bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Joint copy number and mutation phylogeny reconstruction from single-cell amplicon sequencing data

Etienne Sollier[1,2], Jack Kuipers[1,3], Koichi Takahashi[4,5], Niko Beerenwinkel[1,3], and Katharina Jahn[1,3]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[4]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[5]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

October 2021

Abstract

Reconstructing the history of somatic DNA alterations that occurred in a tumour can help understand its evolution and predict its resistance to treatment. Single-cell DNA sequencing (scDNAseq) can be used to investigate clonal heterogeneity and to inform phylogeny reconstruction. However, existing phylogenetic methods for scDNAseq data are designed either for point mutations or for large copy number variations, but not for both types of events simultaneously. Here, we develop COMPASS, a computational method for inferring the joint phylogeny of mutations and copy number alterations from targeted scDNAseq data. We evaluate COMPASS on simulated data and show that it outperforms existing methods. We apply COMPASS to a large cohort of 123 patients with acute myeloid leukemia (AML) and detect copy number alterations, including subclonal ones, which are in agreement with and extend current knowledge of AML development. We further used bulk sequencing and SNP array data to orthogonally validate our findings.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Intratumour heterogeneity plays a key role in the failure of targeted cancer therapies [1]. Obtaining a comprehensive picture of the clonal architecture and the mutational history of a patient's tumour at the timepoint of diagnosis therefore offers great potential to improve treatment choices and predict disease progression. Single-cell DNA sequencing (scDNAseq) generally provides a higher resolution of intratumour heterogeneity than sequencing bulk tumour samples. However, this advancement comes at the cost of higher levels of noise primarily introduced during DNA amplification, an essential preparatory step for scDNAseq. As tumours typically evolve through a combination of single-nucleotide variants (SNVs) and copy number variants (CNVs), it has been a critical limitation that current DNA amplification technologies do not permit the reliable calling of SNVs and CNVs simultaneously from the same cells. Multiple Displacement Amplification (MDA) [2], which is used in most scDNAseq protocols, provides a high coverage and has a low error rate and is therefore well suited to detect SNVs. However, MDA results in amplification biases, which preclude reliable detection of CNVs [3]. Other protocols are better suited to detect CNVs but not SNVs, for example a shallow whole-genome sequencing (WGS) as introduced by 10x Genomics [4]. Recently, a high-throughput microfluidics approach was introduced, which processes thousands of single cells while sequencing only a small set of disease-specific genes [5], and later commercialised by Mission Bio, Inc. as the Tapestri® platform. While the limited physical coverage of the genome is far from ideal for calling copy number events, which can stretch anywhere from a small number of bases to whole chromosomes, this approach allows for the use of targeted PCR in the amplification step which does not introduce the strong amplification biases observed in MDA and therefore allows, in princple, to infer both SNVs and CNVs from the same cells [6].

Method development for inferring the evolutionary history of tumours from scDNAseq data closely followed the technology development (Table 1). Initially, approaches have been developed to reconstruct SNV-based mutation histories [7, 8, 9, 10, 11, 12]. Later methods were introduced that analyse the history of copy number variants [13, 14]. SCARLET [15] was the first method for single-cell data that tried to bridge the gap between SNV- and CNV-based tumour phylogeny reconstruction. It infers an SNV phylogeny with CNV-constrained loss of heterozygosity (LOH), but the CNV tree has to be obtained separately, from different cells of the same tumour. BiTSC^2 [16] is the only existing method that can jointly infer the phylogeny of SNVs and CNVs. Its main drawback is that it assumes that in the absence of copy number events, the coverage is uniform across the genome, which, in our experience, is not the case for amplicon sequencing data. BiTSC^2 also does not model copy number-neutral loss of heterozygosity (CNLOH), and might therefore falsely interpret such events as copy number losses.

Here, we introduce COMPASS (COpy number and Mutation Phylogeny from Amplicon Single-cell Sequencing), a statistical model and inference algorithm that can reconstruct the joint phylogeny of SNVs and CNVs from single-cell amplicon sequencing data. Its key features are that it models amplicon-specific coverage fluctuations and that it can efficiently process high-throughput data of thousands of cells. We show in simulation studies that COMPASS outcompetes BiTSC^2 in tree reconstruction accuracy in all settings with realistic coverage variability. Moreover, COMPASS calls CNVs more conservatively than BiTSC^2 whose false positive rate we find to be up to 32 times higher. We apply COMPASS to a large cohort of 123 patients with acute myeloid leukemia (AML) [17] and orthogonally validate our findings with bulk sequencing and SNP array data.

Results

Probabilistic model for joint SNV and CNV single-cell tumour phylogenies

We have developed COMPASS, a likelihood-based approach to infer the evolutionary tree of somatic events in a tumour from single-cell panel sequencing data. The set of somatic events considered by COMPASS comprises SNVs, CNVs (loss or gain) and CNLOH. COMPASS uses as input the reference and mutated read counts, for each variant in each cell, and the number of reads covering each region (Figure 1). The variant

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Method	SNVs	CNVs	Doublets	SNV Recurrence	SNV loss	Homozygous mutations	Est. max # cells	Est. max # loci
∞SCITE [7, 8]	YES	NO	YES	YES*	YES*	NO	10000	100
SCIΦ [9]	YES	NO	NO	NO	NO	NO	100	1000
OncoNEM [10]	YES	NO	NO	NO	NO	NO	100	100
SiCloneFit [11]	YES	NO	YES	YES	YES	NO	100	100
SPhyr [12]	YES	NO	NO	NO	YES	NO	100	100
SCICoNE [13]	NO	YES	NO	-	-	-	100	-
CHISEL [14]	YES§	YES	NO	-	-	-	1000	-
SCARLET [15]	YES	NO†	NO	NO	YES‡	NO	100	100
BiTSC² [16]	YES	YES¶	NO	NO	NO	YES	100	100
COMPASS	YES	YES	YES	NO	YES‖	YES‖	10000	100

Table 1: List of methods for tumour phylogeny inference from scDNAseq data, with their main features. The maximum number of cells and loci are estimates for reasonable runtimes and performance. *However model selection is not automated. §Can assign SNVs to clones after the CNV-tree is inferred by aggregating all cells assigned to each clone. †Requires CNV tree as input, which must be obtained with another method. ‡If supported by copy-number loss; which could miss CNLOH. ¶Only in regions with SNVs. ‖With copy number loss or CNLOH.

read count is the main source of information to infer SNVs, and the total number of reads in each region is used to detect CNVs. A region is the smallest genomic entity for which a copy number event can be detected by our model. For panel sequencing data, we define regions at the level of individual genes by accumulating read counts of all amplicons targeting the same gene. A region may contain no variant (like region B in the example of Figure 1), one variant (region C) or several variants (region A). When variants are present in a region, the CNV calls are allele-specific and COMPASS takes into account the expected ratio between mutated and wild type read counts. Germline SNPs can also be included in addition to somatic SNVs to improve the CNV inference. When this is done, COMPASS will automatically detect that these variants are present in the non-neoplastic cells and will place them at the root of the tree.

In a tree of somatic evolutionary events, each node implies a genotype, which is obtained by altering the wild type diploid genome by the sequence of events defined by the path from the root to the node. By assigning cells to a genotype associated with a tree node, the likelihood of the observed cell-specific read count profiles can be computed, as is described in the methods. In order to compute the likelihood of the tree of somatic events, COMPASS marginalizes out the assignment of cells to node genotypes, which is much more computationally efficient than sampling the attachments of cells to nodes when the number of cells is high. To account for the major sources of noise in scDNAseq data, COMPASS models sequencing errors, allele-specific dropout rates, and doublets. For tree inference, we define a prior distribution on trees that penalizes the number of nodes and of CNV and CNLOH events to explain the observed sequencing data. A simulated annealing algorithm is then used to infer the tree that maximizes the posterior probability.

