SuperFreq: Integrated mutation detection and clonal tracking in cancer

Christoffer Flensburg[1,*], Tobias Sargeant[2], Alicia Oshlack[3], and Ian Majewski[1,4,*]

[1] Division of Cancer and Haematology, The Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
[2] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Australia
[3] Murdoch Children's Research Institute, The Royal Children's Hospital, Parkville, Australia
[4] Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Australia

[*] Correspondance: flensburg.c@wehi.edu.au; majewski@wehi.edu.au

Abstract

Motivation: Analysing multiple tumour samples from an individual cancer patient allows insight into the way the disease evolves. Monitoring the expansion and contraction of distinct clones helps to reveal the mutations that initiate the disease and those that drive progression; therefore, the ability to identify and track clones using genomics data is of great interest. Existing approaches for clonal tracking typically require the user to combine multiple tools that are not purpose-made. Furthermore, most methods require a matched normal (non-tumour) sample, which limits the scope of application.

Results: We have built superFreq, a cancer exome sequencing analysis tool that calls and annotates somatic SNVs and CNAs and attributes them to clones. SuperFreq makes use of unrelated control samples and does not require matched normal samples. We demonstrate the ability of superFreq to track clones by combining real samples in known proportions to simulating a multi-sample analysis. In addition, we compared superFreq to other somatic SNV callers and CNA callers on exome sequencing data from cancer-normal pairs, including 304 participants gathered from 33 cancer types in The Cancer Genome Atlas (TCGA).

SuperFreq offers a reliable platform to identify somatic mutations and to track clones. SuperFreq recalled 91% of somatic SNVs identified by a consensus of four other methods, with a median of 1 additional somatic SNV per sample that was not found by any other method. CNA calls from superFreq showed good agreement with those generated by Sequenza, or those from ASCAT generated using matched SNP arrays. Using our simulated data set for testing multi-sample clonal tracking, we found that superFreq identified 93% of clones with a cellular fraction of at least 50%, and mutations were assigned to clones with high recall and close to 100% precision. In addition, SuperFreq maintained a similar level of performance for most aspects of the analysis without a matched normal control. SuperFreq is a highly adaptable method and has already been used in multiple different projects.

Availability: SuperFreq is implemented in R and available on github at https://github.com/ChristofferFlensburg/superFreq.

Introduction

Tracking clonal evolution within a cancer can reveal a wealth of information. It can help detect the cause of relapse or drug resistance, identify early driver mutations, or track the course of metastasis. Tracking mutations across multiple samples can also be highly informative, either from patients or from model systems, including animal models of cancer, xenografts or cell lines, which may include many technical replicates or varied experimental conditions. A typical analysis of multiple cancer samples from the same individual involves calling somatic single nucleotide variants (SNVs) (using methods such as multiSNV[1], VarScan 2[2] MuTect[3], SomaticSniper[4] and Strelka[5]) and copy number alterations (CNAs) (using methods such as Sequenza[6], PureCN[7] and ABSOLUTE[8]), then combining the calls within a dedicated clonal tracker (using methods such as PhyloWGS[9], SciClone[10] and PyClone[11]). The analysis will cluster mutations and produce a phylogeny, which reflects the relationship between different clones in the cancer. This multi-step process works well in capable hands, but is sensitive to data quality issues and to parameter choices. In addition, somatic SNV and CNA callers are not optimized for downstream use in clonal tracking, which makes the process of combining the calls challenging.

We have developed superFreq, a pipeline for integrating mutation detection and clonal tracking that is suitable for use with cancer exome data. To achieve reliable clonal tracking, the SNV and CNA calls need to be robust to varying sample quality. We use first order perturbation calculations to propagate uncertainty throughout the analysis, which allows us to take a wide range of error sources into account. SuperFreq also uses a set of reference normal samples (at least 2 samples, 5-10 suggested, preferably sharing technical biases) to improve variance estimates and detect recurring sequencing artefacts. The use of reference normals permits the analysis of cancer samples that lack a suitable matched normal. This is an important consideration, as many cancer samples are collected without a high quality matched normal. While there are many methods that do a subset of these analyses, superFreq provides an integrated platform purpose built for clonal tracking. **Table 1** shows the attributes of superFreq, compared to some commonly used tools for cancer exome analysis.

Below we describe the algorithms underlying superFreq, with additional detail provided in the Supplementary Methods. We demonstrate the performance of superFreq with data from 304 TCGA participants with comparison to other state of the art software. We provide a case study in which superFreq was used for multi-sample clonal tracking using exome data from a patient with acute myeloid leukaemia (AML).

method	call SSNV	SSNV w/o normal	call het SNPs	call SCNA	CNA w/o normal	subclonal SCNA	call clones	multisample	track SCNA
superFreq	yes	yes	yes	yes	yes	yes	yes	yes	yes
mutect2	yes	yes	-	-	-	-	-	-	-
somaticSniper	yes	-	-	-	-	-	-	-	-
strelka	yes	-	-	-	-	-	-	-	-
varScan2	yes	-	-	-	-	-	-	-	-
sequenza	yes	-	yes	yes	yes	-	-	-	-
absolute	-	-	-	yes	-	yes	yes	-	-
pureCN	yes	yes	yes	yes	yes	-	-	-	-
phyloWGS	-	-	-	-	-	-	yes	yes	yes
sciClone	-	-	-	-	-	-	yes	yes	-
pyClone	-	-	-	-	-	-	yes	yes	-

Table 1: *Properties of some of the most used mutation callers and clonal trackers in comparison to superFreq. call SSNV: does the method identify somatic SNVs. SSNV w/o normal: does the method identify somatic SNVs without a matched normal. call het SNPs: does the method identify heterozygous germline SNPs for CNA calling. call SCNA: does the method call somatic CNAs. CNA w/o normal: does the method call CNAs without a matched normal. subclonal SCNA: can the method identify CNAs of multiple different clonalities. call clones: does the method identify clones. multisample: does the method track clones across multiple samples. track SCNA: does the method track CNAs across samples.*

Methods

The overall workflow of superFreq is outlined in **Figure 1** and described below with greater details proved in the supplementary methods.

