Supplementary Note

Accurate identification of single nucleotide variants in whole genome amplified single cells

Xiao Dong[1*], Lei Zhang[1*], Brandon Milholland[1*], Moonsook Lee[1], Alexander Y. Maslov[1], Tao Wang[2] and Jan Vijg[1,3]

[1]Department of Genetics, [2]Department of Epidemiology & Population Health,
[3]Department of Ophthalmology & Visual Sciences, Albert Einstein College of Medicine.

*These authors contributed equally to this work.
Correspondence should be addressed to Jan Vijg (jan.vijg@einstein.yu.edu).

Library preparation and whole-genome sequencing

For the 8 single cells and the kindred clone, 2 µg of DNA from each sample was sheared using Covaris sonication to an average size of 600 bp. Then we performed size-selection ranging from 500 bp to 800 bp on a 1% agarose gel. The size-selected DNA was purified using the MinElute Gel Extraction Kit (Qiagen). We transferred 300 ng of purified DNA to a 0.2-ml PCR tube. PCR-free libraries were constructed following manuals of Accel-NGS 2S DNA Library Kit (Swift Biosciences). Bead-based SPRI cleanups were used to remove oligonucleotides and small fragments after each step in the library construction. The resulting functional libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems). Paired-end sequencing (2×250bp) was performed on HiSeq 2500 sequencers (Illumina) at the Epigenomics Core Albert Einstein College of Medicine. For the bulk cell population, genomic DNA was extracted from the same human fibroblast population as the single cells using DNeasy Blood & Tissue Kit (Qiagen). One µg of purified genomic DNA was fragmented by Covaris sonication to a size of ~400 bp and a PCR-free whole-genome sequencing library was constructed using the KAPA LTP Library Preparation Kit (KAPA Biosystems) with Trueseq adapters (Illumina) at the Epigenomics Core of the Albert Einstein College of Medicine. Paired-end sequencing (2×100 bp) was performed on HiSeq 2500 sequencers (Illumina) at the Epigenomics Core of the Albert Einstein College of Medicine. For clones C1, C2 and C3, DNA was extracted from the clone cultures, each grown from a single cell. Library construction and sequencing were performed on the HiSeq X Ten at the New York Genome Center, New York, NY.

Quantitative PCR for estimation of amplification uniformity

Single cell amplicons, positive and negative controls, and empty Rafts were tested for uniform amplification at 8 separate loci[1,2] (Supplementary Table 2) using real-time PCR (RT-PCR). The 10-µl reaction included 2 ng DNA as template, 1 µM forward primer, 1 µM

reversed primer and 5 µl Fast Sybr Green master mix (Applied Biosystems). Relative abundance of each locus was calculated as,

$$Relative\ uniformity\ value\ (RUV) = 2^{-(Cti-Ct0)} \quad (1)$$

Where, Ct_i denotes the *Ct* value of the locus in sample *i*, and Ct_0 denotes the *Ct* value of the same locus in the unamplified genomic DNA. RUV of unamplified genomic DNA should be 1. An RUV closer to 1 indicates a more uniform amplification of the locus. Primers used for RT-PCR are listed in Supplementary Table 2. RUVs of each locus from single cell amplicons, positive and negative control, and empty Rafts (Methods) are listed in Supplementary Table 1. This test provides a quick and rough estimate of amplification uniformity across the genome. Currently, we consider RUVs of 6 out of the 8 loci in a range of 0.25-4 as satisfactory and 66% of amplicons from single cells are within this range (Supplementary Table 1).

Sanger sequencing

To confirm that somatic SNVs called by SCcaller were present in the single cell amplicons, but not in the bulk DNA, we performed Sanger sequencing on 16 randomly selected variants of the kindred group (Supplementary Table 5). This was done in the kindred cells and clone, as well as the bulk, even for SNVs identified in only one cell. All the somatic SNVs selected were confirmed by Sanger sequencing. Of note, the Sanger sequencing can only confirm whether the mutation is in the amplicon, but cannot rule out the possibility that the mutation is the result of an amplification error. We took advantage of the kindred design to distinguish real SNVs and amplification errors (Fig. 1 a, b).

Accessing sequencing uniformity

To assess sequencing uniformity of the bulk, as well as the clone and single cell samples amplified using SCMDA and HighTemp MDA, we split the genome into 1kb non-overlapping windows. For each window, we calculated the number of reads (mapQ≥20) aligned to it using bedtools (v2.24.0). We used the read counts per window to plot Lorenz curves of the read numbers using the R package “ineq” (Supplementary Fig. 4).

