Genome Medicine

RESEARCH Open Access

Sensitivity to sequencing depth in single-cell cancer genomics

João M. Alves[1,2,3*] and David Posada[1,2,3*]

Abstract

Background: Querying cancer genomes at single-cell resolution is expected to provide a powerful framework to understand in detail the dynamics of cancer evolution. However, given the high costs currently associated with single-cell sequencing, together with the inevitable technical noise arising from single-cell genome amplification, cost-effective strategies that maximize the quality of single-cell data are critically needed. Taking advantage of previously published single-cell whole-genome and whole-exome cancer datasets, we studied the impact of sequencing depth and sampling effort towards single-cell variant detection.

Methods: Five single-cell whole-genome and whole-exome cancer datasets were independently downscaled to 25, 10, 5, and 1× sequencing depth. For each depth level, ten technical replicates were generated, resulting in a total of 6280 single-cell BAM files. The sensitivity of variant detection, including structural and driver mutations, genotyping, clonal inference, and phylogenetic reconstruction to sequencing depth was evaluated using recent tools specifically designed for single-cell data.

Results: Altogether, our results suggest that for relatively large sample sizes (25 or more cells) sequencing single tumor cells at depths > 5× does not drastically improve somatic variant discovery, characterization of clonal genotypes, or estimation of single-cell phylogenies.

Conclusions: We suggest that sequencing multiple individual tumor cells at a modest depth represents an effective alternative to explore the mutational landscape and clonal evolutionary patterns of cancer genomes.

Keywords: Single-cell sequencing, Intratumor genetic heterogeneity, Variant calling, Clonal inference, Tumor phylogenies

Background

Recent advances in next-generation sequencing (NGS) technologies revealed that the large majority of cancer genomes are heterogeneous despite their monoclonal origin, with the continuous expansion of the tumor mass contributing to the accumulation of somatic mutations within malignant cells, hence promoting the proliferation of distinct genetic lineages (i.e., clones) through time [1]. While quantifying this intratumor heterogeneity (ITH) remains a difficult task, as standard methods in cancer genomics generally rely on population-level analysis from bulk experiments, single-cell sequencing (SC-Seq) approaches are now widely viewed as a promising alternative to explore tumor evolution [2]. Indeed, a collection of recent studies have successfully applied SC-Seq to determine the mutational load in individual tumors [3], estimate the frequency of subclones [4], infer evolutionary relationships [5], or explore the role of ITH in metastatic dissemination [6].

Nevertheless, several technical challenges surrounding current SC-Seq methodologies greatly limit our ability to obtain reliable genomic information from single cells. For instance, the multiple rounds of whole genome amplification (WGA) usually required prior to SC-Seq are known to introduce a high number of sequence artifacts that can be confounded with genuine biological variation (see [7] for a detailed review). Other technical errors, such as insufficient physical coverage, uneven genome amplification, and allelic dropout, may also generate substantial artificial variability in cancer genomes, compromising the ability to detect real somatic heterogeneity

* Correspondence: jalves@uvigo.es; dposada@uvigo.es
[1]Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain
Full list of author information is available at the end of the article

from SC-Seq data [8]. As a consequence, alternative strategies are needed in order to eliminate the noise generated during WGA while effectively allowing the quantification of ITH from single cells.

Zhang et al. [9] started addressing some of these issues and demonstrated the efficiency of a census-based strategy for accurate variant detection in single cells. By using multiple cells and trusting only variants detected in at least two single-cell libraries, they detected up to 80% of germline SNPs in the human chromosome 5 with 59 cells sequenced at 0.3× or 22 cells at 1×. Their results suggest that for detecting clonal and subclonal variants in single cells, and given a fixed sequencing effort, it is best to sequence multiple cells (in their case a minimum of 20) at a modest depth (~ 1×).

Here, we further explore the sensitivity of SC-seq to sequencing depth using five publicly available single-cell whole-genome (scWGS) and whole-exome (scWXS) cancer datasets. We expand not only on the scale of the datasets, but also on the scope of the inferences, including copy-number variant detection, clonal inference, and phylogenetic estimation. Altogether, our results suggest that even though sequencing depth does indeed contribute to a better refinement of somatic variant characterization from tumor single cells, sample size plays a more determinant role for a reliable assessment of the general patterns of somatic variation in cancer genomes. For relatively large sample sizes (e.g., ≧ 25 samples), sequencing single cells at modest depths (i.e., 5×) enables a similar description of somatic variation, clonal composition, and evolutionary history compared to sequencing depths one order of magnitude higher.

Methods

Five publicly available sequencing datasets from four single-cell studies were retrieved from the Sequence Read Archive (SRA) in FASTQ format, including four single-cell genomes from a breast cancer patient [5] (we will call this dataset "W4" to indicate the authors and the number of cells), eight single-cell exomes from circulating tumor cells from one lung adenocarcinoma patient [10] ("N8" dataset), 25 single-cell exomes derived from a kidney tumor patient [11] ("X25" dataset), 55 single-cell exomes from a breast cancer patient [5] ("W55" dataset), and 65 single-cell exomes from a single JAK-2 negative neoplasm myeloproliferative patient [12] ("H65" dataset). Normal and tumor bulk WGS/WXS data from the same patients were also retrieved. Normal single cells were only available for the three largest datasets. A list of the individual samples and corresponding accession codes is available in Additional file 1: Table S1.

All the analyses enumerated below are described in detail in the accompanying Additional file 1: Note, including command lines. Both single-cell and bulk reads were aligned to human reference GRCh37 using the *MEM* algorithm in the BWA software [13]. Following a standardized best-practices pipeline [14], mapped reads from all datasets were independently processed by filtering reads displaying low mapping-quality, performing local realignment around indels, and removing PCR duplicates. Raw single-nucleotide variant (SNV) calls for the bulk datasets were obtained using the paired-sample variant-calling approach implemented in the VarDict software [15]. For the N8 dataset, since samples from both primary tumor and metastasis were available, VarDict was run twice, independently for both samples, and the resulting SNVs subsequently merged using the *CombineVariants* tool from the Genome Analysis Toolkit (GATK) [16]. Low-quality SNV calls were removed using the *SelectVariants* tool from GATK. The remaining SNVs were further subdivided into two distinct categories: "germline" SNVs if present in both tumor and normal bulk samples, and "somatic" SNVs if found solely in the tumor bulk samples. Small indels and other complex structural rearrangements were ignored in order to generate a final list of "gold-standard" bulk SNVs. All analyses presented here were based on this set of variants.