Initial data inputs to superFreq: The input to superFreq is a set of indexed BAM files for the samples and reference normals, together with metadata of the samples, reference data (reference genome, optional capture regions) and liberal variant calls in VCF format for the samples.

SNV and small indel quality control: SuperFreq filters variants from the supplied VCF file using base quality, mapping quality, and strandedness. Variants present in the reference normals are removed from the analysis of somatic SNVs, but common population polymorphisms are retained for CNA calling.

Somatic SNV calling and annotation: Variants are identified as somatic if they have a significantly higher variant allele frequency (VAF) in the cancer compared to the normals. If there is a matched normal sample we perform a Fisher exact test on the number of variant and reference reads between the cancer and the matched normal. In the absence of a matched normal, a filter is applied to exclude variants with a population frequency > 0.1% (dbSNP[12] and ExAC[13]). Candidate somatic variants that cluster with the germline in the clonal tracking are marked with the *germlineLike* flag. The somatic SNV assessment is summarised in a quality score *somaticP* between 0 and 1 reflecting the confidence that the variant is somatic. For

downstream analysis of somatic variants, we typically use *somaticP* > 0.5 as cut-off, but it can be adjusted to favor precision or recall. Somatic SNVs are annotated using Ensembl Variant Effect Predictor[14], and candidate driver mutations are highlighted through comparison to the Catalogue Of Somatic Mutations In Cancer[15].

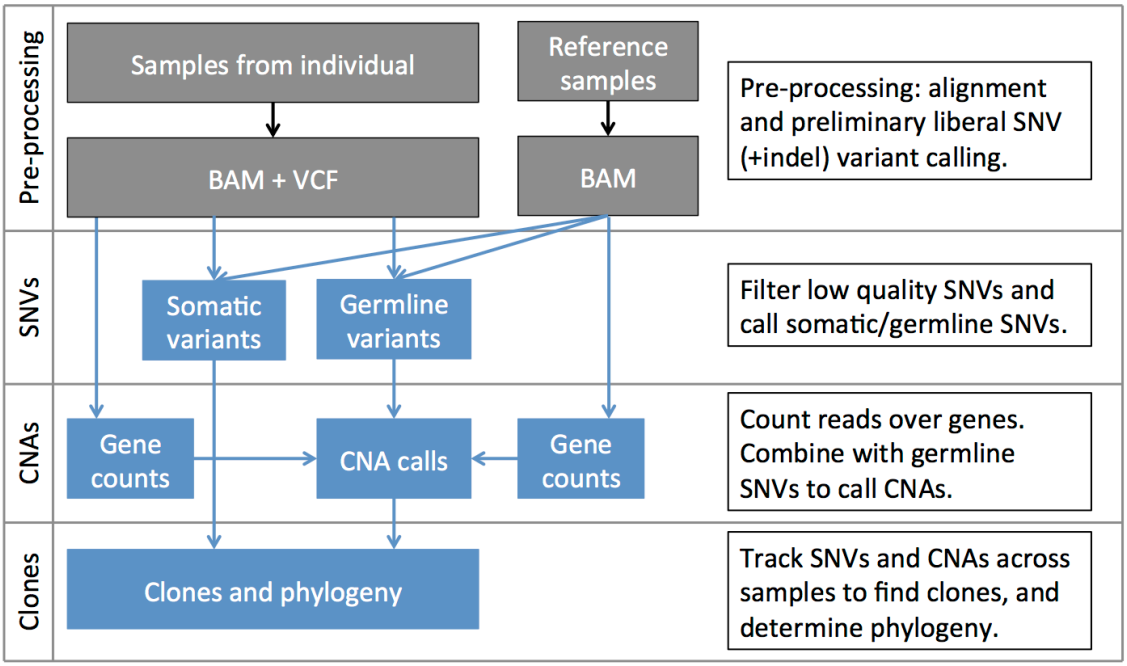

Figure 1: *The workflow of superFreq. The input is aligned* BAM *files from the samples under study, and at least 2 reference normals (5-10 recommended), as well as liberal variant calls. SuperFreq filters the preliminary SNVs for artefacts using quality scores in the* BAM *file, and through comparison to the reference normals. Somatic SNVs are called from the remaining variants, while heterozygous germline SNPs are used for CNA calling. CNAs are identified based on differences in coverage and detecting shifts in allele frequency at heterozygous germline SNPs. Finally, somatic SNVs and CNAs are analysed across samples to designate and track clones.*

CNA calling: SuperFreq uses read coverage and B-allele frequencies (BAFs) at heterozygous germline variants to call CNAs. FeatureCounts[16] is used to determine the read count over each capture region (exon) for each sample. The read counts are corrected for GC-bias and MA-bias against the reference normals. SuperFreq runs limma[17]-voom[18] with sample weights[19] on the corrected counts, comparing each sample one-against-many to the reference normals, resulting in a log fold change (LFC) and t-statistic for each region which can be converted to an uncertainty measure. SuperFreq exploits the expected property that most adjacent capture regions will share the same true LFC, i.e. that the number of true copy number breakpoints is much smaller than the number of capture regions. With that assumption, a median difference between adjacent capture regions larger than expected from the limma-voom variance estimates is a sign of underestimated variance, which is corrected by adding a constant to the variance estimate.

