# Accurate genotyping of single cells with Octopus

**Daniel Cooke** ( dcooke@well.ox.ac.uk )
University of Oxford https://orcid.org/0000-0002-0765-1557
**Gerton Lunter**
University of Oxford https://orcid.org/0000-0002-3798-2058
**David Wedge**
University of Manchester https://orcid.org/0000-0002-7572-3196

**Brief Communication**

**Keywords:** Variant Calling Tool, Amplification Stochasticity, Sequencing Error, Haplotype-based Bayesian Model

**Posted Date:** June 3rd, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-583831/v1

# Accurate genotyping of single cells with Octopus

**Daniel P. Cooke[1,*], Gerton Lunter[3,1], and David C. Wedge[2]**

[1]MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK. [2]Manchester Cancer Research Centre, University of Manchester, Manchester, UK. [3]Department of Epidemiology, University Medical Centre Groningen, Groningen, The Netherlands. [*]Correspondence should be addressed to D.P.C (dcooke@well.ox.ac.uk)

**We describe an extension to our variant calling tool, Octopus (https://github.com/luntergroup/octopus), for single-cell DNA sequencing data. Octopus jointly genotypes cells from a lineage, accounting for amplification stochasticity and sequencing error with a haplotype-based Bayesian model. Octopus is considerably more accurate at genotyping single cells than existing methods.**

Cancers are founded by a single cell that proliferates abnormally as a result of inherited or somatic mutation. Subsequent mutations in the founder's descendants chart the course of the tumor's evolution, including the birth of subclones that characterise intra-tumor heterogeneity and may lead to more aggressive disease and influence clinical decision making[1,2]. Single-cell DNA sequencing offers, in principle, the most detailed picture of a tumor's state and could improve on currently used variant allele frequency (VAF) clustering methods for tumor subclonal reconstruction[3–6].

We previously showed that our haplotype-aware variant calling tool – Octopus – achieves high accuracy in several clinically relevant experimental designs, including somatic mutation calling from bulk tumour sequencing[7,8]. However, none of the calling models previously discussed are appropriate for genotyping single cells as they do not account for common sources of uncertainty found in single-cell whole-genome amplified (scWGA) sequencing data, such as amplification stochasticity that can result in allelic dropout (ADO)[9]. Indeed, due to these artifacts, accurate variant calling of scWGA data is still considered a "grand challenge" of single-cell technology[10].

Similar to previous work[11–16], we designed a calling model for Octopus that accounts for amplification stochasticity in scWGA data (Methods). However, unlike other single-cell methods, Octopus discovers candidate variants with local *de novo* assembly; realigns reads to candidate haplotypes to account for sequencing and alignment errors; leverages physical read linkage information during genotyping; and recalibrates genotype quality scores using machine learning. In addition, Octopus jointly models cell genotypes with a phylogenetic tree prior. Homogeneous samples, including from bulk tissues, may be incorporated to improve accuracy. Variant calls, per-sample genotypes, and local cell phylogenies are inferred with variational Bayes for each called haplotype region. Octopus reports phased SNVs, indels and small complex replacements. Sample genotypes are assigned empirical probability scores with a random forest classifier trained on scWGA sequencing data, none of which is used for performance assessment here.

To assess calling accuracy, we ran Octopus on previously studied human primary fibroblast cells with closely related bulk samples that provide approximate ground truth for the cellular genotypes[12,14,15] (Methods). Octopus had considerably higher genotyping accuracy than SCcaller[12] – the only other single cell method that calls indels – despite SCcaller being given the baseline heterozygous variants to fit its allele bias model (Fig. 1a and Supplementary Table 1). Octopus genotyped $5\times$ more baseline indels correctly than SCcaller (Fig. 1c), and $2\times$ more baseline genotypes overall. The F-measure improvement of Octopus over SCcaller was $37\%$ and $66\%$ in kindred cells IL11 and IL12, respectively (Fig. 1b). The large performance differential for SCcaller is likely driven by higher rates of ADO in IL12, possibly due to IL12 being a doublet[14]. Surprisingly, MonoVar[11] and ProSolo[15] performed better on average than SCcaller despite not calling indels. However, they too had inconsistent accuracy in IL11 and IL12, and Octopus had substantially higher F-measure on average than both ($36\%$ and $31\%$ higher, respectively). Evaluation on the basis of allele match[17] also showed large accuracy improvements for Octopus (Supplementary Table 1).

