Bioinformatics, 32(6), 2016, 808–813
doi: 10.1093/bioinformatics/btv685
Advance Access Publication Date: 20 November 2015
Original Paper

Genome analysis

SomVarIUS: somatic variant identification from unpaired tissue samples

Kyle S. Smith[1,2,3,†], Vinod K. Yadav[1,†], Shanshan Pei[4], Daniel A. Pollyea[1,4], Craig T. Jordan[1,4] and Subhajyoti De[1,2,5,*,†]

[1]Department of Medicine, [2]Department of Pharmacology, [3]Computational Biosciences Training Program, University of Colorado School of Medicine, Aurora, CO, USA, [4]University of Colorado Cancer Center, Aurora, CO, USA and [5]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

†The authors wish it to be known that, in their opinion, the first two authors and last author should be regarded as Joint Authors

Received on July 14, 2015; revised on November 12, 2015; accepted on November 13, 2015

Abstract

Motivation: Somatic variant calling typically requires paired tumor-normal tissue samples. Yet, paired normal tissues are not always available in clinical settings or for archival samples.

Results: We present SomVarIUS, a computational method for detecting somatic variants using high throughput sequencing data from unpaired tissue samples. We evaluate the performance of the method using genomic data from synthetic and real tumor samples. SomVarIUS identifies somatic variants in exome-seq data of $\sim$150 $\times$ coverage with at least 67.7% precision and 64.6% recall rates, when compared with paired-tissue somatic variant calls in real tumor samples. We demonstrate the utility of SomVarIUS by identifying somatic mutations in formalin-fixed samples, and tracking clonal dynamics of oncogenic mutations in targeted deep sequencing data from pre- and post-treatment leukemia samples.

Availability and implementation: SomVarIUS is written in Python 2.7 and available at http://www.sjdlab.org/resources/

Contact: subhajyoti.de@ucdenver.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Assessment of somatic and germ line mutations to tailor personalized diagnosis and treatment is becoming a corner stone for precision medicine in oncology initiatives (Collins and Varmus, 2015; Sboner and Elemento, 2015). Somatic mutations are typically detected by comparing sequencing data from target (e.g. tumor) and matched control tissues (e.g. benign tissue from the same patient). Initiatives such as the ICGC-TCGA DREAM Genomic Mutation Calling Challenge have shown that popular algorithms can detect somatic mutations from paired tumor-normal tissue samples with reasonably high accuracy (Boutros *et al.*, 2014). However, in many practical situations matched control tissues are not available, and it remains challenging to distinguish between somatic and germ line variants in those cases. For instance, matched normal samples are not routinely obtained in clinical care (Meric-Bernstam *et al.*, 2015). Additionally, a considerable proportion of the older FFPE and fresh frozen tumor samples have no matched control tissues available. Popular variant calling tools are not designed to identify somatic and germ line variants from unpaired tissue samples, and there have been only limited efforts to detect somatic mutations from unpaired samples (Sun *et al.*, 2014).

We present SomVarIUS, a computational method for detecting somatic variants using high throughput sequencing data from unpaired tissue samples. SomVarIUS accepts sorted alignment files (.bam) as input, and outputs predicted somatic mutations in the variant call format (.vcf), thereby enabling its easy integration into any standard genome analysis pipeline. It also generates an optional output providing information about the status of known

cancer/disease-associated mutations in the sample. We first describe the method, before showing its application by identifying somatic mutations in FFPE samples, and tracking clonal dynamics of oncogenic mutations in pre- and post-treatment leukemia samples.

2 Methods

The SomVarIUS workflow is shown in Figure 1A. The working principle of the framework is that, non-reference bases in high-throughput sequencing reads mapped to a genomic site are probably of germ line or somatic origin, or indicate sequencing error (Fig. 1B). Therefore, one can identify potential somatic mutations after calculating probabilities that the observed, non-reference bases at that site are unlikely due to sequencing error and are also not germ line SNPs. Accordingly, the SomVarIUS workflow, as outlined in Figure 1A, involves three key steps—initial data pre-processing to prioritize potential variant sites, estimation of the probability of sequencing error and also the probability of observing a germ line SNP at those sites.

