A ratiometric method for detecting selection of somatic mutations in cancer

Daniel Wells
Schuster-Böckler Group, Ludwig Institute of Cancer Research, Univeristy of Oxford

INTRODUCTION

Cancer is among one of the most important diseases worldwide, causing 15% of all deaths and 8% of disability adjusted life years in 2013 (Global Burden of Disease Cancer Collaboration 2015).

The common cause of all common cancers is understood to be somatic mutations which occur during the lifetime of an individual in certain genes (Croce 2008; Stratton et al. 2009; Vogelstein and Kinzler 2004). Hence identifying both the causes of these somatic mutations and the genes involved could enable the prevention and treatment of cancer respectively and so are major goals in the field of oncology. Both of these questions can potentially be addressed by sequencing paired tumour-normal tissue samples. Recent drops in the cost of sequencing have therefore led to a growth of sequencing data from tumour-normal matched pairs. The largest collection of such data is the International Cancer Genome Consortium (ICGC – which includes The Cancer Genome Atlas (TCGA)) which has funding commitments to sequence 25,000 individuals from 50 different cancer (sub)types (Hudson et al. 2010). An additional 25,000 individuals will be sequenced by the end of 2017 as part of a UK government initiative (Genomics England Ltd 2015).

When somatic mutations occur, cells with mutations that confer a growth advantage divide more rapidly and form a clonal population (i.e. are positively selected). Hence, mutations in cancer driver genes are found more frequently in tumour compared to normal tissue. Many studies have therefore used sequencing data to identify cancer driver genes by first estimating the number of expected mutations and then testing for genes which have a higher number of mutations than expected (Lawrence et al. 2014; Lawrence et al. 2013; Kandoth et al. 2013; C. G. A. R. Network 2013; C. G. A. Network 2012; Stephens et al. 2012; Sjöblom et al. 2006; Ding et al. 2008; Tamborero et al. 2013a; Gonzalez-Perez and Lopez-Bigas 2012; Tamborero et al. 2013b; Youn and Simon 2011).

Many of these studies only looked for genes with higher than expected rates of non-synonymous mutation, whereas in this study we also look for those with lower i.e. under-mutated genes. The presence of fewer than expected mutations suggests that these genes are to some extent required for the survival of the cancer, and hence are potential drug targets. These genes would be especially attractive targets if their essentiality was dependent on a cancer-driving mutation as disrupting the essential gene would lead to synthetic lethality - leaving normal cells unharmed. So far efforts to drug classical cancer driver genes have been disappointing due to the difficulty of pharmacologically reactivating tumour suppressor genes or targeting oncogenes with similar active sites to many other proteins (Kaelin 2005; Wang and Giaccone 2011).

However, hypomutated genes are harder to identify than hypermutated genes as there needs to be a minimum level of background mutation to detect a deficit whereas no such threshold exists for hyper-mutated genes, effectively the signal is amplified by the cancer itself.

The second way this work differs from the most popular previous approaches such as MutSig is that we use a ratio-metric method to identify selection. For example MutSig calculates an absolute number of expected mutations per gene, which requires pooling genes with similar replication timing, expression level and chromatin states to achieve an accurate estimation. This is then compared to the observed rate (Lawrence et al. 2014). However with increasing amounts of data it becomes possible to look for evidence of selection via the odds of non-synonymous (amino acid changing) to synonymous (silent) mutations compared to the expected ods. A odds of ~ 1 suggests lack of selection, $<< 1$ purifying (negative) selection, and $>> 1$ posi-

tive selection. This approach is inspired by both the use of dNdS in evolutionary analysis and the 20/20 rule suggested by Vogelstein et al. 2013.

Although there have been previous attempts to look for negative selection using a ratiometric approach many used data from only a single cancer type hence sacrificing power, or they failed to show convincing positive controls, used arbitrary thresholds to create a candidate gene list, failed to test for statistical significance, or suggested unlikely candidates (TTN) with low effect sizes as possible cancer genes (Pyatnitskiy et al. 2015; Zhou et al. 2015; Ostrow et al. 2014; Woo and W.-H. Li 2012).

METHODS

A. Data Sources and Cleaning

Somatic mutation data for each project where available was downloaded from the ICGC data portal (release v20). Technical artefacts were minimised by filtering out mutations with poor mappability (<0.8 average 50bp mappability over a 100bp windows (Derrien et al. 2012)) or with a high frequency in the population (>1% in ExAC v0.3.1 (Exome Aggregation Consortium et al. 2015)). This resulted in a final data set with 1,353,946 somatic mutations from 9,077 patients and 55 projects occurring within the coding sequences of 17,929 genes.

B. Calculation of effect size

We calculate effect size as the odds ratio of observed odds compared to the expected odds where the expected is calculated assuming no selection of mutations. Hence our null hypothesis, H_0, is that the odds ratio = 1.

$$\text{odds ratio} = \frac{\text{observed odds}}{\text{expected odds}} = \frac{N/S}{n/s} \tag{1}$$

N = Number of observed non-synonymous mutations
S = Number of observed synonymous mutations
n = Number of expected non-synonymous mutations
s = Number of expected synonymous mutations

If there is no selection the odds of expected non-synonymous to synonymous mutations in a gene can be modelled by considering the following factors:

Codon composition

Different codons have different number of possible non-synonymous and synonymous sites (Nei and Gojobori 1986). For example all mutations in the codon for tryptophan TGG will by non-synonymous, but all mutations in the third nucleotide of the codon for serine (TCT) will be synonymous as TCC, TCA and TCG also code for serine.

Nucleotide sequence and cancer type

Mutations in different cancer types are caused by different mutagens which mutate some nucleotide sequences more often than others due to their molecular mechanism. For example, in skin cancer there is a high number cytosine-cytosine to thymine-thymine mutations due to ultraviolet light exposure (Alexandrov et al. 2013).

To simplify the model and reduce the number of parameters we assume that this mutation profile is constant across the whole genome, whereas in reality certain types of mutation may be more likely to occur in certain regions due to the physical structure of the genome, with some regions more exposed to mutagenic processes than others.

