THE ASSEMBLATHON

N50 must die?

Genome assembly workshop, Santa Cruz, 3/15/11

twitter: @assemblathon web: assemblathon.org

Should N50 die in its role as a frequently used measure of genome assembly quality? Are there other measures out there? Are there other measures that are better? Read on to find out more.

Overview

Two parts of the talk. First a brief recap of what the Assemblathon is all about and who is in it, then a look at some of the metrics we have been working with.

UC Davis

Genome Center

Ian Korf

Dawei Lin

Keith Bradnam

Aaron Darling

Joseph Fass

Ken Yu

Vince Buffalo

UC Santa Cruz

Haussler Lab

David Haussler

Dent Earl

Benedict Paten

John St. John

Ngan Nguyen

Two groups involved in the evaluation of the Assemblathon. David Haussler's group will talk next.

The challenge

Assemble genome of species ‘A’

112 MB diploid synthetic genome

Reference genome of species ‘B’ available

2 paired-read + 2 mate pairs libraries

Species A and B were created using the EVOLVER suite of software and have an estimated divergence date of 100 million years ago. EVOLVER models many different types of evolutionary change (substitutions, indels, translocations, etc.) and has models for repeat evolution. Synthetic Illumina reads were created from the species A genome. Most groups did not use species B info.

Goals

What makes a good *assembler*?

First goal is most obvious. Expect lots of discussion about this during the workshop.

Goals

What makes a good *assembly*?

This is less obvious. The best assembler in the world doesn't amount to much if other errors are made when putting the final assembly together. One of the checks that we put into the Assemblathon (E. coli contamination of known genome), helps differentiate between good assemblers and good assemblies.

Goals

What makes a good *assembly metric*?

Can we move beyond a reliance on N50? Can we see which measures work well together and which are revealing different aspects of a good genome assembly? Are there easy-to-calculate metrics that perform just as well as complex, computationally-involved metrics?

Assemblathon by the numbers



7 countries




Assemblathon by the numbers

17 teams

Assemblathon by the numbers

62 assemblies

21 by UC Davis

If you want to develop a good assembly metric, it is useful to have both good and bad assemblies to measure. Some of our 21 assemblies are intentionally bad, e.g. not using mate pair information. Bad assemblies should score badly, so these assemblies provide an extra check on how useful our metrics are.

Assemblathon by the numbers

>100 metrics

Many of which are clearly correlated to each other. Altogether, the metrics tackle the ‘what makes a good assembly?’ question from many different angles. Some are simple statistical descriptions of assemblies, and others are new measures that use more complex calculations.

Assemblathon by the numbers

7,949,257,310 bp

This is the sum of all of the assemblies that we have been working with.

Metrics

What are we measuring?

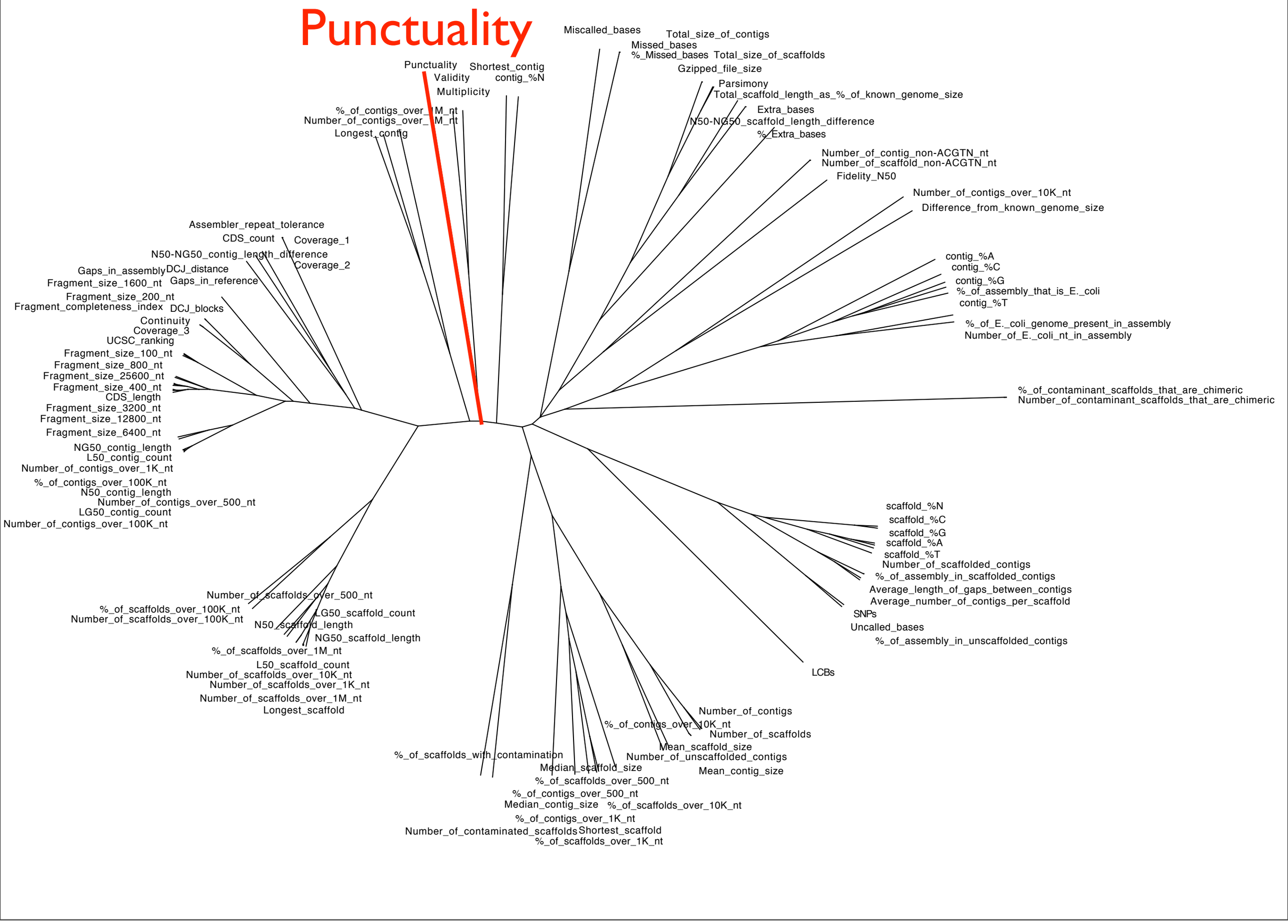
Number of scaffolds
Total size of scaffolds
Total scaffold length as % of known genome size
Longest scaffold
Shortest scaffold
Number of scaffolds > 500 nt
% of scaffolds > 500 nt
Number of scaffolds > 1K nt
% of scaffolds > 1K nt
Number of scaffolds > 10K nt
% of scaffolds > 10K nt
Number of scaffolds > 100K nt
% of scaffolds > 100K nt
Number of scaffolds > 1M nt
% of scaffolds > 1M nt
Mean scaffold size
Median scaffold size
N50 scaffold length
L50 scaffold count
NG50 scaffold length
LG50 scaffold count
scaffold %A
scaffold %C
scaffold %G
scaffold %T
scaffold %N
scaffold %non-ACGTN

N50-NG50 scaffold length difference
contig %A
contig %C
contig %G
contig %T
contig %N
contig %non-ACGTN
Number of contig non-ACGTN nt
Gzipped file size
Difference from known genome size
Punctuality
Number of E. coli nt in assembly
% of assembly that is E. coli
Number of contaminated scaffolds
% of scaffolds with contamination
Number of contaminant scaffolds that are chimeric
% of contaminant scaffolds that are chimeric
% of E. coli genome present in assembly
Extra bases
% Extra bases
Missed bases
% Missed bases
CDS count
CDS length
Number of scaffold non-ACGTN nt
N50-NG50 contig length difference