In the pre-processing step, SomVarIUS applies multiple filters to prioritize variants. The pipeline excludes the reads that have low mapping quality (default ≤ 55), low base quality at that position (default ≤ 13), and those which are not proper pairs (for paired end sequencing). The genomic locations that overlap with SNPs in dbSNP (v142), have low coverage (default ≤ 10), or have very few reads supporting the non-reference allele (default ≤ 4) in the pileup are not evaluated for somatic mutation discovery. Duplicate reads can be optionally removed.

The subsequent step involves additional filters and ascertainment of the variant alleles. Major and minor alleles are identified based on their respective allele frequencies; if multiple non-reference bases are present in the pileup at a genomic location, the top two alleles in terms of allelic abundance are considered. While detecting somatic substitution, the reads that harbor base insertions and deletions in the pileup at that particular genomic site are not considered (e.g. do not contribute to allelic abundance estimation). The non-reference allele is required to have a high mean base quality (default Phred score: ≥ 25). Strand bias is estimated using the published SB score (Guo *et al.*, 2012). Both alleles should be supported by reads mapping to both the forward and reverse strand, and have low strand bias (default: ≥ 0.8) score (Guo *et al.*, 2012), since extreme strand bias typically indicates a potential false positive.

Thereafter, SomVarIUS utilizes information from base quality scores, coverage information and applies Chernoff bound (Chernoff, 1952), to determine the probability of sequencing error at a genomic position using an approach similar to that of Bansal (2010). In brief, let a genomic base position be covered by n reads, of which s reads

