Read clouds uncover variation in complex regions of the human genome

Alex Bishara,[1,6] Yuling Liu,[1,2,6] Ziming Weng,[3] Dorna Kashef-Haghighi,[1] Daniel E. Newburger,[4] Robert West,[3] Arend Sidow,[3,5] and Serafim Batzoglou[1]

[1]*Department of Computer Science, Stanford University, Stanford, California 94305, USA;* [2]*Department of Chemistry, Stanford University, Stanford, California 94305, USA;* [3]*Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA;* [4]*Biomedical Informatics Training Program, Stanford, California 94305, USA;* [5]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA*

Although an increasing amount of human genetic variation is being identified and recorded, determining variants within repeated sequences of the human genome remains a challenge. Most population and genome-wide association studies have therefore been unable to consider variation in these regions. Core to the problem is the lack of a sequencing technology that produces reads with sufficient length and accuracy to enable unique mapping. Here, we present a novel methodology of using read clouds, obtained by accurate short-read sequencing of DNA derived from long fragment libraries, to confidently align short reads within repeat regions and enable accurate variant discovery. Our novel algorithm, Random Field Aligner (RFA), captures the relationships among the short reads governed by the long read process via a Markov Random Field. We utilized a modified version of the Illumina TruSeq synthetic long-read protocol, which yielded shallow-sequenced read clouds. We test RFA through extensive simulations and apply it to discover variants on the NA12878 human sample, for which shallow TruSeq read cloud sequencing data are available, and on an invasive breast carcinoma genome that we sequenced using the same method. We demonstrate that RFA facilitates accurate recovery of variation in 155 Mb of the human genome, including 94% of 67 Mb of segmental duplication sequence and 96% of 11 Mb of transcribed sequence, that are currently hidden from short-read technologies.

[Supplemental material is available for this article.]

Although next-generation sequencing (NGS) technologies have enabled whole-genome sequencing (WGS) of many individuals to identify variation, current large-scale and cost-effective resequencing platforms produce reads of limited length (Shendure and Ji 2008; Metzker 2010); and as a result, variant identification within repeated sequences remains challenging. The 1000 Genomes Project Consortium has reported that nearly 6% of the GRCh37 human genome reference is inaccessible by short-read technologies (The 1000 Genomes Project Consortium 2012). Further studies have shown that as much as 10% of GRCh37 cannot be aligned to for the purpose of accurate variant discovery (Lee and Schatz 2012).

The portion of the human genome that is currently dark to short-read technologies is significant in both its size and phenotypic effect. Recent segmental duplications (also referred to as low copy repeats), consisting of regions >5 kbp in size and >94% sequence identity, have been identified as making up 130.5 Mb, or ~4.35% of the human genome (Bailey et al. 2002). These regions tend to be hotspots of structural and copy number variants (CNVs) (Coe et al. 2014; Chaisson et al. 2015) that in aggregate affect a larger fraction of the genome than that affected by single nucleotide polymorphisms (SNPs) (Conrad et al. 2010). CNVs have been associated with diseases such as autism (Sebat et al. 2007; Pinto et al. 2010), Crohn's disease (Wellcome Trust Case Control Consortium et al. 2010), schizophrenia (Stefansson et al. 2008; McCarthy et al. 2009), and neurocognitive disorders (Coe et al. 2014). However, current short-read technologies are unable to identify precise nucleotide variation in these regions.

In principle, longer sequencing reads provide an opportunity to disambiguate repeated sequences. Technologies such as Pacific Biosciences (PacBio) (McCarthy 2010) and Oxford Nanopore (Ashton et al. 2014) produce long reads, but at much higher per-base error rate. PacBio has been leveraged for improved bacterial reference genome assemblies (Koren et al. 2013) and for targeted de novo assembly of the complex 1.3 Mb of 17q21.31 (Huddleston et al. 2014). However, these technologies are currently substantially lower in throughput and higher in cost than short-read technologies and so cannot currently be used to cost-effectively uncover variation in repeated regions of the genome.

An alternative approach used in LFR (Peters et al. 2012), CPT-seq (Amini et al. 2014), and Illumina TruSeq Synthetic Long-Reads (previously known as Moleculo) (Kuleshov et al. 2014) utilizes accurate short-read sequencing of long DNA fragments in order to obtain long-range information at high nucleotide accuracy. The Illumina TruSeq protocol is able to produce 10-kbp long reads, retaining the benefits of the highly accurate and cost-effective Illumina technology (Kuleshov et al. 2014) and enabling human genome phasing (Kuleshov et al. 2014) and de novo assembly of complex genomes (Voskoboynik et al. 2013; McCoy et al. 2014).

Under Illumina's synthetic long-read protocol, DNA sequencing libraries are prepared as follows: First, the genomic DNA is

[6]**These authors contributed equally to this work.**
Corresponding author: serafim@cs.stanford.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.191189.115. Freely available online through the *Genome Research* Open Access option.

sheared into long (≥10 kbp) fragments and ligated with amplification adapters at both ends; second, these molecules are diluted into wells so that each well receives only a small fraction (1%–2%) of the genome; third, molecules are amplified, sheared into short fragments, and uniquely barcoded within each well (Kuleshov et al. 2014). The individual wells are then pooled and sequenced together. Demultiplexing the resulting reads by well barcode and aligning them to the reference genome yields clusters of short reads, which we call *read clouds*, each of which originated from a single long DNA molecule (Fig. 1A). Additionally, short reads that originate from the endpoints of a read cloud will overlap the original adapters ligated to the long molecules and serve as end-markers of the original long molecule.

A read cloud approach has two key parameters for genome coverage (Fig. 1): coverage of the genome with long DNA fragments, C_F, and coverage of each long fragment with short reads, C_R. The total sequencing depth is then $C = C_F \times C_R$. The choice of C_F and C_R for a given short-read sequencing budget C heavily influences the ability of the read cloud approach to accurately discover variation within a target genome. Both C_F and C have to be sufficiently high; in particular, C_F has to be high enough so that both haplotypes of a diploid genome are covered with a sufficient number of long fragments (Lander and Waterman 1988). The original protocol (McCoy et al. 2014) required for each well to be sequenced at a high depth ($C_R = 50\times$) in order to first de novo assemble synthetic long reads (SLR) of the original source long fragments (Fig. 1B; Voskoboynik et al. 2013). However, performing WGS with this approach requires an exorbitant amount of total sequencing in order to obtain a sufficiently high C_F. For example, if $C_R = 50\times$ (Voskoboynik et al. 2013) and $C_F = 20\times$, $C = 50 \times 20 = 1000\times$, or the equivalent of 33 whole human genomes sequenced at the currently standard 30× coverage.

The alternative strategy to true SLR approaches is to bypass the requirement for actual assembly of the original long fragments and to minimize short-read coverage ($C_R \leq 2\times$). This strategy allows a sufficiently high C_F in order to cover a genome at a reasonable coverage budget C. Choosing $C_R = 1.5$ and $C_F = 20\times$, $C = 1.5 \times 20 = 30\times$, would yield valuable long-range information for the same total sequencing cost as the currently standard short-read WGS approach.

In this work, we present RFA (Random Field Aligner), a novel methodology that utilizes the high C_F, low C_R read cloud approach to confidently map short reads within repetitive regions. In RFA, we directly model the short-read generative process from source long molecules in order to capture the dependencies of short reads through the hidden source long molecules. Using this probabilistic approach, we reduce the problem of finding optimal short-read alignments to optimizing a Markov Random Field (MRF). The resulting alignments tend to cluster the mapped reads into read clouds that fit the properties of the synthetic long-read sequencing protocol. The model naturally favors alignment of a read cloud to the specific copy of a repeated sequence that minimizes the sequence variation of the read cloud to the copy.