Evaluation on synthetic data

We evaluated COMPASS on synthetic data and compared it against BiTSC² [16], which is the only other method that can infer a joint SNV- and CNV-based tumour phylogeny. We also included SCITE [7], an established method of SNV-based tumour phylogeny, in order to highlight the benefits of joint SNV and CNV inference over SNV only inference. We generated data that resembles data produced by the Tapestri® platform, as described in Supplementary Section D.1. We used 2000 cells, 20 regions and trees with different numbers of nodes, SNVs and CNVs, and 2 CNLOH events. The Tapestri® platform produces data where the coverage is not uniform across amplicons, since each pair of primers has its own efficiency (Supplementary Figures C.1 and C.2), so we varied the variance in coverage across regions. We evaluated the performance by MP3 similarity [18] between the inferred and the true tree. The MP3 similarity is defined on mutation trees where each node contains a set of mutations, and can be applied to trees which do not have exactly the same set of mutations. Here, we assigned a unique label to each SNV and to each CNV (defined by the affected region and whether the CNV is a gain or a loss), such that the MP3 similarity captures the correctness of both the detected CNVs and the inferred tree topology.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

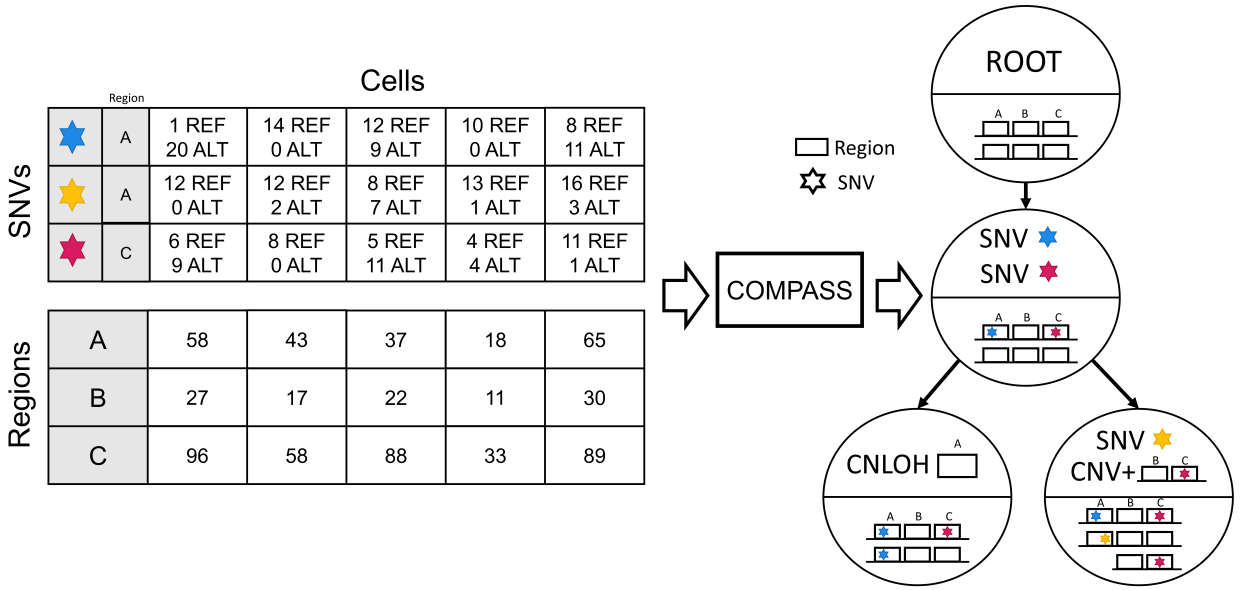

Figure 1: Overview of the input and output of the COMPASS algorithm. Each region (typically a gene) can contain zero, one or several variants. COMPASS needs the number of reads in each region in each cell and the number of reads supporting the reference (REF) and alternative (ALT) allele for each variant. COMPASS infers a tree of somatic events (SNVs, CNVs and CNLOHs). These somatic events imply a genotype for each node, which is depicted in the lower part of each node. Here, the tree contains 2 SNVs in the first clone, then one CNLOH in the bottom-left clone and a copy number gain and an additional SNV in the bottom-right clone.

COMPASS was found to perform best in all settings we analyzed (Figure 2). BiTSC^2 did not perform as well, in particular when the coverage was not uniform across regions and when the number of SNVs was low. SCITE performed well when there were no CNVs and when each node contained exactly one SNV (Figure 2 top right), but its performance dropped when CNVs were present. This is expected, since the true tree cannot be perfectly reconstructed by SCITE in this setting, which highlights the benefit of joint SNV and CNV phylogeny to recover a more comprehensive picture of tumour evolution.

The lower performance of BiTSC^2 can in part be explained by some of the assumptions of its model which may not be applicable to targeted scDNAseq data. First, BiTSC^2 does not allow CNLOH events and might therefore falsely interpret them as copy number losses. In addition, it does not allow losses of the mutated allele and only allows a copy-number gain of the mutated allele when it occurs in the same node as the corresponding SNV. Furthermore, BiTSC^2 only uses the coverage at SNVs to detect CNVs, which might miss CNVs in regions without SNVs. This explains the lower performance of BiTSC^2 when the number of SNVs is low, because in this case there are more regions not covered by SNVs. Moreover, BiTSC^2 assumes that, in the absence of CNVs, all positions in the genome have the same coverage, which is not the case with targeted sequencing. We found that the performance of BiTSC^2 decreased when the coverage was less uniform across regions. In particular, when the variance in coverage is set to the value estimated from real data, the performance of BiTSC^2 drops sharply compared to when the coverage is uniform. Even when we generated data according to the BiTSC^2 model, which is less realistic, an uneven coverage across amplicons was sufficient to prevent BiTSC^2 from recovering the correct tree (Supplementary Figure D.3). Finally, BiTSC^2 was previously only applied to datasets of up to 500 cells and might not scale well to datasets generated by the Tapestri® platform, which often contain up to 10000 cells. In fact, we found that BiTSC^2 performed worse with a higher number of cells, even when we used a high number of iterations in the MCMC sampling (Supplementary Figure D.1). This might be because BiTSC^2 samples attachments of cells to clones using Gibbs sampling instead of marginalizing over the attachments, which might make the MCMC get stuck in local optima when the number of cells is too high. To address this issue, we subsampled the input of BiTSC^2 to 100 cells, which improved its performance. In addition, we note that BiTSC^2 was much slower than COMPASS on large datasets, to the point where its runtime was prohibitively long for datasets of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

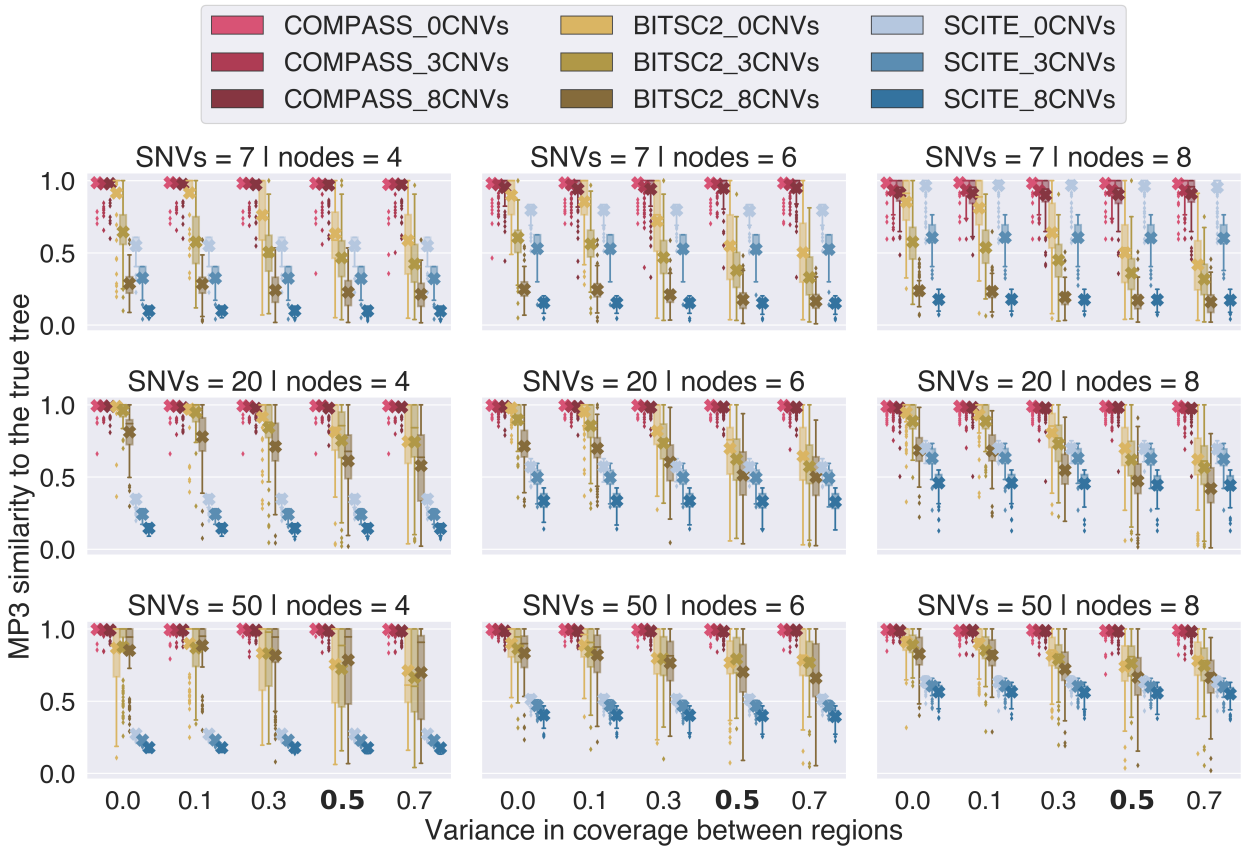

Figure 2: Evaluation of COMPASS, BiTSC² and SCITE on synthetic data, with different variances for the region weights (0: all regions have the same coverage; high values: regions have very different coverages; the highlighted 0.5 corresponds to the value estimated from real data). Here, we used trees with a varying number of nodes (4, 6 or 8), SNVs (7, 20 or 50) and CNVs (0, 3 or 8), and 2 CNLOH events. For each setting, we generated 100 different trees. The crosses represent the mean MP3 similarity to the true tree and the boxplots show the median and quartiles.

