6- First Pass Assembly & QC

Thursday morning

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia

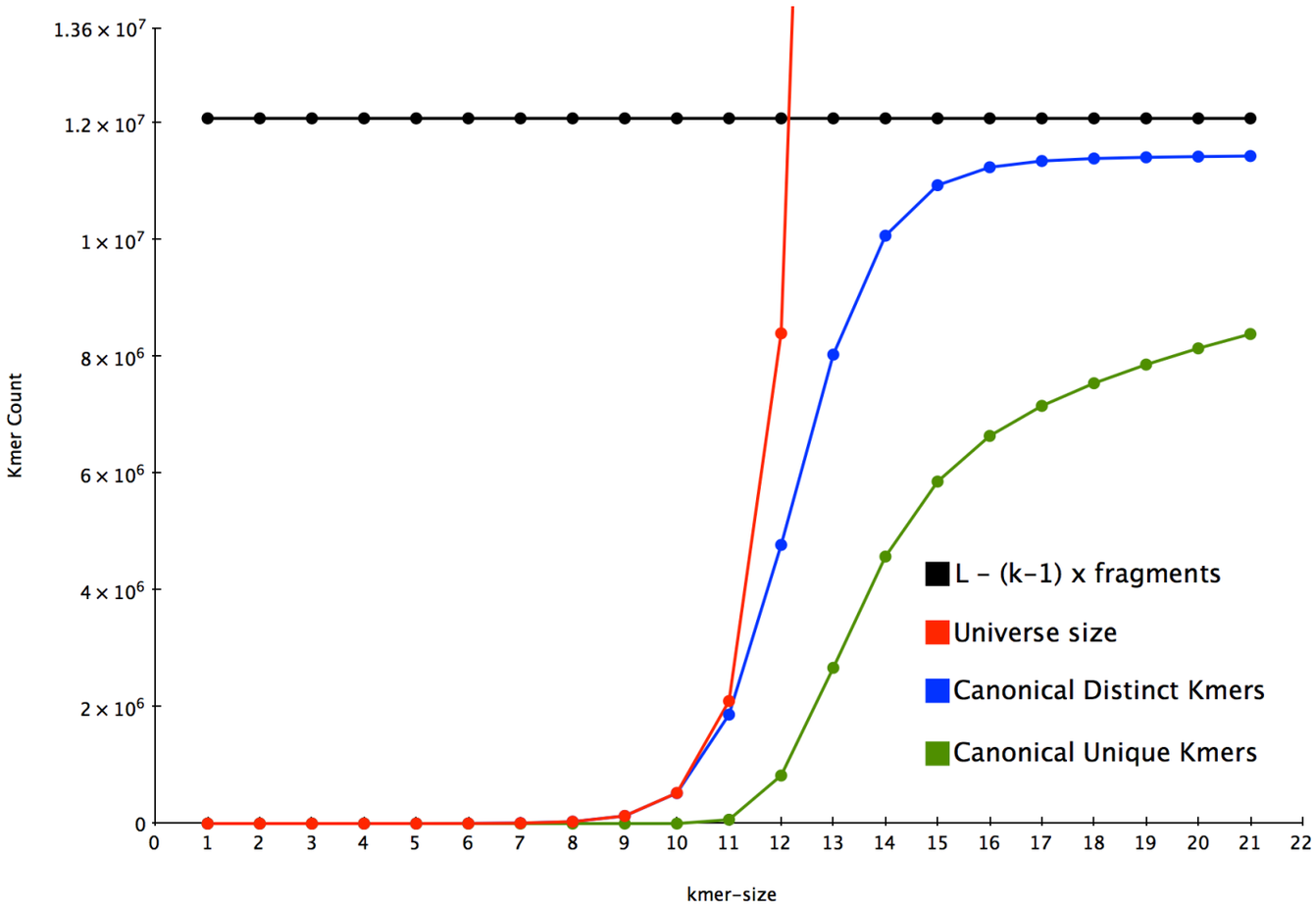




TGAC

The K tradeoff

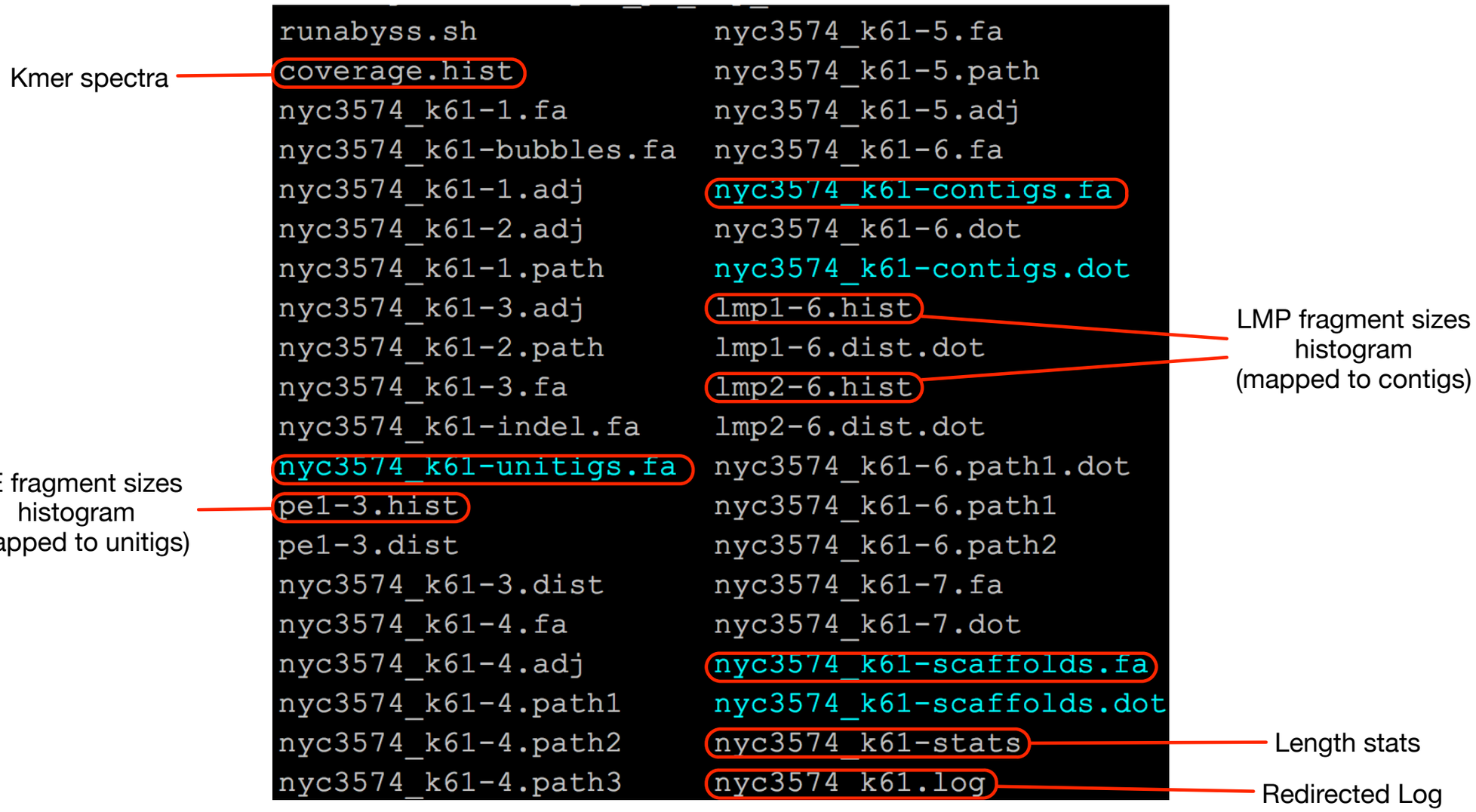
- **Longer kmers** are more **unique** in the target, **disentangling** the graph.
- **Smaller kmers** will **overlap** more often, favouring **contiguity**.
