Pseudo reference approach for comparative genomic studies

Phylogenomic Insights into Mouse Evolution Using a Pseudoreference Approach

Brice A.J. Sarver[1], Sara Keeble[1], Ted Cosart[1], Priscilla K. Tucker[2], Matthew D. Dean[3], and Jeffrey M. Good[1,*]

[1]Division of Biological Sciences, University of Montana, Missoula, MT

[2]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI

[3]Molecular and Computational Biology, University of Southern California, Los Angeles, CA

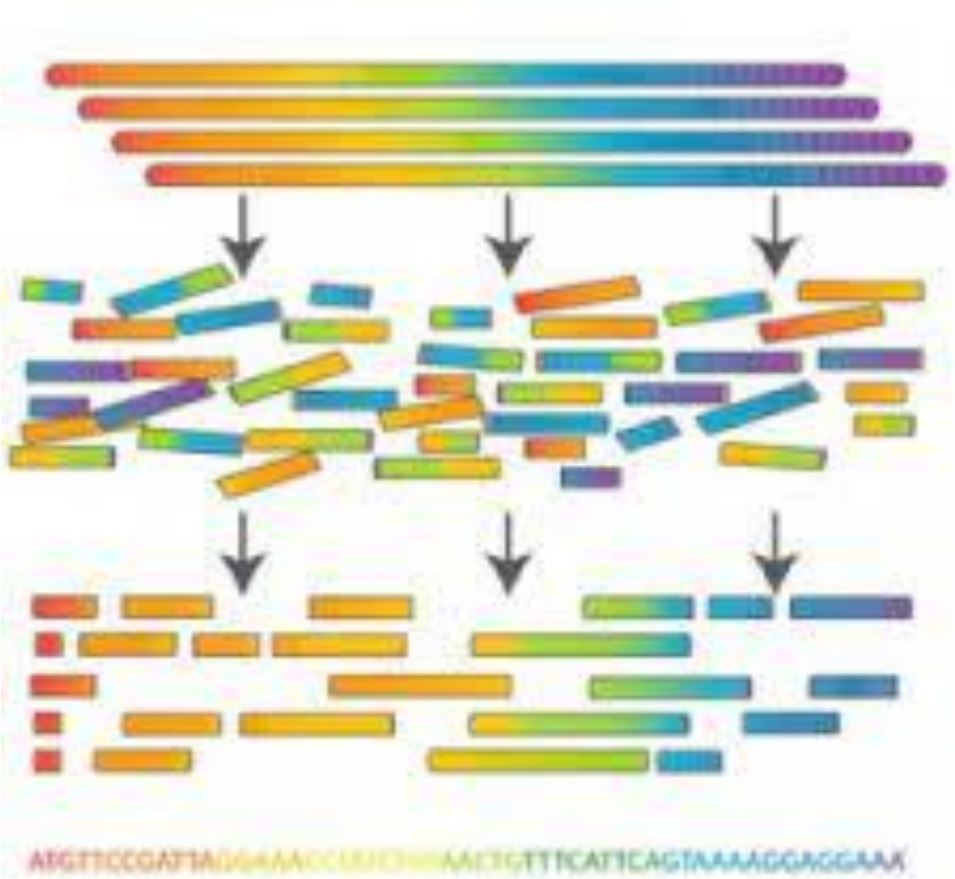
*Corresponding author: E-mail: jeffrey.good@umontana.edu.

Accepted: February 20, 2017

Data deposition: This project has been deposited at the NCBI Sequence Read Archive under the accession PRJNA323493.

