Subject Section

Cell-level somatic mutation detection from single-cell RNA-sequencing

Trung Nghia Vu¹, Ha-Nam Nguyen², Stefano Calza³, Krishna R Kalari⁴, Liewei Wang⁵, Yudi Pawitan¹,*

¹Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden; ²Information Technology Institute, Vietnam National University in Hanoi, 144 Xuan Thuy, Hanoi, Vietnam; ³Department of Molecular and Translational Medicine, University of Brescia, Viale Europa 11, 25125, Brescia, Italy; ⁴Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA; ⁵Department of Molecular Pharmacology & Experimental Therapeutics, Mayo Clinic, 200 First Street SW, Rochester, MN, 55905, USA.

* To whom correspondence should be addressed.

Abstract

Motivation: Both single-cell RNA-sequencing (scRNA-seq) and DNA-sequencing have been applied for cell-level genomic profiling. For mutation profiling, the latter seems more natural. However, the task is highly challenging due to the limited input materials from only two copies of DNA molecules, while whole-genome amplification generates biases and other technical noises. ScRNA-seq starts with a higher input amount, so generally has better data quality. There exists various methods for mutation detection from DNA sequencing, it is not clear whether these methods work for scRNA-seq data.

Results: Mutation detection methods that are developed for either bulk-cell or single-cell DNA sequencing data do not work well for the scRNA-seq data, as they produce substantial numbers of false positives. We develop a novel and robust statistical method — called SCmut — to identify specific cells that harbor mutations discovered in bulk-cell data. Statistically SCmut controls the false positives using the two-dimensional local false discovery rate method. We apply SCmut to several scRNA-seq datasets. In scRNA-seq breast cancer datasets SCmut identifies a number of highly confident cell-level mutations that are recurrent in many cells and consistent in different samples. In a scRNA-seq glioblastoma dataset we discover a recurrent cell-level mutation in the PDGFRA gene that is highly correlated with a well-known in-frame deletion in the gene. To conclude, this study contributes a novel method to discover cell-level mutation information from scRNA-seq that can facilitate investigation of cell-to-cell heterogeneity.

Availability: The source codes and bioinformatics pipeline of SCmut are available at https://github.com/nghiavtr/SCmut.

Contact: Yudi.Pawitan@ki.se

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Cell-to-cell heterogeneity is a common feature in cancer and it has potentially important clinical consequences (Huang, 2009), but it is not possible to study this phenomena using traditional bulk-cell sequencing. Recent advances of single-cell sequencing technologies enable the study of molecular processes at cell level (Navin, 2014; Van Loo and Voet, 2014; Wang and Navin, 2015; Wen and Tang, 2016). Detection of genomic mutations using single-cell DNA-sequencing (scDNA-seq) has been reported for several diseases, for example breast cancer (Wang et al., 2014) and renal carcinoma (Xu et al., 2012). However, with very low input materials coming only from two copies of DNA molecules (Navin, 2014; Van Loo and Voet,
scDNA-seq suffers many problems such as technical errors, amplification bias, low physical coverage, chimeric DNA, non-uniform coverage, allelic drop-out (ADO) event, etc (Wang and Navin, 2015; Van Loo and Voet, 2014). General analysis tools for detecting SNVs from scDNA-seq data that address some of these issues have appeared recently, for example Monova (Zafar et al., 2016).

Single-cell RNA-sequencing (scRNA-seq) has also a considerable development in recent years. Even though a mammalian cell contains a very low amount of RNAs (Wang and Navin, 2015), the number of copies of RNAs in a cell is still much greater than that of DNAs. ScRNA-seq has been widely used in investigating gene expression of cells. The information of single nucleotide variant (SNV) and allele-specific expression (ASE) of single cell from scRNA-seq have also been investigated recently. For example, in (Kim et al., 2015a), the authors predict that only 17.8% stochastic allele-specific expression patterns contribute to biological noise. Similarly, Boret et al. (Boret et al., 2015) report that 76.4% of heterozygous SNVs display stochastic mono-allelic expression in single cells. Recently, Kim et al. (Kim et al., 2015b) study the heterogenous expression of SNVs in a study of patient-derived xenograft (PDX) cells of lung adenocarcinoma.

Bulk-cell RNA-sequencing (bcRNA-seq) from a population of cells has been used to detect genomic variants in many studies (Goya et al., 2010; Tang et al., 2014). For instance, in recent study (Piskol et al., 2013), Piskol et al. report that over 70% of all expressed coding variants are identified from RNA sequencing, and whole exome sequencing (WES) and RNA-seq have comparable numbers of identified exonic variants. So it is natural to investigate genomic variants from the scRNA-seq data. For example, Chen et al. (Chen et al., 2016) investigate the single-cell single nucleotide polymorphisms (SNPs) based on scRNA-seq in colon cancer. However, up to now, to our best knowledge, there are no methods specifically designed to detect cell-level somatic mutations from scRNA-seq.