- Every read produces **L-k+1** kmers.
  - Higher k -> less coverage.
- Every single error affects **k** kmers.
  - Higher k -> more errors.
- A typical choice for 100bp reads is k=71.



TGAC

Running abyss as a first pass assembler

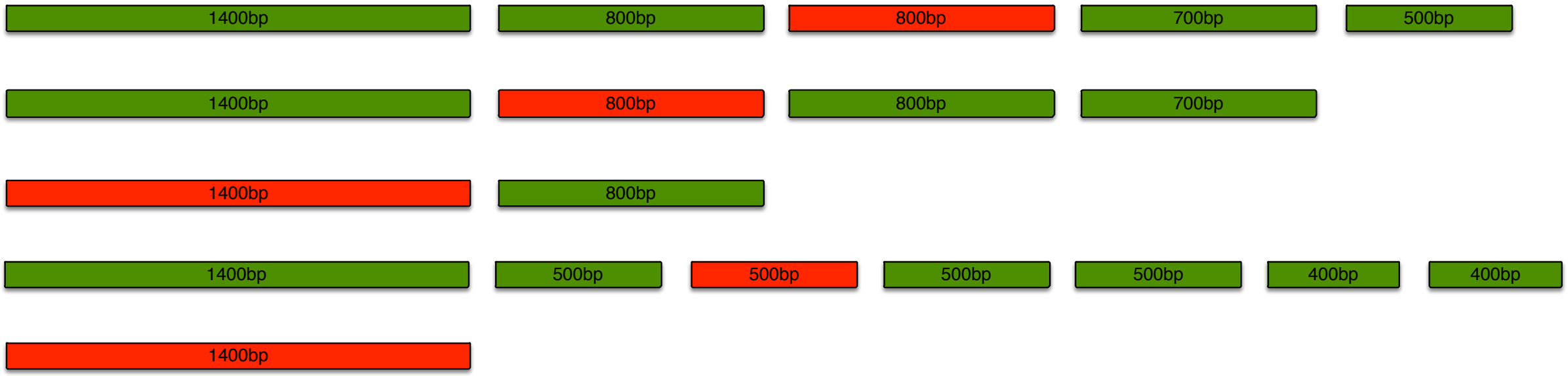
- It runs easily and can use both single and multi-host multiprocessing.
- Creates a ton of useful output, and a nice log.



