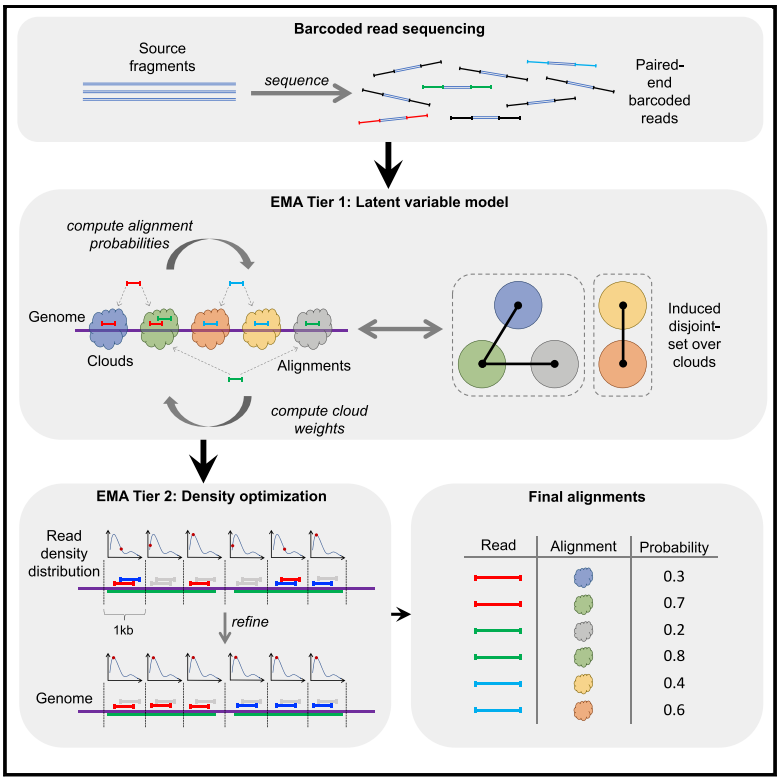
Cell Systems

Report

Statistical Binning for Barcoded Reads Improves Downstream Analyses

Graphical Abstract



Highlights

- We devise a two-tiered statistical binning model to align barcoded reads to the genome
- We can map highly homologous regions to uncover rare variants important in disease
- Our method greatly improves downstream genotyping and haplotyping accuracy
- We determine not only alignments but also interpretable alignment *probabilities*

Authors

Atiya Shajii, Ibrahim Numanagić, Christopher Whelan, Bonnie Berger

Correspondence

bab@mit.edu

In Brief

Researchers are applying barcoded read sequencing to capture longer-range information in the genome at low error rates. We introduce a two-tiered statistical binning model, named EMA, which probabilistically assigns reads to "clouds" and then optimizes read assignments within clouds based on read densities. Unlike previous approaches, our efficient method enables alignment to highly homologous regions of the genome important in disease and substantially improves downstream genotyping and haplotyping. Our method also uncovers rare variants in clinically important genes.

Shajii et al., 2018, Cell Systems 7, 219–226
August 22, 2018
https://doi.org/10.1016/j.cels.2018.07.005

CellPress

Cell Systems
Focus on RECOMB

Report

Statistical Binning for Barcoded Reads Improves Downstream Analyses

Atiya Shajii,[1] Ibrahim Numanagić,[1,2] Christopher Whelan,[3,4,5,6] and Bonnie Berger[1,2,7,*]
[1]Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Data Sciences Platform, Broad Institute, Cambridge, MA, USA
[4]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
[5]Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA
[6]Department of Genetics, Harvard Medical School, Boston, MA, USA
[7]Lead Contact
*Correspondence: bab@mit.edu
https://doi.org/10.1016/j.cels.2018.07.005

SUMMARY

Sequencing technologies are capturing longer-range genomic information at lower error rates, enabling alignment to genomic regions that are inaccessible with short reads. However, many methods are unable to align reads to much of the genome, recognized as important in disease, and thus report erroneous results in downstream analyses. We introduce EMA, a novel two-tiered statistical binning model for barcoded read alignment, that first probabilistically maps reads to potentially multiple "read clouds" and then within clouds by newly exploiting the non-uniform read densities characteristic of barcoded read sequencing. EMA substantially improves downstream accuracy over existing methods, including phasing and genotyping on 10x data, with fewer false variant calls in nearly half the time. EMA effectively resolves particularly challenging alignments in genomic regions that contain nearby homologous elements, uncovering variants in the pharmacogenomically important *CYP2D* region, and clinically important genes *C4* (schizophrenia) and *AMY1A* (obesity), which go undetected by existing methods. Our work provides a framework for future generation sequencing.

INTRODUCTION

Sequencing is the most fundamental operation in genomics, transcriptomics, and metagenomics. As sequencing technologies continue to advance beyond the initial introduction of next-generation sequencing (NGS), we have begun to see the emergence of the so-called "third-generation" sequencing platforms, which seek to improve on the standard short-read sequencing that has thus far been at the heart of most NGS (Mardis, 2017). Several organizations are at the center of this new sequencing revolution, including Pacific Biosciences (Eid et al., 2009), Oxford Nanopore (Wang et al., 2014), and 10x Genomics (Zheng et al., 2016). While the former two have developed sequencing technologies that produce much longer physical reads (e.g., 10–200 kb) at typically higher error rates, the latter is an example of a barcoded sequencing technology, which typically produces short reads (up to 300 bp) with low error rates (Goodwin et al., 2016).

At a high level, barcoded sequencing is any sequencing method where long DNA fragments are sheared and the sheared pieces have some identifier ("barcode") relating them back to the source fragment. These barcodes can be explicit (a physical barcode is ligated to each sheared piece, e.g., as in 10x sequencing) or implicit (the fragments are distributed to identifiable wells, e.g., as in Illumina's TruSeq Synthetic Long-Read sequencing, henceforth referred to as TruSeq SLR). These sheared pieces are then sequenced using standard short-read sequencing, thereby producing barcoded short reads (Figure 1A). Other barcoded sequencing technologies include Illumina's Continuity Preserving Transposition technology (CPT-seq), Complete Genomics' Long Fragment Read technology, Drop-seq, and CEL-Seq2 (Zheng et al., 2016; McCoy et al., 2014; Macosko et al., 2015; Hashimshony et al., 2016; 10x Genomics, 2018). Because they help identify the original source fragment, these barcodes implicitly carry long-range information, which can have a significant impact on alignment and many downstream analyses such as structural variation detection and phasing.

Barcoded reads have several advantages over physically long reads. First, and perhaps most important, barcoded read sequencing is substantially cheaper than long-read sequencing; whereas PacBio's and Oxford Nanopore's sequencing platforms currently cost anywhere from $750 to $1,000 per GB of data, barcoded sequencing is a comparatively cheap add-on to standard short-read sequencing and therefore bears the same cost (e.g., 10x sequencing costs $30 per GB plus a $500 overhead per sample) (Goodwin et al., 2016). Second, the error profile of barcoded reads is very similar to that of standard short reads (roughly 0.1% substitution errors), which enables us to augment the tools and algorithms that have been developed for regular short reads to handle their barcoded counterparts. By contrast, long-read sequencing typically produces high rates of erroneous indels (ranging from 12%–13%), which presents a challenge when trying to use preexisting algorithms. Beyond these