Heterozygous germline SNPs are identified for use in CNA calling. If a matched normal is present, common population variants that are close to 50% VAF in the matched normal are used. If no matched normal is present, then variants with > 1% population allele frequency with a sample VAF between 5% and 95% are used.

The genome is segmented into regions based on the coverage LFC and BAF. The capture regions for each gene are merged and hierarchical clustering is performed. The most similar adjacent segments are merged recursively, with a distance measure comparing LFC and BAF. The ploidy of the sample is then determined from the relative normalisation of the sample with respect to the reference normals. Different segments are allowed different purities for the copy number call. This process is illustrated in the maypole plot in **Supplementary Figure 1**, where ploidy corresponds to a constant shift along the x-axis.

Clonal tracking: The clonality of each somatic SNV is calculated based on the VAF, accounting for local copy number. The clonality of each CNA is tracked over samples, and alterations affecting different alleles are split into separate mutations (e.g. AAB and ABB genotypes). The SNVs and CNAs undergo hierarchical clustering based on the clonality and uncertainty across all samples. The resulting clusters are required to be consistent with a phylogenetic tree. Specifically we require clonal unitarity: that the immediate subclones are not allowed to have a significantly higher summed clonality than that of the parental clone. Inconsistencies are resolved by removing the clone scoring highest in a set of properties typical of false clones, such as constant clonality, high proportion of indels compared to SNVs, or few supporting mutations. The clustering is initially performed with only high confidence somatic mutations. Mutations with lower confidence are then added to the most similar cluster, or discarded if no sufficiently similar cluster is found.

Results

SuperFreq's ability to track clones from exome data relies on using SNV and CNA calls that are generated as part of the analysis pipeline. We therefore assessed these elements of superFreq's performance by comparing to a number of established SNV and CNA callers. To provide a comprehensive sample set, we randomly selected 10 cancer-normal pairs from each of the 33 cancer types included in The Cancer Genome Atlas (TCGA). Of 330 samples that were selected a total of 304 (92%) were successfully downloaded and processed. We also assessed the performance of superFreq without matched normals. Test datasets were established to gauge sensitivity for low purity samples and subclones. To do this we performed in silico dilution and slicing, where we substitute reads from the cancer with those from the normal, either genome-wide or in specific genomic intervals.

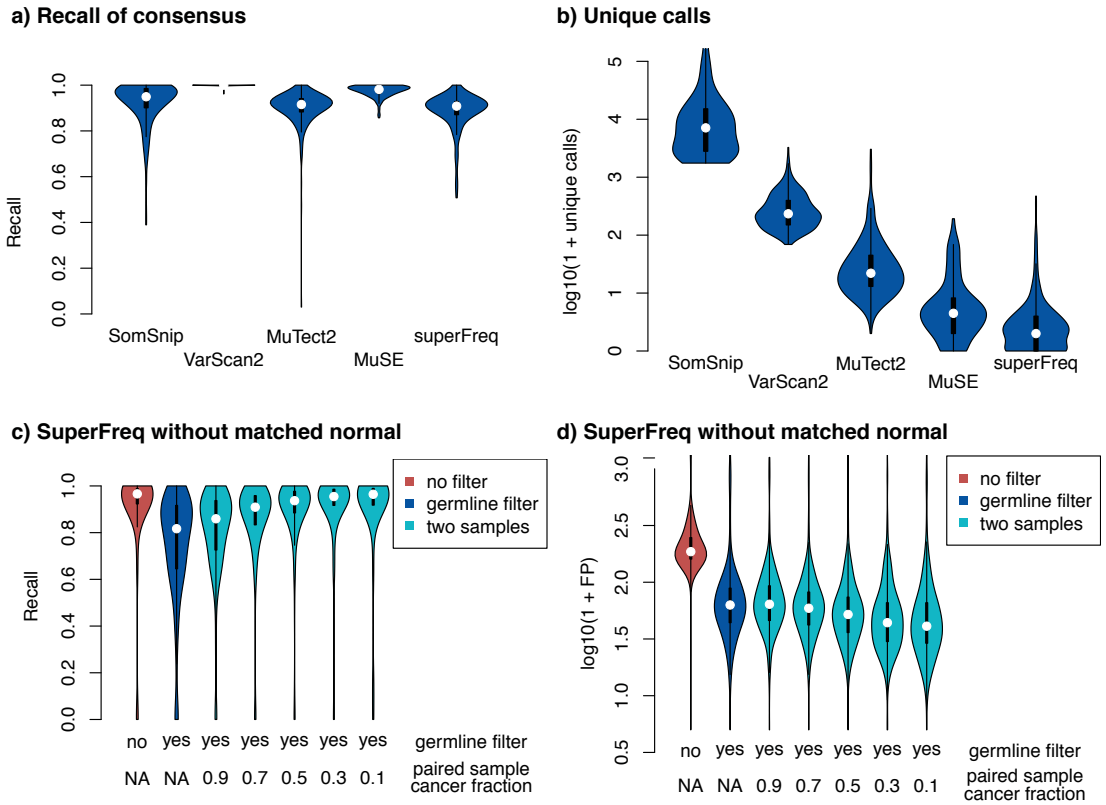

Figure 2 *Precision and recall of somatic SNV calling across 304 TCGA participants and 33 cancer types.* ***a)*** *Recall of somatic SNVs called by the other four callers.* ***b)*** *Number of unique somatic calls generated by each caller.* ***c)*** *Recall of coding somatic SNVs from superFreq without a matched normal, using superFreq cancer-normal analysis as truth. Violins from left: cancer sample alone without filtering on the* germlineLike *flag, cancer sample alone with filtering, cancer sample paired with an in-silico dilution of the cancer and matched normal between 10% and 90%, filtered on the* germlineLike *flag.* ***d)*** *Number of false coding SNV calls in the same sample configurations.*