TGAC

Beware of N50

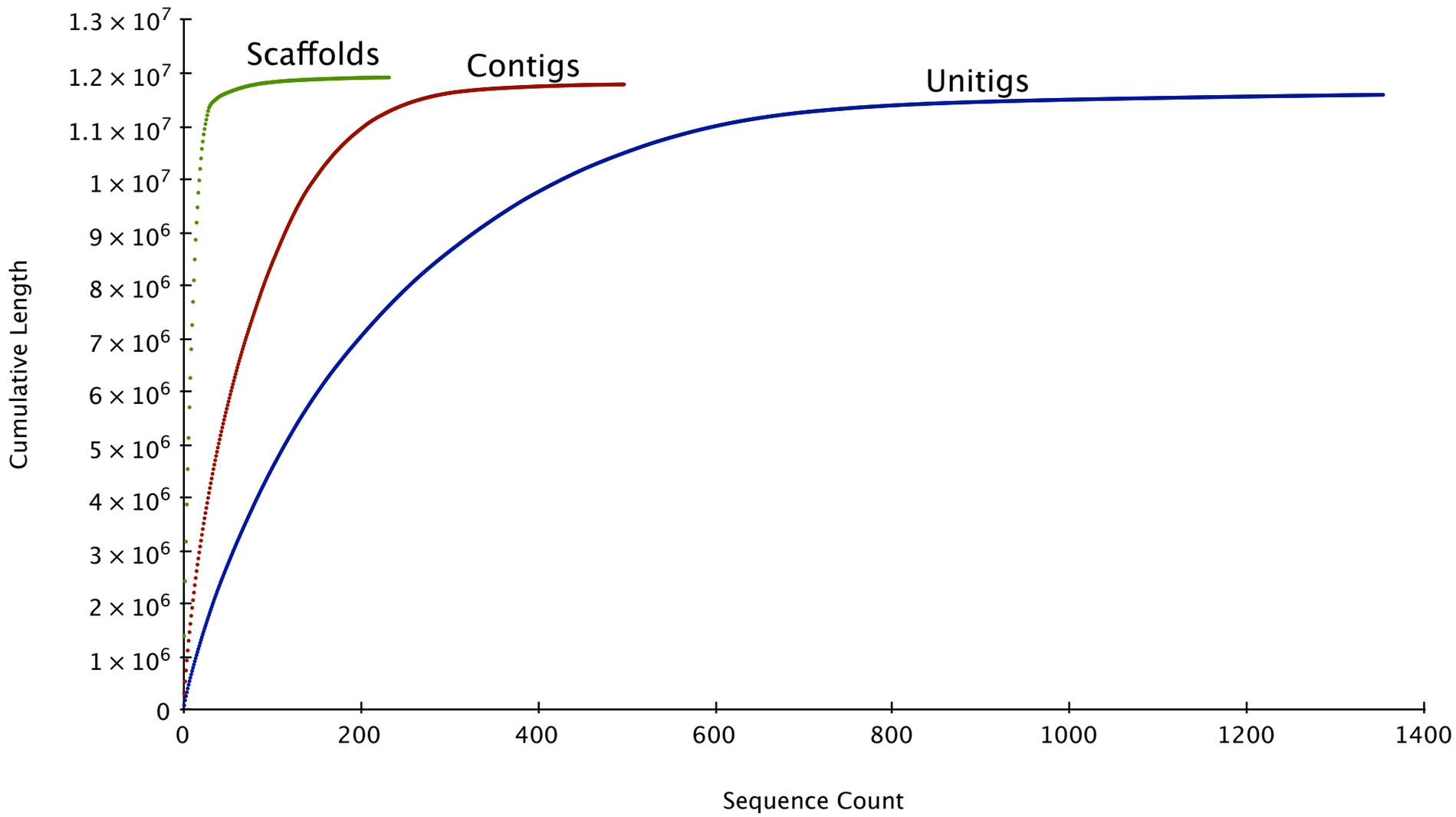
- N50 is the most used metric in assembly world… and it should not be:
  - Using contiguity as primary goal reward “risky joining”.
  - N50 is affected by filtering, and not very sensitive!



