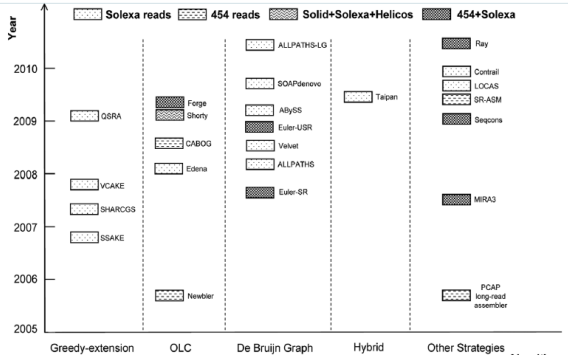
Additional Notes on this morning's Practicals

Hendrik-Jan Megens

WAGENINGEN UNIVERSITY
WAGENINGEN UR

There is an overwhelming number of assemblers to choose from ….



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Zhang et al., 2011, PLoS One 6:e17915

Many complex (and less complex) genomes are being assembled right now, or in the near future – how well do modern assemblers perform on modern data?

GENOME 10K©
THE ASSEMBLATHON

WAGENINGEN UNIVERSITY
WAGENINGEN UR

http://assemblathon.org
http://cnag.bsc.es

Demonstration data sets derived from GAGE

Resource

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

http://genome.cshlp.org/content/early/2012/01/12/gr.131383.111

WAGENINGEN UNIVERSITY
WAGENINGEN UR

GAGE website and characteristics

Genome Assembly Gold-Standard Evaluations

How does GAGE differ from the other assembly bake-offs?

At least two other assembly comparisons have been announced for 2011, the Assemblathon (http://assemblathon.org/) and dnGASP (http://cnag.bsc.es/). Here are some differences:

1. *GAGE is being run by assembly experts.* Our team has assembled hundreds of genomes, and has written some of the leading genome assembly software. We have been evaluating assemblers for more than 10 years. All of the assemblies and the comparisons among them will be conducted by experts.
2. *All natural ingredients.* GAGE will use FOUR different whole-genome shotgun data sets, all from recent sequencing projects. Assemblathon and dnGASP will both use simulated data. Who can say how simulated data relates to real results? We prefer the real thing.
3. *Completely open protocols.* We will compare multiple genome assembly programs, and we will describe all the parameters used to run them. We will also describe all the steps we take in cleaning up the data (pre-processing) and scrubbing the output of the assembly programs (post-processing). All of our results will thus be reproducible by anyone with sufficient computing resources.

WAGENINGEN UNIVERSITY
WAGENINGEN UR

GAGE has tested 8 assemblers

The assemblers

We chose eight of the leading genome assemblers, each of which is able to run large, whole-genome assemblies using Illumina-only short read data:

- ABySS (Simpson et al. 2009)
- ALLPATHS-LG (Gnerre et al. 2011)
- Bambus2 (Koren et al. 2011) (http://www.cbcb.umd.edu/software/bambus).
- CABOG (Miller et al. 2008)
- MSR-CA (http://www.genome.umd.edu/MSR_CA_MANUAL.htm)
- SGA (Simpson and Durbin 2012)
- SOAPdenovo (Li et al. 2010b)
- Velvet (Zerbino and Birney 2008)

WAGENINGEN UNIVERSITY
WAGENINGEN UR

The GAGE website provides four datasets for four genomes of increasing complexity

Table 1. Details of the four next-generation sequence data sets used for the GAGE assembly comparison

Species	*S. aureus*	*R. sphaeroides*	Human Chr14	*B. impatiens*
Size (Mb)	2.90	4.60	88.29	250 (est.)
Read length	101, 37	101	101	124
Fragment size, Library 1	180	180	155	400
Number of reads, Library 1	1,294,104	2,050,868	36,504,800	303,118,594
Fragment size, Library 2	3500	3500	2280–2800	3000–4000
Number of reads, Library 2	3,494,070	2,050,868	22,669,408	129,118,270
Fragment size, Library 3			35 kb	8 kb
Number of reads, Library 3			2,405,064	65,081,280

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Genome Assembly Gold-Standard Evaluations

Main page | Genome Assemblers | Data sets | Recipes | Results | Twitter

What data sets were used in GAGE?

