Chapter 3: Cancer mutation processes

Nigel J. Burroughs

January 25, 2017

1 Introduction

As discussed in chapter 1, cancer is caused by mutation of genes in key pathways, mutations that give a selective growth advantage, in particular escape from the normal checks on unlimited growth. These driver genes can be classified into 3 broad classes of genes

1. Oncogenes. Genes which promote cancer and are thus activated in cancers. These genes tend to have mutations at a small number of specific locations in the gene, Figure 1, sites that enhance the genes activity.

2. Tumour suppressor genes. These genes typically police abnormal behaviour, thus their inactivation in cancer removes a level of control. These genes typically undergo inactivation mutations, missence (terminating transcription) or truncation, Figure 1. Inactivation can typically be achieved through mutation at a larger number of sites than activation.

3. Stability genes. Damage repair mechanisms. Inactivation of these genes will increase the mutation rate.

A major area of cancer research involves identifying these mutations and the associated genes/pathways, discriminating them from the larger number of background mutations (passenger/neutral mutations). The data that is widely available to do this is sequence data, samples from a range of tumour types, e.g. the Cancer Genome Atlas (TCGA) and the Cancer Genome Project. A common approach is to identify genes that are mutated at higher frequency — the logic is that mutations in these genes (at appropriate sites) are under positive selection and so will be mutated at a higher frequency in the samples, i.e. significantly above the background rate. However, the difficulty in identifying these driver mutations is the fact that cancers have different mutation signatures, and different mutation rates.

In the following, in section 2, we examine how mutation signatures are determined for different cancers. Secondly, in section 4 we examine how driver mutations are detected, illustrated by a basic Binomial model and discuss more advanced algorithms that take into account more varaibles. Thirdly, we present an analysis of mutation accumulation, agreeing with a predominantly neutral evolution model. Finally, in section 6 we examine pathways, and the fact that multiple mutations within the same pathway in a given patient are rare.

2 Profiling mutation signatures

So far we have treated mutation as a base independent process, i.e. there is a single rate for mutation at any site. This is a simplification, in particular the base pairings A-T, C-G have different binding energies which affects their predisposition to pair with the wrong base, whilst carcinogens may promote a certain type of base mutation.

This has been analysed through examining mutation types across a range of cancers, detecting (lower dimensional) factors, or patterns in the data [2, 3]. To set the scene consider the possible

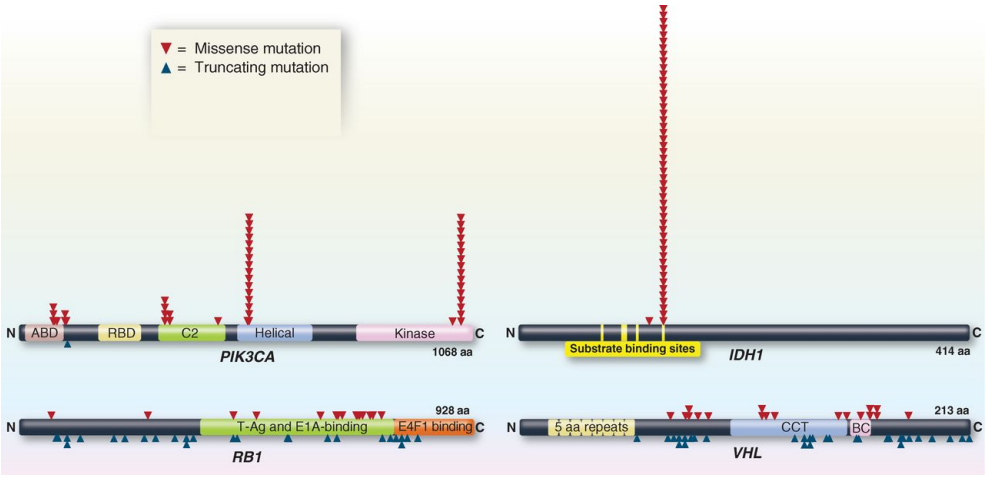

Figure 1: Distribution of mutations in two oncogenes (PIK3CA and IDH1) and two tumor suppressor genes (RB1 and VHL). The distribution of missense mutations (red arrowheads) and truncating mutations (blue arrowheads) are shown. For PIK3CA and IDH1, only 50 randomly selected mutations obtained from the COSMIC database are shown. For RB1 and VHL, all mutations recorded in COSMIC are plotted. aa, amino acids. Reproduced from [1].

mutations of the 4 bases; there are 6 possible mutations of a base pairing (recall DNA is composed of complementary pairs):

$$\begin{aligned} C:G \to A:T, \\ C:G \to G:C, \\ C:G \to T:A, \\ T:A \to A:T, \\ T:A \to C:G, \\ T:A \to G:C \end{aligned} \tag{1}$$

For example, $C:G \to A:T$ means C is changed to A in one strand, and G to T in the complementary strand.

By focussing on the pair, mutation is assumed to be independent of strand. More complex models allow for neighbouring bases (96 signatures = 4 left neighbours $\times$ 4 right neighbours $\times$ 6 types), and positive/negative strand of DNA (192 signatures). The number of mutation events is denoted K.

A mutation signature p is then a probability vector on these types, p_j probability of mutation j (eg $C:G \to G:C$), with $\sum_{j=1}^{K} p_j = 1$. Define Ξ as the set of mutation processes that are active, e.g. sunlight, cigarette smoke, each having a mutation signature, $P_{\cdot k}, k = 1...N$ (dimension $dim(\Xi) = N$), with P_{jk} the probability of mutation j occuring for signature k. Then if a genome is exposed to mutation process k with weight e_k, the expected number of mutations of type j is $\sum_{k=1}^{N} P_{jk} e_k$. (The weight e_k also incorporates the counts for the sites where mutation of type k can occur, thus the analysis below must be done on the same sequence to allow this simplification).

For multiple genomes we can count the number of mutations of each type. Define M_{jg} as the number of mutations of type j for genome g (or part genome). Then we have the factor decomposition[1]

$$M = PE + \Gamma, \tag{2}$$

where E_{kg} is the exposure (or load) of genome g to mutation process k and Γ is noise (measurement errors and noise from other (mutational) processes). The factor model describes a constrained relationship between the observed mutation frequencies and the N possible mutation processes. It

[1]Factor models may be defined with $\mathbf{E}[E] = 0$ in which case we have to accomodate a mean expression.