First, we analytically calculated the expected number of mutations under the null hypothesis by summing for each gene-project combination the number of possible (non)synonymous mutations weighted by the probability of that mutation occurring given the nucleotide sequence:

$$n_{gp} = \sum_{i=1}^{g} \sum_{i=1}^{l} \sum_{x \in (B \setminus B_o)} f_{NS}(c_o, c_n) \frac{\sum t_o \rightarrow t_n}{d_p \sum t_o} \tag{2}$$

Where

$$f_{NS}(c_o, c_n) = \begin{cases} 0 & \text{if } c_o \rightarrow c_n \text{ is synonymous} \\ 1 & \text{if } c_o \rightarrow c_n \text{ is non-synonymous} \end{cases} \tag{3}$$

g = gene
p = project
l = length of gene in base pairs
$B := \{A, C, T, G\}$
B_o = original base
c_o = original codon
c_n = new codon
t_o = original trimer
t_n = new trimer
d_p = number of donors in a project p

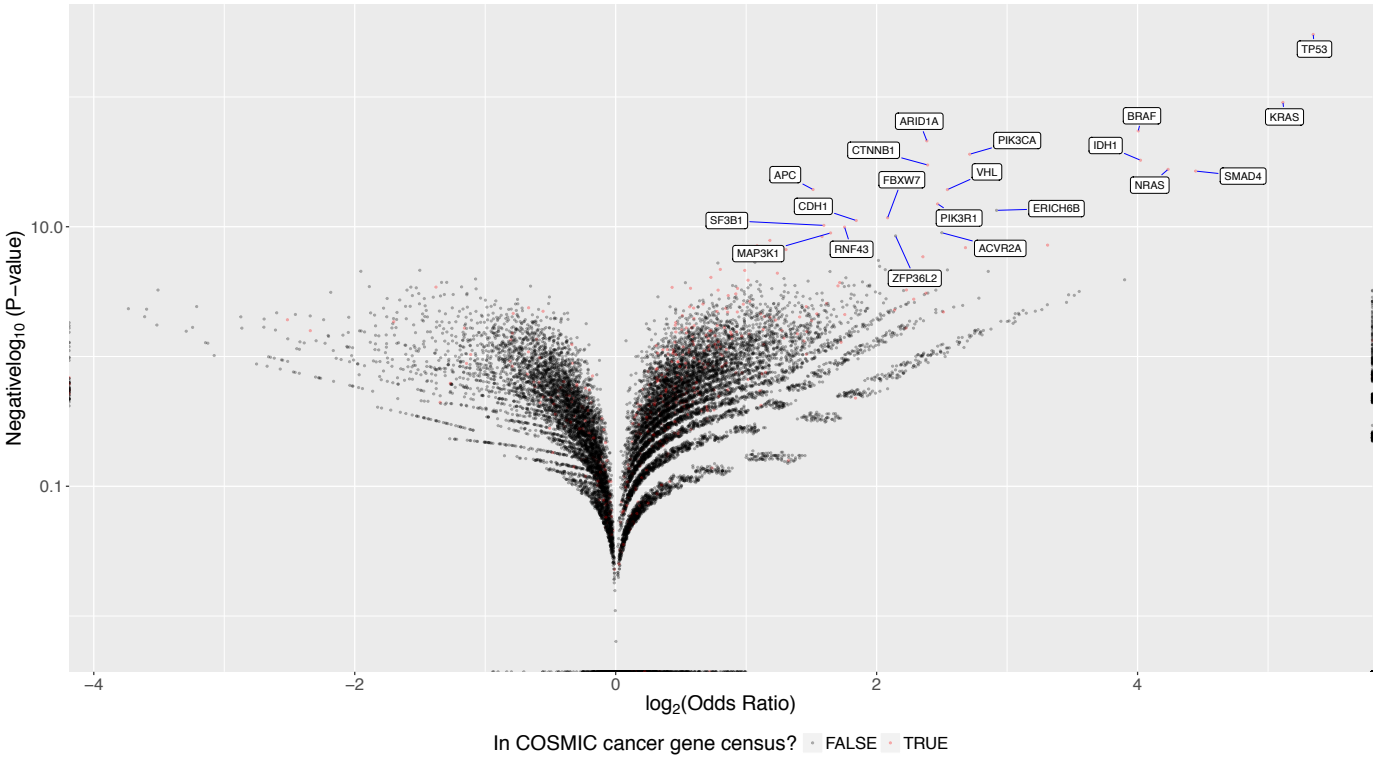

Figure 1 Volcano plot, effect size (clinical significance) on the x-axis and statistical significance on the y-axis. The top 20 most significant genes are labelled.

We also calculated the expected ratio under the null via an empirical approach whereby for each project we shuffled the positions of all mutations within coding regions keeping the mutation type and context constant. We repeated the shuffling 10,000 times and calculated the mean number of observed synonymous $\hat{s}$ and non-synonymous $\hat{n}$ mutations in each gene. The odds $\hat{n}/\hat{s}$ gives an estimate of the expected odds. There was good agreement between these two methods (correlation coefficient = 0.994, supplementary figure 6) and so we used the values from the analytical approach in our analysis due to the lower computational demands.

C. Calculation of statistical significance

We note that for a given gene the number of synonymous mutations observed under neutral selection roughly follows a binomial distribution given the total number of mutations and the probability of synonymous mutation (Greenman et al. 2006).

$$P(S|H_0) \sim B(N+S, (\frac{s}{s+n})) \quad (4)$$

We therefore used a two-sided binomial test to calculate p-values (the probability of observing an S value equal to or more extreme than we did given the null hypothesis) and corrected them for multiple testing by the Benjamini-Hochberg method (figure 2). For $\alpha = 0.001$ and a null odds of 2.6 this test has 80% power to detect an absolute $\log_2$(odds ratio) of -2 or greater for genes with more than 38 total mutations. 64% of the genes in our set have at least this many mutations (supplementary figures 3 and 4).