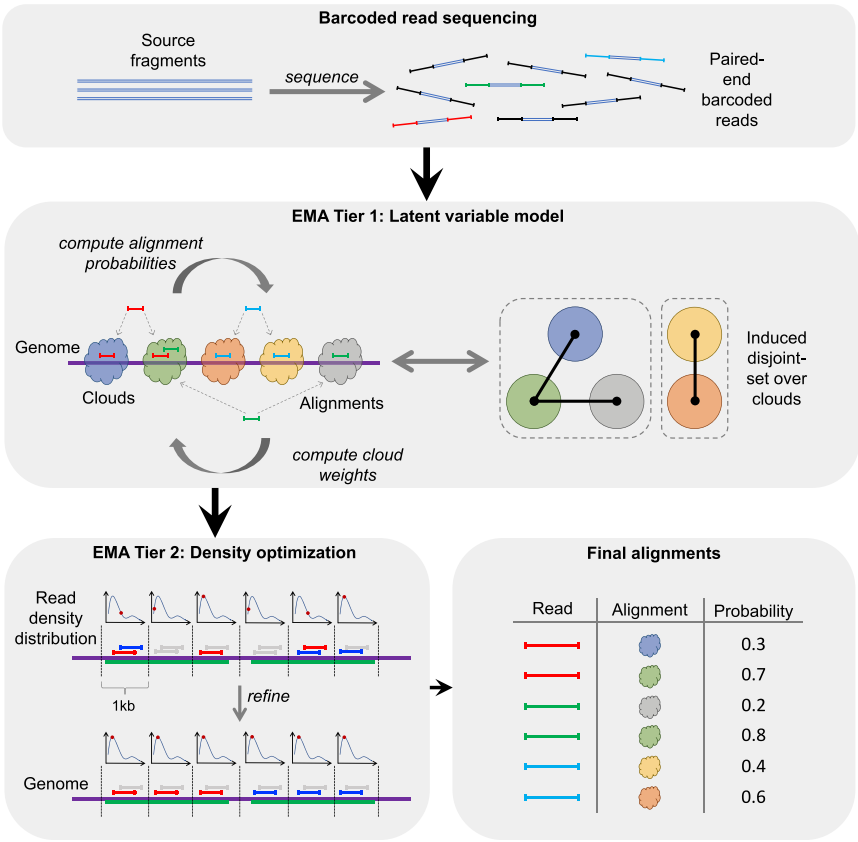


Figure 1. Overview of EMA Pipeline

(A) An idealized model of barcoded read sequencing, wherein some number of unknown source fragments in a single droplet or well are sheared, barcoded, and sequenced to produce barcoded reads.

(B) EMA's "read clouds" are constructed by grouping nearby-mapping reads sharing the same barcode; these clouds represent possible source fragments. EMA then partitions the clouds into a disjoint-set induced by the alignments, where two clouds are connected if there is a read aligning to both; connected components in this disjoint-set (enclosed by dashed boxes) correspond to alternate possibilities for the *same* unknown source fragment. EMA's latent variable model optimization is subsequently applied to each of these connected components individually to deduce each of the potentially many fragments sharing this barcode.

(C) EMA applies a novel read density optimization algorithm to clouds containing multiple alignments of the same read to pick out the most likely alignment, by optimizing a combination of alignment edit distances and read densities within the cloud. The green regions of the genome are homologous, thereby resulting in multi-mappings within a single cloud.

(D) While the read density optimization operates within a single cloud, EMA's latent variable model optimization determines the best alignment of a given read between different clouds and produces not only the final alignment for each read but also interpretable alignment *probabilities* (see Figure S1).

advantages, barcoded reads are compatible with doing hybrid-capture exome sequencing, and not sequencing introns (which is not possible with long-read sequencing technologies, as a contiguous long-read cannot only sequence the exons in a gene). This is a very substantial additional cost-advantage for barcoded reads (exome sequencing can mean a 10- to 20-fold reduction in sequencing cost over whole genome sequencing) (Schwarze et al., 2018). These and other benefits have led to the recent proliferation of barcoded sequencing technologies for various use cases; for example, 10x and TruSeq SLR sequencing for whole genome sequencing, as well as Drop-seq and CEL-Seq2 for single-cell RNA sequencing (RNA-seq), 10x for single-cell VDJ sequencing and so on (Zheng et al., 2016; McCoy et al., 2014; Macosko et al., 2015; Hashimshony et al., 2016; 10x Genomics, 2018). A comprehensive review of many of these methods is available (Ziegenhain et al., 2017). Furthermore, barcoded sequencing is also playing a greater role in downstream applications such as the generation of transcriptomic profiles (Cleary et al., 2017).

As with virtually all sequencing data, the first step in the analysis pipeline for barcoded reads is typically alignment. While barcoded reads can, in theory, be aligned by a standard short-read aligner (e.g., CORA [Yorukoglu et al., 2016] (COmpressive Read-mapping Accelerator), BWA (Burrows-Wheeler Aligner) [Li and Durbin, 2009], Bowtie2 [Langmead and Salzberg, 2012]), this would fail to take advantage of the information provided by the barcodes. An alternative approach (McCoy et al., 2014) is to assemble the reads for each particular barcode and to treat the result as a single "synthetic long read." While this strategy works well for technologies such as TruSeq SLR, in which source fragments are generally sequenced with high coverage, it is not practical when fragments are shallowly sequenced as with 10x, which achieves high coverage not by having high per-barcode coverage but rather by having many barcodes. Also worth noting is the fact that TruSeq SLR's sequencing fragments at high coverage inflate their sequencing costs to be on par with PacBio's and Oxford Nanopore's, whereas 10x circumvents this high cost via shallow fragment sequencing (Goodwin et al., 2016).

Currently, the state-of-the-art in terms of barcoded read alignment employs "read clouds"—groups of reads that share the same barcode and map to the same genomic region—to choose the most likely alignment from a set of candidate alignments for each read. Intuitively, read clouds represent the possible source fragments from which the barcoded reads are derived. The read cloud approach to alignment effectively begins with a standard all-mapping to a reference genome to identify these clouds, followed by an iterative update of the reads' assignments to one of their possible alignments, guided by a Markov random field that is used to evaluate the probability of a given read-to-cloud configuration (taking into account the alignment scores, clouds, etc.). This method was first implemented for Moleculo data as Random Field Aligner (RFA) and was subsequently the foundation for the 10x aligner Lariat, which we compare to extensively in this work (Bishara et al., 2015). Notably, in this framework, clouds are inherently fixed entities to which some number of reads are assigned at any given point, which does not take

into account the fact that reads can have suitable alignments in several different clouds. Since this information can be valuable in downstream analyses such as genotyping, phasing, and structural variation detection, we wish to account for it.

Confounding barcoded read alignment is the fact that multiple fragments can share the same barcode; it is in general not possible to infer the source fragment of a read (and thus its correct alignment within a reference genome) merely by looking at its barcode. In order to deduce the correct placement of a read, and thus its unknown source fragment, all possible alignments of that read need to be examined. Even then, it can be difficult to determine the correct alignment, particularly in homologous regions of the genome that result in multi-mappings within a single cloud.

Here, we propose a general paradigm for barcoded read alignment that newly employs a probabilistic interpretation of clouds: EMerAld, or EMA for short (Figure 1). Our two-tiered statistical binning approach enables the more accurate placement of reads in and within read clouds, which is the critical step in barcoded read alignment. The two tiers consist of (1) a novel latent variable model to probabilistically assign reads to clouds, which introduces the notion of clouds as distributions over generated reads rather than simply fixed groups of reads; and (2) newly exploiting expected read coverage (read density) to resolve the difficult case of multiple alignments of reads *within* clouds. The idea and subsequent observation of the fixed read density distribution within source fragments are novel to EMA and can be utilized by many barcoded read analysis tools: for example, an assembler might use our idea to model the distance between reads within the same source fragment and thus break ties, if any. Note that these ambiguous alignments account for a large fraction of the rare variants that currently cannot be resolved and are of great interest to biologists (Ingelman-Sundberg, 2004; Sekar et al., 2016; Falchi et al., 2014).

