SI Appendix

Ultra-accurate Genome Sequencing and Haplotyping of Single Cells

Wai Keung Chu, Peter Edge, Ho Suk Lee, Vikas Bansal, Vineet Bafna, Xiaohua Huang, and Kun Zhang

Table of Content

I. SI Materials and Methods

I.1 Cell culture and single cells preparation

The primary PGP1 fibroblast cells were harvested at passage 12-13 and used for the SISSOR experiments. The cells were cultured under standard conditions (37 °C and 5% CO_2) in Dulbecco's Modified Eagle's Medium (DMEM; Mediatech) supplemented with 10% fetal bovine serum (Mediatech) and 1% penicillin/streptomycin (MP Biomedicals). The cells were washed with PBS and then detached from the surface of the culture flask using TrypLE Express (Life Technologies). The cells were collected by centrifugation at 1000 rpm for 3 minutes and washed twice with PBS. Finally, the cells were re-suspended in PBS and kept on ice until use.

I.2 Quality control for whole genome amplification

Common qPCR method with specific primers (Table S9) was used to detect and quantify the amount of on-chip DNA amplification. Human PGP1 fibroblast cell line and Alu sequences were used to quantify the amount of DNA amplicons in each sample. Efficiency of DNA denaturation was measured using the ratio of Alu DNA to mitochondrial DNA which does not have tightly wrapped nucleosome and higher-order chromosomal structures. The ratio of Alu to mitochondrial DNA was normalized using human genomic DNA standards. A qPCR reaction with 0.5 μl of sample input in a total volume of 20 μL containing 0.25 μM of both forward and reverse primers, and 1x KAPA SYBR mastermix (Kapa) was carried out as follows: 98°C for 30s, 40 cycles of (98°C for 10s, 58°C for 30s, and 72°C for 45s), 72°C for 2 minutes. The final products were verified using PAGE gel.

I.3 Sequencing library construction using CoRE fragmentation

CoRE (Controlled Random Enzymatic) fragmentation is used to prepare 50% of PGP1 fibroblast cell 1 and 100% of cell 3 sequencing libraries. The CoRE fragmentation process starts by generating a single nucleotide gaps at the locations of uracil incorporated in the MDA amplicons and then creating double stranded DNA from the nicked position (1). After amplicon fragmentation, NEBNext loop adapter and NEBNext sequencing index (New England Biolabs: NEB) were sequentially added by ligation and PCR. 1 μl each of 3.3x exo- Klenow buffer (Epicentre), 39 U/ml uracil-DNA glycosylase (NEB), and 78 U/ml endonuclease IV (NEB) were added to 1 µl of sample collected from each of the 24 chambers and held at 37 °C for 2 hours, then 65°C for 15 minutes. Then, 1 µl each of 100 U/ml exo- Klenow and 1 mM dNTP's (Epicentre) were added and the solution was incubated at 37°C for 1 hour and at 75°C for 15 minutes. After amplicon fragmentation, 44 µl of ligation mix with 57 nM NEBNext adaptor, 1.2 U/ul T4 DNA ligase (Kapa) and 1.2x rapid ligation buffer (Kapa) was added and the solution was incubated at 25°C for 10 minutes. 2 µl USER enzyme (NEB) was added and the solution was incubated at 37°C for 15 minutes to remove the uracil at the NEBNext loop adaptor. Un-ligated NEBNext adaptor was removed by standard 50 µl AMPure XP bead purification (Beckman Coulter), and the sample was eluted from the beads with 20 µl water. Finally, individual NEBNext Sequencing Index was added by PCR in a 30 µl reaction containing the eluted DNA, 0.33 µM each of the forward primer and NEBNext index primer (NEB), 1.67x KAPA SYBR mastermix (Kapa) as follows: 98°C for 30s, 12 cycles of (98°C for 10s, 65°C for 30s, 72°C for 45s), 72°C for 2 minutes. The PCR products were separated using PAGE gel and products in the 300-700 bp range were extracted.

I.4 Sequencing library construction using transposon tagmentation

Transposon tagmentation was used to prepare 50% of PGP1 fibroblast cell 1 and 100% of cell 2 sequencing libraries. The process requires 2nd strand synthesis of MDA amplicons with Pol I treatment to make accessible double stranded DNA for transposases digestion. After the digestion, adapters and sequencing index sequence were sequentially added by PCR. 3 µl of alkaline lysis solution (ALS) (400 mM KOH, 10 mM DTT and 1% Tween20) was added to each 2 µl sample at room temperature for 3 minutes. The tube was then place on an ice block. 3 µl of neutralization solution (NS) (400 mM HCl and 600 mM Tris-HCl) was mixed with the sample to neutralize the solution. A solution containing 10 U of DNA Polymerase I (Invitrogen), 1x Ampligase buffer (Epicentre), 1x NEB buffer 2, 250 ng unmodified random hexamer and 1 mM dNTP's (Epicentre) was added and the solution was incubated at 37°C for 1 hour then at 65°C for 10 minutes. To seal the nicks, 1 U Ampligase (Epicentre) was added and the ligation was carried out at 37°C for 10 minutes, then at 65°C for 10 minutes. The product was purified with standard ethanol precipitation and eluted in 7 µl of water. Each DNA sample was fragmented by adding 1 µl of Nextera transposase (1:50 dilution, Epicentre) and 2 µl of 5x High Molecular Weight buffer and incubating the solution at 55°C for 5 minutes. The transposase was then inactivated by adding 0.05 U of protease (Qiagen) and incubating the solution at 50°C for 10 minute, then at 70°C for 20 minutes. Single-stranded end filling was carried out by adding 5U Exo minus Klenow (Epicentre) and 1 mM dNTP's and incubating the solution at 37°C for 15 minutes, followed enzyme inactivation at 75°C for 20 minutes. Adaptor sequences were added by 7 cycles of PCR using 1x Kapa SYBR fast mastermix, 10 µM each of adapter 1 and barcode adapter, and 7 additional cycles of PCR using 1x Kapa SYBR fast mastermix, 10 µM each of Illumina primer 1 and primer 2. The products were verified using PAGE gel and purified using Ampure XP beads. PCR products were separated and DNA in the 300-700 bp range was selected and extracted for sequencing.

I.5 DNA sequencing using next-generation short-read sequencing

Three sequential rounds of Illumina sequencing yielded higher depth and genomic coverage for the sequencing libraries. We performed the first round of single-end 36 bp or 50 bp sequencing reads to provide some insight into DNA fragment lengths and distribution in different chambers. Around 40 million 36 bp single-end reads were obtained using Genome Analyzer II for each of PGP1#21 and PGP1#22 samples. Around 50 million 50 bp single-end reads were obtained using MiSeq and sample from PGP1#A1. Statistics such as mapping rate, unique rate, fragment size estimation and distribution were determined using the data from the first round of sequencing. The 2nd and 3rd round of single-end or paired-end 100 bp reads provided the coverage and depth required for base calling. For the 2nd round, about 400 million 100 bp single-end reads from each sample were obtained using HiSeq rapid run. The 3rd round of sequencing required to reach over 50% clonal rate was estimated from the unique rate from the 2nd round of sequencing. Depending on the sequencing level needed, more single-end or paired-end 100 bp reads were obtained by a combination of HiSeq rapid and high output runs. Sequencing depth and coverage are summarized in Table S3.

I.6 Genome coverage

Genome coverage was calculated by the percentage of bases in the genome using uniquely mapped reads. Coverage from individual single cell libraries ranged from 54.9%-73.6% of the roughly 2.9 Gb reference GRCh37 genome (Table S4). The difference in genome coverage from the three cells summed up to a total of 94.9% coverage of the genome.

I.7 Segmentation of SISSOR fragments using HMM

Single-stranded DNA fragments from a single-cell genome are randomly distributed in MDA reaction chambers in our SISSOR device. The sequencing library from each chamber is presumed to contain no more than one single haplotype. However, amplification bias introduced during whole genome amplification could result in uneven genome coverage. Using a hidden Markov model (HMM), the aligned sequencing reads from the individual MDA chamber are joined into a continuous segment, which we call a SISSOR fragment, based on the read depth and proximity in the localized genomic region.

The CSHL code (2) was used to divide the whole human genome into 50,000 bins of ~60 kb. A string of 50,000 digits was created for each chamber and input to an HMM (smooth.discrete of mhsmm package (3) in R). Read depths above the threshold in a bin create the "observed" state. The numbers of reads below the threshold create the empty (unobserved) state alternatively. Begin and end positions of each segment are determined by the bin boundaries. The total number of mappable reads from each individual sequencing library divided by 50,000 bins equals to the minimum threshold for segmentation. In this two-state model, we set the value of such bin to "1" for the observed state and "0" for the empty state. Bins containing centromere regions and other bad bins defined by CSHL are removed (set state to "0") before and after Viterbi decoding. We set the probability of moving from i to j in one time step $Pr(j|i) = Pi,j$ as 0.01 and staying as 0.99 for the transition matrix, where i is the row and j is the column. The emission matrix is also a two by two matrix with 0.99 as the probability for emitting 1 (above threshold) in observed state and the probability for emitting 0 (below threshold) in the empty state. The result of this two-state smoothing process closely resembled the start and end of dense reads. Five fold of the minimum thresholds was used. This HMM is expected to smooth the sparse but long DNA fragments into a continuous "observed" segment. We extracted the reads from SAM/BAM within the fragment boundaries and verified the HMM visually using SeqMonk and/or direct plotting with R (Fig. S4 and S5).

I.8 SNV calling algorithm

I.8.1 SNV calling overview

The goal of SNV calling with SISSOR data is twofold: first, to determine the best consensus call (SNV or reference) for every genomic position, given read data for every chamber, second, to determine the best call for each individual SISSOR chamber in light of information from the other chambers. For instance, if the same SNV is observed in multiple chambers, then the confidence of the SNV call in each chamber is higher than if the SNV were only observed in one chamber. Similarly, the confidence for that SNV in the consensus genotype over all chambers is higher if it is observed multiple times. In general, a group of reads observed in a chamber at a genomic position is generated from one of four strands: the forward or reverse strand from one haplotype, or the forward or reverse strand from the other haplotype (hereafter referred to as strands 1, 2, 3 and 4, respectively). Traditional SNV calling algorithms are unsuitable for the task of calling variants from multiple groups of reads amplified from separate DNA strands, especially across multiple cells and in light of protocol-specific error modalities. For this reason, we designed a custom consensus SNV calling approach for the SISSOR method. The variant caller accounts for multiple sources of errors besides sequencer error, including error introduced during MDA from the Phi29 enzyme, and the occurrence of multiple source DNA strands being amplified in the same chamber. Given sets of read observations from different chambers, the variant caller considers every possible way in which the four single strands of DNA for each genotype could be distributed across chambers. Given multiple single-cell libraries, it considers all combinations of events in each cell that could result in the combined set of data. These events are modeled in a likelihood framework to make a Bayesian calculation for the posterior probability that a SNV is present. The variant caller is implemented in python and takes as input a multi-sample pileup of all the chamber bam files, generated with samtools mpileup (4). The caller makes the following primary assumptions: reads are correctly mapped, variant calls at different genomic positions are independent, and the genotypes of each cell in the multiple-cell case are the same. To make use of reads amplified from strands smaller than the ~60 kb HMM window sizes, all chambers with read coverage ≥3 were considered in the model (and genomic positions with more than 4 such chambers in a cell were left uncalled, in keeping with the diploid model). The following sections describe the consensus SNV calling model.

I.8.2 SNV calling parameters

The SNV caller models the experimental workflow of the SISSOR method. As such, it requires knowledge of various library-specific parameters. We estimated parameters for the model either empirically from the data or based on prior studies. The prior probability of a genotype, $P(G)$, was estimated using the method described by Li et al (5). We refer to the set of genotype priors as P_G. We denote the probability of sampling a fragment from a given chamber i as $P_s[i]$. $P_s[i]$ was assumed to be proportional to the relative genomic coverage of a chamber:

$$P_s[i] = (1 - P_s[\emptyset]) * \frac{cov(i)}{\sum_{j\in 1..24} cov\,(j)}$$

$P_s[\emptyset]$ represents the probability that a strand is not sampled, and was estimated to be consistent with the distribution of strand depth (the number of chambers at a given position containing fragments). Let S[i] be the number of genomic positions with exactly i chambers with reads. Let the 4-tuple $C = (c_1, c_2, c_3, c_4)$represent a *chamber configuration* of 4 distinct DNA strands to chambers, with $c_1, c_2, c_3, c_4 \in \{1,2,...,24, \emptyset\}$. Let Č refer to the set of all possible configurations, and let $Č_i \subset Č$ be the set of chamber configurations such that exactly i of (c_1, c_2, c_3, c_4) are not equal to $\emptyset$. If we assume that strand coverage per position results from independent trials depending only on the overall probability of drawing exactly that many strands, we can describe a likelihood for the observed strand coverages:

$$\mathrm{P}(\mathrm{S}\,|P_s) = \prod_{i=0}^{4} \left(\sum_{C\in Č_i} P_s\,[c_1] * P_s[c_2] * P_s[c_3] * P_s[c_4] \right)^{S[i]}$$

We selected $P_s[\emptyset] = 0.81$ as the approximate value that maximizes this likelihood, and this result was consistent with estimates based on the difference of the total coverage from theoretical perfect 4-strand coverage. We use ε to refer to the probability of error in base-calling. It is described as a constant variable for simplicity but in general represents the per-base quality score of a read position. We use $P_m[x]$ to denote the probability that x fraction of reads in a chamber are "noise" of a minority base resulting from MDA amplification. Assuming the X chromosome should be monoallelic except for MDA error, we estimated $P_m[x]$ as the distribution of the fraction of the second-most-common allele for each position on the X chromosome. Although this accounts for noise from secondary bases due to MDA, it is known that the consensus error rate from MDA is on the order of $1 * 10^{-5}$ (6). We let $\omega = 1 * 10^{-5}$ represent the probability that the majority base in a chamber (the "consensus") was changed as a result of MDA amplification. We use $P_p[x]$ to denote the probability that x fraction of reads in a chamber originate from a given parent. $P_p[x]$ accounts for the possibility of strands from different haplotypes occuring at the same position in the same chamber. P_p was estimated using the logic that the fragments of the hemizygous X chromosome can be shuffled to random positions to simulate a separate homologous chromosome. By overlapping the original fragments with the shuffled fragments we can simulate strand overlaps in a diploid case. With this in mind, we sampled coverages x_1 and x_2 of independent random positions from the X chromosome $10^8 1 * 10^8$ times, and used the distribution of the value $\frac{x_1}{x+x_2}$ as an estimate of P_p. In the following formulation, we use $\pi = \{P_G, P_s, \varepsilon, P_m, \omega, P_p\}$ to refer to the entire collection of parameters.

I.8.3 SNV calling framework

We begin by considering a single genomic position, with some number of observed reads aligned to the position in each of 24 chambers. Let a_i represent the pileup of observed bases in chamber i, and $a_{i,j}$ denote the base $\in \{A, C, G, T\}$ observed in chamber i at the jth read (in chamber i's base pileup). The set of observed data is $A = [a_1, a_2, \dots, a_{24}]$. Let G denote the true genotype of the individual at the site, so G can be homozygous in the reference or alternate allele, or heterozygous. Using Bayes' rule, we can compute the posterior probability of a specific genotype in terms of the probability of the data given each genotype:

$$P(G|A,\pi) = \frac{P(A|G,\pi)P(G)}{\sum_G P\,(A|G,\pi)Pr(G)}$$

Because of DNA strand loss and uneven amplification, the data for many positions may be insufficient to assign a diploid genotype even though it is highly likely that a specific allele is present. For this reason, we computed all alleles $\alpha \in \{A, C, G, T\}$ such that the probability that α is present, regardless of genotype, is greater than a threshold τ:

$$P(\alpha|A,\pi) = \sum_{G,\alpha\in G} P\,(G|A,\pi)$$

$$\boldsymbol{return}\,\{\alpha \mid P(\alpha|A,\pi) > \tau\}$$

I.8.4 Likelihood of data in all chambers

In order to calculate $P(A|G,\pi)$, it is necessary to account for every configuration in which the 4 strands can be distributed amongst 25 chambers (treating $\emptyset$, or "unsampled", as a 25th chamber). Let $C = (c_1, c_2, c_3, c_4)$ be the chambers corresponding to the four strands where c_i can take values from 1-25. We can compute the probability of this configuration as the product of the probabilities of sampling strands in those chambers:

$$P(C) = \prod_{i\in c} P_s\,[i]$$

Given SISSOR read data A for a single position, Let $N(A) \subseteq \check{C}$ denote the set of configurations that could have generated A with non-zero probability.

