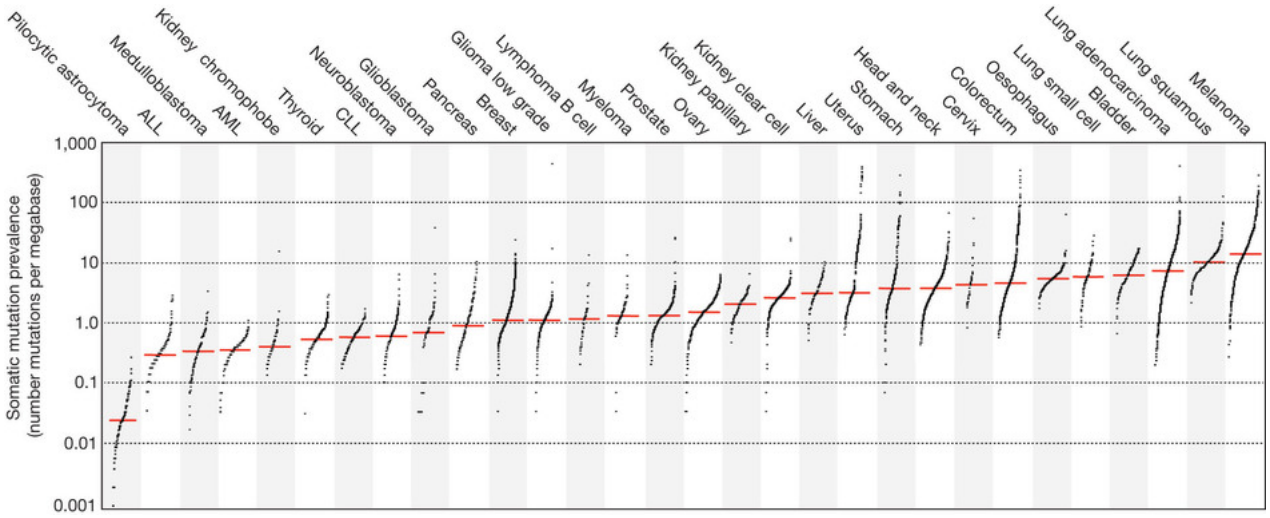

Figure 5: The number of mutations observed per megabase across different human cancers. Reproduced from [5]. ITSELF FROM AN EARLIER PAPER?

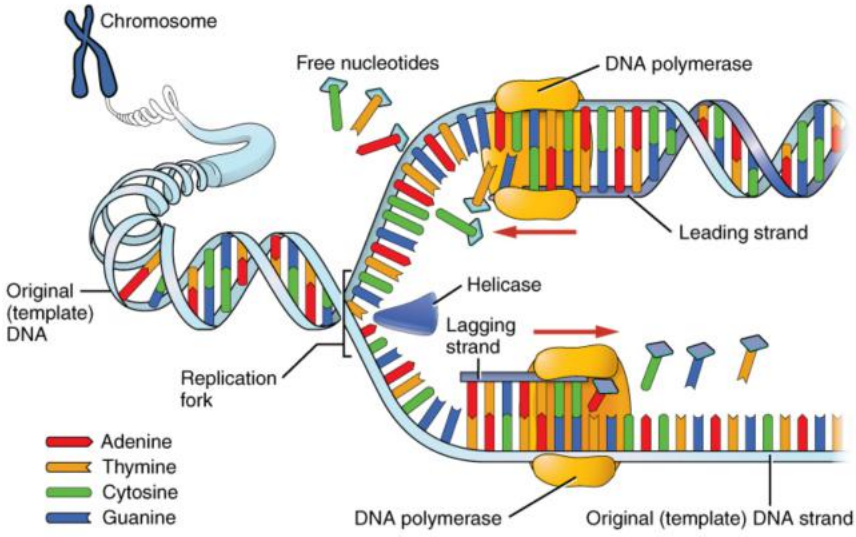

Figure 6: Copying of DNA involves unbinding (unzipping) of the double helix to give two (complementary) templates to copy. The duplication machinary of cells can in fact only synthesise 5' to 3'; hence the complex staggering in one of the copies. Reproduced from AboutBiology http://biology.about.com/od/cellularprocesses/ss/DNA-Replication.htm#step1