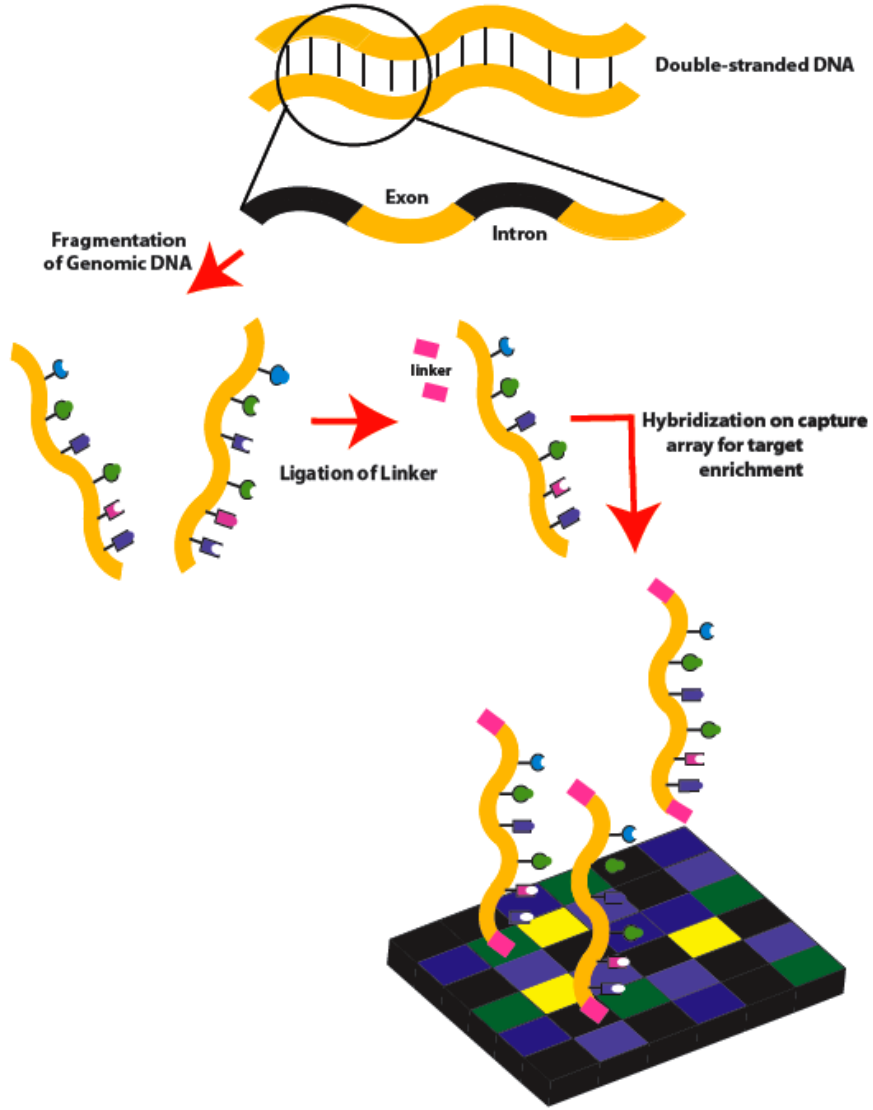
Comparative genomics using whole genomes

- De novo assemblies can be done with Pac-Bio long reads
- Very high coverage short-reads (~80x)
- Optimal but expensive
- Not feasible for large data sets (specious groups)
- Not feasible for large complex genomes (*e.g.* salamanders)