Somatic SNVs

We compared the somatic SNVs called by superFreq on cancer-normal pairs to calls available through the Genomics Data Commons generated with MuSE[20], SomaticSniper[4], Mutect2[3] and Varscan2[2]. We use consensus calls as a proxy for true somatic variants. We reasoned that, somatic variant calls made by only one out of the five methods are more likely to be false positives. The calls from each variant caller were compared to the consensus calls from the other four methods. SuperFreq detected a median of 91% of variants that were called by the other four callers across the 304 test samples from TCGA (**Figure 2a**). Mutect2 had a similar median (92%), while the other callers were more sensitive, with 95% for SomaticSniper, 98% for MuSE and 100% for Varscan2. However, superFreq only called a median of 1 somatic SNV that was not called by any other method, which was considerably lower than all other methods (**Figure 2b**). MuSE called a median of 3.5 unique variants, Mutect2 called a median of 21, while Varscan2 and

SomaticSniper called 230 and 7100 unique variants respectively. Further distributions for the number of somatic variants called by permutations of two or three callers are shown in **Supplementary Figure 2** using UpSetR[21]. These results highlight the design of superFreq to prioritise accurate variant calls, which is an important consideration for clonal tracking.

Somatic SNV without a matched normal

When a matched normal is not available, superFreq uses population frequencies and clonal tracking to improve the detection of somatic SNVs. We estimated the recall and false positive rate for calls generated without a matched normal by comparing to a truth set generated with matched normal controls. When run with no matched normal superFreq identified a median of 97% of protein changing somatic SNVs detected in the truth set (**Figure 2c**). However, a median of 185 additional protein changing somatic SNVs were also called (**Figure 2d**), which were largely composed of rare germline variants. We next filtered the calls without matched normal using the superFreq *germlineLike* flag, which identifies variants that are present clonally in all samples. The germline filter reduced the median number of false calls to 62, but also lowered the median sensitivity to 82%. This drop in sensitivity is due to clonal somatic mutations being mistaken for germline variants in high purity cancer samples.

If multiple cancer samples are available that differ in tumour purity, they can be used to distinguish germline variants from somatic variants. To simulate this process, we diluted the cancer sample in-silico with sequence data from the matched normal to produce samples with lower tumour purity (10%-90% of the original cancer sample). We analysed the original cancer sample together with the diluted sample and the availability of a sample with lower tumour content helped to separate somatic variants from germline variants. Adding a matched sample with 70% of the original purity and filtering on the *germlineLike* flag brought the median recall rate up to 91% with a median of 58 false calls.

CNA calling

SuperFreq monitors B-allele frequency and shifts in coverage compared to the reference normals to generate allele specific CNA calls. Clonalities of the CNA calls are determined for each segment independently. At the clonal tracking stage, the CNAs are clustered together with the somatic SNVs. We generated copy number calls on the 304 TCGA samples and assessed the performance of superFreq in comparison to calls from Sequenza[6], another exome analysis tool, and ASCAT[22]. The ASCAT calls were generated from matched SNP array data and lifted over from hg19 with `segment_liftover`[23]. Sequenza and ASCAT compare the cancer to a matched normal sample, while superFreq instead uses reference normals. Despite these different approaches, superFreq, Sequenza and ASCAT produced calls that were generally similar, in terms of relative DNA abundance. There was slightly higher agreement between Sequenza and ASCAT (median 93% of the genome), than between superFreq and ASCAT (median 90%) (blue in **Figure 3a**). We repeated the comparison with stricter criteria, requiring similar segments with the same allelic

copy number call (red in **Figure 3a**), which depends on segmentation, ploidy and B-allele frequency. In this comparison we saw large discrepancies between methods, with a median of 13% of the genome in agreement between Sequenza and ASCAT. SuperFreq and ASCAT had a closer agreement with a median of 30%, but with a wide spread across participants.

As superFreq compares the read depth to the reference normals, the algorithm is largely unchanged when running without a matched normal, except that heterozygous germline SNPs are identified directly from the cancer sample instead of from the matched normal. To demonstrate this we compared the superFreq copy number calls on the cancer sample alone to the calls in the matched cancer-normal analysis. The copy number calls by superFreq without a matched normal are very similar to those with a matched normal, with close to 100% agreement for the relative DNA abundance and a median 90% agreement for the stricter comparison (**Supplementary Figure 3**). Similarly, the comparison to ASCAT shows virtually indistinguishable results to those generated with a matched normal.

Next we compared the ploidy calls between superFreq and ASCAT (**Figure 3b**). The majority of the participants had a ploidy call close to 2 using either method. Participants with a high ploidy call in superFreq also had a high ploidy call in ASCAT, but there was a subset of patients with moderate or low ploidy calls in superFreq that had high ploidy calls in ASCAT. Comparing superFreq to Sequenza revealed a similar pattern (**Supplementary Figure 4**). A critical difference between the methods is that ASCAT and Sequenza assume that all CNAs have the same clonality, whereas superFreq accommodates subclonal CNAs. The single clone models used by ASCAT and Sequenza fail to accommodate subclonal CNAs at the true ploidy, but sometimes find a good fit at a higher ploidy. Indeed we observe that participants with large difference in ploidy between superFreq and ASCAT typically have subclones identified by superFreq. An example of a sample with low ploidy and subclones identified by superFreq, but with a high ploidy in Sequenza is demonstrated in **Supplementary Figures 5-6**.

In order to assess superFreq's sensitivity to CNAs covering small genomic regions and those present at low purity, we diluted the cancer sample with the matched normal to simulate lower purity CNAs. We also generated sliced samples in which reads from set regions of the cancer sample replaced those in the normal sample. In this way we created samples with CNAs spanning specific genomic regions, where we could control the size, and could also approximate lower tumour purity. Using the superFreq cancer-normal calls as truth, we measured the rate of recall from superFreq run on sliced and diluted samples as function of the size and clonality of the CNA. We see that above 10Mbp and 30% clonality, almost all CNAs are called, and there is then decreasing sensitivity for smaller events and lower purity. (**Figure 3d**).