The differences in sensitivity between methods may be partly explained by their response to heterozygous alleles affected by amplification stochasticity (Fig. 1d). As the most sensitive method, Octopus showed good recall across a wide VAF range in both IL11 and IL12, and was particularly robust to frequent ADO in IL12. Notably, Octopus calls clearly displayed characteristic trimodal scWGA VAF distributions[20], despite not explicitly modelling this. In contrast, SCcaller showed poor recall for ADO variants and strong VAF bias; it had good recall for high-VAF variants, but not for low VAF variants. MonoVar also had poor sensitivity for ADO variants.

Having established genotyping accuracy for clonal variants, we next sought to evaluate calling accuracy for subclonal mutations in the study population, because power to impute genotypes from joint calling with related samples, whether single cells or bulk, is reduced for these variants. To do so, we identified $120$ putative subclonal variants in kindred cells IL11 and IL12 (Methods). Octopus correctly called $39\%$ more curated subclonal mutations in both IL11 and IL12 than MonoVar and $200\%$ more than SCcaller despite making the fewest subclonal mutation calls overall (Fig. 1e). Notably, just $2\%$ of Octopus calls were subclonal compared with $87\%$ for SCcaller, $53\%$ for MonoVar, and $56\%$ for Prosolo (Fig. 1e and Supplementary Fig. $1$).