By thinking of clouds not as arbitrary clusters of reads but rather as distributions, EMA's latent variable model (tier I) is able to generate more accurate alignments and to newly assign interpretable probabilities to its alignments, which greatly improves downstream analyses. Genotyping improvements and, more generally, the resolution of distant homologs (e.g., longer than fragment length or interchromosomol) stem from our probabilistic interpretation. We demonstrate EMA's performance by evaluating downstream genotyping and phasing accuracy first using real 10x data. We discovered that roughly 20% of all reads in our datasets had multiple suitable alignments and were therefore able to be targeted by EMA's optimization algorithm. We also found that genotypes called from EMA's alignments contained over 30% fewer false positives than those called from 10x's Lariat aligner—and at the same time contained fewer false negatives—on independent 10x datasets of NA12878, NA24149, NA24143, and NA24385. The National Institute of Science and Technology's "Genome in a Bottle" high-confidence variant calls were used as a gold standard. EMA also improved phasing performance by reducing switch errors and producing larger phased blocks.

Focusing on uncovering novel biology, we additionally demonstrate that—through its read density optimization (tier II)—EMA improves alignments in several clinically important and highly homologous genes: *CYP2D6/CYP2D7* (of great pharmacogenomic importance [Ingelman-Sundberg, 2004]), *C4* (linked to schizophrenia [Sekar et al., 2016]), and *AMY1A* (conjectured association with obesity [Falchi et al., 2014]). We discovered using EMA and validated through published studies (Jain et al., 2017; Mostovoy et al., 2016; Pendleton et al., 2015) several variants in these regions that go undetected by Lariat and BWA. Moreover, we sought to demonstrate that our approach generalizes to other barcoded sequencing technologies by applying it to TruSeq SLR data as well as CPT-seq data, where we observe similar results. Intuitively, EMA's ability to handle shorter homologies (i.e., those within a cloud) leads to these novel findings.

In addition to achieving superior accuracy, providing interpretable probabilities, and uncovering novel biology, the EMA pipeline is up to 2× faster than Lariat—which translates into days faster for typical 10x datasets—and does not add any memory overhead to the alignment process. Thus, we expect the algorithms introduced here to be a fundamental component of barcoded read methods in the future.

RESULTS

Experimental Setting

We first compared the performance of EMA against Lariat (10x Genomics, 2017) (10x's own aligner and a component of the Long Ranger software suite) and BWA-MEM ("Maximal Exact Matches") (Li and Durbin, 2009) (which does not take advantage of barcoded data and was therefore used as a baseline for what can be achieved with standard short reads). In order to benchmark the quality of the aligners, we examined downstream genotyping accuracy, alignments in highly homologous regions, and downstream phasing accuracy.

We ran each tool on four 10x *Homo sapiens* datasets for NA12878, NA24149, NA24143, and NA24385 and used the corresponding latest NIST GIAB (Zook et al., 2014, 2016) high-confidence variant calls as a gold standard for each. For both EMA and BWA, we performed duplicate marking after alignment using Picard's MarkDuplicates tool (URL: https://broadinstitute.github.io/picard/), with barcode-aware mode enabled in the case of EMA; Long Ranger performs duplicate marking automatically. Genotypes were called by HaplotypeCaller from the Genome Analysis Toolkit (GATK) (McKenna et al., 2010; DePristo et al., 2011) with default settings, while phasing was done by HapCUT2 (Edge et al., 2016) in barcode-aware mode. Genotyping accuracies were computed using RTG ("Real Time Genomics") Tools (Cleary et al., 2014). We also ran EMA and Lariat on a much higher coverage NA12878 dataset ("NA12878 v2") to test genotyping accuracy at high coverage as well as scalability.

To test EMA's improvements on other barcoded read sequencing technologies, we ran EMA and BWA on an NA12878 TruSeq SLR dataset (Bishara et al., 2015) as well as an NA12878 CPT-seq dataset (Amini et al., 2014). All analyses in this paper were performed with respect to the GRCh37 human reference genome.

EMA Improves Downstream Genotyping Accuracy

EMA's genotyping accuracy surpasses that of other aligners (Figure 2). We found that for each of the four 10x *H. sapiens* datasets, EMA produced 30% fewer false positive variant calls compared to Lariat and produced fewer false negative calls as

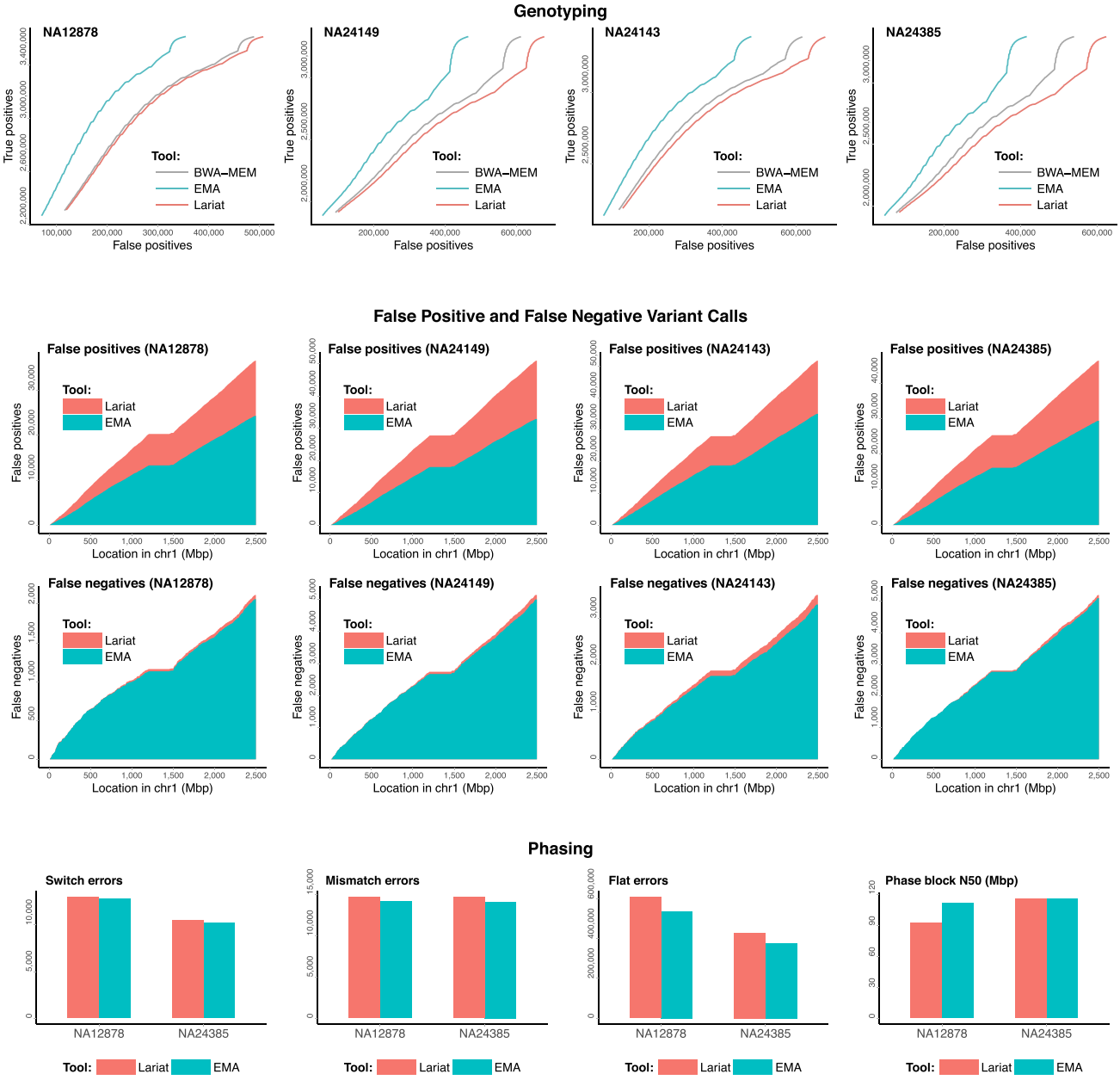


Figure 2. Genotyping and Phasing Results for Each Aligner

The top row shows true positive variant calls as a function of false positives for alignments produced by EMA (turquoise), Lariat (orange), and BWA-MEM (gray) on the well-studied samples NA12878, NA24149, NA24143, and NA24385. Genotype confidences are determined by the genotype quality (GQ) annotations generated by GATK's HaplotypeCaller. The middle two rows contain cumulative histograms of false positives (top) and false negatives (bottom) throughout chromosome 1 for each dataset, for both EMA (turquoise) and Lariat (orange). EMA achieves more than a 30% average improvement over the other methods in terms of eliminating erroneous variant calls. The bottom row shows EMA and Lariat's phasing results for several metrics: switch errors, mismatch errors, flat errors (Edge et al., 2016), and phase block N50 (lower is better for the first three, while higher is better for the last). EMA outperforms Lariat in phasing on every metric.

