COMMENTARY

Haplotype resolution at the single-cell level

Andrew C. Adey[a,b,1]

One of the most rapidly advancing areas in genomics in recent years has been the explosion of platforms to assess properties at the single-cell level. The driving force behind these advancements has been a renaissance of sorts in how we think about the architecture and heterogeneity present in a population of cells (1, 2). While these concepts are not new, the advancement of technologies to assess latent cell heterogeneity has enabled this new era. One of the earliest iterations of single-cell "omics" was the sequencing of the genome itself, which has seen a rapid proliferation of technological advancements (3), including a novel platform by Chu et al. (4) presented in PNAS. Over the years, a number of studies have utilized single-cell genome sequencing to reveal novel insights into a wide range of biological systems from somatic mutation profiling, both in (5) and out (6) of the cancer context, to the determination of meiotic recombination rates (7).

To date, multiple platforms have been developed to sequence the genome of individual cells, each aimed at addressing specific types of variation. High-cell-count, low-depth strategies are tailored to assess copy number alterations: gains or losses of stretches of DNA up to entire chromosomes (8). However, these methods fail to capture point variation, single-nucleotide variants or very short insertions or deletions (indels), that may play a critical role in tumor clonal progression, or somatic variation in other contexts (9, 10). Strategies to ascertain point variation in single cells, which will be focused on in this commentary, involve the amplification of DNA and then sequencing to a high depth of coverage for each cell, resulting in higher costs and, typically, reduced cell numbers (3).

One of the primary challenges faced by these higher coverage methods is that they rely on even amplification of each copy of the genome. Unfortunately, stretches of DNA often fail to amplify (Fig. 1*A*). When this occurs on both homologous copies of a locus, it results in a complete lack of coverage and all variants in that region will be missed. When it occurs on only one of the two copies, it results in something called allelic dropout, where only one homolog's sequence variants are ascertained, making it appear as though all variants in that region are homozygous, when, in fact, a number of variants may be heterozygous. Furthermore, there will likely be additional heterozygous variants present only on the allele that dropped out that will be missed.

A second challenge is that polymerases are not perfect. During the amplification of DNA, errors can occur that are frequently impossible to distinguish from real heterozygous variants or even homozygous variants (if the other allele dropped out and the error occurred early on in the amplification process). This is particularly challenging for the ascertainment of de novo variants that may be observed in only a single cell and is particularly valuable for studies that rely on sporadic variants for lineage tracing in cancer or development. While several methods have been developed to reduce the mutational burden (11, 12), they still occur and are unable to be distinguished from a true de novo variant.

A third shortcoming that is shared between single-cell sequencing techniques and the majority of bulk cell genome sequencing is that these platforms are unable to discriminate between the homologous copies of chromosomes that occur in any nonhaploid organism. This shortcoming is exactly that: a short length of sequence contiguity that prevents the association of variants with one another on the same homolog, or haplotype. Haplotypes play an integral part in population genetics, linkage and association studies, clinical genetics, and even for the study of *cis*-regulatory effects on allele-specific gene expression.

Over the years, a number of strategies have been deployed to experimentally haplotype-resolve, or phase, genomes (as opposed to haplotype inference, which relies on population allele frequencies, and is often unable to phase rare variants) (13). Much like single-cell genome sequencing, these techniques break down into two categories: obtaining a relatively low length of haplotype contiguity (thousands to tens of thousands of base pairs) with dense sampling within those regions (>90% of variants sequenced) (14, 15) or

[a]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR 97239-3098; and [b]Knight Cardiovascular Institute, Oregon Health & Science University, Portland, OR 97239-3098

Author contributions: A.A. wrote the paper.

The author declares no conflict of interest.

See companion article on page 12512.

[1]Email: adey@ohsu.edu.