Then,

$$P(A|G,\pi) = \sum_{C\in N(A)} P\,(A|C,G,\pi)P(C|G) = \left(\sum_{C\in N(A)} \prod_{j=1}^{24} P\,(a_i|C,G,\pi)\right) Pr[C|G]$$

In the case of multiple cells, the data from different cells is independent conditional on the genotype G. To generalize to multiple (in our case, $n = 3$) cell samples, we change A to be of length $24n$ and refer to the observed data across all $24n = 72$ chambers. We redefine Č for multiple cells as the nth Cartesian power of Č in the single-cell case, or the set of unique n-tuples of 4-tuples that combines one single-cell strand configuration of each cell. An n-cell configuration C describes one way to combine one single-cell configuration from each cell, and Č (in the n-cell case) describes every such way. The probability of an n-cell configuration is equal to the product of the constituent single-cell strand configurations' probabilities.

I.8.5 Likelihood of data in one chamber

Now we consider how to calculate $P(a_i|C_i, G)$, or the likelihood of the observed chamber data (read bases) given the genotype and knowledge of which strands are present (strand configuration). Let $g_1, g_2 \in \{A, C, G, T\}$ denote the allelic values of G currently being considered. First, we define the probability of seeing a read base $a_{i,j}$ given that it originated from genotype allele g:

$$P(a_{i,j}|g, \pi) = \begin{cases} (1-\varepsilon) & if a_{i,j} = g \\ \varepsilon & otherwise \end{cases}$$

We address each possible case for chamber-strand configurations separately. We use K_1 to represent the set of configurations in which 1 strand falls into chamber i. We use K_2 to represent the set of configurations in which 2 or more strands fall into chamber i, and more than 1 distinct allele is present.

In the case where there is only one strand allele present, we take the product of the probabilities of observing each base $a_{i,j}$ given that the true allele is g. g refers to the allele of G that is present in chamber i as a result of configuration C_i. We assume that MDA error changes the majority allele from g to a different base b (the *consensus* allele) with probability ω. We assume that MDA also introduces j "noise" bases of a base ~b with probability $\frac{j}{n}$, and otherwise the base is b with probability $\frac{n-j}{n}$. $P_m\left[\frac{j}{n}\right]$ is the probability that j of the n bases are noise.

$$\Omega[b] = \begin{cases} 1-\omega & if b = g \\ \dfrac{\omega}{3} & otherwise \end{cases}$$

$$P(a_i|C_i \in K_1, G, \pi) = \sum_{b\in\{A,C,G,T\}} \Omega[b] \sum_{j=1}^{n} P(a_i|\mathrm{b}, \pi, noise = j) P_m\left[\frac{j}{n}\right]$$

To compute the probability of chamber data given that the consensus allele is b, we sum over all possible proportions of allele mixture from MDA:

$$P(a_i|b, \pi, noise = j) = \prod_{k=1}^{n}\left(\left(\frac{j}{n}\right) P(a_{i,k}|\sim\mathrm{b}, \pi) + \left(\frac{n-j}{n}\right) P(a_{i,k}|\mathrm{b}, \pi)\right)$$

We now consider the case where there are multiple strands with different alleles occurring in the same chamber. This is modeled similarly to MDA noise. To compute the total likelihood of an allele call c_i given a genotype, we sum over all possible proportions of strand mixture, with j representing reads originating from parental allele 1 and $n - j$ representing reads originating from parental allele 2:

$$P(a_i|C_i \in K_2, G, \pi) = \sum_{j=1}^{n}\left[P_p\left[\frac{j}{n}\right] \prod_{k=1}^{n}\left(\left(\frac{j}{n}\right) P(a_{i,k}|g_1, \pi) + \left(\frac{n-j}{n}\right) P(a_{i,k}|g_2, \pi)\right)\right]$$

The $P_p\left[\frac{j}{n}\right]$ term accounts for the probability of occurrence of a parental allele in the given fraction. k

represents the current index in the set of base calls for chamber i. The term $\left(\frac{j}{n}\right)P\left(a_{i,k}\middle|g_1\right)$ represents the case that $a_{i,k}$ was generated by g_1 from strand 1 (probability $\frac{j}{n}$ that $a_{i,k}$ came from strand 1, times $P\left(a_{i,k}\middle|g_1\right)$ the probability of allele $a_{i,k}$ given that it came from g_1). The next two terms represent analogous information for the case that $a_{i,k}$ came from strand 2. To reduce computation, we assume that MDA noise and strand overlap do not occur in the same chamber. Computation was further minimized by constraining the domains of P_m and P_p to ≤ 20 evenly spaced bins.

I.8.6 Likelihood of an allele in a chamber

Along with a consensus call across many chambers, we also called the most likely allele occurring in a specific chamber, in light of information from other chambers. This is done in a similar fashion to computing the most likely genotype G. Consider a single chamber i for which we want to determine the allele present (if any) on the original strand. We denote the assignment of an allele $\alpha \in \{A, C, G, T\}$ to chamber i as α_i. We want to choose the most likely α_i:

$$\max_{\alpha} \; P(\alpha_i|A,\pi)$$

As before, we use Bayes' rule:

$$P(\alpha_i|A,\pi) = \frac{P(A|\alpha_i,\pi)P(\alpha_i)}{\sum_{\alpha_i} P\,(A|\alpha_i,\pi)P(\alpha_i)}$$

Computing $P(A|\alpha_i,\pi)$ follows similarly to computing $P(A|G,\pi)$, except that we sum the likelihood of chamber data over all genotypes and configurations in which chamber i contains allele α.

I.9 Haplotype assembly

Haplotype assembly requires a large set of high-confidence heterozygous SNVs. For the purpose of haplotype assembly, we used a set of heterozygous SNVs from 60× Illumina WGS of PGP1f cells (Encode phase 3, ENCSR674PQI) (7). The original VCF containing SNV calls was lifted over to hg19 using CrossMap and sorted with vcftools (8, 9). After this, the heterozygous SNV calls were filtered for coverage ≥10 and quality score ≥30. Reference and variant calls in each SISSOR chamber were grouped into haplotype fragments if they fell inside the boundaries of the same called fragment. Chamber calls that differed from the majority base in the chamber's base pileup were filtered out (e.g. in unusual cases where data for individual chambers does not fit cleanly into the diploid base-calling model). Only base calls with coverage ≥5 $>$ 5that overlapped the set of heterozygous SNVs were retained, and quality scores of allele calls in haplotype fragments were fixed to 20. Four post-processing steps were applied to increase haplotype accuracy: first, fragments were filtered out if more than 5% of base calls were mixed alleles, which indicate strand overlap from different haplotypes or similar error. Then, fragments were split at spans of multiple mixed-allele base calls (more than 25% of calls having a mixed allele in a span of ≥3 heterozygous SNV locations). Then, fragments that were highly discordant to other fragments were filtered out (≥30% rate of switch errors of any length across all overlaps to other fragments). Finally, a haplotype fragment was split if it had a switch error of length 2 SNVs or greater with respect to multiple overlapping fragments. If it was ambiguous which fragment was the source of the error (for instance, in the case of only two overlapping fragments), multiple fragments were split. Following these fragment processing steps, $\geq$ 3The processed fragments were assembled into haplotype blocks with HapCUT2 (10). SNVs were pruned at a HapCUT2 SNV confidence level of 0.95, blocks were split at a switch confidence level of 0.95, and a standard discrete pruning heuristic was applied (11).

I.10 Accuracy of haplotypes

To assess haplotype accuracy, it is important to compare against a confident reference haplotype. We compared against haplotypes assembled from BAC clones (12). To maximize the accuracy of the BAC clone haplotypes, the original BAC fragments were filtered for heterozygous SNVs present in the PGP1f Illumina WGS dataset used to generate SISSOR haplotype fragments (7). In the same fashion as raw SISSOR fragments were processed, BAC clones were split at positions where 2 or more heterozygous SNVs were switched with respect to other clones. After this, the processed BAC clones were assembled into haplotype blocks using HapCUT2 (10) and pruned at high stringency: SNVs were removed at HapCUT2 SNV confidence level of 0.9999, blocks were split at switch confidence level of 0.9999, and a standard discrete pruning heuristic was applied (11). Accuracy of SISSOR haplotypes was assessed by all-pairs comparison of SISSOR haplotype blocks to these high-stringency BAC haplotype blocks. Accuracy was measured using the concept of switch and mismatch errors (also called long and short switches, respectively) (13) . Looking at positions shared by a single SISSOR haplotype block and a single BAC haplotype block, a switch error is defined as a heterozygous SNV position where the phase in the SISSOR haplotype block is different than the BAC reference with respect to the previous shared position (called in both haplotypes). Two switch errors occurring in a row are instead called a single mismatch error, which results in a difference in phase of only one SNV with respect to the BAC reference. The mismatch discordancy rate is defined as the fraction of compared positions that had a mismatch error. The switch discordancy rate is similarly defined, but the denominator is slightly smaller as it does not count first and last compared SNVs in a block (these are always called mismatch errors if the phase differs). The term "discordancy rate" is used instead of "error rate" because it is assumed that the BAC haplotypes, while accurate, may have non-negligible error.

I.11 Same haplotype strand pairing

Fragments were assigned to haplotypes by matching them back to the assembled haplotype blocks. A fragment was required to match the assembled block with 80% accuracy or greater and contain at least 2 haplotype-informative calls at heterozygous SNV positions. After assignment, all base calls (with calls different from the pileup majority base filtered out) inside overlapping fragments were analyzed and a position was classified as a *strand-match* if both fragments had the same call and as a *strand-mismatch* if the fragments had different calls. Strand-mismatched positions were quantified for the purpose of estimating the effects of errors from MDA, DNA damage, and other sources. Strand-matched positions in adjacent chambers (chambers 1 and 2, chambers 2 and 3, ... chambers 23 and 24) were discarded, because cases of DNA leakage were observed where DNA from a single strand leaked to physically adjacent chambers and generated a false haplotype-paired strand. The remaining strand-matched calls are of higher confidence than other calls because of their haplotype support, so these calls were tested for concordance against a curated set of SNV and reference calls for PGP1 (described below). Strand-matched calls between strands in different cells that differed from the PGP1 reference were used to estimate the maximum error rate for strand-matched calls in SISSOR technology since these calls are shared by the cell line. Strand-matched calls between strands in the same cell that differed from the PGP1 reference were analyzed as potential *de novo* variants specific to the cell.

I.12 Accuracy of SNV calling

SNV (and reference) calls from SISSOR were compared to a dataset obtained by combining multiple sources for PGP1. First, raw BAMs from a 60× Illumina WGS sequencing of PGP1f cells (Encode phase 3, ENCSR674PQI) were used to generate calls at every genomic position using Freebayes with the standard_filters and report_monomorphic options (7). These calls were lifted over to hg19 with CrossMap and sorted with vcftools (8, 9). The single-nucleotide calls in this dataset were filtered for those matching a CGI WGS dataset for PGP1, to filter for only high-quality calls shared by both samples (14). The resulting intersected dataset had 2.7 billion reference calls and 3.0 million SNV calls. This dataset served as the basis for comparison (SISSOR calls were compared against positions called in the intersected dataset). Variants observed in BAC sequencing libraries (12) served as an extra source for validation; calls that differed from the intersected data but were seen in BAC were considered to be correct.

I.13 Workflow management

The complete workflow for variant calling, haplotype assembly, haplotype strand pairing, and accuracy calculations was managed with Snakemake (15).

I.14 Differences to LFR technology

LFR technology uses reproducibly phased variants in the two replicate libraries to distinguish *de novo* and cell line specific mutations from false positive variants (1). SISSOR essentially calculates the sequencing error rate from phased variants in two individual SISSOR libraries. We distinguished *de novo* variants in a single cell by correctly calling the reference base in the SISSOR libraries from the other cells. We also distinguished cell line specific variants by calling the identical consensus call in a third chamber in the other SISSOR libraries.

II. Supplementary Figures

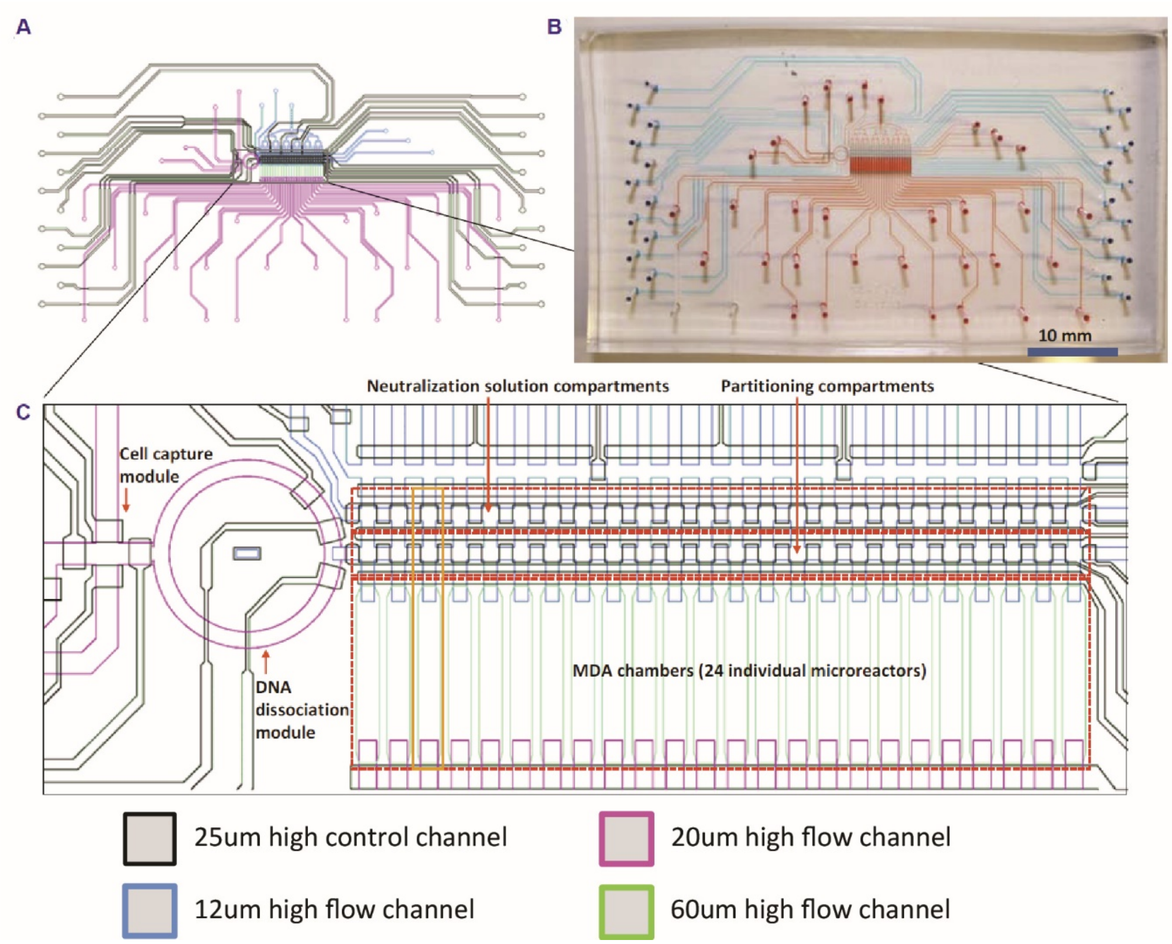
A
B
10 mm
C
Neutralization solution compartments
Partitioning compartments
Cell capture module
DNA dissociation module
MDA chambers (24 individual microreactors)
25um high control channel
20um high flow channel
12um high flow channel
60um high flow channel

Fig. S1. SISSOR device.
(A) The overall CAD design. (B) An image of a functional SISSOR device filled with dye solutions (red: fluidic channels; blue: the valves and valve lines). (C) A zoom-in view of the regions with functional modules. Highlighted in an orange box is a single unit with the neutralization, partitioning and MDA chambers. Black lines are valves and valve lines. Blue lines are 12 µm (H) x 120 µm (H) domed channels. Magenta lines are 25 µm (H) x 250 µm (W) domed channels. Green lines are 60 µm (H) x 200 µm (W) rectangle MDA chambers with vertical side walls.