Using other variant callers

For Haplotypecaller, we used the options -stand_call_conf 30 –stand_emit_conf 10 as recommended by its developers.

For MuTect[3], (version 1.1.7), we input single cell or clone data as the tumor (--input_file:tumor) and bulk cell population data as the normal (--input_file:normal). The dbSNP (build 138) data was also used as input to MuTect (--dbsnp). Candidate somatic SNVs overlapping with dbSNP (build 144), located on sex chromosomes or with sequencing depth less than 20 were discarded.

For VarScan[4], (version 2.3.8), we first mpileuped single cells and bulk data using samtools. The somatic command for VarScan, with the default settings, was used for

calling somatic SNVs from each single cell or clone against the bulk with a minimum sequencing depth of 20. Candidate somatic SNVs reported by VarScan overlapping with dbSNP (build 144), located on sex chromosomes or with variant reads in the bulk were discarded.

For Monovar[5], we first mpileuped the single cell data using samtools (-BQ0 -d10000 -f referencegenome.fa -q 40, as recommended) and then called SNVs (-p 0.002 -a 0.2 -t 0.05, as recommended). Somatic SNVs were identified as not present in the bulk, not overlapping with dbSNP (build 144), located on autosomes, and with a sequencing depth equal or more than 20.

Calling somatic SNVs from clones

We used three software tools, MuTect[3], VarScan[4] and Unifiedgenotyper[6], as described above, to call somatic SNVs from the clones using the bulk cell population as control, and took the intersection of the three call sets; these we considered to be high-quality somatic SNV calls and we reported them in the main text (Fig. 3). This strategy of taking the intersection is a common practice, for example, in the 1000 Genomes Project[7].

Determining true mutations and amplification artifacts

For our kindred cells and clone, TPs and FPs were determined as follows. For each germline SNP detected by a variant caller in a kindred cell (Fig. 2a,b), we calculated the variant read counts in the kindred clone, the other kindred cell and the bulk (mapQ≥40). When the three numbers were all ≥ 1, we assigned the germline SNP call as a TP, and when all equal to 0, it was assigned as a FP. For each somatic SNV called by a variant caller in a kindred cell (Fig. 2d,e), we calculated the number of variant reads in the kindred clone (mapQ≥40). When at least 4 reads in the clone, a threshold based on commonly used criteria for calling mutations from a single dataset[4], we considered it to be a TP, and when there was 0 variant-supporting read in the clone, a FP. This assignment left some somatic SNVs with an unknown TP/FP status. We approximated the proportion of TPs and FPs with an unknown TP/FP status as that of the SNVs with clear TP/FP status described above.

For public single cell sequencing data, TPs and FPs were determined as follows. For germline SNP calling (Fig. 2c), SNPs presented in dbSNP and covered in bulk (minimum 4 variant supporting reads), were considered as germline SNPs. The number of false positives was approximated as the number of SNVs not present in dbSNP and not present in bulk (minimum 1 variant supporting read). This is because according to Hazen et al.[8] the real somatic SNVs in this cell type is about a hundred per cell and we detected 1,000 to 30,000 depending on the variant caller applied, so the vast majority of these variant calls are FPs. For somatic SNV calling (Fig. 2f), SNVs called by a variant caller in one kindred cell and also present in the other two kindred single cells (minimum 1 variant supporting read) in the MALBAC dataset[9], were considered as overlapping somatic SNV calls in all kindred cells.

Estimating the number of somatic SNVs per cell

To estimate the number of somatic SNVs per cell, we adjusted the raw number of somatic SNVs for sequencing depth and coverage. To do this, we divided the raw number of somatic SNVs by the number of base pairs in the genome with sequencing depth ≥ 20 in both a single cell and its bulk (Supplementary Table 3), and multiplied by the total number of bases in the genome.

Somatic SNV distribution across genomic features

Gene annotations were downloaded from the Ensembl Biomart (hg19). DNase I hypersensitive peaks and TF peaks were generated by ENCODE projects and downloaded from the UCSC genome browser. Repeat annotations were downloaded from the RepeatMasker website (http://www.repeatmasker.org). Germline variants were downloaded from the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

We used bedtools (v2.24.0) to annotate somatic SNVs on the above genomic features. Level of enrichment of SNVs on a genomic region was presented as the ratio between observed number of somatic SNVs and the expected number from genome-wide average after adjusting sequencing depth and coverage (Supplementary Fig. 10a).

