Cell

Resource

Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm

Jianbin Wang,[1,5] H. Christina Fan,[1,5,6] Barry Behr,[2] and Stephen R. Quake[1,3,4,*]
[1]Department of Bioengineering
[2]Departments of Obstetrics and Gynecology
[3]Department of Applied Physics
[4]Howard Hughes Medical Institute
Stanford University, Stanford, CA 94305, USA
[5]These authors contributed equally to this work
[6]Present address: ImmuMetrix LLC, 552 Del Rey Avenue, Sunnyvale, CA 94085, USA
*Correspondence: quake@stanford.edu
http://dx.doi.org/10.1016/j.cell.2012.06.030

SUMMARY

Meiotic recombination and de novo mutation are the two main contributions toward gamete genome diversity, and many questions remain about how an individual human's genome is edited by these two processes. Here, we describe a high-throughput method for single-cell whole-genome analysis that was used to measure the genomic diversity in one individual's gamete genomes. A microfluidic system was used for highly parallel sample processing and to minimize nonspecific amplification. High-density genotyping results from 91 single cells were used to create a personal recombination map, which was consistent with population-wide data at low resolution but revealed significant differences from pedigree data at higher resolution. We used the data to test for meiotic drive and found evidence for gene conversion. High-throughput sequencing on 31 single cells was used to measure the frequency of large-scale genome instability, and deeper sequencing of eight single cells revealed de novo mutation rates with distinct characteristics.

INTRODUCTION

Gametogenesis is a biological process by which precursor cells undergo cell division and differentiation to form mature haploid gametes. Human gametogenesis occurs by mitotic division of gametogonia, followed by meiotic division of gametocytes into various gametes. During this process, the gamete genome experiences both programmed and spontaneous changes, among which meiotic recombination shuffles the two haploid somatic genomes to create a unique hybrid haploid genome for each gamete cell, while accumulated replication errors contribute point mutations that may affect the gametes' functionality. This results in an enormous variety of new genomes being created in the gametes, thereby enabling one's children to add to the genetic diversity of the human race in a more complex manner than by simply mixing and matching entire parental chromosomes. The genome-wide recombination activity and de novo mutation rate have been directly characterized in many model organisms. However, it has been unclear how an individual human's genome is edited during gametogenesis.

Using pedigree data and statistical methods, deCODE (Kong et al., 2010) and the International HapMap Consortium (International HapMap Consortium, 2005) have been able to create high-resolution recombination maps at the population level. However, such maps only show average results across a population and cumulative results throughout evolutionary history (Jeffreys et al., 2005), and it is not clear what the relationship is between these population maps and the personal recombination processes for any given individual, especially because these focus only on meiotic products that yield successful offspring (Tiemann-Boege et al., 2006). The 1000 Genome Project measured the mutation rate in two family trios (Conrad et al., 2011). However, their results are limited to measuring only a single meiosis per individual, and in general, such an approach probes only viable offspring, is limited by the number of offspring per family, and requires access to parental genome data.

Here, we describe a single-cell whole-genome analysis method to characterize the genomic changes from gametogenesis. Using this technique, we analyzed the whole genomes of >100 single human sperm cells. Recombination data from 91 single sperm cells presented a comprehensive landscape of personal recombination activity. Genome-wide meiotic drive and gene conversion were also directly tested. Single-cell whole-genome sequencing further revealed primary information about human sperm genome instability and mutation rate.

RESULTS

Microfluidic Single-Sperm Whole-Genome Amplification

We developed a strategy to perform parallel analysis of the haploid genomes of many individual sperm cells by employing

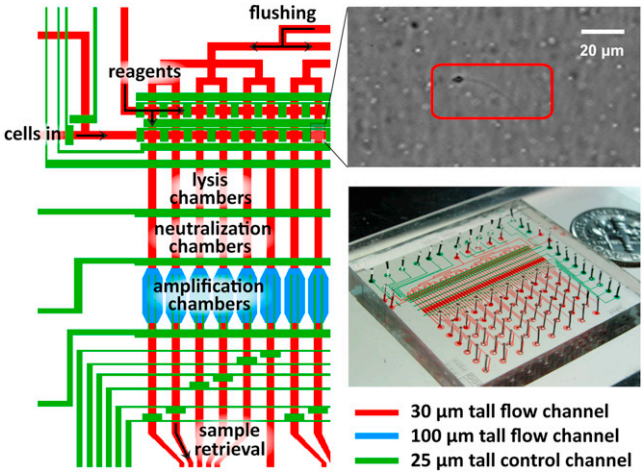


Figure 1. Microfluidic Device Designed for the Whole-Genome Amplification from Single Sperm Cells

Device layout and operation pipeline are slightly modified from a similar device used to measure haplotype. A single sperm cell highlighted by the red square is recognized microscopically and captured in the cross region. In the overview image of the device, control channels are filled with green dye, and flow channels are filled with red dye.

See also Table S1.

single-sperm whole-genome amplification on a microfluidic device (Figure 1). Previously, we used microfluidic automation to perform whole-genome haplotype analysis by amplifying individual chromosomes at a rate of one cell per device (Fan et al., 2011) and demonstrated high-fidelity single-chromosome amplification. We have now extended that principle both in parallelization and in complexity of the starting material. The device described here enables the random dispensing of cell aliquots into 48 separate chambers, leading to typically half of them holding exactly one cell. We performed high-fidelity amplification of the entire genome in each chamber, followed by whole-genome genotyping and high-throughput sequencing analyses.

We collected a sperm sample from a 40-year-old Caucasian individual (P0) whose genome has been sequenced (Pushkarev et al., 2009), clinically annotated (Ashley et al., 2010), and haplotype phased (Fan et al., 2011). The patient has healthy offspring and normal clinical semen analysis results. Before the amplification reaction, we verified which microfluidic chambers held sperm cells with optical microscopy (Figure 1). With the products of each of the 125 single-cell amplification attempts, we performed 46 loci genotyping TaqMan PCR to evaluate the amplification performance (a total of 5,750 PCR reactions, a subset of which is shown in Figure 2A). Across the 125 samples, the mean call rate is 76.5% (4,398 out of 5,750), and 98 samples yielded call rates >70%, indicating effective whole-genome amplification (Figure 2B). Eight samples gave signals in <30% of the PCR assays (Figure 2A, chamber 11), suggesting amplification failure or misidentification of sperm cells by imaging. Because of the haploid nature of sperm cells, amplification products from single sperm cells should give only homozygous genotyping results, regardless of the polymorphism status of the diploid genome. As expected, 99.4% of the positive PCR reactions yielded signals from only one allele, and the allele combinations from multiple amplification products at each position match the genomic genotype at that locus. The 26 heterozygous calls (0.6% of 4,398) reside in 11 of the 125 single-cell experiments (ranging from 1 to 7 per cell), and we interpreted these heterozygous calls as the consequence of multiple cells in the chamber or other DNA contamination (Figure 2A, chamber 23). These results show that it is possible to obtain large numbers of high-quality single-cell genome amplification products by using an automated microfluidic device, and the products can be used for downstream genomic analysis (Table S1 available online).

Whole-Genome Sperm Typing from 91 Single Cells Gave a Personal Recombination Map

We selected 93 amplification products with high yield and no heterozygous genotyping calls for an additional round of MDA, followed by Illumina Omni1S whole-genome genotyping (Table S1). Each single cell yielded successful calls at ~30%–50% of the 1.2 million SNPs tested (Figure 2C), of which 83.2% were called as homozygous. The lower call rate on the bead array as compared to genotyping PCR is due to amplification bias from MDA. The abundance variation across different regions of the genome exceeds the dynamic range of microarray, and the underrepresented loci are not detected. TaqMan PCR, which has much larger dynamic range, gave >70% call rate, and this reveals the true extent of coverage of the amplification products. The heterozygous false positive rate is due to similar effects. Within the 0 to ~3 Illumina signal intensity spectrum, the mean intensity of homozygous calls was 1.27, whereas the mean of heterozygous calls was 0.12, which is barely above the default noise cutoff value of 0.1. These results, together with those from qPCR, reveal that the heterozygous calls are false positives due to low signal intensity. To improve the genotyping accuracy, we applied a stringent noise cutoff on the raw genotyping calls to remove the low-intensity signals and hence eliminate the heterozygous calls.

By mapping the genotyping results from each sperm cell to the two somatic haplotypes obtained by microfluidic direct deterministic phasing (DDP) of single lymphocytes (Fan et al., 2011), we detected single-chromosome deletions in two cells (Figure 5A), whereas the other 91 cells gave a total of 2,075 autosomal crossover events (22.8 ± 0.4 SE [±3.7 SD] in each sperm) (Figure 2D and Table S2). The sizes of crossovers range from a few hundred base pairs to >1 Mbp, with 59%, 37%, and 13% of the total events localized to intervals of 200 kb, 100 kb, and 30 kb, respectively, comparing to 70%, 51%, and 20% from previous Hutterite pedigree data for the same intervals. The fact that P0 has a low number of heterozygous loci in the genotyping panel, in combination with the genotype calling rate, contributed to the slightly lower resolution of our data. The collection of all of these recombination events yields a personal recombination map for P0. To our knowledge, this is the first reported high-resolution genome-wide personal recombination map for an individual.

