Assemblytics: a web analytics tool for the detection of assembly-based variants

Maria Nattestad and Michael C. Schatz

Table of Contents

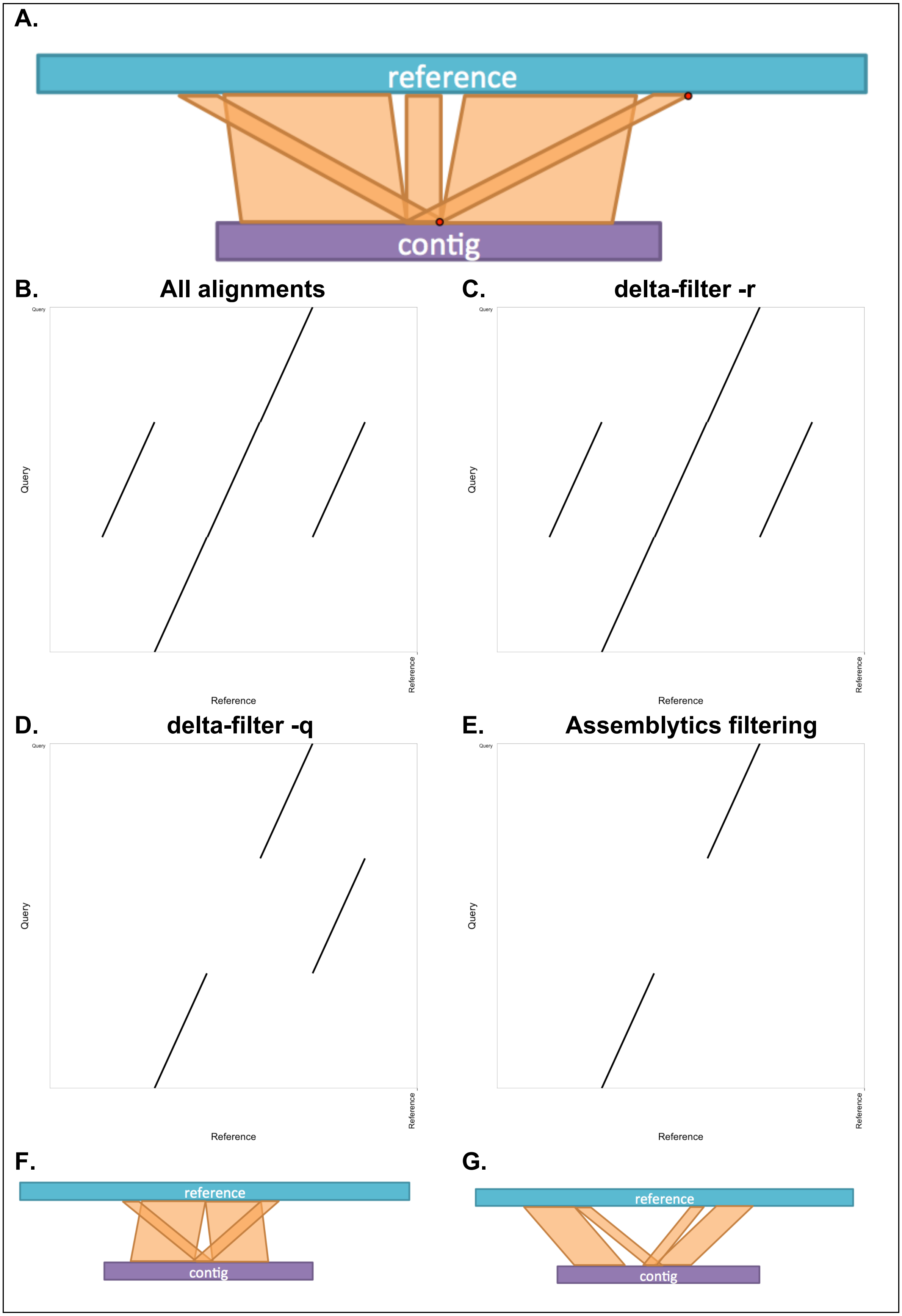
Supplementary Note 1: Unique Anchor Filtering
Supplementary Figure 1: Schematic Diagram of Unique Anchor Filtering
Supplementary Note 2: Assemblytics Accuracy with Simulated Genomes
Supplementary Figure 2: Recall of Simulated Structural Variants
Supplementary Note 3: Assemblytics Web Interface
Supplementary Figure 3: Assemblytics Starting Page
Supplementary Figure 4: Assemblytics of MHAP *Homo sapiens* Assembly (brief)
Supplementary Note 4: Assemblytics Analysis of MHAP Genome Assemblies
Supplementary Figure 5: Assemblytics of MHAP *H. sapiens* Assembly (detailed)
Supplementary Figure 6: Assemblytics of MHAP *A. thaliana* Assembly
Supplementary Figure 7: Assemblytics of MHAP *D. melanogaster* Assembly
Supplementary Figure 8: Assemblytics of MHAP *S. cerevisiae* Assembly
Supplementary Figure 9: Assemblytics of MHAP *E. coli* Assembly

Supplementary Note 1: Unique Anchor Filtering

Assemblytics filters out repetitive sequences that have no unique anchoring alignments. Without such filtering, repetitive alignments could cause many false structural variations between each copy of the repeat. **Supplementary Figure 1 A-E** demonstrate a case where the repeat's "true alignment" could be chosen as the middle alignment based on the alignments of the neighboring sequences on the contig, but often there is no clear choice of consistent alignment (**Supplementary Figure 1 F and G**) and mutations could also be obscuring the mapping.

In an effort to conservatively call variants around repeats, Assemblytics does not attempt to choose the alignment for each repeat but rather filters out all alignments that do not contain at least a minimum amount (default: 10 kb) of uniquely anchoring sequence on the contig. The minimum anchoring sequence length can also be adjusted at runtime for different genome complexities and/or sequencing technologies. Note that this does not remove all sequences containing repeats but rather all sequences consist of only repeats without any unique sequence to anchor them.

The sequences and delta files used in this figure are available at:
http://qb.cshl.edu/assemblytics/reproducibility


A.
reference
contig
B.
All alignments
Query
Query
Reference
Reference
C.
delta-filter -r
Query
Query
Reference
Reference
D.
delta-filter -q
Query
Query
Reference
Reference
E.
Assemblytics filtering
Query
Query
Reference
Reference
F.
reference
contig
G.
reference
contig

Supplementary Figure 1. Each repetitive element in a genome assembly can map ambiguously to multiple locations in the reference genome. *Delta-filter* is a component of MUMmer designed to filter repetitive alignments (Kurtz, *et al.* 2004; Phillippy, *et al.* 2008). It uses a longest-increasing subsequence (LIS) dynamic programming algorithm to identify subsets of long high identity alignments while penalizing overlapping alignments. In contrast, Assemblytics includes a filtering step to eliminate repeats without substantial unique anchoring sequence (by default 10kb). **A.** This simulated example demonstrates a case in which a 20 kb sequence in the contig (query) matches three locations in the reference exactly, except the addition of a single nucleotide (indicated by the red point) that matches better to the reference repeat on the right. **B.** Dot plot showing all alignments in the raw, unfiltered delta file output from *nucmer*. **C.** Dot plot showing results of applying *delta-filter -r* (same as unfiltered). **D.** Dot plot showing results of applying *delta-filter -q*. The 1 bp addition to a 20 kb repeat is enough to make *delta-filter -q* choose the third alignment. **E.** Dot plot showing results of the unique anchor filtering in Assemblytics. The repeat alignments are all removed since none of these include a unique anchor sequence of at least 10 kb that only aligns to a single position in the reference. Assemblytics tags any structural variants within this gap as repeat expansions for contractions, depending on whether the size of the gap between the two surrounding alignments is greater in the query or the reference, respectively. No variant is reported unless the size of the gap changes, so repeats themselves are not reported as structural variants, only changes in their size are reported as repeat expansions (increases in size) or repeat contractions (decreases in size). **F, G.** Diagrams showing additional ambiguous examples of repeats mapping to multiple locations.

Supplementary Note 2: Assemblytics Accuracy with Simulated Variants

To assess the accuracy of structural variation calling with Assemblytics, we embedded simulated insertions and deletions at known positions in the hg19 human reference genome. We then aligned the original reference genome sequence as well as the de novo assembly of the human genome assembled with MHAP presented by Berlin *et al* (2015) to this mutated reference in order to see if we could recover the simulated variants. The variants were created in the reference to allow an actual assembly to be used as the query without simulating reads, as the long read sequencers and long read assemblies have complex error models.