In this study, we show that mutation detection methods that are developed for either bulk-cell or single-cell DNA sequencing data do not work well for the scRNA-seq data, as they produce too many false positives. We propose a novel statistical method – called SCmut – to identify cells that harbor mutations discovered in bulk-cell data. In brief, the method first collects somatic mutations from bulk-cell DNA sequencing (bcDNA-seq) of tumor and matched germline tissues. Then, combining with the collection of SNVs of single cells extracted from scRNA-seq, SCmut statistically detects the somatic mutations at cell level using the two-dimensional local false discovery rate (2D local fdr) method. We apply the method to several scRNA-seq datasets from (i) two breast cancer patients in a recent study (Chung et al., 2017), (ii) two sets of cells from the breast cancer cell-line MDA-MB-231, and (iii) one set of glioblastoma cells. In (i) the discovered cell-level mutations are well separated between tumor and non-tumor cells, and in (ii) the mutations are replicated in two independent datasets. In the glioblastoma dataset (iii), we discover a cell-level mutation that is highly correlated with a well-known 24bp-in-frame deletion in the PDGFRα gene. The cell-level mutation information can be used to support the characterization of cell-to-cell heterogeneity in cancer.

2 Methods

The analysis pipeline is presented in Figure 1. First, the FASTQ files of scRNA sequencing and bcDNA sequencing are put through preprocessing steps for alignment, duplicate removal, recalibration, etc to generate aligned sequences in BAM files. Next, the DNA samples of tumor and germline are used to obtain somatic mutations. Then, variant calling is implemented to all data samples of both single-cell and bulk-cell to get the list of single-nucleotide variants (SNVs). Finally, statistical methods are applied to the SNV list to discover cell-level mutations. Details of each step are presented in the following sections.

2.1 Data preprocessing

For DNA-seq data, which are the WES data in our examples, the FASTQ files are mapped to human hg19 annotation of Ensembl GRCh37.75 using BWA (Li and Durbin, 2009) version 0.7.10 to achieve aligned reads (BAM files). After mapping, duplicate reads are marked and removed to reduce biases from library preparation, e.g., PCR artefacts using Picard (http://broadinstitute.github.io/picard/) version 2.3.0. Realignment around indels (GATK BaseRecalibrator) is applied to deal with the problems of over- or under-estimated scores caused by errors of sequencing machines. These two last steps are applied with the supports of known variant sites from phase I of 1000 Genomes Project and dbSNP-138 (Sherry et al., 2001). All the tools of GATK are implemented in GATK version 3.6.

To process RNA-seq data, the FASTQ files are also mapped to human hg19 annotation of Ensembl GRCh37.75, but using TopHat (Trapnell et al., 2009) version 2.0.12 and Bowtie2 (Langmead et al., 2009) version 2.2.3 to create BAM files. Then the follow-up processes are generally similar to the processing workflow of DNA-seq data. However, to avoid possible specific pitfalls of RNA-seq data, such as sequences overlapping into the intronic regions, after the step of removing read duplicates, an extra step (GATK Split’N’Trim) is applied. In this step, reads marked with ‘N’ symbol are eliminated and sequences overlapping regions into the intronic regions are hard-clipped.

After the preprocessing phase, the reads of RNA-seq and DNA-seq data are aligned and summarized for downstream analysis.

2.2 Somatic mutation detection from bcDNA-seq and variant calling

From the bcDNA-seq of tumor tissue and matched germline, the somatic mutations can be discovered by any somatic mutation detection methods, such as Mutect (Cibulskis et al., 2013) or VarScan (Koboldt et al., 2012), etc. For the breast-cancer and GBM patients data, we detect somatic mutations with the support of the databases of known SNP and indels from phase I of 1000 Genomes Project and dbSNP-138 (Sherry et al., 2001).

Next, all samples of both single cells and bulk tissue are put through variant calling using SAMtools (Li et al., 2009) version 1.3 followed by VarScan (Koboldt et al., 2012) version 2.3.7. An SNV is retained for further analysis only if it has at least (i) 5 supporting reads, (ii) 1% variant frequency, and (iii) 15 average quality score, for at least one sample. For each valid SNV, we compute the cell-level statistics, including total reads and variant-allele frequency (VAF).