Personal Recombination Map Recapitulates Population Results at a Broad Scale

At a genome-wide scale, the recombination rate of 22.8 ± 0.4 SE (±3.7 SD) events per cell agrees well with the average male

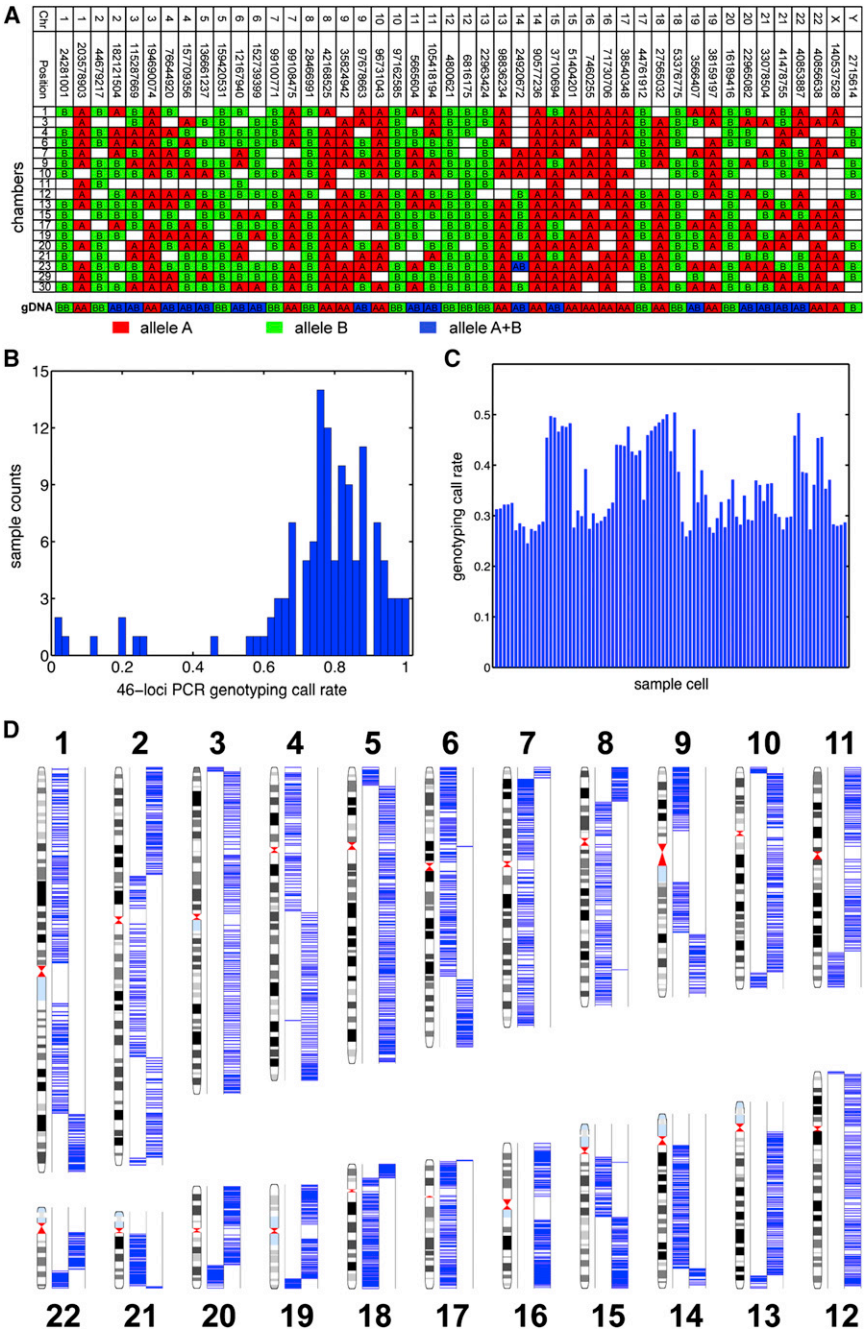


Figure 2. Whole-Genome Single-Sperm Typing

(A) Evaluation of amplification performance using 46 loci PCR. This table represents results from a subset of sperm cells being amplified. Each row represents the content from a microfluidic chamber, and each column represents a locus, with specified chromosome number and coordination (NCBI b36). The genotypes of genomic DNA control are also shown. The two alleles of a SNP are highlighted in red and green. Heterozygous loci are labeled in blue. Sample 11 shows a genotyping profile similar to no-template WGA control, indicating misidentification of sperm cell before amplification. Sample 23 shows heterozygous genotype on chromosome 14 and sex chromosome, suggesting multiple cells during amplification.

(B) 46 loci PCR genotyping call rates.

(C) Whole-genome genotyping call rates of 91 single sperm samples from Illumina HumanOmni1S Bead Array.

(D) Detection of recombination from a single sperm sample. The two columns in each chromosome represent the two somatic haplotypes, and blue lines show the genotyping calls of heterozygous SNPs from the sample. Each switch of haplotype block indicates a recombination event.

See also Table S2.

results implied from other methods, such as cytological imaging (49.8 ± 0.4 SE [±4.3 SD] MLH1 loci within the tetraploid spermatocytes [Sun et al., 2004] and data inference (24.0 ± 0.2 SE [±2.7 SD] from Caucasian pedigrees [Cheung et al., 2007]). The slightly lower recombination level in P0 is consistent with P0's genotype of *RNF212* (T/T at rs3796619), which is associated with a 5% lower recombination level than average (Kong et al., 2008). When comparing the number of recombination events within each chromosome, we found similar discrepancies between chromosome length in base pairs and recombination rates (Table S3), as has been previously reported by both cytological and pedigree studies (Sun et al., 2004; International HapMap Consortium, 2005).

Nonuniformity in the probability of recombination events also occurs within each chromosome. Our data show telomere-weighted distributions that are qualitatively similar to those found in population genetics studies (Kong et al., 2010; Myers et al., 2005). With a 5 Mb window size, we detected a correlation of 0.85 between P0 and deCODE male data and 0.76 between P0 and HapMap data, whereas the correlation between deCODE male and HapMap data is 0.85 (Figures 3 and S1). We observed an 87 Mb median distance between adjacent recombination events, comparing with the 49 Mb expected value after we randomly shuffled the recombination events (permutation test, $p < 10^{-4}$), which demonstrates positive recombination interference, as has been previously observed (Sun et al., 2004). Taken together, P0's personal recombination map shows that recombination events within an individual recapitulate the general broad-scale features from population data. Our results experimentally demonstrate general concordance between an individual and the population average, which can be thought of as an analogy to the ergodic principle from statistical physics.

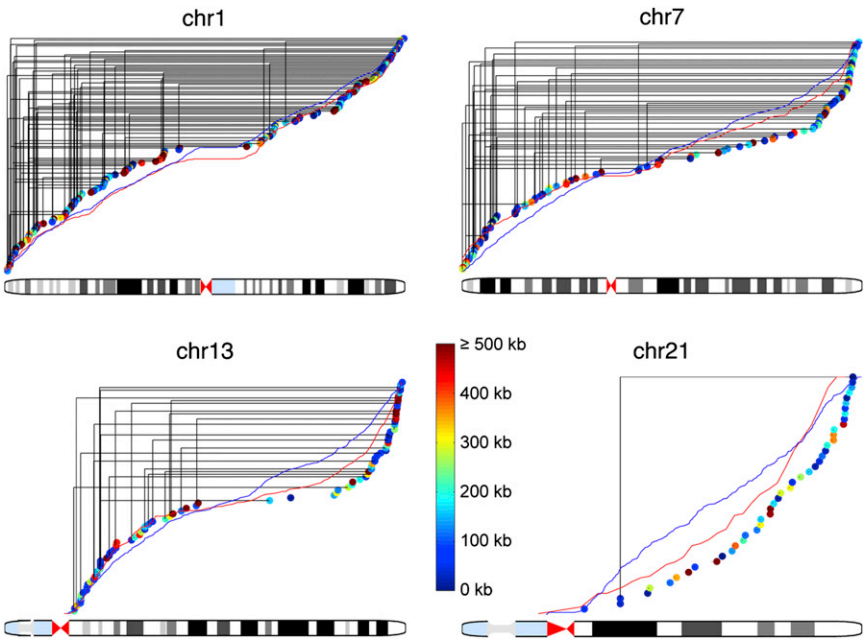


Figure 3. Recombination Map from Chromosomes 1, 7, 13, and 21

Each dot represents a recombination event, with color code for resolution. Solid black lines connect recombination events from the same sperm cell. Red and blue lines show the cumulative recombination rates from deCODE (male) and HapMap, respectively.

See also Figure S1 and Table S3.

High-Resolution Analysis Revealed Personal Specific Recombination Activity

When one compares our results and the population data at higher resolution, differences emerge. The telomere-weighted bias is stronger in our results than in HapMap or deCODE data, resulting in large regions without recombination near the centromere (Figure S1). For example, no recombination was detected within any ~8 Mb region symmetrically crossing the 17 metacentric chromosome centromeres in P0 (p value 0.028 based on deCODE male data). The relative activities on the p arms of some chromosomes are also higher than population-wide results (Figure S1). These differences suggest some potential individual-specific features that may be diluted by population-wide averaging, and we therefore performed a more extensive comparison at a finer scale. A sliding window of 2 Mb was applied to P0's recombination map with 1 Mb increments, and the resulting windows for which P0's recombination rate was at least triple the genome-wide average (3 cM/Mb) were compared with deCODE male activity. Within the total of 66 such windows, 3 showed significantly higher activity than the deCODE male data in the corresponding regions. We refined the boundaries of these regions and summarized the activities in Table 1 ("Sliding Window Scanning").

Both the deCODE and HapMap projects have made extensive catalogs of recombination hot spots at the population level (Kong et al., 2010; International HapMap Consortium, 2005). Previous sperm studies have demonstrated that some particular hot spots are used idiosyncratically among individuals but have not had the ability to measure genome-wide activity for an individual (Tiemann-Boege et al., 2006; Webb et al., 2008). Data from a Hutterite pedigree suggested interindividual variation in hot spot usage (Coop et al., 2008) and supported a hypothesis that the meiosis-specific histone methyltransferase PRDM9 may act as a universal regulator for recombination distribution (Baudat et al., 2010). Polymorphisms in *PRDM9*, to some extent, correlate with the level of historical hot spot usage. However, the small number of meioses that each individual has in the pedigree data, as well as uncertainty from statistical haplotype inference, led to extensive overlapping of the hot spot usage percentage between individuals (95% confidence interval of single measurement covering ±25%–40%). Consequently, the power of PRDM9 explaining hot spot usage variation is still under debate.

Sanger sequencing showed that P0 has the homozygous A/A *PRDM9* genotype (allele naming from Baudat et al., 2010), which correlates with the highest historical hot spot usage. We employed the likelihood method from the Hutterite study (Coop et al., 2008) on the portion of P0's recombination data that matched their criteria (specifically, the 274 events with 30 kb or smaller size) and determined that only 58% of P0's recombination events coincide with HapMap hot spots. The ten times larger sample size in our data led to higher accuracy than the previous results, revealed by our 95% confidence interval of hot spot overlap fraction as ±10%. These high-accuracy measurements of P0's usage of historical hot spots reveals that, even with the most active and hot-spot-correlated variant of PRDM9, an individual still generates a substantial proportion of recombination events outside of historical hot spots.