RESULTS

We observe 75 statistically significant genes (q<0.1) under positive selection with large effect sizes (mean $\log_2$(odds ratio) = 1.9, figure 1). This gene-set is highly enriched for the classic canonical oncogenes and tumour suppressor genes including TP53, KRAS and BRAF (table 1); in total 40 out of the 75 are known cancer genes. We observe fewer (n=12) statistically significant genes under negative selection (table 2), which could be due to low power (supplementary figure 3).

Some of these genes are clearly false positives - some due to incomplete filtering of technical artefacts (NELFE, ERICH6B), and others such as the olfactory receptors have very low expression and by function are not likely to be involved in cancer (Lawrence et al. 2013). There are also some genes which are not anno-

tated as known cancer genes in COSMIC (Forbes et al. 2015) but are likely true positives: ZFP36L2, ZFPM1 (FOG1) and ACVR2A (Hodson et al. 2010; Zheng and Blobel 2010; Deacu et al. 2004)

EXTENSIONS AND LIMITATIONS

There are a number of ways this method could be improved as well as some more fundamental limits.

This method only considers mutations in coding regions due to both the difficulty of assigning functional impact to a non-coding mutation as well as that only 24% of the donors had whole genome sequencing available. However, it has been suggested that non-coding mutations do play a role in oncogenesis (Melton et al. 2015; Khurana et al. 2016).

We consider all non-synonymous mutations equal whereas actually, some are more likely to have a functional impact than others (by amino acid type or evolutionary conservation). In addition, we consider all synonymous mutations to be non functional however they can have functional consequences for example changing splice sites or miRNA binding (Supek et al. 2014). Indeed, BCL2L12 which was previously identified as both having a functional synonymous mutation and being under negative selection ranks 24th in our list of 'negatively selected' genes (Gartner et al. 2013; Zhou et al. 2015).

We calculate an average odds ratio per gene, but this could average out smaller regions under selection It may be possible to separately test for clustering of mutations in the sequence or in 3D protein structure (Kamburov et al. 2015).

We did not have the capacity to re-call the mutations from the raw sequencing files, and so these mutations are called by different programs using different parameters (Alioto et al. 2015). Furthermore sequenced tumours are actually a mix of tumorous and non-tumorous stomal tissue i.e. are not 100% pure which weakens the signal in the data. Additionally we only look at the longest transcript for each gene but this may not be the most functionally important transcript.

We have shown that using a ratiometric approach can identify selection in cancer genomes and with additional data we expect that we will be able to call with more confidence genes under negative selection specifically.

LITERATURE CITED

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. (2013). "Signatures of mutational processes in human cancer". In: *Nature* 500.7463, pp. 415–421. DOI: 10.1038/nature12477.

Alioto, Tyler S, Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D Eldridge, Eivind Hovig, Lawrence E Heisler, Timothy A Beck, Jared T Simpson, Laurie Tonon, et al. (2015). "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing". In: *Nature communications* 6. DOI: 10.1038/ncomms10001.

Chernick, Michael R and Christine Y Liu (2002). "The Saw-Toothed Behavior of Power Versus Sample Size and Software Solutions". In: *The American Statistician* 56.2, pp. 149–155. DOI: 10.1198/000313002317572835.

Croce, Carlo M. (2008). "Oncogenes and Cancer". In: *New England Journal of Medicine* 358.5, pp. 502–511. DOI: 10.1056/NEJMra072367.

Deacu, Elena, Yuriko Mori, Fumiaki Sato, Jing Yin, Andreea Olaru, Anca Sterian, Yan Xu, Suna Wang, Karsten Schulmann, Agnes Berki, Takatsugu Kan, John M. Abraham, and Stephen J. Meltzer (2004). "Activin Type II Receptor Restoration in ACVR2-Deficient Colon Cancer Cells Induces Transforming Growth Factor-Beta Response Pathway Genes". In: *Cancer Research* 64.21, pp. 7690–7696. DOI: 10.1158/0008-5472.CAN-04-2082.

Derrien, Thomas, Jordi Estellé, Santiago Marco Sola, David G. Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca (2012). "Fast Computation and Applications of Genome Mappability". In: *PLoS ONE* 7.1, pp. 1–16. DOI: 10.1371/journal.pone.0030377.

Ding, Li, Gad Getz, David A Wheeler, Elaine R Mardis, Michael D McLellan, Kristian Cibulskis, Carrie Sougnez, Heidi Greulich, Donna M Muzny, Margaret B Morgan, et al. (2008). "Somatic mutations affect key pathways in lung adenocarcinoma". In: *Nature* 455.7216, pp. 1069–1075. DOI: 10.1038/nature07423.

Exome Aggregation Consortium et al. (2015). "Analysis of protein-coding genetic variation in 60,706 humans". In: *bioRxiv*. DOI: 10.1101/030338.