To our knowledge, RFA is the first attempt to take advantage of the long-range information present in shallow read cloud sequencing to improve the resulting short-read alignments and also to use read clouds to directly genotype an individual. Prior implementations of read clouds to provide molecular-phased genotypes for a single individual require known genotypes as input (Kitzman et al. 2011; Amini et al. 2014; Kuleshov et al. 2014) and typically align the resulting read clouds with standard short-read aligners in order to observe the allele at a known SNV within each read cloud. As genotypes are typically determined with a standard whole-genome 30× shotgun sequencing, in which a short-read workflow would be used, variants in complex regions would remain unresolved.

We demonstrate the utility of our approach using shallow-sequenced read clouds ($C_R = 1.5\times$) obtained from the Illumina TruSeq synthetic long-read protocol (henceforth referred to as TruSeq read clouds to avoid confusion with the Illumina product that uses deep sequencing to assemble synthetic long reads). We tested our approach on simulated read cloud wells, on TruSeq read cloud libraries for the cell line GM12878 for which assembled synthetic long reads are also available for direct validation (Genomes Moleculo NA12878, 2014, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/), and on a high coverage cancer sample that we sequenced. Evaluation of the results confirmed that our method accurately recovers precise nucleotide variation within a significant fraction of the human genome that was previously dark to current short-read technologies. We are able to leverage the read cloud strategy to recover this variation at a fraction of the cost of the original protocol and eliminate the need for first assembling synthetic long reads.

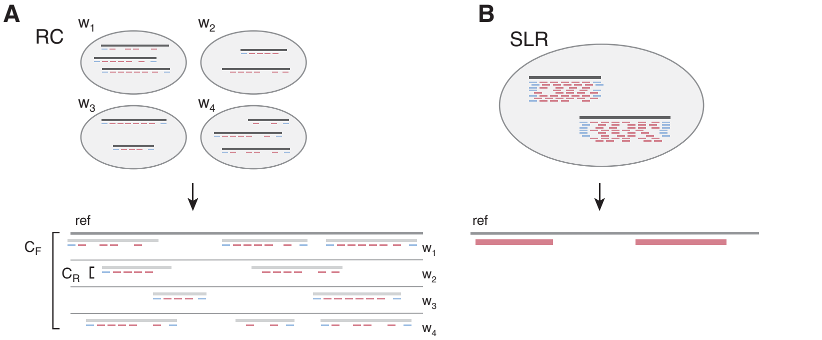

Figure 1. Read clouds (RC) and synthetic long reads (SLR) obtained by Illumina TruSeq Synthetic Long-Read sequencing. Each well initially contains long molecules that represent a small fraction of the target genome; reads from each long molecule are separated in genomic coordinates within the target genome, and therefore, clusters of such reads (read clouds) are formed with each cluster originating from one source fragment. Blue reads denote end-markers of the source fragments and may not always be present as sequenced short reads. (*A*) In the RC approach, long fragments from several wells w_n are sequenced to a shallow depth and aligned to the reference to obtain read clouds. Pooling of reads across several read clouds allows inference of the variation in the underlying long fragments. (*B*) In the SLR approach, long fragments are sequenced to a much higher depth to enable de novo assembly of synthetic long reads. For the same total sequencing budget *C*, the RC approach covers proportionally more target genome space than the SLR approach.

Results

Overview of algorithms

In order to accurately align short reads resulting from a synthetic long-read protocol, we developed a probabilistic framework to model the process by which read clouds are generated. In this framework, each well contains a set of hidden source long fragments, $\boldsymbol{M}$, that generate the short-read fragments, $\boldsymbol{R}$. Read alignment is then the problem of jointly aligning all the reads of a given well to the target genome to maximize the probability of the observed reads, $P(\boldsymbol{R})$. The probability distribution over the reads in each well can be written as

$$P(\boldsymbol{R}) = \sum_{\boldsymbol{M}} P(\boldsymbol{R}|\boldsymbol{M})P(\boldsymbol{M}),$$

where $P(\boldsymbol{R}|\boldsymbol{M})$ is the short-read generative process from long fragments; $P(\boldsymbol{M})$ is our prior belief over possible hidden long fragment configurations; and the sum is over all possible hidden configurations. To make computation of $P(\boldsymbol{R})$ tractable, we developed a heuristic to first determine the candidate long fragments in each well, and with these seeds, construct a Markov Random Field (MRF) in which each candidate long molecule induces a single potential function over the reads (see Methods).

RFA leverages our model of the read cloud generation process to produce unique alignments to specific copies of repeated sequences (Fig. 2). Wells from the sample are first aligned to the reference using an existing short-read aligner. The features of uniquely mapped read clouds are used to learn $P(\boldsymbol{M})$, which captures protocol properties such as the long fragment size distribution. Each well is then aligned separately using the following steps: First, we use an existing short-read aligner to produce multiple candidate alignments for short reads and to determine the positions of potentially sampled long fragments $\boldsymbol{M}$; second, we perform approximate inference on our model to identify a maximum a posteriori (MAP) assignment

$$\boldsymbol{r}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{r}} \sum_{\boldsymbol{M}} P(\boldsymbol{R} = \boldsymbol{r}|\boldsymbol{M})P(\boldsymbol{M}). \quad (1)$$

Last, we use this MAP assignment to compute probability queries for both short-read alignment confidence and long fragment mappability. Simulations indicated this approach to be highly accurate and efficient in contrast to sampling approaches to compute the marginal, $P(R_{\mathrm{n}})$, for each read. The precise definition of our framework, together with details for efficient identification of the MAP assignment $\boldsymbol{r}^{\mathrm{MAP}}$ in Equation 1 and computation of the queries, is described in Methods.

Alignment accuracy in simulations

In order to determine the utility of RFA over the standard short-read alignment approach, we simulated read clouds from synthetic long read data as described in Methods. From these data, we used the following four different sets of alignments and quality scores that serve as comparison points:

- *Baseline* represents the standard method of aligning short-reads without the benefit of information from a long read process.
- *Naive* represents the naive approach of first creating an abbreviated reference corresponding to candidate long molecules within a well and then realigning the short reads directly to this reference.
- *RFA* is our method that utilizes the same abbreviated reference as in *naive*, and subsequently uses probabilistic inference over an MRF in order to realign the short reads to this reference.