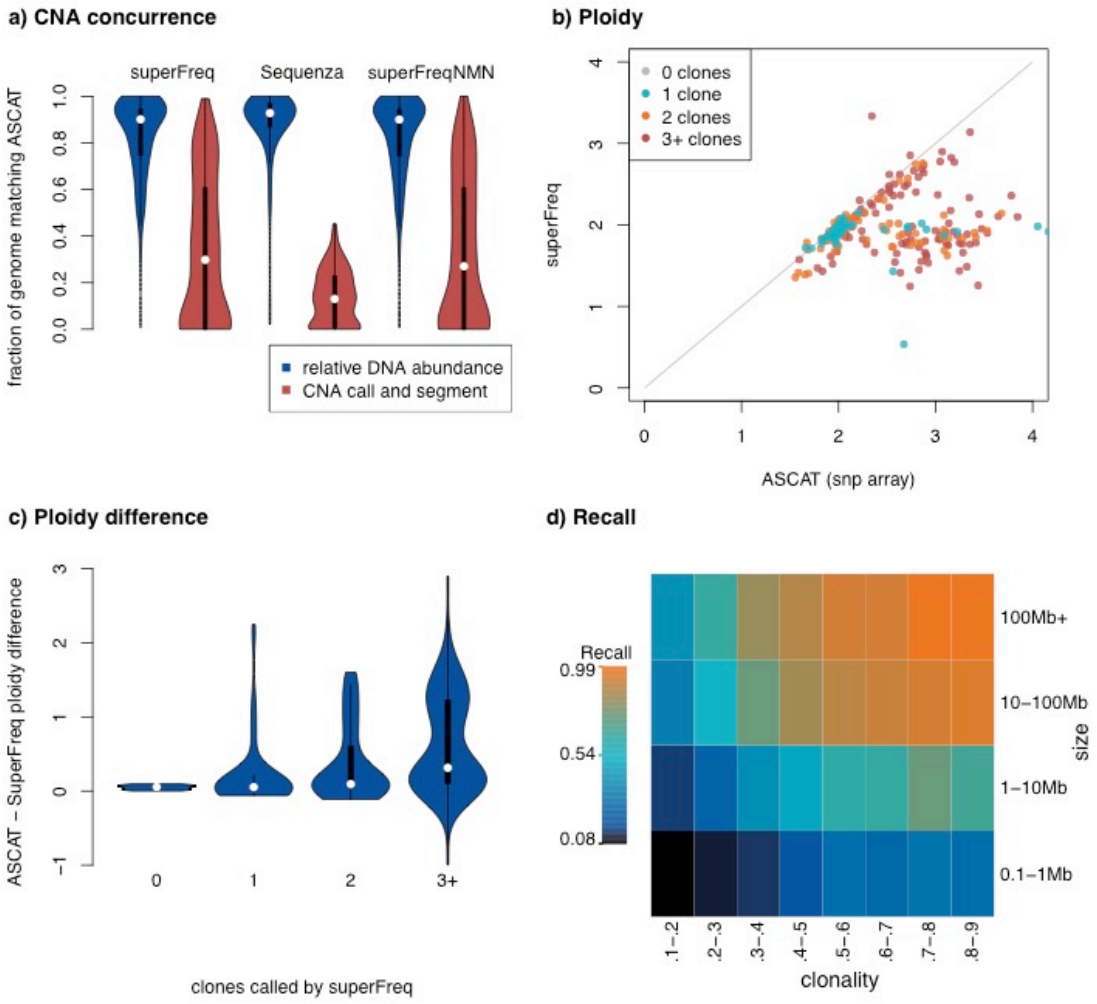

Figure 3: *Comparison of somatic CNA calls for 304 TCGA participants across 33 cancer types.* ***a)*** *Fraction of the genome where two methods agree on the relative DNA abundance (blue) and allele sensitive absolute CNA call and segmentation (red). SuperFreqNMN denotes superFreq run with no matched normal.* ***b)*** *The ploidy calls from ASCAT (SNP array) and superFreq (exomes), with the number of cancer clones called by superFreq shown by colour.* ***c)*** *Difference between ASCAT ploidy call and superFreq ploidy call as function of number of clones called by superFreq.* ***d)*** *Recall of superFreq for somatic CNAs in diluted and sliced samples, with the original cancer-normal used as truth. Each bin is based on at least 100 copy number calls.*

Clonal tracking

SuperFreq identifies clones by hierarchical clustering of the somatic SNVs and CNAs, using a distance measure based on estimated clonality and uncertainty across all samples. This allows designation of any number of clones, as long as each clone has support from at least one mutation.

To assess clonal tracking performance, we generated a test dataset that simulates a multi-sample analysis. To generate the test dataset, we diluted and sliced the cancer

and normal samples to produce a set of three matched cancer samples with four clones as shown in **Figure 4a**. Each of the four clones carries mutations from the cancer sample, but only in a subset of chromosomes. Only participants that had a cancer clone detected in the cancer-normal analysis were included. We then ran superFreq on the matched samples and measured the rate of recall of clones, as a function of the maximum clonality (**Figure 4b** and **c**). We also assessed how many mutations were correctly attributed to each clone. An example participant is shown in **Supplementary Figures 7-9**.

To put the results in context, we compared the performance of superFreq to SciClone. For this comparison we used the superFreq somatic SNV and CNA calls as input for SciClone. SciClone has relatively strict requirements for somatic SNVs, with a default requirement for at least 10 high quality SNVs in regions with normal diploid copy number. As some of the cases did not meet this requirement, the default filters were gradually relaxed until the algorithm could be executed.

