Published online 19 December 2017

Nucleic Acids Research, 2018, Vol. 46, No. 4 ***1661–1673***
doi: 10.1093/nar/gkx1266

A survey of localized sequence rearrangements in human DNA

Martin C. Frith[1,2,3,*] and Sofia Khan[3]

[1]Artificial Intelligence Research Center, AIST, Tokyo 135-0064, Japan, [2]Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8562, Japan and [3]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), AIST, Tokyo 169-8555, Japan

Received August 30, 2017; Revised November 13, 2017; Editorial Decision December 06, 2017; Accepted December 07, 2017

ABSTRACT

Genomes mutate and evolve in ways simple (substitution or deletion of bases) and complex (e.g. chromosome shattering). We do not fully understand what types of complex mutation occur, and we cannot routinely characterize arbitrarily-complex mutations in a high-throughput, genome-wide manner. Long-read DNA sequencing methods (e.g. PacBio, nanopore) are promising for this task, because one read may encompass a whole complex mutation. We describe an analysis pipeline to characterize arbitrarily-complex 'local' mutations, i.e. intrachromosomal mutations encompassed by one DNA read. We apply it to nanopore and PacBio reads from one human cell line (NA12878), and survey sequence rearrangements, both real and artifactual. Almost all the real rearrangements belong to recurring patterns or motifs: the most common is tandem multiplication (e.g. heptuplication), but there are also complex patterns such as localized shattering, which resembles DNA damage by radiation. Gene conversions are identified, including one between hemoglobin gamma genes. This study demonstrates a way to find intricate rearrangements with any number of duplications, deletions, and repositionings. It demonstrates a probability-based method to resolve ambiguous rearrangements involving highly similar sequences, as occurs in gene conversion. We present a catalog of local rearrangements in one human cell line, and show which rearrangement patterns occur.

INTRODUCTION

Genomes mutate and evolve by base substitutions, deletions, inversions, tandem duplications, and more complex events (1), up to large-scale chromosome shattering (2). DNA sequence rearrangements contribute to a wide range of medical disorders (1,3), and somatic rearrangements serve as hallmarks in tumor genomes (4,5).

The methods for studying structural variations have long been laborious, time-consuming or low resolution, e.g. fluorescence *in situ* hybridization (FISH), comparative genomic hybridization (CGH) or genome-wide SNP arrays (reviewed in (6)). Short-read DNA sequencing can provide high-throughput base-level resolution; however, the read length limits the types of structural variation detected, and limits reliable alignment to repetitive regions (7), including subtelomeric regions that are known to harbor rearrangements (8).

A recent study surveyed 'large' structural variation in 689 people, at ∼5 kb resolution, by both long-insert and linked-read whole-genome sequencing (3). On one hand, they found many novel 'complex' variants: most (97%) are intrachromosomal, and they were classified into 16 types. On the other hand, their data cannot resolve micro-complexity (<5 kb), so they likely underestimate the complexity and diversity of complex variants.

Long-read DNA sequencing technologies provide read lengths to span tens of kilobases and beyond, which enables investigation of complex rearrangements spanned by one read. Long-read DNA sequencing has been used on its own (9,10), or in combination with short reads (11), to investigate structural variation. The present tools for structural variant calling solely from long read data focus on insertions, deletions, and inversions (SMRT-SV (9)) and additionally translocations (MultiBreaks-SV (12) and PBHoney (13)).

In this study, we describe a method for finding arbitrarily-complex 'local' rearrangements, i.e. intrachromosomal rearrangements spanned by one DNA read. (We ignore pure deletions, which are relatively common and well-studied.) Our approach has two unusual features. First, we consider the ancestral state of each rearrangement, which is necessary in general to characterize them accurately. It also leads to a better understanding of 'insertions'. Second, our approach is based on probabilities. There is often ambiguity in how to align rearranged fragments of a read, for example in gene conversion (replacement of a DNA segment by

[*]To whom correspondence should be addressed. Tel: +81 3 3599 8001; Fax: +81 3 5530 2061; Email: mcfrith@edu.k.u-tokyo.ac.jp

a duplicate of a similar segment). Such ambiguity can be resolved by calculating the probabilities of alternative alignments, based on the probabilities of substitutions and gaps due to sequencing error. These two features come together in an overlooked classical algorithm, the Repeated Matches Algorithm (14), which finds the most probable alignments of derived to ancestral sequences allowing for any number of duplications, deletions, repositionings, and insertions.

This method is applied to both MinION and PacBio DNA reads from the same human cell line. We describe error rates, and the troubling presence of dubious rearrangements that are likely artifacts of the DNA sequencing process. By semi-manual analysis, we extract 166 likely-real rearrangements, many of which are found in both MinION and PacBio. This reveals several recurring patterns of rearrangement.

Very recently, two methods appeared for finding structural variants in long reads: NanoSV (15) and NGMLR/Sniffles (Sedlazeck *et al.* https://doi.org/10.1101/169557). Our study is largely complementary to those. Whereas they characterize variants automatically, we draw pictures of them and manually survey the types of rearrangement. We find intricate rearrangements, such as heptuplication, localized shattering, and gene conversion, not mentioned by them. Finally, those studies do not consider ancestral arrangements, and NGMLR/Sniffles does not use fitted probabilities.

The remainder of this Introduction section describes the relevant aspects of DNA sequence evolution and probability-based alignment.

Comparing rearranged sequences

In order to find rearrangements between two sequences, we must first determine ‘equivalent’ positions between them, i.e. align them. To clarify this, let us consider how DNA sequences can mutate:

- Substitution of bases.
- Deletion of one or more consecutive bases.
- Duplication of sequence.
- Spontaneous generation: insertion of sequence not descended from any ancestor. This is unusual. An example is retrotranscription and insertion of RNA poly-A tails, which are not descended from any ancestral sequence.
- Repositioning of sequence segments, e.g. inversion or translocation.
- Insertion of sequence from another (e.g. viral) genome.

One mutational event may produce several of these effects, e.g. repositioning and deletion and duplication (1).

Let us now consider sequence evolution with complex mutations. Figure 1 shows an ancestral sequence (A) evolving into two derived sequences (B and C). If we wish to find equivalent positions between B and C, this means identifying bases that descend from the same base in their most recent common ancestor. In general, this is hard: for example, it is hard to avoid an incorrect alignment between segment 3 in B and segment 4 in C.

The key observation is that it’s fundamentally easier to compare a derived sequence to an ancestral sequence, e.g. B to A. This is because *every* part of the derived sequence (apart from spontaneously-generated parts) is descended from a *unique* part of the ancestral sequence(s). Thus, a reasonable approach is to seek an optimal division of B into parts, where each part is aligned to a most-similar part of A. It may be objected that we rarely have an ancestral sequence, but this can be worked around using outgroups (see below).

Classic probability-based alignment

The classic approach to sequence alignment, which ignores rearrangement, is based on probabilities (14). First, we must determine the probabilities of each kind of substitution (e.g. `c`→`t`), and insertion and deletion probabilities, for a specific kind of sequence comparison. Four examples are shown in Table 1. Then, we can align sequences *while considering these probabilities*: in other words, prefer alignments with higher probability (14). It stands to reason that this approach will tend to produce more accurate alignments (than alignment by ad hoc criteria not based on the rates of substitution, insertion and deletion).

Probability-based alignment with rearrangements

Thus, we would like to find an optimal division of a derived sequence into parts, and optimal alignment of each part to ancestral sequence(s) based on probabilities of substitution, insertion and deletion. Fortunately, there is an algorithm to do exactly this: the Repeated Matches Algorithm (in section 2.3 of (14)). This algorithm seems to do more work than classic alignment (because it finds an optimal segmentation *and* alignment), but its computational cost is not significantly greater (14). It is too slow for human-genome-scale data, so fast heuristic approximations are needed. One such heuristic is `last-split` (16).