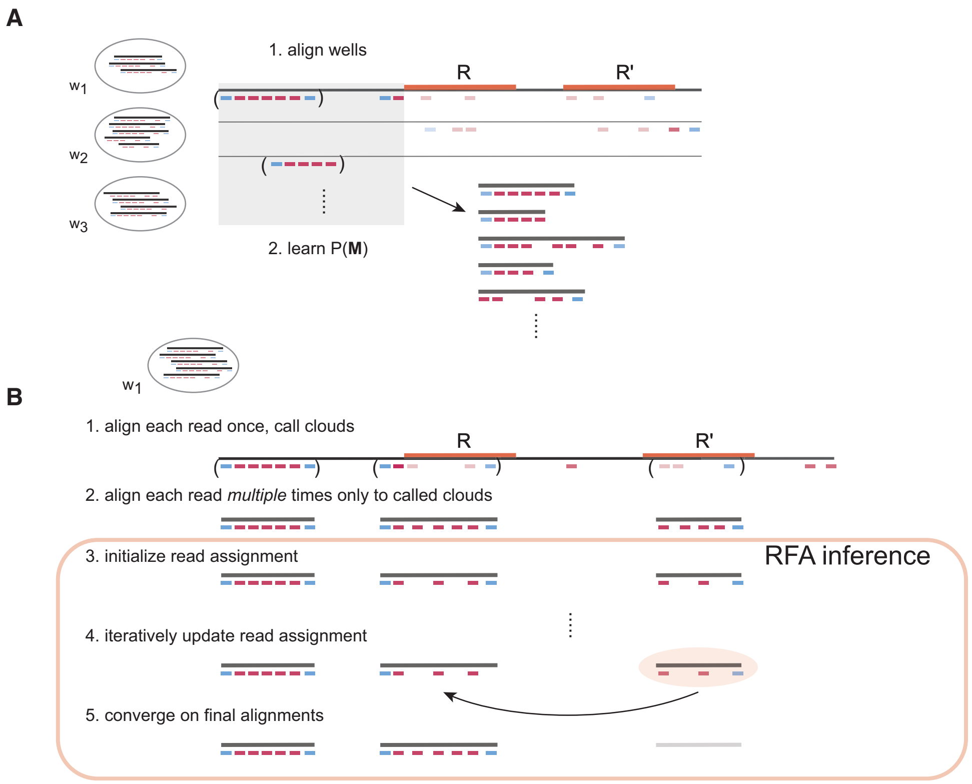
A
1. align wells
R
R'
w1
w2
w3
2. learn P(M)
B
w1
1. align each read once, call clouds
R
R'
2. align each read multiple times only to called clouds
3. initialize read assignment
RFA inference
4. iteratively update read assignment
5. converge on final alignments

Figure 2. RFA overview. (*A*) Wells w_{n} from the sample are first aligned to the reference using an existing short-read aligner, and uniquely mapped read clouds are used to learn a prior $P(\boldsymbol{M})$, which captures protocol properties such as the long fragment size distribution. (*B*) Each well is aligned separately with the aid of a short-read aligner to determine candidate source long fragment locations as well as multiple candidate short-read alignments to the long fragments. Finally, MAP inference is performed to converge on optimal alignments. In this example, RFA successfully determines the correct repeat copy R that overlaps with a source long fragment.

- *Oracle* represents the theoretical upper limit of first aligning the reads to the abbreviated reference using the short-read aligner while allowing for multiple mappings, and then picking the true mapping for each read, if that mapping was returned by the short-read aligner.

We simulated alignment of 1000 wells and computed results of each of the preceding approaches after we filtered reads with a MAPQ less than 10 (corresponding to <90% mapping confidence). The percentage of reads in each well correctly placed by RFA is very similar to Oracle, whereas the Baseline and Naive approaches have significantly lower accuracy (Fig. 3A). RFA confidently maps an additional 2.9% (out of 3.2% from Oracle) of the total reads over the Baseline approach. The results are similar when restricted to reads that were multimapped in the abbreviated reference, which signifies the usefulness of RFA over the naive approach (Fig. 3B). Our probabilistic approach with RFA is able to achieve 92% of the Oracle performance and place an additional 29% of multimapped reads confidently over the naive approach. RFA's error rate, computed by assessing the number of incorrectly placed multimapped reads with >90% confidence, was ~1% across all simulated wells.

Characterization of recovered regions

Our approach accurately aligns short reads to 90.6% (155 Mb) of the 171 Mb of the human genome deemed inaccessible to short-read technologies by the 1000 Genomes Project Consortium. To understand the nature of these recovered regions, we tabulated their RepeatMasker (Smit et al. 1996) and segmental duplication annotations (Bailey et al. 2001; Kent et al. 2002). We are able to recover the majority of previously ambiguous elements across all repeat categories (90%), including 94% of 67 Mb of the segmental duplications and 96% of 11 Mb of transcribed sequences that fall within the 171 Mb of currently inaccessible sequence (Table 1). Both high copy repetitive elements as well as long stretches of ambiguous segmental duplications with high sequence identity are accurately illuminated by our approach.

To quantify gene content in the 155 Mb of inaccessible regions that we recovered with RFA, we used annotations from GENCODE (Harrow et al. 2012). We found that these regions cover >80% of the length of 2740 genes, and >50% of the length of 4510 genes, with both significant enrichment and depletion by type (Table 2). Examining these genes by family also shows significant enrichment and depletion (Table 3). A previous study that used copy number genotyping to examine gene families falling into paralogous regions (Sudmant et al. 2010) found several families to be highly variable and dynamic between individuals and populations. Of the highly dynamic gene families previously identified, we found *ANKRD, ZNF,* endogenous ligands*, CD*, *RBM*, and *RNF* families to be enriched in our recovered gene set. Preliminary analysis indicates that the gene content in these regions is both highly variable and also heavily understudied due to the limitations of existing NGS technologies. The nucleotide-level population variation that our method uncovers could facilitate functional annotation and disease association in these gene-rich repeat regions.

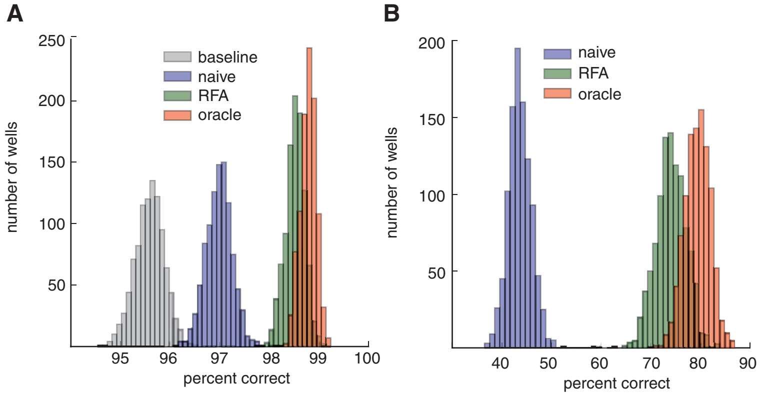

Figure 3. Histograms of simulation results across 1000 wells. Each point in the histogram represents the result of a single simulated well. (*A*) All reads. (*B*) Only reads that were multimapped in the abbreviated reference. RFA confidently maps an additional 2.9% (out of 3.2% from Oracle) of the total reads over the Baseline approach, and achieves 92% of the Oracle performance.

SNV identification within recovered regions

We next applied RFA to two shallow-sequenced TruSeq read cloud samples ($C_R = 1.5\times$). We identified SNVs with these new alignments and performed validation in order to determine our variant discovery accuracy within repeat regions. For each sample, we aligned each well separately using RFA, applied the GATK (DePristo et al. 2011) pipeline to the resulting alignments of all wells simultaneously to recover candidate germline variants, and used computed queries to further filter these candidates to produce variants (see Supplemental Material for details). We compared the resulting calls with variant calls performed using the same GATK pipeline without application of RFA or read cloud information.

NA12878 sample

We first applied RFA to three shallow TruSeq read cloud libraries for the HapMap sample NA12878. The first library ($C_R = 1.5\times$, $C_F = 6.2\times$) (Kuleshov et al. 2014) had sufficiently different coverage properties from the other two ($C_R = 0.5\times$, $C_F = 12.5\times$ for each lane), which were provided to us much later after the protocol had been extensively optimized for whole-genome haplotyping. Each well of the more recent libraries also contained a significantly higher fraction of the genome than the original library (3.4% versus 1.6%).