well. Interestingly, BWA-MEM (which does not take barcodes into account) performed marginally better than Lariat here. Nevertheless, EMA also outperforms BWA-MEM, attaining the fewest false positive and false negative variant calls between the three aligners on each dataset. To verify that EMA's superior accuracy scales to higher coverage datasets, we tested it on a high-coverage NA12878 dataset (Supplementary Figure S3). EMA attains an even more substantial improvement on the high-coverage dataset, eliminating nearly 37% of Lariat's false positives and 6% of its false negatives.

When run on TruSeq SLR and CPT-seq data, we did not observe any significant differences in genotyping accuracy

between EMA and BWA. This finding is likely due to the fact that these platforms divide the source fragments into just 384 and 9,128 wells ("barcodes"), respectively, limiting the utility of the barcodes in unambiguous regions of the genome, which is primarily what our NIST GIAB gold standard consists of. However, for both technologies, we did observe improvements in resolving ambiguous regions of the genome, which we detail below.

Overall, we found that typically ~20% of all reads in our various datasets had multiple suitable alignments and were therefore able to be targeted by EMA's two-tiered statistical binning optimization algorithm. These are precisely those reads that are most challenging to align and can occur in clinically important regions of the genome, as we next demonstrate.

EMA Improves Alignments and Analysis of Highly Homologous Regions

Among the principal promises of barcoded read sequencing is better structural variation detection, which invariably requires resolving alignments in homologous regions. One such important region is the *CYP2D* region in chromosome 22, which hosts *CYP2D6*—a gene of great pharmacogenomic importance (Ingelman-Sundberg, 2004)—and the two related and highly homologous regions *CYP2D7* and *CYP2D8*. The high homology between *CYP2D6* and *CYP2D7* makes copy number estimation and variant calling in this region particularly challenging. Indeed, the majority of aligners misalign reads in this region. The difficulty is especially evident in NA12878, which, in addition to the two copies of both *CYP2D6* and *CYP2D7*, contains an additional copy that is a fusion between these two genes (Twist et al., 2016), as well as *CYP2D7* mutations that introduce even higher homology with the corresponding *CYP2D6* region. Especially problematic is exon/intron 8 of *CYP2D6*, where many reads originating from *CYP2D7* end up mapping erroneously (see Figure 3 for a visualization). Even the naive use of barcoded reads is not sufficient: both homologous regions in *CYP2D* are typically covered by a single cloud. For example, Lariat performs no better than BWA in this region (Figure 3). For these reasons, we chose to evaluate EMA in *CYP2D* to benchmark its accuracy in such highly homologous regions.

As can be seen in Figure 3, EMA's statistical binning strategy significantly smooths out the two problematic peaks in *CYP2D6* and *CYP2D7*. This technique enabled us to detect three novel mutations in *CYP2D7* (Figure 3), which exhibit high homology with the corresponding region in *CYP2D6*. Thus, all reads originating from these loci get misaligned to *CYP2D6*, especially if one only considers edit distance during the alignment (as Lariat and BWA do). Such misalignments are evident in the "peaks" and "holes" shown in Figure 3. We additionally cross-validated this region with the consensus sequence obtained from available NA12878 assemblies (Jain et al., 2017; Mostovoy et al., 2016; Pendleton et al., 2015) and confirmed the presence of novel mutations. Notably, we found similar enhancements in other clinically important and highly homologous genes: *C4* and *AMY1A*, as depicted in the same figure.

In addition, the copy number derived from EMA's alignments in this problematic region (spanning from exon 7 up to exon 9 in *CYP2D6* and *CYP2D7*) was closer to the "expected" copy number by 20% compared to the copy number derived from Lariat's alignments (we used Aldy [Numanagić et al., 2018] to obtain this data). We further ran Aldy on our high-coverage NA12878 v2 dataset, where it correctly detected the *3/*68+*4 allelic combination on both EMA's and Lariat's alignments, and EMA's overall copy number error over the whole region was around 4% better than Lariat's. Finally, statistical binning did not adversely impact phasing performance in this region as we were able to correctly phase *CYP2D6*4A* alleles in our NA12878 sample from EMA's alignments.

To demonstrate the generalizability of our paradigm to other similar barcoded sequencing technologies, we tested it on TruSeq SLR (BioProject: PRJNA287848) and CPT-seq (BioProject: PRJNA241346) data, where the bin distributions follow a similar pattern as 10x's. We alone were able to detect the same novel *CYP2D7*, *C4*, and *AMY1A* variants in an NA12878 TruSeq SLR dataset (even with shallow coverage) and to detect the *CYP2D7* variants in an NA12878 CPT-seq dataset, as shown in the right-hand side of Figure 3.

EMA Improves Downstream Phasing

We applied the state-of-the-art phasing algorithm HapCUT2 (Edge et al., 2016), which supports 10x barcoded reads, to phase (i.e., link variants into haplotypes) the variants called by GATK for both EMA's and Lariat's alignments. We evaluated our results with the phasing metrics defined in the HapCUT2 manuscript. As shown in Figure 2, EMA provides more accurate phasing with respect to every metric in comparison to Lariat.

EMA Is Computationally More Efficient

Runtimes and memory usage for each aligner are provided in Table 1 for our small and large NA12878 datasets. These times include alignment, duplicate marking, and any other data post processing (e.g., BAM sorting and merging). The reported memory usages are per each instance of the given mapper. We found that EMA scales better than Lariat: specifically, we observe a 1.5x speedup on our smaller dataset and a nearly 2x speedup on our larger one, over Lariat's runtimes. We ran EMA on a total of four high-coverage datasets and have observed that EMA scales linearly in the size of the dataset.

Runtime and memory usages on two NA12878 datasets ("NA12878"—used also in Figure 2—is about 287 GB of raw data; "NA12878 v2" is about 823 GB). Numbers in parentheses indicate the performance of the aligner alone (i.e., without sorting, merging, or duplicate marking). For the small dataset, each mapper was allocated 40 Intel Xeon E5-2650 CPUs @ 2.30 GHz. For the large dataset, each was allocated 48 Intel Xeon E5-2695 CPUs @ 2.40 GHz. Memory measurements include only the actual aligner's memory usage and do not include the memory requirements of pre- and post-processing steps as they are virtually the same for all methods. BWA-MEM was used only as a baseline on the smaller dataset.

