Published online 31 July 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 12* **1–6**
doi:10.1093/nar/gkn000

GENOMICS AND BIOINFORMATICS
Article title

Brian K. Mannakee[1] and Ryan N. Gutenkunst[2*]

[1]University of Arizona Mel and Enid Zuckerman College of Public Health and [2]University of Arizona Department of Molecular and Cellular Biology

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

Text. Text.

INTRODUCTION

Cancer develops as the result of the accumulation of somatic mutations and clonal selection of cells with mutations that confer a selective advantage on the cell. Understanding the forces that shaped the evolutionary history of a tumor, the mutations that are responsible for its growth, the rate at which mutations are occurring, or how much genetic diversity is likely present in the tumor, requires accurate variant calling, particularly at low variant allele frequency (Williams et al., 2016, Bozic et al., 2016, Williams et al., 2018). Accurate variant identification is also critical in optimizing the treatment regime for an individual patients disease (Ding et al., 2012, Mardis, 2012, Chen et al., 2013, Borad et al., 2014, Findlay et al., 2016). Low frequency mutations present a significant problem for current mutation calling methods because their signature in the data is difficult to distinguish from the noise introduced by Next Generation Sequencing (NGS), and this problem increases as sequencing depth increases.

Methods for identifying true somatic mutations - i.e. variant calling - from NGS data are an active area of research in bioinformatics. The earliest widely used somatic variant callers aimed specifically at tumors, Mutect1 and Varscan2, used a combination of heuristic filtering and a model of sequencing errors to identify and score potential variants, setting a threshold for that score designed to balance sensitivity and specificity (Koboldt et al., 2012, Cibulskis et al., 2013). Subsequent research gave rise to a number of alternate variant calling strategies including haplotype based callers (Garrison and Marth, 2012), joint genotype analysis (SomaticSniper, JointSNVMix2, Seurat, and CaVEMan,MuClone)(Larson et al., 2012, Roth et al., 2012, Christoforides et al., 2013, Jones et al., 2016, Dorri et al., 2019), allele frequency based analysis (Strelka, MuTect, LoFreq, EBCall, deepSNV, LoLoPicker, and MuSE)(Saunders et al., 2012, Wilm et al., 2012, Shiraishi et al., 2013, Gerstung et al., 2012, Carrot-Zhang and Majewski, 2017, Fan et al., 2016), and a mixture of ensemble and deep learning methods (MutationSeq, SomaticSeq, SNooPer, and BAYSIC). All of these methods have varying levels of complexity, and some are focused on specific types of data. The one thing they all have in common is that they either implicitly or explicitly assume that the probability of a mutation occuring at a particular site is proportional to the overall mutation rate, and the same at every site in the genome.

Single nucleotide substitions, i.e. simple mutations, arise in tumors at a rate and at genomic locations driven by two main processes. The first is the spontaneous accumulation of mutations that occurs in all dividing tissues, and has a characteristic mutation signature that describes the probability of mutation in a given genomic context (Nik-Zainal et al., 2012, Alexandrov et al., 2015, Lee-Six et al., 2018). The second, and far more complex, process is the accumulation of mutations through exposure to mutagens or degradation - via mutation or deletion - of cellular machinery responsible for the identification and repair of damage or replication errors. Many mutagens and DNA repair mechanism defects also have highly specific mutation signatures, such that they can be identified by observing the mutations in the tumor (Alexandrov et al., 2013, Helleday et al., 2014, Nik-Zainal et al., 2016, Kandoth et al., 2013, Alexandrov et al., 2016).

Here we present an algorithm for estimating the prior probability of mutation at a given site using the observed mutation spectrum of the tumor as well as its mutation rate, and show that the addition of this prior to the MuTect variant calling model produces a superior variant classifier in both simulated and real tumor data. We then extend the method with an application of the local false discovery rate by computing the probability that a site is non-null under an assumption of clonal expansion with either early or small selective differences between clones. We provide a simple implementation in R that takes MuTect caller output as input, and returns the posterior probability that a site is variant for every site observed by MuTect.