By applying RFA to these three lanes, we recovered an additional 50,314 variants (35,092 heterozygous) over those found using baseline short-read alignments within the 171 Mb of highly repeated sequences. We validated these recovered variants against three different sources of long-read sequencing data available for NA12878 (Table 4; see Supplemental Material for validation details):

1. A separate whole-genome sample that had been sequenced with high coverage assembled Illumina TruSeq synthetic long reads ($C_R = 50\times$, $C_F = 29\times$; Genomes Moleculo NA12878, 2014, op. cit.), which we aligned to the reference using BWA-MEM (Li 2013). The long reads provided sufficient coverage to directly validate 33,984 of our recovered SNV calls, with a validation rate of 99.5% for our homozygous calls and 92.0% for our heterozygous calls.
2. A set of BACs targeting large high-identity duplications in 15q13.3 that had been sequenced with PacBio (Antonacci et al. 2014). We used BWA-MEM to align the assembled

Table 1. Placement of confident alignments within previously inaccessible 171 Mb (6% of GRCh37) by element class

Element class	Frequency (%)	Illuminated (%)
All	100.0	90.6
Annotated	88.4	90.5
Segdup	43.4	93.8
LINE	35.2	88.3
SINE	14.2	92.7
Gene	7.0	95.5
LTR	6.3	92.3
Simple repeat	4.6	85.9
Satellite	2.3	88.1
Low complexity	1.6	89.0

Frequency is with respect to only previously inaccessible regions and percent illuminated is the fraction of that originally ambiguous class which was correctly and confidently aligned to in simulation.

BACs to the reference, and validated 301 of our recovered SNV calls overlapping the aligned regions. Of the variants that we identified within these regions, 97.6% of our homozygous calls and 53.1% of our heterozygous calls were validated. A validation rate of roughly ~50% for heterozygous calls is expected since these clones are known to originate from a single haplotype in these regions.

3. A set of 103 fosmids that had been randomly selected across the whole genome (Weisenfeld et al. 2014) and sequenced with PacBio. We used BWA-MEM to align the assembled fosmids to the reference and validated 127 variants. Of the variants that we identified within the regions overlapped by these fosmids, 100% of our homozygous calls and 80% of our heterozygous calls were validated.

Invasive ductal carcinoma sample

We then applied RFA to high coverage sequence data obtained from a fresh frozen sample of a grade 3, *ERBB2* amplified, estrogen receptor (ER) negative invasive ductal carcinoma (IDC). Eight Moleculo libraries of total long fragment coverage $C_F = 44\times$ were sequenced at $C_R = 1.8\times$ to produce read clouds with a total short-read coverage of $C = 78.5\times$. The high sequence coverage in this IDC sample allowed us to discover a total of 3,286,470 SNVs. Of these variants, 197,529 (6%) were found in the RFA alignments only and not in the baseline, representing a significant, previously hidden fraction of variation. We randomly selected a subset of the RFA-only variants within the 171 Mb of inaccessible sequence that was amenable to multiplex PCR validation. Of 346 submitted, 323 (93.4%) were validated as displaying an alternate allele frequency (AAF) of at least 0.1 (see Supplemental Material for validation details). Although these validated variants required at least one unique primer (up to one mismatch), and therefore are in principle biased to be within repeat regions with at least one unique nucleotide in range of the amplicon that distinguishes the copies, they provide additional support that matches our variant discovery accuracy measured in NA12878.

We determined the placement of our discovered variants in the IDC sample within each annotation category across the whole genome (Fig. 4). A comparison of the placement of our SNVs in the IDC sample against known, high sequence-identity segmental duplications indicates that a significant portion lie within dense stretches of recovered low copy repeats. We further examined the placement of the identified SNVs across all repeated sequence classes and found that 116,173 (59%) lie within segmental duplications and 16,521 (8%) lie within gene regions, with the rest residing within lower complexity elements (Table 5). We examined both the set of variants that were filtered (by the GATK pipeline) from the baseline run and recovered in the RFA run, as well as the set of variants which were unique to just the RFA run. The majority of variants (55%) were unique to RFA and showed no significant signal in the original baseline short-read alignments.

RFA is able to confidently align reads to 171 Mb of the human genome previously inaccessible to short-read technologies. Validation with assembled synthetic long reads, PacBio assembled clones, as well as PCR and deep targeted sequencing confirm that these alignments can be used for accurate nucleotide-level variant discovery. Examination of both alignments and SNVs in the IDC sample indicates that RFA enables variant discovery across all classes of repeated sequence, including genes, high copy repetitive elements, and high sequence identity segmental duplications.

Discussion

In this work, we demonstrate the ability of RFA to leverage read clouds to accurately map short reads in 171 Mb of the human genome that is inaccessible to variant calling using short reads. Our method enables the use of the TruSeq synthetic long-read protocol in a cost-effective, read cloud setting, in which each long molecule is covered lightly with reads ($C_R \leq 2\times$) rather than requiring long molecules to be assembled into synthetic long reads through deep sequencing ($C_R = 50\times$). Using a probabilistic approach, we jointly map all reads of a well to the reference genome to produce confident unique mappings that subsequently enable accurate discovery of novel nucleotide variation in this previously dark portion of the human genome.

Our approach models the features of a given synthetic long-read protocol in $P(\boldsymbol{M})$. For TruSeq read clouds, we chose to decompose and learn this function with empirical distributions,

Table 2. Gene type breakdown in recovered regions

Type	Count	Enrichment	*P*-value
IG-V gene	50	+	2.44×10^{-15}
Pseudogene	1612	+	2.44×10^{-15}
miRNA	399	+	2.44×10^{-15}
IG-D gene	37	+	2.44×10^{-15}
IG-V pseudogene	49	+	7.33×10^{-15}
IG-J gene	7	+	2.74×10^{-3}
IG-C gene	4	+	
IG-J pseudogene	2	+	
Processed transcript	45	+	
Polymorphic pseudogene	5	+	
TR-J gene	6	+	
IG-C pseudogene	1	+	
3′ overlapping ncRNA	1	+	
Protein coding	825	−	2.99×10^{-67}
Antisense	219	−	1.45×10^{-16}
snoRNA	48	−	2.39×10^{-8}
Sense intronic	16	−	6.36×10^{-8}
snRNA	89	−	4.02×10^{-4}
Sense overlapping	4	−	2.98×10^{-2}
rRNA	23	−	
Misc RNA	115	−	
lincRNA	457	−	

Type of the 4510 genes with >50% of the length covered by recovered alignments. Corresponding enrichment (+) and depletion (−) are shown with only significant *P*-values shown here.

Table 3. Gene family breakdown in recovered regions

Family	Count	Enrichment	*P*-value
ZNF	722	+	9.79×10^{-12}
SLC	456	+	2.09×10^{-9}
endogenous ligands	235	+	7.67×10^{-6}
CD	388	+	1.44×10^{-5}
I-set domain containing	161	+	1.58×10^{-4}
BTBD	134	+	1.63×10^{-4}
WDR	262	+	6.17×10^{-4}
OR2	113	+	1.73×10^{-3}
OR5	112	+	1.93×10^{-3}
RNF	274	+	3.29×10^{-3}
EF-hand	225	+	3.60×10^{-3}
PLEKH	207	+	3.99×10^{-3}
TTC	112	+	2.77×10^{-2}
RBM	213	+	3.74×10^{-2}
ANKRD	242	+	4.05×10^{-2}
V-set domain containing	163	+	4.43×10^{-2}
tRNA	612	–	0.0
scRNA	866	–	5.15×10^{-55}
VN1R	112	–	4.00×10^{-18}
snRNA	67	–	2.01×10^{-14}
RPL	243	–	7.89×10^{-13}
VN2R	20	–	4.74×10^{-6}
rRNA	34	–	1.34×10^{-2}

Families with enrichment and depletion are shown.

but replacing it would allow our approach to be readily adapted to other synthetic long read technologies.

