Assembly lab post-mortem

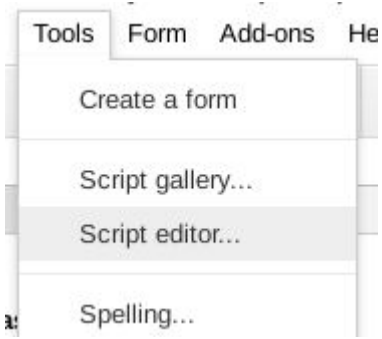
(8 mins, before the Stacks lab)

- Stats of your assemblies
- more k values in SPAdes
- Was the top assembly unbeatable?

	assembler used	Dataset (500k or full? pacbio?)	Value(s) of k, if applicable	number of scaffolds/contigs (larger than 0 bp)	total size of assembly (Kbp) ("Kbp" means: times 1000 base pairs)	scaffold NG50 (Kbp)	scaffold NGA50 (Kbp) (if you have not computed this metric you are not eligible for the contest!)	contig N50 (Kbp)	(op nur pro pre Pro
Maggie Sefton	SPAdes	Full+Pacbio	5,7,11,15,21,25,33,55,69,77				running	running	
Abby Schiff	SPAdes	full + pacbio	default	53	4063	806	806	950	
Gema Alama	Spades	Full+Pacbio	21,33,55	36	4054	806	806	939	
Jeremias Brand	Spades	full + pacbio	default	53	4063		806	950	
Mario Vicente	spades	full + pacbio	default	53	4063	806	806	950	
Peter Christ	SPAdes	full+pac	default	53	4059	806	806	950	
Rosario Castañeda	SPAdes	full + pacbio	multi-k	54	4058	806	806	939	
Sergio	SPAdes	full + pacbio	multi-k	35	4059	806	806	950	
Tim Nice	SPAdes	full + pacbio	27,39,51,63,75	44	4060	810	747	810	
Tim Nice	SPAdes	full + pacbio	21,33,55,77	46	4060	807	747	807	
Krystyna Cwiklinski	SPAdes	Full +pacbio	21,31,33,55,77	50	4060	806	746	806	
Reinder Radersma	SPAdes	Full+Pacbio	21,33,55,75	45	4065	810	746	810	
Greg McCracken	SPAdes	full+pacbio			4053		638	638	
Stefan Ciaghi	SPAdes	full + pacbio1	21,33,55,77,127	62	4049	638	638	638	
Hugo, Aurelie, Andrea	SPAdes	full + pacbio	27,31,37	55	4042	637	637	637	
Sandra Lorena Ament	SPAdes	full + PacBio	21,33,55,77	62	4055	637	637		
Jamie S.	SPAdes	500 + 2 x pb	31	102	4038	274	566		
Willian Silva	SPAdes	500k + pacbio	default	48	4045	274	470	470	
Beatriz Willink	SPAdes	full+pacbio	25,37,59,81	51	4054	428	428	428	
Reinder Radersma	SPAdes	500k + pacbio	default	59	4060	418	370	481	

147 entries (10 more than last year!)

