Metagenomics
Course materials

Mihai Pop
http://www.cbcb.umd.edu/~mpop
(mpop@umiacs.umd.edu)

Data

- All data discussed here are in: ftp://ftp.cbcb.umd.edu/pub/data/MetAmos-demo/MetAmos-demo.zip
- Note: file is big (4 GB)

MetAMOS

- metAMOS page: http://www.cbcb.umd.edu/software/metamos
- metAMOS automatically helps you install a number of other utilities
- metAMOS is under active development – it should improve fast in the near future
- better documentation will be forthcoming

VALET – metagenomic assembly validation

- git clone https://github.com/jgluck/VALET.git
- cd VALET/
- ./setup.sh

- # Let your shell know where to find the VALET pipeline.
- export VALET=`pwd`/src/py/

running metAMOS

- initialize directory

```
initPipeline -d MetAmos-demo/SRS014465 \
    -1 SRS014465.denovo_duplicates_marked.trimmed.1.fastq \
    -2 SRS014465.denovo_duplicates_marked.trimmed.2.fastq -i 100:500
```

- run assembly

```
runPipeline -d MetAmos-demo/SRS014465 -a soapdenovo
```

- Note: you can run different assembly tools with the same command

prerun assemblies

- SRS014465 – HMP vaginal sample

- SRS014465_hmp – original HMP assembly
- SRS014465_soap – metAMOS assembly using soapdenovo
- SRS014465_spades – metAMOS assembly using Spades

Validating assemblies with VALET

- pipeline.py -q -a SRS014465_hmp/SRS014465.scaffolds.fa \
  -o SRS014465_hmp_valet \
  -1 SRS014465.denovo_duplicates_marked.trimmed.1.fasty \
  -2 SRS014465.denovo_duplicates_marked.trimmed.2.fastq \
  --window-size 100 --min-coverage 10 --coverage-multiplier 0.0 \
  --threads 4 --ignore-ends 100 --min-contig-length 300

- pipeline.py -q -a SRS014465_soap/Postprocess/out/proba.ctg.fa \
  -c SRS014465_soap/Postprocess/out/proba.ctg.cvg \
  -o SRS014465_soap_valet \
  -1 SRS014465.denovo_duplicates_marked.trimmed.1.fasty \
  -2 SRS014465.denovo_duplicates_marked.trimmed.2.fastq \
  --window-size 100 --min-coverage 10 --coverage-multiplier 0.0 \
  --threads 4 --ignore-ends 100 --min-contig-length 300

precomputed results

- SRS014465_hmp_valet – original HMP assembly
- SRS014465_soap_valet – metAMOS assembly using soapdenovo
- SRS014465_spades_valet – metAMOS assembly using Spades

Exploring the assembly

- Pretty reports (ASMDIR is the name of directory from initPipeline):
  - ASMDIR/Postprocess/out/html/summary.html
- Output contigs and scaffolds
  - ASMDIR/Postprocess/out/proba.ctg.fa
  - ASMDIR/Postprocess/out/proba.scf.fa
- Genes
  - ASMDIR/Postprocess/out/proba.orf.faa
  - ASMDIR/Postprocess/out/proba.orf.fna
- Variation motifs
  - ASMDIR/Postprocess/out/proba.motifs.fa
- Taxonomically broken up contigs
  - ASMDIR/Postprocess/out/class.classified
    [note: taxonomic level for classification can be changed at runtime]

Which assembly is better?