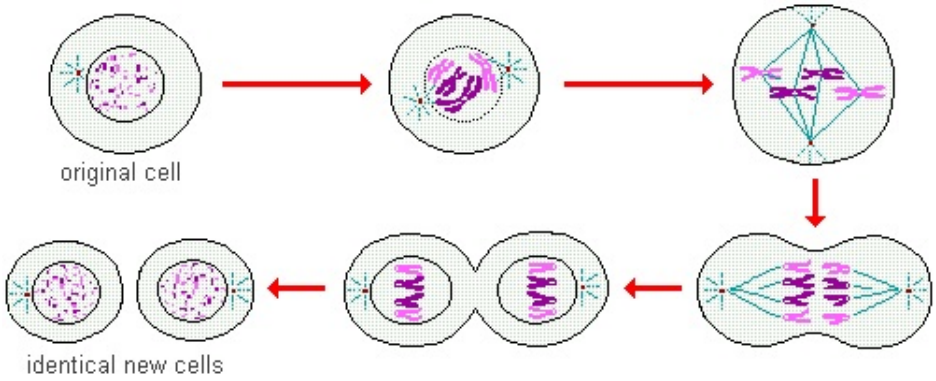

Figure 7: Duplicated chromosomes (held together) then go through a sorting process to ensure each daughter cell receives a copy of all the chromosomes. This is acheived by forming the spindle, a macroscopic molecular machine made of microtubule fibres that emanate from poles at either end of the cell and catch the chromosomes. The duplicated chromosomes are then separated when all chromosome pairs are captured. From http://anthro.palomar.edu/biobasis/bio_2.htm.

2 Basic biology

2.1 The human genome

The blue-print of life is DNA, a sequence of 4 letters, A, T, C, G, that encode all the genes. The human genome comprises 23 sequences (chromosomes) of total length 10^{11}. DNA is double standed with complementary pairing A-T, C-G, forming a double helix, each strand having a direction, i.e. the ends are labelled 3' and 5'. Replication of the DNA then occurs through unbinding giving 2 strands, synthesis of complementary strands then giving near identical double stranded sequences, Figure 6. Replication is not perfect, with copying errors occuring with a frequency of about 1 error in 10^{10}. This is far more accurate than occurs in bacteria due to sophisticated error correction processes.

Chromosomes are then separated mechanically to the two daughter cells, Figure 7. This also has a small error rate, with daughter cells occasionally inheriting incomplete or duplicated chromosomes. This is often used as a detection method for cancerous cells, i.e. whether the karyotype looks normal Figure 8.

If stretched out, human DNA is about 4 meters in length - this is packed inside a cell nucleus that is about a micron in diameter. This involves incredible packing and organisation. Chromosomes are thus only 1/2 DNA [], the rest is accessary proteins responsible for DNA maintenance, proteins that pack DNA and regulatory proteins. This chromosome packing is dynamic, the degree of packing locally to a gene in fact regulating the access to that gene, and thus gene expression. This alteration of gene expression through packing, and modification of proteins around the gene, is called *epigenetics*.

3 Mutations in cancer

The genome with 23 pairs of chromosomes is a means to package the genes, the working component of the genome. Thus, rearrangements of the genes can occur that retain viability, rearrangments being common in cancer making the cancer genome hard to interpret, and its evolution from the original host difficult to unravel. The mutation processes that are observed to occur in cancer are:

- Point mutations (or single base substitutions (SBS) or single nucleotide variants (SNV)). This occurs through copying errors at a rate of 1 error in 10^{-10} bases [7], where the incorrect base is inserted in the copied sequence. This type of mutation is the main focus of analysis, since it is the most well understood and can be modelled using branching processes, see Chapter 2. The genetic code translating the 4 nucleotide bases to the 20 amino acids (protein sequence)