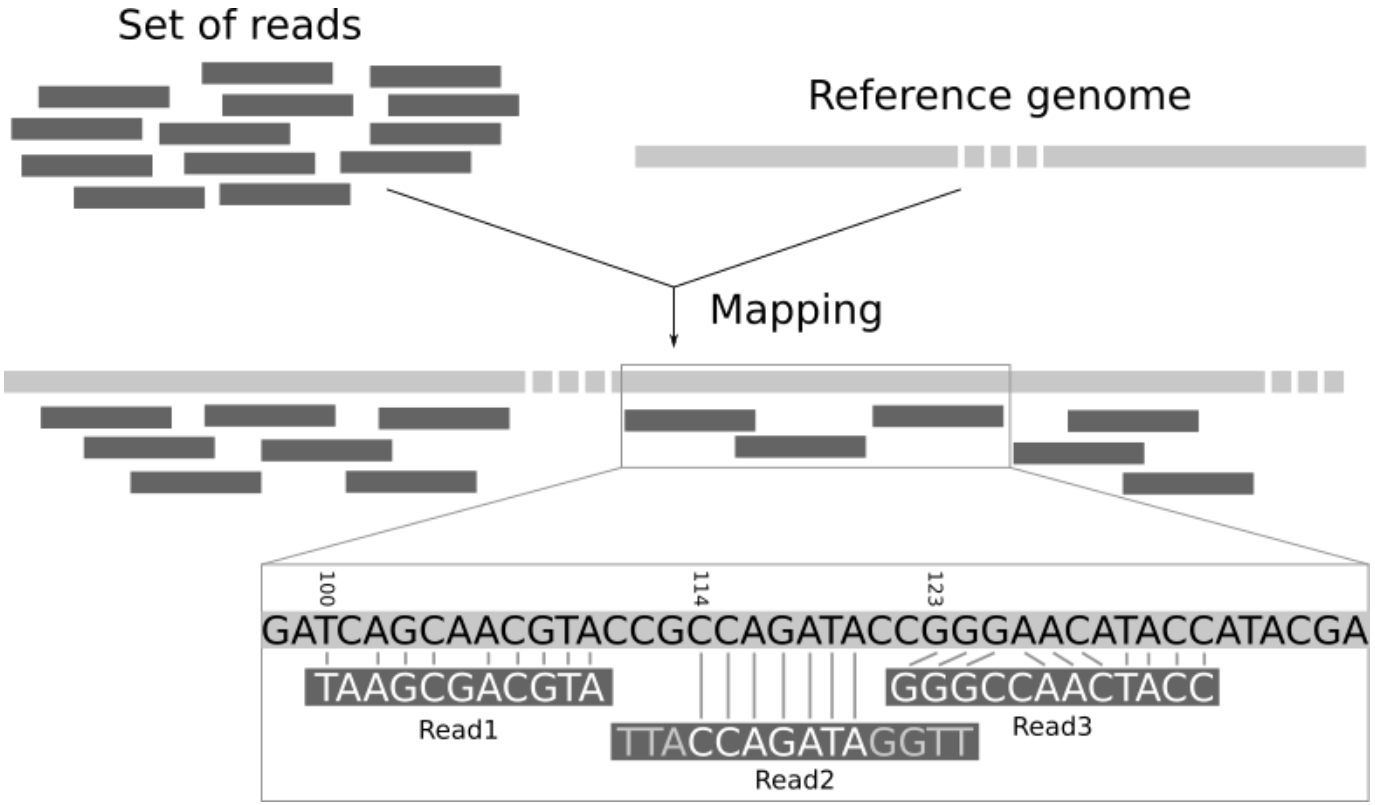
Partitioning approaches as alternatives

- Restriction site associated sequences
- Targeted capture
- Transcriptomic
- Whole genomes short-reads (lower coverage)



Partitioned genomic data challenges

- Rely on a high-quality reference genome ideally from the same species or closely related lineage
- Mapping to a divergent reference can generate a number of systematic biases
- Genotypes may converge towards the reference, resulting in an overestimated similarity between subject and reference sequences in divergent regions.



“Pseudogenomes”

- Reference genomes that incorporate sample-specific variation.
- Allows annotation to be carried over from a reference while accounting for sequence divergence during the mapping stage.

House mouses

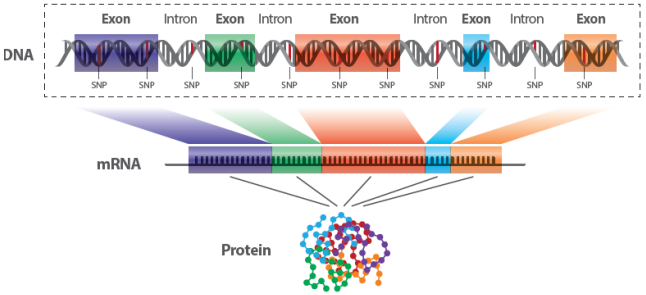
- Radiation of ~ 38 species that share a common ancestor ~7.5 mya
- Reference genome of *Mus musculus* second best in quality
- Unresolved evolutionary relationships among some lineages
- Uncertainty of how much phylogenetic discordance there is across the house mouse genome due to ILS or introgression

Objectives

1) Developed a scalable pseudoreference approach to iteratively incorporate sample-specific variation into a reference and thereby reduce the effects of systematic mapping bias in downstream analyses
2) Evaluate the performance of the newly developed pseudoreference approach on 10 species of *Mus*
3) Resolved phylogenetic relationships while assessing phylogenomic discordance by ILS and introgression

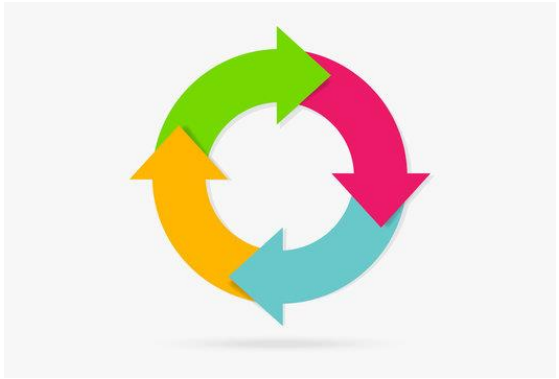
Methods

1 Exome capture sequencing



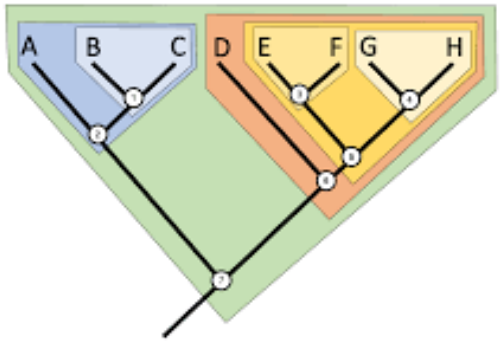
- Sequences the protein-coding regions of the genome
- Cost-effective alternative to whole-genome sequencing

