www.sciencemag.org/cgi/content/full/338/6114/1622 /DC1

Supplementary Materials for

Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell

Chenghang Zong, Sijia Lu, Alec R. Chapman, X. Sunney Xie

*To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu

Published 21 December 2012, *Science* **338**, 1622 (2012)
DOI: 10.1126/science.1229164

This PDF file includes:

Supplementary Online Material

"Genome-Wide Detection of Single Nucleotide and Copy Number Variations of a Single Human Cell"

Chenghang Zong[1,†], Sijia Lu[1,†#], Alec R. Chapman[1,2,†], X. Sunney Xie[1,*]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
[2]Program in Biophysics, Harvard University, Cambridge, MA 02138, USA

[†]These authors contributed equally to the work.

[#]Current address: Yikon Genomics Inc., 1 China Medical City Ave, TQB building 5th floor, Taizhou, Jiangsu, China

[*]To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu

Sample preparation before whole genome amplification

The SW480 colorectal adenocarcinoma cell line is obtained from American Type Culture Collection (ATCC, Rockville). SW480 cells are maintained in ATCC-formulated Leibovitz's L-15 Medium supplemented with 10% fetal bovine serum (ATCC), 100 I.U./ml Penicillin and 100 μg/ml Streptomycin (ATCC). The cells are treated with 0.25% Trypsin-EDTA, followed by washing and dilution in PBS. Single cells are mouth pipetted into individual PCR tubes. After briefly spinning down the single cells to the bottom of PCR tubes, 5μl of freshly prepared cell lysis buffer containing QIAGEN protease is prepared according to manufacturer's specifications and added into each tube. The lysis of the single cell is performed by the following temperature steps: 50°C 3 hours, 75°C 20 minutes, 80°C 5 minutes. Isolation of single cells can also be performed with other techniques such as laser dissection, microfluidic devices, or flow cytometry. Avoiding DNA contamination from environment and operators is critical for single cell SNV analysis.

Single cell whole genome pre-amplification with Multiple Annealing and Looping Based Amplification Cycles (MALBAC)

The amplification procedure starts with a pre-amplification step, in which multiple annealing and looping based amplification cycles (MALBAC) is used to create overlapped shotgun amplicons covering most of genome. In the second step, regular PCR is used to further amplify the amplicons for next generation sequencing. The common 27-nucleotide sequence in MALBAC primer is GTG AGT GAT GGT TGA GGT AGT GTG GAG.

For MALBAC, 30ul of amplification buffer (20mM Tris-Cl (pH 8.8), 10mM $(NH_4)_2SO_4$, 10mM KCl, 3mM MgSO4, 0.1% Triton X-100, 0.32uM primers) is added into PCR tubes each containing a lysed single cell, followed by incubating the tube at 94°C for 3 mins to melt the double-stranded genome DNA into single stranded form. After melting, the single-stranded genomic DNA molecules are immediately quenched on ice to increase the efficiency of primer binding. 2.5 Units of Bst large fragment (NEB), or 2 Units of Bst large fragment supplemented with 0.8 Units of Pyrophage 3173 exo- (Lucigen) are then added into each PCR tubes and the following temperature steps are performed: 10°C—45secs, 20°C—45secs, 30°C—45secs, 40°C—45secs, 50°C—45secs, 65°C—2mins, 94°C—20s; The tubes are then quickly quenched on ice.

After quenching on ice, the same polymerase mix is added to provide enzyme for the next round of amplification. The following thermo-cycle is performed before quenching the reactions on ice: 10°C—45secs, 20°C—45secs, 30°C—45secs, 40°C—45secs, 50°C—45secs, 65°C—2mins, 94°C—20secs, 58°C—20secs. The above cycles are repeated 5 times.

Further Amplification by regular PCR for next generation sequencing

We further amplified the product from the MALBAC pre-amplification by regular PCR using the following thermal cycle repeated 18 times: 94°C—20secs, 59°C—20secs, 65°C—1min, 72°C—2mins. It generates about 3μg of double-strand DNA products from a single cell for preparing sequencing libraries for next-generation sequencing.

The amplification products have a size distribution of ~500bp to 1500bp. Quantitative PCR is performed to check for the amplification uniformity for 16 different loci, each on a different chromosome (Table S3). Out of the 16 randomly selected loci, 14 of them are amplified with reasonable Ct number in qPCR, which is consistent with the ~90% coverage of the whole genome at 30x sequencing depth of single cells (Table S4).

The amplified DNA products are then used for preparing sequencing libraries for Illumina and SOLiD sequencing platforms. The DNA product from MALBAC amplification can be directly used in constructing the sequencing library for both Illumina and SOLiD with standard procedures. For Illumina sequencing, ~3μg DNA material are provided to the vendor for standard library preparation and sequencing. For SOLiD sequencing, we performed the library preparation and sequencing following the standard protocol of SOLiD 4 system. 3μg DNA material is used as the starting material. We used EZ bead system for emulsion PCR and enrichment. The platform for Illumina sequencing is Hiseq-2000. All sequencing experiments are listed in Table S1.

GC bias and correlated bias in MALBAC amplification

GC bias in MALBAC amplification is given in Fig. S1. The amplification products have an average of 46% GC content, slightly higher than the 42% of human genome, which indicates MALBAC slightly favors GC over AT. The correlation of the amplification efficiency between two cells is shown at different binning window in Fig. S2. The low correlation coefficient of 0.31 at 1-kilobase window shows that MALBAC does not have significant local bias. The correlation coefficient of 0.79 at a 200-kilobase window indicates the existence of correlated variations, though the amplitude of variation is within 2 fold.

Analysis of single nucleotide variations of SW480 with bulk and single cell sequencing data

The raw sequencing data generated by Illumina is mapped by BWA (*33*). After mapping, the duplicate reads are removed. The SNVs are called using GATK (*34*) to generate the raw SNVs. By comparing to the bulk data, we identify the common SNVs and use these SNVs as training set to train a Gaussian mixture model using GATK. The SNVs are categorized into different tranches based on the log odds ratio of being a true positive vs. false positive using the trained Gaussian model (*34*).

We identified 2,842,162 SNVs by bulk sequencing of single cell derived clone. Detailed analysis using SNPEff (*35*) is attached at the end of SOM.