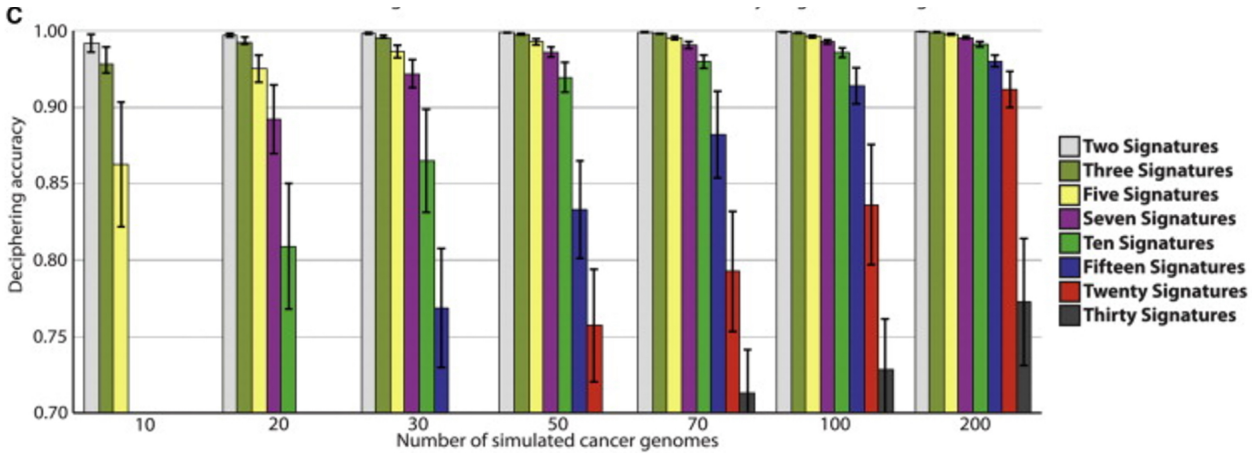

Figure 2: The number of signatures that can be detected increases with the number of simulated genomes. Reproduced from [2].

is unusual in that the data (M, size $K \times G$) is on the left hand side, and both P (size $K \times N$) and E (size $N \times G$) are unknown. Effectively the data M is being decomposed into a lower number of patterns, or factors P. Thus, N must be small, typically less than both the number of genomes G and number of mutation types K. To further constrain the problem the weights are often taken as orthogonal, $EE^T = I$. There is an identifiability problem in that if P, E is a solution then for any orthogonal transformation U, PU^T, UE is also a solution. Model selection can be used to ascertain the number of signatutes N.

There are a number of techniques to solve (2), from Bayesian methods to nonnegative matrix factorization (NMF) [4]. An analysis using NMR (on Euclidean metric $|M-PE|^2$, so noise Γ assumed Gaussian with equal variance) demonstrated the relationship between the number of signatures that are detectable and the quantity of data, Figure 2, and as the number of mutations increases accuracy increases, [2].

Applied to cancer data, this analysis detects mutational signatures; in breast cancer using the 196 possible mutation types (mutation rate depends on base pair and 2 immediate neighbours) 4 signatures are detected, Figure 3. Signature 3 is likely the base mutation rate, whilst the other 3 due to exposure to cancinogens.

An analysis across multiple cancer genomes demonstrates that different cancers have different profiles of mutation (different signature weights), Figure 4. The loads, or weights E_{kg} can be analysed for trends by type of cancer that genome g belongs to, and other variables such as strand, expression level of the associated gene etc. This demonstrated that there is only one signature that has a correlation with the age of detection of the cancer suggesting that this signature represents accumulation of mutations with age. The other signatures, that don't correlate with age, are thus influenced by other factors, eg concentration of exposure to a cancinogen.

This demonstrates that different cancers have different mutation signatures, e.g. smoking is a leading cause of lung cancer and is associated with a particular signature characterised by exposure to smoke.

3 The genetic code: synonymous and non-synonymous mutations

The genome encodes the genes through the 4 letter alphabet (A,T,C,G). However, these need to be *translated* into proteins; proteins are the functional entities, eg mechanical motors, enzymes. Proteins are encoded by 20 letters (the amino acids), so to encode these the DNA sequence is blocked into units of 3 (the codons). A gene sequence starts with the START codon (AUG, Methionine), and

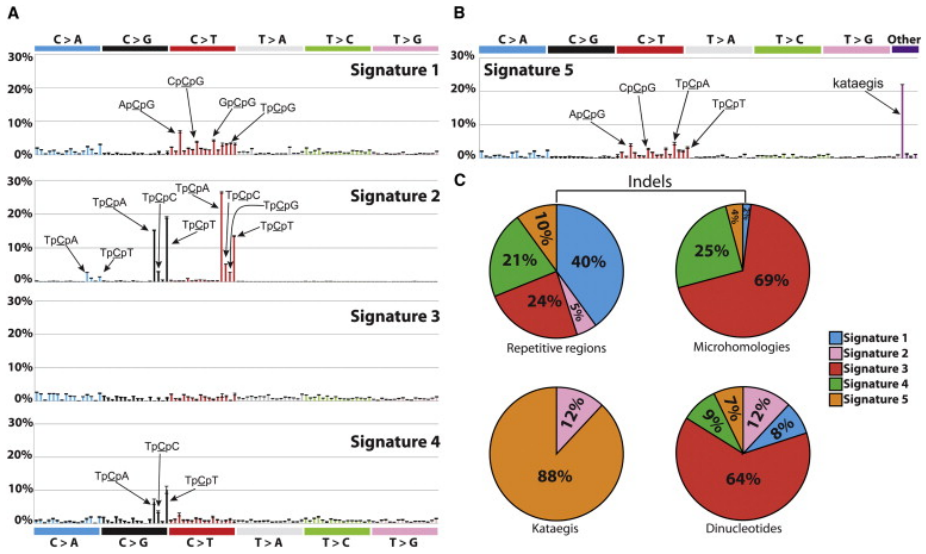

Figure 3: Mutational signatures found in human breast cancer. (A) Four mutational signatures deciphered from the base substitutions (including their immediate 3 and 5 neighbours) identified in 21 breast cancer genomes. Each signature is displayed according to the 96 substitution classification defined by the substitution class and a sites two immediate (3′ and 5′) neighbours. The probability bars for the six types of substitutions are displayed in different colours. The mutation types are on the horizontal axes, whereas vertical axes depict the percentage of mutations attributed to a specific mutation type. (B) A fifth mutational signature identified when kataegis (localised hypermutation), dinucleotide substitutions, and indels at microhomologies and at mono or polynucleotide repeats are added as mutation types. (C) Total contributions of mutations of the five signatures for kataegis, dinucleotide substitutions, and indels in the 21 breast cancer genomes. Reproduced from [2].