*To whom correspondence should be addressed. Email: rgutenk@email.arizona.edu

MATERIALS AND METHODS

Base probability model

At every site in the genome with non-zero coverage, Next Generation Sequencing (NGS) produces a vector $\mathbf{x} = (\{b_i\},\{q_i\}), i = 1 \ldots D$ of base calls and their associated quality scores, where D is total read depth. The goal is to use $\mathbf{x}$ to select between competing hypotheses;

$$\begin{aligned} \mathbf{H_0}: &\quad \text{Alt allele} = m; \quad \nu = 0 \\ \mathbf{H_1}: &\quad \text{Alt allele} = m; \quad \nu = \hat{f}, \end{aligned}$$

where ν is the variant allele frequency, $\hat{f}$ is the maximum likelihood estimate of ν given data $\mathbf{x}$, i.e. the ratio of the count of variant reads and total read depth, and m is any of the 3 possible alternative non-reference bases. For a given read with base b_i and q-score q_i, the density function under a particular hypothesis is defined as

$$\mathrm{f}_{\nu,m}(b_i,q_i) = \begin{cases} \nu\frac{10^{-q_i/10}}{3} + (1-\nu)(1-10^{-q_i/10}) & b_i = \text{reference} \\ \nu(1-10^{-q_i/10}) + (1-\nu)\frac{10^{-q_i/10}}{3} & b_i = m \\ \frac{10^{-q_i/10}}{3} & otherwise. \end{cases}$$

The likelihood under the hypothesis is then $\mathcal{L}_{\nu,m}(\mathbf{x}) = \prod_{i=1}^{D} \mathrm{f}_{\nu,m}(x_i)$. MuTect reports the log likelihood ratio $\log(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})/\mathcal{L}_{\nu=0,m}(\mathbf{x}))$ as either TLOD or t_lod_fstar depending on the version. By fixing the threshold posterior odds at two, the site-specific mutation probability a constant $\mathrm{p}(M) = \mu = 3\mathrm{e}{-6}$, and $\mathrm{p}(m\,|\,M)$ the prior probability of mutation to specific allele m constant $\mathrm{p}(m\,|\,M) = \mu/3 = 1\mathrm{e}{-6}$, they derive a TLOD threshold of 6.3 for classifying a site as a somatic variant.

Estimation of the mutation signature.

While variant calling algorithms typically assume a constant probability of mutation at every site in the genome, work by Alexandrov and others show that the random mutation generating process actually varies from site to site in a nucleotide context specific manner. We develop a model of the prior probability of mutation to allele m conditional on the observed genomic context $\mathrm{p}(m,M\,|\,C)$, and demonstrate an empirical Bayes method for computing this probability from MuTect output. The prior probability $\mathrm{p}(m,M\,|\,C)$ can be decomposed as

$$\mathrm{p}(m,M\,|\,C) = \mathrm{p}(m\,|\,C)\mathrm{p}(M\,|\,C) = \mathrm{p}(m\,|\,C)\mathrm{p}(C\,|\,M)\frac{p(M)}{p(C)}$$

since the probability of a mutation at a site and the probability that it is to allele m are independent conditional on the context. Here $\mathrm{p}(M) = \mu$ as above, and the empirical distribution of contexts $\mathrm{p}(C)$ is the fraction of the genome made up of each context. We model $\mathrm{p}(C\,|\,M)$ as a multinomial distribution with parameter $\boldsymbol{\pi} = \{\pi_i\}, i = 1 \ldots 96$. Mutations are drawn from this multinomial distribution such that $\mathrm{p}(C = i\,|\,M) = \pi_i$. The final quantity $\mathrm{p}(m\,|\,C)$, the probability of mutation to m given a particular three letter context, is a function of $\boldsymbol{\pi}$. We are left to estimate only the vector of probabilities $\boldsymbol{\pi}$.

$$\begin{aligned} C\,|\,M,\boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{\pi}) \\ \boldsymbol{\pi}\,|\,\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}). \end{aligned}$$

The posterior distribution of $\boldsymbol{\pi}$ is $\boldsymbol{\pi}\,|\,C,\boldsymbol{\alpha} \sim \text{Dirichlet}(\mathbf{C} + \boldsymbol{\alpha})$, where $\mathbf{C} = (C_1,\ldots,C_{96})$ are the counts of mutations present in the tumor for each of the 96 contexts. We compute an empirical bayes estimate of $\boldsymbol{\pi}$ by choosing $\mathbf{C}$ as the set of mutations assigned a TLOD by MuTect above some threshold, which we choose as 10. We show through extensive simulation that our estimate of $\boldsymbol{\pi}$ converges quickly its true simulated value.

