Supplementary Information

The Terabase Search Engine: a large-scale relational database of short-read sequences

Richard Wilton, Sarah J. Wheelan, Alexander S. Szalay, and Steven L. Salzberg

Appendix A1. Binary run-length encoded edit (BRLEE) string format.

The goals of the BRLEE encoding are to conserve space in representing a read (Q) sequence and to make it possible to reconstruct a Q sequence given a BRLEE and a start location in a reference (R) sequence.

A total length (number of BRLEE bytes) is associated with each BRLEE string. The length may be zero, but the meaning of a zero-length BRLEE string is implementation-dependent.

Bits 6 and 7 of each BRLEE represent a BRLEE byte type; the meaning of bits 0 through 5 depends on the byte type:

bits	value	description	bits 0..5
6..7	00	match	run length
	01	gap in Q (deletion from R)	run length
	10	mismatch	bits 0..2: symbol 1; bits 3..5: symbol 2
	11	gap in R (insertion into R)	bits 0..2: symbol 1; bits 3..5: symbol 2
0..5		(according to bits 6..7)	

In this way, the BRLEE format uses run lengths to record matching symbols and deletions in Q (that is, symbols skipped in R). It uses 3-bit fields to record mismatched symbols and insertions.

Run length format. Run length is accumulated by concatenating the 6-bit values (bits 0..5) from successive BRLEE bytes in big-endian order. Run-length accumulation is terminated either by a byte with a different type (bits 6..7) or by the end of the BRLEE string.

Symbol format. Symbols are represented in a BRLEE as 3-bit values:

```
000 (null)
001 (null)
010 (null)
011 N
100 A
101 C
110 G
111 T
```

There is no length or count associated with a series of symbols. Instead, the end of the series is indicated either with a null symbol value (e.g., 000) or by encountering a different BRLEE byte type.

Unmapped reads. For reads not associated with a reference-sequence location, the BRLEE string encodes the entire Q sequence as a series of 2-bit or 3-bit symbols. The 3-bit representation above is used when the sequence contains one or more Ns. When an unmapped Q sequence contains only A, C, G, and T, two bits per symbol are used:

```
00 A
01 C
10 G
11 T
```

The BRLEE string thus consists of the bytes that contain the bitmapped symbols, preceded by a single byte that indicates whether a 2-bit or 3-bit representation is used. Since a mapping cannot start with a gap in the query sequence, the first BRLEE byte is formatted as a "gap in Q" byte type:

bits	value	description
6..7	01	("gap in Q")
5..5	0	2 bits/symbol (4 symbols/byte)
	1	3 bits/symbol (2 symbols/byte)
3..4	00	
0..2		number of symbols in the final BRLEE byte

For 2-bit encoded symbols, each bitmap byte has the following format:

bits	
6..7	symbol 4
4..5	symbol 3
2..3	symbol 2
0..1	symbol 1

For 3-bit encoded symbols, the bitmap is:

bits	
6..7	00
3..5	symbol 2
0..2	symbol 1

Examples. The first three examples illustrate how a BRLEE is constructed as a string of edit differences between Q and R sequences. The final example represents an unmapped Q string encoded using two bits per symbol:

len	binary	hex	description
1	00 010000	10	16 matching symbols
2	00 000001 00 100100	01 24	100 matching symbols
4	00 001010 10 110101 00 000001 00 100010	0A B5 01 22	10 matches, 2 mismatches (CG), 98 matches
4	01 000011 10011100 00010111 00010111	43 9C 17 17	ATCGTCCATCC

Appendix A2. Binary run-length encoded quality (BRLEQ) string format.

The goal of the BRLEQ encoding is to conserve space in representing Phred-formatted base quality scores for a read (Q) sequence.

Base quality score (BQS) quantization. The BRLEQ format records quantized base quality scores using 3-bit (8-bin) resolution. Empirically, this degree of quantization has been demonstrated to preserve all or nearly all of the information available in the raw (6-bit) BQSs. That is, when compared with raw BQS values, BQS values reconstructed from 3-bit quantized values do not introduce significant inaccuracies when they are used by variant callers or other analysis tools.

The BRLEQ binning algorithm operates by placing each of the BQS values for a given read into one of eight bins. The mapping of BQS values to bins is thus determined dynamically for each read. This tends to preserve an even distribution of BQS values to bins even for reads with a small number of different BQS values or a narrow range of BQS values.

Binned BQS values in a BRLEQ string may be represented as either 2-bit or 3-bit values. Bin numbers are assigned so that the lowest (2-bit) values represent higher BQS; this heuristic favors a more compact representation because higher BQS are more common than lower BQS in raw sequencer data.

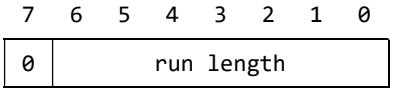
BRLEQ encoding. A total length (number of BRLEQ bytes) is associated with each BRLEQ. The length may be zero, but the meaning of a zero-length BRLEQ string is undefined.