Ambiguity and alignment probabilities

Alignments have varying degrees of ambiguity. For example, suppose that our ancestral sequence has several identical copies of a retrotransposon, one of which has been retrotransposed (i.e. duplicated and inserted at a random location) in the derived sequence. While it is true that this insertion descends from a unique part of the ancestral sequence, this unique ancestral part is unknowable.

By using the Repeated Matches Algorithm’s probabilistic interpretation, we can quantify such ambiguity (16). In the above retrotransposon example, the algorithm would align the insertion to an arbitrarily-chosen copy in the ancestral sequence, and annotate these bases with low correct-alignment probabilities (high ‘mismap’ probabilities). This is useful, and the best that can be done.

MATERIALS AND METHODS

DNA read data

The MinION (R9.4) dataset (14 million reads, 90 billion bases) is from `rel3` at https://github.com/nanopore-wgs-consortium/NA12878 (Jain *et al.* https://doi.org/10.1101/128835). The PacBio (P5-C3) dataset (30 million reads, 197

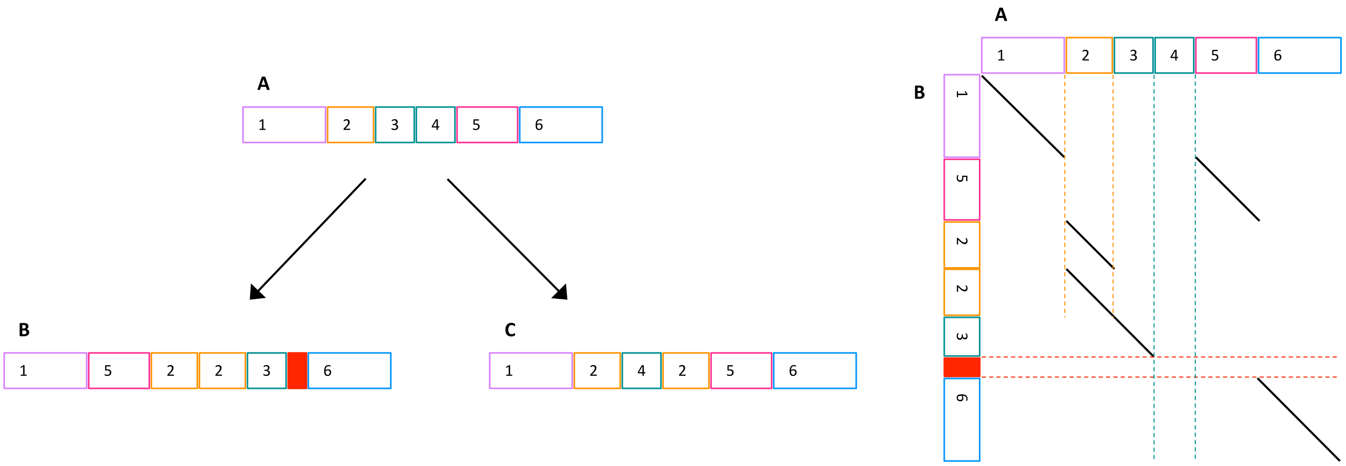

Figure 1. Left: sketch of sequence evolution by complex mutations. (**A**) An ancestral sequence; (**B**) and (**C**) are derived. Same-color blocks (e.g. 3 and 4 in A) indicate similar sequences. The solid red block is 'spontaneously generated', i.e. not descended from any ancestral sequence. Right: alignment between A and B. Dashed lines indicate: duplication (vertical orange), deletion (vertical teal), and spontaneous generation (horizontal red).

Table 1. Probabilities of substitution, deletion, and insertion

Sequence Comparison		Substitutions				Del open	Del extend	Ins open	Ins extend
		a	c	g	t				
Human R9.4 versus reference genome	a	.27	.0019	.019	.0017				
	c	.0019	.19	.00064	.0040	.067	.44	.017	.48
	g	.019	.00064	.18	.0013				
	t	.0017	.0040	.0013	.30				
		a	c	g	t				
Human P5-C3 (columns) versus reference genome (rows)	a	.29	.0035	.0010	.00082				
	c	.0024	.20	.0022	.0020	.045	.15	.073	.35
	g	.0020	.0016	.19	.0022				
	t	.0012	.0015	.0013	.30				
		a	c	g	t				
Human versus chimp	a	.29	.00070	.0025	.00066				
	c	.00070	.20	.00073	.0025	.0017	.73	.0017	.73
	g	.0025	.00073	.20	.00070				
	t	.00066	.0025	.00070	.29				
		a	c	g	t				
Human versus gorilla	a	.29	.00098	.0033	.00084				
	c	.00098	.20	.0010	.0033	.0021	.88	.0021	.88
	g	.0033	.0010	.20	.00098				
	t	.00084	.0033	.00098	.29				

billion bases) is from accession SRX627421 (11). Note that the PacBio reads are a few years older, so this may not be a fair comparison of the technologies as they are now.

Alignment of DNA reads to the `hg38` human genome

The two datasets, R9.4 and P5-C3, were analyzed separately in this and all following steps. First, the probabilities of substitution, insertion, and deletion were determined by `last-train`. Then, the reads were aligned to the genome by `last-split` (exact commands in the Supplement).

Outgroup alignments

The `hg38` human genome was aligned to genomes of chimp (17) and gorilla (18). The precise methods, and resulting alignments, are available at https://github.com/mcfrith/last-genome-alignments.

Finding rearrangements

Rearrangements were found by these steps:

1. In each read separately, identify 'local' rearrangements, i.e. intrachromosomal rearrangements encompassed by one read.
2. Cluster overlapping rearrangements (from multiple reads) into 'rearranged regions'.
3. Discard each rearranged region that is not covered by an unbroken alignment of chimp or gorilla DNA to the reference human genome. This aims to get rearrangements where the reference human genome's arrangement is ancestral. (It could fail to do so, if the derived arrangement was present in the common ancestor of these species.)

The software is available at: https://github.com/mcfrith/local-rearrangements.

Re-aligning non-orphan rearrangements

The rearrangements include many orphans present in just one read (more likely to be artifacts), and fewer non-orphans present in multiple reads (more likely to be real). In order to characterize the non-orphans more accurately, these reads were re-aligned to the genome more slowly-and-sensitively, by using `lastal` options `-m50 -d90` (exact commands in the Supplement).

Genome annotations

Annotations of the `hg38` human genome were taken from its `refFlat.txt` (19), `rmsk.txt` (20), and `simpleRepeat.txt` (21) files, in the UCSC genome database (22).

RESULTS

Sequencing error rates

The rates of substitution, insertion, and deletion relative to a reference human genome (`hg38`) were estimated with `last-train` (23). These rates include both sequencing errors and real genome differences, but greatly exceed the well-established rates of real difference, so are mostly error (Table 1).

Both R9.4 and P5-C3 have more indel than substitution errors. R9.4 has more deletions (3.9×) than insertions. P5-C3 has more insertions (4.3×) than R9.4, but they tend to be shorter. It has fewer deletions (0.7×), also shorter.

A preliminary result for R9.4 (not shown) had an almost-symmetric substitution matrix, therefore the training was redone assuming a symmetric matrix (Table 1). P5-C3 lacks this symmetry. R9.4 has higher rates of transition errors, especially `a`↔`g`.

We also estimated the rates of substitution, deletion and insertion for chimp and gorilla DNA versus human (in order to make outgroup alignments). Compared to the long-read sequencing errors, there are much fewer insertions and deletions, but they tend to be longer. The rates of transition (`a`↔`g` and `c`↔`t`) are higher than those of P5-C3, but lower than those of R9.4. The transversion rates are mostly lower than those of the long reads.

Orphan rearrangements

Because each dataset has many-fold coverage of the genome, we expect each rearrangement to be covered by multiple DNA reads. Surprisingly, however, there are many orphan rearrangements appearing in just one read.

The R9.4 dataset has 33 695 rearranged regions, of which 31 600 have just one rearranged read. Tandem duplications (TDs) are frequent: 29 339 of the orphans, and 1169 of the non-orphans, are TDs. A (pseudo)random sample of TD and non-TD orphans is shown in Figure 2. The non-TD orphans often have a zigzag pattern, implying triplication.

