CSH PRESS GENOME RESEARCH

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent A. Earl, Keith Bradnam, John St. John, et al.

Genome Res. published online September 16, 2011
Access the most recent version at doi:10.1101/gr.126599.111

THE ASSEMBLATHON

Home
News
Rules
Download data
Timetable
Contact us

Home

Reidar Andreson
Journal Club – 26.09.2011

What is the Assemblathon?

The Assemblathon is a set of periodic collaborative efforts that all help improve methods of genome assembly. It will hopefully become an annual event that will spur improvements in this computationally intensive field. The overall goal of each Assemblathon event is to have participating groups try to use their own software to each assemble one or more genomes that the organizers of the Assemblathon will make available (see the rules page for more details of the latest challenge). All participants will have the same

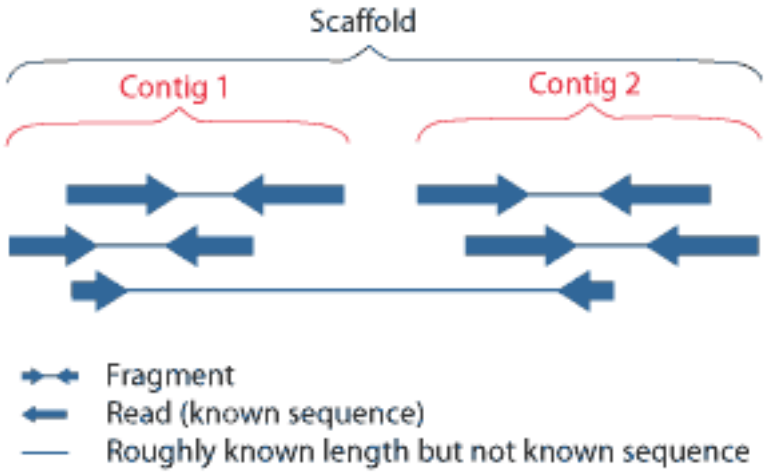
Genome assembly

- Assembly – *in silico* merging and ordering shorter sequence fragments (**reads**) sampled from a set of larger sequences in order to reconstruct contiguous sequence fragments (**contigs**)

- Scaffolds – ordered and oriented contigs creating larger sequences

• Multi-stage pipelines of assemblers:

  ✓ Correcting reads measurement errors
  ✓ Constructing contigs
  ✓ Resolving repeats
  ✓ Scaffolding contigs



Assembler types

- First assemblers using overlaps or string graphs and small input volumes: Phrap, GigAssembler, Celera, ARACHNE, and Phusion
- Overlap graph approach: Edena and Newbler, SGA
- Word look-up tables to greedily extend reads: SSAKE, SHARCGS, VCAKE, OligoZip, PRICE, and Monument
- Word look-up and de Bruijn graphs: Euler, AllPaths, and Velvet
- Optimal memory usage (fitting whole genomes): ABySS, Meraculous, SOAPdenovo, and Cortex

Assembly accuracy assessment

- Calculation of contig/scaffold length summary statistics:
  - N50
  - Total sequence lengths
  - Total number of produced sequences
  - Read coverage

  Additional methods:
- Reference genome
- Independent experimental proof
- Transcriptome information
- Closely related and well-sequences species (patterns of indels)
- Use of “core genes”

Why genome simulation?

- Possibility to use genome with no reasonable homology to anything other than own-created out-group genomes – fair blind test

- Define the proportions of the genome (eg. size) -> maximum number of participants covered

- They could simulate diploid genome with precisely and fully known haplotypes

Tasks

- Simulation of genome using Evolver (Edgar R, Asimenos G, Batzoglou S, Sidow A. http://www.drive5.com/evolver/)
- Read simulation using their own tool SimSeq (https://github.com/jstjohn/SimSeq)
- 17 different groups providing *de novo* assemblies
- Multiple sequence alignment with Cactus for assessing assembly relationships with simulated sequences
- Additional BLAST verifications

Genome simulation

- Evolver input: human chr13 (95.6 Mb), non-N, divided into 4 chromosomes
- ~200 my to generate most recent common ancestor (MRCA), ~50 my for two lineages