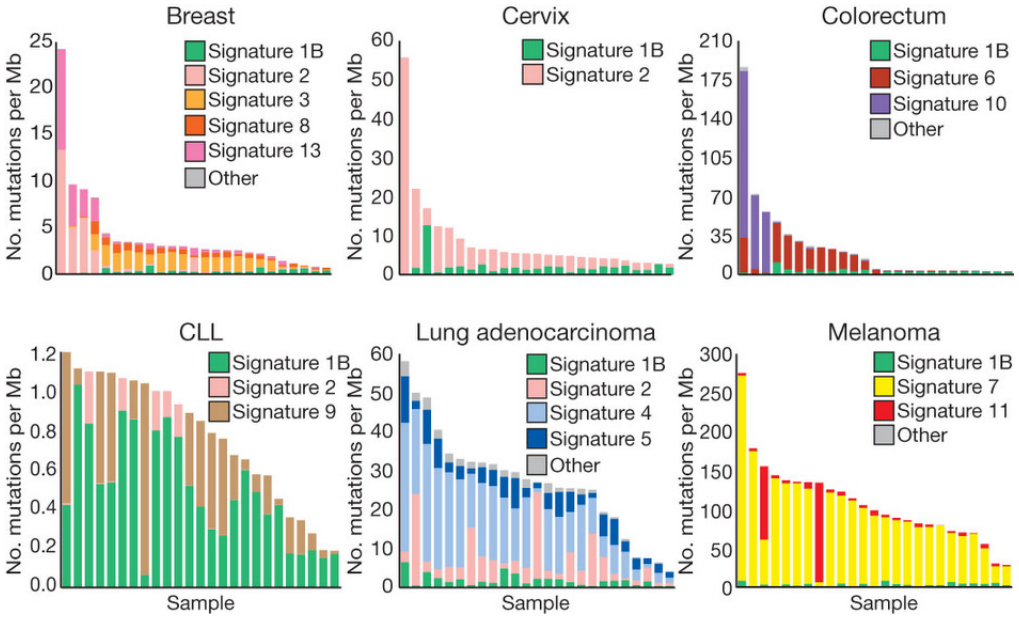

Figure 4: The contributions of mutational signatures to individual cancers of selected cancer types generated from an analysis of 30 cancers. Each bar represents a typical selected sample from the respective cancer type and the vertical axis denotes the number of mutations per megabase for that sample. The proportion of different mutation signatures is shown in colour. Reproduced from [3]; see this article for additional cancers.

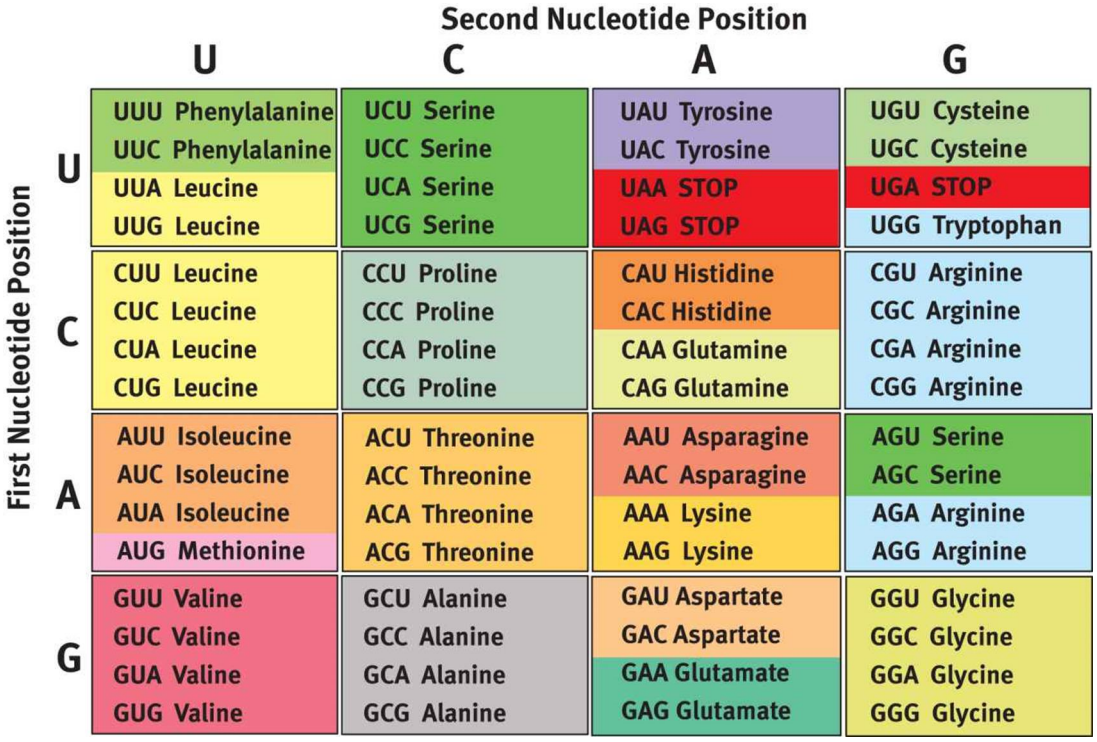

Figure 5: The genetic code enocoding amino acids by DNA.

terminates with a STOP codon. Since there are 64 combinations of 4 letters of length 3, thus the genetic code (DNA to amino acids) is redundant, Figure 5.

Mutations of the DNA sequence may now have different effects on the proein. Three types of mutation are,

- Silent/synonymous. This mutation does not alter the amino acid, i.e. the protein sequence is identical. Typically, mutation of the third base is silent.
- Nonsence. The amino acid is altered.
- Missence. The protein sequence terminates, i.e. the codon is mutated from an amino acid to a STOP codon.

A non-synonymous mutation is a mutation that is not synonymous.

4 Detecting driver mutations

There are a few methods to identify driver mutations, the simplest being a Binomial test to identify genes with increased mutation frequency. There are numerous factors that can improve the power of the test if taken into account, including, [5],

- Mutation signatures. Mutation can be caused by a range of factors, from a background rate to exposure to carcinogens. Thus, cancers have specific mutational profiles.
- Mutation patterns. Oncogenes and tumour suppressor genes are subject to different types of mutation.
- Replication timing and expression of the gene. The mutation rate decreases with gene expression [**?**] and increases the later the gene is replicated in cell division, because of the reduced time that the replication machinary has to correct errors [**?**, **?**].

- Mutation to function. Relating mutations to function, and thus the potential of a mutation to affect growth.

4.1 A basic model: frequency tests

Consider a simple model of mutation with a (background) site mutation rate b. Thus, a gene of length L will acquire mutations at rate bL.