The P5-C3 dataset has 196 527 rearranged regions, of which 183 653 have just one rearranged read. TDs are again frequent: 154 487 orphans (and 3143 non-orphans) are TDs. A sample of these orphans is also shown in Figure 2. Each TD orphan has a very small duplication (compared to R9.4), flanking an unaligned part of the read. The non-TD orphans often have a bizarre fountain-like appearance (Figure 2M, O), with interspersed fragments from opposite strands.

Table 2. Number of rearrangements of each type

Type	Total	Found with	
		R9.4	P5-C3
Non-tandem duplication (same strand)	26	26	14
Non-tandem duplication (opposite strand)	21	21	13
Non-tandem double duplication	2	2	0
Inverted tandem duplication	6	6	5
Inversion	8	8	3
Inversion next to deletion	12	12	6
Inversion in deletion	8	8	1
Non-tandem duplication (same strand) with deletion	6	6	1
Non-tandem duplication (opposite strand) with deletion	6	6	2
Localized shattering	4	4	1
Conversion	4	4	1
Tandem multiplication	54	42	24
Unique	9	8	2
TOTAL	166	153	73
Unclear	21	2	19
Dubious	36	9	29

A plausible explanation of these orphan rearrangements is that they are artifacts of the DNA sequencing process.

Supported rearrangements

Non-orphan rearrangements were analyzed further. Most non-orphans are tandem duplications and/or inside tandem repeats: the latter is not surprising, because tandem repeats have high rates of tandem duplication and gene conversion (24). To avoid manual inspection of numerous similar rearrangements, we discarded tandem duplications (but not triplications etc.), and rearranged regions $>$ 50% covered by tandem repeats.

For the R9.4 dataset, this left 164 rearranged regions, which were manually inspected (Supplemental figures). A few were judged unclear or dubious (Table 2), usually due to doubt that the reference human genome's arrangement is ancestral. (Gaps ≤ 50 bp were allowed in the ape-human alignments, and sometimes these gaps coincided suspiciously with small rearrangements.)

For the P5-C3 dataset, it is harder to separate plausible from dubious rearrangements. There are more dubious rearrangements, and the fountain-like ones are especially troublesome, because they are large (e.g. 10 kb) and often overlap other rearrangements. We therefore applied a stricter criterion, retaining only rearranged regions that have at least four reads with non-TD rearrangements. This left 121 rearranged regions (Supplemental figures), of which 48 were judged unclear or dubious (Table 2), usually because they have multiple overlapping fountain-like rearrangements.

Ultimately, this semi-manual analysis produced 166 local rearrangements judged to be clear and reliable (Table 2). One argument for their reliability is that each case has more than one read with the *same* rearrangement (Supple-

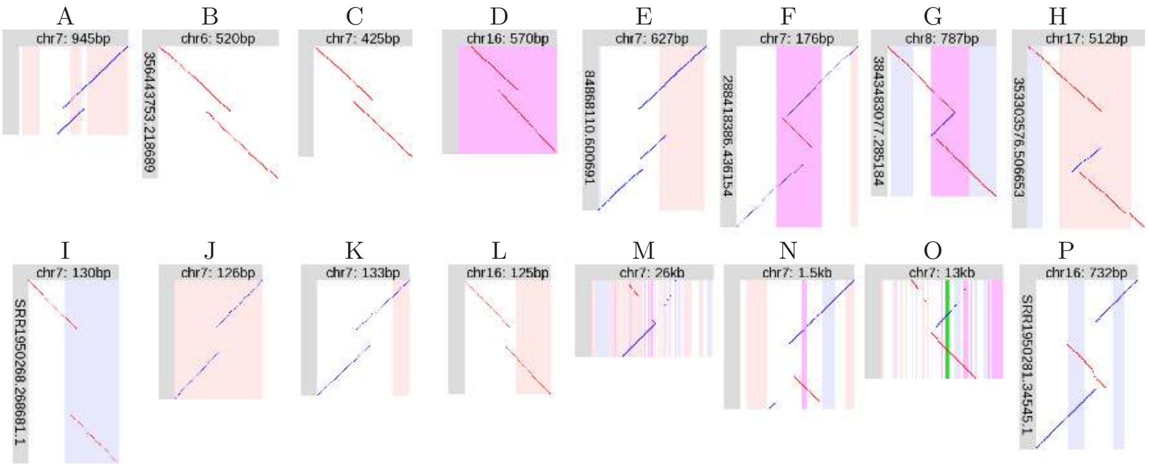

Figure 2. Orphan rearrangements. (**A**–**D**) R9.4 tandem duplications (TDs). (**E**–**H**) R9.4 non-TD rearrangements. (**I**–**L**) P5-C3 tandem duplications. (**M**–**P**) P5-C3 non-TD rearrangements. Diagonal lines indicate alignments between a segment of a DNA read (vertical), and a segment of the reference human genome (horizontal). Red lines indicate same-strand alignments; blue lines indicate opposite-strand alignments. The vertical stripes indicate features in the reference genome; pink: forward-strand transposable element, blue: reverse-strand transposable element, purple: low-complexity or tandem repeat, green: exon.

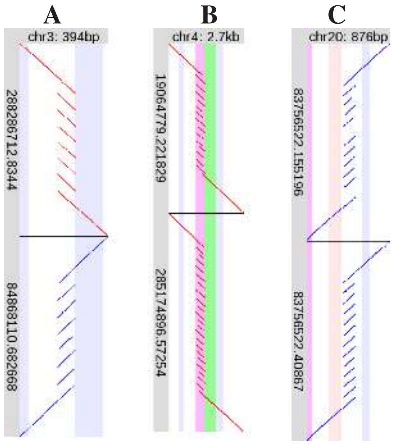

Figure 3. Examples of tandem multiplication. Please see the description of Figure 2. Here, each of (**A**), (**B**) and (**C**) shows two R9.4 DNA reads, one above the other.

mental figures). An example is shown in Figure 3A: this shows two reads, one from each DNA strand, with the same tandem heptuplication. In contrast, there are many cases where fountain-like rearrangements overlap each other, but they are different (e.g. Supplementary Figure S252, Figure S276).

Another argument for the reality of these rearrangements is that 60 of them are found by both R9.4 and P5-C3, independent sequencing technologies. In fact, the P5-C3 rearrangements are mostly a subset of the R9.4 rearrangements, suggesting that with R9.4 we find almost all local rearrangements present in this genome. The main exception is the tandem multiplication category, where P5-C3 has 12 rearrangements not found in R9.4 (Table 2). This is related to our removal of rearrangements in tandem repeats: most of the 12 would be in tandem repeats, except for one or two reads with dubious alignments that enlarge the rearranged region (e.g. Supplementary Figure S186). On the other hand, the reason for finding only a subset of the rearrangements in P5-C3 is (at least partly) the stricter criterion used to discard dubious rearrangements.

Zygosity

A further argument for the reality of these rearrangements is from zygosity. Since the DNA reads are from diploid cells (with two copies of each chromosome), if a rearrangement is real, it should be either homozygous (in both copies of the chromosome), or heterozygous (in just one copy). By examining all the DNA reads that align to a rearranged region, we can judge its zygosity.

Many of the rearranged regions have a clear zygosity. For example, the tandem multiplication in Figure 3A is heterozygous: about half the R9.4 DNA reads align straight across this region without rearrangement (Supplementary Figure S167). This is confirmed by P5-C3 reads (Supplementary Figure S168). In contrast, Figure 3B shows two different rearrangements in the same region, implying that both chromosomes are rearranged, so there should be no unrearranged DNA reads here. Indeed, there are none (Supplementary Figure S171). Some homozygous examples are: Figure 4H (Supplementary Figure S95), Figure 4L (Supplementary Figure S108), Figure 4N (Supplementary Figure S123), and Figure 4 P (Supplementary Figure S143). Some heterozygous examples are: Figure 3C (Supplementary Figure S189), Figure 4A (Supplementary Figure S4), Figure 4B (Supplementary Figure S10), and Figure 4J (Supplementary Figure S119).