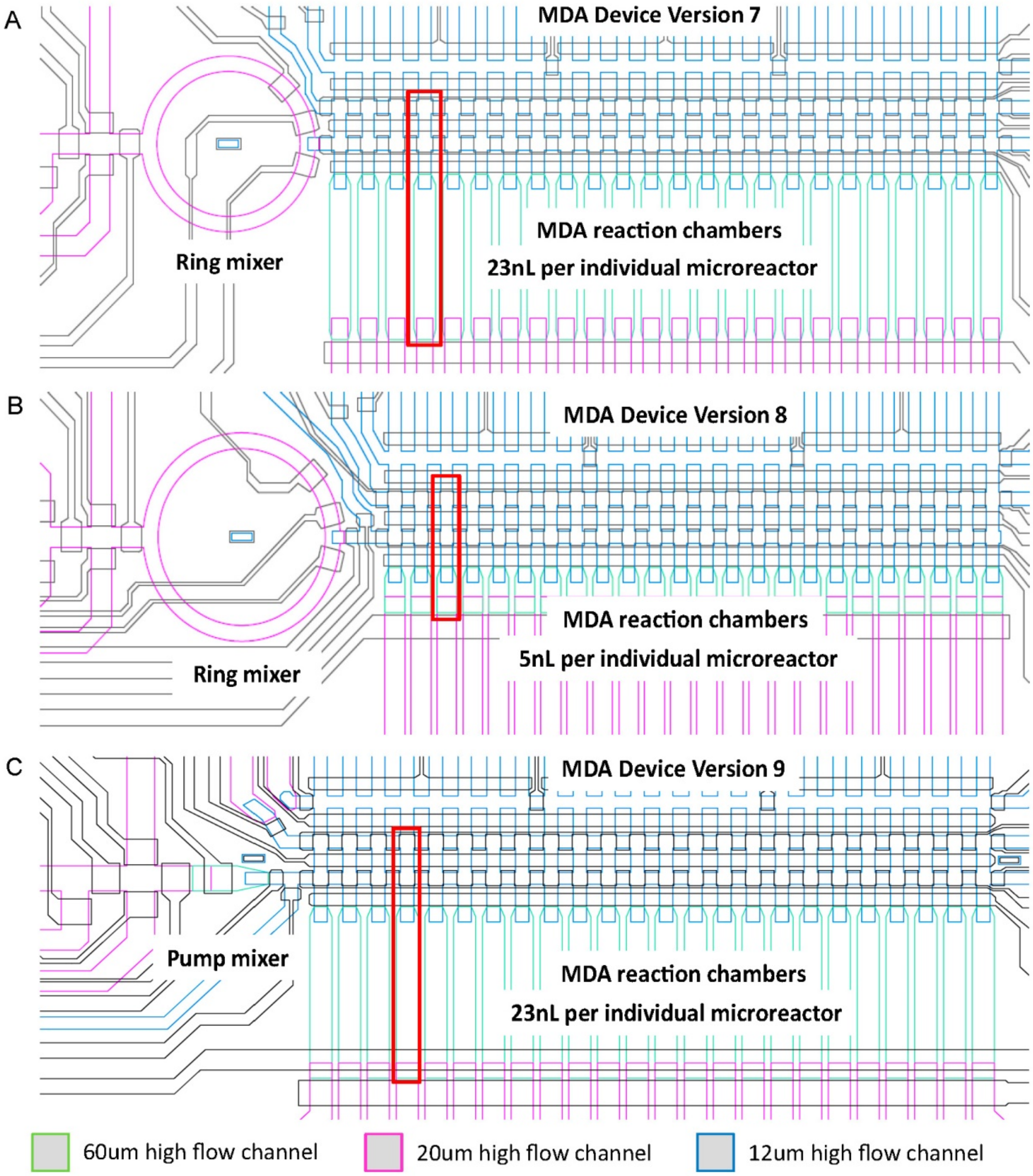

Fig. S2. Multiple versions of tested SISSOR devices.

(A) Optimized SISSOR device with a rotary mixer and ~20nL MDA chambers. (B) SISSOR device with reduced volume in MDA reaction chambers. (C) SISSOR device with a mixer actuated by the PDMS valve above the denaturation chamber.

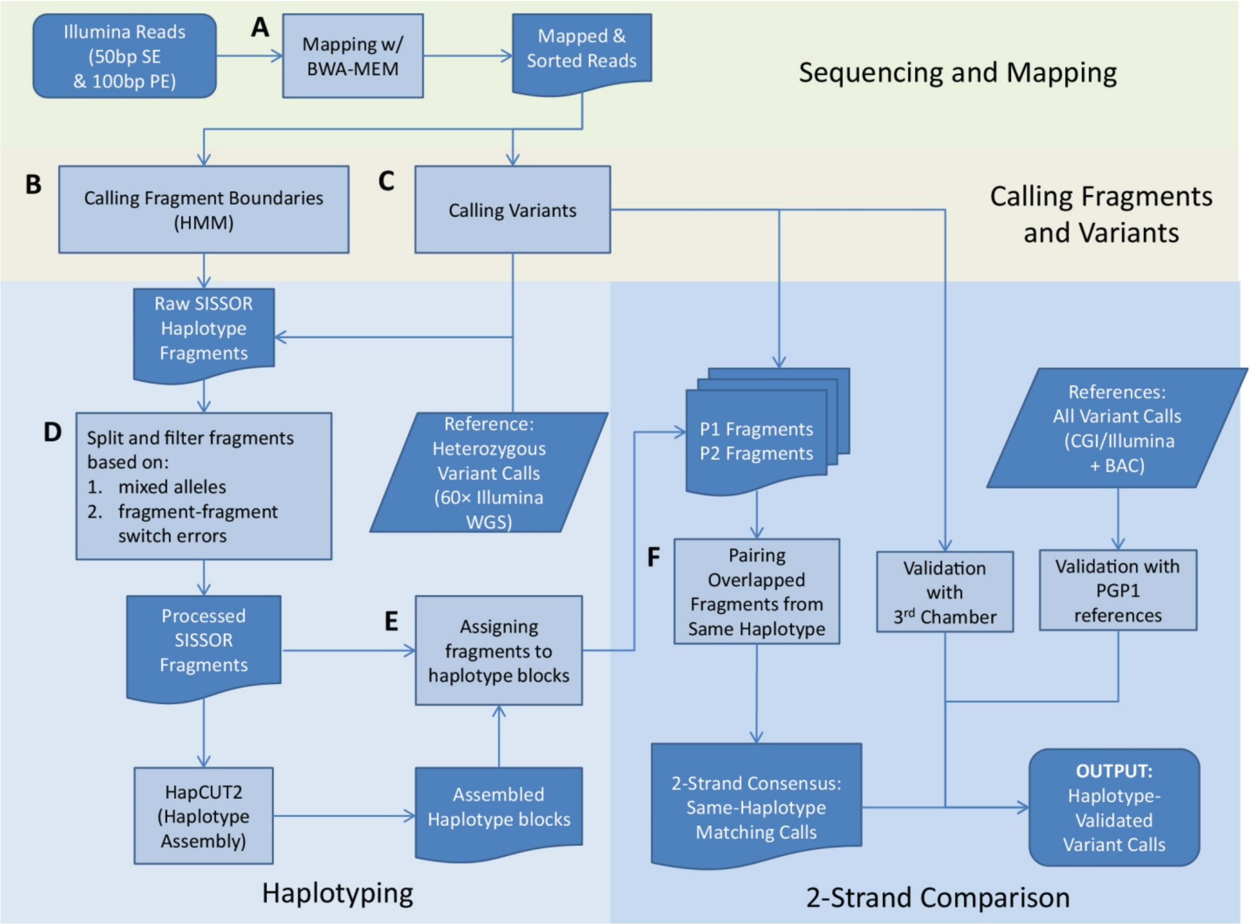
Illumina Reads (50bp SE & 100bp PE)
A
Mapping w/ BWA-MEM
Mapped & Sorted Reads
Sequencing and Mapping
B
Calling Fragment Boundaries (HMM)
C
Calling Variants
Calling Fragments and Variants
Raw SISSOR Haplotype Fragments
D
Split and filter fragments based on:
1. mixed alleles
2. fragment-fragment switch errors
Reference: Heterozygous Variant Calls (60× Illumina WGS)
P1 Fragments
P2 Fragments
References: All Variant Calls (CGI/Illumina + BAC)
F
Pairing Overlapped Fragments from Same Haplotype
Validation with 3rd Chamber
Validation with PGP1 references
Processed SISSOR Fragments
E
Assigning fragments to haplotype blocks
HapCUT2 (Haplotype Assembly)
Assembled Haplotype blocks
2-Strand Consensus: Same-Haplotype Matching Calls
OUTPUT: Haplotype-Validated Variant Calls
Haplotyping
2-Strand Comparison

Fig. S3. Schematic overview of data analysis.

(A) Routine sequencing and mapping. (B) Determination of SISSOR fragments. (C) Variant calling with novel algorithm (SI Appendix, Supplementary methods). Haplotyping: (D) Processing SISSOR fragments and (E) assigning fragments to haplotypes. Two-strand comparison: (F) Pairing overlapped positions in fragments.

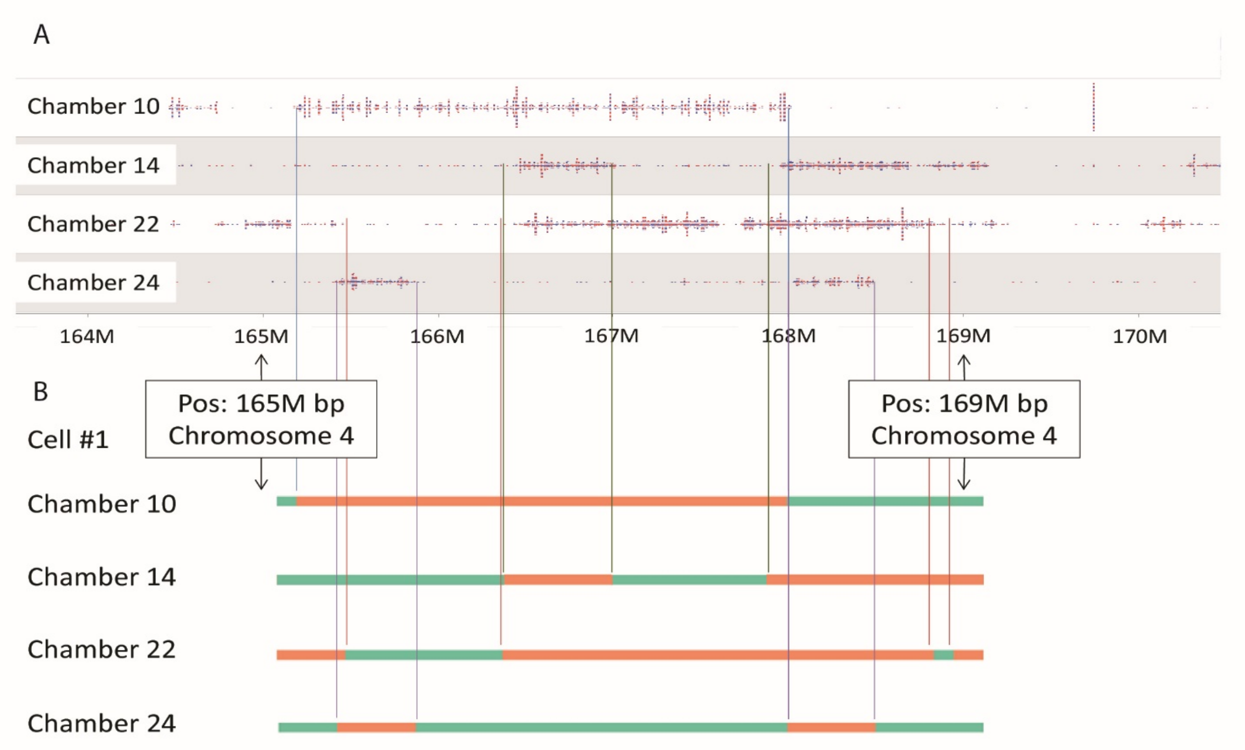

Fig. S4. Visualization of reads and single-stranded fragments within a 4 Mbp region of Chromosome 4 in four selected chambers.

Genomic positions of reads and fragments were visualized using SeqMonk and HMM SISSOR fragments extraction. (A) Dense sequencing reads depicted the replica of single-stranded DNA. Red and blue lines are forwardly and reversely mapped reads created by MDA. (B) Fragment size and boundaries were determined by the number of read counts in 50k variable bins and hidden Markov model (mshmm in R). Two fragments from the same parental origin appeared at the same genomic position validates single stranded DNA amplification.

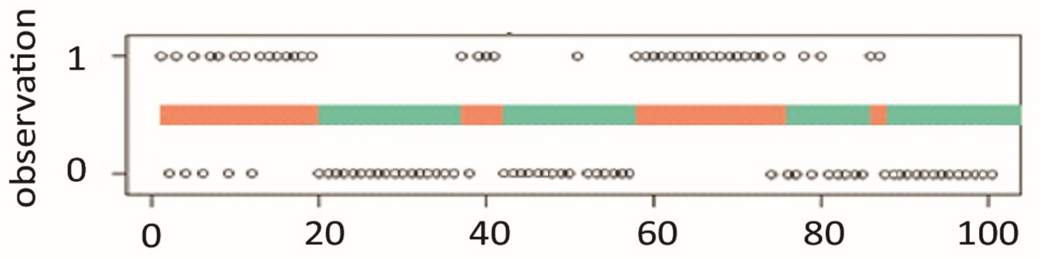

Fig. S5. Schematic view of segmentation by HMM on 100 bins.
The first 100 bins of chromosome 4 in cell 1 chamber 22 are selected to illustrate the segmentation result. Circles depict the observations of each bin, where state “0” and “1” respectively represent the number of reads below and above the threshold (> 5x average reads per bin). Two color fragments, directly plotted by R, showed the calculated hidden states. The discrete states were smoothed and four SISSOR fragments are created by HMM in this example.

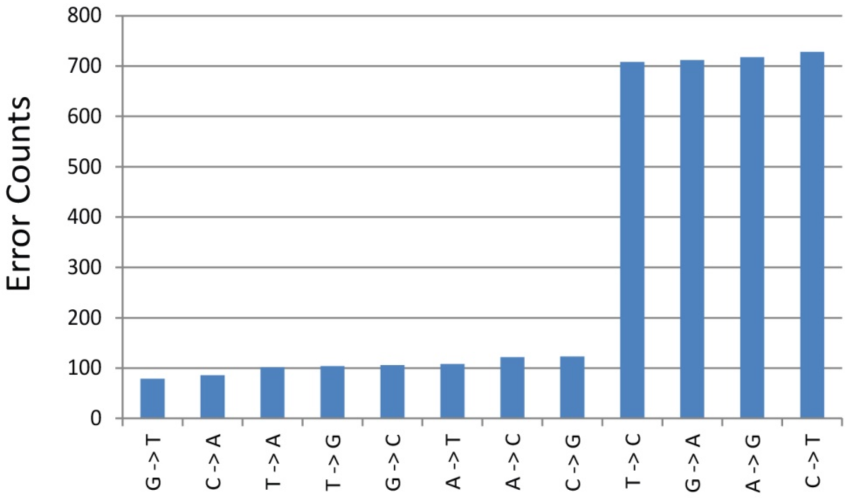

Fig. S6. Quantification of 12 nucleotide substitutions in all same-cell strand-strand mismatches. High ratio of transition to transversion (~3.45) usually indicates higher accuracy in SNPs discovery (16).