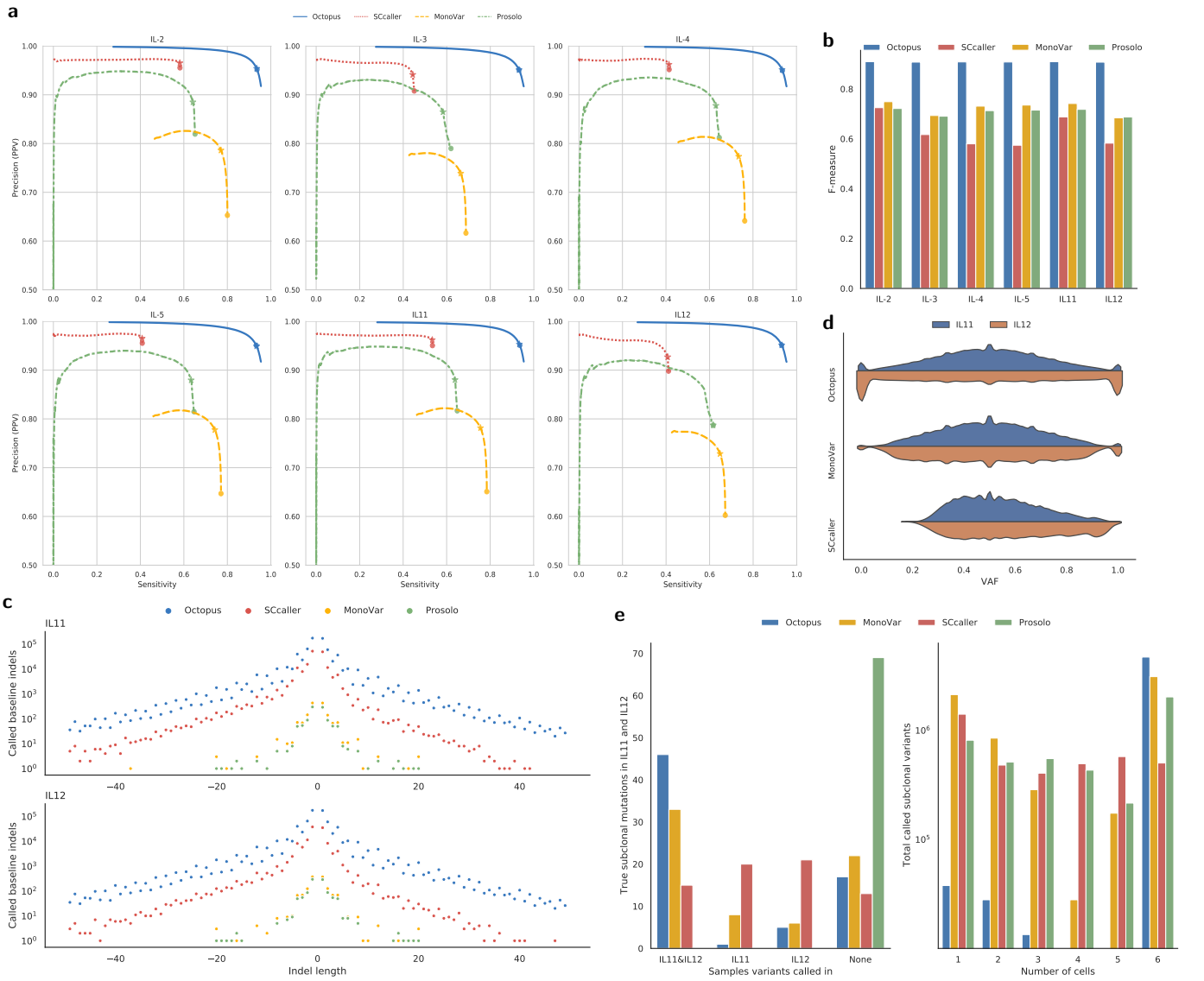


**Fig. 1 | Variant calling accuracy in single cells. a** Precision-recall curves for six human primary fibroblast cells using clonal variants detected in matched bulk samples by Strelka2[18] as the ground truth. Calls were evaluated with RTG Tools vcfeval[19]. Fields used to score variants were RFGQ (Octopus), GQ (SCcaller), QUAL (MonoVar), and GQ (Prosolo). Dots indicate default or recommended PASS values. Stars indicate the score with the maximum F-measure. **b** Maximum F-measures for all variants in each cell. **c** Indels in the baseline genotyped correctly. The maximum indel size is limited to 49bp by Strelka2. MonoVar and Prosolo do not call indels, but still call some baseline indels due to representation differences. **d** VAFs of biallelic heterozygous variants, determined by allele depths (AD) reported by the caller. Prosolo is not shown as it does not report AD. In the truth set, $55\%$ (2,427,808/4,453,494) of variants were biallelic heterozygotes. **e** True subclonal mutations in kindred cells IL11 and IL12 and all called subclonal mutations.

To further assess somatic mutation genotyping accuracy, we called variants in a previously studied triple-negative breast cancer dataset comprising 16 G2/M tumor cells (aneuploids a1-a8 and hypodiploids h1-h8) and 16 matched normal cells (n1-n16)[5]. We chose not to label normal samples for Octopus in order to test ability to recover this natural grouping (Methods). Octopus genotype calls were sufficiently accurate that hierarchical clustering on somatic calls with default filters not only partitioned normal and tumor cells correctly, but also robustly captured clonal loss-of-heterozygosity (LOH) events and the two aneuploid subclones originally identified[5,11,13] (Fig. 2a). Despite the random forest not being trained on subclonal somatic variants, clustering was broadly consistent when limited to cancer-related and all genes, with different levels of filtering (remarkably, even when using all 163,957 unfiltered somatic calls), and alternative clustering methods (Supplementary Fig. 2). Notably, subclonal aneuploid cells a1, a4, a6 were always clustered together, as were a2, a5, and a7. Cell a8 was inconsistently clustered between the aneuploid and hypodiploid groups, reflecting the uncertain origin of this cell[5,13]. In contrast, clustering of MonoVar and Prosolo genotypes failed to consistently partition normal and tumor cells – with and without filtering – and did not reveal LOH regions (Fig. 2b,2c). Furthermore, SCcaller only called $4$ non-singleton somatic mutations.

