Targets

- Bioinformatics
- Plus Comp. Bio
- BMC bioinformatics
- Nature Scientific Reports.
- Genome Biology (published MuSE (Fan et al. 2016))
- Nucleic Acids Research published EBCall in 2013

Introduction

Cancer is an evolutionary process, and understanding initiation, progression, and metastasis will require applications of evolutionary theory. One of the major tools in the evolutionary theory toolbox is the allele frequency spectrum. This allele frequency spectrum is constructed from single nucleotide variant calls in the tumor.

If tumors, as evidence suggests (Marc J Williams et al. 2016; Marc J Williams et al. 2018; Bozic, Gerold, and Nowak 2016), evolve essentially neutrally, then even driver mutations can't be expected to rise to high frequency during tumor evolution. As a result, finding mutations important to progression, resistance, and metastasis requires finding lower frequency mutations. Tumor heterogeneity has been associated with prognosis (1-4 in chuang paper) and the evolutionary trajectory helps identify the number of tumor subclones and their selective advantage.

The variant allele frequency spectrum that is currently used most often in cancer is truncated at a level above 5-10% because of difficulties in identifying low frequency variants.

- There are two main tracks in variant calling.
  - Heuristic filters
  - Statistical models of sequencing error
- We focus here on a model of mutation probability, including but not limited to sequencing error.
- Many types of callers, all assume there is no biological preference for mutation at a given site. Any site specific estimates are site specific sequencing/alignment error models(Xu 2018).
- Mutect2, FreeBayes and others are haplotype based callers
- Callers with site specific variant probabilities generate them either from other samples or through deep sequencing (deepSNV,EBCall,LoLoPicker). They are essentially generating a site specific sequencing error model, not a site specific probability of mutation
- Need to think about how the method applies to UMI (barcode) based sequencing, which are mostly deep targeted MuSE is continuous time

markov evolutionary model, still assuming no biological difference in site specific mutation probability(Fan et al. 2016)
- very little attention to the statistical model, either in competition or development
- there is useful biology.....
  - (Temko et al. 2018) links between mutational processes and driver mutations
  - (Van den Eynden and Larsson 2017) mutational signature critical for estimating selection
  - (Kandoth et al. 2013; Alexandrov et al. 2013) Underlying mutational processes generate tumor and tumor type specific mutation signatures
- Rather than using a constant probability for mutation, as other variant callers do, we convert that to an average or expected mutation probability, and compute the probability conditional on context and genome composition
- Poisson models make similar assumptions about the probability of an allele at a site. (Illumina technical note https://www.illumina.com/Documents/products/technotes/technote_so
- we simulate neutral tumor evolution, and assign vafs using a Beta(1,6) distribution
  - if M(f) is proportional to 1/f, then an exponential distribution is implied (Tarabichi et al. 2018; Marc J. Williams et al. 2017)(and the answering note by De, which also has a strong argument about why we need lower frequencies to do evolutionary inference). We choose a beta distribution to draw vafs and tuned to achieve a slightly fatter distribution in the 2-5% range in which we are most interested.
- Need a list of why evolutionary inference on tumors is important. Resistance, virulence(heterogeneity), biology (mutation rate/signature/microenvironment).

Results

MuTect computes the probability of a mutation from reference allele r to base m as a function of base calls b, estimated allele frequencies f, and per base error probabilities e. The probability that a given base is correctly called can be written as

$$P(b_i \mid e_i, r, m, f) = \begin{cases} f\frac{e_{b_i}}{3} + (1-f)(1-e_{b_i}) & b_i = r \\ f(1-e_{b_i}) + (1-f)\frac{e_{b_i}}{3} & b_i = m \\ \frac{e_{b_i}}{3} & otherwise. \end{cases}$$

Now consider two models for the data. Model M_0 in which there are no variants at a site, and M_f^m where allele m is present at allele fraction f. Assuming reads are independent the likelihood of the model given the data is

$$\mathcal{L}(M_f^m) = P(\{b_i\} \mid \{e_{b_i}\}, r, m, f) = \prod_{i=1}^{d} P(b_i \mid e_{b_i}, r, m, f)$$

and the probability of M_f^m can be written

$$P(m, f \mid \{b_i\}, \{e_{b_i}\}, r) = \mathcal{L}(M_f^m) \frac{P(m, f)}{P(\{b_i\} \mid \{e_{b_i}\}, r)}.$$