Data cleaning

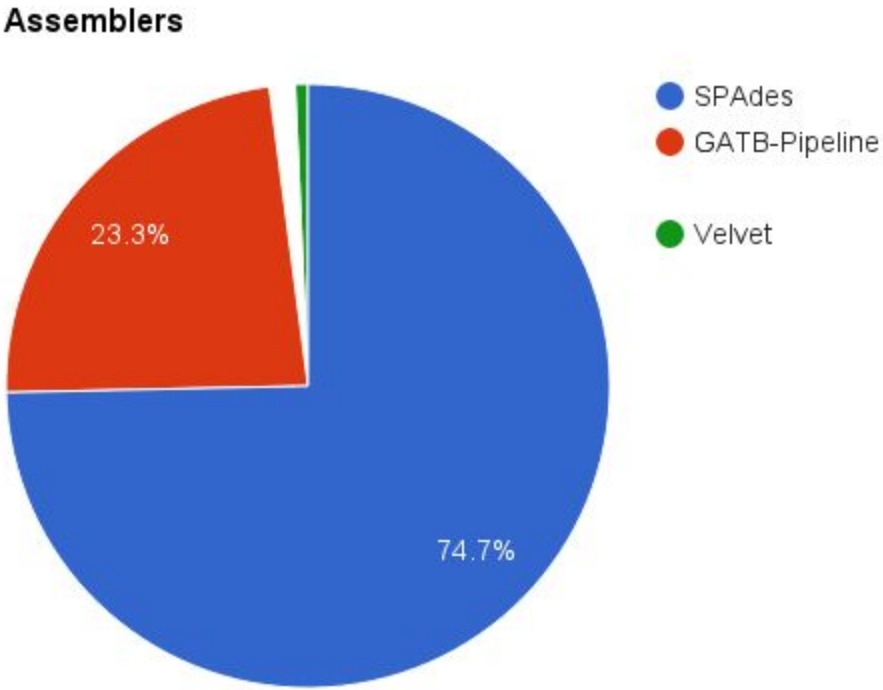


Code.gs

```
function assembler(name) {
  if (name.toLowerCase().indexOf("spades") > -1)
    return "SPAdes";
  if (name.toLowerCase().indexOf("gatb") > -1)
    return "GATB-Pipeline";
  return name;
}
```

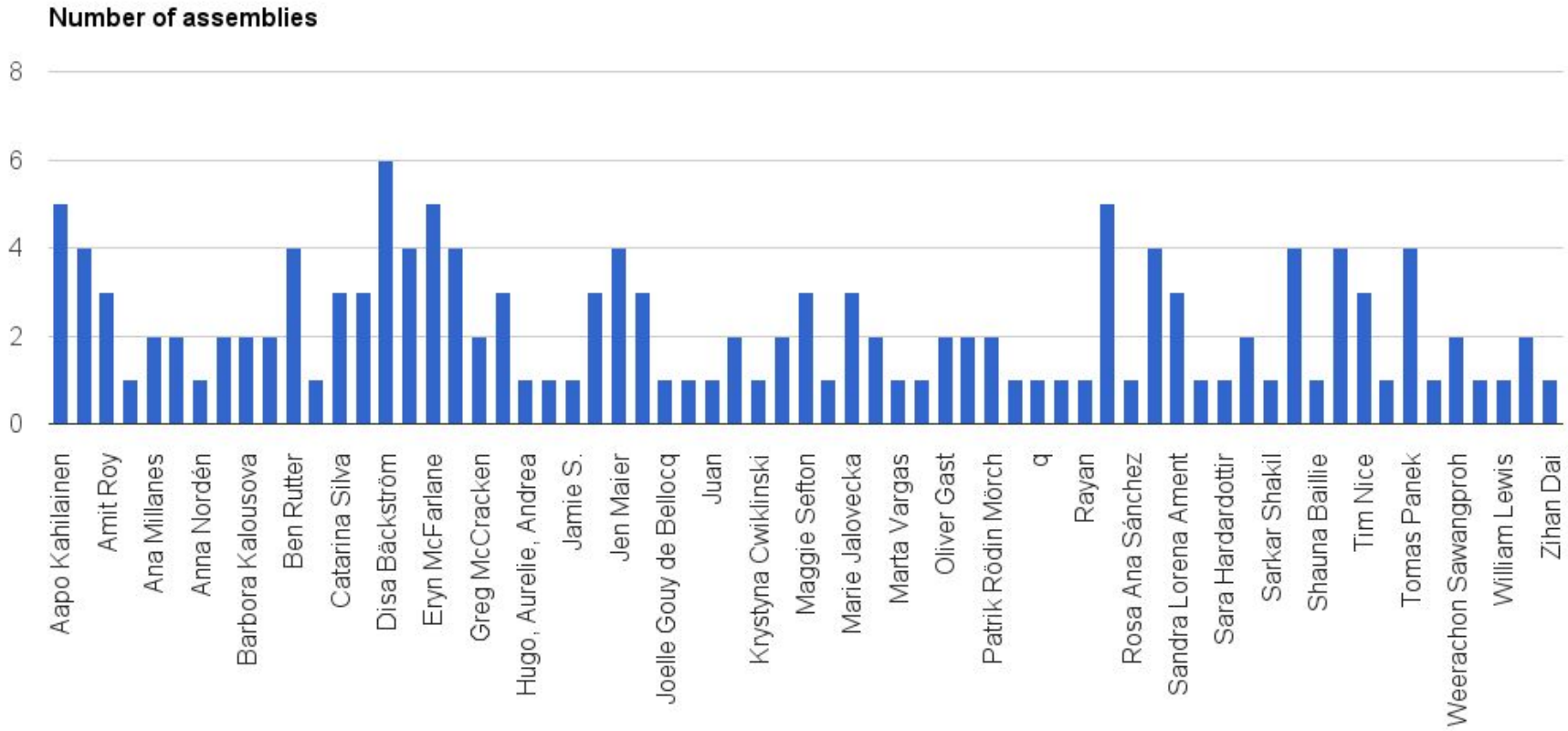
=assembler(C2)