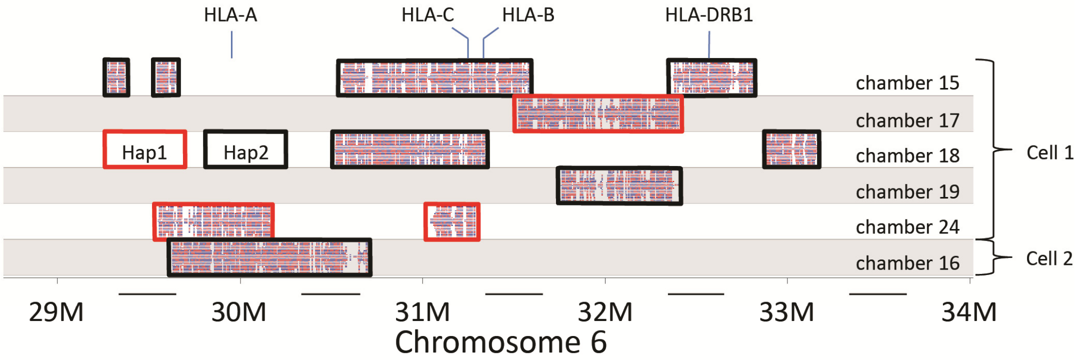

Fig. S7. Phasing SISSOR fragments to human leukocyte antigen (HLA) region (28.5-33.5 Mbp on Chromosome 6).

Representative SISSOR fragments overlapping the HLA regions visualized using SeqMonk. Heterozygous SNPs were matched to PGP1 haplotype library and were counted against two parental origins (Hap1/Hap2). High connectivity in long fragments >500 kb enables correctly phased haplotype.

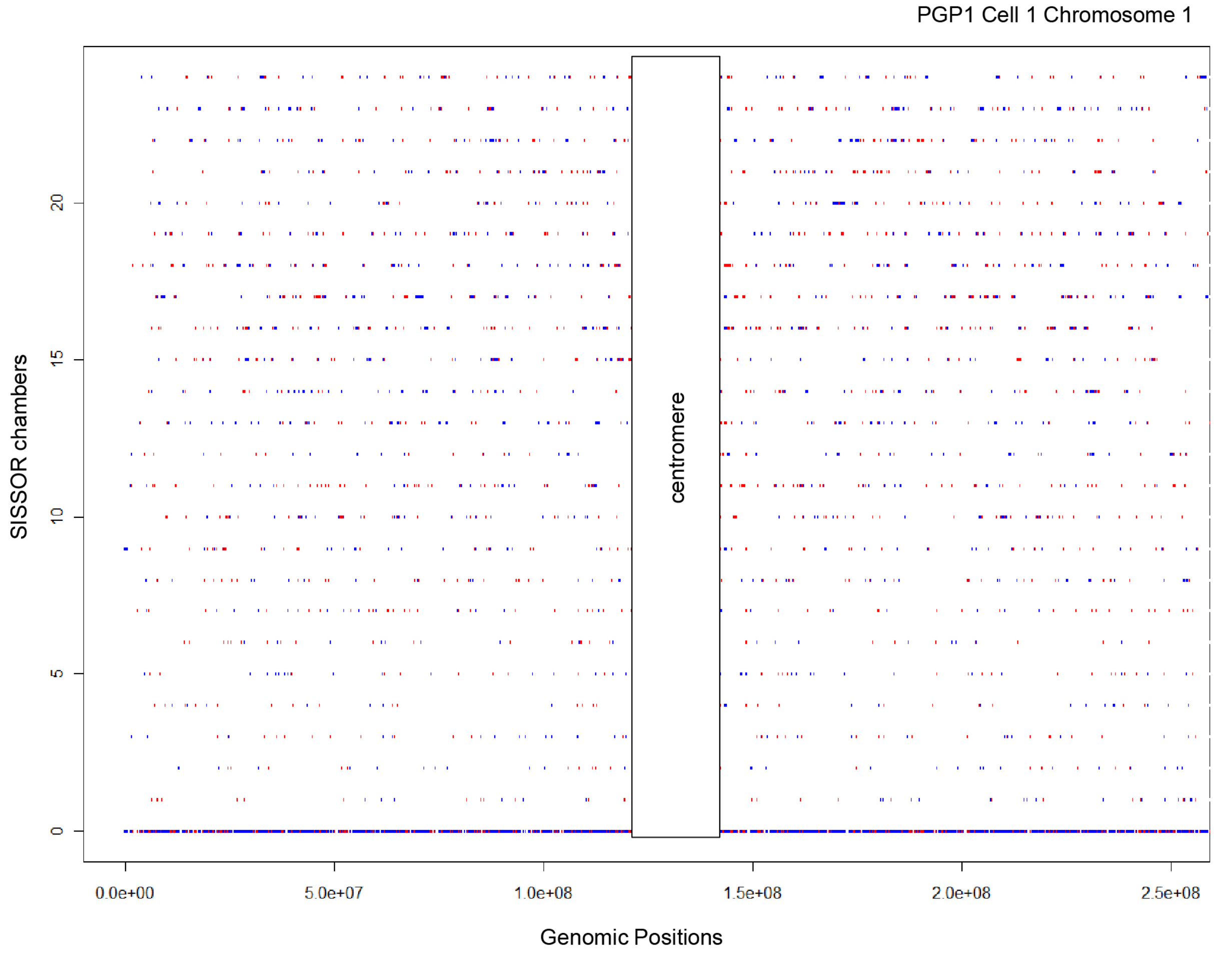

Fig. S8. Distribution of unprocessed sequencing reads on chromosome 1 of PGP1 Cell 1. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). The first 3916 bins, representing chromosome 1, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

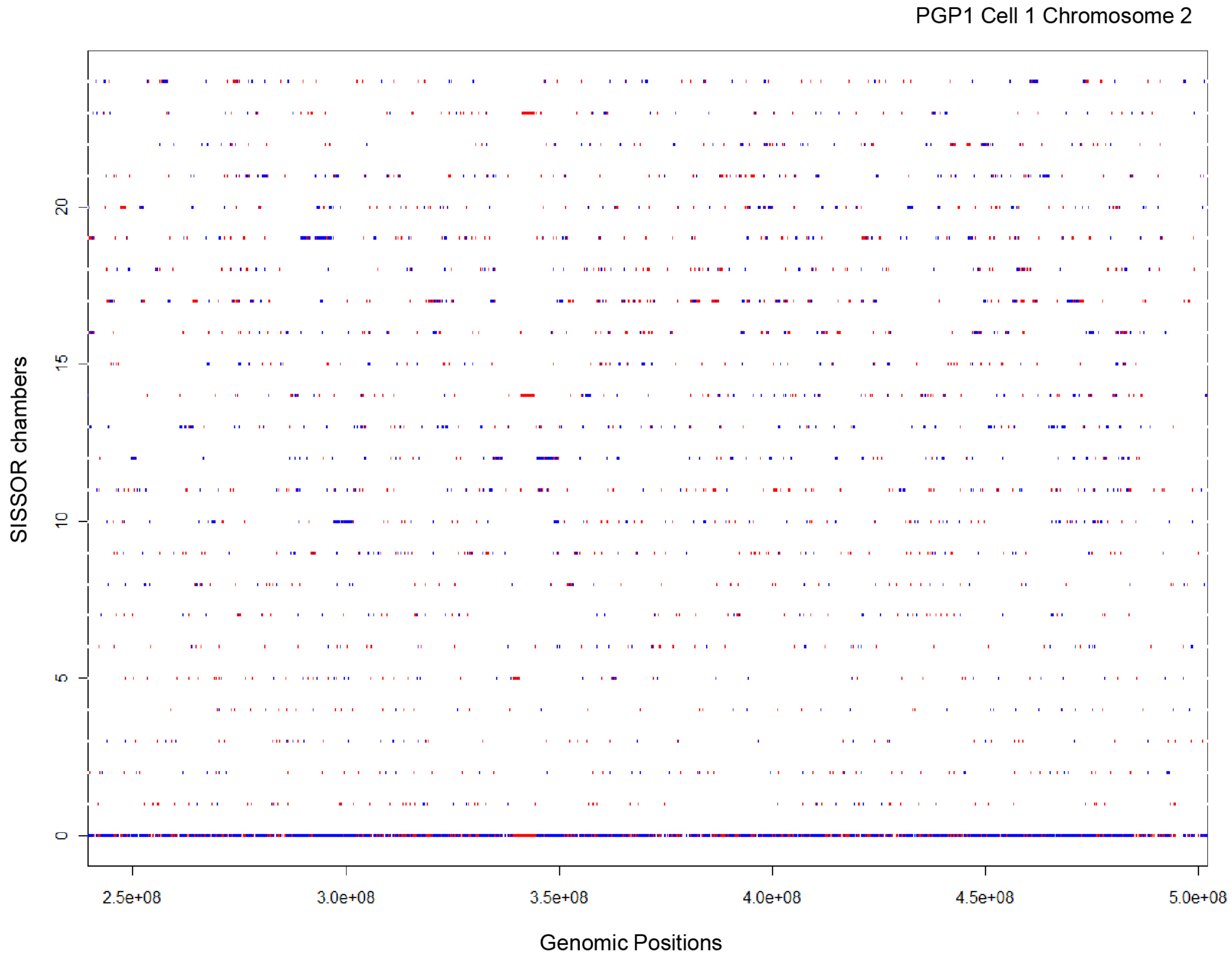

Fig. S9. Distribution of unprocessed sequencing reads on chromosome 2 of PGP1 Cell 1. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 4235 bins, representing chromosome 2, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

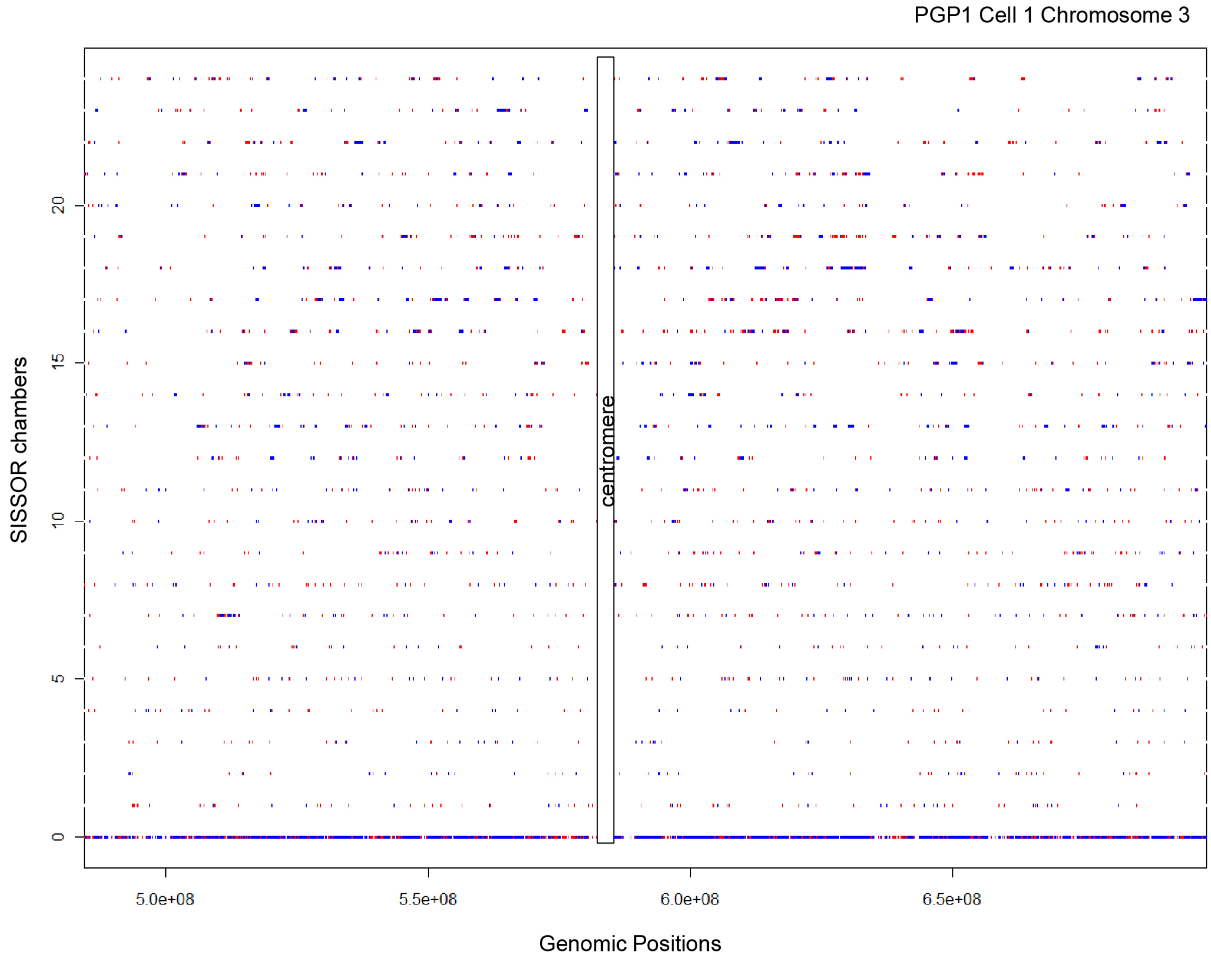

Fig. S10. Distribution of unprocessed sequencing reads on chromosome 3 of PGP1 Cell 1. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3520 bins, representing chromosome 3, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

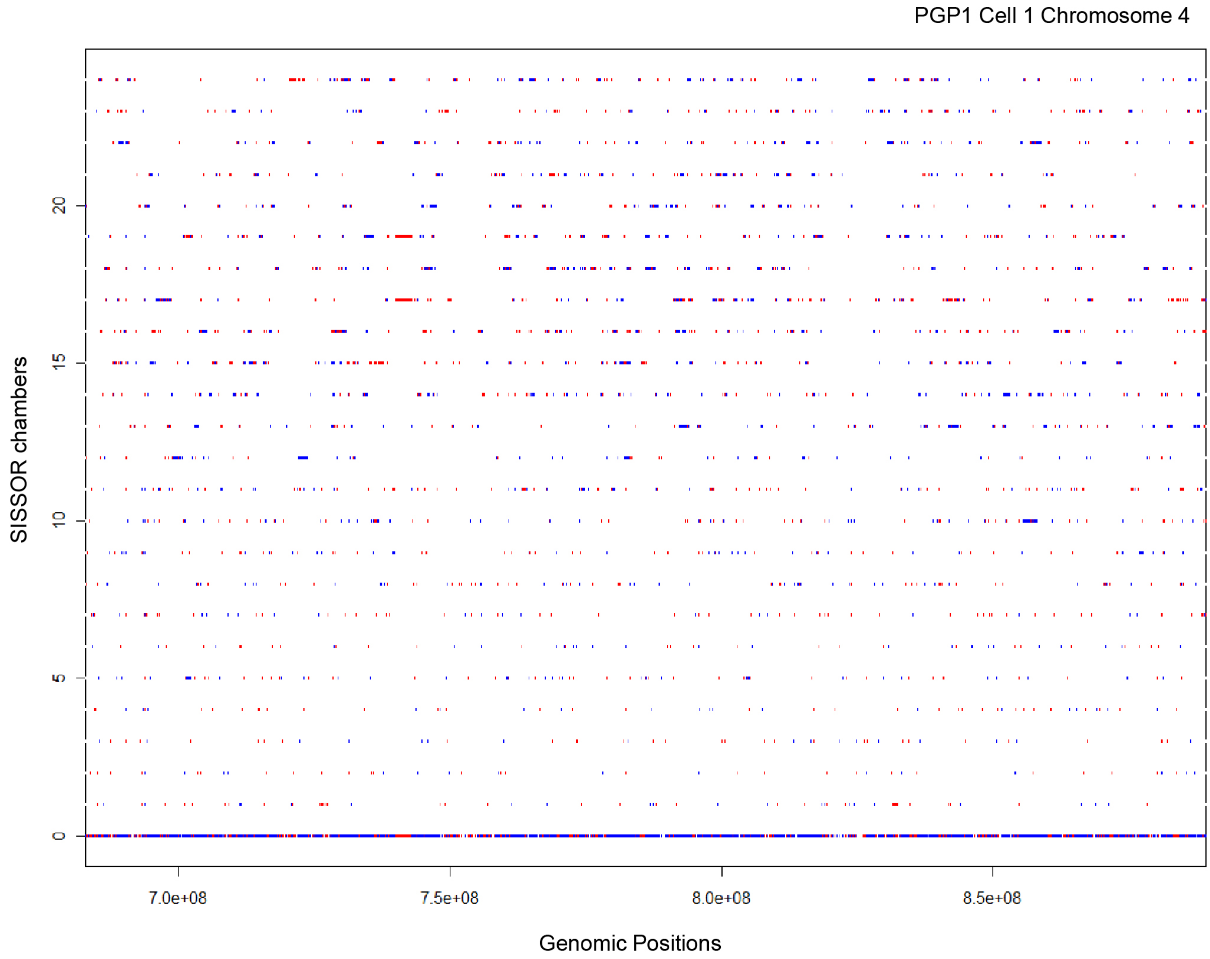

Fig. S11. Distribution of unprocessed sequencing reads on chromosome 4 of PGP1 Cell 1. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3373 bins, representing chromosome 4, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