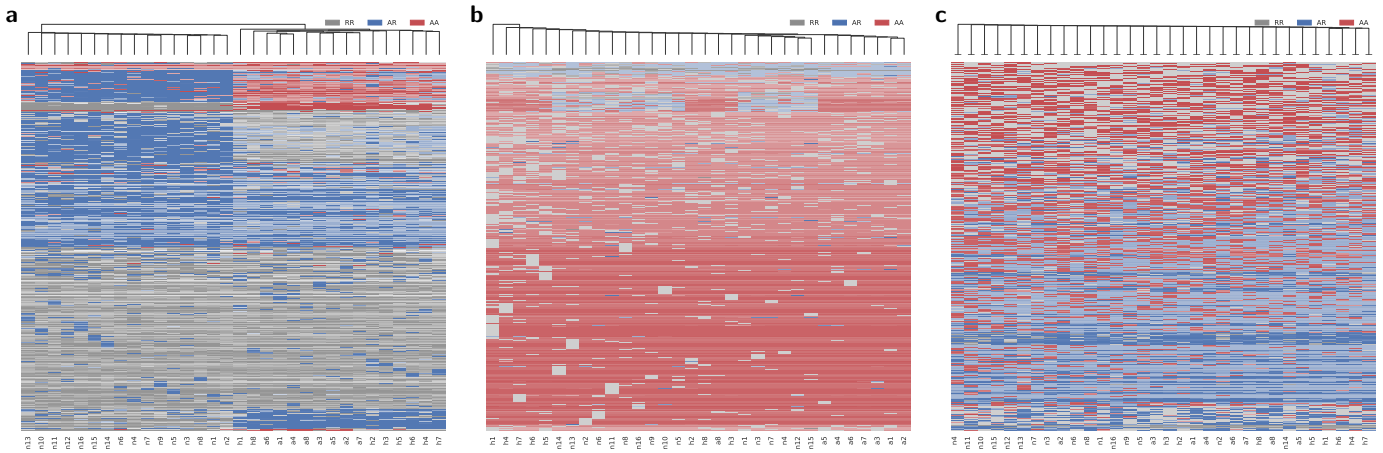


**Fig. 2 | Somatic alterations in triple-negative breast cancer single cells.** Hierarchal clustering of biallelic somatic mutation genotype calls in cancer related genes of 16 triple-negative breast cancer and 16 matched normal cells. Cells are colored by genotype call (RR: homozygous reference; AR: heterozygous; AA: homozygous mutant) and shaded by genotyping confidence (GQ). Homozygous tumor cells indicate loss of heterozygosity (LOH). Only calls with at least 2 cells called somatic were included. **a** 3409/163,957 PASS Octopus calls correctly partition normal and tumor cells, group anueploid subclone cells, and clearly show LOH regions, including those where a somatic mutation has also occured (RR in normal cells, AA in tumour cells). Sites with heterozygous normals and homozygous tumor cells have high uncertainty due to confounding LOH and ADO signals. LOH sites with subsequent somatic mutation are more confidently called. **b** 1792/1,117,566 filtered MonoVar calls. **c** 6605/37,003,396 filtered Prosolo calls. Both MonoVar and Prosolo fail to partition normal and tumor cells.

Octopus identified clinically relevant mutations not previously reported in this patient, including a homozygous - indicating LOH - clonal splice acceptor gain SNV in TP53 (chr17:g.7673609C>A; COSV52664150); clonal homozygous frameshift indels in FAS (chr10:g.89014337A>-), SMARCD1 (chr12:g.50087382TT>-), ETV6 (chr12:g.11891601->A), and TCF7L2 (chr10:g.113089447->C); a clonal in-frame insertion in DDX10 (chr11:g.108917908->TGA); and a subclonal frameshift deletion in MYO5A (chr15:g.52336478T>-). These mutations were confirmed in the bulk samples for this patient[5].

Our analyses demonstrate that haplotype-based modelling with biologically realistic phylogenetic tree priors leads to considerable improvements in genotyping accuracy of single cells, resulting in novel clinical insights. Several previous methods have proposed sophisticated nonlinear interpolation models conditioned on known heterozygous sites to account for amplification stochasticity[12,14]. Octopus does this implicitly – without variant databases – simultaneously accounting for sequencing and alignment error, by direct modelling of haplotypes.

Future versions of Octopus may address current limitations of the method. While the model presented is capable of explicitly calling zygosity changes, this feature is currently disabled by default as it increases runtime, particularly for amplifications. In addition, the model assumes each sample is homogenous, but this may not be true for bulk samples, and is also violated by cell doublets. Finally, retraining the forest model on subclonal variants would likely result in better calibrated quality scores for somatic genotypes, particularly in LOH regions. However, we believe the version of Octopus presented here represents significant progress in the "grand challenge" of single cell variant calling.

**Acknowledgements.** This work was supported by The Wellcome Trust Genomic Medicine and Statistics PhD Program (grant nos. 203735/Z/16/Z to D.P.C). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

**Author contributions.** D.P.C designed and implemented the algorithm, did the analysis, and wrote the paper. D.C.W provided data that supported the development of the algorithm. D.C.W. and G.L supervised the project.

**Ethics declaration.** The authors declare no competing interests.

## References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–13. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).
2. Fittall, M. W. & Van Loo, P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Med* **11,** 20. ISSN: 1756-994X (Electronic) 1756-994X (Linking) (2019).
3. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–4. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2011).
4. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148,** 873–85. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2012).

5. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–60. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2014).

6. Tarabichi, M. *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat Methods.* ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/s41592-020-01013-2 (2021).