There is no end-of-string delimiter. To decompress a BRLEQ string, all three of the following must be known:

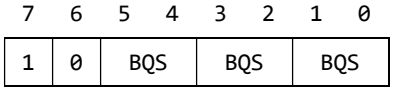
- minimum BQS (in the original BQS string)
- maximum BQS (in the original BQS string)
- string length of the original BQS string (same as the length of the associated read sequence)

There is no length or count associated with a series of bin values. Instead, every BRLEQ byte that contains bin values is assumed to contain either two 3-bit values or three 2-bit values. When the last byte in a BRLEQ contains bin values, the total number of BQSs in the reconstructed binary string is assumed to be the same as the length of the corresponding read sequence.

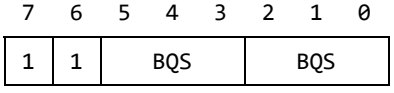
BRLEQ byte types. Bits 6 and 7 of each BRLEQ byte indicate the byte type, which determines the meaning of the remaining bits in the byte. Bit 7 determines whether a byte contains a run length or BQS bin values; when bit 7 is set, bit 6 determines the number of bits in each bin value:



(run length)



(three 2-bit values)



(two 3-bit values)

The largest run length that can fit into a single byte is 127 (7 bits). For longer runs, the bits are concatenated from successive bytes, with the high-order bits of the run length derived from the first byte encountered.

The bin value for a run is determined by the most recently encountered previous bin value. (Every BRLEQ must therefore start with a set of explicit bin values.)

BQS bin values are stored in little-endian order.

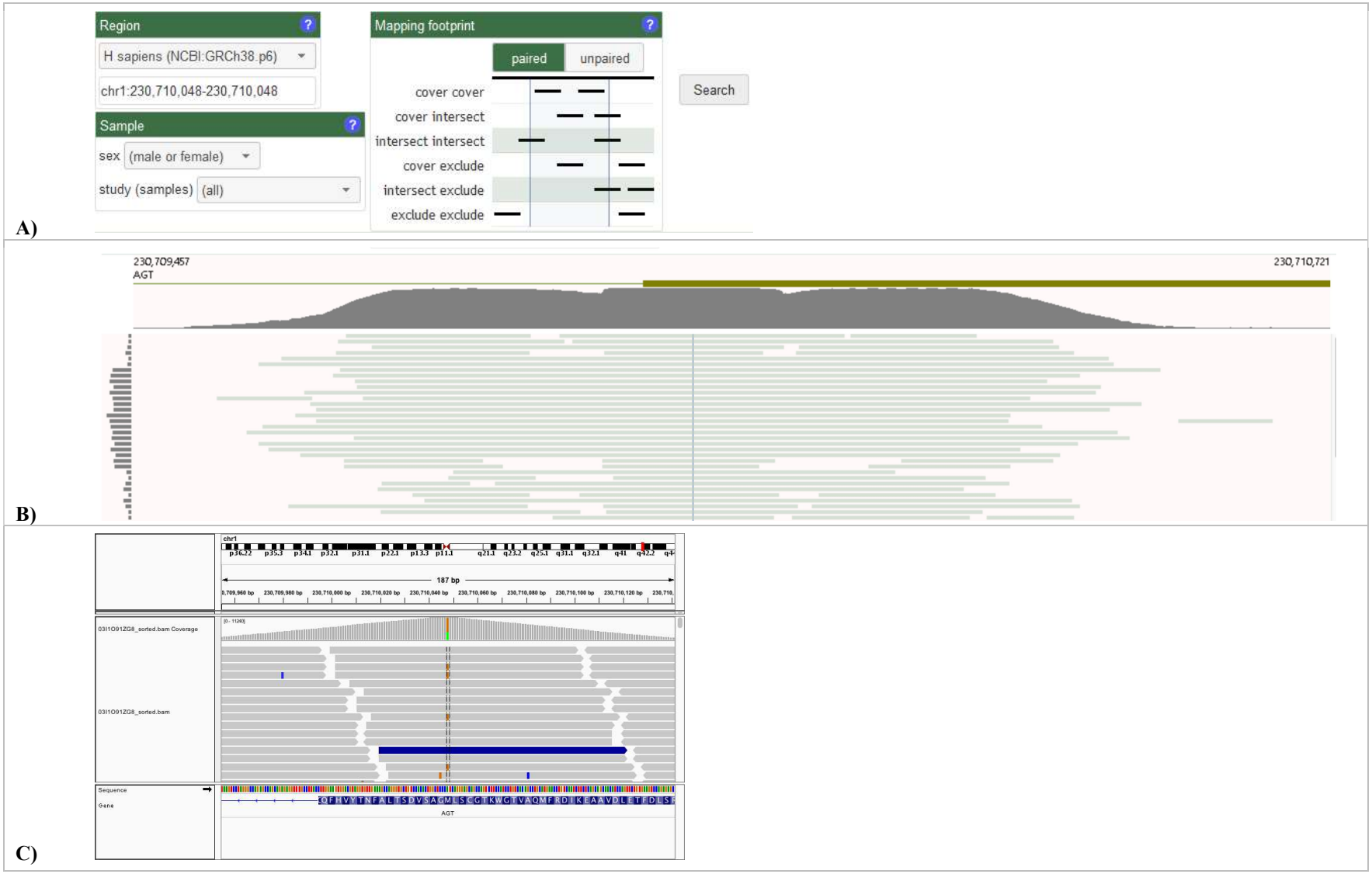
Examples. The first and last examples illustrate alternate encodings for the same set of BQS values. In cases where the final byte may be encoded either as a run length or a set of bin values, the bin-value byte type is preferred:

bin number	**binary**	**hex**	
1 1 1 1	10 01 01 01 10 00 00 01	95 81	(preferred)
1 1 1 1	10 01 01 01 0 0000001	95 01	
1 1 1 1 1 1 1	10 01 01 01 0 0000101	81 05	
3 1 2	10 10 01 11	A7	
4 1 2	11 001 100 10 00 00 10	CC 82	
4 1 5 5 5 5	11 001 100 11 101 101 11 101 101	CC EB EB	(preferred)
4 1 5 5 5 5	11 001 100 11 101 101 0 00000010	CC EB 02	