The phylogeny of the simulated haploid genomes. The root genome derives from human chromosome 13. The *α1 and α2 haplotypes form the diploid genome from which we generated reads. The β and β2 haplotypes form a diploid out-group genome that was made available to the assemblers.*

A

Genome	Mb	GC (%)	Reps (%)	Reps 100mer (%)	Chr	Subs	Dels	Inv	Moves	Copy	Tandem	Chr Split	Chr Fuse
Input	95.6	38.8	7.1 / 42.3*	0.8	4	–	–	–	–	–	–	–	–
MRCA	109.4	39.9	6.9	0.3	2	35.9e+06	2.47e+06	11,701	4,714	14,644	1.16e+06	2	4
α	112.4	40.0	7.5	0.3	3	9.70e+06	6.72e+05	3,325	1,369	4,151	3.13e+05	1	0
α_1	112.5	40.0	7.5	0.3	3	1.97e+05	13,528	54	34	83	6,436	0	0
α_2	112.5	40.0	7.5	0.3	3	1.97e+05	13,834	61	31	80	6,494	0	0
β	112.3	40.0	6.8	0.3	2	9.71e+06	6.74e+05	3,313	1,325	4,043	3.14e+05	0	0
β_1	112.4	40.0	6.8	0.3	2	1.97e+05	13,632	64	26	82	6,354	0	0
β_2	112.4	40.0	6.8	0.3	2	1.97e+05	13,621	71	35	79	6,445	0	0

B

Comparison	SNPs	Substitutions	∑ Subs	Indels	∑ Indels	Inversions
$\alpha_1 \alpha_2$	439,385	441,796	444,247	29,972	521,142	115

Genome simulation statistics. (A) Event numbers are between the previous branch point and the named node. Mb: size of the genome in megabases; GC: percentage GC content; Reps: percent of the genome masked by the union of tandem repeats finder and RepeatMasker, *is the published value for chromosome 13 (Dunham et al. 2004); Reps 100mer: percent repetiteveness of the sequence and its reverse complement for 100-mers calculated with the tallymer tool (Kurtz et al. 2008); Chr: number of chromosomes; Subs: number of substitution events; Dels: number of deletion events; Inv: number of inversion events; Moves: number of translocations; Copy: number of DNA segmental duplications; Tandem: number of tandem repeat insertions; Chr Split: number of chromosome fission events; Chr Fuse: number of chromosome fusion events. (B) Differences between haplotypes *α1* and α2 as determined by inspection of the Evolver pairwise alignment. SNPs: count of single nucleotide polymorphisms; Subs: count of substitutions, including SNPs*; α* Subs: sum of the lengths of all substitutions; Indels: count of insertion deletion events*; α* Indels: sum of the lengths of all insertion deletion events; Inv: the sum of number of inversions invoked in each of the *α1* and *α2* Evolver steps.

Read simulation (1)

- One combined short read dataset with multiple read libraries for the Illumina Hi-seq 2000 platform
- No suitable software for that (Illumina in-house tools, dwgsim, metasim, PEMer, ReSeqSim, SimNext, Flux Simulator, Mason)
- Wrote their own simulator SimSeq that combines Illumina mate-pair and paired-end read modeling with their own empirical error models trained on Illumina data
- Added 3 copies of E. coli sequence to the two haplotype sequences (~5% bacterial contamination rate)

Read simulation (2)

- Paired-end sampling (rand. fragments uniformly)
  - 200 & 300 bp insert +/- 20 & 30 standard deviation
    - 2 x 100 bp
    - 22,499,731 read pairs (~40x coverage)
    - 0.01 probability of being a duplicate

- Mate-Pair sampling
  - 3 & 10 kb loop length +/- 300 & 1000 standard deviation
    - 2 x 100 bp
    - 500 bp loop fragmentation size +/- 50 bp
    - 11,249,866 read pairs (~20x coverage)
    - 0.05 & 0.08 probability of being a duplicate

Read simulation (3)

- Base-level error model – dependent on the position within the read and the underlying reference base
  - Human mitochondiral genome Illumina reads assembled with MIA
  - Mapped reads to assembly with BWA (default settings)
  - Kept all alignments with mapq score > 10
  - Using Phred to create empirical distribution of scores
  - Each position in read got quality score

