

6- First Pass Assembly & QC Hands On

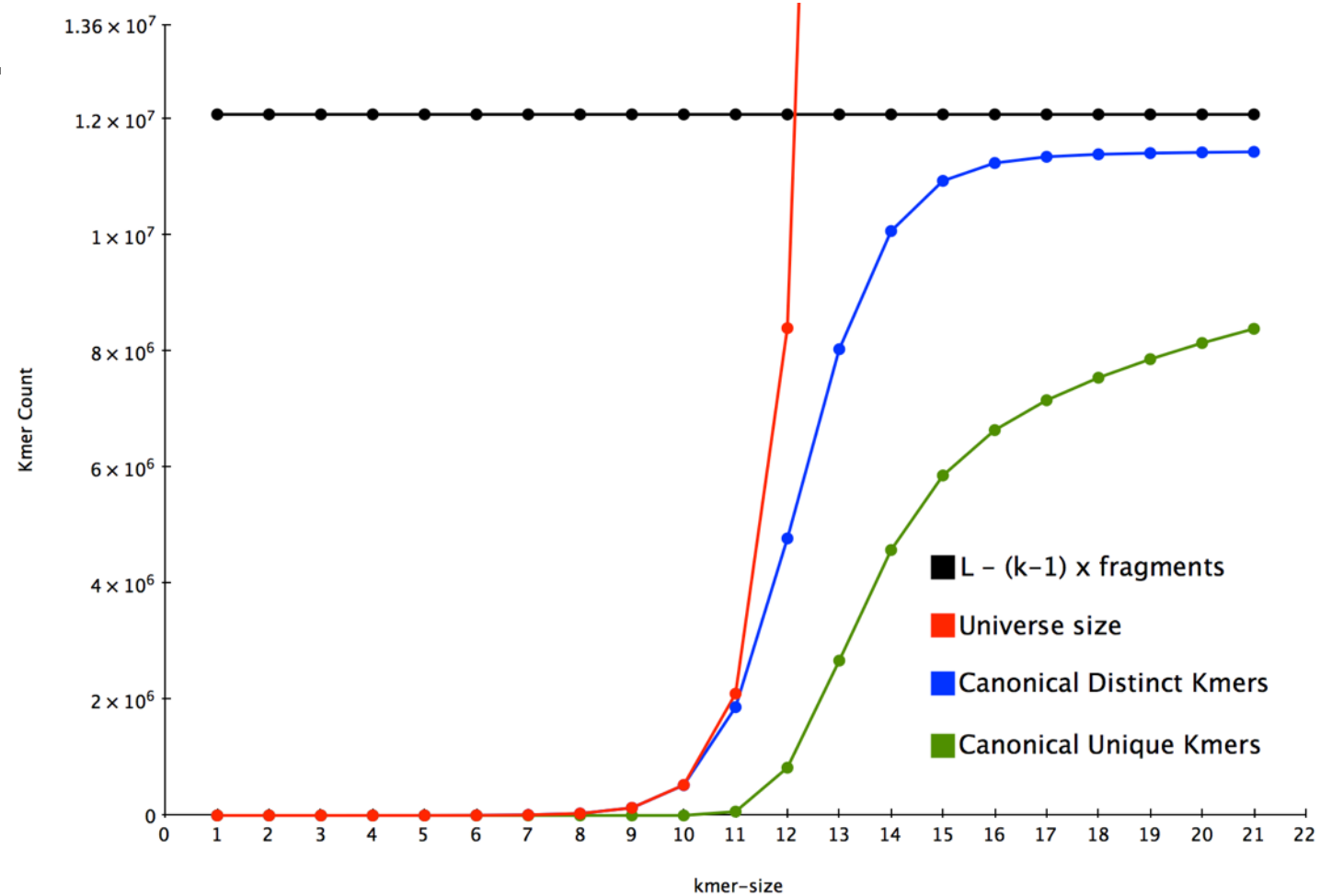
Thursday morning

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia / Jon Wright



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 $k \rightarrow$ less coverage.
- Every single error affects k kmers.
 - Higher $k \rightarrow$ more errors.



- A typical choice for 100bp reads is $k=71$.