Although we provide proof of concept of our method for discovering copy-specific SNVs on a whole-genome human sample, there are numerous additional potential applications of aligning read clouds. In future work, our improved alignments can be leveraged to discover other types of variation, including indels and larger structural rearrangements present in these regions. Our preliminary analyses of the recovered alignments in the human genome also suggest a copy-rich structure that has yet to be resolved accurately beyond illuminating SNV variation.

Our approach can be used to resolve ambiguity in complex regions of high sequence identity. It could be useful in resolving ambiguity in homologous regions, such as MCS and LRC, for which the updated GRCh38 human reference genome contains multiple alternative haplotype paths (Church et al. 2011). Aligning with RFA may enable determination of the correct haplotype paths of these variable regions. Our approach may also prove useful for subspecies discovery and quantification in metagenomic samples for which previous work had much higher sequencing requirements using TruSeq-assembled synthetic long reads (Sharon et al. 2015). There has been previous work to leverage PacBio sequencing to accurately resolve RNA transcripts (Sharon et al. 2013). With some modeling extensions, RFA can potentially be used to align directly to an available transcriptome (of which many transcripts will have high sequence identity) or for discovery of novel transcripts with previously unknown alternative splicing sites.

Our current method has several limitations. First, alignment generation with standard short-read aligners limits our method's ability to fully illuminate high copy-number elements because they may not generate any concordant paired alignments for reads originating from high-copy regions. If the true candidate alignment for a particular read is not produced by short-read alignment during pass2, then our method will be unable to find the true mapping for that read. However, our approach naturally lowers the quality scores for alignments in high-copy regions such that they do not affect further analyses. Second, although our inference procedure performed well in practice, it may benefit from an improved proposal distribution to explore the state space of possible alignment configurations more effectively and further close the gap with respect to Oracle performance. Finally, breaking the independence assumption to allow variant information to be shared across wells may improve variant discovery accuracy and recall.

Of the 130.5 Mb that are identified as high sequence identity segmental duplications (Bailey et al. 2002), a significant portion is currently hidden from short-read technologies, and the effect of variation in these regions is largely unknown. High-identity segmental duplications are highly variable between individuals, and copy number variants (CNVs) within these regions are strongly correlated with increased sequence identity (Redon et al. 2006). Nonallelic homologous recombination between LCRs as well as *Alu* repetitive elements can result in a variety of balanced and unbalanced structural alteration events (Redon et al. 2006; Sasaki et al. 2010; Ou et al. 2011). These regions are enriched for transcript content, and a subset of these regions consists of multicopy genes also known to vary widely in copy number between populations and individuals (Sudmant et al. 2010). Illumination of these highly dynamic regions will enable discovery of variation across individuals, and ultimately, functional annotation and phenotype association.

Methods

In order to leverage read clouds for accurate read alignment, we model the long read generative process to capture the dependencies between the resulting short reads through the hidden long fragments. Using a standard short-read aligner to provide seeded candidate long molecules and short-read alignments, we are then able to reduce the problem of finding optimal alignments to optimizing over an MRF. We first define our model and our objective function and then show how we can use this framework to obtain alignments and their respective mapping quality scores.

Definitions

For wells $\boldsymbol{W} = \{1, 2, \ldots, W\}$, we set up the following system to describe the long read process. Each $w \in \boldsymbol{W}$ contains a set of read *fragments* $\boldsymbol{R}_w = \{R_{wn} | n = 1, 2, \ldots, N_w\}$. Each read fragment, R_{wn}, can potentially have multiple alignments, and we denote $\boldsymbol{A}_{wn} = \{a_{wnk} | k = 1, 2, \ldots, K_{wn}\}$ as the set of its possible alignments with respect to the reference genome. We incorporate our knowledge of the synthetic long-read process by modeling the set of original long molecules $\boldsymbol{M}_w = \{M_{wi} | i = 1, 2, \ldots, I_w\}$ that generated the short reads.

For each well w, the process can be described by the following (w is omitted for clarity):

Table 4. Validation results of our NA12878 SNVs

Validation set	Overlapping SNVs	Homozygous (accuracy %)	Heterozygous (accuracy %)
TruSeq SLR	33984	9651 (99.5%)	24333 (92.0%)
15q13.3 PacBio BACs	301	126 (97.6%)	175 (53.1%)
103 PacBio fosmids	127	77 (100.0%)	50 (80%)

Validation rates of our SNVs that overlap long sequencing reads from high coverage TruSeq-assembled synthetic long reads (SLR), PacBio-assembled BACs targeting high identity duplications in 15q13.3, and 103 PacBio-assembled fosmids randomly selected across the whole genome.