We then analyzed the reference human genome for the PRDM9 13 bp degenerate DNA sequence motif, which was previously shown to be enriched in HapMap hot spots (Myers et al., 2008, 2010). The motif is significantly ($p < 10^{-3}$) enriched in P0 recombination regions compared to the genome background. However, 50 out of 162 recombination regions smaller than 30 kb do not contain the motif. When we focused on recombination smaller than 10 kb, the enrichment was not significant ($p = 0.29$) due to the low motif occurrence. We performed a de novo motif search within those regions without the 13 bp motif. All five hits reside in transposon sequences and are significantly enriched in P0's recombination regions ($p < 0.05$ by simulation). This is consistent with the PRDM9 motif, which is also often located in transposon regions. These results suggest that PRDM9 binding may not be directly required for recombination, and other regulatory mechanisms may exist, such as homologous DNA pairing within transposons.

Among the 2,075 recombination events in P0, 940 overlap with at least one another event. These 940 overlapping events form

Table 1. Examples of Individual Specific Recombination Active Regions in P0

Chr	Start	End	Size (kb)	P0 Events	deCODE Male (cM)/p Value[a]	HapMap Sex-ave (cM)/p Value[a]	P0 Activity by Allelic PCR (cM)	P0 Activity by 2 Loci Typing (cM)
2 Mb Sliding Window Scanning (66 Comparisons)								
20	58,753,528	59,651,027	897	12	$0/< 4.7 \times 10^{-15}$	$3.3753/3.8 \times 10^{-3}$	—	—
10	20,012,520	21,553,899	1541	6	1.175/0.049	$0.7817/5.7 \times 10^{-3}$	—	—
11	132,525,133	133,419,195	1393	14	3.97[b]/9.5×10^{-4}	$2.3364/1.9 \times 10^{-6}$	—	—
P0 Self-Overlapping Sets (324 Comparisons)								
16	7,988,699	7,990,230	1.5	2	0.12/1.00	0.0407/0.21	0.77	0.84
9	1,864,696	1,868,831	4.1	2	0.019[b]/0.05	0.1090/1.00	—	0.81
20	58,753,528	59,651,027	897	12	$<1 \times 10^{-6}/< 1.5 \times 10^{-12}$	3.3753/0.019	—	—
15	21,445,700	21,732,000	286	3	$<1 \times 10^{-6}/< 1.5 \times 10^{-12}$	1.2007/1.00	—	—
Non-Hot-Spot Overlapping[c] (135 Comparisons)								
3	197,249,108	197,250,198	1.1	1	0.0035[b]/1.00	0.0027/0.33	ND[d]	—
4	18,404,324	18,406,601	2.3	1	$<1 \times 10^{-6}/< 1.2 \times 10^{-4}$	0.0072/0.88	1.0	1.2

See also Figure S2.

[a]p value calculated as the chance of expecting equal or more recombination events from historical data than in P0 from 91 meioses using binomial statistics; further adjusted with Bonferroni correction.

[b]2002 deCODE data.

[c]Only showing those with sperm typing results.

[d]Recombination activity not detected in P0.

324 distinct sets, with 2–14 overlapping events in each set. A simulation based on HapMap activities showed a significantly higher level of self-overlapping in P0 (permutation test, p value = 0.001), suggesting that these recombination clusters are new hot spots. To confirm that P0 does have high recombination activities within these regions, we selected two regions with manageable sizes for allelic PCR and 2 loci digital haplotyping (Figure S2) and independently verified their high activities in P0 (Table 1, "Self-Overlapping Sets"; Chr16: 7,988,699–7,990,230 and Chr9: 1,864,696–1,868,831). By comparing to the deCODE male data, we found that most of these clusters are also active in the population. However, three regions showed significant higher activities than deCODE (Table 1, "Self-Overlapping Sets"). Considering the small number of recombination events that we detected in P0 comparing with the historical hot spots pool, such a high level of overlap demonstrates P0's preference for only a subset of historical hot spots.

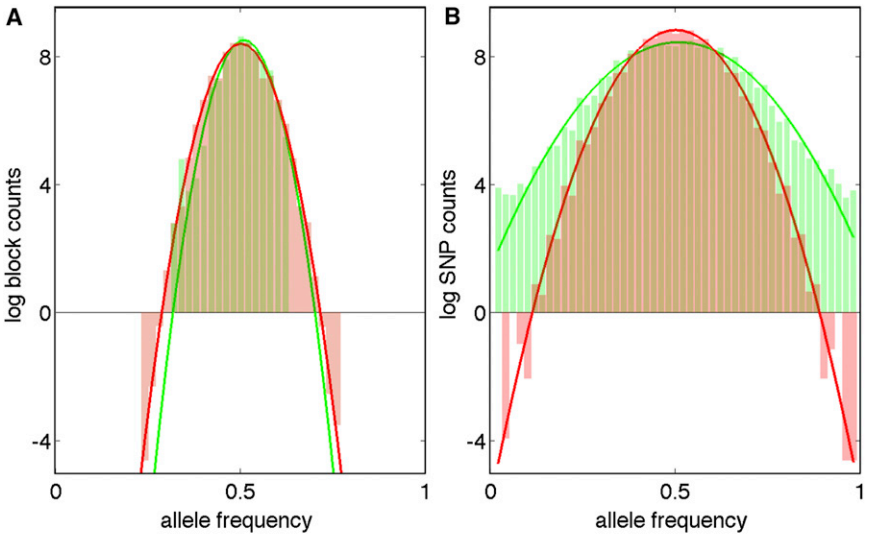
Of P0's recombination events, 135 do not overlap with any HapMap hot spots. Despite being all singlets, 38 of these events showed statistical significance relative to the activities measured in the deCODE male data, even after multiple comparison adjustment. Such a set as a whole is likely enriched with new recombination spots that can serve as targets for further analysis with traditional sperm typing methods. To demonstrate this, we selected two further regions for allelic specific PCR sperm typing (Figure S2) and discovered that one of them is a new personal hot spot (Table 1, "Non-Hot-Spot Overlapping"; Chr3:197,249,108–197,250,198 and Chr4:18,404,324–18,406,601).

Meiotic Drive and Gene Conversion

Mendel's laws propose that the two alleles at a genetic position are transmitted to offspring with equal probability. However, results from specific regions and the whole genome have suggested transmission biased toward one allele (Williams et al., 1993; Zöllner et al., 2004), an effect that can, in part, be explained by the phenomenon of meiotic drive and that can be directly tested in our data.

We first investigated whether the meiotic drive happens at the whole-chromosome level. Because of the general absence of recombination near centromeres, we can accurately define the haplotype across these regions, where kinetochores assemble for mechanical segregation. None of the 22 autosomes had a transmission ratio that significantly deviated from an equal distribution ($p > 0.7$, binomial distribution). Pearson's correlation test between different chromosomes did not detect any cotransmission of centromere haplotypes. Then, we divided the whole genome into 100 kb haploblocks and studied whether any block showed meiotic drive. Even though many blocks had some evidence for bias, none of them reached genome-wide significance level (Figure 4A). Together with the centromere data, our haplotype block results demonstrate that meiotic drive does not appear as a large haplotype. We then turned to measure the transmission ratio of individual SNPs, where we found an obvious difference between our data and simulations of equal transmission (Figure 4B). A putative reason for this pattern is gene conversion.

Meiotic gene conversion is the transfer of information between homologs without reciprocal recombination. Although effectively contributing to genome diversity equivalently as two closely spaced recombination crossovers, gene conversion is less well studied in humans due to its small size relative to genetic marker density. Gene conversion at specific loci has been studied by sperm typing and population genetics data (Gay et al., 2007; Jeffreys and May, 2004), but direct whole-genome measurements have not been conducted for humans.



C

cell	23	24	27	28	101	113	135	136
recombination	24	21	21	18	30	27	27	27
gene conversion	10	10	5	13	6	10	15	12
SNP typed	68690	71931	68594	69828	48387	43096	48583	46591

Figure 4. Meiotic Drive and Gene Conversion

(A and B) Allele frequency histograms of 100 kb haplotype blocks (A) and individual heterozygous SNP (B). Green columns represent experiment data, and red columns represent simulation results assuming no transmission distortion. Solid lines are normal distribution fitting results in log scale.

(C) Gene conversion statistics of single cells.

See also Table S4.

As shown in Figure 2D, some SNPs have genotypes that are opposite to the haplotype in which they resided; therefore, they serve as good candidates for gene conversion detection. To eliminate potential errors in genotyping, we performed high-throughput sequencing on eight of these cells (Table S1, samples 23, 24, 27, 28, 101, 113, 135, and 136). Six to eight × average coverage was obtained with Illumina 2 × 100 read pairs from each sample, covering ~30%–50% of the haploid genome. The less than expected physical coverage based on Poisson statistics is mainly due to amplification bias from MDA. Because the sperm genomes are haploid, one can make highly confident allele calls with substantially lower coverage than the 30 × standard genome sequencing depth. To test the accuracy of this genotype calling method, we performed quality control analysis by mapping the sequencing data to the two P0 somatic haplotypes. We correctly detected 184 of the 193 crossover events in these eight cells without false positives, and the nine missing events all reside near the tips of the chromosomes and had low sequencing coverage.

For each gene conversion candidate SNP covered by high-throughput sequencing, we compared the genotypes of the same SNP across different single cells as well as P0 genomic DNA sequencing data (Pushkarev et al., 2009) to confirm genotyping and haplotyping accuracy. From the 568 candidates, we confirmed 90 converted SNPs (Table S4). Most gene conversions presented as single SNP, whereas five groups of nearby SNPs gave gene conversion regions whose sizes range from 1 to 22 kbp. This size range is comparable to what was found in yeast (Mancera et al., 2008), but not in human (Jeffreys and May, 2004). More interestingly, when we aligned the converted SNP to historical recombination hot spots, only 10 out of the 90 SNPs reside in hot spot regions. This is substantially different than the 58% hot spot overlapping of P0 recombination events. We did not find a strict relationship between gene conversion and recombination level, but generally, cells with more crossovers ended with fewer gene conversions or vice versa (Figure 4C).

Single-Cell Sequencing Results Reveal Sperm Genome Instability

We chose eight sperm cells that clearly passed the SNP-PCR assay (“normal”) and 23 further sperm cells that had marginal or failing scores on the assay (“abnormal” and not within the 93 samples for the recombination study) for high-throughput sequencing (Table S1) and obtained 0.02 × coverage of the genome. After mapping the sequence reads to the human reference genome, we found a discrete distribution of relative sequencing tag density in each chromosome in which chromosomes were typically either present at a uniform level or completely absent (Figure 5B and S3). All eight of the “normal” cells and 17 “abnormal” cells exhibited such patterns with one of the two sex chromosomes missing, and another four “abnormal” cells had clear aneuploidy. Two cells displayed complex, continuous distributions of chromosome representation (Figures 5B and S3). Additional genotyping results confirmed the sequencing findings. The results of these six abnormal cells cannot be explained by the known bias mechanisms in MDA (Marcy et al., 2007), and our previous study on single-chromosome amplification showed no bias for particular chromosomes or sharp coverage drops in any region (Fan et al., 2011). Therefore, the most likely source of missing sequencing reads in the present results is genomic abnormality in the individual sperm cells. The six abnormal samples (Figure 5B), together with the other two samples from the recombination analysis (Figure 5A), represent ~7% of the 116 single cell amplifications with high-resolution analysis, which agrees with literature results on aneuploidy of ~2%–10% measured with FISH (Luetjens et al., 2002; Macklon et al., 2002).

