Assemblathon 1

A competitive assessment of de novo short read assembly methods

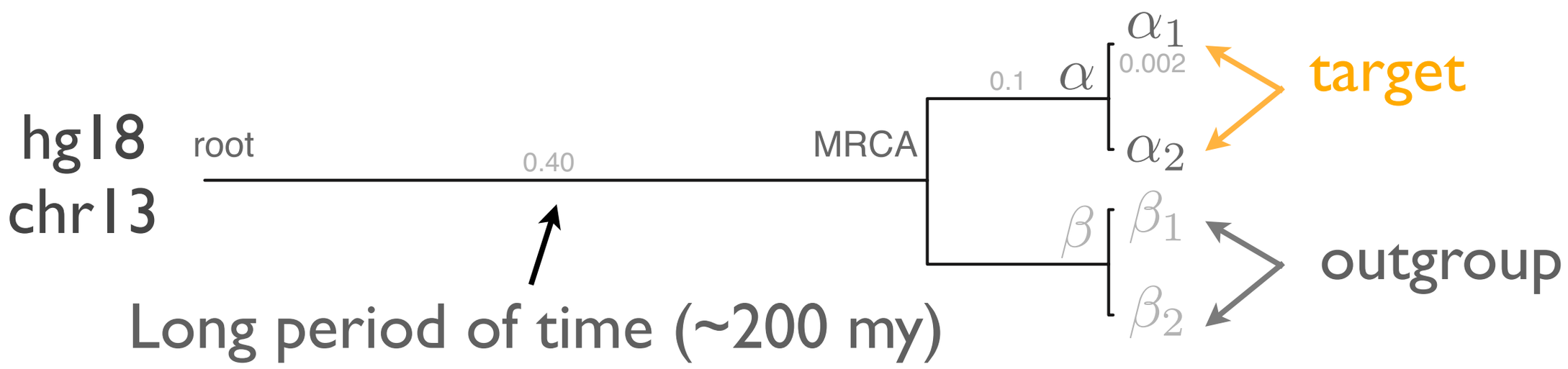
Benedict Paten

Assemblathon 1

- Project to assess **de novo assembly** with **short read** sequencing technology
- Motivated by needs of Genome 10k
- Assemblers invited to compete blind
- Evaluation performed by UC Davis and UCSC, who did not contribute assemblies*

*actually, UCSD contributed a few default parameter assemblies using popular programs

Assemblathon 1



- Dataset: a simulated vertebrate genome, at 1/10th scale
  - Used Evolver - complex simulation tool from Arend Sidow and Robert Edgar
  - Started with hg18 chr13 and "evolved" the above tree
  - Eventual genome had 3 chromosomes, ~120 megabases total length
  - Diploid - 2 simulated haplotypes, with 0.002 subs/site difference
  - Provided outgroup genome to assemblers

Assemblathon 1

- From the simulated genome two Illumina Hi-Seq paired reads types were simulated:
  - "paired ends" = 80X, with 200, 300 bp inserts
  - "mate-pairs" = 40X, 3,000 and 10,000 bp inserts
- Various appropriate errors in the reads were simulated
- ~5% *E. coli* contamination
- Total of 120X for the sample.
  - Removing contamination gives overall 55X per haplotype

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	?
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	Yes
DCISU	Department of Computer Science, Iowa State University	1	PCAP	No
IoBUGA	Computational Systems Biology Laboratory, University of Georgia, USA	3	Seqclean, SOAPdenovo	?
UCSF	UC San Francicso, USA	1	PRICE	?
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	?
BGI	BGI, Shenzhen China	1	SOAPdenovo	No
Broad	Broad Institute	1	ALLPATHS-LG	No
nVelv	—	6	Velvet	No
nCLC	—	12	CLC	No
aABySS	—	6	ABySS	No

MSA

- For each assembly we form a multiple sequence alignment using Cactus* between:
  - the two haplotypes
  - the bacterial contamination
  - the assembly
- To broadly confirm each analysis we used BLAST to align to each haplotype in turn

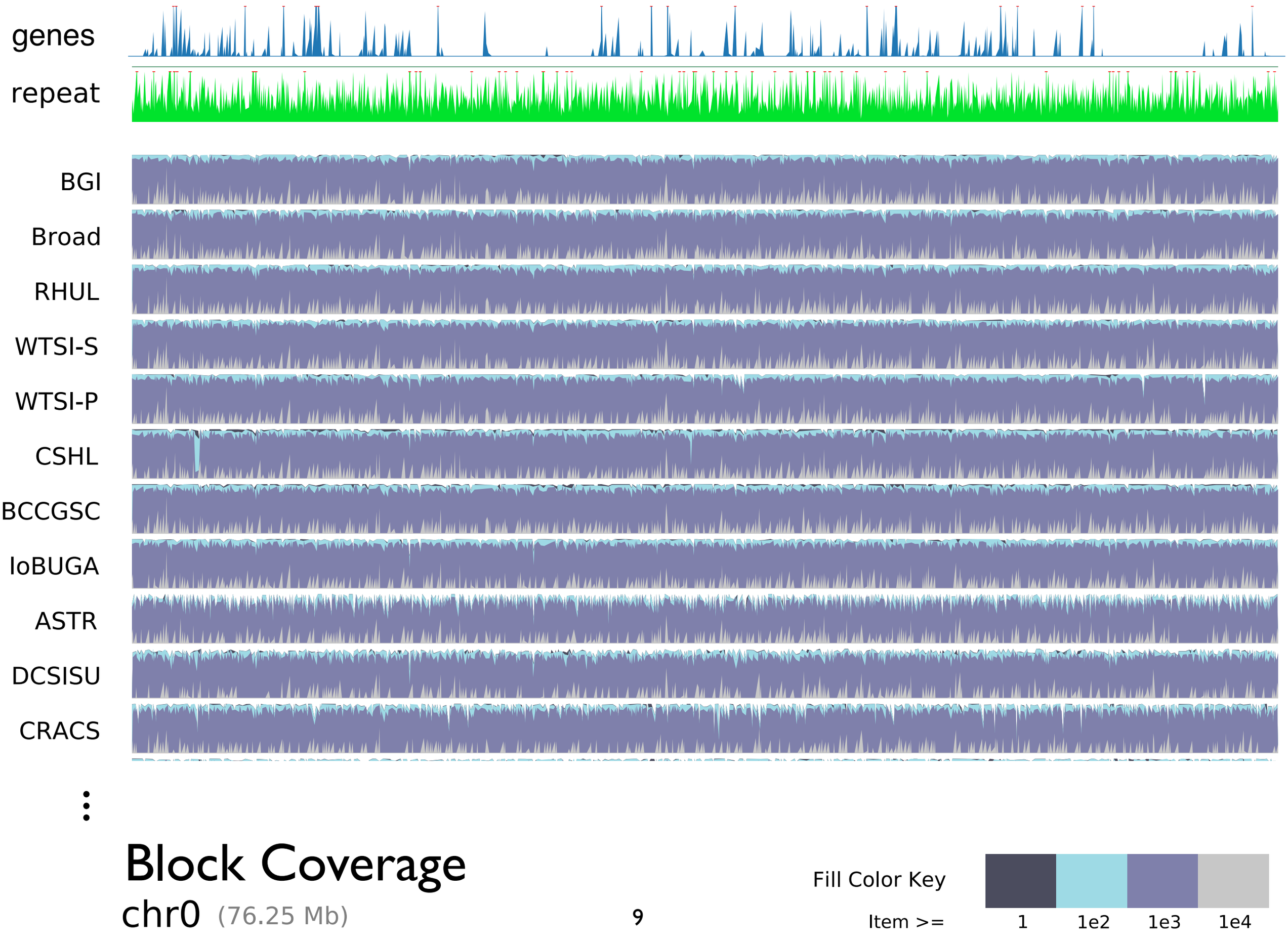
*__Cactus: Algorithms for genome multiple sequence alignment__
Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans,Daniel Zerbino and David Haussler, *Genome Research*, September 2011

Coverage

ID	Hap Total (%)	Hap α_1 (%)	Hap α_2 (%)	Bac (%)	CDS (%)	Unmapped
BGI	98.8	98.9	98.8	0.0	97.8	2.637e+05
BCCGSC	98.7	98.7	98.7	99.9	97.9	6.549e+06
WTSI-P	98.7	98.7	98.7	99.8	97.7	5.369e+06
RHUL	98.5	98.5	98.5	100.0	97.7	4.961e+06
CSHL	98.5	98.6	98.5	99.9	97.8	7.811e+06
Broad	98.3	98.4	98.3	68.9	97.5	3.538e+06
IoBUGA	98.3	98.3	98.3	4.8	97.4	7.821e+05
WTSI-S	97.8	97.8	97.8	99.1	95.2	4.948e+06
EBI	97.7	97.7	97.7	0.9	97.4	4.577e+05
nABySS	97.5	97.5	97.5	99.8	97.7	1.111e+07
DOEJGI	97.3	97.4	97.3	99.5	93.8	5.304e+06
nCLC	97.2	97.2	97.2	99.8	96.2	5.673e+06
nVelv	96.5	96.6	96.5	99.8	97.1	8.028e+06
CRACS	96.3	96.3	96.3	99.8	95.8	5.265e+06
IRISA	95.7	95.6	95.7	99.7	95.2	4.968e+06
DCSISU	94.3	94.3	94.2	99.5	93.6	6.259e+06
ASTR	90.9	90.9	90.9	100.0	92.9	5.176e+06
GACWT	86.4	86.4	86.4	0.0	88.9	2.053e+06
UCSF	83.7	83.7	83.7	0.0	88.3	1.837e+06
CIUoC	78.5	79.0	78.1	0.6	85.4	3.638e+05