Forbes, Simon A., David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Min-

	Number of Mutations				
Gene	Synonymous	Nonynonymous	Log_2(odds ratio)	q-value	Known cancer gene
TP53	23	2566	5.35	7.24×10^{-300}	TRUE
KRAS	8	914	5.11	1.46×10^{-87}	TRUE
BRAF	12	583	4.01	1.13×10^{-51}	TRUE
ARID1A	51	652	2.39	5.52×10^{-43}	TRUE
PIK3CA	26	636	2.71	2.16×10^{-33}	TRUE
IDH1	7	340	4.02	8.42×10^{-30}	TRUE
CTNNB1	31	553	2.39	3.12×10^{-27}	TRUE
NRAS	5	280	4.23	4.52×10^{-25}	TRUE
SMAD4	4	284	4.45	2.78×10^{-24}	TRUE
APC	63	703	1.51	7.56×10^{-17}	TRUE
VHL	18	232	2.54	7.56×10^{-17}	TRUE
PIK3R1	14	250	2.47	1.53×10^{-12}	TRUE
ERICH6B	8	165	2.92	5.77×10^{-11}	FALSE
FBXW7	17	250	2.08	2.40×10^{-9}	TRUE
CDH1	24	206	1.84	8.14×10^{-9}	TRUE
SF3B1	29	317	1.60	6.38×10^{-8}	TRUE
RNF43	23	223	1.75	1.22×10^{-7}	TRUE
ACVR2A	8	149	2.50	1.02×10^{-6}	FALSE
MAP3K1	24	224	1.65	1.04×10^{-6}	TRUE
ZFP36L2	13	100	2.15	2.83×10^{-6}	FALSE
PBRM1	24	247	1.58	3.37×10^{-6}	TRUE
ATRX	44	407	1.18	1.15×10^{-5}	TRUE
RB1	19	200	1.66	2.56×10^{-5}	TRUE
BAP1	20	148	1.66	3.30×10^{-5}	TRUE
HRAS	3	66	3.31	4.35×10^{-5}	TRUE
ARID2	50	343	1.08	5.09×10^{-5}	TRUE
ZFPM1	18	111	1.75	5.69×10^{-5}	FALSE
CASP8	17	151	1.72	7.35×10^{-5}	TRUE
SPOP	5	97	2.68	7.78×10^{-5}	TRUE
GATA3	34	161	1.31	1.28×10^{-4}	TRUE

Table 1 Top 25 most significant genes with evidence for positive selection

	Number of Mutations				
Gene	Synonymous	Nonynonymous	Log_2(odds ratio)	q-value	Known cancer gene
NELFE	32	39	−1.50	1.05×10^{-2}	FALSE
SPATA3	23	17	−1.95	1.21×10^{-2}	FALSE
AP1B1	57	53	−1.07	3.90×10^{-2}	FALSE
OR1M1	42	30	−1.29	5.78×10^{-2}	FALSE
WNT3A	36	30	−1.33	5.78×10^{-2}	FALSE
SLC8A3	109	160	−0.652	8.00×10^{-2}	FALSE
GINS4	18	13	−1.89	8.13×10^{-2}	FALSE
HOXD3	40	37	−1.18	8.34×10^{-2}	FALSE
FUS	27	34	−1.38	8.39×10^{-2}	TRUE
ARSA	27	19	−1.57	8.39×10^{-2}	FALSE
OR2F2	27	20	−1.53	8.97×10^{-2}	FALSE
KCNK1	46	44	−1.07	9.86×10^{-2}	FALSE
KRTAP19-4	8	2	−3.51	1.08×10^{-1}	FALSE
UHRF2	19	22	−1.60	1.08×10^{-1}	FALSE
OR10H3	15	7	−2.18	1.31×10^{-1}	FALSE
OR1G1	40	35	−1.14	1.42×10^{-1}	FALSE
REEP2	25	18	−1.51	1.42×10^{-1}	FALSE
NOP9	30	33	−1.25	1.52×10^{-1}	FALSE
TRIB2	28	25	−1.31	1.52×10^{-1}	FALSE
OR8B12	38	34	−1.14	1.53×10^{-1}	FALSE
FGF21	21	16	−1.56	1.58×10^{-1}	FALSE
KRTAP27-1	24	27	−1.34	1.64×10^{-1}	FALSE
MTUS2	135	245	−0.508	1.75×10^{-1}	FALSE
BCL2L12	23	23	−1.40	1.77×10^{-1}	FALSE
OR4K17	38	45	−1.04	1.92×10^{-1}	FALSE
MYL9	17	12	−1.75	1.96×10^{-1}	FALSE
VPS72	19	21	−1.47	1.96×10^{-1}	FALSE
PCDHGA5	68	86	−0.752	2.05×10^{-1}	FALSE
SPAG7	17	15	−1.65	2.16×10^{-1}	FALSE
NKAPL	35	60	−0.999	2.16×10^{-1}	FALSE

Table 2 Top 25 most significant genes with evidence for negative selection

jie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott, and Peter J. Campbell (2015). "COSMIC: exploring the world's knowledge of somatic mutations in human cancer". In: *Nucleic Acids Research* 43.D1, pp. D805–D811. DOI: 10.1093/nar/gku1075.

Gartner, Jared J. et al. (2013). "Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma". In: *Proceedings of the National Academy of Sciences* 110.33, pp. 13481–13486. DOI: 10.1073/pnas.1304227110.

Genomics England Ltd (2015). *The 100,000 Genomes Project*. URL: http://www.genomicsengland.co.uk/the-100000-genomes-project/ (visited on 03/04/2016).

Global Burden of Disease Cancer Collaboration (2015). "The global burden of cancer 2013". In: *JAMA Oncology* 1.4, pp. 505–527. DOI: 10.1001/jamaoncol.2015.0735.

Gonzalez-Perez, Abel and Nuria Lopez-Bigas (2012). "Functional impact bias reveals cancer drivers". In: *Nucleic Acids Research* 40.21, e169. DOI: 10.1093/nar/gks743.

Greenman, Chris, Richard Wooster, P. Andrew Futreal, Michael R. Stratton, and Douglas F. Easton (2006). "Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer". In: *Genetics* 173.4, pp. 2187–2198. DOI: 10.1534/genetics.105.044677.

Hodson, Daniel J, Michelle L Janas, Alison Galloway, Sarah E Bell, Simon Andrews, Cheuk M Li, Richard Pannell, Christian W Siebel, H Robson MacDonald, Kim De Keersmaecker, et al. (2010). "Deletion of the RNA-binding proteins ZFP36L1 and ZFP36L2 leads to perturbed thymic development and T lymphoblastic leukemia". In: *Nature immunology* 11.8, pp. 717–724. DOI: 10.1038/ni.1901.

Hudson, Thomas J, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. (2010). "International network of cancer genome projects". In: *Nature* 464.7291, pp. 993–998. DOI: 10.1038/nature08987.

Kaelin, William G (2005). "The concept of synthetic lethality in the context of anticancer therapy". In: *Nature reviews cancer* 5.9, pp. 689–698. DOI: 10.1038/nrc1691.

