We present a survey of single-cell whole-genome amplification (WGA) methods, including degenerate oligonucleotide–primed polymerase chain reaction (DOP-PCR), multiple displacement amplification (MDA), and multiple annealing and looping–based amplification cycles (MALBAC). The key parameters to characterize the performance of these methods are defined, including genome coverage, uniformity, reproducibility, unmappable rates, chimera rates, allele dropout rates, false positive rates for calling single-nucleotide variations, and ability to call copy-number variations. Using these parameters, we compare five commercial WGA kits by performing deep sequencing of multiple single cells. We also discuss several major applications of single-cell genomics, including studies of whole-genome de novo mutation rates, the early evolution of cancer genomes, circulating tumor cells (CTCs), meiotic recombination of germ cells, preimplantation genetic diagnosis (PGD), and preimplantation genomic screening (PGS) for in vitro–fertilized embryos.
INTRODUCTION

Individual cells are the fundamental units of life. DNA that carries genetic information exists as single molecules in individual cells. In biology and medicine, there is often a need to characterize genomes of individual cells for several reasons: (a) Some cells are precious and exist in low numbers [for example, human oocytes and circulating tumor cells (CTCs)]; (b) every cell is unique in its genome (for example, every sperm cell of an individual is different because of recombination); (c) the genomes of individual cells undergo stochastic changes with time, and hence single cells’ genomes at particular times can reveal their temporal evolution; and (d) the genomes of individual cells in the same sample are heterogeneous (as in the case of primary cancer tissues), and consequently one often needs to determine the distribution rather than the average of a large ensemble of cells.

Changes in the genome of a single cell include single-nucleotide variations (SNVs) and structural variations, which result in copy-number variations (CNVs). SNVs are single-base insertions, deletions, or mutations, either transitions (e.g., C→T, A→G) or transversions (e.g., A→T, C→G). Structural variations are genomic rearrangements, including insertions or deletions (indels), duplications, inversions, and translocations. Structural variations at large genome scales are linked to CNVs. Because of technical difficulties (chimeras; see below), structural variations are difficult to observe at the single-cell level, whereas CNVs ranging in size from hundreds of kilobases to megabases can be detected relatively easily.

The copy number of a particular gene in a human somatic cell is normally two because of the two alleles (one from each parent). Although it had long been known that gene copy numbers could deviate from two in humans, genome-wide measurements of CNVs (27, 56) prompted extensive investigations of the biology and clinical consequences of CNVs, which often result from double-strand breaks that cannot be repaired perfectly (21).

We note that epigenetics (e.g., the methylation of DNA) represents another type of genomic change. The single-cell methylome has been determined (19), but this topic is beyond the scope of our review.

Advances in next-generation DNA sequencing technologies have enabled individual human genomes to be sequenced at affordable costs (4, 43, 69, 72). With the development of single-cell whole-genome amplification (WGA) techniques—i.e., amplification of all DNA molecules after the cell has been lysed—single-cell genomics has emerged as an exciting field of its own. In this article, we review the principles and compare the performance of the existing methodologies for single-cell WGA. We note that, in parallel with single-cell genomics, single-cell transcriptomics has seen rapid developments as well; these techniques are beyond the scope of this article, and we refer readers to other recent reviews for more information (55, 57, 60). We provide examples to illustrate the utility of single-cell WGA in biomedical applications, including fundamental research on genome stability and the biology of meiosis, as well as clinical perspectives for cancer diagnosis and reproductive medicine.

SINGLE-CELL WHOLE-GENOME AMPLIFICATION METHODS

Although sequencing individual DNA molecules with lengths of thousands of bases has become possible (20), currently there is not a method that can collect and sequence all DNA fragments from a single cell. Therefore, sequencing an entire genome of a cell requires WGA. Given the trace amount of DNA from a single human cell (a few picograms), extreme care should be taken to avoid contamination. Indeed, a major difficulty in single-cell genomics is substantial contamination from
the environment and operators. Our experience is that WGA should be conducted in a dedicated clean room with controlled air pressure and quality. As such, the bacterial contamination can be controlled below 0.1% of the amount of DNA in a human cell (78). Alternatively, microfluidic devices can be used to minimize contamination (14, 31, 68), which is particularly important for WGA of a bacterial cell.

Regardless of whether the WGA is done in a polymerase chain reaction (PCR) tube or a microfluidic device, the chemistry is the most important aspect. Several different chemistries are available for single-cell WGA, and here we review three major ones: the degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR) (61), multiple displacement amplification (MDA) (12), and multiple annealing and looping–based amplification cycles (MALBAC) (78).

Degenerate Oligonucleotide–Primed Polymerase Chain Reaction (DOP-PCR)

PCR has had a major impact on biology and medicine in the past 30 years (54). It offers single-copy sensitivity—that is, a single copy of DNA can generate a signal detectable by PCR. To detect the existence of one mutation in a particular gene, one can design a PCR primer to amplify the gene locus. Attempts have been made to use PCR to amplify the entire genome using a set of primers. Primer extension preamplification PCR (PEP-PCR) was among the first single-cell WGA methods used (3, 45, 77), followed by the more widely adopted DOP-PCR (61) (Figure 1).