Human reproduction is well known to be inefficient, with monthly fecundity rates of only 30%–40%, and a large number of conceptions fail before the women are aware of the pregnancy (Macklon et al., 2002). This early determination of pregnancy fate was further confirmed by results showing the ability to predict embryo development by the four-cell stage, before embryonic genome activation (EGA) (Wong et al., 2010). The importance of cytokinesis dynamics in embryo development strongly suggests genome integrity as a key factor, as genome instability

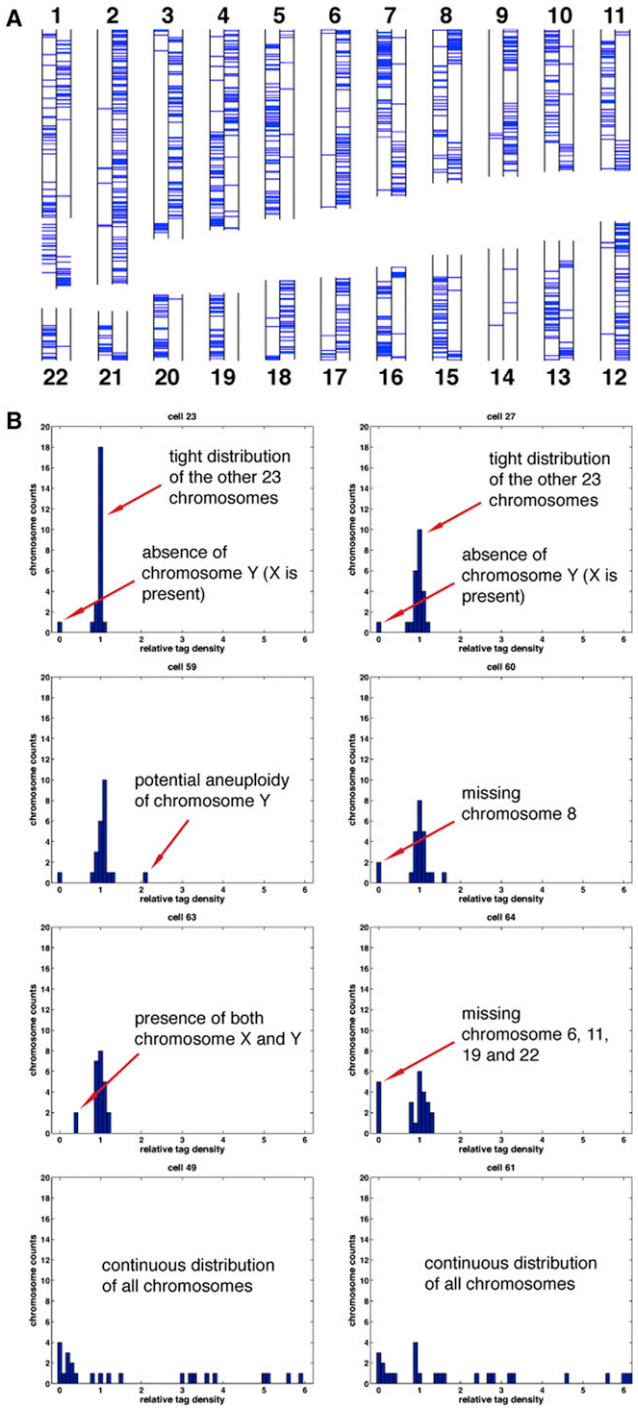


Figure 5. Germline Genome Instability

(A) Whole-genome genotyping results of cell 112. Two columns in each chromosome represent the two haplotypes, and each horizontal bar shows the genotype of a SNP. Chromosome 14 showed very low call rates, suggesting its complete deletion.

(B) Cells 23 and 27 are shown as normal controls, with 23 chromosomes clustered by normalized tag density and one sex chromosome dropped. Cells 59, 60, 63, and 64 had whole-chromosome aneuploidy. Cells 49 and 61 displayed complex, continuous distributions of chromosome representation. See also Figure S3.

will induce cell-cycle arrest. Although the aneuploidy rate for oocytes (20%–30%) is higher than that found in sperm (2%–10%), male genome defects are still a significant contribution to conception failure. Even if the embryo does develop correctly, gamete genome abnormality may impose increased risk to certain diseases. For example, the large-scale deletion of chromosome 13 long arm (13q) found in two of our sperm samples (Figure S3B) may induce 13q deletion syndrome with malformations of craniofacial region and skeletal abnormalities (Quélin et al., 2009).

De Novo Mutations in Primary Sperm Cells

Sequencing data from the gene conversion study also offered the opportunity to measure de novo germline mutations. The recombination detection performed above demonstrated robust genotyping by single-cell sequencing, and we further evaluated the error rate for mutation detection. We selected high-confidence homozygous positions in the P0 somatic genome based on previous sequencing and genotyping (Pushkarev et al., 2009) and calculated the first alternate allele calling frequency in sperm sequencing reads at the same positions. Histogramming these frequency data revealed a decreasing number of positions extending from the perfect agreement side of the discordance axis (Figure S4A). This long tail of background noise represented an amplification/sequencing error rate of 2.7×10^{-4} per read per position.

A distinct group of loci with 100% discordance with somatic DNA clearly stands out from the amplification error background (Figure S4A). These data are not statistically consistent with any of the measured amplification or sequencing errors and are strong candidates for de novo mutations in the sperm. After excluding signals from repetitive regions or with low alignment confidence, we detected 25–36 candidate point mutations in each sperm cell (Tables 2 and S5). We selected 19 mutations for PCR-Sanger sequencing and were able to obtain PCR products from 16 regions. The Sanger results from these 16 regions all confirmed our original calls, thus ruling out the possibility of sequence or alignment errors. Because these loci are inconsistent with the statistical distribution of amplification errors, we conclude that they are de novo mutations.

P0's mutation rate ($2–4 \times 10^{-8}$) is higher than that obtained from genome-sequenced pedigree data ($\sim 1 \times 10^{-8}$) (Conrad et al., 2011), but it is consistent with evolutionary studies, which have revealed ~4–5× more mutations in male than in female, possibly due to the larger number of germline cell divisions in male (Crow, 2000; Makova and Li, 2002). The results from the pedigree study identify the variation of germline mutation levels transmitted to each offspring but are not able to identify the source of such variation. Our results from the eight individual sperm cells have a high degree of internal consistency between their respective mutation levels (Figure S4B), which suggests inter- rather than intraindividual variation. Within each cell, most mutations reside in intergenic or intronic regions (Table 2). However, we detected three missense mutations, a category that was not observed in the pedigree genomes. The transition-to-transversion ratio of P0 mutations is 5.6, as compared to a population average of 2.1. The main reason of more transition than transversion is generally thought to be deamination of

Table 2. Sperm Mutation Properties

Sperm ID	23	24	27	28	101	113	135	136
Physical coverage (Mbp)	931	987	966	954	1272	1231	1250	1382
Mutation counts	36	30	35	33	27	34	25	31
Mutation rates ($\times 10^{-8}$)	3.8	3.0	3.6	3.5	2.1	2.8	2.0	2.2
Ts:Tv	11.0	6.5	4.0	2.7	12.5	7.5	11.5	3.4
CpG	0.06	0.07	0.09	0.15	0.04	0.09	0	0.10
Coding-missense	1	0	0	1	0	1	0	0
Coding-synonymous	0	0	0	0	0	0	0	1
UTR	0	0	1	0	1	0	0	0
Noncoding genes	0	1	0	0	0	0	0	0
Intronic	18	7	16	6	13	16	9	11
Intergenic	17	22	18	26	13	17	16	19

See also Figure S4 and Table S5.

methylated cytosine, primarily at CpG and potentially in other sequence contexts. The higher level of transition we observed is consistent with this, as 21% of C→T mutations correlated with CpApG, though only 8% were at CpG sites.

DISCUSSION

Despite the advances in personal genomics thus far (Ashley et al., 2010; Levy et al., 2007; Pushkarev et al., 2009; Wheeler et al., 2008), gamete genome variation within individuals, especially fine-scale personal recombination activity and germline mutation rates, has been as yet generally inaccessible. Bulk analysis of sperm cells with PCR offers high-resolution and sensitivity (Jeffreys et al., 2005; Webb et al., 2008) and has been used to demonstrate variable usage of historical recombination hot spots but is limited to investigating focused areas within the genome. Cytological approaches can be used to study recombination-related effects in individuals, but these studies use gamete progenitor cells instead of sperm and have several limitations. First, the sample collection requires invasive biopsies. Second, the analysis targets the synaptonemal complexes in spermatocytes, so each progenitor cell analyzed by this method predicts an average result of putative recombination from four future sperm cells. Third, cytological staining does not allow high-resolution molecular analysis such as genotyping or sequencing.

There has been increasing interest in performing single-cell genome analysis in human cancers, and one can compare the methods and results used in cancer with those used here for human gamete genomes. One group used FACS to sort individual nuclei from human breast tumors (Navin et al., 2011). The genomes from these nuclei were amplified in microliter volumes and lightly sequenced to ~0.2 × coverage. This data was sufficient to construct a rough cell lineage map but did not allow calling of individual bases; rather, low-resolution structural variants were used. Another group used mouth pipetting to isolate individual cells from hematopoietic and kidney tumor (Hou et al., 2012; Xu et al., 2012), whose genomes were then amplified in microliter volumes. Rather than performing whole-genome analyses, these samples were then put through exome amplification and sequencing—effectively obtaining 30 × coverage of only 1% of the genome. That data was also used to establish lineage relationships between the cells, this time on the basis of point mutations. Their work reveals one of the challenges of performing single-cell analyses on diploid genomes: only 57% of the diploid calls were correct. Without the ability to examine a significant proportion of the whole genome, the studies mentioned above had to rely on high mutation rate to distinguish single cells. As a consequence, none of the methods have been applied to samples other than late-stage cancers.