To detect driver genes, the idea is to identify genes that are acquiring mutations at a faster rate. If N_j mutations are observed for gene j, we want to determine those genes where N_j is significantly larger than this model predicts.

Under the background model, the number of times gene j is mutated is Binomially distributed $N_j \sim B(G, bL_j)$. Thus, given the background probability b, the probability of observing a frequency as high as N_j can be computed, and the p-value determined, denoted $p_j = P(X \geq N_j | X \sim B(G, bL_j))$. Using a suitable threshold (e.g. 1%), the genes that have a significantly higher frequency of mutation can be determined. Since we are doing multiple tests, a multiple testing proceedure should also be used, e.g. control the false discovery rate (FDR=FP/(FP+TP), false positives FP, true positives TP) using the standard Benjamini and Hochberg False Discovery Rate procedure, or Bonferroni correction.

This model can easily be improved on, allowing for different mutation rates per base. The point mutation rate (base substitution rate) is also dependent on the sequence, a large number of models being defined in the phylogenetics literature, e.g. Jukes Cantor (all bases substitutions equally weighted), Kimura model (transitions and transversions have different rates). Other models allow for neighbouring bases.

4.2 Estimating the background mutation rate

The above test requires a good estimate of the background mutation rate b. This is in fact hard to estimate because it is infact sample specific. For example, a cancer that mutates the error correcting mechanisms will increase the background error rate.

To estimate the background rate, mutation events that are neutral need to be identified. This is often taken as a combination of silent mutations (mutations that do not affect the protein sequence), non-coding mutations and mutations in genes that are known not to affect fitness. However, silent mutations are not necessarily neutral because translation efficiency depends on the codon, whilst access to the gene by the transcription rate (copying DNA gene sequence into RNA) may be sequence dependent, e.g. through chromosome packing. The background rate of other types of mutations such as indels, gene duplications are even harder to estimate.

The background rate of point mutation is also patient and cancer specific, Figure 6

4.3 Improved driver detection using context and frequency

MutSigCV. [6]. A popular algorithm accounts for these and other factors by stratifying the data - mutation type/signature and patient specific rates, and estimating the background by pooling across similar genes. The latter is performed by using a Euclidean distance in covariates, specifically gene expression and time of replication. These two variables explain most of the variation of the mutation rate across the genome, Figure 7.

Applying MutSigCV to lung cancer reduced the number of significantly mutated genes from 415 to 11 - most of the genes in this short list have been previously reported to be mutated in squamous cell lung cancer (TP53, KEAP1, NFE2L2, CDKN2A, PIK3CA, PTEN, RB111,16) or other tumor types (MLL2, NOTCH1, FBXW7), [6].

MADGiC. Algorthims that incorporate additional information such as mutation context have also been developed [5], context refering to the use of scoring methods that estimate how likely that mutation affects gene function. This can be estimated using sequence conservation across related species, e.g. the SIFT algorithm [7].

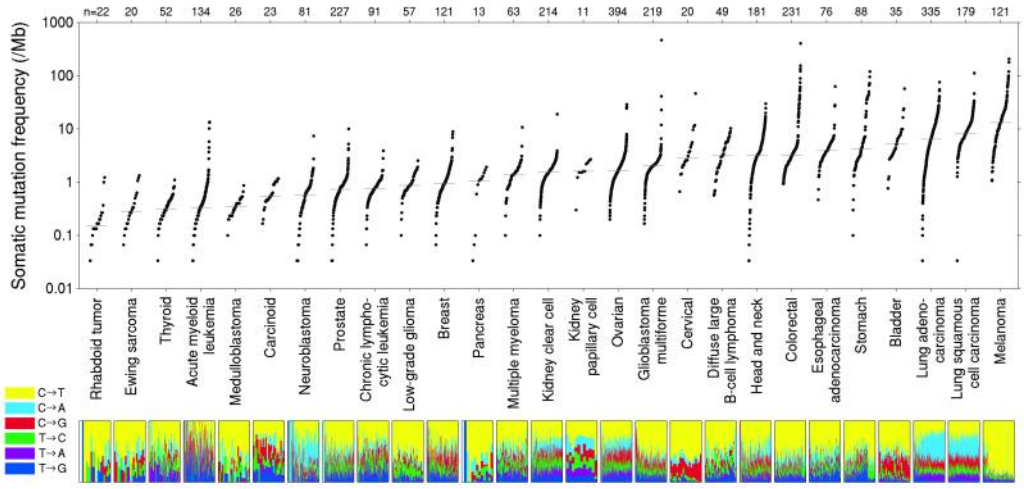

Figure 6: Somatic mutation frequencies observed in exomes (coding) from 3,083 tumor-normal pairs. Each dot corresponds to a tumor-normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumor types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in hematological and pediatric tumors, and the highest (right) in tumors induced by carcinogens such as tobacco smoke and UV light. Mutation frequencies vary more than 1000-fold between lowest and highest mutation rates across cancer and also within several tumor types. The lower panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left. Reproduced from [6]

Methods, such as linear regression to model trends, and basing a test on a model of mutation patterns should be more powerful than pooling similar genes but have not been done as far as I am aware.

5 Selection and Neutral evolution of cancer (from data)

Under the lineage or clonal model of cancer, a mutation that appears at time t when the population size is $N(t)$ will, under neutral evolution, retain its relative frequency (up to fluctuations). Thus, the relative frequency $f = 1/(\pi N(t))$ (c.f. branching processes), where π is the average copy number per cell. This provides a map between the relative frequency of an allelle (mutant gene) and time, which under neutral evolution model can be used to map the accumulation of mutations, [8].

Define $M(t)$ as the number of mutations in the population at time t. Then, the rate of mutations accumulating in the population is given by,

$$\frac{dM(t)}{dt} = \mu\pi\lambda(t)N(t)$$

where μ is the mutation probability and λ the growth rate, possibly time dependent. However, not all cells generate a surviving lineage (see for example Chapter 2, branching processes), so $\lambda(t)$ is the effective growth rate giving $N(t) = N(0)exp\int_0^t \lambda(t)dt$. There will be a small population of mutations in cells that fail to generate a surviving lineage that we ignore. Thus,

$$M(t) = \mu\pi \int_{t_0}^{t} \lambda(t)N(t)dt$$

the mutations that have accumulated between $[t_0, t]$.