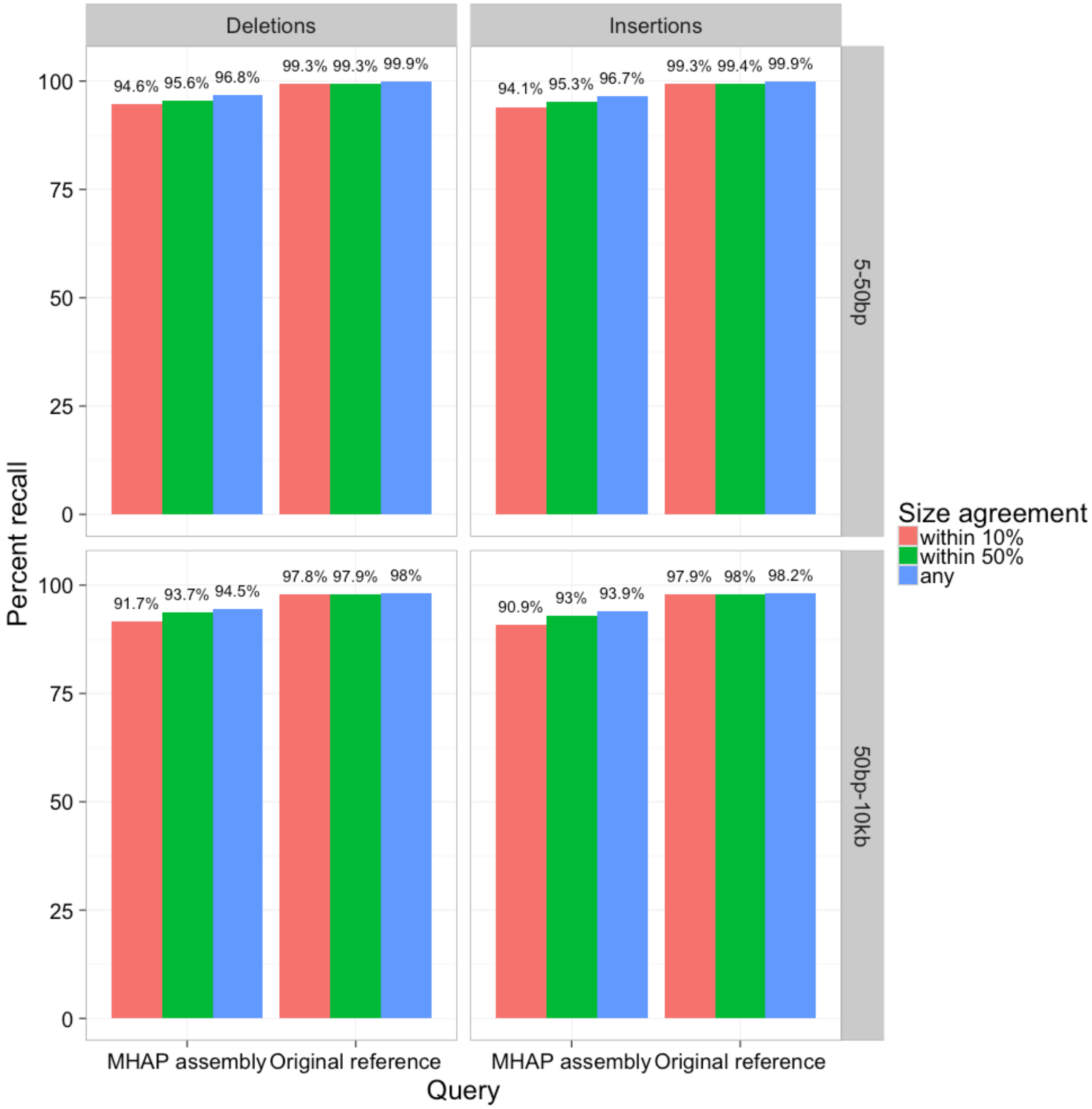
Variants were simulated at every 100 kbp, except where the 10kb sequences upstream and downstream contain a series of 100 or more Ns indicating problematic genomic regions with missing sequence; these are skipped. Variant sizes are chosen randomly from a uniform distribution in the range 50 bp to 10 kbp, and separately for the range of 5 bp to 50 bp to test whether Assemblytics performs differently on structural versus smaller variants. *Nucmer* was used for alignment and Assemblytics for variant-calling, and then *bedtools* was used to detect how many of the simulated variants were recovered. For insertions, the reference coordinates indicate only a single base-pair, so these were extended with a window of 10 bp to either side to allow for small arbitrary shifts in the local alignments.

The recall of the variants with the original reference and the MHAP assembly are presented in **Supplementary Figure 2** and shows that the vast majority of events can be detected. On inspection, variants that cannot be detected are those within long repetitive elements. The recall scores using the MHAP assembly are lower than for the original reference, as expected, since some sequences in the genome are hard to assemble and may not be represented in the MHAP assembly.

Since the MHAP assembly is from a different individual than the hg19 reference genome there are many real variants between these so the precision cannot be fully evaluated with this dataset. Instead we evaluated the precision of variants in the original reference with respect to the modified reference sequences. For deletions, 80 out of 28006 Assemblytics calls (0.286%) are false positives and among the insertions, 114 out of 28125 Assemblytics calls (0.405%) are false positives.

The simulated genomes fasta and delta files used in this figure are available at
http://qb.cshl.edu/assemblytics/reproducibility

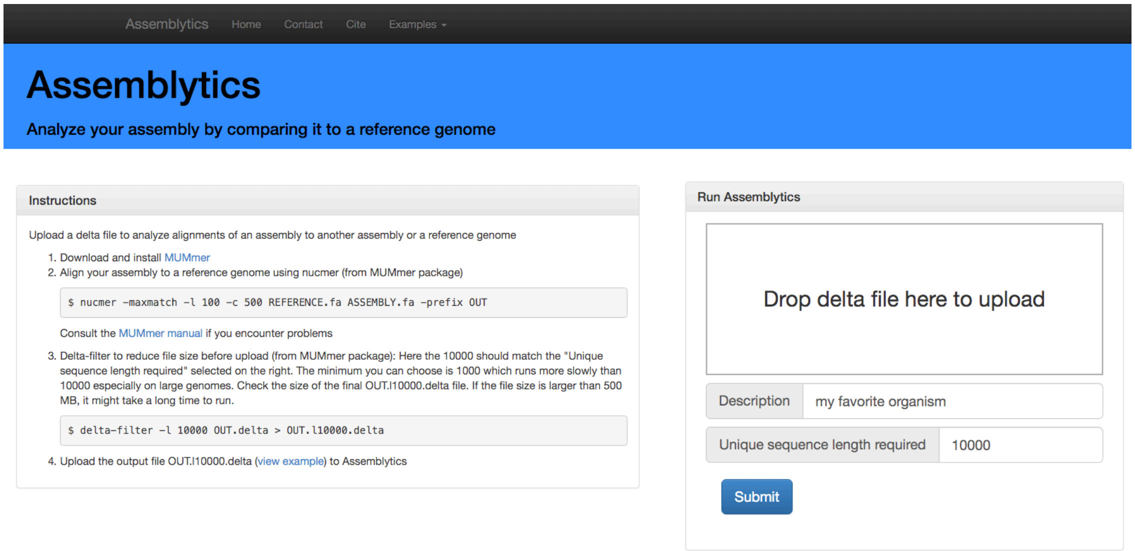
A.



Supplementary Figure 2. Simulation results for insertions and deletions in the size range 5 bp to 50 bp (top) and in the size range 50 bp to 10 kbp (bottom). Each of the four plots contains 28522 simulated variants, 1 at every 100 kbp simulated in the reference and using either the MHAP human assembly or the un-mutated hg19 reference.

Supplementary Note 3: Assemblytics Web Interface

A.



Assemblytics

Analyze your assembly by comparing it to a reference genome

Instructions

Upload a delta file to analyze alignments of an assembly to another assembly or a reference genome

1. Download and install MUMmer
2. Align your assembly to a reference genome using nucmer (from MUMmer package)

```
$ nucmer -maxmatch -l 100 -c 500 REFERENCE.fa ASSEMBLY.fa -prefix OUT
```

   Consult the MUMmer manual if you encounter problems

3. Delta-filter to reduce file size before upload (from MUMmer package): Here the 10000 should match the "Unique sequence length required" selected on the right. The minimum you can choose is 1000 which runs more slowly than 10000 especially on large genomes. Check the size of the final OUT.l10000.delta file. If the file size is larger than 500 MB, it might take a long time to run.

```
$ delta-filter -l 10000 OUT.delta > OUT.l10000.delta
```