In this study, we applied microfluidics to single-cell whole-genome amplification. This technique not only enables great parallelization, but also improves amplification performance. MDA is sensitive to environmental contamination, and extensive sample purification is required for traditional bench-top whole-genome amplifications (Hou et al., 2012; Woyke et al., 2011; Xu et al., 2012). More sensitive assays even revealed contamination in the MDA reagents (Blainey and Quake, 2011). By incorporating the amplification into microfluidic chips, we reduced the reaction volume and, hence, the contamination by ~1,000 fold.

Amplification error has been a concern for single-cell whole-genome analysis. Previous microliter volume single-cell exome studies have shown 2–3 × 10^{-5} false discovery rates from MDA (Hou et al., 2012; Xu et al., 2012). Using our microfluidic approach on haploid cells, we have reduced the error rate to 4 × 10^{-9} with 5 × coverage (binomial probability with per read error rate). An important feature of single molecule MDA is its repetitive usage of the originating genuine template molecule. Even if an amplification error happens in the initial stage, there will still be a large fraction of products preserving the correct base information from the original template, and the power of statistics from multiple coverage discriminates these errors from true genomic variation.

Using this microfluidic MDA approach, we reported the first genome-wide single-cell analysis of human sperm. We were able to create a personal recombination map for an individual and to measure the rate of de novo mutations in this individual's germline. The advantage of sampling a large set of meioses from a single individual for fine-scale analysis allowed us to uncover

individual specific features potentially buried under population data. P0's preference for a subset of historical hot spots suggests how individual features contribute to the population diversity and a potential solution for the hot spot paradox. We propose that this partially overlapping feature is also the general pattern in individuals: everyone is using a different subset of the historical hot spots. While some hot spots are dying in some people, new recombination activities evolve to refill the hot spot pool; the partially overlapping patterns of individuals give rise to the population results, with hot spots (still active in many people) and deserts (used by fewer people). Support for this theory comes from single-cell analysis. Whereas P0 has, on average, 58% overlap with the historical hot spots, this ratio ranges from 0 to 100% for his single cells (Figure S2D). The partially overlapping patterns between individual cells produce P0's personal recombination landscape.

Transmission distortion has long been known, but the key factors behind it are not clear. Biased segregation during meiosis, differing ability to achieve fertilization, and differing postzygotic viability can all contribute to this phenomenon. Specifically, if meiotic drive exists, the molecular mechanism is not known. Our data from 91 cells showed that meiotic drive does not generally appear as whole haplotype blocks but may occur at individual SNP loci. The most intuitive explanation for this result would be gene conversion. Indeed, we found 5–15 gene conversions in each genome-sequenced cell. This represents a lower bound for the total number of conversions in each single human sperm because there is a limited heterozygous SNP density. If both crossover events and gene conversion originate from double-strand breaks and share a recombination mechanism, then they should have the same hot spot overlapping ratio. If we match the number of gene conversions at hot spots and further assume that there are 1.5 million heterozygous SNP in the genome, the total number of gene conversions in a single cell would be ~250–800, which is 10–35× the number of crossovers. Previous sperm typing studies have suggested 4–15× the number of gene conversions over crossovers, based on data from three hot spots (Jeffreys and May, 2004). But this value apparently changes across the genome (Gay et al., 2007).

Evolutionary studies have estimated the germline mutation level (Makova and Li, 2002), but recent results from the 1000 Genome Project (Conrad et al., 2011) are not consistent with the previous findings. The combination of data from our study and the 1000 Genome Project suggests that the germline mutation rate can vary greatly among different individuals, but not among different cells from the same individual. This may explain why the male mutation rate is not always higher than the female. DNA methylation also affects genome instability (Li et al., 2012) and C→T point mutation levels but in opposite ways. A fine-tuned methylation level is therefore required for high-quality sperm genome. The high germline mutation rate at CpA regions (Conrad et al., 2011; Miyoshi et al., 1992) at least suggests a methylation profile that is different from the somatic genome. The fact that cytosine deamination is less well repaired at CpA than at CpG also explains our findings (Wang and Edelmann, 2006).

The ability to study a large number of single sperm cells has offered several new insights in meiosis. Studying the germline genome is but one application of single-cell genomics, and we expect that the method described here will find applications in many other fields, including cancer, aging, immunology, and developmental biology.

EXPERIMENTAL PROCEDURES

Sample Collection

Semen sample was collected from P0 in Stanford Reproductive Endocrinology and Infertility Center and analyzed with a computer-assisted semen analyzer following clinical standards.

Single-Cell Whole-Genome Amplification

Whole genomes from P0's single sperm cells were amplified on a microfluidic device using multiple strand displacement amplification (Repli-g midi, QIAGEN), yielding a gain of ~10^4-fold. Amplification products from single sperm cells were subjected to 46 loci TaqMan genotyping PCR assays (Applied Biosystems) to evaluate the amplification performance.

Public Data Set Access

Human reference genome sequence and annotation were downloaded from UCSC Genome Bioinformatics (http://genome.ucsc.edu/). The P0 somatic genome and genotyping data were from a previous study. Population recombination data were from HapMap Project (http://hapmap.ncbi.nlm.nih.gov/, rel22) and deCODE genetics (http://www.decode.com/).

Personal Recombination Map

93 samples with high-amplification efficiency were reamplified by MDA and genotyped on Illumina's Omni1S Bead Array. Raw genotyping data were processed by Illumina GenomeStudio for genotype calling and were further filtered to remove low-intensity calls. Haploid genome from each sperm cell was aligned to the two P0 somatic haplotypes (Fan et al., 2011). Recombination events were called by a MATLAB script and further manually confirmed. Distribution of P0 recombination events along the genome were compared with population-wide data from deCODE (male noncarrier) and the HapMap Project. Statistical analysis with population data was based on binomial distribution, followed by Bonferroni correction. Recombination frequency in selected regions with recurrent crossover events was measured using allelic specific digital PCR (BioMark, Fluidigm).

Meiotic Drive Measurement

Centromere boundaries were extracted from UCSC chromosome band table and were extended by 5 Mbp to include enough heterozygous SNP. Chromosome haplotype was defined as the haplotype of its centromere. Allele frequency of each autosome was measured based on recombination data of 91 single cells. Pearson's correlation coefficient was calculated for each pair of chromosomes to detect potential cotransmission of chromosome haplotypes. The whole genome was further divided in 100 kbp blocks, and the haplotype of each block was assigned based on chromosome-wide recombination results. Blocks overlapping with recombination events were excluded to avoid haplotype ambiguity within the recombination region. Binomial distribution was used for no distortion simulation. We studied the allele frequency of single SNP with the same approach.

Gene Conversion

For each of the eight single cells, SNPs with alleles pointing to the other haplotype of the block in which they reside were extracted for further examination. The same eight cells were sequenced on Illumina HiSeq2000 with 2 × 100 paired-end read option. Raw sequencing data were processed by Illumina software and aligned to human reference genome hg18 with BWA. Alignment was refined with GATK realignment tool and piled up with samtools. For each gene conversion candidate SNP covered by high-throughput sequencing, we set three requirements for gene conversion calling. The SNP must have sequencing support for its allele call; the SNP must have sequencing support from other single cells or P0 genomic DNA for its heterozygosity; and the SNP must have support from other single cells for its haplotyping. SNP that failed to

meet any of the above three requirements would be cataloged as low coverage or genotyping error, low heterozygosity, and haplotyping error.

Single-Cell Genome Instability Measurement

Single sperm samples with bulk sperm genomic DNA were subjected to multiplex Illumina library construction with NuGEN Encore NGS kit and briefly shotgun sequenced on an Illumina GAII with 1 × 36 read option. Raw sequencing data were processed by Illumina software and aligned to human reference genome hg18 with CASAVA 1.6.0. Sequencing tag density in every 500 kb nonoverlapping window was counted and normalized with sperm genomics DNA control. Genome abnormality was analyzed based on sequencing tag density distribution.

De Novo Mutation Detection

To eliminate random errors induced by MDA and sequencing, we used a binomial test to detect high-confidence mutations. Specifically, false positive rate from MDA and sequencing was measured by using high-confidence homozygous loci in P0 from genotyping and somatic sequencing. The probability of observing a given number of mutation calls under a given sequencing depth was calculated for each position with a binomial distribution.

ACCESSION NUMBERS

All sequencing data from this study are deposited in NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA053375.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.06.030.

ACKNOWLEDGMENTS

We would like to thank Jessica Melin and the Stanford Microfluidic Foundry for fabrication of the microfluidic devices. We thank Janice Gebhardt and the Stanford IVF Laboratory for semen sample processing. We thank Norma Neff, Gary Mantalas, and Ben Passarelli for assistance in sequencing experiments and data processing. We also thank Jad Kanbar for help with PCR experiments. The project was supported by CCNE-T, a grant funded by NCI-NIH (grant number U54CA151459). J.W. was supported by a scholarship from the Chinese Scholarship Council. H.C.F was supported by a scholarship from the Siebel Foundation.

J.W., H.C.F., and S.R.Q. designed research. B.B. coordinated sperm sample collection. H.C.F. designed microfluidic device. J.W. and H.C.F. performed research. J.W., H.C.F., and S.R.Q. analyzed data and wrote the paper. All authors discussed results and commented on the paper.

Received: March 22, 2012
Revised: May 31, 2012
Accepted: June 13, 2012
Published: July 19, 2012

REFERENCES

Ashley, E.A., Butte, A.J., Wheeler, M.T., Chen, R., Klein, T.E., Dewey, F.E., Dudley, J.T., Ormond, K.E., Pavlovic, A., Morgan, A.A., et al. (2010). Clinical assessment incorporating a personal genome. Lancet *375*, 1525–1535.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840.

Blainey, P.C., and Quake, S.R. (2011). Digital MDA for enumeration of total nucleic acid contamination. Nucleic Acids Res. *39*, e19.

Cheung, V.G., Burdick, J.T., Hirschmann, D., and Morley, M. (2007). Polymorphic variation in human meiotic recombination. Am. J. Hum. Genet. *80*, 526–530.

Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project. (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.

Coop, G., Wen, X., Ober, C., Pritchard, J.K., and Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science *319*, 1395–1398.

Crow, J.F. (2000). The origins, patterns and implications of human spontaneous mutation. Nat. Rev. Genet. *1*, 40–47.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nat. Biotechnol. *29*, 51–57.

Gay, J., Myers, S., and McVean, G. (2007). Estimating meiotic gene conversion rates from population genetic data. Genetics *177*, 881–894.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., et al. (2012). Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell *148*, 873–885.

