De novo transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set: Supplementary Information

Sonal Singhal

singhal@berkeley.edu

Museum of Vertebrate Zoology
University of California, Berkeley
3101 Valley Life Sciences Building
Berkeley, California 94720-3160

Department of Integrative Biology
University of California, Berkeley
1005 Valley Life Sciences Building
Berkeley, California 94720-3140

1 Tables

individual	lineage	latitude	longitude	Locality
SS34	*C. rubrigularis* N	-16.617	145.458	Mount Harris
SS35	*C. rubrigularis* N	-16.617	145.458	Mount Harris
SS37	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS40	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS41	*C. rubrigularis* N	-16.611	145.452	Mount Harris
SS48	*C. rubrigularis* S	-17.694	145.694	S. Johnstone River, Sutties Gap Rd
SS50	*C. rubrigularis* S	-17.694	145.694	S. Johnstone River, Sutties Gap Rd
SS52	*C. rubrigularis* S	-17.660	145.722	S. Johnstone River, Sutties Gap Rd
SS56	*C. rubrigularis* S	-17.678	145.710	S. Johnstone River, Sutties Gap Rd
SS57	*C. rubrigularis* S	-17.678	145.710	S. Johnstone River, Sutties Gap Rd
SEW08448	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SEW08452	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS135	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS136	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS138	*L. coggeri* C	-16.976	145.777	Lake Morris Rd
SS64	*L. coggeri* N	-16.579	145.315	Mount Lewis
SS65	*L. coggeri* N	-16.572	145.322	Mount Lewis
SS67	*L. coggeri* N	-16.578	145.308	Mount Lewis
SS72	*L. coggeri* N	-16.585	145.289	Mount Lewis
SS74	*L. coggeri* N	-16.584	145.302	Mount Lewis
SS54	*L. coggeri* S	-17.660	145.722	S. Johnstone River, Sutties Gap Rd
SS59	*L. coggeri* S	-17.700	145.693	S. Johnstone River, Sutties Gap Rd
SS60	*L. coggeri* S	-17.700	145.693	S. Johnstone River, Sutties Gap Rd
SS62	*L. coggeri* S	-17.676	145.713	S. Johnstone River, Sutties Gap Rd
SS63	*L. coggeri* S	-17.628	145.740	S. Johnstone River, Sutties Gap Rd
SS25	*S. basiliscus* C	-17.295	145.712	Butchers Creek
SS28	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS29	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS30	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS32	*S. basiliscus* C	-17.299	145.701	Butchers Creek
SS127	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS128	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS129	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS130	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd
SS131	*S. basiliscus* S	-18.199	145.849	Kirrama Range Rd

Table 1: Individuals included in this study and their associated locality data; individuals are accessioned at the Museum of Vertebrate Zoology at University of California, Berkeley.

filtering type	rate
duplication	1.4 ± 0.2%
contamination	0.4 ± 1.1%
low-complexity reads	0.004 ± 0.003%
merging reads	68.7 ± 4.7%

Table 2: Quality control filtering and their rates for raw data, summarized across seven lineages.

database	annotated contigs	unique, annotated contigs
A. carolinensis	23804	12218
G. gallus	22324	11146
UniProt90 database	26089	12324
Ensembl 9-species database	25838	NA
Ensembl 54-species database	26601	NA

Table 3: Number of contigs annotated according to different reference databases for a randomly selected assembly.

assembly	initial chimerism	final chimerism	initial stop codons	final stop codons
C. rubrigularis, N	4.6%	0.0%	2.6%	0.6%
C. rubrigularis, S	3.7%	0.0%	2.8%	0.8%
L. coggeri, N	10.3%	0.0%	3.3%	1.1%
L. coggeri, C	5.5%	0.0%	3.1%	1.0%
L. coggeri, S	3.9%	0.0%	3.3%	1.0%
S. basiliscus, C	4.4%	0.0%	2.6%	0.6%
S. basiliscus, S	4.0%	0.0%	2.8%	0.7%

Table 4: Prevalence of chimerism, or percentage of contigs that appeared to consist of multiple genes misassembled together, and stop codons, or percentage of contigs that had nonsense mutations, in assemblies, summarized across seven lineages both before and after the data were run in the annotation pipeline.

coverage	number of contigs within lineage	number of contigs between lineages
10x	3326 ± 494	2606 ± 399
20x	1888 ± 316	1439 ± 245
30x	1311 ± 245	981 ± 178
40x	994 ± 190	741 ± 133
50x	808 ± 157	602 ± 108

Table 5: Number of annotated contigs which have given coverage for each individual; shown for one randomly selected lineage-pair.