Returning to the model above, we can calculate the log posterior odds in favor of $\mathbf{H_1}$ as

$$\log_{10}\left(\frac{(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})\mathrm{p}(m,M\,|\,C)}{(\mathcal{L}_{\nu=0,m}(\mathbf{x})(1-\mathrm{p}(m,M\,|\,C))}\right) = \text{TLOD} + \log \text{ prior odds},$$

and the posterior odds ratio in favor of $\mathbf{H_1}$ as

$$10^{(\text{TLOD}+\log \text{ prior odds})}.$$

We show via extensive simulation that under the assumption that the mutation signature describes the biological process generating mutations in a tumor, our posterior odds ratio is a better classifier than MuTect at any threshold. Unfortunately, this increased classification performance comes at a cost in terms of calibration, The probabilities from this model are substantially worse than those from Mutect, such that threshold selection is essentially impossible and precision/recall is substantially worse for this model. There are several distributional assumptions in this model that effect model calibration, and we address all of them below.

Estimation of the mutation rate.

As discussed above, MuTect fixes the site-specific mutation probability at $\mathrm{p}(M) = \mu = 3\mathrm{e}{-6}$. All variant callers we are aware of either fix this parameter μ or allow the user to input the value, but there is no way to really know this value until the variant allele frequencies have been observed. As with estimation of the mutation signature, we can use high confidence mutations and a model of tumor evolution to compute the tumor-specific mutation rate. Bozic et al. (2016) show that for any variant allele frequency α, the total number of mutations with frequency greater than α and less than 0.25 is

$$N(\alpha) = N\mu\left(\frac{1}{\alpha} - \frac{1}{0.25}\right)$$

Where N is the total number of sites sequenced and μ is the per-site mutation probability. By selecting α such that we are highly confident in all calls at frequencies greater than α, we can compute μ and recompute the odds in favor of $\mathbf{H_1}$.

Tumor simulations.

We simulated realistic variant sites and allele frequencies using a branching process to simulate neutral evolution with no death. Variants were selected from TCGA and PCAWG variant files(dates). Whole genome (100X depth), and whole exome (500X depth) reads from the GRCH38 reference genome with VarSim (Mu et al., 2015), and aligned them to GRch38 with BWA (Li and Durbin, 2009), both with default parameters. Variants were spiked to create tumors with Bamsurgeon with default parameters (Ewing et al., 2015), and called with MuTect 1.1.7 (Cibulskis et al., 2013) with the following parameters:

```
java -Xmx24g -jar $MUTECT_JAR --analysis_type MuTect --reference_sequence $ref_path \
      --dbsnp $db_snp \
      --enable_extended_output \
      --fraction_contamination 0.00 \
      --tumor_f_pretest 0.00 \
      --initial_tumor_lod -10.00 \
      --required_maximum_alt_allele_mapping_quality_score 1 \
      --input_file:normal $tmp_normal \
      --input_file:tumor $tmp_tumor \
      --out $out_path/$chr.txt \
      --coverage_file $out_path/$chr.cov
```

Variants identified by MuTect are labelled as to whether they pass all MuTect filters, pass all filters *other* than the evidence threshold tlod_f_star, or fail to pass any filter other than tlod_f_star. Variants that pass all filters or fail only tlod_f_star are then passed to method for prior estimation and rescoring.

Real tumor data.

Acute Myeloid Leukemia We downloaded the whole genome sequence for aml31 BKM: citation and download date (Griffith et al., 2015). We merged the gold and platinum lists, and define is *not present* any variant for which deep sequencing was performed and zero alternate reads were observed. This is a very conservative metric.

What I call the cell paper Not sure if this will get used.

RESULTS

Results subsection one

Text. Text.

Results subsection two

Text. Text (see Table 1).

Text. Text (see Figure 2a).

Text. Text.

Results subsection three

Text. Text.

DISCUSSION

Discussion subsection one

Text. Text.

Table 1. This is a table caption

Col. head 1	Col. head 2 (%)	Col. head 3 (s^{-1})	Col. head 4 (%)	Col. head 5 (s^{-1})
Row 1	Row 1	Row 1	–	–
Row 2	Row 2	Row 2	Row 2	Row 2

This is a table footnote

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Discussion subsection two

Text. Text.

Text. Text.

Discussion subsection three

Text. Text.

Text. Text.

Text. Text.