Across 289 test datasets generated from TCGA samples, we found that superFreq detected 93% of the simulated clones with a maximum clonality above 50%, compared to 67% of clones detected by SciClone (**Figure 4b**). When considering participants with tumour purity above 75%, superFreq detected four clones in 52% of cases, and three or more clones in 79% of cases, compared to 35% and 64% for SciClone (**Figure 4c**), but SciClone achieved higher sensitivity when considering cases with tumour purity below 30%. SuperFreq had a lower false positive rate, calling a false clone in less than 10% of cases, whereas this was slightly higher for SciClone at 19% (**Figure 4d**).

When considering clones that were correctly identified, we can measure precision and recall of the mutations contributing to the clone. For clones above 50% clonality, superFreq recalled a median of 59% of the mutations with 100% precision, while SciClone recalled 14% of mutations, also with 100% precision (**Figure 4e-f**). This showcases the power of superFreq's two-step clustering, where high quality anchor mutations are used to define the clone before gathering lower quality mutations. In contrast, low quality mutations are discarded by SciClone, explaining the lower level of recall.

We next assessed clonality calling in the absence of a matched normal control. In this analysis superFreq had slightly lower recall for clones above 50% clonality, dropping from 93% to 80%. When considering cases with high tumour purity, superFreq still recalled three or more clones in 67% of cases. There was a marked increase in the fraction of cases in which a false clone was called, increasing from 10% to 80%. The median recall rate of mutations remained similar, but with a small drop in median precision from 100% to 89%.

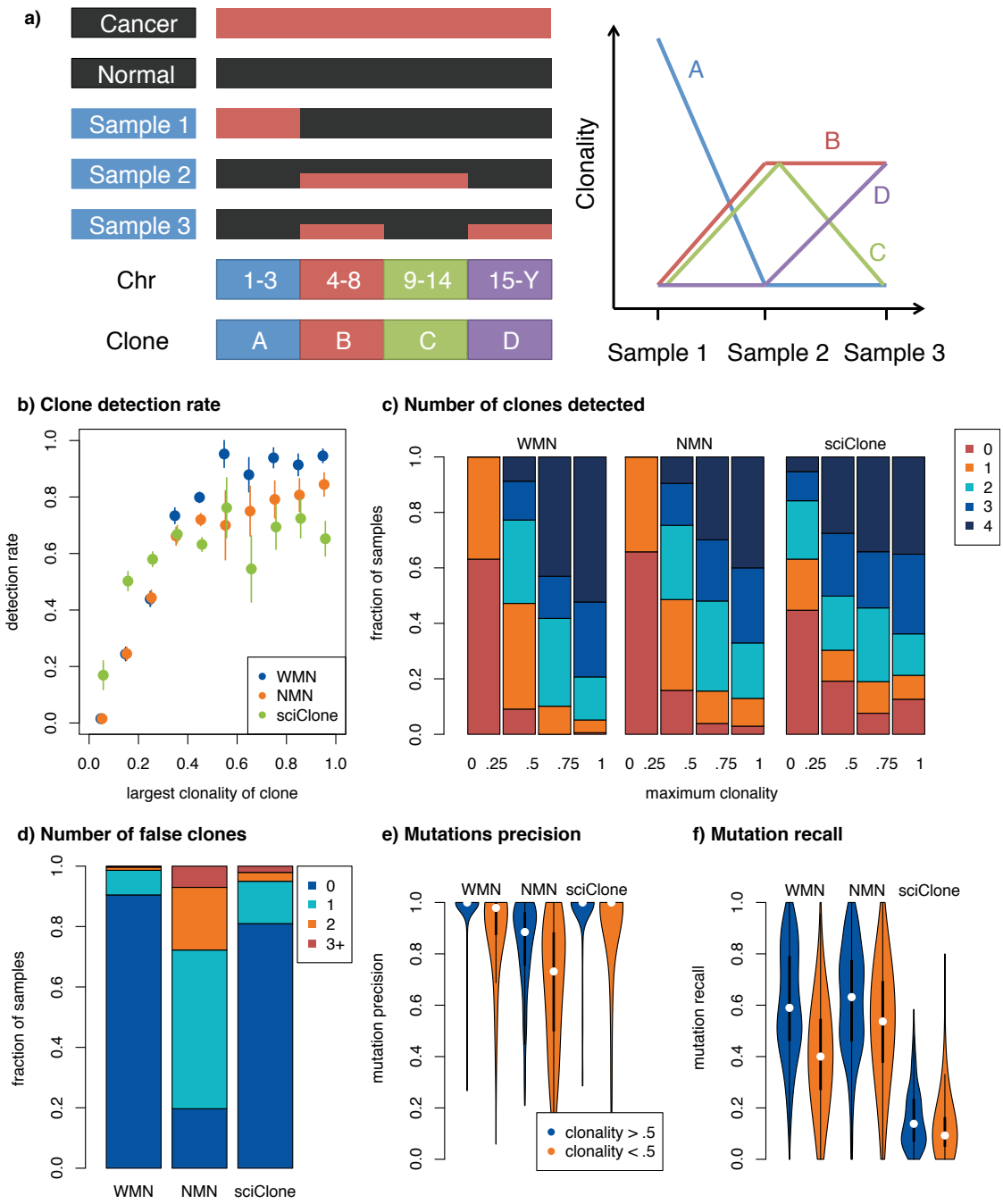

Figure 4: *Precision and recall of superFreq clonal tracking.* ***a)*** *Overview of simulations. As illustrated on the left, the genome is divided into four regions (chr 1-3, 4-8, 9-14 and 15-Y), and the cancer and normal samples are blended to create three samples that contain four clones supported by mutations that reside in different regions of the genome. The expected clonalities across the three samples are shown to the right.* ***b)*** *Sensitivity to find clones with a matched normal (WMN) or no matched normal (NMN) in superFreq and SciClone, as function of maximum clonality.* ***c)*** *Recall of the four simulated clones, binned on the purity of the original cancer sample.* ***d)*** *Number of false clones called.* ***e)*** *Fraction of mutations associated with a clone originating from the expected chromosomes in panel* ***a****.* ***f)*** *Fraction of mutations called in the cancer-normal recalled by the called clones.*