For the special case of exponential growth, λ constant, we have the Luria-Delbruck model (accumulation of mutations in bacteria, [9])

$$N(t) = N_0 e^{\lambda t}, M(t) = \mu\pi N_0 \left(e^{\lambda t} - e^{\lambda t_0}\right)$$

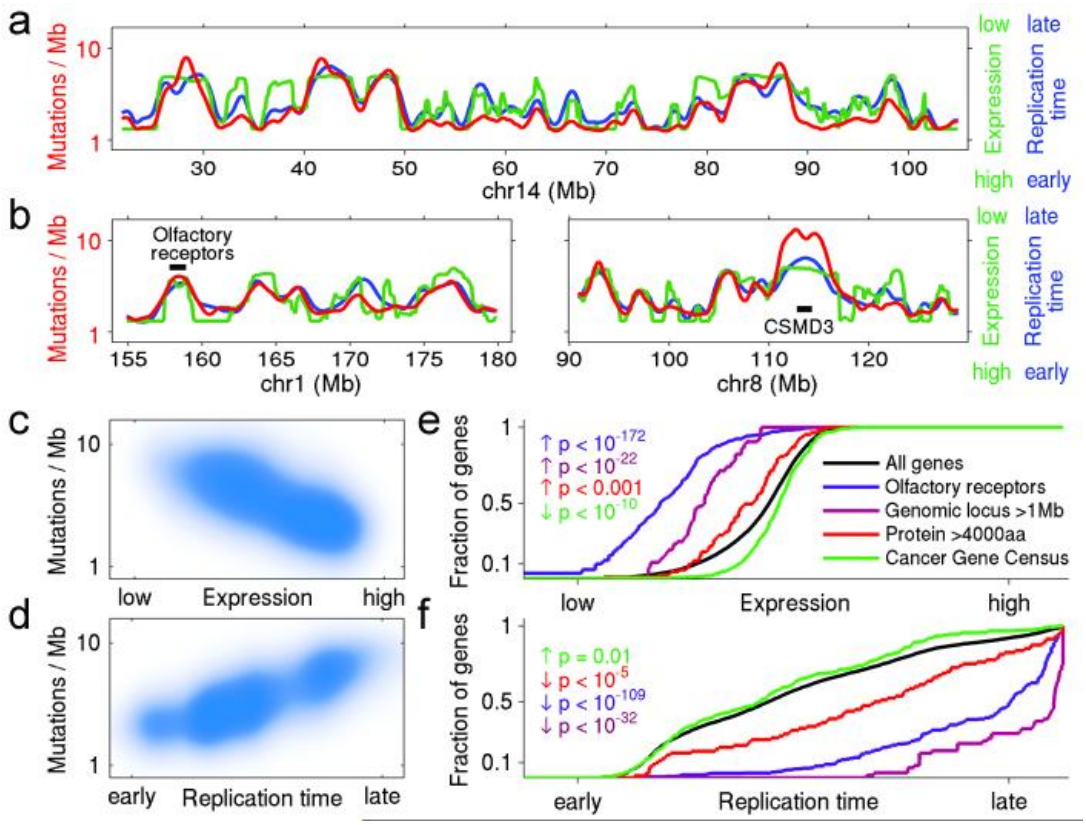

Figure 7: Mutation rate varies widely across the genome and correlates with DNA replication time and expression level. **(a,b)** Mutation rate, replication time, and expression level plotted across selected regions of the genome. Red shows total noncoding mutation rate calculated from whole-genome sequences of 126 samples (excluding exons). Blue shows replication time. Green shows average expression level across 91 cell lines in the Cancer Cell Line Encyclopedia (CCLE), determined by RNA sequencing. (Note that low expression is at the top of the scale and high expression at the bottom, in order to emphasize the mutual correlations with the other variables). Shown are **(a)** entire chromosome 14 and **(b)** portions of chromosomes 1 and 8, with the locations of two specific loci: a cluster of 16 olfactory receptors on chr1 and the gene CSMD3 on chr8. These two loci have very high mutation rates, late replication times, and low expression levels. (The local mutation rate at CSMD3 is even higher than predicted from replication time and expression, suggesting contributions from additional factors, perhaps locally increased DNA breakage: the locus is a known fragile site). **(c,d)** Correlation of mutation rate with expression level and replication time, for all 100 Kb windows across the genome. **(e,f)** Cumulative distribution of various gene families as a function of expression level and replication time. Olfactory receptor genes, genes encoding long proteins (¿4,000aa) and genes spanning large genomic loci (¿1Mb) are significantly enriched towards lower expression and later replication. In contrast, known cancer genes (as listed in the Cancer Gene Census) trend toward slightly higher expression and earlier replication. Reproduced from [6] where additional information can be found.

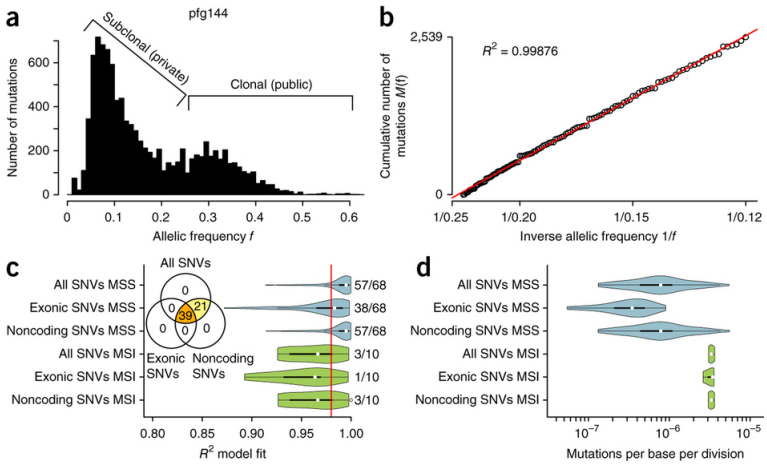

Figure 8: Neutral evolution across the genome of gastric cancers. **a** A large number of coding and noncoding mutations can be identified using whole-genome sequencing. Here the number of mutant alleles is plotted against the frequency of occurence in the data (allelic frequency). **b** All detected mutations precisely accumulate as 1/f over the specified range following the neutral model in this example. Reproduced from [8].

which gives, on substituing for t in favour of f,

$$M(1/f) = \mu\left(\frac{1}{f} - \pi N_0 e^{\lambda t_0}\right)$$

i.e. the accumulated mutations $M(1/f)$ with frequency in $[f, f_{max}]$, $f_{max} = 1/(\pi N_0 e^{\lambda t_0})$. Thus, under the assumptions of exponential growth and neutral evolution (no benefit or cost to mutations) the cumulant $M(1/f)$ is linear in $1/f$.

This linearity is infact borne out by the data for certain cancers [8], Figure 8. This is not surprising as passenger mutations will numerically dominate, which are neutral. The gradient gives an estimate of the mutation proability per division per base μ, Figure 8D, which differs between coding (exonic) and non-coding regions of the genome.