Block

- *Blocks* are a maximal gapless alignment of a set of homologous sequences.
  - In this case of the haplotypes and the given assembly
- Due to polymorphism present in the two haplotypes, blocks tend to be short
  - Median block size ~ 4 Kb


genes
repeat
BGI
Broad
RHUL
WTSI-S
WTSI-P
CSHL
BCCGSC
IoBUGA
ASTR
DCSISU
CRACS
Block Coverage
chr0 (76.25 Mb)
9
Fill Color Key
Item >=
1
1e2
1e3
1e4

repeat

RHUL

BCCGSC

DOEJGI

IRISA

nCLC

nABySS

CSHL

CRACS

Broad

WTSI-S

Bacterial Contamination (this region ~2 Mb of 4.75 Mb)

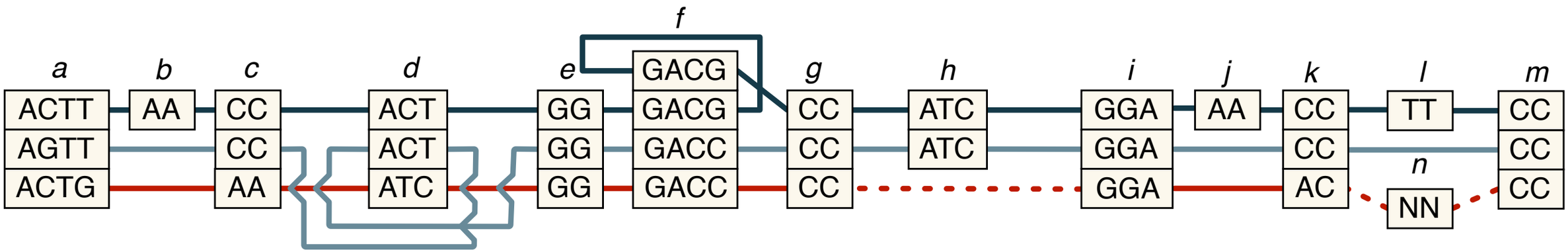
Block Coverage

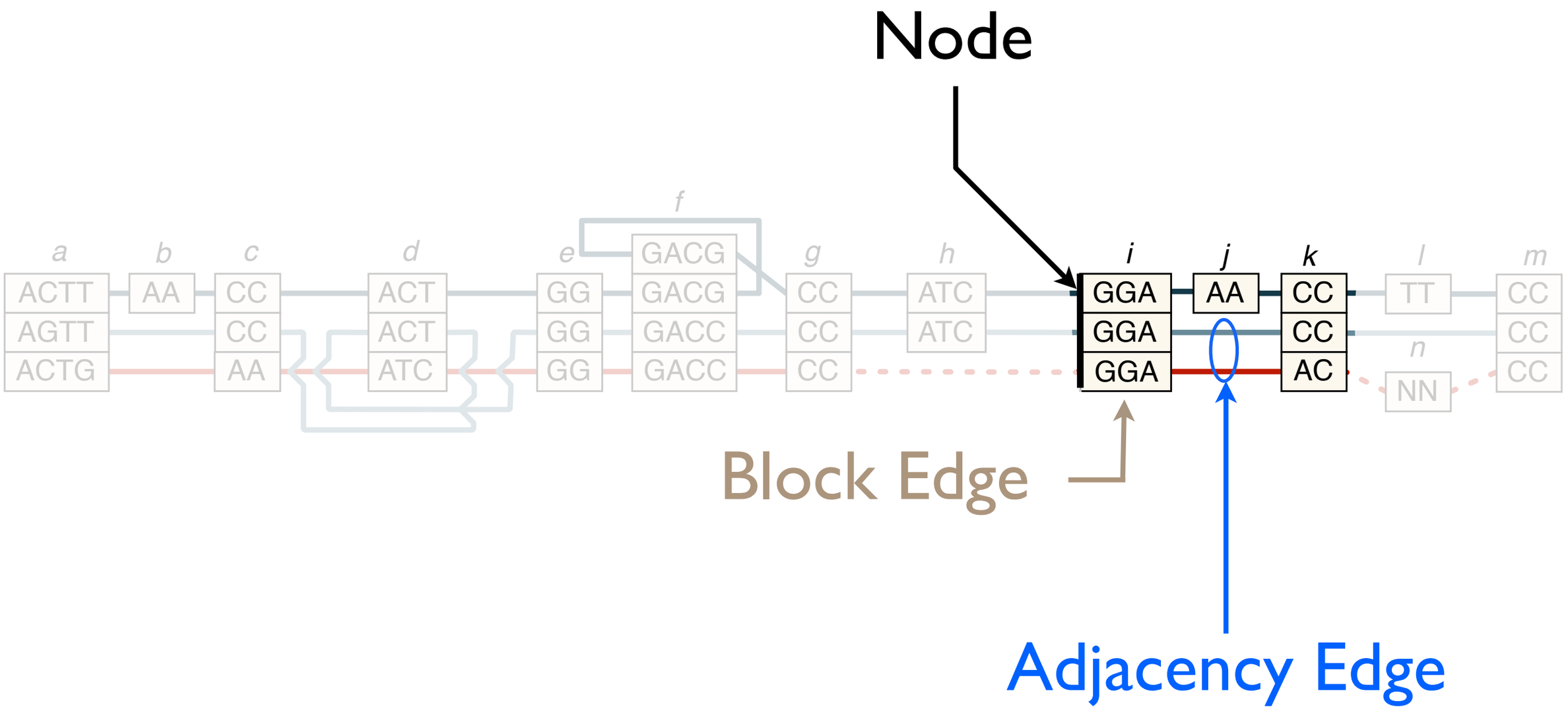
Fill Color Key

I0	Item >=	1	1e2	1e3	1e4	1e5

Adjacency Graph

- *Block Edge* (which you just saw)
- Nodes — ends of blocks edges
- *Adjacency Edge* — collections of connections between ends of blocks, representing connectivity of sequences


a
b
c
d
e
f
g
h
i
j
k
l
m
n
ACTT
AGTT
ACTG
AA
CC
CC
AA
ACT
ACT
ATC
GG
GG
GG
GACG
GACG
GACC
GACC
CC
CC
CC
ATC
ATC
GGA
GGA
GGA
AA
CC
CC
AC
TT
NN
CC
CC
CC


Node
a
b
c
d
e
f
g
h
i
j
k
l
m
n
ACTT
AGTT
ACTG
AA
CC
CC
AA
ACT
ACT
ATC
GG
GG
GG
GACG
GACG
GACC
GACC
CC
CC
CC
ATC
ATC
GGA
GGA
GGA
AA
CC
CC
AC
TT
NN
CC
CC
CC
Block Edge
Adjacency Edge

Adjacency Graph

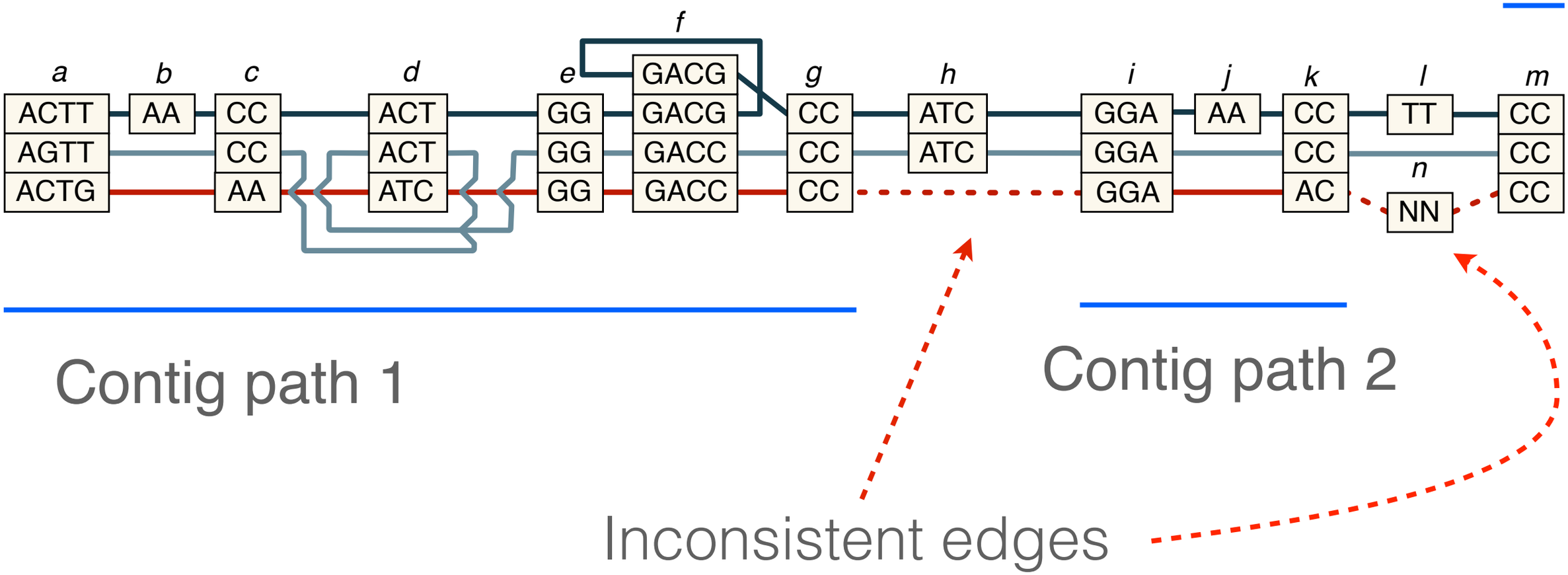
- *Thread* — A path of alternating adjacency edges and block edges
- *Consistent edge* — an edge that is labelled with segments from one or both of the Haplotypes and an assembly sequence
- *Contig Path* — A maximal subthread of an assembly thread in which all edges are consistent