2 QC & Iterative mapping

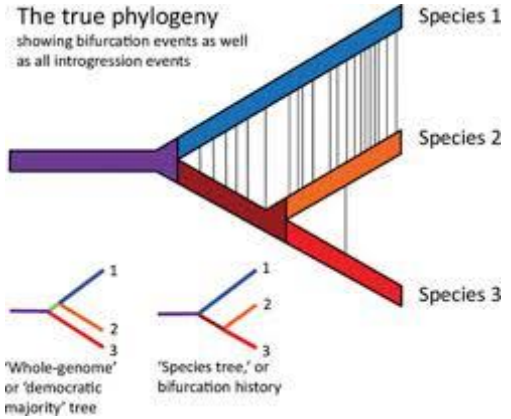


- Quality check of the reads
- Assemble the sample genome by iteratively incorporate sample-specific variation into a reference.

3 Phylogenetic inference

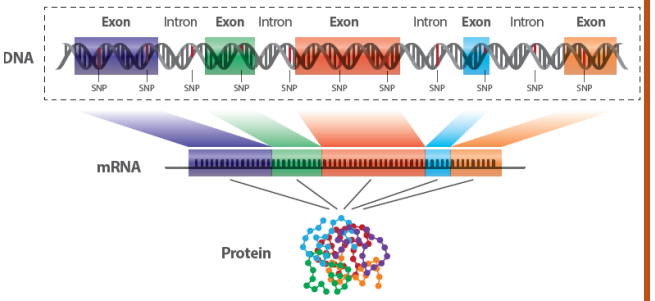


4 Introgression analysis



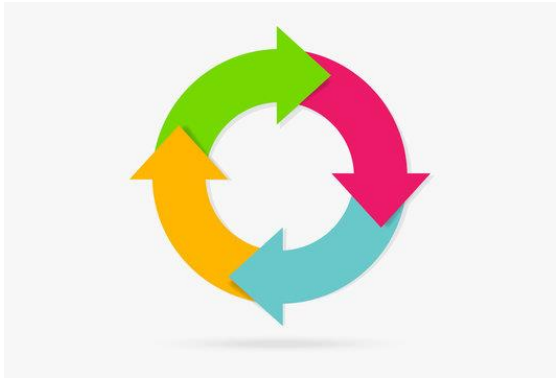
Methods

1 Exome capture sequencing



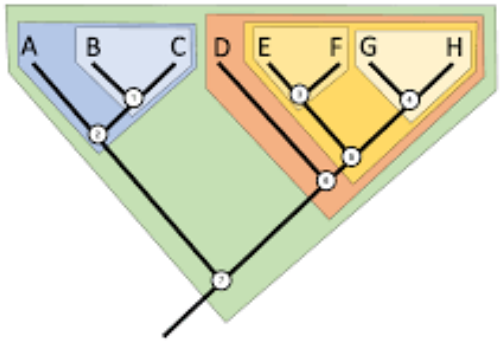
- Sequences the protein-coding regions of the genome
- Cost-effective alternative to whole-genome sequencing

2 QC & Iterative mapping

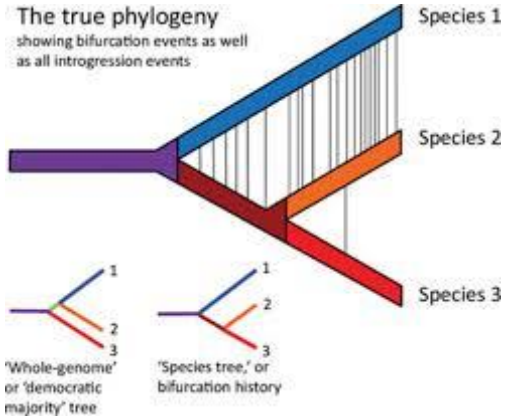


- Quality check of the reads
- Assemble the sample genome by iteratively incorporate sample-specific variation into a reference.