We can also express this probability in terms of the model M_0

$$1 - P(m, f \mid \{b_i\}, \{e_i\}, r) = \mathcal{L}(M_0) \frac{1 - P(m, f)}{P(\{b_i\} \mid \{e_{b_i}\}, r)}.$$

Taking the log of the ratio of the two previous equations gives the log odds in favor of M_f^m, and some cancellation yields

$$LOD_T(m, f) = \log_{10} \left(\frac{\mathcal{L}(M_f^m) P(m, f)}{\mathcal{L}(M_0^m)(1 - P(m, f))} \right).$$

A classifier for variants is constructed by selecting an odds threshold δ_T and labeling variants satisfying the condition

$$LOD_T(m, f) = \log_{10} \left(\frac{\mathcal{L}(M_f^m) P(m, f)}{\mathcal{L}(M_0^m)(1 - P(m, f))} \right) \geq \log_{10} \delta_T$$

as true variants, and rejecting them otherwise. Note that the expression for $LOD_T(m, f)$ can be further factorized as the sum of the log-likelihood ratio of the two models and the log odds of the prior for M_f^m. Current variant callers calculate this prior by assuming the allele and its frequency are independent, and that $f \sim \mathrm{U}(0, 1)$, so that $P(f) = 1$. If all substitutions are equally likely, then $P(m) = \mu/3$ where $\mu = 3 \times 10^{-6}$, the estimated per-base mutation rate in tumors. Given these assumptions the log prior odds is a constant, and the classifier can be re-written as

$$LOD_T(m, f) = \log_{10} \left(\frac{\mathcal{L}(M_f^m)}{\mathcal{L}(M_0^m)} \right) \geq \log_{10} \delta_T - \log_{10} \left(\frac{P(m)}{1 - P(m)} \right) \geq \theta_T.$$

If $\delta_T = 2$, i.e the odds in favor of M_f^m is 2, then $\theta_T = 6.3$, and this is the threshold implemented in MuTect 1.

The conditional probability that a mutation to allele m will occur given a specific genomic context C, $P(m \mid C)$ can be computed from the empirical data in Figure **??**, but $P(M \mid C)$ can not be. Using Bayes rule we can rewrite$P(m \mid C)$ as

$$P(m \mid C) = P(C \mid m)\frac{P(m)}{P(C)}.$$

Now $P(C \mid m)$ is the mutation spectrum of the tumor, $P(m) = \mu$, and $P(C)$ can be estimated as the frequency of context C in the genome. The new expression for the log odds is

$$LOD_T(m, f) = \log_{10}\left(\frac{\mathcal{L}(M_f^m)P(m \mid C)}{\mathcal{L}(M_0^m)(1 - P(m \mid C))}\right).$$

Sensitivity in real data

We examined two real tumor datasets in which variants had been validated by deep targeted resequencing (M. Griffith et al. 2015; W. Shi et al. 2018). M. Griffith et al. (2015) performed whole genome sequencing of an acute myeloid leukemia to a depth of ~312X, called variants with seven different variant callers and validated over 200,000 variants by targeted re-sequencing to a depth of ~1000X. This led to a platinum set of variant calls containg 1,343 SNVs. We obtained BAM files from this experiment and called variants using MuTect 1.1.7, then compared the sensitivity of the calls between MuTect and our method (Figure 1A). At any relevant threshold our method is slightly more sensitive than MuTect. MuTect is unable to recover 100% of the calls due to hueristic filtering and other differences between MuTect and the other variant callers used.

W. Shi et al. (2018) performed multi-region sequencing of 6 breast tumors to evaluate the effects of variant calling and sequencing depth on estimates of tumor heterogeneity, validating 1,385 somatic SNVs. As with the leukemia we obtained BAM files for this experiment and compared our method to raw MuTect calls (Figure 1B). We again find that our method is more sensitive than MuTect across the full range of relevant thresholds.