DISCUSSION

EMA's unique ability to assign interpretable probabilities to alignments has several benefits, the most immediate of which is that it enables us to set a meaningful confidence threshold on alignments. Additionally, these alignment probabilities can be incorporated into downstream applications such as genotyping, phasing, and structural variation detection. We demonstrate

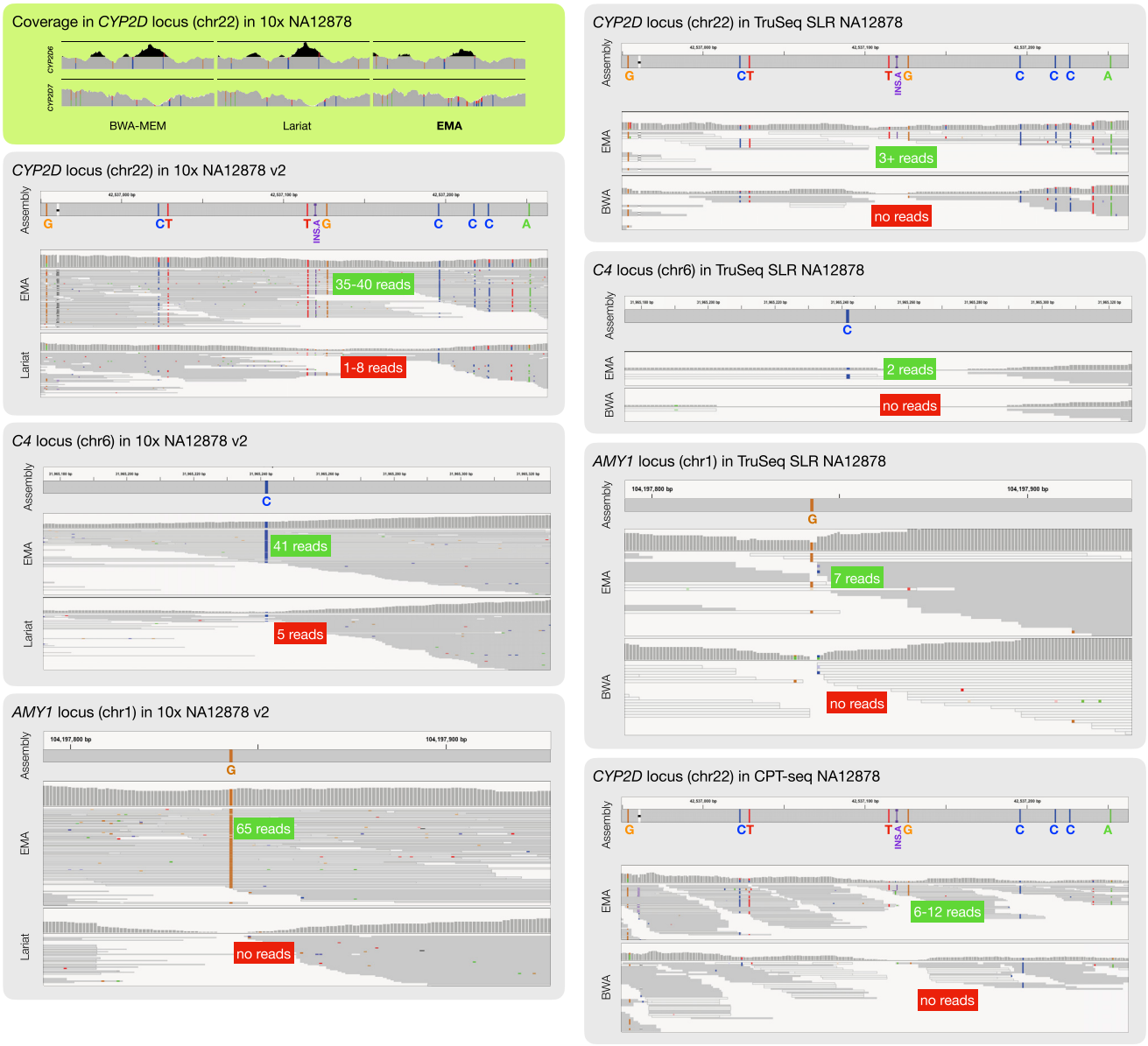


Figure 3. Positive Effect of EMA's Statistical Binning in the Clinically Important Genes *CYP2D6*, *CYP2D7*, *C4*, and *AMY1A*

The green inset shows the read coverage for the region around exon/intron 8 of *CYP2D6* (top row) and *CYP2D7* (bottom row). Spurious coverage peaks (i.e., increases in observed coverage likely to be false) in *CYP2D6* are shaded black. EMA is clearly able to remove the problematic peaks and correctly assign them to *CYP2D7*. The insets below show the newly assigned mappings to *CYP2D7*: EMA's alignments agree with the assembly consensus sequence (observe the insertion and two neighboring SNPs detected by EMA). By contrast, both Lariat and BWA-MEM aligned virtually no reads to this region and were thus unable to call these mutations. Analogous images are shown for *C4* and *AMY1A*, as well as for TruSeq SLR and CPT-seq data.

this feature here by computing mapping qualities based on these probabilities, which consequently enhance genotyping and phasing. Nevertheless, specialized algorithms centered around these probabilities are also conceivable.

Moreover, EMA is able to effectively discern between multiple alignments of a read in a single cloud through its read density optimization algorithm. This capability addresses one of the weaknesses of barcoded read sequencing as compared to long-read sequencing; namely, that only a relatively small subset of the original source fragment is observed—and, more specifically, that the order of reads within the fragment is not known—making it difficult to produce accurate alignments if the fragment spans homologous elements. By exploiting the insight that read densities within a fragment follow a particular distribution, EMA more effectively aligns the reads produced by such fragments, which can overlap regions of phenotypic or pharmacogenomic importance, such as *CYP2D*, *C4*, or *AMY1*, as we demonstrated. In summary, EMA's first tier (latent variable model) helps resolve the case of distant homologs, and its second tier, the case of proximal homologs.

Table 1. Runtime and Memory Usages on Two NA12878 Datasets

	NA12878		*NA12878 v2*	
Tool	Time (hh:mm)	Mem./core (GB)	Time (hh:mm)	Mem./core (GB)
EMA	14:58 (10:40)	5.4	28:30 (17:45)	8.7
Lariat	21:49 (12:45)	7.0	54:53 (26:01)	8.2
BWA-MEM	14:49 (9:52)	5.5		

"NA12878"—used also in Figure 2—is about 287 GB of raw data; "NA12878 v2" is about 823 GB. Numbers in parenthesis indicate the performance of the aligner alone (i.e., without sorting, merging, or duplicate marking). For the small dataset, each mapper was allocated 40 Intel Xeon E5-2650 CPUs @ 2.30GHz. For the large dataset, each was allocated 48 Intel Xeon E5-2695 CPUs @ 2.40GHz. Memory measurements include only the actual aligner's memory usage and do not include the memory requirements of pre- and post-processing steps, as they are virtually the same for all methods. BWA-MEM was used only as a baseline on the smaller dataset.

As we usher in the next wave of the NGS technologies, barcoded read sequencing will undoubtedly play a central role, and fast and accurate methods for aligning barcoded reads, such as EMA, will ultimately prove invaluable in downstream analyses.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Standard Data Preprocessing
  - Latent Variable Model for Aligning Barcoded Reads to Clouds
  - Algorithm 1
  - Read Density Optimization to Handle Multi-Mappings in a Single Cloud
  - Algorithm 2
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and five data files and can be found with this article online at https://doi.org/10.1016/j.cels.2018.07.005.

ACKNOWLEDGMENTS

We thank Chad Nusbaum, Eric Banks, as well as the rest of the GATK SV Group from the Broad Institute for providing us with data samples and many valuable suggestions. Also, we thank Jian Peng for helpful discussions, as well as Lillian Zhang for her help in evaluating EMA's performance. Finally, we thank the Vancouver Prostate Centre for providing infrastructure to evaluate EMA. A.S., I.N., and B.B. are partially funded by the NIH grant GM108348. This content is solely the responsibility of the authors and does not reflect the official views of the NIH.Editor's note: An early version of this paper was submitted to and peer reviewed at the 2018 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at Cell Systems.