For two-cell analysis, we first use GATK to call raw SNVs. The raw SNVs shared by two cells are collected as SNVs for downstream filtering. We applied the same GATK analysis to the raw SNVs shared by two cells. After training the Gaussian mixture model, we recalibrate the variants. The SNVs with the best true variant versus false variant ratio are collected as shown in Table 2 (Maintext).

For three-cell analysis, we use GATK to call raw SNVs for each single cell. Following that, we apply a pairwise T-test of forward and reverse reads to filter out common systematic errors (*36*). The raw SNVs is analyzed by following steps: Step1: the SNVs shared by at least two cells are collected. This list of SNVs are further filtered if the third cell has > 5x reads but without sequencing this SNV once. The filtering is based on one-sided binomial test. The number of heterozygous and homozygous SNVs is given in Table 2 (Maintext). Step 2: we filtered out the SNVs that appear in the bulk with at least two reads, which will exclude SNVs called by bulk sequencing as well as the false negatives missed by bulk SNV calling. Using the excluded SNVs, we determined our SNV detection efficiency (Table 2, Maintext). Step 3: In order to identify and remove the systematic errors due to amplification as well as sequencing, we sequenced two unrelated cells. We filtered out the SNVs that also appear in the unrelated cells for at least once. This filtering procedure efficiently removes systematic amplification and sequencing errors occurred in MALBAC. Step 4: the mapping scores for the reads at the position of a SNV and close to the position were used to eliminate SNVs from the poorly mapped regions; we further eliminated those from homopolymer, tandem repeat, significantly GC-biased regions or the regions with very high local density of mutations. 12 false positives identified in Step 4 were randomly selected and proven not being false negatives by Sanger sequencing of C4, C5 and C6. The above filtering procedure is given in Figure 4B. The false SNVs with reads in only once cell locate on the x-, y- or z- axis. The false SNVs with reads in two cells and not in the third cells locate on xy-, xz- and yz-planes. The SNVs with p-value > 0.1 for all three cells are potentially true SNVs. To avoid the crowdedness in Figure 4B, we did not show the SNVs detected in the bulk (identified in Step 2) and the false positives filtered out by comparing to two other unrelated cells in Step 3. The residual false SNVs identified in Step 4 are plotted as green dots and the final 35 newly acquired SNVs are plotted as red dots in Figure 4B. Kindred cells C4, C5 and C6 were amplified by MALBAC. The amplified DNA was used for PCR amplification of eight randomly selected from the 35 newly acquired SNVs. Sanger sequencing of these PCR products confirmed that these SNVs are true positives. We also Sanger sequenced the bulk DNA to confirm that these SNVs were absent in the bulk, hence they are not false negatives in SNV calling based on next-generation sequencing of the bulk. Sanger sequencing trace files are attached.

Determining copy number variations

A two-step process was used to determine copy number variations in single cells from coverage data in 200kb bins. In order to determine an appropriate normalization factor for our coverage, we first use a hidden Markov model (HMM) to determine likely diploid regions by comparing the coverage normalized by total reads to a MALBAC amplified normal blood cell. The coverage is then normalized by the amplification bias after the smoothing with a cubit spline. The emissions are a binary vector indicating whether the cancer single cell had higher coverage than the normal cell. The model contains three states corresponding to one, two, or three copies.

The model was parameterized by two constants $a = 10^{-2}$ and $b = 10^{-1}$ corresponding to the rate per bin in which abnormal copy numbers are expected to begin and end, respectively. The diploid probability is then $p_d = b/(a + b)$. We take the prior probabilities to be:
$[(1-p_d)/2, p_d, (1-p_d)/2]$.
The transition matrix is taken to be:

$$\begin{pmatrix} 1-b & a/2 & 0 \\ b & 1-a & b \\ 0 & a/2 & 1-b \end{pmatrix}$$

For a state with copy number s, the probability observing more reads in the cancer cell is calculated as $p(X_1*s/2 > X_0)$, where X_1 and X_0 are random variables distributed according the observed distribution of coverage in the normal cell. The Viterbi algorithm is used to calculate a most likely state path. This provided a normalization factor for the coverage so that regions determined to be diploid would have a mean coverage of 2.

The second step aimed to more precisely identify variations over a wider range of copy numbers. Six hidden states are allowed, corresponding to 0 to N = 5 copies, and the emissions are the normalized coverage.

The emission probabilities for the two copy state are assumed to follow the distribution observed in the diploid regions determined in the previous step, denoted by p_2. The emission probabilities p_1 for the single copy state should satisfy $p_2 = p_1*p_1$, where * denotes convolution. This relationship is readily inverted by taking the square root in

Fourier space. Once p_1 is obtained, emission probabilities for higher copy states can be determined from the relation $p_n=p_{n-1}*p_{1.}$

A transition matrix similar to the previous step is used in which all transitions to an abnormal copy number state had probability a/N, and transitions from an abnormal state to the diploid state had probability b.

The Viterbi algorithm is again used to determine the most likely state path. The regions determined to be diploid could then be used to provide an updated normalization factor used to repeat this step. A small number of iterations ($\lesssim 5$) generally provide sufficient convergence.

In addition to the CNVs of three cells shown in Figure 3 of the maintext, those of five more cells are shown in Figure S4.

MDA amplification and read histogram

The MDA amplification is performed with both GE Genomiphi kit and Qiagen Repli-g kit. Q-PCR results with 16 pairs of random primers show that their performances for single cells from the same cell line are similar. Fig. S3 shows the sequencing result of one of those MDA amplifications of single SW480 cells. The observation of large fluctuations in the distribution of reads indicates that extensive bias is present in MDA amplification. This sample was amplified by GE kit.

Calculation of allele dropout rate

Here we denote the allele dropout rate as α and N as the number of the heterozygous SNV positions that have enough reads covered for SNV analysis. The number N is related to the coverage of amplification methods.

For MDA sequencing data, we found 172,565 SNVs that are called homozygous SNVs based on single cell data, but are heterozygous SNVs based on bulk data. This indicates allele dropout of the reference allele. Here we assume the cell is diploid for simple formulation, and then we have $N\alpha(1-\alpha)=172{,}565$, where α is allele dropout probability and *N* is the number of heterozygous SNVs whose loci are covered by reads. In MDA, we also called 93,140 SNVs as heterozygous, which follows $N(1-\alpha)(1-\alpha)=93{,}140$. With the two above equation, we estimate the allele dropout α of MDA is 0.65. Similarly, we estimated the allele dropout rate for MALBAC is ~1%.