Sensitivity and specificity in simulated data

In order to describe the operating characteristics of our score as a classifier compared to MuTect, we simulated six tumors (see methods), three 100X whole genomes and three 500X whole exomes, with three differnent mutation spectra(methods). In WES simulations the relatively smaller number of variants, and consequent lower number of very low frequency variants, causes the methods to perform similarly, but our method is slightly more sensitive and has slightly higher AUROC than raw MuTect scores. The large number of mutations present and at low frequency in whole genome simulations provide a clearer demonstration of the benefits of the method. The portion of the ROC curve for our method is substantially higher than the curve for MuTect, and the MuTect curve is

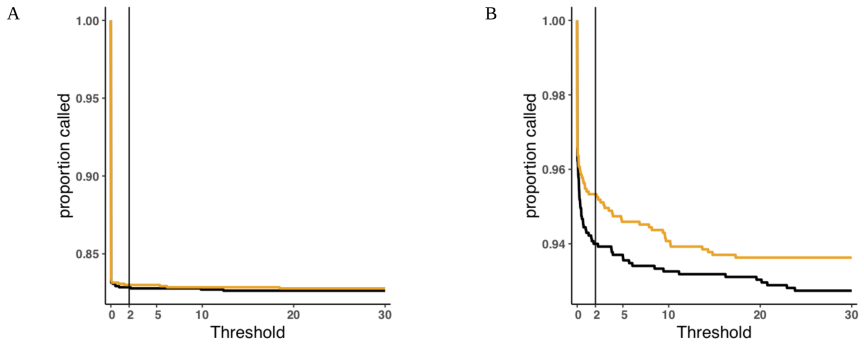

Figure 1: Sensitivity in real tumors. A) AML31 platinum SNV calls (M. Griffith et al. 2015). B) Validated SNV in 6 breast cancers(W. Shi et al. 2018).

essentially linear, is due to the effect of the prior. The prior is lowering scores of false positive mutations and raising the scores of true positives in this region. (This is super inelegant{bkm}).

Convergence of the prior to simulated target distributions.

In both whole genome (Figure 3) and whole exome (supplement) simulations, the estimated mutation spectrum is very close to the simulated spectrum. The conditional probability of mutation at a given site averaged over all sites is 3e-6 (the $P(m) = \mu$ used by MuTect; important that this is averaged over every site in the genome. The probability here includes estimates of the context content of the genome $P(m \mid C) = P(C \mid m) * P(m)/P(C)$), but our method overweights some contexts and underweights others in line with the data generating distribution. (I think I need an exome too. I have the B figure, but need to generate the C figure{bkm}) Supplementary figures for other target distributions? Or a different type of figure than we have here? Or something else? We get what we would expect with other simulated spectra. The prior is as sharp or diffuse as the data generating process.

- The performance of the method is always better, but the amount of benefit is directly tied to the concentration of the spectrum

Discussion

- Must include a strong argument for better real tumor validation sets. Focus on false negatives as well as false positives. The aml31 paper gets alot

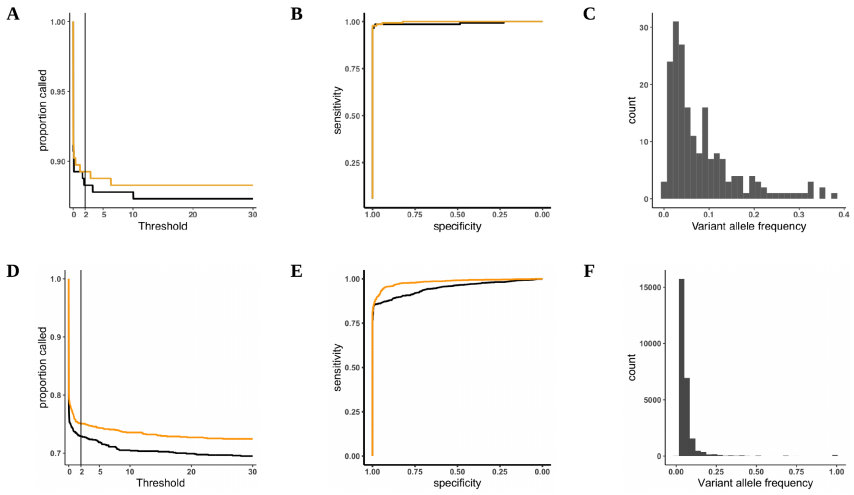

Figure 2: Sensitivity in simulated tumors. A-C) Whole exome simulation. D-F) Whole genome simulation

of them, but if they had for instance just used mutect to identify any potential variant that passed all other heuristic filters they would have a better sense of false negative rates.

