nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-019-0366-2

Linked-read analysis identifies mutations in single-cell DNA-sequencing data

Craig L. Bohrson[1,2], Alison R. Barton[1,2], Michael A. Lodato[3,4,5], Rachel E. Rodin[3,4,5,6], Lovelace J. Luquette[1,2], Vinay V. Viswanadham[1,2], Doga C. Gulhan[1], Isidro Cortés-Ciriano[1,7], Maxwell A. Sherman[1], Minseok Kwon[1], Michael E. Coulter[3,4,5,6], Alon Galor[1], Christopher A. Walsh[3,4,5] and Peter J. Park[1*]

Whole-genome sequencing of DNA from single cells has the potential to reshape our understanding of mutational heterogeneity in normal and diseased tissues. However, a major difficulty is distinguishing amplification artifacts from biologically derived somatic mutations. Here, we describe linked-read analysis (LiRA), a method that accurately identifies somatic single-nucleotide variants (sSNVs) by using read-level phasing with nearby germline heterozygous polymorphisms, thereby enabling the characterization of mutational signatures and estimation of somatic mutation rates in single cells.

Comprehensive profiling of genetic mutations by whole-genome sequencing has aided in answering fundamental questions in biology and medicine. Applied to single cells, it has the unique potential of revealing genetic variation within an organism at unmatched resolution. With novel protocols for generating single-cell libraries and reduced sequencing costs, large-scale single-cell DNA-sequencing studies of genetic variation have now become feasible, utilizing whole exomes[1–3], low-coverage whole genomes (for copy number analysis)[4], and, more recently, higher-coverage whole genomes[5,6]. However, widespread application of single-cell whole-genome sequencing has been hindered by the technical errors associated with whole-genome amplification (WGA) and the lack of computational methods that can properly remove such errors.

There are three main single-cell WGA protocols in use: (1) degenerate oligonucleotide-primed PCR; (2) multiple annealing and looping-based amplification cycles[7]; and (3) multiple displacement amplification (MDA)[8]. A fourth method, linear amplification via transposon insertion, has been recently described but is not yet widely used[9]. Although degenerate oligonucleotide-primed PCR and multiple annealing and looping-based amplification cycles generate reproducible coverage profiles and have been used for detection of large-scale copy number variation[10,11], their coverage is confined to a subset of the genome and they are characterized by high single-nucleotide error rates (approximately 1 error in 10^4 bases)[12]. For single-cell sSNV detection, MDA tends to produce the lowest false positive and false negative rate, owing to the fidelity of the ϕ29 polymerase (approximately 1 error in 10^6–10^7 bases)[12,13], a relatively low allelic dropout rate, and high genomic coverage. As such, MDA has become the most common WGA method in studies aiming to identify sSNVs[10,11].

However, identifying sSNVs in single-cell sequencing data acquired using MDA remains difficult. Despite its relative fidelity, ϕ29 is expected to produce hundreds to thousands of polymerase errors in the first replication of the genome alone. Additionally, genomic DNA may be damaged during cell lysis or other sample preparation steps before MDA and may be unfaithfully copied early on during amplification. In particular, heat is known to induce cytosine deamination[14–16]; deaminated cytosine created during cell lysis may introduce a substantial burden of artifactual C>T mutation calls. As an exponential amplification process, MDA will cause early errors to propagate. Although these artifacts theoretically will have a lower variant allele fraction (VAF) than expected for an sSNV, high variance in VAF due to allelic dropout and amplification bias may cause a fraction of artifactual calls to reach VAFs comparable to those of true mutations in the final amplification product.

This high level of technical noise has made the validation of putative sSNVs important. Given that MDA consumes the original genome of the cell, past studies have aimed to confirm that candidate sSNVs from individual cells are present in other cells from the same organism. This can be done in bulk tissue by using an orthogonal technology such as droplet digital PCR or amplicon sequencing, or by looking for sSNVs that are shared across multiple single cells[5,7]. While this strategy is viable for the validation of mosaic or subclonal sSNVs, it cannot validate those detected in only one cell (singletons). This precludes analysis of sSNVs occurring post-mitotically in differentiated single cells, and sSNVs shared by arbitrarily small cell populations.

In this report, we present LiRA, which applies read-backed phasing with nearby heterozygous germline single-nucleotide polymorphisms (SNPs) to identifying sSNVs in whole-genome sequencing data derived from amplified single-cell genomes. Although the idea of phasing has been used to identify somatic mutations in cancer[17], somatic mosaicism[18–20], and de novo germline mutations[20], this is the first method that applies this principle to single-cell data. Also, although leveraging heterozygous variants to improve single-cell variant calling has been previously described and implemented as SCcaller[21], no existing sSNV caller makes use of read- or mate-pair-backed phasing. By ensuring that candidate mutation calls of interest are consistent with the haplotypes implied by nearby

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [2]Bioinformatics and Integrative Genomics PhD Program, Harvard Medical School, Boston, MA, USA. [3]Division of Genetics and Genomics, Manton Center for Orphan Disease, and Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA, USA. [4]Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, USA. [5]Broad Institute, Cambridge, MA, USA. [6]Program in Neuroscience and Harvard/MIT MD-PHD Program, Harvard Medical School, Boston, MA, USA. [7]Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, UK. *e-mail: peter_park@hms.harvard.edu