CONCLUSION

Text. Text.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Conflict of interest statement. None declared.

REFERENCES

Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48(3):238–244, jan 2016. ISSN 1061-4036. doi: 10.1038/ng.3489. URL http://www.nature.com/doifinder/10.1038/ng.3489.

Ivana Bozic, Jeffrey M. Gerold, and Martin A. Nowak. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Computational Biology*, 12(2):e1004731, 2016. ISSN 15537358. doi: 10.1371/journal.pcbi.1004731.

Marc J Williams, Benjamin Werner, Timon Heide, Christina Curtis, Chris P Barnes, Andrea Sottoriva, and Trevor A Graham. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50(June):895–903, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0128-6. URL http://dx.doi.org/10.1038/s41588-018-0128-6.

J. Ding, A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon, S. Aparicio, and S. P. Shah. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, 28(2):167–175, jan 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr629. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr629.

Elaine R. Mardis. Applying next-generation sequencing to pancreatic cancer treatment. *Nature Reviews Gastroenterology & Hepatology*, 9(8):477–486, 2012. ISSN 1759-5045. doi: 10.1038/nrgastro.2012.126. URL http://www.nature.com/doifinder/10.1038/nrgastro.2012.126.

Xiang Chen, Elizabeth Stewart, Anang A. Shelat, Chunxu Qu, Armita Bahrami, Mark Hatley, Gang Wu, Cori Bradley, Justina McEvoy, Alberto Pappo, Sheri Spunt, Marcus B. Valentine, Virginia Valentine, Fred Krafcik, Walter H. Lang, Monika Wierdl, Lyudmila Tsurkan, Viktor Tolleman, Sara M. Federico, Chris Morton, Charles Lu, Li Ding, John Easton, Michael Rusch, Panduka Nagahawatte, Jianmin Wang, Matthew Parker, Lei Wei, Erin Hedlund, David Finkelstein, Michael Edmonson, Sheila Shurtleff, Kristy Boggs, Heather Mulder, Donald Yergeau, Steve Skapek, Douglas S. Hawkins, Nilsa Ramirez, Philip M. Potter, John A. Sandoval, Andrew M. Davidoff, Elaine R. Mardis, Richard K. Wilson, Jinghui Zhang, James R. Downing, and Michael A. Dyer. Targeting Oxidative Stress in Embryonal Rhabdomyosarcoma. *Cancer Cell*, 24(6):710–724, 2013. ISSN 15356108. doi: 10.1016/j.ccr.2013.11.002. URL http://dx.doi.org/10.1016/j.ccr.2013.11.002.

Mitesh J. Borad, Mia D. Champion, Jan B. Egan, Winnie S. Liang, Rafael Fonseca, Alan H. Bryce, Ann E. McCullough, Michael T. Barrett, Katherine Hunt, Maitray D. Patel, Scott W. Young, Joseph M. Collins, Alvin C. Silva, Rachel M. Condjella, Matthew Block, Robert R. McWilliams, Konstantinos N. Lazaridis, Eric W. Klee, Keith C. Bible, Pamela Harris, Gavin R. Oliver, Jaysheel D. Bhavsar, Asha A. Nair, Sumit Middha, Yan Asmann, Jean Pierre Kocher, Kimberly Schahl, Benjamin R. Kipp, Emily G. Barr Fritcher, Angela Baker, Jessica Aldrich, Ahmet Kurdoglu, Tyler Izatt, Alexis Christoforides, Irene Cherni, Sara Nasser, Rebecca Reiman, Lori Phillips, Jackie McDonald, Jonathan Adkins, Stephen D. Mastrian, Pamela Placek, Aprill T. Watanabe, Janine LoBello, Haiyong Han, Daniel Von Hoff, David W. Craig, A. Keith Stewart, and John D. Carpten. Integrated Genomic Characterization Reveals Novel,

Therapeutically Relevant Drug Targets in FGFR and EGFR Pathways in Sporadic Intrahepatic Cholangiocarcinoma. *PLoS Genetics*, 10(2), 2014. ISSN 15537390. doi: 10.1371/journal.pgen.1004135.