2.3 Two-dimensional local false discovery rate

Ideally, a variant caller with its statistical method should achieve a high specificity and minimize the number of false positive mutations as possible. However, as we show in Section 3.1, when applied to scRNA-seq data, the traditional methods designed for DNA sequencing data produce high false positive rates. Therefore, we introduce a statistical approach to overcome this issue.

To get a procedure that is both efficient and has a good control of the false positives, we adapt a 2D false discovery rate (fdr2d) procedure, originally developed for analysis of microarray data (Piffer et al., 2006). Let denote the total reads by \( z \) and the VAF by \( y \), measured for each SNV from each cell. The \( fdr2d \) based on \( z = (z_1, z_2) \) is defined as

\[
\text{fdr2d}(z_1, z_2) = \tau_0 \frac{f_1(z_1, z_2)}{f(z_1, z_2)}
\]

where \( f_1(z) \) is the 2D-density function of the statistic from the null variants, and \( f(z) \) the marginal density from all sites. The parameter \( \tau_0 \) is the
Cell-level mutation detection from scRNA-seq

Fig. 1. The pipeline for detecting cell-level mutation from scRNA-seq data. First, the FASTQ files of scRNA-seq and bcDNA-seq are put through preprocessing steps for alignment and clean-up to create aligned sequences in BAM files. Next, the somatic mutations are detected from bcDNA-seq data, and both single-cell and bulk-cell data are put through variant calling procedures. Suppose the data contain $n$ single cells and the number of obtained SNVs is $M$. Finally, statistical methods, particularly the two-dimensional local false discovery rate, are used to identify cell-level mutations. The bottom-right panel is an example of cell-level mutations discovered by SCmut from the single cells of primary tumor of patient BC03 in this study. The contour map represents the statistics from the permutation in 2D local fdr method, and each filled-circle point presents a mutation of a single cell. The red dots and blue triangles indicate the tumor cell and non-tumor cell, respectively. The significant cell-level mutations with $fdr2d < 0.2$ and $fdr2d < 0.05$ are marked by orange (light) and brown (dark) squares, respectively.

The fdr2d measures the relative contributions of null SNVs to the observed density at $z$, so it measures the rate of false discoveries if we declare the sites with observed value $z = (z_1, z_2)$ to be mutations. Thus we can control the false positives directly by limiting the estimated fdr2d. Since the non-mutations are not known, a key step of the method is to generate $z$ by Monte Carlo sampling from non-bc-mutation sites in the bcDNA data. A non-parametric smoothing procedure is then used to estimate fdr2d. Denote by SNV$_1, \ldots, $ SNV$_m$ the bc-mutation sites from the bcDNA-seq data. The observed statistics are $z$ values from these $m$ SNVs across all single cells 1, ..., $n$. Let $Z$ be the $m \times n$ matrix of observed $z$ values. For convenience, assume each cell of the matrix contains the pair of statistics $(z_1, z_2)$. The data required to estimate $f_0(z)$ are based on $K$ random samples, each of size $m$, of the null SNVs, i.e. the non-bc-mutation sites. As for the bc-mutations we limit to SNVs with VAF = 0 in the germline, since somatic mutations are not likely to have any variant in germline. Denote these samples as $Z^*_1, \ldots, $ $Z^*_K$, representing samples of $Z$ under the null hypothesis of no mutation. In all of the examples in this paper we use $K = 100$ samples.

In principle, we could use nonparametric density smoothing to estimate $f(z)$ from the observed $z$'s, and $f_0(z)$ from $Z^*_1, \ldots, $ $Z^*_K$, then compute the fdr2d($z$) by simple division. However, in practice this approach is problematic: at the edges of the distribution of $z$ the ratio is noisy, and to control the noise, different amounts of smoothing are required for the two functions. Statistically it is better to estimate $r(z) = K f_0(z) / (f_0(z) + K f_0(z))$, as the target parameter, and compute the fdr2d as

$$fdr2d(z) = \pi_0 \frac{r(z)}{K (1 - r(z))} \quad (2)$$
is the proportion of million reads. The scRNA-seq as 'failures', so that false positives, or $<Z_0 \leq Z$ is typically used in our examples. Note that $z$ as 'successes' and the is too conservative as a function of $z$. Further details of the 2D local fdr approach are given in the Supplementary report.

After the fdr2d estimation, each observed candidate of the cell-level mutations has a corresponding fdr2d value (Pawitan et al., 2005). The threshold of fdr2d < 0.2 is typically used in our examples. Note that fdr2d is not a P-value, so it does not follow the usual reasoning for p-value thresholds. For example, if we report 10 significant mutations with fdr2d < 0.2, then we expect only < 2 false positives, or > 8 true positives. So, fdr2d < 0.2 is a reasonable cutoff, while fdr2d < 0.05 is too conservative and would lead to unnecessarily low sensitivity.