AUTHOR CONTRIBUTIONS

A.S. and I.N. developed the core methods and performed the experiments. B.B. oversaw and guided the method development process and experimentation. A.S., I.N., and B.B. collectively wrote the manuscript. C.W. provided expertise on 10x data and rare variants, as well as several datasets for experimentation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 3, 2018
Revised: May 3, 2018
Accepted: July 10, 2018
Published: August 22, 2018

REFERENCES

10x Genomics (2017). What is long ranger?, https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger/.

10x Genomics (2018). Sequencing. https://www.10xgenomics.com/solutions/vdj/, 2018. V(d)j.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J.O., Vijayan, K., et al. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat.Genet. *46*, 1343–1349.

Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D.E., West, R., Sidow, A., and Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. Genome Res. *25*, 1570–1580.

Cleary, B., Cong, L., Cheung, A., Lander, E.S., and Regev, A. (2017). Efficient generation of transcriptomic profiles by random composite measurements. Cell *171*, 1424–1436.

Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Nohzadeh-Malakshah, S., Rathod, M., et al. (2014). Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. J. Comput. Biol. *21*, 405–419.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Edge, P., Bafna, V., and Bansal, V. (2016). Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. *27*, 801–812.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, John, Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Falchi, M., El-Sayed Moustafa, J.S., Takousis, P., Pesce, F., Bonnefond, A., Andersson-Assarsson, J.C., Sudmant, P.H., Dorajoo, R., Al-Shafai, M.N., Bottolo, L., et al. (2014). Low copy number of the salivary amylase gene predisposes to obesity. Nat. Genet. *46*, 492–497.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. Genome Biol. *17*, 77.

Ingelman-Sundberg, M. (2004). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): Clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J. *5*, 6–13.

Jain, M., Koren, S., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., et al. (2017). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The sequence alignment/map format and samtools,. Bioinformatics *25*, 2078–2079.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Mardis, E.R. (2017). DNA sequencing technologies: 2006-2016. Nat. Protoc. *12*, 213–218.

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., and Fiston-Lavier, A.S. (2014). Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PloS ONE *9*, e106689.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., Dakula, Z., et al. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. Nat. Methods *13*, 587–590.

Numanagić, I., Malikić, S., Ford, M., Qin, X., Toji, L., Radovich, M., Skaar, T.C., Pratt, V.M., Berger, B., Scherer, S., et al. (2018). Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. Nat. Commun. *9*, 828.

Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A., et al. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat. Methods *12*, 780–786.

Schwarze, K., Buchanan, J., Taylor, J.C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet. Med.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177.

Twist, G.P., Gaedigk, A., Miller, N.A., Farrow, E.G., Willig, L.K., Dinwiddie, D.L., Petrikin, J.E., Soden, S.E., Herd, S., Gibson, M., et al. (2016). Constellation: A tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. NPJ Genom. Med. *1*, 15007.

Wang, Y., Yang, Q., and Wang, Z. (2014). The evolution of nanopore sequencing. Front. Genet *5*, 449.

Yorukoglu, D., Yu, Y.W., Peng, J., and Berger, B. (2016). Compressive mapping for next-generation sequencing. Nat. Biotech. *34*, 374–376.

Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. Nat. Biotechnol. *34*, 303–311.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell rna sequencing methods. Mol. Cell *65*, 631–643.

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data *3*, 160025.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotech. *32*, 246–251.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
NA12878 WGS (10x)	10x Genomics	https://support.10xgenomics.com/genome-exome/datasets/2.1.0/NA12878_WGS_210
NA12878 WGS v2 (10x)	10x Genomics	https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2
NA12878 WGS (Illumina TruSeq Synthetic Long-Read)	Bishara et al. (2015)	BioProject: PRJNA287848
NA12878 WGS (CPT-seq)	Amini et al. (2014)	BioProject: PRJNA241346
Software and Algorithms		
Long Ranger	10x Genomics	https://support.10xgenomics.com/genome-exome/software/downloads/latest
BWA	Li and Durbin (2009)	https://github.com/lh3/bwa
GATK	Broad Institute	https://software.broadinstitute.org/gatk/
HapCUT2	Edge et al. (2016)	https://github.com/vibansal/HapCUT2
Picard	Broad Institute	https://broadinstitute.github.io/picard/
Samtools	Li et al. (2009)	https://github.com/samtools/samtools
RTG Tools	Cleary et al. (2014)	https://github.com/RealTimeGenomics/rtg-tools
EMA	This paper	https://github.com/arshajii/ema
Other		
NIST GIAB	Zook et al. (2016)	http://jimb.stanford.edu/giab/
NA24149, NA24143 and NA24385 WGS (10x)	Broad Institute	http://ema.csail.mit.edu

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Bonnie Berger (bab@mit.edu).

METHOD DETAILS

General barcoded read sequencing begins with splitting the source DNA into long fragments (10–200kb) where each such fragment is assigned some barcode (e.g. a short 16bp DNA sequence in 10x sequencing). These fragments are sheared and each sheared piece has the assigned barcode ligated to it (or, alternatively, resides in an identifiable well), whereupon standard short-read sequencing is applied to the sheared pieces. As a result, barcoded reads have the same low error rates as typical Illumina whole-genome sequencing reads. An idealization of this process is illustrated in Figure 1A.

Standard Data Preprocessing

The first stage in the alignment process is to preprocess the data and to identify the barcodes. Currently, EMA uses an in-house 10x barcode preprocessor, which extracts and corrects the barcodes from the raw data. Data from many other barcoded read technologies (e.g. TruSeq SLR) can be preprocessed in a more straightforward manner, as the barcodes are given as well identifiers for each read, meaning the preprocessing stage consists of a simple demultiplexing step.

For 10x data preprocessing we largely follow the same practices used by 10x Genomics' WGS software suite, Long Ranger. The purpose of this preprocessing is to:

- extract the barcode from the read sequence,
- error-correct the barcode based on quality scores and a list of known barcode sequences,
- and group reads by barcode into "barcode buckets" to enable parallelism during alignment.

In summary, in the barcode extraction stage, we remove the 16bp barcode from the first mate of each read pair, and trim an additional 7bp to account for potential ligation artifacts resulting from the barcode ligation process during sequencing (the second mate shares the same barcode as the first mate). Subsequently, we compare each barcode to a list B of known barcodes to produce a count for each barcode in B, and compute a prior probability for each based on these counts (specifically, this prior is proportional to the fraction of times we see the given barcode in our data). Note that this list is provided by 10x Genomics, and is designed so that no two barcodes are Hamming-neighbors of one another. Now for each barcode b not appearing in B, we examine each of its Hamming-1 neighbors b' and, if b' appears in B, compute the probability that b' was the true barcode based on its prior and the quality score of the changed base. Similarly, for each b appearing in B, we consider each Hamming-2 neighbor b' and compute the probability that b' was the true barcode in an analogous way. The reason we examine the Hamming-2 neighbors of barcodes that are already in our whitelist is because it is possible, albeit unlikely, that two sequencing errors in the barcode changed it from a barcode on the list to another *also* on the list (in practice we found this Hamming-2 correction step to be largely unnecessary as described below, but it is performed in Lariat's data preprocessing nonetheless). Lastly, we employ a probability cutoff on the barcodes, and thereby omit the barcodes of reads that do not meet this cutoff. Any read not carrying a barcode after this stage is aligned with a standard WGS mapper such as CORA or BWA.

While in standard read alignment parallelism can be achieved at the read-level, for barcoded read alignment we can only achieve parallelism at the barcode-level. Therefore, the last preprocessing step is to group reads by barcode into some number of buckets. Each such bucket contains some range of barcodes from B, which are all grouped together within the bucket. This enables us to align the reads from each bucket in parallel, and to merge the outputs in a post-processing step.