Participants

ID	Affiliations	Entries	Software	Used β
ASTR	Agency for Science, Technology and Research, Singapore	1	PE-Assembler	No
WTSI-P	Wellcome Trust Sanger Institute, UK	2	Phusion2, phrap	No
EBI	European Bioinformatics Institute, UK	2	SGA, BWA, Curtain, Velvet	No
WTSI-S	Wellcome Trust Sanger Insitute, UK	4	SGA	No
CRACS	Center for Research in Advanced Computing Systems, Portugal	3	ABySS	Yes
BCCGSC	BC Cancer Genome Sciences Centre, Canada	5	ABySS, Anchor	No
DOEJGI	DOE Joint Genome Insititute, USA	1	Meraculous	No
IRISA	L'IRISA (Institut de recherche en informatique et systèmes aléatoires), France	5	Monument	No
CSHL	CSHL (Cold Spring Harbor Laboratory), USA	2	Quake, Celera, Bambus2	No*
DCISU	Department of Computer Science, Iowa State University	1	PCAP	No
IoBUGA	Computational Systems Biology Laboratory, University of Georgia, USA	3	Seqclean, SOAPdenovo	No
UCSF	UC San Francicso, USA	1	PRICE	Yes
RHUL	Royal Holloway, University of London, UK	5	OligoZip	No
GACWT	The Genome Analysis Centre, Sainsbury Laboratory, and Wellcome Trust Centre for Human Genetics, UK	3	Cortex_con_rp	No
CIUoC	Department of Computer Science, University of Chicago, USA	1	Kiki	No
BGI	BGI, Shenzhen China	1	SOAPdenovo	No
Broad	Broad Institute	1	ALLPATHS-LG	No
nVelv	—	6	Velvet	No
nCLC	—	9	CLC	No
nABySS	—	6	ABySS	No

Groups that submitted assemblies. The first 17 rows in the table correspond to entries submitted by participants in the competition. Assemblies with IDs beginning with "n," (for naïve), were generated by organisers of the competition to demonstrate the performance of popular programs run with variations on their default parameters.
*CSHL.1 used the *β* genome though that team's top assembly, CSHL.2, which is referred to in the main paper as CSHL, did not.

N50 and NG50

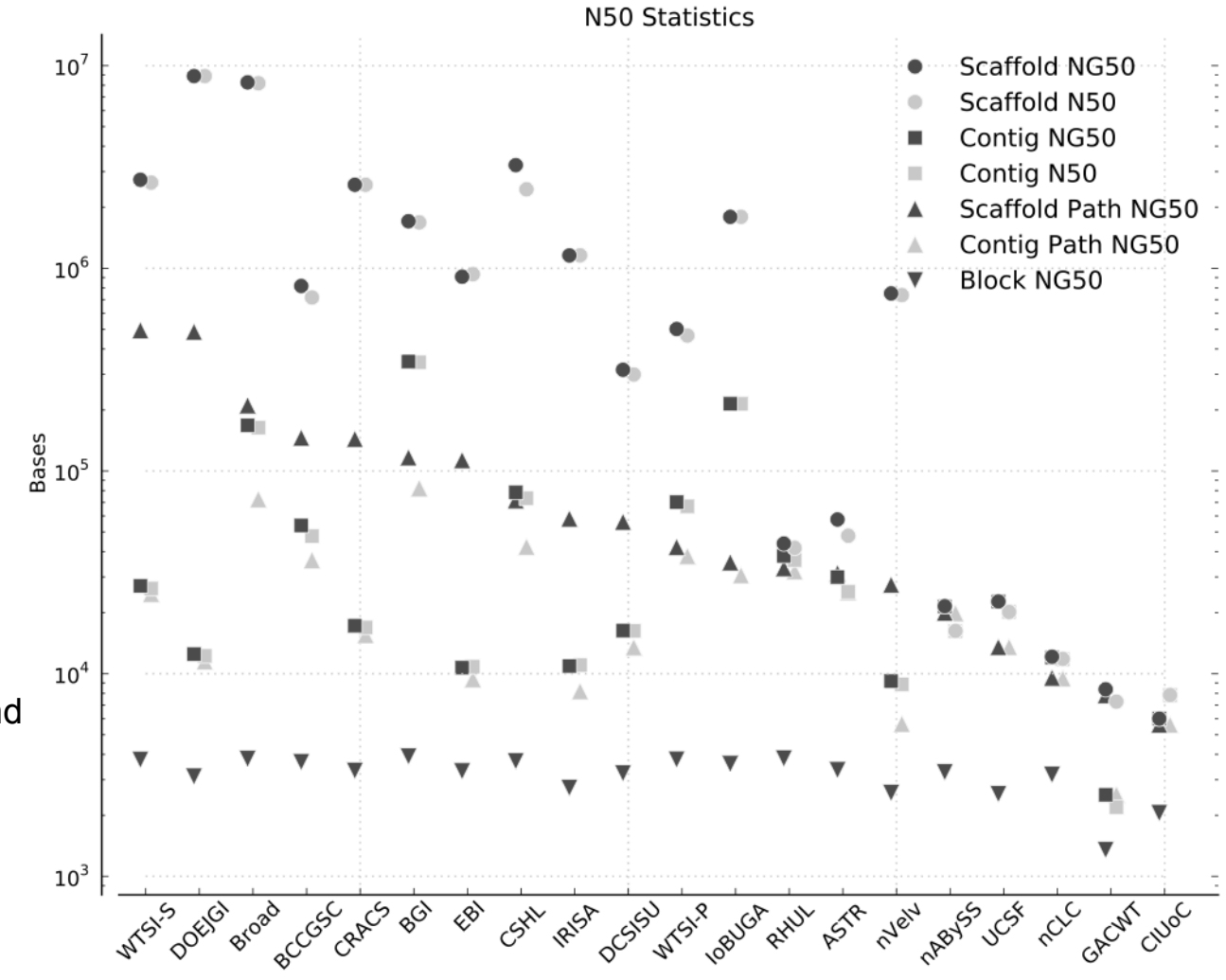
N50 – measure of the average length of a set of sequences, with greater weight given to longer scaffolds using total length of assembly