- Try

  $ module load bioinfo/assemblystats

  $ assembly_stats.pl ASMDIR/Postprocess/out/proba.scf.fna

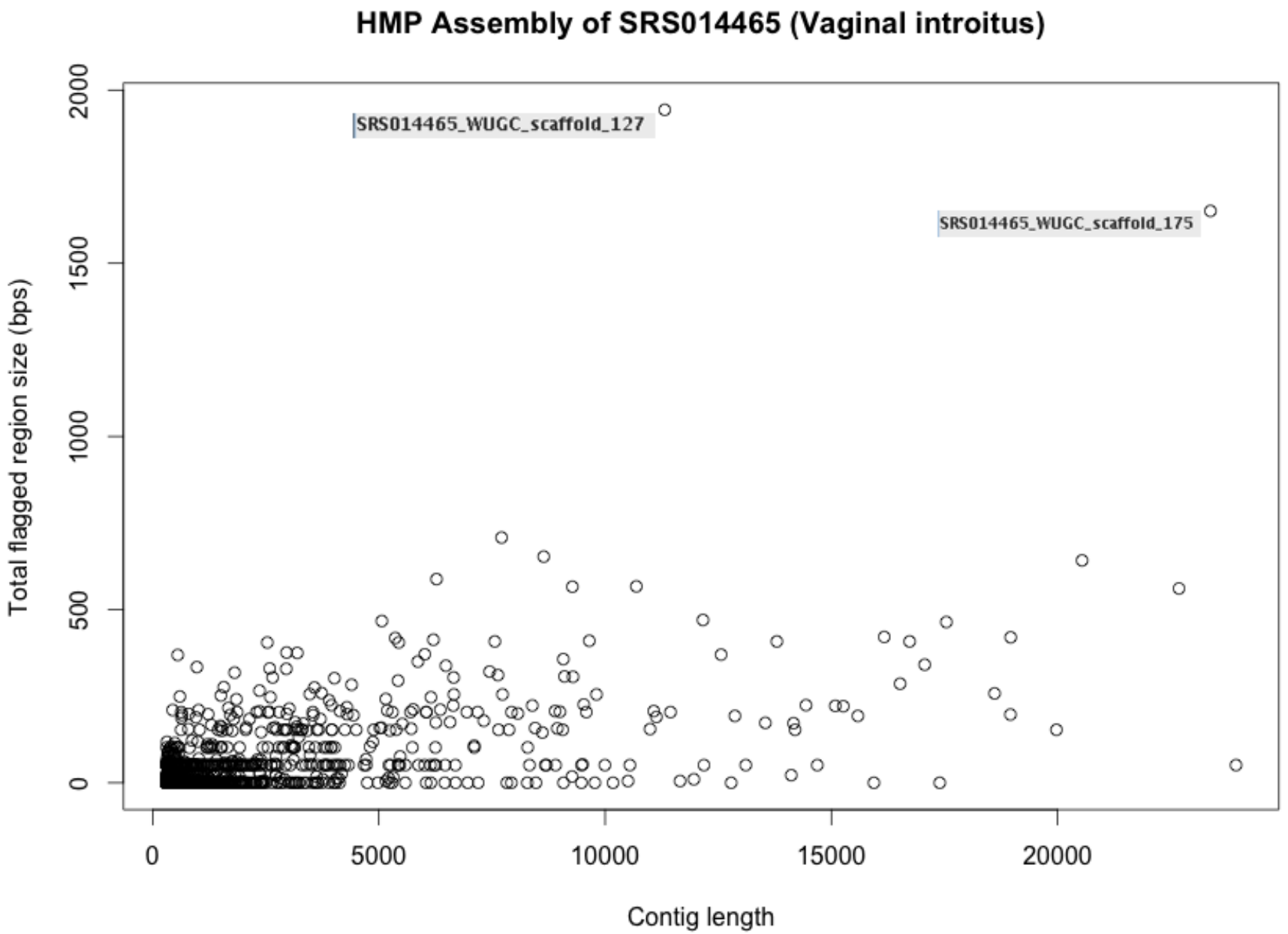
- Explore metAMOS logs:
  - ASMDIR/Logs/COMMANDS.log