4. Upload the output file OUT.l10000.delta (view example) to Assemblytics

Run Assemblytics

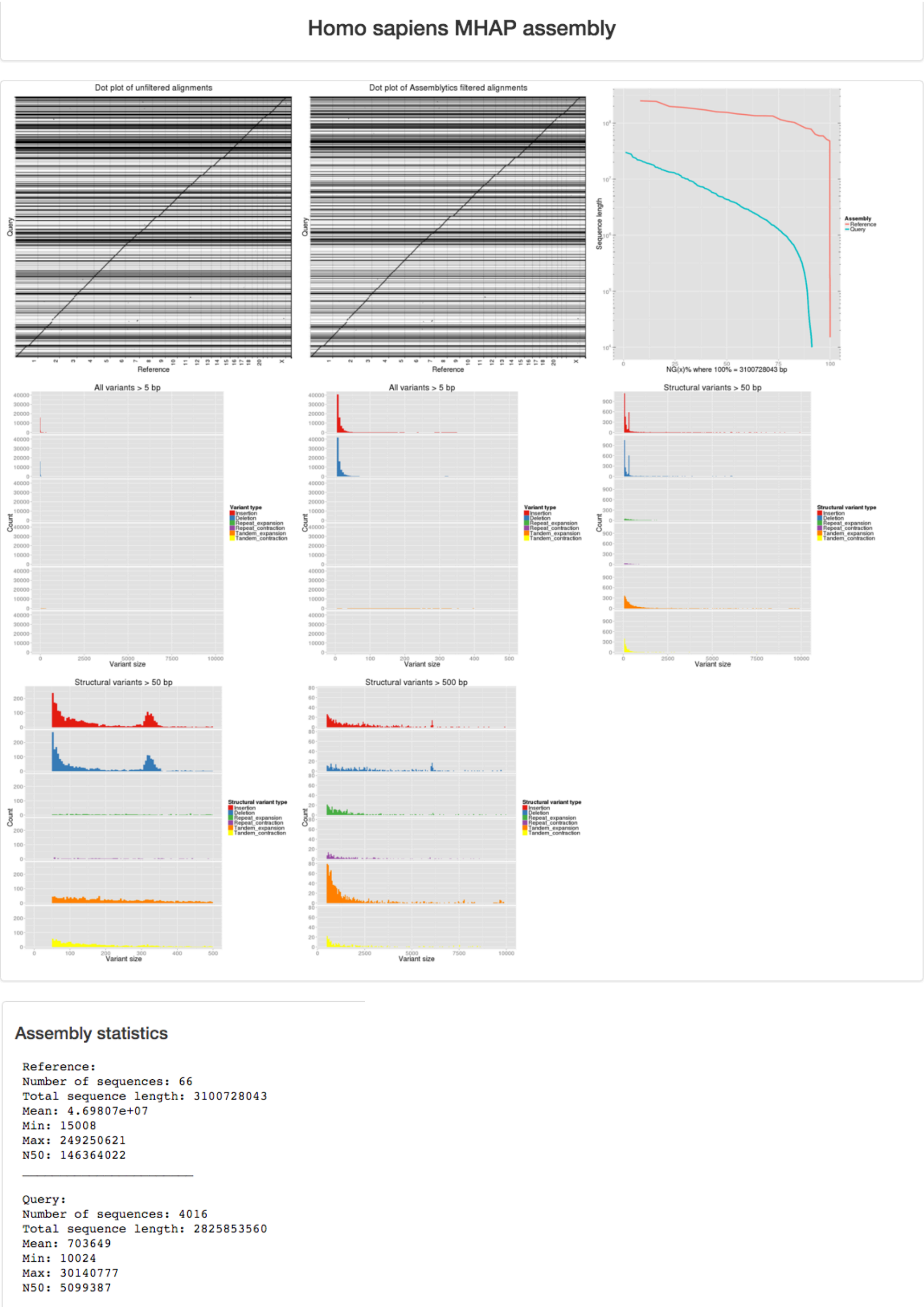
Drop delta file here to upload

Description: my favorite organism

Unique sequence length required: 10000

Supplementary Figure 3. Front page of Assemblytics web application where users can upload a delta file from *Nucmer*, assign it a identifier, and choose the unique sequence length required to anchor alignments. The page includes instructions for running *nucmer*, and after users click the submit button any problems with the uploaded input is addressed before running Assemblytics. Examples are also available in the top navigation bar showing the output for five different organisms: Human, Drosophila, Arabidopsis, E. coli, and yeast, each assembled with MHAP and presented by Berlin et al (2015).

A.

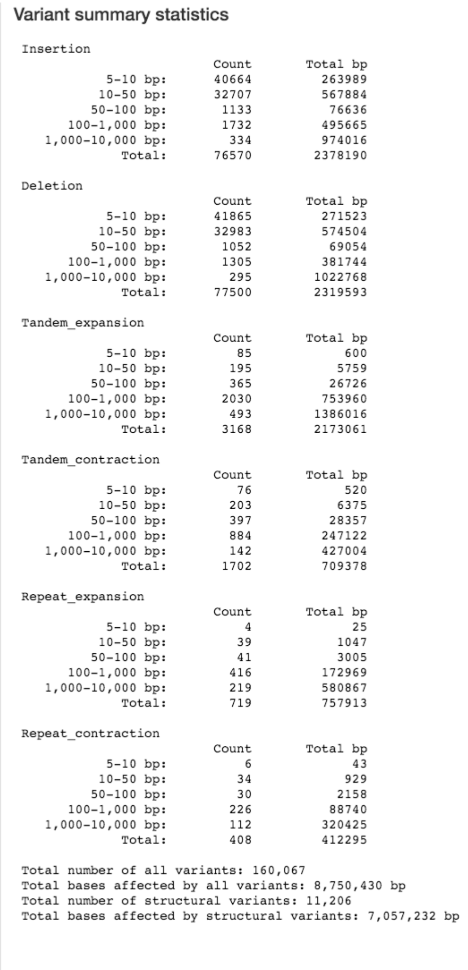


Homo sapiens MHAP assembly

Assembly statistics

```
Reference:
Number of sequences: 66
Total sequence length: 3100728043
Mean: 4.69807e+07
Min: 15008
Max: 249250621
N50: 146364022
_______________________

Query:
Number of sequences: 4016
Total sequence length: 2825853560
Mean: 703649
Min: 10024
Max: 30140777
N50: 5099387
```

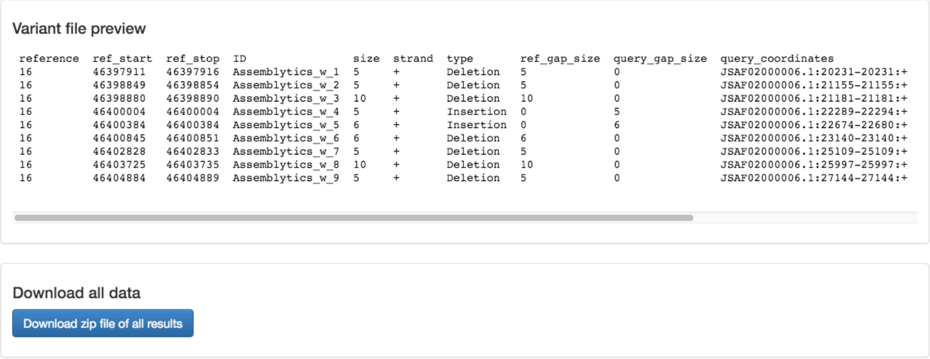
B.



Variant summary statistics

```
Insertion
                       Count       Total bp
          5-10 bp:     40664         263989
         10-50 bp:     32707         567884
        50-100 bp:      1133          76636
     100-1,000 bp:      1732         495665
  1,000-10,000 bp:       334         974016
           Total:      76570        2378190

Deletion
                       Count       Total bp
          5-10 bp:     41865         271523
         10-50 bp:     32983         574504
        50-100 bp:      1052          69054
     100-1,000 bp:      1305         381744
  1,000-10,000 bp:       295        1022768
           Total:      77500        2319593

Tandem_expansion
                       Count       Total bp
          5-10 bp:        85            600
         10-50 bp:       195           5759
        50-100 bp:       365          26726
     100-1,000 bp:      2030         753960
  1,000-10,000 bp:       493        1386016
           Total:       3168        2173061

Tandem_contraction
                       Count       Total bp
          5-10 bp:        76            520
         10-50 bp:       203           6375
        50-100 bp:       397          28357
     100-1,000 bp:       884         247122
  1,000-10,000 bp:       142         427004
           Total:       1702         709378

Repeat_expansion
                       Count       Total bp
          5-10 bp:         4             25
         10-50 bp:        39           1047
        50-100 bp:        41           3005
     100-1,000 bp:       416         172969
  1,000-10,000 bp:       219         580867
           Total:        719         757913

Repeat_contraction
                       Count       Total bp
          5-10 bp:         6             43
         10-50 bp:        34            929
        50-100 bp:        30           2158
     100-1,000 bp:       226          88740
  1,000-10,000 bp:       112         320425
           Total:        408         412295

Total number of all variants: 160,067
Total bases affected by all variants: 8,750,430 bp
Total number of structural variants: 11,206
Total bases affected by structural variants: 7,057,232 bp
```

C.



Variant file preview

```
reference  ref_start  ref_stop  ID                size  strand  type       ref_gap_size  query_gap_size  query_coordinates
16         46397911   46397916  Assemblytics_w_1  5     +       Deletion   5             0               JSAF02000006.1:20231-20231:+
16         46398849   46398854  Assemblytics_w_2  5     +       Deletion   5             0               JSAF02000006.1:21155-21155:+
16         46398880   46398890  Assemblytics_w_3  10    +       Deletion   10            0               JSAF02000006.1:21181-21181:+
16         46400004   46400004  Assemblytics_w_4  5     +       Insertion  0             5               JSAF02000006.1:22289-22294:+
16         46400384   46400384  Assemblytics_w_5  6     +       Insertion  0             6               JSAF02000006.1:22674-22680:+
16         46400845   46400851  Assemblytics_w_6  6     +       Deletion   6             0               JSAF02000006.1:23140-23140:+
16         46402828   46402833  Assemblytics_w_7  5     +       Deletion   5             0               JSAF02000006.1:25109-25109:+
16         46403725   46403735  Assemblytics_w_8  10    +       Deletion   10            0               JSAF02000006.1:25997-25997:+
16         46404884   46404889  Assemblytics_w_9  5     +       Deletion   5             0               JSAF02000006.1:27144-27144:+
```

Download all data

Supplementary Figure 4. Results page of Assemblytics showing the results of aligning the *Homo sapiens* assembly from the MHAP study to the hg19 reference genome. **A.** Plots include two dot plots (before and after filtering), a query vs. reference length comparison showing cumulative sequence lengths, and five variant size distribution plots. Larger versions of these plots are shown in **Supplementary Figure 5**. Basic assembly statistics are also shown including N50. **B.** Variant summary statistics broken down by variant type and size class to show the exact counts and numbers of bases in the genome affected. **C.** A preview of the variants file in bed format, and the button where users can download all their data including the Assemblytics unique anchor filtered delta file, the full bed file of variants, and all the plots and summary tables shown on this page.