John M Findlay, Francesc Castro-Giner, Seiko Makino, Emily Rayner, Christiana Kartsonaki, William Cross, Michal Kovac, Danny Ulahannan, Claire Palles, Richard S Gillies, Thomas P Macgregor, David Church, Nicholas D Maynard, Francesca Buffa, Jean-Baptiste Cazier, Trevor A Graham, Lai-Mun Wang, Ricky A Sharma, Mark Middleton, and Ian Tomlinson. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nature Communications*, 7, 2016. doi: 10.1038/ncomms11111. URL https://www.nature.com/articles/ncomms11111.pdf.

D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, mar 2012. ISSN 1088-9051. doi: 10.1101/gr.129684.111. URL http://www.ncbi.nlm.nih.gov/pubmed/22300766 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3290792 http://genome.cshlp.org/cgi/doi/10.1101/gr.129684.111.

Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3): 213–219, 2013. ISSN 1087-0156. doi: 10.1038/nbt.2514. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3833702&tool=pmcentrez&rendertype=abstract http://www.nature.com/doifinder/10.1038/nbt.2514.

Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. jul 2012. URL http://arxiv.org/abs/1207.3907.

David E Larson, Christopher C Harris, Ken Chen, Daniel C Koboldt, Travis E Abbott, David J Dooling, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England)*, 28(3):311–7, feb 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr665. URL http://www.ncbi.nlm.nih.gov/pubmed/22155872 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3268238.

A. Roth, J. Ding, R. Morin, A. Crisan, G. Ha, R. Giuliany, A. Bashashati, M. Hirst, G. Turashvili, A. Oloumi, M. A. Marra, S. Aparicio, and S. P. Shah. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, 28(7):907–913, apr 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts053. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts053.

Alexis Christoforides, John D. Carpten, Glen J. Weiss, Michael J. Demeure, Daniel D. Von Hoff, and David W. Craig. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics*, 14:302, 2013. ISSN 14712164. doi: 10.1186/1471-2164-14-302.

David Jones, Keiran M. Raine, Helen Davies, Patrick S. Tarpey, Adam P. Butler, Jon W. Teague, Serena Nik-Zainal, and Peter J. Campbell. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Current Protocols in Bioinformatics*, 56(1):15.10.1–15.10.18, dec 2016. ISSN 19343396. doi: 10.1002/cpbi.20. URL http://doi.wiley.com/10.1002/cpbi.20.

Fatemeh Dorri, Sean Jewell, Alexandre Bouchard-Côté, and Sohrab P. Shah. Somatic mutation detection and classification through probabilistic integration of clonal population information. *Communications Biology*, 2(1):44, dec 2019. ISSN 2399-3642. doi: 10.1038/s42003-019-0291-z. URL http://www.nature.com/articles/s42003-019-0291-z.

Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England)*, 28(14):1811–7, jul 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts271. URL http://www.ncbi.nlm.nih.gov/pubmed/22581179.

Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, 40(22):11189–11201, 2012. ISSN 03051048. doi: 10.1093/nar/gks918.

Yuichi Shiraishi, Yusuke Sato, Kenichi Chiba, Yusuke Okuno, Yasunobu Nagata, Kenichi Yoshida, Norio Shiba, Yasuhide Hayashi, Haruki Kume, Yukio Homma, Masashi Sanada, Seishi Ogawa, and Satoru Miyano. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research*, 41(7):e89, 2013. ISSN 03051048. doi: 10.1093/nar/gkt126.

Moritz Gerstung, Christian Beisel, Markus Rechsteiner, Peter Wild, Peter Schraml, Holger Moch, and Niko Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3(May):811–818, 2012. ISSN 20411723. doi: 10.1038/ncomms1814. URL http://dx.doi.org/10.1038/ncomms1814.

Jian Carrot-Zhang and Jacek Majewski. LoLoPicker: detecting low allelic-fraction variants from low-quality cancer samples. *Oncotarget*, 8 (23):37032–37040, 2017. ISSN 1949-2553. doi: 10.1101/043612. URL www.impactjournals.com/oncotarget%0Awww.impactjournals.com/oncotarget/.

Yu Fan, Liu Xi, Daniel S.T. Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*, 17(1):178, 2016. ISSN 1474760X. doi: 10.1186/s13059-016-1029-6. URL http://dx.doi.org/10.1186/s13059-016-1029-6.