The single-cell BAM files were independently downscaled to 25, 10, 5, and 1× sequencing depth using Picard [17]. For each depth level, ten technical replicates were generated for statistical validation, resulting in a total of 6280 BAM files. Single-cell SNV calls were obtained from the original and down-sampled single-cell BAM files using Monovar [18], a variant caller specifically designed for single-cell data, under default settings. Single-cell variant-calling performance was evaluated by estimating the proportion of "gold-standard" germline and somatic bulk SNVs identified in the down-sampled single-cell datasets (germline and somatic recall, respectively). To further characterize the effect of sequencing depth on single-cell variant calling, we determined the fraction of somatic SNVs found in the down-sampled single-cell replicates that were also identified in the original single-cell datasets ("somatic precision"). In addition, we repeated the recall analysis focusing only on the somatic SNVs already described in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [19] and on the non-synonymous SNVs previously detected (Additional file 1: Table S2).

Single-cell copy-number variants (CNVs) were identified with Ginkgo [20] using variable-length bins of around 500 kb. After binning, data for each cell was normalized and segmented using default parameters. Sensitivity was evaluated by assessing the recall of the CNVs and segment breakpoints at the different sequencing depths.

Clonal genotypes were estimated from the somatic SNVs using the Single-Cell Genotyper (SCG) [21] (Additional file 1: Note), and their recall across sequencing depth was measured with the adjusted Rand Index [22], a version of the Rand Index corrected for chance [23]. The Rand-Index is a popular statistical measure of the similarity between two data clusterings (corresponding here to groups of mutations, or clones). In addition, clonal trees were also inferred from the somatic SNVs with OncoNEM [24]. Using a similar approach to Ross and Markowetz [24], the pairwise cell shortest-path distance was used to measure the consistency in tree reconstruction across the different sequencing depths. Furthermore, maximum-likelihood single-cell phylogenies were estimated from the SNVs using SiFit [25]. In this case, phylogenetic recall across sequencing depth was measured using the standard Robinson-Foulds tree distance [26]. In addition, we also calculated the homoplasy index (HI), a measure of the amount of homoplasy on a tree, using the phangorn R package [27]. The HI is one minus the ratio between the minimum number of changes required and the actual number observed [28].

Statistical significance for the differences in recall or HI for the experiments described above were assessed using Tukey's HSD test with a family-wise error rate of 0.05 in R. See the Additional file 1: Note for a detailed description.

Results

Genome coverage

Genome coverage (percentage of the reference genome covered by ≥ 1 read) for the single-cell down-sampled datasets decreased non-linearly with lower sequencing depths, in particular when moving from 5× to 1× (Fig. 1).

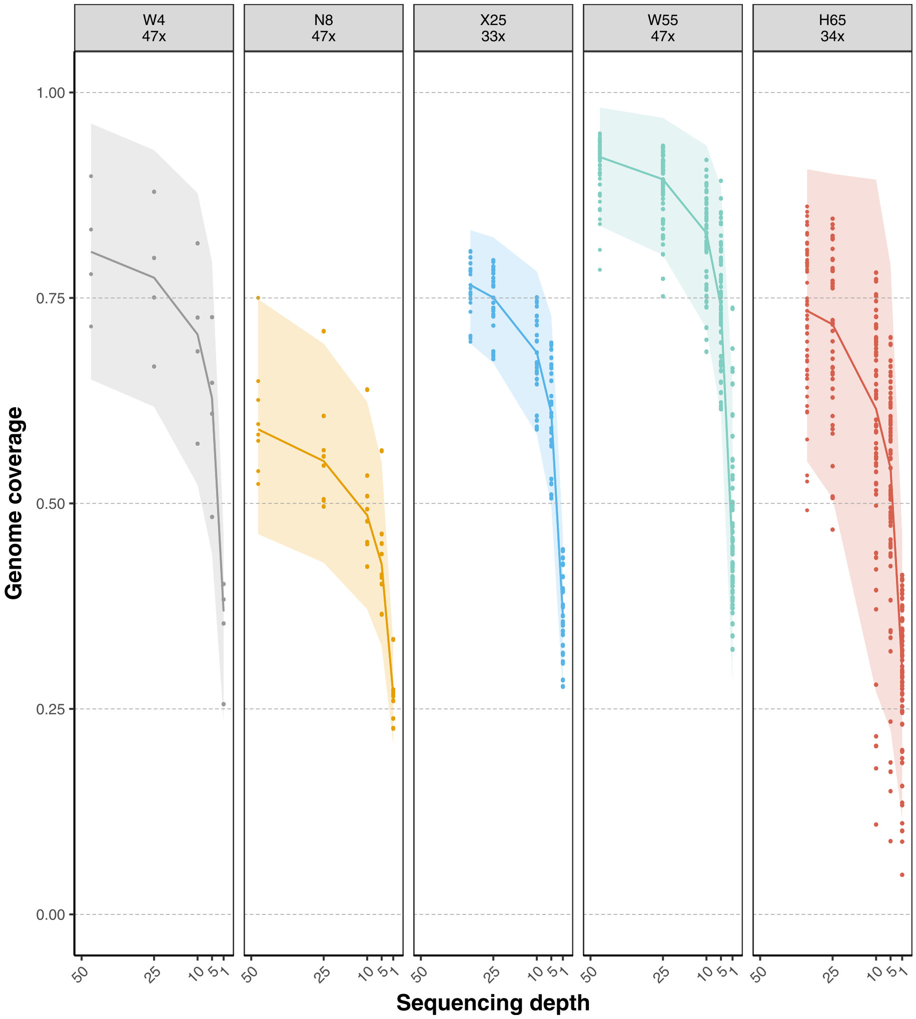

Fig. 1 Genome coverage and sequencing depth in the down-sampled single-cell datasets. Each panel depicts a single-cell dataset (e.g., W4) where the number in the header indicates its original sequencing depth (e.g., 47×). *Solid lines* represent the average genome coverage (proportion of bases covered by at least one read, measured per nucleotide) obtained for the different replicates at the different down-sampled depths. *Dots* correspond to single cells. *Shaded areas* indicate the standard deviation from the mean