size of those generated by the Tapestri® platform (Supplementary Figure D.7).

We also ran simulations with only SNVs (Supplementary Figure D.4). SCITE performed much better in this context, but the model used by COMPASS contains several improvements over SCITE which allow COMPASS to outperform SCITE in some settings, even in the absence of CNV and CNLOH events. In particular, COMPASS outperforms SCITE when the subclones have very different sizes because COMPASS allows each node to have its own attachment probability.

All methods were found to be very robust to the presence of doublets in the data (Supplementary Figure D.2). Only when the doublet rate is very high does the performance drop, and this can be alleviated by using the models of COMPASS or SCITE which explicitly account for doublets (at the cost of an increased computational time).

We found that the performance of COMPASS decreases with an increasing number of copy number events, especially in settings where the number of events approaches the number of nodes in the tree (Figure 2). As described in the Methods section, COMPASS first infers the best tree without CNVs, identifies regions whose coverage at one node differs from their coverage at the root, and selects those regions as candidate regions which might harbour a CNV. COMPASS then looks for the best tree, but allowing only CNVs in these selected regions. This approach drastically reduces the number of false positive CNV calls, but decreases the sensitivity to detect subclonal CNVs. If a CNV is located in a subclone which contains an SNV or LOH event, the subclone will be present in the tree without CNVs, and the corresponding region should be selected, enabling the detection of this CNV. However, if a subclone is only defined by a CNV, it will

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be missing from the tree without CNVs, and the CNV will not be detected. To quantify this phenomenon, we say that a CNV is supported by SNVs if the CNV is in a node that contains an SNV, a CNLOH or a CNV resulting in a LOH, or it has a descendant containing such an event. As expected, the false negative rate of COMPASS for CNVs not supported by SNVs is high (Table 2), but for CNVs supported by SNVs it is much lower than that of BITSC^2. The decreased ability of COMPASS to detect CNVs in subclones not supported by SNVs is counterbalanced by a very low false positive rate. In contrast, BiTSC^2 is significantly less conservative in calling CNVs and has a very high false positive rate.

	COMPASS	BiTSC^2
FPR, overall	0.01	0.32
FNR, overall	0.23	0.83
FNR, CNVs supported by SNVs	0.10	0.80
FNR, CNVs not supported by SNVs	0.73	0.92

Table 2: False positive rate (FPR) and false negative rate (FNR) of CNV detection by COMPASS and BiTSC^2 (regardless of the position in the tree). Here, we used the simulation setting that is closest to the real AML data analysed in this work (trees with 8 nodes, 7 SNVs, 3 CNVs, 2 CNLOHs and a variance in region coverage of 0.5) and averaged the results over 100 random trees. CNVs supported by SNVs are CNVs which are located in a node which contains a SNV or a LOH, or which has a descendant containing such an event.

Correlations between the coverage at different amplicons

When there are no CNVs, we would expect the sequencing depth on each amplicon to be independent. However, we observed strong correlations between the relative sequencing depth on different amplicons (Supplementary Figures C.3 and C.4). Such correlations in Tapestri® data have not been reported before. The biological explanation for these correlations is not clear, but they have the potential to confound the CNV inference, since we could interpret the two main clusters as two different clones with very different copy number profiles. However, these correlations are independent from the actual clonal architecture of the tumour, so by jointly inferring SNVs and CNVs, only the true CNVs should be detected. We simulated data with such correlations between the coverage of different regions, and verified that these correlations did not affect the results of our method (Supplementary Figure D.5).

Overview of CNVs and CNLOHs detected in real AML data

We applied COMPASS to the cohort of 123 AML patients that were previously analyzed with the Tapestri® platform [17]. These samples were sequenced using two different panels: 67 samples with a 50-amplicon panel covering 19 genes and 53 samples with a 279-amplicon panel covering 37 genes (the genes covered by the panels are shown in Supplementary Table 1). In total, COMPASS detected CNVs in 16 samples (Table 3) and CNLOH events in 35 samples. Not surprisingly, more CNVs were detected in samples analyzed with the 279-amplicon panel than with the smaller 50-amplicon panel. For example, we found 4 samples with CNVs on chromosome 8 with the larger panel, but the smaller panel does not contain any amplicon on this chromosome. The most common CNV is a loss of the *EZH2* gene on chromosome 7, which we detect in 8 samples (6.5% of the cohort). Loss of chromosome 7 (or deletion of the long arm of chromosome 7) is indeed known to be common in AML [19]. In AML, mutations in the *TP53* gene are known to be associated with a complex karyotype (multiple CNVs) [20]. We tested if we could also find such an association based on our analysis. Reassuringly, we indeed detected mutations in *TP53* in 9 samples, 5 of which also had a CNV event ($p = 0.0013$, Fisher's exact test, one-sided).

The gene for which we detected the largest number of CNLOH events was *FLT3* ($N = 12$ samples, 9.8%), followed by *SRSF2* ($N = 7$, 5.7%) and *RUNX1* ($N = 4$, 3.3%). *FLT3* is known to be commonly affected by CNLOH in AML, especially when there is an internal tandem duplication [21, 22]. CNLOH on *RUNX1* is also known to be frequent [23]. However, *SRSF2* has not been reported to be frequently affected by CNLOH. The amplicons on *SRSF2* present in the panel have a very low efficiency, resulting in a very high dropout rate, which might be falsely interpreted by COMPASS as a LOH. Interestingly, all 3 samples which had the

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

JAK2 pV617F mutation also had a CNLOH (in patients AML-02, AML-89 and AML-92, Supplementary Figures F.12 to F.14), in agreement with previous reports of CNLOH for this mutation [24].

Orthogonal validation of COMPASS-derived CNV and CNLOH calls with bulk data

Bulk targeted sequencing covering 297 genes was available for 85 samples of the 123 samples. We used CNVkit [25] to detect CNVs in these samples. In addition, we had SNP array data available for 32 samples, for which we used ASCAT [26] to detect both CNVs and CNLOHs. These bulk data provide an opportunity to orthogonally validate the CNV and CNLOH calls of COMPASS with more established (but lower resolution) approaches. We restricted the validation to events present in more than 50% of the cells, since bulk data cannot reliably detect CNV and CNLOH events present in a small percentage of the cells.

Among the 16 samples for which we detected CNVs, bulk SNP array data was available for 2 of them and bulk targeted sequencing for 6 of them. Among these 8 samples, all of the CNVs present in a majority of the cells identified by COMPASS were also detected by bulk sequencing, except for one in sample AML-60-001 (Figures 3 and 4, and Supplementary Figures E.1 to E.12).

In the 85 samples with bulk sequencing, only 4 contained CNVs detected in bulk data on regions covered by the single-cell panel that were not detected by COMPASS (Supplementary Figures E.13 to E.18), and some of them might be false positive calls in the bulk data. Among the 32 samples for which reliable SNP array data was available, six of them contained CNLOH events detected by both COMPASS and ASCAT, two samples contained CNLOH events detected by COMPASS but not ASCAT, and four samples contained CNLOH events detected by ASCAT but not COMPASS, but those were either in regions not targeted by any amplicons or where we did not detect any SNVs (Supplementary Figures E.19 to E.26).