Serena Nik-Zainal, LudmilB. Alexandrov, DavidC. Wedge, Peter VanLoo, ChristopherD. Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, LucyA. Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, RichardRance, KingWai Lau, LauraJ. Mudie, Ignacio Varela, DavidJ. McBride, GrahamR. Bignell, SusannaL. Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, PatrickS. Tarpey, HelenR. Davies, Elli Papaemmanuil, PhilipJ. Stephens, Stuart McLaren, AdamP. Butler, JonW. Teague, Göran Jönsson, JudyE. Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerød, Andrew Tutt, JohnW.M. Martens, SamuelA.J.R. Aparicio, Åke Borg, AnneVincent Salomon, Gilles Thomas, Anne-Lise Børresen-Dale, AndreaL. Richardson, MichaelS. Neuberger, P.Andrew Futreal, PeterJ. Campbell, and MichaelR. Stratton. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993, may 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.04.024. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412005284.

Ludmil B Alexandrov, Philip H Jones, David C Wedge, Julian E Sale, Peter J Campbell, Serena Nik-Zainal, and Michael R Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47 (12):1402–1407, 2015. ISSN 1061-4036. doi: 10.1038/ng.3441. URL http://www.nature.com/doifinder/10.1038/ng.3441.

Henry Lee-Six, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, Brian J. P. Huntly, Inigo Martincorena, Elizabeth Anderson, Laura O'Neill, Michael R. Stratton, Elisa Laurenti, Anthony R. Green, David G. Kent, and Peter J. Campbell. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478, sep 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0497-0. URL http://www.nature.com/articles/s41586-018-0497-0.

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, jan 2013. ISSN 22111247. doi: 10.1016/j.celrep.2012.12.008. URL http://www.sciencedirect.com/science/article/pii/S2211124712004330?via%3Dihub.

Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*, 15(9):585–598, jul 2014. ISSN 1471-0056. doi: 10.1038/nrg3729. URL http://www.nature.com/doifinder/10.1038/nrg3729.

Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B. Alexandrov, Sancha Martin, David C. Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B. Brinkman, Sandro Morganella, Miriam R. Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E. Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A. Foekens, Moritz Gerstung, Gerrit K. J. Hooijer, Se Jin Jang, David R. Jones, Hyung-Yong Kim, Tari A. King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A. Purdie, Keiran Raine, Kamna Ramakrishnan, F. Germán Rodríguez-González, Gilles Romieu, Anieta M. Sieuwerts, Peter T. Simpson,

Rebecca Shepherd, Lucy Stebbings, Olafur A. Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G. Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T. Ueno, Christos Sotiriou, Alain Viari, P. Andrew Futreal, Peter J. Campbell, Paul N. Span, Steven Van Laere, Sunil R. Lakhani, Jorunn E. Eyfjord, Alastair M. Thompson, Ewan Birney, Hendrik G. Stunnenberg, Marc J. van de Vijver, John W. M. Martens, Anne-Lise Børresen-Dale, Andrea L. Richardson, Gu Kong, Gilles Thomas, and Michael R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605): 47–54, 2016. ISSN 0028-0836. doi: 10.1038/nature17676. URL http://www.nature.com/doifinder/10.1038/nature17676.

Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D. M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502 (7471):333–339, oct 2013. ISSN 0028-0836. doi: 10.1038/nature12634. URL http://www.nature.com/doifinder/10.1038/nature12634.