2 Figures

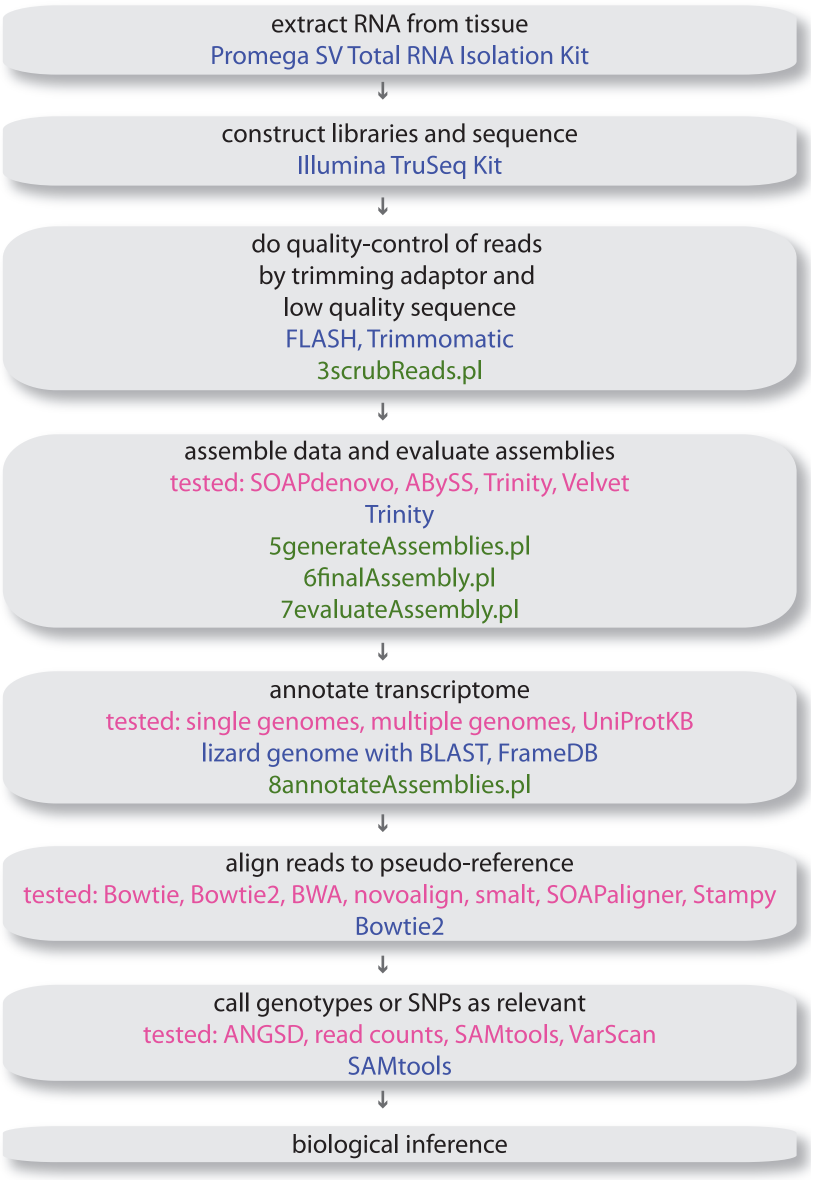

extract RNA from tissue
Promega SV Total RNA Isolation Kit
construct libraries and sequence
Illumina TruSeq Kit
do quality-control of reads
by trimming adaptor and
low quality sequence
FLASH, Trimmomatic
3scrubReads.pl
assemble data and evaluate assemblies
tested: SOAPdenovo, ABySS, Trinity, Velvet
Trinity
5generateAssemblies.pl
6finalAssembly.pl
7evaluateAssembly.pl
annotate transcriptome
tested: single genomes, multiple genomes, UniProtKB
lizard genome with BLAST, FrameDB
8annotateAssemblies.pl
align reads to pseudo-reference
tested: Bowtie, Bowtie2, BWA, novoalign, smalt, SOAPaligner, Stampy
Bowtie2
call genotypes or SNPs as relevant
tested: ANGSD, read counts, SAMtools, VarScan
SAMtools
biological inference

Figure 1: Pipeline used in this work, annotated to show (1) different approaches tested [pink], (2) the approach used for the final analysis [blue], and (3) scripts used, as named in the DataDryad package [green].