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.

```
runabyss.sh
coverage.hist
nyc3574_k61-1.fa
nyc3574_k61-bubbles.fa
nyc3574_k61-1.adj
nyc3574_k61-2.adj
nyc3574_k61-1.path
nyc3574_k61-3.adj
nyc3574_k61-2.path
nyc3574_k61-3.fa
nyc3574_k61-indel.fa
nyc3574_k61-unitigs.fa
pe1-3.hist
pe1-3.dist
nyc3574_k61-3.dist
nyc3574_k61-4.fa
nyc3574_k61-4.adj
nyc3574_k61-4.path1
nyc3574_k61-4.path2
nyc3574_k61-4.path3
nyc3574_k61-5.fa
nyc3574_k61-5.path
nyc3574_k61-5.adj
nyc3574_k61-6.fa
nyc3574_k61-contigs.fa
nyc3574_k61-6.dot
nyc3574_k61-contigs.dot
lmp1-6.hist
lmp1-6.dist.dot
lmp2-6.hist
lmp2-6.dist.dot
nyc3574_k61-6.path1.dot
nyc3574_k61-6.path1
nyc3574_k61-6.path2
nyc3574_k61-7.fa
nyc3574_k61-7.dot
nyc3574_k61-scaffolds.fa
nyc3574_k61-scaffolds.dot
nyc3574_k61-stats
nyc3574_k61.log
```

Kmer spectra — **coverage.hist**

PE fragment sizes histogram (mapped to unitigs) — **pe1-3.hist**

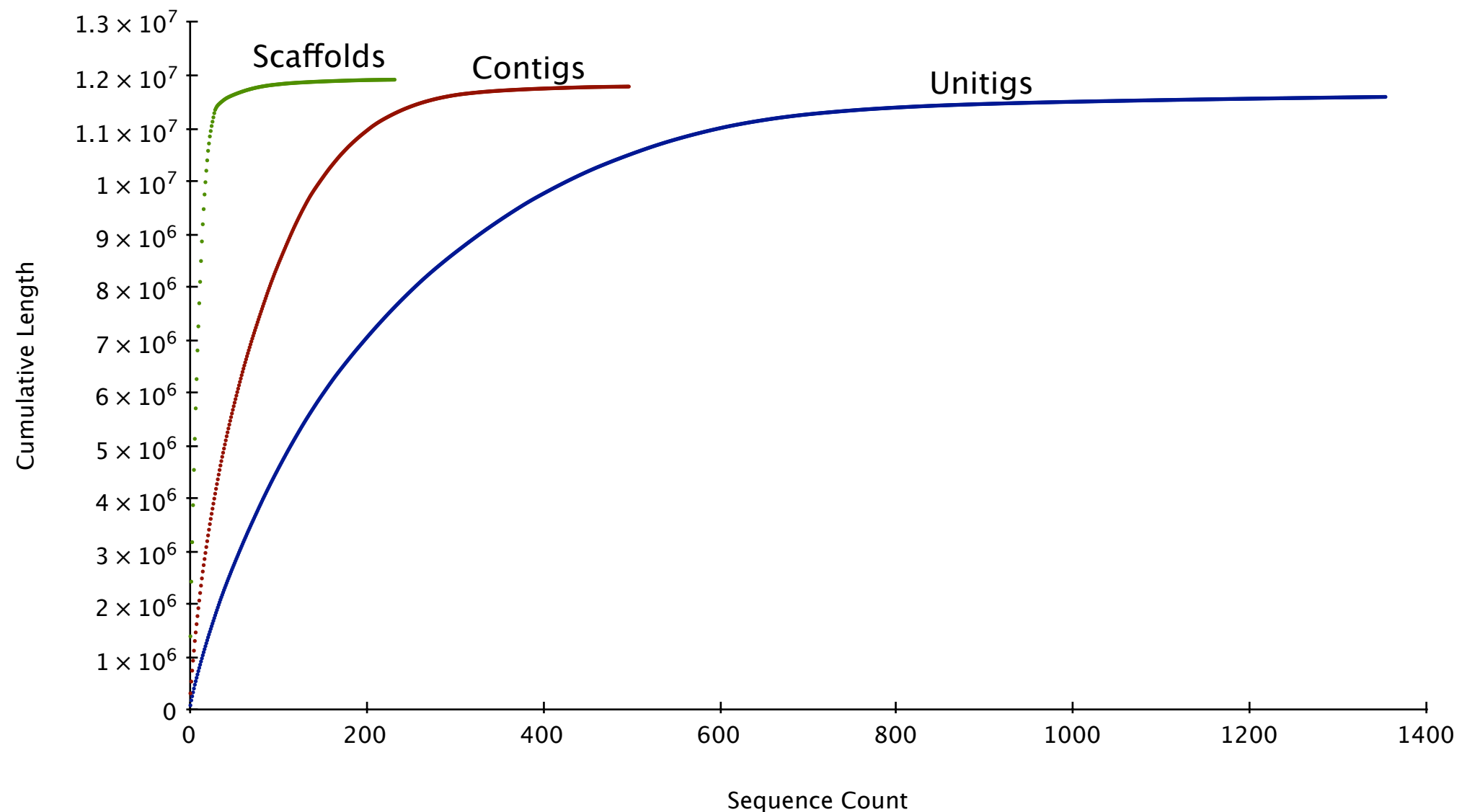
LMP fragment sizes histogram (mapped to contigs) — **lmp1-6.hist**, **lmp2-6.hist**