3 Phylogenetic inference

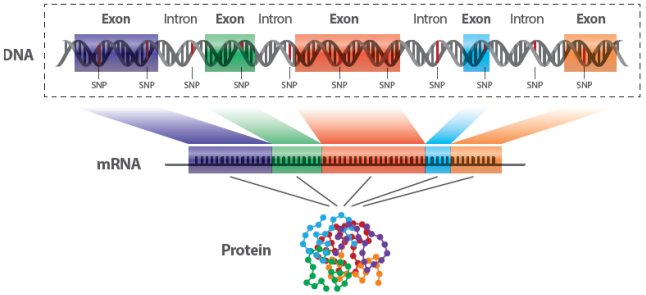


4 Introgression analysis



Methods

1 Exome capture sequencing



- Sequences the protein-coding regions of the genome
- Cost-effective alternative to whole-genome sequencing

- Illumina sequencing libraries
- Ten species of *Mus*

Methods

2 QC & Iterative mapping

- Quality check of the reads
- Assemble the sample genome by iteratively incorporate sample-specific variation into a reference.

expHTS

========

Python application for "Experimental High Throughput Sequencing"

1. Screen for contaminants (minimally PhiX) Using a mapping based approach
2. Deduplicate Reads Using Super-Deduper
3. Quality Trimming (polyA/T) trimming Using Scickle
4. Overlap paired-end reads Using Flash2
5. Modify read names, filter for too-short, QA
6. Produce multi-sample reports

Methods

2 QC & **Iterative mapping**

- Quality check of the reads
- Assemble the sample genome by iteratively incorporate sample-specific variation into a reference.

PSEUDO-IT

Pseudo-it: Reference-guided genome assembly with iterative mapping

install with bioconda | Platforms noarch | version v3.1.1 | Last updated 06 Mar 2023 | downloads 4k | codecov 90% | license GPL-3.0

For the first iteration, pseudo-it performs the following steps:

1. Map provided reads (FASTQ) to **provided reference genome (FASTA).**
2. Call variants on mapped reads.
3. Generate a consensus FASTA file by inserting the called variants into the original sequence.

For each subsequent iteration, the previous iteration's consensus FASTA file serves as the new reference for read mapping:

1. Map provided reads (FASTQ) to **previous iteration consensus sequence (FASTA).**
2. Call variants on mapped reads.
3. Generate a new consensus FASTA file by inserting the called variants into the previous iteration's sequence.

Each iteration should allow for more reads to be mapped and more variation to be incorporated into the assembly.

Methods

2 QC & **Iterative mapping**

- Quality check of the reads
- Assemble the sample genome by iteratively incorporate sample-specific variation into a reference.

PSEUDO-IT

Pseudo-it: Reference-guided genome assembly with iterative mapping

install with bioconda | Platforms noarch | version v3.1.1 | Last updated 06 Mar 2023 | downloads 4k | codecov 90% | license GPL-3.0

- Python script that coordinates the running of other software
- Runs the following common bioinformatic software:

1. BWA for read mapping.
2. GATK for variant calling.
3. samtools for handling mapped reads (BAM).
4. Picard Tools for handling mapped reads (BAM).
5. bedtools for soft-masking FASTA files.
6. bcftools for handling VCF files and generating consensus FASTA and .chain files.

Step by step

1) map reads to the genome: Cleaned sequence reads are aligned to the reference genome using **BWA (Burrows-Wheeler Aligner) with its MEM algorithm**

2) Eliminating Duplicate Reads: After mapping, duplicate reads are identified using **Picard**. Duplicate reads are typically result from PCR amplification.

3) Multiply Mapped Reads: For reads that can map to multiple locations in the genome, only the location with the best mapping quality is kept for further analysis.

4) Indel Realignment: Regions of the genome where insertions or deletions (indels) occur are realigned to make sure these variations are accurately represented.

5) Calling Variants: HaplotypeCaller in the GATK (Genome Analysis Toolkit) is used to identify single nucleotide variants (SNVs). This tool looks for differences between the mapped reads and the reference genome to call variants.

6) Filtering Variants: Eliminates variants that 1) have less de 30 of Quality score and 2) have less than 5 reads of sequencing depth.

7) Updating Reference with Variants: The identified variants are incorporated back into the reference genome using a tool called **FastaAlternateReferenceMaker** from GATK.