### 2.4 Datasets

**Breast cancer patient dataset:** The full dataset from Chung et al. (Chung et al., 2017) contains 11 breast cancer patients from different (predicted) molecular subtypes Luminal A, Luminal B, HER2-enriched and Basal-like. We select two patients BC03 (HER2-enriched) and BC07 (Basal-like) because they have scRNA-seq data of the primary tumor and lymph-node tissues. In addition we also collect the bulk-cell whole exome sequencing (bcWES) data of the primary tumor, lymph node and the matched blood. In bcRNA-seq, bulk RNAs were extracted from pooled cells or tumor tissues. In addition we also collect the bulk-cell whole exome sequencing because they have scRNA-seq data of the tumor and lymph-node tissues. The further details of the datasets are referred to the original paper.

**Breast cancer cell-line dataset:** The dataset includes a batch of 96 scRNA-seq samples from triple-negative breast cancer cell-line (MDA-MB-231) (control group), and another batch of 96 scRNA-seq samples from the same cell-line but treated with metformin (treated group). Single cells were captured and sequenced using a combination of Fluidigm protocol and Illumina HiSeq machine. There are on average 4.9 million read-pairs per cell, with read length 100bp. Two cell groups were sequenced in separate batches, thus making them fully independent. After removing empty-cell wells (the negative controls), there remain 82 and 88 cells in the control group and the treated group, respectively. Since there are no available DNAseq of the cancer cell-line and matched normal germline, we obtain 99 confirmed-somatic mutations of MDA-MB-231 cell-line from the COSMIC database (Forbes et al., 2017) for downstream analysis (available in Table S1 of the supplementary report).

**Glioblastoma dataset:** This dataset contains 96 cells from a primary brain tumor of a glioblastoma multiforme (GBM) patient (patientID SF10282) from a recent study (Müller et al., 2016). Libraries of single cells were captured and prepared on the Fluidigm C1 system then sequenced on HiSeq 2500 (Illumina) using paired-end 100bp protocol. The bcWES data is 100x for tumors and 50x for blood samples. The authors also separated the cells into tumor and non-tumor (lymphocyte) cells. The further details of the datasets are referred to the original paper.
3 Results

3.1 Challenges of mutation detection from scRNA-seq

3.1.1 Concordance of VAFs between scRNA-seq and bulk-cell sequencing data

To assess the quality of scRNA-seq, we first check their concordance with the more established bcDNA-seq and bcRNA-seq. Figure 2 shows the result for primary breast-cancer from patient BC03, comparing the VAFs from scRNA-seq (pooled across 33 cells) against the VAFs from bcRNA-seq and bcDNA-seq (bcWES). Only common variants (present in ≥ 50% cells) are included. The correlation is high (r = 0.89) with bcDNA-seq (panel b), and even higher (r = 0.96) with bcRNA-seq (panel a). Thus, despite the high level of noise, scRNA-seq data can capture the underlying variant information that exists in bulk-cell data.

3.1.2 High level of noise in scRNA-seq data

The high proportion of stochastic mono-allelic expression of SNVs is well known in scRNA-seq data (Borel et al., 2015). Figure 3 displays the total reads (coverage) and VAFs of the data from the primary breast cancer of patient BC03. We first apply Mutect (Cibulskis et al., 2013) to discover the mutations in the tumor from the bcWES data. The mutated sites are highlighted in the plots using red/blue colors, where the red is for the mutated sites from tumor cell and the blue is for the ones from non-tumor cell. The distributions of the non-mutated sites (grey circles) in the scRNA-seq (panel a) and bcWES (panel b) are similar. However, the distributions of the mutated sites from tumor cell and the blue is for the ones from non-tumor cell. For bulk-cell data such high VAF would be a strong evidence for mutation sites, but there is very little overlap between these SNVs and the bulk-cell mutation calls. In (d), the distribution of the mutations in bcWES are detectable as the extreme points of the distribution. But, in (c), those mutation sites lie in the middle of the distribution. These features highlight the difficulty of mutation detection from scRNA-seq data alone, as the approach used for in bcWES is not likely going to work.