7. Cooke, D. P., Wedge, D. C. & Lunter, G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology.* ISSN: 1546-1696. doi:10.1038/s41587-021-00861-3 (2021).

8. Cooke, D. P., Wedge, D. C. & Lunter, G. Benchmarking small-variant genotyping in polyploids. *bioRxiv,* 2021.03.29.436766 (2021).

9. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* **43,** e143. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2015).

10. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21,** 31. ISSN: 1474-760X (Electronic) 1474-7596 (Linking) (2020).

11. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant detection in single cells. *Nat Methods* **13,** 505–7. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2016).

12. Dong, X. *et al.* Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* **14,** 491–493. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2017).

13. Singer, J., Kuipers, J., Jahn, K. & Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nat Commun* **9,** 5144. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking) (2018).

14. Luquette, L. J., Bohrson, C. L., Sherman, M. A. & Park, P. J. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. *Nat Commun* **10,** 3908. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking) (2019).

15. Lähnemann, D. *et al.* ProSolo: Accurate Variant Calling from Single Cell DNA Sequencing Data. *bioRxiv,* 2020.04.27.064071 (2020).

16. Zafar, H., Navin, N., Chen, K. & Nakhleh, L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res* **29,** 1847–1859. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (2019).

17. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37,** 555–560. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2019).

18. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods.* ISSN: 1548-7105 (Electronic) 1548-7091 (Linking). doi:10.1038/s41592-018-0051-x (2018).

19. Cleary, J. G. *et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv.* doi:10.1101/023754 (2015).

20. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350,** 94–98. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2015).

21. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7,** 19. ISSN: 1472-6750 (Electronic) 1472-6750 (Linking) (2007).

22. Lin, C., Lu, J., Wei, Z., Wang, J. & Xiao, X. Optimal algorithms for selecting top-k combinations of attributes: theory and applications. *The VLDB Journal* **27,** 27–52. ISSN: 0949-877X (2018).

23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *bioRxiv* (2013).

24. Koster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34,** 3600. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (2018).

25. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17,** 261–272. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2020).

26. Waskom, M. *et al. mwaskom/seaborn: v0.8.1 (September 2017)* version v0.8.1. Sept. 2017. doi:10.5281/zenodo.883859. <https://doi.org/10.5281/zenodo.883859>.

27. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47,** D941–D947. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2019).

## Methods

**Read pre-processing for MDA chimeras.** DNA synthesized by MDA is known to contain sequence chimeras formed due to erroneous template switching, predominantly resulting in inverted repeats in sequencing reads[21]. These artefacts are problematic for variant calling in Octopus because the haplotype likelihood model does not account for these kind of errors, and base qualities in chimeric sequence are often high.

To limit the impact of these artefacts during genotyping, we added additional read pre-processing transformations to Octopus that mask (set base qualities to 0) read sequences deemed to be MDA related chimeras. Specifically, we align soft-clipped sequence with proximal inverted reference sequence and reference sequence 5' of the clip start. Currently, only clipped sequence with perfect alignments are masked.

**Single cell phylogeny prior.** We model cell phylogeny as binary trees consisting of nodes $\rho_1, \ldots, \rho_T$. For mathematical simplicity, the tree structure can be described using an ancestry matrix $A$ where

$$A_{ij} = \begin{cases} 1, & \text{if } \rho_j \text{ is ancestor of } \rho_i \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq i, j \leq T$. Then $\mathbb{P} = (A, \rho_1, \ldots, \rho_T)$ defines the phylogeny, and the probability of genotypes $\boldsymbol{g} = (g_1, \ldots, g_T)$ conditional on $\mathcal{P}$ is

$$p(\boldsymbol{g} \mid \mathbb{P}, \mathcal{M}_g, \mathcal{M}_s) = p(\boldsymbol{g}_1 \mid \mathcal{M}_g) \prod_{1 \leq i < j \leq T}^{T} p(\boldsymbol{g}_i \mid \boldsymbol{g}_j, \mathcal{M}_s)^{A_{ij}} \quad (1)$$