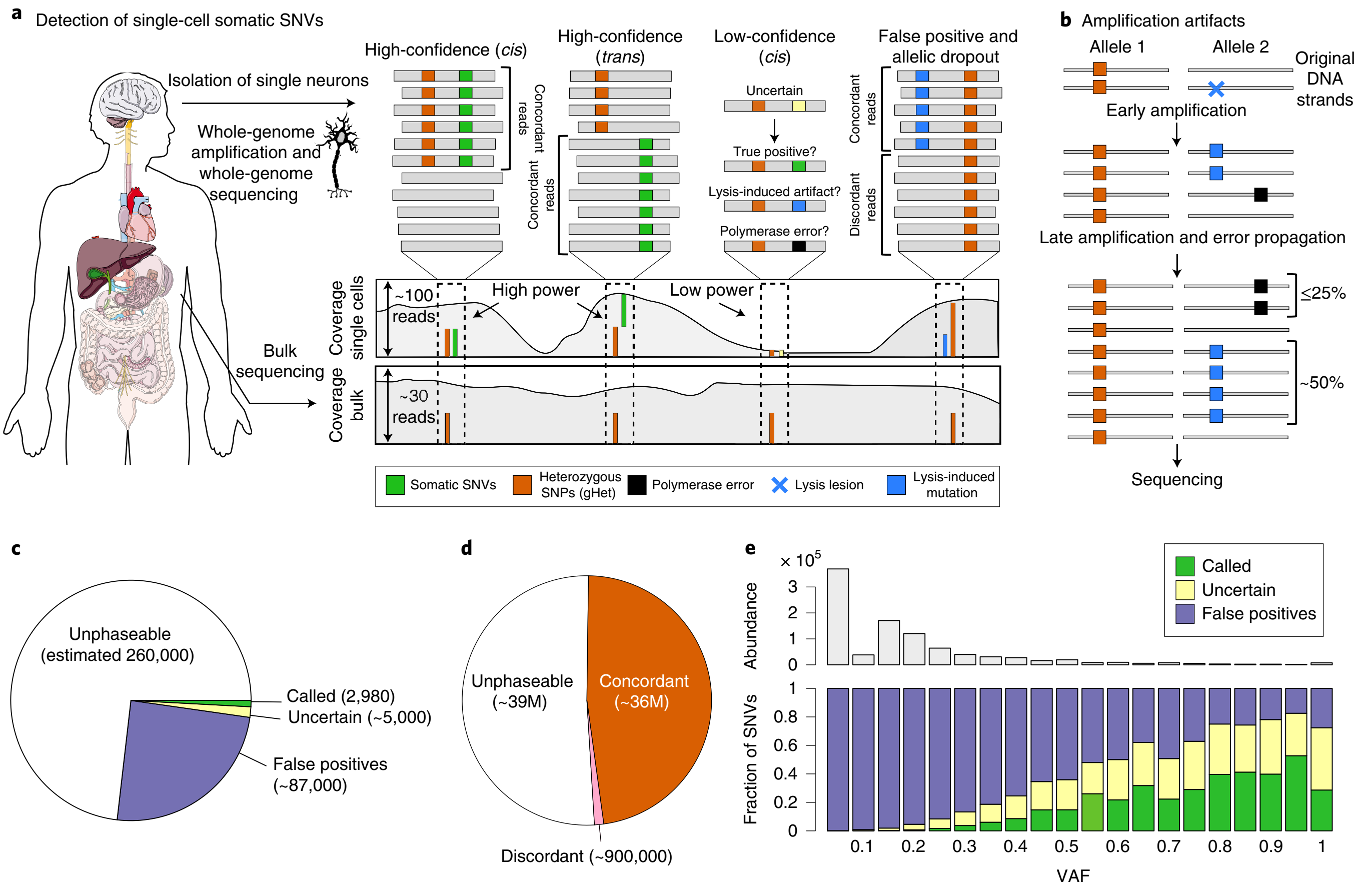

Fig. 1 | Overview of LiRA. a, Methodology for identifying false positive sSNVs. LiRA analyzes reads and mate-pair reads that cover the positions of an sSNV and a gHet (spanning reads). 'Concordant reads' support the gHet allele (*Alt/Ref* in *cis/trans*) and the sSNV alt call. 'Discordant reads' support the gHet allele but the reference base at the sSNV position. Some images are adapted from the Servier Medical Art PowerPoint Image Bank (https://smart.servier.com/) under a Creative Commons license (https://creativecommons.org/licenses/by/3.0/). **b**, Model for how the linked read pattern specific to false positives arises from DNA lesions or polymerase errors. A lesion may be present on and copied from one strand of input DNA (blue), or ϕ29 polymerase may mispair a base (black). Both errors are exponentially amplified. Since polymerase errors are introduced after the first round of amplification at the earliest, they are expected to appear in ≤25% of gHet-linked reads, whereas lesion-derived artifacts are expected to appear in ~50%. **c**, Classification of candidate sSNVs in LiRA. Most sSNV candidates (estimated ~260,000, 73%) are too far away from a gHet to be covered by the same read or mate pair. Over the powered fraction (27%, ~95,000), most (92%, ~87,000) are filtered as false positives due to the presence of at least one discordant read covering the sSNV position and each linked gHet. In the remaining subset, most (63%, ~5,000) do not meet LiRA's quality thresholds, and 2,980 (37%, 0.8% overall) are reported as LiRA sSNV calls. **d**, Phasing of gHets. Just under half of gHets are close enough to other gHets to be linked, and only 2% are filtered (erroneously) as false positives. **e**, Call status of candidate sSNVs in LiRA by VAF. Most sSNV candidates are low VAF; LiRA filters almost all low VAF sSNV candidates. As VAF increases, sSNV candidates are more frequently called, but a substantial proportion of high VAF candidates are still false positives.

heterozygous germline SNPs, LiRA removes amplification-associated variants and achieves unparalleled specificity.

Results

LiRA aims to provide robust validation for a subset of candidate sSNVs occurring near polymorphic germline heterozygous SNPs (gHets) (Fig. 1a). The key insight underlying LiRA is that false positive calls are derived from factors specific to one strand of DNA, while sSNVs, as fixed mutations, are derived from both strands of one chromosome. Single reads as well as mate pairs covering the genomic positions of a candidate sSNV and the linked locus of a nearby gHet ('spanning reads') can distinguish these two scenarios. For example, for a true sSNV occurring on the same haplotype as the linked gHet, every spanning read will contain both the gHet and the sSNV. We call these reads 'concordant' (Fig. 1a). In contrast, for a false positive, the set of spanning reads will contain a mixture of reads containing only the gHet and reads containing both the gHet and the sSNV. For false positives derived from DNA damage, the reads containing the reference allele, which we call 'discordant reads' (Fig. 1a), originate from the undamaged strand of the same chromosome, and for polymerase errors, from faithfully copied strands of the same chromosome (Fig. 1a,b).

In LiRA, we first utilize the Genome Analysis Toolkit (GATK)[22] to identify as many candidate sSNVs as possible; in principle, any variant caller with high sensitivity could be used in this first step. Candidate sSNVs are identified as any variants that are phased with gHets (any population-polymorphic heterozygous SNV calls made in bulk) and not present in matched bulk sequencing data. We identify sSNVs and gHets that are supported by the same read or mate pairs; any sSNVs that are found with any discordant reads are filtered and referred to as 'LiRA false positives' (Supplementary Fig. 1; for a case with multiple gHets linked to an sSNV, see Supplementary Fig. 2).