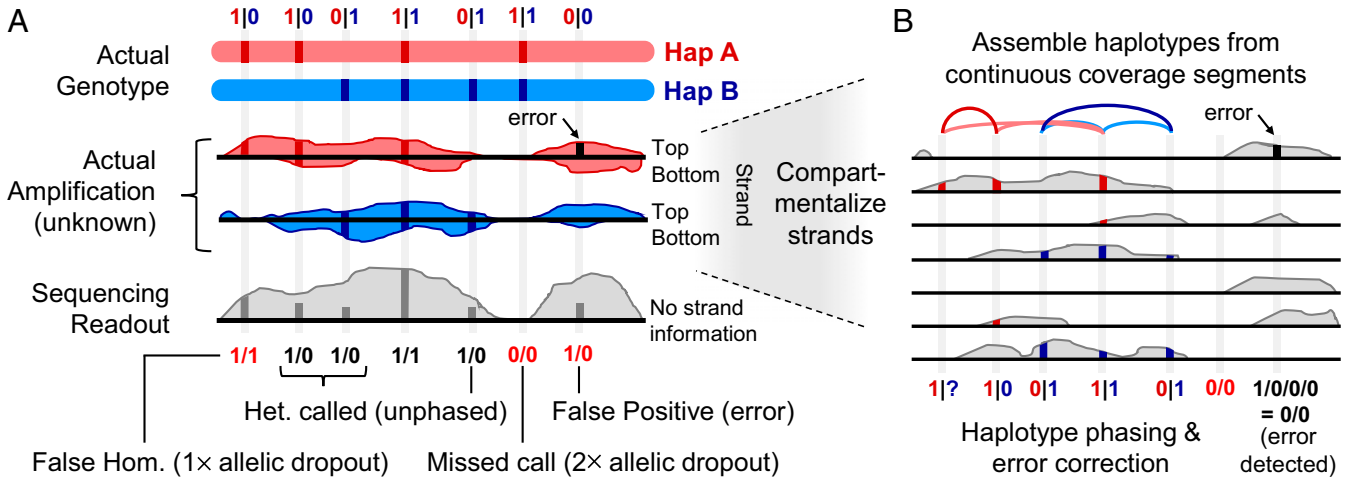

Fig. 1. (*A*) Common challenges of identifying single-nucleotide variants (SNVs) from single-cell genome sequencing data. Failure of individual strands to amplify can result in allelic dropout and false homozygous (Hom.) calls. Even when both alleles are amplified, the haplotype (Hap) assignment is not possible. Lastly, errors can look identical to heterozygous (Het.) SNVs and cannot be detected. (*B*) Diluting individual strands into separate amplification compartments enables the association of variants to the same haplotype (haplotype phasing; dark loops at the top indicate the association of heterozygous variants to a haplotype), as well as the detection of errors when only one of the four amplified strands harbors a given variant.

a high length of contiguity (millions of base pairs) but at the expense of sampling density such that relatively few variants are phased from each stretch (<10%) (16–19). This does include a method that utilized mitotic chromosome sorting of single cells (20); however, this approach produced very sparse information and was not suitable for variant calling from the microarray or sequencing that was performed, instead relying on previous genotype calls.

In PNAS, Chu et al. (4) build upon these existing concepts of haplotype resolution and apply them to single-cell genome sequencing (11). In the new microfluidics-based technology, single-stranded sequencing using microfluidic reactors (SISSOR), the DNA of a single cell is isolated and denatured, which also breaks up the DNA into megabase pair-sized segments. This material is then distributed into a set of 24 individual nanoliter compartments, such that the probability of having two strands from the same region of the genome making their way into the same compartment is low. This limited dilution strategy has been a staple of haplotype resolution technologies (13); however, such efforts typically utilize multiple cells' worth of DNA and no denature strands before compartmentalization.

The resulting readout of SISSOR is 24 libraries with interspersed islands of sequence coverage throughout the genome (Fig. 1*B*). Each island is the result of a long stretch of single-stranded DNA making its way into the compartment and undergoing amplification. Therefore, each island is highly likely to be sequence-derived from a single strand of a single haplotype, enabling all variants that are called within the island to be associated with one another on the same haplotype.