formatted assembler	assembler used
SPAdes	SPAdes
GATB-Pipeline	GTAB-Pipeline
GATB-Pipeline	GATB
SPAdes	Spades
SPAdes	SPAdes

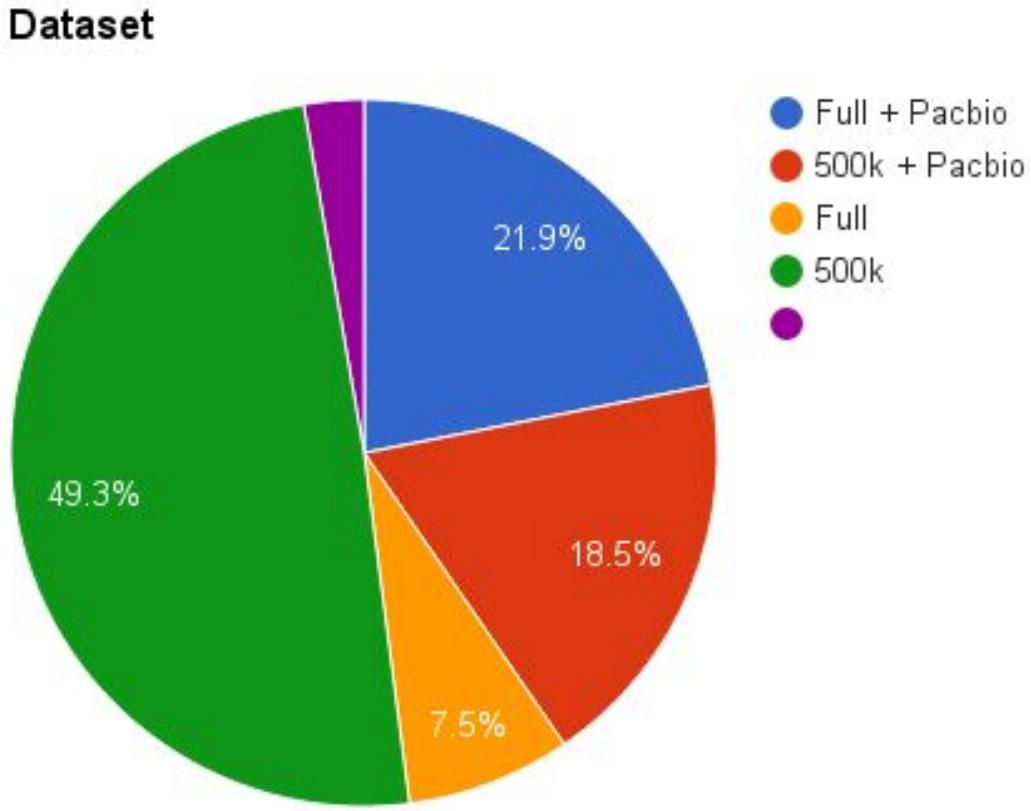


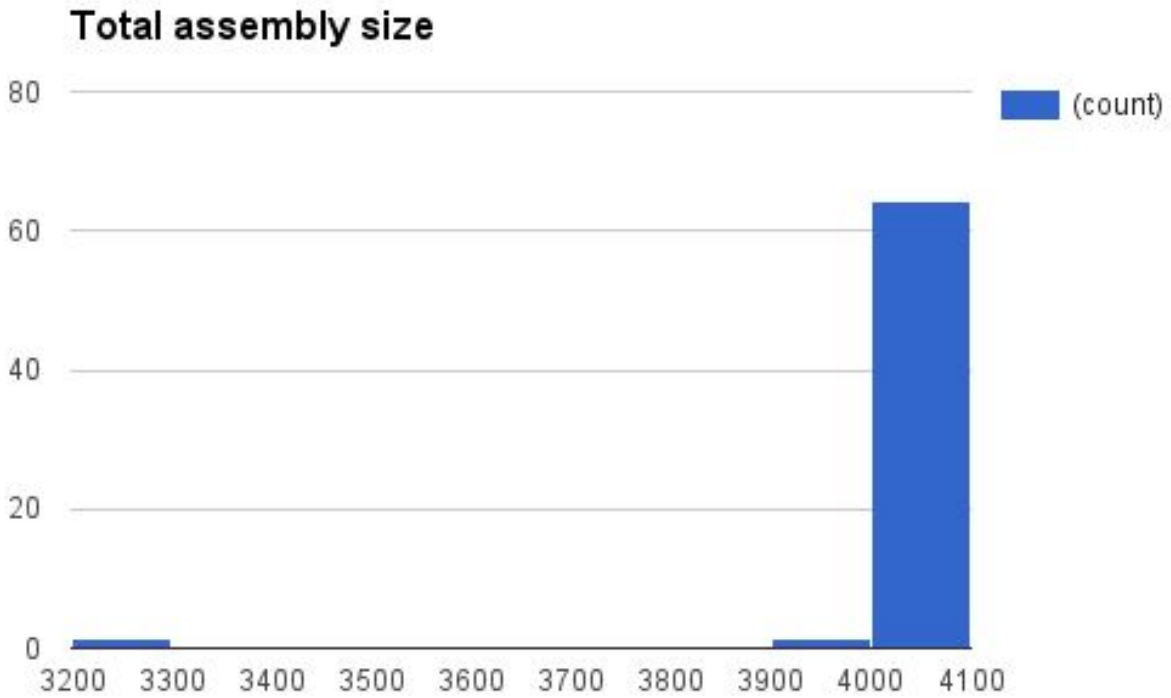
Could also have tried:

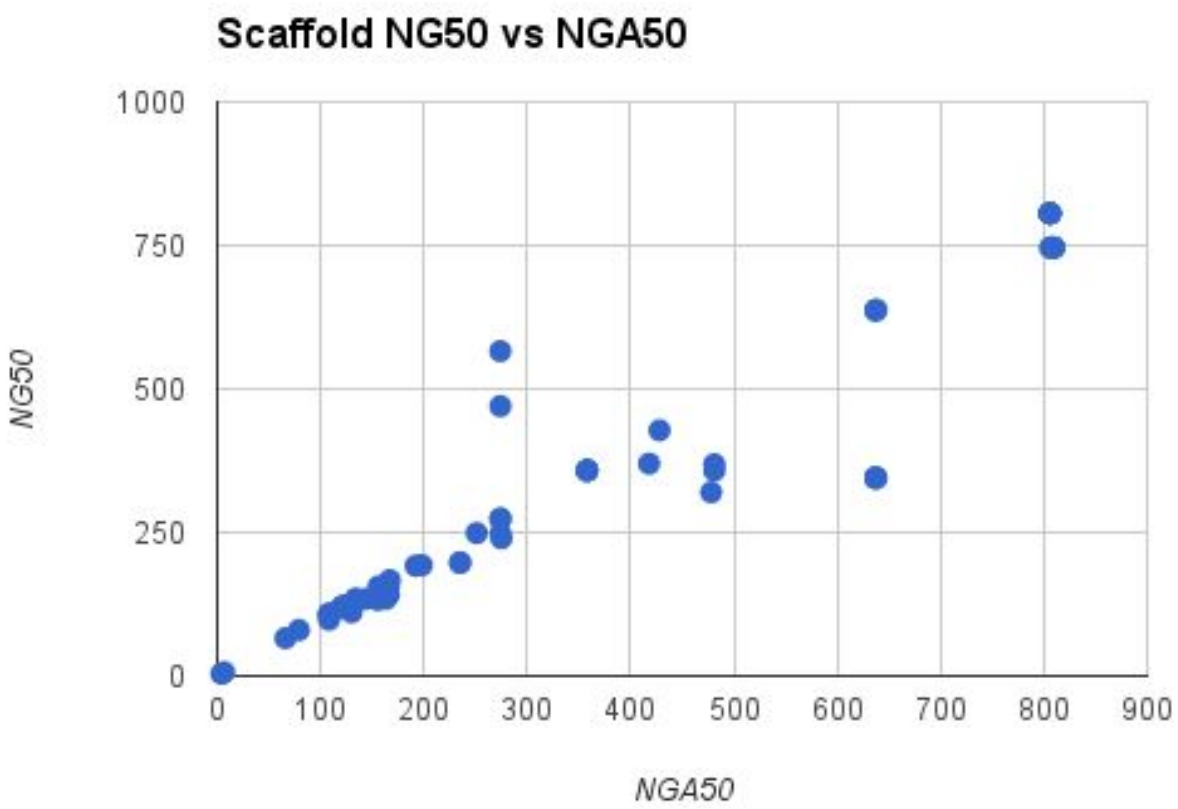
- MaSuRCA
- ABySS
- Megahit