![Figure 1](https://example.com/dop-PCR.png)

**Figure 1**

Schematic of the degenerate oligonucleotide–primed polymerase chain reaction (DOP-PCR), which uses degenerate oligonucleotide primers for whole-genome amplification (45). Additional abbreviation: ssDNA, single-stranded DNA.
The principle of DOP-PCR is to use degenerate primers containing a random six-base sequence at the 3’ end and a fixed sequence at the 5’ end. For the initial amplification, the primers bind to the DNA template at a low annealing temperature. Strand extension is then achieved at a raised temperature. During the second stage of PCR amplification, the previous products are amplified with a primer targeting the 5’ fixed sequence at a higher annealing temperature. The concentrations of the primers and polymerase directly affect the result of DOP-PCR. DOP-PCR has been used to amplify picogram quantities of human genomic DNA for genotypic analyses (7).

DOP-PCR often yields low genome coverage, which is pertinent to the exponential amplification of PCR. Any small differences in the amplification factors among different sequences are exponentially enlarged, causing overamplified regions and underamplified regions in the genome, and hence low coverage. Although lacking completeness in accessing the whole genome, DOP-PCR can be well suited for measuring CNVs on a large genomic scale with large bin sizes (1 million bases) (47).

Multiple Displacement Amplification (MDA)

MDA was developed in 2001 by Lasken and coworkers (12) using a random hexamer as a primer and ϕ29 DNA polymerase, a highly processive DNA polymerase (5) with strong strand displacement activity. ϕ29 has a high replication fidelity because of its 3’→5’ exonuclease activity and proofreading activity (17, 51). Under isothermal conditions, MDA extends the random primers and produces branched structures, which are extended by other primers and eventually form multibranched structures (Figure 2). The DNA fragments are 50–100 kb long.

MDA offers much higher genome coverage than DOP-PCR. However, like DOP-PCR, MDA is an exponential amplification process. This results in sequence-dependent bias, causing overamplification in certain genomic regions and underamplification in other regions. However, such sequence-dependent bias of MDA is not reproducible along the genome from cell to cell, rendering CNV measurements noisy and normalization ineffective. Nevertheless, MDA has been widely applied since its invention (37).
Annealing of primers

Denaturing

Quenching at 0°C

Extension at 65°C

Melting at 94°C

(m+1) × n

(m × n)^2

MALBAC primer

Genomic DNA

Semiamplicon

Full amplicon

MALBAC product to be sequenced

Figure 3

Schematic of multiple annealing and looping–based amplification cycles (MALBAC). Random primers with a fixed sequence are used in a temperature cycle in which only the original genomic DNA and semiamplicons are linearly amplified, and full amplicons are protected from further amplification by the formation of DNA loops owing to the complementarity of the fixed sequences at the 3′ and 5′ ends. The DNA loops are polymerase chain reaction (PCR) amplified at the final stage (78). Here, \( m \) is the number of temperature cycles (\( m = 0 \sim 10 \)), and \( n \) is the number of primers bound; \((m + 1) \times n\) is the number of semiamplicons present at the \( m \)th cycle, and \( m \times n^2 \) is the number of full amplicons generated in the \( m \)th cycle.

Multiple Annealing and Looping–Based Amplification Cycles (MALBAC)

MALBAC was first reported in 2012 by Zong et al. (78) for single-cell WGA; this method has the unique feature of quasi-linear amplification, which reduces the sequence-dependent bias exacerbated by exponential amplification. It has recently been applied to single-cell transcriptome measurement as well (6). The key to MALBAC is to not make copies of copies, and instead make copies only of the original genomic DNA by protecting the amplification products (Figure 3). The specially designed MALBAC primers have a common 27-nucleotide sequence at the 5′ end and 8 random nucleotides at the 3′ end, which can evenly hybridize to the template when the temperature is lowered (to 15–20°C). At the beginning, semiamplicons with variable lengths (0.5–1.5 kb) are made when the temperature is elevated (to 70–75°C). The semiamplicons are melted off from the templates (at 95°C), from which full amplicons are made with complementary ends, causing the formation of hairpins when the temperature is lowered (to 58°C), preventing their further amplification. This cycle is repeated 8–12 times. The quasi-linear amplification at these first few cycles is critical for avoiding the sequence-dependent bias exacerbated by exponential
amplification. MALBAC uses a thermally stable DNA polymerase with strand displacement activity. This important preamplification stage is then followed by exponential amplification of the full amplicons by PCR, generating the amount of DNA required for next-generation sequencing.

MALBAC is not a mere combination of DOP-PCR and MDA, but is fundamentally different because of its quasi-linear, as opposed to exponential, amplification. This results in two major advantages: accuracy for CNV detection and a low false negative rate for SNV detection.

MALBAC is not free from sequence-dependent bias. Unlike MDA, however, MALBAC’s sequence-dependent bias is reproducible along the genome from cell to cell. Therefore, signal normalization for CNV noise reduction can be carried out. As shown below, after signal normalization with a reference cell, MALBAC offers the best CNV accuracy.