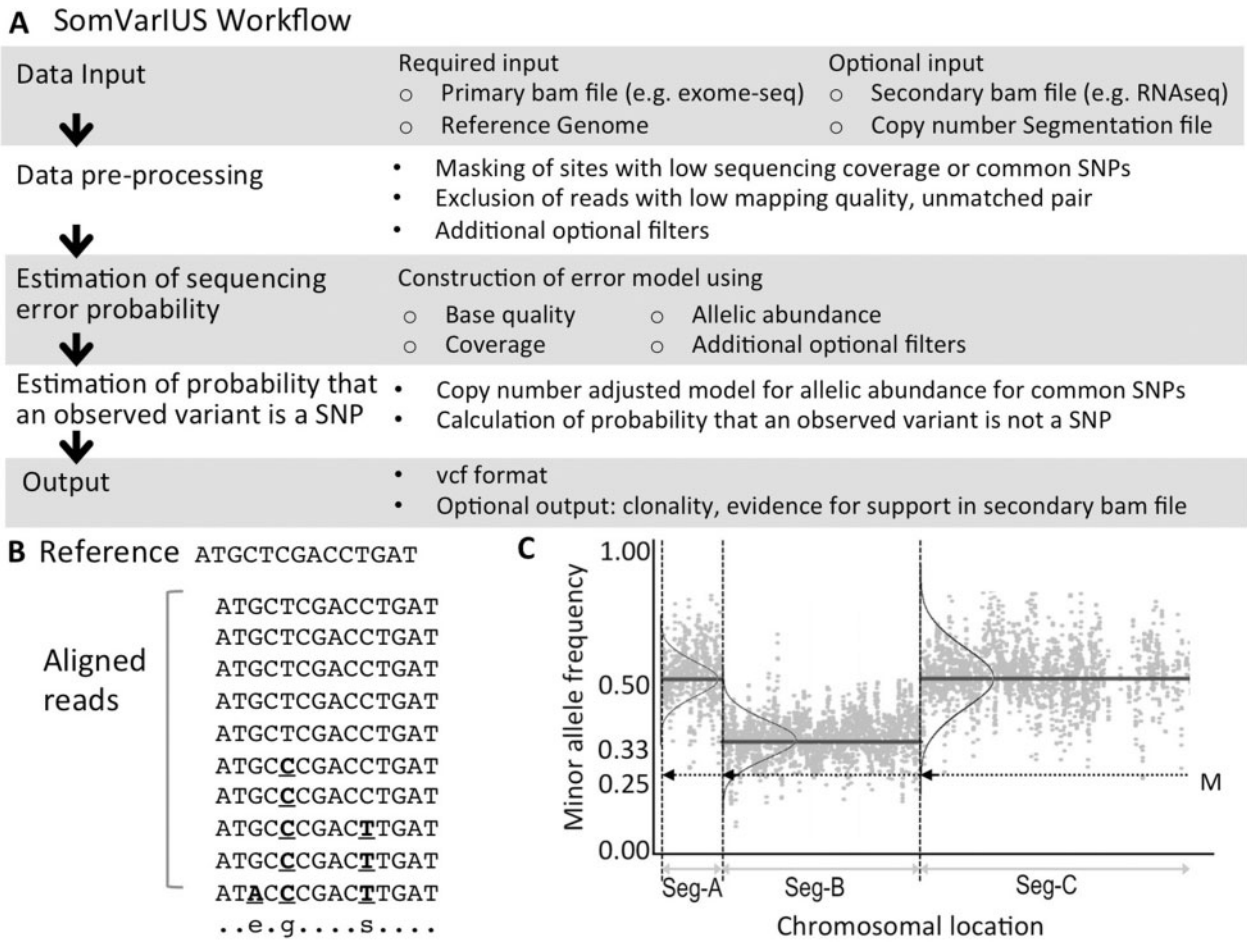

Fig. 1. (**A**) A schematic representation of SomVarIUS workflow. (**B**) A schematic representation showing reads mapped to the reference genome, where sites of potential sequencing error (*e*), germ line SNP (*g*) and somatic mutation(*s*) are marked. (**C**) Allelic abundance of germ line SNPs (grey dots) is shown for copy neutral (Seg-A and Seg-C) and copy number altered (Seg-B) regions. This is taken into consideration while assessing whether a variant with allelic abundance *M* would be considered as of potential somatic or germ line origin

are supporting the alternate allele. Let $\{Q_i\}$ be the quality scores of the base calls for the n reads at that position where $1 \le i \le n$. Assuming base calls are independent between reads, and by modeling their sequencing error probabilities as n independent Bernoulli trials, we can compute the mean expected sequencing error at that genomic position as $\mu = \Sigma_i 10^{-0.1 \times Q_i}$. The probability of the sum $X = X_1 + X_2 + \cdots + X_n$ of n independent Bernoulli random variables deviating from its mean μ is bounded by Chernoff's equation (Bansal, 2010; Chernoff, 1952).

$$\Pr[X > (1+\delta)\mu] < \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

If we apply the transformation, $s = (1+\delta)\mu$, then this probability, denoted as p_s indicates the upper bound of the probability of observing s out of n reads covering the base position to support the alternate allele due to sequencing error alone. In SomVarIUS, a threshold for probability of sequencing error, given the other filters mentioned above, helps prioritize high confidence biological variants.

Next, SomVarIUS estimates the probability of observing a potential germ line SNP given the allelic abundance of the major and minor alleles, in addition to information about known germ line heterozygous SNPs in the neighborhood (Fig. 1C). It first interrogates the known SNP positions from the input file, and identifies those, which are heterozygous in the sample based on detection of both the reference and alternate alleles. We used the SNPs from dbSNP v142 after excluding the COSMIC variants. But a user may choose to provide an alternate SNP file (e.g. 1000 Genome SNPs with population allele frequency >5%) as input. A SNP position is classified as heterozygous if both the reference and alternate alleles are detected with sufficient read support (e.g. default: >4 reads supporting both alleles; or higher than that expected due to sequencing error). SomVarIUS then utilizes those positions to model the distribution of allelic abundance for the heterozygous SNPs in the sequencing data using a beta-binomial distribution, where for the read count for the alternate allele, Y, the model is