This efficiency of amplifying both alleles is contributed by the multiple annealing and amplification cycles in MALBAC.

Examination of mutation rate based on bulk sequencing

To verify the mutation rate measured by the single cell sequencing data, we sequenced the bulk from the original SW480 culture with 30x sequencing depth (Fig. S5). By comparing with the single-cell derived bulk, we found 754 SNVs that exist in single cell derived bulk but not in the heterogeneous bulk. Previous records show that the SW480 cell line has propagated for at least 150 generations. Therefore the mutation rate is at maximum ~5 mutations per cell division, which agrees well with our measurement based on single cell sequencing data. This estimation of the upper bound of the mutation rate includes possible selection occurring during the culturing propagation.

Percentage of genome coverage as a function of sequencing depth

For bulk, MALBAC and MDA sequencing data, the percentage of genome coverage for lower depth can be extrapolated from the coverage at 25x mean sequencing depth by constructing a probability transition matrix. In Figure S6, the fraction of the genome with the ≥ 1x coverage (blue) and ≥ 6x coverage (red) is plotted as the function of the mean sequencing depth. This figure can be used to estimate the sequencing cost as the function of coverage gain.

Anomalous transition transversion ratio

We observe n_{ts}= 8 transitions. We regard the number of transitions as a Poisson random variable with parameter $\mu_{ts.}$ The posterior probability for μ_{ts} is given by:

$$p(\mu_{ts}|n_{ts}) = p(n_{ts}|\mu_{ts})p(\mu_{ts})/p(n_{ts})$$

Assuming a uniform prior for μ_{ts}, we obtain:

$$p(\mu_{ts}|n_{ts}) \sim {\mu_{ts}}^{n_{ts}} e^{-\mu_{ts}}$$

We observe n_{tv} = 27 transversions and obtain a similar expression for the posterior distribution of μ_{tv}. By random sampling from the distributions, we obtain the posterior distribution for the ratio of transitions to transversions shown in Figure S7. The 95% confidence interval is (0.14, 0.63). The tstv ratio of 2.01 observed for all mutations in the bulk sample is well outside of this interval.

A low tstv ratio has been observed in ErBb2+ cells (*37*). However, as SW480 does not exhibit known ErBr2+ mutations, the observed value of 0.3 for these new SNVs indicates that other mechanisms may dominate the mutation spectrum. Here we speculate that during rapid growth of the cell line, replication errors can be the dominant source of mutations rather than DNA damage by which transition mutations are favored. This mechanism can lead to a much lower tstv ratio.

The raw sequencing data and Sanger sequencing data are available online.

FIGURES:

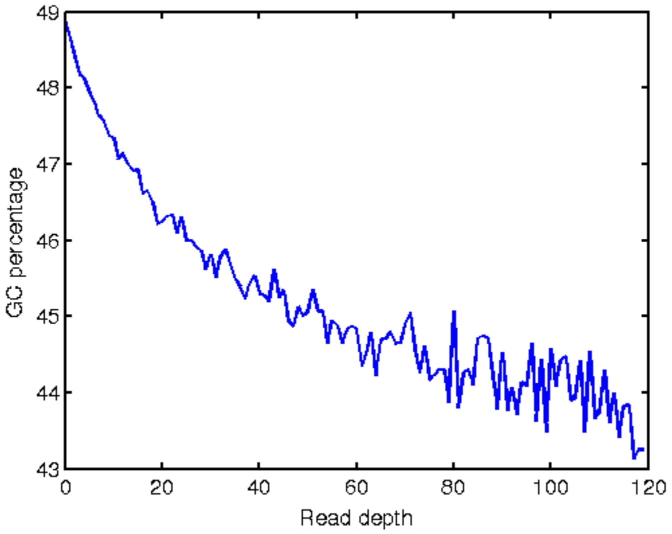


Figure S1: GC percentage for the positions with different read depth.

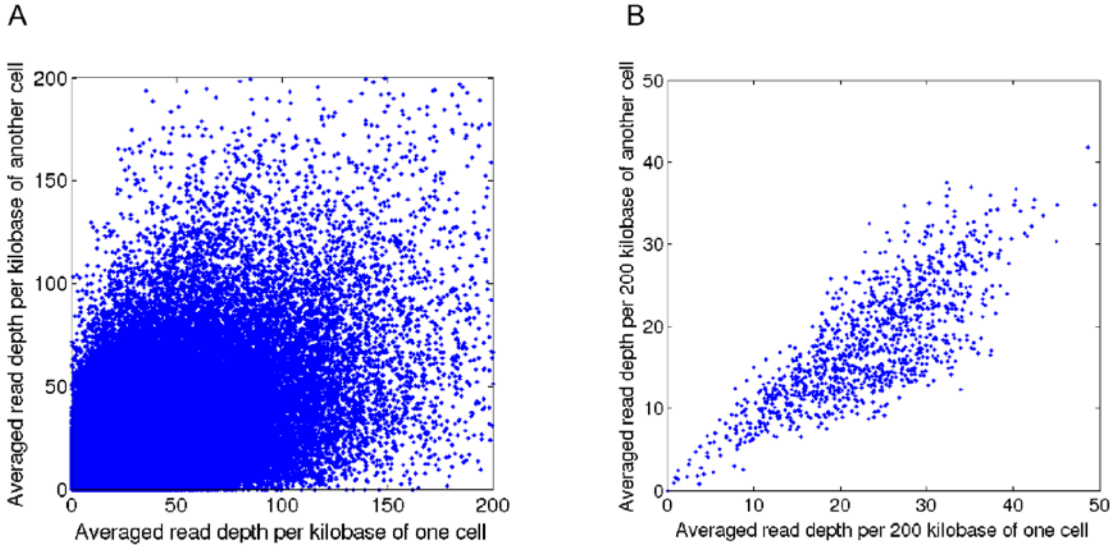


Figure S2: Correlation plot of read density at different binning windows between two cells. (A) The binning window is 1 kilobase. The correlation coefficient is 0.31. (B) The binning window is 200 kilobases. At this scale, the correlation coefficient is 0.79, which indicates correlated variation in coverage. We also note that majority of points with averaged depth < 10x is due to chromosome loss in this cancer cell line and the majority of the data distribute between 10 and 50, indicating that the correlated variations in coverage are small. The data are calculated only using chromosome 1.