TGAC

Contiguity stats

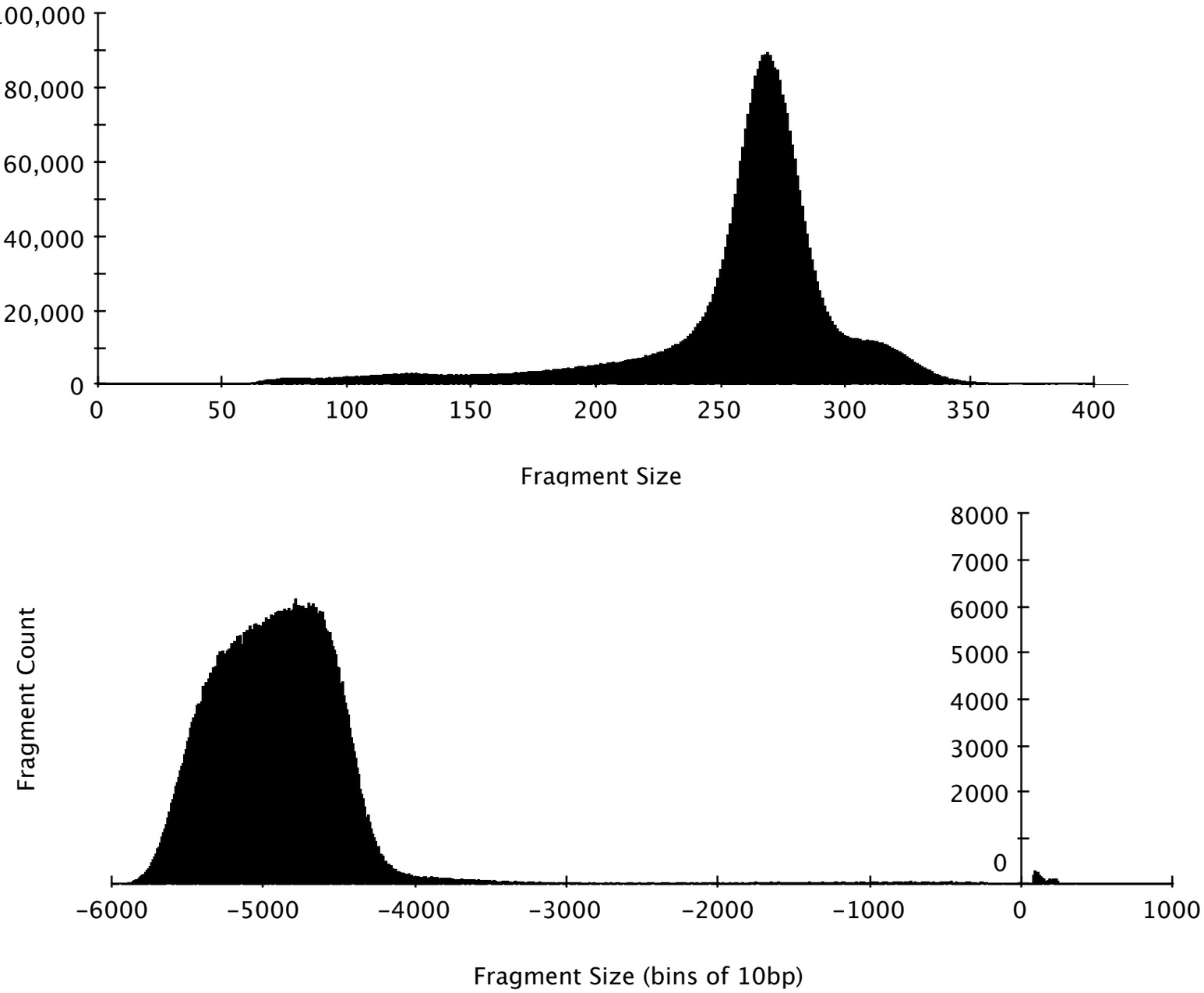
n	n:200	n:N50	min	N80	N50	N20	max	sum	
10773	1353	144	200	11170	25592	44031	96106	11.6e6	nyc3574_k61-unitigs.fa
8880	497	53	200	32554	66307	139116	322315	11.79e6	nyc3574_k61-contigs.fa
8615	232	8	200	269923	551245	1029531	1372216	11.74e6	nyc3574_k61-scaffolds.fa



- Don’t forget to check your “Ns” !!!