$$Y|\alpha, \beta \sim BetaBinomial(n, \alpha, \beta)$$

where n is the number of reads covering the base position, and α, β are beta binomial model parameters. The model parameters (α, β) are estimated using the method of moments (Tripathi *et al.*, 1994). SomVarIUS takes into account local copy number status implicitly based on allelic abundance of heterozygous germ line SNPs for the entire chromosomes or genomic segments within each chromosome. For instance, at a triploid locus allelic abundance for the germ line, minor allele is expected to be ~0.33 and not 0.5, as expected in the diploid regions. If a copy number segmentation file is provided as an input (optional), the beta-binomial distribution is fitted to each genomic segment to account for locus-specific copy number status and other biases. Next, this distribution is used to evaluate the germ line mutation status of other candidate positions in the genomic segment. If at a candidate genomic position, the allelic abundance of the minor allele is at the tail of the beta-binomial distribution for the underlying genomic segment, this position will have a small probability (p_g) to be a germ line variant. The beta-binomial allows us to model over-dispersion of allelic abundance in the sequencing data based on information from the genomic segment. However, if the number of heterozygous common SNPs in the segment is small (<5), p_g is computed based on binomial distribution or beta-binomial constructed using heterozygous germ line SNPs built from the entire chromosome. A threshold for p_g i.e. probability of observing a germ line SNP, given the other filters mentioned above, helps prioritize variants that are unlikely to be germ line variants.

We classify a genomic position to harbor somatic mutation, if that position has a small sequencing error probability (default: $p_s < 1E\text{-}05$) as well as a small probability to be a germ line variant ((default: $p_g < 5E-02$). The output somatic variants are reported in the vcf format. SomVarIUS also provides additional information, which could be useful in applied settings. First, SomVarIUS accepts an optional high throughput sequencing dataset (e.g. RNAseq, WGS, WES, or targeted deep sequencing) for the same sample, as optional input to add further evidence in support of the variant allele calling. It has been demonstrated that an integrated analysis of RNA and DNA can help detect mutations with high sensitivity and precision (Radenbaugh *et al.*, 2014), and our results also corroborate that notion. Second, when a catalog of known oncogenic/disease associated variants (e.g. mutations conferring drug-resistance) is provided as an optional input, SomVarIUS interrogates those positions, and if the disease-associated alleles are detected in the sample, they are reported in a separate vcf file. The output file reports the variants with appropriate annotations (i.e. even if they are present below detection threshold, or are potentially germ line). Finally, it models allele frequency distributions of the somatic mutations using a Gaussian mixture model, fitted using an expectation maximization algorithm, and reports if the observed mutations are likely to be clonal or sub-clonal. Tissue purity estimates, if provided as optional input can help to provide a refined estimate. SomVarIUS is written in Python 2.7. The default parameter settings and filters are meant to improve the quality of predictions, but those options can be changed in a user-defined way depending on the application.

We evaluated SomVarIUS using data from synthetic tumor datasets and real tumor sample from the Cancer Genome Atlas (Weinstein *et al.*, 2013). For the TCGA samples, we compared our results with those reported by Mutect (Cibulskis *et al.*, 2013) and VarScan2 (Koboldt *et al.*, 2012). In certain cases, VarScan2 classified a variant as '*germ line*' even if the position is bi/multi-allelic in the tumor but mono-allelic in the paired normal sample; we reclassified them as somatic mutations.

3 Results

3.1 Assessment using synthetic tumor mutation data

We first evaluated the utility of SomVarIUS using synthetic tumor genomics data with spiked in 'true' somatic mutations using published approaches (Boutros *et al.*, 2014). In brief, we selected a normal, non-malignant, blood-derived high coverage (~150×) whole exome sequencing dataset for the sample NA12878 from the 1000 Genomes Project. Next, using this as a template for bamsurgeon (Ewing *et al.*, 2015) we spiked in 192 somatic mutations to create a synthetic tumor genome (Fig. 2A), such that it had ~60% tumor