In some cases the zygosity is less clear, due to imperfect and ambiguous alignments. In a few cases (Supplementary Figure S125, Supplementary Figure S131) the rearrangement is present in much less than half of the reads: these rearrangements may exist in cell subpopulations.

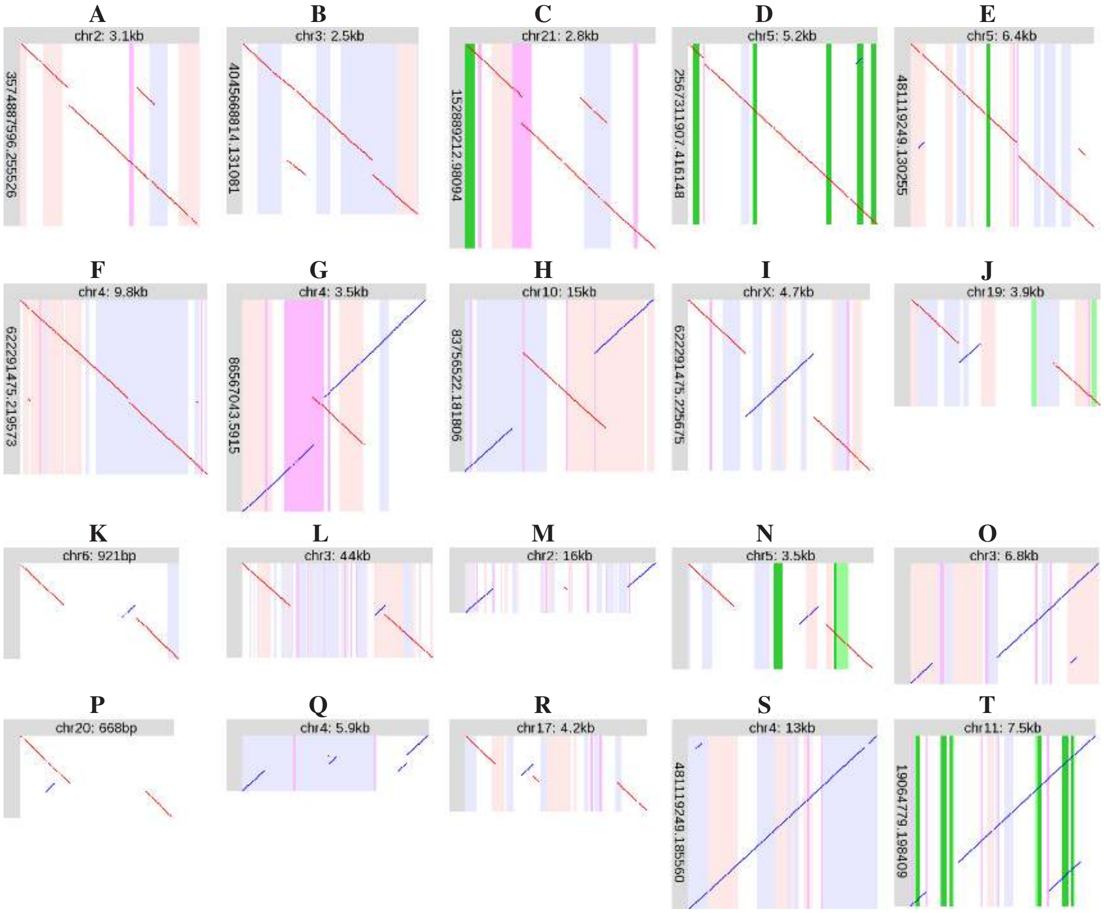

Figure 4. Examples of recurring rearrangement patterns. Diagonal lines indicate alignments between a segment of a DNA read (vertical), and a segment of the reference human genome (horizontal). Red lines indicate same-strand alignments; blue lines indicate opposite-strand alignments. The vertical stripes indicate features in the reference genome; pink: forward-strand transposable element, blue: reverse-strand transposable element, purple: low-complexity or tandem repeat, green: exon, dark green: protein-coding sequence. Some of the alignments (diagonal lines) are tiny: it may help to view this on a screen and zoom in.

Recurring patterns of rearrangement

Many of the rearrangements strikingly resemble each other, i.e. there are recurring rearrangement motifs. Some of these motifs (generally, the simpler ones) match complex-variant classes from the limited-resolution study of Collins *et al.* (3).

Tandem multiplication. The most frequent type of rearrangement is tandem multiplication (Table 2). The degree of multiplication varies greatly, e.g. duplication (most common), heptuplication (Figure 3A), and more (Figure 3B and C). The degree of multiplication is sometimes unclear, especially in tandem repeats: this is presumably due to ambiguity in alignment of tandemly-repeating sequences. The degree is often heterozygous, e.g. Figure 3B shows a smaller allele with ∼14-fold multiplication and a larger allele with ∼22-fold multiplication.

The tandem multiplication in Figure 3B is immediately upstream of the transcription start site of LINC01596 (long intergenic non-protein coding RNA 1596), so is likely to have a profound effect on the expression of this transcript.

Figure 3C illustrates a problem that occurs with all types of rearrangement, but especially tandem multiplications. While the lower DNA read shows 12-fold tandem multiplication, the upper read has two missing fragments: these two parts of the read are unaligned. A plausible explanation is that the alignment method simply failed to align these fragments, perhaps because of sequencing errors. Such missing fragments can often be rescued by re-aligning the DNA read with more slow-and-sensitive heuristics.

Non-tandem duplication. Another frequent pattern is non-tandem, but localized, duplication (Table 2). The duplicated segment can be inserted in its original orientation (Figure 4A–C), or in the opposite orientation (Figure 4D). These rearrangement patterns were reported previously (11).

There are two cases of double non-tandem duplication (Figure 4E–F). In both cases, two non-adjacent segments have been duplicated, and inserted adjacent to each other roughly half-way between the duplication sources.

Inverted tandem duplication. Another recurring pattern is inverted tandem duplication (Figure 4G). This superficially resembles the zigzag pattern of orphan rearrangements (Figure 2F–H), but is duplication rather than triplication.

There are some near-tandem inverted duplications (e.g. Supplementary Figure S65), so perhaps there is no sharp distinction between non-tandem opposite-orientation duplications (Figure 4D) and the present category.

Inversion. Inversions can occur without deletion (Figure 4I), but more often occur next to (Figure 4J–L) or in/between (Figure 4M–N) deletions. This fits with previous reports that inversions are often flanked by gaps (25–27).

The example in Figure 4N deletes the second-last protein-coding exon of SPINK14 (serine peptidase inhibitor, Kazal type 14) (11). Examination of all reads aligned here (Supplementary Figure S123) indicates that this rearrangement is homozygous.

The inversion in Figure 4H is flanked by oppositely-oriented L1PA7 LINEs, suggesting that it was caused by homologous recombination between these LINEs. Figure 4H shows a small deletion to the left of the inversion, and a small duplication to the right: this is gene conversion (replacement of a sequence by a duplicate of a similar sequence), which is a typical effect of homologous recombination.

Non-tandem duplication with deletion. Another recurring pattern is non-tandem (but localized) duplication, combined with deletion at the duplication target site. Here too, the duplicated segment can be inserted in its original orientation (Figure 4O), or in the opposite orientation (Figure 4P).

Localized shattering. A more complex rearrangement motif is 'localized shattering' (Figure 4Q–R). This is so named because it looks like the ancestral sequence was shattered into multiple fragments, of which some were lost, and the others rejoined in random order and orientation. Accordingly, this motif exhibits deletion and repositioning but not duplication. Another diagnostic feature is that some of the rearranged fragments come from adjacent parts of the ancestral sequence. If the fragments came from randomly and independently chosen parts of the ancestral sequence, it is unlikely that they would be adjacent. Shattering naturally explains this adjacency.

All of these diagnostic features are also present in the inversion-next-to-deletion pattern, which might thus be regarded as a special case of localized shattering.