For sample AML-59-001, COMPASS inferred a tree containing two main clones, each of which has a different mutation in the *RUNX1* gene (Figure 3). In addition, the dominant clone has one deletion of *EZH2* on chromosome 7, and one amplification of *WT1* on chromosome 11, while the smaller clone has a loss of *TP53* on chromosome 17, which results in a LOH for one germline variant (the sample might also have a somatic mutation on *TP53* not captured by the panel). The ASCAT profile inferred from SNP array data also contains the deletion on chromosome 7 and the amplification of chromosome 11, but does not contain any loss on chromosome 17. This is expected, as this deletion is only present in 5% of the cells and hence is unlikely to be detected from a bulk sample. This example supports the correctness of the upper part of our tree since the CNVs found in the dominant clone are also detected with an orthogonal method. It also shows that our method can take advantage of the single-cell resolution of the Tapestri® data to uncover CNVs in small subclones which are missed by bulk methods, and to reveal the order of SNV and CNV events in a branching clonal architecture.

The inferred tree for sample AML-99-001 displays a linear evolution and contains one amplification of two genes on chromosome 8, as well as a CNLOH of *RUNX1* on chromosome 21 (Figure 4). One germline variant on *RAD21* was covered by the targeted panel, which improves the reliability of the CNV call on chromosome 8, since it is based both on the total coverage in the region and the allelic fraction of the *RAD21* SNP. The ASCAT profile inferred from SNP array data also contains this copy number gain on chromosome 8 and the CNLOH on chromosome 21, again supporting our tumour phylogeny. Interestingly, there are 5 longitudinal samples available for this patient, and we detect the copy number gain on chromosome 8 and the CNLOH on *RUNX1* in all 5 of them, although the CNLOH on *RUNX1* is only present in a small subclone on the fourth and fifth samples (Supplementary Figures F.8 to F.11).

Two independent deletions on chromosome 17

Sample AML-101-001 provides an interesting illustration of the benefits of joint SNV and CNV phylogeny inferred from single-cell data for understanding the evolution of a tumour. This sample contains two different mutations in *TP53* (on the two different allelic copies) and COMPASS inferred two independent deletions on chromosome 17 (Figure 5). In the first deletion, all three genes present in the panel on this chromosome

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Patient	Panel	CNV
AML-07	50-amplicon	+1 FLT3 (chr13) -1 RUNX1 (chr21) -1 U2AF1 (chr21)
AML-39	50-amplicon	-1 EZH2 (chr7)
AML-59	50-amplicon	-1 EZH2 (chr7) +1 WT1 (chr11) -1 TP53 (chr17)
AML-60	50-amplicon	+1 ASXL1 (chr20)
AML-73	50-amplicon	-1 RUNX1 (chr21)
AML-42	279-amplicon	-1 TET2 (chr4)
AML-78	279-amplicon	-1 EZH2 (chr7) -1 RAD21 (chr8) -1 MYC (chr8) -1 ETV6 (chr12)
AML-79	279-amplicon	-1 EZH2 (chr7) -1 TP53 (chr17) -1 SETBP1 (chr18) +1 RUNX1 (chr21) +1 U2AF1 (chr21)
AML-83	279-amplicon	+1 RAD21 (chr8) +1 MYC (chr8)
AML-98	279-amplicon	-1 EZH2 (chr7)
AML-99	279-amplicon	+1 RAD21 (chr8) +1 MYC (chr8)
AML-101	279-amplicon	-1 EZH2 (chr7) -1 TP53 (chr17) -1 NF1 (chr17) -1 PPM1D (chr17)
AML-103	279-amplicon	-1 TET2 (chr4)
AML-107	279-amplicon	-1 EZH2 (chr7) -1 ETV6 (chr12) -1 FLT3 (chr13)
AML-111	279-amplicon	-1 EZH2 (chr7)
AML-117	279-amplicon	+1 RAD21 (chr8) +1 MYC (chr8) -1 TP53 (chr17)

Table 3: List of CNVs detected in the cohort. 67 samples were sequenced with a 50-amplicon panel and 56 were sequenced with a larger 279-amplicon panel. The sign indicates whether the CNV is a loss (-1) or a gain (+1).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

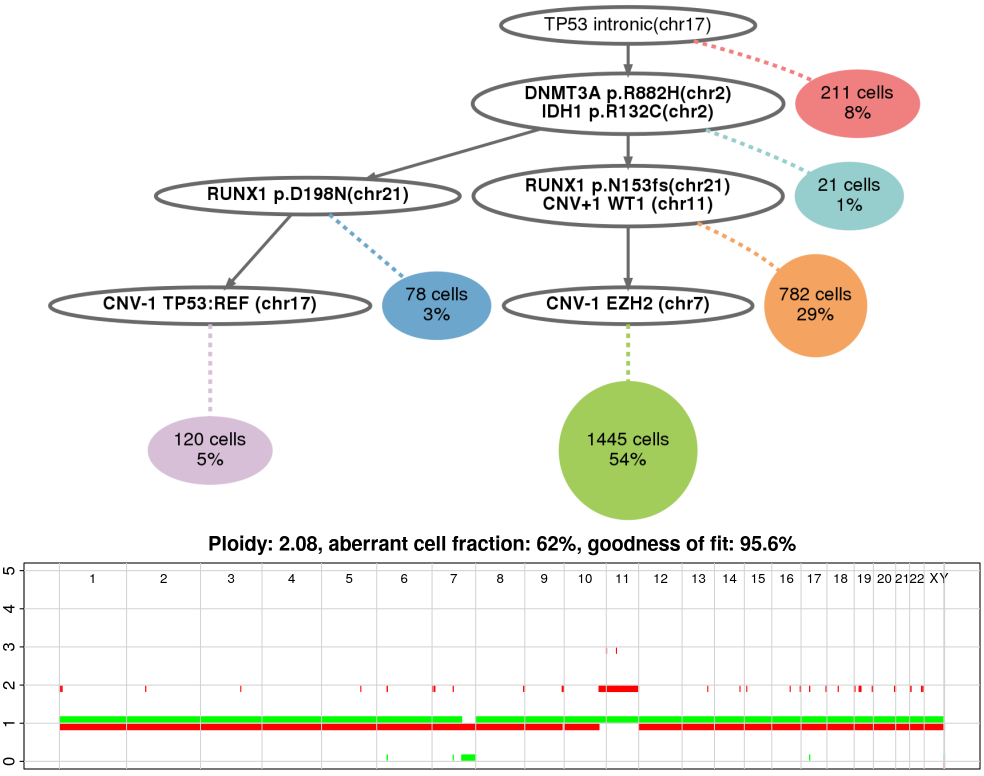

Figure 3: Top: inferred tree for sample AML-59-001, where the dominant clone contains a copy number loss on *EZH2* (chr7) and a copy number gain on *WT1* (chr11), and a small subclone contains a copy number loss on *TP53* (chr17). A germline TP53 variant is shown at the root because it is taken into account by COMPASS to detect the copy number loss. The notation CNV-1 TP53:REF indicates that one copy of the reference allele is lost. Bottom: allelic copy number profile inferred from bulk SNP array data using ASCAT [26]. Both CNVs in the dominant clone are confirmed, but the bulk SNP array data cannot detect a CNV present in only 5% of the cells.

(*TP53*, *NF1* and *PPM1D*) were lost, and in the second deletion only *TP53* and *NF1* were lost. Such double *TP53* mutations are not rare in AML, although they are less common than one *TP53* mutation followed by a LOH [27]. Once both *TP53* alleles are mutated we would not expect any additional fitness advantage from losing one copy, whereas here if two deletions on chromosome 17 were independently selected, it seems likely that this deletion drives oncogenesis. A possible explanation is that the fitness advantage provided by these deletions on chromosome 17 does not come from the loss of *TP53*, but rather from the loss of *NF1*. *NF1* codes for the protein neurofibromin, which is a GTPase activating protein that can accelerate the hydrolysis of RAS-bound GTP into GDP, thus downregulating the RAS pathway. Consequently, a loss of *NF1* could result in an increased activity of the RAS pathway [28]. This proposed mechanism would be consistent with the fact that there are two additional clones which also contain mutations upregulating the RAS pathway (mutations in *KRAS* and *PTPN11*). Thus, this would be a case where there are 4 co-existing clones with different genotypes, but all of these genotypes have the same consequence on the RAS pathway. In this example, integration of SNVs and CNVs into the phylogeny is critical because based on the coverage information alone, it would not be possible to detect that two different copies of *TP53* are lost independently.