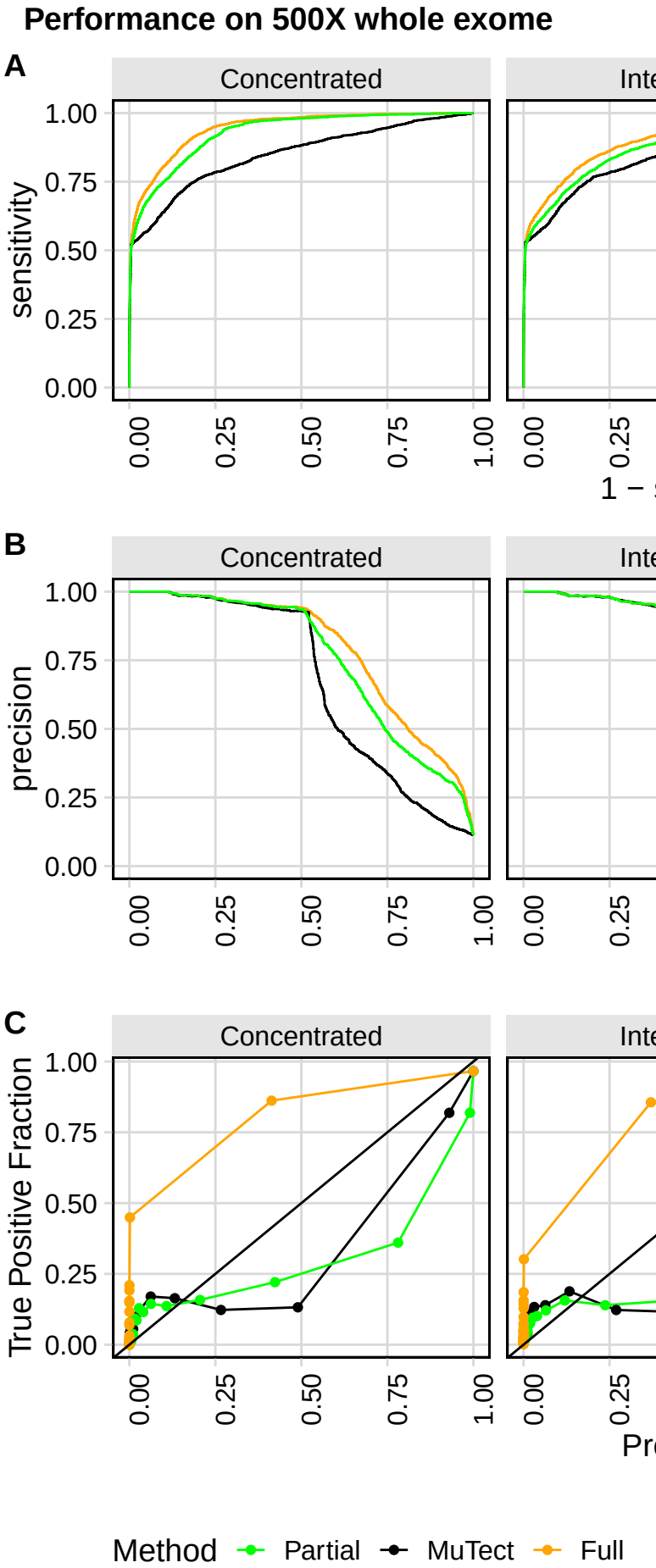
L. B. Alexandrov, Y. S. Ju, K. Haase, P. Van Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, A. Fujimoto, H. Nakagawa, T. Shibata, P. J. Campbell, P. Vineis, D. H. Phillips, and M. R. Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, nov 2016. ISSN 0036-8075. doi: 10.1126/science.aag0299. URL http://www.ncbi.nlm.nih.gov/pubmed/27811275 http://www.sciencemag.org/cgi/doi/10.1126/science.aag0299.