Question: How does this link to the branching process caculation in Ch 2?

6 Mutation exclusion along pathways

There are many reports that genes in the same pathway are typically not mutated in a tumour (mutual exclusion of driver mutations by pathway); this can be interpreted as a consequence of Darwinian selection since the selective advantage of disrupting a pathway has already been acquired by the first mutation, and thus additional mutations are unlikely to be under positive selection (no growth/survival advantage) and may even confer a disadvantage (as most mutations are detrimental).

A likelihood ratio test (for mutation independence) developed in [10] showed a high level of pathway exclusivity and also indicated high levels of co-occurence of mutation amongst different key pathways. Deviations from these were interpreted as tissue specific variations. This test however ignored a number of issues, including that mutation rates differ amongst tumour samples (i.e. samples are not identically distributed), whilst there are observation errors (false positive, false negatives in assignment of whether a gene is mutated).

6.1 Differential mutation rates: DISCOVER

Differential mutation rates are accounted for in [11] in their statistical test (called DISCOVER); their analysis suggests that previous analyses have missed a number of mutually exclusive genes,

whilst co-occurance can be explained by chance alone. The importance of allowing for differential mutation rates is that mutation in a pair of genes is more likely (and thus overlap of mutated gene sets) in tumours with high mutation rates.

Let p_{ij} be the probability of a mutation in gene $i = 1...n$, in tumour $j = 1...m$. If mutation is independent for genes i_1, i_2, then the probability of a mutation in genes i_1 and i_2 is $p_{i_1\,j}p_{i_2\,j}$. The number of tumours that have a mutation in both genes then follows a Poisson-binomial distribution, the distribution for the sum of independent, non-identically distributed Bernoulli random variates with success probabilities $p_1...p_n$. Using this as the null distribution, pairs (sets) of genes that have extreme probabilities can thus be identified.

Let X_{ij} be the indicator variable for a mutation in gene $i = 1..n$, tumour $j = 1..m$; the likelihood then reads (data x, so $x_{ij} = 1$ if gene i in tumour j is mutated),

$$L(p; X = x) = \prod_i \prod_j (x_{ij}p_{ij} + (1 - x_{ij})(1 - p_{ij}))$$

There are nm parameters for nm data points, so the form of p_{ij} needs to be constrained to reduce parameters otherwise the data will be overfitted.

DISCOVER, [11], uses a maximum entropy method, subject to constraints that the total number of mutations per gene is reproduced, and the total number of mutations per tumour. These constraints are identical to those used in the contigency table tests (Fisher exact test), preserving row and column sums. Specifically parameters are estimated from the following optimisation problem:

Estimate $\hat{p}_{ij}$ by maximising the entropy,

$$H(p) = -\sum_i \sum_j (p_{ij} log\, p_{ij} + (1 - p_{ij}) log(1 - p_{ij}))$$

subject to the constraints,

$$\sum_j p_{ij} = \sum_j x_{ij},\ \sum_i p_{ij} = \sum_i x_{ij} \tag{3}$$

The constraints mean that the expected number of mutations of gene i is equal to that observed, $E[X_{i\cdot}] = \sum_j p_{ij} = \sum_j x_{ij}$ and similarly for the expected number of mutations in sample j.

The maximum extropy optimisation occurs in $nm - n - m$ dimensions, which implies that $\hat{p}_{ij}$ is in fact parametrised by $n + m$ parameters. By using Lagrange multipliers $(\mu_i, \lambda_i, i = 1...n)$ the constrained optimisation of H is equivalent to the unconstrained minimisation of $H' = -H + \sum_i \mu_i \left(\sum_j p_{ij} - x_{i\cdot}\right) + \sum_j \lambda_j \left(\sum_i p_{ij} - x_{\cdot j}\right)$ where $x_{\cdot j} = \sum_i x_{ij}$ etc. Taking derivatives gives,

$$\frac{\partial H'}{\partial p_{ij}} = log\, p_{ij} - log\,(1 - p_{ij}) + \mu_i + \lambda_j$$

from which we obtain the parametrisation, [11],

$$\hat{p}_{ij} = \frac{1}{1 + e^{\mu_i + \lambda_j}},$$

which are estimated by substitution into H', and minimising, essentially minimising the deviation from the constraints (3). There is no closed form for the parameter estimates μ_i, λ_j; numerical solutions can be obtained, e.g. using the quasi-Newton method (surface is approximated by a quadratic and the Hessian is estimated from sequential steps).

Given this null model (mutation is independent across all genes), sets of genes with extreme p-values can thus be identified. Different statistics can be used (and have been used in the literature) to analyse deviations from the null model, *coverage* (the number of tumours that have a mutation in at least one of the genes), *exclusivity* (the number of tumours with exactly 1 gene mutated) and *impurity* (the number of tumours with 2 or more genes mutated). These statistics are all Poisson-binomial distributed in this model (Proof). For exclusivity, the number of tumours with exactly

one gene mutated, N, has a null distribution with the success probabilities p_j (the probability p_j of tumour j having exactly 1 gene mutated) given by,

$$p_j = \sum_{i=1}^{n} p_{ij} \prod_{k \neq i} (1 - p_{kj})$$

The null distribution for N can thus be computed and p-values determined for any pair of genes (or set of genes). If the observed N is large, the pair (set) have significantly exclusivity, if N is low the pair (set) have significant co-occurence.

In practice a multiple testing correction needs to be used, e.g. Bonferroni [] and Benjamini-Hochberg []; however in the case of a discrete test statistic these techniques would be conservative and direct enumeration of the tail is used following [**?**] for contingency tables.

Results: comparison across different methods

There are a number of different methods for analysis of mutual exclusivity, based on contingency tables or likelihood methods. A comparison using simulated data (using the exclusivity model and random model of section 6.2 demonstrates DISCOVER has the highest sensitivity (detects correctly those pairs that are simulated under the exclusivity model) and a false positive rate that is close to that expected[2], Figure 9. Some of these models are discussed in the Supplement of [12].

Results: mutual exclusivity in cancer data

For different classes of tumours, each is analysed separately (stratify analysis into homogeneous groups).

In 3386 tumours across 12 cancer types, analysing 404 mutations (across 374 genes), pairwise co-occurence (for genes not on the same chromosome) was not detected (FDR 1%), whilst exclusivity was detected amongst 181 pairs (FDR 1%), corresponding to 107 genes. The corresponding binomial test (identical mutation rates across tumours) reduced this to 3 pairs. Testing exclusivity amongst functional groups (from the STRING interaction network []) demonstrated these pairs group into functional sets. The most significant groups are given in Figure 10.