TGAC

Fragment Sizes



TGAC

Read mapping stats

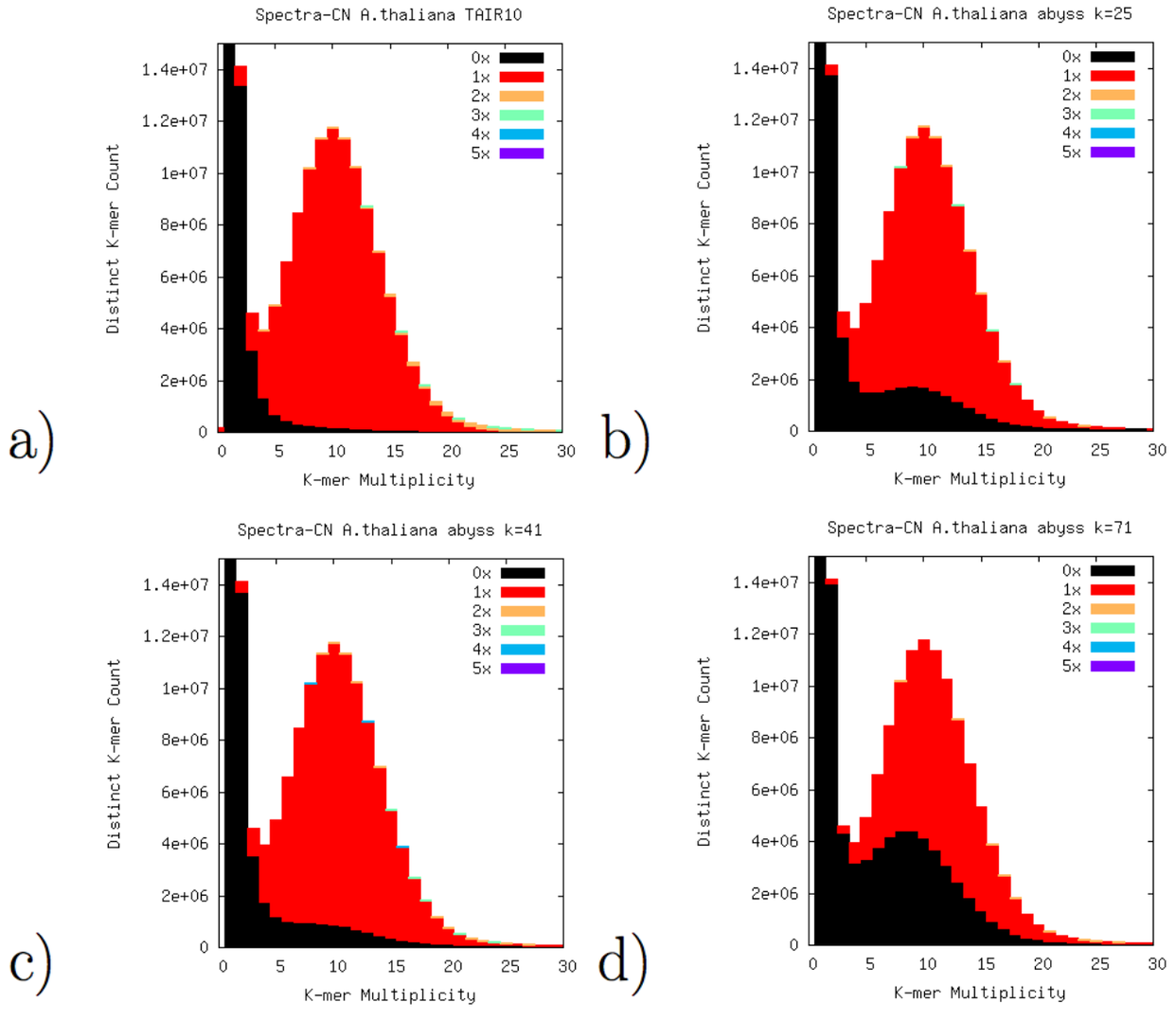
```
abyss-map  -j150 -l61   /scratch/clavijob/yeast_tests/diploid/s_3_1_seq
uence.txt /scratch/clavijob/yeast_tests/diploid/s_3_2_sequence.txt nyc3
574_k61-3.fa \
               |abyss-fixmate   -h pe1-3.hist \
               |sort -snk3 -k4 \
               |DistanceEst  -j150 -k61 -l61 -s200 -n10   -o pe1-3.dis
t pe1-3.hist
Building the suffix array...
Building the Burrows-Wheeler transform...
Building the character occurrence table...
Mateless         0
Unaligned    71619  1.14%
Singleton   516328  8.19%
FR         4018144  63.8%
RF              28  0.000444%
FF            8285  0.131%
Different  1686337  26.8%
Total      6300741
```

Read mapping stats

```
abyss-map  -j150 -l61   /scratch/clavijob/yeast_tests/diploid/LIB3796_c
lipped_A_R1.fastq /scratch/clavijob/yeast_tests/diploid/LIB3796_clipped
_A_R2.fastq nyc3574_k61-6.fa \
                |abyss-fixmate   -h lmp1-6.hist \
                |sort -snk3 -k4 \
                |DistanceEst --dot  -j150 -k61 -l61 -s200 -n10   -o lmp
1-6.dist.dot lmp1-6.hist
Building the suffix array...
Building the Burrows-Wheeler transform...
Building the character occurrence table...
Mateless         0
Unaligned   127754  6.8%
Singleton   828893  44.1%
FR            3191  0.17%
RF          668696  35.6%
FF           20536  1.09%
Different   230815  12.3%
Total      1879885
```

Checking content inclusion using KAT

- Just compare the frequency of kmers in the assembly to the reads spectrum.