MALBAC exhibits the lowest false negative rates for SNV calling. However, MALBAC has a higher false positive rate for SNV detection than MDA because the DNA polymerase currently used has a lower fidelity than the φ29 polymerase.

Characterization and Comparison of Whole-Genome Amplification Methods

It is necessary to provide a critical comparison of different WGA methods. Here we choose key features to compare, such as coverage, uniformity, reproducibility, allele dropout rate, false positive rate, chimera rate, and unmappable rate, as defined below. We compared the performance of five commercial single-cell WGA kits that use one of the three methods discussed above: the Sigma-Aldrich GenomePlex Single Cell Whole Genome Amplification Kit (DOP-PCR), the Qiagen REPLI-g Single Cell Kit (MDA), the General Electric (GE) illustra Single Cell GenomiPhi DNA Amplification Kit (MDA), the Yikon Genomics Single Cell Whole Genome Amplification Kit (MALBAC), and the Rubicon Genomics PicoPLEX WGA Kit (MALBAC-like).

De Bourcy et al. (11) recently compared several WGA methods for single-cell analyses of Escherichia coli. For mammalian cells, issues related to CNVs and heterozygosity of alleles are germane to genomic analyses, which prompted us to perform a comparison of human cells. We note that human cancer cell lines often exhibit aneuploidy, which seriously complicates analyses of genome coverage and allele dropout rate. Here, we chose a diploid human cell line, BJ primary human foreskin fibroblast. The lack of aneuploidy of this cell line makes it an ideal system to characterize amplification performance.

To perform this comparison, we amplified several single cells and sequenced to exclude cells in metaphase (in which DNA is replicated). The amplified DNA products from each cell were fragmented by sonication to ~250 base pairs. The sequencing libraries were prepared by using the NEBNext Ultra DNA Library Prep Kit before sequencing on an Illumina HiSeq X Ten platform with 150-base paired-end reads.

Table 1 summarizes our characterization of key parameters that are defined and discussed below for the five tested kits for DOP-PCR, MDA, and MALBAC. Below, we define the parameters used in the comparison and present and discuss the results.

Coverage. At this point there has not yet been a complete human genome sequence—even the reference genome has gaps. In comparing the genome coverage of single cells with different WGA kits, we used the bulk sequencing without amplification at 30× depth as the reference and assumed it has 100% coverage.

The comparison was done using the total raw data of 80–100 Gb (30×) for each cell (Table 1). The single-cell sequencing data from certain kits have a large fraction of unmappable reads, which is of course undesirable. Comparison at identical sequencing depths would be impractical.
Table 1  Comparison of key parameters of five single-cell whole-genome amplification kits

<table>
<thead>
<tr>
<th>Kit</th>
<th>WGA method</th>
<th>Raw data</th>
<th>Coverage</th>
<th>CV</th>
<th>Reproducibility</th>
<th>ADO</th>
<th>FPR</th>
<th>CR</th>
<th>UMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma-Aldrich</td>
<td>DOP-PCR</td>
<td>79 Gb</td>
<td>39%</td>
<td>0.14</td>
<td>0.93</td>
<td>76%</td>
<td>9.6 × 10⁻⁴</td>
<td>15%</td>
<td>64%</td>
</tr>
<tr>
<td>Qiagen</td>
<td>MDA</td>
<td>78 Gb</td>
<td>84%</td>
<td>0.21</td>
<td>0.68</td>
<td>33%</td>
<td>1.3 × 10⁻⁴</td>
<td>2%</td>
<td>18%</td>
</tr>
<tr>
<td>GE</td>
<td>MDA</td>
<td>105 Gb</td>
<td>82%</td>
<td>0.17</td>
<td>0.31</td>
<td>38%</td>
<td>8.2 × 10⁻⁵</td>
<td>3%</td>
<td>44%</td>
</tr>
<tr>
<td>Yikon</td>
<td>MALBAC</td>
<td>96 Gb</td>
<td>72%</td>
<td>0.10</td>
<td>0.87</td>
<td>21%</td>
<td>3.8 × 10⁻⁴</td>
<td>5%</td>
<td>22%</td>
</tr>
<tr>
<td>Rubicon</td>
<td>MALBAC-like</td>
<td>80 Gb</td>
<td>52%</td>
<td>0.13</td>
<td>0.98</td>
<td>28%</td>
<td>2.4 × 10⁻⁴</td>
<td>13%</td>
<td>66%</td>
</tr>
</tbody>
</table>

Kits: Sigma-Aldrich, GenomePlex Single Cell Whole Genome Amplification Kit; Qiagen, REPLI-g Single Cell Kit; GE, illustra Single Cell GenomiPhi DNA Amplification Kit; Yikon, Single Cell Whole Genome Amplification Kit; Rubicon, PicoPLEX WGA Kit. Abbreviations: ADO, allele dropout; CR, chimera rate; CV, coefficient of variation; DOP-PCR, degenerate oligonucleotide–primed polymerase chain reaction; FPR, false positive rate; MALBAC, multiple annealing and looping–based amplification cycles; MDA, multiple displacement amplification; UMR, unmappable rate; WGA, whole-genome amplification.

because of the high cost. Considering that the coverage is dependent on the actual sequencing depth, Figure 4 gives the genome coverage of the single cells referenced to the bulk as a function of the effective sequencing depth.