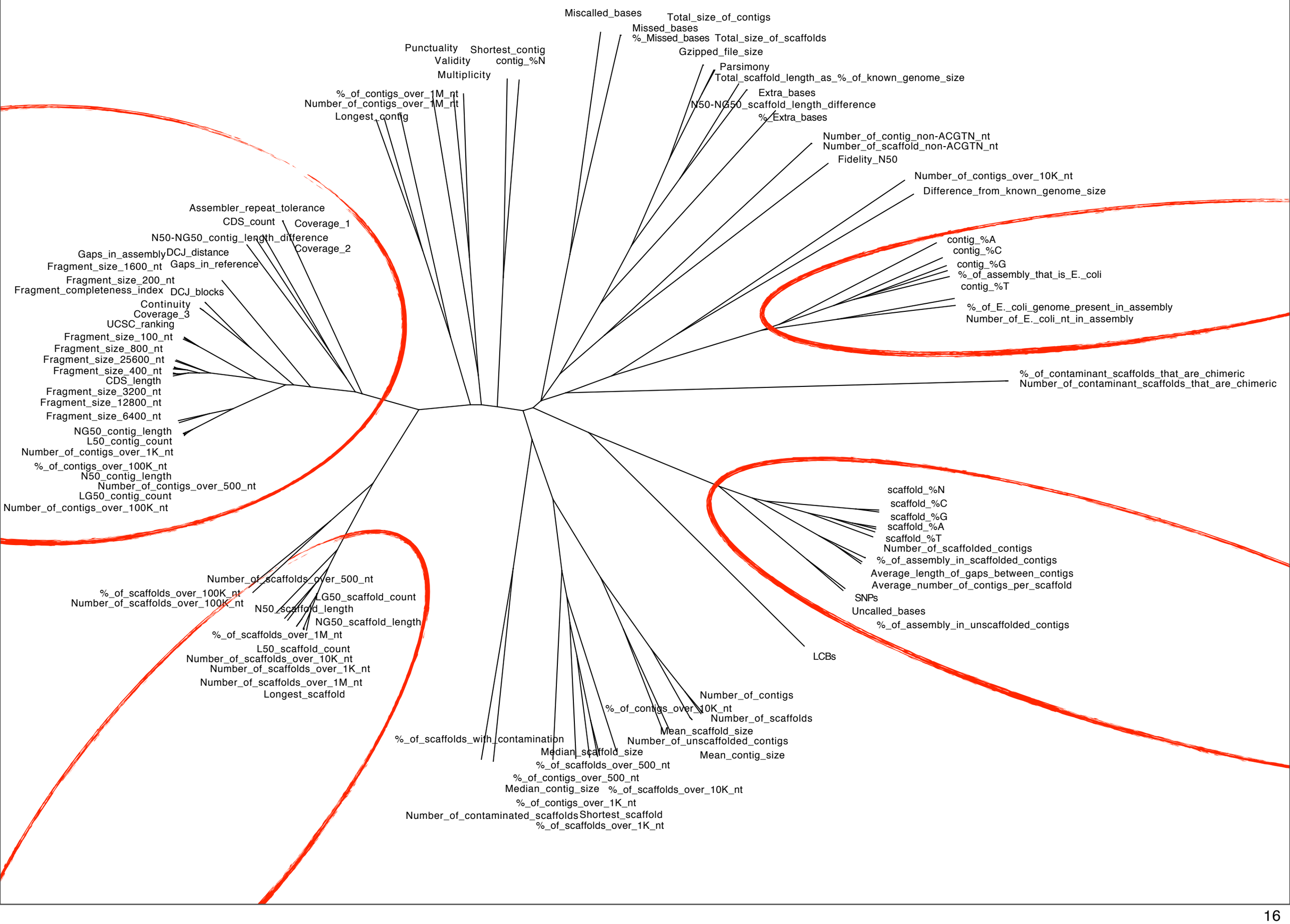
Fragment size 100 nt
Fragment size 200 nt
Fragment size 400 nt
Fragment size 800 nt
Fragment size 1600 nt
Fragment size 3200 nt
Fragment size 6400 nt
Fragment size 12800 nt
Fragment size 25600 nt
Fragment completeness index
Assembler repeat tolerance
Coverage 1
Coverage 2
Coverage 3
Continuity
Fidelity N50
Validity
Multiplicity
Parsimony
UCSC ranking
LCBs
DCJ distance
DCJ blocks
SNPs
Miscalled bases
Uncalled bases

% of assembly in scaffolded contigs
% of assembly in unscaffolded contigs
Average number of contigs per scaffold
Average length of gaps between contigs
Number of contigs
Number of scaffolded contigs
Number of unscaffolded contigs
Total size of contigs
Longest contig
Shortest contig
Number of contigs > 500 nt
% of contigs > 500 nt
Number of contigs > 1K nt
% of contigs > 1K nt
Number of contigs > 10K nt
% of contigs > 10K nt
Number of contigs > 100K nt
% of contigs > 100K nt
Number of contigs > 1M nt
% of contigs > 1M nt
Mean contig size
Median contig size
N50 contig length
L50 contig count
NG50 contig length
LG50 contig count
Gaps in reference
Gaps in assembly

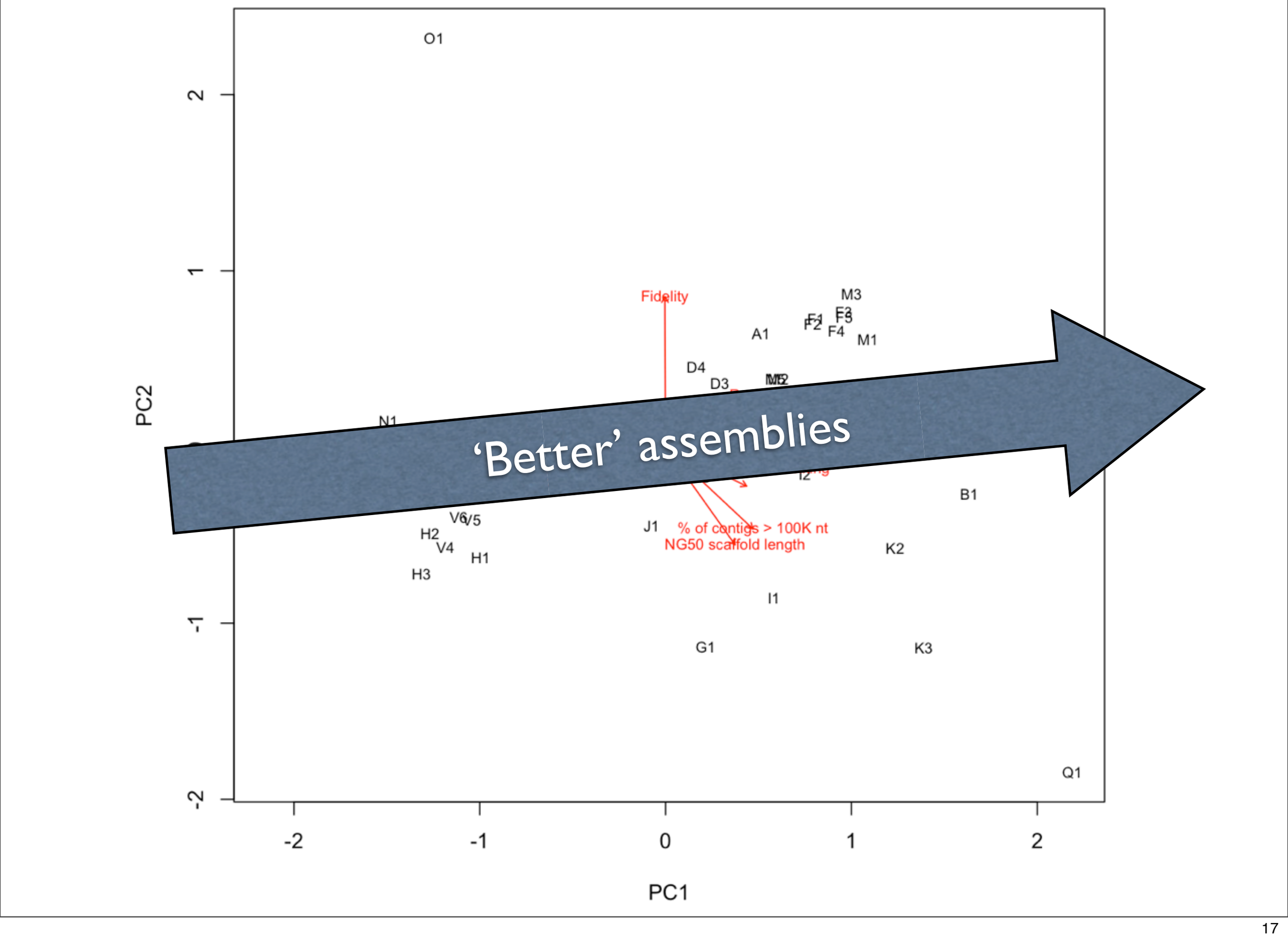
These are all the UC Davis metrics that we have been working with. Some are simple, others quite complex. Many are closely related to each others.



You can see how well any two metrics agree and from the strength of the correlation, produce a distance, and therefore produce a tree. One metric was based on how long it took groups to submit an assembly. Thankfully, this doesn't seem to correlate with any quality metrics. So groups who submitted later, did not seem to gain any obvious advantage.



Many metrics are clearly unrelated to each other, but some metrics group very closely together. Not always in the most expected fashion. E.g. % of assembly that is E. coli is highly correlated with contig %G



Principal components analysis can be used to see if assemblies can be grouped in any meaningful way. Overall, the first principal component (based on 9 selected metrics) seems to do a good job at ordering assemblies based on their overall quality.