Kamburov, Atanas, Michael S. Lawrence, Paz Polak, Ignaty Leshchiner, Kasper Lage, Todd R. Golub, Eric S. Lander, and Gad Getz (2015). "Comprehensive assessment of cancer missense mutation clustering in protein structures". In: *Proceedings of the National Academy of Sciences* 112.40, E5486–E5495. DOI: 10.1073/pnas.1516373112.

Kandoth, Cyriac, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. (2013). "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471, pp. 333–339. DOI: 10.1038/nature12634.

Khurana, Ekta, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A Rubin, and Mark Gerstein (2016). "Role of non-coding sequence variants in cancer". In: *Nature Reviews Genetics* 17.2, pp. 93–108. DOI: 10.1038/nrg.2015.17.

Lawrence, Michael S, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484, pp. 495–501. DOI: 10.1038/nature12912.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457, pp. 214–218. DOI: 10.1038/nature12213.

Melton, Collin, Jason A Reuter, Damek V Spacek, and Michael Snyder (2015). "Recurrent somatic mutations in regulatory regions of human cancer genomes". In: *Nature genetics* 47.7, pp. 710–716. DOI: 10.1038/ng.3332.

Nei, M and T Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." In: *Molecular Biology and Evolution* 3.5, pp. 418–426.

Network, Cancer Genome Atlas et al. (2012). "Comprehensive molecular characterization of human colon and rectal cancer". In: *Nature* 487.7407, pp. 330–337. DOI: 10.1038/nature11252.

Network, Cancer Genome Atlas Research et al. (2013). "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia". In: *The New England journal of medicine* 368.22, p. 2059. DOI: 10.1056/NEJMoa1301689.

Ostrow, Sheli L., Ruth Barshir, James DeGregori, Esti Yeger-Lotem, and Ruth Hershberg (2014). "Cancer Evolution Is Associated with Pervasive Positive Se-

lection on Globally Expressed Genes". In: *PLoS Genet* 10.3, pp. 1–11. DOI: 10.1371/journal.pgen.1004239.

Pyatnitskiy, Mikhail, Dmitriy Karpov, Ekaterina Poverennaya, Andrey Lisitsa, and Sergei Moshkovskii (2015). "Bringing Down Cancer Aircraft: Searching for Essential Hypomutated Proteins in Skin Melanoma". In: *PLoS ONE* 10.11, pp. 1–14. DOI: 10.1371/journal.pone.0142819.

Sjöblom, Tobias et al. (2006). "The Consensus Coding Sequences of Human Breast and Colorectal Cancers". In: *Science* 314.5797, pp. 268–274. DOI: 10.1126/science.1133427.

Stephens, Philip J, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, et al. (2012). "The landscape of cancer genes and mutational processes in breast cancer". In: *Nature* 486.7403, pp. 400–404. DOI: 10.1038/nature11017.

Stratton, Michael R, Peter J Campbell, and P Andrew Futreal (2009). "The cancer genome". In: *Nature* 458.7239, pp. 719–724. DOI: 10.1038/nature07943.

Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner (2014). "Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers". In: *Cell* 156.6, pp. 1324–1335. DOI: 10.1016/j.cell.2014.01.051.

Tamborero, David, Abel Gonzalez-Perez, and Nuria Lopez-Bigas (2013a). "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes". In: *Bioinformatics* 29.18, pp. 2238–2244. DOI: 10.1093/bioinformatics/btt395.

Tamborero, David, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. (2013b). "Comprehensive identification of mutational cancer driver genes across 12 tumor types". In: *Scientific reports* 3. DOI: 10.1038/srep02650.

Vogelstein, Bert and Kenneth W Kinzler (2004). "Cancer genes and the pathways they control". In: *Nature medicine* 10.8, pp. 789–799. DOI: 10.1038/nm1087.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler (2013). "Cancer Genome Landscapes". In: *Science* 339.6127, pp. 1546–1558. DOI: 10.1126/science.1235122.

Wang, Yisong and Giuseppe Giaccone (2011). "Challenges in Cancer Molecular Targets and Therapeutics". In: *Frontiers in Oncology* 1.4. DOI: 10.3389/fonc.2011.00004.

Woo, Yong H and Wen-Hsiung Li (2012). "DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes". In: *Nature communications* 3, p. 1004. DOI: 10.1038/ncomms2502.

Youn, Ahrim and Richard Simon (2011). "Identifying cancer driver genes in tumor genome sequencing studies". In: *Bioinformatics* 27.2, pp. 175–181. DOI: 10.1093/bioinformatics/btq630.

Zheng, Rena and Gerd A. Blobel (2010). "GATA Transcription Factors and Cancer". In: *Genes and Cancer* 1.12, pp. 1178–1188. DOI: 10.1177/1947601911404223.

Zhou, Zhan, Yangyun Zou, Gangbiao Liu, Jingqi Zhou, Shiming Zhao, Zhixi Su, and Gu Xun (2015). "A New Mutation-Profile-Based Method for Understanding the Evolution of Cancer Somatic Mutations". In: *bioRxiv*. DOI: 10.1101/021147.