Single-nucleotide variants

SNV detection

The observed decline in genome coverage was logically reflected in the proportion of bulk germline and somatic SNVs found in the single-cell down-sampled datasets ("germline and somatic recall"), which decreased significantly (Tukey HSD p value < 0.05) at lower sequencing depths (Fig. 2). The germline recall decrease was much less pronounced when the number of cells was large (≥25). Thus, for the X25, W55, and H65 datasets the germline recall was at 1× as high as 70–80%, and at 5× close to 100% (Fig. 2a). On the other hand, when only four or eight cells were available (W4, N8 datasets), the germline recall at 1× decreased dramatically to 5–13%. The somatic recall rate was, as expected, much more modest than for the germline variants (Fig. 2b). The effect of sequencing depth was significant in practically every case. Notably, the fraction of SNVs found in the down-sampled replicates that were also identified in the original single-cell datasets ("somatic precision"; Fig. 2c) was much less affected by sequencing depth, with many non-significant variations between "contiguous" levels of coverage (i.e., 1–5, 5–10, 10–25×).

Interestingly, a significant amount of somatic variants was detected exclusively in the single-cells (i.e., absent in the bulk), particularly at higher sequencing depths (Additional file 1: Figure S1A). However, the overall variant quality scores for these calls were much lower than for those shared with the bulk dataset (Additional file 1: Figure S1B), suggesting that most might be untrustworthy.

COSMIC and non-synonymous SNV detection

Moreover, the somatic recall specific for COSMIC somatic variants (Fig. 3a, b) decreased very rapidly and significantly (p value < 0.05) with lower sequencing depths for the smallest datasets (W4 and N8), but not as abruptly for the larger ones, in particular for the X25 and W55 datasets. For example, for the latter the recall was already around 70% at only 5×. A statistically significant trend was also observed for non-synonymous SNVs, which were very difficult to detect at 5× or 1× only for the smaller datasets (Fig. 3c, d). For larger sample sizes, the non-synonymous SNV recall rate was already above 70% at 1×.

SNV genotyping

The recall for single-cell SNV genotyping also dropped significantly at lower sequencing depth for all datasets (Fig. 4). Nevertheless, at 5×, 60–90% of the genotypes identified in the original single-cell datasets were already recovered without error. Importantly, discordant SNV genotype calls were relatively infrequent, and differences between depth levels and datasets were usually due to different amounts of missing calls.

Copy-number variants

Single-cell copy-number profiles were remarkably consistent across sequencing depth (Fig. 5). Breakpoint detection was slightly better for higher sequencing depths, but always quite accurate. For example, more than 70% of the CNV breakpoints inferred from the original dataset were already detected at 1× in all datasets. Moreover, CNV genotype calls were not affected by sequencing depth. Indeed, at 1× the CNV genotype recall was already > 99% for all datasets.

Clonal genotypes

Clonal inference recall by SCG [21], as measured by the adjusted Rand Index, was not affected by sequencing depth

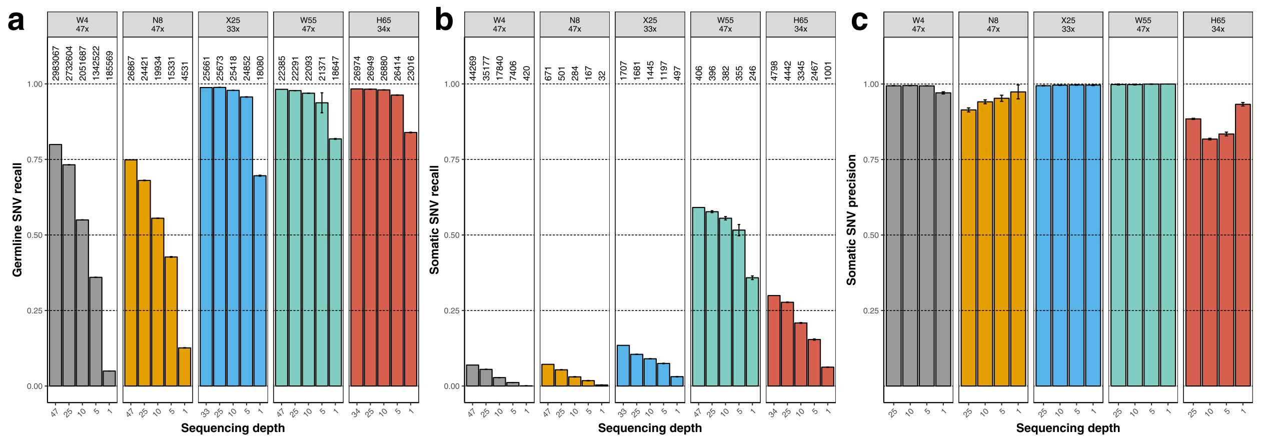

Fig. 2 SNV recall and precision in single-cells. **a** Barplots illustrating the proportion of bulk germline variants called in the single-cell down-sampled datasets (germline recall). **b** Proportion of bulk somatic variants identified in the single-cell down-sampled datasets (somatic recall). **c** Proportion of somatic variants called in the down-sampled datasets that were also identified in the original single-cell dataset (somatic precision). *Error bars* indicate 95% confidence intervals. *Numbers above bars* indicate number of calls (for **b** and **c** these numbers are the same)