Hap 1

Hap 2

Assembly

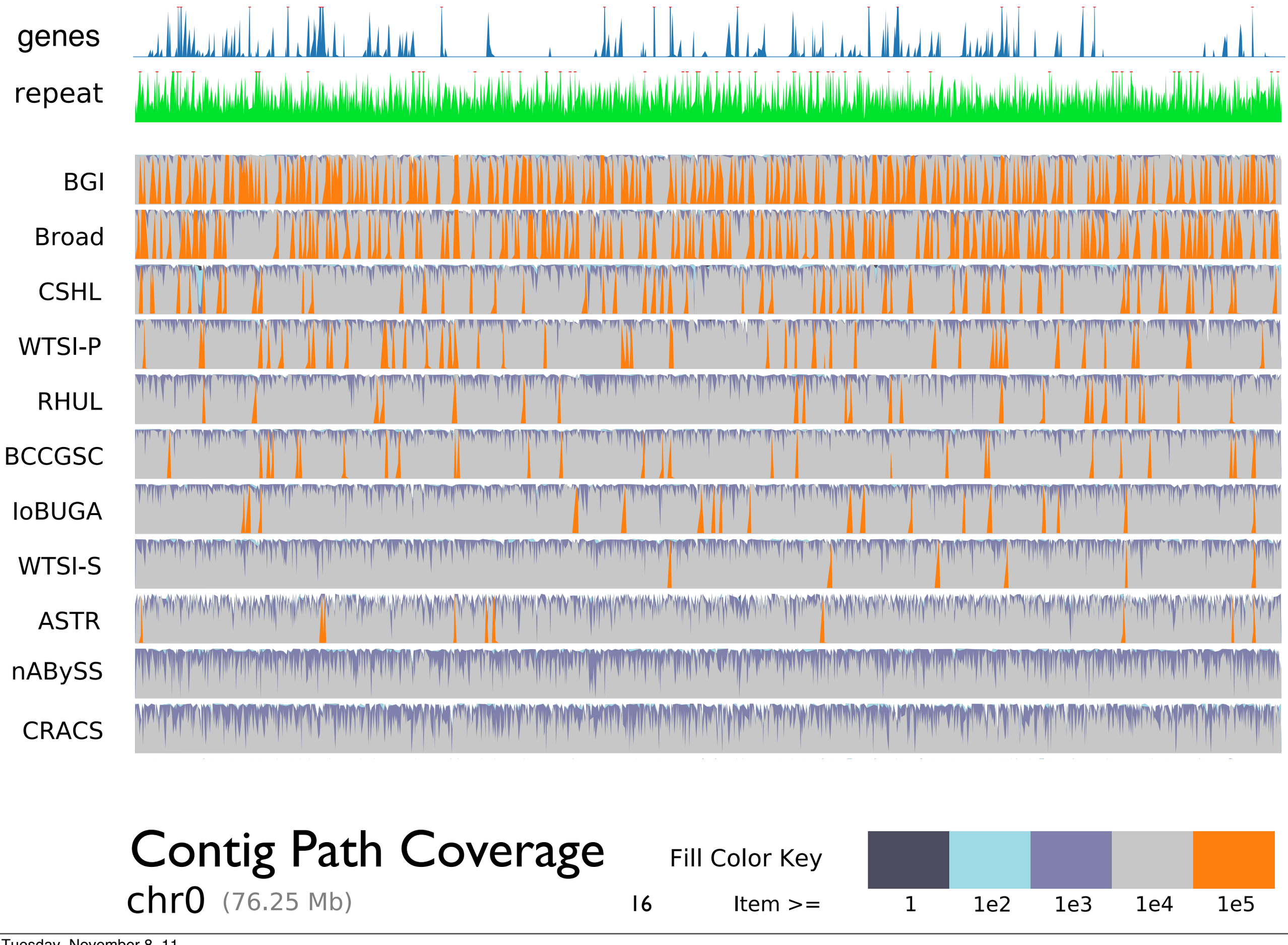
Contig path 3



Contig path 1

Contig path 2

Inconsistent edges


genes
repeat
BGI
Broad
CSHL
WTSI-P
RHUL
BCCGSC
IoBUGA
WTSI-S
ASTR
nABySS
CRACS
Contig Path Coverage
chr0 (76.25 Mb)
Fill Color Key
16
Item >=
1
1e2
1e3
1e4
1e5