SUPPLEMENTARY FIGURES

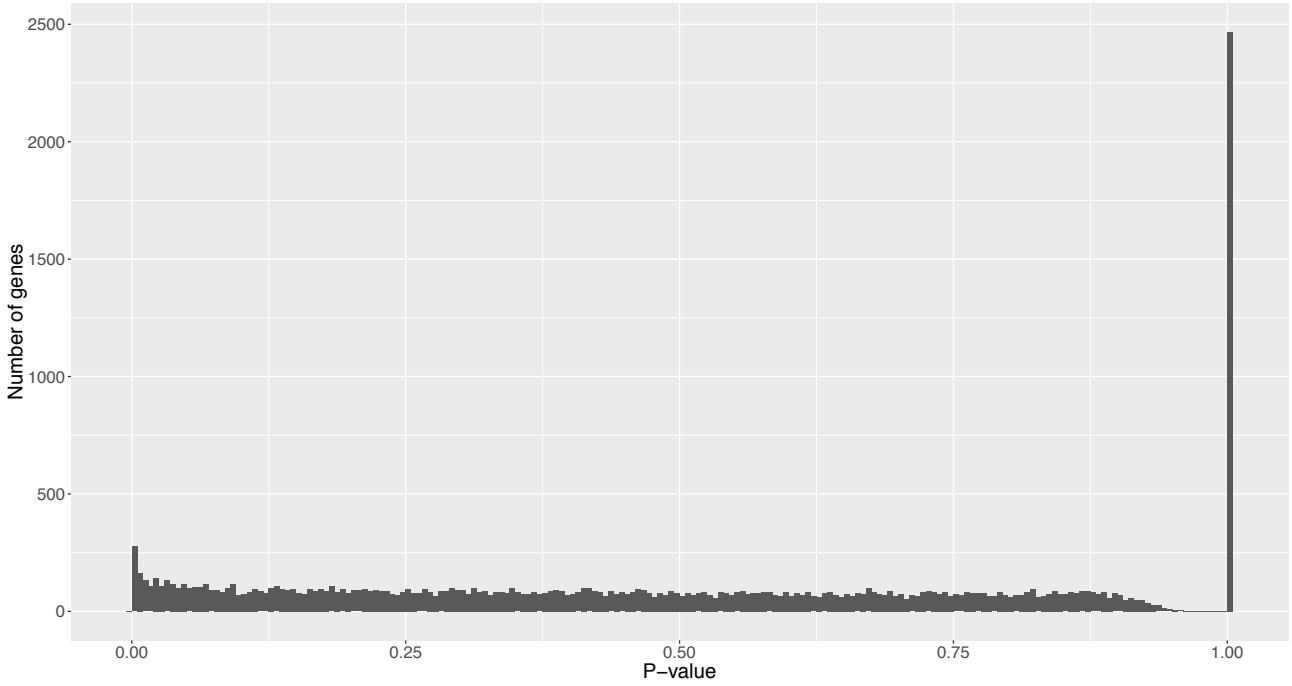

Figure 2 Distribution of p-values. The depletion of values just before 1 is due to the discrete nature of the binomial distribution this is based on.

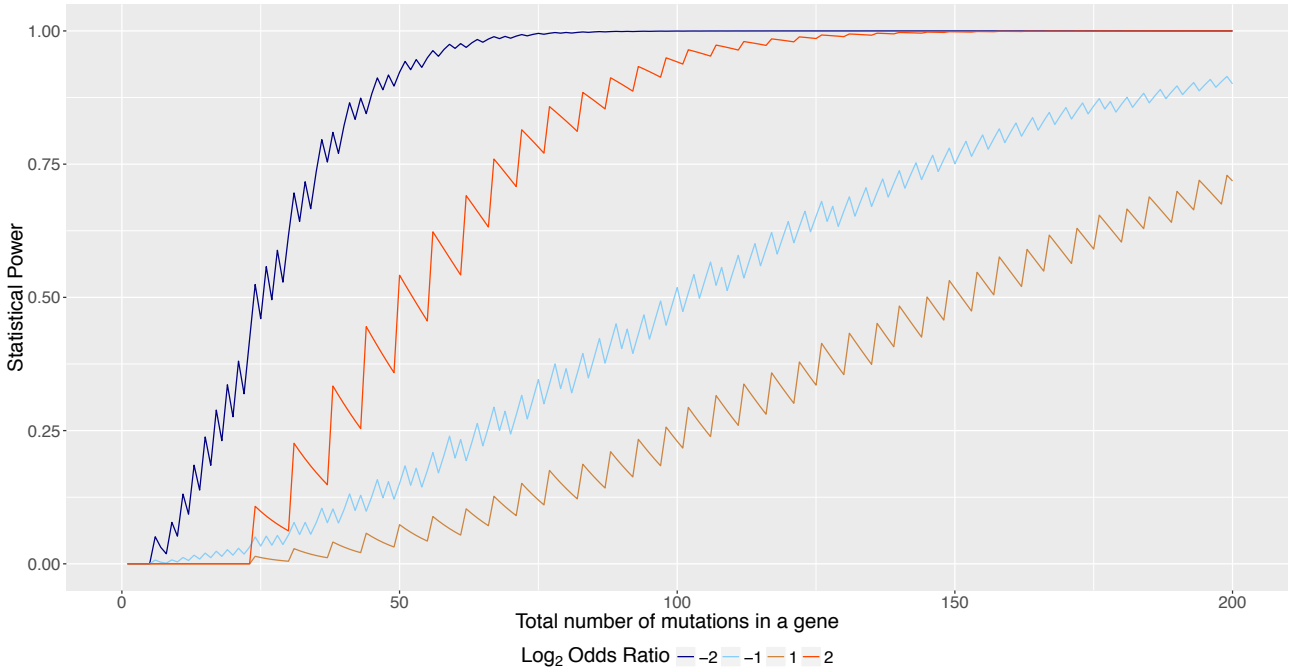

Figure 3 Power curve for differing effect sizes and number of mutations per gene. The saw tooth effect is due to the discrete nature of the binomial distribution (Chernick and Liu 2002). Although the test has greater power to detect negative effect sizes for genes with equal number of mutations, positively selected genes will have a higher number of mutations than neutral or negatively selected genes.