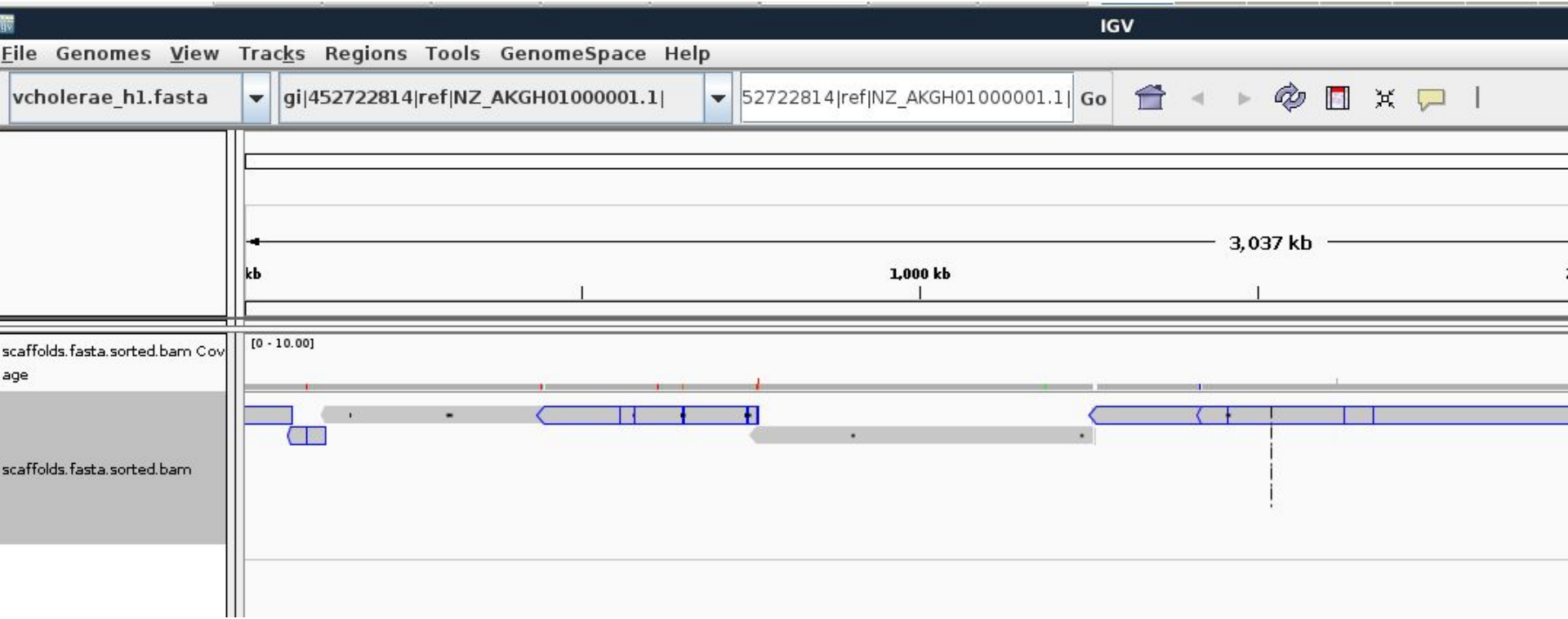


last year, record was 8 assemblies (Luca, with minia + velvet)