Conversion. Another well-known pattern of DNA sequence change is gene conversion, where one sequence tract is replaced by a duplicate of a similar tract (Figure 4S–T).

The example in Figure 4T involves protein-coding genes. The three green stripes on the left indicate the three exons of HBG1 (hemoglobin subunit gamma 1), and those on the right indicate HBG2 (hemoglobin subunit gamma 2). Here, the first two exons of HBG1 have been replaced by the equivalent exons of HBG2. Because HBG1 and HBG2 are very similar, this changes just one amino acid: a threonine is converted to an isoleucine. This substitution is present in dbSNP (rs1061234).

At the DNA level, the segments of HBG1 and HBG2 that participate in this conversion differ by just one length-4 indel and a dozen substitutions. (The exact number of substitutions depends on the exact endpoints of the conversion, which are highly ambiguous.) Thus, these DNA reads could all-too-easily be aligned to the HBG locus co-linearly, without rearrangement. This example showcases the power of probability-based alignment, which calculates that the rearranged alignment is much more probable. The most informative alignment part (of one read) is shown in Figure 5: a co-linear alignment would align this part of the read to HBG1.

```
read    gaaacGAATaacaaaaatAG---aaagGtgaaagttctttctacCgA
HBG2    aaaacGAATaacaaaaatAGgggaaagGtgaaagttctttctacCgA
HBG1    aaaac----aacaaaaatGAgggaaagAtgaaagttctttctacTgG
```

Figure 5. A segment of a DNA read aligned to HBG1 and HBG2. It is likely that this part of the read is paralogous to HBG1, and should rather be aligned to HBG2.

Unique rearrangements

Some rearrangements were judged to be unique, and not follow any recurring pattern. The eight cases found in the R9.4 dataset are described here.

Inverted tandem duplication with deletion. Figure 6A shows an inverted tandem duplication with adjacent deletion. The duplication and deletion have roughly the same size, suggesting gene conversion. However, the duplicated and deleted sequences lack similarity, which argues against conversion.

Non-tandem duplication next to deletion. Figure 6B shows a non-tandem duplication, combined with a deletion adjacent to the duplication *source*.

Non-tandem duplication with Alu deletion. Figure 6C shows a non-tandem duplication, and near the duplication target site is a precise (or near-precise) deletion of an AluY element. Perhaps this deletion occured separately and independently of the duplication. Precise deletion of transposable elements is unusual (28): maybe the reference human genome's condition is not ancestral, even though it is shared by chimp and gorilla (Supplementary Figure S159).

Inversion with tandem repeat expansion. Figure 6D shows an inversion that overlaps a period-45 tandem repeat (purple stripe). The tandemly repeated region has been partly duplicated. Since tandem repeats frequently expand, perhaps this inversion and expansion occurred independently.

Inversion with deletion and duplication. Figure 6E superficially resembles inversion and gene conversion by homologous recombination (Figure 4H). However, the duplicated and deleted regions are dissimilar and of different lengths (the deletion is ∼600 bp and the duplicated region is ∼930 bp), which argues against conversion and recombination. Therefore, we suggest this rearrangement should be classified separately. Though it is unique in our study, this delINVdup pattern has been seen in other human DNA (29).

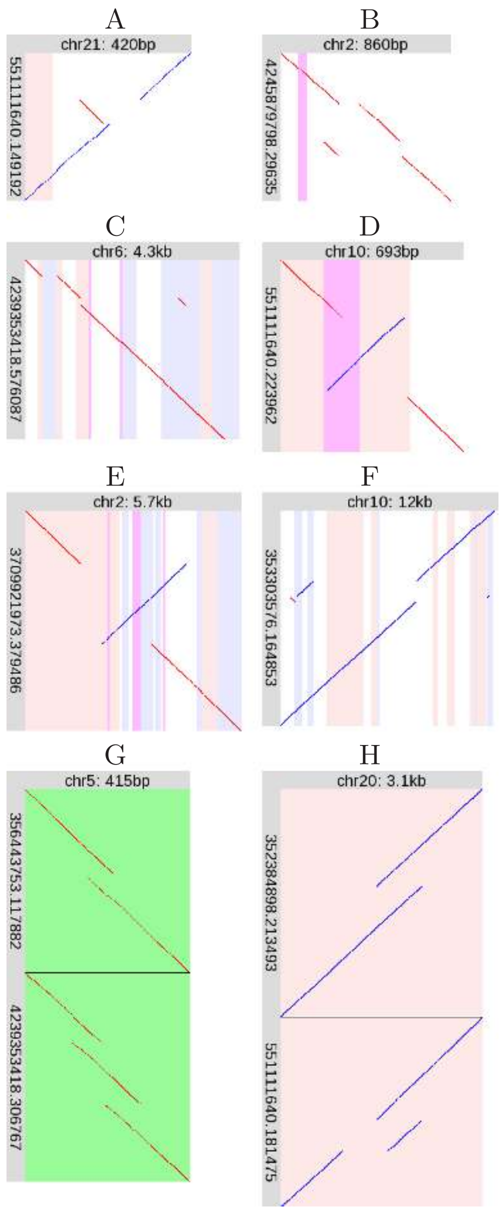

Figure 6. Unique rearrangements. Please see the description of Figure 4.

Multiple non-tandem duplication with conversion. Figure 6F shows non-tandem duplication, with three duplicated fragments. The genomic sources of the two left-hand fragments are separated by 116 bp, and the right-hand fragment's source is 114 bp. Moreover, the left-hand separation and right-hand fragment are in equivalent regions of Alu elements, suggesting that conversion contributed to this complex rearrangement.

Heterozygous adjacent tandem duplications. Figure 6G shows two (heterozygous) rearrangements (relative to the reference genome) at the same location, in the 3′UTR of AHRR (aryl-hydrocarbon receptor repressor). One is a tandem duplication, the other is two adjacent tandem duplications: the latter pattern is unique in this study.

Tandem duplication next to heterozygous deletion. Figure 6H shows two (heterozygous) rearrangements at the same location, within centromeric alpha satellite DNA. It appears that both rearrangements share the same tandem duplication, but one has an adjacent deletion.

Repeat associations

Many of the rearrangements are strikingly associated with repeats. Firstly, tandem multiplications often overlap tandem repeats (e.g. Figure 3B), as expected.

Secondly, some inversions and conversions are attributable to homologous recombination between repeats, also as expected. Figure 4S shows conversion between L1PA3 LINEs. In Figure 4J, the inverted segment is near-precisely flanked by oppositely-oriented AluSx SINEs, though this does not explain the deletion.

Thirdly, the target sites of some non-tandem duplications coincide with simple repeats (Figure 4C–E). An explanation may lie in the fact that simple repeats can be fragile (24).

Fourthly, rearrangement edges often coincide with edges of transposon fragments. In Figure 4G, the duplication's right end coincides with the 3′-end of an L1 LINE fragment (that lacks the 3′-most 5729 bases of a full-length L1). In Figure 4N, the rightmost deletion edge is ∼6 bp downstream of the 5′-end of a (5′-truncated) MIR. The inverted fragment's right edge is ∼20 bp downstream of the 3′-end of a (3′-truncated) MamSINE1. In Figure 4L, the deletion's left edge is ∼4 bp upstream of the 3′-end of a MIR (that lacks the 3′-most 126 bp of a full-length MIR). In Figure 4M, the leftmost deletion edge is ∼3 bp upstream of the 5′-end of an L1 (that lacks the first 6216 bp of a full-length L1). The rightmost edge is ∼30 bp upstream of the 5′-end of an L2 (that lacks the first 3303 bp of a full-length L2). There are similar coincidences in Figures 4E, O, R and 6D.

Some tandem multiplications also coincide with edges of transposon fragments. The multiplication in Figure 3A is next to the 3′-end of a (3′-truncated) MLT1H2 LTR. The multiplication in Figure 3C sits neatly between a MER103C transposon (left) and an L2 LINE (right).