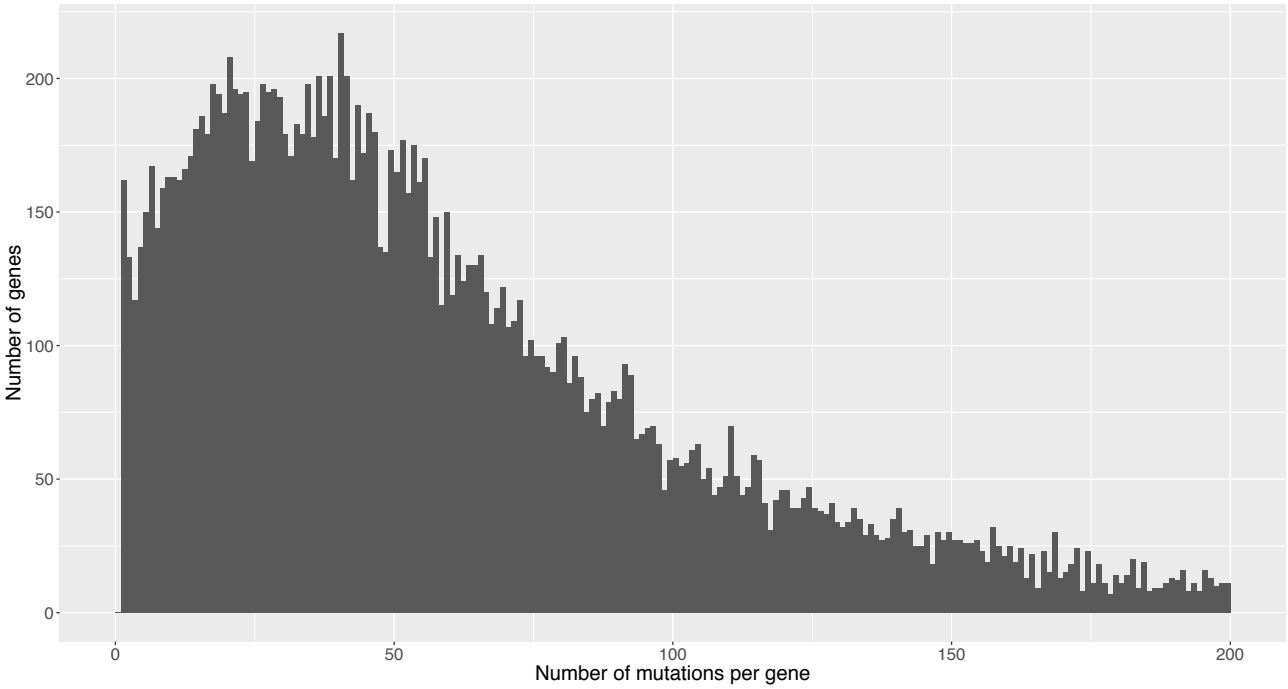

Figure 4 Distribution of mutations per gene (values beyond 200 not shown).

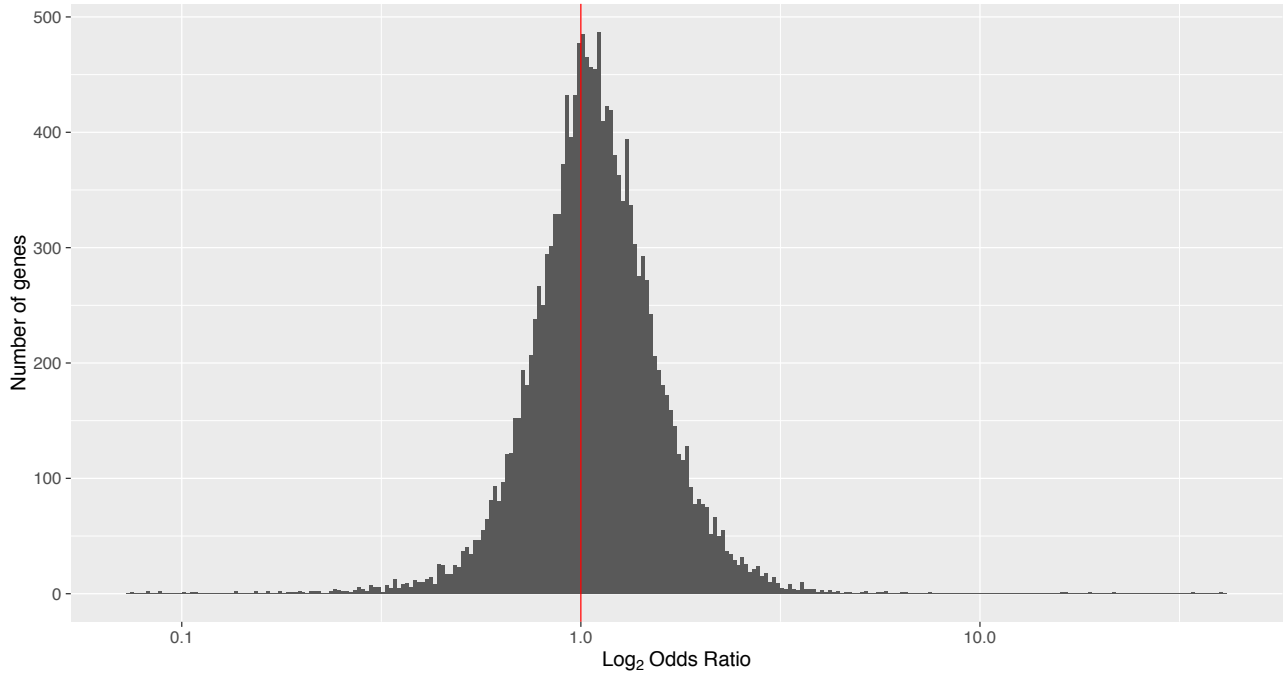

Figure 5 Distribution of $\log_2$ odds ratios.