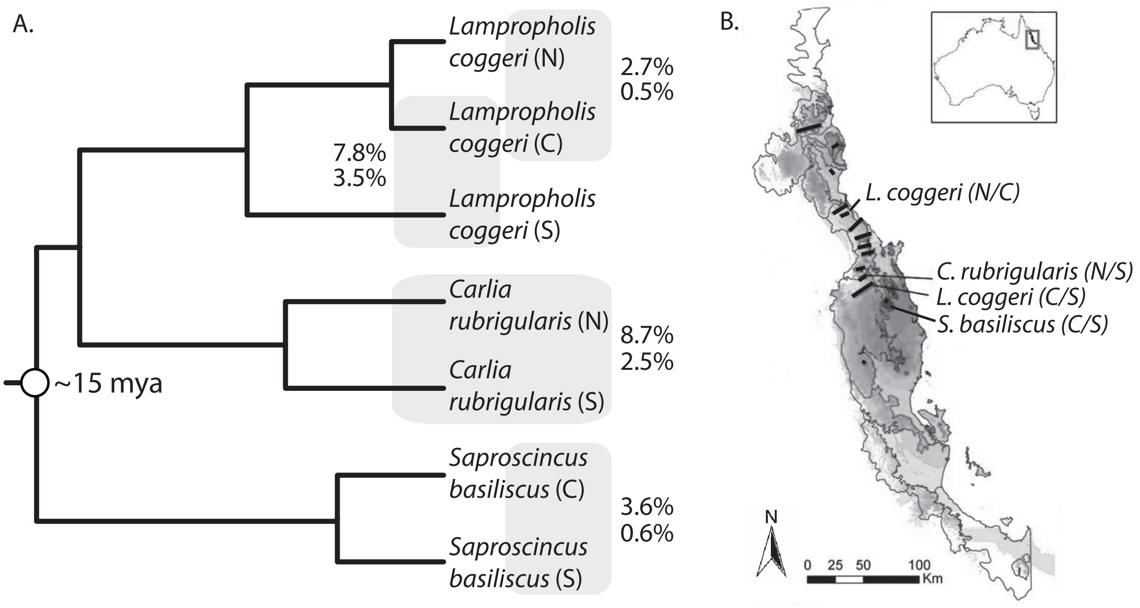


Figure 2: A. Phylogeny of the lineages studied in this work. Boxes indicate contacts studied; the top percentage reflects the mitochondrial divergence between lineages and the bottom is nuclear. B. A map of the Australian Wet Tropics, with all identified contact zones represented by black lines. Contacts of interest in this study are labelled.