The low effective sequencing depth for the Sigma-Aldrich DOP-PCR kit was mainly due to the high number of universal adapter reads. For Rubicon’s MALBAC-like kit, in addition to the primer and adapter reads, the high number of unmappable reads includes large contributions from the small DNA fragments (<30 base pairs) inserted into the sequencing adapters, which were too short to be mapped to the human reference genome. The coverage of Qiagen’s MDA kit is higher than that of Yikon’s MALBAC kit, which may result from MDA’s lack of reproducibility of the sequence-dependent bias for the two different alleles in the diploid cell (36). Suffice it to say...
that MDA and MALBAC kits have comparable levels of coverage, both of which are significantly higher than the coverage of DOP-PCR.

**Uniformity.** Uniform amplification is important for accurate measurements of CNV. Figure 5a shows the raw read density of 23 chromosomes, clearly illustrating the sequence-dependent bias along the genome. The variation is at the 1,000-kb scale.

If the variation along the genome is the same from cell to cell, then a normalization factor can be calculated by averaging the read density of several cells in each bin and using this to normalize the single-cell data. If the purpose of the investigation is to measure CNV, then reference cells without aneuploidy should be chosen. Figure 5b shows the data normalized by the average of three cells with no aneuploidy.

We characterize the uniformity by coefficients of variation of the read density (the genome-wide standard deviation divided by the mean). The normalization is done by the average of a few reference cells (as in Figure 5b) in a certain bin size (1,000 kb). Figure 6a shows the uniformity comparison of the five kits.

It is evident from Figure 5a that, of the five methods, DOP-PCR gives the flattest CNV raw data without normalization. MDA creates variations along the genome that are not reproducible from cell to cell and cannot be smoothed via normalization. MALBAC’s sequence-dependent bias is reproducible from cell to cell, giving the flattest CNV after normalization.

**Reproducibility.** Reproducibility is important in single-cell measurements, and measurement noise needs to be minimized. We use Pearson’s cross-correlation coefficient of the read densities along the genome between two identical cells to characterize the reproducibility.

Figure 6b shows our measurements of the cell-to-cell reproducibility of a WGA method by the cross-correlation coefficients for different WGA methods. It is apparent that the reproducibility of DOP-PCR and MALBAC kits is better than that of MDA methods that exhibit stochasticity in WGA.

**Allele dropout rate.** The allele dropout rate is one of the most important characteristics of WGA, particularly for medical applications. Allele dropout arises from uneven WGA, which needs to be improved. If a diploid cell has a heterozygous mutation, the lack of amplification in one of the two alleles causes allele dropout (Figure 7a). Allele dropout is the primary cause of false negatives of SNV calling. The allele dropout rate is measured by the ratio of the undetected and the actual heterozygous SNVs in a single cell. The latter is often approximated by the bulk measurement of identical cells.

Figure 6c shows our measurements of the allele dropout rates of the five kits. It is evident that the MALBAC methods have lower allele dropout rates than the other WGA methods.

**False positive rate.** A false positive of base calling can arise from either a sequencing error or an amplification error. Whereas random sequencing errors can be avoided with high sequencing
Access provided by Harvard University on 09/01/15. For personal use only.

### a

Sigma-Aldrich

Qiagen

GE

Yikon

Rubicon

#### Chromosome

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

### b

Sigma-Aldrich, CV = 0.14

Qiagen, CV = 0.21

GE, CV = 0.17

Yikon, CV = 0.10

Rubicon, CV = 0.13
Figure 6
Comparison of key parameters of the five commercial single-cell whole-genome amplification kits, using data generated from three single cells from the BJ primary human foreskin fibroblast cell line. (a) Coefficient of variation, calculated as the ratio of the standard deviation of read density to its average in three individual cells. (b) Reproducibility, measured by Pearson’s cross-correlation coefficient of the read densities between two identical cells. (c) Allele dropout rate, defined as the percentage of homozygous sites in the single-cell samples where the bulk sample is heterozygous at the same nucleotide site. (d) False positive rate, defined as the number of heterozygous site calls in the single cell divided by the number of sites in the bulk sample that are homozygous or have a different allele at the same nucleotide site. (e) Chimera rate, defined as the number of reads that are improperly connected (including abnormal fragment size and interchromosomal connection) divided by the total number of mappable reads. (f) Fraction of reads that are unmappable to the reference genome.

depth, WGA errors often dominate, occur during the first or second cycle of exponential amplification, and are dependent on the error rate of the DNA polymerase (Figure 7b). Figure 6d compares the false positive rates of the WGA kits. The Qiagen and GE kits perform better than the MALBAC approaches, mainly because of the higher accuracy of the φ29 polymerase in the former.
Allele dropout

Figure 7

Schematics for three types of errors that can arise in whole-genome amplification. (a) Allele dropout. The lack of amplification of only one of the two alleles makes a heterozygous point mutation appear to be a homozygous one, causing false negative errors. (b) False positive errors arising from point mutations generated by the whole-genome amplification. (c) Chimera formation, in which different parts of the original genome are artificially connected by whole-genome amplification.