Scaffold Path

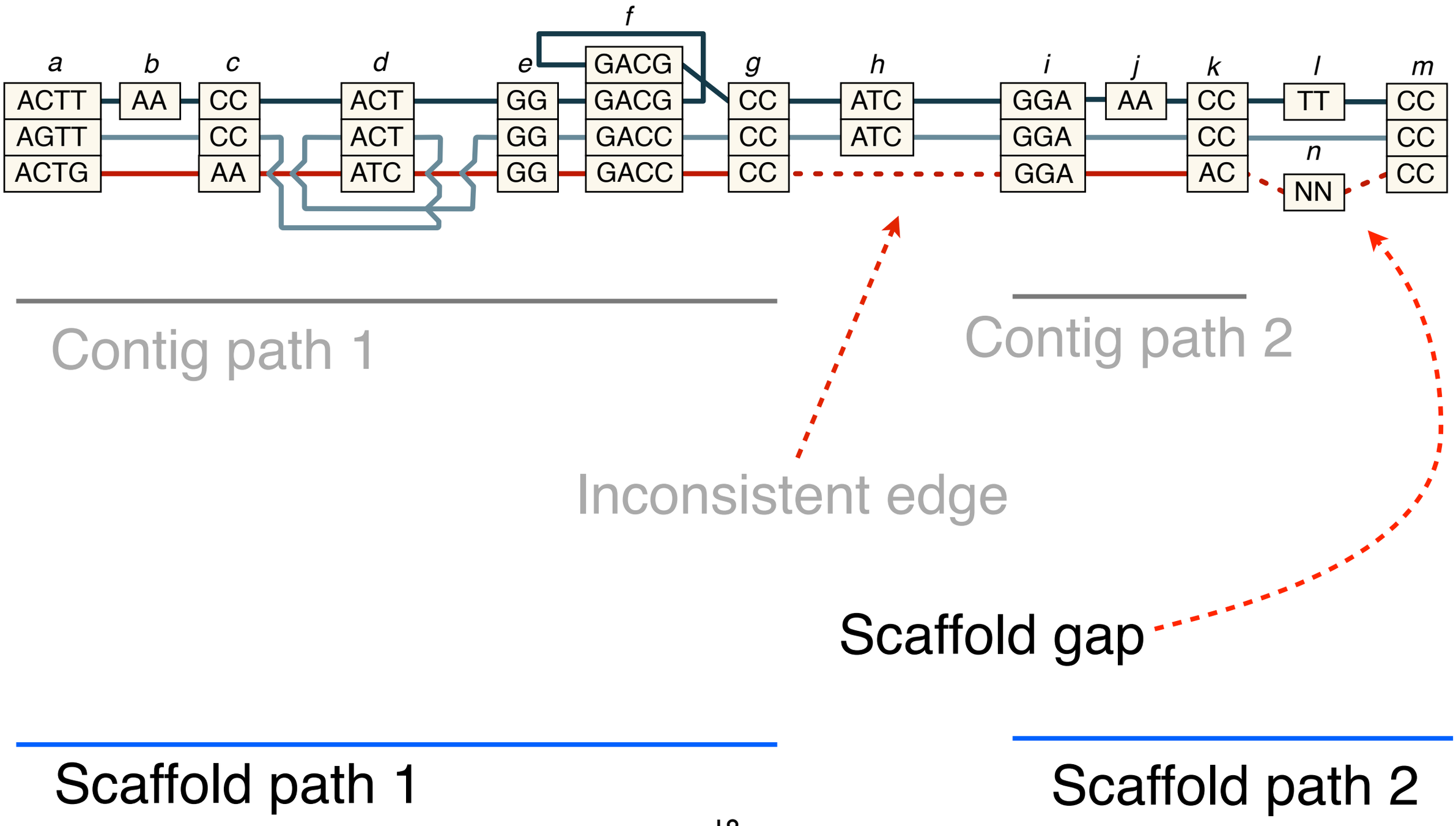
…ACTGACTG NNNNNN ACTGACTG…

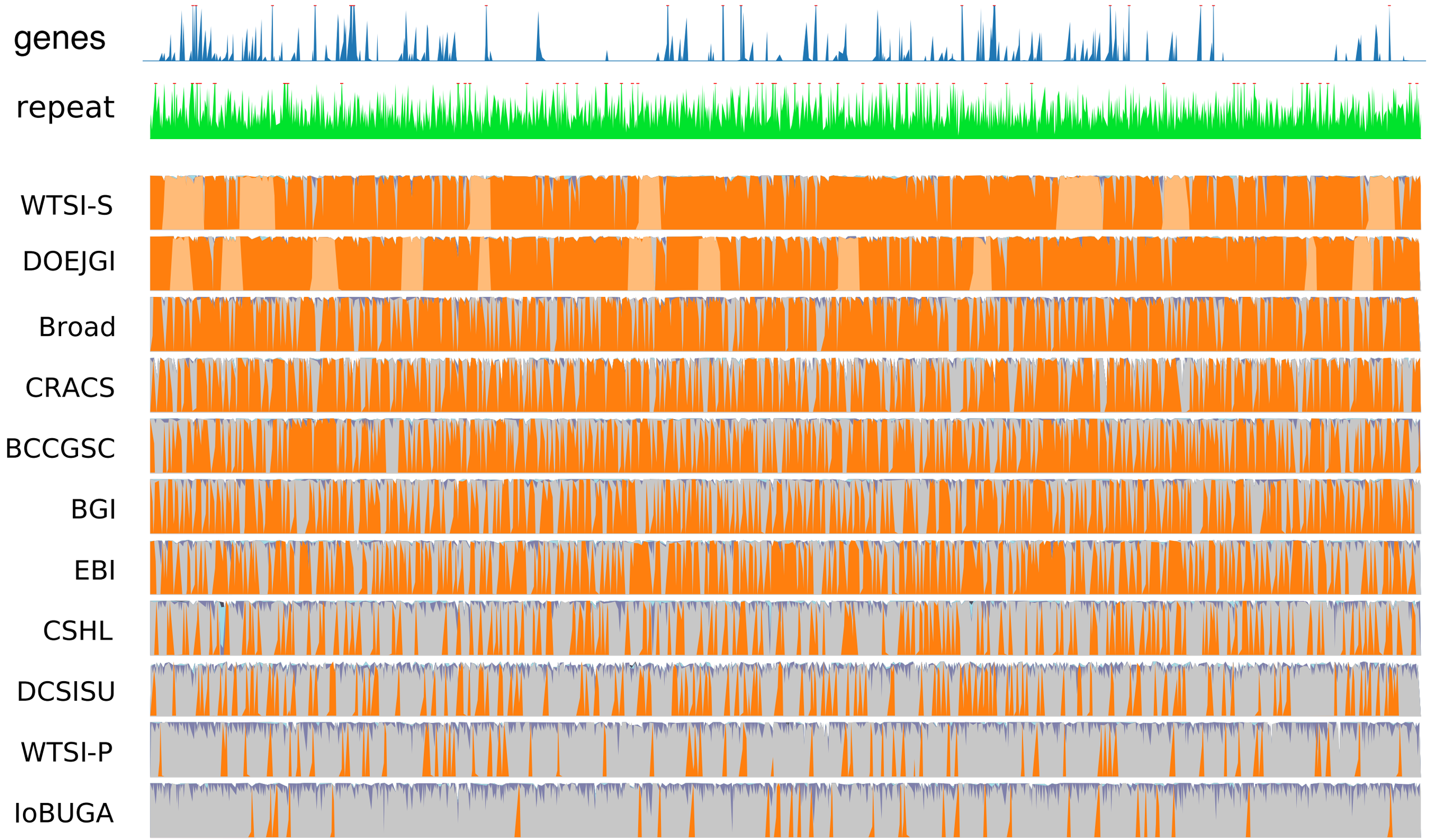
- *Scaffold Gap* — a subgraph representing an indel and containing an assembly segment that is labelled with wildcard characters (N's)
- *Scaffold Path* — a maximal subthread of an assembly thread in which all edges are *consistent* or part of a *scaffold gap* subgraph

Hap 1

Hap 2

Assembly





Scaffold Path Coverage

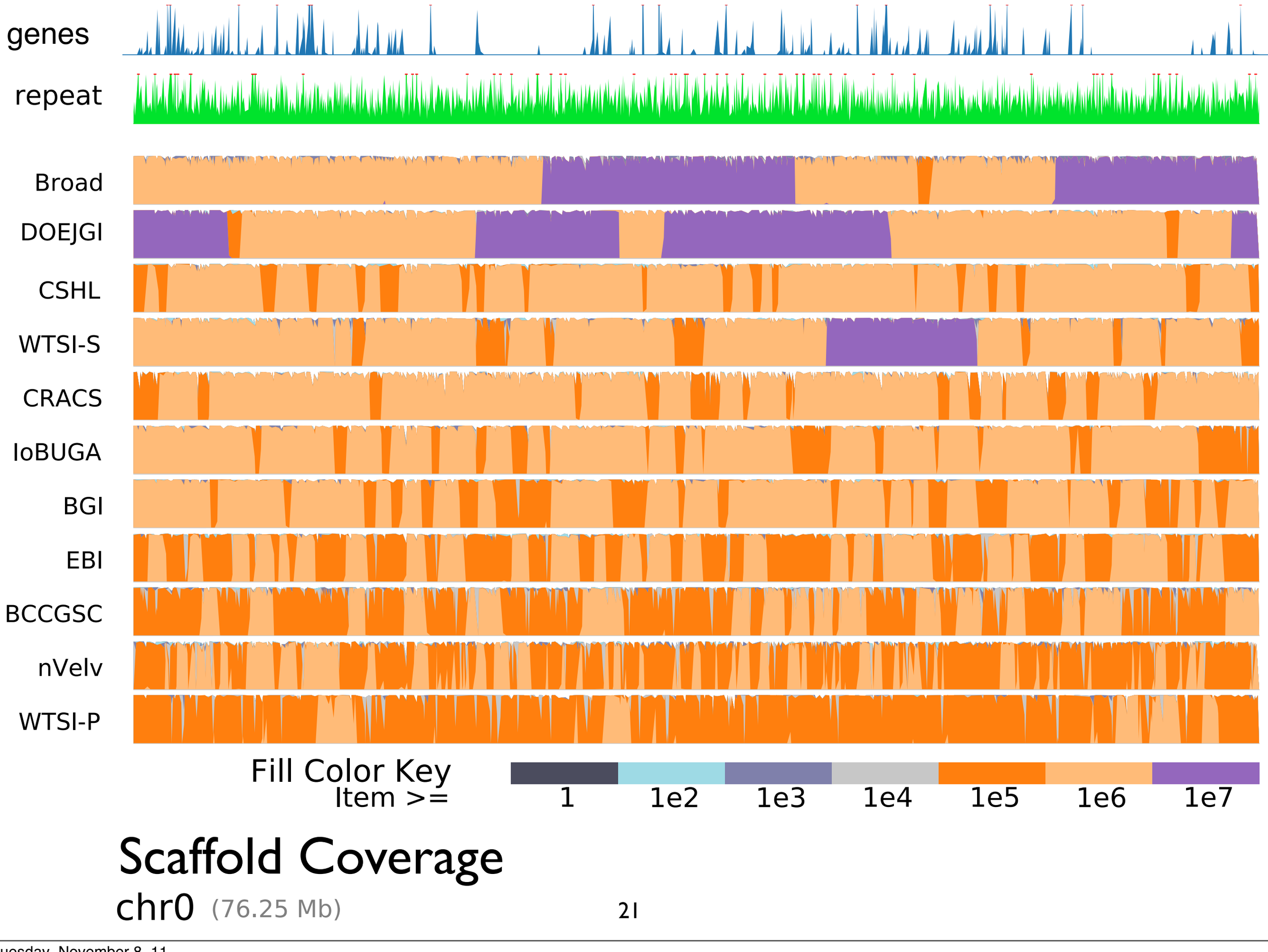
chr0 (76.25 Mb)

Fill Color Key

Item >= 1 1e2 1e3 1e4 1e5 1e6

Raw Scaffolds

- The aspirations of assemblers



Scaffold Coverage

chr0 (76.25 Mb)