Four different whole-genome shotgun data sets were studied, all from recent sequencing projects:

1. ***Staphylococcus aureus*: Data download page**
   - Library 1: Fragment
     - Avg Read length: 101bp
     - Insert length: 180bp
     - # of reads: 1,294,104
     - Fastq read file 1
     - Fastq read file 2
   - Library 2: Short jump library
     - Avg Read length: 37bp
     - Insert length: 3500bp
     - # of reads: 3,494,070
     - Fastq read file 1
     - Fastq read file 2
2. ***Rhodobacter sphaeroides*: Data download page**
   - Library 1: Fragment
     - Avg Read length: 101bp
     - Insert length: 180bp
     - # of reads: 2,050,868
     - Fastq read file 1
     - Fastq read file 2
   - Library 2: Short jump library
     - Avg Read length: 101bp
     - Insert length: 3500bp
     - # of reads: 2,050,868
     - Fastq read file 1
     - Fastq read file 2
3. **Human Chromosome 14: Data download page**
   - Library 1: Fragment
     - Avg Read length: 101bp
     - Insert length: 155bp
     - # of reads: 36,504,800
     - Fastq read file 1
     - Fastq read file 2
   - Library 2: Short jump library
     - Avg Read length: 101bp
     - Insert length: 2283-2803bp
     - # of reads: 22,669,408
     - Fastq read file 1
     - Fastq read file 2
   - Library 3: Long jump library
     - Avg Read length: 76-101bp
     - Insert length: 35,295-35,318bp
     - # of reads: 2,405,064
     - Fastq read file 1
     - Fastq read file 2
4. ***Bombus impatiens* (bumblebee): Data download page**
   - Library 1: Fragment
     - Avg Read length: 124bp
     - Insert length: 400bp
     - # of reads: 303,118,594
     - Fastq read file 1
     - Fastq read file 2
   - Library 2: Short jump library 1
     - Avg Read length: 124bp
     - Insert length: 3-4Kbp
     - # of reads: 129,118,270
     - Fastq read file 1_1 , 1_2
     - Fastq read file 9_1 ,9_2
   - Library 3: Short jump library 2
     - Avg Read length: 124bp
     - Insert length: 8Kbp
     - # of reads: 65,081,280
     - Fastq read file 1
     - Fastq read file 2

WAGENINGEN UNIVERSITY
WAGENINGEN UR

The *S. aureus* and HChr14 assemblies have been benchmarked

Table 2. Assemblies of *S. aureus* (genome size 2,872,915)

	Contigs				Scaffolds			
Assembler	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	302	29.2	19	24.8	246	34	1	28
ALLPATHS-LG	**60**	96.7	20	**66.2**	**12**	1,092	**0**	**1,092**
Bambus2	109	50.2	190	16.7	17	1,084	**0**	1,084
CABOG	Could not run: incompatible read lengths in one library							
MSR-CA	94	59.2	34	48.2	17	**2,412**	3	1,022
SGA	252	4.0	**10**	4.0	456	208	1	208
SOAPdenovo	107	**288.2**	65	62.7	99	332	8	284
Velvet	162	48.4	42	41.5	45	762	17	126

Table 4. Assemblies of human chromosome 14 (ungapped size 88,289,540)

	Contigs				Scaffolds			
Assembler	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	51,924	2.0	**704**	2.0	51,301	2.1	**9**	2
ALLPATHS-LG	4529	36.5	2760	21.0	**225**	**81,647**	45	**4702**
Bambus2	13,592	5.9	11,943	4.3	1792	324	143	161
CABOG	**3361**	**45.3**	3181	**23.7**	479	393	597	26
MSR-CA	30,103	4.9	5550	4.3	1425	893	1068	94
SGA	56,939	2.7	981	2.7	30,975	83	19	79
SOAPdenovo	22,689	14.7	6424	7.4	13,502	455	268	214
Velvet	45,564	2.3	4910	2.1	3,565	1190	9156	27

WAGENINGEN UNIVERSITY
WAGENINGEN UR

… as are the other two. The bumble bee genome could not be assembled by all

Table 3. Assemblies of *R. sphaeroides* (genome size 4,603,060)

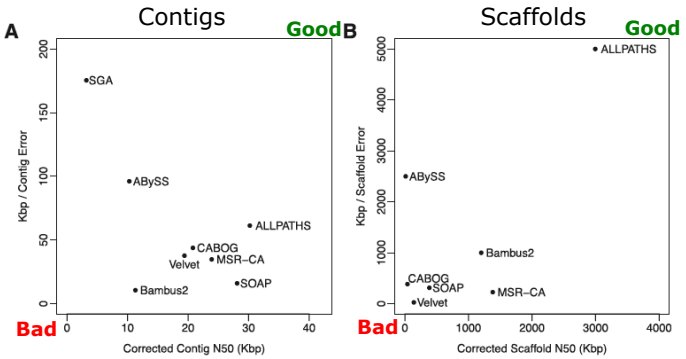
	Contigs				Scaffolds			
Assembler	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	1915	5.9	76	4.2	1701	9	3	5
ALLPATHS-LG	204	42.5	49	**34.4**	**34**	**3192**	**0**	**3192**
Bambus2	**177**	93.2	373	12.8	92	2439	2	2419
CABOG	322	20.2	44	17.9	130	66	5	55
MSR-CA	395	22.1	52	19.1	43	2,976	5	2966
SGA	3067	4.5	**12**	2.9	2096	51	**0**	51
SOAPdenovo	204	**131.7**	422	14.3	166	660	3	658
Velvet	583	15.7	43	14.5	178	353	6	270

Table 5. Assemblies of the bumble bee, *B. impatiens* (estimated size 250 Mb)