In this work, Chu et al. (4) demonstrate SISSOR in a proof-of-principle experiment in which three individual cells of a fibroblast cell line were sequenced to an average depth of 65-fold. This coverage captured ∼64% of the genome, which the authors ascribe to fragment loss during partitioning based on the lack of a systematic bias in regional dropout, suggesting future improvements to fragment distribution may result in more comprehensive coverage. For the rest of the genome, SISSOR produced islands of coverage with an N50 of 1 Mbp; that is, 50% of the total size of all coverage islands is present in islands that are at least 1 million bp in length, with the largest observed island containing an impressive 9 Mbp of strand- and haplotype-specific DNA. Typically, methods that produce contiguity at this range fall into the sparse category of haplotype-resolution techniques (13); however, the dense coverage and long range produced by SISSOR position it as a best-of-both-worlds technology. Indeed, the authors were able to achieve an N50 haplotype block length of over 7 Mbp, with the largest phased block extending over 28 Mbp at a phasing accuracy comparable to that of exiting platforms. This was accomplished by leveraging information from all three cells, thus overcoming the relatively low percentage of the genome covered in each cell, which resulted in haplotype phasing for nearly the entire human genome (2.63 billion bp). The authors further demonstrate the haplotype resolution of the highly polymorphic HLA region, a locus with clinical benefits. Even if one were to ignore the single-cell nature of SISSOR, it stands as a powerful haplotype-phasing technology.

While the haplotype-resolution aspect of SISSOR is valuable, the real power, from the single-cell genomics perspective, is that the strand amplification information is no longer completely unknown. This enables a much higher level of accuracy that can be achieved in instances where both strands from the same haplotype harbor the same variant, as the probability of the same mutation occurring in two separate amplification reactions is extremely low. Similarly, variants that are observed on one strand, but not the other, of the same haplotype, or variants that are not at an allele frequency of 1, are likely errors that can be discarded (Fig. 1*B*). Based on these principles, Chu et al. (4) developed a novel variant-calling algorithm and then compared their calls for bases covered by all three cells that were also covered by bulk cell genome sequencing that served as a reference. The vast majority were concordant; however, 19 discordant calls were identified: 10 of which were validated by the sequencing of BAC clones, 5 of which were confirmed as de novo mutations between the lineage used for the reference genome sequencing versus those in this study, and 4 of which were unaccounted for that are potential errors, for a maximum error rate of 1×10^{-8}.

It is important to note that this work by Chu et al. (4) is a proof-of-principle study, performed on a low number of cells from an immortalized cell line as opposed to a primary tissue sample. Furthermore, there are several key limitations, the first of which is the relatively low coverage of the genome, with any individual cell missing a third of the genome on average. The second major limitation is the cost per cell, in both labor (construction of 24 libraries per cell) and price (again, 24 libraries per cell), which is likely the driving reason why only three cells were profiled. While additional microfluidics integration may reduce this burden, as suggested by the authors, the sequencing cost will still be relatively high at 65-fold coverage per cell, thus making it only reasonable to deploy in studies for which high accuracy and haplotype information are worth the added cost.

1 Trapnell C (2015) Defining cell types and states with single-cell genomics. *Genome Res* 25:1491–1498.
2 Chen X, et al. (2016) Single-cell analysis at the threshold. *Nat Biotechnol* 34:1111–1118.
3 Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: Current state of the science. *Nat Rev Genet* 17:175–188.
4 Chu WK, et al. (2017) Ultraaccurate genome sequencing and haplotyping of single human cells. *Proc Natl Acad Sci USA* 114:12512–12517.
5 Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472:90–94.
6 McConnell MJ, et al. (2013) Mosaic copy number variation in human neurons. *Science* 342:632–637.
7 Lu S, et al. (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338:1627–1630.
8 Vitak SA, et al. (2017) Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* 14:302–308.
9 Xu X, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148:886–895.
10 Lodato MA, et al. (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350:94–98.
11 Fu Y, et al. (2015) Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc Natl Acad Sci USA* 112:11923–11928.
12 Dong X, et al. (2017) Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* 14:491–493.
13 Snyder MW, Adey A, Kitzman JO, Shendure J (2015) Haplotype-resolved genome sequencing: Experimental methods and applications. *Nat Rev Genet* 16:344–358.
14 Kitzman JO, et al. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* 29:59–63.
15 Adey A, et al. (2013) The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* 500:207–211.
16 Peters BA, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487:190–195.
17 Selvaraj S, R Dixon J, Bansal V, Ren B (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31:1111–1118.
18 Amini S, et al. (2014) Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat Genet* 46:1343–1349.
19 Zheng GXY, et al. (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 34:303–311.
20 Fan HC, Wang J, Potanina A, Quake SR (2011) Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* 29:51–57.