Blocks

BGI

Broad

RHUL

Contig Paths

BGI

Broad

CSHL

Scaffold Paths

WTSI-S

DOEJGI

Broad

Scaffolds

Broad

DOEJGI

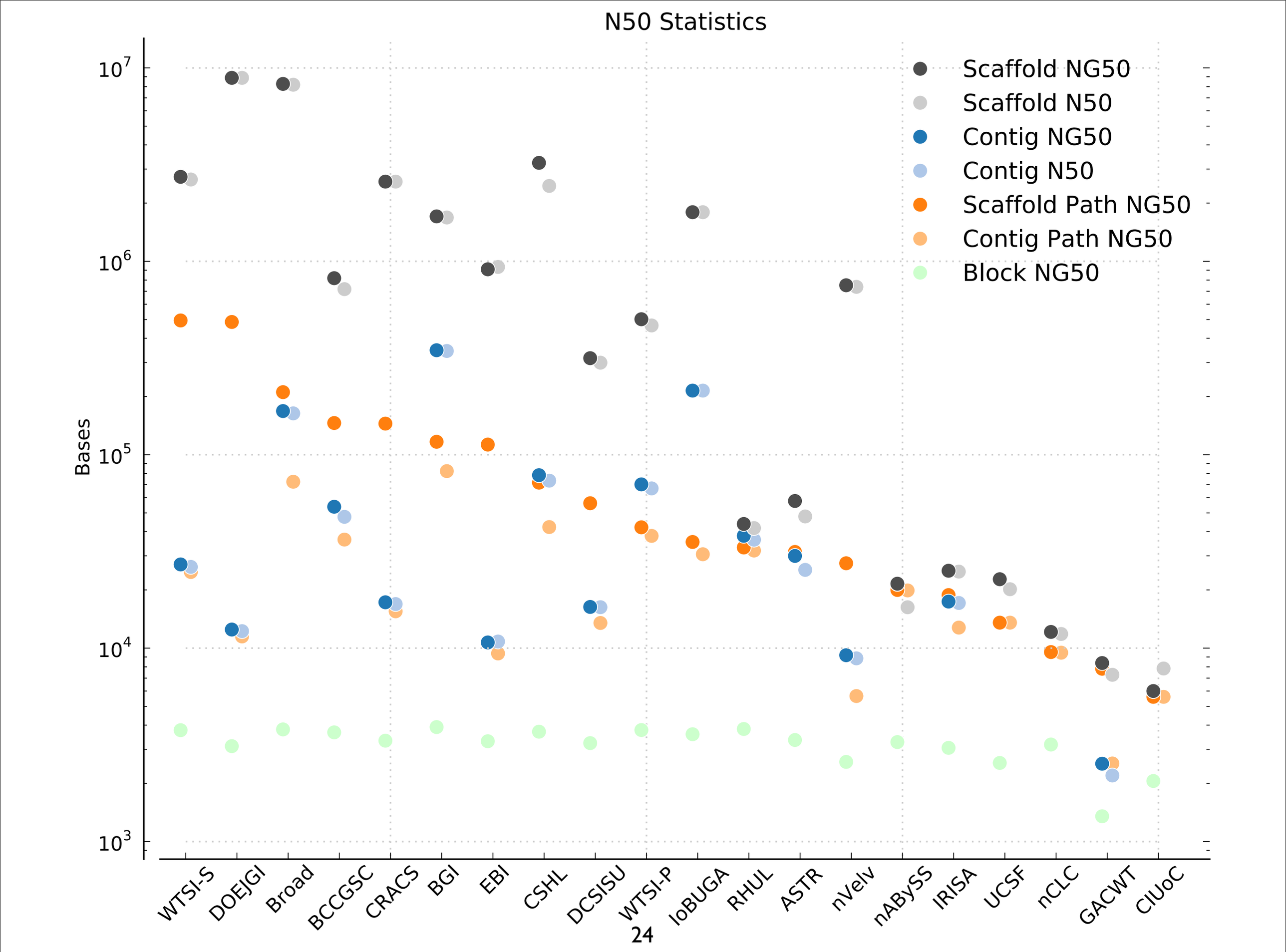
CSHL

Fill Color Key
Item >=

1 1e2 1e3 1e4 1e5 1e6 1e7

We define the NG50 (G for genome) identically to the commonly used N50 for contigs and scaffolds, except that we estimate the length of the genome being assembled as being equal to the average of the two haplotypes.

The Scaffold Path NG50, Contig Path NG50 and Block NG50 values are identical to the NG50s, except that they are computed over the set of scaffold paths, contig paths and blocks, respectively.

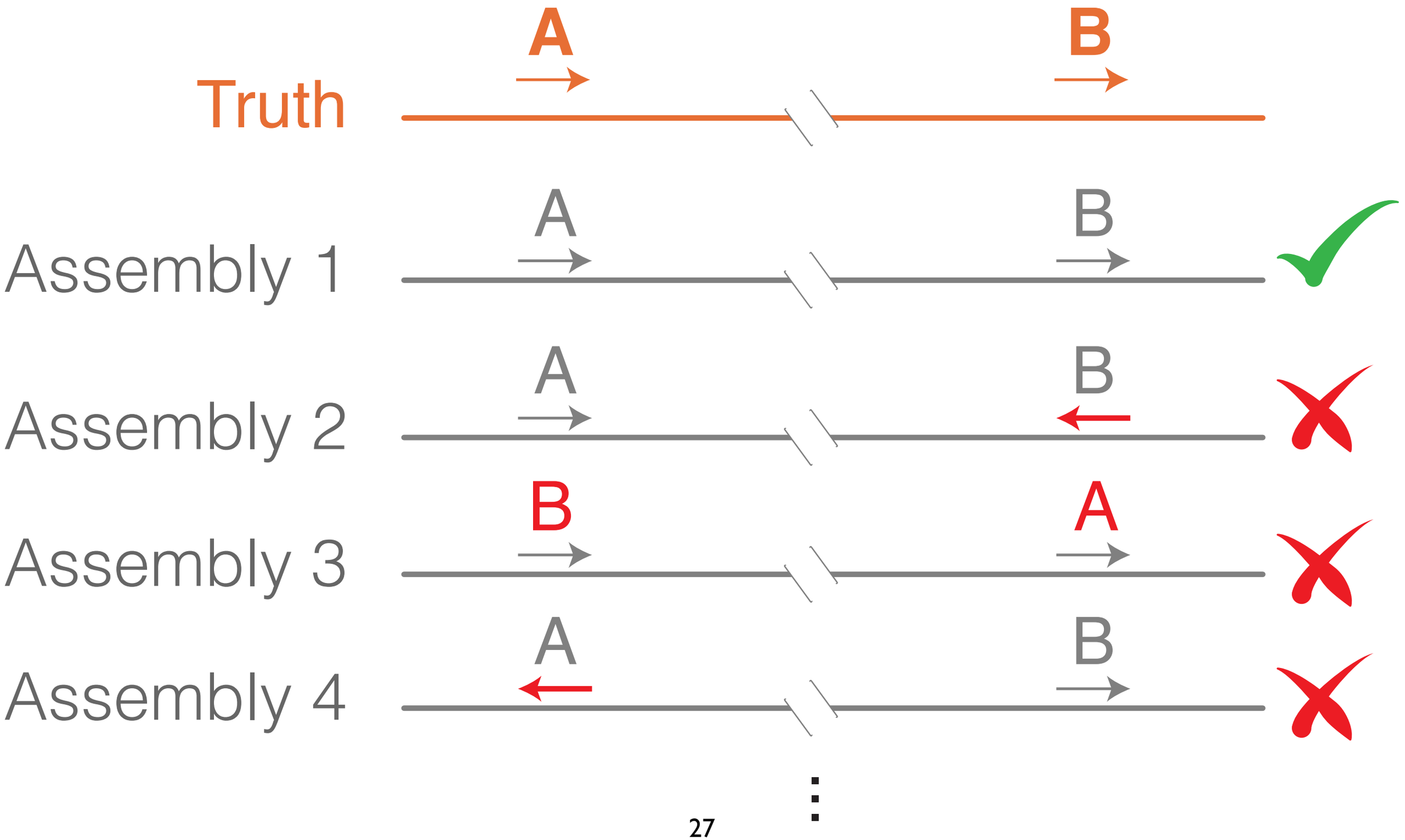

N50 Statistics
Scaffold NG50
Scaffold N50
Contig NG50
Contig N50
Scaffold Path NG50
Contig Path NG50
Block NG50
Bases
10^7
10^6
10^5
10^4
10^3
WTSI-S
DOEJGI
Broad
BCCGSC
CRACS
BGI
EBI
CSHL
DCSISU
WTSI-P
IoBUGA
RHUL
ASTR
nVelv
nABySS
IRISA
UCSF
nCLC
GACWT
CIUoC