Length stats — **nyc3574_k61-stats**

Redirected Log — **nyc3574_k61.log**

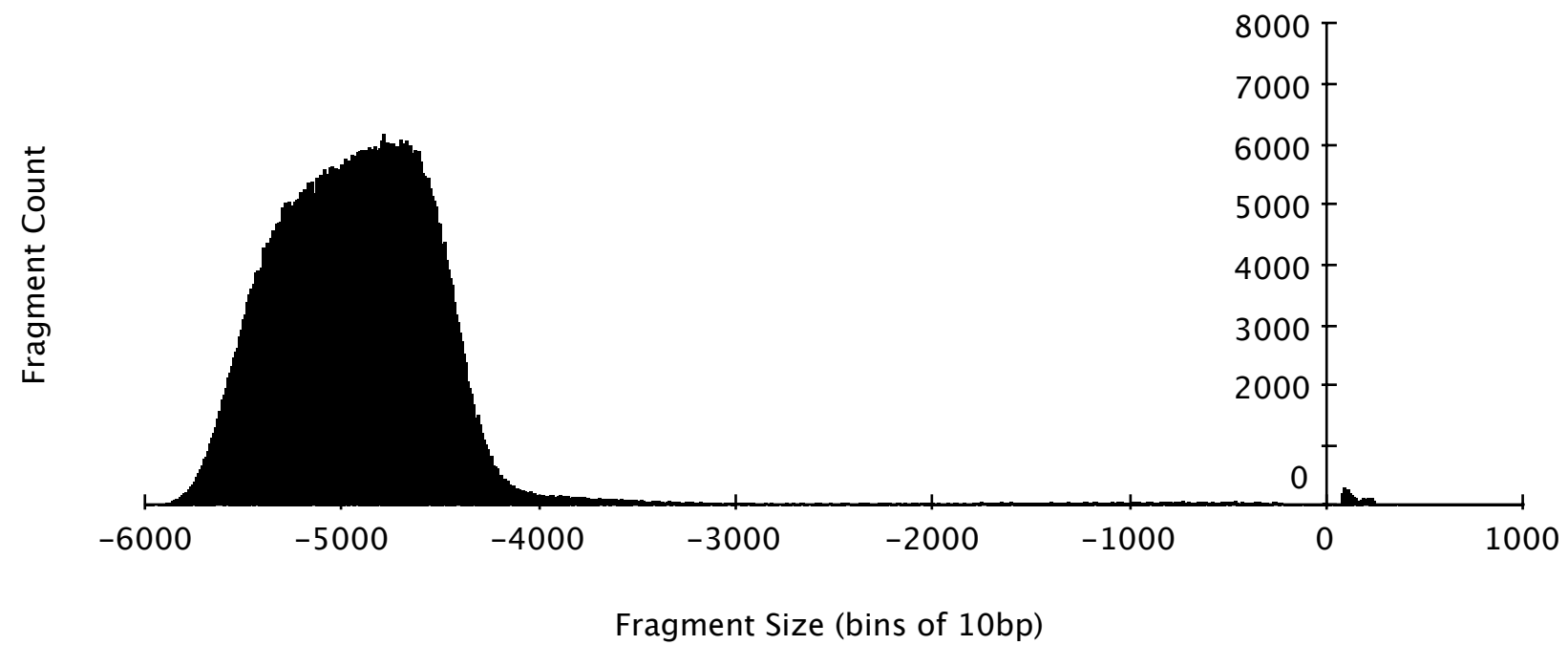
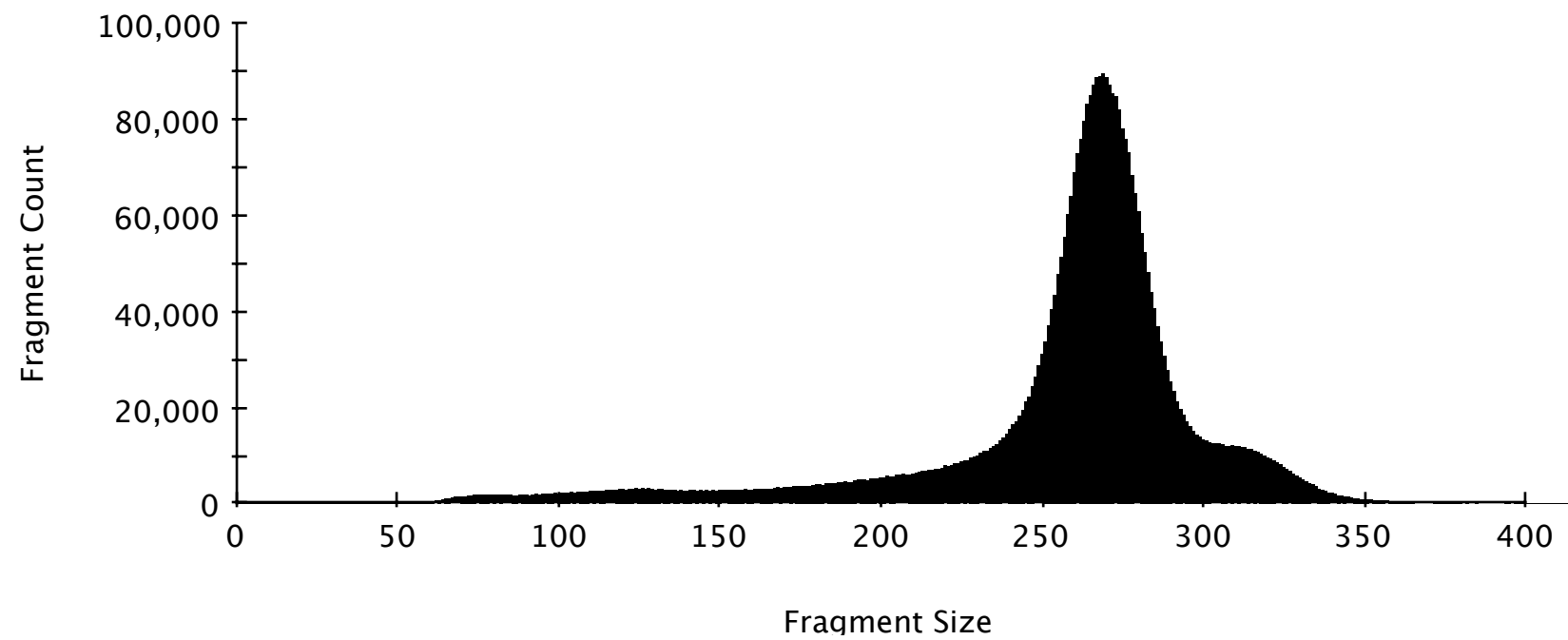
Contiguity stats