We note that the Hamming-2 search takes a substantial fraction of the total time, but is often unnecessary: on a large 980GB 10x dataset, only 276 out of almost 1.5 billion reads are affected by the Hamming-2 correction (amounting to <0.0001% overall effect). Thus, it is safe to skip the Hamming-2 correction step. Nevertheless, we applied Hamming-2 correction on all our datasets for the sake of consistency with Lariat. Finally, EMA offers a parallelized barcode correction implementation, which significantly speeds up the overall pipeline.

Latent Variable Model for Aligning Barcoded Reads to Clouds

Here we employ a latent variable model for determining the optimal assignment of reads to their possible clouds. A "cloud" is defined to be a group of nearby alignments of reads with a common barcode, thereby representing a possible source fragment (Bishara et al., 2015). We consider all the reads for an individual barcode simultaneously, all-mapping and grouping them to produce a set of clouds for that barcode (Figure 1B). The clouds are deduced from the all-mappings by grouping any two alignments that are on the same chromosome and within 50kb of one another into the same cloud, which is the same approach employed by Lariat (for TruSeq SLR or CPT-seq data, we use 15kb as a cutoff; this is a tuneable parameter that can be adjusted depending on the underlying technology). While this heuristic works well in the majority of cases, it can evidently run into issues if, for example, a single read aligns multiple times to the same cloud. We address such cases below, but assume in the subsequent analysis that clouds consist of at most one alignment of a given read.

As notation, we will denote by c the set of alignments contained in a given cloud. We restrict our analysis to a single set of clouds $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ that corresponds to a connected component in the disjoint-set over clouds induced by alignments, as shown in Figure 1B (i.e. two clouds c_i and c_j will be connected if there is a read that has an alignment to both c_i and c_j). Conceptually, the clouds in $\mathcal{C}$ can be thought of as alternate possibilities for the *same* latent source fragment. By definition, for any given read aligning to some cloud in $\mathcal{C}$, we will have to consider only the clouds in $\mathcal{C}$ when determining the best alignment for that read, so we focus on each such set of clouds separately. Note that we make the same implicit assumption made by Lariat: namely that distinct fragments sharing a common barcode (i.e. fragments in the same droplet/well) do not overlap on the genome. In reality, there is nothing preventing this from happening, but we can see that it occurs rarely since fragments are effectively sampled uniformly from the entire genome. If we partition the 3Gb genome into 100kb bins (as a reasonable upper bound on mean fragment length) and assume a droplet/well contains about 10 fragments (also a reasonable bound), we can observe that only about $1 - \prod_{i=1}^{10}(1 - (i - 1/3\text{Gb}/100\text{kb})) \approx 0.15\%$ will contain overlapping fragments, where (as an approximation) we assume fragments overlap if they are contained in the same bin. By comparison, about 5%–6% of all 10x reads are usually left without a barcode after standard barcode correction, so the additional 0.15% is rather marginal. A second possible undesirable scenario would be if two fragments with the same barcode originated from distinct but homologous regions. In this case, we would find two connected clouds in the disjoint-set that we would wrongly consider to be alternate possibilities of a *single* fragment. Nevertheless, we expect our optimization algorithm to handle this situation gracefully, assigning some probability to each read of mapping to either homolog (arguably, this is more of a problem with the previously employed read cloud methods wherein reads are only ever assigned to a single cloud).

For $\mathcal{C} = \{c_1, \ldots, c_n\}$, let C_i denote the event that cloud c_i represents the true source fragment. Since the clouds $c_1, \ldots, c_n$ are different possibilities for the same source fragment, we have $\Pr(C_i \cap C_j) = 0$ $(i \neq j)$ and $\sum_{i=1}^{n} 1(C_i) = 1$ (where $1(\cdot) \in \{0, 1\}$ is an indicator for the specified event). We assume uniform priors on the clouds so that $\Pr(C_i) = (1/n)$ (while it is possible to devise a prior that takes into account features such as cloud length, we observed a large variance between clouds in our datasets that renders this unhelpful). Now, a cloud c_i can be conceptualized as an entity that generates some number of reads K_i, parameterized by some weight θ_{c_i}, so that we can say $K_i \sim \text{Cloud}(\theta_{c_i})$ for some unknown "cloud" distribution over generated reads. We make the key assumption that, in expectation, $\Pr(C_i|\theta_{c_i}) \propto K_i \propto \theta_{c_i}$ for all $c_i \in \mathcal{C}$. In other words, if a cloud is expected to have generated a large number of reads, then the probability that the cloud represents a true source fragment is high. Let $\boldsymbol{\theta} = (\theta_{c_1}, \ldots, \theta_{c_n})$ be the vector of cloud

weights. We assume the cloud weights are normalized so that $\Pr(C_i/\theta_{c_i}) = \theta_{c_i}$, and that they are drawn from a uniform Dirichlet distribution so that $\boldsymbol{\theta} \sim \mathrm{Dir}(1)$. Consider now the probability γ_{r,c_i} that a read r truly originates from cloud c_i (denoted as an event by Γ_{r,c_i}) given the cloud parameters $\boldsymbol{\theta}$ (i.e. $\Gamma_{r,c_i}/\boldsymbol{\theta} \sim \mathrm{Ber}(\gamma_{r,c_i})$, where $\mathrm{Ber}(p)$ is the Bernoulli distribution with parameter p). By Bayes' rule, we can say:

$$\gamma_{r,c_i} = \Pr(\Gamma_{r,c_i}/\boldsymbol{\theta}) = \frac{1}{Z_{\mathcal{C}}}\Pr(\boldsymbol{\theta}/\Gamma_{r,c_i})\Pr(\Gamma_{r,c_i}),$$

where $Z_{\mathcal{C}}$s (and variants thereof) are normalization constants that are the same for each $c \in \mathcal{C}$. Since Γ_{r,c_i} occurs if and only if C_i occurs, we have

$$\gamma_{r,c_i} = \frac{1}{Z_{\mathcal{C}}}\Pr(\boldsymbol{\theta}/C_i)\Pr(\Gamma_{r,c_i}).$$

Applying Bayes' rule again to $\Pr(\boldsymbol{\theta}/C_i)$ and using the fact that both $\Pr(\boldsymbol{\theta})$ and $\Pr(C_i)$ are uniform, we obtain

$$\gamma_{r,c_i} = \frac{1}{Z_{\mathcal{C}}}\frac{\Pr(\boldsymbol{\theta})\Pr(C_i/\boldsymbol{\theta})}{\Pr(C_i)}\Pr(\Gamma_{r,c_i}) = \frac{1}{Z'_{\mathcal{C}}}\Pr(C_i/\boldsymbol{\theta})\Pr(\Gamma_{r,c_i}) = \frac{\theta_{c_i}}{Z'_{\mathcal{C}}}\Pr(\Gamma_{r,c_i}),$$

where $Z'_{\mathcal{C}} = [\Pr(C_i)/\Pr(\boldsymbol{\theta})]Z_{\mathcal{C}}$. Note that $\Pr(\Gamma_{r,c_i})$ is a prior on the probability that r truly originates from c_i, which is not dependent on the barcode but rather only on edit distance, mate alignment, and mapping quality as in standard short-read alignment. Henceforth, we refer to $\Pr(\Gamma_{r,c_i})$ as $\gamma^{(0)}_{r,c_i}$, so that $\Gamma_{r,c_i} \sim \mathrm{Ber}(\gamma^{(0)}_{r,c_i})$.