NG50 – identical to N50 except using total length of genome (average of *α*1 & *α*2)

Contig path – maximal subsequences of contigs that are entirely consistent with *α1,2.*

Scaffold path – maximal concatenations of contig paths and scaffold breaks that maintain correct order and orientation.

Block – maximal gapless alignment of a set of sequences (by MSA)



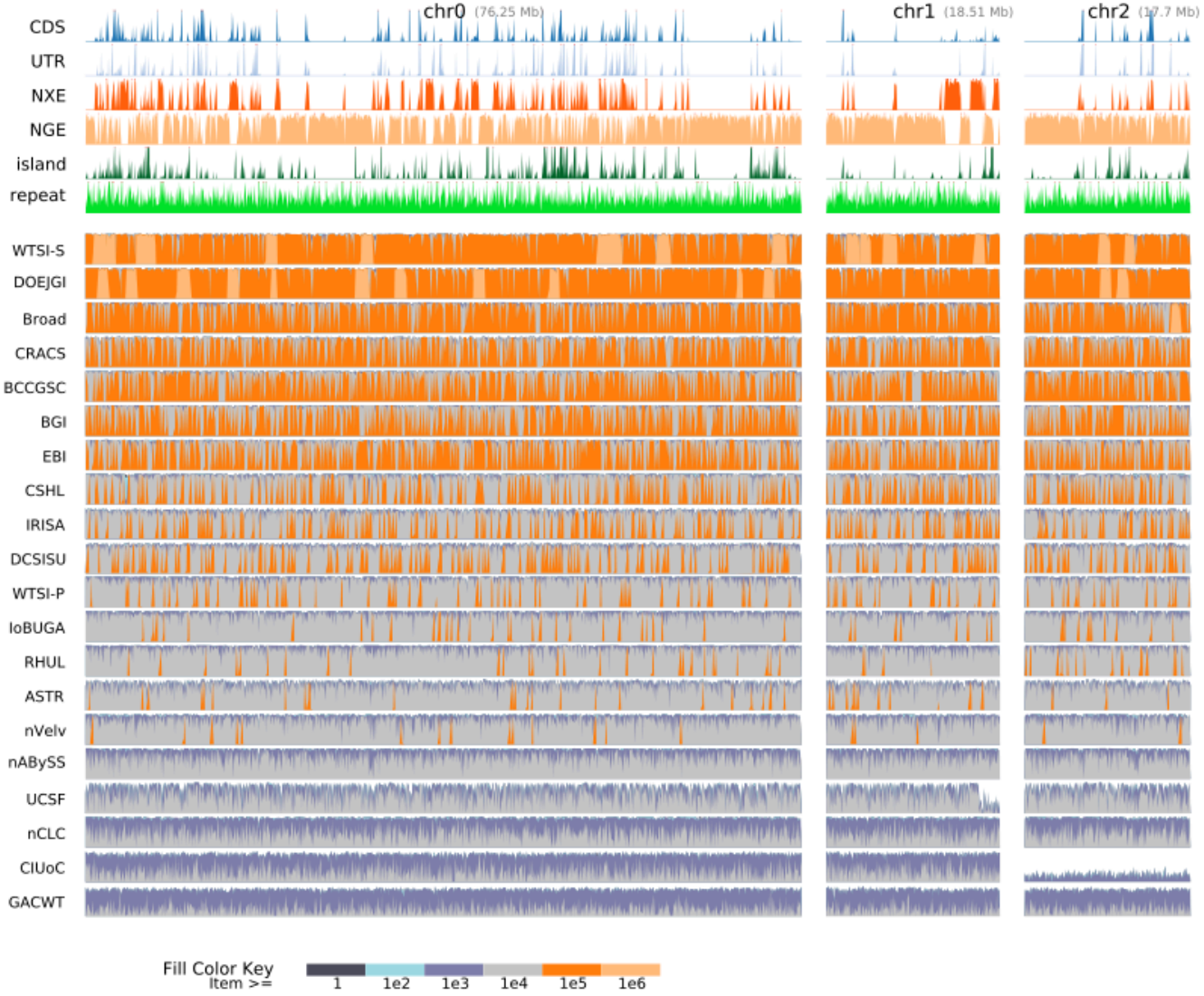
N50 statistics. Assemblies are sorted left to right in descending order by scaffold path NG50. Data points for each assembly are slightly offset along the x-axis in order to show overlaps.

Coverage

Assembly coverage along haplotype α*1* stratified by scaffold path length weighted overall coverage.

The top 6 rows show density plots of annotations. CDS: coding sequence; UTR: untranslated region; NXE: non-exonic conserved regions within genes; NGE: non-genic conserved regions; island: CpG islands; repeats: repetitive elements

For example, the left most light-orange block of the WTSI-S assembly row represents a region of haplotype α1 that is almost completely covered by a scaffold path from the WTSI-S assembly greater than one megabase in length.



Coverage statistics

ID	Hap Total (%)	Hap α_1 (%)	Hap α_2 (%)	Bac (%)	Genic (%)	Unmapped
BGI	98.8	98.9	98.8	0.0	92.7	2.637e+05
BCCGSC	98.7	98.7	98.7	99.9	88.9	6.546e+06
WTSI-P	98.7	98.7	98.7	99.8	75.0	5.369e+06
RHUL	98.5	98.5	98.5	100.0	67.4	4.961e+06
CSHL	98.5	98.6	98.5	99.9	89.1	7.815e+06
Broad	98.3	98.4	98.3	68.9	93.8	3.538e+06
IoBUGA	98.3	98.3	98.3	4.8	92.8	7.822e+05
WTSI-S	97.8	97.8	97.8	99.1	91.8	4.948e+06
EBI	97.7	97.7	97.7	0.9	88.5	4.553e+05
nABySS	97.5	97.5	97.5	99.8	57.2	1.111e+07
DOEJGI	97.3	97.4	97.3	99.5	92.3	5.304e+06
nCLC	97.2	97.2	97.2	99.8	55.4	5.673e+06
nVelv	96.5	96.6	96.5	99.8	84.8	8.028e+06
CRACS	96.3	96.3	96.3	99.8	90.2	5.265e+06
DCSISU	94.3	94.3	94.2	99.5	79.0	6.259e+06
IRISA	93.7	93.7	93.7	99.7	88.1	5.426e+06
ASTR	90.9	90.9	90.9	100.0	68.5	5.175e+06
GACWT	86.4	86.4	86.4	0.0	48.0	2.053e+06
UCSF	83.7	83.7	83.7	0.0	59.6	1.822e+06
CIUoC	78.5	79.0	78.1	0.6	48.9	3.638e+05