Dataset
21.9%
18.5%
7.5%
49.3%
Full + Pacbio
500k + Pacbio
Full
500k


Total assembly size
80
60
40
20
0
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
(count)


Scaffold NG50 vs NGA50
1000
750
500
250
0
NG50
0
100
200
300
400
500
600
700
800
900
NGA50

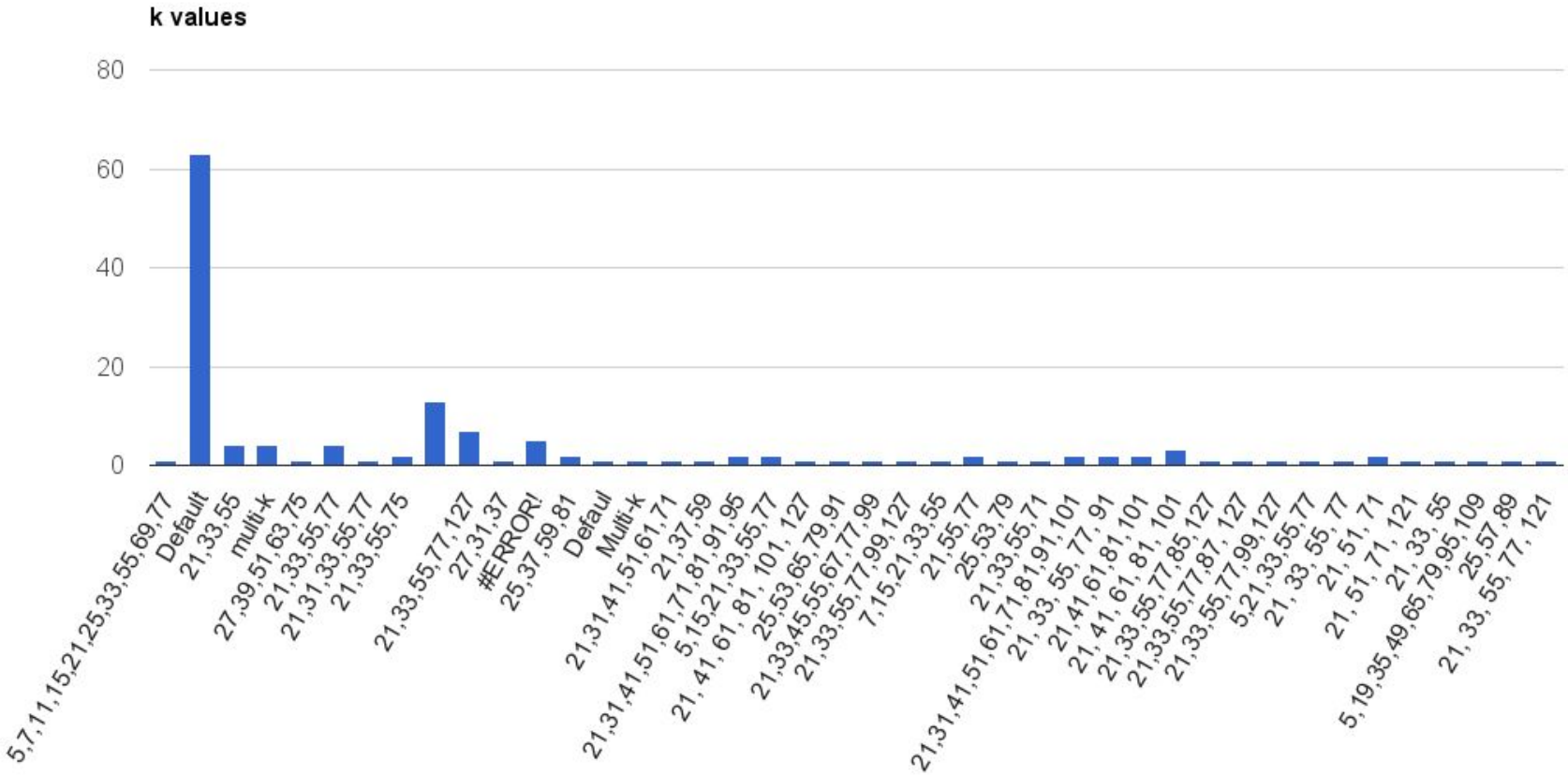

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
vcholerae_h1.fasta
gi|452722814|ref|NZ_AKGH01000001.1|
52722814|ref|NZ_AKGH01000001.1|
Go
3,037 kb
kb
1,000 kb
scaffolds.fasta.sorted.bam Cov
age
[0 - 10.00]
scaffolds.fasta.sorted.bam