N50

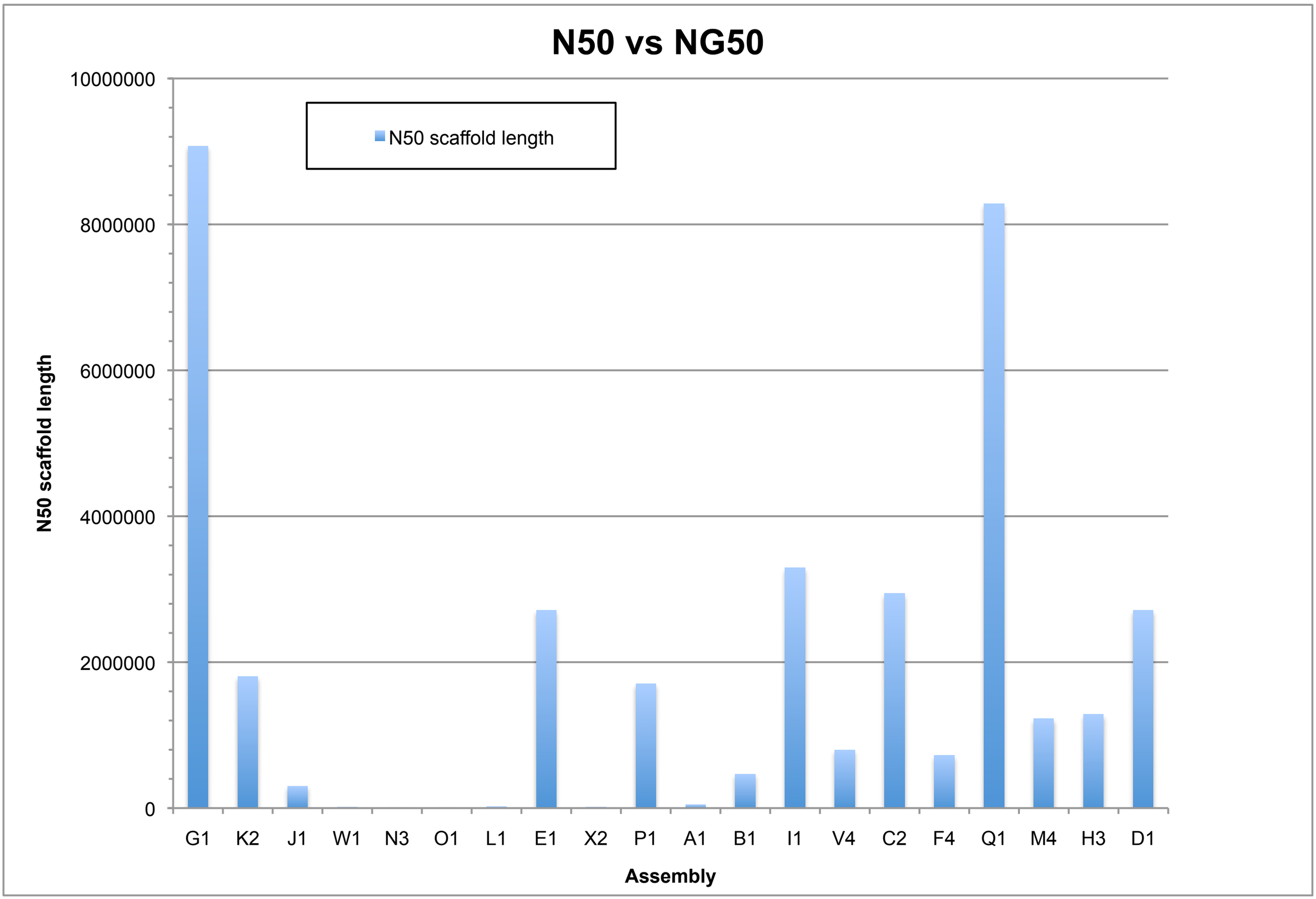
Scaffold/contig length at which you have covered 50% of total assembly length

A reminder of the most traditional assembly metric. This measure is biased if you choose to exclude lots of short sequences from your assembly. It also can not be fairly compared to other assemblies which are different in size.