Supplementary Note 4: Assemblytics Analysis of MHAP Genome Assemblies

The five assemblies presented in the MHAP paper by Berlin et al (2015) were aligned using *nucmer* against the following reference genomes:

Human: hg19 version used in 1000 Genomes project.
Arabidopsis: TAIR10
Drosophila:
ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.09_FB2016_01/fasta/dmel-all-chromosome-r6.09.fasta.gz
Saccharomyces:
http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_Current_Release.tgz
E. coli: http://www.ncbi.nlm.nih.gov/nuccore/NC_000913

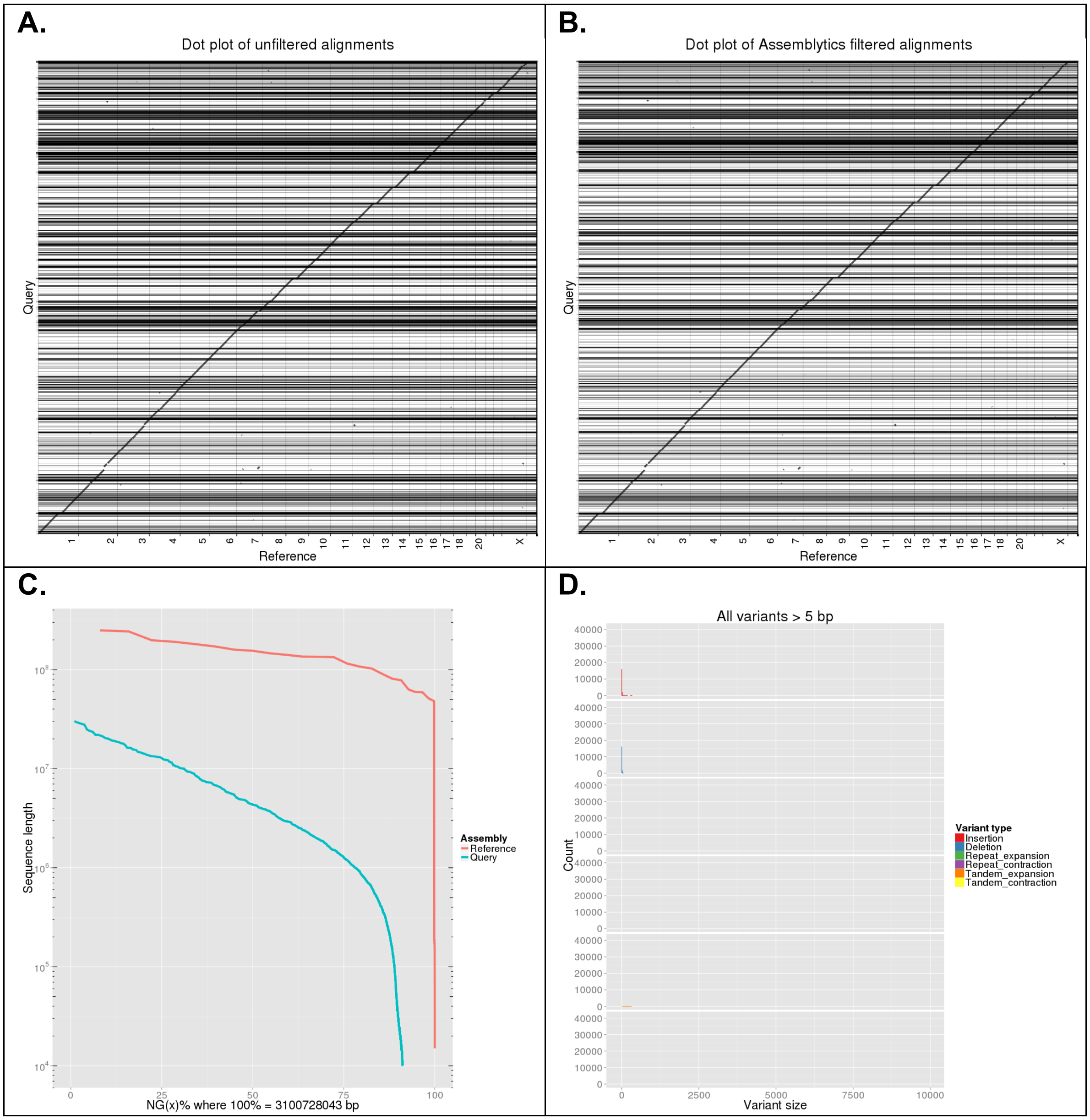
The summary statistics of the variants are presented in **Supplementary Table 1**. The full Assemblytics reports are presented in **Supplementary Figures 5-9** and are on the Assemblytics website.

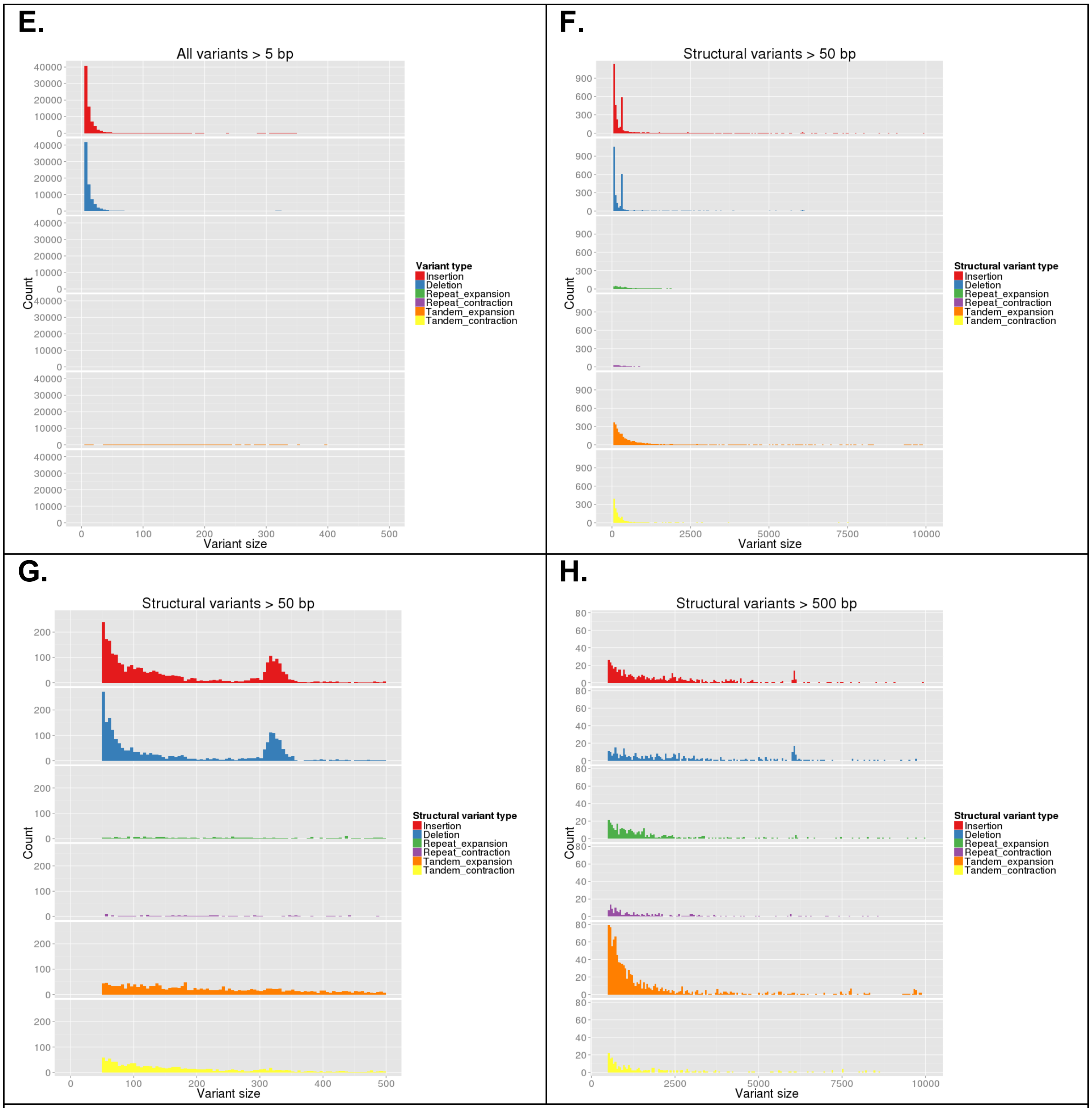
All delta files from this section are available at:
http://qb.cshl.edu/assemblytics/reproducibility

Species	**Reference genome size**	**Structural variants (>50bp)**	**Total variants (>5 bp)**
Homo sapiens	3,100,728,043	11,206	160,067
Drosophila melanogaster	139,047,227	204	1,189
Arabidopsis thaliana	119,667,750	2,501	23,035
Saccharomyces cerevisiae	12,157,105	42	166
Escherichia coli	4,641,652	4	4