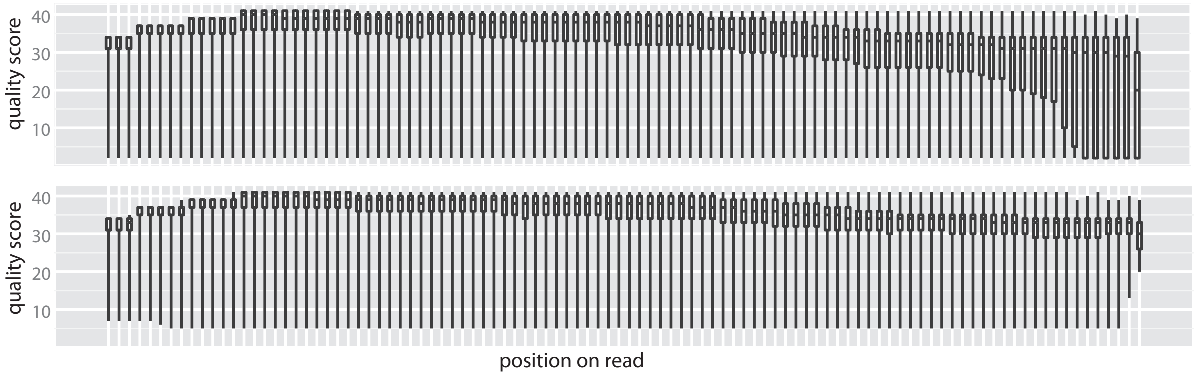


Figure 3: Quality scores in Phred along a read; top graph shows quality prior to cleaning and filtering, bottom shows quality after cleaning.

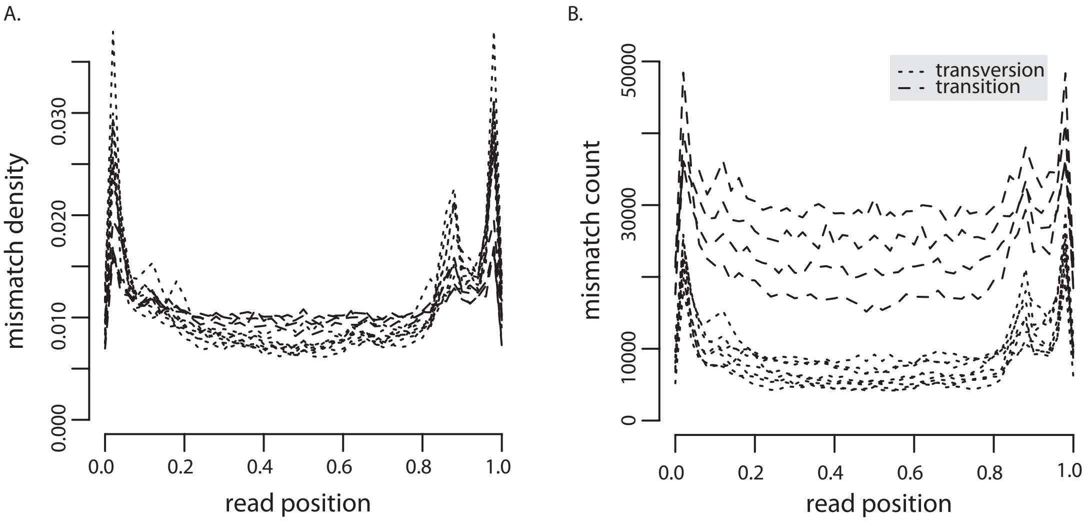


Figure 4: Identified mismatches between reads from a randomly-selected individual and the reference sequence, A. expressed in raw numbers and B. as a density distribution.