TGAC

Different assemblies and pre-processing

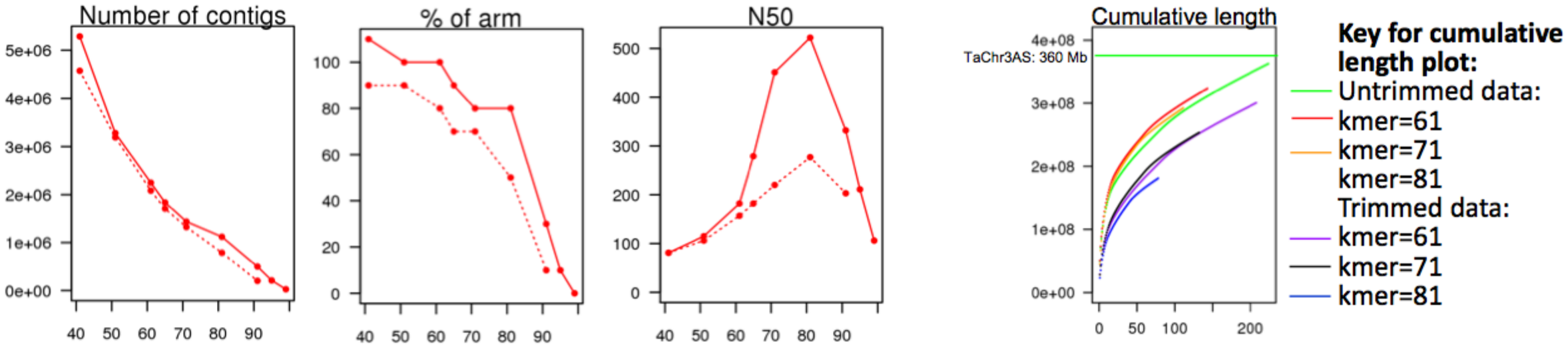


Figure 5 *Length and coverage-based metrics for chromosome 3AS assembled using kmer sizes between 41 and 99 as indicated on the x axis: —— untrimmed data, ------ trimmed reads (to quality score of Q30)*

Assembly	Number Of Hits To ESTs
3AS untrimmed	1539
3AS trimmed	1132

TGAC

Look for expected content

- Just BLAST the output to check what it looks like.
- But you can also try finding genes/markers/ETS:

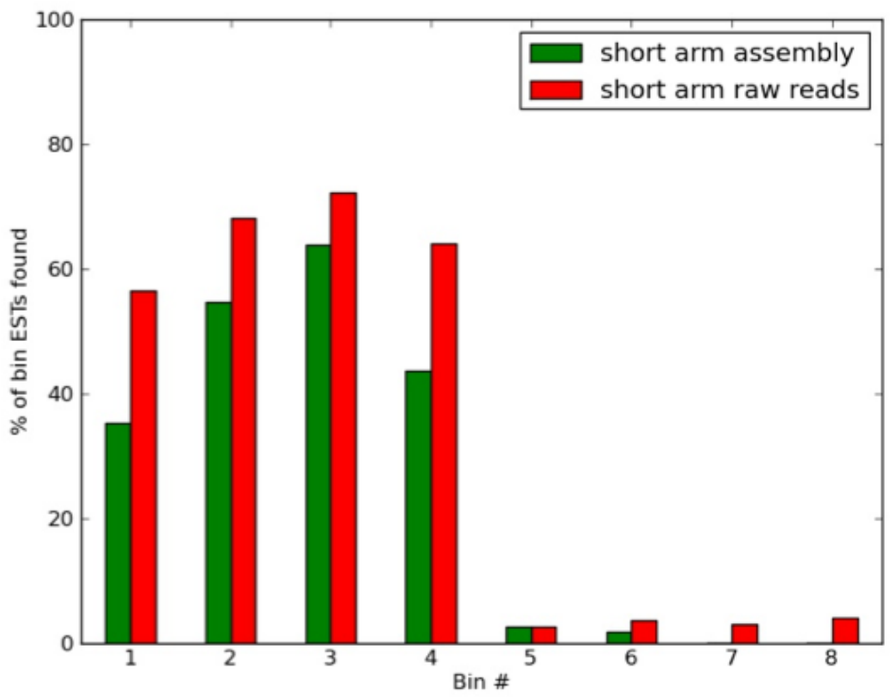


Figure 7: short arm EST recovery

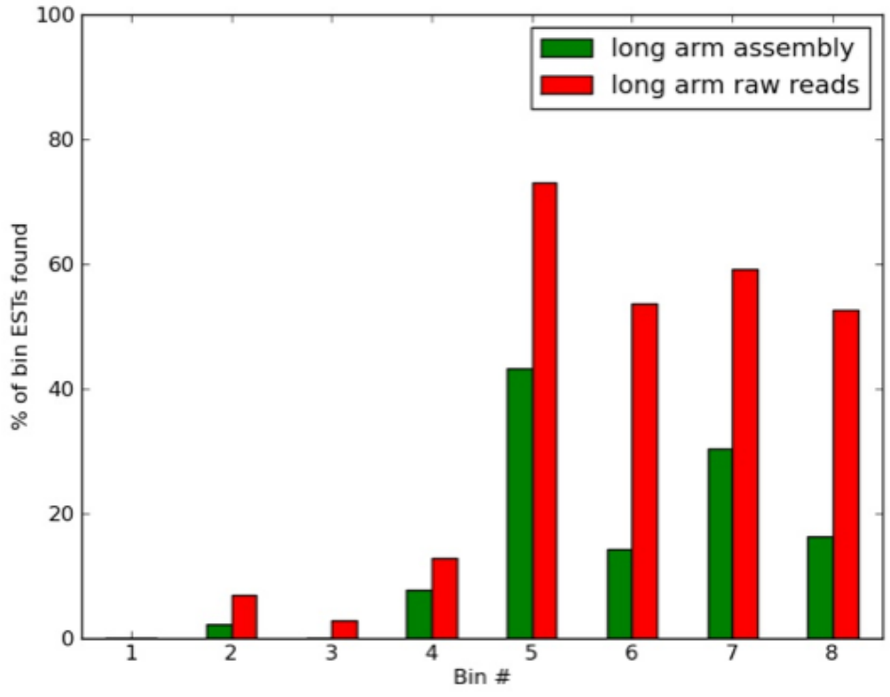
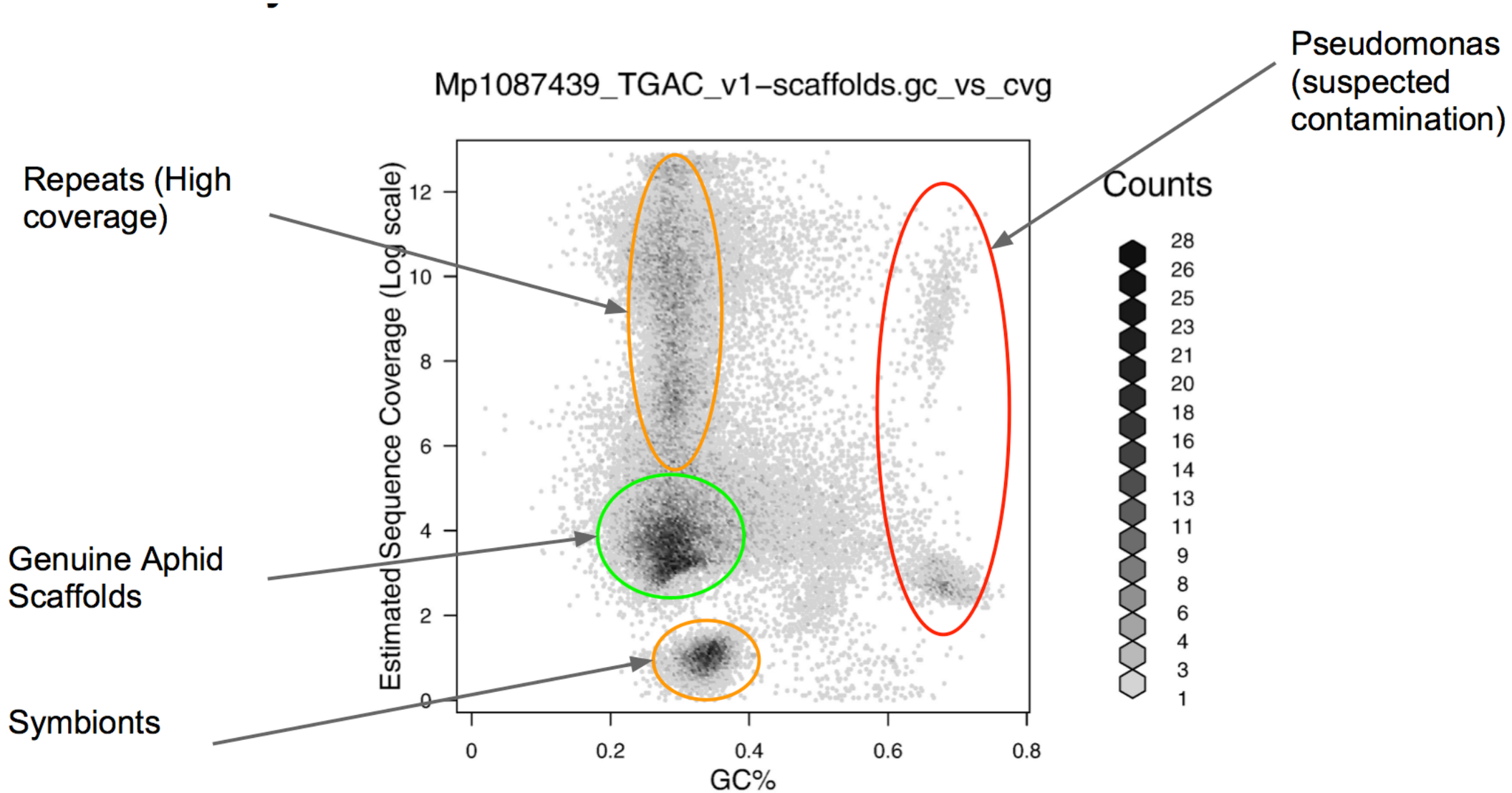