Chimera rate. Chimera formation is an artifact of WGA that generates reads that can be mapped to different parts of the genome that are not physically linked (Figure 7c). It prevents the identification of true structural variations, and single-cell structural variations are still challenging to detect owing to this technical problem. The chimera rate is defined as the percentage of reads that exhibit improper connections in the reference genome, i.e., those that show an abnormal fragment size or indicate interchromosomal connections. Figure 6e shows the chimera rates of the WGA kits. The chimera rates of the Sigma-Aldrich and Rubicon kits are much higher than those of the other kits.

Unmappable rate. The unmappable rate is the probability that junk sequences are generated in the WGA process, which can arise from the formation of primer dimers, short DNA fragments, and/or nonspecific incorporations. A large fraction of unmappable reads reduces the cost effectiveness of genome sequencing and the completeness of the genome coverage. Figure 6f shows the unmappable rates of the WGA kits. The Qiagen and Yikon kits showed the lowest unmappable rates.

We would like to point out that all of the comparisons described above were done using the existing kits available to us and that each would improve with time. As important as the specific results is how we performed the comparison. It is fair to say that each method or kit has its own advantages and disadvantages, and that selection should be based on specific applications.

For accurate CNV determination, DOP-PCR and MALBAC kits offer the highest reproducibility and lowest coefficients of variation. MALBAC kits and Qiagen’s MDA kit are well suited for simultaneous measurements of SNVs and CNVs. For calling de novo SNV, the false negative and false positive rates need to be considered. Yikon’s MALBAC kit has the lowest false negatives (the allele dropout rate), while Qiagen’s MDA kit has the lowest false positive rate. When sequencing cost effectiveness is considered, a low unmappable rate is desirable for high effective coverage.
We performed our comparison using small sample tubes instead of microfluidic devices; the performance parameters might change under microfluidic conditions and might be different for bacterial (small) and human (large) genomes. Although expensive, a similar comprehensive comparison of the chemistries for human genomes would be worth doing under microfluidic conditions.

**STUDY OF GERM CELLS**

**Phasing of Human Genomes with Single-Cell Genomics**

One central goal of human genetics is to characterize the genetic variations of individual human beings and study their correlations with function-related traits (1, 44). Human genomes are diploid in nature, and human genes often contain multiple coding and regulatory elements with variations among different populations. Different phenotypic consequences can often arise depending on whether the genetic variations are associated on the same chromosome (in *cis*) or on opposite homolog chromosomes (in *trans*) (2, 62). One example is the phenomenon of compound heterozygosity (28, 66), in which the two heterozygous nonsynonymous mutations in a certain gene could result in either one normal version (as in *cis*) or no normal version (as in *trans*) of the gene. A complete understanding of genetic variations and their consequences in human diseases cannot be obtained without resolving the combination of genetic variations at different loci on the same chromosome—that is, the haplotypes (16, 29). Without the haplotype information, the description of individual human genomes is incomplete, and the functional interpretation can be error prone.

However, owing to the limited read lengths of the currently available sequencing techniques, it is often challenging to separate the two homolog chromosomes in genetic analyses to obtain the association information between the genetic variations. Molecular cloning technologies have been developed for phasing individual genomes (33, 38, 59, 76). However, the cloning procedures are often lengthy and labor intensive and therefore may not be scalable with the increasing demand for individual genome sequencing.

Single-cell genome sequencing provides an alternative way to obtain phasing information on human individuals. Individual metaphase chromosomes have been isolated from actively dividing cells by techniques such as laser microdissection (41), microfluidics (14), or flow cytometry (74) before being whole-genome amplified and genotyped to obtain the association information for genetic variations on individual chromosomes. Peters et al. (50) further developed a low-cost DNA haplotyping process by diluting genomic DNA into 384 wells, each containing subcellular-level genomic DNA. DNA molecules from each well were then separately whole-genome amplified and sequenced to obtain the phasing information. Hou et al. (24), Kirkness et al. (32), and Lu et al. (40) used whole-genome-amplified individual germ cells for haplotype analyses, further increasing the haplotype block size to the chromosome-spanning level (Figure 8). These two categories of haplotyping approaches do not require molecular cloning, cell culturing, or sophisticated instrumentation for chromosome isolation, potentially enabling haplotype-resolved comprehensive genetic studies and clinical applications in the near future.

**Study of Meiotic Recombination in Human Germ Cells**

Meiotic recombination is essential to the proper segregation of homolog chromosomes, which results in the exchange of genetic information through crossover events and creates diversity for evolution (8). Abnormality in generating crossovers of homolog chromosomes is the leading
cause of miscarriage and birth defects (13). Population analyses such as linkage disequilibrium and pedigree studies are widely used in studying the highly uneven pattern of meiotic recombination across the human genome. However, population analyses yield results that are averaged among individuals and affected by evolutionary pressures, therefore often masking the quickly evolving or individual-specific features of the recombination-active regions (23, 30, 34).