Appendix A3. Use case: prevalence of a SNP

Because the samples in the TSE database represent a wide geographic distribution of normal individuals, it is possible to assess the overall frequency of a common variant such as SNP rs699, which is located at position 230,710,048 on chromosome 1 and has a high alternate allele frequency (global MAF A=0.2949).

The following figure shows how the search region is specified as the one-base "range" chr1:230,710,048-230,710,048, and reads with intersect/exclude or intersect/intersect geometries are requested. The TSE returns a map showing where each genome sample in the database contains reads that intersect the specified genome position. Over 21,000 reads correspond to these parameters and are converted to SAM format by the TSE, from which they can be directly imported into the Integrative Genome Viewer (IGV). Visualization of the reads in IGV confirms the presence of this SNP in 35% of the reads.



A) The TSE searches for all paired-end reads where one or both mappings intersect the position of a SNP at chr1:230,710,048. B) The TSE displays a coverage summary for the specified reads; the left-hand side shows one horizontal coverage map for each WGS sample in the TSE database. C) IGV displays the reads surrounding the SNP in the TSE-generated SAM file.

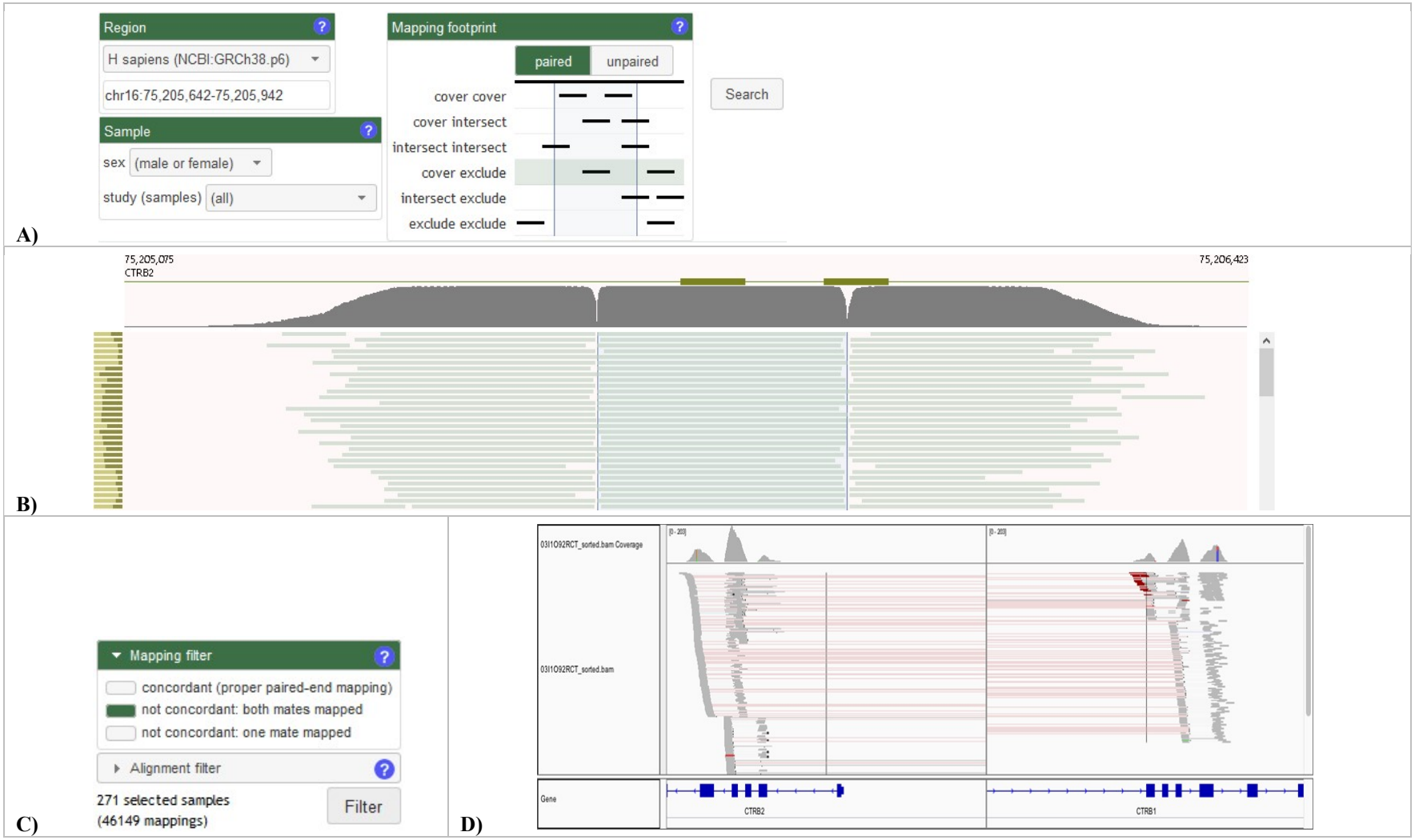
Appendix A4. Use case: identifying an inversion