Now we can form a prior $\boldsymbol{\theta}^{(0)} = (\theta^{(0)}_{c_1}, \ldots, \theta^{(0)}_{c_n})$, which is intuitively the initial vector of cloud weights. If we are given a set of alignment probabilities and a "current" $\boldsymbol{\theta}$ estimate $\boldsymbol{\theta}^{(t)} = (\theta^{(t)}_{c_1}, \ldots, \theta^{(t)}_{c_n})$ (initially $t = 0$), we can iteratively compute a better estimate $\boldsymbol{\theta}^{(t+1)}$ using the fact that $\theta_{c_i} \propto K_i$ in expectation:

$$\begin{aligned}\theta^{(t+1)}_{c_i} &= \frac{1}{|\mathscr{R}|}\mathbb{E}(K_i) = \frac{1}{|\mathscr{R}|}\mathbb{E}\left(\sum_{r\in\mathscr{R}} 1\left(\Gamma_{r,c_i}\right)/\boldsymbol{\theta}^{(t)}\right)\\ &= \frac{1}{|\mathscr{R}|}\sum_{r\in\mathscr{R}}\Pr\left(\Gamma_{r,c_i}\middle|\boldsymbol{\theta}^{(t)}\right)\\ &= \frac{1}{|\mathscr{R}|}\sum_{r\in\mathscr{R}}\gamma^{(t)}_{r,c_i},\end{aligned}$$

where $\mathscr{R}$ is the set of reads mapping to any cloud in $\mathcal{C}$, and the $1/|\mathscr{R}|$ factor ensures that $\sum_{c\in\mathcal{C}}\theta_c = 1$. This latent variable model formulation naturally leads to an expectation-maximization algorithm—one of the widely used ways of maximizing likelihood in such models—for determining the cloud weights and, thereby, the final alignment probabilities $\gamma^{\star}_{r,c_i}$. An implementation of this algorithm is given in Algorithm 1 (in practice we use $T = 5$ EM iterations).

Algorithm 1

Barcoded read alignment via expectation − maximization

Require: $\mathscr{R}$, $\mathcal{C}$

Ensure : $\gamma^{\star}_{r,c}$ for each $r \in \mathscr{R}$, $c \in \mathcal{C}$

$\gamma^{(0)}_{r,c} \leftarrow \Pr(r \in c)$, $\forall\ r \in \mathscr{R}$, $c \in \mathcal{C}$

$\theta^{(0)}_{c} \leftarrow \frac{1}{|\mathcal{C}|}$, $\forall\ c \in \mathcal{C}$

for $t \in \{0, 1, \ldots, T-1\}$ **do**

E step : $\gamma^{(t+1)}_{r,c} \leftarrow \Pr\left(r \in c_i | \theta^{(t)}_{c}\right)$ $\forall\ r \in \mathscr{R}$, $c \in \mathcal{C}$

M step : $\theta^{(t+1)}_{c} \leftarrow \frac{1}{|\mathscr{R}|}\sum_{r\in\mathscr{R}}\gamma^{(t)}_{r,c}$ $\forall\ r \in \mathscr{R}$

end for

$\gamma^{\star}_{r,c} \leftarrow \gamma^{(T)}_{r,c}$ $\forall\ r \in \mathrm{R}$, $c \in \mathrm{C}$

Each of the described variables and their interactions with one another is summarized in Supplementary Figure S1. Once we determine the final alignment probabilities through this method (as in Figure 1D), we use them to compute mapping qualities ("MAPQs"), which are a standard per-alignment metric reported by all aligners and are frequently used by downstream analysis pipelines. Specifically, we take the MAPQ to be the minimum of the alignment probability, the barcode-oblivious alignment score

and the MAPQ reported by BWA-MEM's API (which is used in EMA's current implementation to find candidate alignments). Importantly, we also report the actual alignment probabilities determined by EMA via a special standard-compliant SAM tag, so that they are available to downstream applications.

Read Density Optimization to Handle Multi-Mappings in a Single Cloud

While the 50kb-heuristic described above is typically effective at determining the clouds, it does not take into account the fact that a single read may align multiple times to the same cloud (which can occur if a cloud spans two or more homologous regions). In such cases, rather than simply picking the alignment with lowest edit distance within the cloud, as is the current practice, we propose a novel alternative approach that takes into account not only edit distance but also read *density*. We take advantage of the insight that there is typically only a single read pair per 1kb bin in each cloud; the exact distribution of read counts per 1kb bin is shown in Supplementary Figure S2. Now consider the case where one of our source fragments spans two highly similar (homologous) regions, and thereby produces a cloud with multi-mappings, as depicted in Figure 1C. If we pick alignments solely by edit distance, we may observe an improbable increase in read density (as shown in the figure). Consequently, we select alignments for the reads so as to minimize a combination of edit distance *and* abnormal density deviations.

Specifically, consider any cloud with multi-mappings consisting of a set of reads $R = \{r_1,\ldots,r_n\}$, and denote by A_r the set of alignments for read $r \in R$ in the cloud. Additionally, let $a_r \in A_r$ denote the currently "selected" alignment for r. We will initially partition the cloud, spanning the region from its leftmost to its rightmost alignment, into the set of bins $B = \{b_1, \ldots, b_n\}$ of equal width w, where each bin b_i covers the alignments whose starting positions are located in the interval $[i \cdot w, (i+1) \cdot w)$, as shown in Figure 1C. In practice, we set w to 1kb. Denote by C_{b_i} the random variable representing the number of reads in bin b_i, where C_{b_i} is drawn from the bin density distribution CloudBin(i). Lastly, let γ_{a_r} denote the prior probability that alignment a_r is the true alignment of read r based on edit distance and mate alignments alone. Our goal is to maximize the objective:

$$\left[\prod_{r\in R}\gamma_{a_r}\right] \cdot \left[\prod_{b_i \in B} \Pr\left(C_{b_i} = \sum_{r\in R} 1(a_r \in b_i)\right)^{\alpha}\right],$$

where α is a parameter that dictates the relative importance of the density probabilities compared to the alignment probabilities (we use $\alpha = 0.05$ in practice). We determine the distribution CloudBin(i) of each C_{b_i} beforehand by examining uniquely-mapping clouds that we are confident represent the true source fragment. Taking the logarithm, this objective becomes:

$$J(a_{r_1}, \ldots, a_{r_n}) = \sum_{r\in R} \log \gamma_{a_r} + \alpha \sum_{b_i \in B} \log \Pr\left(C_{b_i} = \sum_{r\in R} 1(a_r \in b_i)\right).$$

We optimize J through simulated annealing by repeatedly proposing random changes to a_r and accepting them probabilistically based on the change in our objective (the corresponding algorithm is described in Algorithm 2, in which K is the number of simulated annealing iterations, and $\tau(\cdot)$ defines the annealing schedule, which can be taken to be an exponentially decreasing function).

Algorithm 2

Read density optimization via simulated annealing

Require : $R; A_r \ \forall r \in R$

Ensure : $a_r^{\star} \ \ \forall \ r \in R$

$a_r \leftarrow \text{random}(A_r) \ \forall \ r \in R$

$z \leftarrow J(a_{r_1}, \ldots, a_{r_n})$

for $k \in \{1, \ldots, K\}$ **do**

$r' \leftarrow \text{random}(\{r \in R : |A_r| > 1\})$

$a_r' \leftarrow \text{random}(A_r \setminus \{a_r\})$

$z' \leftarrow J(a_{r_1}, \ldots, a_{r'}, \ldots, a_{r_n})$

if $z' > z$ **or** $\exp\left(-\frac{z - z'}{\tau(k)}\right) > \text{random}([0,1))$ **then**

$a_r \leftarrow a_r'$

$z \leftarrow z'$

end if end for

$a_r^{\star} \leftarrow a_r \ \forall \ r \in \text{R}$

We apply the preceding latent variable optimization algorithm to deduce optimal alignments *between* clouds and, if necessary, use this statistical binning algorithm to find the best alignments *within* a given cloud.

DATA AND SOFTWARE AVAILABILITY

EMA's full source, links to all datasets used and detailed guidelines for reproducing our results are available online at http://ema.csail.mit.edu and https://github.com/arshajii/ema.