Single-germ-cell genome sequencing offers a novel approach for studying meiotic recombination at the level of individual human beings. Each germ cell is unique in genome construction because of the differences in recombination, and sequencing multiple single germ cells from an individual allows investigators to map individual recombination events and study germline genome instability. Single-cell studies have been published using donor sperm (40, 68) as well as oocytes (24) for mapping recombination in individual human beings. Wang et al. (68) utilized a microfluidic device to separate and sequence single sperm, thereby mapping the recombination distributions and studying the gene conversion and de novo mutation events of an individual. Lu et al. (40) sequenced ∼100 single sperm cells using the MALBAC method and mapped the recombination events of an individual at high resolution (Figure 9). This study revealed the correlation of a decreased crossover frequency with an increase of autosomal aneuploidy rate from human sperm, recapitulating the importance of meiotic crossover in maintaining genome stability in germ cells.
Hou et al. (24) utilized the first and the second polar bodies of individual oocytes to determine the crossover maps of female individuals and to study the chromatid interference effect in oocytes. This study also utilized the information of the two polar bodies to deduce the genomes of the oocyte pronuclei, enabling noninvasive preimplantation genetic diagnosis (PGD) and preimplantation genomic screening (PGS) for in vitro–fertilized embryos (Figure 10).

Preimplantation Genetic Diagnosis and Genomic Screening in In Vitro Fertilization

In vitro fertilization is a technique of assisted reproductive technology by which an egg is fertilized by sperm outside the human body. The fertilized egg is cultured in vitro for 2–6 days before being implanted into the uterus with the intention of creating a successful pregnancy. Nucleic acid...
First polar body

Second polar body

Female pronucleus predicted

Female pronucleus confirmed

Total chromatid copy numbers conserved in the whole oocyte

Chromosome

Healthy baby

Birth difficulty or baby with genetic disorder

Baby with dominant disease allele
materials obtained from the in vitro–fertilized embryos can be used to perform PGD and PGS to avoid the inheritance of pathogenic mutations and chromosome abnormalities, respectively. PGD/PGS can be performed on the two polar bodies of the oocytes, one of the blastomeric cells from day-3 embryos, or several trophectoderm cells from day-5 blastocyst embryos, in which only one or a few cells can be obtained for PGD/PGS purposes. Performing WGA on these materials enables comprehensive chromosome analyses on various genome analytical platforms, such as a comparative genomic hybridization array (52), single-nucleotide polymorphism (SNP) array (63), or multiplex quantitative PCR (65). The rapid development of high-throughput sequencing techniques further reduced the cost and increased the precision and resolution of chromosome-level PGD/PGS (24, 26, 71). Treff et al. (64) utilized targeted high-throughput sequencing to perform PGD of monogenic disease, and efforts were made to perform both point mutation PGD and chromosomal PGS in the same embryo (10, 58). By using MALBAC next-generation sequencing, we were able to report a fully integrated pipeline for combined monogenic PGD and chromosomal PGS and a live birth free of a specific pathogenic mutation carried by the child’s parents (22). We envision that the advancement of WGA techniques will enable clinical trials with fully integrated pipelines for combined monogenic PGD and chromosomal PGS in the near future.

STUDY OF GENOME EVOLUTION IN CANCER
Cancer is a genomic disease (67). Recent advances in single-cell WGA methods have allowed determination of CNVs at a large scale (48, 70) and at the single-gene level (15) as well as SNVs of a single cancer cell (18, 70, 75), in primary tissue, or in CTCs in blood (9, 39, 49).

A pair of companion studies used single-cell WGA and exome sequencing to examine tumor heterogeneity (25, 73). These studies found hundreds of SNVs—which differ from cell to cell, revealing the genetic complexity of these tumors—and applied population genetic analysis to study their clonal composition. However, our independent analysis of these data indicates that the vast majority of these SNVs are due to sequencing artifacts or possibly contaminations, and only a handful, if any, may be attributed to true heterogeneity of the tumor samples (A. Chapman, C. Zong & X.S. Xie, unpublished results). Of the 711 tumor-specific SNVs reported in one of the studies (25), we found that 42% of them were actually present at a lower level in the normal tissue as well. These are likely to be SNPs that were mistakenly not identified in the normal tissue (false negatives) rather than tumor-specific SNVs. Furthermore, 58% of the remaining SNVs reported were present in dbSNP, a database of germline mutations known to occur in the human population. Because new mutations in a tumor arise randomly, it would be highly unlikely that a significant number of them happen to coincide with known mutations present in the general population. Instead, an alternative explanation for these mutations could be contaminations from the operators or other human sources. Indeed, we observed that of those remaining SNVs not found in dbSNP, 84% were found in unrelated samples—either normal control cells from the same study or cells from the companion study. In summary, more than 96% of the SNVs identified as being tumor specific in that report (25) appear likely to be contaminant or sequencing artifacts.