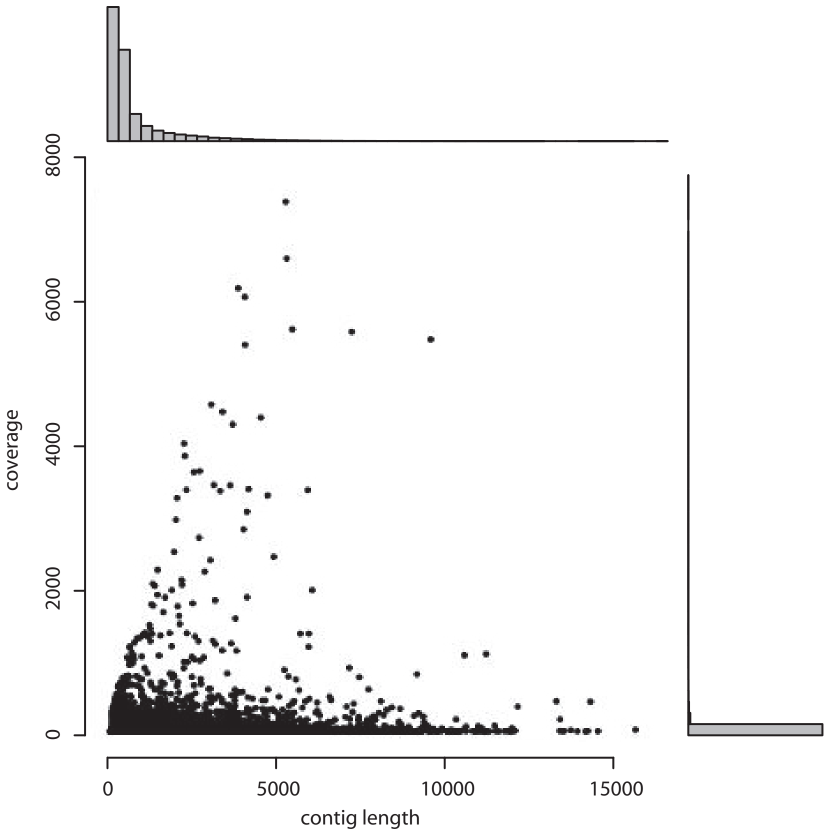
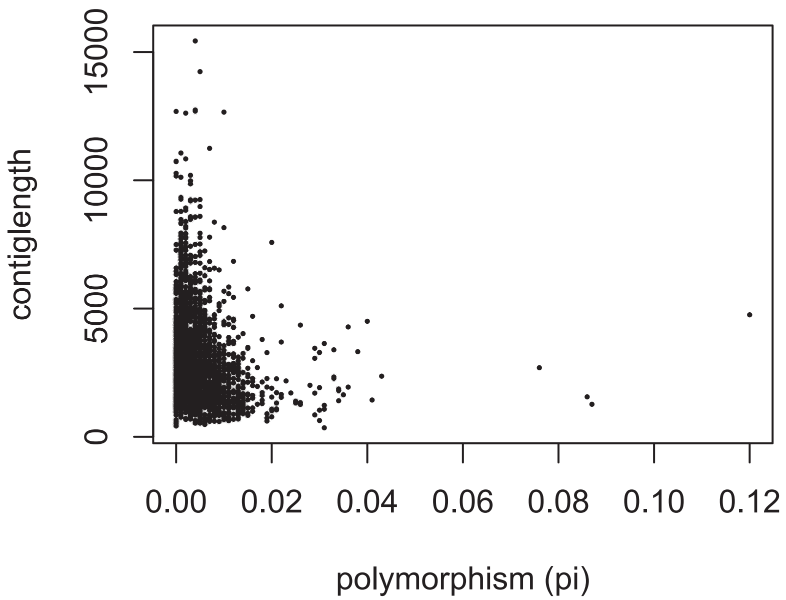


Figure 5: Correlation between contig length and coverage for a randomly-selected final assembly.


15000
10000
5000
0
contiglength
0.00
0.02
0.04
0.06
0.08
0.10
0.12
polymorphism (pi)