- Why are false negative rates important?
- heterogeneity
- selection inference
- rare but druggable variant identification

Methods

Variant allele frequency distribution

- The allele frequency spectrum of a particular tumor is determined by intrinsic factors including mutation rate and the action of natural selection.
- The theoretical neutral distribution is $M(f) \approx 1/f$ (Bozic, Gerold, and Nowak 2016), which creates a roughly decreasing exponential shape on $[0, 1]$ for allele frequency.
- We chose a Beta(1,6) distribution to simulate a roughly neutral evolutionary trajectory while providing a significant fraction of variants in the 1% - 5% range where discrimination is most difficult.
- 20% of variants have frequency between .017 and .057. 50% are less than .1

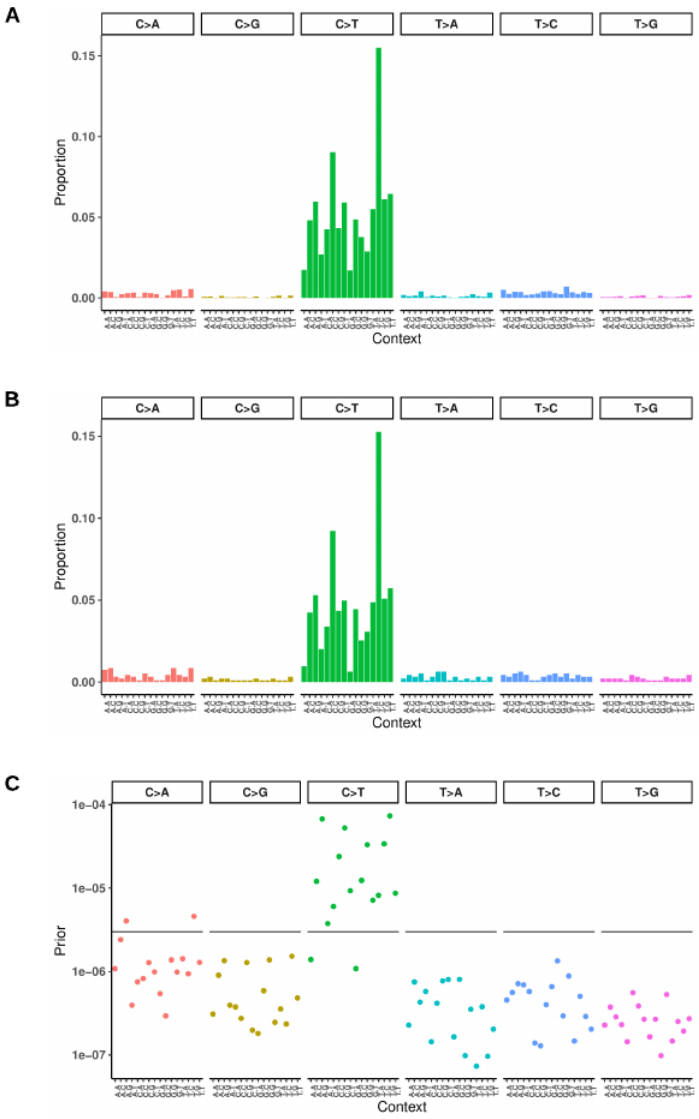

Figure 3: Prior probability of mutation estimated from high confidence calls. A) The simulated mutation spectrum (1,7,11). B) The maximum likelihood estimate of the data generating distribution (Dirichlet). C) The conditional probability of mutation at a site given its genomic context (bar at 3e-6, the global estimate of mutation rate)

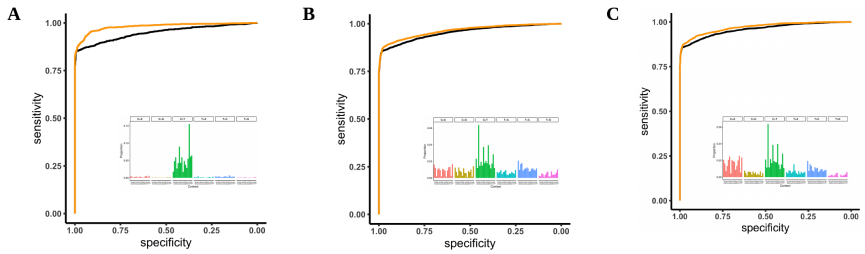

Figure 4: Effect of spectrum concentration on results in WGS. A) 1,7,11 B) 1,3,5 C) 1,4,5