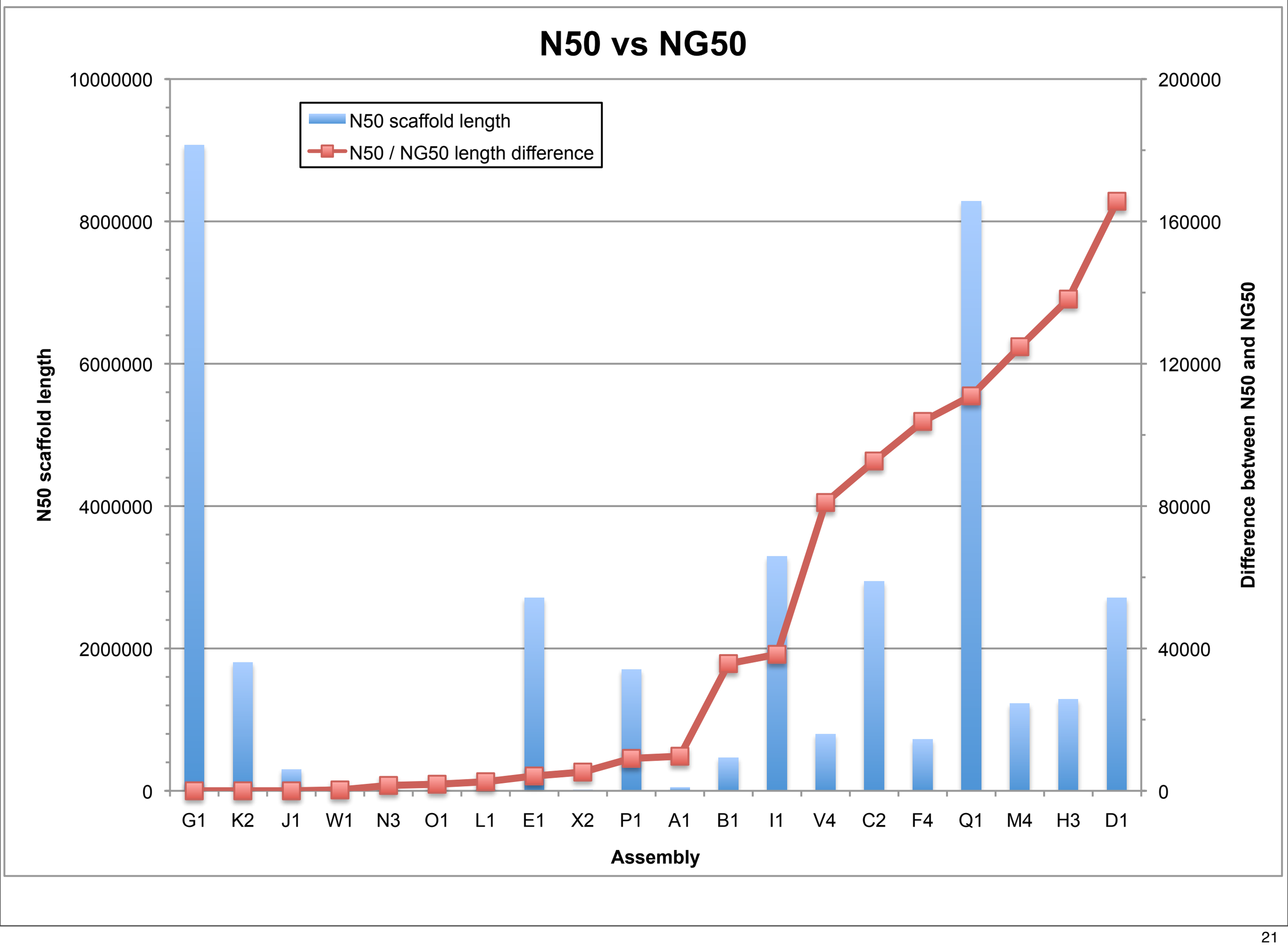
NG50

Scaffold/contig length at which you have covered 50% of total *genome* length

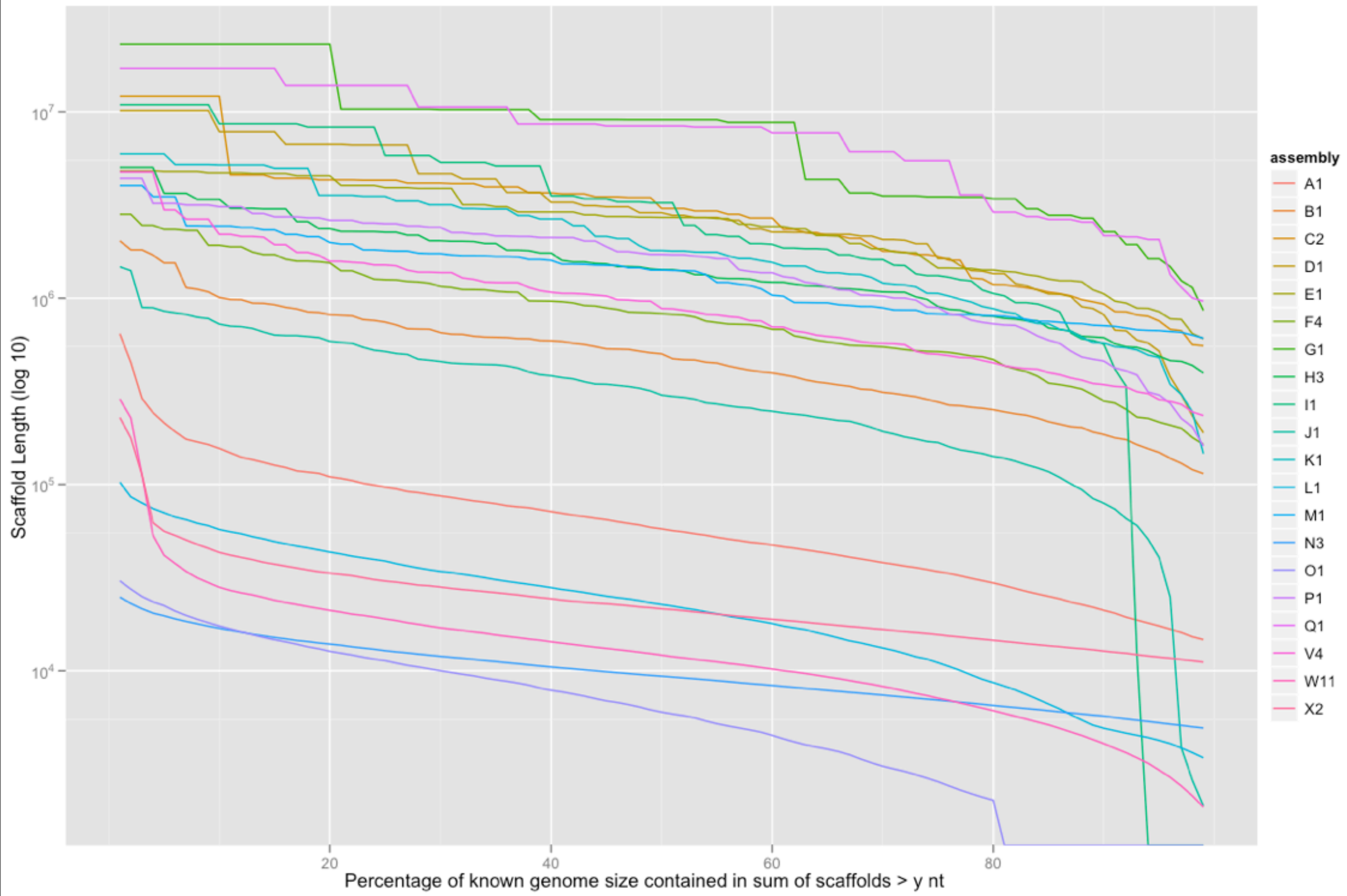
We prefer a measure that we are calling ‘NG50’. All calculations use the length of the known genome as the denominator. Can now compare assemblies to each other.



Huge variation in assemblies when looking at N50. G1 and Q1 are clear winners by this metric. Only showing best assembly from each team.



Second series (in red) shows difference between N50 and NG50. Assemblies are ordered in increasing N50/NG50 difference. Some assemblies have no difference, some have large differences (over 160 kbp).



This NG(X) graph shows the values not just of NG50, but all values from NG1 through to NG99. Y-axis is on a log scale. This graph allows you compare all assemblies in a visual manner. Total area under the curve could be used as another assembly metric.

Fragment analysis

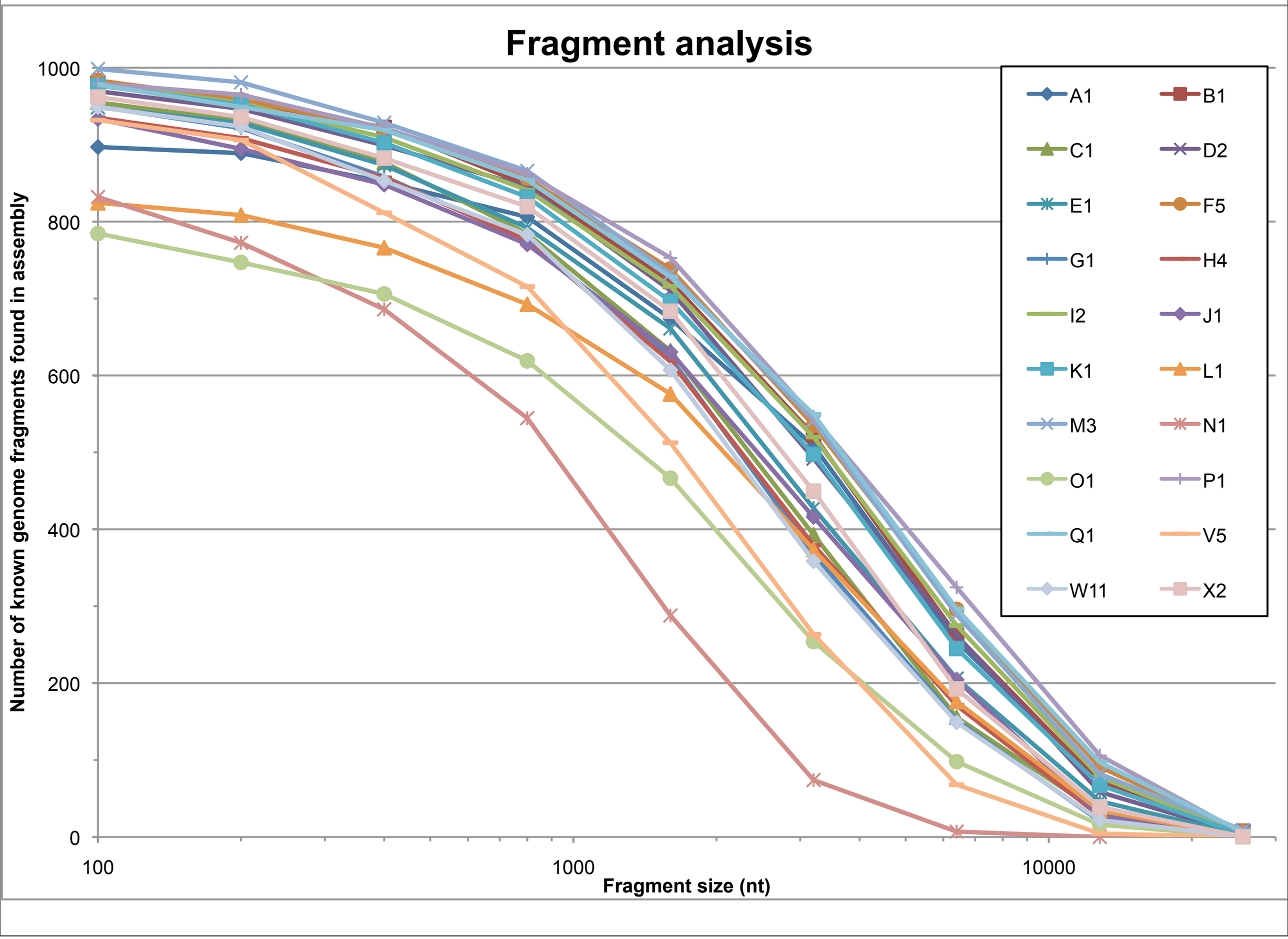
Count how many randomly chosen fragments from species A genome can be found in assembly

Use BLASTN, 95% identity over 95% length

Repeat at different fragment sizes

Repeat for both species A haplotypes

We use haplotype A1 and A2 separately, and average the results. Can then plot the number of matching fragments as a function of fragment length. Easy to visually compare lots of assemblies at once.

