Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Three vertebrate species

Multiple international teams

Many different assemblers

To answer:

- which is the best assembler?
- How can you predict what makes a good assembly?

The Assemblathon 2 Cast

Budgeriagar
Melopsitaccus undulatus
1.2 Gbp genome

<u>Data:</u>
285X Illumina p.e. & m.p
16X 454
10X PacBio

Lake Malawi cichlid
Maylandia zebra
1.0 Gbp genome

<u>Data:</u>
192X Illumina p.e. & m.p.

Boa constrictor
Boa constrictor
1.6 Gbp genome

<u>Data:</u>
125X Illumina p.e. & m.p.

Table 1 Assemblathon 2 participating team details

Team name	Team identifier	Number of assemblies submitted			Sequence data used for bird assembly	Institutional affiliations	Principal assembly software used
		Bird	Fish	Snake			
ABL	ABL	1	0	0	4 + I	Wayne State University	HyDA
ABySS	ABYSS	0	1	1		Genome Sciences Centre, British Columbia Cancer Agency	ABySS and Anchor
Allpaths	ALLP	1	1	0	I	Broad Institute	ALLPATHS-LG
BCM-HGSC	BCM	2	1	1	4 + I + P[1]	Baylor College of Medicine Human Genome Sequencing Center	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap, CrossMatch, Velvet, BLAST, and BLASR
CBCB	CBCB	1	0	0	4 + I + P	University of Maryland, National Biodefense Analysis and Countermeasures Center	Celera assembler and PacBio Corrected Reads (PBcR)
CoBiG[2]	COBIG	1	0	0	4	University of Lisbon	4Pipe4 pipeline, Seqclean, Mira, Bambus2
CRACS	CRACS	0	0	1		Institute for Systems and Computer Engineering of Porto TEC, European Bioinformatics Institute	ABySS, SSPACE, Bowtie, and FASTX
CSHL	CSHL	0	3	0		Cold Spring Harbor Laboratory, Yale University, University of Notre Dame	Metassembler, ALLPATHS, SOAPdenovo
CTD	CTD	0	3	0		National Research University of Information Technologies, Mechanics, and Optics	Unspecified
Curtain	CURT	0	0	1		European Bioinformatics Institute	SOAPdenovo, fastx_toolkit, bwa, samtools, velvet, and curtain

What they did with the assemblies

Analyses

Contiguity

Presence of Core Genes (CEGMA)

Alignment of fosmids to scaffolds

- COMPASS analysis of validated fosmid regions (VFRs)
- Short range accuracy in VFRs