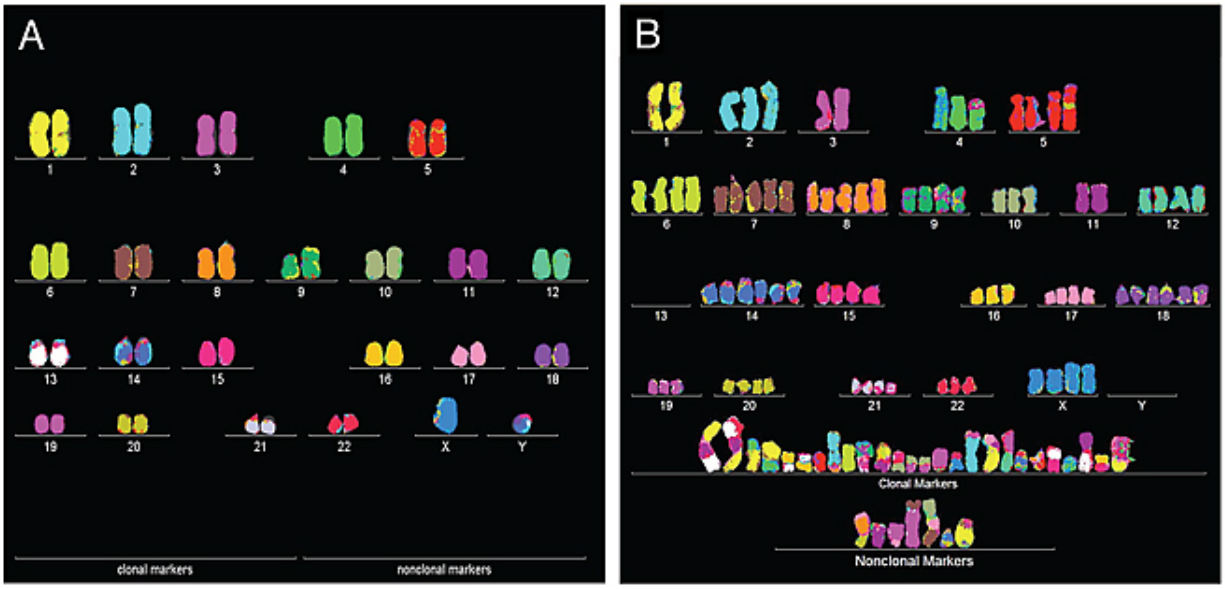

Figure 8: Staining chromosomes with different dyes highlights the orderly nature of the normal human karyotype (left), that is, humans have precisely two copies of each chromosome with no leftovers. A bladder cancer cell (right) has extra copies of some chromosomes, a few missing normal chromsomes, and a lot of hybrid or marker chromosomes, which characterize cancer cells.

is redundant and also encodes a STOP codon, terminating the translation into a protein sequence. This means that some mutations do not alter the protein sequence, so are essentially silent mutations (dont affect protein function), giving rise to **synonymous** (silent) and **non-synonymous** mutations (changing the protein sequence and thus potentially its function). The mutated DNA sequence may also generate nonsence proteins through, for instance, premature termination of the protein through mutation to a STOP codon (nonsence mutations). Only some of the mutations will in fact affect the growth rate of the clone; these are the driver mutations that incur a selective advantage, whilst all others are **passenger mutations**, going along for the ride.

- Indels - insertions and deletions. These are insertions or deletions of small parts of the geneome that can result in a frameshift of the downstream gene sequence if the insertion, or deletion length is not divisible by 3.
- Abnormal numbers of chromosomes. This arises through loss or gain of additional chromosomes during cell replication, i.e. results from errors in segregation.
- Translocations, gene amplifications, gene duplications, gene deletions, Figure 9. Large parts of one chromosome can be moved to another, recopied (duplication, amplification) or deleted; such changes are quantified as copy number aberrrations (CNA). This can result in fusion of genes as the translocations can occur part through a gene. Such large changes can lead to large jumps in the clones' selective advantage as a large number of genes are affected, and thus these alteration can give rise to jumps in evolution.

Evolution is thus a combination of SNVs and more exteme mutation events, the latter allowing evolutionary jumps, the cell behaviour undergoing drastic changes under such mutations. Under SNVs alone, evolution can be viewed as a sequence of small alterations in cell function. Since SNVs are the dominant mutation type in cancer, Figure 9, this is often taken as a justification for ther analysis alone. This certainly simplifies the modelling.