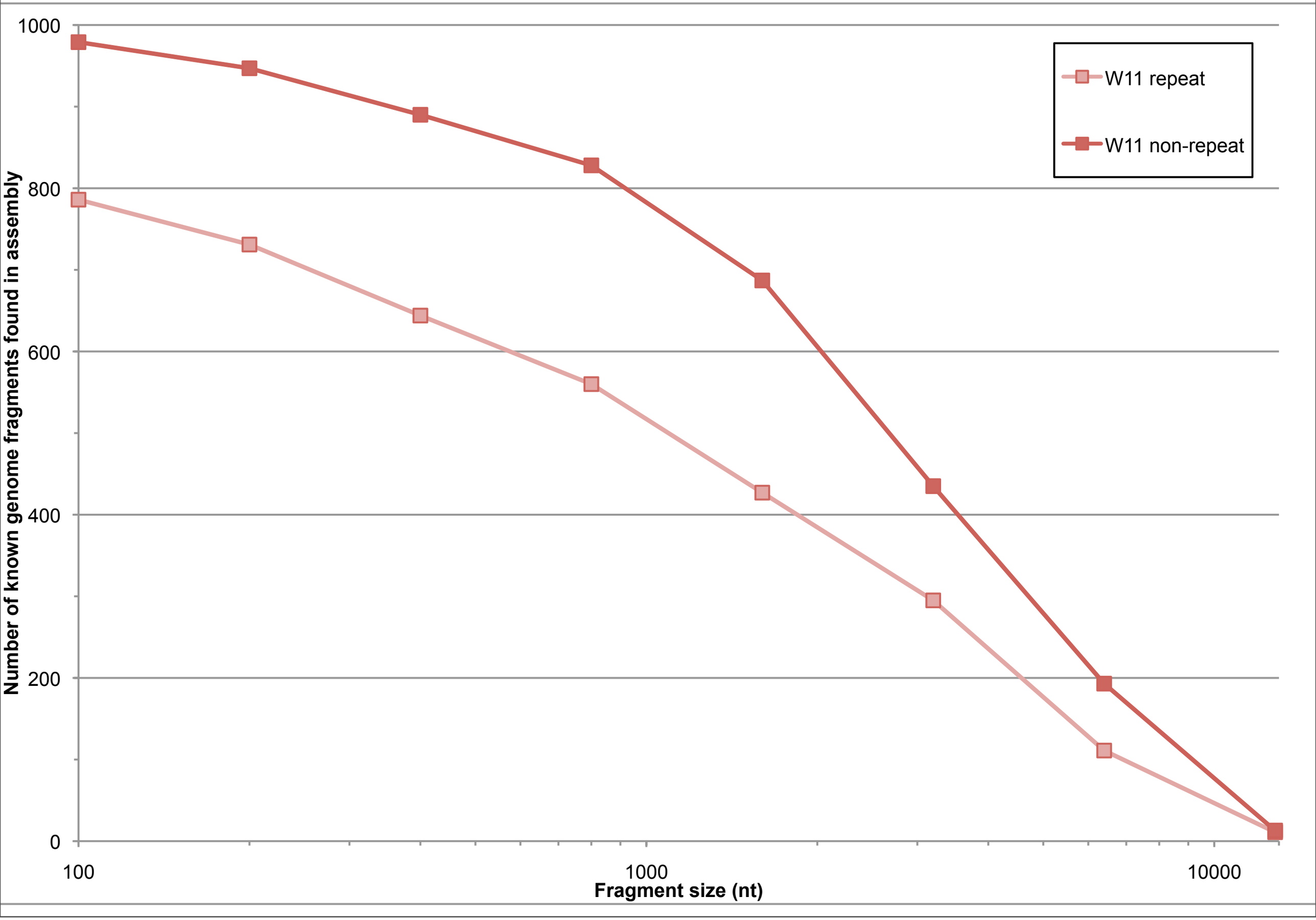


x-axis is log scale. Assemblies do relatively well at containing near-perfect regions of the known genome as long as fragments are short. Longer fragments can not be found because of errors and the problem of dealing with haplotype differences. The total area under the curve serves as a useful metric.

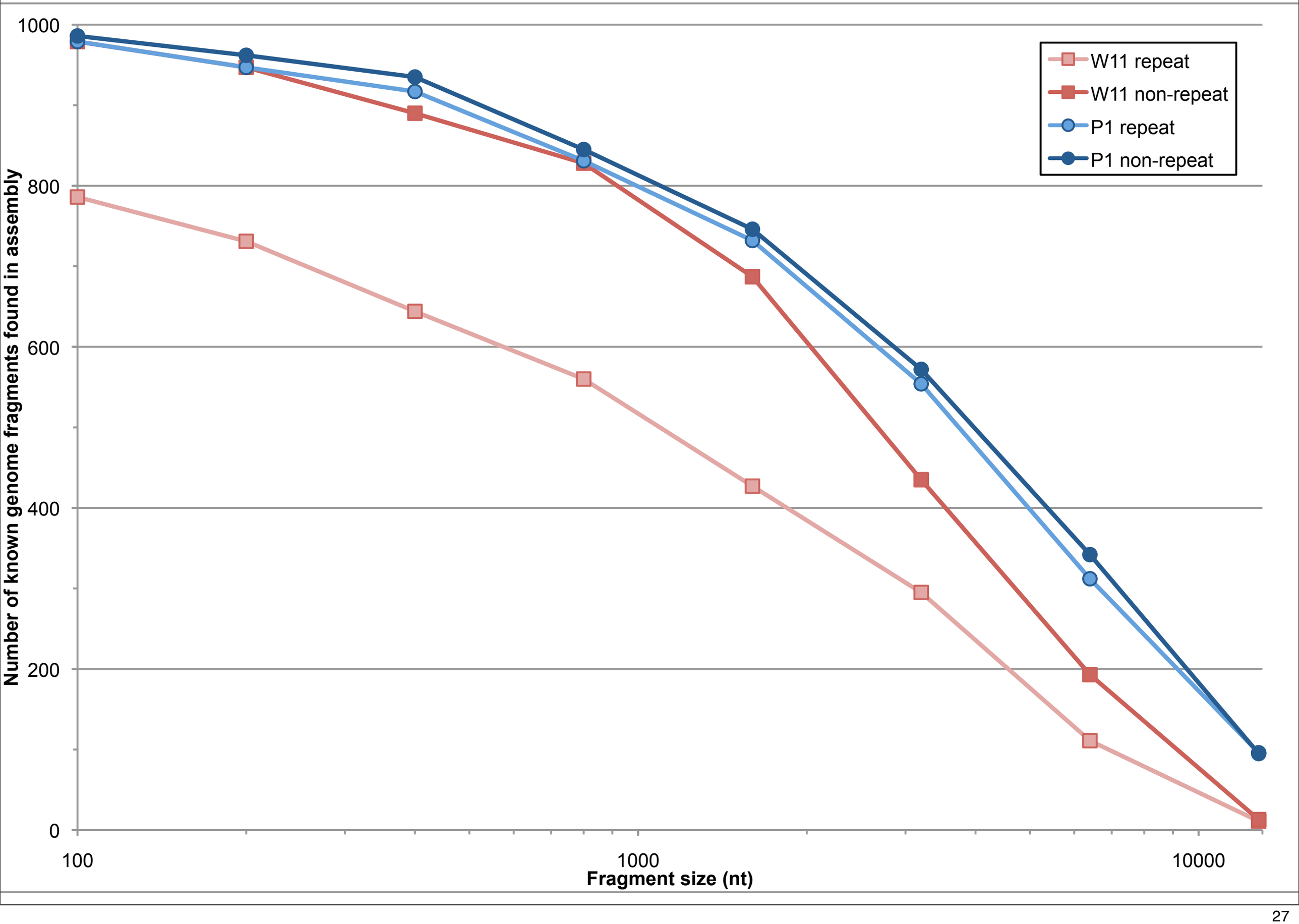
Repeat analysis

Choose fragments that either overlap or don't overlap a known repeat

Can separate out fragments on the basis of whether they overlap any sort of repeat in the known genome. This allows us to see how well different assemblers deal with assembling repeats. About 6% of the species A genome is a repeat of one kind or the other (from homopolymer runs up to transposons).



The (intentionally bad) W11 assembly does a poor job at coping with repeats, and the two lines on the above graph are far apart. The assembly contains fewer fragments of the species A genome that overlap known repeats. The sum difference between the repeat and non-repeat lines also makes for a useful metric.

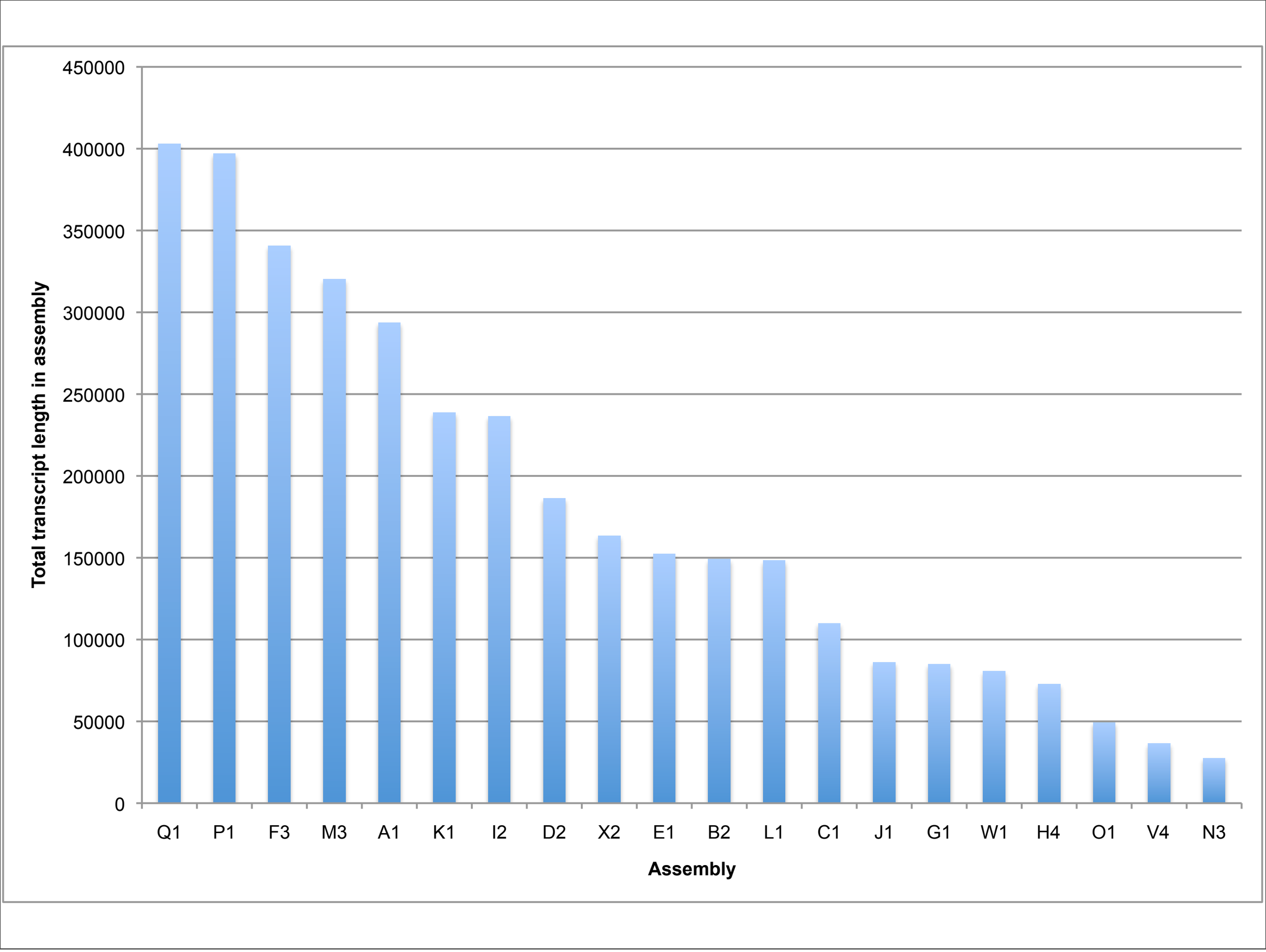


In contrast, the P1 assembly does a very good job at coping with repeats, and the two lines on the above graph are very close together.