Optical map analysis

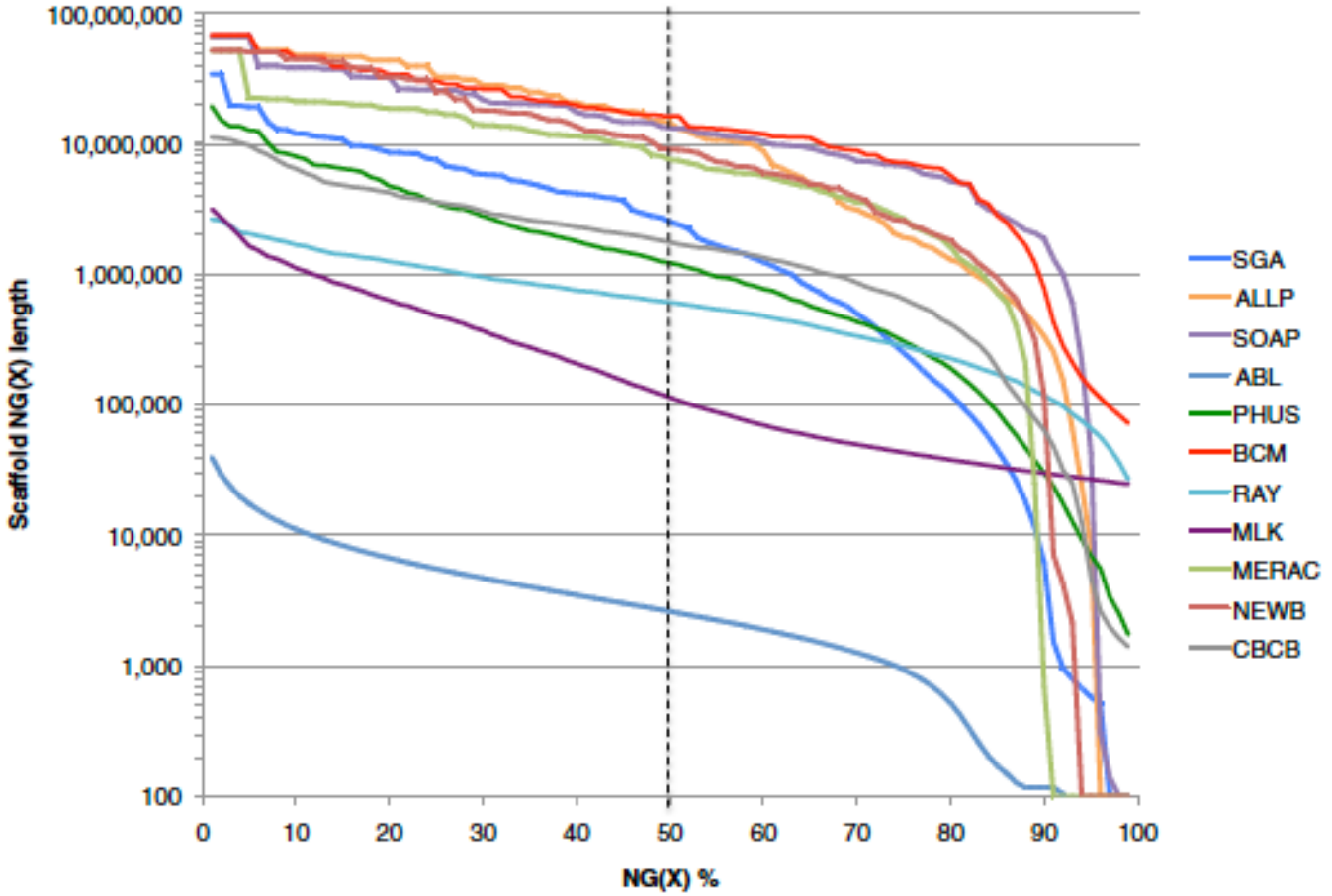
REAPR analysis

These metrics were incorporated

