Published in final edited form as:

Nat Genet. 2018 December ; 50(12): 1626–1628. doi:10.1038/s41588-018-0219-4.

Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomics data

Abdul Balaparya¹, Subhajyoti De^{1,*}

¹Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901, USA.

Tumors arise as a result of an evolutionary process operating on somatic cells within tissues, and thus identifying the modes of cancer evolution is of fundamental interest. Analyzing whole exome and genome sequencing data for 904 samples from 14 different cancer types and cohorts, Williams et al,¹ reported that mutant allele frequencies of about a third (323/904) of the samples follow a simple power-law distribution (mean R²: ~0.9 – 0.98 for each cohort) and inferred that subclones in those tumors are under neutral evolution. In this study the authors used mutant allele frequency from whole genome and exome sequencing data as a direct proxy for relative fraction of tumor cells carrying that mutation, and fitted power law directly to the distributions of mutant allele frequency estimates. This can introduce different types of biases. Various steps of the genomic pipeline including tumor DNA extraction, PCR amplification, and sequencing protocols introduce variability² such that allelic proportional read counts do not automatically indicate clonal proportion, and instead follow a distribution with the underlying clonal proportion as a prior in diploid regions. Additional variation in allele frequency estimates could arise due to copy number alterations and pathological tumor purity. Williams et al,¹ selected only high purity tumors (> 70%), and analyzed regions with copy number absolute log R ratio > 0.5, but these filters were lenient, and they did not explicitly include purity and ploidy estimates in their model. As we show below, collectively these attributes can contribute to moderate level of over-dispersion of allele frequency distribution of somatic mutations, and present challenges while fitting power law distribution.

Analyzing copy number and exome-sequencing data for >300 paired tumor-normal samples from the Cancer Genome Atlas (TCGA)³ bladder, lung, and colon cancers, and focusing on the genomic regions with absolute log R ratio > 0.5 (see Supplementary Information for details of methods), we observe that a majority of them have greater dispersion in read count based allelic frequency distribution of germ line SNPs in the tumor genomes compared to that in the matched normal genomes. This is partly due to low allele frequency somatic subclonal copy number events (e.g. absolute log R ratio > 0.5) in the tumor genomes, as

Permission Reprints and permissions information is available at www.nature.com/reprints.

*Authors for correspondence. subhajyoti.de@rutgers.edu.

Author contributions A.B. contributed to data analysis. S.D. conceived the project, contributed to data analysis, interpreted the results, and wrote the manuscript.

Code and data availability Method and supporting data are available as Supplementary Information.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Competing interests The authors declare no competing interests.

shown in Figure 1A-B. We then examined the effects of dispersion of somatic allele frequency distribution in tumor genomes on the power law fit. Following an established approach⁴, we fitted beta binomial model to the allele frequency distributions of germ line variants to the TCGA samples to account for over-dispersion and estimated model parameters, and then applied similar dispersion parameters to realistically model read-count based allelic frequencies for somatic variants in tumor genomes with one, two, or a few sub-clones. Using representative examples we demonstrate that power law distribution show excellent fit to the allele frequency distributions (R^2 : 0.9 – 99) within the allele frequency range (0.12-0.25) used by Williams et al.¹, even when the allele frequency data was derived from simulated tumor genomes with 1-3 sub-clones and 80X median coverage (Figure 1C-D). Our results were not sensitive to model selection. For instance, we also obtained good fit to power law when over-dispersion was ignored and binomial distribution was used to simulate allele frequency distributions in tumor genomes with 2-3 sub-clones and 80X median coverage (Figure 1E-F). Synthetic genomes may not capture different levels of complexities of biological data from real human tumors. Therefore, we further obtained data on the estimates of the number of clones in the same tumor from a published study⁵, which took into account point mutations, purity estimates, and copy number alterations to estimate clonality. Among the TCGA tumors samples identified by Williams et al.¹ to have strong signatures of neutral evolution ($R^2_{all} > 0.99$), 27/79 (35%) and 21/79 (26%) were reported to have 2-4 clones (oligoclonal) and 8 clones (polyclonal), respectively. These observations suggest that significant fit to power law within the allele frequency range (0.12-0.25) could be achieved under multiple alternate hypotheses, and null hypothesis may not be reliably rejected. Thus, caution should be taken while inferring neutral evolution in the light of the complexity of genomic data. It is plausible that each subclone comprises of populations of tumor cells with genetic heterogeneity among themselves, but the TCGA data did not have sufficient depth of coverage and resolution to address that.

Second, depth of coverage for whole exome and genome sequencing data of majority of the cohorts chosen including the TCGA³ only allows detecting somatic mutations with allele frequency >10% (or >5% in some cases). At this allele frequency resolution it is possible to detect only the principal clonal branches, and the depth of sequencing coverage in this data is not suitable for detecting the ‘twigs’ and ‘leaves’. For instance, at allele frequency threshold of 0.1, it is only possible to detect heterozygous mutations in clones that constitute >20% of the tissue mass, even if issues due to tumor purity and variation in read counts are adjusted for. So, beyond the major branches, this data provides no information about fractal topology of the twigs and leaves in the evolutionary tree in Figure 5 of Williams et al.¹, and no inference about $M(f)$ vs $1/f$ could be made for those sub-regions of the tree.

Nonetheless, Williams et al.¹ have provided a valuable conceptual framework to integrate tumor genomic data with theoretical population dynamics to identify and compare modes of tumor evolution, which would stimulate further advances in this area, and may have future implications for clinical management of cancer patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

SD acknowledges support from NIH grant P30CA072720.

References:

1. Williams MJ, Werner B, Barnes CP, Graham TA & Sottoriva A Identification of neutral tumor evolution across cancer types. *Nat. Genet* 48, 238–244 (2016). [PubMed: 26780609]
2. Sims D, Sudbery I, Iltott NE, Heger A & Ponting CP Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet* 15, 121–32 (2014). [PubMed: 24434847]
3. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45, 1113–1120 (2013). [PubMed: 24071849]
4. Bansal V A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 26, i318–24 (2010). [PubMed: 20529923]
5. Andor N et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* 22, 105–113 (2016). [PubMed: 26618723]

Figure 1.

A) Copy number alterations, and the extent dispersion of germ line and somatic allele frequency distributions in a representative tumor TCGA-05-4432. The dotted lines indicate log R ratio 0.5 and -0.5. B) Histogram showing distributions of non-reference allele frequency estimates for heterozygous germ line SNPs in the tumor and paired normal genomes in the genomic regions with absolute log R ratio ≥ 0.5 for TCGA-05-4432. C-D) Histogram showing the somatic allele frequency (*f*) distributions for simulated tumors with 1 and 3 sub-clones, respectively, modeled using beta binomial distribution to account for over-dispersion. Model parameters are shown in the legend. Cumulative frequency distributions of the somatic variants against 1/*f* are shown in the inset. E-F) Histogram showing the somatic allele frequency (*f*) distributions for simulated tumors with 2 and 3 sub-clones, respectively, ignoring over-dispersion and modeled using binomial distribution. Model parameters are shown in the legend. Cumulative frequency distributions of the somatic variants against 1/*f* are shown in the inset.