Application to single-neuron sequencing data. We applied our method to single-neuron data (~45×) from phenotypically normal individuals[5]. We found that a substantial portion of sSNV candidates are close enough to gHet sites to be subjected to LiRA analysis (27% overall, 9–44% in individual cells; Fig. 1c and Supplementary Table 1). Among those identified, 92% are filtered as LiRA false positives (87–96% across cells; Fig. 1c,d and Supplementary Table 1). Applying the same procedure across gHet–gHet pairs, we found that only 2% of gHets are filtered as LiRA false positives (1–4% across cells; Fig. 1d and Supplementary Table 1). This stark difference in results between gHet–gHet and candidate sSNV–gHet pairs suggests that filtering sSNVs using LiRA removes false positives while excluding minimal true variation, and that standard genotypers cannot be used to call sSNVs in single cells without considerable filtering and validation.

After filtering based on the presence of discordant reads, we determined the quality of the remaining variants based on a measure called 'composite coverage', defined as the minimum spanning read depth across bulk and single-cell sequencing data (Supplementary Figs. 1 and 2). As the bulk coverage observed at the same locus increases, it becomes increasingly likely that a candidate sSNV is not a missed germline variant. As single-cell coverage increases, so does confidence that discordant reads are truly absent in the MDA amplification product and not simply missed due to undersampling. LiRA approaches this issue by finding a composite coverage threshold that controls the estimated false discovery rate (FDR) at a tolerable level (default of 10%). sSNVs with support equal to or greater than the threshold are called, whereas those with subthreshold support are called as 'uncertain' (Fig. 1c,d).

To determine an appropriate threshold, we took advantage of the fact that, in the absence of false positives, the estimated genome-wide sSNV rate should not depend on composite coverage. LiRA first measures the distribution of composite coverage values at all genomic positions linked to any gHet site on both autosomal alleles (Supplementary Fig. 3); it then uses this and the distribution of the composite coverage values of sSNV candidates to compute the composite coverage-specific genome-wide sSNV rates (Supplementary Figs. 4 and 5). LiRA then models the observed relationship between the somatic mutation rate and the composite coverage as the mixture of two components: (1) an exponentially decaying error component; and (2) an approximately constant true mutation component (Supplementary Figs. 4 and 5). The utility of the model lies in the fact that the fitted true mutation component gives an estimate of the genome-wide sSNV rate, and its value relative to the error component gives an estimate of the FDR at each level of sSNV quality. This information is used to assign false positive probabilities to individual mutations, and to ascertain the overall FDR expected across a set of calls at various thresholds.

We found that this procedure identifies many uncertain sSNVs among the candidate set of non-discordant sSNV calls (63% overall, 20–81% across cells; Fig. 1c and Supplementary Table 1). Also, we found that while higher VAF calls are more likely to be called by LiRA, uncertain and LiRA false positives still frequently appear at high VAF values (Fig. 1e). This suggests that read-level phasing by LiRA adds substantial specificity to mutation calling in single-cell sequencing data acquired using WGA.

Importantly, the two-component model used by LiRA is fitted for each cell individually, and thus can account for variable artifactual burdens among samples. Instead of choosing a universal cutoff across all cells involved in a study, LiRA chooses a threshold for each so that the FDR is controlled at a specified level. In the Lodato et al.[5] data, this proved to be an important consideration, since the ratio of errors to predicted true sSNVs (a measure of artifactual burden) varied widely across cells (Supplementary Table 2).

In terms of the genome-wide sSNV rate (Supplementary Fig. 4), we found that LiRA-estimated rates were generally consistent with those found by a previous study[23], which measured the sSNV rate at a subset of genomic loci in the frontal cortex using an orthogonal method that avoided single-cell sequencing or WGA (Supplementary Table 2). This suggests that LiRA accurately accounts for the heterogeneity in power across the genome introduced through MDA-related coverage non-uniformity. Although only a small fraction (Fig. 1c,d) of initial candidate sSNVs are eventually output as LiRA calls, we found that LiRA retains sufficient sensitivity for downstream analysis. For Lodato et al.[5], we detected an average of 83 sSNVs per cell, which extrapolates to 919 sSNVs per cell genome-wide (Supplementary Table 2).

Comparison to other variant callers. To confirm the accuracy of LiRA's sSNV calls, we compared the VAF distribution of LiRA sSNV calls to that of LiRA false positives, LiRA uncertain calls, and gHets (Fig. 2a). True sSNVs should be characterized by a VAF distribution similar to that of gHets, whereas false positives should have lower VAFs, owing to their origin in progressively later rounds of amplification or on one strand of DNA. Accordingly, we found that LiRA calls had a distribution nearly identical to that of gHets, whereas LiRA false positives and, to a lesser extent, LiRA uncertain calls were skewed toward lower VAF values. This remained true when the VAF distributions were split by mutational type (C>A, C>G, C>T, T>A, T>C, T>G) (Supplementary Fig. 6). Overall, these results were consistent with the notion that LiRA distinguishes bona fide fixed heterozygous sSNVs from amplification-induced artifacts.

To compare LiRA to other methods for calling sSNVs, we compared the VAF distribution of LiRA calls to those produced by SCcaller[21], Monovar[24], GATK[22], VarScan[25], and MuTect[26] (Fig. 2a). Unlike the LiRA-derived or germline VAF distribution, calls reported by other variant callers produced VAF distributions skewed toward low VAF calls and inconsistent with the VAF distribution of gHets. As a further comparison, we intersected call sets from each of these packages with phaseable sSNV candidates that LiRA could analyze (Fig. 2b). We found that all methods tested had a substantial burden of LiRA false positives, indicating the presence of discordant reads, and uncertain variants (Fig. 2b), suggesting that LiRA achieves a much lower FDR in calling sSNVs from WGA-amplified single cells.