Somatic SNVs in expressed genes

Transcriptome data were taken from the ENCODE projects under experiment IDs ENCFF640FPG and ENCFF704TVE (https://www.encodeproject.org/experiments/ENCSR797BPP/), which correspond to human dermal fibroblasts, the same cell type we used. The data include FPKM values for genes.

To test whether mutation frequency in exomes correlate with gene expression level we performed a permutation test as follows (Supplementary Fig. 10b). We sampled a random gene set with the same number of genes harboring somatic SNVs and calculated the average FPKM value for the random gene set. The sampling was repeated 2,000 times, and this generated 2,000 average FPKM values of random gene sets (the solid black line in Supplementary Fig. 10b). The average FPKM value of the mutant gene set was also calculated and compared to these 2,000 average FPKM values. The *P* value was estimated as,

$$P = \frac{N_{le} + 1}{N_{run} + 1} \quad (2)$$

where N_{le} denotes the number of random gene sets with a smaller or equal average FPKM compared to the mutant, and N_{run} denotes the number of permutations, i.e., 2000 in this test.

Supplementary Table 1. Quantitative PCR at 8 loci for SCMDA amplicons of single human dermal fibroblasts.

Sample	Type	Yield μg	Relative uniformity value (Supplementary Information)							
			site1	site2	site3	site4	site5	site6	site7	site8
h1	Single cell	14.4	1.11	0.14	1.01	0.38	0.19	1.31	0.28	0.50
h2	Single cell	15.2	1.02	0.25	0.09	0.53	0.70	0.60	0.15	0.38
h3	Single cell	17.2	1.68	0.21	1.17	0.06	0.28	0.76	0.49	0.66
h4	Single cell	17.5	0.69	0.11	0.73	0.12	0.34	0.72	0.57	0.52
h5	Single cell	14.6	10.04	0.04	0.01	0.32	0.02	0.02	0.01	0.03
h6	Single cell	16.4	0.84	0.36	0.91	0.18	0.33	0.99	0.27	0.63
h7	Single cell	16.4	0.53	0.16	0.98	0.28	0.29	0.31	0.16	0.34
h8	Single cell	17.0	1.05	0.14	1.20	0.39	0.47	0.88	0.35	0.58
h9	Single cell	15.2	1.31	0.25	1.14	0.33	0.17	0.75	0.22	0.63
h10	Single cell	15.6	0.95	0.23	0.90	0.54	0.51	0.60	0.36	1.19
h11	Single cell	16.4	1.09	0.22	1.81	0.41	0.47	1.16	0.39	0.77
h12	Single cell	16.2	0.50	0.17	0.25	0.33	0.24	0.52	0.16	0.85
h13	Single cell	14.8	1.08	0.33	1.24	0.32	0.51	0.58	0.40	0.60
h14	Single cell	16.7	0.62	0.21	0.57	0.18	0.25	0.89	0.28	0.60
h15	Single cell	14.8	1.59	0.33	1.50	0.21	0.68	1.24	0.32	1.03
h16	Single cell	16.4	0.41	0.21	0.00	0.37	0.01	0.41	0.17	0.32
h17	Single cell	13.8	0.73	0.15	1.47	0.38	0.51	0.61	0.25	0.27
h18	Single cell	13.4	0.28	0.06	0.40	0.21	0.43	0.63	0.21	0.21
h19	Single cell	15.5	0.28	0.06	8.94	0.13	0.01	0.02	0.00	0.09
h20	Single cell	12.7	1.89	0.24	0.57	0.34	0.80	0.88	0.50	0.93
h21	Single cell	13.0	0.78	0.08	0.01	0.45	0.18	1.24	0.02	0.19
h22	Single cell	12.7	1.43	0.11	0.51	0.78	0.68	0.36	0.13	0.54
h23	Single cell	13.1	1.41	0.39	0.88	0.62	0.68	1.21	0.36	0.90
h24	Single cell	12.8	1.28	0.87	0.75	0.96	0.40	3.53	0.32	0.84
h25	Single cell	12.2	0.99	0.20	0.74	0.24	0.39	0.31	0.05	0.41
h26	Single cell	13.2	0.02	0.33	1.58	0.69	0.52	0.76	0.27	1.41
h27	Single cell	12.8	1.44	0.05	0.00	0.57	0.01	0.01	0.49	1.21
h28	Single cell	14.4	0.48	0.27	0.90	0.09	0.15	0.57	0.19	0.81
h29	Single cell	13.9	0.81	0.39	0.58	0.51	0.42	1.17	0.46	0.77
h30	Single cell	14.2	0.95	0.22	0.97	0.71	0.24	1.64	0.35	0.66
h31	Single cell	14.4	0.65	0.15	1.78	0.42	0.44	0.78	0.37	1.04
h32	Single cell	13.6	0.93	0.36	0.42	0.30	0.45	0.91	0.19	0.96
h33	Single cell	14.3	0.75	0.20	1.02	0.32	0.48	0.61	0.47	1.02
h34	Single cell	13.7	0.58	0.33	1.90	0.45	0.39	2.27	0.36	0.44
h35	Single cell	14.3	2.24	0.33	5.16	0.44	0.40	1.69	0.51	1.20
h36	Single cell	14.2	0.98	0.08	0.02	0.01	0.24	0.91	0.76	0.22
h37	Single cell	13.7	1.35	0.31	1.75	0.67	0.33	1.11	0.36	0.42
h38	Single cell	12.4	1.10	0.04	0.00	0.28	0.00	4.01	0.22	0.34
h39	Single cell	12.1	0.47	0.53	0.00	0.09	0.05	0.23	0.05	0.11
h40	Single cell	12.2	0.81	0.25	0.19	0.51	0.16	0.21	0.35	0.52
h41	Single cell	12.2	0.66	0.09	2.16	0.47	0.64	0.83	0.13	0.50
h42	Single cell	11.9	0.39	0.40	0.09	0.14	0.86	0.23	0.19	0.59
h43	Single cell	12.1	0.45	0.27	0.00	0.95	0.50	0.45	0.11	0.03
h44	Single cell	11.5	0.70	0.06	0.75	1.20	0.48	1.52	0.06	0.34
empty raft1	Empty raft*	1.0	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.00
empty raft2	Empty raft	1.1	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.00
empty raft3	Empty raft	0.9	0.00	0.00	0.01	0.00	0.01	0.04	0.00	0.00
empty raft4	Empty raft	1.0	0.00	0.00	0.01	0.00	0.01	0.04	0.00	0.00
pc1	Positive control**	12.6	1.70	0.52	0.93	0.61	0.48	1.25	0.65	0.66
pc2	Positive control	11.5	1.16	0.27	1.04	0.71	0.40	0.95	0.31	0.43
nc1	Negative control***	0.6	0.00	0.00	0.01	0.00	0.02	0.07	0.00	0.00