Here, $\mathcal{M}_g$ is a model of germline mutations, $\mathcal{M}_s$ is a model of somatic mutations and

$$p(\boldsymbol{g}_i \mid \boldsymbol{g}_j, \mathcal{M}_s) \propto p(|\boldsymbol{g}_i|, |\boldsymbol{g}_j|) \sum_{\kappa_1=1}^{|\boldsymbol{g}_j|} \cdots \sum_{\kappa_{|\boldsymbol{g}_i|}=1}^{|\boldsymbol{g}_j|} \prod_{k=1}^{|\boldsymbol{g}_i|} p(\boldsymbol{g}_{ik} \mid \boldsymbol{g}_{j\kappa_k}, \mathcal{M}_s) \quad (2)$$

where $|\boldsymbol{g}_i|$ is the ploidy of genotype $\boldsymbol{g}_i$ and $p(|\boldsymbol{g}_i|, |\boldsymbol{g}_j|)$ is the probability of copy number change which we define as

$$p(|\boldsymbol{g}_i|, |\boldsymbol{g}_j|) = \lambda^{||\boldsymbol{g}_i| - |\boldsymbol{g}_j||}$$

where $\lambda$ is a configurable parameter to the model.The somatic mutation model used to assign probabilities $p(\boldsymbol{g}_{ik} | \boldsymbol{g}_{jm}, \mathcal{M}_s)$ – the haplotype $\boldsymbol{g}_{ik}$ conditional on the haplotype $\boldsymbol{g}_{jm}$ is the same pair HMM model previously described for *de novo* and somatic mutations[7].

**Genotype model.** To account for amplification stochasticity, we model reads using a finite mixture distribution:

$$p(r \mid g, \pi) = \sum_k \pi_k \, p(r \mid g_k)$$

where $p(r \mid g_k)$ is the likelihood for haplotype $g_k$ as computed by the previously described pair HMM[7]. We assume that haplotype mixtures are distributed $\pi \sim \text{Dirichlet}(\alpha)$, which allows modelling of bulk samples by adjusting the concentration parameter $\alpha$ as the read counts become Binomially distrbuted in the limit as $\alpha \to \infty$. The default value of $\alpha$ is 5.

The complete joint likelihood function expresses that samples are randomly sampled from clones on the phylogenetic tree proportionally to the size of the clone.

$$p(\boldsymbol{R} \mid \boldsymbol{g}, \boldsymbol{\pi}, \psi) = \sum_{t=1}^{T} \psi_t \prod_{s=1}^{S} p(\boldsymbol{R}_s \mid \boldsymbol{g}_t, \boldsymbol{\pi}_{st})$$

Here, we assume that the clone mixture frequencies are also distributed $\psi \sim \text{Dirichlet}(\beta)$. The concentration parameter $\beta$ controls the dispersion of samples across clones and is configurable by the user. The default value is 20.

The posterior distribution of all latent variables is given by Bayes law:

$$p(\boldsymbol{g}, \boldsymbol{\pi}, \psi \mid \boldsymbol{R}, \mathcal{M}_{\mathbb{P}}) \propto p(\boldsymbol{g}, \boldsymbol{\pi}, \psi \mid \mathcal{M}_{\mathbb{P}}) \, p(\boldsymbol{R} \mid \boldsymbol{g}, \boldsymbol{\pi}, \psi)$$

where $\mathcal{M}_{\mathbb{P}}$ denotes the phylogenetic prior previously described.

As the posterior cannot be computed exactly, we approximate the full posterior with variational Bayes (VB) by introducing binary indicator variables $z$ and $w$ for read-haplotype and sample-clone assignments (Supplementary Fig. 3). For example, the per-read likelihood function becomes

$$p(r_{sn}, z_{sn} \mid g, \pi, w) = \prod_k \pi_k^{z_{snk}} \, p(r_{sn} \mid g_k)^{w_{st} z_{snk}}$$

The posterior value of $w_{st}$ is the *responsibility* clone $t$ in $\mathbb{P}$ takes for sample $s$ and $\sum_t w_{st} = 1$, for all $s$. We use the resulting VB inferences to compute a lower-bound on the model evidence, $p(\boldsymbol{R} \mid \mathcal{M}_{\mathbb{P}})$.

Interestingly, in the case where $T = 1$ (i.e., the phylogeny consists of just a single clone), the model is equivalent to the subclone model previously described for bulk tumour genotyping[7]. In practice, we actually defer to the subclone model in this case as the implementation is slightly faster.