Ranking of key metrics for each assembly using z-scores

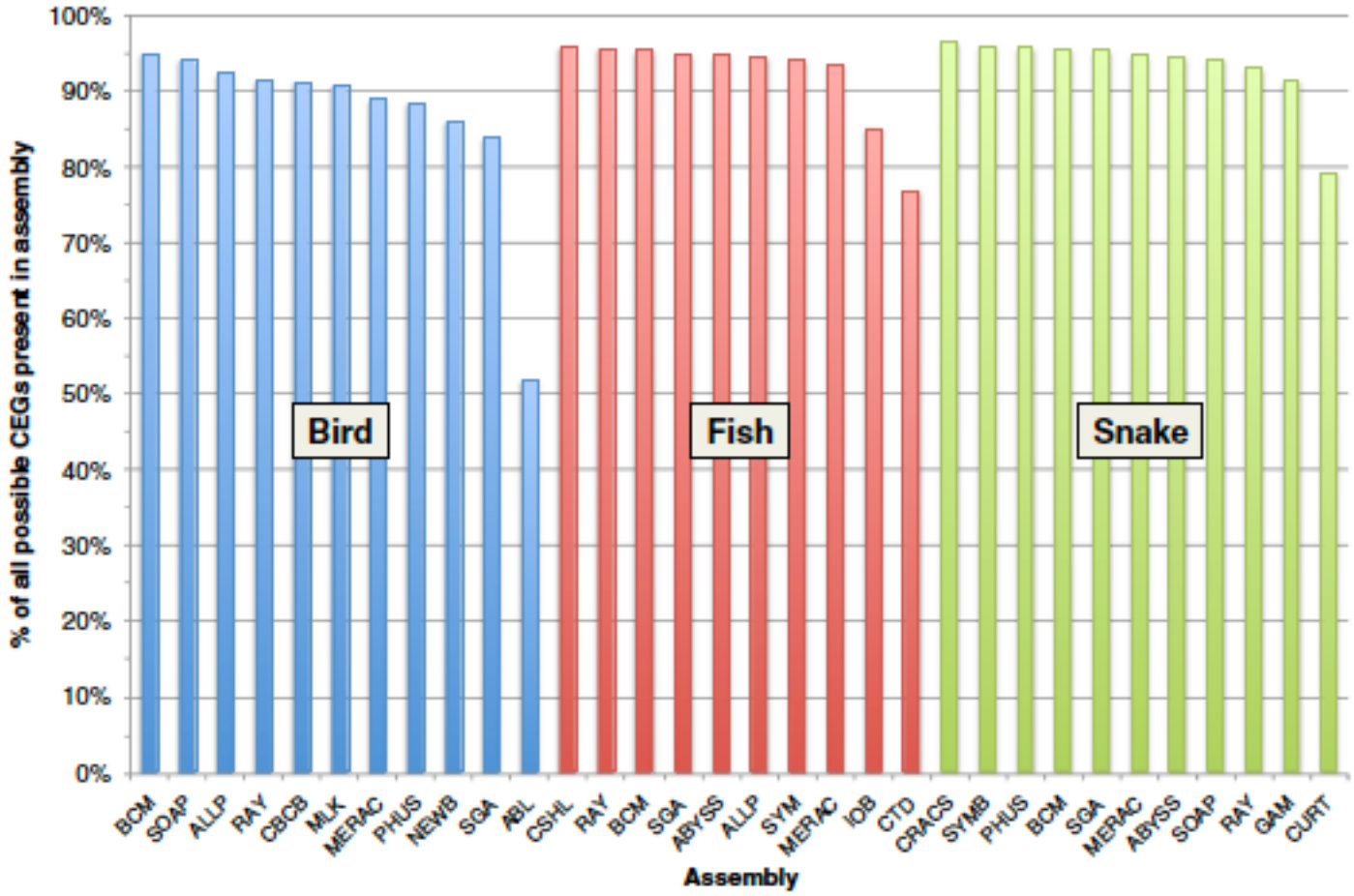
Analysis of metrics using linear regressions