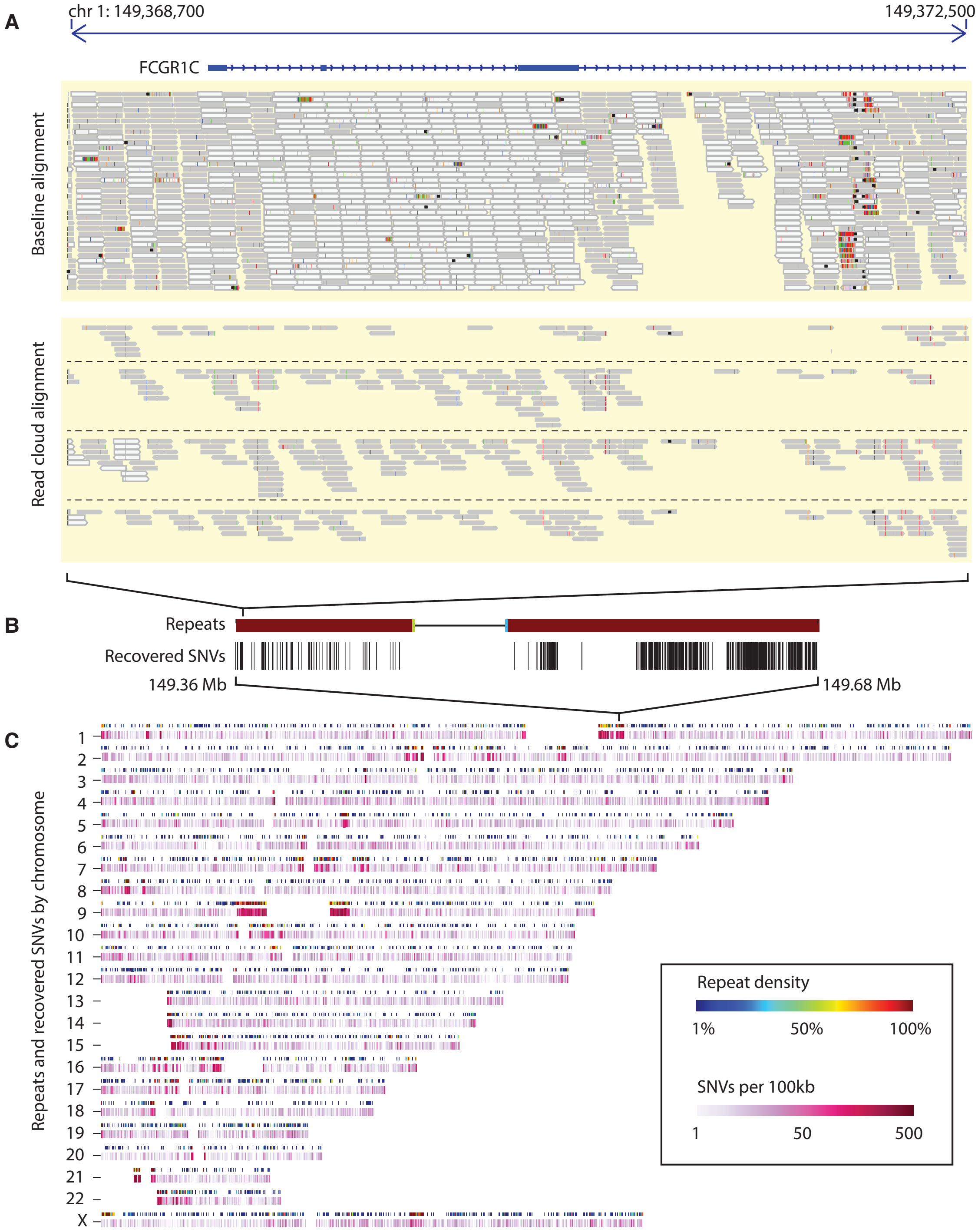

Figure 4. Whole-genome SNV calling on the IDC sample. (*A*) Comparison of the initial baseline short-read alignments of all the wells merged together with four wells aligned with RFA (from two distinct haplotypes), in a region overlapping the *FCGR1C* gene. (*B*) Placement of recovered SNVs within the surrounding 300-kbp region. (*C*) Density of recovered SNVs throughout the whole genome (*bottom* track), by chromosome, compared to density of segmental duplications (*top* track). Long clustered regions of recovered SNVs coincide with dense regions of annotated segmental duplications.

1. A set of hidden random vectors denoting long molecules that generated the short reads:

$$\boldsymbol{M} = \{M_c | c = 1, 2, \ldots, C\}.$$

To describe M_c, we feature it as (L_c, $\boldsymbol{X_c}$, λ_c, B_c, S_c, E_c). L_c is the aligned position in the reference; $\boldsymbol{X_c}$ represents the hidden sequence of the molecule; and S_c is the size of the molecule. λ_c controls the short-read emission random process. B_c captures the presence of the end-markers and can be stored as a conditional probability table (CPT). B_c, S_c, and λ_c can all interact depending on the particular long-read protocol, so we model them jointly as $P(B_c, S_c, \lambda_c)$. For modeling convenience, E_c

Table 5. Placement of SNVs from the IDC sample within previously inaccessible 171 Mb (6% of GRCh37) by element class

Element class	Recovered SNVs (heterozygous)	Unique SNVs (heterozygous)
All	89,568 (49,184)	107,961 (82,113)
Annotated	86,889 (47,308)	105,407 (80,237)
Segdup	41,050 (27,105)	75,123 (57,822)
LINE	36,375 (17,825)	31,309 (24,404)
SINE	15,383 (6992)	17,090 (12,516)
LTR	7816 (4405)	11,010 (8256)
Gene	5817 (3831)	10,704 (8485)
Simple repeat	1759 (916)	3112 (2143)
Satellite	4055 (2754)	2222 (1789)
Low complexity	380 (242)	844 (632)

The whole-genome SNV numbers are from our high coverage IDC sample and were obtained using the same criteria as in validation. Recovered SNVs were present and filtered in the original Bowtie 2 run but predicted in the RFA run. Unique SNVs are exclusive to just the RFA run.

denotes whether this source molecule existed in this well, and $E_c \sim \mathrm{Ber}(p^e)$.

2. A set of partially observed random vectors denoting read fragments:

$$\boldsymbol{R} = \{R_n | n = 1, 2, \ldots, N\}.$$

We represent R_n as ($\boldsymbol{o}_n$, A_n) where $\boldsymbol{o}_n$ is the observed information of the read such as nucleotide sequence and base quality scores, and A_n is the hidden alignment and captures the differences between the reference and the read for this particular alignment.

The distribution over the read fragments, over which we seek to optimize, can be described by

$$P(\boldsymbol{R}) = \sum_{\boldsymbol{M}} P(\boldsymbol{M}, \boldsymbol{R}), \quad (2)$$

where the sum is over all possible hidden long fragment configurations.

Long-read alignment seeding

Although the original long molecules themselves are completely latent, we can predetermine a set of partially observed candidate long molecules

$$\{M_c = (L_c = l_c, \boldsymbol{X_c}, \lambda_c, B_c, S_c = s_c, E_c) | c = 1, 2, \ldots, C\}$$

as follows (Fig. 5). Reads are aligned to the full hg19 reference (*pass1*) using a short-read aligner allowing at most one alignment per read. We pass through the resulting alignments and use a heuristic set of rules to detect candidate clouds corresponding to clusters of reads. We group reads that are within 3.5 kbp of one another into the same cluster and require a cluster to have a minimum of six reads to constitute a candidate cloud. We adopt a simple heuristic to leverage the TruSeq long-fragment end-marker information to split a candidate cloud in two in the cases in which two long fragments happen to be within 3.5 kbp of one another (without mixing overlap) and would otherwise be clustered by our simple rule above. The choice of 3.5 kbp as the distance between reads was made using training simulations to minimize the number of superfluous candidate clouds (not representing true sampled long fragments) while also minimizing false negative clouds so that true candidates are not missed. After candidate cloud calling, we then create an abbreviated reference containing a separate contig for each candidate long molecule. All reads are then realigned to this abbreviated reference (*pass2*) allowing multiple candidates per read.

The resulting alignments serve as seeded domains A_n for each read R_n. Definition of the abbreviated reference helps in filtering spurious alignments to distant homolog sequences, which should not be considered as potential source molecules. In practice, we found that a lenient threshold for calling candidate clouds in pass1 was sufficient to retain nearly all the true candidate clouds containing ambiguous reads.

Read graph construction

In order to perform inference, we make the following simplifying assumptions:

1. The underlying sample sequence in one well does not provide any information about the underlying sequence in other wells such that

$$P(\boldsymbol{R}) = \prod_{w} P(\boldsymbol{R}_w).$$

2. Within each well, the long molecules do not overlap in the same genomic region in the reference genome.

The process to generate read clouds in one well (dilution, shear into short fragments, and barcoding) is unlikely to affect the observed reads in any other well. However, since the wells share the same underlying germline genome, the observed short reads in each well are actually not independent and interact through the unobserved underlying sample sequence. Although our first assumption in principle limits the ability of our model to leverage all available information, it is a significant simplification that allows us to efficiently perform alignment on each well separately. Regarding the second assumption, there is actually a nontrivial chance two fragments sample overlapping genomic coordinates in a single well, equal to the well genome coverage, which was ~2% in our samples. However, the end-markers from TruSeq allow us to detect most of these collisions when they do occur. We estimated the end-marker efficiency for our samples to be $p = 0.77$; assuming independence of the end-markers of two overlapping fragments allows us to detect $1 - p^2 = 95\%$ of collisions. There is an additional, subtle assumption here: In order for fragments to be nonoverlapping from one another, they would have to be sampled from a nonuniform distribution such as the "parking process" (Krapivsky 1992). However, when the total fragment length is small compared to the total target genome length, the uniform distribution of sampled fragments (Lander and Waterman 1988) is a good approximation of the parking process (Batzoglou et al. 1999). All together, we approximate the long fragments as being nonoverlapping, but otherwise independently drawn, and decompose read assignment scores by candidate long molecules for tractable computation

$$P(\boldsymbol{M}_w) \approx \prod_{c} P(M_{wc}).$$

The assumptions above allow us to decompose the global probabilistic function in Equation 2 to

$$P(\boldsymbol{R}) = \prod_{c} \sum_{M_c} P(M_c) P(\boldsymbol{R}_c | M_c), \quad (3)$$

where c indexes over clouds, and $R_n \in \boldsymbol{R}_\mathrm{c}$ if there is at least one alignment in A_n within the coordinate range $[l_c, l_c + s_c)$. The sum over M_c is over all possible hidden values of the domain of the features defined in the previous section. This results in a Markov Random Field (MRF) with each long molecule inducing a potential function $\phi_\mathrm{c}(\boldsymbol{R}_c)$ of the form above $[\varphi_\mathrm{c}(\boldsymbol{R}_c) = \sum_{M_c} P(M_c) P(\boldsymbol{R}_c | M_c)]$ such