	Contigs			Scaffolds		
Assembler	Num	N50 (kb)	E-size (kb)	Num	N50 (kb)	E-size (kb)
ALLPATHS-LG	Could not run: incompatible library types					
CABOG	22,107	23.5	34.2	**1191**	1125	1367
MSR-CA	21,885	32.4	46.9	2551	1246	1528
SGA	Program crashed: cause unclear					
SOAPdenovo	**15,957**	**57.1**	**78.2**	5800	**1374**	**1608**
Velvet	Program crashed: insufficient memory (256 GB)					

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Graphical evaluation of assemblers:



WAGENINGEN UNIVERSITY
WAGENINGEN UR

The GAGE-paper provides various assembly quality evaluations, including comparison to a published reference

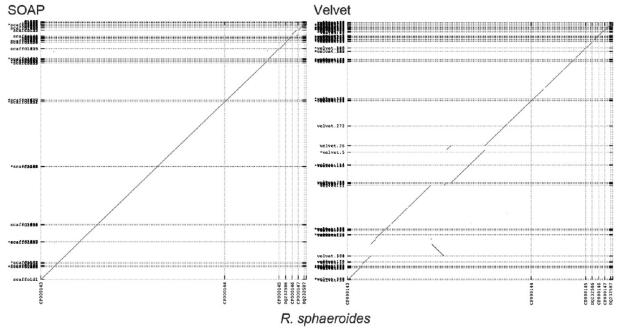


Figure 2. A dot-plot comparison of the SOAPdenovo and Velvet scaffolds of *R. sphaeroides*. The finished reference chromosomes are plotted on the *x*-axis and the assembly scaffolds on the *y*-axis. Dotted lines indicate scaffold or chromosome boundaries. The apparent rearrangement at the *top right* of the SOAPdenovo plot is an artifact of the circular reference plasmid.

WAGENINGEN UR

Each assembler can introduce unique artifacts

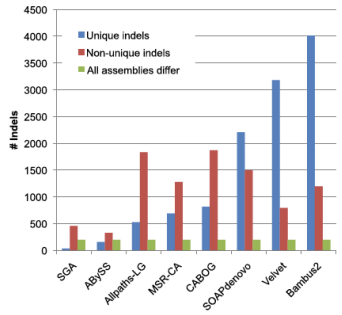
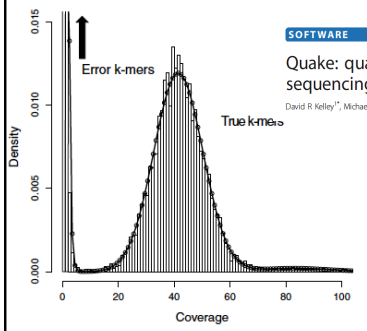


Figure 5. Comparison of insertion and deletion errors among all eight assemblers for human chromosome 14. (Blue) The indel errors >5 bp in length that are unique to each assembler. (Red bars) Indel errors made by at least one other assembler. (Green bars) Indels shared by all assemblers, which might represent true differences between the target genome and the reference.

Data cleaning and error correction: among the most important steps



Kelley et al. Genome Biology 2010, 11:R116
http://genomebiology.com/2010/11/1/R116

Genome Biology

SOFTWARE Open Access

Quake: quality-aware detection and correction of sequencing errors

David R Kelley*, Michael C Schatz, Steven L Salzberg

A likelihood model is applied in error correction

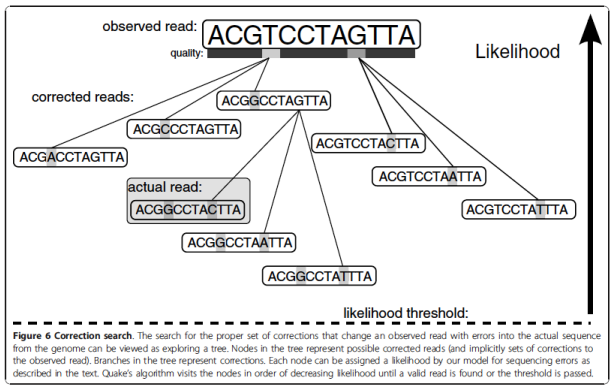


Figure 6 Correction search. The search for the proper set of corrections that change an observed read with errors into the actual sequence from the genome can be viewed as exploring a tree. Nodes in the tree represent possible corrected reads (and implicitly sets of corrections to the observed read). Branches in the tree represent corrections. Each node can be assigned a likelihood by our model for sequencing errors as described in the text. Quake's algorithm visits the nodes in order of decreasing likelihood until a valid read is found or the threshold is passed.

Of course, everything you learn today on *De Novo* assembly, may be completely obsolete next year….

Practical this morning:

- GAGE datasets: *Staphylococcus areus* and human chromosome 14
- Pre-cleaned and pre-corrected datasets
- SOAPdenovo assembler
- Basic evaluation (N50; MUMmer).

WAGENINGEN UNIVERSITY
WAGENINGEN UR