3.1.3 Traditional methods designed for bulk-cell sequencing data

To investigate further, we use Mutect (Cibulskis et al., 2013), a widely used bulk-cell method. We select mutations in the single cells of the primary breast tumor of patient BC03. In particular, the scRNA-seq sample of each single cell (treated as tumor sample) and the bcWES of the blood sample (normal sample) are put through the software. We collect the detected mutations from all single cells and plot them in Figure 4a. The grey circles of the plot are non-mutated sites, and the red dots and blue triangles are the mutations from the tumor and non-tumor cells respectively. There is a total of 25,265 cell-level mutations from 24,469 mutation sites, i.e., an average of 25,265/24,469=1.03 cells have mutations per mutation site, or almost all mutations are singletons (seen only in one cell). The called mutations cover the full range of VAFs and total reads above a certain value; this is an expected feature of the bulk-cell method. However, there is very little overlap between these SNVs and the bulk-cell mutation calls. In (d), the mutation sites in bcWES are detectable as the extreme points of the distribution. But, in (c), those mutation sites lie in the middle of the distribution. These features highlight the difficulty of mutation detection from scRNA-seq data alone, as the approach used for in bcWES is not likely going to work.

3.2 Cell-level mutations in breast cancer patient BC03

We apply SCmut to detect mutations from single cells of both primary tumor and lymph node tissues from breast cancer patient BC03. Both primary tumor and lymph node cells have a high level of heterogeneity, where a high proportion of different immune cells (≥ 50%) infiltrated into the tissues (Table S2). The types of cells (tumor or non-tumor) are identified in the original study (Chung et al., 2017); we use this information to assess the specificity of the mutation calls, since we do not expect the non-tumor cells to have mutations as in tumor cells. First, we discover 371 somatic mutations from the primary tumor using the bcWES data, then, a total of 1,253,869 SNVs detected across the single cells and the bulk cells are used to identify cell-level mutations. The results of the fdr2d method for the single cells from primary tumor dataset are presented in the panel "Cell-level mutation detection" in Figure 1, and re-plotted in Figure S1. The contour map of the plot represents the fdr2d estimate, and each point represents SNVs from a single cell limited to the 371 bc-mutation sites above. The red and blue points indicate the tumor cell
and non-tumor cell respectively. The significant cell-level mutations (with \( \text{fdr2d} < 0.2 \)) are marked by the orange squares. Most of the detected significant mutations with \( \text{fdr2d} < 0.2 \) are from tumor cells, and all satisfying \( \text{fdr2d} < 0.05 \) (marked by the brown squares) are tumor cells.

The results indicate the high specificity of SCmut, since few of the cells are from the non-tumor cells. Looking at the top left part of the panel, SCmut does not call SNVs with VAF = 1 as significant mutations, as would be expected in bulk-cell analysis. In single-cell data, these observations are quite common (Figure 3 a and c) due to stochastic monoallelic expression (Borel et al., 2015). Thus, SCmut is robust to the intrinsic noise of the single-cell data.

The same pipeline is used to discover cell-level mutations in the cancer tissue from the lymph node. We compare the significant mutations (Figure 2S) with those from the primary tumor. The top ten most frequent among the significant mutations are presented in Figure 5, the full list is supplied in Figure S3 of the Supplementary document. Each rectangle of the heatmap represents the information of the mutation (row) in a single cell (column). Seven mutated genes detected from bcWES data are detected in single cells of either primary tumor or lymph node (\( \text{fdr2d} < 0.2 \): MT-RNR2, MT-RNR1, MT-ND5, MT-TL, HUWE1, TMEM219 and INTS8. Among those, only the mutation of gene MT-RNR2 at position 2602 of the mitochondria chromosome is replicated in both primary tumor and lymph node.

All cell-level mutations detected with \( \text{fdr2d} < 0.05 \) are from gene MT-RNR2 of the tumor cells in both tissues (Figure 5). MT-RNR2 encodes the humanin, an anti-apoptotic peptide that can prevent the translocation of Bcl2-associated X protein (Bax) from the cytosol to mitochondria to suppress apoptosis (Guo et al., 2003). It can play a role in regulating cell survival and apoptosis via interacting with insulin-like growth factor-binding protein 3 (IGFBP3) (Ikenoue et al., 2003). Apoptosis is an important pathway in breast cancer where the increase of apoptosis is associated with malignant tumors due to increased proliferation, high grade and negativity for oestrogen receptors of breast tumors, and worse survival (Parton et al., 2001). The shared mutations discovered in the primary tumor and the lymph node would identify the clone(s) that have very likely migrated from the primary tissue to the lymph node. This is of potential clinical significance, as these clones have thus already shown a local metastatic potential.

Similar analyses are applied to the single cell datasets of the primary tumor and the lymph node tissues of patient BC03. As displayed in Figure S4 of the Supplementary report, there are few mutations frequently detected in both primary tumor and lymph node such as PSMD7(16:74339229), POLR2L(11:8424185) and SFT2D1(6:166755986). However, none of the mutations are statistically significant with \( \text{fdr2d} < 0.2 \) in either primary tumor (Figure S5) or lymph node (Figure S6).