Which assembly is better...in detail

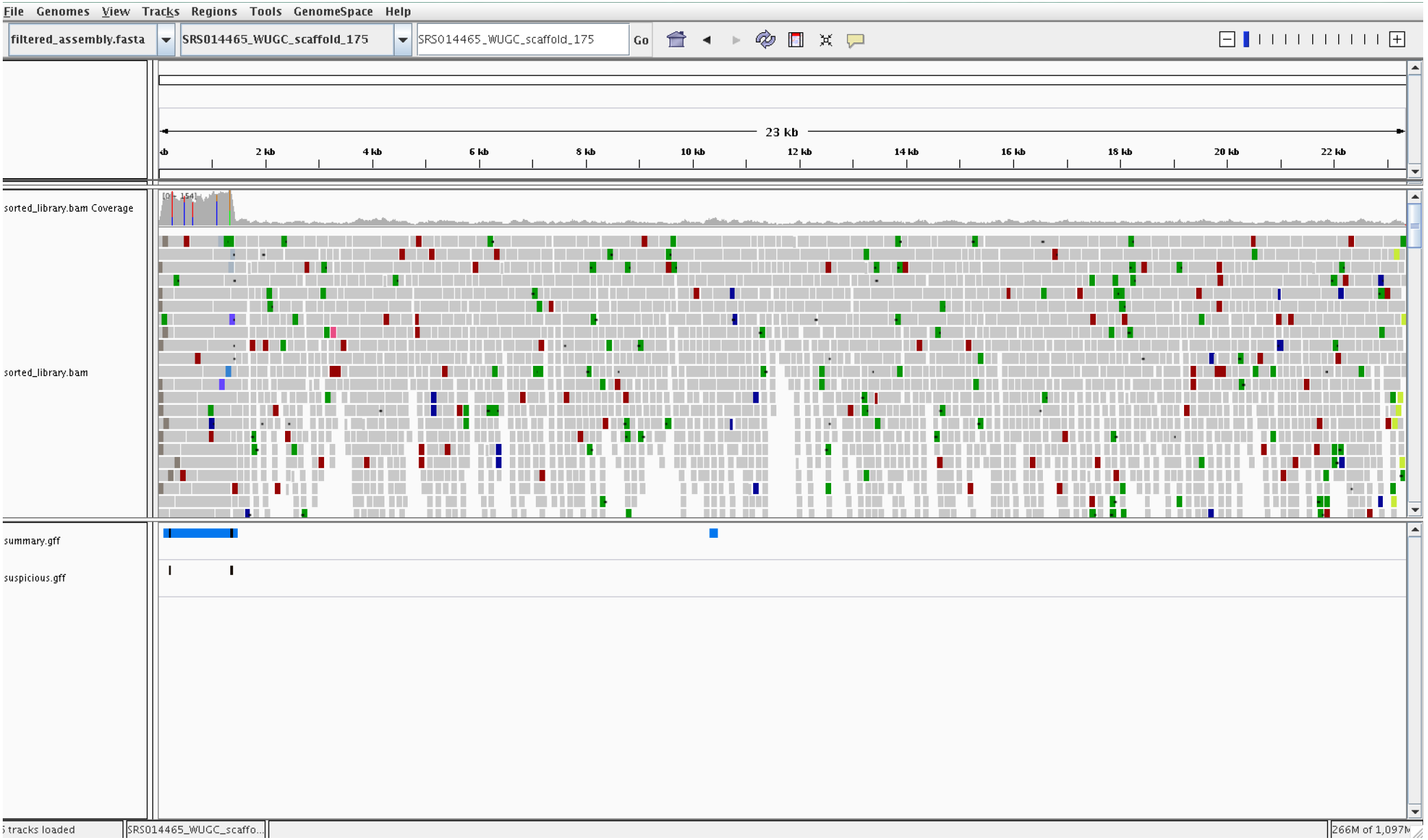
- Look into VALET output
  lap/output.sum
  summary.tsv