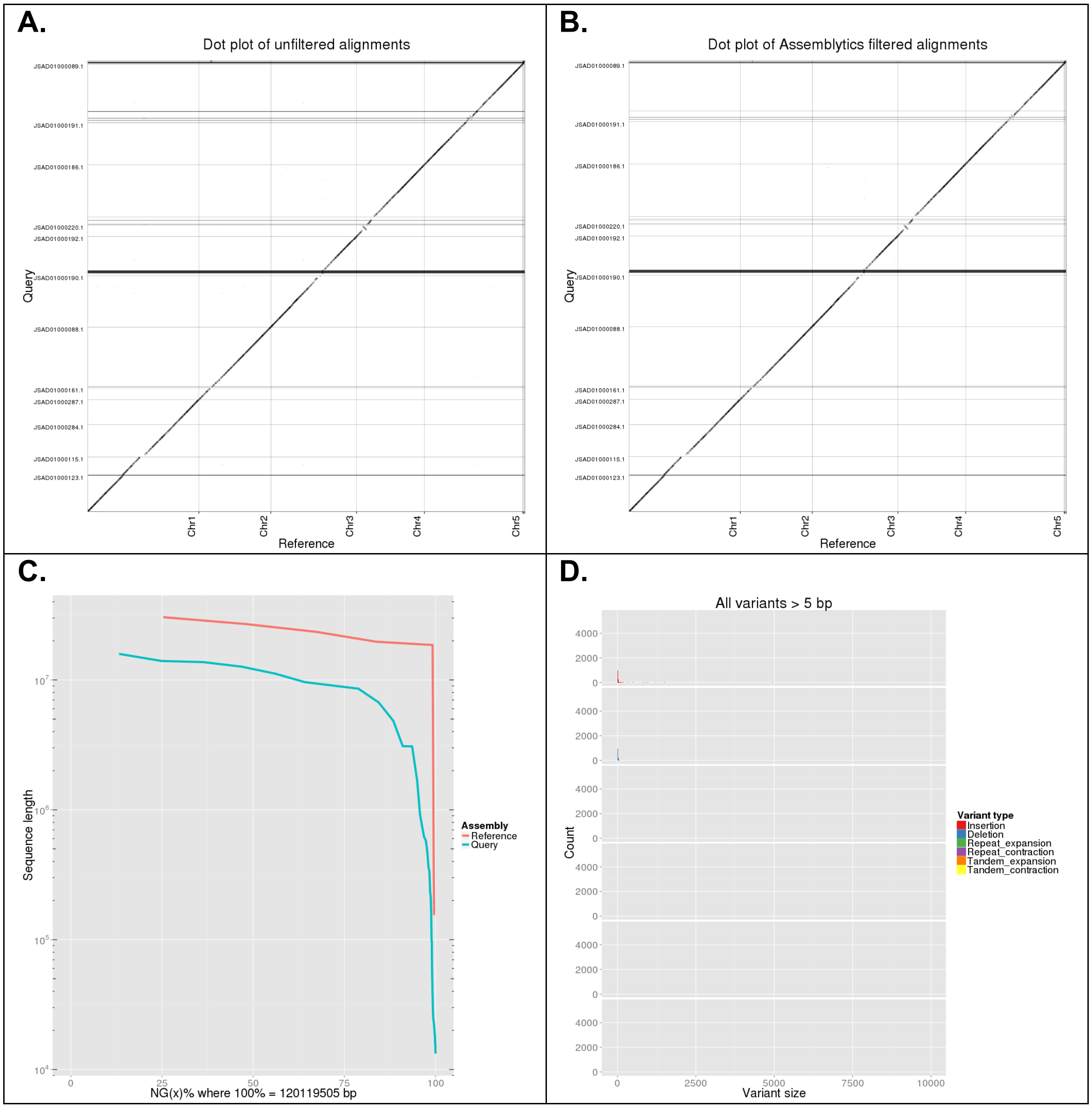
Supplementary Table 1. Each assembly from the MHAP study was aligned to its respective reference genome using nucmer, and structural variants were called using Assemblytics. The count of total variants includes the structural variants.

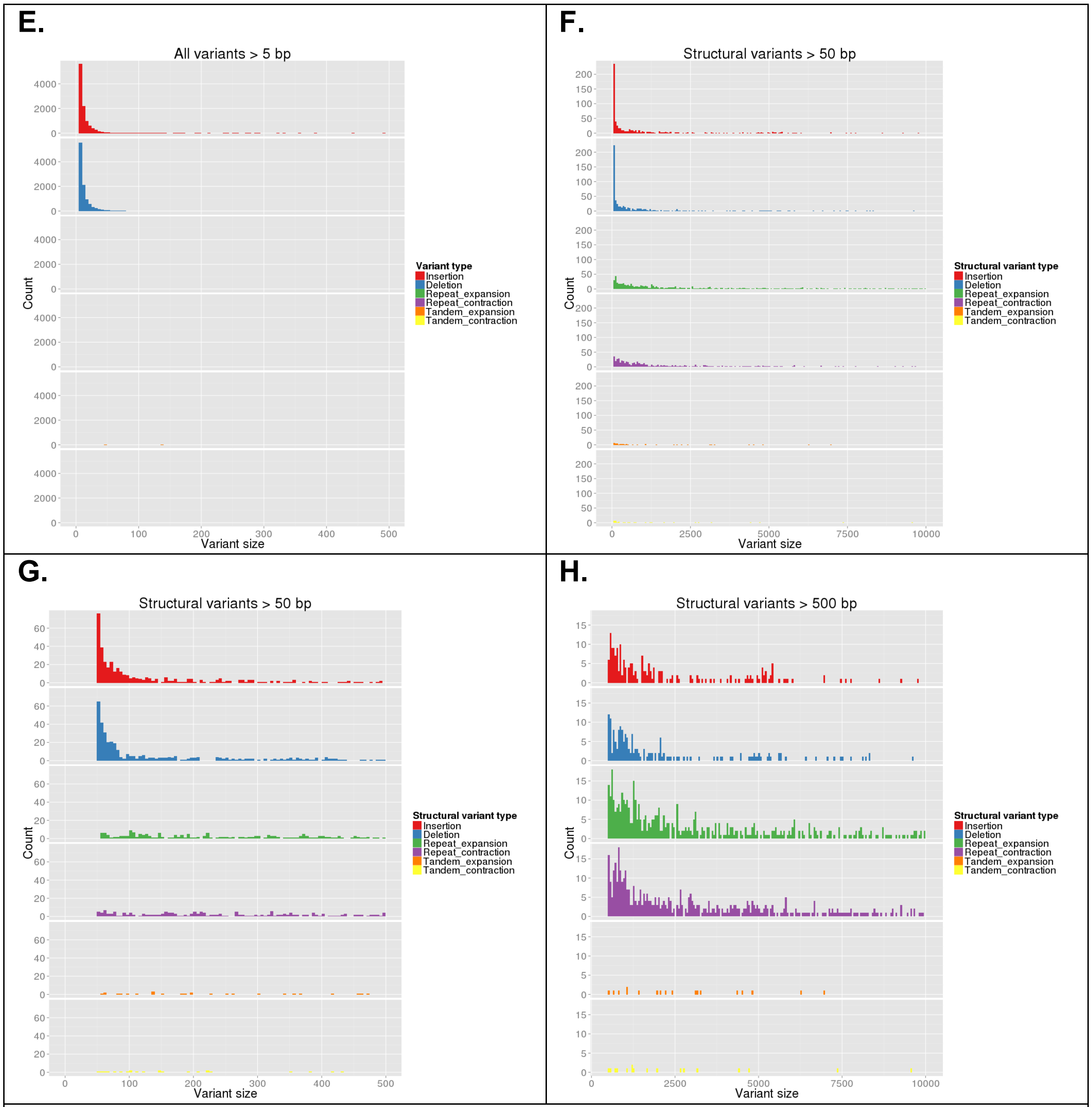
In the human assembly there is a noticeable peak of insertions and deletions around 320 bp. Extracting all deletions in the size range 300-350 bp from the variants output bed file, we see that 579 out of 604 (96%) of these intersect with known Alu elements in the RepeatMasker database of repeats for the hg19 reference genome using bedtools. This is consistent with a high level of Alu elements being deleted while there is a smaller background level of other variants that happen to fall within this size range. Since insertions of Alu elements would not be already marked in the RepeatMasker database for the reference genome, we instead extracted the sequences from the fasta file for all 588 insertions between 300 and 350 bp in size and used BLAST to align them to the Human ALU repeat elements database, where 548 (93%) of these aligned. Thus we conclude that the majority of the enrichment of insertions and deletions around 320 bp are due to jumping Alu elements.