A major problem with variational Bayes is that the inference procedure can become stuck at a local maximum, which may result in inaccurate inferences. A second problem is that the number of genotype combinations grows exponentially in the number of clones, so it is not practical to evaluate all combinations in general. To address both of these problems, we propose a subset of $m$ genotype combinations, where $m$ is configurable, the are ideally close to the highest probability genotype combinations of the full model. To do so, we first run the population model[7] on all samples, and then cluster samples based on their 'population' marginal genotype posterior distributions using *k*-medoids (where $k = T$). Next, we run the individual genotype model on each cluster by merging reads for each sample in the cluster. Finally, we select the best $m$ combinations, using a variant of the top-$k$ selection algorithm[22], where combinations are ranked by the sum of their individual model genotype posteriors. As the seeds to the VB model are simply the initial genotype posteriors, we just use the best $l$ ranked genotype combinations as seeds, where $l$ is configurable but $l \leq m$.

Another limitation of the model is that samples are assumed to originate from a single clone on the phylogeny, which means that cell doublets and subclonal bulk samples are not modelled.

**Cell calling model.** In order to to call variants and genotypes in the samples we apply the model described above. However, because this model is conditioned on a particular phylogeny, and since the number of trees is on the order $\mathcal{O}(T!!)$, we use a greedy iterative strategy to infer the phylogeny. Specifically, at each locus, we evaluate the model under both the trivial case of a single clone and the one possible phylogeny with two clones. If the evidence of the latter is greater than the former, then we evaluate the two possible phylogenies with three nodes (linear and branching). We continue by extending the phylogeny with the highest evidence with a single new node in all ways – but ignoring isomorphisms – until either the evidence decreases or a configurable maximum number of nodes is reached.

Next, we calculate posteriors for all evaluated phylogenies using Bayes law assuming a uniform prior, and compute haplotype and genotype posteriors required by all Octopus calling models (for haplotype reduction and phasing) by marginalising over phylogenies. For example, the per-sample genotype posteriors are given by

$$p_s(g \mid \boldsymbol{R}) = \sum_{\mathbb{P}} p(\mathcal{M}_{\mathbb{P}} \mid \boldsymbol{R}) \sum_t w_{st \mid \mathbb{P}} \, p_t(g \mid \boldsymbol{R}, \mathcal{M}_{\mathbb{P}})$$

Finally, the maximum a posteriori (MAP) phylogeny is called

$$\hat{\mathbb{P}}_{\text{MAP}} = \arg\max_{\mathbb{P}} p(\mathcal{M}_{\mathbb{P}} \mid \boldsymbol{R})$$

and the clone responsibilities from this model, $w_{st \mid \hat{\mathbb{P}}_{\text{MAP}}}$, are used to assign samples to clones in the tree – the MAP clone is chosen – and thereby also call genotypes.

**Training Octopus' random forest.** We trained Octopus' random forest model using a previously described scWGA dataset[20]. Specifically, we used the 15 single neuronal nuclei from "Brain B" (SRA BioProject PRJNA245456) for training. Octopus was used to generate baseline clonal genotypes from heart and cortex bulk samples from the same individual. Somatic mutations called in the cortex sample ($2954$ variants) were ignored.

**Clonal mutations in fibroblast cells.** We downloaded paired-end fastq files from SRA (BioProject PRJNA305211) and mapped reads to GRCh38 (specifically hs38DH) with BWA-MEM[23].

To generate baseline genotypes, we joint called matched bulk samples Hunamp and IL1C using Strelka2[18] (v2.9.10) in germline mode. We then generated high-confidence clonal regions by removing loci where inconsistent genotypes were called in each sample.

We evaluated calling accuracy by comparing caller and baseline calls in the high-confidence regions using RTG Tools vcfeval[19] (v3.12). To compare by allele match, we ran vcfeval with the *--squash-ploidy* option. Precision-recall curves were plotted by setting the *--score-field* option to vcfeval and using the resulting *weighted_roc.tsc.gz* file.

We ran Octopus (v0.7.4), SCcaller (v2.0.0), MonoVar (v0.0.1), and Prosolo (v0.6.1) in a similar way to a previous analysis of this dataset[15]. For Octopus, we used the *cell* calling model to jointly call all cells and bulk clones C1-3. The *--sample-dropout-concentration* option was set to 50 for bulk clones. For SCcaller, we called variants in each cell and merged bulk clones for the matched bulk sample. Heterozygous SNVs found in the baseline set were given to the *--snp_in* option. We also needed to use custom code to fix indel calls as these are reported in invalid VCF by SCcaller. Prosolo was also provided with the merged bulk clone. Since SCcaller and Prosolo do not jointly call cells, we genotyped each cell independently and merged calls with bcftools merge. All analysis was coded into a Snakemake[24] workflow (https://github.com/luntergroup/single-cell).