Discussion

We have developed COMPASS, a probabilistic model for inferring clonal phylogenies based on point mutations and copy number events from single-cell DNAseq data. COMPASS is geared towards the use of read count data from high-throughput amplicon-based sequencing, for example, as generated by the MissionBio Tapestri® platform. Unlike BiTSC^2 which is currently the only other method to infer tumour phylogenies

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

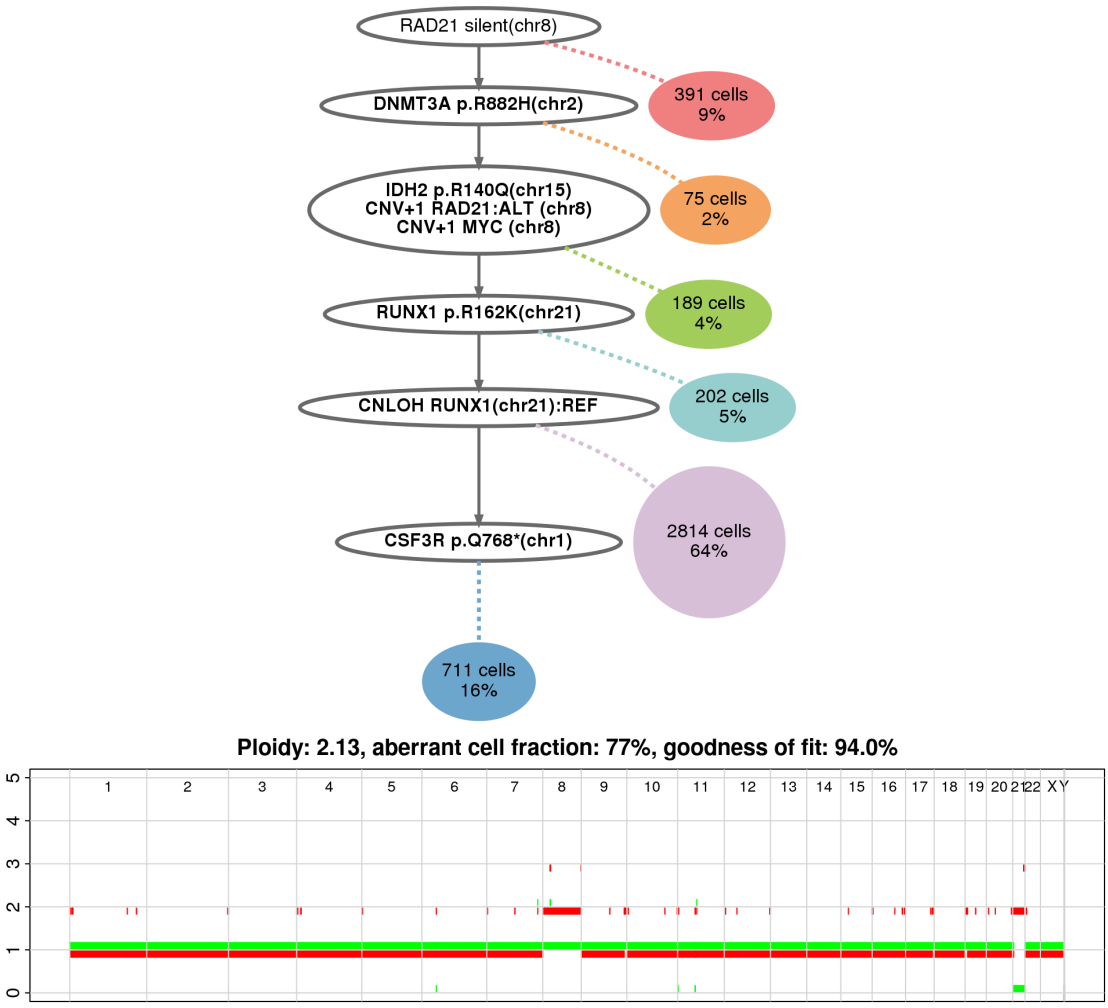

Figure 4: Top: inferred tree for sample AML-99-001, where most neoplastic cells have a copy number gain on chromosome 8 and a CNLOH on *RUNX1* (chr21). Bottom: allelic copy number profile inferred from bulk SNP array data using ASCAT [26]. The CNV on chromosome 8 and the CNLOH on chromosome 21 are confirmed.

based on SNVs and CNVs from single-cell sequencing data, COMPASS can also detect copy-neutral loss of heterozygosity, an important prognostic marker in AML. Our simulation experiments illustrate two further key advantages of COMPASS over BiTSC². First, COMPASS is able to process larger datasets with thousands of cells, and second, it is more robust to systematic local coverage fluctuations between amplicons that are independent of copy number changes. These patterns are most likely introduced by variability in primer pair efficiency in the targeted amplicon-based sequencing. In general, SNVs are easier to call from deep targeted sequencing than copy number states. This stands in contrast to shallow WGS which is better suited for detecting large CNVs than SNVs. Looking at CNV detection alone, our results show that COMPASS outcompetes BiTSC² with regard to both false negative and false positive rate with the latter being up to 32 times higher for BiTSC². We also observe that it is particularly challenging to detect subclones characterized only by CNVs, and in our simulations, COMPASS mostly fails to detect them. This is analogous to how CHISEL can only detect SNVs in subclones containing CNVs [14], for shallow WGS data. The same trend is observed for BiTSC² but less pronounced due to the overall less conservative CNV calling strategy of this method. In practice, many of the subclones seemingly characterized only by CNVs will in fact be supported by SNVs located outside the small set of genes currently targeted in high-throughput assays. Therefore sequencing a larger part of the genome will likely reduce the number of these hard to detect CNVs.

We applied COMPASS to a large real-world dataset of 123 AML samples. Previously, clonal architecture of

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

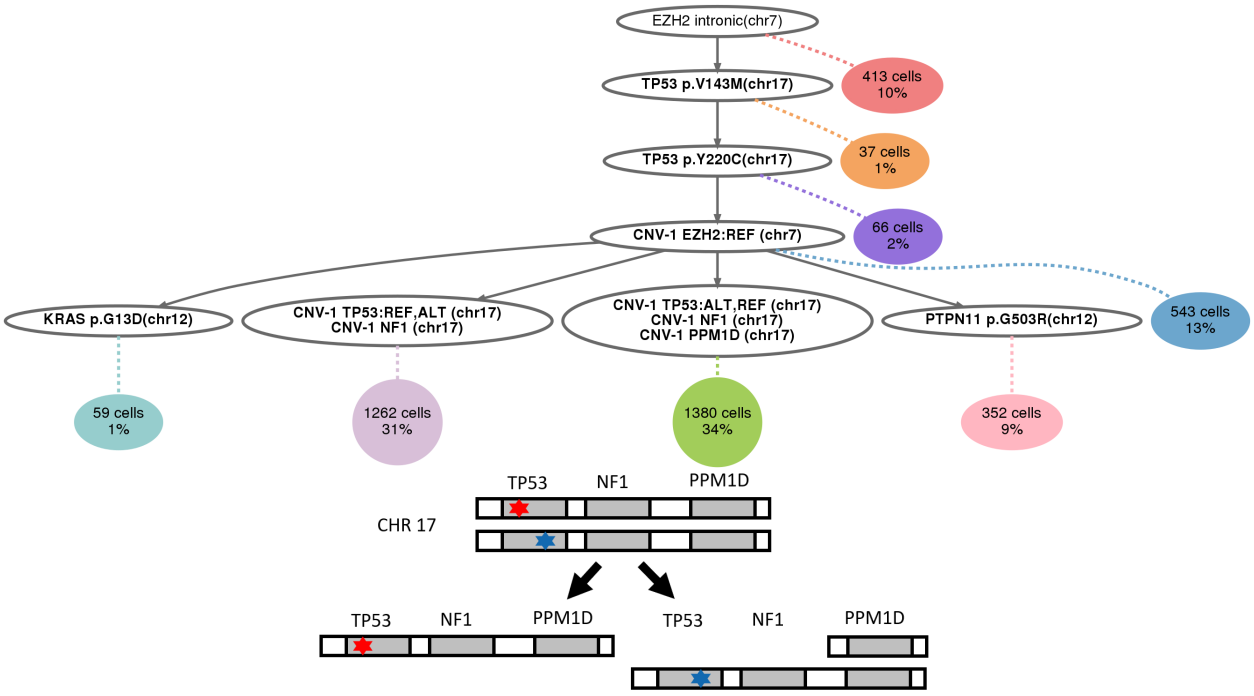

Figure 5: Top: Inferred tree for sample AML-101-001. The notation CNV-1 TP53:ALT,REF indicates that the copy of TP53 that was lost contained the mutation pY220C, but not the mutation pV143M. Conversely, CNV-1 TP53:REF,ALT indicates that the pV143M mutation was lost, but not pY220C. Bottom: sketch describing the corresponding two independent deletions on chromosome 17.

these samples was only inferred based on SNVs. Jointly analysing SNVs and CNVs with COMPASS allowed for a more complete characterization of the clonal heterogeneity in these samples. The main scientific advance provided by COMPASS is the ability to delineate the order of SNV, CNV and CNLOH in a branching evolution pattern, which can help analyze the fitness of clones that are evolving in parallel. Among our most striking findings are a strong association of CNVs with mutations in TP53, and deletions to be most common on chromosome arm 7q. Both findings are in agreement with the current knowledge of AML development. We were also able to orthogonally validate CNVs detected with COMPASS using bulk SNP array data.