We now give a few examples of new information available from single-cell genomics.

Measurement of Spontaneous Whole-Genome Mutation Rate
In calling SNVs from a single cell by WGA (24, 68, 73), one challenge is high false negative rates caused by allele dropout, and another is false positives associated with amplification and sequencing errors (either random or systematic) (42). Figure 11 shows 35 unique SNVs in a human cancer cell line (SW480) newly acquired during 20 cell divisions. Adjusting for a 70% detection efficiency for
Detecting newly acquired single-nucleotide variations (SNVs) with no false positives and estimation of the mutation rate of a human cancer cell line (SW480). (a) Experimental design. A single ancestor cell was chosen and cultured for ∼20 generations. The vast majority of cells were used to extract DNA for bulk sequencing to represent the ancestral cell’s genome. A single cell from this culture was chosen for another expansion of four generations. The kindred cells were isolated for single-cell whole-genome amplification. Single-cell samples C1, C2, and C3 were used for high-throughput sequencing. (b) Locations of the 35 newly acquired SNVs on the chromosomes of a single cell. (c) Next-generation sequencing data of a newly acquired SNV. The SNV (C→G) existed in the high-throughput data of all three kindred cells but not in the bulk data. Additional abbreviation: MALBAC, multiple annealing and looping–based amplification cycles. Figure adapted from Reference 78.

heterozygous SNVs, Zong et al. (78) estimated that ∼49 mutations occurred in the 20 generations, yielding a mutation rate of ∼2.5 nucleotides per cell generation, which is consistent with the estimate based on the bulk data.

At this rather slow spontaneous mutation rate, cancer development would take a long time. What happens at the very beginning of cancer development has been a long-standing question, and single-cell WGA is well suited to help answer it.

Does Copy-Number Variation Precede Single-Nucleotide Variation?

In a bulk tissue sample at an early stage of cancer development, detection of abnormal CNVs is often difficult, especially when the number of cells with abnormal CNVs is small. Single-cell genomic analysis is essential in evaluating the relationship, if any, between CNVs and SNVs.
Hyperproliferation
Adenomatous polyps
High-grade dysplasia
Adenocarcinoma
Invasive cancer

Abnormal cell growth
We have carried out single-cell genomic analyses of colonoscopy biopsies at different adenoma stages (L. Huang, S. Ding & X.S. Xie, unpublished results). Some single cells in stage II adenoma have CNVs in the tumor suppressor gene \( \text{APC} \) (a reduction in copy number from two to one) as well as SNVs in numerous cancer-related genes. Interestingly, we found that although some single cells exhibited CNV reduction in \( \text{APC} \) without SNVs, all single cells with SNVs that have been reported as somatic mutations of colon and other cancers in the Catalogue of Somatic Mutations in Cancer (COSMIC) database showed the CNV reduction in \( \text{APC} \). Moreover, we did not see any single cell that has colon cancer–related SNVs but no CNVs in an adenoma. Figure 12 shows such data for the COSMIC gene \( \text{KCNB1} \), which encodes an ion channel. These data indicate that the CNV in \( \text{APC} \) precedes the SNV in colon cancer development, at least in the particular colony examined.

Furthermore, we found that the single cells from the same adenoma exhibited CNV patterns reproducible among all the cells, indicating that these cells might be derived from a single stem or progenitor cell in which the CNVs first arose. Thus, we have established a correlation between the SNVs and CNVs and propose that the SNVs are generated as a consequence of abnormal CNVs in the genome, or arise after the abnormal CNVs are acquired in the original populating cell. If confirmed to be general for driving SNVs, this result could have significant implications for the genesis of cancer. Suffice it to say that this experiment underscores the importance of single-cell genomics in understanding cancer.

### Circulating Tumor Cells

Originating from primary tumors, CTCs enter peripheral blood and seed metastases, which account for 90% of cancer-related deaths. The genome sequencing of CTCs could offer noninvasive prognosis or even diagnosis but has been hampered by low single-cell genome coverage of scarce CTCs.

Ni et al. (49) applied MALBAC for WGA of single CTCs from lung cancer patients and observed characteristic cancer-associated SNVs and indels in exomes of CTCs. These mutations provided information needed for individualized therapy (53), such as drug resistance and phenotypic transition, but were heterogeneous from cell to cell.