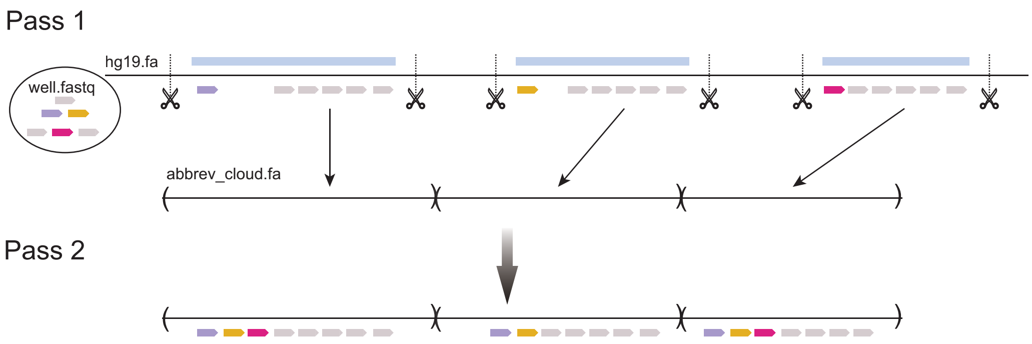

Figure 5. Abbreviated reference framework. The framework for generating putative long reads and associated short-read alignments to these segments: (1) Reads are aligned to hg19 at most once in pass1; (2) putative long read segments are identified and spliced together to create an abbreviated reference; and (3) reads are aligned again in pass2 to this abbreviated reference allowing multiple mappings.

that the distribution of read assignments can be represented as

$$P(\boldsymbol{R}) = \frac{1}{Z}\prod_{c} \varphi_c(\boldsymbol{R}_c).$$

For a particular read assignment, $\boldsymbol{R}_c = (\boldsymbol{o}_c, \boldsymbol{a}_c)$, L_c, B_c, S_c, and E_c can be trivially inferred to result in only one significant value in each of their domains (shown as constants below). To compute the desired potential value, $P(\boldsymbol{R}_c)$, we still have to integrate or sum over the domain of λ_c and $\boldsymbol{X_c}$. Our prior of long molecule features decomposes as

$$\begin{aligned} P(M_c) &= P(L_c = l_c, \boldsymbol{X_c}, \lambda_c, B_c, S_c = s_c, E_c = e_c) \\ &= P(E_c = e_c)P(B_c = b_c, S_c = s_c, \lambda_c|E_c)P(L_c = l_c|E_c) \prod_{X_l \in \boldsymbol{X_c}} P(\boldsymbol{X_c}|E_c) \end{aligned} \tag{4}$$

E_c is a binary random variable, which is 1 if any reads are assigned to M_c and 0 otherwise. Note that $P(L_c, B_c, S_c, \lambda_c, \boldsymbol{X_c}|E_c = 0) = 1$. $P(X_1|E_c = 1) \sim \mathrm{Cat}(\theta^l)$ captures our belief of the SNP mutation rate at position l in the reference genome (due to germline plus somatic variation). In our implementation, we estimated a shared parameter by determining the germline SNP rate at unique regions in the genome. We chose $P(L_c|E_c = 1)$ to be uniform, but the bias of a particular protocol to sample certain long fragments over others could be captured in this function. Decomposition of the prior as in Equation 4 allows us to compute the potential $\phi_c(\boldsymbol{R}_c)$ efficiently in linear time by variable elimination of $\boldsymbol{X_c}$ and λ_c.

Our approximate inference procedure relied on the ability to compute the score of an updated assignment efficiently using the decomposition across candidate long molecules in Equation 3 and is explained in the Supplemental Section MAP Inference Procedure. Although we obtained good results by decoupling the wells, this assumption can be loosened to allow interaction between wells in order to share information and improve performance. Each well has information about the set of underlying variants $\boldsymbol{X}$, but contains a different set of observations through its subset of reads. Allowing wells to share their belief about $\boldsymbol{X}$ would affect the joint alignment and in principle may result in improved performance. Although our assumption that long fragments in each well do not overlap in genomic coordinates was useful in deriving our model, it is often not true in practice. Fortunately, the presence of the end-marker short reads in TruSeq sequencing allowed us to detect most of the cases when sampled long fragments overlapped. The respective reads were excluded from our analysis since we do not have the ability to tell from which long fragment the short reads originated in the case of an overlap.

Query computation

Read quality scores

We compute maximum a posteriori (MAP) read alignments (see Supplemental Material for details) and infer quality scores for the alignments as follows. We assume that once the MAP assignment converges, reads only have local interactions such that

$$P(R_n) = P(R_n|\boldsymbol{R}_{l-} = \boldsymbol{r}_{l-}^{\mathrm{MAP}}),$$

where $\boldsymbol{R}_{l-}$ denotes the set of reads which do not overlap R_n in the identified MAP assignment. We can then compute

$$\begin{aligned} P(R_n|\boldsymbol{R}_{l-}) &= \sum_{\boldsymbol{R}_{l-n}} P(\boldsymbol{R}_l|\boldsymbol{R}_{l-}) \\ &= \sum_{\boldsymbol{R}_{l-n}} \sum_{\boldsymbol{X}} P(\boldsymbol{R}_l, \boldsymbol{X}|\boldsymbol{R}_{l-}) \\ &= \sum_{\boldsymbol{R}_{l-n}} \sum_{\boldsymbol{X}} P(\boldsymbol{R}_l|\boldsymbol{X}, \boldsymbol{R}_{l-})\mathrm{P}(\boldsymbol{X}|\boldsymbol{R}_{l-}), \end{aligned}$$

where $\boldsymbol{R}_{l-n}$ denotes $\boldsymbol{R}_l$ excluding R_n; and $\boldsymbol{X}$ is the union of sequence at all the potential locations $\boldsymbol{R}_l$ could be mapped to. The small sizes of both $\boldsymbol{R}_{l-n}$ and $\boldsymbol{X}$ allow this quantity to be computed efficiently. The computed $P(R_n|\boldsymbol{R}_{l-})$ can then be converted to a MAPQ score for that read.

With our method, alignments converge in the MAP assignment such that certain candidate long molecules are empty, and the corresponding candidate alignments within these coordinates can be eliminated from the domains $\boldsymbol{A}_n$. If the resulting domains have only one alignment active, as is often the case for low copy repeats, then the quality score will be high. However, the quality score for a recovered alignment will be lowered if it falls in elements that are repeated with high sequence identity within the resulting active long molecules, in which cases $P(R_n|\boldsymbol{R}_{l-})$ will be naturally lowered.