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” !!!

Fragment Sizes



Read mapping stats

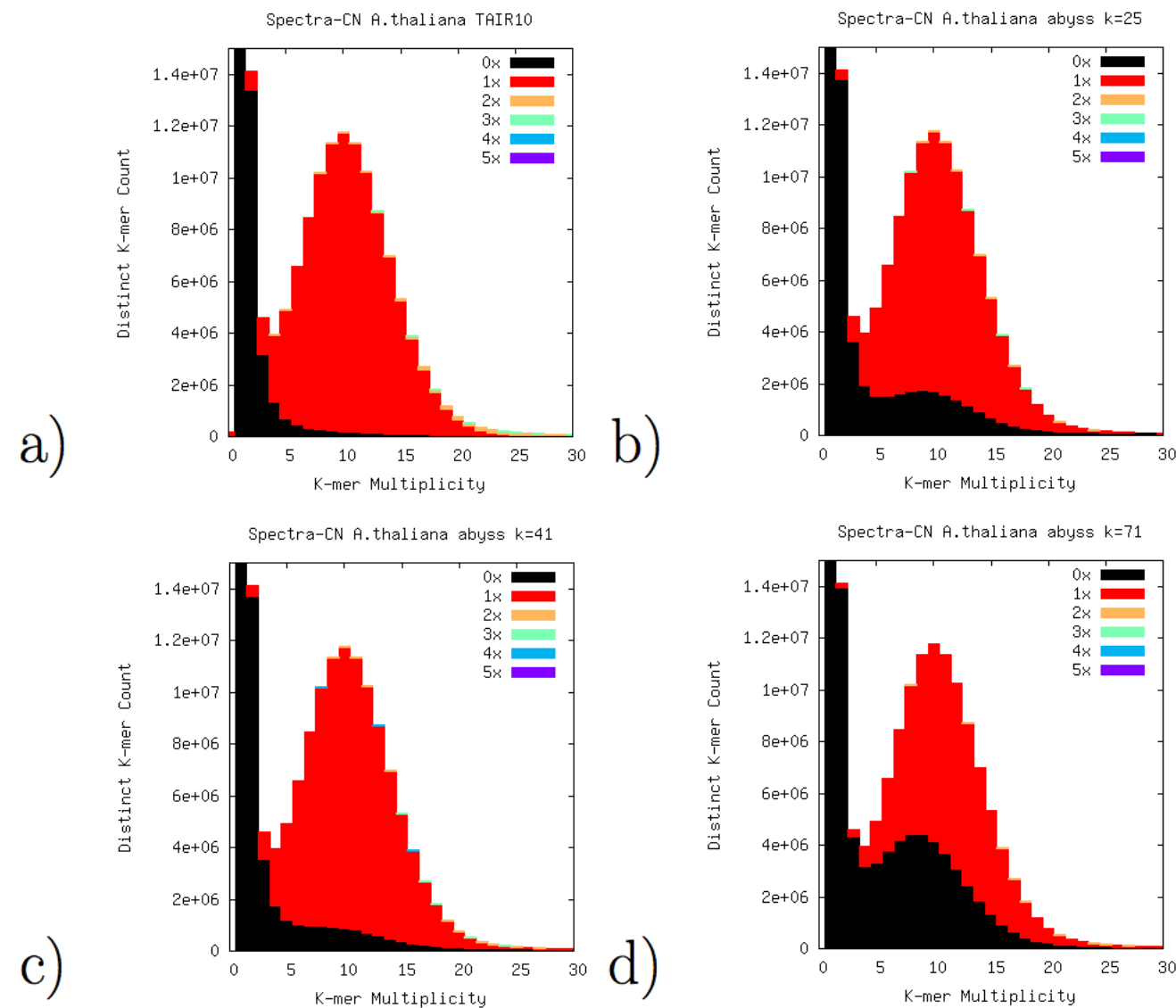
```
abyss-map -j150 -l61 /scratch/clavijob/yeast_tests/diploid/s_3_1_sequence.txt /scratch/clavijob/yeast_tests/diploid/s_3_2_sequence.txt nyc3574_k61-3.fa \  
          |abyss-fixmate -h pe1-3.hist \  