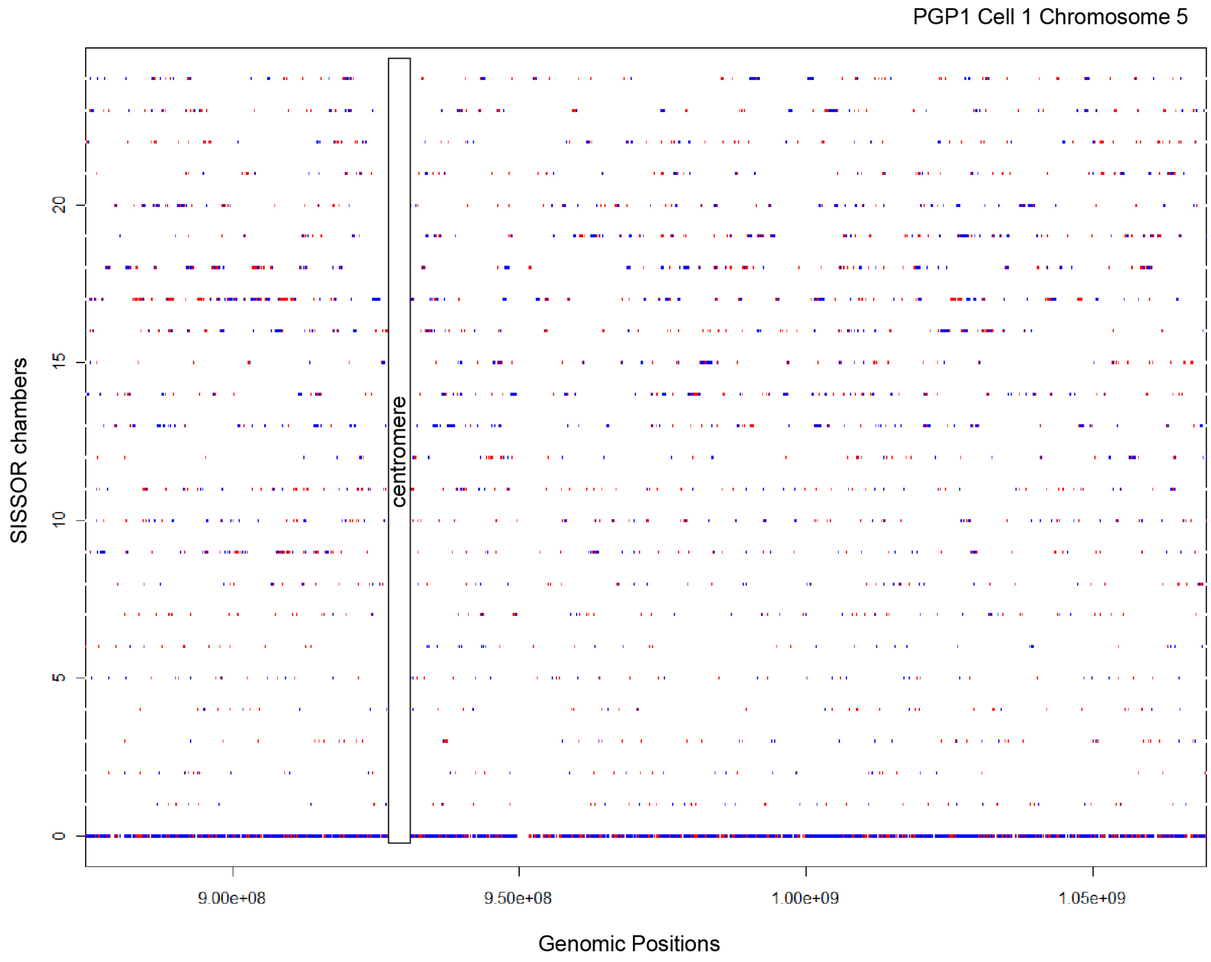

Fig. S12. Distribution of unprocessed sequencing reads on chromosome 5 of PGP1 Cell 1. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3158 bins, representing chromosome 5, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

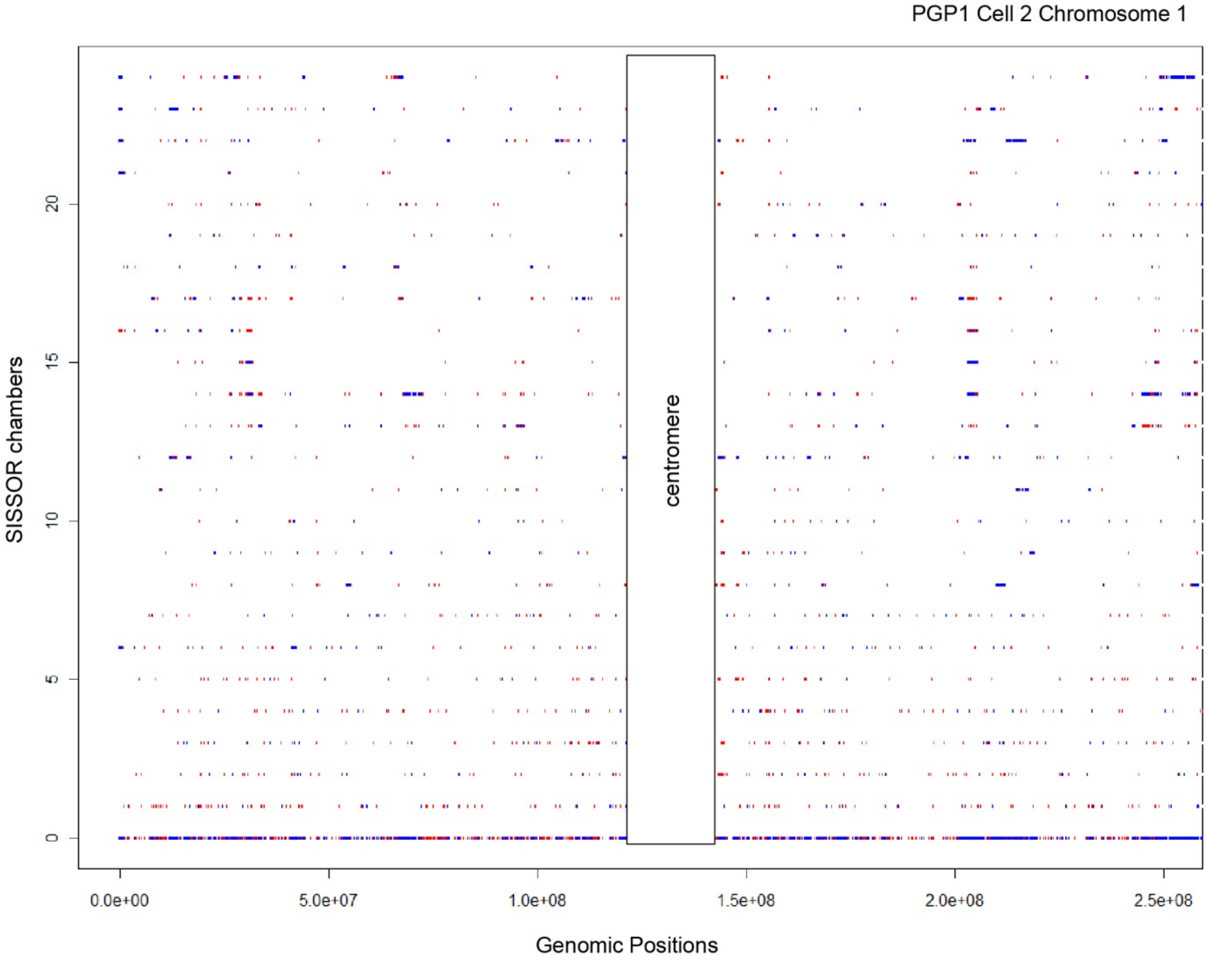

Fig. S13. Distribution of unprocessed sequencing reads on chromosome 1 of PGP1 Cell 2.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). The first 3916 bins, representing chromosome 1, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

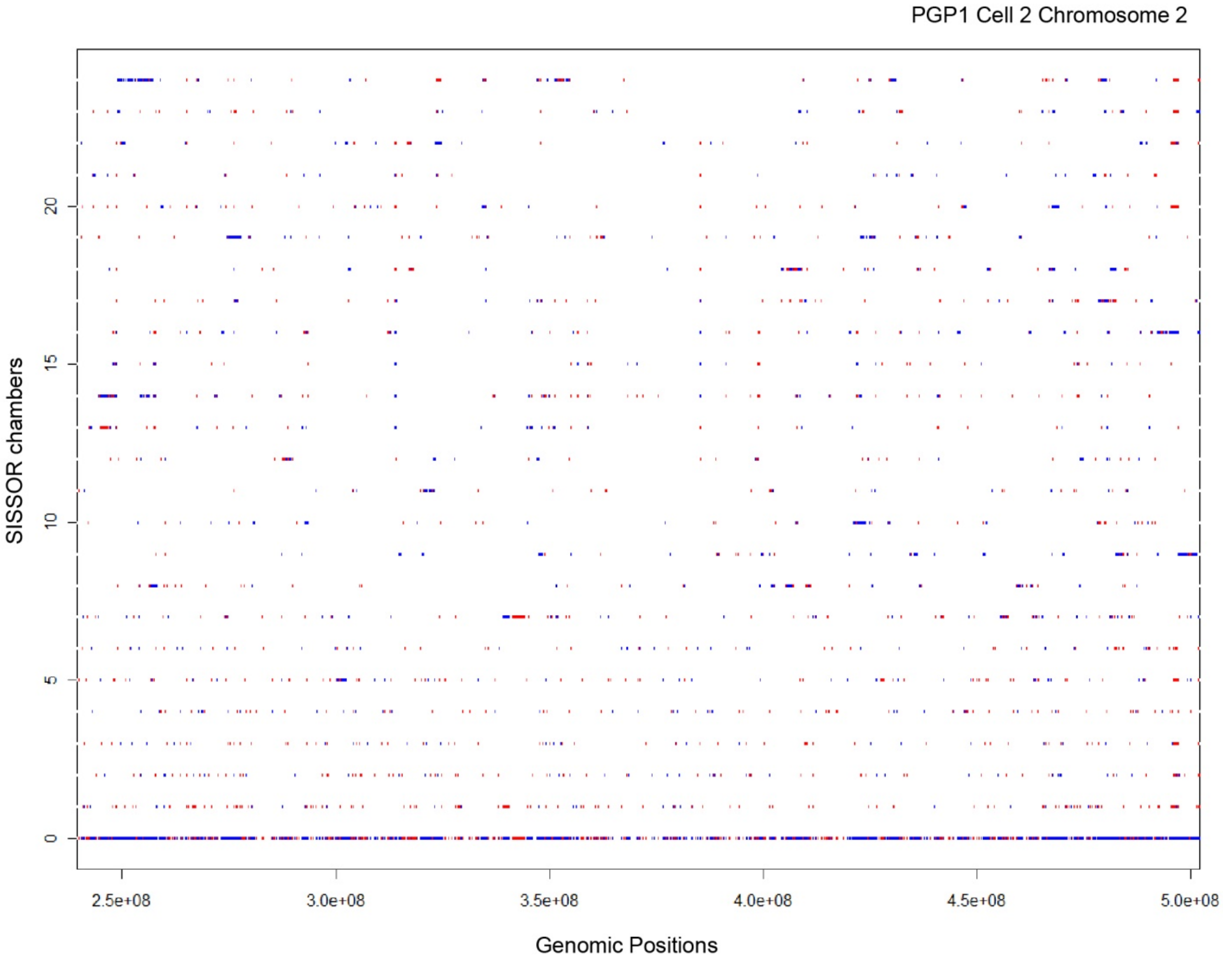

Fig. S14. Distribution of unprocessed sequencing reads on chromosome 2 of PGP1 Cell 2.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 4235 bins, representing chromosome 2, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

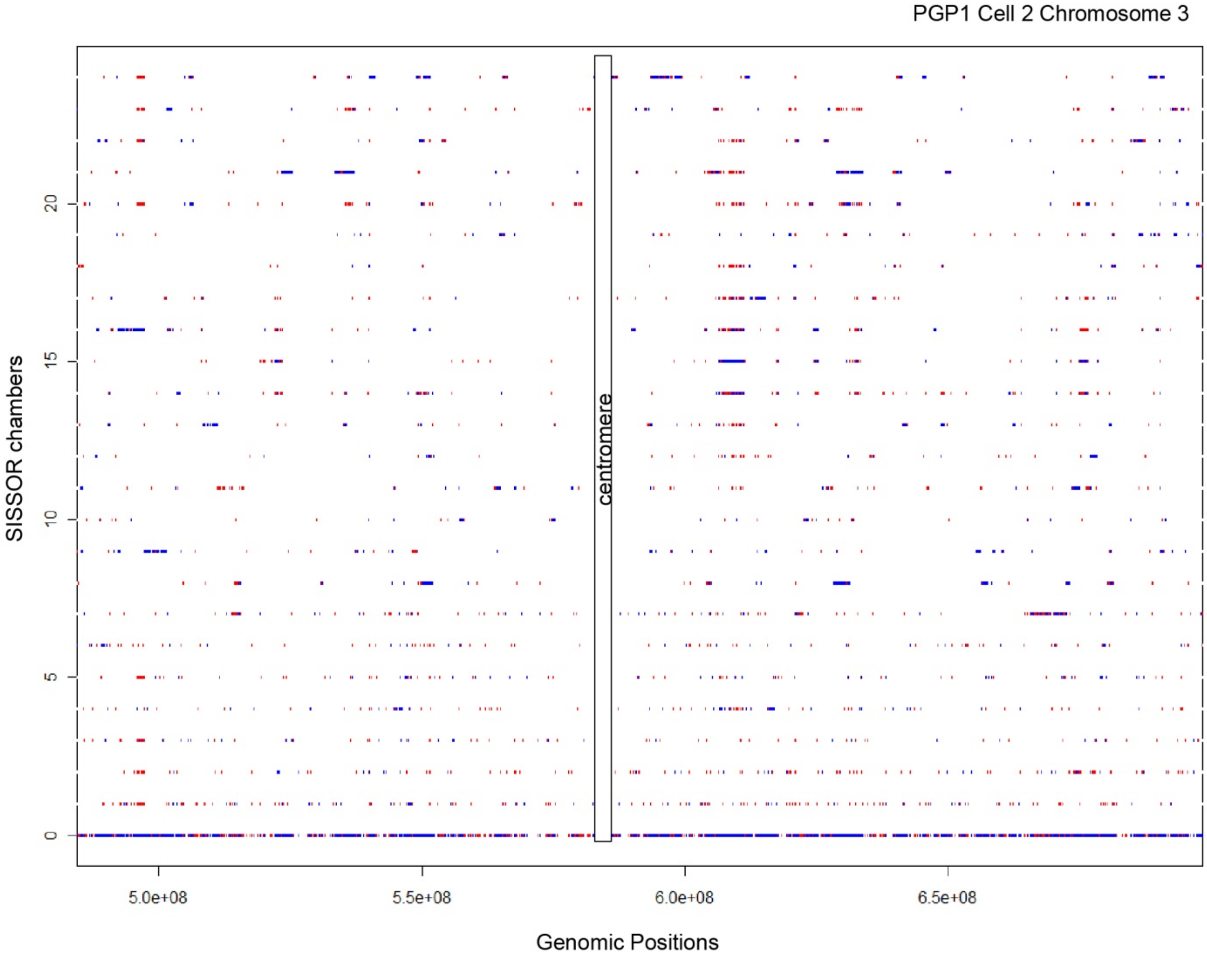

Fig. S15. Distribution of unprocessed sequencing reads on chromosome 3 of PGP1 Cell 2.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3520 bins, representing chromosome 3, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