* Empty rafts had no cells on it.

** 1 ng of human genomic DNA in 2.5 μl PBS was amplified as a positive control.

*** 2.5 μl of PBS without any template was amplified as a negative control.

Supplementary Table 2. Primers for qPCR of the 8 loci.

Species	Chr	Forward	Reverse
human	1p	TTTGATGGAGAAATCCGAGG	CTGACTCGGAGAGCAGGAC
human	1q	GGTAGGATGATTCTAGAATGCCA	GCCCAAATTGGCTTCTTTTT
human	4p	AACTGAATGGCAGTGAAAACA	CCCTAGCCTGTCATTGCTG
human	4q	TATAGCCCACCTGACCCAAG	CTGTCATCACTGTCTACTTCCTCTC
human	10p	GTTCTGCTGCCTCTACACAGG	ATCCTTCTGTGAACTCTCAAATCC
human	10q	CTTCCTGACCTGTTTGCAGT	CTTCAGTGCACAGAATGCAG
human	12p	CCACACACTCTGGTTTTATAAAGC	TTTTTCTCCTGCATCCATGG
human	12q	TCCTCATTGTTGGGGATGAT	TGGCCAAAAATAGAAGCCAT

Supplementary Table 3. Coverage statistics and sequencing depth (mapQ≥20).