          |sort -snk3 -k4 \  
          |DistanceEst -j150 -k61 -l61 -s200 -n10 -o pe1-3.dist  
t 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.



Different assemblies and pre-processing

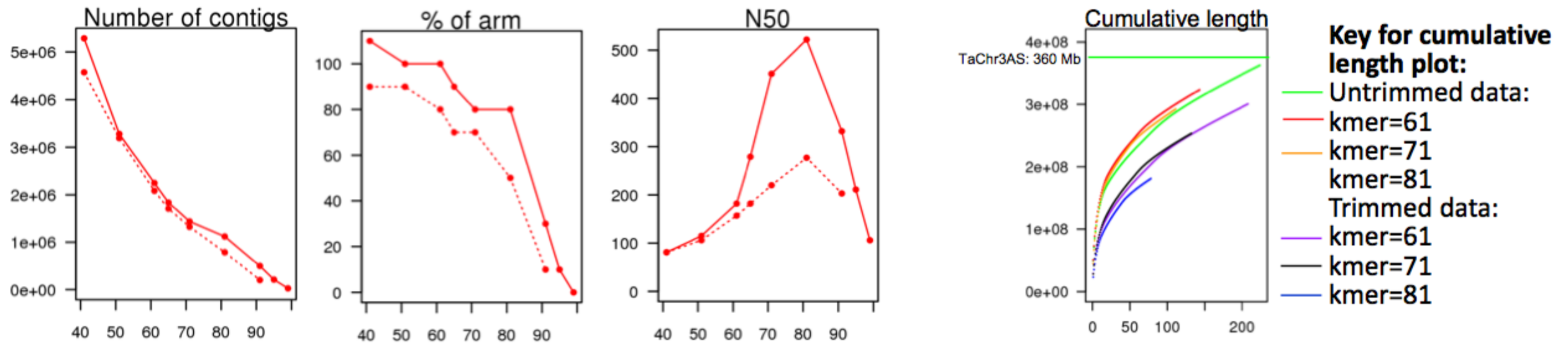


Figure 5 Length and coverage-based metrics for chromosome 3AS assembled using kmer sizes between 41 and 99 as