3.3 Comparisons with other methods

3.3.1 False positive rates

We further compare SCmut to widely used bulk-cell mutation detection methods Mutect (Cibulskis et al., 2013) and VarScan2 (Koboldt et al., 2012), and Monovar (Zafar et al., 2016), a SNV detection method designed for scDNA-seq data. We use the results from patient BC03, and first compare the false positive rates (FPRs). The non-tumor cells identified in the original study (Chung et al., 2017) are used as negative controls, so we can estimate FPRs from the mutation calls on these cells.

First we apply these two bulk-cell methods to discover somatic mutations from each non-tumor single cell of patient BC03 using its scRNA-seq sample (treated as tumor) and the bcWES sample of the patient’s blood (normal). Since there are no available p-value from results of Mutect, we infer this value from the log odds (LOD score) (Cibulskis et al., 2013) of tumor. The LOD score is constructed from the likelihood ratio between the signal (true variant) and noise. Twice the log likelihood value is approximately \( \chi^2 \) with one degree of freedom (Pawitan, 2013), so we can compute the p-value for each SNV. Following the requirement of the normal sample to carry somatic mutations with high confidence (Cibulskis et al., 2013), we keep only the sites with significant log odds in the normal (LODnormal \( \geq 2.3 \)).

Similarly, for VarScan2 we collect only somatic p-values from the sites with no variant in germline. For SCmut, the p-value is computed from the connection between the global FDR and p-value (Pawitan et al., 2005). P-values for SCmut are collected from both the breast and lymph-node samples.

Monovar is run for the set of single cells from the tumor sample of patient BC03. Since Monovar was designed for scDNA-seq data, we adapt some tuning parameters to make them more appropriate for scRNA-seq data. First, allelic drop-out event is common \((\sim 20\%)\) in scDNA-seq (Zafar et al., 2016), but not in single-cell RNA-seq (where the drop-out event refers to the RNA transcripts, not alleles). We set the ‘prior probability for allelic drop-out parameter \( \alpha \) to zero. For the ‘prior probability for false-positive error’, which is suitable for the sequencing error rate of the RNA-seq data from Illumina HiSeq, we set \( \beta = 0.003 \) (Schimer et al., 2016; Wall et al., 2014; McBryer et al., 2012). The default values are applied for the other parameters. Following the original study (Zafar et al., 2016), from the set of SNVs called by Monovar, somatic mutations are filtered by the bulk-cell germline variants. Since Monovar does not provide p-values for SNVs of cells, we compute the p-values from the reported likelihoods of genotypes as follows. Following a recent study (Singer et al., 2018), we first transform back the normalized and Phred-scaled likelihoods for genotypes supplied by Monovar. Each SNV site, which is assumed biallelic by Monovar, from a single cell data \( D \), has the likelihood values of three genotypes \( g \), including \( L_0 \) (heterozygous variant) and \( L_2 \) (homozygous variant), where \( L_1 \equiv P(D|g = i) \) or \( L_2 \equiv P(D|\bar{i}) \) in short. Thus, the posterior probability \( P(g = 0|D) \) of the wild-type can be computed as

\[
P(g = 0|D) = \frac{P(D|g = 0)P(0)}{\sum_{i=0}^{2} P(D|i)P(i)}
\]

The genotype prior \( P(i) \) for a single cell is taken from formula (11) of Zafar et al’s for the number of cells \( m = 1 \). So, we have \( P(0) = \frac{1}{2} (1 - \theta) \) and \( P(2) = \frac{1}{2} (1 - \theta) \), where \( \theta = 0.001 \) is the population-level mutation rate. We consider \( P(g = 0|D) \) as the local false discovery
rate of the mutation calls, which can then be converted into p-value, exactly as we have done for SCmut.

Figure 4b displays the FPR curves of the these methods. The y-axis presents the observed FPR under a certain p-value threshold in the x-axis, an unbiased method should have its FPR close to the diagonal line. Both Mutect and Varscan2 have a very high FPR (> 0.90) even at low p-value threshold (< 0.1). As already described previously, this result again highlights the challenge of scRNA-seq data for the traditional mutation detection methods for bulk-cell data. Monovar has a better FPR curve, but still very high (FPR > 0.7) at the same p-value threshold (< 0.1). The FPR of SCmut tracks the target diagonal line closely, indicating that it is unbiased.