The TSE can be searched for structural variations such as inversions in a specific region of the reference genome. For example, a common genomic inversion can be found on chromosome 16 that spans the genes CRTB1 and CRTB2. This inversion is the major allele, occurring in 95% of the global population[1,2].

In the following figure, reads in the TSE that involve the inversion are identified by defining a search region just within the 5′ end of the inversion and requesting reads with the cover/exclude geometry; i.e., paired-end reads with one mate mapped entirely within the inversion and near its 5′ end, and with the opposite mate mapped entirely outside it. In individuals without the inversion, we expect to see such pairs mapped concordantly; in individuals with the inversion, we expect such pairs to map with an inferred fragment length roughly equal to the length of the inversion (that is, one mate maps to the reference genome just beyond the 5′ end of the inversion sequence in the reference genome, and the opposite mate maps near the 3′ end of the inversion sequence). Thus, by filtering out concordant mappings from the search results, the TSE returns reads only from individuals in which the inversion occurs. The mappings form two peaks when visualized in IGV.

[1] (http://invfestdb.uab.cat HsInv0030)

[2] Rosendahl J, Kirsten H, Hegyi E, et al. (2018). Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut* **0**:1–9. DOI:10.1136/gutjnl-2017-314454

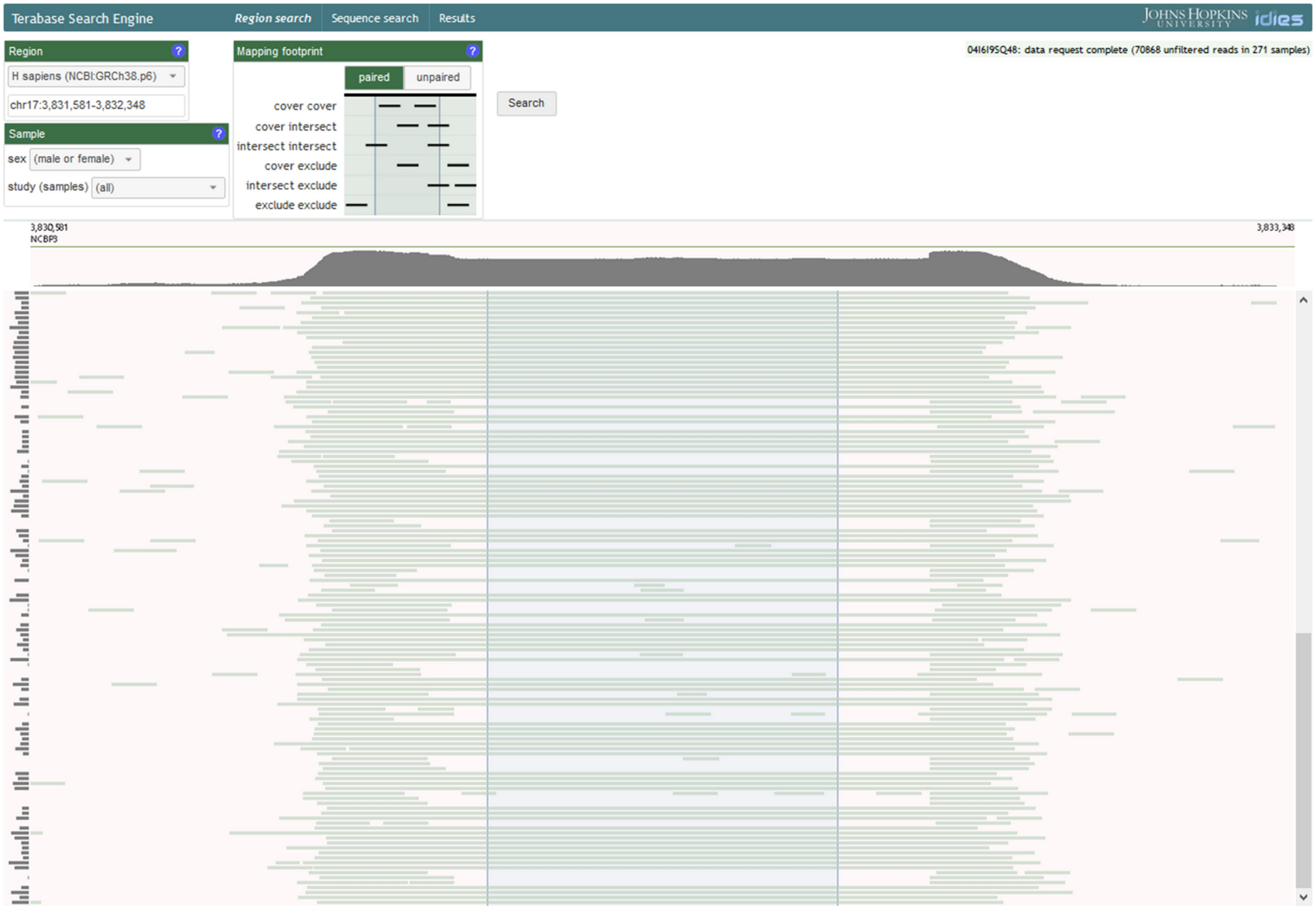


A) The TSE searches for all paired-end reads where the mates straddle the 5′ breakpoint. B) The coverage summary displayed by the TSE. C) A request to return only non-concordant mappings, where both reads in a pair are mapped but do not conform to the expected distance and orientation constraints. D) IGV displays the reads from the TSE-generated SAM file.

Appendix A5. Use case: prevalence of an uncatalogued deletion

As part of a clinical study (unpublished), a previously-uncatalogued deletion of about 780bp in the region chr17:3,831,581-3,832,348 was observed in WGS samples for four individuals, all from the same family, affected with an inherited disorder. The immediate question was whether this deletion might be the cause of their disorder. One of the ways to answer this question is to determine whether or not the deletion also appears in WGS samples for unaffected individuals.

The prevalence of this deletion can be assessed using a TSE region search with all possible mapping footprint configurations (see next page). Upon running this search, it became evident that many normal individuals have a deletion in this same area. This evidence allowed us to determine that the deletion is not the cause of the genetic disorder in the original family.