Scaffold NGX

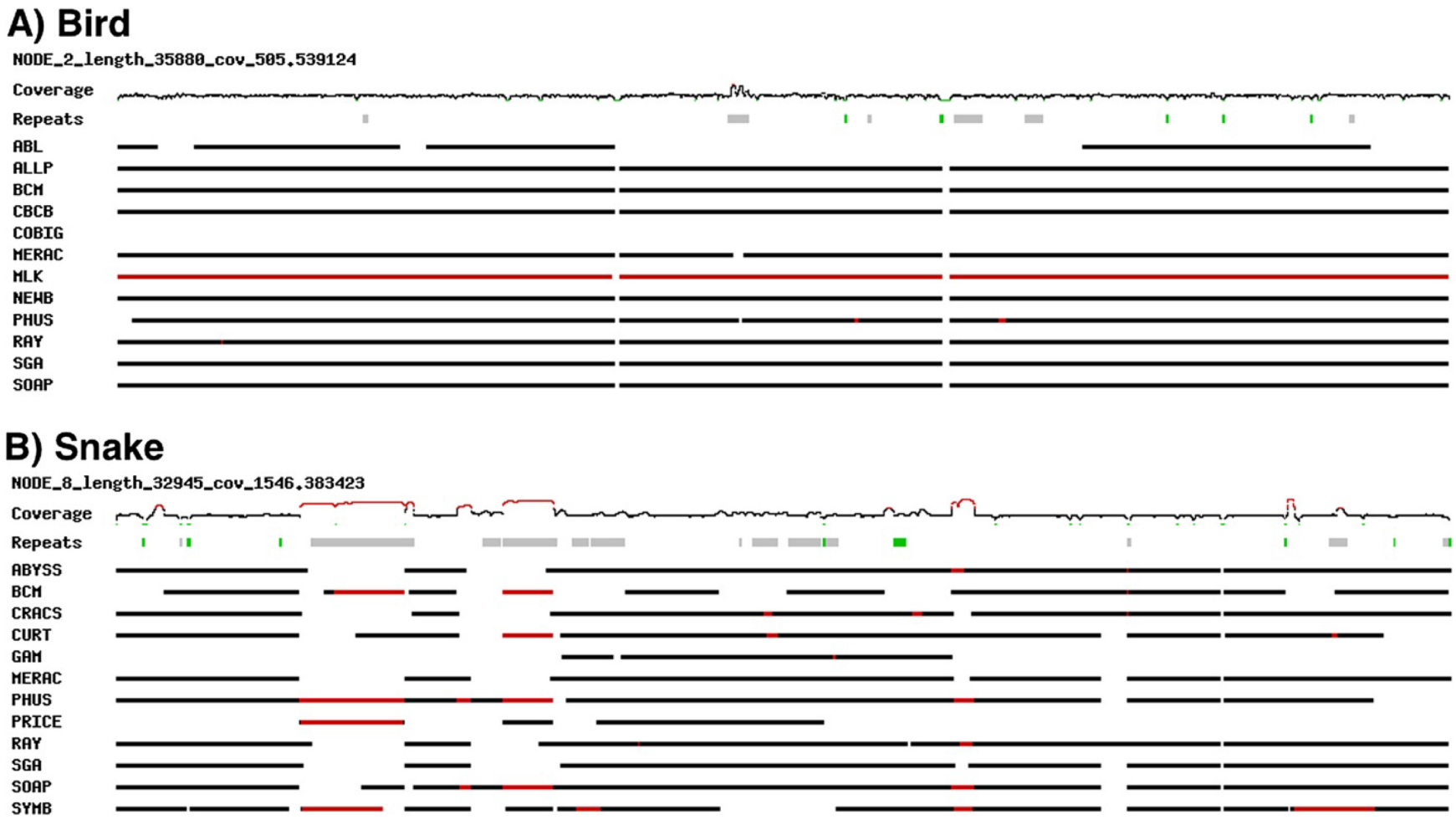


NG50 varied widely

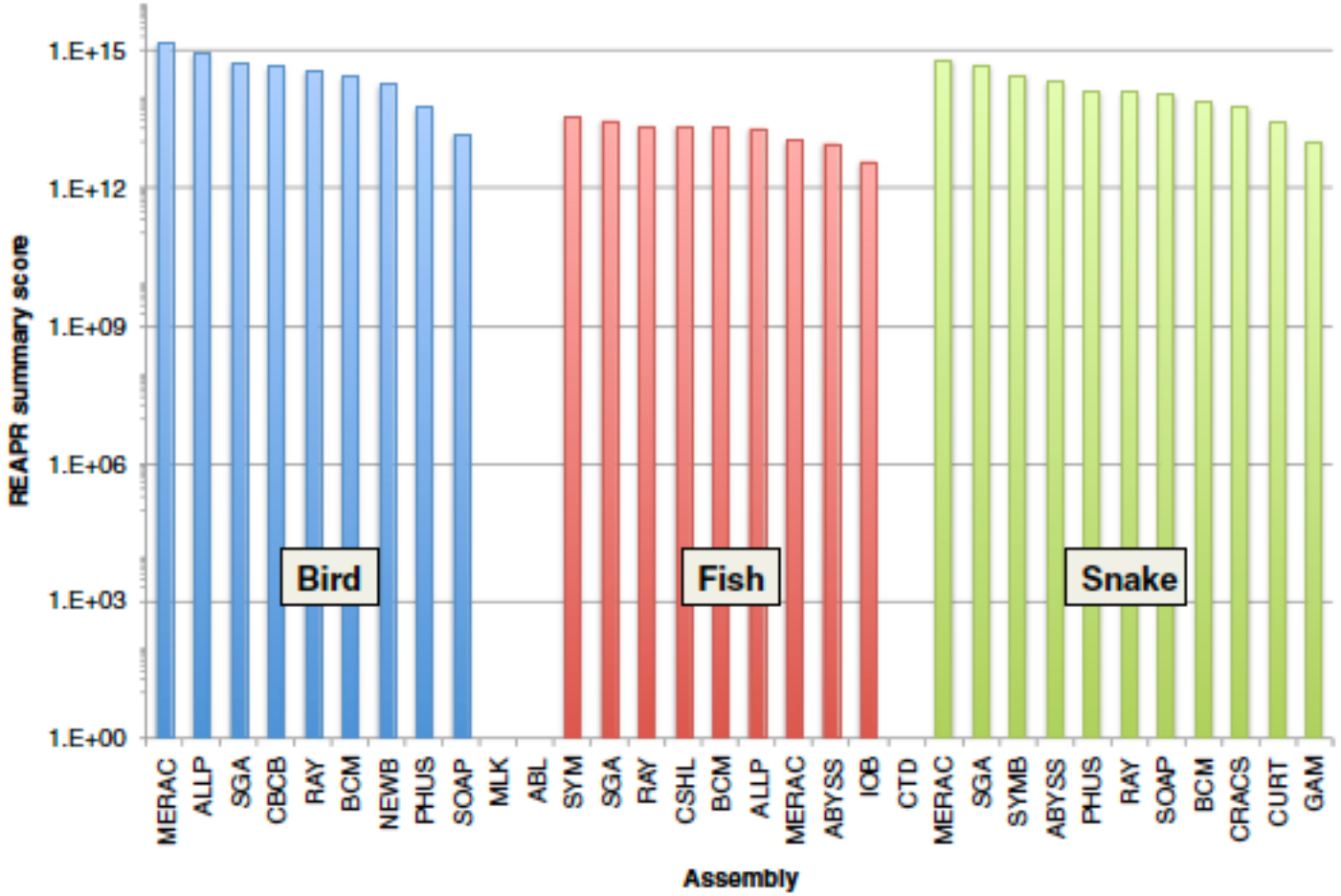
Presence of Core Genes



Fosmid Coverage for Bird and Snake Genomes