To get a fairer comparison with SCmut, we further restrict the comparison to the mutations that overlap with the somatic mutations from bcWES. Figure 4c presents the cell-level mutation status after the restriction. There remain 132, 89 and 58 single-cell mutations from Mutect, Varscan2 and Monovar, respectively. As a result, the FPRs of these methods (Figure 4e) are similar to those without restriction (Figure 4b). This result again indicates that the cell-level mutations are over-detected by these methods.

3.3.2 Recovery sensitivity

We use the information of VAF from the bulk-cell sequencing to compare the recovery sensitivity of SCmut to the other methods. The VAF of a mutation from bcWES reflects the fraction of tumor cells with the mutation; the latter is observed in single-cell data. Therefore, we can use this correspondence for assessing the sensitivity of the methods. The mutations called by SCmut (fd2d < 0.2), Mutect and Monovar that are concordant to the calls of bcWES are collected from the primary tumor data of patient BC03. To avoid the effects of copy number variants, we collect the data of copy number variants from Supplementary Data 2 of the original study (Chung et al., 2017) and exclude all mutations in regions not having two copies in the tumor sample. Figure 4e and f show the results. For SCmut, the observed VAF estimated from the cells with the mutation calls is highly concordant (r = 0.89) with the expected fraction from the bcWES. The correlation is significantly higher than that from Mutect (r = 0.17) and Monovar (r = 0.20). Thus, for the cell-level mutations, SCmut shows a better recovery sensitivity than Mutect and Monovar.

3.4 Cell-level mutations in the breast cancer cell-line

We apply SCmut to the breast cancer cell-line (MDA-MB-231) datasets which have highly homogeneous cell populations. Results of fd2d are presented in Figure S8 for the control group and Figure S9 for the treated group. A total of 99 somatic mutations (in exon region) from the COSMIC database are used as the bc-mutation sites. We observe 26 and 34 SNVs that overlap with the COSMIC sites in the control group and the treated group, respectively. All mutations in the control group are replicated in the treated group (Figure S9). There is a high concordance in the coverage of mutation between two groups (Figure 6, panel a). Moreover, some mutations with high coverage from genes CNIH4, PAK1IP1 and SNRPC can be preserved up to more than 90% of cells, indicating positive controls (Figure S10). We compare the recurrences of significant cell-level mutations (fd2d < 0.2) between two groups by their proportion, i.e. the proportion of cells sharing the same mutation site, in panel b. The minor variation of the mutations to the diagonal line indicates a high correlation of the recurrent proportions between two groups (r = 0.98). Thus, there are no significant effects of the neofusion on the somatic mutations of MDA-MB-231 cell-line. In other words, the cell-level mutations detected by SCmut are consistent between two homogeneous cell populations of the breast cancer cell-line.

3.5 Cell-level mutations in the glioblastoma dataset

The results of SCmut to the glioblastoma dataset are given in Figures S11 and Figure S12 of the Supplementary report. SCmut detects a total of 104 cell-level mutations with fd2d < 0.2. We discover one highly recurrent mutation at chr4:55,133,837, inside the PDGFA gene, and found in 31 single cells. Intriguingly this mutation is highly correlated with a well-known 248p- in-frame deletion in exon 7 of PDGFRA, which is also recurrent with many cells expressing PDGFRA (Muller et al., 2016). The corresponding VAFs, shown in Figure 7a, have a Pearson correlation of (0.9). While we do not understand its biological significance, statistically the high correlation between these two events indicates the good sensitivity of SCmut for detecting the mutation events.

We further investigate the clinical impact of the top three recurrent mutated genes discovered by SCmut, including PDGFA, DYNC1L12 and CHD6 (Figure S13). We extract the mutation status of these genes as called by Mutect in a glioblastoma study TCGA-GBM (Brennan et al., 2013) from the TCGA project (https://portal.gdc.cancer.gov/). Figure 7b shows that these mutations together are associated with poor overall survival (P-value =2e-04). The results for individual genes are given in Figure S13 of the Supplementary report.

4 Discussion and Conclusion

We have proposed a novel method (SCmut) to identify cell-level mutations from scRNA-seq. We present the challenges of identifying mutations from single cells, showing high levels of noise and discordances between the single-cell and bulk-cell data. Traditional mutation detection methods developed for bulk-cell sequencing data are shown to produce substantial number of false positives if applied to scRNA-seq data. We use the two-dimensional local false discovery rate statistic to deal with the multiple testing issue and control the false positives.