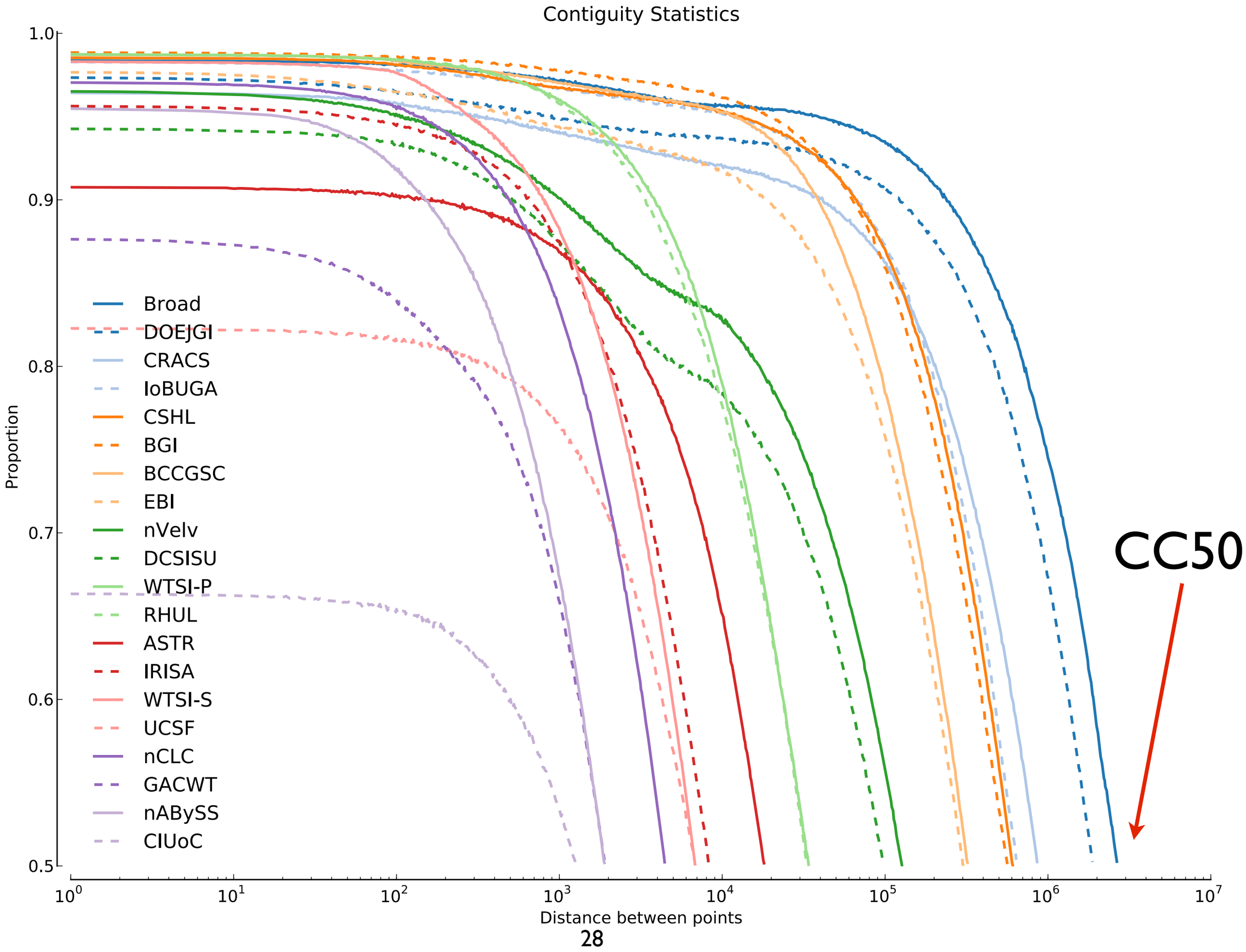
Error Subgraphs

- We also defined "error subgraphs" of the MSAs, including:
  - insertions
  - deletions
  - simulataneous insertions and deletions
  - inter- and intra-chromosomal non-linear rearrangements

ID	Intra chromosomal joins	Inter chromosomal joins	Insertions	Deletions	Insertion and deletion	Insertion at ends	$\sum$ errors
DOEJGI	21	160	55	108	40	72	456
WTSI-S	6	191	56	76	19	127	475
Broad	75	161	524	379	9	96	1,244
CRACS	684	309	198	123	50	305	1,669
nABySS	17	48	208	188	63	1,207	1,731
BGI	368	288	355	639	98	130	1,878
EBI	459	567	126	547	55	317	2,071
RHUL	691	349	172	264	26	1,050	2,552
ASTR	2,062	198	106	225	71	141	2,803
BCCGSC	349	289	248	229	107	1,640	2,862
IRISA	67	171	521	993	44	2,061	3,857
DCSISU	1,411	955	330	953	108	560	4,317
WTSI-P	1,940	449	1,851	289	87	279	4,895
CSHL	395	338	413	3,285	219	491	5,141
IoBUGA	919	330	1,663	2,933	356	108	6,309
nCLC	23	64	2,359	2,237	68	2,532	7,283
GACWT	757	730	905	1,292	216	4,722	8,622
nVelv	2,885	455	1,473	2,838	306	669	8,626
CIUoC	1,205	684	1,189	2,026	65	6,113	11,282
UCSF	2,725	2,396	5,825	6,156	988	6,722	24,812

Long Range Contiguity




Contiguity Statistics
1.0
0.9
0.8
0.7
0.6
0.5
Proportion
Broad
DOEJGI
CRACS
IoBUGA
CSHL
BGI
BCCGSC
EBI
nVelv
DCSISU
WTSI-P
RHUL
ASTR
IRISA
WTSI-S
UCSF
nCLC
GACWT
nABySS
CIUoC
CC50
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Distance between points

Etc.

- Additionally we analysed:
  - Copy number variation
  - Base calling
  - Gene and repeat subregion assembly
  - Evidence for phasing by the assemblies
  - ...

Rankings

ID	Overall	CPNG50	SPNG50	Struct.	CC50	Subs.	Copy Num.	Cov. Tot.	Cov. CDS
BGI	34	1 (8.23e+04)	6 (1.17e+05)	6 (1878)	7 (5.66e+05)	9 (1.20e-05)	2 (6.75e-03)	1 (98.8)	2 (97.8)
Broad	37	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)	7 (97.5)
WTSI-S	46	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)	13 (95.2)
CSHL	50	3 (4.23e+04)	8 (7.17e+04)	14 (5141)	6 (6.11e+05)	7 (1.04e-05)	6 (4.94e-02)	4 (98.5)	2 (97.8)
BCCGSC	56	5 (3.64e+04)	4 (1.46e+05)	10 (2862)	8 (3.22e+05)	11 (1.32e-05)	15 (1.17e-01)	2 (98.7)	1 (97.9)
DOEJGI	56	15 (1.15e+04)	2 (4.86e+05)	1 (456)	2 (1.89e+06)	3 (4.43e-07)	7 (5.42e-02)	11 (97.3)	15 (93.8)
RHUL	58	6 (3.20e+04)	12 (3.31e+04)	8 (2552)	14 (1.59e+04)	5 (3.52e-06)	5 (4.77e-02)	4 (98.5)	4 (97.7)
WTSI-P	63	4 (3.80e+04)	10 (4.21e+04)	13 (4895)	12 (3.41e+04)	14 (1.48e-05)	4 (4.38e-02)	2 (98.7)	4 (97.7)
EBI	64	17 (9.39e+03)	7 (1.13e+05)	7 (2071)	9 (3.04e+05)	6 (5.20e-06)	1 (3.59e-03)	9 (97.7)	8 (97.4)
CRACS	64	11 (1.55e+04)	5 (1.45e+05)	4 (1669)	4 (8.61e+05)	2 (3.81e-07)	12 (6.82e-02)	14 (96.3)	12 (95.8)
IoBUGA	71	7 (3.06e+04)	11 (3.54e+04)	15 (6309)	5 (6.47e+05)	16 (3.80e-05)	3 (8.38e-03)	6 (98.3)	8 (97.4)
nABySS	94	10 (1.99e+04)	15 (2.00e+04)	5 (1731)	16 (6.97e+03)	15 (1.81e-05)	19 (3.17e-01)	10 (97.5)	4 (97.7)
DCSISU	101	12 (1.35e+04)	9 (5.61e+04)	12 (4317)	11 (9.84e+04)	12 (1.37e-05)	13 (6.91e-02)	16 (94.3)	16 (93.6)
ASTR	105	8 (2.53e+04)	13 (3.14e+04)	9 (2803)	13 (1.81e+04)	10 (1.28e-05)	18 (2.88e-01)	17 (90.9)	17 (92.9)
nCLC	107	16 (9.47e+03)	18 (9.54e+03)	16 (7283)	18 (4.36e+03)	8 (1.11e-05)	8 (5.61e-02)	12 (97.2)	11 (96.2)
IRISA	111	14 (1.28e+04)	16 (1.88e+04)	11 (3857)	15 (8.28e+03)	13 (1.41e-05)	14 (7.26e-02)	15 (95.7)	13 (95.2)
nVelv	111	18 (5.65e+03)	14 (2.75e+04)	18 (8626)	10 (1.27e+05)	18 (6.21e-05)	10 (6.22e-02)	13 (96.5)	10 (97.1)
UCSF	141	12 (1.35e+04)	17 (1.35e+04)	20 (24812)	17 (6.78e+03)	20 (1.21e-04)	17 (2.30e-01)	19 (83.7)	19 (88.3)
GACWT	148	20 (2.53e+03)	19 (7.82e+03)	17 (8622)	19 (2.60e+03)	17 (3.86e-05)	20 (3.46e-01)	18 (86.4)	18 (88.9)
CIUoC	153	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)	20 (85.4)

Conclusions

- We demonstrated that the best teams were able to assemble:
  - ~100 Kb regions without error or gaps (contig path analysis)
  - 1 Mb regions without error, but with gaps (scaffold path analysis)
- Huge differences between assemblies
- Some metrics correlated, but every assembler had areas of weakness
- Path N50s and simple N50s correlate i.e. in this case, you could usefully, though imperfectly, compare N50 values.