6.2 A generative model for exclusivity

NOTE: matrix X is transposed between these papers.

A model for mutually exclusive gene sets (MEGS), with associated likelihood was given in [12]. This used a likelihood ratio test to determine whether, for a given gene set, mutation conformed to the MEGS model or to a random mutation model; their test MEGSA performed well in comparisons Figure 9 with high sensitivity and low false positives.

MEGS set

Consider a set of m genes that conform to MEGS, with coverage γ, i.e. a sample conforms to MEGS with probability γ. Let genes (within MEGS set) have relative mutation rates p_j, $\sum_{j=1}^{m} p_j = 1$. Then for samples conforming to MEGS, one gene is selected according to weights p_j.

Random model

All samples are then subject to random mutation probabilities π_j

Let X be the mutation status, $X_{ij} = 1$ if gene j is mutated in sample i. The likelihood then reads,

$$L(\gamma, p, \pi; X) = \prod_{i=1}^{N} \left((1-\gamma) \prod_{j=1}^{m} \pi_j^{X_{ij}} (1-\pi_j)^{1-X_{ij}} + \gamma \sum_{k=1}^{m} p_k I_{X_{ik}=1} \prod_{j \neq k} \pi_j^{X_{ij}} (1-\pi_j)^{1-X_{ij}} \right) \quad (4)$$

where the first term is the random model (probability $1-\gamma$) and the second is the MEGS (with gene k selected) and additional mutations from the random model.

[2]The confidence level α controls the type I error of the test, i.e. the error of rejecting the null hypothesis when it is in fact true. The tests should have false positive rate (FPR) $\leq \alpha$.

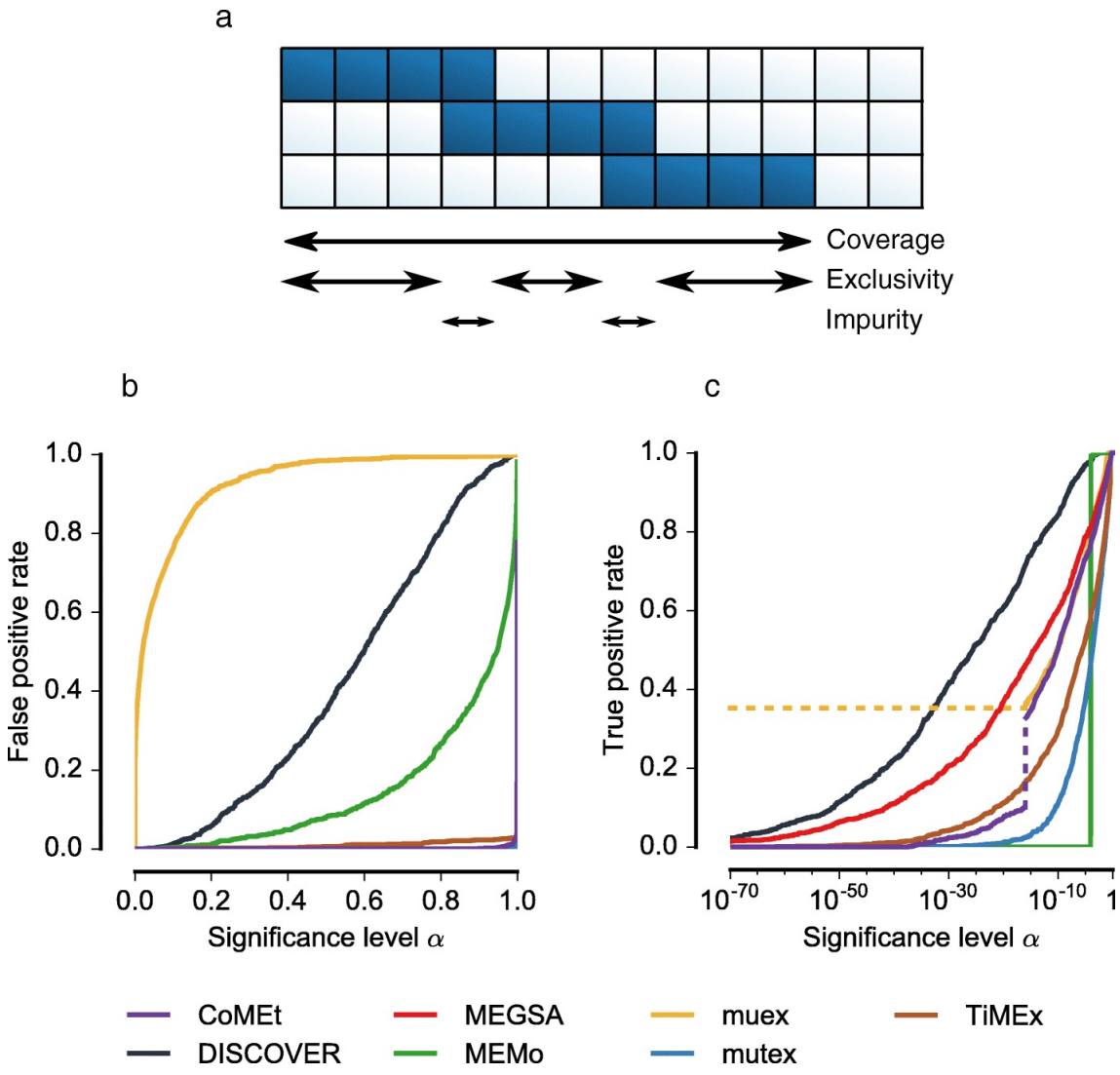

Figure 9: Comparison between different tests for detection of mutual exclusivity of mutation. **a**. Three alternative statistics for measuring the degree of mutual exclusivity within a group of genes-coverage, exclusity and impurity (see text). **b**. Specificity and sensitivity analysis of mutual exclusivity tests. The false positive rate and true positive rate plotted against the confidence level α. Reproduced from [11] where details can be found.