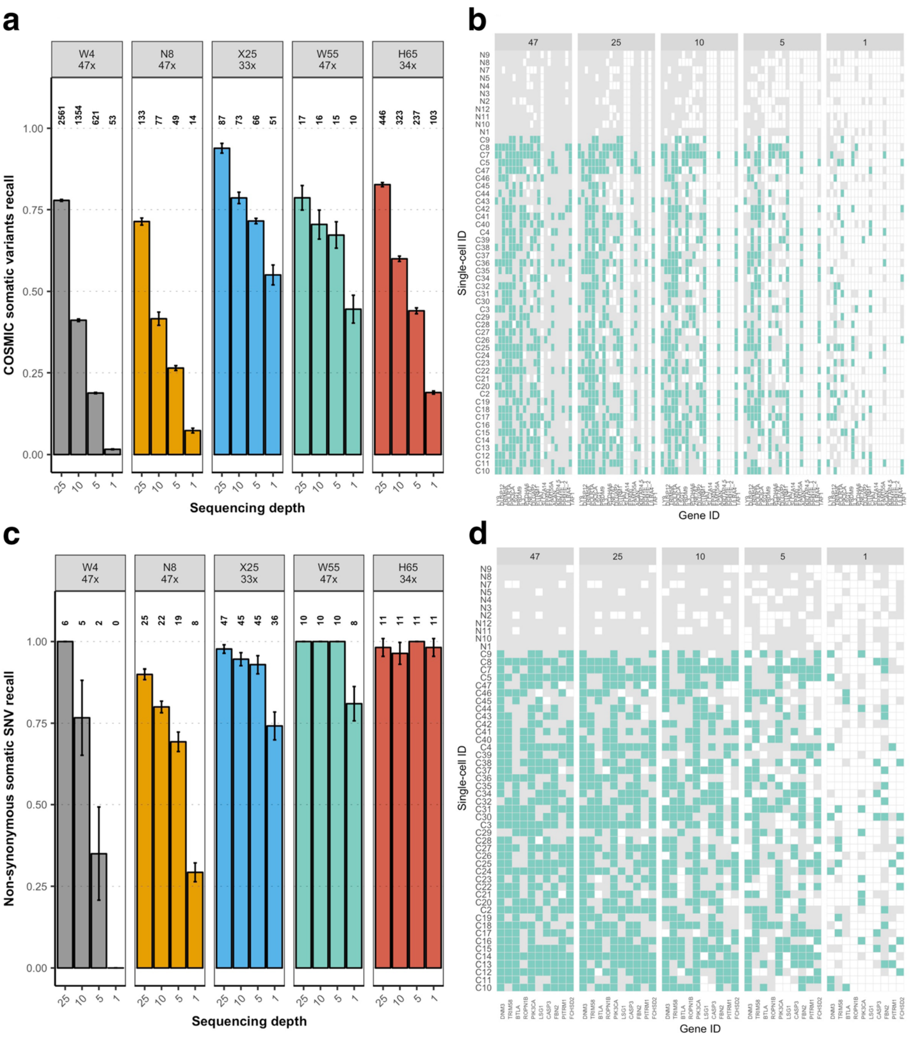

Fig. 3 COSMIC and non-synonymous somatic SNV recall in single cells. **a** Barplots indicate the proportion of bulk COSMIC somatic variants detected in the single-cell datasets (COSMIC recall). *Error bars* indicate 95% confidence intervals. *Numbers above bars* indicate number of variants called. **b** Presence–absence profile of COSMIC SNVs across cells for replicate 1 of the W55 dataset. Colors illustrate mutation status: mutated allele, *green*; reference allele, *grey*; missing data, *white*. **c** Barplots indicate the proportion of bulk non-synonymous somatic variants detected in the single-cell datasets (non-synonymous recall). **d** Presence–absence profile of non-synonymous SNVs across cells for replicate 1 of the W55 dataset. Colors illustrate mutation status: mutated allele, *green*; reference allele, *grey*; missing data, *white*

in the smallest and largest datasets (W4, N8, and H65) where the number of inferred clones was always one (data not shown), but decreased to a different extent at lower depths in the X25 and W55 datasets (Fig. 6a, b). Indeed, despite the improvements observed at sequencing depths beyond 5× for the X25 dataset, the distinct clonal clusters of the W55 dataset were only distinguishable at 25×.

Clonal trees

In contrast to the smallest datasets, the recall of the clonal trees inferred by OncoNEM [24] (Fig. 7a, b) was maintained or decreased slightly—not significantly in multiple occasions—at lower sequencing depths for the larger datasets (X25, N55, and H65) where the number of potential phylogenetic solutions is much bigger.

Single-cell phylogenies

SiFit [25] single-cell phylogenies were also very stable at sequencing depths equal to or larger than 5× (Fig. 8a, b). In most instances the differences due to depth were not statistically significant. At 1×, in some cases the inferred phylogeny displayed healthy cells intermixed with tumor cells, likely due to poor resolution. Nevertheless, this effect disappeared at 5× and beyond, when all tumor cells always clustered together in a single clade, as expected. Despite the observed stability in tree topology, variants present in all cells in the original single-cell datasets

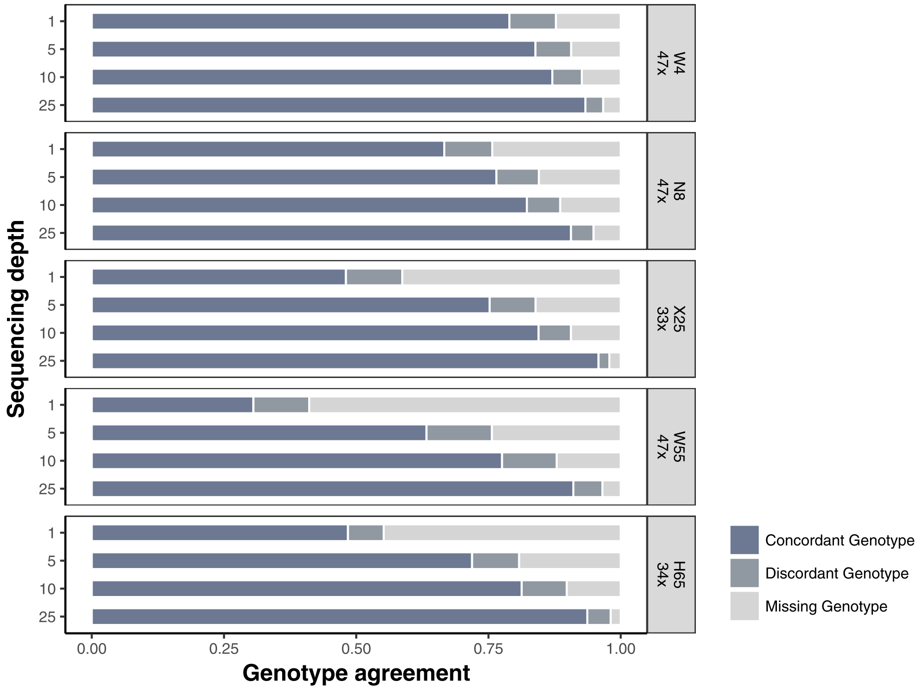

Fig. 4 SNV genotype recall in single cells. *Horizontal bars* represent the proportion of concordant (*dark blue*), discordant (*dark gray*), and missing (*light gray*) SNV genotype calls (homozygous for the reference allele, heterozygous or homozygous for the alternative allele) for the down-sampled datasets

increasingly became subclonal at lower depths (Additional file 1: Figure S2). The amount of homoplasy was, however, generally constant across sequencing depths with the exception of 1×, where for the larger datasets (X25, W55, H65) there was a significant decrease of the HI scores (Additional file 1: Figure S3).