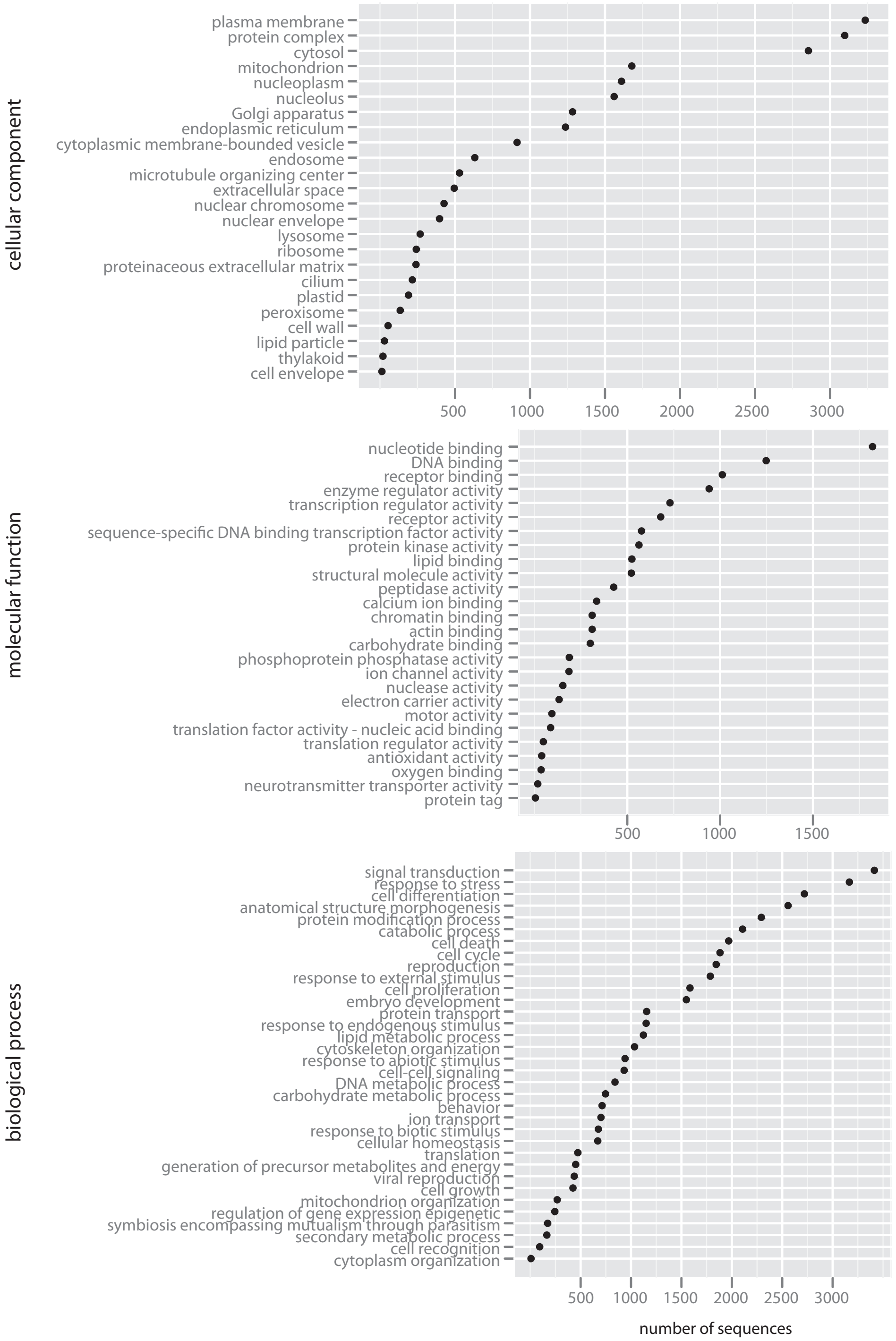


Figure 7: Gene ontology for annotated contigs for a randomly-selected lineage, with respect to cellular component, biological process, and molecular function.

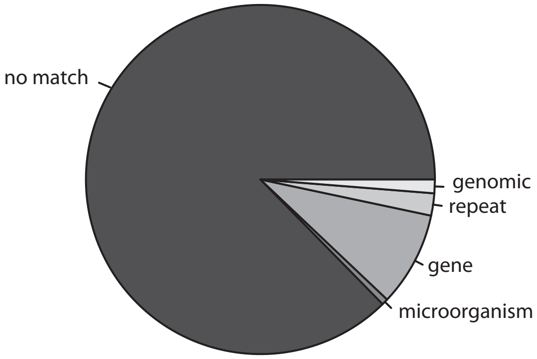


Figure 8: Identify of unannotated contigs from a randomly selected assembly, as identified from a BLAST search to the NCBI 'nr' nucleotide database.

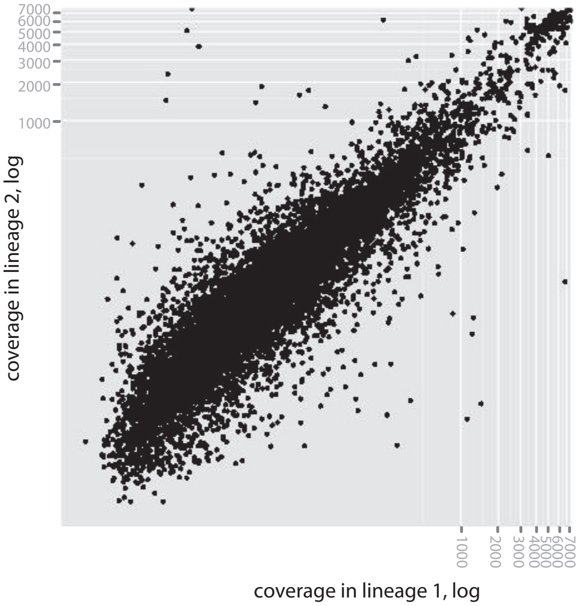


Figure 9: Correlation in coverage between homologous, annotated contigs for a randomly-selected lineage-pair.