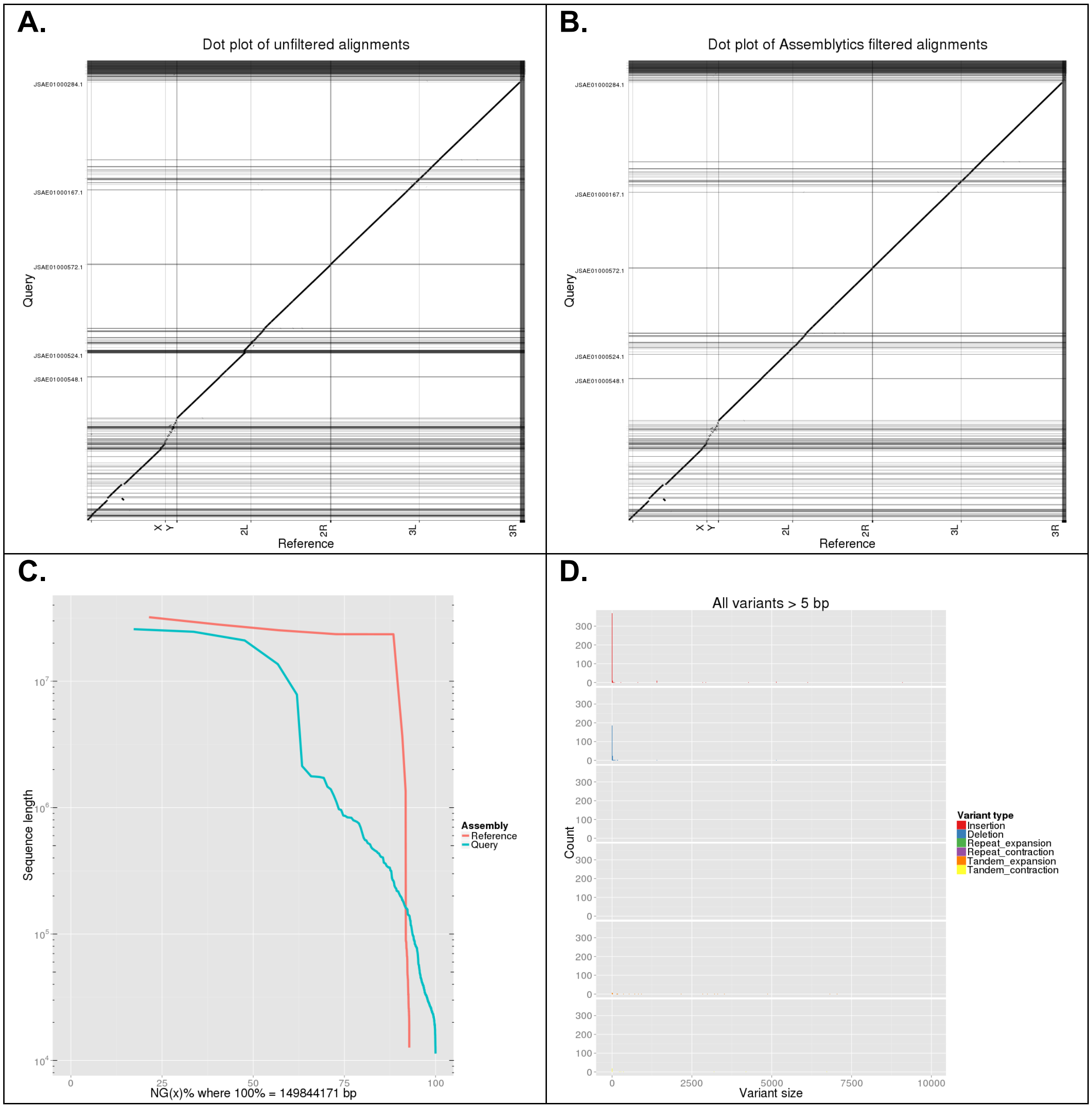


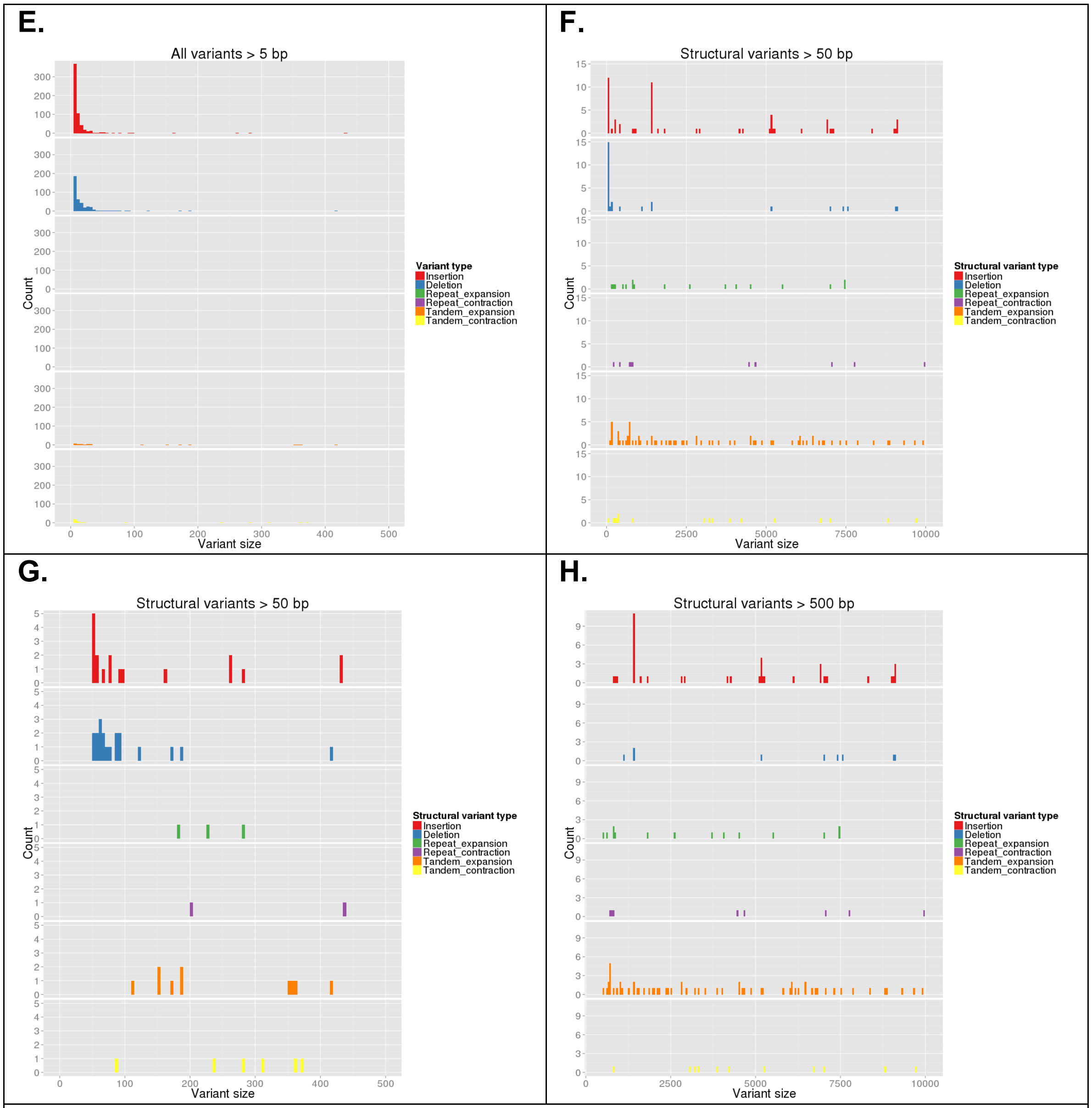
Supplementary Figure 5. Assemblytics results for the *Homo sapiens* genome assembly from the MHAP study. **A.** Dot plot showing alignments of the assembly (query) to the reference before filtering by Assemblytics. **B.** Dot plot after Assemblytics unique anchor filtering. **C.** Cumulative sequence length plot of both the reference and the query. **D.** Variant type and size histogram of all variants above 5 bp. **E.** Variant type and size histogram of all variants 5-500 bp. **F.** Variant type and size histogram of all variants above 50 bp. **G.** Variant type and size histogram of all variants 50-500 bp. **H.** Variant type and size histogram of all variants above 500 bp.


A.
Dot plot of unfiltered alignments
JSAD01000089.1
JSAD01000191.1
JSAD01000186.1
JSAD01000220.1
JSAD01000192.1
JSAD01000190.1
JSAD01000088.1
JSAD01000161.1
JSAD01000287.1
JSAD01000284.1
JSAD01000115.1
JSAD01000123.1
Query
Chr1
Chr2
Chr3
Chr4
Chr5
Reference
B.
Dot plot of Assemblytics filtered alignments
JSAD01000089.1
JSAD01000191.1
JSAD01000186.1
JSAD01000220.1
JSAD01000192.1
JSAD01000190.1
JSAD01000088.1
JSAD01000161.1
JSAD01000287.1
JSAD01000284.1
JSAD01000115.1
JSAD01000123.1
Query
Chr1
Chr2
Chr3
Chr4
Chr5
Reference
C.
10^7
10^6
10^5
10^4
Sequence length
0
25
50
75
100
NG(x)% where 100% = 120119505 bp
Assembly
Reference
Query
D.
All variants > 5 bp
4000
2000
0
4000
2000
0
4000
2000
0
4000
2000
0
4000
2000
0
4000
2000
0
Count
0
2500
5000
7500
10000
Variant size
Variant type
Insertion
Deletion
Repeat_expansion
Repeat_contraction
Tandem_expansion
Tandem_contraction