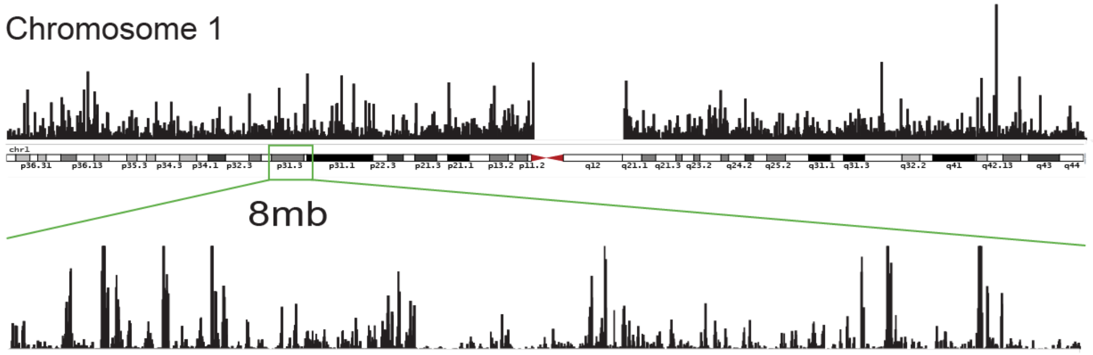


Figure S3: The histogram of reads (chromosome 1) of MDA amplification of a single SW480 cell. The mean sequencing depth is 25x.

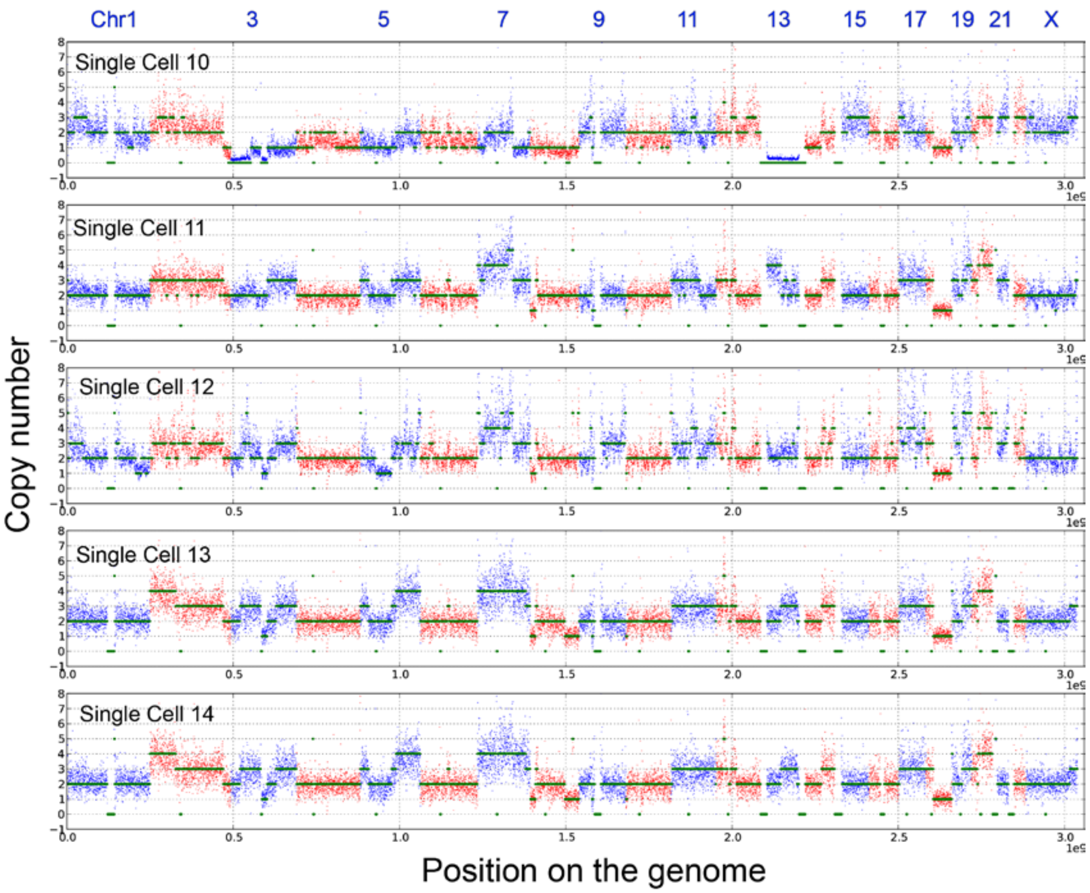


Figure S4: CNVs of the five additional single SW480 cells. Single cell 10 exhibits significant loss of chromosome 3 and entire 13. This observation indicates the unstable nature of chromosomal instability, which could lead to the cell death.