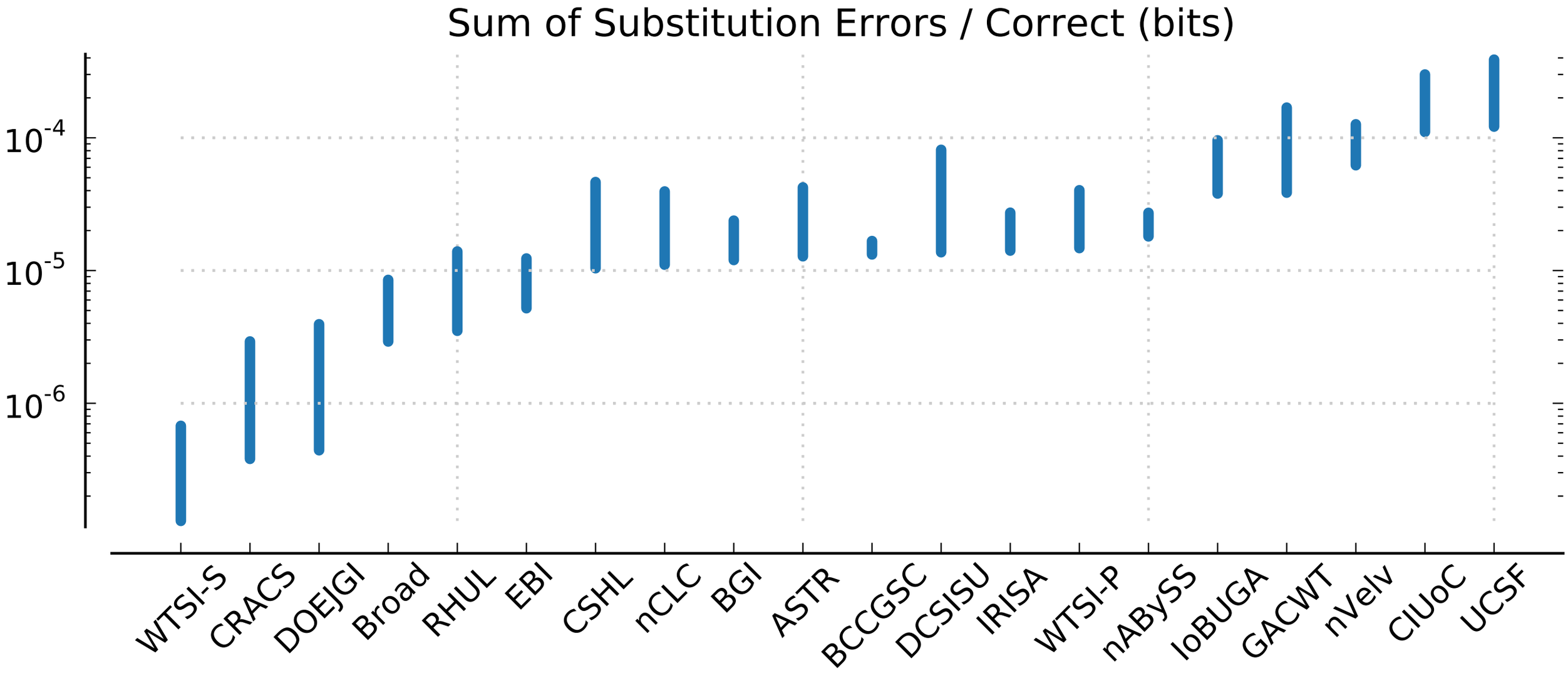
Future Work

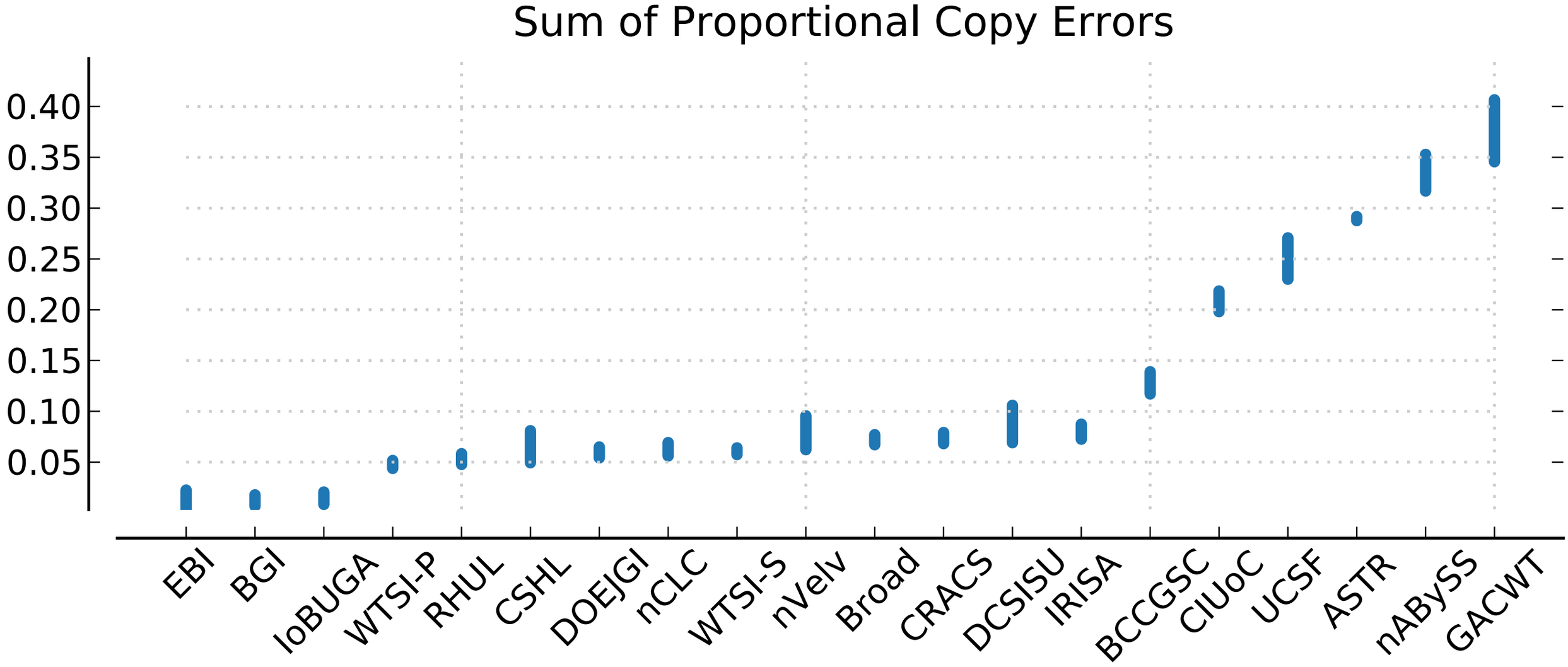
- Community now hard at work on Assemblathon 2:
  - Is using three biological datasets
  - Explores different read technologies
  - Is at scale
  - Meeting on Assemblathon 2 on Saturday afternoon