Sample id	Type	Protocol	Average depth	% genome			
				≥ 20x	≥ 10x	≥ 5x	at 0x
IL11	kindred cell	SCMDA	26.5	54.9%	78.9%	87.3%	9.3%
IL12	kindred cell	SCMDA	26.9	49.3%	70.0%	80.9%	11.0%
IL1C	kindred clone	-	31.6	82.6%	90.1%	90.9%	8.7%
IL2	single cell	SCMDA	26.3	57.7%	82.3%	88.8%	9.1%
IL3	single cell	SCMDA	27.1	52.8%	72.9%	82.3%	11.2%
IL4	single cell	SCMDA	17.9	35.5%	68.8%	84.1%	9.6%
IL5	single cell	SCMDA	17.8	34.7%	68.5%	84.3%	9.5%
HL1	single cell	HighTemp MDA	42.8	65.1%	79.2%	85.6%	10.1%
HL2	single cell	HighTemp MDA	40.9	60.6%	76.0%	83.9%	10.4%
clone1	single cell clone	-	29.5	82.1%	89.6%	90.5%	9.0%
clone2	single cell clone	-	33.6	84.9%	90.0%	90.6%	9.0%
clone3	single cell clone	-	27.6	80.0%	89.3%	90.5%	9.0%
Bulk control	bulk	-	43.9	88.1%	90.0%	90.5%	9.1%

Supplementary Table 4. *Pearson* correlation coefficients of observed and estimated major allele fractions* in cell IL11 and IL12.

Sample ID	λ** = 2000		λ = 10000 (SCcaller default)	
	10 fold cross validation	leave-100-out cross validation	10 fold cross validation	leave-100-out cross validation
IL11	0.5	0.55	0.49	0.53
IL12	0.8	0.84	0.76	0.79

* The data were estimated on a randomly selected region, chr1:100,000,000-110,000,000.

** λ denotes half of window width for smoothing (Methods).

Supplementary Table 5. Sanger sequencing validation of randomly selected somatic SNVs.

Chr	Position	Reference	Genotype in bulk cell population	Single cell genotype from Sanger	Called by SCcaller		Sanger sequencing		
					IL11, kindred cell	IL12, kindred cell	IL11, kindred cell	IL12, kindred cell	IL1C, kindred clone
6	124299091	T	T/T	T/C	TRUE	-	TRUE	TRUE	TRUE
9	5732840	G	G/G	G/T	TRUE	TRUE	TRUE	TRUE	TRUE
10	8488243	T	T/T	T/C	TRUE	TRUE	TRUE	TRUE	TRUE
9	4461559	C	C/C	C/A	TRUE	-	TRUE	TRUE	TRUE
1	152685505	C	C/C	C/T	TRUE	TRUE	TRUE	TRUE	TRUE
12	14908833	A	A/A	A/G	TRUE	TRUE	TRUE	TRUE	TRUE
2	142675185	T	T/T	T/C	TRUE	-	TRUE	TRUE	TRUE
4	132380455	T	T/T	T/C	-	TRUE	TRUE	TRUE	TRUE
8	112315076	A	A/A	A/C	-	TRUE	TRUE	TRUE	TRUE
5	158512199	G	G/G	G/A	-	TRUE	TRUE	TRUE	TRUE
1	77255928	G	G/G	G/A	-	TRUE	TRUE	TRUE	FALSE
2	115020724	A	A/A	A/G	TRUE	TRUE	TRUE	TRUE	FALSE
6	136645743	T	T/T	T/G	TRUE	-	TRUE	TRUE	FALSE
7	76872746	C	C/C	C/A	TRUE	TRUE	TRUE	TRUE	FALSE
20	60057506	G	G/G	G/T	TRUE	-	TRUE	TRUE	FALSE
21	28506056	C	C/C	C/A	TRUE	-	TRUE	TRUE	FALSE

References

1 Gundry, M., Li, W., Maqbool, S. B. & Vijg, J. Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic acids research* **40**, 2032-2040, doi:10.1093/nar/gkr949 (2012).

2 Hosono, S. *et al.* Unbiased whole-genome amplification directly from clinical samples. *Genome research* **13**, 954-964, doi:10.1101/gr.816903 (2003).

3 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

4 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

5 Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant detection in single cells. *Nature methods*, doi:10.1038/nmeth.3835 (2016).

6 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

7 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

8 Hazen, J. L. *et al.* The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. *Neuron* **89**, 1223-1236, doi:10.1016/j.neuron.2016.02.004 (2016).

9 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 (New York, N.Y.)* **338**, 1622-1626, doi:10.1126/science.1229164 (2012).