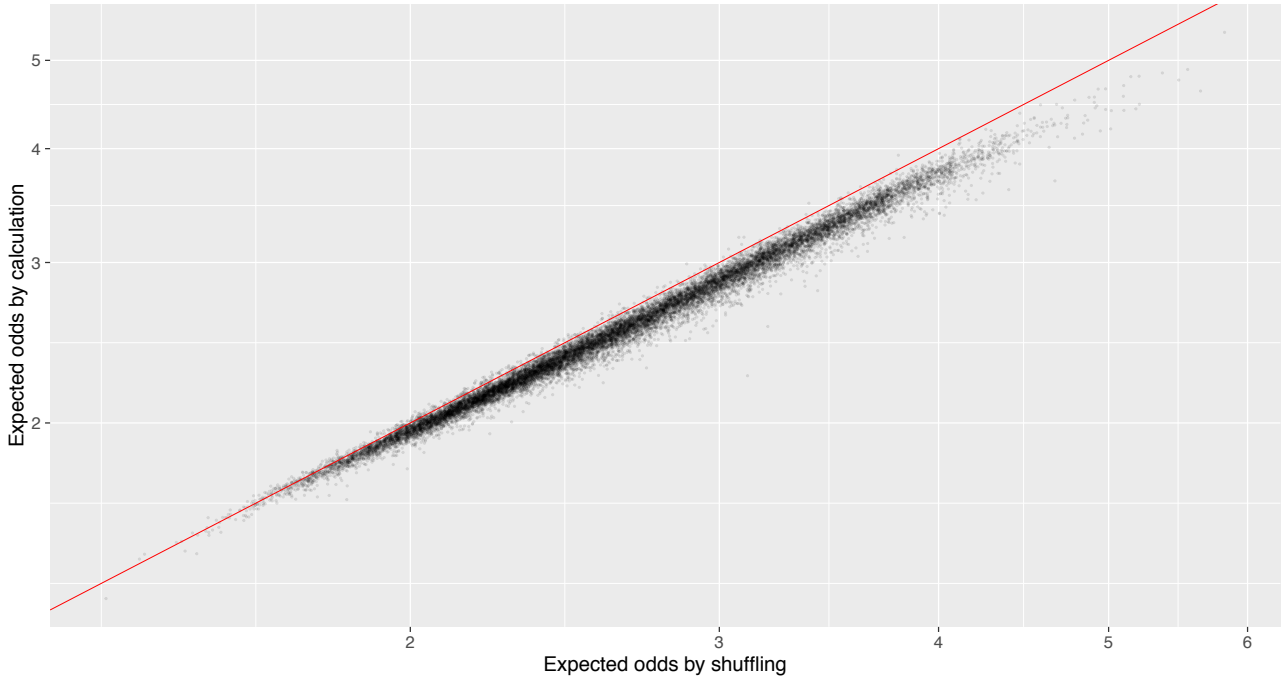

Figure 6 Correlation between analytically and empirically calculated expected ratio of non-synonymous mutations to synonymous mutations. The empirical approach results in slightly higher odds on average.

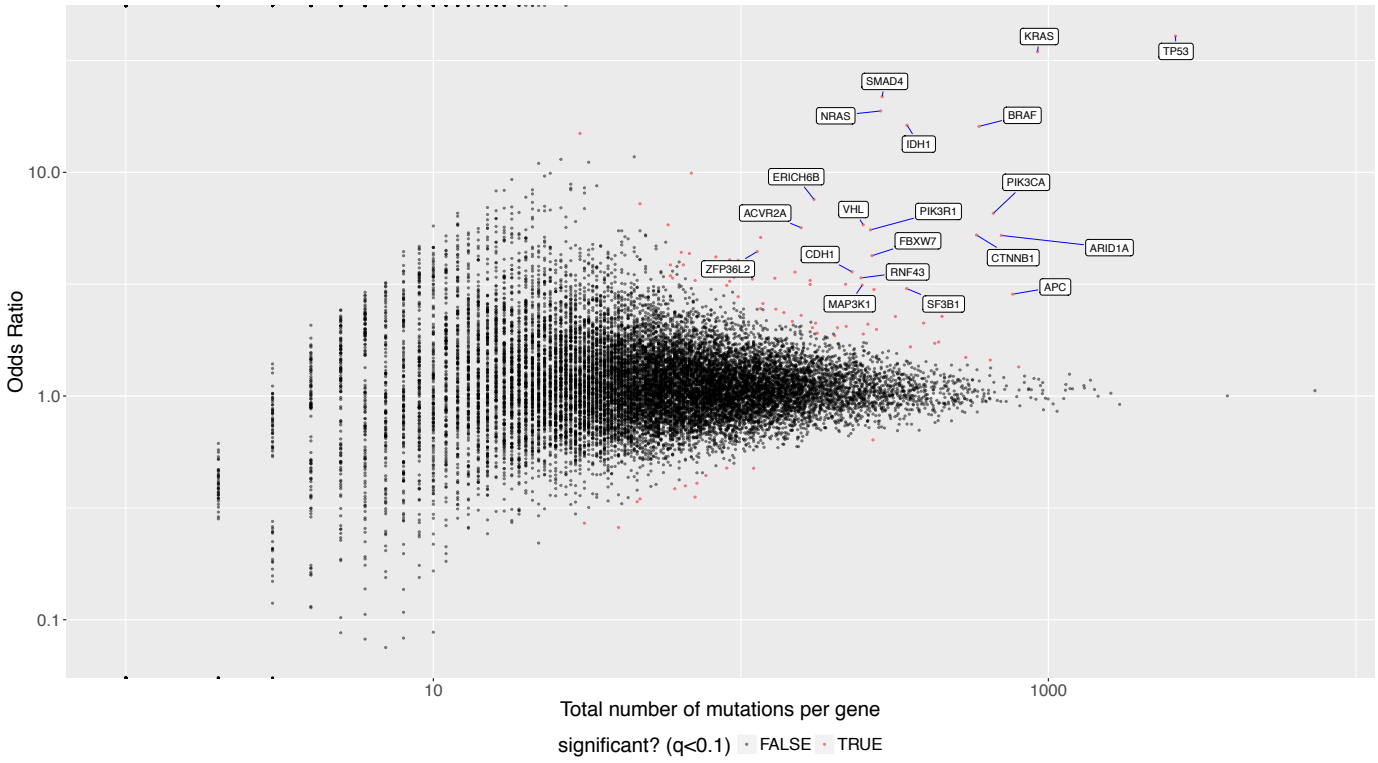

Figure 7 Funnel plot showing the decrease in variation of odds ratio with increasing number of mutations.