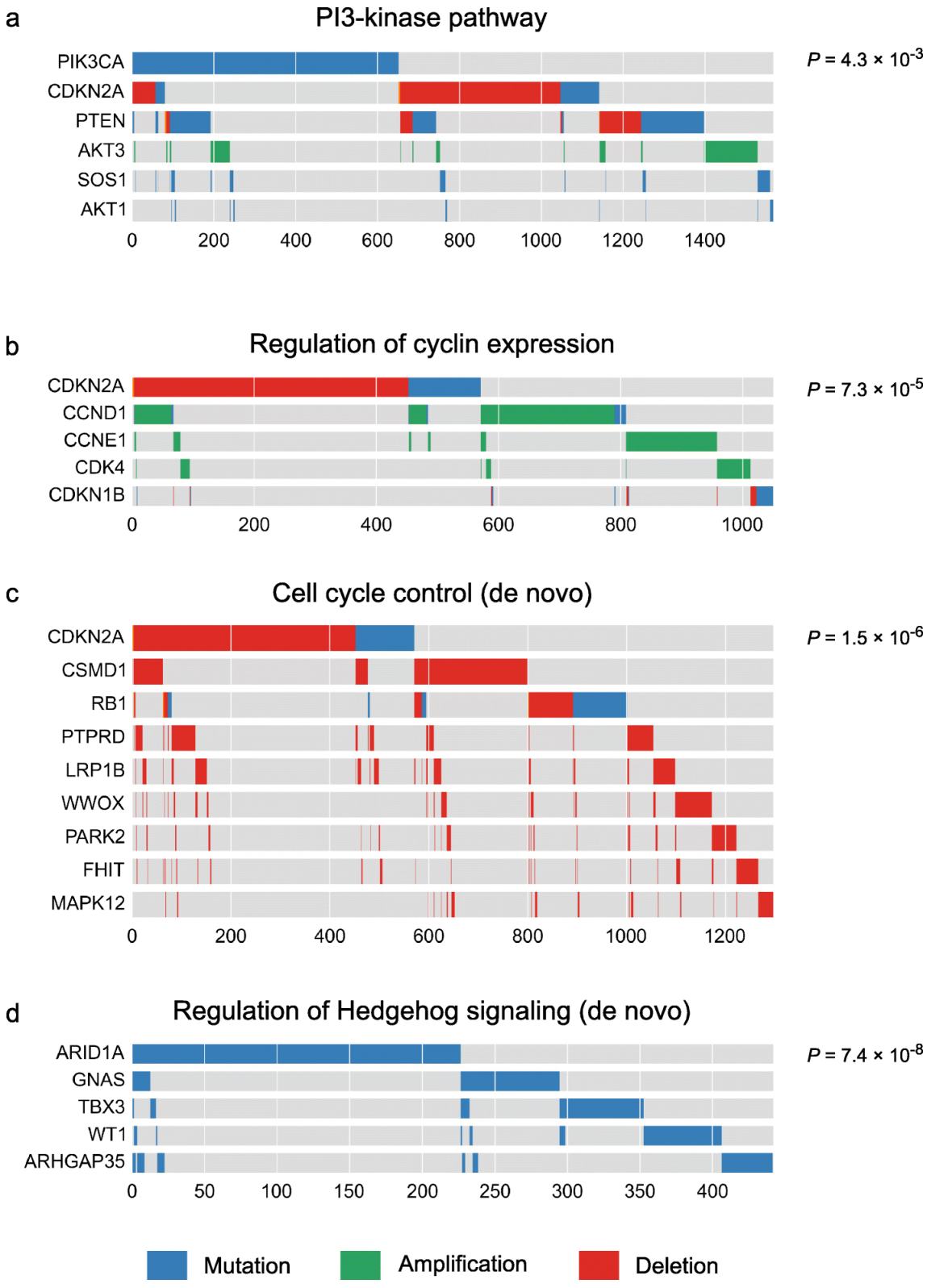

Figure 10: Examples of exclusive gene sets detected using DISCOVER. The p-values were computed using DISCOVERs group-based test. **a.** PI3-kinase, a signalling pathway important in cell division. **b.** Cyclin regulation. Cyclins control progression through the cell cycle. **c.** Cell cycle control. **d.** Hedgehog signalling. A differentiation pathway. Reproduced from [11].

The Likelihood ratio test (LRT) then determines if the MEGS model is preferred (for given gene set, $i = 1..m$), testing the null hypothesis $H_0 : \gamma = 0$ against $H_1 : \gamma > 0$ (nested models). In practice the number of nuisance parameters is reduced by assuming $p \propto \pi$. The log LRT statistic $S = 2\left(logL(\hat{\gamma}_1, \hat{\pi}_1; X) - logL(\gamma = 0, \hat{\pi}_0; X)\right)$ (maximum likelihood estimates $\hat{\gamma}_k, \hat{\pi}_k$ on hypothesis k) is a mixture of chi-squared distributions, $0.5\chi_0^2 + 0.5\chi_1^2$ (a point mass at zero and a 0.5 probability as χ_1^2 corresponding to the degree of freedom changes 0 and 1 respectively).

To test the global null hypothesis - no gene sets are mutually exclusive, [12] use a permutation strategy and reduce the search space over gene sets by testing the most significant gene combinations (based on lower order gene sets, a multiple search path algorithm).

Remark: With a computationally tractable likelihood, methods such as Markov chain Monte Carlo are applicable, and may well be more efficient at sampling the combinatorial space.

6.3 Discussion: Analysis by different groupings

The grouping of mutations, into genes, genes into pathways respectively, reflect the fact that mutations affect function that is has a complex relationship to sequence.

Working at the level of genes has the advantage that genes are well defined.

Pathways, which are representations of genes that work together, are however in a network, so a gene can be in multiple pathways. Hence, pathway exclusion is only an approximation to identifying mutations that give approximately orthogonal selection pressures.

References

[1] 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.

[2] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports*, 3(1):246 – 259, 2013.

[3] Ludmil B. Alexandrov, others, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[4] D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6):788–791, 1999.

[5] K D Korthauer and C Kendziorski. MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, 31(10):1526–1535, May 2015.

[6] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, Adam Kiezun, Peter S Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H Ramos, Trevor J Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M Dulak, Jens Lohr, Dan-Avi Landau, Catherine J Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A McCarroll, Jaume Mora, Ryan S Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B Gabriel, Charles W M Roberts, Jaclyn A Biegel, Kimberly Stegmaier, Adam J Bass, Levi A Garraway, Matthew Meyerson, Todd R Golub, Dmitry A Gordenin, Shamil Sunyaev, Eric S Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[7] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(8):1073–1081, June 2009.

[8] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 3:238–244, 2016.

[9] S E Luria. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, November 1946.

[10] Chen-Hsiang Yeang, Frank McCormick, and Arnold Levine. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J*, 22:2605–2622, 2008.

[11] Sander Canisius, John W. M. Martens, and Lodewyk F. A. Wessels. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome biology*, 17, 2016. Copyright - Copyright BioMed Central 2016; Last updated - 2017-01-04.

[12] Xing Hua, Paula L Hyland, Jing Huang, Lei Song, Bin Zhu, Neil E Caporaso, Maria Teresa Landi, Nilanjan Chatterjee, and Jianxin Shi. MEGSA: A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations. *The American Journal of Human Genetics*, 98(3):442–455, March 2016.