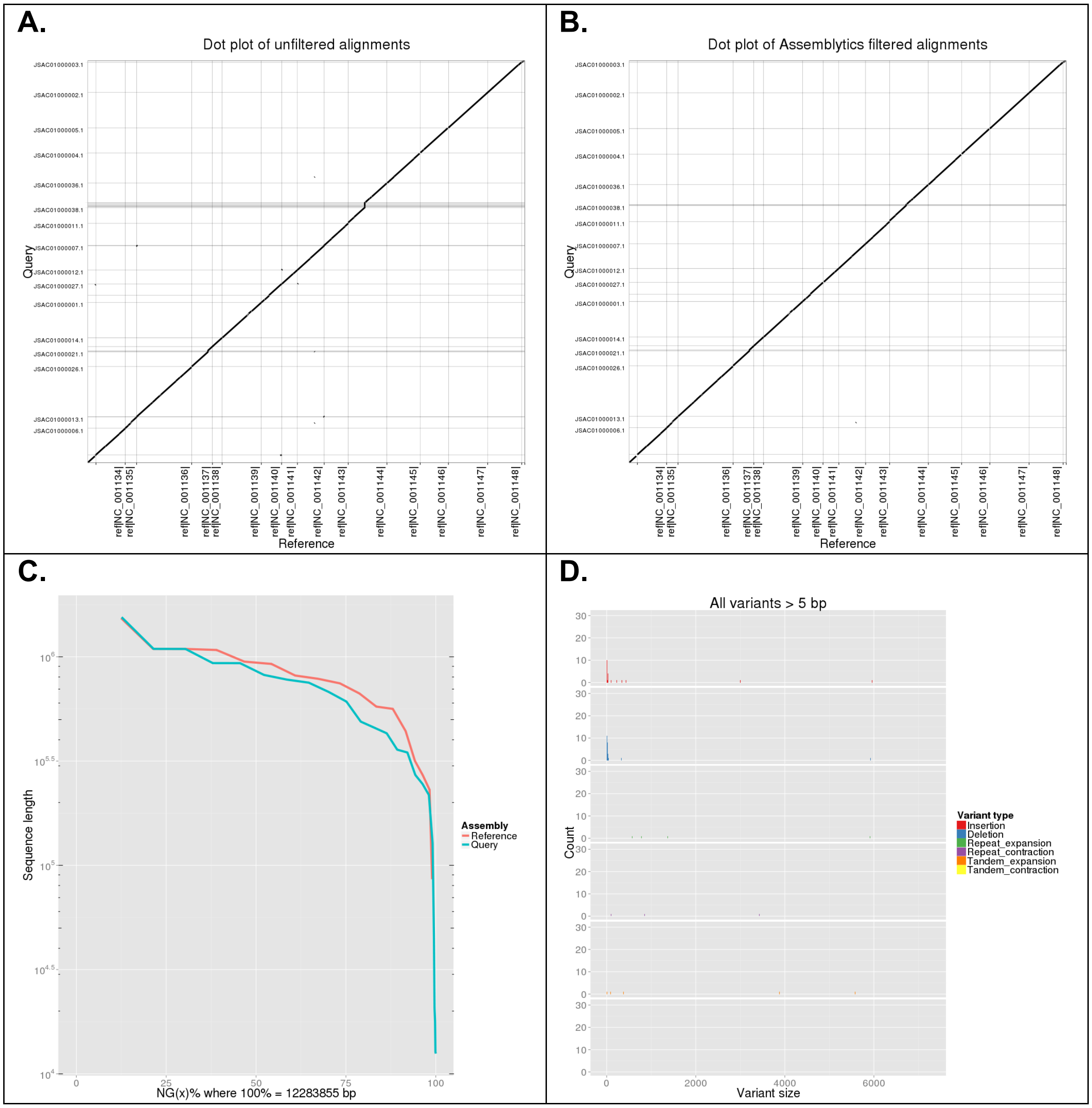


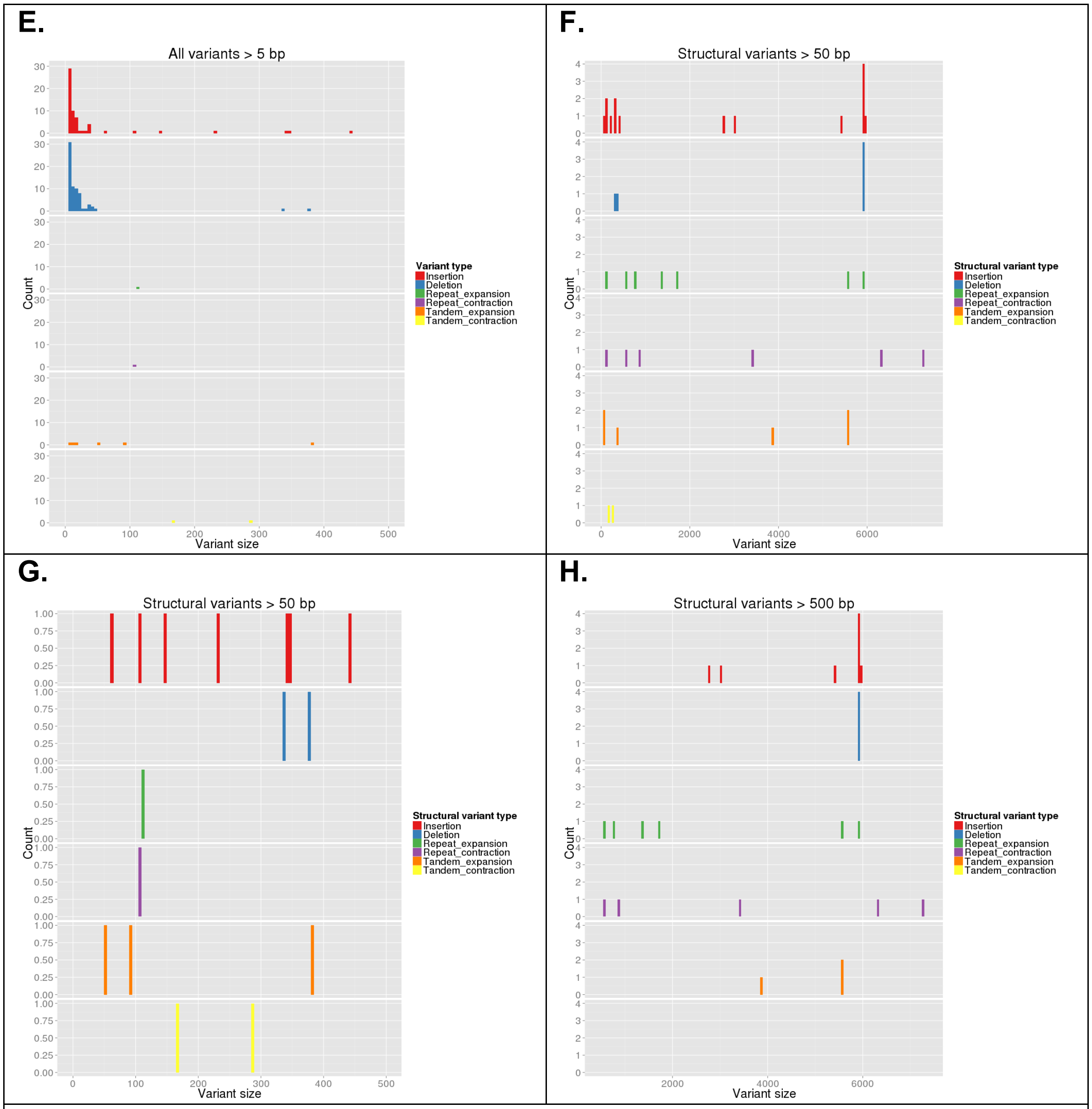
Supplementary Figure 6. Assemblytics results for the *Arabidopsis thaliana* genome assembly from the MHAP study. **A.** Dot plot showing alignments of the assembly (query) to the reference before filtering by Assemblytics. **B.** Dot plot after Assemblytics unique anchor filtering. **C.** Cumulative sequence length plot of both the reference and the query. **D.** Variant type and size histogram of all variants above 5 bp. **E.** Variant type and size histogram of all variants 5-500 bp. **F.** Variant type and size histogram of all variants above 50 bp. **G.** Variant type and size histogram of all variants 50-500 bp. **H.** Variant type and size histogram of all variants above 500 bp.


A.
Dot plot of unfiltered alignments
Query
JSAE01000284.1
JSAE01000167.1
JSAE01000572.1
JSAE01000524.1
JSAE01000548.1
X
Y
2L
2R
3L
3R
Reference
B.
Dot plot of Assemblytics filtered alignments
Query
JSAE01000284.1
JSAE01000167.1
JSAE01000572.1
JSAE01000524.1
JSAE01000548.1
X
Y
2L
2R
3L
3R
Reference
C.
Sequence length
10^7
10^6
10^5
10^4
0
25
50
75
100
NG(x)% where 100% = 149844171 bp
Assembly
Reference
Query
D.
All variants > 5 bp
Count
300
200
100
0
0
2500
5000
7500
10000
Variant size
Variant type
Insertion
Deletion
Repeat_expansion
Repeat_contraction
Tandem_expansion
Tandem_contraction