Another validation of LiRA is the comparison of the single-nucleotide substitution types between LiRA sSNV calls and LiRA false positives. High-quality GATK calls (marked with the 'PASS' flag) found to be false positives have different mutational frequencies from LiRA calls ($P < 10^{-5}$, Fisher's exact test). The false positives were depleted in C>G, T>A, T>C, and T>G mutations while enriched in C>T and C>A calls, which have been associated with artifactual sSNV calls in previous studies[14–16,27] (Fig. 2c). The largest depletion was observed for T>G (~320% higher in LiRA calls); the largest enrichment was observed for C>A (~50% lower in LiRA calls; Fig. 2c). These results suggest that the two sets originate from different underlying processes. An expanded analysis over trinucleotide context upheld these observations (Supplementary Fig. 7).

Application to cancer genomic data. A highly specific set of somatic mutations in single neurons obtained by LiRA has allowed us to discover the association between the number of mutations and age, identify the underlying mutational processes, and estimate the rate of mutagenesis[28]. To investigate its applications to cancer genomics, we analyzed a single-cell exome sequencing dataset of bladder cancer and normal tissues from Li et al.[29]. This analysis revealed that, when a relatively large number of mutations are shared across cells (as is the case in cancer), LiRA can be used to confidently determine with a small number of reads whether an sSNV is present or absent from a cell. In approaches agnostic to linkage, not detecting a mutation in a cell could be due to dropout or the fact that the mutation is really absent. Thus, in previous approaches, calling the status of a cell lacking evidence of an sSNV has involved models where the probability that a mutation is truly absent depends on the depth of

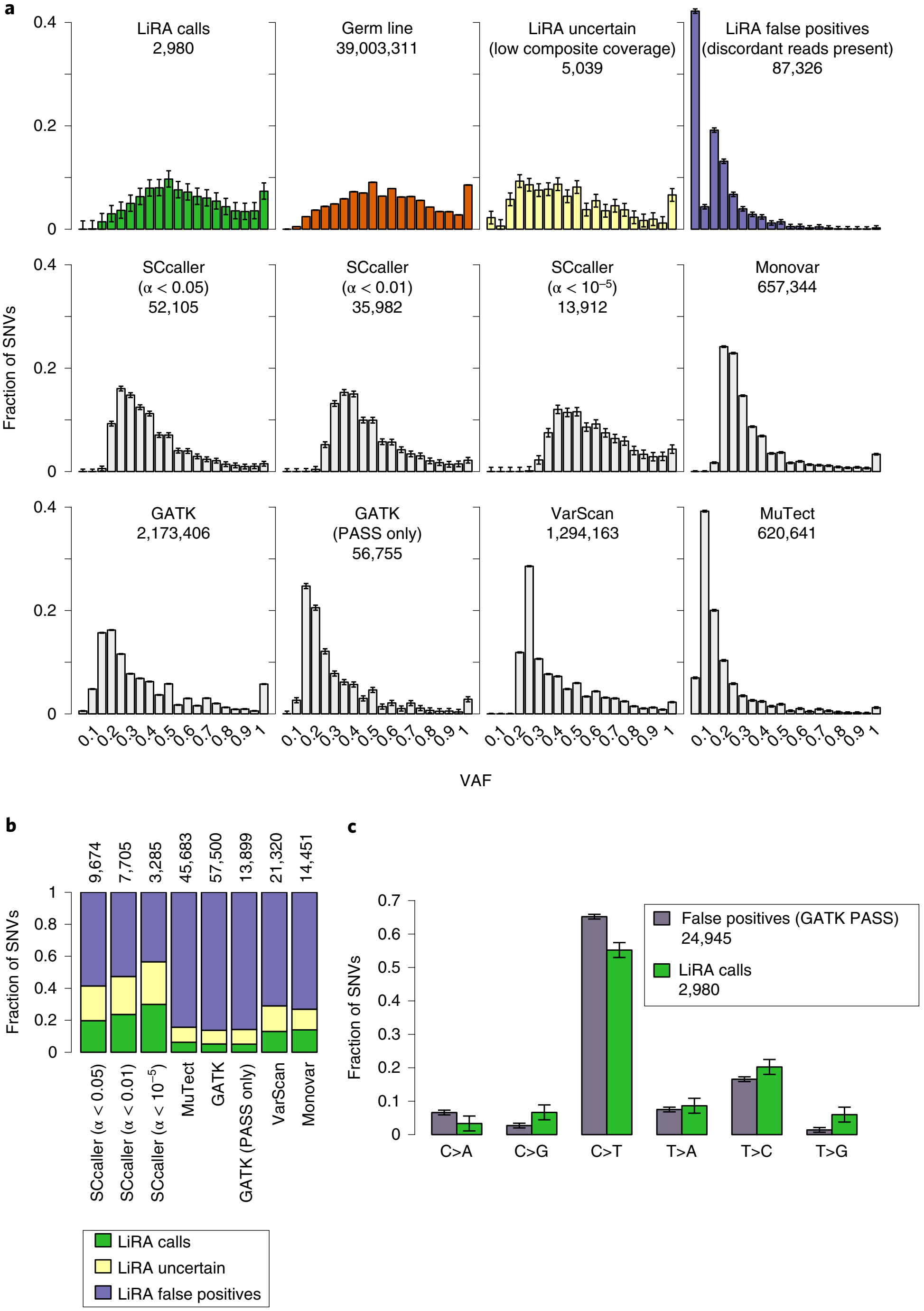

Fig. 2 | Performance of LiRA compared to other calling methods. a, Comparison of the VAF of LiRA high-confidence calls, uncertain sSNVs, and false positives to germline mutations and other calling methods. LiRA calls have a VAF distribution indistinguishable from that of heterozygous germline polymorphisms, while LiRA uncertain mutations and false positives are moderately and severely skewed toward low VAF values, respectively. Other single-cell variant calling methods also produce VAF distributions skewed toward low VAF values. Accepting only PASS mutations after variant quality score recalibration in GATK does not change this. In SCcaller, α is the probability that a candidate sSNV is an amplification artifact, and a set of calls is obtained by accepting only those with α less than a user-set threshold. Lowering α mitigates but does not remove skewing toward low VAF values. The 99% simultaneous CIs on frequency are shown and the total number of calls made is listed below each label. **b**, Call status of sSNVs called by other methods in LiRA. Calls made by single-cell variant calling methods contain many variants filtered as false positives in LiRA. Accepting only PASS mutations after variant quality score recalibration in GATK does not change this. In SCcaller, lowering α lowers the proportion of the variants identified in LiRA as false positives, but the proportion remains high. The 99% simultaneous CIs are shown and the size of the LiRA-intersection is listed above each bar. **c**, Comparison of sSNV types between LiRA false positives and LiRA calls. Well-supported LiRA false positives, distinguished as those that are marked as 'PASS' by GATK, differ significantly from LiRA calls in mutational spectra. The 99% simultaneous CIs are shown.

sequencing at the locus and the overall rate of allelic dropout[24,29]. In contrast, with LiRA, we can confidently call absence with just one read spanning an sSNV locus and a nearby gHet since, as we have demonstrated over gHets, the rate of errors producing discordant reads is low (Fig. 1d; 98% of linked gHets are linked with only concordant reads).