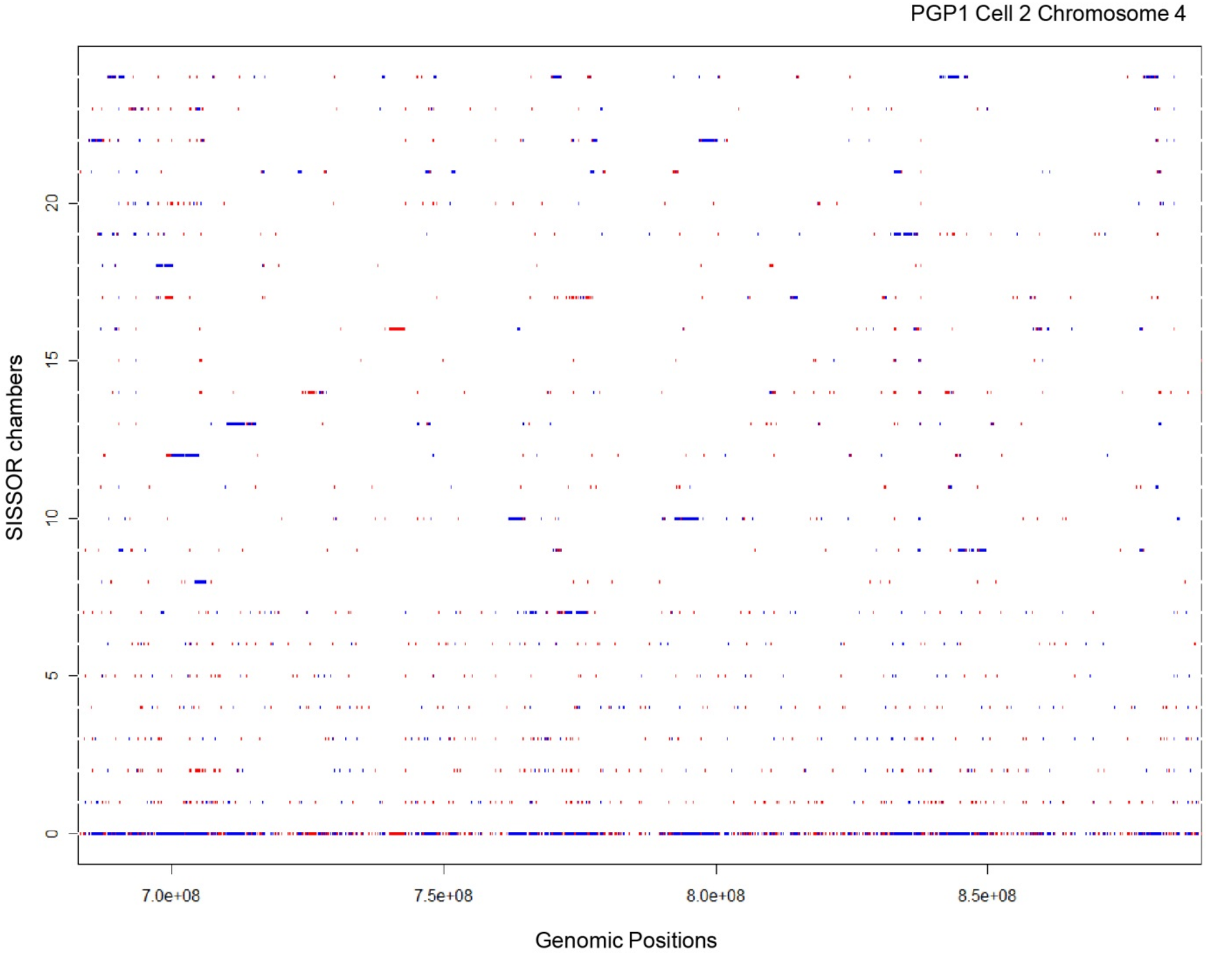

Fig. S16. Distribution of unprocessed sequencing reads on chromosome 4 of PGP1 Cell 2.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3373 bins, representing chromosome 4, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

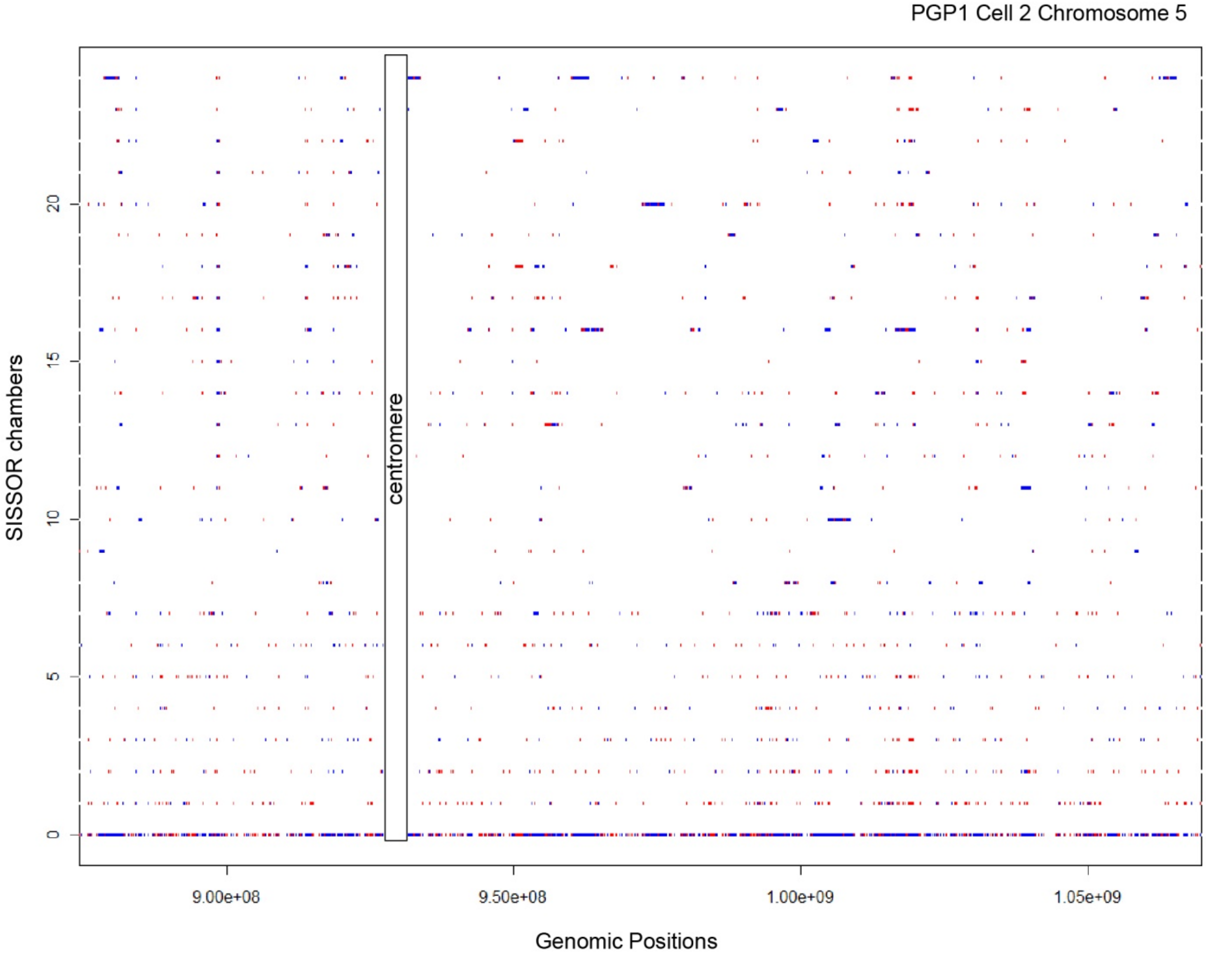

Fig. S17. Distribution of unprocessed sequencing reads on chromosome 5 of PGP1 Cell 2.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3158 bins, representing chromosome 5, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

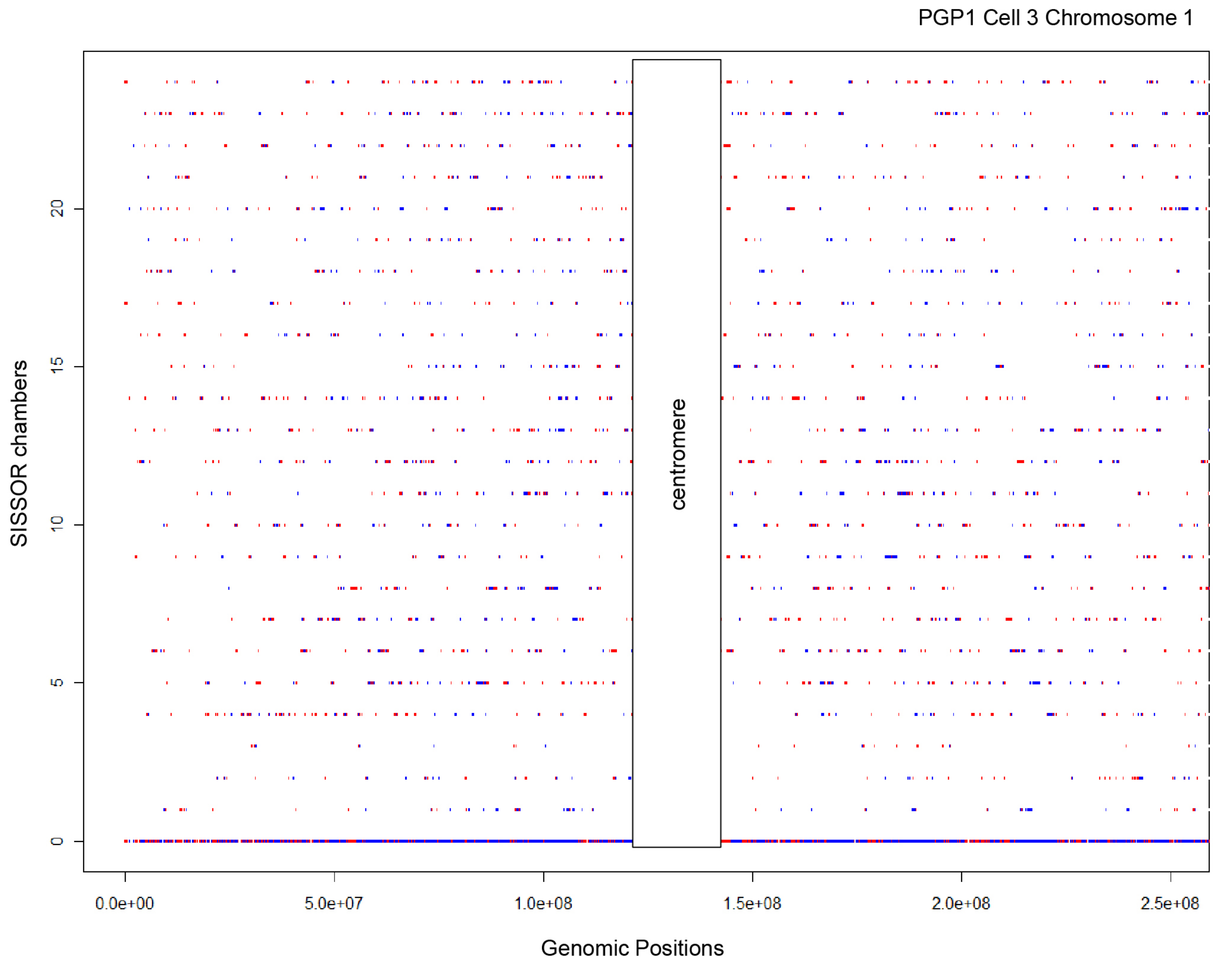

Fig. S18. Distribution of unprocessed sequencing reads on chromosome 1 of PGP1 Cell 3. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). The first 3916 bins, representing chromosome 1, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

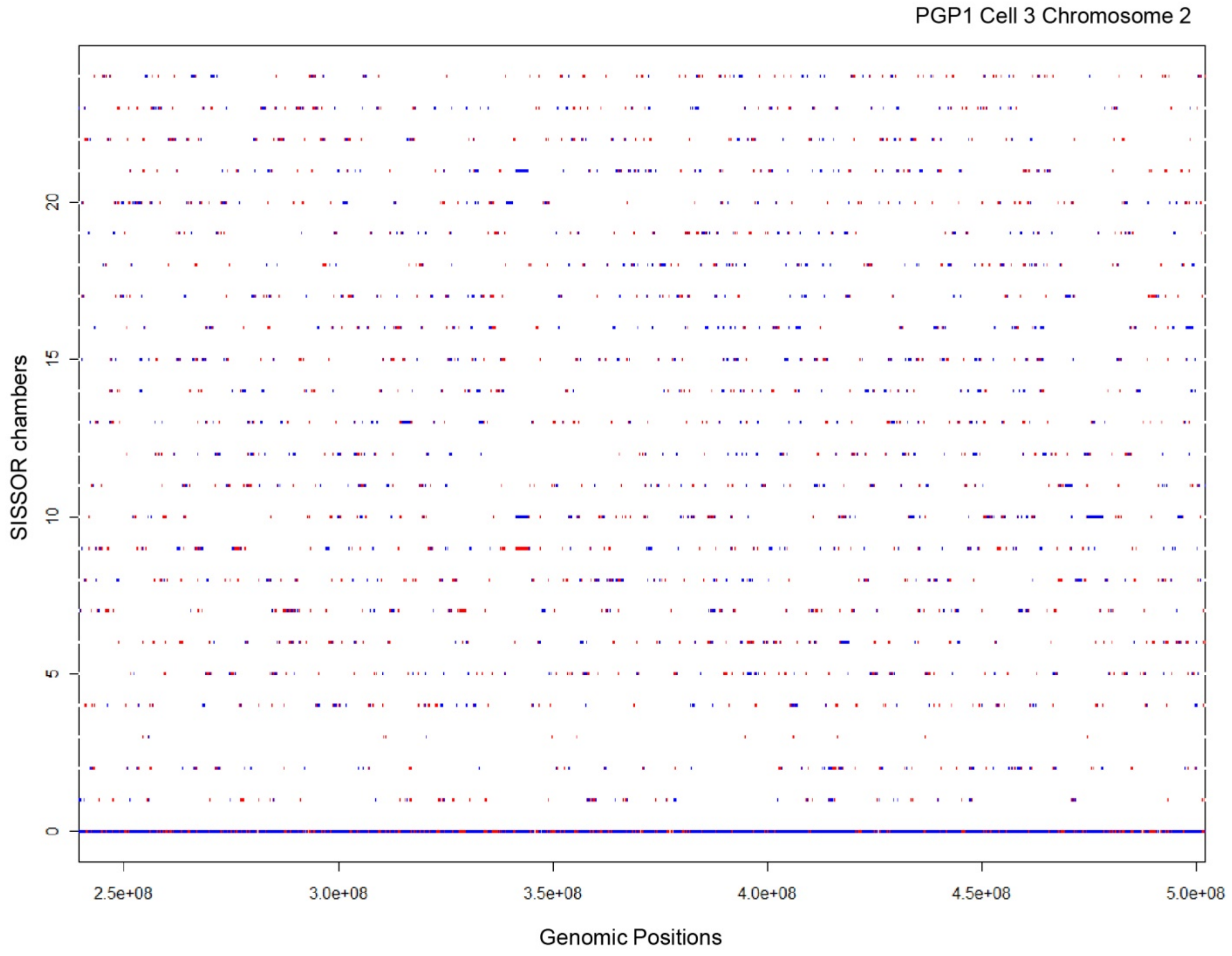

Fig. S19. Distribution of unprocessed sequencing reads on chromosome 2 of PGP1 Cell 3.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 4235 bins, representing chromosome 2, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