TSE Region search for possible deletion at chr17:3,831,581-3,832,348.

Appendix A6. Use case: variations in the length of a tandem repeat

An abnormality in the length of a tandem repeat of the hexamer GGGGCC in C9orf72 is the most common inherited cause of human amyotrophic lateral sclerosis (ALS, or Lou Gehrig's Disease)[3,4]. Although the fragment length of the paired-end reads in the current TSE database is far too short to measure the 1000-fold expansion of the hexanucleotide that is associated with clinical disease, it is still possible to evaluate the frequency with which the number of copies of this tandem repeat differs from the human reference genome in the WGS samples represented in the TSE database.

The human reference genome GRCh38 contains only 3 repeats of this hexamer at chr9: 27,573,529-27,573,546. To search for reads in the TSE samples that originated from this region, a query must include reads that map to the region as well as reads whose opposite mates map within the expected paired-end fragment length from the region. (In the latter case, a read derived from the region may differ sufficiently from the reference genome that it has no valid high-scoring alignment even though its opposite mate maps within the expected distance from the region.)

In effect, we are interested in searching for reads whose mappings lie within a region that extends upstream and downstream of the region of interest:

```
one mate is upstream                ======          =======
one mate is downstream                              =======          ======
GRCh38.p6 chr9:27573529-27573546  ----------------------===----------------------
```

Since the average fragment length (TLEN) in the current TSE WGS samples is about 300bp, the boundaries of the region to search are:

```
(27573546 - 300) - 50 = 27573196
(27573529 + 300) + 50 = 27573879
```

This set of reads is specified using "cover-cover" and "cover-exclude" mapping footprints in the TSE Region search web page. The result can be exported in the TSE Results web page in SAM format for subsequent analysis (see next page).

A few Linux shell commands suffice to generate a rough estimate of the frequency with which the tandem repeat occurs in the reads in this region:

```
cat <(awk -F'GGGGCC' '{print NF-1}' 13I6IJ0NVP.sam) <(awk -F'CCCCGG' '{print NF-1}' 13I6IJ0NVP.sam) |
 sed '/^0$/d' | sort -n | uniq -c
```

This gives the following result (column 1 shows the number of reads, column 2 the number of repeat copies within the read):