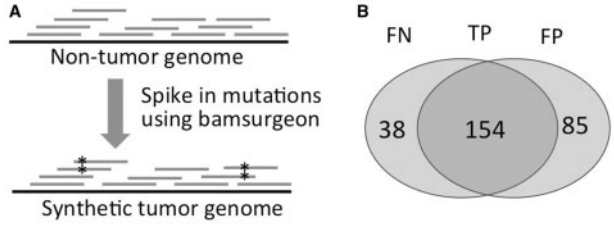

Fig. 2. (**A**) Synthetic tumor genomes with spiked in somatic mutations were generated using bamsurgeon. (**B**) Summary statistics showing the extent of precision and recall rates of the SomVarIUS predictions

purity and contained minor sub-clones with ~40% clonal contributions. Details of the spiked in somatic mutations are provided in the Supplementary Table S1. Using default parameter settings of SomVarIUS v1.1 on this synthetic tumor sample, we were able to detect 154 (80.2%) true positives, but also there were 85 false positives and 38 false negatives (Fig. 2B). Majority of the false positives was misclassified germ line variants that had abnormally imbalanced read support for the two alleles (Supplementary Table S1), while most of the false negative cases had low base quality for the variant alleles in the synthetic tumor sample. This analysis indicates that, SomVarIUS can detect somatic mutations with reasonable accuracy from a high coverage, synthetic, unpaired tumor dataset. VarScan, in the single sample variant calling mode, called 276 990 variants, which included all of the 192 spiked mutations, and also all the predicted somatic mutations reported by SomVarIUS.

To assess the performance of SomVarIUS for low coverage samples, we selected another synthetic tumor sample with ~30× whole genome sequencing data from the ICGC-DREAM Mutation Calling Challenge (synthetic tumor dataset 2; Boutros *et al.*, 2014). This synthetic sample was monoclonal and had ~80% tumor purity. Coefficient of variation of allelic abundance of reference and alternate alleles at known germ line heterozygous SNP positions was high, which is typical of low coverage datasets. Therefore, we ran SomVarIUS using a weaker p_g threshold ($p_g < 2E-01$). Out of 4 102 true single nucleotide somatic variants spiked in autosomal chromosomes in the synthetic genome, we were able to detect 2968 variants (72.35%) using SomVarIUS unpaired sample analysis. Majority of the spiked in 'true' somatic mutations had allelic abundance within a very narrow range around ~0.35. SomVarIUS also reported 15 433 false positives, majority of which (>90%) were misclassified germ line SNPs (i.e. variant also detected in the normal sub-bam) that had skewed allelic abundance of the reference and alternate alleles. A vast majority of these misclassified germ line variants had allelic abundance for the minor allele much lower than the true somatic mutations in the sample (>90% somatic mutations had allelic abundance between 0.33 and 0.375). So in retrospect, the low precision was not necessarily a consequence of the weak p_g threshold; it would be difficult to filter the germ line derived false positives in a low coverage dataset.

Taken together, our analysis using synthetic tumor data suggests that SomVarIUS can detect somatic mutations with a reasonable recall rate from unpaired tumor samples, but samples with sufficient sequencing coverage would be recommended to obtain high recall as well as a high precision-rate. In some cases, it is possible to improve the performance further by tailoring the filters for a given dataset. For instance, in the NA12878 analysis, both precision and recall rates improved when only high coverage positions (depth of coverage >50; read support for alternate allele >10) were considered. But, assuming that, in general, no or limited prior information is available for the unpaired samples (except probably depth of coverage, average sequencing quality, etc.), we do not report these results. But, whenever possible using additional information (e.g. pathological tumor purity and clonality) can potentially help adjust the parameter settings and improve the predictions further.