- visualize assembly with IGV
  igv.sh -b IGV.batch

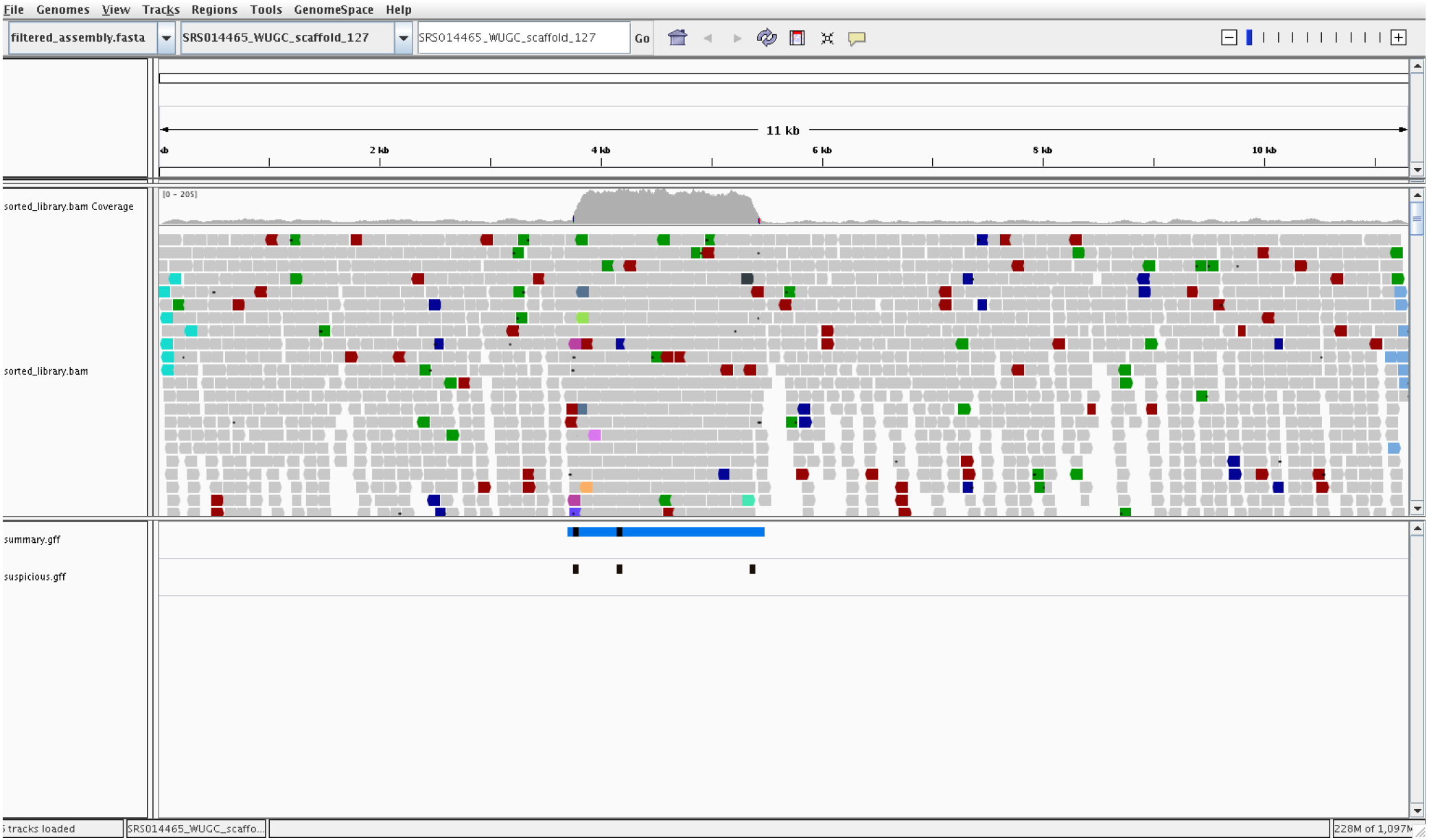
Contig size vs. mis-assembly rate



Miss. 1


File Genomes View Tracks Regions Tools GenomeSpace Help
filtered_assembly.fasta
SRS014465_WUGC_scaffold_175
SRS014465_WUGC_scaffold_175
Go
23 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
16 kb
18 kb
20 kb
22 kb
sorted_library.bam Coverage
sorted_library.bam
summary.gff
suspicious.gff
tracks loaded
SRS014465_WUGC_scaffo...
266M of 1,097M

Miss. 2


File Genomes View Tracks Regions Tools GenomeSpace Help
filtered_assembly.fasta
SRS014465_WUGC_scaffold_127
SRS014465_WUGC_scaffold_127
Go
11 kb
2 kb
4 kb
6 kb
8 kb
10 kb
sorted_library.bam Coverage
[0 - 205]
sorted_library.bam
summary.gff
suspicious.gff
tracks loaded
SRS014465_WUGC_scaffo...
228M of 1,097M