International HapMap Consortium. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

Jeffreys, A.J., and May, C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. *36*, 151–156.

Jeffreys, A.J., Neumann, R., Panayi, M., Myers, S., and Donnelly, P. (2005). Human recombination hot spots hidden in regions of strong marker association. Nat. Genet. *37*, 601–606.

Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Gudbjartsson, D.F., Jonsdottir, G.M., Gudjonsson, S.A., Sverrisson, S., Thorlacius, T., et al. (2008). Sequence variants in the RNF212 gene associate with genome-wide recombination rate. Science *319*, 1398–1401.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Li, J., Harris, R.A., Cheung, S.W., Coarfa, C., Jeong, M., Goodell, M.A., White, L.D., Patel, A., Kang, S.-H., Shaw, C., et al. (2012). Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. PLoS Genet. *8*, e1002692.

Luetjens, C.M., Rolf, C., Gassner, P., Werny, J.E., and Nieschlag, E. (2002). Sperm aneuploidy rates in younger and older men. Hum. Reprod. *17*, 1826–1832.

Macklon, N.S., Geraedts, J.P.M., and Fauser, B.C.J.M. (2002). Conception to ongoing pregnancy: the 'black box' of early pregnancy loss. Hum. Reprod. Update *8*, 333–343.

Makova, K.D., and Li, W.-H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. Nature *416*, 624–626.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L.M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature *454*, 479–485.

Marcy, Y., Ishoey, T., Lasken, R.S., Stockwell, T.B., Walenz, B.P., Halpern, A.L., Beeson, K.Y., Goldberg, S.M.D., and Quake, S.R. (2007). Nanoliter reactors improve multiple displacement amplification of genomes from single cells. PLoS Genet. *3*, 1702–1708.

Miyoshi, Y., Ando, H., Nagase, H., Nishisho, I., Horii, A., Miki, Y., Mori, T., Utsunomiya, J., Baba, S., Petersen, G., et al. (1992). Germ-line mutations of the APC gene in 53 familial adenomatous polyposis patients. Proc. Natl. Acad. Sci. USA *89*, 4452–4456.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. *40*, 1124–1129.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science *327*, 876–879.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Pushkarev, D., Neff, N.F., and Quake, S.R. (2009). Single-molecule sequencing of an individual human genome. Nat. Biotechnol. *27*, 847–850.

Quélin, C., Bendavid, C., Dubourg, C., de la Rochebrochard, C., Lucas, J., Henry, C., Jaillard, S., Loget, P., Loeuillet, L., Lacombe, D., et al. (2009). Twelve new patients with 13q deletion syndrome: genotype-phenotype analyses in progress. Eur. J. Med. Genet. *52*, 41–46.

Sun, F., Oliver-Bonet, M., Liehr, T., Starke, H., Ko, E., Rademaker, A., Navarro, J., Benet, J., and Martin, R.H. (2004). Human male recombination maps for individual chromosomes. Am. J. Hum. Genet. *74*, 521–531.

Tiemann-Boege, I., Calabrese, P., Cochran, D.M., Sokol, R., and Arnheim, N. (2006). High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. PLoS Genet. *2*, e70.

Wang, J.Y.J., and Edelmann, W. (2006). Mismatch repair proteins as sensors of alkylation DNA damage. Cancer Cell *9*, 417–418.

Webb, A.J., Berg, I.L., and Jeffreys, A. (2008). Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc. Natl. Acad. Sci. USA *105*, 10471–10476.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Williams, C., Davies, D., and Williamson, R. (1993). Segregation of delta F508 and normal CFTR alleles in human sperm. Hum. Mol. Genet. *2*, 445–448.

Wong, C.C., Loewke, K.E., Bossert, N.L., Behr, B., De Jonge, C.J., Baer, T.M., and Reijo Pera, R.A. (2010). Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. Nat. Biotechnol. *28*, 1115–1121.

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., and Cheng, J.-F. (2011). Decontamination of MDA reagents for single cell whole genome amplification. PLoS ONE *6*, e26161.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell *148*, 886–895.

Zöllner, S., Wen, X., Hanchard, N.A., Herbert, M.A., Ober, C., and Pritchard, J.K. (2004). Evidence for extensive transmission distortion in the human genome. Am. J. Hum. Genet. *74*, 62–72.

Cell

EXTENDED EXPERIMENTAL PROCEDURES

Microfluidic Device Design, Fabrication, and Operation

The microfluidic device was made of polydimethylsiloxane (PDMS) and was fabricated using soft lithography by the Stanford Microfluidic Foundry. The two-layered device had rectangular 25 μm tall control channels at the bottom and rounded flow channels at the top. The device was bonded to a glass slide coated with a thin layer of PDMS. Flow channels were 30 μm tall and reaction chambers were 100 μm tall. All channels are 150 μm wide. A membrane valve was formed when a control channel crossed over with a flow channel and was actuated when the control channel was pressurized at ~20 psi. The area of each valve was 150 μm x 150 μm. Membrane valves were controlled by external pneumatic solenoid valves that were driven by custom electronics connected to the USB port of a computer. A Matlab program was written to interface with the valves.

Sperm Sample Preparation

P0 semen sample was collection by masturbation in Stanford Andrology Clinics and processed by standard protocol for semen analysis. Sperm cells were washed with DPBS twice at 400 g 10 min and can be used for amplification directly or cryopreserved for future use. The cryopreservation was based on a modified protocol from Stanford Fertility and Reproductive Medicine Center. Briefly, sperm cells were dilution to 12 × 10^6 per ml with DPBS and then mixed gently with cryoprotectant (TEST-modified Yolk Buffer with glycerol, Irvine Scientific, cat #9971) at 1:1 ratio. 0.5 ml of sperm/cryoprotectant mixture was aliquoted to 1 ml cryovials. The cryovials were immersed in liquid nitrogen for 10 min and transferred to a final storage tank with liquid nitrogen. To revive the cells, samples were thawed in a 37°C water bath. The thawed sample was transferred to a 5 ml tube and 1.0 ml DPBS was gently laid on top. The tube was left at 37°C for 1 hr to allow sperm cell to swim up into the DPBS layer. The top 0.75 ml DPBS with cells were retrieved and washed with DPBS at 400 g 10 min. The cryopreservation protocol was developed for in vitro fertilization and thus the general cell quality after thawing is guaranteed.

Multiple Strand Displacement Amplification

Prior to the loading of cell suspension, the cell-partitioning channel of the device was treated with Pluronic F127 (0.2% in DPBS) for 10 min. Sperm cell concentration was adjusted to 4-8 × 10^5 per ml with cell dilution buffer (1 mM EDTA, 1% BSA, 1% Triton X-100 in DPBS). The addition of the cell dilution buffer immobilized the sperm cells. Cells were introduced into the partition channel and separated into forty-eight 1.7 nl compartments by actuating a series of valves along the channel. 0.5% trypsin and 100 mM DTT in cell dilution buffer (1.7 nl) was introduced to lyse the cells and to digest chromosomal proteins at 40°C on a flat-topped thermal cycler. Ten minutes later, denaturation buffer (QIAGEN's Repli-g Midi kit's buffer DLB supplemented with 1.5% Tween-20) (3.5 nl) was introduced. The DNA was denatured at 40°C for 10 min. This was followed by the introduction of neutralization solution (Repli-g kit's stop solution) (3 nl) and incubation at room temperature for 10 min. A mixture of reaction buffer (QIAGEN's Repli-g Midi Kit), phi29 polymerase (QIAGEN's Repli-g Midi Kit), 1X protease inhibitor cocktail (Roche) and 0.5% Tween-20 (40 nl) was fed in. The total volume per reaction was 50nl and the device was placed on the flat-topped thermal cycler set at 32°C for about 16 hr. Amplification product from each chamber was retrieved from its corresponding outlet by flushing the chamber with TE buffer (pH 8.0) supplemented with 0.2% Tween-20. About 5 μl of products were collected from each chamber. Products were incubated at 65°C for 3 min to inactivate the phi29 enzyme.

Initial Genotyping with 46 Loci

To determine the identity of chromosomes in each chamber, we performed TaqMan PCR using a set of 46 genotyping assays (2 assays per autosome and 1 assay per sex chromosome) on the products of each chamber on the 48.48 Genotyping Dynamic Array (Fluidigm). The assays used are listed in our previous study (Fan et al., 2011).

Whole-Genome Sperm Genotyping

In order to generate sufficient materials for genotyping array experiments, DNA products from the microfluidic device were amplified a second time in 10 μl volume using the Repli-g Midi Kit's protocol for amplifying purified genomic DNA. The amplification product from each sperm sample was genotyped on the HumanOmni1S BeadChips (Illumina). Genomic DNA and somatic haplotype samples of P0 were genotyped on the same array during previous whole genome haplotyping study.

Genotyping raw signals were processed by Illumina Genome Studio for initial genotype calls. We further wrote custom scripts to filter the calls. Briefly, for each SNP, we take the genotyping results from all samples. Assuming heterozygous calls are low intensity false positive, we raised the noise cut-off to intensity of the heterozygous calls and filtered the results. We also re-called some SNP based on the R and Theta values from the Genome Studio output. SNPs with Theta lower than 0.1 or higher than 0.9, and R higher than highest value of existing calls will be called with corresponding genotypes.

Only the informative heterozygous SNPs were used for downstream analyses. For each sperm sample, we aligned the genotyping calls to the two somatic haplotypes. Recombination events were detected by comparing the somatic haplotype alignment consensus within a 15-SNP sliding window. We also manually examined all the samples to confirm the recombination calls.

Recombination Data Sources

Recombination frequencies and hot spot activities were downloaded from the website of the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) and deCODE genetics (http://www.decode.com/addendum/). For each arbitrary region along the genome, the expected recombination level was calculated using the Binomial Distribution with the historical recombination frequency.

Interference Measurement

We employed previously published strategy (Mancera et al., 2008) to measure the recombination interference. For all the recombination events on the same chromosome, we kept the positions but shuffled the sample assignment. Thus we maintained the recombination activity inhomogeneity along the chromosome, but removed potential interference. The median distance between adjacent recombination was used for comparison. The shuffling process was repeated 10,000 times and P-value was calculated as the chance of observing larger median inter-event distance from the shuffling than our data.