3.2 Comparative assessment using TCGA data

Next, we used genomic data from actual tumor samples for further evaluation of our method. We first detected somatic variants in exome-seq data for 8 unpaired tumor samples taken from the TCGA using SomVarIUS default parameter settings, and then compared our calls with those made using tumor-normal pairs by the TCGA MuTect pipeline for the same samples (Cibulskis *et al.*, 2013; Weinstein *et al.*, 2013) (Fig. 3A and Supplementary Table 2). The samples had 60–70% tumor purity, and median depth of sequencing coverage of ~200×. For a fair comparison, we only analyzed those genomic positions that had sufficient depth of coverage for variant calling. Unpaired sample variant calling detected 99.4% of the germ line mutations reported by TCGA. Moreover, unpaired and paired variant calls resulted in detection of 651 and 649 somatic mutations, respectively, of which 441 (441/651 = 67.7%) were shared (Fig. 3B). Of the remaining 210 potential somatic mutations reported by SomVarIUS, 98 were flagged as germ line SNPs, and 112 were not detected by TCGA. We investigated them further using independent approaches.

Since experimentally validated somatic mutations were not reported for these samples, for comparative evaluation we identified somatic variants by another paired sample variant caller, VarScan2 (Koboldt *et al.*, 2012), using exome-seq data from same 8 tumor-normal pairs. We found 462 (462/651 = 70.9%) variants shared between VarScan2 and SomVarIUS, when genomic positions that had sufficient depth of coverage (default ≥ 10×) were analyzed. Remaining 189 potential somatic mutations reported by SomVarIUS, 78 were flagged as germ line SNPs, and the VarScan2 did not detect 111. Out of 441 variants shared by MuTect and SomVarIUS, VarScan identified 348 (348/441 = 75%) variants. Interestingly, out of 112 variants that were only found by SomVarIUS but not by MuTect, VarScan2 found 62 (62/112 = 55%) of them. In comparison, of the 104 variants that were detected by MuTect only, VarScan2 detected only 34 (34/104 = 32%).

When RNAseq data from the same samples were included, we found additional support for the somatic mutations called by our method (Fig. 3C). Of the 112 potential somatic mutations reported by us and not detected by the TCGA, 79 had sufficient coverage in the RNAseq data, and of them 34 (43%) had evidence for expression of the somatic mutant allele. Importantly, the potential somatic mutations reported by SomVarIUS and not by TCGA had significantly higher support in the RNAseq data (34/79) compared to those reported by TCGA and not by SomVarIUS (24/68; Fisher's exact test; *P*-value: 5.36E − 04).

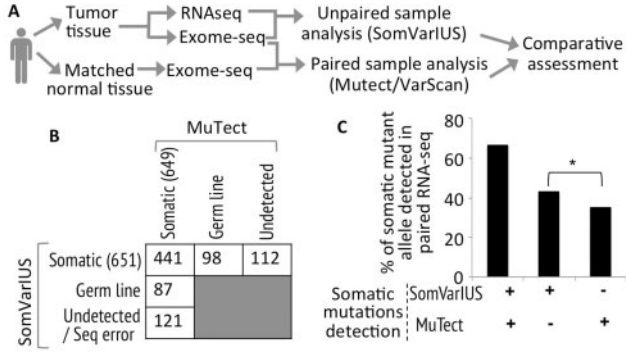

Fig. 3. (**A**) The workflow for comparative assessment of unpaired sample somatic variant call by SomVarIUS with paired sample somatic variant calls. (**B**) Summary statistics of the extent of overlap in somatic variant calls between paired and unpaired sample analyses. (**C**) Detection of mutant allele in the matched RNAseq data for the somatic mutations detected by both and/or either methods. The mutations reported by both methods were more likely to have support for the mutant allele in the paired RNAseq data compared to those predicted by only one method. The somatic mutations reported by only SomVarIUS were significantly more likely to have support for the mutant allele in the paired RNAseq data relative to those reported by only MuTect

The above analyses indicate that, SomVarIUS can detect somatic mutations using information from high coverage unpaired samples, more than 2/3 of which are concordant with paired-sample variant calls. Results from VarScan2 and RNAseq analyses indicate that this might be an under-estimate, and the true proportion of somatic mutations detected by SomVarIUS might be even higher.