indicated on the x axis: — untrimmed data, ---- trimmed reads (to quality score of Q30)

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

Look for expected content

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

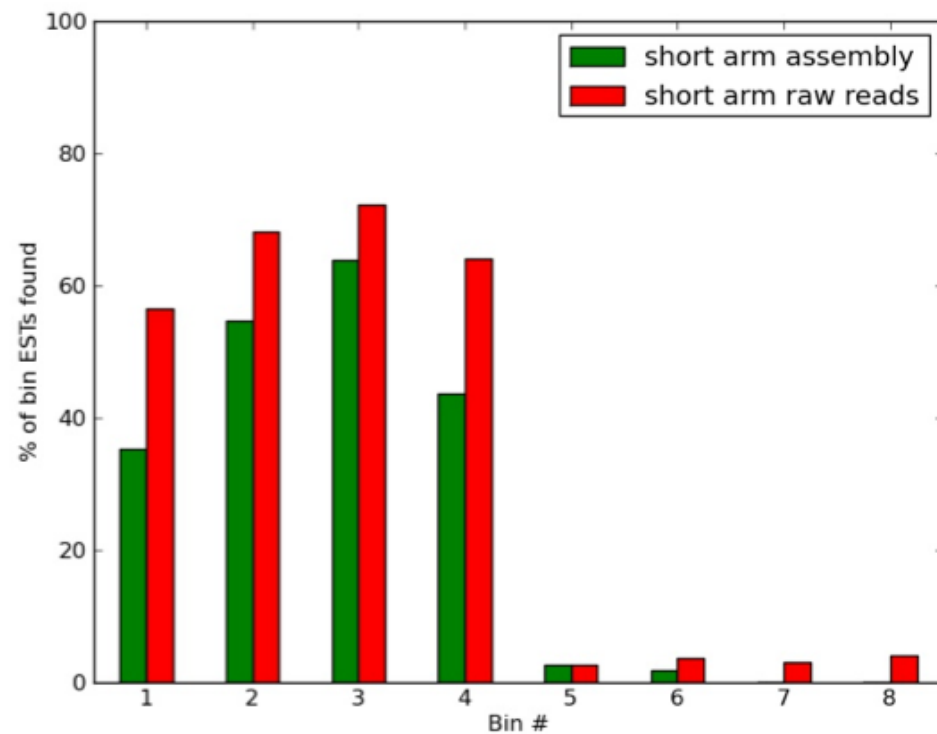


Figure 7: short arm EST recovery

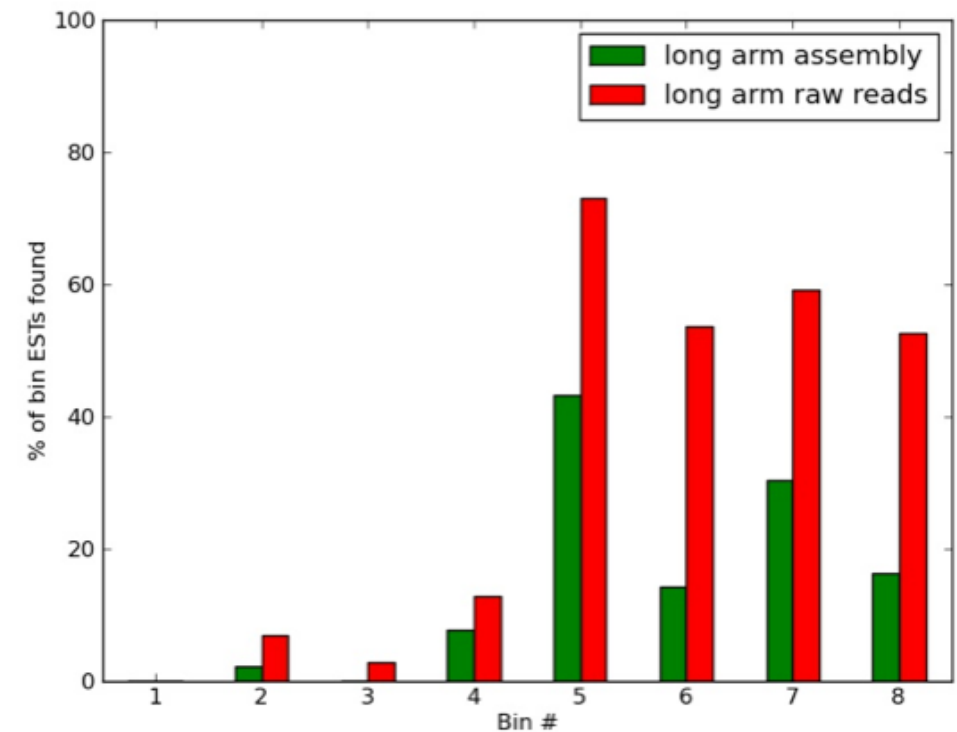