The duplication source in Figure 4C near-perfectly overlaps an MLT1B LTR, suggesting retrotransposition of this element. This LTR has many mutations relative to the MLT1B consensus sequence, which enables unambiguous alignment, but suggests that it has lost retrotranspositional activity.

Retrogene reduplication

The non-tandem duplication in Figure 4B perhaps warrants further description. The duplication source is in a white region flanked by blue stripes. This white region is a retrotransposed copy of RPL23A (ribosomal protein L23a) (30), which was inserted in a Charlie5 transposon (blue stripes). Thus, the rearrangement in Figure 4B is a reduplication of most of this retrogene.

There are many retrotransposed RPL23A copies in the genome (30), so we may wonder if this alignment is correct and unambiguous. This retrogene (the white region in Figure 4B) has 90.2% identity to its parent mRNA (30). The inserted part of the DNA read in Figure 4B aligns here with

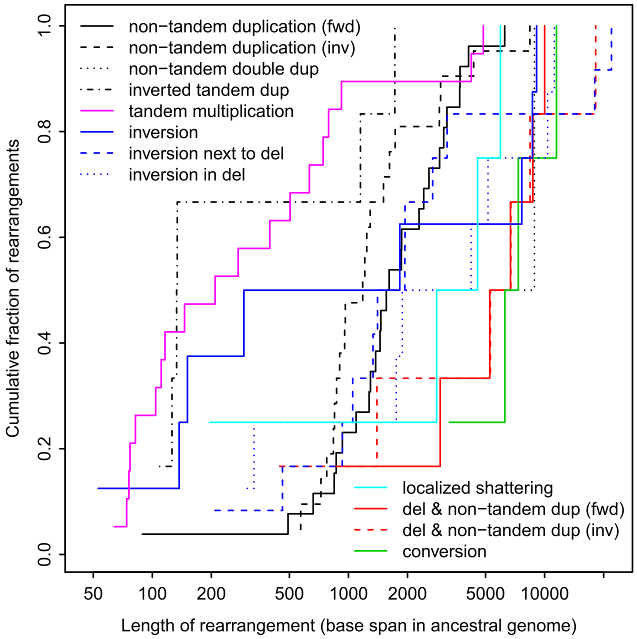

Figure 7. Lengths of rearrangements in the R9.4 dataset.

192 matches, 15 mismatches and 64 gaps. This alignment has much higher probability than the second-best hit (187 matches, 20 mismatches, 82 gaps), to a retro-RPL23A on chromosome 19 with 97.7% identity to the parent mRNA. The insertion aligns to the parent mRNA with 186 matches, 21 mismatches and 85 gaps, suggesting that it is not a direct retrotransposition of the parent gene.

Rearrangement lengths

The rearrangement lengths (span in ancestral genome) vary between 50 and 20 000 bases (Figure 7). The upper bound is limited by the read lengths. The lower bound presumably reflects the difficulty of aligning short, error-prone sequence fragments: shorter rearrangements exist (27). Since the figure shows *ancestral* lengths, it is not surprising that tandem multiplications tend to be shorter, and rearrangements involving deletion tend to be longer. The conversions are longer-range than most of the other rearrangements (but there are just four of them).

Genomic context of rearrangements

The rearrangements occur throughout the genome, with some clumping, e.g. a conversion near a non-tandem duplication in chromosome 9 (Supplementary Figure S1) (31). Most of the tandem multiplications are near telomeres, regions known to be rearrangement hot spots (8). Tandem multiplications tend to occur in fast-evolving sequence (Supplementary Figure S2) (32) with above-average GC content (Supplementary Figure S3), perhaps because they tend to overlap tandem repeats.

Eight rearrangements overlap large recently-duplicated segments ('segmental duplications') (Supplementary Table S1) (33): 2 conversions, 2 tandem multiplications, 3 non-tandem duplications (same strand) and 1 inversion. The association with conversions is not surprising, because conversion occurs between similar sequences. One tandem multiplication (Supplementary Figure S192) is an expansion of tandemly-duplicated segments.

Comparison to previously-found inversions

The P5-C3 reads are from a previous study (11), which published a table of 40 inversions. Fourteen of them match our rearrangements: 2× inversion, 6× inversion next to deletion, 3× inversion in deletion, 2× unique (Figure 6D–E), and 1× 'dubious' (reference human genome not ancestral). So only a minority of each inversion catalog is common to both.

Seventeen of our inversions were not found in the previous study. The reason is unclear; however, among the 17 are our six longest (>8 kb) and three shortest (<200 bp) inversions.

We miss 26 of the previous inversions, for clear reasons. In 23 cases, the reference human genome lacks an unbroken alignment to chimp or gorilla, thus failing our ancestrality criterion. One other case could not be transferred between genome versions. (The previous study used an older reference human genome.) The remaining two cases are not 'local' rearrangements. One of them (in chromosome 1q25) is basically an inversion, but a fragment of the DNA read next to the inversion aligns to a distal location. The other (in chromosome 9q34) does not look like an inversion: it seems to be an interchromosomal rearrangement involving chromosome 22.

Since we miss most of the previously-found inversions due to our ancestrality requirement, the reader may wonder whether this requirement is too strict. For inversions without deletion (Figure 4I), it is indeed too strict: the alignments reliably indicate such inversions regardless of whether the reference genome's arrangement is ancestral. For inversions with deletion (Figure 4J–N), however, ancestrality is important. Suppose the reference genome has an inversion with deletion, relative to the DNA reads. The regions of DNA reads that have been deleted in the genome should be unaligned, but if any part of the deleted sequence has a sufficiently-similar paralog anywhere else in the genome, the reads will be wrongly aligned to this paralog. Thus, while the ancestrality requirement could be relaxed for some types of rearrangement, in general it is necessary.

Comparison to the database of genomic variants

Finally, we compared the R9.4 rearrangements to structural variants in the Database of Genomic Variants (DGV release: 31 August 2016) (34). The boundaries of both our rearranged regions and DGV variants are not necessarily precise, so we sought overlaps. About 100 R9.4 rearranged regions (65%) roughly coincide with DGV variants (Figure 8). By this criterion, DGV lacks most of our tandem multiplications (30/42) and inversions (6/8).

Many R9.4 rearranged regions coincide near-exactly with DGV variants (Figure 8), but the descriptions often differ. For instance, the double duplication in Figure 4F coincides (intersection over union = 0.986) with variant esv3324, a

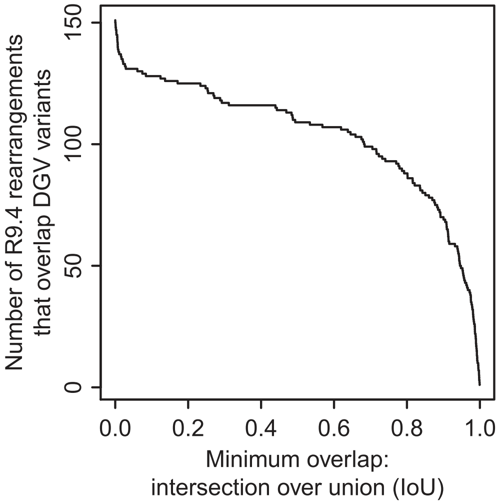

Figure 8. Number of R9.4 rearrangements that overlap variants in DGV.

'CNV loss'. This variant was found from short DNA reads (35): it looks as though they correctly found the adjacency of the two duplicated fragments, but incorrectly interpreted this as deletion of the interval between the two duplication sources. This illustrates the advantage of longer reads, which reveal the whole rearrangement.

Diverse R9.4 rearrangements coincide with DGV 'inversions', including Figure 4D (IoU 0.999), Figure 4E (IoU 0.996), Figure 4L (IoU 0.999), and Figure 6F (IoU 0.99). These do all involve inverted fragments, but to describe, e.g. Figure 6F as 'inversion' is at best incomplete and misleading. Short reads can show that neighboring parts of a sample genome map to opposite strands of a reference genome, but this is a feature of diverse types of rearrangement.