Coverage statistics for the top assembly from each team. Hap Total: overall coverage, Hap α*1: percent coverage for Haplotype α1, Hap β2: percent coverage for Haplotype β2, Bac: percent coverage of the bacterial contamination, Genic: percent coverage of the coding sequences (176 genes in total, >=95% coverage), Unmapped: number of unmapped bases, many corresponding to short contigs.*

Coverage statistics

ID	Hap Total (%)	Hap α_1 (%)	Hap α_2 (%)	Bac (%)	Genic (%)	Unmapped
BGI	98.8	98.9	98.8	0.0	92.7	2.637e+05
BCCGSC	98.7	98.7	98.7	99.9	88.9	6.546e+06
WTSI-P	98.7	98.7	98.7	99.8	75.0	5.369e+06
RHUL	98.5	98.5	98.5	100.0	67.4	4.961e+06
CSHL	98.5	98.6	98.5	99.9	89.1	7.815e+06
Broad	98.3	98.4	98.3	68.9	93.8	3.538e+06
IoBUGA	98.3	98.3	98.3	4.8	92.8	7.822e+05
WTSI-S	97.8	97.8	97.8	99.1	91.8	4.948e+06
EBI	97.7	97.7	97.7	0.9	88.5	4.553e+05
nABySS	97.5	97.5	97.5	99.8	57.2	1.111e+07
DOEJGI	97.3	97.4	97.3	99.5	92.3	5.304e+06
nCLC	97.2	97.2	97.2	99.8	55.4	5.673e+06
nVelv	96.5	96.6	96.5	99.8	84.8	8.028e+06
CRACS	96.3	96.3	96.3	99.8	90.2	5.265e+06
DCSISU	94.3	94.3	94.2	99.5	79.0	6.259e+06
IRISA	93.7	93.7	93.7	99.7	88.1	5.426e+06
ASTR	90.9	90.9	90.9	100.0	68.5	5.175e+06
GACWT	86.4	86.4	86.4	0.0	48.0	2.053e+06
UCSF	83.7	83.7	83.7	0.0	59.6	1.822e+06
CIUoC	78.5	79.0	78.1	0.6	48.9	3.638e+05

Coverage statistics for the top assembly from each team. Hap Total: overall coverage, Hap α*1: percent coverage for Haplotype α1, Hap β2: percent coverage for Haplotype β2, Bac: percent coverage of the bacterial contamination, Genic: percent coverage of the coding sequences (176 genes in total, >=95% coverage), Unmapped: number of unmapped bases, many corresponding to short contigs.*