Ni et al. (49) also discovered that, unlike the highly heterogeneous point mutations, the CNV patterns of CTCs are reproducible from CTC to CTC within a patient and even within different patients of the same cancer type, but are distinctly different among different cancer types (Figure 13). Furthermore, the reproducible CNV patterns of CTCs are similar to those of the metastatic tumor. This result raised intriguing questions about the genesis of metastasis. It is evident that gains and losses in copy numbers of certain chromosome regions are selected for...
Figure 13
Reproducible copy-number variation (CNV) patterns of seven lung cancer patients. Patient 1 (P1) experienced a phenotypic transition from adenocarcinoma (ADC) in the lung to small-cell lung cancer (SCLC) in the metastatic liver, patients 2–6 (P2–6) have ADC, and patient 7 (P7) has a mix of ADC and SCLC. (a) Reproducible CNV patterns of eight circulating tumor cells (CTCs) in P1. The patterns are different from that of the primary tissue (first row) but the same as that of the metastatic tissue (second row). (b) CNV patterns of CTCs from the seven patients. (c) Clustering analyses of CTCs based on the CNVs. CTCs from P1 and P7 (SCLC patients) were well separated from CTCs from P2–P6 (ADC patients). The $y$ axis is the cluster distance constructed using Ward’s method based on Euclidean distances between the patients’ CNVs. Figure adapted from Reference 49.

metastases. The finding that the CNV patterns of CTCs are cancer or tissue dependent offers the potential for noninvasive cancer diagnosis based on the CNV patterns.

DISCLOSURE STATEMENT
S.L. and X.S.X. are coauthors on a patent applied for by Harvard University for MALBAC technology and are cofounders of Yikon Genomics.

ACKNOWLEDGMENTS
We thank Chenghang Zong for his contribution to the invention of MALBAC. We thank Yaqiong Tang, Dandan Cao, and Hongshan Guo for technical assistance in data collection and Guangyu Zhou for help with data analysis. We are indebted to our collaborators Fuchou Tang, Jie Qiao, Xiaohui Ni, Fan Bai, Wei Fan, Liya Xu, Jie Wang, Ning Zhang, Youyong Lu, Liying Yan, Yu Hou, Zhe Su, Yan Gao, Shigang Ding, Hejun Zhang, Ruiqiang Li, Mingyu Yang, Jinsen Li, Jessica Sang, and Yanyi Huang for their contributions to the work reviewed in this article. We are grateful for funding from the National Natural Science Foundation of China (21327808), US National Institutes of Health (NIH) National Human Genome Research Institute grants (HG005097-1 and HG005613-01), and an NIH Director’s Pioneer Award (5DP1CA186693) to X.S.X. as well as a National Cancer Institute grant (5R33CA174560). The work at the Biodynamic Optical Imaging Center (BIOPIC) has been supported by Peking University 985 special funding for collaboration with hospitals and funding from Guangxi Wuzhou Zhongheng Group Co., Ltd.

LITERATURE CITED
7. Cheung VG, Nelson SF. 1996. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *PNAS* 93:14676–79
Contents

A Mathematician’s Odyssey
  *Walter Bodmer* ................................................................. 1

Lessons from modENCODE
  *James B. Brown and Susan E. Celniker* .................................. 31

Non-CG Methylation in the Human Genome
  *Yupeng He and Joseph R. Ecker* ........................................... 55

Single-Cell Whole-Genome Amplification and Sequencing:
  Methodology and Applications
  *Lei Huang, Fei Ma, Alec Chapman, Sijia Lu, and Xiaoliang Sunney Xie* .... 79

Unraveling the Tangled Skein: The Evolution of Transcriptional
  Regulatory Networks in Development
  *Mark Rebeiz, Nipam H. Patel, and Veronica F. Hinman* .................. 103

Alignment of Next-Generation Sequencing Reads
  *Knut Reinert, Ben Langmead, David Weese, and Dirk J. Evers* ............ 133

The Theory and Practice of Genome Sequence Assembly
  *Jared T. Simpson and Mihai Pop* .......................................... 153

Addressing the Genetics of Human Mental Health Disorders
  in Model Organisms
  *Jasmine M. McCammon and Hazel Sive* .................................... 173

Advances in Skeletal Dysplasia Genetics
  *Krista A. Geister and Sally A. Camper* ................................... 199

The Genetics of Soft Connective Tissue Disorders
  *Olivier Vanakker, Bert Callewaert, Fransiska Malfait, and Paul Coucke* .... 229

Neurodegeneration with Brain Iron Accumulation: Genetic Diversity
  and Pathophysiological Mechanisms
  *Esther Meyer, Manju A. Kurian, and Susan J. Hayflick* .................... 257

The Pathogenesis and Therapy of Muscular Dystrophies
  *Simon Guiraud, Annemieke Aartsma-Rus, Natassia M. Vieira, Kay E. Davies,
  Gert-Jan B. van Ommen, and Louis M. Kunkel* ........................... 281
Detection of Chromosomal Aberrations in Clinical Practice: From Karyotype to Genome Sequence
Christa Lese Martin and Dorothy Warburton ......................................................... 309

Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality
David M. Evans and George Davey Smith ............................................................ 327

Eugenics and Involuntary Sterilization: 1907–2015
Philip R. Reilly ........................................................................................................ 351

Noninvasive Prenatal Genetic Testing: Current and Emerging Ethical, Legal, and Social Issues
Mollie A. Minear, Stephanie Alessi, Megan Allyse, Marsha Michie, and Subhashini Chandrasekharan ......................................................... 369

Errata
An online log of corrections to *Annual Review of Genomics and Human Genetics* articles may be found at http://www.annualreviews.org/errata/genom