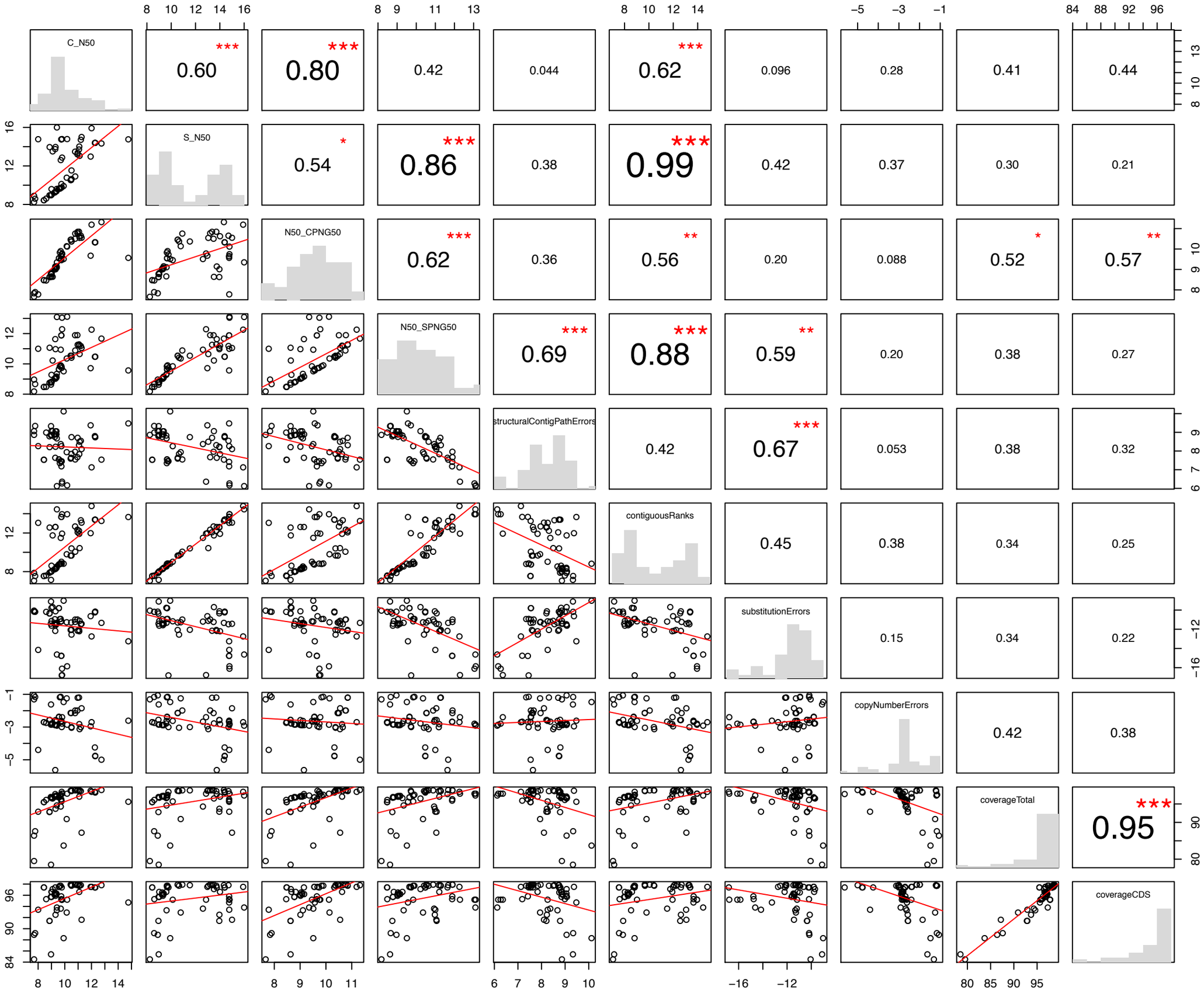
Thank you!

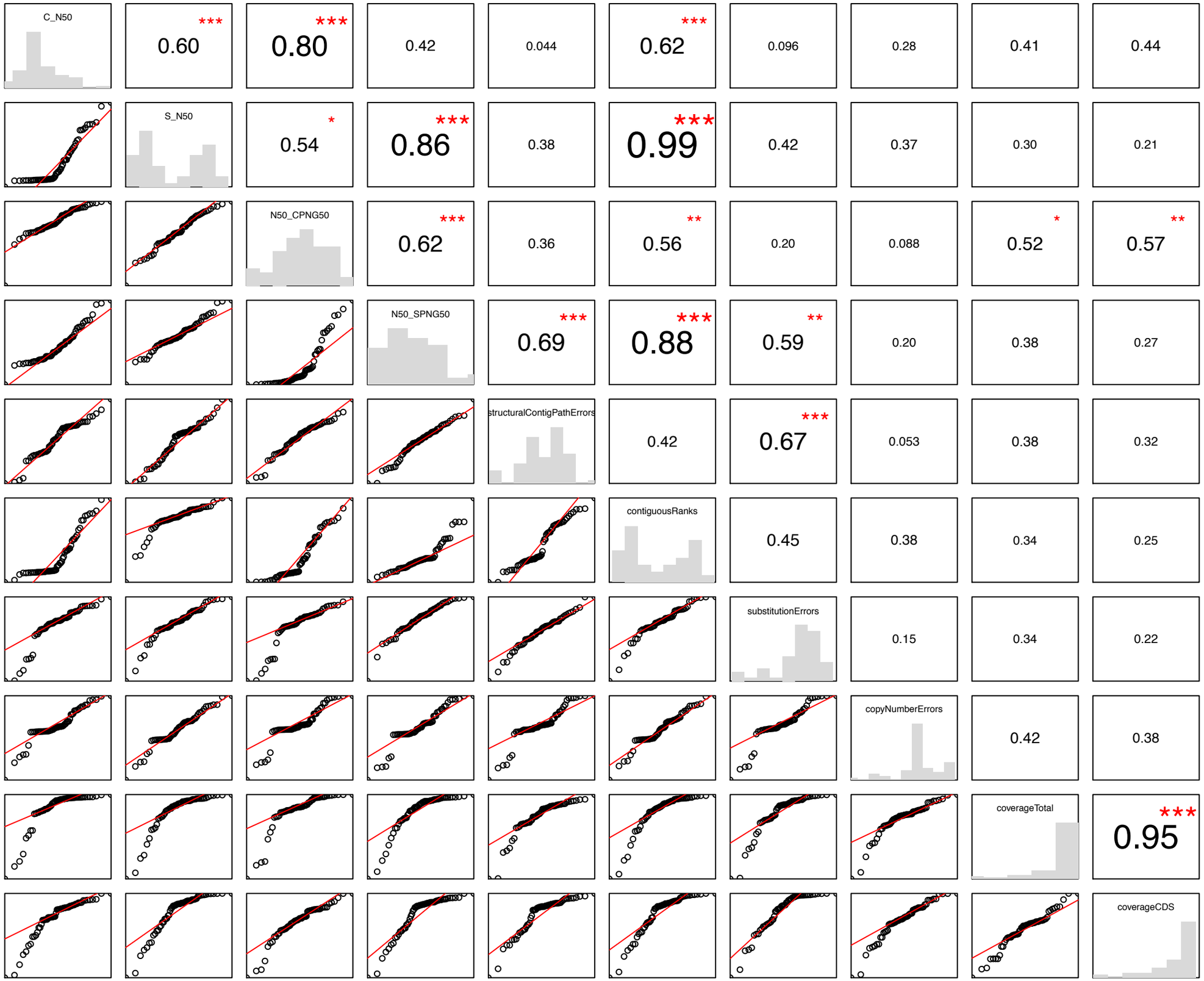
Dent Earl, Keith Bradnam, John St. John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R. Zerbino,Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne,Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T. Simpson,Nuno A. Fonseca, İnanç Birol, T. Roderick Docking, Isaac Y. Ho,Daniel S. Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis,Delphine Naquin, Nicolas Maillet, Michael C. Schatz, David R. Kelley,Adam M. Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu,Wen-Chi Chou, Anuj Srivastava, Timothy I. Shaw, J. Graham Ruby,Peter Skewes-Cox, Miguel Betegon, Michelle T. Dimon, Victor Solovyev,Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez,Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li,Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum,Dariusz Przybylski, Filipe J. Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall,Paul J. Kersey, Richard Durbin, Shaun D. Jackman, Jarrod A. Chapman,Xiaoqiu Huang, Joseph L. DeRisi, Mario Caccamo, Yingrui Li, David B. Jaffe,Richard E. Green, David Haussler, Ian Korf

Article at: (http://genome.cshlp.org/content/early/2011/11/02/gr.126599.111.abstract)


Sum of Substitution Errors / Correct (bits)
10^{-4}
10^{-5}
10^{-6}
WTSI-S
CRACS
DOEJGI
Broad
RHUL
EBI
CSHL
nCLC
BGI
ASTR
BCCGSC
DCSISU
IRISA
WTSI-P
nABySS
IoBUGA
GACWT
nVelv
ClUoC
UCSF


Sum of Proportional Copy Errors
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
EBI
BGI
IoBUGA
WTSI-P
RHUL
CSHL
DOEJGI
nCLC
WTSI-S
nVelv
Broad
CRACS
DCSISU
IRISA
BCCGSC
CIUoC
UCSF
ASTR
nABySS
GACWT


C_N50
S_N50
N50_CPNG50
N50_SPNG50
structuralContigPathErrors
contiguousRanks
substitutionErrors
copyNumberErrors
coverageTotal
coverageCDS
0.60 ***
0.80 ***
0.42
0.044
0.62 ***
0.096
0.28
0.41
0.44
0.54 *
0.86 ***
0.38
0.99 ***
0.42
0.37
0.30
0.21
0.62 ***
0.36
0.56 **
0.20
0.088
0.52 *
0.57 **
0.69 ***
0.88 ***
0.59 **
0.20
0.38
0.27
0.42
0.67 ***
0.053
0.38
0.32
0.45
0.38
0.34
0.25
0.15
0.34
0.22
0.42
0.38
0.95 ***


C_N50
0.60 ***
0.80 ***
0.42
0.044
0.62 ***
0.096
0.28
0.41
0.44
S_N50
0.54 *
0.86 ***
0.38
0.99 ***
0.42
0.37
0.30
0.21
N50_CPNG50
0.62 ***
0.36
0.56 **
0.20
0.088
0.52 *
0.57 **
N50_SPNG50
0.69 ***
0.88 ***
0.59 **
0.20
0.38
0.27
structuralContigPathErrors
0.42
0.67 ***
0.053
0.38
0.32
contiguousRanks
0.45
0.38
0.34
0.25
substitutionErrors
0.15
0.34
0.22
copyNumberErrors
0.42
0.38
coverageTotal
0.95 ***
coverageCDS