Further examples include Figure 4M, which coincides (IoU 0.995) with a 'CNV loss', a correct but incomplete description. The conversions in Figure 4S (IoU 0.938) and Figure 4T (IoU 0.956) coincide with 'CNV deletions'. Figure 4B (IoU 0.981) and Figure 6B (IoU 0.9) coincide with 'tandem duplications'. The localized shattering in Figure 4Q coincides (IoU 0.943) with a 'CNV gain+loss', which is consistent but vague.

In summary, our long read analysis clarifies previously-reported variants.

DISCUSSION

Long DNA reads have many dubious rearrangements

Both datasets (R9.4 and P5-C3) have many orphan rearrangements, that occur in just one read and are not supported by other reads. These might be artifacts of the sequencing process, or real rearrangements present in a fraction of the cells (e.g. dead or dying cells). That the frequencies and types of orphan rearrangements differ between the two datasets argues that many are artifacts. For R9.4, we can distinguish plausible from dubious rearrangements provided we have enough sequencing coverage that real rearrangements are expected to be represented in more than one read. For P5-C3, it was harder to distinguish plausible from dubious rearrangements, especially because of the dubious fountain-like rearrangements, which are large and often overlap other rearrangements.

Some, but not all, orphan rearrangements have characteristic patterns (e.g. zigzag, fountain-like) not seen in the supported rearrangements. So, it may be possible to automatically recognize and discard them. This could be a moving target, however, if artifact patterns change with different versions of sequencing technology.

Another approach would be to automatically recognize if more than one DNA read has the same rearrangement, not merely overlapping rearrangements. This is not trivial, because two reads with the same rearrangement can have non-identical alignments due to sequencing error (e.g. Figure 3C).

Rearrangement patterns

Nearly all the supported rearrangements belong to recurring patterns/motifs. One pattern can manifest at a wide range of scales, e.g. hundreds of bp to tens of kb (Figure 4J–L).

The most common type of local rearrangement is tandem multiplication (especially duplication). This is very common in tandem repeats, as expected. Non-tandem duplications are the second-most frequent, but more complex patterns occur too, e.g. double duplication (Figure 4E–F), localized shattering (Figure 4Q–R), and a fascinating unique rearrangement (Figure 6F). We have established here that such complex rearrangements can be detected in long DNA reads.

Gene conversions were also identified: although this is a classically-known phenomenon, it has surely been underreported in high-throughput sequencing studies, due to the ambiguity in aligning these sequences. Probability-based alignment is a powerful way to resolve this kind of ambiguity.

Some patterns that might be expected do not occur. In particular, there are no cases of translocation (movement of a DNA fragment), without deletion or duplication.

None of the supported rearrangements have large unaligned 'insertions' in the DNA reads (apart from cases like Figure 3C, which are likely alignment failures rather than true insertions). This fits the hypothesis that most sequence is descended from ancestral sequence, not spontaneously generated; it also supports the reality of these rearrangements. Thus the term 'insertion' is potentially misleading: it is not symmetric with 'deletion', and is usually duplication (or deletion in the reference). On the other hand, orphan rearrangements often have unaligned parts of the reads (Figure 2).

Possible rearrangement mechanisms

A classic rearrangement-causing mechanism is homologous recombination. Some of the rearrangements are typical examples: conversion (Figure 4S–T) and inversion between oppositely-oriented repeats (Figure 4H). Other rearrangements are harder to explain. The inversion in Figure 4J lies between oppositely-oriented AluSx elements, suggesting recombination, but it is unclear how the adjacent deletion arose. The inversion in Figure 6E overlaps oppositely-oriented Alus, but perhaps that is a coincidence: this inversion may be better explained by template switching during DNA replication.

It has been suggested that template switching by DNA polymerase, during DNA replication, is a cause of complex rearrangements. Template switching may be triggered by a broken replication fork, in a mechanism termed MMBIR (microhomology-mediated break-induced replication) (1), or a stalled replication fork, in a mechanism termed FoSTeS (fork stalling and template switching) (36). The suggestion is that multiple template switches can occur, producing rearrangements with any number of duplications, deletions, and repositionings (1).

Some of the rearrangements are best explained by shattering of DNA into fragments, which then rejoined in random order and orientation, with some fragments lost (e.g. Figure 4Q–R). The main reason to favor this explanation over template switching is that some of the rearranged fragments come from adjacent parts of the ancestral sequence: we know of no reason why template switching would exhibit this tendency. Localized shattering and rejoining can be caused by radiation, such as cosmic rays and alpha particles from radon gas (37–40).

The rearrangements are often associated with repeats. In some cases, this is expected: tandem repeats have high rates of duplication and conversion, similar repeats are prone to recombination, and simple repeats can be fragile. It is less clear, however, why rearrangement edges are often close to the edges of transposon fragments. The edge of a transposon fragment is necessarily the edge of a previous rearrangement or deletion which created that fragment: so maybe these sites are fragile.

Chromothripsis

There is great interest in chromothripsis: massive genomic rearrangement acquired in a single event (5). Several possible causes have been suggested, including radiation (2,5). Our alignment method should work well for chromothriptic DNA reads, because the reference genome is likely to be approximately ancestral. Inferring chromothripsis from these alignments, if the event is not encompassed by one read, is a non-trivial problem addressed elsewhere (15). We do detect 'localized shattering': it is unclear to us whether this is distinct from chromothripsis.

Aligning rearranged sequences

Although the Repeated Matches Algorithm was described decades ago (and `last-split` has been publicly available since 2013), its suitability for elucidating rearrangements has not been recognized. This algorithm simultaneously optimizes the division of a query sequence into (any number of) parts, and the alignment of each part to reference sequences, based on probabilities of substitution, insertion, and deletion. Here, we have demonstrated that it can find intricate rearrangements in long DNA reads, and handle ambiguous situations like gene conversion.

We have emphasized the importance of an ancestral reference sequence. The main problems with a non-ancestral reference are: reference-specific deletions and duplications. If there is a reference-specific deletion, the corresponding region of the query sequence should not be aligned, but if any part of it has a sufficiently similar paralog elsewhere in the reference, it will be aligned. Here, 'sufficiently' similar is determined by the probabilities of substitution, insertion, and deletion used for alignment (Table 1). Thus, lower sequencing error rates would reduce this problem.

Our approach cannot align fragments shorter than some minimum length (which depends on the genome size and sequencing error rates). For example, it will not align a 12 bp fragment to a human genome, because such an alignment is too insignificant to distinguish from chance matches. However, it is possible to detect inversions <12 bp (27). The Repeated Matches Algorithm cannot do this because it assumes that rearranged fragments can be anywhere in the genome with uniform probability. Perhaps it could be modified to favor aligning nearby parts of a query sequence to nearby parts of the reference (41).

`last-split` is a useful approximation to the Repeated Matches Algorithm, but better approximations (in terms of accuracy and speed) are surely possible. One idea is to run the full Repeated Matches Algorithm on complex rearrangements tentatively identified by `last-split`.

Ancestral genomes are ideal reference genomes

If we could reconstruct the most recent common ancestor of all extant human genomes, it would be an ideal reference for human nucleotide sequences. Almost every part of any human DNA read would be descended from a unique part of the reference. This would solve the worst problems with currently-used reference genomes.

Identifying rearrangements from alignments

Although the alignment procedure can handle arbitrary rearrangements, this study only examined 'local' rearrangements. A local rearrangement has a tightly-defined location: a start and an end coordinate in one chromosome. This property was used to: gather overlapping rearrangements in different DNA reads, check these regions' alignments to chimp and gorilla, draw pictures of these regions, and compare them to DGV.

If we consider non-local rearrangements, even the concept of a discrete rearrangement becomes unclear. Rearrangement *events* are discrete, but they may interact to produce a complex pattern of sequence change. For example, two overlapping inversions produce a combined rearrangement that does not resemble two distinct rearrangements. We would ideally like to infer the history of rearrangement events from the final sequence, but this is a deep problem, especially if one event can be complex (e.g. shattering).

Our analysis pipeline is semi-manual: it makes pictures of rearranged regions for human interpretation. This is appropriate for exploring what kinds of (real and artifactual) rearrangements exist.