3.3 Detection of somatic mutations in FFPE samples

A sizable proportion of the older FFPE tumor samples have no matched control tissues available. Furthermore these samples can have degraded DNA, which can influence sequencing and mutation detection. We evaluated SomVarIUS variant calls for the FFPE samples, to assess if the performance was comparable to that observed for the fresh frozen samples. We detected somatic variants in exome-seq data for 3 FFPE tumor samples (2 colon carcinoma and 1 bladder cancer; Supplementary Table S3) taken from the TCGA using SomVarIUS v1.1 at default parameter settings. 2339 somatic mutations were detected (TCGA-A6-6781: 1527; TCGA-A6-3809: 623; TCGA-BL-A131: 189).

Next, we decided to compare and contrast our results with paired-tissue somatic variant calls for the same samples. Even though the FFPE sections of these samples were not used by the TCGA for mutation identification, their fresh frozen sections were sequenced and analyzed by the TCGA mutation-calling pipeline for paired tumor-normal tissue somatic mutation detection. SomVarIUS detected 2339 somatic variants in these samples, and of them 1167(49.9%) were identified by the TCGA paired tissue analysis as well. Alternate alleles for ~40% of the somatic mutations reported by both methods were also supported by paired RNAseq data. In contrast, 32–37% of those detected by only one method were supported by RNAseq data.

The extent of overlap in somatic variant calls between the two approaches was slightly lower compared to that observed in the previous section. There were multiple reasons for this observation. First, we called somatic mutations from FFPE sections, but paired tissue variant calls were available for the fresh frozen sections of the same tumor samples. FFPE and fresh frozen sections represent different parts of the same tumor samples, and intra-tumor genetic heterogeneity could potentially contribute to slightly different genetic make up of these sections, even in the same tumor. Second, nucleic acids extracted from FFPE samples are fragmented and chemically modified, such that DNA/RNA extraction and sequencing are challenging and error-prone. Despite these limitations, the observed overlap in somatic mutation calls between the two approaches highlights the utility of our method for detecting somatic mutations in unpaired FFPE tumor samples. Nonetheless, given these issues analysis of the FFPE samples would probably benefit from higher sequencing coverage.

3.4 Treatment associated clonal dynamics in leukemia

We performed targeted deep sequencing of pre- and post-treatment clinical samples from 5 acute myeloid leukemia patients undergoing an experimental therapy using a gene panel covering 33 frequently mutated cancer genes. Paired normal samples were not available for these cases. Tumor purity was measured by the percent bone marrow aspirate blast count in the pre- and post-treatment samples. Average depth of sequencing in samples analyzed was ~5000×. We identified potential somatic mutations in several leukemia-associated cancer genes using SomVarIUS, and investigated their change in allele frequencies between pre- and post-treatment conditions (Fig. 4). The allelic abundance of ATM mutations in patient 1 and 5 and RUNX1 mutations in patient 2 and 3 increased in post-treatment samples (≥5% change in allelic abundance). These changes were unlikely due to difference in percent blast count (Supplementary Table S4). Our analysis indicates that mutations in ATM and RUNX1 were present in the pre-treatment stages, albeit at lower allele frequencies, and that perhaps treatment induced clonal competition and selection lead to preferential expansion of clones carrying these mutations. The patterns of clonal evolution observed here are similar to that reported elsewhere (Ding *et al.*, 2012), and indicate that subclonal cancer gene mutations present additional challenges for precision medicine in oncology initiatives.

4 Discussions

SomVarIUS can be useful in a variety of clinical or research settings, where matched normal samples are not yet routinely collected (e.g. precision medicine initiatives) or for archival samples (e.g. FFPE and fresh frozen tumor samples stored in the tissue banks). Filtering of common, germ line SNPs can help minimize accidental release of genetic information (e.g. haplotype blocks), which could be used to trace patient identity, while still enabling data sharing within the cancer research community.