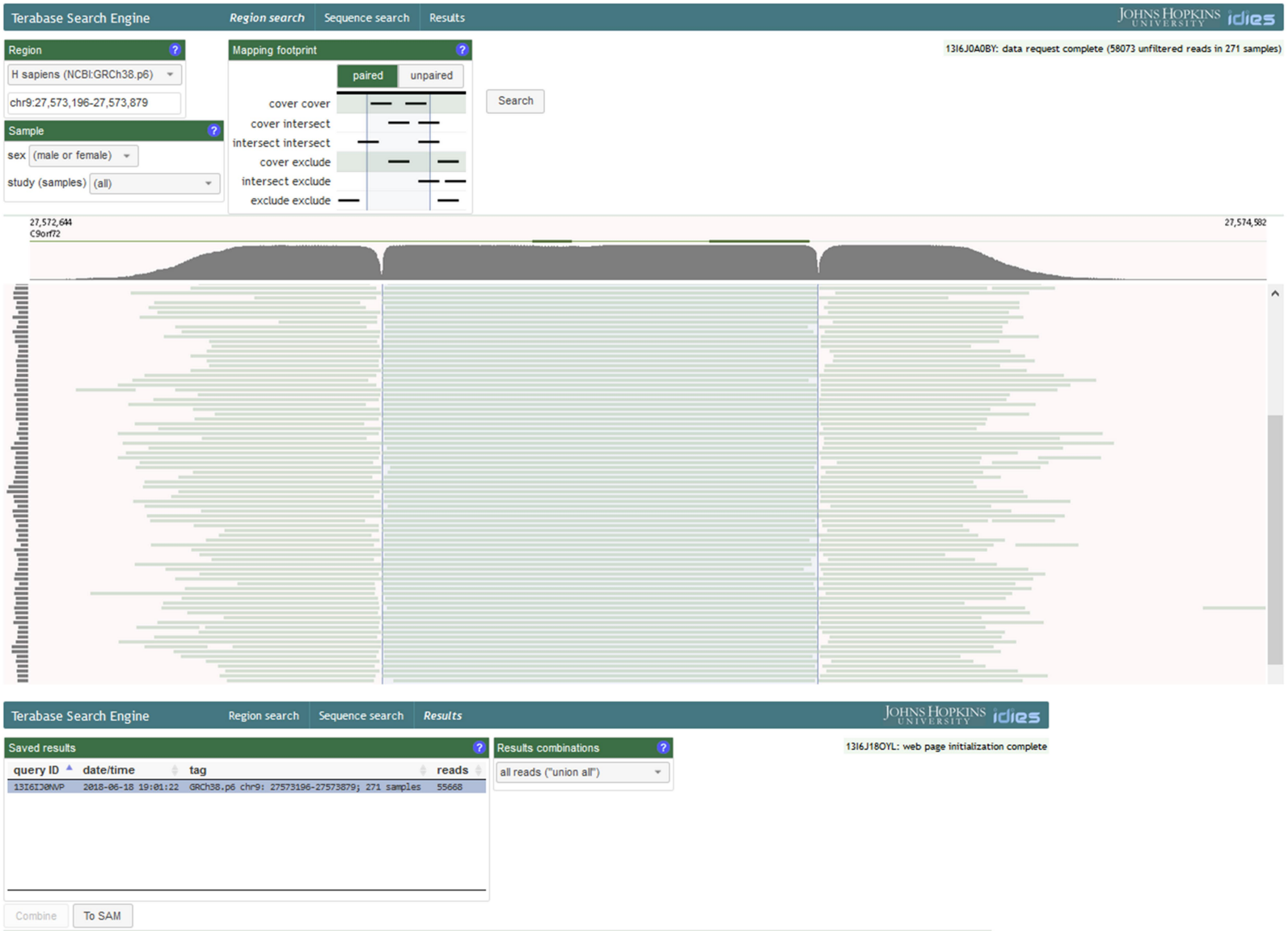
```
9155 1
2687 2
1410 3
1840 4
 610 5
 526 6
 417 7
 260 8
 134 9
  49 10
  27 11
   8 12
```

Of course, this is far from a rigorous analysis — but it takes only one minute to extract and summarize the pertinent reads from the WGS samples in the TSE database.

---

[3] DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**:245–256.

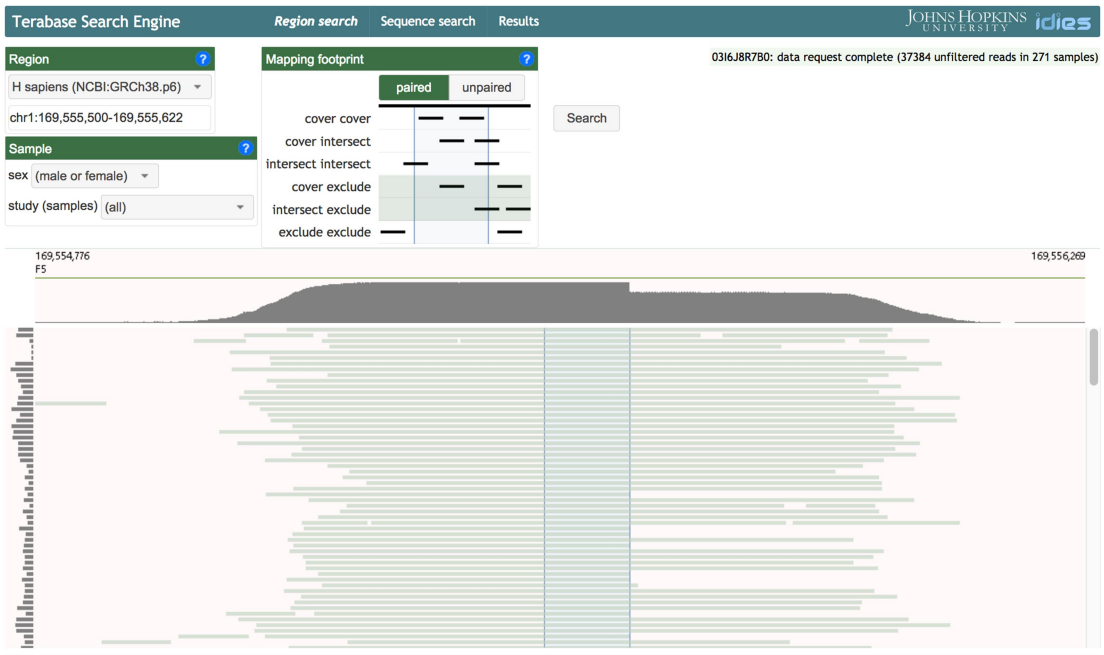
[4] Renton AE, Majounie E, Waite A, et al. (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**:257–268.