In the cancer data, LiRA identified a high resolution 'scaffold' of sSNVs over which we had confident linked positive and null mutation calls. We then extended this by identifying unlinked mutations that had a pattern of support across cells more associated with a LiRA-identified mutation than was expected by chance (Methods). This analysis resulted in the identification of several non-synonymous mutations and one nonsense mutation not found in the original study, in addition to recapitulating the clustering results of Li et al.[29] (Supplementary Fig. 8). The majority of these mutations were also not found using Monovar when this dataset was used in its validation[24]. Among the new mutations was a non-synonymous mutation in *SYTL3*, a gene that has been previously implicated in bladder cancer through its involvement in the Rab pathway via interaction with Rab27 (ref. [30]).

Discussion

Our results show that LiRA represents an advance in sSNV calling in single cells, especially with respect to singletons. Whereas existing variant callers produce variants with very high FDRs, LiRA produces a set of high-precision calls that display the characteristics of fixed, heterozygous sSNVs. Although there is a limitation of observing the single-cell genome only around gHets, LiRA still produces a sufficient number of accurate calls from which insights on biological processes can be gathered. In future studies, the utility of LiRA over single-cell sequencing data could be improved further when used in combination with longer reads or synthetic long reads such as those provided by the 10x Genomics platform, greater depth of coverage, or greater heterozygosity in the diploid genome. While the last factor is not easily modifiable in humans, mice, or other model organisms, crossing distantly related strains may yield very high rates of heterozygosity and greatly improve LiRA's power.

In theory, there are error modes in amplification that would cause false positives to escape LiRA's filtering steps. LiRA relies on both strands of a single chromosome being subject to relatively even amplification. If present, strand dropout or severe non-uniformity in strand-specific amplification could cause DNA lesions or polymerase errors to appear as fixed mutations in single-cell sequencing data using LiRA. Shorter or more heterogeneous amplicon sizes in MDA might worsen this effect, as might cell lysis protocols other than the alkaline-based one used previously[5,28]. Although we cannot technically rule out strand dropout, the quality of the two-component model fitted across cells (Supplementary Fig. 5), as well as the other properties of LiRA calls, suggest that this process is of negligible effect size.

Overall, as a new approach to single-cell analysis, LiRA provides a window into the mutational processes within a cell, including rate and characteristics of mutagenesis, leading to new insights into cell aging, lineage, and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0366-2.

Received: 22 January 2018; Accepted: 1 February 2019;
Published online: 18 March 2019

References

1. Leung, M. L., Wang, Y., Waters, J. & Navin, N. E. SNES: single nucleus exome sequencing. *Genome Biol.* **16**, 55 (2015).
2. Xu, X. et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
3. Hou, Y. et al. Single-cell exome sequencing and monoclonal evolution of a *JAK2*-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
4. Baslan, T. et al. Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
5. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
6. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
7. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
8. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
9. Chen, C. et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189–194 (2017).
10. Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genomics Hum. Genet.* **16**, 79–102 (2015).
11. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
12. de Bourcy, C. F. A. et al. A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
13. Esteban, J. A., Salas, M. & Blanco, L. Fidelity of phi 29 DNA polymerase: comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.* **268**, 2719–2726 (1993).
14. Fryxell, K. J. & Zuckerkandl, E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**, 1371–1383 (2000).
15. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405–3410 (1974).
16. Frederico, L. A., Kunkel, T. A. & Shaw, B. R. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**, 2532–2537 (1990).
17. Usuyama, N. et al. HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. *Bioinformatics* **30**, 3302–3309 (2014).
18. Freed, D. & Pevsner, J. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.* **12**, e1006245 (2016).
19. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
20. Ramu, A. et al. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10**, 985–987 (2013).
21. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
22. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
23. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
24. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant detection in single cells. *Nat. Methods* **13**, 505–507 (2016).
25. Koboldt, D. C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
26. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
27. Chen, L., Liu, P., Evans, T. C. Jr. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752–756 (2017).
28. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
29. Li, Y. et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience* **1**, 12 (2012).
30. Ho, J. R. et al. Deregulation of Rab and Rab effector genes in bladder cancer. *PLoS ONE* **7**, e39469 (2012).

Acknowledgements

This work was mainly supported by the training grant in Bioinformatics and Integrative Genomics from the National Human Genome Research Institute (grant no. T32HG002295 to C.L.B., A.R.B., L.J.L., and V.V.), a Brain Somatic Mosaicism Network grant from the National Institute of Mental Health (grant no. U01MH106883 to P.J.P., C.A.W.), and Ludwig Center at Harvard Medical School (P.J.P.). I.C.-C. received funding from the European Union (Marie Curie Skłodowska-Curie grant agreement no. 703543).

Author contributions

C.L.B. and M.A.L. conceived the project and P.J.P. supervised it. C.L.B. developed the algorithm. C.L.B., A.R.B., M.K., and A.G. generated the alignment and performed the variant calling. A.R.B., M.A.L., R.E.R., L.J.L., V.V., D.C.G., I.C.-C., M.A.S., M.K., M.E.C., and C.A.W. suggested impactful improvements to LiRA and aided in evaluating its performance. C.L.B. wrote the manuscript supervised by P.J.P., with input from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0366-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.J.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

Methods

Variant calling of candidate sSNVs and population-polymorphic gHets. The GATK Haplotype Caller best practice pipeline[22] with default parameters was used to call variants jointly on single-cell and bulk sequencing data from each individual. To maximize sensitivity, all variants reported in the output variant call format, regardless of the FILTER column flag, were considered. Candidate sSNVs were identified as calls with no alternate allele-supporting reads in bulk and at least one alternate allele-supporting read in a single cell, as specified in the variant call format. Polymorphic gHets were identified as variants found with non-zero population frequencies in the 1000 Genomes Project database[31] as annotated in the Single Nucleotide Polymorphism Database version 147 (http://www.ncbi.nlm.nih.gov/SNP), and called with a '0/1' heterozygous genotype in bulk.