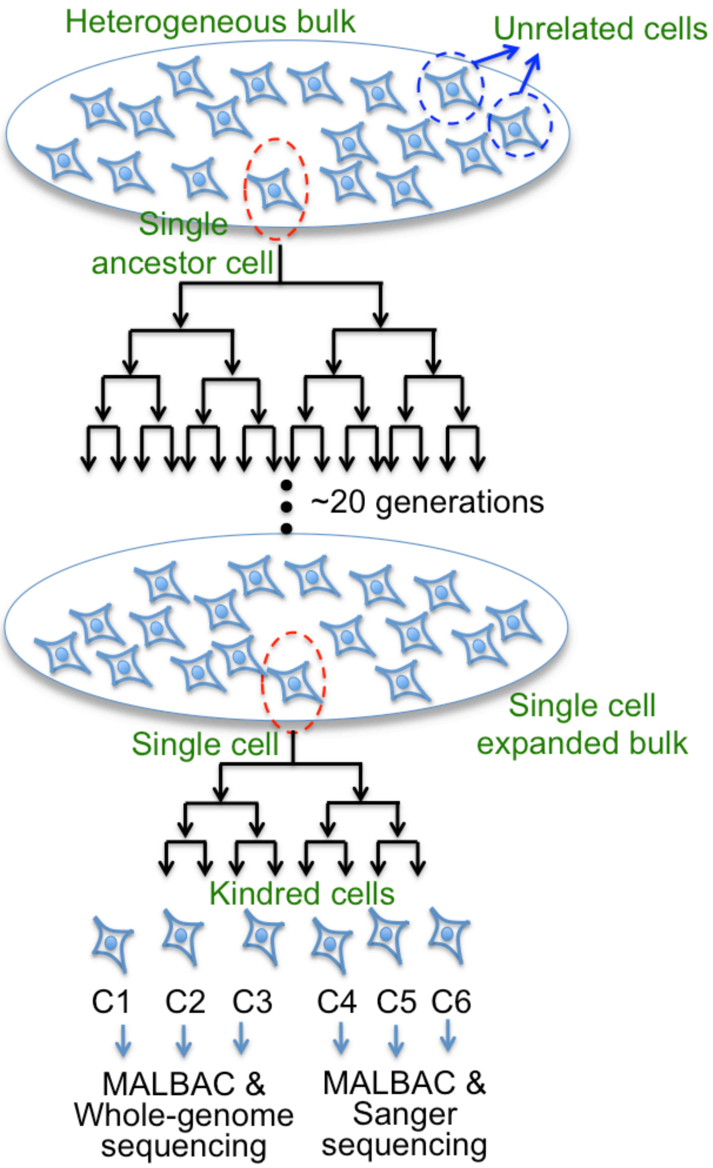


Figure S5: The experimental scheme that includes the original heterogeneous bulk and two unrelated cells. The heterogeneous bulk is sequenced and compared with single cell expanded bulk. The two unrelated cells were amplified with MALBAC and used for identifying the systematic errors.

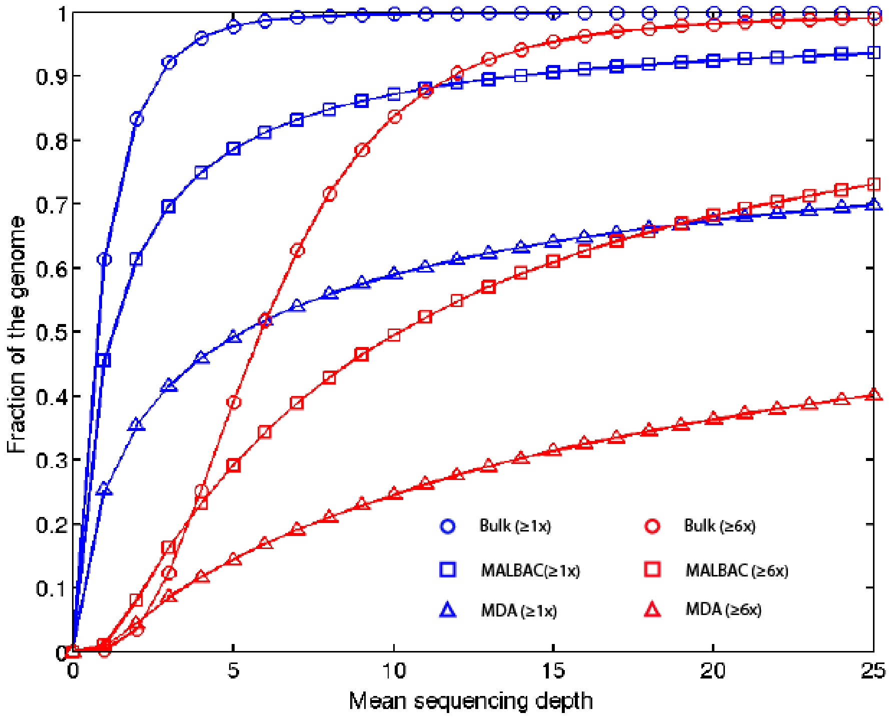


Figure S6: The fraction of the genome with the ≥ 1x coverage (blue) and ≥ 6x coverage (red) vs. the mean sequencing depth. At 25x mean sequencing depth, ~74% of the genome is covered with ≥ 6x reads for MALBAC, which corresponds to ~76% SNV detection efficiency as shown in Table 1; ~40% of the genome is covered with ≥ 6x reads for MDA, which corresponds to ~40% SNV detection efficiency. At 25x mean sequencing depth, the fraction of the genome with ≥ 1x coverage approaches to a plateau for both MALBAC and MDA, which indicates the further increase of mean sequencing depth will not significantly improve the genome coverage.

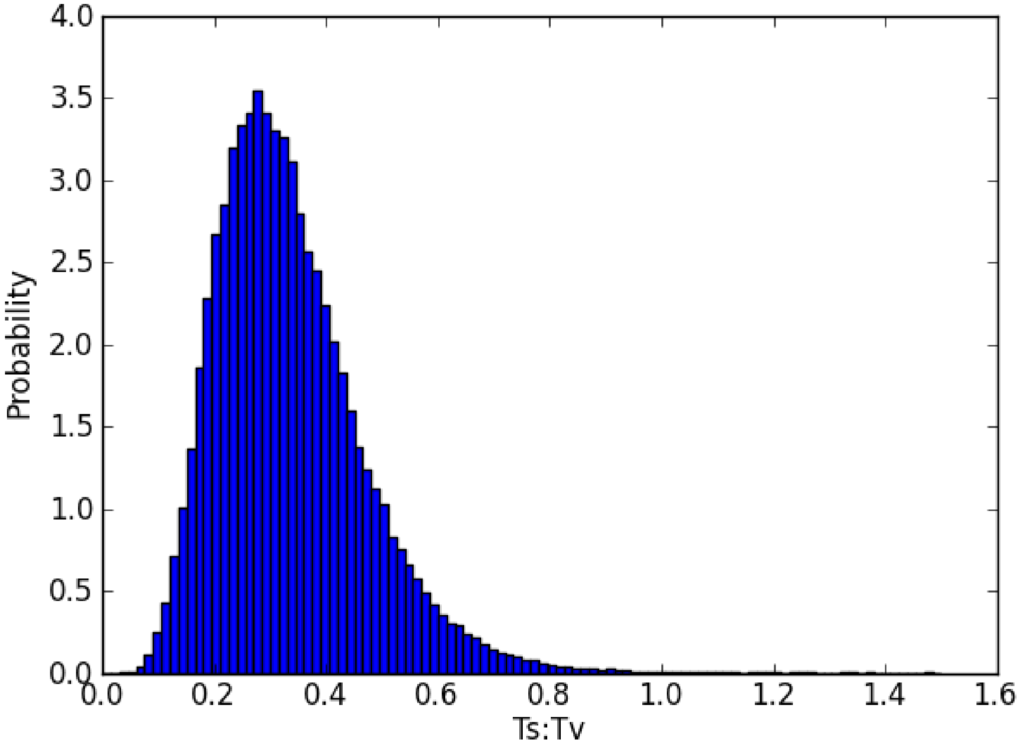


Figure S7: Posterior probability distribution of the tstv ratio. Given that we observe 8 transitions and 27 transversions, there is a 95% probability that the true tstv for our cell line lies between 0.14 and 0.63.