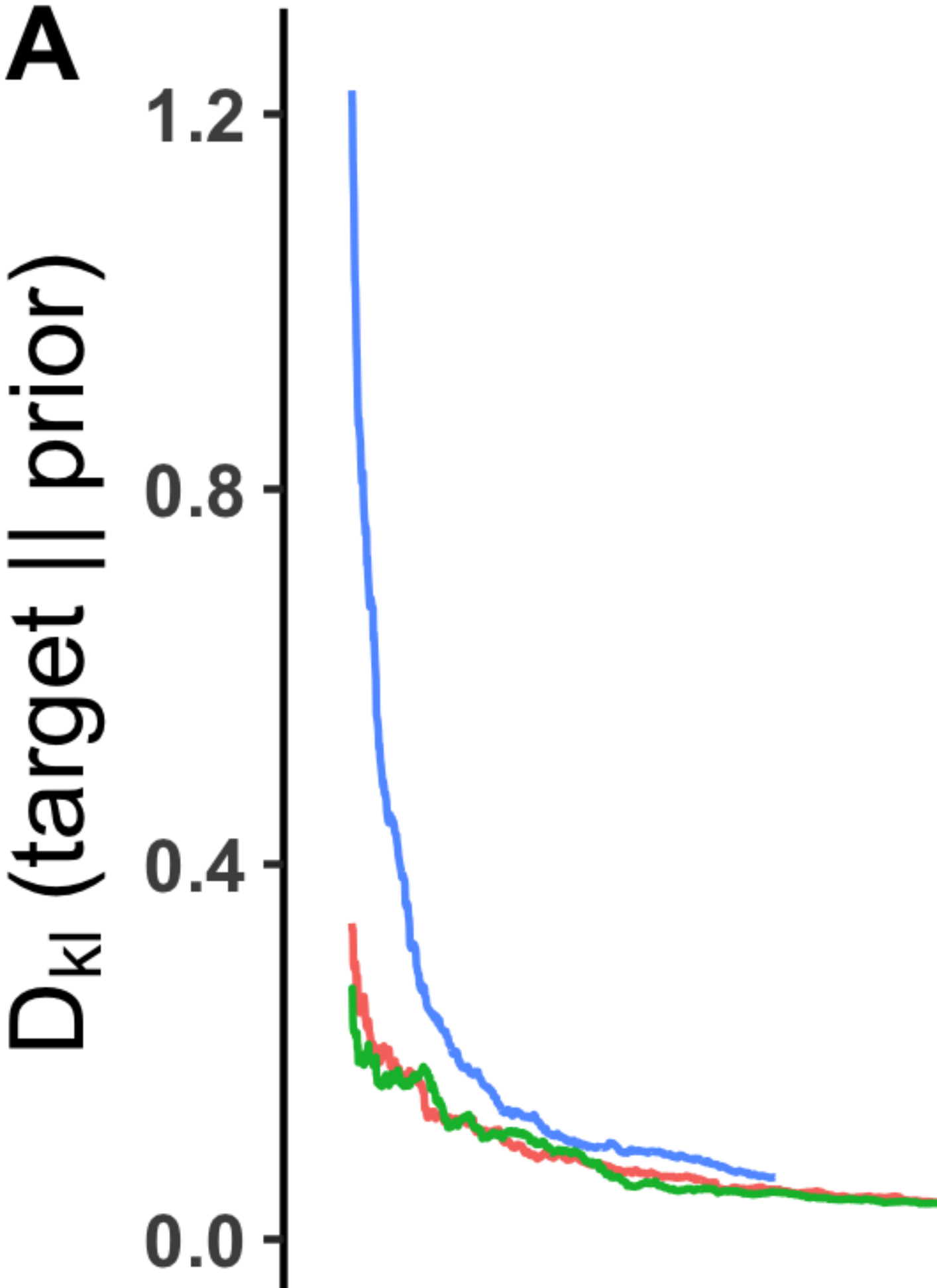
J. C. Mu, M. Mohiyuddin, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, 31(9):1469–1471, may 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu828. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu828.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, jul 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324.

Adam D Ewing, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y Sabelnykova, Michael R Kellen, Thea C Norman, David Haussler, Stephen H Friend, Gustavo Stolovitzky, Adam A Margolin, Joshua M Stuart, Paul C Boutros, Chenhao Li, Denis Bertrand, Niranjan Nagarajan, Qing-Rong Chen, Chih-Hao Hsu, Ying Hu, Chunhua Yan, Warren Kibbe, Daoud Meerzaman, Kristian Cibulskis, Mara Rosenberg, Louis Bergelson, Adam Kiezun, Amie Radenbaugh, Anne-Sophie Sertier, Anthony Ferrari, Laurie Tonton, Kunal Bhutani, Nancy F Hansen, Difei Wang, Lei Song, Zhongwu Lai, Yang Liao, Wei Shi, José Carbonell-Caballero, Joaquín Dopazo, Cheryl C K Lau, Justin Guinney, Michael R Kellen, Thea C Norman, David Haussler, Stephen H Friend, Gustavo Stolovitzky, Adam A Margolin, Joshua M Stuart, and Paul C Boutros. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nature Methods*, 12(7):623–630, jul 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3407. URL http://www.nature.com/articles/nmeth.3407.

Malachi Griffith, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, Ha X. Dang, Lee Trani, David E. Larson, Ryan T. Demeter, Michael C. Wendl, Joshua F. McMichael, Rachel E. Austin, Vincent Magrini, Sean D. McGrath, Amy Ly, Shashikant Kulkarni, Matthew G. Cordes, Catrina C. Fronick, Robert S. Fulton, Christopher A. Maher, Li Ding, Jeffery M. Klco, Elaine R. Mardis, Timothy J. Ley, and Richard K. Wilson. Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems*, 1(3):210–223, 2015. ISSN 24054712. doi: 10.1016/j.cels.2015.08.015. URL http://dx.doi.org/10.1016/j.cels.2015.08.015.




A
1.2
0.8
0.4
0.0
D_{KL} (target || prior)