Discussion

In this study we aimed to characterize the impact of sequencing depth in single-cell cancer genomics studies. Undeniably, here we have used five datasets with specific characteristics like number of mutations, number of clones, tissue of origin, genomic target, sequencing depth, or amplification bias. In consequence, although some general patterns seem to be more or less clear, care must be taken in generalizing our findings as particular trends may vary for other cancer datasets.

With this caveat in mind, our downsampling experiments suggest that, overall, larger sequencing depths for small numbers of cells (eight or less) might lead to relevant improvements. In contrast, for relatively large datasets (25 or more cells), our results indicate that sequencing single cells at moderate depths (i.e., 5×) should represent a reasonable approach to characterize the genomic diversity and evolution of tumors, including the identification of putative driver alterations. This is in line with the results of Zhang et al. [9], who showed that for variant detection it is better to have multiple cells sequenced at low depth, given a fixed sequencing effort.

Unsurprisingly, all recalls (SNVs, CNVs, clones, phylogenies) showed some kind of decrease at smaller sequencing depths. In many cases the drop was statistically significant despite being of small magnitude. Notably, for the larger datasets (and by large here we mean—only—dozens of cells), the impact of sequencing depth was much smaller, with the exception of the H65 dataset. This particular dataset, albeit being the largest, displays a very heterogeneous genome coverage for the single cells sampled which may have mislead some of the analyses. Indeed, genome coverage bias has been shown to contribute to a lower sensitivity to detect variants [9], hence potentially explaining some of the somewhat discordant results of the H65 dataset.

In any case, bulk germline SNVs were relatively easy to identify for the three largest datasets even at low sequencing depth. This was indeed expected since germline variants should be present in the vast majority, if not all, of tumor cells. Nevertheless, when the number of single cells was small, the effect of sequencing depth on germline SNV recall was much more pronounced and reached a limit of ~ 75% at the highest sequencing depth (i.e., 47×) reinforcing the idea that, due to the inherent bias in single-cell genome amplification, broader sampling effort should be favored over increased sequencing depth in variant detection analysis [9].

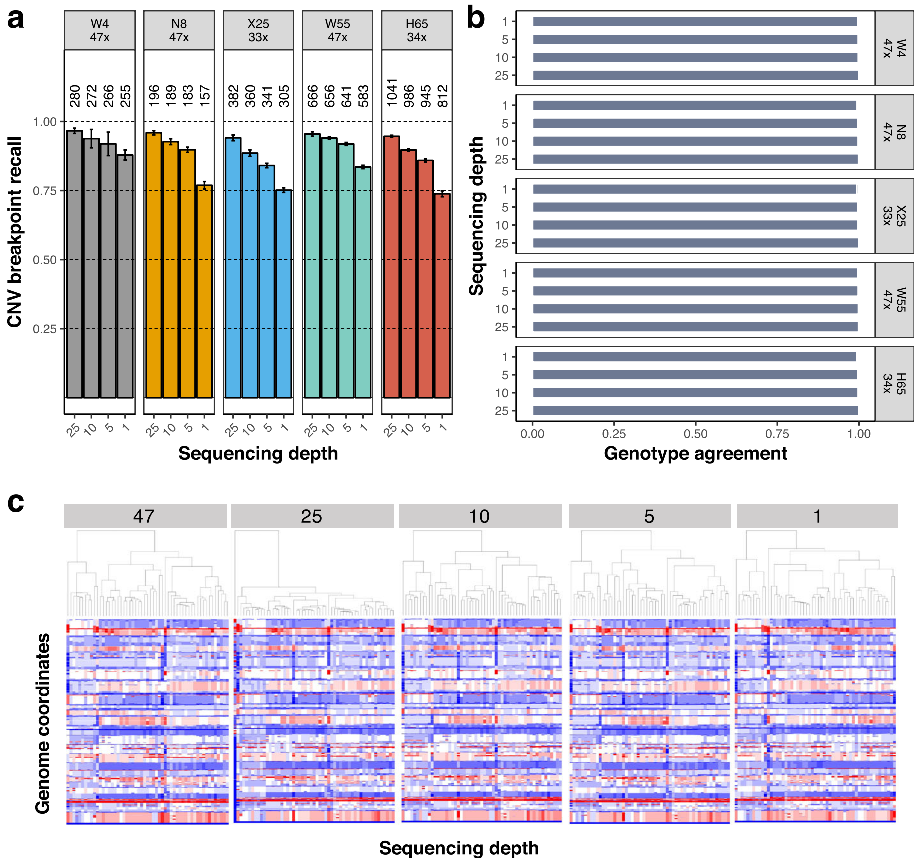

Fig. 5 CNV recall in single cells. **a** Barplots indicate the proportion of CNV breakpoints detected in the down-sampled datasets that were also called in the original single-cell dataset (recall). *Numbers above bars* indicate number of breakpoints detected. *Error bars* indicate 95% confidence intervals. **b** *Horizontal bars* represent the proportion of concordant (*dark blue*), discordant (*dark gray*), and missing (*light gray*) CNV genotype calls. **c** Copy-number profiles at different depths for replicate 1 of the W55 dataset. Distinct colors represent the CN configuration: CN gain, *red*; CN loss, *blue*

While somatic SNVs were much more difficult to detect, it should be highlighted that the number of somatic mutations detected at 5× were usually at the same order of magnitude as the number of mutations detected at higher sequencing depths, except for the smaller datasets. Still, for the smallest dataset analyzed (W4), the high number of somatic SNVs detected at 5× (7406) seem plenty enough to conduct many subsequent analyses, like clonal inference or phylogeny reconstruction.