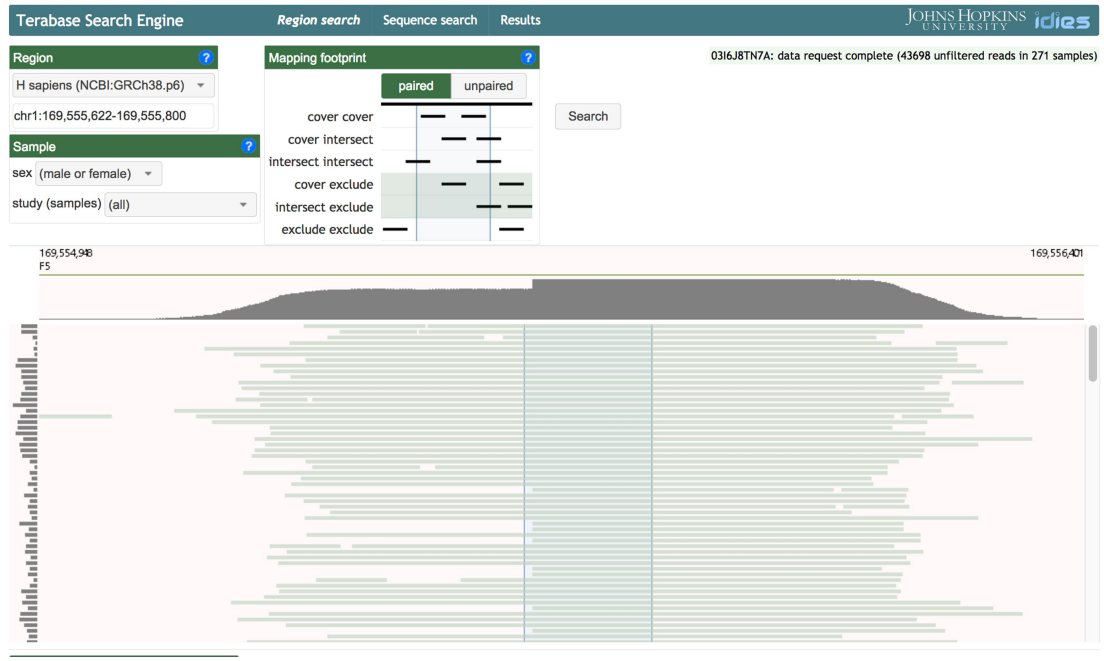


TSE Region search and Results for reads that may contain a tandem repeat at chr9: 27,573,529-27,573,546.

Appendix A7. Use case: prevalence of a polymorphic L1 (LINE-1) insertion

Roughly 37% of the human population has a polymorphic L1 element at chr1:169,555,622 (1000 genomes release 20130502). Querying for reads starting on either side of this position, and filtering to retain read pairs that are "not concordant: both mates mapped" and "not concordant: one mate mapped," results in 9625 and 8321 read pairs from the 5′ and 3′ flanking regions, respectively:





The coverage profiles reflect the structural variation common to a large percentage of the mapped reads. In this example, 54.4% of the mates from the upstream sequence and 37.5% of the mates from the downstream sequence align to the L1 sequence, confirming the presence of this polymorphic element in many of the samples.

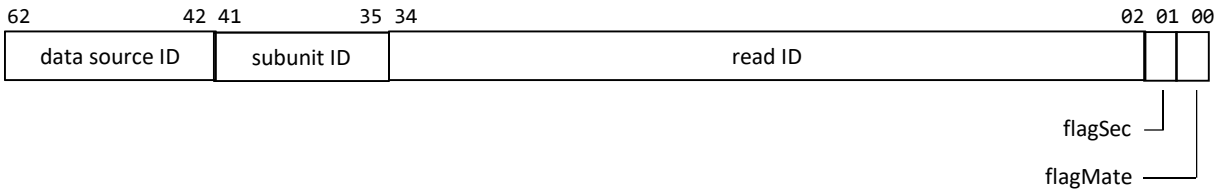


Figure S1. Read sequence identifier format. The low-order bit ("flagMate") indicates whether a read represents mate 1 or mate 2 of a paired-end read. Bit 1 ("flagSec") is set when a read's mapping is flagged as a secondary mapping by the read aligner. Bits 2 through 34 contain a unique numeric identifier assigned sequentially by the read aligner. Bits 35 through 41 and 42 through 62 identify the data source of the read; in the TSE database, each such "data source" represents a reference to a WGS (whole-genome sequencing) run. The high-order bit (bit 63) is unused.