While we focus here on AML, COMPASS is not generally restricted to this cancer type. For example, more complex copy number events like whole genome duplications which are uncommon in AML but occur in many solid tumours could be easily modelled in COMPASS by adding one event which doubles every copy number.

Methods

Probabilistic model

COMPASS defines a probability distribution over trees of somatic events (SNVs, CNVs, CNLOHs). The prior on trees penalizes the number of nodes and CNV/CNLOH events, and the likelihood takes into account both the total number of reads in each region and the number of reads at each locus supporting the mutated vs wild type allele. The two components of this likelihood are described in more detail below. A simulated annealing algorithm is used to infer the tree with the highest probability. A complete description of the probabilistic generative process defined for COMPASS is provided in the Supplementary Material section A.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Likelihood for the number of reads in each region

We model the read count in each region with a negative binomial distribution (Gamma-Poisson), which we parameterize with a mean μ and inverse dispersion parameter θ. This corresponds to sampling the read counts from a Poisson distribution, where the rate of the Poisson distribution is first sampled from a Gamma distribution with shape parameter θ and scale parameter $\frac{\mu}{\theta}$. Compared to a Poisson distribution whose variance would be equal to the mean μ, the Gamma-Poisson distribution has a higher variance $\mu + \frac{1}{\theta}\mu^2$.

When there is a total of D_j reads in cell j, reads have a probability ρ_k to fall into region k (in the absence of CNVs). If cell j is attached to node σ_j, which has a copy number of $c_k(\sigma_j)$, for region k, then the expected read count for region k in cell j is $\mathbb{E}(D_{kj}) = D_j \frac{c_k(\sigma_j)}{2}\rho_k$, which leads to the following likelihood:

$$P(D_{kj} \mid D_j \frac{c_k(\sigma_j)}{2}\rho_k, \theta) = \frac{\Gamma(D_{kj}+\theta)}{\Gamma(D_{kj}+1)\Gamma(\theta)} \left(\frac{\theta}{\theta + D_j \frac{c_k(\sigma_j)}{2}\rho_k}\right)^{\theta} \left(\frac{D_j \frac{c_k(\sigma_j)}{2}\rho_k}{\theta + D_j \frac{c_k(\sigma_j)}{2}\rho_k}\right)^{D_{kj}} \tag{1}$$

Likelihood for the number of mutated reads at each variable locus

COMPASS does not take as input called genotypes, but instead works directly with the allelic read counts, similar to SCIΦ [9]. Even when the coverage is low, COMPASS can harness all of the available information while taking the uncertainty into account. In addition, a copy number alteration can lead to one allele having a higher copy number than the other, resulting in an unbalanced allelic proportion, which can be detected from the allelic read counts (Supplementary Figure C.9), making the CNV inference more precise. We model allelic read counts with a beta-binomial distribution to account for overdispersion. Let D be the sequencing depth at a position and A be the counts of alternative reads, f be the frequency of the alternative nucleotide and ω be the concentration parameter. The beta-binomial likelihood is given by

$$P(A \mid D, f, \omega) = \binom{D}{A} \frac{B(A+\omega f, D-A+\omega(1-f))}{B(\omega f, \omega(1-f))} \tag{2}$$

where B is the beta function.

Let $c^{(r)}$ and $c^{(a)}$ be the number of copies of the reference and alternative allele, respectively. The true proportion of the alternative nucleotide is $\frac{c^{(a)}}{c^{(r)}+c^{(a)}}$. Let ε be the sequencing error rate. If we exclude the two other nucleotides different from ref and alt, the proportion of alternative reads should be $\frac{c^{(a)}}{c^{(r)}+c^{(a)}}(1-\varepsilon) + \frac{c^{(r)}}{c^{(r)}+c^{(a)}}\varepsilon$. Each of the two alleles can independently be dropped out: Let k and l, respectively, be the number of reference and alternative allele copies which got amplified. We observed that in real data, different variants had different dropout rates (Supplementary Figure C.8), so we allowed in our model each variant i to have its own dropout rate μ_i, which is inferred using an EM algorithm described below. Taking into account all of the dropout possibilities, the probability of the observed read counts for cell j at locus i is

$$P\left(A_{ij} \mid D_{ij}, c^{(r)}, c^{(a)}, \mu, \varepsilon, \omega_{\text{hom}}, \omega_{\text{het}}\right) = \sum_{\substack{0\le k\le c^{(r)} \\ 0\le l\le c^{(a)} \\ (k,l)\neq(0,0)}} \binom{c^{(r)}}{k}\binom{c^{(a)}}{l} \mu_i^{c^{(r)}+c^{(a)}-k-l}(1-\mu_i)^{k+l} P\left(A_{ij} \mid D_{ij}, \frac{l}{k+l}(1-\varepsilon) + \frac{k}{k+l}\varepsilon, \omega(k,l)\right) \tag{3}$$

where $\omega(k,l) = \omega_{\text{hom}}$ if $k=0$ or $l=0$ and $\omega(k,l) = \omega_{\text{het}}$ otherwise (the overdispersion is higher in case of heterozygosity).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Marginalization over the attachments of cells to nodes

Instead of sampling the attachments of cells to nodes as part of the MCMC scheme, we compute the likelihood of a tree by marginalizing over the attachments and we only sample trees as in SCITE [7]. This makes the inference much faster when the number of cells is high, which is typically the case for Tapestri® data. Unlike SCITE, COMPASS does not use a uniform prior over nodes in the marginalization, but instead learns the probability π_n to sample a cell from a node n. This improves the inference, especially when some clones are much smaller than others, and this is feasible for Tapestri® data because we have a high number of cells and a small number of nodes. The node weights π_n are learnt using an EM procedure described below. If σ_j denotes the node to which cell j is attached, then the likelihood of a tree can be written as

$$P(\boldsymbol{D}, \boldsymbol{A} \mid \mathcal{T}) = \prod_{\text{cell } j} \sum_{\sigma_j} \pi_{\sigma_j} \prod_{\text{region } k} P(D_{kj} \mid c_k(\sigma_j), \rho_k) \prod_{\text{locus } i} P(A_{ij} \mid D_{ij}, c_i^{(r)}(\sigma_j), c_i^{(a)}(\sigma_j), \mu_i) \tag{4}$$

Doublets

Doublets can optionally be modelled. In case they are included, we compute separately the probability of a cell to attach to a single node, and to attach to a doublet, and we mix them with the doublet probability δ, as is done in ∞SCITE [8].

The general formula of Equation (4) remains valid, but the attachment σ_j of cell j can either be a single node or a pair of nodes. In case σ_j is a single node n, the probability to attach to it is $P(\sigma_j) = (1-\delta)\pi_n$. In case σ_j is a doublet (n, n'), the probability to attach to it is $P(\sigma_j) = \delta\pi_n\pi_{n'}$. The genotype of a doublet is computed by adding the copy numbers of the alleles of the two nodes, and averaging the copy numbers of the regions. If we explicitly separate singlets from doublets, we obtain

$$P(\boldsymbol{D}, \boldsymbol{A} \mid \mathcal{T}) = \prod_{\text{cell } j} \left((1-\delta) \sum_n \pi_n P(\boldsymbol{D_j}, \boldsymbol{A_j} \mid n)) + \delta \sum_n \sum_{n'} \pi_n \pi_{n'} P(\boldsymbol{D_j}, \boldsymbol{A_j} \mid n, n') \right) \tag{5}$$

$$\begin{aligned} P(\boldsymbol{D}, \boldsymbol{A} \mid \mathcal{T}) = \prod_{\text{cell } j} \Bigg(& (1-\delta) \sum_n \pi_n \prod_{\text{region } k} P(D_{kj} \mid c_k(n), \rho_k) \prod_{\text{locus } i} P(A_{ij} \mid D_{ij}, c_i^{(r)}(n), c_i^{(a)}(n), \mu_i) \\ & + \delta \sum_n \sum_{n'} \pi_n \pi_{n'} \prod_{\text{region } k} P(D_{kj} \mid \frac{c_k(n) + c_k(n')}{2}, \rho_k) \\ & \prod_{\text{locus } i} P(A_{ij} \mid D_{ij}, c_i^{(r)}(n) + c_i^{(r)}(n'), c_i^{(a)}(n) + c_i^{(a)}(n'), \mu_i) \Bigg) \end{aligned} \tag{6}$$

Tree prior

The tree prior penalizes the number of nodes in the tree, as well as the number of LOH and CNV events.