DNA Extraction from P0 Blood and Sperm Cells

We collected whole blood from P0 and isolated lymphocytes in a previous whole genome sequencing study. DNA was extracted from P0 lymphocytes using QIAamp DNA Blood Mini Kit (QIAGEN) following manufacturer's instruction. DNA was also extracted from P0 sperm cells using the same method with the following modification: supplement 40 mM DTT in the cell lysing step and extent the 56°C incubation to 1 hr. DNA was further quantified with NanoDrop and a TaqMan PCR assay on *MBNL2* gene (ultraconserved element 356, 13q32.1). Primer_F: CTCACCTATCCACAATGCAA; Primer_R: GGGATTCAAGCGAATTAACA; Probe: AGGTGCATCAT GGGAACGGC.

PRDM9 Sequencing

The zinc finger coding region of *PRDM9* was amplified from P0's gDNA, two somatic haplotypes and two single sperm samples with the same primers and condition as previous study (Baudat et al., 2010). Sanger sequencing results from MCLAB were aligned to different *PRDM9* alleles for genotyping.

Recombination Frequency by Allelic Specific PCR

We designed allelic specific primers with the 3′ end aligning to the SNP being typed (Figure S2A). By using different combinations of allelic specific primers for the two heterozygous SNPs flanking a recombination region, we specifically amplified the somatic and recombined haplotypes separately. We pre-amplified both blood and sperm DNA for 15 cycles with the following protocol: 50 μl ABI AmpliTaq Gold PCR reaction with 30 ng DNA ($\sim 10^4$ copies of haploid genome equivalent) and 200 nM primers. 95°C 5 min; 15 cycles of 95°C 30 s, 55°C 30 s, 72°C 2 min; 72°C 10 min. Pre-amplification products were diluted 1:20 with Buffer EB (QIAGEN) without purification. We then used TaqMan digital PCR (Fluidigm 12.765 Digital Array) on one of the two SNPs to quantify the amplified haplotypes. Each amplification product was series diluted with Buffer EB to have ~100 amplifiable molecules in each digital PCR panel (765 chambers). Recombination frequencies were calculated using the copy number of somatic and recombined haplotypes with blood sample as negative control.

Recombination Frequency by 2 Loci Digital Haplotyping

With allelic specific TaqMan assays for each of the two SNPs flanking a recombination region, we performed 2-loci digital haplotyping (Menzel et al., 2010) using microfluidic digital PCR (Fluidigm 48.770 Digital Array, Figure S2B-C). After extensive dilution, the two alleles, one from each SNP should be present in the same chambers if they are in coupling phase, or in different chambers in they are in repulsion phase. Since the method detects the two SNPs separately, we are not limited by the amplicon size but rather the fragment size of DNA samples. The false positive signals mainly came from the co-occupancy of multiple DNA template molecules in the same chamber. We used the Poisson distribution model to estimate this false positive level and subtracted it from the observed signals. Recombination frequency was estimated as the ratio between the numbers of recombined molecules and total molecules. Approximately 1,500 haploid genome equivalent DNA were analyzed for each region.

High-Throughput Sequencing of Single Sperm Samples

We selected 31 sperm samples with a sperm genomic DNA control for 1x36 shotgun sequencing on Illumina GAII. 23 samples had relatively lower amplification efficiency according to 46-loci PCR, while the other 8 samples had high call rates. Second-round amplified materials from these samples were first digested with S1 nuclease (1 μg DNA with 1 unit S1 nuclease in a 40 μl reaction, 37°C 30 min) and then fragmented with Covaris S2 system per manufacturer's instruction. Fragmented DNA was subjected to Illumina sequencing library preparation with NuGEN Encore NGS Multiplex System I per manufacturer's instruction. Sequencing libraries with 8 different barcodes were pooled together with equal molarity and quantified with digital PCR. Each library was sequenced on one lane on the flow cell with 36 bases read length. Image analysis, base calling, and alignment were performed using Illumina's CASAVA 1.6.0. The last 32 bases without barcode from each read were aligned to the human genome (Build 36.1).

We also sequenced 8 single cell samples on Illumina HiSeq 2000 with 2x100 read length. Sequencing results were aligned to the human genome (Build 36.1) with BWA and further locally re-aligned with GATK. Read pileup was done with samtools. Given the

haploid nature of the sperm samples, we called the genotypes using 95% majority of the base calls, with base alignment quality (BAQ) score > = 30.

Meiotic Drive Detection

Haplotype of a chromosome is defined as the haplotype of its centromere region (5 Mbp extended from the chromosome band on both sides). Haplotype of each chromosome, each 100 kbp block and each SNP was called based on genotyping results and recombination map. Simulation assuming no transmission distortion was performed with binomial distribution.

Gene Conversion Detection

For each of the 8 single cells, haplotype blocks boundaries were assigned by crossovers. SNPs having the opposite haplotype as their own blocks were picked out for further examination. For each gene conversion candidate SNP covered by high-throughput sequencing, we set three requirements for gene conversion calling. The SNP must have sequencing support for its allele call; the SNP must have sequencing support from other single cells or P0 genomic DNA for its heterozygosity; the SNP must have support from other single cell for its haplotyping. SNP failed to meet any of the above three requirements would be cataloged as low coverage or genotyping error, low heterozygosity and haplotyping error.

Single-Cell Genome Instability Measurement

To compare the coverage of the different chromosomes, a sliding window of 500 kb was applied across each chromosome, except in regions of assembly gaps and microsatellites, and the number of sequence tags falling within each window was counted, and the median value was chosen to be the representative of the chromosome. We normalized the tag density first to genomic DNA control, then to the median sequence tag density among autosomes. The normalized values were used for aneuploidy detection (Figures 5B and S3A).

De Novo Mutation Detection

To measure the de novo mutation rate, we obtained sequencing results of P0 somatic genome and called the genotypes with GATK. Genotypes from somatic genome and single sperms were co-aligned to detect mutation. Single molecule MDA accuracy was evaluated by comparing the single sperm sequence results with high confident (genotyping score > = 100 in Phred scale) homozygous somatic positions. The distribution of the discordance ratio suggests a separate group of positions with 100% discordance from the somatic genome, which outstands out of the long tail of errors (Figure S4A).

SUPPLEMENTAL REFERENCES

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nat. Biotechnol. *29*, 51–57.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L.M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature *454*, 479–485.

Menzel, S., Qin, J., Vasavda, N., Thein, S.L., and Ramakrishnan, R. (2010). Experimental generation of SNP haplotype signatures in patients with sickle cell anaemia. PLoS ONE *5*, e13004.

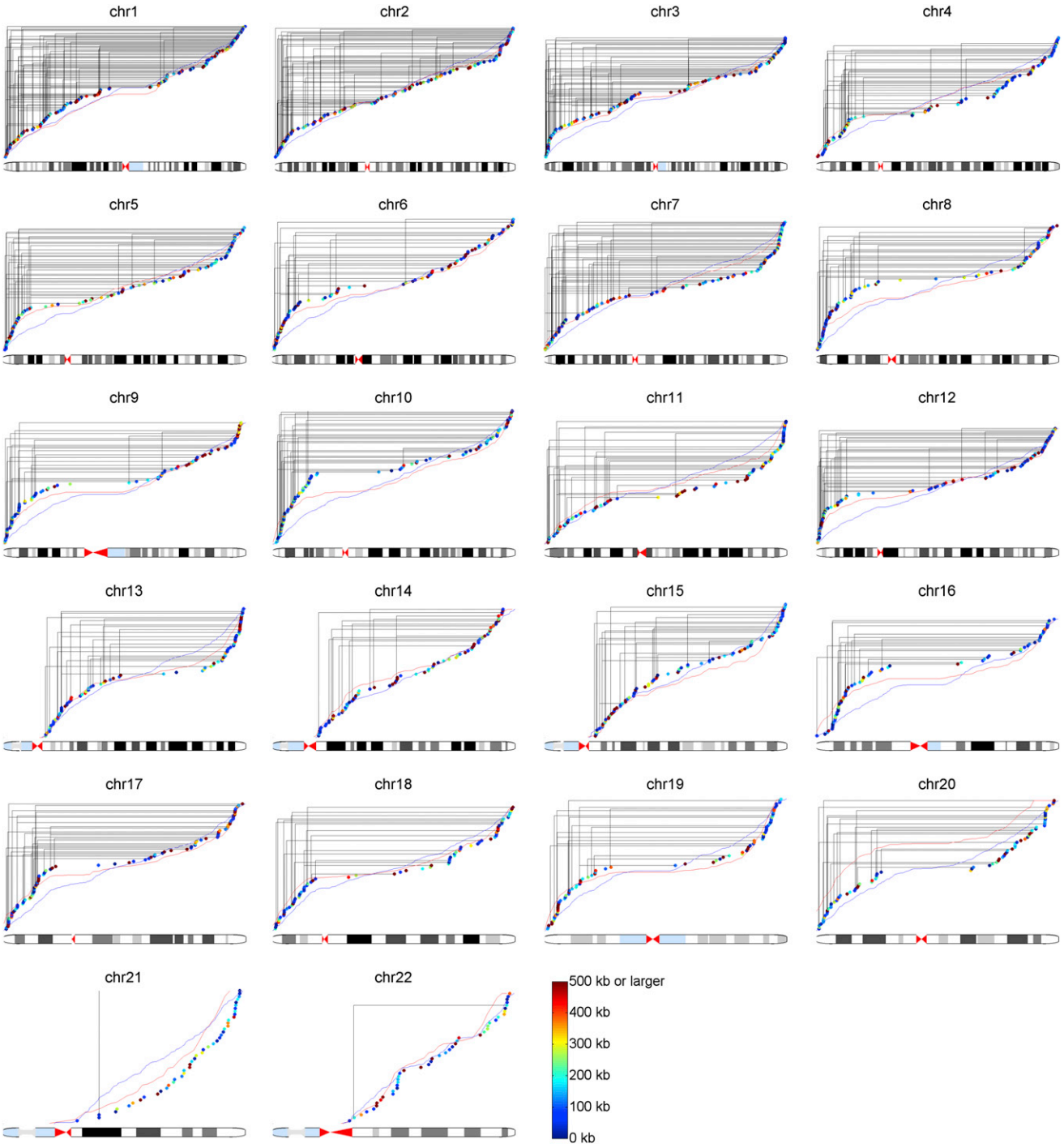


Figure S1. Personal Recombination Map for Each Autosomal Chromosome, Related to Figure 3

Each dot represents a recombination event with color code for resolution. Solid black lines connect recombination events from the same sperm cell. Red and blue lines show the cumulative recombination rates from deCODE (male) and HapMap, respectively.