Identification of candidate variants for LiRA analysis. sSNV–gHet and gHet–gHet pairs that had at least two reads or mate pairs supporting both variant loci underwent analysis by LiRA. Included reads were required to have maximum mapping quality score (60), to map in a proper pair (SAM flag 2), and to have no indel or base-clipping CIGAR operations.

Read-backed phasing of variant pairs. In LiRA, the relative phasing of two SNVs, that is, whether they are derived from the same (*cis*) or homologous (*trans*) chromosomes, defines the pattern of alternate (A) and/or *Ref* (R) allele support in discordant and concordant reads spanning the two loci. There are four possible patterns of support (SNV1/SNV2): R-R; A-R; R-A; A-A. If two SNVs (SNV1 and SNV2) are linked in *cis*, concordant reads for both SNVs will show A-A, and discordant reads R-A and A-R for SNV1 and SNV2, respectively. Alternatively, if two SNVs are linked in *trans*, concordant reads will show A-R and R-A, and discordant reads for both SNVs will show R-R.

Phasing in LiRA was done by simple majority of counted of reads. SNV1 was linked to SNV2 in *cis* if the count of A-A reads outnumbered A-R reads, and otherwise was linked in *trans*.

Filtering non-concordant sSNV candidates and computation of composite coverage. Following variant phasing, discordant and concordant read counts were obtained for each variant pair (sSNV–gHet and gHet–gHet) over single-cell data and bulk data. Any sSNV candidate or gHet not in a pair that showed complete concordance was filtered, and the composite coverage was computed over the remaining set.

Composite coverage was computed as follows: we first considered each discordant read-free sSNV–gHet pair, and measured the pairwise composite coverage as the minimum of the concordant read count in single cells and the unmutated haplotype in bulk (for sSNV–gHet pairs) or the concordant read count in bulk (for gHet–gHet pairs). The unmutated haplotype in bulk, for a *trans*-linked sSNV–gHet, was R-R, and for a *cis*-linked sSNV–gHet, R-A.

We then measured the composite coverage as the pairwise composite coverage observed for each sSNV, where a maximum was taken if an sSNV or gHet was linked with multiple gHets.

Power estimation. For the *Ref* and *Alt* allele of each gHet, we extracted the set of all supporting mate-pair reads from bulk and single-cell sequencing data. Then, we measured the minimum coverage between the bulk and single-cell supporting reads at all genomic positions covered by at least two reads in both sets. This gave the hypothetical composite coverage value an sSNV–gHet pair would have received had it occurred at one of the positions covered on the chromosomal haplotype corresponding to the gHet allele under consideration (either *Ref* or *Alt*).

Some positions were close enough to multiple gHets to receive more than one hypothetical composite coverage value on one or both haplotypes. In cases where the gHets themselves could be linked directly in the same reads, to compute an overall value for these sites we took the maximum composite coverage observed across all pairs. However, in cases where a position was close enough to two gHets to be linked with both, but the gHets were too far from each other to be covered by any spanning reads, it was unclear from read data alone which composite coverage values corresponded to the same haplotype. To resolve this, we used SHAPEIT version 2 (ref. [32]), with default parameters on bulk samples for each neuron donor, to determine the haplotype of the *Ref* and *Alt* allele of each gHet, and used this information to transform measurements specific to variant alleles (*Ref*/*Alt*) into measurements specific to haplotype.

Overall, this analysis yielded a map between the location of a hypothetical sSNV (genomic position and chromosomal copy) and the composite coverage with which it would have been detected.

Aggregate power calculation. To calculate the relationship between the estimated somatic mutation rate and composite coverage, we obtained the aggregated counts of the total number of loci at which a hypothetical sSNV could have been detected at particular composite coverage values ≥2 (P_c, aggregate power). These counts were adjusted to account for two confounding factors: (1) loss of power due to non-artifact-driven discordant read observations; and (2) loss of power due to the random occurrence of bulk-alternate reads supporting sSNV calls due to technical noise.

To account for the first factor, we reasoned that as composite coverage increases, so should the probability that a discordant read will be observed due to technical noise. This would reduce the power to detect sSNVs by some amount yet unaccounted for, since SNV pairs with discordant reads are excluded from LiRA in the first step. Our approach to this issue was to measure the rate at which gHet–gHet pairs were observed with discordant reads as a function of composite coverage (d_c) and to adjust the aggregate power at each composite coverage value down by the fraction of those we predicted to lose.

To account for the second factor, we predicted that as composite coverage in bulk sequencing increased, so would the probability that bulk reads would support an sSNV because of random sequencing error. We approximated this probability using half the rate at which a third allele is observed in bulk sequencing data at gHets (for example, a read supporting T at a C/G heterozygous site). We found that this quantity had coverage dependence; however, the relationship was complex and the rate did not consistently increase with coverage. Because of this, we used a fixed rate (b) computed and applied across all bulk coverage values to adjust P_c.

Overall, equation (1) describes the adjustment, which was completed over single cells individually:

$$P_c^{\text{adjusted}} = (1-b)\,(1-d_c)\,P_c^{\text{original}} \tag{1}$$

Rate calculation and two-component model. To obtain estimates of and bounds on the observed somatic mutation rate at different composite coverage values, we used a Beta distribution with Jeffreys prior: $\mathrm{B}(M_c+1/2,\, P_c - M_c+1/2)$, where M_c is the number of mutations with composite coverage c and P_c is the adjusted count of the number of loci with power to detect a mutation with composite coverage c. This gave the mutation rate in sSNVs per base pair (bp), and we converted this to sSNVs billion bp (Gbp) by multiplying by 10^9. For each cell, we modeled the expected value of this Beta distribution, the average somatic mutation rate measured at each value of composite coverage, $\mathrm{B}(M_c+1/2)/(P_c+1)$, as the mixture of an error (E) and a 'true mutation' (T) component. The error component we fitted had the form:

$$E(c) = Kp^{(c-2)}\,;\, p<1 \tag{2}$$

Visually, a decaying exponential appeared to fit the data well at low composite coverage values (Supplementary Fig. 5), consistent with a high burden of false positive calls at that level of quality. Theoretically, if we assume an initial burden of K errors at $c=2$, and that the probability of sampling a concordant read given a variant is truly discordant is p, then the error abundance as a function of composite coverage takes exactly this form.