The penalty for the number of nodes is proportional to the number of mutations in the tree because if the tree contains many mutations, it is more likely that each node will contain several mutations.

When several CNV or CNLOH events affect contiguous regions, they are counted as one event, because such events typically affect large genomic regions, often whole chromosomes. The same penalty is used for CNLOH and copy number losses resulting in a LOH, but a lower penalty is used for CNVs which do not result in a LOH, because they have a lower impact on the likelihood. The penalty for CNLOH and CNV events has an affine relationship with the number of cells, because when more cells are present, such events have a higher impact on the likelihood, but we also need a minimum evidence to be able to detect such events.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Since COMPASS allows the inclusion of germline variants in the tree (to improve the inference of CNVs, in case they are part of a region affected by a loss or amplification), mutations which are not at the root of the tree are penalized. Optionally, COMPASS can take as input the frequency of variants in the 1000 Genomes database. Variants present in this database are penalized more heavily (proportionally to their population frequency) for not being at the root, since they are more likely to be germline variants.

The prior of a tree depends on parameters for the penalties p_1, p_2, p_3 and p_4, which are chosen empirically. The main one is p_2 which controls the addition of CNLOH and CNV events. Its default value works well on MissionBio datasets, but might have to be adjusted in case there are too many false positives or false negatives on other datasets. The default value of p_1 is low because the node probabilities already remove most of the benefits of having additional nodes, but it could be increased if we wanted to reduce the number of nodes in the tree. The values of p_3 and p_4 are not critical because this part of the prior does not play a significant role in most cases. The formula for this log-prior is:

$$\begin{aligned}\log(P(\mathcal{T})) = &- p_1 n_{\text{muts}} n_{\text{nodes}} \\ &- (1500 + n_{\text{cells}}) p_2 \left(n_{\text{CNLOH}} + n_{\text{CNV\_LOH}} + \frac{1}{2} n_{\text{CNV\_no\_LOH}} \right) \\ &- \sum_{\text{locus } i} \mathbb{1}_{i \text{ not attached at the root}} \left(p_3 + p_4 \text{freq}_{\text{1000Genomes}}(i) \right) \\ &+ \text{Constant}\end{aligned}$$

Simulated annealing

Even though the number of mutations with targeted DNA sequencing is small, the tree space is still very large, which precludes an exhaustive search over the whole tree space. Consequently, we use a simulated annealing (SA) approach. At each iteration, we start from a tree $\mathcal{T}$ and propose a new tree $\mathcal{T}'$ by sampling it from a proposal distribution $q(\mathcal{T}, \mathcal{T}')$. The MCMC moves are described in the Supplementary Material section B. Then, we compute the likelihood of the new tree, and accept the new tree with probability $\min\left\{1, \exp\left(\frac{\log(P(\mathcal{T})P(D|\mathcal{T}))}{T}\right)\right\}$ where T is a temperature parameter. Otherwise, we reject the new tree and start a new iteration from tree $\mathcal{T}$. The temperature is progressively lowered, which prevents being stuck in a local optimum initially.

In practice, we first run SA without CNVs. That way, we can identify the cells that are attached to the root as non-neoplastic cells, and use those cells to estimate the weight of each region ρ_k, which is the probability for a read to fall into region k for a diploid cell without any CNVs. In addition, in the inferred tree without CNVs, we look for regions which have a lower or higher average normalized sequencing depth in some nodes compared to the root, and we select those regions as potential regions which might harbour copy number variants. Then, we run the SA with CNVs, but we restrict the addition of CNV events to the selected regions. We also exclude regions which have a very low amplification rate from the CNV inference, as their sequencing depth is very unreliable. This selection might lead to false negative CNVs, but reduces the number of false positives and decreases the number of iterations required in the SA, since it reduces the set of possible events that can be proposed.

Estimation of the node probabilities and dropout rates

The model contains two parameters which need to be estimated: the weight π_n of each node n and the dropout rate μ_i of each variant i. Ideally, we would like to marginalize over these parameters. However, the space is too large to integrate over, and sampling these parameters with the MCMC would be very inefficient: when a new tree is proposed, the old parameters might not work well for this new tree, which would lead to the tree being refused with a very high probability. Alternatively, we could jointly propose a new tree and new node weights and dropout rates, but the probability to obtain good parameters would be extremely low.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thus, instead of marginalizing over the node probabilities and dropout rates, we use the parameters which maximize the posterior probability. This can be efficiently performed with an EM algorithm, which has to be performed inside each MCMC step. We have two types of latent variables: the attachments of cells to nodes, σ_j, and for each cell j and each locus i, the number of reference and alternative alleles that did not get dropped out, $C_{ij}^{(r)}$ and $C_{ij}^{(a)}$. We use a beta prior centered on 0.05 for the dropout rates and a flat Dirichlet prior $D(1,\ldots,1)$ for the node weights.

During the E-step, we compute the probabilities Q of the latent variables, given the current parameters.

$$Q(\sigma_j = n) = P(\sigma_j = n \mid \boldsymbol{D_j}, \boldsymbol{A_j}, \boldsymbol{\pi}, \boldsymbol{\mu}) = \frac{\pi_n P(\boldsymbol{D_j}, \boldsymbol{A_j} \mid \sigma_j = n, \boldsymbol{\pi}, \boldsymbol{\mu})}{\sum_{n'} \pi_{n'} P(\boldsymbol{D_j}, \boldsymbol{A_j} \mid \sigma_j = n', \boldsymbol{\pi}, \boldsymbol{\mu})} \tag{7}$$

$$\begin{aligned} Q(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid \sigma_j = n) &= P(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid D_{ij}, A_{ij}, \sigma_j = n, \boldsymbol{\mu}) \\ &= \frac{\mu_i^{c_i^{(r)}(n)+c_i^{(a)}(n)-k-l}(1-\mu_i)^{k+l} P(A_{ij} \mid D_{ij}, k, l)}{\sum_{k',l'} \mu_i^{c_i^{(r)}(n)+c_i^{(a)}(n)-k'-l'}(1-\mu_i)^{k'+l'} P(A_{ij} \mid D_{ij}, k, l)} \end{aligned}$$

During the M-step, we update the parameters (node probabilities π_n and dropout rates μ_i) in order to maximize the sum of the log-prior and of the expected hidden log-likelihood.

$$\pi_n = \frac{1}{n_{\text{cells}}} \sum_{\text{cell } j} Q(\sigma_j = n) \tag{8}$$

$$\mu_i = \frac{\alpha - 1 + \sum_{\text{node } n} \sum_{\text{cell } j} Q\left(\sigma_j = n\right) \sum_{k,l} Q\left(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid \sigma_j = n\right) \left(c_i^{(r)}(n) + c_i^{(a)}(n) - k - l\right)}{\alpha + \beta - 2 + \sum_{\text{node } n} \sum_{\text{cell } j} Q\left(\sigma_j = n\right) \left(c_i^{(r)}(n) + c_i^{(a)}(n)\right)} \tag{9}$$

Data availability

The single-cell DNA sequencing data of Morita et al. [17] is available on the SRA under the project ID PRJNA648656.

Code availability

COMPASS has been implemented in C++ is freely available under a GPL3 license at `https://github.com/cbg-ethz/COMPASS`.

References

[1] Nicholas Mcgranahan and Charles Swanton. "Clonal heterogeneity and tumor evolution: past, present, and the future". *Cell* 168 (2017). DOI: 10.1016/j.cell.2017.01.018.

[2] Roger Lasken. "Genomic DNA amplification by the multiple displacement amplification (MDA) method". *Biochemical Society Transactions* 37 (2009). DOI: 10.1042/BST0370450.