Cloud quality scores

For each cloud, we can directly compute $P(\boldsymbol{R}_c) = \sum_{M_c} P(M_c)P(\boldsymbol{R}_c|M_c)$ for the identified MAP assignment by variable elimination as described in the previous section. This quantity in log-space is directly proportional to the number of variants predicted in a long fragment for a given assignment, $\boldsymbol{R}_c$. Low-quality long molecules may correspond to novel copy number variants in the sample that are not present in the reference and whose sequence identity with the reference homolog where they map is unusually low. We exclude short reads within clouds with low quality scores from our variant calling pipeline. A pooling strategy that collects all such read clouds to assemble additional potential copy number variants in the target genome is an interesting subject for future work.

Our read quality score, $P(R_n|\boldsymbol{R}_{l-} = \boldsymbol{r}_{l-}^{\mathrm{MAP}})$, does not capture cases in which the alignment may be unique within an active long molecule, but all the reads of that molecule could map nearly equally well to an inactive long molecule. We approximated $P(E_c|\boldsymbol{R}_{c-} = \boldsymbol{r}_{c-}^{\mathrm{MAP}})$ (which cannot be computed exactly due to the size of $\boldsymbol{R}_c$), and chose an appropriate cutoff using training simulations for excluding alignments in order to minimize error. We can approximate it effectively by identifying the set of assignments to inactive clouds $\{r_{c_1}, \ldots, r_{c_k}\}$, computing $P(\boldsymbol{X}_c|\boldsymbol{r}_c)$ for each one, and then renormalizing. This is an inexpensive way to identify the peaks of the distribution since upon convergence, the assumption is that interactions between reads of different long molecules are mostly decoupled. The result can be used to exclude long molecules for which all the member read fragments can map equally well in another source molecule, c', thus making them indistinguishable.

Simulations of read clouds

We used a nonparametric approach in order to simulate read clouds and capture biases intrinsic to the sequencing process (see Supplemental Fig. 1). We first aligned all the wells from our IDC sample and generated candidate clouds using the same criteria described in pass1. We removed outlier clouds in regions with low mappability (corresponding to repeats) or where we observed unusually high coverage likely due to copy number variants. The remaining set of read clouds was assumed to be representative of true sequenced read clouds that are present across the whole genome. For each read cloud in this set, we created a stencil of the short-read positions relative to the start of the cloud to be used as a "cookie cutter" for generating short reads in a simulated read cloud. In order to simulate base substitution errors in reads, we fit a first-order model of the read bases and base quality scores from our sequenced sample. We introduced a substitution error with probability proportional to the simulated base quality score. To simulate reads from a well, we choose a location uniformly at random in the reference and draw a stencil that has not yet been used in order to create a sampled long fragment.

We repeat the process until the well contains the expected genome coverage (2% in our samples). The use of this empirical simulation methodology, rather than use of a model-based simulation, enables us to more accurately capture intrinsic biases present in read-cloud sequencing and to better estimate the true accuracy of our alignment approach. Still, we recognize that this simulation strategy does not capture any sequence-specific biases of the read-cloud and sequencing protocols.

TruSeq instantiation

To align short reads to the reference, we chose to use Bowtie 2 (Langmead and Salzberg 2012) for its high efficiency and accuracy and for its ability to output multiple candidate alignments for each read. We restricted the number of alignments for each read to be 15. Reads with possibly more alignments were unlikely to be informative in the placement of other reads.

We found that the distributions of B_c, S_c, and λ_c for TruSeq read clouds were difficult to parameterize. The read-cloud size and density varied greatly conditioned on the presence of the end markers, so we factorized $P(B_c,S_c,\lambda_c)$ as $P(B_c)P(S_c,\lambda_c|B_c)$. These functions were estimated with Kernel Density Estimates (KDEs) on features extracted from the valid clouds described in the simulations section.

Implementation and availability

RFA is a Python package that leverages the short-read aligner Bowtie 2. It is open source and freely available at http://readclouds.stanford.edu and in the Supplemental Material. To align a single lane of sequenced read clouds, a subset of the wells are aligned using the default Bowtie 2 settings, and features of uniquely mapped clouds are extracted to learn the cloud model $P(\boldsymbol{M})$. To achieve practical runtimes on large genomes, our implementation required the use of a compute cluster to parallelize alignment across wells. Each well is aligned in the following main steps: (1) Align all reads to the whole reference once; (2) determine abbreviated reference and align all reads to this abbreviated reference allowing multiple candidates; (3) build in memory structures and the read MRF graph; (4) perform MAP inference to determine optimal alignments; and (5) compute alignment quality scores and generate the final alignment files. For the first lane of NA12878 sequencing ($C_R = 1.5\times$, $C_F = 6.2\times$), for which we initially designed our implementation, steps 1–5 took 40, 76, 57, 29, and 49 CPU hours, respectively, across all 384 wells. Our method requires about 6× more total CPU time than generating baseline alignments (step 1). In future software releases, steps 1–3 may be folded into a single step and further optimized, and steps 3–4 may be implemented in a compiled language.

Data access

All raw sequence reads for the eight Illumina TruSeq libraries for the IDC sample as well as the three libraries for NA12878 have been submitted to NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA287848.

Acknowledgments

We thank Illumina, Inc., for contributing the two additional lanes of TruSeq read-cloud sequencing for NA12878. This paper was selected for oral presentation at RECOMB 2015, and an abstract is published in the conference proceedings. A.B. has been supported in part by National Institutes of Health (NIH) training grant T32HG000044. This work has been supported in part by NIH grant # R01CA183904.

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

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

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M, et al. 2014. Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* **46:** 1293–1302.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2014. Minion nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33:** 296–300.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current Human Genome Project assembly. *Genome Res* **11:** 1005–1017.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Batzoglou S, Berger B, Mesirov J, Lander ES. 1999. Sequencing a genome by walking with clone-end sequences: a mathematical analysis. *Genome Res* **9:** 1163–1174.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al.

2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, et al. 2011. Modernizing reference genome assemblies. *PLoS Biol* **9:** e1001091.

Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46:** 1063–1071.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464:** 704–712.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24:** 688–696.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006.

Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, et al. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* **29:** 59–63.

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol* **14:** R101.

Krapivsky PL. 1992. Kinetics of random sequential parking on a line. *J Stat Phys* **69:** 135–150.

Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, Kertesz M, Snyder M. 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32:** 261–266.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2:** 231–239.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28:** 2097–2105.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv. preprint arXiv* 1303.3997.

McCarthy A. 2010. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol* **17:** 675–676.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, et al. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41:** 1223–1227.

McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier AS. 2014. Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9:** e106689.

Metzker ML. 2010. Sequencing technologies—the next generation. *Nat Rev Genet* **11:** 31–46.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L, et al. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* **21:** 33–46.

Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487:** 190–195.

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, et al. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466:** 368–372.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Sasaki M, Lange J, Keeney S. 2010. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol* **11:** 182–195.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J. 2007. Strong association of de novo copy number mutations with autism. *Science* **316:** 445–449.

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31:** 1009–1014.

Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M, Thomas BC, Burstein D, Tringe SG, et al. 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res* **25:** 534–543.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135–1145.

Smit AF, Hubley R, Green P. 1996. *Repeatmasker open-3.0*. http://www.repeatmasker.org/.

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* **455:** 232–236.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330:** 641–646.

Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ. 2013. The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2:** e00569.

Weisenfeld NI, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C. 2014. Comprehensive variation discovery in single human genomes. *Nat Genet* **46:** 1350–1355.

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, et al. 2010. Genome-wide association study of CNVS in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464:** 713–720.

Received February 14, 2015; accepted in revised form August 14, 2015.