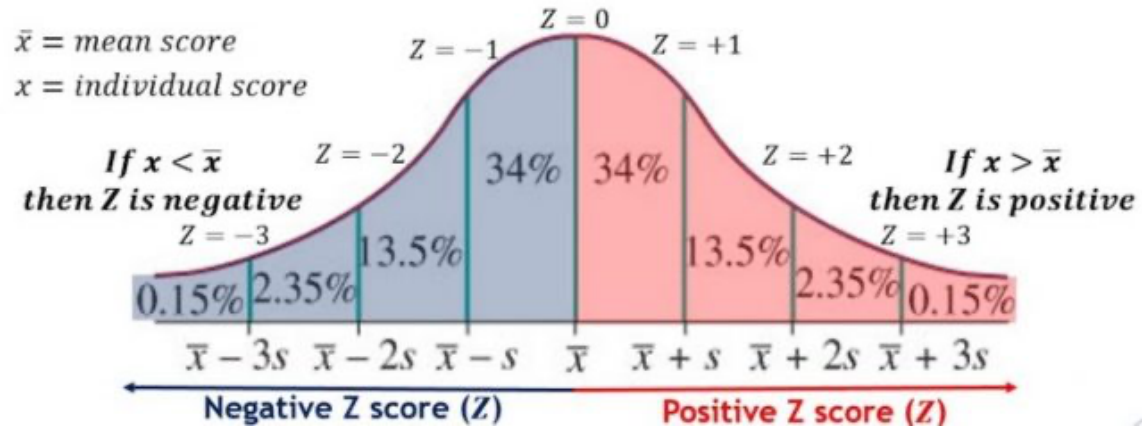
REAPR – remapping paired reads back to assembly to obtain base-by-base metrics



Hunt et al. ***REAPR: A universal tool for genome assembly evaluation.*** 2013. *Genome Biology*