In relation to this, it is important to highlight that, aside from sample size and sequencing depth, somatic variant detection can additionally be affected by the choice of thresholds during variant calling. Indeed, conservative thresholds may prevent the discovery of true mutations due to excessive filtering, whereas relaxed thresholds may cause an increase of false-positive calls. Determining the best parameters for filtering variants is, therefore, difficult. Most studies analyzing SC-Seq data have relied on "hard" filtering thresholds for a minimum depth of coverage (e.g., > 10 reads; e.g., [5]). Here, a similar filtering strategy would prove too stringent for most down-sampled datasets. To allow proper comparisons among the different depth levels we decided not to use a minimum depth threshold. Instead, we required each variant to be detected in at least two single cells. Such a consensus strategy has already been shown to be quite efficient [9, 18].

Remarkably, the somatic single-cell SNV precision was, in general, very robust to sequencing depth, suggesting that lower depths do not result in new calls that would not have been made at higher depths. Intuitively, this observation makes perfect sense since at lower sequencing depths the variants detected tend to be the clonal ones (i.e., variants shared by the majority of the single cells sampled) whereas the detection of low-frequency mutations required higher read depths (data not shown).

One might be worried, however, about missing putative driver mutations, but our results suggest that, as far as the number of single cells is reasonably large (here 25 or more), most COSMIC somatic variants can be detected at modest sequencing depths (here 5× or more).

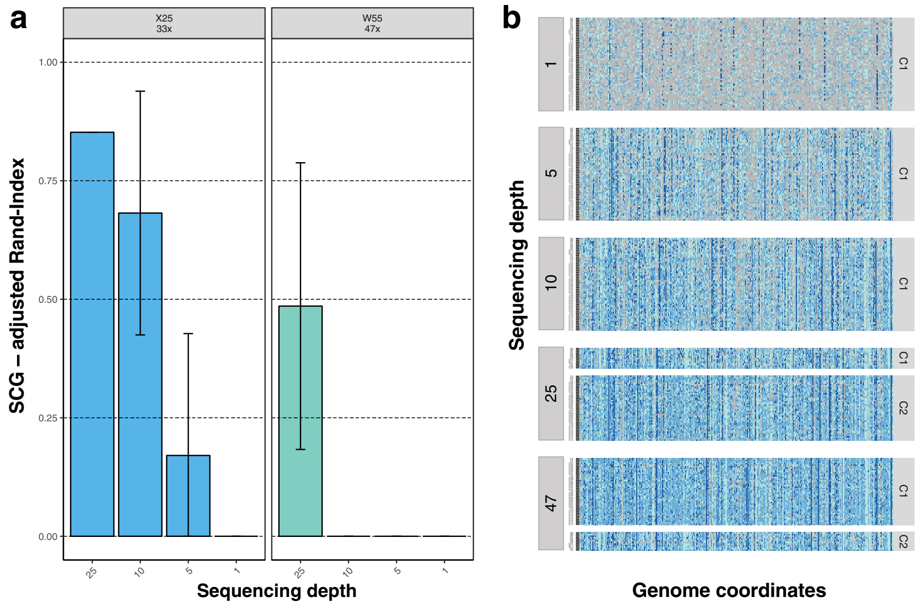

Fig. 6 Clonal inference recall from single cells. **a** Barplots show the adjusted Rand Index between the clonal clusters inferred from the original dataset and the clonal clusters from the down-sampling experiments. *Error bars* indicate 95% confidence intervals. **b** SCG clonal clusters at different sequencing depths for replicate 1 of the W55 dataset. Genotypes for each locus highlighted in different colors: Homozygous reference, *light blue*; heterozygous, *blue*; homozygous alternative, *dark blue*; missing, *light gray*)

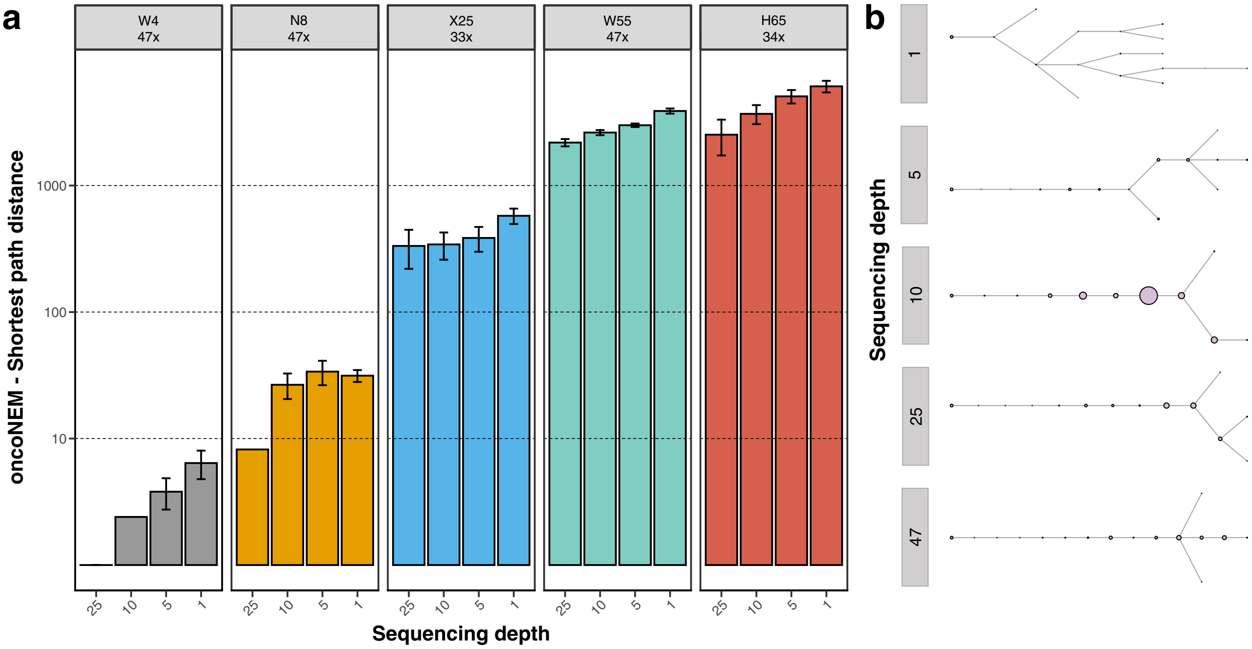

Fig. 7 Clonal tree recall from single cells. **a** Barplots show the pairwise cell shortest-path distance between the OncoNEM clonal tree inferred from the original dataset and the OncoNEM clonal trees inferred from the down-sampling experiments. *Error bars* indicate 95% confidence intervals. **b** OncoNEM clonal trees at different depths for replicate 1 of the W55 dataset