Summary of results

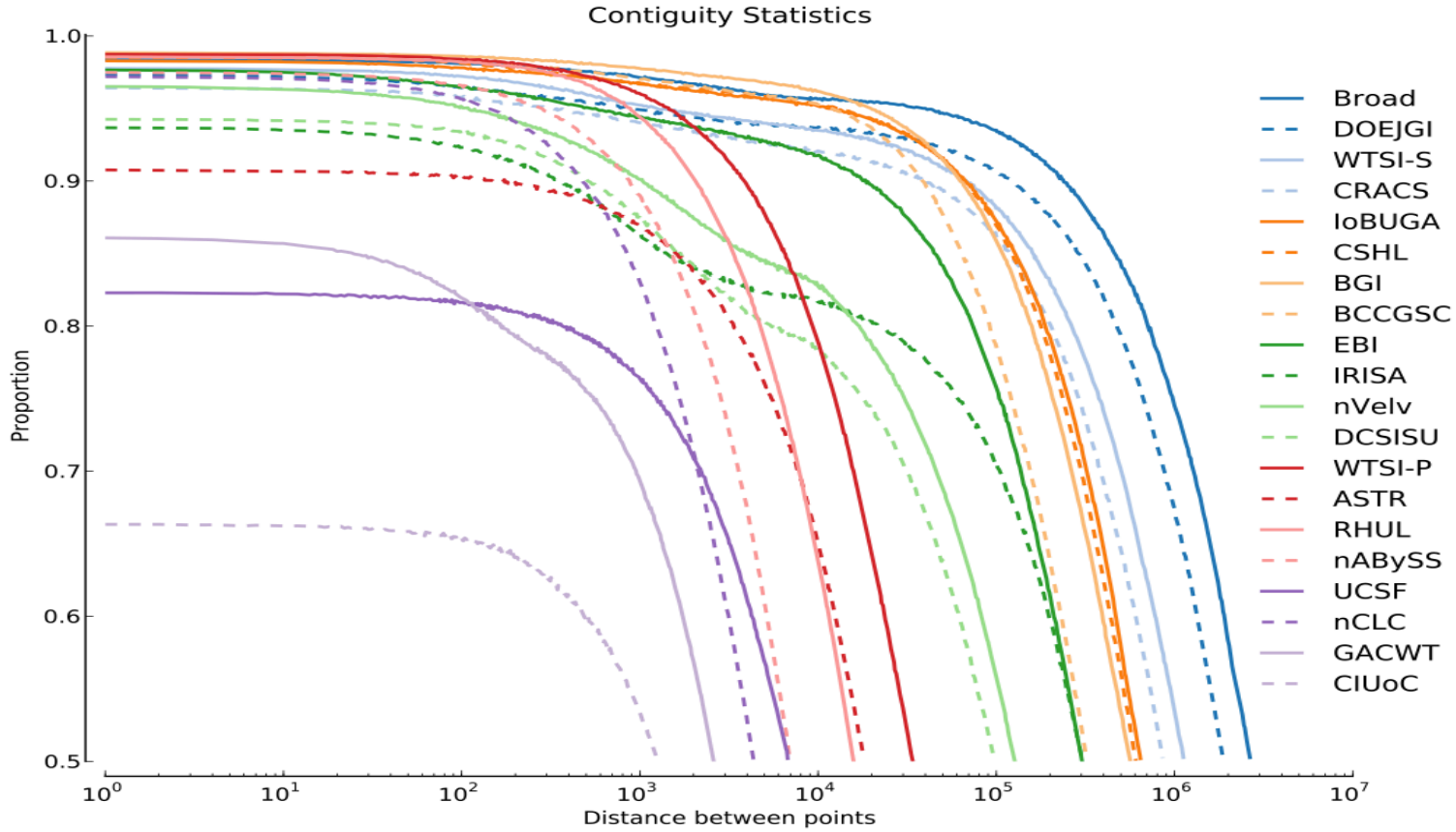
ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy Num.	Cov. Tot.	Cov. Genic
Broad	31	2 (7.25e+04)	3 (2.11e+05)	3 (1244)	1 (2.66e+06)	4 (2.92e-06)	11 (6.71e-02)	6 (98.3)	1 (93.8)
BGI	37	1 (8.23e+04)	6 (1.17e+05)	6 (1878)	7 (5.66e+05)	11 (1.20e-05)	2 (6.75e-03)	1 (98.8)	3 (92.7)
WTSI-S	38	9 (2.48e+04)	1 (4.95e+05)	2 (475)	3 (1.14e+06)	1 (1.30e-07)	9 (5.74e-02)	8 (97.8)	5 (91.8)
DOEJGI	44	14 (1.15e+04)	2 (4.86e+05)	1 (456)	2 (1.89e+06)	3 (4.43e-07)	7 (5.42e-02)	11 (97.3)	4 (92.3)
CSHL	57	3 (4.23e+04)	8 (7.17e+04)	14 (5146)	6 (6.11e+05)	9 (1.02e-05)	6 (4.95e-02)	4 (98.5)	7 (89.1)
CRACS	58	11 (1.55e+04)	5 (1.44e+05)	4 (1666)	4 (8.61e+05)	2 (3.81e-07)	12 (6.82e-02)	14 (96.3)	6 (90.2)
BCCGSC	60	5 (3.63e+04)	4 (1.46e+05)	10 (2867)	8 (3.22e+05)	8 (7.00e-06)	15 (1.17e-01)	2 (98.7)	8 (88.9)
EBI	64	16 (9.39e+03)	7 (1.13e+05)	7 (2055)	9 (3.04e+05)	6 (5.17e-06)	1 (3.56e-03)	9 (97.7)	9 (88.5)
IoBUGA	65	7 (3.06e+04)	12 (3.54e+04)	15 (6310)	5 (6.47e+05)	15 (3.80e-05)	3 (8.38e-03)	6 (98.3)	2 (92.8)
RHUL	71	6 (3.20e+04)	13 (3.31e+04)	8 (2551)	15 (1.59e+04)	5 (3.52e-06)	5 (4.77e-02)	4 (98.5)	15 (67.4)