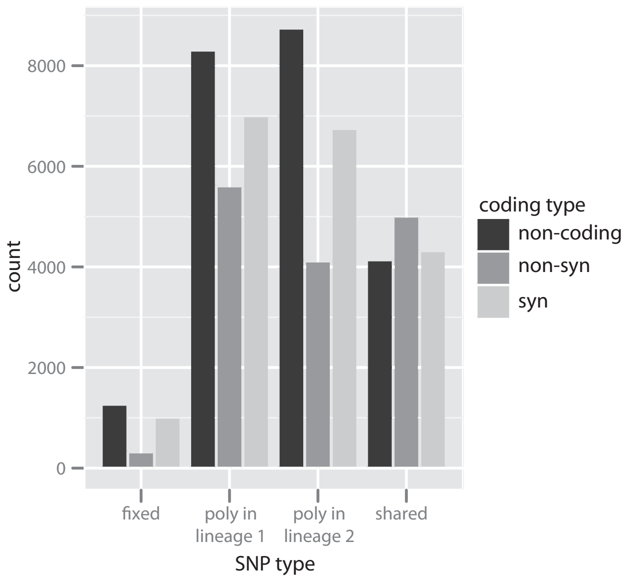


Figure 10: Summary of SNPs found in a randomly-selected lineage-pair, annotated with respect to SNP and coding type.

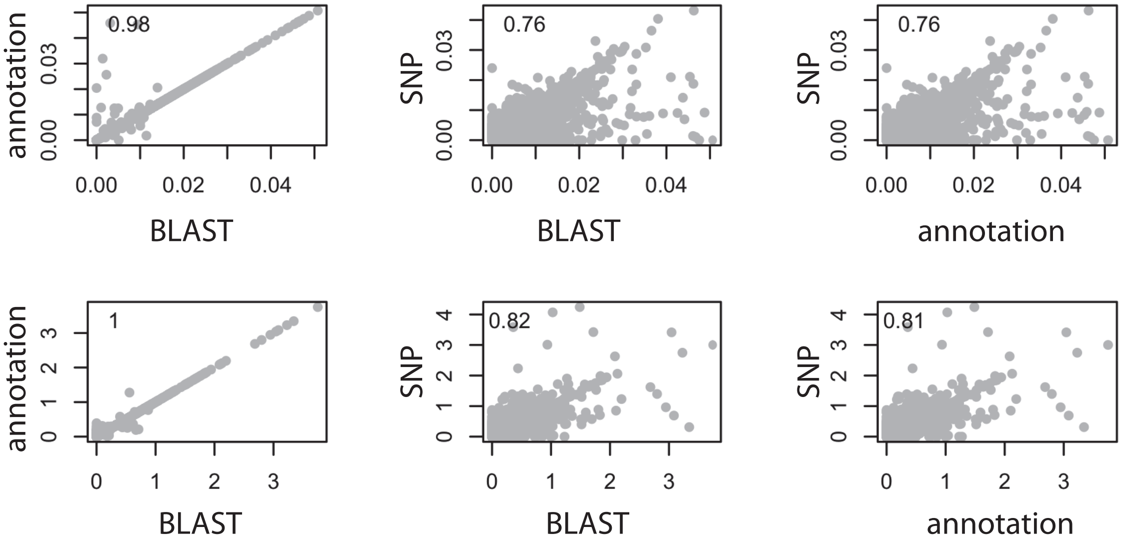


Figure 11: Top row shows correlation in sequence divergence and bottom row shows correlation in inferred $\frac{dN}{dS}$ ratios for homologs for a randomly-selected lineage-pair for three methods of homolog discovery: annotation, in which contigs which share the same annotation are inferred to be homologous, BLAST, in which reciprocal best-hit BLAST is used to identify homologs, and SNP methods, in which variant information is used to reconstruct one homolog with respect to another.