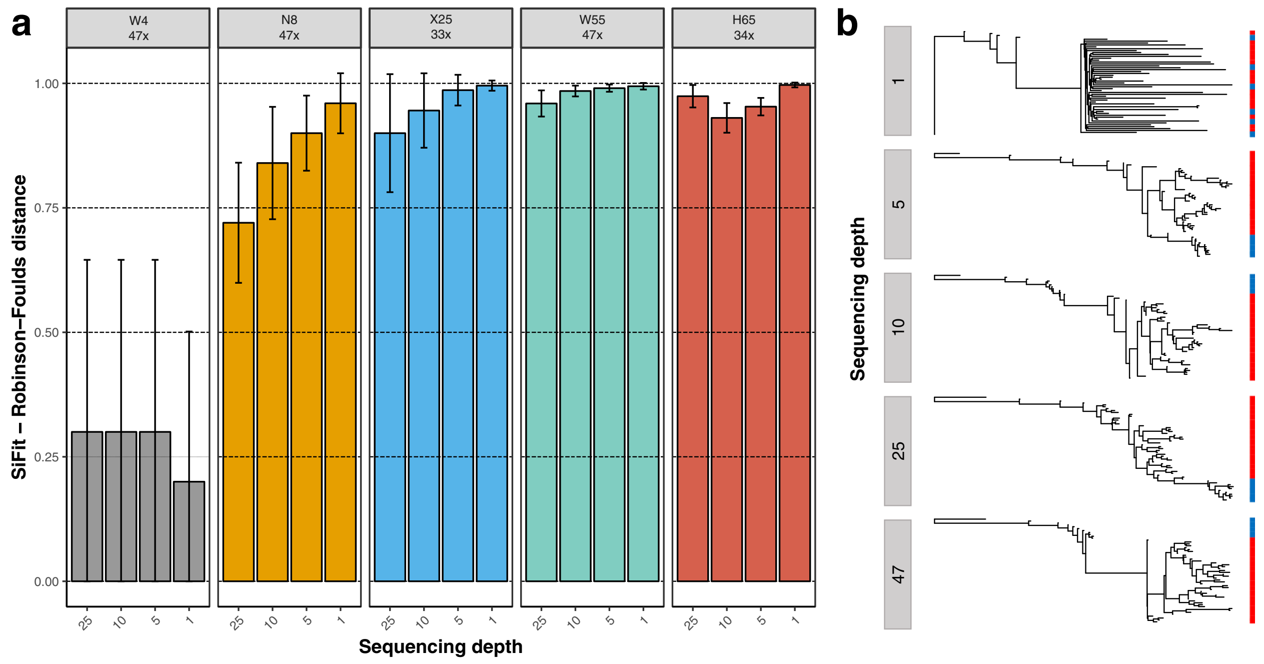

Fig. 8 Single-cell tree recall. **a** Barplots show the Robinson-Foulds distance between the SiFit single-cell tree inferred from the original dataset and the SiFit single-cell trees inferred from the down-sampling experiments. *Error bars* indicate 95% confidence intervals. **b** SiFit unrooted single-cell trees at different depths for replicate 1 of the W55 dataset, . Distinct colors represent cell type: healthy cell, *blue*; tumor cell, *red*

Similar results were also observed for the somatic non-synonymous variants, suggesting that, in principle, many relevant variants in single-cell genomes are likely to be detected at modest sequencing depths.

Obviously, assigning particular genotypes to the individual cells is a much more involved task than just detecting variants. Importantly, for SNV genotyping, reducing sequencing depth generally resulted in an increased amount of missing data in the single-cell genotype matrix, rather than different genotype calls.

Moreover, and in agreement with previous studies [20, 29], CNV characterization from single cells was also very robust to sequencing depth, with all down-sampled datasets showing remarkable preservation of CNV breakpoints. Furthermore, CNV genotype assignment was insensitive to the variation in the sequencing depths explored. In general, the copy-number analysis of single-cell libraries can be confounded by amplification bias. However, previous studies suggest that amplification biases are randomly distributed and sufficiently separated throughout the genome [30] as to not affect CNV calling at the level of resolution chosen here (500-kb bins). Popular single-cell amplification methods like multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) usually generate amplicons of around 10–100 kb and 1–5 kb, respectively; therefore, we do not expect many false positive CNV calls [31]. Yet, we acknowledge that our choice of bin size may have prevented the identification of small CNVs [20].

It is relatively well established that an accurate identification of clonal genotypes can be very important to understand tumor dynamics and genomic architecture [32–34]. For the datasets analyzed here, our results suggest that SC-Seq depth does not affect the identification of tumor clones when the genomic variability between malignant cells is small (i.e., displaying limited clonal population genetic diversification). However, the same was not true for tumors comprising a larger number of subclones, where the different clonal genotypes were only distinguishable at higher sequencing depths. While these results are not necessarily surprising, as clonal identification remains a complex problem even for bulk sequencing data [35, 36], they seem to suggest that higher sequencing coverage is ultimately required to resolve fine-scale clonal structure in more heterogeneous tumors.

Finally, in our evolutionary analyses, we observed a moderate impact of sequencing depth with respect to the estimated phylogenetic relationships of the inferred clones and single cells. Perhaps due to the uncertainty stemming from significant amounts of missing data, datasets down-sampled to 1× resulted in phylogenetic trees with healthy cells intermingled with tumor cells, which can be safely considered as artifacts. While the amount of homoplasy was lower at 1×, this was likely an effect of the smaller amount of variant calls per cell at such a low depth. Otherwise, tree topologies at 5× seemed quite similar to those inferred at higher depths, suggesting that relatively few clonal variants might be enough to resolve the topology of the single-cell trees. Note that the topology does not include branch lengths, whose accurate estimation might require higher sequencing depths.