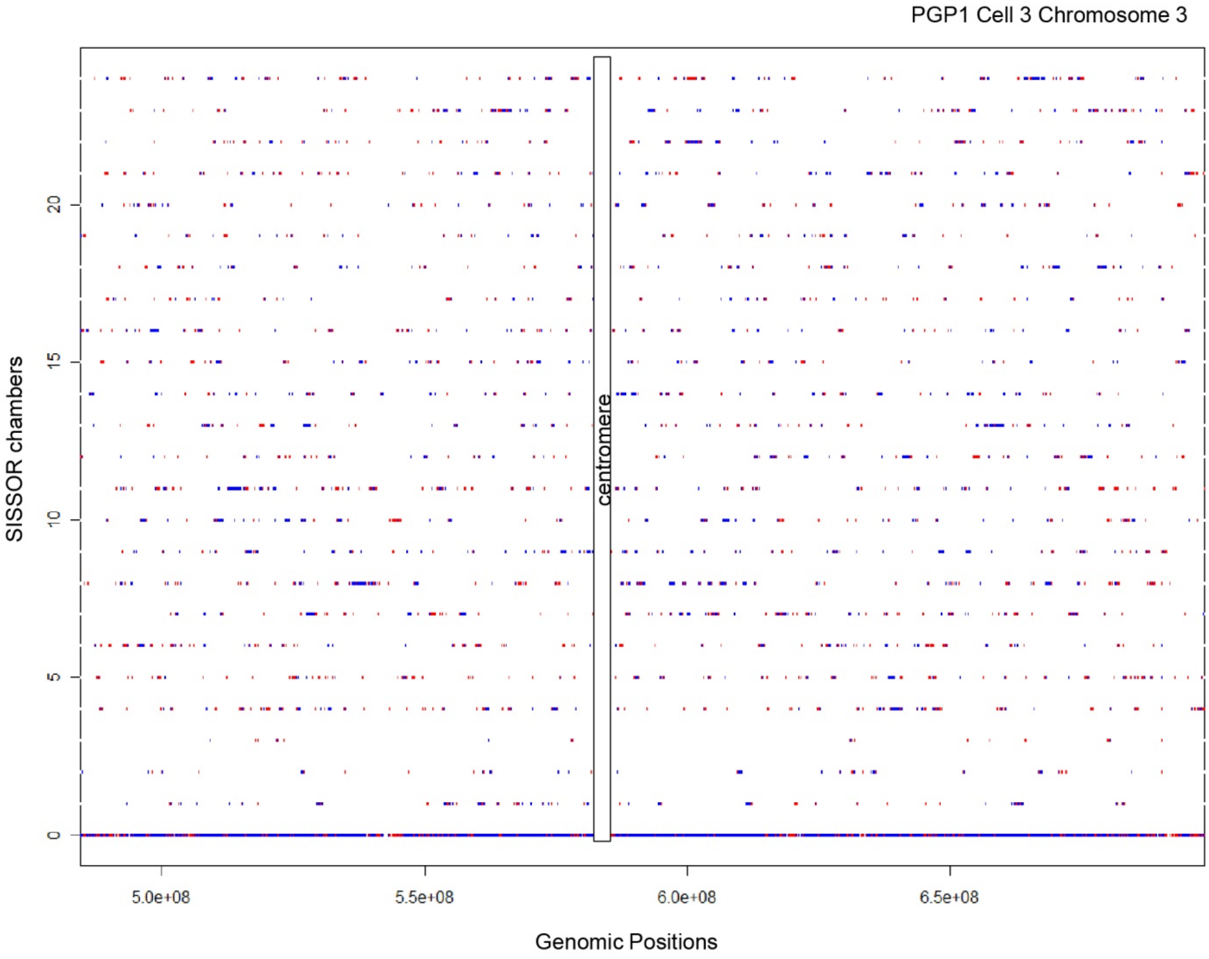

Fig. S20. Distribution of unprocessed sequencing reads on chromosome 3 of PGP1 Cell 3.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3520 bins, representing chromosome 3, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

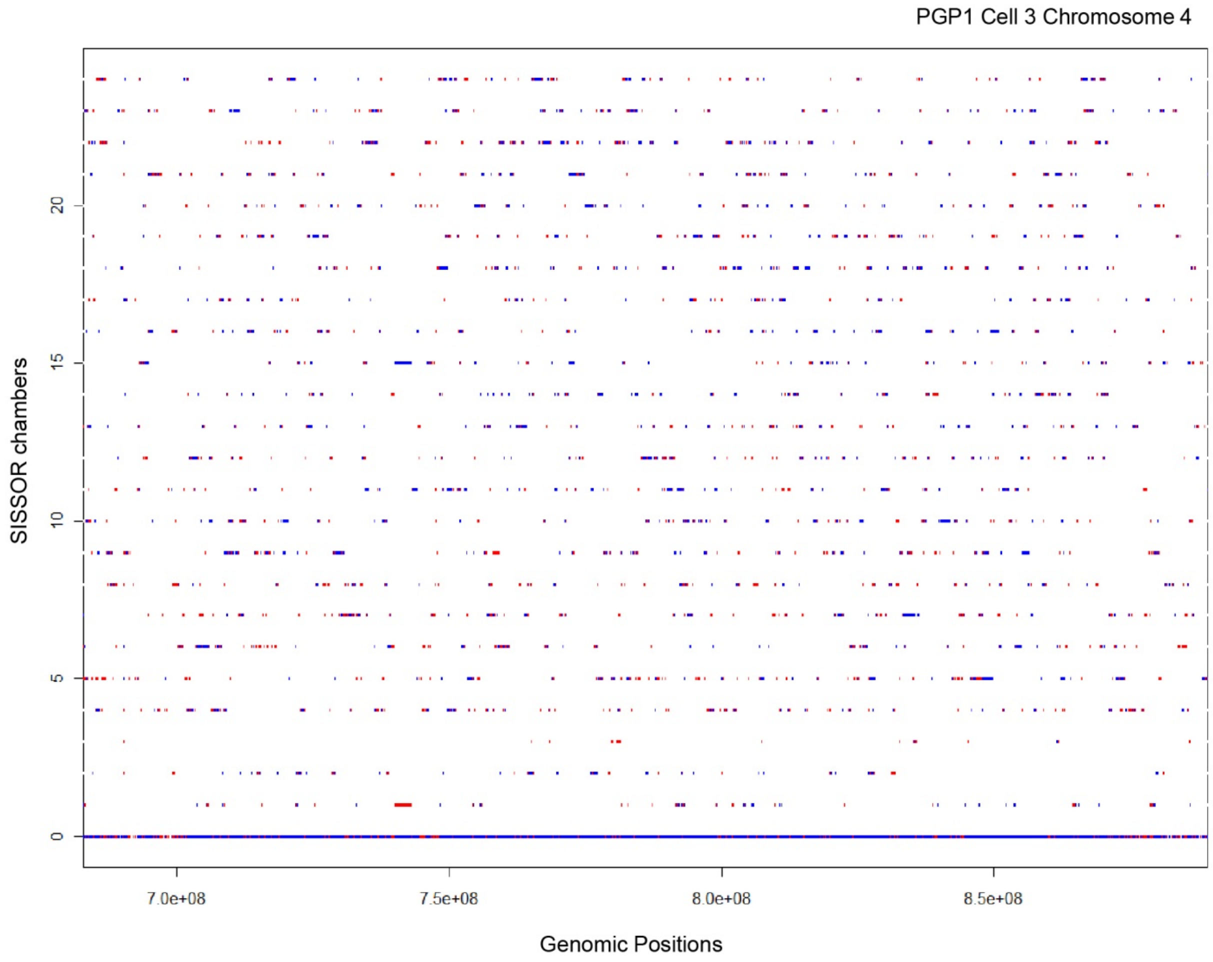

Fig. S21. Distribution of unprocessed sequencing reads on chromosome 4 of PGP1 Cell 3.
The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3373 bins, representing chromosome 4, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

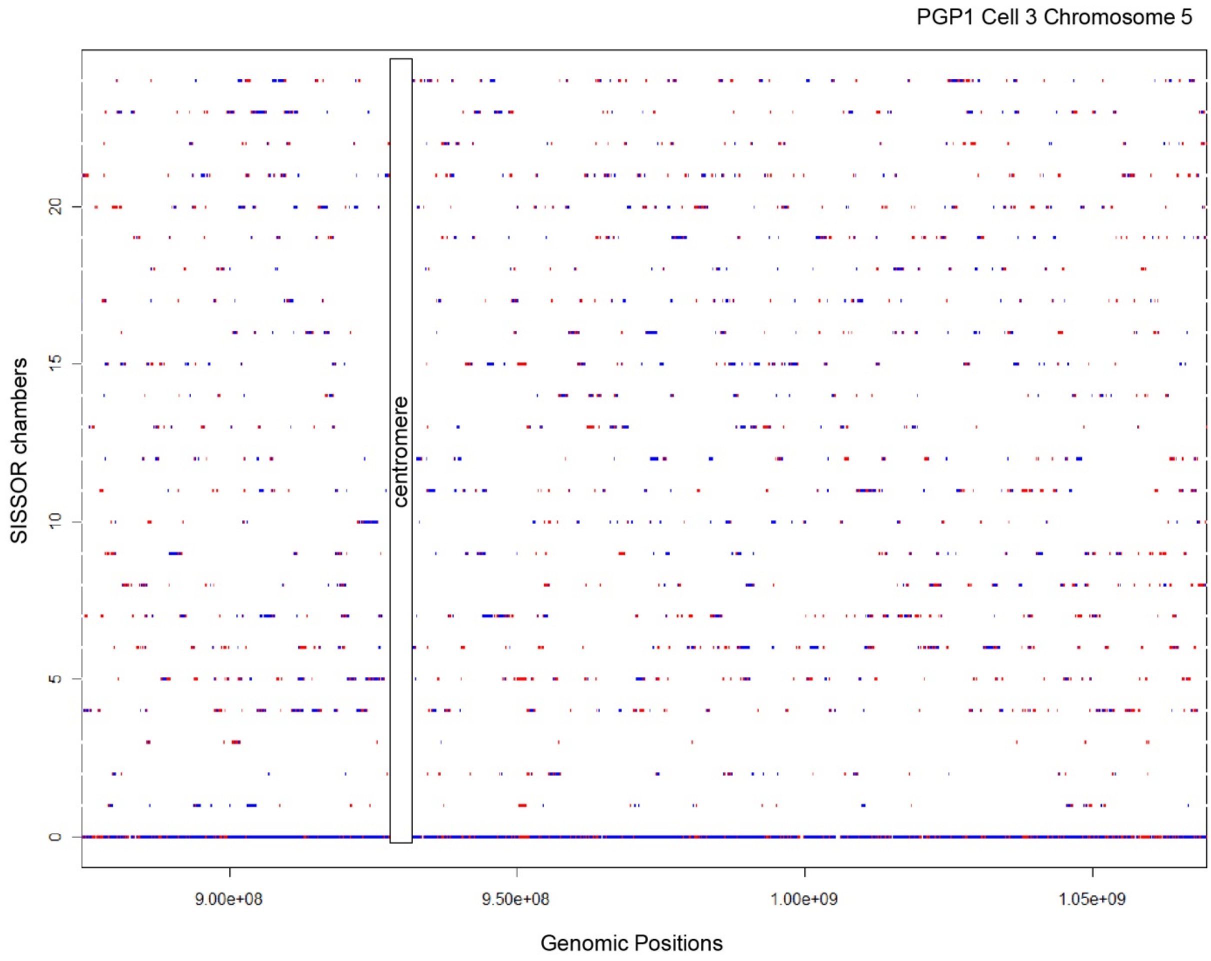

Fig. S22. Distribution of unprocessed sequencing reads on chromosome 5 of PGP1 Cell 3. The entire human genome was divided into 50,000 bins by variable bin method and normalized reads per bin were calculated (2). 3158 bins, representing chromosome 5, were displayed here. Bins in red color represented the normalized number of reads per bin and were subjected to the segmentation algorithm. Bins in blue color represented five times higher than normalized reads per bin. The whole chromosome coverage, where the highest value of each bin at each position, was collapsed into a single track at the bottom (chamber 0).

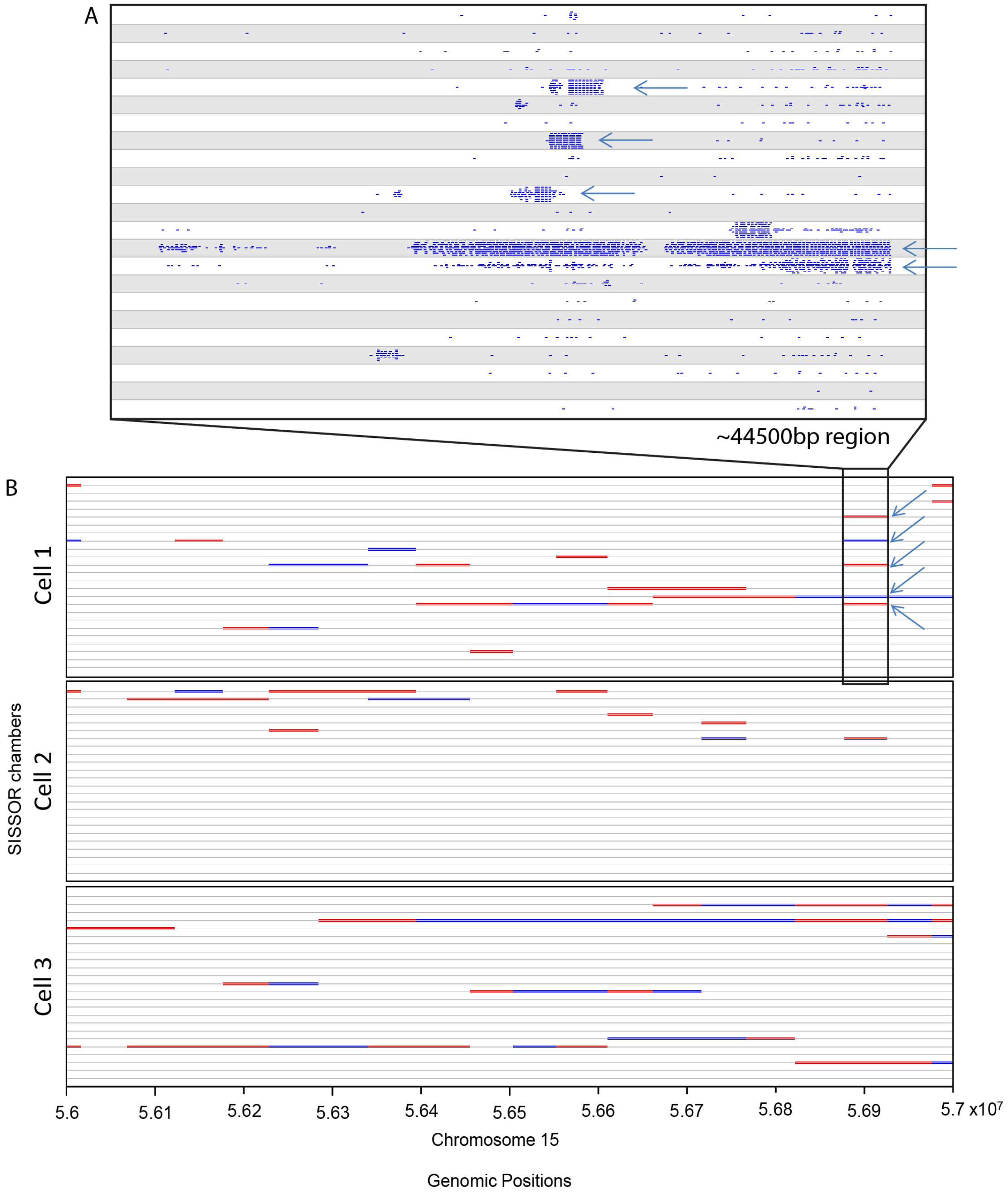

Fig. S23. Distribution of unprocessed sequencing reads in a 1Mbp region on chromosome 15. (A) Mapped locations of unprocessed reads in a single bin were visualized by SeqMonk. (B) Distribution of unprocessed reads per bin was shown in a 1Mbp region on chromosome 15. Bins in red and blue colors represented the normalized and the more than five times higher than normalized reads per bin respectively. Maximum 4-strand coverage at each position was possible in each single cell. In the case of 4+ coverage of a single bin (arrows pointed to the chambers in cell 1), we observed some fragments that were smaller than the bin size and occupied different regions within this single bin. The distribution of these smaller fragments may result in the observation of 4+ coverage in a single bin region.