8) File Processing: Additional processing steps, like **indexing**, **merging**, and **sorting** the sequencing files, are performed using **SAMtools**

Step by step

PSEUDO-IT

Pseudo-it: Reference-guided genome assembly with iterative mapping

1) map reads to the genome

2) Eliminating Duplicate Reads

3) Multiply Mapped Reads

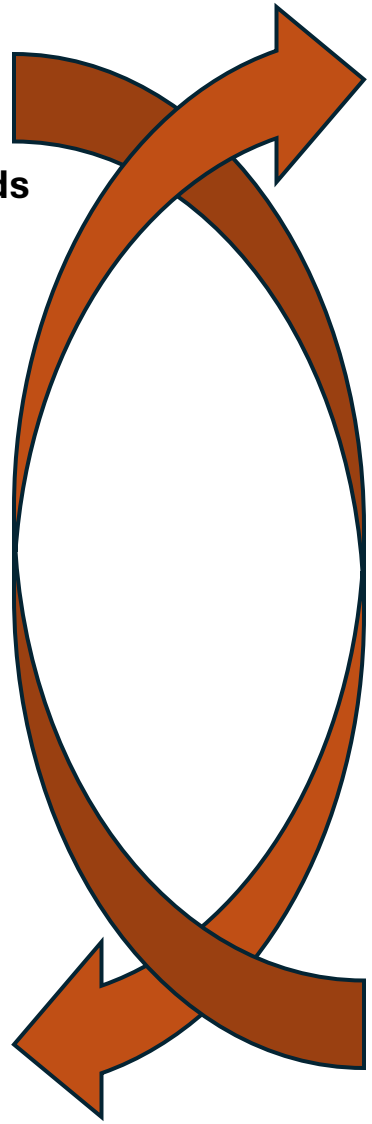
4) Indel Realignment

5) Calling Variants

6) Filtering Variants

7) Updating Reference with Variants

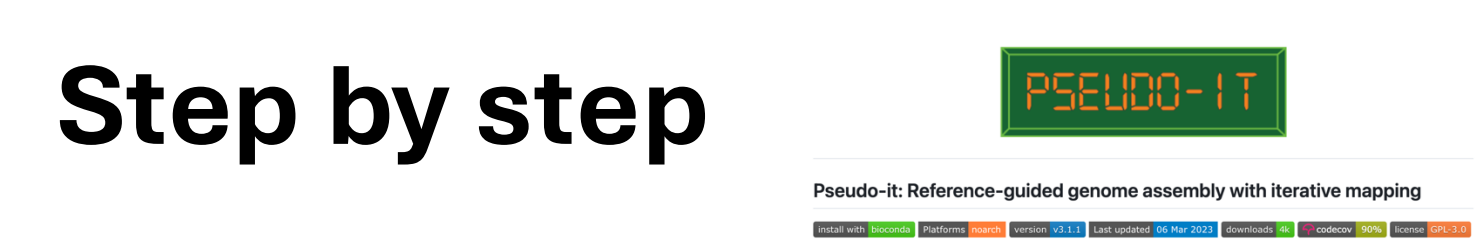
8) File Processing



Step by step

PSEUDO-IT

Pseudo-it: Reference-guided genome assembly with iterative mapping



1) map reads to the genome

2) Eliminating Duplicate Reads

3) Multiply Mapped Reads

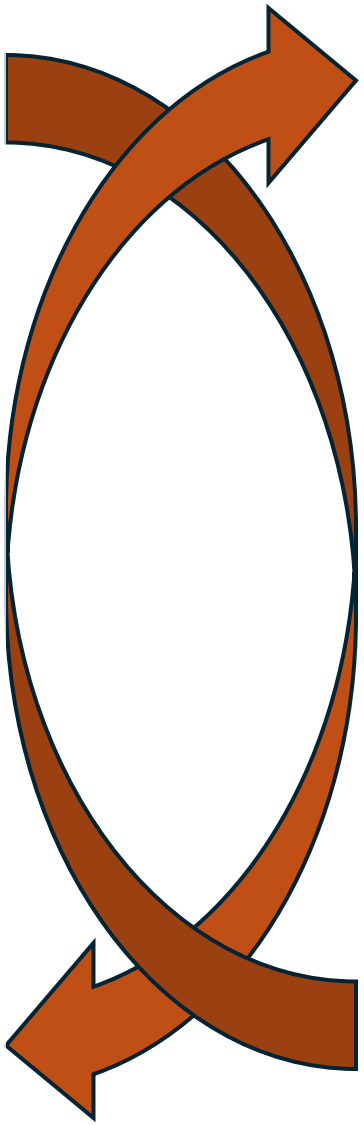
4) Indel Realignment

5) Calling Variants

6) Filtering Variants

7) Updating Reference with Variants

8) File Processing



Final steps:

- Exclude positions with insufficient data
- Additional variant calling in all positions of the genomes (not only where it differences with the reference.
- Hard masking ambiguous positions (fills them with Ns)