For breast cancer patient BC03, we discover one mutation from the human genome, an associated apoptosis gene in the mitochondrial chromosome, highly preserved in the tumor cells of both the primary tumor and lymph node. In addition, our results show that the detected cell-level mutations are well separated for tumor cells from non-tumor cells in the highly heterogeneous patient-derived cell populations, and consistent in the homogeneous cell-line populations. For the glioblastoma data example, we discover a cell-level mutation that is highly correlated with a well-known in-frame deletion, while the 3 top-ranking cell-level mutated genes are associated with poor patient survival.
Mutation detection from scRNA-seq data has some limitations. First, the cell-level mutations must be in theastic regions. This is a general disadvantage of all approaches to detect mutations from RNA sequencing or whole exome sequencing data. Second, the procedure is highly dependent on the quality of the alignment and hence the completeness of the transcriptome annotation. Third, the stochastic mono-allelic expression (Borel et al., 2015) might limit the expression of the mutation sites in single cells. Finally, the detection sensitivity of a mutation is determined by the corresponding gene expression in the cell. An important mutation is statistically detectable from scRNA-seq only if it belongs to a highly expressed gene. It is challenging to assess cell-level mutations in genes with low or no expression, since the low expression could be a loss-of-function effect, but could also be due to the low coverage of scRNA-seq data, leading to false negatives. Hence, in order to detect cell-level mutation events, we recommend high coverage scRNA-seq. Further discussion about the detection sensitivity and the coverage threshold of SCmut is presented in the Supplementary document.

It is worth noting that SCmut focuses on detection of somatic SNVs where the sites are homozygous in normal sample but heterozygous in the tumor sample. Other types of variants such as single-nucleotide polymorphism (variants between normal samples) and homozygous SNVs (variants that are heterozygous in the normal sample but homozygous in the tumor sample) are not in the scope of this study. To conclude, this study demonstrates that cell-level mutations can be detected from scRNA-seq data using SCmut. The identified mutations specific to cells can facilitate the characterization of the cell-to-cell heterogeneity, for instance in identifying tumor/non-tumor cells, assessing individual drug-response, profiling cell subclones, etc.

Acknowledgements

This work is partially supported by funding from the Swedish Cancer Foundation, the Swedish Research Council (VR) and the Swedish Foundation for Strategic Research (SSF). We thank Prof. Hao-Ook Lee for the breast cancer data. We acknowledge the TCGA Research Network, National Cancer Institute and National Human Genome Research Institute, Bethesda, MD, USA, for access to the TCGA breast cancer data. We also acknowledge the CINECA, the Italian computing centre for supporting computation of this study. We acknowledge Dr. Aaron Diaz, the UCB Professor for access permission to the brain tumor data set deposited at the European Genome-phenome Archive under accession number EGAS00001001900.

References


Fig. 7. Analysis of cell-level mutations of the glioblastoma dataset. (a) The plot of VAFs between the point mutation chr5:133,837 and the in-frame 24bp deletion in gene PDGFRα, having a high Pearson correlation of 0.91. The VAF of the deletion is the proportion of reads that support the deletion event. (b) The overall survival of the TCGA-GBM patients with mutations in the genes discovered by SCmut, including PDGFRα, DYNC1LI2 and CHD6 (slashed curve) vs the wild-type group (solid).

Mutation detection from scRNA-seq data has some limitations. First, the cell-level mutations must be in theastic regions. This is a general disadvantage of all approaches to detect mutations from RNA sequencing or whole exome sequencing data. Second, the procedure is highly dependent on the quality of the alignment and hence the completeness of the transcriptome annotation. Third, the stochastic mono-allelic expression (Borel et al., 2015) might limit the expression of the mutation sites in single cells. Finally, the detection sensitivity of a mutation is determined by the corresponding gene expression in the cell. An important mutation is statistically detectable from scRNA-seq only if it belongs to a highly expressed gene. It is challenging to assess cell-level mutations in genes with low or no expression, since the low expression could be a loss-of-function effect, but could also be due to the low coverage of scRNA-seq data, leading to false negatives. Hence, in order to detect cell-level mutation events, we recommend high coverage scRNA-seq. Further discussion about the detection sensitivity and the coverage threshold of SCmut is presented in the Supplementary document.

It is worth noting that SCmut focuses on detection of somatic SNVs where the sites are homozygous in normal sample but heterozygous in the tumor sample. Other types of variants such as single-nucleotide polymorphism (variants between normal samples) and homozygous SNVs (variants that are heterozygous in the normal sample but homozygous in the tumor sample) are not in the scope of this study. To conclude, this study demonstrates that cell-level mutations can be detected from scRNA-seq data using SCmut. The identified mutations specific to cells can facilitate the characterization of the cell-to-cell heterogeneity, for instance in identifying tumor/non-tumor cells, assessing individual drug-response, profiling cell subclones, etc.