Z-score

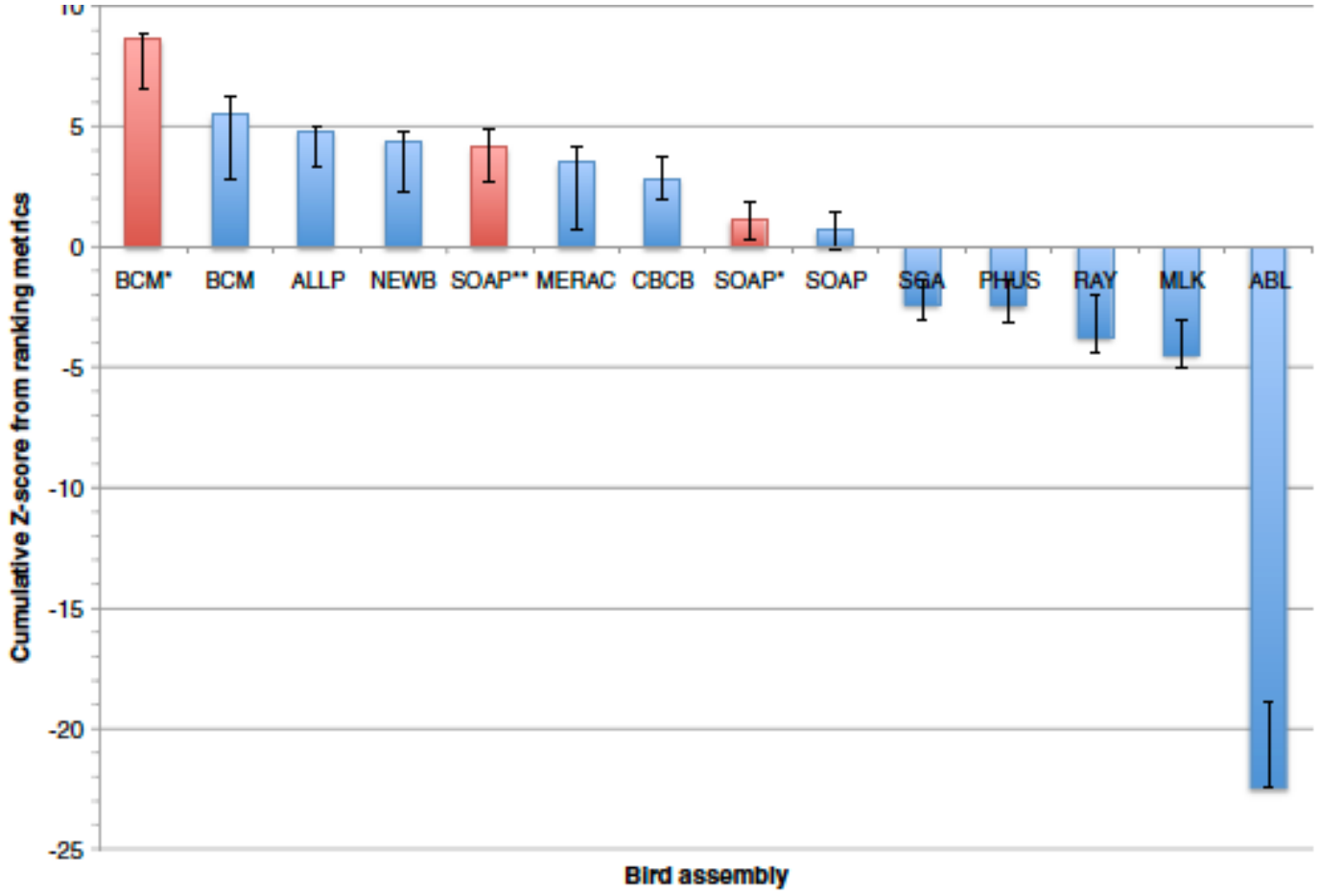
- Indicates the number of standard deviations a measurement is from the sample mean
- Z-scores for each key metric, then removed each metric and recalculated the z-score to provide error bars (resampling)
- Rewards/penalizes high/low scores in any one metric.