PacBio-only assembly

- We had around 10x PacBio (both fastq files combined)
- Canu assembler (super-recent, came out last week)

assembly with Canu: N50 = 7 Kbp, total size = 1.2 Mbp

- Canu wants > 25x coverage


k values
80
60
40
20
0
5,7,11,15,21,25,33,55,69,77
Default
21,33,55
multi-k
27,39,51,63,75
21,33,55,77
21,31,33,55,77
21,33,55,75
21,33,55,77,127
27,31,37
#ERROR!
25,37,59,81
Defaul
Multi-k
21,31,41,51,61,71
21,37,59
21,31,41,51,61,71,81,91,95
5,15,21,33,55,77
21, 41, 61, 81, 101, 127
25,53,65,79,91
21,33,45,55,67,77,99
21,33,55,77,99,127
7,15,21,33,55
21,55,77
25,53,79
21,33,55,71
21,31,41,51,61,71,81,91,101
21, 33, 55, 77, 91
21,41,61,81,101
21, 41, 61, 81, 101
21,33,55,77,85,127
21,33,55,77,87,127
21,33,55,77,99,127
5,21,33,55,77
21, 33, 55, 77
21, 51, 71
21, 51, 71, 121
21, 33, 55
5,19,35,49,65,79,95,109
25,57,89
21, 33, 55, 77, 121

Giving more k values to SPAdes

Default:

# misassemblies	**2**
# misassembled contigs	**2**
Misassembled contigs length	**1098337**
# local misassemblies	**7**
# unaligned contigs	0 + 0 part
Unaligned length	0
Genome fraction (%)	**98.976**
Duplication ratio	1.003
# N's per 100 kbp	0.00
# mismatches per 100 kbp	**4.10**
# indels per 100 kbp	**4.72**
Largest alignment	945333
NA50	**806420**
NGA50	**806420**

Custom list of k values (17,21,33,55,79)

# misassemblies	**1**
# misassembled contigs	**1**
Misassembled contigs length	**900932**
# local misassemblies	**6**
# unaligned contigs	0 + 0 part
Unaligned length	0
Genome fraction (%)	**99.023**
Duplication ratio	1.003
# N's per 100 kbp	0.00
# mismatches per 100 kbp	**3.38**
# indels per 100 kbp	**5.21**
Largest alignment	945346
NA50	**746038**
NGA50	**746038**

This is why a single metric, no matter how good it is (NGA50), is not a reliable way to compare assemblies. *Apologies to those who tuned SPAdes with more k values and didn't win the contest. You are the heroes the assembly lab needs, not the ones it deserves.*

Rayan’s attempt

- Started with SPAdes full+pacbio, default k values
- SSPACE-Longread scaffolding with same pacbio data

# misassemblies	**5**
# misassembled contigs	**4**
Misassembled contigs length	1415180
# local misassemblies	**15**
# unaligned contigs	0 + 0 part
Unaligned length	0
Genome fraction (%)	**99.184**
Duplication ratio	1.003
# N's per 100 kbp	108.42
# mismatches per 100 kbp	3.90
# indels per 100 kbp	4.51
Largest alignment	1190600
NA50	**973806**
NGA50	**973806**

:/

For reference, SPAdes default:

# misassemblies	**2**
# misassembled contigs	**2**
Misassembled contigs length	1098337
# local misassemblies	**7**
# unaligned contigs	0 + 0 part
Unaligned length	0
Genome fraction (%)	**98.976**
Duplication ratio	1.003
# N's per 100 kbp	0.00
# mismatches per 100 kbp	**4.10**
# indels per 100 kbp	**4.72**
Largest alignment	945333
NA50	**806420**
NGA50	**806420**