TABLES

Sample index	Amplification method	Averaged coverage depth	Percentage Genome coverage	Notes*
SW480 bulk (Heterogeneous culture)	None	33.7x	92%**	I
SW480 bulk (Single cell derived)	None	28.5x	92%	I
Single cell 1	MALBAC	29.2x	93%	I
Single cell 2	MALBAC	30.3x	93%	I
Single cell 3	MALBAC	25.2x	88%	I
Single cell 4	MALBAC	26.2x	84%	I
Single cell 5	MALBAC	17x	79%	I
Single cell 6	MDA	24.8x	73%	I
Single cell 7	MALBAC	0.54x	33%	S
Single cell 8	MALBAC	0.88x	45%	S
Single cell 9	MALBAC	0.72x	40%	S
Single cell 10	MALBAC	0.61x	30%	S
Single cell 11	MALBAC	0.78x	40%	S
Single cell 12	MALBAC	1.74x	52%	S
Single cell 13	MALBAC	1.08x	48%	S
Single cell 14	MALBAC	0.50x	29%	S

Table S1: The summary of sequencing runs and data in this paper. The CNVs of single cell 7, 8 and 9 are plotted in Figure 3 (Maintext). The CNVs of single cell 10 to 14 are plotted in Figure S3. For single cell 3, 4 and 5, we used restriction enzyme to remove the MALBAC primer before library construction.

*S--SOLiD sequencing platform, I--Illumina sequencing platform

**Single SW480 cells have an incomplete genome (91.6% of the reference genome) due to loss of chromosomal fragments and unmapped regions. The coverage of single cell sequencing data is renormalized by this percentage.

	Chr	Position	Ref	Alt	ID	Region
1	chr1	88893337	G	T	novel	INTERGENIC
2	chr1	104637194	G	T	novel	INTERGENIC
3	chr1	107284358	C	A	novel	INTERGENIC
4	chr1	153059496	C	A	novel	INTERGENIC
5	chr2	7157323	C	A	novel	INTRON (RNF144A)
6	chr2	16431259	G	C	novel	INTERGENIC
7	chr2	57736349	C	G	novel	INTERGENIC
8	chr2	67286710	G	T	novel	INTERGENIC
9	chr2	158670022	T	C	novel	INTRON (ACVR1)
10	chr2	178037523	T	C	novel	INTERGENIC
11	chr2	229487751	C	A	novel	INTERGENIC
12	chr3	39345479	G	A	novel	INTERGENIC
13	chr4	44847660	C	G	novel	INTERGENIC
14	chr5	152232349	G	T	novel	INTERGENIC
15	chr6	98784811	G	C	novel	INTERGENIC
16	chr6	98784812	C	T	novel	INTERGENIC
17	chr6	159891473	G	T	novel	INTERGENIC
18	chr8	4396528	C	A	novel	INTRON (CSMD1)
19	chr8	15580190	G	T	novel	INTRON (TUSC3)
20	chr8	62158474	G	A	novel	INTERGENIC
21	chr8	89071435	G	C	novel	INTRON (MMP16)
22	chr9	1127914	C	G	novel	INTERGENIC
23	chr9	17942197	G	C	novel	INTERGENIC
24	chr11	58037258	C	G	novel	UPSTREAM (OR10W1)
25	chr11	105503786	C	A	novel	INTRON (GRIA4)
26	chr11	130731295	G	T	novel	INTERGENIC
27	chr13	88503683	C	G	novel	INTERGENIC
28	chr13	92134551	C	T	novel	INTRON (GPC5)
29	chr14	75662387	G	T	novel	INTERGENIC
30	chr15	42544253	G	T	novel	INTRON (TMEM87A)
31	chr15	47652403	G	A	novel	INTRON (SEMA6D)
32	chr16	62200248	A	T	novel	INTERGENIC
33	chr16	65916168	G	A	novel	INTERGENIC
34	chr17	43766883	A	T	novel	INTRAGENIC (MIR4315-1)
35	chr18	61495436	C	A	novel	INTERGENIC

Table S2: The list of unique and nascent SNVs identified from sibling cells analysis.

Chr1	AGGAAAGGCATACTGGAGGGACAT TTAGGGATGGCACCACACTCTTGA
Chr2	TCCCAGAGAAGCATCCTCCATGTT CACCACACTGCCTCAAATGTTGCT
Chr3	TCAAGTTGCCAGCTGTGGCTGTAT AGAAGGGCATTTCCTGTCAGTGGA
Chr4	ATGGGCAAATCCAGAAGAGTCCAG CCATTCACTTCCTTGGAAAGGTAGCC
Chr5	AATAGCGTGCAGTTCTGGGTAGCA TTCACATCCTGGGAGGAACAGCAT
Chr6	TGAATGCCAGGGTGAGACCTTTGA TGTTCATTATCCCACGCCAGGACT
Chr7	ACCAAAGGAAAGCCAGCCAGTCTA ACTCCACAGCTCCCAAGCATACAA
Chr9	TCCCAGCTCTCTCTCTTGCATCTT AGTGAAGCTGGTGTATGCAGAGGT
Chr12	AGAGGGCTGCTTTATGCAGGTG CTACATTTGGGTCTTTGCTGCCATG
Chr13	AGCAGCCCCAGGCAGAT CGGAGAGGACGGTCACGTTTAC
Chr14	CGGAGAGGACGGTCACGTTTAC CGTGGGAGTTTTGAAATGCGATGT
Chr15	CCTGTCTCTGCTCCTGCG TGCACACATGCACAGTGGAG
Chr16	CTCCAAGGTTCTGCAGCCTC GGTATGACTACACATTCAGGCTGG
Chr17	GTGGTACATAGTGCATGGTCCG GGCGACATACCCCAACTTCATAAG
Chr18	CGTTCTTAGGACCAAAGGGCTG CCAGCATCCATGTCTCTGCAC
Chr19	GCCCAGAGCGCCTGA CCAGCCCCTGGACCACT

Table S3: Primers for quantitative PCR to test the uniformity of single cell whole genome amplification. The positions are randomly selected from the genome.