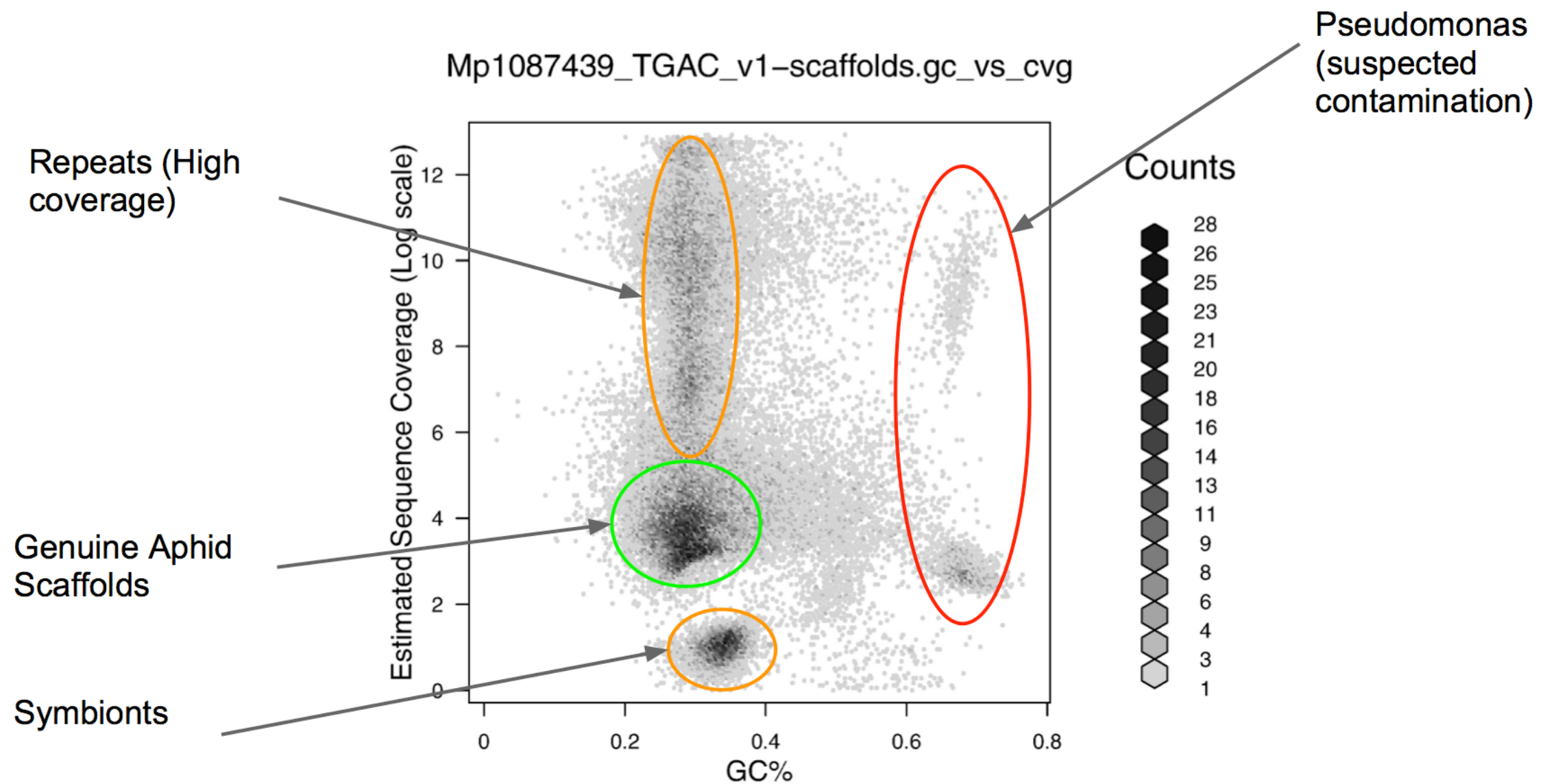


Figure 8: long arm EST recovery

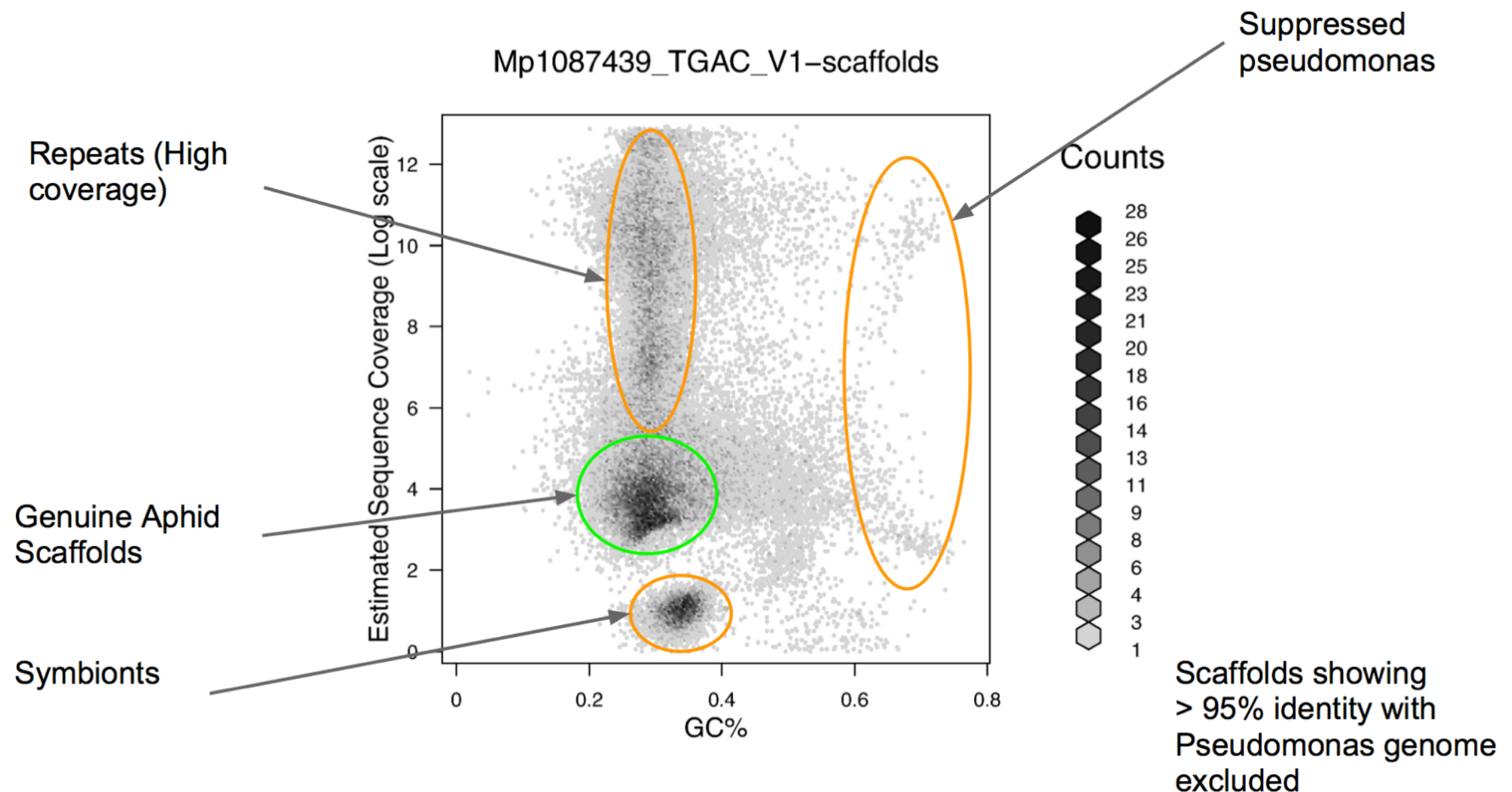
Look for contaminants

- Contaminants including symbionts, mitochondria, chloroplast.



Look for contaminants

- Contaminants including symbionts, mitochondria, chloroplast.



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?