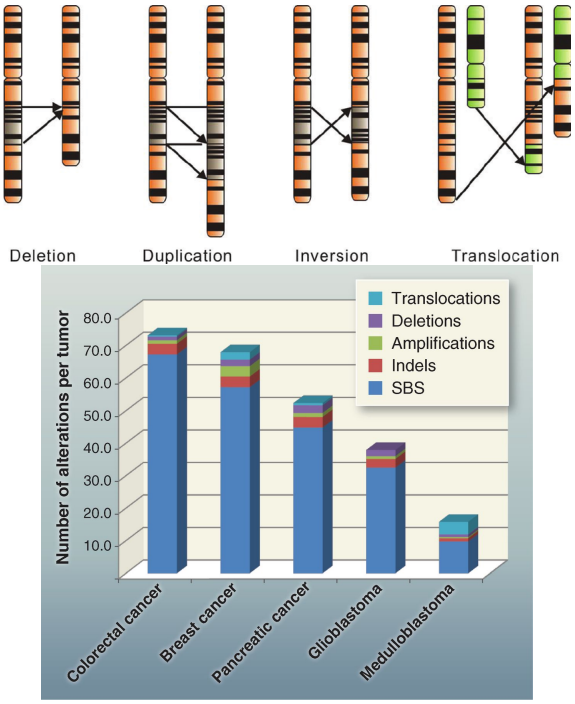

Figure 9: Types of mutations affecting protein-coding genes in selected tumors. **A** Schematic showing 4 of the mutation events that can affect whole genes, reproduced from []. **B** Average number and types of genomic alterations per tumor, including single-base substitutions (SBS), small insertions and deletions (indels), amplifications, and homozygous deletions, as determined by genome-wide sequencing studies. For colorectal, breast, and pancreatic ductal cancer, and medulloblastomas, translocations are also included. Reproduced from [2], where source data can be found.

In addition, other processes that can alter cell function are epigenetic changes (a change in the packing of the DNA, eg through a change in DNA methylation) and an increase in the mutation rate, through mutation in the proteins that duplicate DNA or correct copying errors. Extreme mutation phenotypes are also observed [] - kataegis, chromothripsis, chromoplexy.

4 Challenges

• How does cancer evolve through a sequence of mutations? What are the key drivers and their effects?

Determining how cancer arises through a sequence of mutations, and the growth/survival advantages these mutations endow, is a major question in cancer research. This is a hard question because of a number of factors: (i) there are a large number of ways the genome can be mutated (point mutation, indels, chromosome rearrangements, chromosome number alterations), Figure 9, (ii) the driver mutations are hidden in a larger number of passenger (neutral) mutations, (iii) mutation to function is notorious difficult, and there are typically a large number of alternative pathways to the same cancer type.

• How important is cancer genome heterogeneity, e.g. in progression and treatment.

Tumours evolve as a mixture of clones competing in the tissue environment. Heterogeneity is a clear problem in therapy and there are correlates with poor prognosis. Thus, a leading challenge is how to design optimal therapies for a given cancer.

• What is the optimal treatment strategy for a given cancer (genome)?

The best treatment is influenced by multiple factors - the genome, the maturity of the cancer, whilst other factors such as time of delivery of therapy (chronotherapy) also come into play.

• **How is cancer data best analysed (what sorts of techniques and models), given the complexity of cancer and limitations of data acquisition?**

This is now a major theme of big data, given the massive amount of sequencing data that is available.

References

[1] Sin Jones, Wei-dong Chen, Giovanni Parmigiani, Frank Diehl, Niko Beerenwinkel, Tibor Antal, Arne Traulsen, Martin A. Nowak, Christopher Siegel, Victor E. Velculescu, Kenneth W. Kinzler, Bert Vogelstein, Joseph Willis, and Sanford D. Markowitz. Comparative lesion sequencing provides insights into tumor evolution. *Proceedings of the National Academy of Sciences*, 105(11):4283–4288, 2008.

[2] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[3] Armitage P and Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer*, 8(1):1–12, 1954.

[4] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[5] LudmilB. Alexandrov, others, PeterJ. Campbell, and MichaelR. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[6] Ludmil B. Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J. Campbell, Paolo Vineis, David H. Phillips, and Michael R. Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 2016.

[7] Thomas A. Kunkel, , and Katarzyna Bebenek. Dna replication fidelity. *Annual Review of Biochemistry*, 69(1):497–529, 2000. PMID: 10966467.