qPCR	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr9
Single Cell	22.5	24.4	36.5	24.8	25.4	26.9	25.5	25.2
Positive Control	22.2	24.5	29.6	24.4	24.4	24.4	25.5	25.1
qPCR	Chr12	Chr13	Chr13	Chr15	Chr16	Chr17	Chr18	Chr19
Single Cell	25.8	23.9	39.0	24.7	21.3	24.3	24.2	27.0
Positive Control	26.4	26.3	30.0	23.8	20.0	25.8	23.7	22.5

Table S4: quantitative PCR result of a typical single cell whole genome amplification reaction using MALBAC. Shown here are the Ct numbers of randomly selected 16 loci each on a different chromosome. The single cell results are consistent with the positive control containing 500pg of DNA as starting materials. 14 out of the 16 loci are amplified evenly. The qPCR result is consistent with the ~90% genome coverage with 30x sequencing depth for single cells. Ct numbers of negative controls are all larger than 30 cycles for the above q-PCR primer pairs.

References:

33. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (Jul 15, 2009).

34. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491 (May, 2011).

35. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly* **6**, 80 (Apr-Jun, 2012).

36. F. Meacham *et al.*, Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics* **12**, 451 (2011).

37. S. Liu *et al.*, Genetic instability favoring transversions associated with ErbB2-induced mammary tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3770 (Mar 19, 2002).

Analysis of SNVs in SW480 cell line (SNPEff)

Number of positions: 2,842,162

Number of variants: 2,843,040

Annotated SNVs in dbSNP-135: 2,753,827 (96.9%)

Change rate details

Chromosome	Length	Changes	Change rate
1	249250621	254045	981
2	243199373	261615	929
3	198022430	165236	1198
4	191154276	187436	1019
5	180915260	182693	990
6	171115067	220826	774
7	159138663	136455	1166
8	146364022	129470	1130
9	141213431	124489	1134
10	135534747	157540	860
11	135006516	127358	1060
12	133851895	121265	1103
13	115169878	114028	1010
14	107349540	103886	1033
15	102531392	73470	1395
16	90354753	76987	1173
17	81195210	64743	1254
18	78077248	73674	1059
19	59128983	51789	1141
20	63025520	68350	922
21	48129895	48190	998
22	51304566	31055	1652
X	155270560	68440	2268
Total	3036303846	2843040	1067

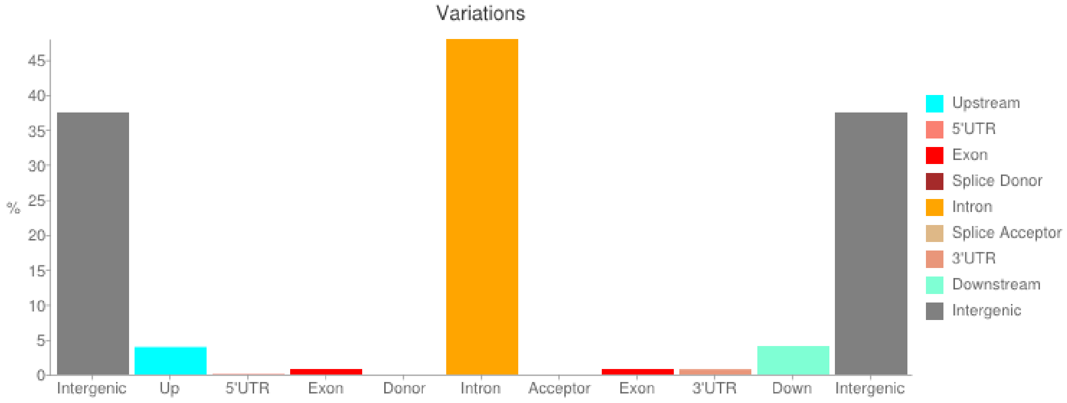
Number of effects by functional class

Type	Count	Percent
MISSENSE	12123	46.245%
NONSENSE	124	0.473%
SILENT	13968	53.282%

Number of effects by type and region

Type	Count	Percent
DOWNSTREAM	172688	4.116%
EXON	8459	0.202%
INTERGENIC	1574015	37.518%
INTRAGENIC	193872	4.621%
INTRON	2014129	48.009%
NON_SYNONYMOUS_CODING	12073	0.288%
NON_SYNONYMOUS_START	4	0%

SPLICE_SITE_ACCEPTOR	63	0.002%
SPLICE_SITE_DONOR	83	0.002%
START_GAINED	753	0.018%
START_LOST	19	0%
STOP_GAINED	124	0.003%
STOP_LOST	27	0.001%
SYNONYMOUS_CODING	13955	0.333%
SYNONYMOUS_STOP	13	0%
UPSTREAM	167337	3.989%
UTR_3_PRIME	33428	0.797%
UTR_5_PRIME	4278	0.102%



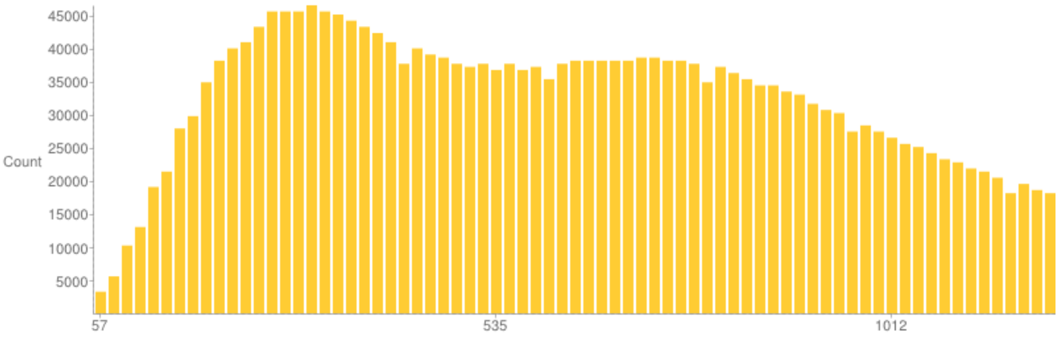
Base changes (SNPs)