Results: Iterative mapping

- reads were more confidently placed with each pseudoreference, resulting in an increase in usable bases and fewer reads discarded due to low mapping quality

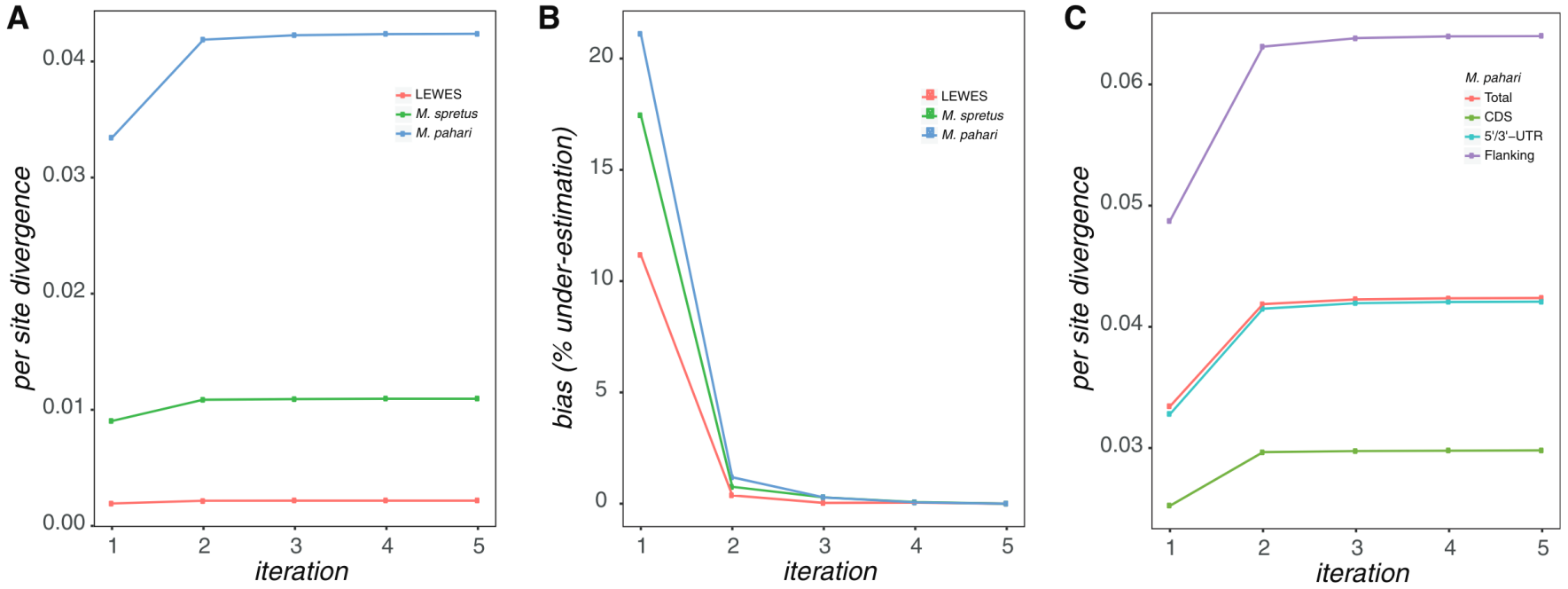
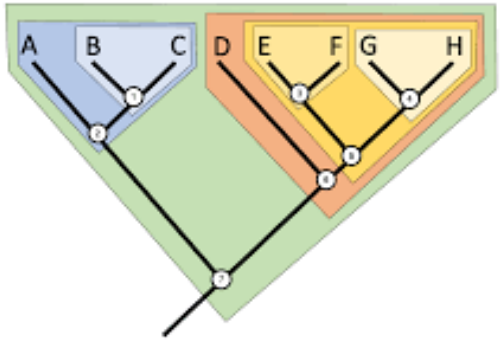


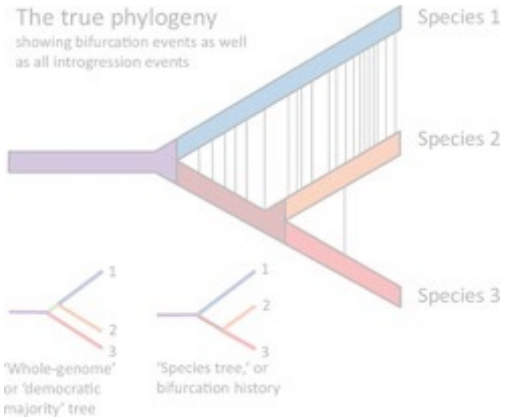
Fig. 1.—Reference bias and sequence divergence. (*A*) Per-site sequence divergence per iteration using confidently called positions on Chromosome 1 for *M. m. domesticus* (dom^{LEWES}), *M. spretus*, and *M. pahari*. (*B*) The bias in divergence estimates (% under-estimation) at each iteration relative to the per-site divergence of the sample's five iteration pseudoreference using the same data. (*C*) Per-site divergence for *M. pahari* partitioned by protein-coding sequence (CDS), untranslated exonic regions (5′/3′-UTR) and flanking sequences.