Gene finding

How many genes are present in each assembly?

The Species A genome contains 176 genes. For many end users of a genome assembly the genes might be all they care about, and it doesn't matter how long contigs are...as long as they contain a full-length gene.

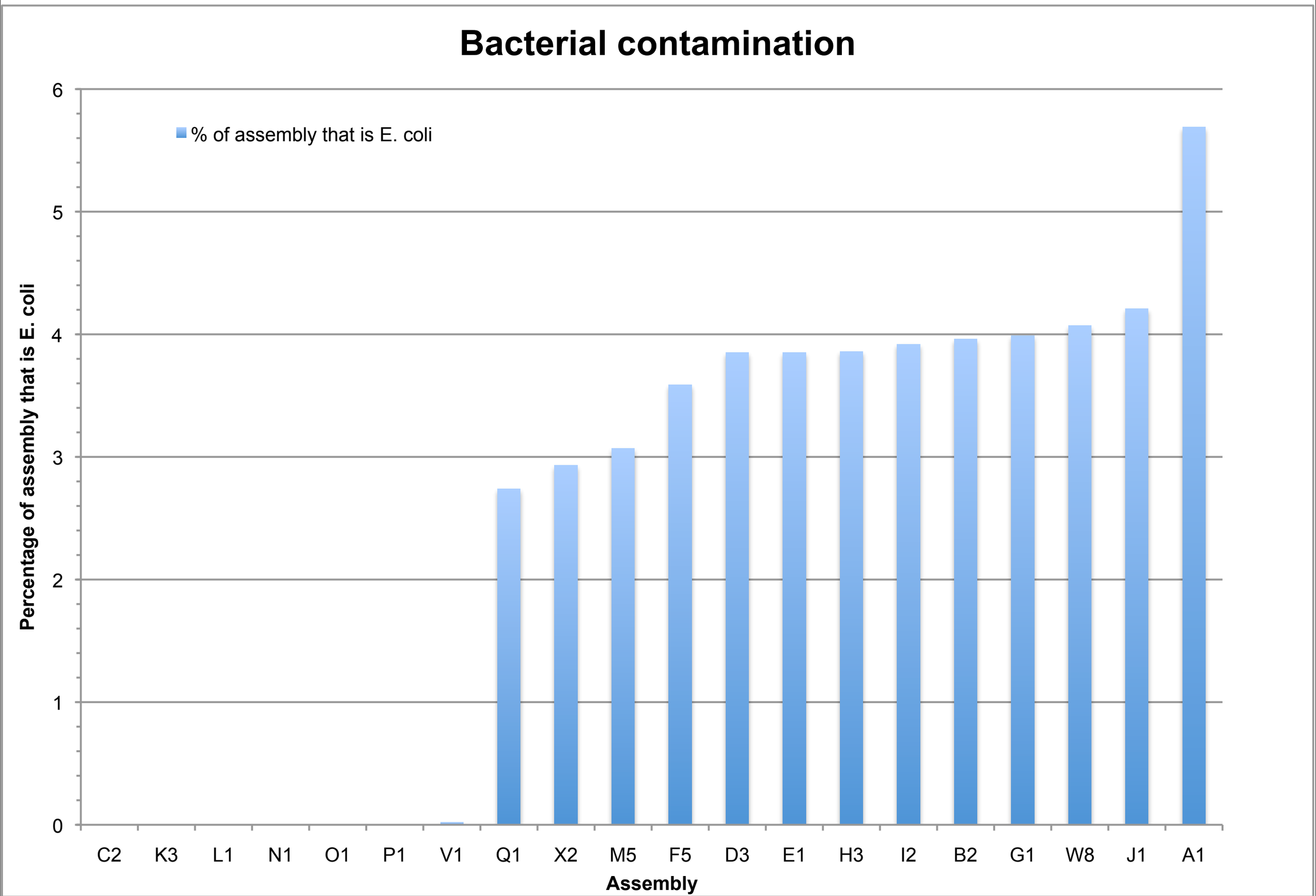


This shows the sum length of all 176 genes that can be found in each assembly.

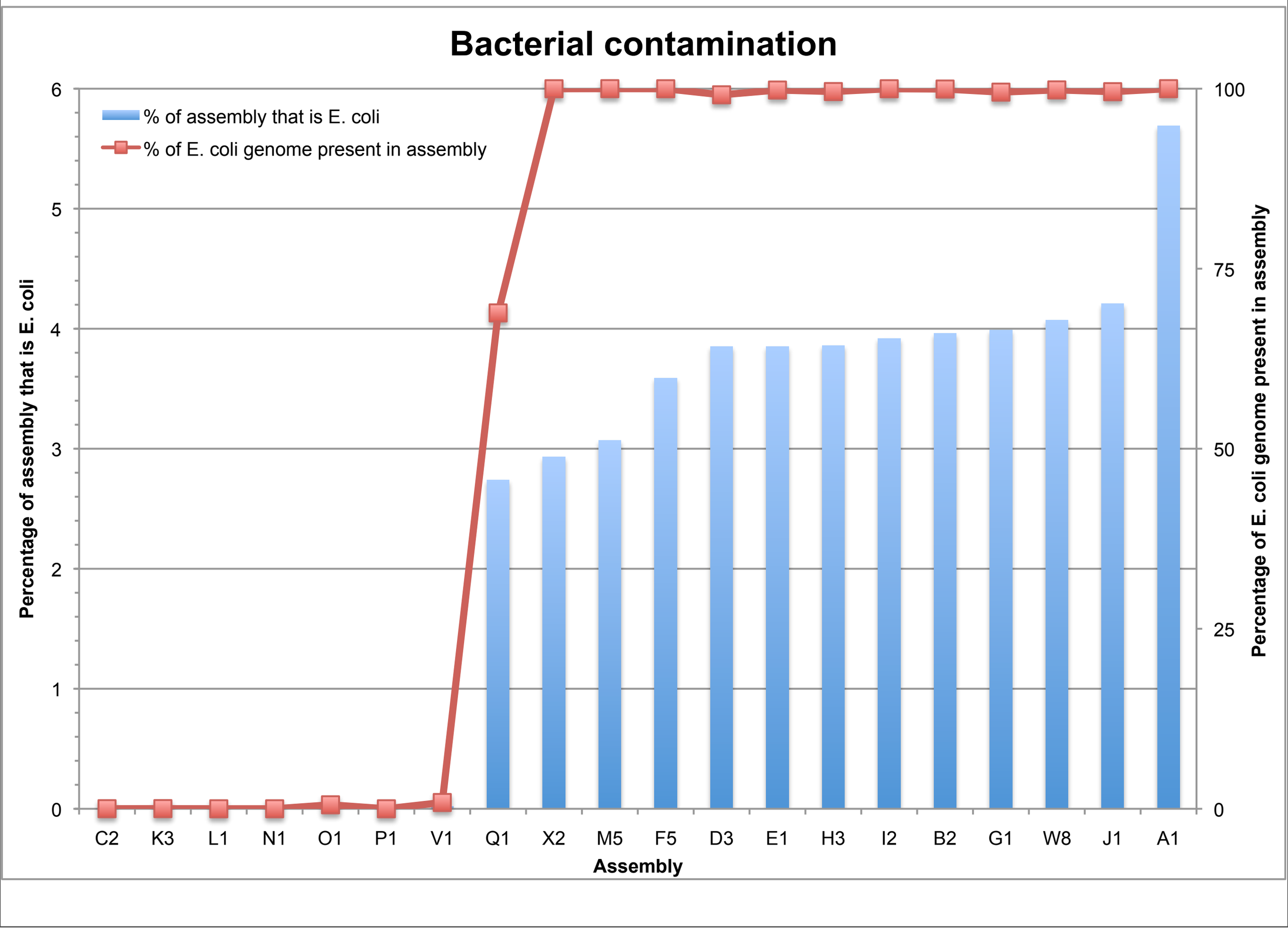
Contamination

"all libraries will contain some bacterial contamination"

This is a line from the README file that accompanied the set of reads that participants could download. About 5% of the reads were taken from the E. coli genome sequence, representing contamination of libraries.