SomVarIUS is fast, and provides an option to run the framework using user-defined parameter settings, which adds to its versatility. It analyzed sorted and indexed bam files for 8 TCGA samples in 2 h time on a 3.2 GHz, 16 Gb RAM quad core machine. It also has a limited number of dependencies, and open-source architecture, which can be of added benefit. It is easy to integrate it into any standard genome analysis pipeline. For instance, it might be possible to prioritize high confidence variants using alternative variant callers, and use it to identify those of potential somatic origin. SomVarIUS takes advantage of availability of additional high throughput sequencing data (e.g. RNAseq, exome-seq, targeted sequencing) data from the same samples (which are sometimes available in clinical settings) to provide additional evidence for the candidate mutations. It also provides additional annotation for clinically relevant variants (e.g. drug resistant mutations).

We note that, SomVarIUS is not designed to identify somatic mutations that reach allelic frequency higher than the ploidy-adjusted heterozygous germ line SNPs (e.g. homozygous mutations with allelic abundance >0.5 in diploid regions). Furthermore, it may not be able to optimally distinguish somatic mutations from germ line variants and/or sequencing errors in low coverage sequencing data. However these limitations are inherent to low coverage unpaired samples, and may not be a major concern for clinical samples that are typically sequenced at high depth. Ongoing projects such as UK10K study, when completed, will also be able to provide an even

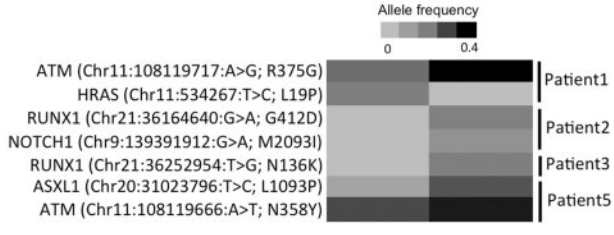

Fig. 4. Clonal dynamics of cancer gene mutations in five AML patients detected using targeted deep sequencing are shown. Only those mutations that have ≥5% allele frequency change between pre- and post-treatment condition are reported. For patient 4, we did not found any somatic mutations that have ≥5% allele frequency change between the pre- and post-treatment stages

denser catalog of SNPs, which will be helpful. Nevertheless, we recommend using samples with sufficient sequencing coverage (e.g. $\geq 100\times$) for somatic mutation detection from unpaired samples with SomVarIUS. We will also improve the clonality assessment using copy number information, similar to that implemented elsewhere (Miller *et al.*, 2014; Qiao *et al.*, 2014; Roth *et al.*, 2014). Finally, SomVarIUS is not intended to replace the paired-sample variant calling tools; rather it provides a useful utility for identifying potential somatic mutations when paired tissues are not available.

Acknowledgements

The authors would like to thank Debashis Ghosh, James Costello, Aikchoon Tan, Brent Pedersen, and members of the Computational Biosciences Graduate Program at University of Colorado School of Medicine for helpful discussion.

Funding

Boettcher Foundation Webb-Waring grant (to S.D.).

Conflict of interest: none declared.

References

Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Boutros,P.C. *et al.* (2014) Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nat. Genet.*, **46**, 318–319.

Chernoff,H. (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, **23**, 493–507.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

Ding,L. *et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.

Ewing,A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, **12**, 623–630.

Guo,Y. *et al.* (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Meric-Bernstam,F. *et al.* (2015) A decision support framework for genomically informed investigational cancer therapy. *J. Natl. Cancer Inst.*, **107**, 1–9.

Miller,C.A. *et al.* (2014) SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.*, **10**, e1003665.

Qiao,Y. *et al.* (2014) SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol.*, **15**, 443.

Radenbaugh,A.J. *et al.* (2014) RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PloS One*, **9**, e111516.

Roth,A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.

Sboner,A. and Elemento,O. (2015) A primer on precision medicine informatics. *Brief. Bioinf.*, **17**, 145–153.

Sun,J.X. *et al.* (2014) A computational method for somatic vs germline variant status determination from targeted next-generation sequencing of clinical cancer specimens without a matched normal control. In: *Proceedings of the 105th Annual Meeting of the American Association for Cancer Research* 74, Abstract 1893.

Tripathi,R.C. *et al.* (1994) Estimation of parameters in the beta binomial model. *Ann. Inst. Statist. Math.*, **46**, 317–331.

Weinstein,J.N. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.