**Subclonal mutations in fibroblast cells.** We found putative subclonal mutations in fibroblast cells IL11 and IL12 by calling somatic mutations unique to the kindred bulk IL1C (i.e. not present in bulk clones C1-3), using the population bulk Hunamp as a normal. Specifically, we ran Strelka2 in somatic mode on Hunamp and IL1C. As Strelka2 does not support multi-sample somatic calling, we also ran Octopus in cancer mode on all bulk samples, and selected passing somatic mutations called by both Octopus and Strelka2 in IL1C, but ignored those genotyped for the variant in C1-3 by Octopus. Another reason that we decided to include Octopus to assist producing subclonal calls was that Strelka2 showed surprisingly low precision in these data; Strelka2 called 19,415 somatic mutations in IL1C, compared with 4891 by Octopus. To eliminate the possibility of including mutations *de novo* in IL1C (during expansion), we intersected the putative bulk set with the union of IL11 and IL12 calls from all callers. In summary, there were $377$ mutations called unique to IL1C by Octopus, $363$ of these were also called by Strelka2, and $120$ of these were found in IL11 or IL12 by at least one single cell caller. This implies $\sim 250$ somatic mutations in IL1C that occurred during expansion after divergence from IL11 and IL12, which is not unreasonable given $\sim 5$ cell divisions.

We then performed a three-way genotype intersection – IL11 and IL12 calls from each caller, along with the putative subclonal mutations – using Starfish (https://github.com/dancooke/starfish).

**Somatic mutations in triple-negative breast cancer.** We downloaded single-end fastq files from SRA (BioProject PRJNA168068) and mapped reads to GRCh38 (specifically hs38DH) with BWA-MEM[23].

To call somatic variants, we ran callers in the same way as described for the fibroblast cells (for SCcaller and Prosolo, we constructed a bulk sample by merging normal cells), and filtered these based on somatic status. Specifically, for Octopus we filtered by the SOMATIC flag; for SCcaller, we filtered using the SO flag and a minimum read depth of 20 (as recommended); for MonoVar and Prosolo, which do not specifically flag somatic mutations, we simply selected mutations genotyped for a variant in the tumor cells but not in the normal cells. We note that, unlike a previous evaluation of MonoVar on this dataset[11], we do not use the bulk samples to filter somatic variant calls in order to better determine raw calling performance. In addition, we only filtered MonoVar calls by QUAL as the filtering approach described in Zafar et al.[11] relies on a closed-source script.

Hierarchical clustering of biallelic somatic calls was done using SciPy[25] via Seaborn[26]. Specifically, called genotypes were encoded into a real number using the function

$$f(g, GQ) = \begin{cases} GQ - 1, & \text{if } g = \text{RR} \\ 1 + GQ, & \text{if } g = \text{AR} \\ 3 + GQ, & \text{if } g = \text{AA} \end{cases}$$

and Euclidean distances were used for UPGMA clustering (Fig. 2) and Ward clustering (Supplementary Fig. 2). Missing genotypes were converted to homozygous reference (RR) and assigned a GQ of $3$. Somatic mutations were identified in COSMIC cancer consensus genes[27].

To validate novel mutations called by Octopus, we ran Strelka2 in somatic mode on the matched tumor-normal paired bulk samples from the same patient. Unfortunately, we were not able to use these calls as a general ground truth to evaluate somatic mutation calling accuracy as the callset was highly imprecise - there were 18,363 PASS somatic calls; almost two orders of magnitude higher than would be expected for a breast cancer exome footprint. Manual review of some of these calls showed clear support for many called somatic mutations in the normal bulk and tumor cells. These observations further highlight the challenging nature of this single-end sequencing dataset.

**Code availability.** Octopus source code and documentation is freely available under the MIT licence from https://github.com/luntergroup/octopus. Snakemake code used for data analysis is available from https://github.com/luntergroup/single-cell.

**Data availability.** All raw data evaluated is publically available: the fibroblast data is available from SRA under BioProject PRJNA305211; the triple-negative breast cancer data is available from SRA under BioProject PR-JNA168068.

**Author contributions.** D.P.C designed and implemented the algorithm, conceived and performed the analysis, and wrote the manuscript. D.C.W and G.L supervised the project.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf
- supptable1.xlsx