Application

To demonstrate how superFreq can be applied, we present an analysis of a patient with AML with relapsed disease[24]. For this patient, samples were available at diagnosis, at relapse, and from purified lymphocytes (as a matched normal control). We ran superFreq using default parameters with 10 normal samples from the same study as reference normals.

A single, dominant cancer clone (blue) was detected at diagnosis, which was not evident in the matched normal control (**Figure 5**). Candidate somatic variants are prioritised, based on variant effect and comparison to the COSMIC database. In this case four mutations were detected in COSMIC census genes, namely *RUNX1*, *SF3B1*, *FOXP1* and *LONP1*. We also detected a copy number neutral loss of heterozygosity event on chr21 (designated chr21 AA), which extends over 31Mbp and includes *RUNX1*. The *RUNX1* mutation has a VAF significantly larger than 50%, which indicates that this mutation preceded the copy number event and suggests both alleles of *RUNX1* have been inactivated. At relapse a new subclone (red) emerges from the diagnostic clone, which indicates that these cells possess a selective advantage. The relapse-specific subclone was present at around 40% clonality, it carries five protein altering SNVs, none of which are featured in the COSMIC census, together with loss of a 10Mbp segment on chr11 (designated "10Mbp A (11)"). Closer examination of the CNA on chr11 revealed loss of 50 genes, which included the key tumour suppressor gene *WT1*, frequently mutated in AML, which may contribute to the outgrowth of these cells.

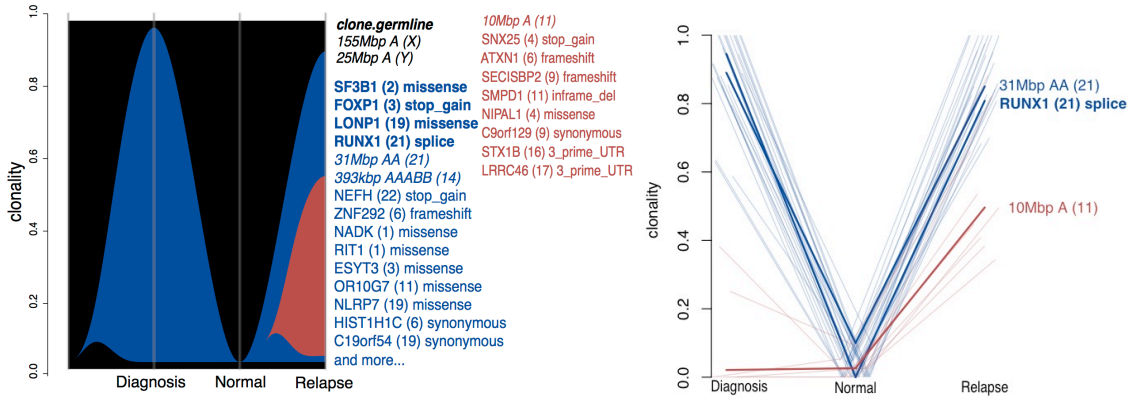

Figure 5: *Example data from AML.084[24]. A river plot (left panel) shows the phylogeny and clonality of called clones and their evolution over time. The mutations from each clone are listed in matching colour, sorted by severity with the chromosome indicated in brackets. The right panel shows the clonalities of the somatic mutations classified into the main cancer clone (blue) with the CNA in chr21 highlighted, and the relapse specific clone (red) with the CNA in chr11 highlighted.*

Discussion

Recapitulating the evolutionary history of a cancer from sequencing data can be tremendously insightful, but is technically challenging. Selecting high quality somatic variants appropriate for clonal tracking is a significant barrier. This can be further compounded by technical imperfections in the sample data or the absence of a high quality matched normal. SuperFreq addresses these challenges; it provides a single workflow that performs somatic SNV filtering, CNA calling and clonal tracking, without requiring a matched normal. Variants are annotated for their impact on the gene/protein as well as against population and cancer databases to aid interpretation and to highlight potential driver mutations.

We opted to test superFreq on data sourced from all TCGA projects to cover a wide range of cancer types and sample quality issues, which are difficult to simulate in silico. Working with real samples meant we did not have a simulation truth, and instead assessed how similar results were between different analysis methods. We found good overlap between the somatic SNV calls from superFreq with four other established methods. As expected, superFreq was the most conservative method in the comparison, closely followed by Mutect2 and MuSE. SuperFreq also generated CNA calls that were in general agreement with those from Sequenza and ASCAT. SuperFreq was more conservative in calling high ploidy, often calling multiple clones instead. The conservative approach to calling is essential to determine the clonal architecture accurately.

To assess the ability of superFreq to reconstruct a clonal history, we developed an approach where we sliced and blended data from TCGA samples to produce sets of samples with an expected clonal structure, with mutations partitioned to specific genomic intervals. This was expected to be a challenging data set to analyse, but superFreq managed to identify 93% of clones above 50% clonality, with fewer than 10% of cases having false clones. This shows that superFreq can reliably perform somatic mutation calling and subsequent clonal tracking with high sensitivity and low false positive rate, without filtering or tuning from the user.