WTSI-P	74	4 (3.80e+04)	11 (4.21e+04)	13 (4895)	13 (3.41e+04)	14 (1.48e-05)	4 (4.38e-02)	2 (98.7)	13 (75.0)
DCSISU	99	12 (1.35e+04)	10 (5.61e+04)	12 (4319)	12 (9.75e+04)	13 (1.37e-05)	13 (6.91e-02)	15 (94.3)	12 (79.0)
nABySS	100	10 (1.99e+04)	16 (2.00e+04)	5 (1731)	16 (6.97e+03)	7 (5.96e-06)	19 (3.17e-01)	10 (97.5)	17 (57.2)
IRISA	103	17 (8.20e+03)	9 (5.82e+04)	11 (3725)	9 (3.04e+05)	17 (3.99e-05)	14 (7.61e-02)	16 (93.7)	10 (88.1)
ASTR	106	8 (2.52e+04)	14 (3.13e+04)	9 (2818)	14 (1.81e+04)	12 (1.28e-05)	18 (2.88e-01)	17 (90.9)	14 (68.5)
nVelv	114	18 (5.65e+03)	15 (2.75e+04)	18 (8626)	11 (1.27e+05)	18 (6.21e-05)	10 (6.22e-02)	13 (96.5)	11 (84.8)
nCLC	115	15 (9.47e+03)	18 (9.54e+03)	16 (7283)	18 (4.36e+03)	10 (1.11e-05)	8 (5.61e-02)	12 (97.2)	18 (55.4)
UCSF	138	12 (1.35e+04)	17 (1.35e+04)	20 (24987)	17 (6.84e+03)	20 (1.21e-04)	17 (2.30e-01)	19 (83.7)	16 (59.6)
GACWT	149	20 (2.53e+03)	19 (7.82e+03)	17 (8622)	19 (2.60e+03)	16 (3.86e-05)	20 (3.46e-01)	18 (86.4)	20 (48.0)
CIUoC	152	19 (5.60e+03)	20 (5.60e+03)	19 (11282)	20 (1.27e+03)	19 (1.11e-04)	16 (1.98e-01)	20 (78.5)	19 (48.9)