	A	C	G	T
A	0	118387	469023	104246
C	121476	0	120145	486978
G	487501	120767	0	121679
T	104394	470293	118151	0

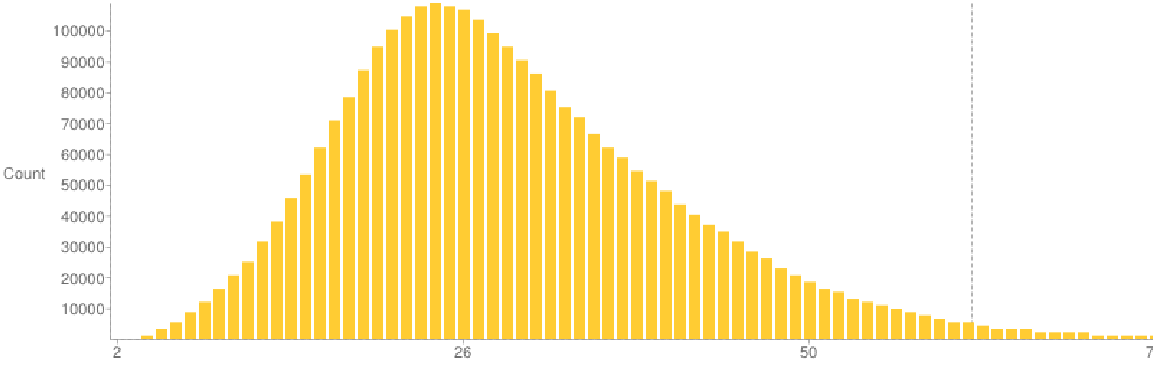
Ts/Tv (transitions / transversions)

Transitions	1913795
Transversions	929245
Ts/Tv ratio	2.0595

Quality distribution (x: QUAL score)



Coverage distribution (x: read depth)



References

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183 (2002). doi:10.1126/science.1070919 Medline

2. G. W. Li, X. S. Xie, Central dogma at the single-molecule level in living cells. *Nature* **475**, 308 (2011). doi:10.1038/nature10315 Medline

3. S. Negrini, V. G. Gorgoulis, T. D. Halazonetis, Genomic instability: An evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220 (2010). doi:10.1038/nrm2858 Medline

4. C. Lengauer, K. W. Kinzler, B. Vogelstein, Genetic instabilities in human cancers. *Nature* **396**, 643 (1998). doi:10.1038/25292 Medline

5. S. Yachida *et al.*, Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114 (2010). doi:10.1038/nature09515 Medline

6. P. J. Campbell *et al.*, The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109 (2010). doi:10.1038/nature09460 Medline

7. Y. M. Lo *et al.*, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010). doi:10.1126/scitranslmed.3001720 Medline

8. J. O. Kitzman *et al.*, Noninvasive whole-genome sequencing of a human fetus. *Sci. Transl. Med.* **4**, 137ra76 (2012). doi:10.1126/scitranslmed.3004323 Medline

9. S. Nagrath *et al.*, Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450**, 1235 (2007). doi:10.1038/nature06385 Medline

10. E. K. Hanson, J. Ballantyne, Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. *Anal. Biochem.* **346**, 246 (2005). doi:10.1016/j.ab.2005.08.017 Medline

11. M. L. Metzker, Sequencing technologies: The next generation. *Nat. Rev. Genet.* **11**, 31 (2010). doi:10.1038/nrg2626 Medline

12. H. C. Fan, J. Wang, A. Potanina, S. R. Quake, Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51 (2011). doi:10.1038/nbt.1739 Medline

13. N. Navin *et al.*, Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90 (2011). doi:10.1038/nature09807 Medline

14. M. Gundry, W. G. Li, S. B. Maqbool, J. Vijg, Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res.* **40**, 2032 (2012). doi:10.1093/nar/gkr949 Medline

15. Y. Hou *et al.*, Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873 (2012). doi:10.1016/j.cell.2012.02.028 Medline

16. J. Wang, H. C. Fan, B. Behr, S. R. Quake, Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402 (2012). doi:10.1016/j.cell.2012.06.030 Medline

17. L. Zhang *et al.*, Whole genome amplification from a single cell: Implications for genetic analysis. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5847 (1992). doi:10.1073/pnas.89.13.5847 Medline

18. H. Telenius *et al.*, Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718 (1992). doi:10.1016/0888-7543(92)90147-K Medline

19. F. B. Dean, J. R. Nelson, T. L. Giesler, R. S. Lasken, Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095 (2001). doi:10.1101/gr.180501 Medline

20. K. Zhang *et al.*, Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680 (2006). doi:10.1038/nbt1214 Medline

21. K. Lao, N. L. Xu, N. A. Straus, Whole genome amplification using single-primer PCR. *Biotechnol. J.* **3**, 378 (2008). doi:10.1002/biot.200700253 Medline

22. W. Dietmaier *et al.*, Multiple mutation analyses in single tumor cells with improved whole genome amplification. *Am. J. Pathol.* **154**, 83 (1999). doi:10.1016/S0002-9440(10)65254-6 Medline

23. X. Xu *et al.*, Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886 (2012). doi:10.1016/j.cell.2012.02.025 Medline

24. R. Beroukhim *et al.*, The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899 (2010). doi:10.1038/nature08822 Medline

25. P. J. Stephens *et al.*, Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27 (2011). doi:10.1016/j.cell.2010.11.055 Medline

26. P. J. Rochette, N. Bastien, J. Lavoie, S. L. Guérin, R. Drouin, SW480, a p53 double-mutant cell line retains proficiency for some p53 functions. *J. Mol. Biol.* **352**, 44 (2005). doi:10.1016/j.jmb.2005.06.033 Medline

27. D. MacArthur, Methods: Face up to false positives. *Nature* **487**, 427 (2012). doi:10.1038/487427a Medline

28. P. C. Blainey, S. R. Quake, Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.* **39**, e19 (2011). doi:10.1093/nar/gkq1074 Medline

29. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667 (1998). Medline

30. J. C. Roach *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636 (2010). doi:10.1126/science.1186802 Medline

31. D. F. Conrad *et al.*, 1000 Genomes Project, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712 (2011). doi:10.1038/ng.862 Medline

32. D. L. Altshuler *et al.*, 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010). doi:10.1038/nature09534 Medline

33. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754 (2009). doi:10.1093/bioinformatics/btp324 Medline

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

35. P. Cingolani *et al*., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80 (2012). Medline

36. F. Meacham *et al*., Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011). doi:10.1186/1471-2105-12-451 Medline

37. S. Liu *et al*., Genetic instability favoring transversions associated with ErbB2-induced mammary tumorigenesis. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3770 (2002). doi:10.1073/pnas.052710299 Medline