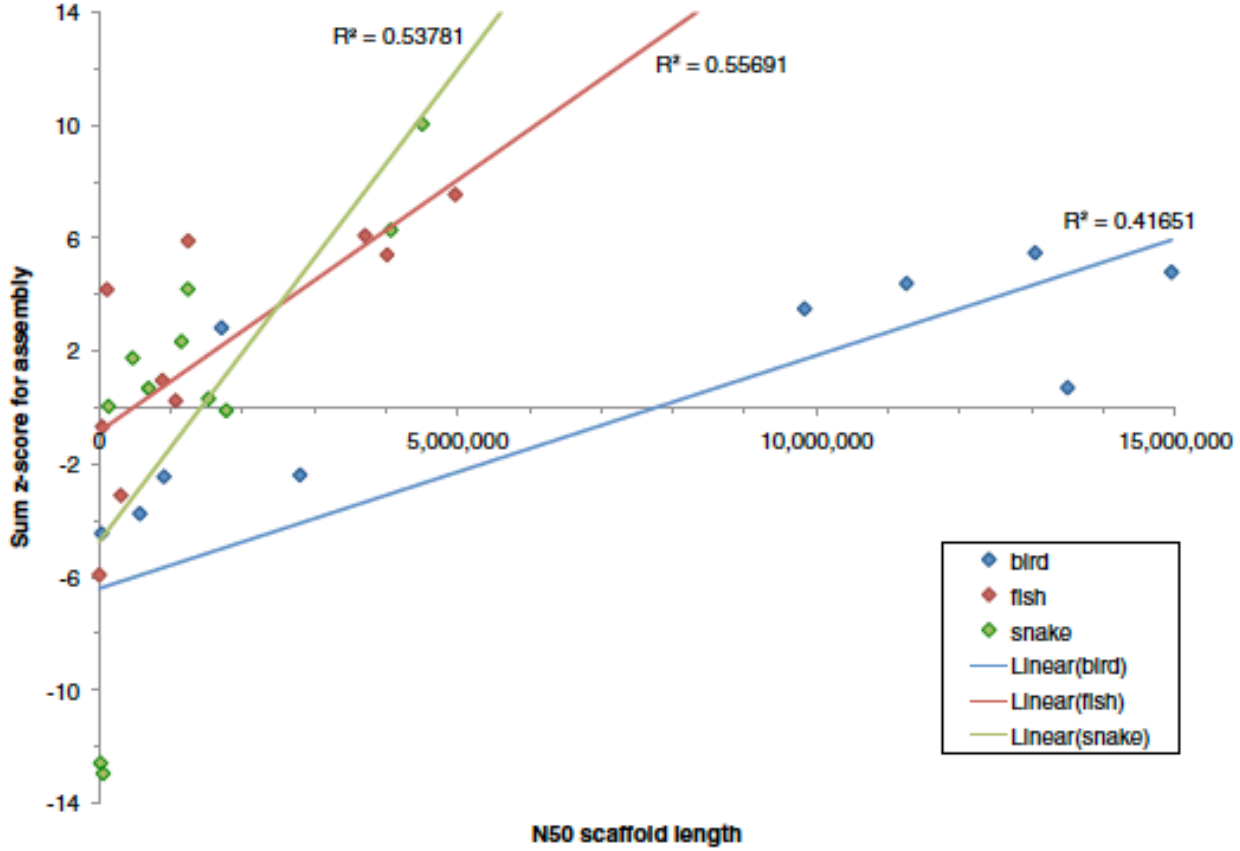
some metrics used:

- scaffold NG50
- contig NG50
- CEGMA – percentage of core genes mapped
- Fosmid coverage
- and more…

Cumulative z-score rankings for the bird



Correlation between z-score ranking and scaffold N50



All are significant – but the effect sizes differ!

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

1. Don't trust the results of a single assembly.
   - Generate several assemblies with different assemblers and/or parameters.
2. Assemblers that work well with data from one species may not work well in another.
3. Scaffold contiguity may not be related to gene coverage.
4. Heterozygosity of target genome is a big issue.
5. Don't place too much trust in a single metric.