SuperFreq employs a two-step process for clustering mutations into clones: The most confident calls, the anchor mutations are first clustered, and lower confidence calls are then matched and associated with clusters. This allows for accurate clone identification with a relatively high recall of mutations. Indeed, superFreq recalled just above half of the available mutations, which was significantly higher than SciClone, which has relatively strict inclusion criteria, and had a median recall of 14%.

A key aspect to maximise sensitivity while limiting false calls throughout the analysis was to maintain accurate error estimates. Using a single value as the error estimate allowed us to propagate the error throughout the analysis and allowed us to account for more potential error sources. The error estimates inform multiple

steps in the analysis, such as copy number segmentation and calling, mutation clustering into clones, and classification of anchor mutations.

Matched normal samples are commonly used to filter out recurring technical noise and germline variants. SuperFreq accounts for samples that lack a matched normal by adopting a set of reference normals that allows us to identify and filter recurring technical noise. In fact, the reference normals provide improved noise filtering even in the presence of a matched normal, which likely contributes to the low rate of unique somatic SNV calls (**Figure 2b**). SuperFreq also uses clonal tracking to remove germline variants. Germline variants are expected to be present at 100% clonality in all samples, while somatic variants follow the cancer purity. In this way, impurity in the cancer sample helps superFreq separate germline from somatic variants. We found that the availability of a second, matched cancer sample with a modest reduction in tumour purity (~70% of the original) improved our ability to recall somatic SNVs, increasing the median recall from 82% to 91% and provided a small reduction in the false positive rate.

SuperFreq was designed to detect and track somatic mutations in exomes, and it has been applied to study breast cancer metastasis[25], lung cancer xenografts[26] and myeloid leukaemia[27]. We have extended the functionality beyond exomes, and have shown that superFreq can be applied to study small capture sets[28] and low pass whole genomes[29]. We aim to reduce runtime to allow analysis of full depth whole genomes at scale, and have also produced promising results when applying superFreq to transcriptome sequencing data.

Availability

SuperFreq is available as an R package on github:
https://github.com/ChristofferFlensburg/superFreq/
Results from the TCGA analysis and code to reproduce the figures are available at:
https://gitlab.wehi.edu.au/flensburg.c/superFreqPaper

Acknowledgements

We wish to thank beta testers and users for feedback and patience throughout the ongoing development process of superFreq. This includes Aliaksei Holik, Göknur Giner, Jan Schröder, Christophe Lefevre, Gil Hornung, Kirill Tsyganov, Arun Ramani, and more.

The results here are based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The combination of donors, research teams and data sharing facilities at GDC provide an invaluable resource for cancer genomics research. We showed an example from a public diffuse large B-cell lymphoma dataset [phs000450.v2.p1], part of the Slim Initiative for Genomic Medicine

(SIGMA), a joint U.S.-Mexico project funded by the Carlos Slim Health Institute. Another example was from the Epigenetic studies in Acute Myeloid Leukemia [phs001027], which was supported by K08CA169055 (F.E. Garrett-Bakelman), Starr Cancer Consortium I4-A442 (A.M. Melnick, R. Levine and C.E. Mason) and LLS SCOR 7006-13 (A.M. Melnick).

This work was supported by grants from the Australian National Health and Medical Research Council (NHMRC) (Project Grant to IJM 1145912; Independent Research Institutes Infrastructure Support Scheme grant 9000220), the Cancer Council Victoria (grant-in-aid to IJM 1124178), a Victorian State Government Operational Infrastructure Support (OIS) grant; a Victorian Cancer Agency fellowship (to IJM) and a Felton Bequest to IJM. We also wish to acknowledge the generous support of Mr. Malcolm Broomhead who provided philanthropic support for the research.

References

1. Josephidou M, Lynch AG, Tavare S. multiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Res*. 2015;43(9):e61.
2. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
3. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.
4. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28(3):311-317.
5. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28(14):1811-1817.
6. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*. 2015;26(1):64-70.
7. Riester M, Singh AP, Brannon AR, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source Code Biol Med*. 2016;11:13.
8. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*. 2012;30(5):413-421.
9. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol*. 2015;16:35.
10. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*. 2014;10(8):e1003665.
11. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014;11(4):396-398.

12. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311.
13. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291.
14. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
15. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*. 2015;43(Database issue):D805-811.
16. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930.
17. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.
18. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29.
19. Liu R, Holik AZ, Su S, et al. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res*. 2015;43(15):e97.
20. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*. 2016;17.
21. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017;33(18):2938-2940.
22. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2018;173(7):1823.
23. Gao B, Huang Q, Baudis M. segment_liftover : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]. *F1000Research*. 2018;7(319).
24. Li S, Garrett-Bakelman FE, Chung SS, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat Med*. 2016;22(7):792-799.
25. Savas P, Teo ZL, Lefevre C, et al. The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program "CASCADE". *PLoS Med*. 2016;13(12):e1002204.
26. Weeden CE, Holik AZ, Young RJ, et al. Cisplatin Increases Sensitivity to FGFR Inhibition in Patient-Derived Xenograft Models of Lung Squamous Cell Carcinoma. *Mol Cancer Ther*. 2017;16(8):1610-1622.
27. Sanders MA, Chew E, Flensburg C, et al. Germline loss of MBD4 predisposes to leukaemia due to a mutagenic cascade driven by 5mC. *bioRxiv*. 2017.
28. Flensburg C, Sargeant T, Bosma A, et al. Dynamic changes in clonal architecture during disease progression in follicular lymphoma. *bioRxiv*. 2017.
29. Kim EJY, Anko ML, Flensberg C, et al. BAK/BAX-Mediated Apoptosis Is a Myc-Induced Roadblock to Reprogramming. *Stem Cell Reports*. 2018;10(2):331-338.