[3] Lei Huang, Fei Ma, Alec Chapman, Sijia Lu, and Xiaoliang Xie. "Single-cell whole-genome amplification and sequencing: methodology and applications". *Annual Review of Genomics and Human Genetics* 16 (2015). DOI: 10.1146/annurev-genom-090413-025352.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[4] Timour Baslan, Jude Kendall, Linda Rodgers, Hilary Cox, Mike Riggs, Asya Stepansky, Jennifer Troge, Kandasamy Ravi, Diane Esposito, Bv Lakshmi, et al. "Genome-wide copy number analysis of single cells". *Nature protocols* 7 (2012). DOI: 10.1038/nprot.2012.039.

[5] Maurizio Pellegrino, Adam Sciambi, Sebastian Treusch, Robert Durruthy-Durruthy, Kaustubh Gokhale, Jose Jacob, Tina X Chen, Jennifer A Geis, William Oldham, Jairo Matthews, et al. "High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics". *Genome research* 28 (2018). DOI: 10.1101/gr.232272.117.

[6] Chrysanthi Ainali, Manimozhi Manivannan, Sombeet Sahu, Adam Sciambi, and Anup Parikh. "Subclonal identification of driver mutations and copy number variations from single-cell DNA sequencing of tumors". *Journal of Biomolecular Techniques* 31 (2020).

[7] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. "Tree inference for single-cell data". *Genome Biology* 17 (2016). DOI: 10.1186/s13059-016-0936-x.

[8] Jack Kuipers, Katharina Jahn, Benjamin Raphael, and Niko Beerenwinkel. "Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors". *Genome Research* 27 (2017). DOI: 10.1101/gr.220707.117.

[9] Jochen Singer, Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. "Single-cell mutation identification via phylogenetic inference". *Nature Communications* 9 (2018). DOI: 10.1038/s41467-018-07627-7.

[10] Edith Ross and Florian Markowetz. "OncoNEM: inferring tumor evolution from single-cell sequencing data". *Genome Biology* 17 (2016). DOI: 10.1186/s13059-016-0929-9.

[11] Hamim Zafar, Nicholas Navin, Ken Chen, and Luay Nakhleh. "SiCloneFit: bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data". *Genome Research* 29 (2019). DOI: 10.1101/gr.243121.118.

[12] Mohammed El-Kebir. "SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error". *Bioinformatics* 34 (2018). DOI: 10.1093/bioinformatics/bty589.

[13] Jack Kuipers, Mustafa Tuncel, Pedro Ferreira, Katharina Jahn, and Niko Beerenwinkel. "Single-cell copy number calling and event history reconstruction" (2020). DOI: 10.1101/2020.04.28.065755.

[14] Simone Zaccaria and Ben Raphael. "Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL". *Nature Biotechnology* 39 (2021). DOI: 10.1038/s41587-020-0661-6.

[15] Gryte Satas, Simone Zaccaria, Geoffrey Mon, and Ben Raphael. "SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses". *Cell Systems* 10 (2020). DOI: 10.1016/j.cels.2020.04.001.

[16] Ziwei Chen, Fuzhou Gong, Lin Wan, and Liang Ma. "BiTSC2 : Bayesian inference of Tumor clonal Tree by joint analysis of Single-Cell SNV and CNA data" (2020). DOI: 10.1101/2020.11.30.380949.

[17] Kiyomi Morita, Feng Wang, Katharina Jahn, Tianyuan Hu, Tomoyuki Tanaka, Yuya Sasaki, Jack Kuipers, Sanam Loghavi, Sa Wang, Yuanqing Yan, Ken Furudate, Jairo Matthews, Latasha Little, Curtis Gumbs, Jianhua Zhang, Xingzhi Song, Erika Thompson, Keyur Patel, Carlos Bueso-Ramos, and Koichi Takahashi. "Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics". *Nature Communications* 11 (2020). DOI: 10.1038/s41467-020-19119-8.

[18] Simone Ciccolella, Giulia Bernardini, Luca Denti, Paola Bonizzoni, Marco Previtali, and Gianluca Della Vedova. "Triplet-based similarity score for fully multilabeled trees with poly-occurring labels". *Bioinformatics* 37 (2020). DOI: 10.1093/bioinformatics/btaa676.

[19] Rui Zhang, Young-Mi Kim, Xianfu Wang, Yan Li, Xianglan Lu, Andrea R. Sternenberger, Shibo Li, and Ji-Yun Lee. "Genomic Copy Number Variations in the Myelodysplastic Syndrome and Acute Myeloid Leukemia Patients with del(5q) and/or -7/del(7q)". *International Journal of Medical Sciences* 12 (2015). DOI: 10.7150/ijms.12612.

[20] David Bowen, Michael Groves, AK Burnett, Y Patel, C Allen, C Green, Rosemary Gale, R Hills, and David Linch. "TP53 gene mutation is frequent in patients with acute myeloid leukemia and complex karyotype, and is associated with very poor prognosis". *Leukemia* 23 (2008). DOI: 10.1038/leu.2008.173.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] D Stirewalt, Era Pogosova-Agadjanyan, K Tsuchiya, J Joaquin, and Soheil Meshinchi. "Copy-neutral loss of heterozygosity is prevalent and a late event in the pathogenesis of FLT3/ITD AML". *Blood Cancer Journal* 4 (2014). DOI: 10.1038/bcj.2014.27.

[22] Eric Severson, Ethan Sokol, Russell Madison, Daniel Duncan, Amanda Hemmerich, Claire Edgerly, Richard Huang, Nicholas Britt, Jo-Anne Vergilio, Julia Elvin, Prasanth Reddy, Pratheesh Sathyan, Brian Alexander, Jeffrey Ross, Siraj Ali, and Shakti Ramkissoon. "Loss of heterozygosity of FLT3-ITD is common in acute myeloid leukemia and may be a more consistent prognostic marker than FLT3-ITD allele frequency". *Blood* 134 (2019). DOI: 10.1182/blood-2019-131248.

[23] Anna Stengel, Wolfgang Kern, Manja Meggendorfer, N Nadarajah, Karolina Perglerova, Torsten Haferlach, and C Haferlach. "Number of RUNX1 mutations, wild-type allele loss and additional mutations impact on prognosis in adult RUNX1 mutated AML". *Leukemia* 32 (2017). DOI: 10.1038/leu.2017.239.

[24] Thiago Noronha, Miguel Mitne, and Maria Chauffaille. "JAK2-mutated acute myeloid leukemia: comparison of next-generation sequencing (NGS) and single nucleotide polymorphism array (SNPa) findings between two cases". *Autopsy and Case Reports* 9 (2019). DOI: 10.4322/acr.2018.084.

[25] Eric Talevich, Alan Shain, Thomas Botton, and Boris Bastian. "CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing". *PLoS Computational Biology* 12 (2016). DOI: 10.1371/journal.pcbi.1004873.

[26] Peter Loo, Silje Nord, Ole Lingjærde, Hege Russnes, Inga Rye, Wei Sun, Victor Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles Perou, Anne-Lise Børresen-Dale, and Vessela Kristensen. "Allele-specific copy number analysis of tumors". *Proceedings of the National Academy of Sciences of the United States of America* 107 (2010). DOI: 10.1073/pnas.1009843107.

[27] Elsa Bernard, Yasuhito Nannya, Robert Hasserjian, Sean Devlin, Heinz Tuechler, Juan Medina-Martínez, Tetsuichi Yoshizato, Yusuke Shiozawa, Ryunosuke Saiki, Luca Malcovati, Max Levine, Juan Arango, Yangyu Zhou, Francesc Solé, Catherine Cargo, Detlef Haase, Maria Creignou, Ulrich Germing, Yanming Zhang, and Elli Papaemmanuil. "Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes". *Nature Medicine* 26 (2020). DOI: 10.1038/s41591-020-1008-z.

[28] Ashley Ward, Benjamin Braun, and Kevin Shannon. "Targeting oncogenic Ras signaling in hematologic malignancies". *Blood* 120 (2012). DOI: 10.1182/blood-2012-05-378596.

Author information

Affiliations

Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

Etienne Sollier, Katharina Jahn, Jack Kuipers & Niko Beerenwinkel

Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Etienne Sollier

SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Katharina Jahn, Jack Kuipers & Niko Beerenwinkel

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.06.475205; this version posted April 22, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Koichi Takahashi

Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Koichi Takahashi

Contributions

KJ and JK conceptualized the study. ES, KJ and JK designed the methodology. ES implemented the methodology. ES and KT performed the biological analysis. KJ, JK and NB provided supervision. ES wrote the original draft. All authors reviewed and approved the paper.

Corresponding author

Correspondence to Katharina Jahn.

Ethics declaration

The authors declare no competing interests.