We found that $p=1/2$ resulted in good fits; this suggested that the artifacts causing an excess of mutations at low composite coverage values originated from lesions present on the original DNA before any amplification. In this scenario, half of the reads from the linked germline haplotype (Fig. 1b) are expected to support the artifactual call.

The 'true' component $T(c)$ fit by the model was practically constant (Supplementary Fig. 5); to improve the quality of fitting, it was computed using a bootstrapped set of germline variants. The procedure used was as follows: (1) a set of germline variants of size equal to the size of the sSNV set for that cell ($c \geq 2$) from those found in DR-free gHet–gHet pairs was randomly selected, constraining the loci distance and orientation (*cis*/*trans*) distribution to be as close as possible to that observed in the somatic set. The later constraints were chosen because we reasoned that the distance between linked SNVs and the orientation (through alignment-mediated reference bias) almost certainly should affect composite coverage; (2) the rate using P_c and the composite coverage distribution for these sampled gHets was computed; (3) the bootstrap rate $B(c)$ (the 'true' component) was computed by averaging over 100 instances of (1–2).

Overall, the model $R(c)=E(c)+T(c)=K_1E(c)+K_2T(c)$ was fitted using the R function nlm.fit, where the $f(K_1,K_2)$, the square error from the Beta mean weighted by the inverse variance, $v_c=\mathrm{var}(B(M_c+½,\, P_c - M_c+1/2))^{-1}$, was minimized:

$$f\,(K_1,K_2) = \sum_c v_c \left(K_1E\,(c) + K_2T\,(c) - \frac{M_c+\frac{1}{2}}{P_c+1} \right)^2$$
$$v_c = \frac{(P_c+1)^2\,(P_c+2)}{\left(M_c+\frac{1}{2}\right)\left(P_c - M_c+\frac{1}{2}\right)} \tag{3}$$

K_1 and K_2 were constrained to be positive by imposing a large penalty on the objective function for K_1 or K_2 less than 0.

$T(c)$, the expected dependence between composite coverage and the observed sSNV rate for true heterozygous mutations, was used to estimate the genome-wide sSNV rate. In the ideal scenario, $T(c)$ should be constant, but we found that in some cells, especially at high composite coverage values, $T(c)$ was variable, often increasing dramatically in tandem with the observed sSNV rate (for example, UMB1465-18, UMB1465-47, UMB1465-51, and UMB4638-2; Supplementary Fig. 5). At these high composite coverage values, the number of gHet–gHet pairs used to construct $T(c)$, the number of sSNVs, and the aggregate power were very low; we attribute this phenomenon to noise introduced by these low counts in the power adjustment. Thus, to remove this effect, we used the value of the fit $T(c)$ curve at the lowest composite coverage value ($K_2T(2)$) to estimate the genome-wide rate.

To obtain bounds on the genome-wide rate, we sampled our Beta model of the observed genome-wide somatic mutation rate $\mathrm{B}(M_c + 1/2, P_c - M_c + 1/2)$ at each value of c and refitted our model on the result 100 times. The bounds we report on the genome-wide somatic mutation rate correspond to the minimum and maximum values of $K_2T(2)$ obtained over these samples; as such, they constitute a 98% confidence interval (CI).

Computation of estimated FDR and choosing a threshold (c*) for *c*. Given the model fit, the estimated FDR for the mutations detected was calculated as follows: (1) the FDR as a function of composite coverage was computed: $\mathrm{FDR}(c) = E(c)/(E(c) + T(c))$. This also allowed us to compute, for each value of c, the corresponding Phred quality score (Q) for somatic mutations with that level of support, $Q = -10\log_{10}(\mathrm{FDR}(c))$; (2) the estimated number of sSNVs ($M_c^{(t)}$) and false positives ($M_c^{(e)}$) as a function of c were computed: $M_c^{(t)} = ((1 - \mathrm{FDR}(c))M_c)$, $M_c^{(e)} = ((1 - \mathrm{FDR}(c))M_c)$; (3) the estimated aggregate FDR when thresholding at c_m was computed as:

$$\mathrm{FDR}_{\mathrm{agg}}(c_m) = \frac{\sum_{c \geq c_m} M_c^{(e)}}{\sum_{c \geq c_m} (M_c^{(e)} + M_c^{(t)})} \quad (4)$$

Choose c^* as the minimum value c_m so that $\mathrm{FDR}_{\mathrm{agg}}(c_m) \leq 0.1$.

Somatic variant calling with GATK Haplotype Caller, SCcaller, VarScan, Monovar, and MuTect. SCcaller[21], MuTect[26], GATK[22], VarScan[25], and Monovar[24] were used with default parameters to call somatic variants from single cells and bulk control data. For SCcaller, we filtered variants at three different (artifact likelihood)/(heterozygous variant) likelihood thresholds: 10^{-5}, 0.01, and 0.05; in all cases we filtered calls to include only sSNVs with an allelic fraction ≥ 1/8 (from the SCcaller github page; https://github.com/biosinodx/SCcaller). GATK does not report somatic calls outright but rather reports genotypes and quality metrics for single cells and bulk samples separately. As such, we used the following filters to call somatic variants: '0/1' or '1/1' genotype in a single cell; '0/0' bulk genotype; no supporting reads in bulk; and maximum bulk genotype quality (99). We also analyzed the set of variants meeting these criteria and annotated as 'PASS'. For Monovar, we filtered raw variants using the procedure described by Zafar et al.[24]. First, we removed candidate sSNVs with < 6× coverage or > 2 *Alt* reads in bulk. Next, we removed sites within 10 bp of each other that were only detected in single cells. Finally, we removed sSNV candidates with < 10× coverage and < 3 *Alt* reads, and those with a VAF < 10% or < 15% when coverage was between 20 and 100 or over 100, respectively.