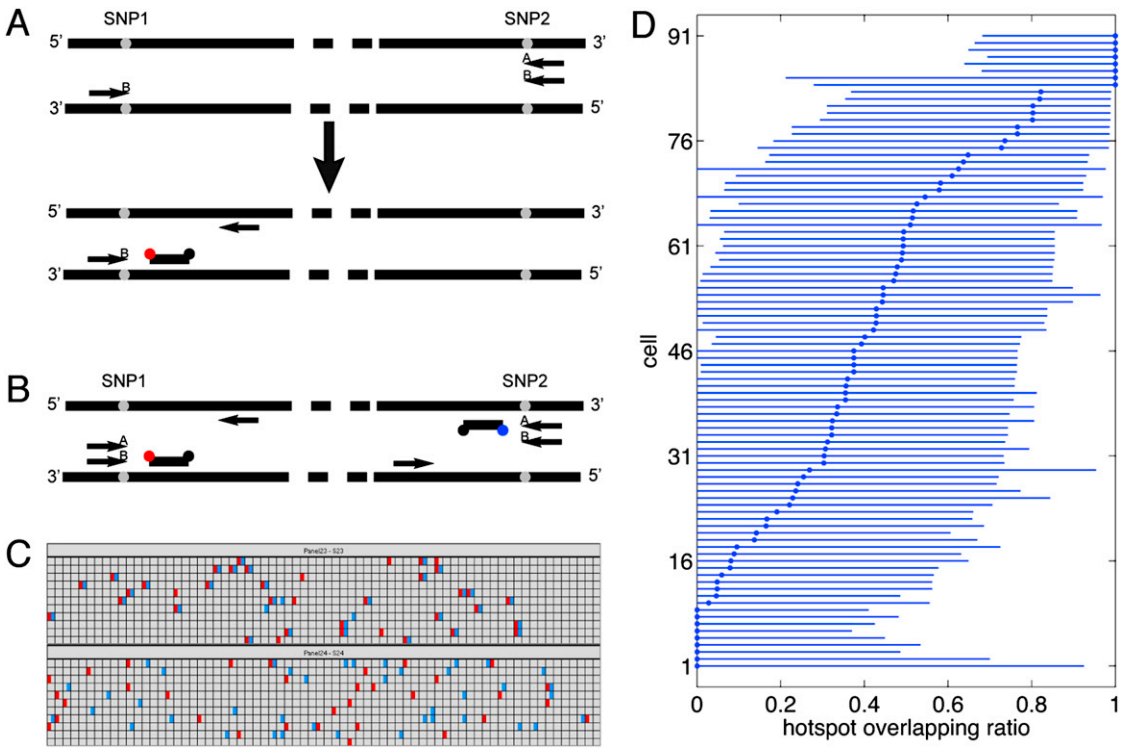


Figure S2. Individual Specific Recombination Hot Spots, Related to Table 1

(A) Somatic or recombined haplotypes were first amplified with different combinations of allelic specific primers (upper panel). 'Primer A' and 'Primer B' represent the two different allele specific primers at each locus. Amplified haplotypes were further quantified with TaqMan assay specific for one allele of one SNP using digital PCR (lower panel).

(B) Scheme of 2-loci allelic specific TaqMan PCR. For each SNP, only one allele is detected at one time. The probes for the two PCR amplicons have different colors (red as FAM and blue as HEX). The combination of allelic specific primers from the two SNP can detect alleles in either coupling (e.g., SNP1-A with SNP2-A) or repulsion (e.g., SNP1-A with SNP2-B) phase.

(C) Digital haplotyping results of P0 blood DNA from a region on chromosome 16. Upper panel shows results from SNP1-A-FAM and SNP2-A-HEX assays, which detect alleles in coupling phase. Lower panel shows results from SNP1-A-FAM and SNP2-B-HEX assays, which detect alleles in repulsion phase. The two chambers with both alleles detected in the lower panel are due to multiple template molecules occupation (1.87 expected from Poisson Distribution).

(D) Historical hot spots overlapping ratio for each single cell. The maximum likelihood estimate is shown as a circle and the 95% confidence intervals are shown by the horizontal lines.

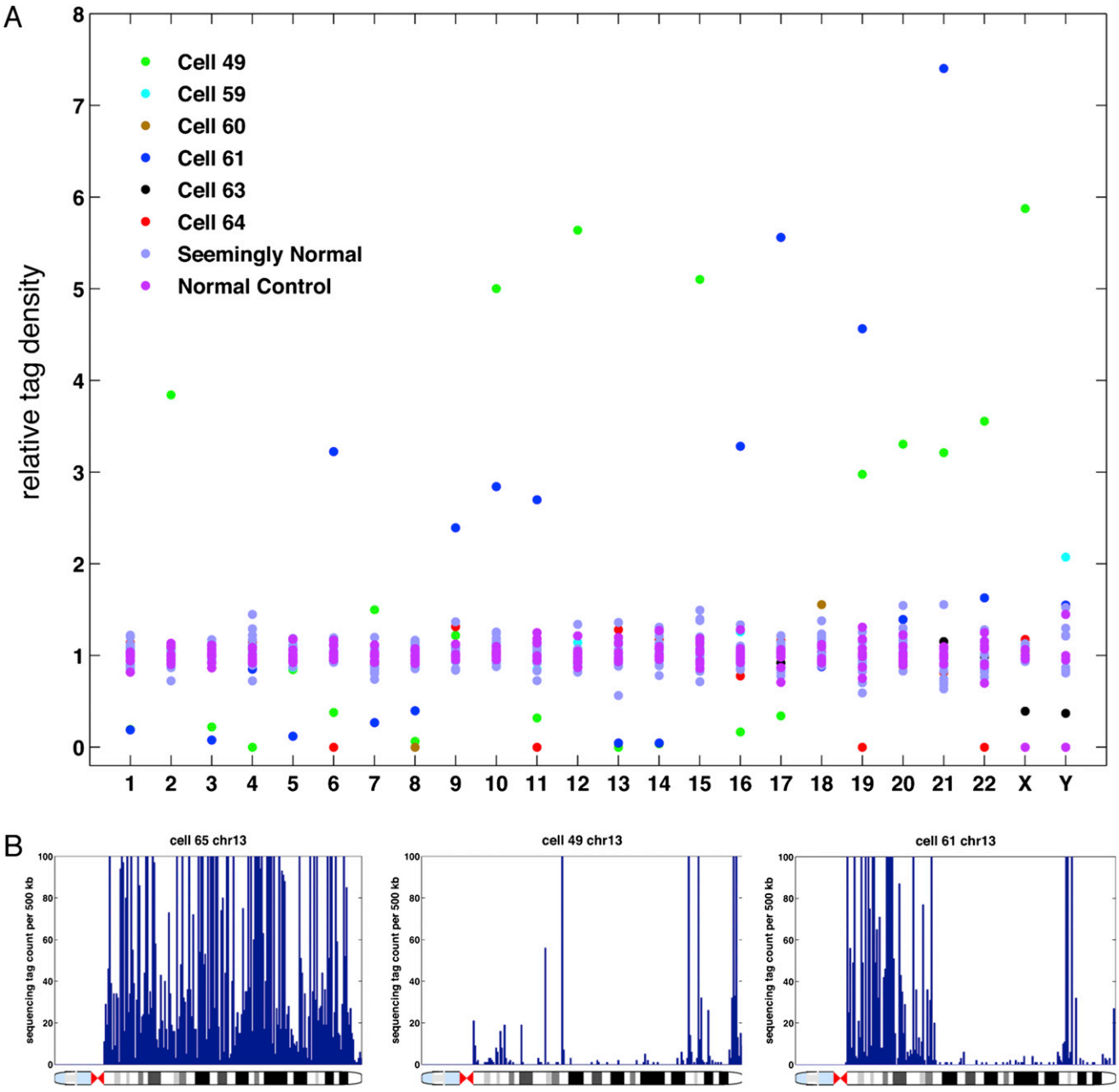


Figure S3. Germline Genome Instability, Related to Figure 5

(A) General characteristics of large-scale genome instability by single cell whole-genome sequencing results.

(B) Whole-genome sequencing results of cell 49 and 61 showed large-scale deletion in chromosome 13. Cell 65 on the left is shown as normal control.

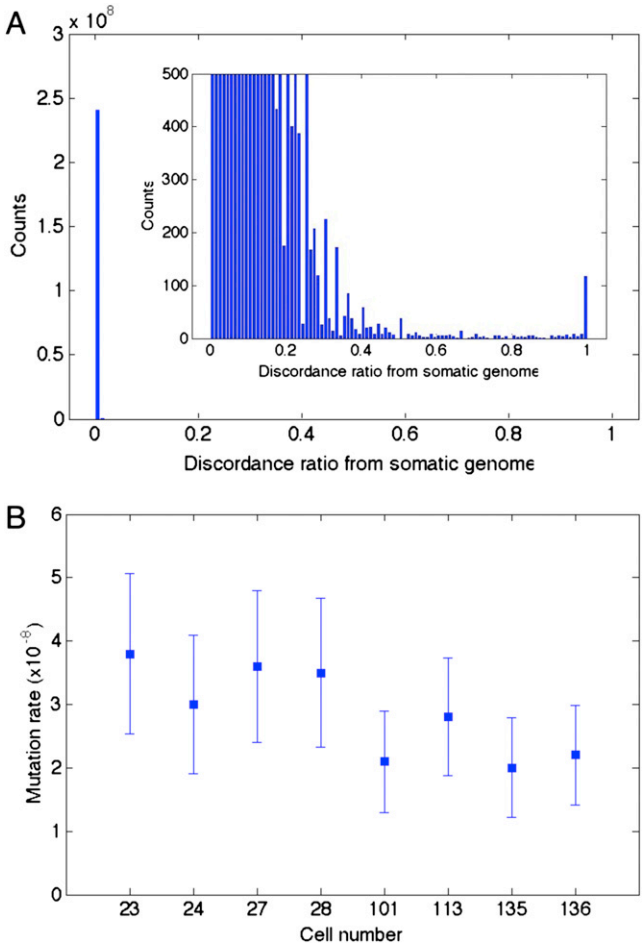


Figure S4. De Novo Germline Mutation by Single-Sperm Whole-Genome Sequencing, Related to Table 2

(A) Allele discordance ratio of sperm MDA products against somatic genome (insert as *y* axis zoom in). The peak at 100% discordance illustrates a distinct group of loci standing out of the amplification errors background tail.

(B) Mutation rates from 8 single cells have overlapping 95% confidence interval.