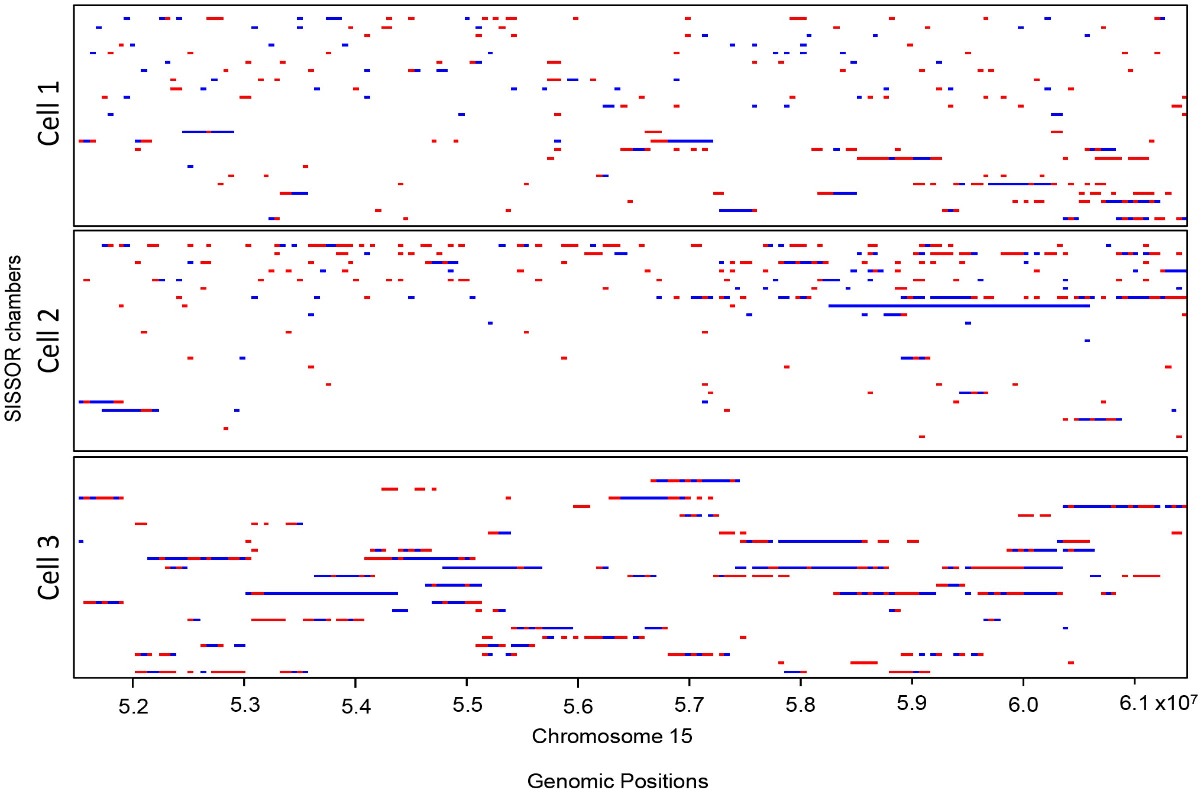

Fig. S24. Distribution of unprocessed sequencing reads in a 10Mbp region on chromosome 15. Zoom-out (10x) view of the same genomic region as shown in Fig. S23. Coverage of some SISSOR fragments was visible from the unprocessed sequencing data.

III. Supplementary Tables

Table S1. Summary of some selected PGP1 fibroblast cells for low depth sequencing

Libraries[1]	Raw Reads	preseq Extrapolated Coverage[2]	Device	Mixing with ALS	ALS [KOH] final	CoRE or Tag[3]	Added	PCR	Mapping Rate	Clonal Rate
PGP1#1	61M	400M	v7	5min	357mM	CoRE	dUTP	11	97.36%	34.7%
PGP1#3	23M	75M	v7	10min	133mM	CoRE	dUTP	9	98.15%	75.7%
PGP1#6	37M	167M	v7	5min	357mM	CoRE	dUTP	11	97.58%	59.6%
PGP1#7	32M	208M	v7	5min	240mM	CoRE	dUTP	11	99.80%	55.4%
PGP1#15	16M	633M	v7	1min	240mM	Tag	X	8+9	91.29%	97.3%
PGP1#18	7M	52M	v9	5min	240mM	Tag	X	8+9	98.87%	70.6%
PGP1#21	47M	641M	v7	10min	357mM	Tag	dUTP	7+7	97.38%	45.3%
PGP1#21	40M	1057M	v7	10min	357mM	CoRE	dUTP	12	97.84%	19.9%
PGP1#22	57M	789M	v7	10min	357mM	Tag	dUTP	5+7	98.80%	21.6%
PGP1#23	43M	626M	v8	10min	340mM	Tag	dUTP	5+7	97.26%	34.0%
PGP1#24	39M	865M	v8	10min	340mM	Tag	N8 primer	5+7	95.96%	23.1%
PGP1#25	47M	517M	v8	10min	340mM	Tag	N9 primer	5+7	96.69%	35.2%
PGP1#26	17M	206M	v8	10min	340mM	Tag	N10 primer	5+7	96.30%	39.5%
PGP1#27	44M	957M	v9	10min	340mM	CoRE	N8 primer + dUTP	12	84.84%	18.1%
PGP1#28	2M	239M	v7	8min	357mM	CoRE	dUTP	12	82.77%	3.6%
PGP1#29	8M	385M	v9	5min	340mM	CoRE	dUTP	12	92.10%	13.0%

[1] Libraries prepared for low depth sequencing. PGP1#21 and PGP1#22 were further sequenced in high depth and renamed as PGP1 Cell 1 and PGP1 Cell 2 in our qualitative analysis.

[2] Numbers of base covered extrapolated at 3 billion total base with preseq (17).

[3] Sequencing library construction using CoRE fragmentation or transposon tagmentation.

Table S2. Summary of Sequencing Data from 3 PGP1 fibroblast cells

	DNA Bases Sequenced (Gb)	Base Coverage on Human Genome	SNP Total Count	Mappable Rate of Sequencing Reads	Unique Rate of Sequencing Reads
PGP1 Cell 1	190	73.6%	1083220	97.3%	60.8%
PGP1 Cell 2	198	54.9%	692831	98.8%	26.5%
PGP1 Cell 3	141	62.8%	664898	92.6%	35.6%
PGP1 Combined	529	94.9%	--	--	--

Table S3. Number of sequencing reads mapped to SISSOR fragments and the human genome in cell 3.

Chamber	Unique Reads in Fragment	All Unique Reads
24	1,813,125	3,101,368
23	2,421,665	3,894,385
22	3,366,394	4,480,834
21	6,060,740	7,525,072
20	7,695,503	9,063,380
19	11,783,636	12,983,541
18	12,556,521	13,785,402
17	11,301,391	12,443,243
16	26,324,802	29,653,811
15	24,114,819	26,419,863
14	27,235,070	30,883,732
13	26,621,859	29,668,057
12	36,945,999	41,482,629
11	31,198,047	33,821,487
10	34,703,295	39,150,198
9	4,039,953	5,201,503
8	31,802,553	35,339,846
7	27,879,285	31,421,882
6	30,459,420	34,661,523
5	26,215,007	29,447,284
4	16,434,969	18,851,221
3	21,832	347,626
2	5,427,028	6,740,287
1	962,272	1,606,947
Total	407,385,185	461,975,121
	Total Positions Removed	54,589,936
	% removed	11.8%

Table S4. Summary of HMM SISSOR fragments from 3 PGP1 fibroblast cells

	Raw DNA Fragments in SISSOR Chambers				
	Number of DNA Fragments	Fragment Average Size (kb)	N50 Fragment Length (kb)	Biggest Fragment Length (Mb)	4-strand Coverage for DNA Fragments
PGP1 Cell 1	8310	415	589	5.1	30.3%
PGP1 Cell 2	7123	658	1040	8.5	41.2%
PGP1 Cell 3	16517	479	727	5.9	69.6%

Table S5. Tabulated data in cross chamber base calling algorithm

Phred-scaled Quality	10	30	50	70	90	110	130	150
Calls above cutoff[1]	2.10E+09	2.10E+09	2.09E+09	2.05E+09	1.33E+09	1.33E+09	1.30E+09	6.98E+08
Calls seen in References[2]	2.05E+09	2.05E+09	2.05E+09	2.01E+09	1.31E+09	1.31E+09	1.28E+09	6.89E+08
Mismatch References[3]	122852	45563	26663	12152	4614	2027	1396	510
SNV Matches	1704182	909054	889605	613669	379925	357653	177096	144378
False Positive Rate[4]	5.47E-05	1.35E-05	5.07E-06	1.42E-06	1.03E-06	5.40E-07	1.31E-07	2.13E-07
False Discovery Rate [4]	6.72E-02	4.77E-02	2.91E-02	1.94E-02	1.20E-02	5.64E-03	7.82E-03	0.000992
Error rate[4]	5.98E-05	2.22E-05	1.30E-05	6.05E-06	3.52E-06	1.55E-06	1.18E-06	8.44E-07

[1] Unique base called in SISSOR.
[2] Regions covered by both SISSOR and the combined coverage of CGI, WGS and BAC references (7, 12, 14).
[3] Base call (SNV or reference) disagreed with CGI, WGS and BAC references.
[4] These are upper bounds for each statistic; Calculated against CGI/WGS+BAC data set as ground truth:

- False Positive Rate = FP / (FP + TN)
- False Discovery Rate = FP / (TP + FP)
- Error Rate = (FP + FN) / (FP + TP + FN + TN)

where:

- FP = called SNV allele, CGI+WGS called 0/0
- TP = called SNV allele, CGI+WGS called 0/1 or 1/1 (same SNV)
- FN = called hg19 reference allele, CGI+WGS called 1/1
- TN = called hg19 reference allele, CGI+WGS called 0/0 or 0/1

Table S6. Summary of Strand-to-Strand mismatch base consensus

	All Cell[1]
Total strand-strand mismatch	10,975
Total strand-strand match	654,210,076
Mismatch rate	1.68E-05

[1] Two strands of identical haplotype matching in any cell. (Unique haploid positions)

Table S7. Summary of error rate analysis from strand-strand consensus

	Same Cell[1]	All Cell[2]	Cross Cell[3]
Total Unique Positions (DP>=5) [4]	153,115,359	461,395,530	355,226,678
Positions included in CGI/WGS reference[5]	150,976,835	455,719,630	351,156,792
SNP counts	118308	357722	273683
Difference to CGI/WGS reference[5]	98	115	19
Difference to CGI/WGS/BAC reference[6]	94	102	9
Difference to CGI/WGS/BAC/3rd chamber[7] and unconfirmed variants	68	72	4
Error rate (upper bound) [8]	4.50E-07	1.58E-07	1.14E-08
False Discovery rate	5.07E-04	1.79E-04	1.46E-05

[1] Two strands of identical haplotype only in the same cell. (Unique haploid positions)

[2] Two strands of identical haplotype matching in any cell. (Unique haploid positions)

[3] Two strands of identical haplotype only in between two different cells. (Unique haploid positions)

[4] SISSOR coverage

[5] CGI/WGS reference coverage

[6] Combined CGI/WGS reference to BAC reference (12)

[7] Internal reference from SISSOR

[8] Maximum error rate in SISSOR

Table S8. Summary of differences in individual cells

	Cell 1	Cell 2	Cell 3
Total Unique Positions (DP>=5) [1]	53,956,666	70,285,423	30,654,766
Positions included in CGI/WGS reference[2]	53,203,331	69,220,980	30,306,948
SNP counts	41400	54832	23477
Difference to CGI/WGS reference[2]	14	75	9
Difference to CGI/WGS/BAC reference[3]	14	71	9
Difference to CGI/WGS/BAC/3rd chamber[4] and unconfirmed variants	14	45	9
Error rate (upper bound) [5]	2.63E-07	6.50E-07	2.97E-07

[1] SISSOR coverage
[2] CGI/WGS reference coverage
[3] Combined CGI/WGS reference to BAC reference (12)
[4] Internal reference from SISSOR
[5] Maximum error rate in SISSOR

Table S9. qPCR primers used for quality control

Name	Forward	Reverse
Human Alu	CTGGGCGACAGAACGAGATTCTAT	CTCACTACTTGGTGACAGGTTCA
Human Mito	CCCCACAAACCCCATTACTAAACCCA	TTTCATCATGCGGAGATGTTGGATGG

IV. Reference

1. Peters BA, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487(7406):190–195.
2. Baslan T, et al. (2012) Genome-wide copy number analysis of single cells. *Nat Protoc* 7(6):1024–1041.
3. O’Connell J, Hojsgaard S (2011) Hidden Semi Markov Models for Multiple Observation Sequences: The mhsmm Package for R | O’Connell | Journal of Statistical Software. *J Stat Softw* 39(4). Available at: https://www.jstatsoft.org/article/view/v039i04 [Accessed August 12, 2016].
4. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma Oxf Engl* 25(16):2078–2079.
5. Li R, et al. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19(6):1124–1132.
6. Paez JG, et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* 32(9):e71.
7. Consortium TEP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
8. Zhao H, et al. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30(7):1006–1007.
9. Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinforma Oxf Engl* 27(15):2156–2158.
10. Edge P, Bafna V, Bansal V (2016) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*:gr.213462.116.
11. Duitama J, et al. (2012) Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res* 40(5):2041–2053.
12. Lo C, et al. (2013) On the design of clone-based haplotyping. *Genome Biol* 14:R100.
13. Kuleshov V (2014) Probabilistic single-individual haplotyping. *Bioinforma Oxf Engl* 30(17):i379-385.
14. Ball MP, et al. (2012) A public resource facilitating clinical use of genomes. *Proc Natl Acad Sci U S A* 109(30):11920–11927.
15. Köster J, Rahmann S (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinforma Oxf Engl* 28(19):2520–2522.
16. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
17. Daley T, Smith AD (2013) Predicting the molecular complexity of sequencing libraries. *Nat Methods* 10(4):325–327.