Conclusions

Single-cell DNA sequencing is expected to be key to obtain accurate inferences of the clonal architecture of tumor samples, which shall ultimately prove crucial to compare models of cancer evolution, trace cell lineages, measure mutation rates, and decipher cell clones responsible for metastatic dissemination and drug resistance [2, 37, 38]. While recent experimental and analytical improvements have improved the quality of single-cell DNA sequencing data [9, 18, 20, 21, 25, 39–41], the costs associated with sequencing multiple single-cell genomes or exomes at high depths are still largely prohibitive. Our results support the idea that sequencing multiple individual tumor cells at a modest depth, such as 5×, may help circumvent this limitation at least for the type of analyses implemented here. Finally, the results obtained here might be extrapolatable to some extent to non-tumor single-cell genomes.

Additional file

Additional file 1: Supplementary note containing all information required to generate the results presented in this manuscript. **Tables S1** and **S2. Figures S1–S3.** (PDF 483 kb)

Abbreviations

CNV: Copy-number variant; COSMIC: Catalogue of somatic mutations in cancer; GATK: Genome analysis toolkit; HI: Homoplasy index; ITH: Intratumor genomic heterogeneity; MALBAC: Multiple annealing and looping-based amplification cycles; MDA: Multiple displacement amplification; NGS: Next-generation sequencing; SC-Seq: Single-cell sequencing; SNV: Single-nucleotide variant; WGA: Whole-genome amplification; WGS: Whole-genome sequencing; WXS: Whole-exome sequencing

Acknowledgements

We would like to thank Sereina Rutschmann, Harald Detering, Laura Tomás, and Sara Rocha for their comments on earlier versions of the manuscript. We also thank the anonymous reviewers for their useful suggestions.

Funding

This work was supported by the European Research Council (ERC-617457-PHYLOCANCER awarded to D.P.) and by the Ministry of Economy and Competitiveness—MINECO (BFU2015-63774-P awarded to D.P.) D.P. receives further support from the Galician government.

Availability of data and materials

A list of the public single-cell datasets analyzed is in Additional file 1: Table S1. The downsampled replicates are available from the authors on request.

Authors' contributions

DP conceived the project. DP designed and JMA performed the analyses. JMA and DP wrote the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain. [2]Biomedical Research Center (CINBIO), University of Vigo, Vigo, Spain. [3]Galicia Sur Health Research Institute, Vigo, Spain.

Received: 13 November 2017 Accepted: 29 March 2018
Published online: 16 April 2018

References

1. Gerlinger M, Swanton C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. Br J Cancer. 2010;103: 1139–43.
2. Navin NE. Cancer genomics: one cell at a time. Genome Biol. 2014;15:452.
3. Potter NE, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, Titley I, et al. Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res. 2013;23:2115–25.
4. Hughes AEO, Magrini V, Demeter R, Miller CA, Fulton R, Fulton LL, et al. Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. PLoS Genet. 2014;10:e1004462.
5. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.
6. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015;526:131–5.
7. Van Loo P, Voet T. Single cell analysis of cancer genomes. Curr Opin Genet Dev. 2014;24:82–91.
8. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell. 2015;58:598–609.
9. Zhang C-Z, Adalsteinsson VA, Francis J, Cornils H, Jung J, Maire C, et al. Calibrating genomic and allelic coverage bias in single-cell sequencing. Nat Commun. 2015;6:6822.
10. Ni X, Zhuo M, Su Z, Duan J, Gao Y, Wang Z, et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc Natl Acad Sci U S A. 2013;110:21083–8.
11. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148:886–95.
12. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148:873–85.
13. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;1303:3997v1.
14. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1–33.
15. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44:e108.
16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
17. Picard software. http://broadinstitute.github.io/picard.Accessed 12 Apr 2018.
18. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat Methods. 2016;13:505–7.
19. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45: D777–D83.

20. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, et al. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015;12:1058–60.
21. Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, et al. Clonal genotype and population structure inference from single-cell tumor sequencing. Nat Methods. 2016;13:573–6.
22. Hubert L, Arabie P. Comparing partitions. J Classification. 1985;2:193–218.
23. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846.
24. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. 2016;17:69.
25. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome Biol. 2017;18:178.
26. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53:131–47.
27. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2010;27:592–3.
28. Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. Syst Zool. 1969;18:1.
29. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, et al. Scalable whole-genome single-cell library preparation without preamplification. Nat Methods. 2017;14:167–73.
30. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.
31. Sherman MA, Barton AR, Lodato MA, Vitzthum C, Coulter ME, Walsh CA, et al. PaSD-qc: quality control for single cell whole-genome sequencing data using power spectral density estimation. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx1195.
32. Alves JM, Prieto T, Posada D. Multiregional tumor trees are not phylogenies. Trends Cancer Res. 2017;3:546–50.
33. Kuipers J, Jahn K, Beerenwinkel N. Advances in understanding tumour evolution through single-cell sequencing. Biochim Biophys Acta. 1867;2017:127–38.
34. Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F. Cancer evolution: mathematical models and computational inference. Syst Biol. 2015;64:e1–25.
35. Turajlic S, McGranahan N, Swanton C. Inferring mutational timing and reconstructing tumour evolutionary histories. Biochim Biophys Acta. 1855;2015:264–75.
36. Beerenwinkel N, Greenman CD, Lagergren J. Computational cancer biology: an evolutionary perspective. PLoS Comput Biol. 2016;12:e1004717.
37. Tsoucas D, Yuan G-C. Recent progress in single-cell cancer genomics. Curr Opin Genet Dev. 2017;42:22–32.
38. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell. 2018;172:205–17. e12
39. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017;356:189–94.
40. Borgström E, Paterlini M, Mold JE, Frisen J, Lundeberg J. Comparison of whole genome amplification techniques for human single cell exome sequencing. PLoS One. 2017;12:e0171566.
41. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14:491–3.