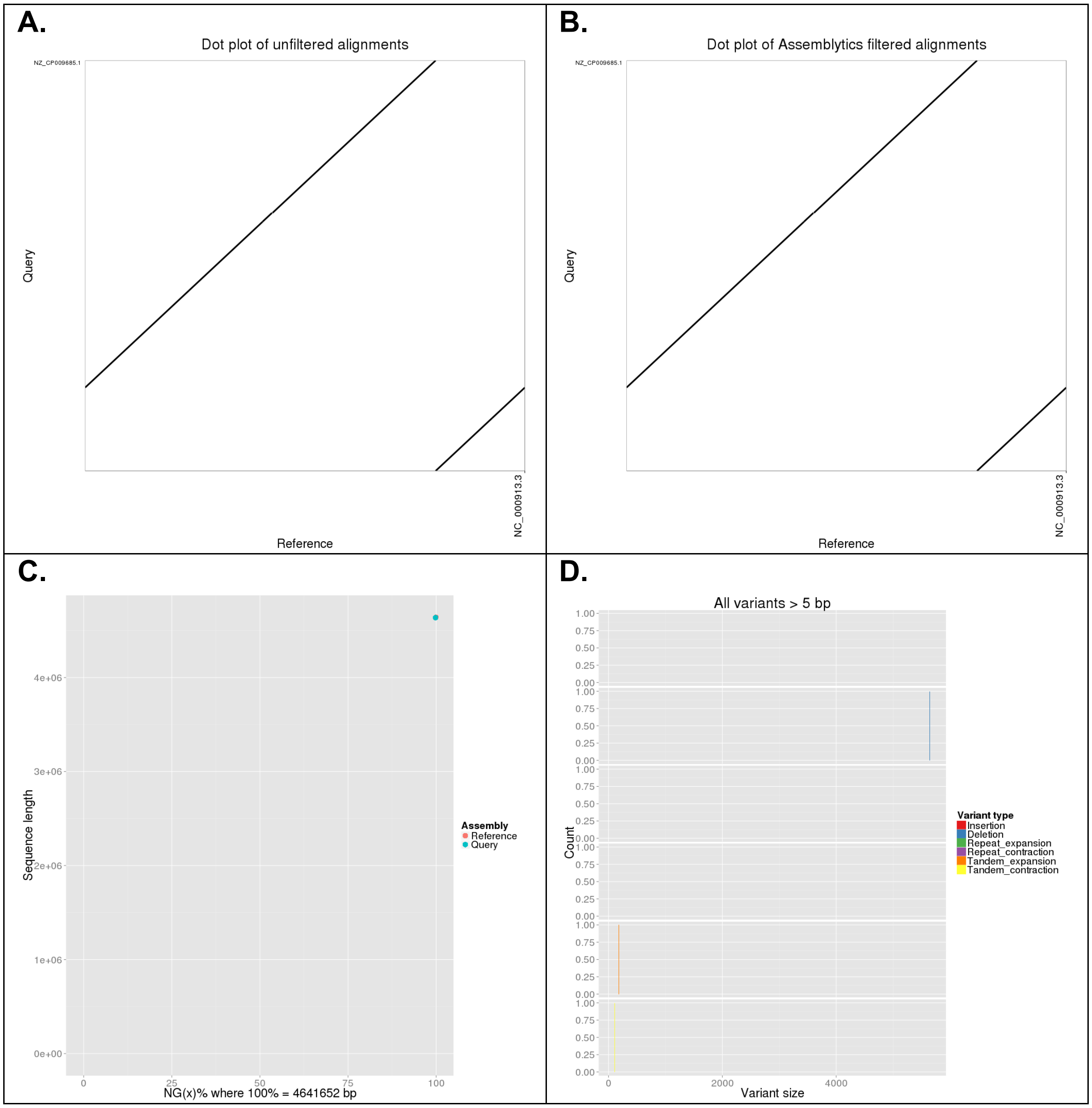


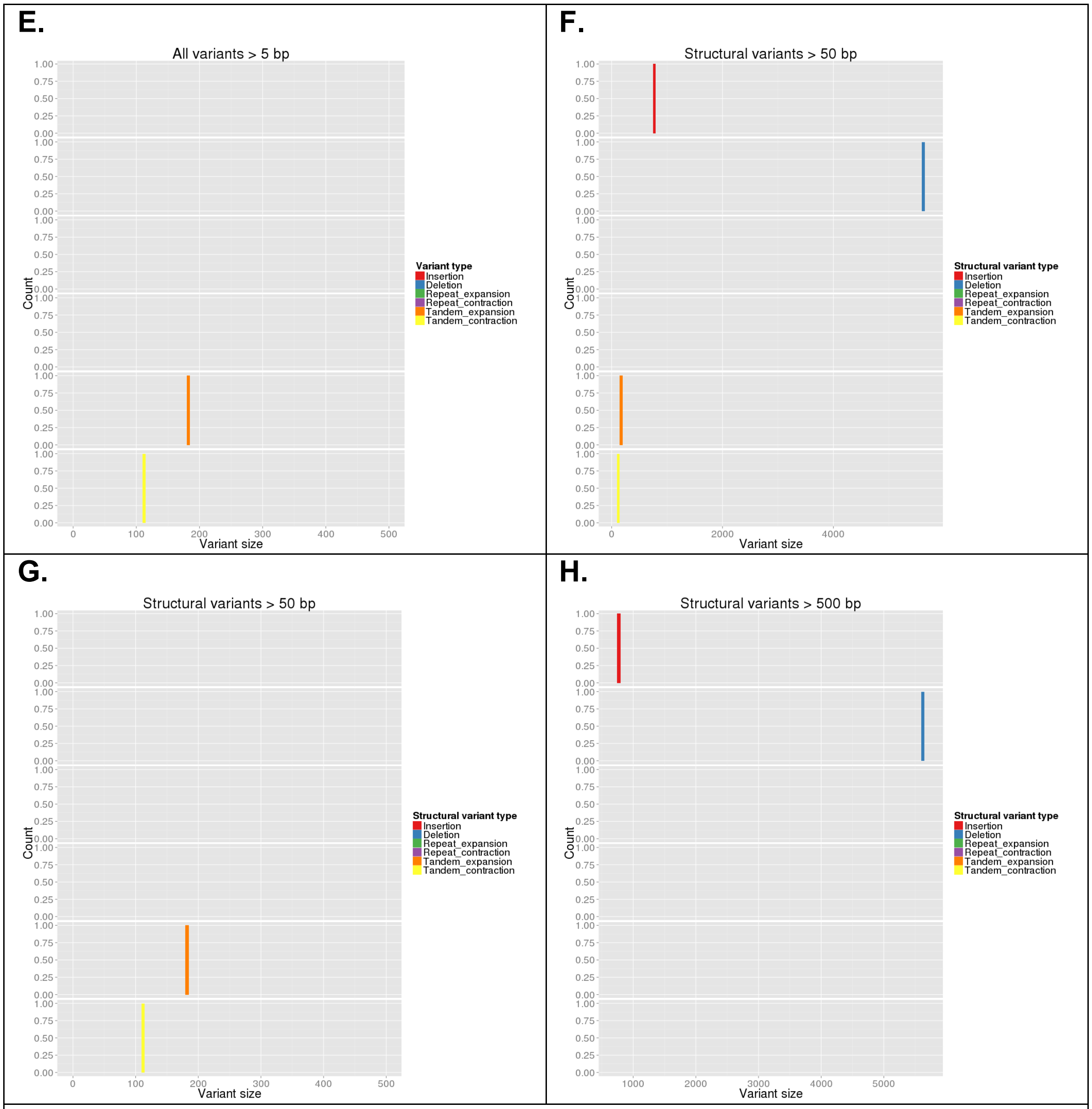
Supplementary Figure 7. Assemblytics results for the *Drosophila melanogaster* genome assembly from the MHAP study. **A.** Dot plot showing alignments of the assembly (query) to the reference before filtering by Assemblytics. **B.** Dot plot after Assemblytics unique anchor filtering. **C.** Cumulative sequence length plot of both the reference and the query. **D.** Variant type and size histogram of all variants above 5 bp. **E.** Variant type and size histogram of all variants 5-500 bp. **F.** Variant type and size histogram of all variants above 50 bp. **G.** Variant type and size histogram of all variants 50-500 bp. **H.** Variant type and size histogram of all variants above 500 bp.


A.
Dot plot of unfiltered alignments
Query
JSAC01000003.1
JSAC01000002.1
JSAC01000005.1
JSAC01000004.1
JSAC01000036.1
JSAC01000038.1
JSAC01000011.1
JSAC01000007.1
JSAC01000012.1
JSAC01000027.1
JSAC01000001.1
JSAC01000014.1
JSAC01000021.1
JSAC01000026.1
JSAC01000013.1
JSAC01000006.1
Reference
ref|NC_001134|
ref|NC_001135|
ref|NC_001136|
ref|NC_001137|
ref|NC_001138|
ref|NC_001139|
ref|NC_001140|
ref|NC_001141|
ref|NC_001142|
ref|NC_001143|
ref|NC_001144|
ref|NC_001145|
ref|NC_001146|
ref|NC_001147|
ref|NC_001148|
B.
Dot plot of Assemblytics filtered alignments
Query
JSAC01000003.1
JSAC01000002.1
JSAC01000005.1
JSAC01000004.1
JSAC01000036.1
JSAC01000038.1
JSAC01000011.1
JSAC01000007.1
JSAC01000012.1
JSAC01000027.1
JSAC01000001.1
JSAC01000014.1
JSAC01000021.1
JSAC01000026.1
JSAC01000013.1
JSAC01000006.1
Reference
ref|NC_001134|
ref|NC_001135|
ref|NC_001136|
ref|NC_001137|
ref|NC_001138|
ref|NC_001139|
ref|NC_001140|
ref|NC_001141|
ref|NC_001142|
ref|NC_001143|
ref|NC_001144|
ref|NC_001145|
ref|NC_001146|
ref|NC_001147|
ref|NC_001148|
C.
Sequence length
10^6
$10^{5.5}$
10^5
$10^{4.5}$
10^4
0
25
50
75
100
NG(x)% where 100% = 12283855 bp
Assembly
Reference
Query
D.
All variants > 5 bp
Count
30
20
10
0
Variant size
0
2000
4000
6000
Variant type
Insertion
Deletion
Repeat_expansion
Repeat_contraction
Tandem_expansion
Tandem_contraction



Supplementary Figure 8. Assemblytics results for the *Saccharomyces cerevisiae* genome assembly from the MHAP study. **A.** Dot plot showing alignments of the assembly (query) to the reference before filtering by Assemblytics. **B.** Dot plot after Assemblytics unique anchor filtering. **C.** Cumulative sequence length plot of both the reference and the query. **D.** Variant type and size histogram of all variants above 5 bp. **E.** Variant type and size histogram of all variants 5-500 bp. **F.** Variant type and size histogram of all variants above 50 bp. **G.** Variant type and size histogram of all variants 50-500 bp. **H.** Variant type and size histogram of all variants above 500 bp.


A.
Dot plot of unfiltered alignments
NZ_CP009685.1
Query
NC_000913.3
Reference
B.
Dot plot of Assemblytics filtered alignments
NZ_CP009685.1
Query
NC_000913.3
Reference
C.
4e+06
3e+06
2e+06
1e+06
0e+00
Sequence length
0
25
50
75
100
NG(x)% where 100% = 4641652 bp
Assembly
Reference
Query
D.
All variants > 5 bp
1.00
0.75
0.50
0.25
0.00
Count
0
2000
4000
Variant size
Variant type
Insertion
Deletion
Repeat_expansion
Repeat_contraction
Tandem_expansion
Tandem_contraction



Supplementary Figure 9. Assemblytics results for the *Escherichia coli* genome assembly from the MHAP study. **A.** Dot plot showing alignments of the assembly (query) to the reference before filtering by Assemblytics. **B.** Dot plot after Assemblytics unique anchor filtering. **C.** Cumulative sequence length plot of both the reference and the query. **D.** Variant type and size histogram of all variants above 5 bp. **E.** Variant type and size histogram of all variants 5-500 bp. **F.** Variant type and size histogram of all variants above 50 bp. **G.** Variant type and size histogram of all variants 50-500 bp. **H.** Variant type and size histogram of all variants above 500 bp.