Estimation of the abundance of unphaseable sSNVs. An sSNV is considered by LiRA only when there are bulk reads or mate pairs spanning the sSNV position and linked gHet allele. As such, it is not possible to directly count the total number of sSNV candidates LiRA would analyze if it had power over the entire genome. Instead, we estimate this number by dividing the number of LiRA sSNV candidates by the fraction of GATK somatic calls (PASS only) to which LiRA could be applied.

Analysis of false positive and LiRA call sSNV mutation type and trinucleotide context. False positives were limited to those also called as somatic by our filtering of GATK variant calls and annotated with 'PASS' in the variant call format filter column. This provided a stronger comparison between false positives and LiRA calls since it removed many low-quality sSNV candidates that had very different VAF, mutation type, and trinucleotide context distributions (data not shown). These excluded calls were likely enriched for errors stemming from sequencing errors rather than WGA.

Method for obtaining histogram error bars. The 99% CIs for the frequencies of SNVs for the VAF distribution bins (Fig. 2a–c and Supplementary Fig. 6), mutation type (Fig. 2c), and trinucleotide context (Supplementary Fig. 7) were computed using the multinomialCI function ($\alpha = 0.01$) from the MultinomialCI package in R on counts pooled across all cells.

Analysis of the Hoang et al.[23] data. The mutation frequencies for bottleneck sequencing of BRA04, BRA05, and BRA06 were obtained from Table S9 in Hoang et al.[23]. These values were converted into sSNVs/Gbp by multiplying by 10^9.

Analysis of the Li et al.[29] bladder cancer exome data. FASTQ files for 55 bladder cancer cells, 12 normal cells, bulk normal, and 2 bulk cancer samples were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and aligned to the Genome Reference Consortium Human Build 38 using BWA-MEM version 0.7.17-r1188 (ref. [33]). The GATK Haplotype Caller best practice pipeline[22] with default parameters was used to call variants jointly on these samples. LiRA was then used to call sSNVs in all single cells relative to the bulk normal sample.

Following analysis by LiRA, sSNVs called as passing or uncertain in any cell were grouped and the status of each sSNV in each cell was queried. For each sSNV, in cells where no reads covered the sSNV-linked haplotype, no genotype call was made. In cells where at least one read covered the sSNV-linked haplotype and this showed no evidence of the sSNV (the 'null haplotype'), sSNVs were called as absent. In cells with at least one supporting read for the sSNV (not necessarily spanning a nearby gHet), a call was made. Further, sSNVs were only considered if: (1) at least two cells showed only the null haplotype (at least two spanning reads); (2) at least two cells showed only the sSNV-positive haplotype (at least two spanning reads); and (3) the sSNV was called as passing or uncertain in more cells than it was called as a false positive. For a small number of cells, there were false positive calls among this set of considered sSNVs, and no genotype call was made in these cases.

To construct an expanded set of calls, a 'rescue matrix' of low-precision calls was created by considering all remaining sites reported by the GATK Haplotype Caller. Over each single cell, a call was assigned a value of 1 if it had an *Alt* depth of at least 1, or zero otherwise. We then performed pairwise Fisher's exact tests (of pairwise complete observations) between the cell calls for sSNVs in the rescue matrix and the cell calls for sSNVs in the set reported by LiRA, aiming to select a set of sSNVs from the rescue matrix that had unexpectedly high correlation in calls over cells with sSNVs reported by LiRA.

To account for multiple hypothesis testing, we first applied a procedure where we computed the minimum possible P value that could be obtained given the fixed marginals for each pairwise test. If this value was above the Bonferroni threshold at the 0.05 significance level for the number of tests being performed (n, threshold: $0.05/n$), it was excluded. Because this lowered the number of tests considered and thus raised the threshold, we repeated this process until a stable set of tests was obtained (that is, until all tests in principle could return a significant result).

We then obtained a set of 'rescued' sSNVs from the rescue matrix by considering those found to be unexpectedly associated with a LiRA sSNV, controlling FDR < 0.1 by the Benjamini–Hochberg procedure. To verify the validity of this, we performed the same procedure on 100 row-permuted rescue matrices and found that a non-zero number of sSNVs were rescued only rarely (3 out of 100) under random expectation. In contrast, we found that 57 were rescued over the real data. This set combined with the set of sSNVs (17) reported by LiRA are used in the heatmap presented in Supplementary Fig 8.

Statistics and reproducibility. We used Fisher's exact test to compare the distribution of mutation types between LiRA filtered false positives and LiRA calls.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability

LiRA is available at https://github.com/parklab/LiRA.

Data availability

LiRA was applied to single-neuron and bulk sequencing data collected from the postmortem brain, heart (UMB1465 and UMB4638), and liver (UMB4643) tissue of three individuals. These data were acquired as part of a previous study[5] and are available in the NCBI SRA under accession nos. SRP041470 (UMB1465) and SRP061939 (UMB4638 and UMB4643). The neuron counts by individual were: UMB1465 (16); UMB4638 (10); and UMB4643 (10).

References

31. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
32. Delaneau, O. et al. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat. Commun.* **5**, 3934 (2014).
33. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).

nature research

Corresponding author(s): Peter J. Park

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

We analyzed publicly available single-cell DNA sequencing data (from our previous paper, Lodato et al, Science, 2015). The sample size of our study was determined by the number of available cells (36 cells from 3 donors). This was sufficient to measure the performance of our method., in that it provided us with enough total sSNV candidates to demonstrate differences between sSNVs ultimately called by LiRA and those identified as false positives.

2. Data exclusions

Describe any data exclusions.

No data were excluded.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

Our findings are computational results. All of the bioinformatic pipelines used to generate these results have been run multiple times and verified to produce the same results.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not relevant to our study, as this was a methods development paper.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Not relevant to our study as there was no group allocation.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

LiRA is command line tool written in python and R, and makes use of several existing software tools for analyzing next-generation sequencing data.

The following software packages were used by LiRA in this study:

SHAPEIT2 v2 r900
samtools 1.2
bcftools 1.2
GATK HaplotypeCaller 2.8
bedtools v2.23.0

LiRA is available at https://github.com/parklab/LiRA.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eurkaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.