Simulated tumors spectra

- 100X whole genome and 500X whole exome for each of three signatures
- Real mutations from TCGA and PCAWG
- 1,7,11 UV (Very concentrated at C>T)
- 1,4,5 Tobacco (Slight concentation at C>A and C>T)
- 1,3,5 Breast (diffuse)

Simulated bam files

- 100X normal and 500X exome reads simulated with VarSim/art (Mu et al. 2015) (default parameters?) and aligned with BWA (H. Li and Durbin 2009).(default parameters)
- Variants spiked with Bamsurgeon with default parameters (Ewing et al. 2015).
- Variants called with MuTect 1.1.7 with specific parameters (Cibulskis et al. 2013). (list them, just copy in as code is what I would prefer to see if I was reading the paper).

```
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
```

Figures

References

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Peter J. Campbell, and Michael R. Stratton. 2013. "Deciphering Signatures of Mutational Processes Operative in Human Cancer." *Cell Reports* 3 (1). Cell Press: 246–59. doi:10.1016/j.celrep.2012.12.008.

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

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples." *Nature Biotechnology* 31 (3). Nature Publishing Group: 213–19. doi:10.1038/nbt.2514.

Ewing, Adam D, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, et al. 2015. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection." *Nature Methods* 12 (7). Nature Publishing Group: 623–30. doi:10.1038/nmeth.3407.

Fan, Yu, Liu Xi, Daniel S.T. Hughes, Jianjun Zhang, Jianhua Zhang, P. Andrew Futreal, David A. Wheeler, and Wenyi Wang. 2016. "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). Genome Biology: 178. doi:10.1186/s13059-016-1029-6.

Griffith, Malachi, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, et al. 2015. "Optimizing Cancer Genome Sequencing and Analysis." *Cell Systems* 1 (3). Elsevier Inc.: 210–23. doi:10.1016/j.cels.2015.08.015.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. "Mutational landscape and significance across 12 major cancer types." *Nature* 502 (7471). Nature Research: 333–39. doi:10.1038/nature12634.

Li, H., and R. Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25 (14). Oxford University Press:

1754–60. doi:10.1093/bioinformatics/btp324.

Mu, J. C., M. Mohiyuddin, J. Li, N. Bani Asadi, M. B. Gerstein, A. Abyzov, W. H. Wong, and H. Y. K. Lam. 2015. "VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications." *Bioinformatics* 31 (9). Oxford University Press: 1469–71. doi:10.1093/bioinformatics/btu828.

Shi, Weiwei, Charlotte K Y Ng, Raymond S Lim, Lajos Pusztai, Jorge S Reis-Filho, Christos Hatzis, Tingting Jiang, et al. 2018. "Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity In Brief Article Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity." *CellReports* 25: 1446–57. doi:10.1016/j.celrep.2018.10.046.

Tarabichi, Maxime, Iñigo Martincorena, Moritz Gerstung, Armand M Leroi, Florian Markowetz, Paul T Spellman, Quaid D Morris, Ole Christian Lingjærde, David C Wedge, and Peter Van Loo. 2018. "Neutral tumor evolution?" *Nature Genetics*. doi:10.1038/s41588-018-0258-x.

Temko, Daniel, Ian P M Tomlinson, Simone Severini, Benjamin Schuster-böckler, and Trevor A Graham. 2018. "The effects of mutational processes and selection on driver mutations across cancer types." *Nature Communications* 9. Springer US: 1857. doi:10.1038/s41467-018-04208-6.

Van den Eynden, Jimmy, and Erik Larsson. 2017. "Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric." *Frontiers in Genetics* 8: 74. doi:10.7908/C11G0KM9.

Williams, Marc J, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. 2016. "Identification of neutral tumor evolution across cancer types." *Nature Genetics* 48 (3). Nature Research: 238–44. doi:10.1038/ng.3489.

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

Williams, Marc J., Benjamin Werner, Chris P. Barnes, Trevor A. Graham, and Andrea Sottoriva. 2017. "Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures." *Nature Genetics* 49 (9): 1289–91. doi:10.1038/ng.3877.

Xu, Chang. 2018. "A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data." *Computational and Structural Biotechnology Journal* 16. The Authors: 15–24. doi:10.1016/j.csbj.2018.01.003.