Assemblies either filtered out all contamination or they ended up with the E. coli genome sequence.



The second series (in red) shows how much of the E. coli genome was present in each assembly. In most cases E. coli contamination produced contigs that consisted of just bacterial sequence. But some assemblies also had chimeric contigs with a mix of E. coli and species A sequence.

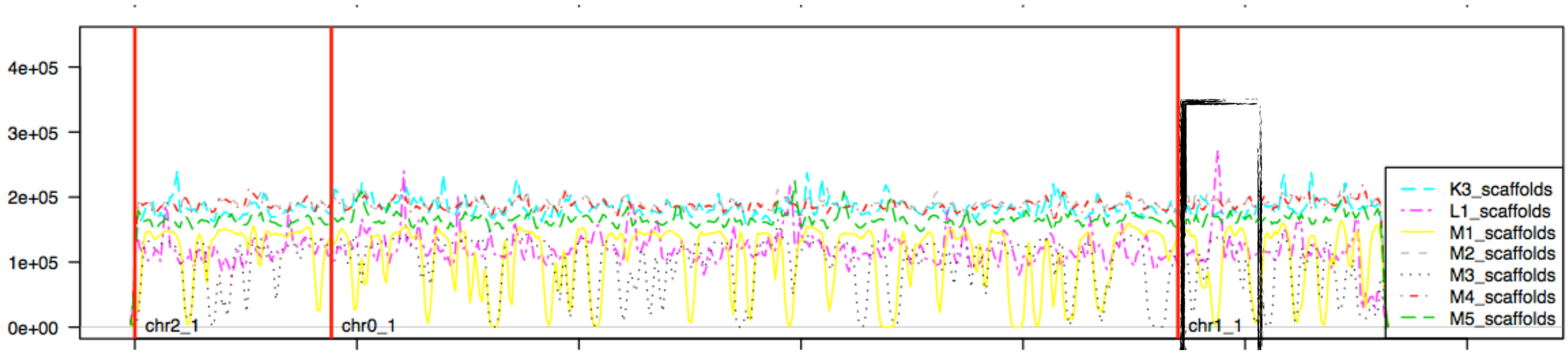
Mauve analysis

Uses whole genome alignment to reveal:

Miscalled bases
Uncalled bases
Missing bases
Extra bases
Misassemblies
Double cut & join distance

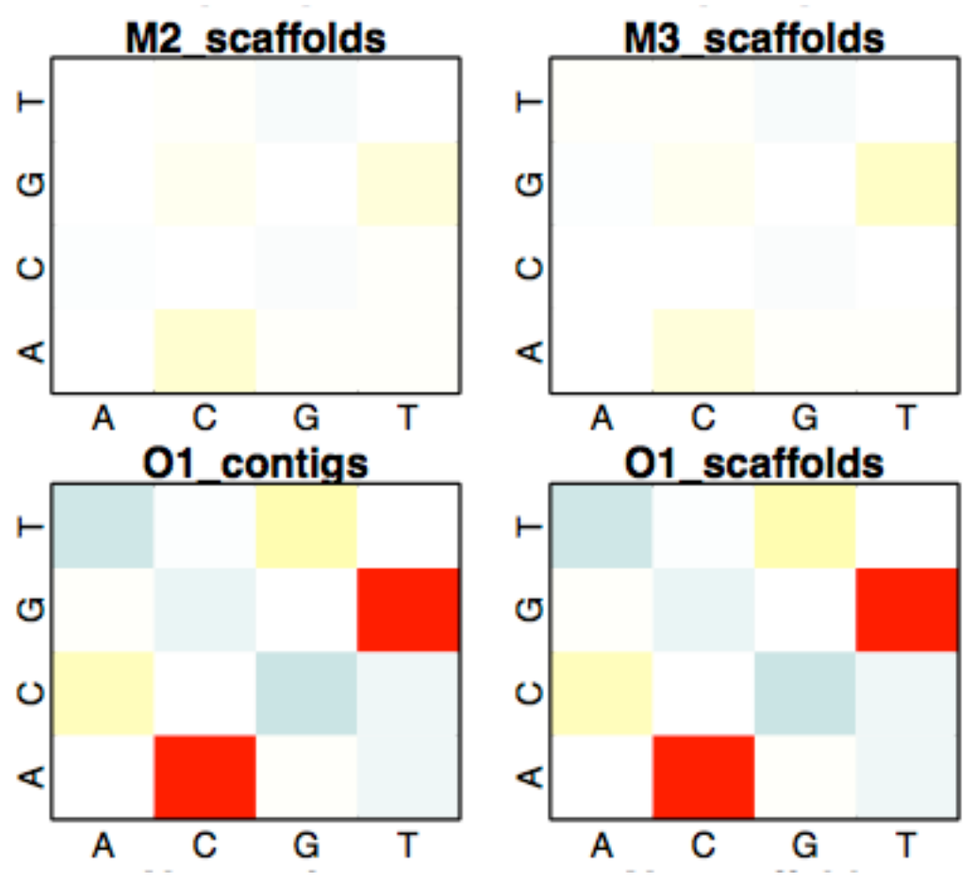
Mauve analysis by Aaron Darling.

Density of miscalled bases



Each panel shows analysis of 7 assemblies. Red lines denote boundaries between the 3 chromosomes of species A. Black box shows an area where many assemblies had a peak of miscalled bases.

Direction of miscalled bases



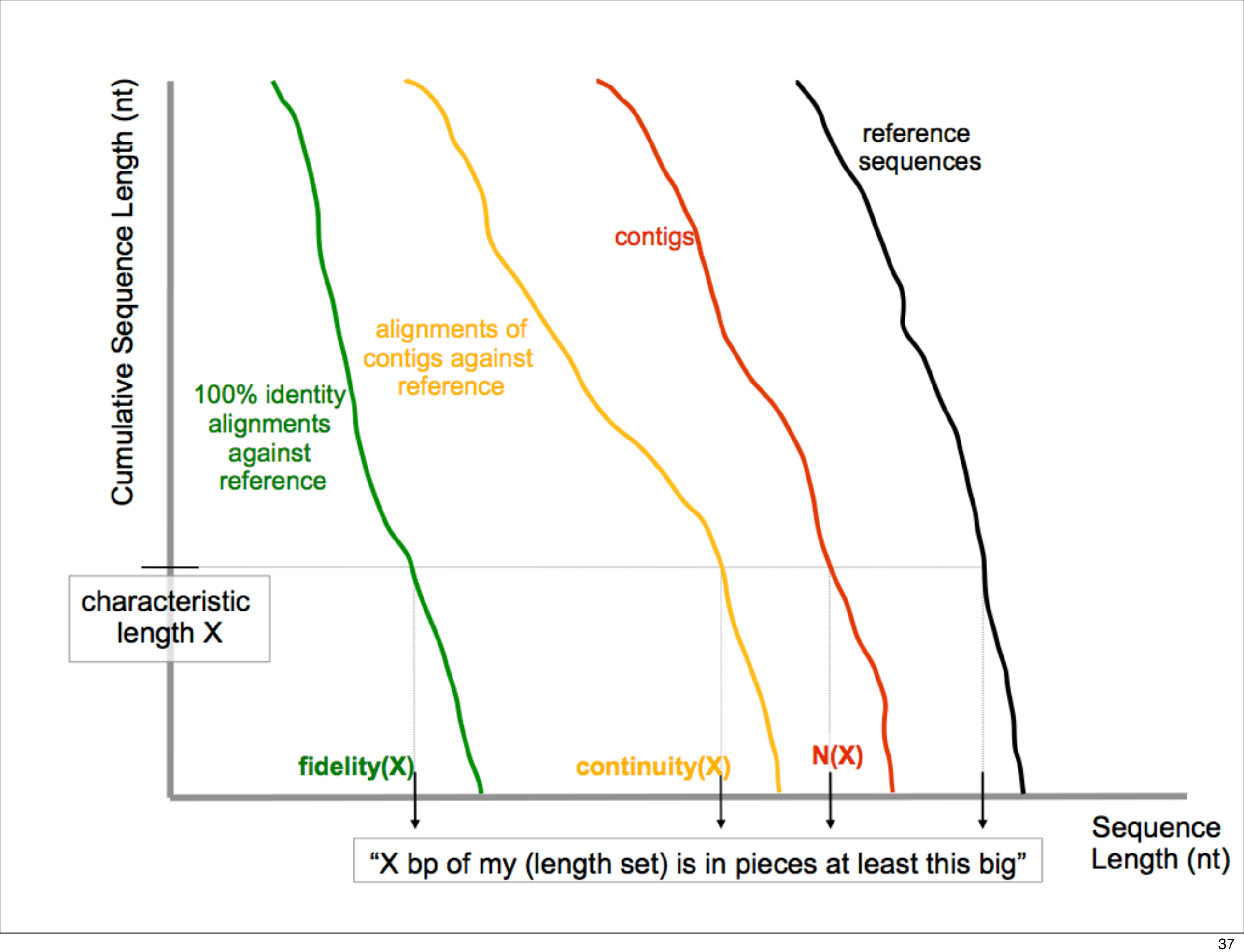
Most miscalled bases had no bias, but some assemblies had miscalled bases that were more likely to be a specific other base as this heat map shows.