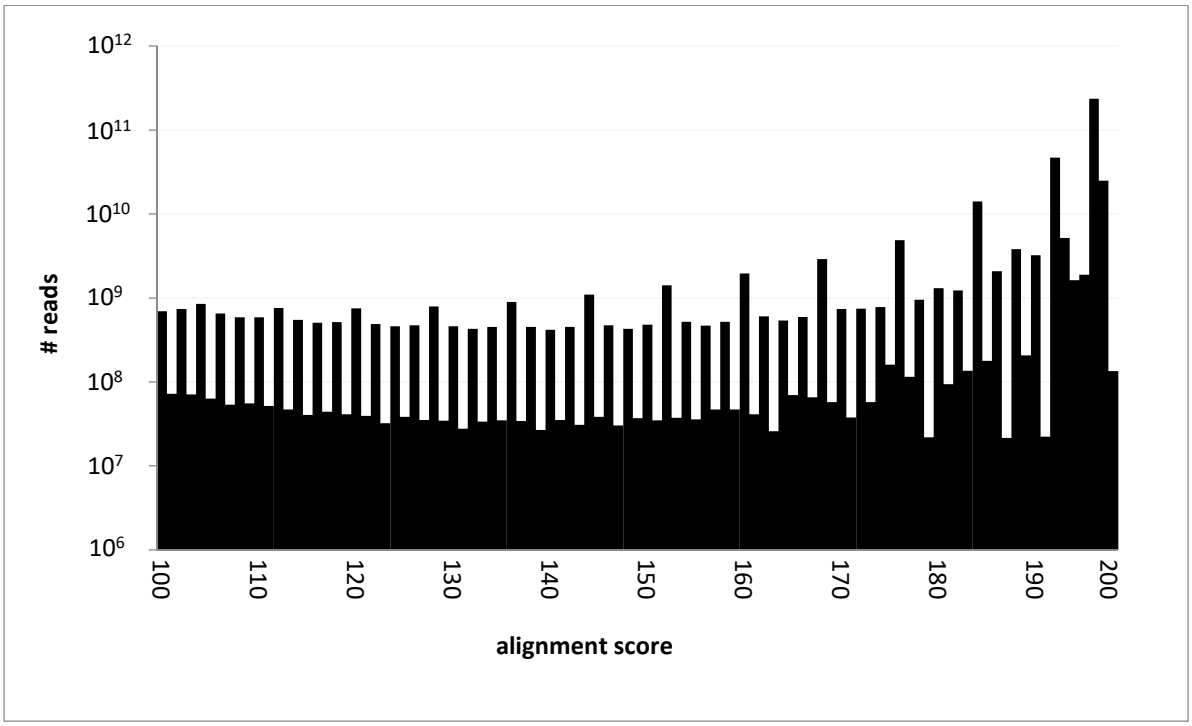


Figure S2. Distribution of alignment scores for mapped reads. The average alignment score for mapped reads is 193 (standard deviation 17).

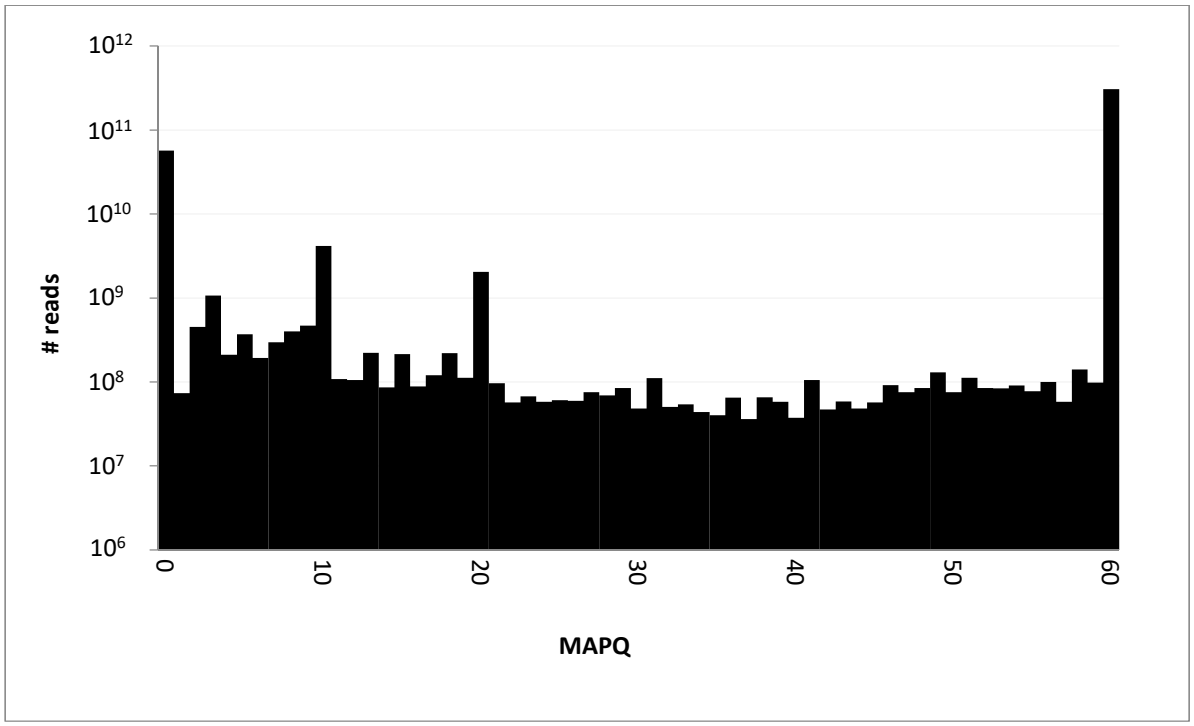


Figure S3. Distribution of mapping quality scores (MAPQ) for mapped reads. The average MAPQ value is 50.