Figure 8: long arm EST recovery

TGAC

Look for contaminants

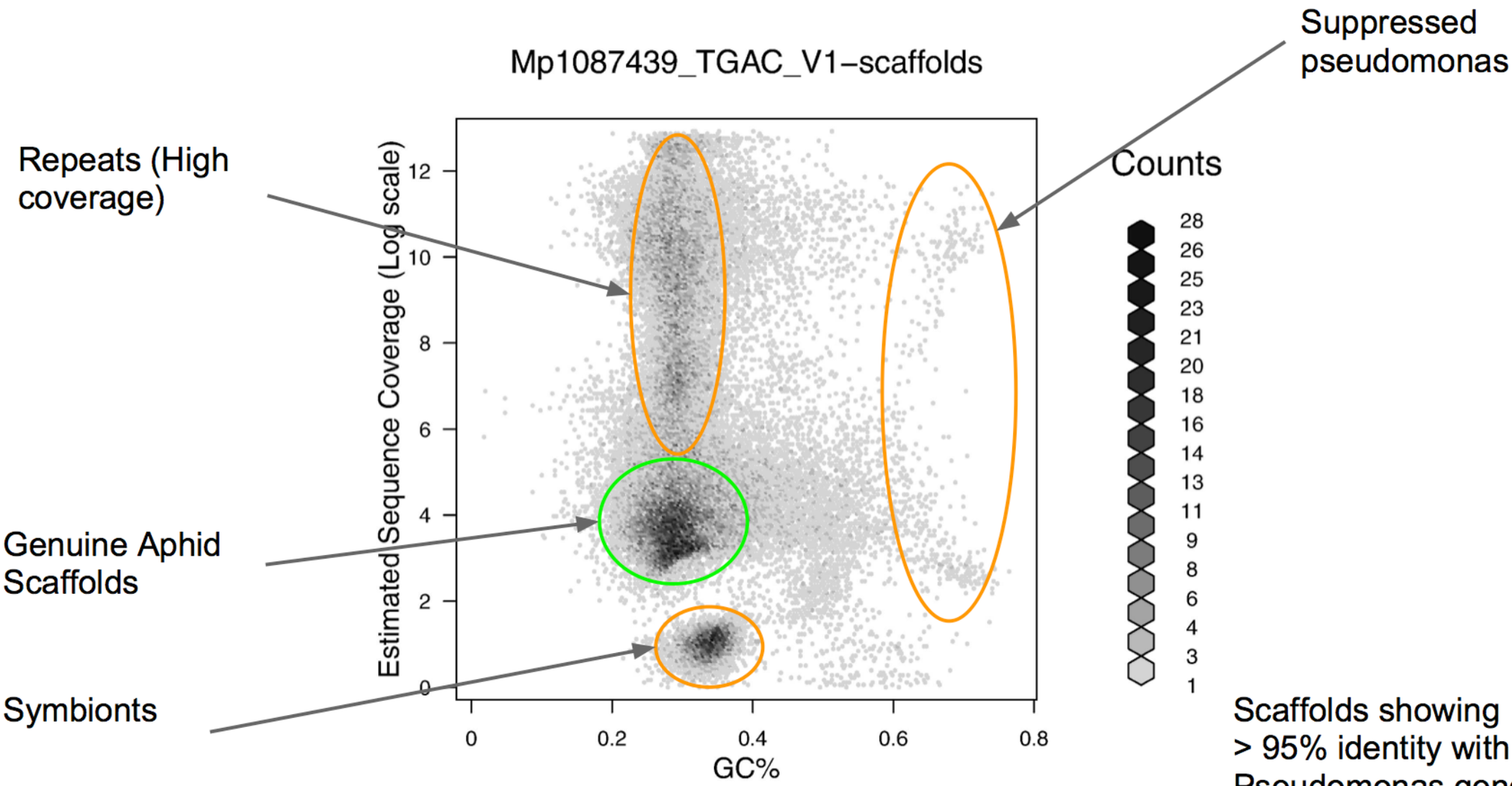
- Contaminants including symbionts, mitochondria, chloroplast.



TGAC

Look for contaminants

- Contaminants including symbionts, mitochondria, chloroplast.



TGAC

What is the output of your first-pass assembly?

- Knowledge about the used datasets:
  - Are they clean? Can they be better?
  - How each one performs.
  - How they all interact.
- Knowledge about the target:
  - How it's "complications" affect assembly. (also the datasets?)
  - Repeat structure?
- Refined choice of K.
- Baseline metrics (to be improved).

Questions?