It would be convenient to automatically infer one or two rearranged alleles in each region. This is not trivial, because different DNA reads that cover the same rearrangement can have different alignments (e.g. Figure 3C). This problem would be reduced with lower rates of sequencing error, because the different DNA reads would align more consistently (and presumably more accurately).

An alternative approach is to first perform reference-free assembly of the reads, then seek rearrangements in the assembly relative to a reference genome (42). It is not obvious which approach is better. Comparing unassembled reads to the reference genome is more direct: there is no question of assembly error, and we can examine the alignments of the original reads.

Prospects with improved sequencing accuracy

Improvements in sequencing accuracy would lead to: faster and more accurate alignment, more complete characterization of intricate rearrangements (e.g. Figure 3C) with lower coverage, and ability to find shorter rearranged fragments (up to a point). It would be easier to judge if two reads have the same rearrangement, thus filter artifacts. There would be less misalignment to paralogs when orthologs are missing from the reference.

CONCLUSIONS

This study demonstrates an approach to find arbitrarily-complex (but localized) sequence rearrangements in long DNA reads. Its key ideas are to exploit the asymmetric relationship between ancestral and derived sequences, and the use of probabilities to resolve ambiguities due to repeats and paralogs. The long reads studied here have many dubious (likely-artifactual) rearrangements, but it is possible to discriminate them from likely-real rearrangements. Almost all the real rearrangements belong to recurring motifs. Although simpler rearrangements are more common, complex and intricate rearrangements can be found: including localized shattering, which resembles radiation damage. Such rearrangements may be powerful homoplasy-free phylogenetic markers. It will be fascinating to survey many more genomes, from healthy and diseased cells, in a variety of organisms.

AVAILABILITY

Methods and alignments are available at https://github.com/mcfrith/last-genome-alignments. The software is available at: https://github.com/mcfrith/local-rearrangements.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Risa Kawaguchi for suggestions on displaying rearrangements, and Anish Shrestha for comments on the manuscript.

FUNDING

Japan Society for the Promotion of Science [KAKENHI 26700030]. Funding for open access charge: Japan Society for the Promotion of Science [KAKENHI 26700030].

Conflict of interest statement. None declared.

REFERENCES

1. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
2. Zhang,C.Z., Leibowitz,M.L. and Pellman,D. (2013) Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev.*, **27**, 2513–2530.
3. Collins,R.L., Brand,H., Redin,C.E., Hanscom,C., Antolik,C., Stone,M.R., Glessner,J.T., Mason,T., Pregno,G. and Dorrani,N.et al. (2017) Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.*, **18**, 36.
4. Greer,S.U., Nadauld,L.D., Lau,B.T., Chen,J., Wood-Bouwens,C., Ford,J.M., Kuo,C.J. and Ji,H.P. (2017) Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med*, **9**, 57.
5. Stephens,P.J., Greenman,C.D., Fu,B., Yang,F., Bignell,G.R., Mudie,L.J., Pleasance,E.D., Lau,K.W., Beare,D., Stebbings,L.A. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.
6. Feuk,L., Carson,A. R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat.Rev. Genet.*, **7**, 85–97.
7. Treangen,T.J. and Salzberg,S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat.Rev. Genet.*, **13**, 36–46.
8. Linardopoulou,E.V., Williams,E.M., Fan,Y., Friedman,C., Young,J.M. and Trask,B.J. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, **437**, 94–100.
9. Chaisson,M.J., Huddleston,J., Dennis,M.Y., Sudmant,P.H., 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.
10. Merker,J.D., Wenger,A.M., Sneddon,T., Grove,M., Zappala,Z., Fresard,L., Waggott,D., Utiramerur,S., Hou,Y., Smith,K.S. *et al.* (2017) Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.*, doi:10.1038/gim.2017.86.
11. Pendleton,M., Sebra,R., Pang,A.W., Ummat,A., Franzen,O., Rausch,T., Stutz,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.
12. Ritz,A., Bashir,A., Sindi,S., Hsu,D., Hajirasouliha,I. and Raphael,B.J. (2014) Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, **30**, 3458–3466.
13. English,A.C., Salerno,W.J. and Reid,J.G. (2014) PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, **15**, 180.
14. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge.
15. Cretu Stancu,M., van Roosmalen,M.J., Renkens,I., Nieboer,M.M., Middelkamp,S., de Ligt,J., Pregno,G., Giachino,D., Mandrile,G., Espejo Valle-Inclan,J. *et al.* (2017) Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.*, **8**, 1326.
16. Frith,M.C. and Kawaguchi,R. (2015) Split-alignment of genomes finds orthologies more accurately. *Genome Biol.*, **16**, 106.
17. The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
18. Gordon,D., Huddleston,J., Chaisson,M.J., Hill,C.M., Kronenberg,Z.N., Munson,K.M., Malig,M., Raja,A., Fiddes,I., Hillier,L.W. *et al.* (2016) Long-read sequence assembly of the gorilla genome. *Science*, **352**, aae0344.
19. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.

20. Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.
21. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
22. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
23. Hamada,M., Ono,Y., Asai,K. and Frith,M.C. (2017) Training alignment parameters for arbitrary sequencers with LAST-TRAIN. *Bioinformatics*, **33**, 926–928.
24. Richard,G.F., Kerrest,A. and Dujon,B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.*, **72**, 686–727.
25. Chaisson,M.J., Raphael,B.J. and Pevzner,P.A. (2006) Microinversions in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 19824–19829.
26. Kolb,J., Chuzhanova,N.A., Hogel,J., Vasquez,K.M., Cooper,D.N., Bacolla,A. and Kehrer-Sawatzki,H. (2009) Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome Res.*, **17**, 469–483.
27. Hou,M., Yao,P., Antonou,A. and Johns,M.A. (2011) Pico-inplace-inversions between human and chimpanzee. *Bioinformatics*, **27**, 3266–3275.
28. Ray,D.A., Xing,J., Salem,A.H. and Batzer,M.A. (2006) SINEs of a nearly perfect character. *Syst. Biol.*, **55**, 928–935.
29. Brand,H., Collins,R.L., Hanscom,C., Rosenfeld,J.A., Pillalamarri,V., Stone,M.R., Kelley,F., Mason,T., Margolin,L., Eggert,S. *et al.* (2015) Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am. J. Hum. Genet.*, **97**, 170–176.
30. Baertsch,R., Diekhans,M., Kent,W.J., Haussler,D. and Brosius,J. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics*, **9**, 466.
31. Kin,T. and Ono,Y. (2007) Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics*, **23**, 2945–2946.
32. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
33. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
34. MacDonald,J.R., Ziman,R., Yuen,R.K., Feuk,L. and Scherer,S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
35. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
36. Lee,J.A., Carvalho,C.M. and Lupski,J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, **131**, 1235–1247.
37. Moore,S., Stanley,F.K. and Goodarzi,A.A. (2014) The repair of environmentally relevant DNA double strand breaks caused by high linear energy transfer irradiation–no simple task. *DNA Repair (Amst.)*, **17**, 64–73.
38. Holley,W.R. and Chatterjee,A. (1996) Clusters of DNA damage induced by ionizing radiation: formation of short DNA fragments. I. Theoretical modeling. *Radiat. Res.*, **145**, 188–199.
39. Rydberg,B. (1996) Clusters of DNA damage induced by ionizing radiation: formation of short DNA fragments. II. Experimental detection. *Radiat. Res.*, **145**, 200–209.
40. Rothkamm,K., Kuhne,M., Jeggo,P.A. and Lobrich,M. (2001) Radiation-induced genomic rearrangements formed by nonhomologous end-joining of DNA double-strand breaks. *Cancer Res.*, **61**, 3886–3893.
41. Faust,G.G. and Hall,I.M. (2012) YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics*, **28**, 2417–2424.
42. Liu,S., Huang,S., Rao,J., Ye,W., Krogh,A., Wang,J. and The Genome Denmark Consortium. (2015) Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. *Gigascience*, **4**, 64.