BWA analysis

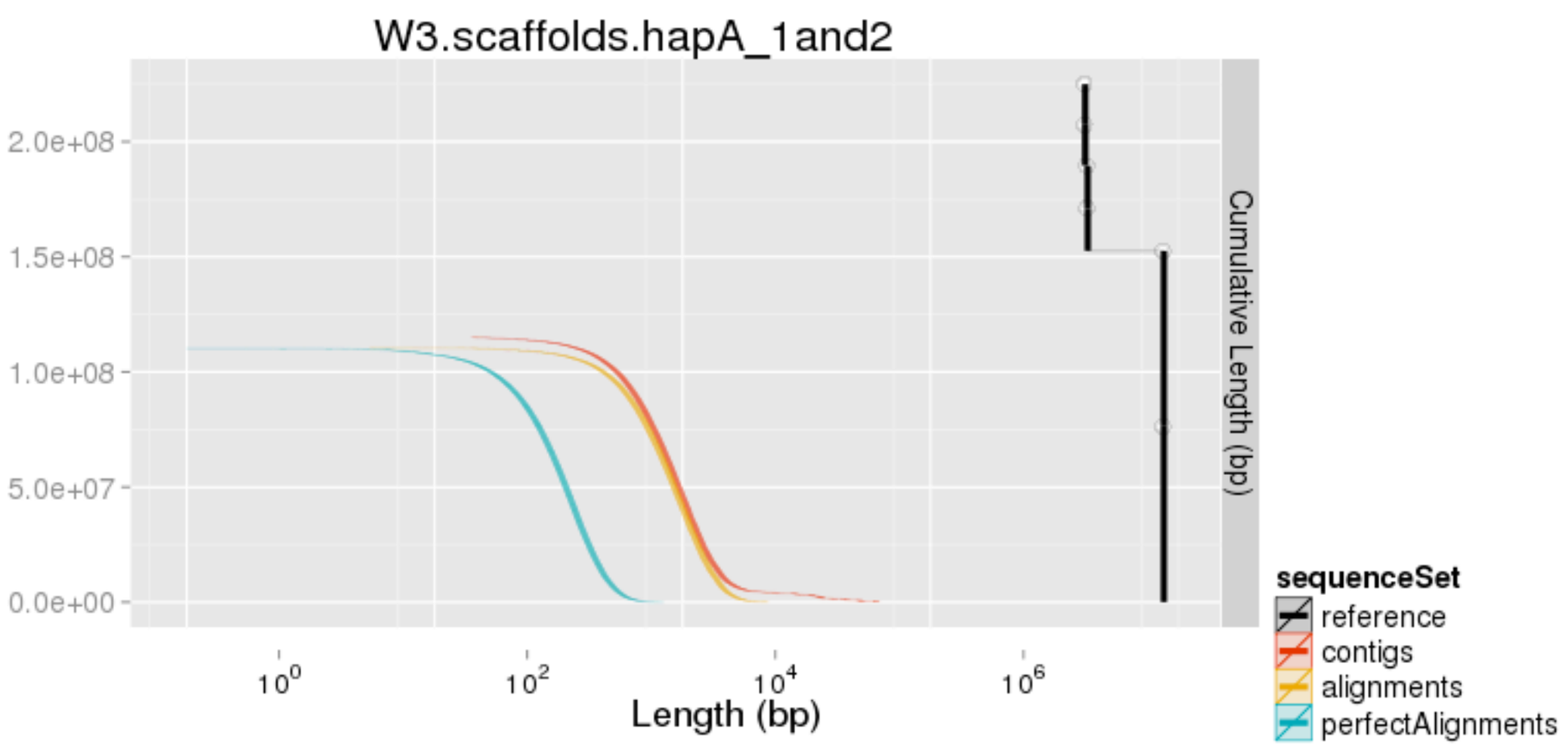
Align contigs to genome to reveal:

What fraction align?
What fraction align *perfectly*?
Lengths of *coverage islands*
Measures of validity, multiplicity, parsimony,

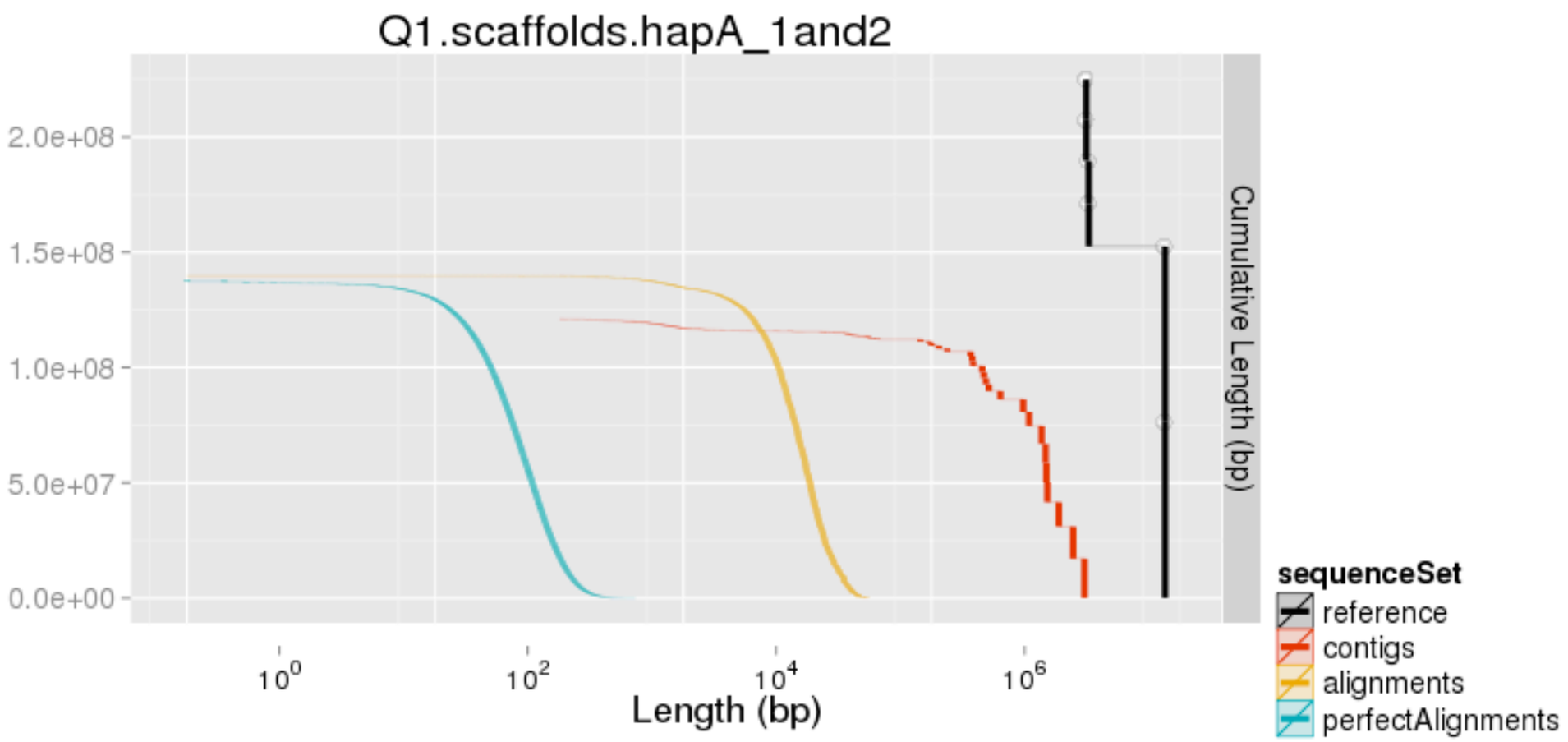
Burrows-Wheeler Aligner analysis by Joseph Fass. Coverage islands = contiguously aligned segments. Validity = sum of aligned lengths divided by sum of contig/scaffold lengths (i.e. what fraction can be aligned). Multiplicity = sum of aligned lengths divided by sum of coverage island lengths. Parsimony = validity divided by multiplicity. A score of 1 is perfect for these 3 measures.



For any set of sequences (contigs, aligned portion of contigs, or perfectly aligned portion of contigs), can calculate distribution of lengths and calculate measures that are analogous to N50. Continuity N50 = the N50 value for all alignable regions of contigs/scaffolds.



This is a deliberately poor assembly, with short contigs. Most contigs align well (close distance between red and yellow lines) but don't all align perfectly. Height on Y-axis denotes total size of each set of sequence lengths.



A better assembly with much longer scaffolds, and longer lengths of aligned sequence.

Summary

What have we learned?

Easy to produce a lot of data!

Use of any single metric can be misleading

Good assemblies tend to score well across most, but not all, metrics

Good assembler != good assembly

End users may care little about all of this except for wanting to see the sequence of their gene of interest.