Rankings of the top assembly from each team in eight categories. For each category all the received assemblies were ranked. The sum of the rankings from each category was then used to create an overall rank for the assemblies, the top (lowest number) ranked assembly from each group was then selected for inclusion in this manuscript. Numbers are ranks, with values shown in parentheses. Overall: sum of all rankings (possible range 8-160), CPNG50: Contig path NG50, SPNG50: Scaffold path NG50, Struct.: Sum of structural errors, CC50: length for which half of any two valid columns in the assembly are correct in order and orientation, Subs.: Total substitution errors per correct bit, Copy Num.: Proportion of columns with a copy number error, Cov. Tot.: Overall Coverage, Cov. Genic: Coverage within coding sequences.

Discussion

- Basic N50 statistic correlates well with path and contiguity metrics and can be useful comparing assemblies created even different programs
- Contig path NG50 is weakly correlated with scaffold path (R2 = 0.38) and contiguity (R2 = 0.31)
- Some methods are good on certain categories (BGI), but others in various categories (Broad)
- Assemblathon 2:
  - At least one mammalian genome scale data set
  - Should feature real data for evaluating metrics
  - Expand to other sequencing technologies

Long-range contiguity



The proportion of correctly contiguous pairs as a function of their separation distance. Each line represents the top assembly from each team. Correctly contiguous 50 (CC50) values are the lowest point of each line. The legend is ordered top to bottom in descending order of CC50. Proportions were calculated by taking 100 million random samples and binning them into 2,000 bins, equally spaced along a log10 scale, so that an approximately equal number of samples fell in each bin.