Downstream analysis with ”psuedogenomes”

3 Phylogenetic inference



4 Introgression analysis



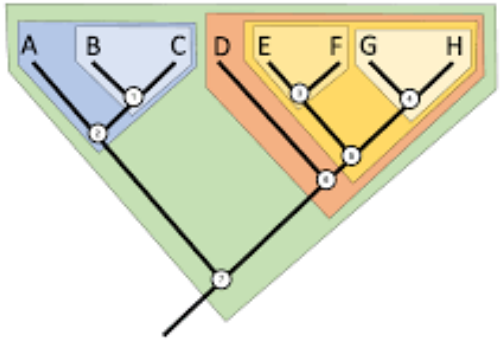
Two phases appproach

1) **Concatenated phylogeny for each chromosome:**

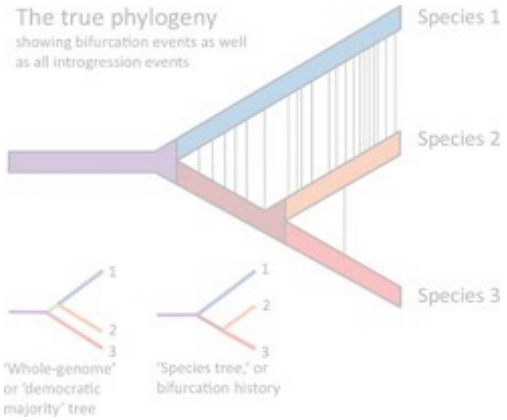
2) **Discordance analysis**

Downstream analysis with ”psuedogenomes”

3 Phylogenetic inference



4 Introgression analysis



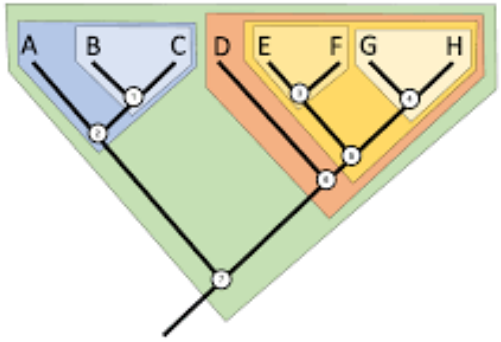
Two phases appproach

1) **Concatenated phylogeny:** phylogeny was estimated using a concatenated alignment of gene sequences

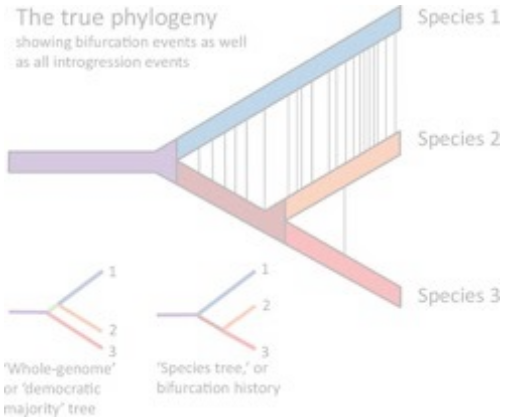
- Extracted exons
- BioMart: Ortholog transcripts
- TranslatorX: transcripts alignments
- Phyutility: transcripts were concatenated by chromosome into a supermatrix.
- RAxML: Maximum likelihood (ML) trees were estimated
- FigTree, and discordance among the trees (from different chromosomes) was assessed using Robinson-Foulds distances

Downstream analysis with ”psuedogenomes”

3 Phylogenetic inference



4 Introgression analysis



Two phases appproach

2) Discordance analysis: accounted for discordance between gene trees and species trees

- Estimated species tree using ASTRAL (breaks down gene trees into quartets).
- Phylogenetic trees were estimated for different window sizes across the genome (100 kbp to 5 Mbp) using RaxML.
- The species trees inferred from all windows were combined in ASTRAL.
- Quartet discordance, which measures how much gene trees disagree with the species tree, was used to quantify phylogenetic discordance.

Results

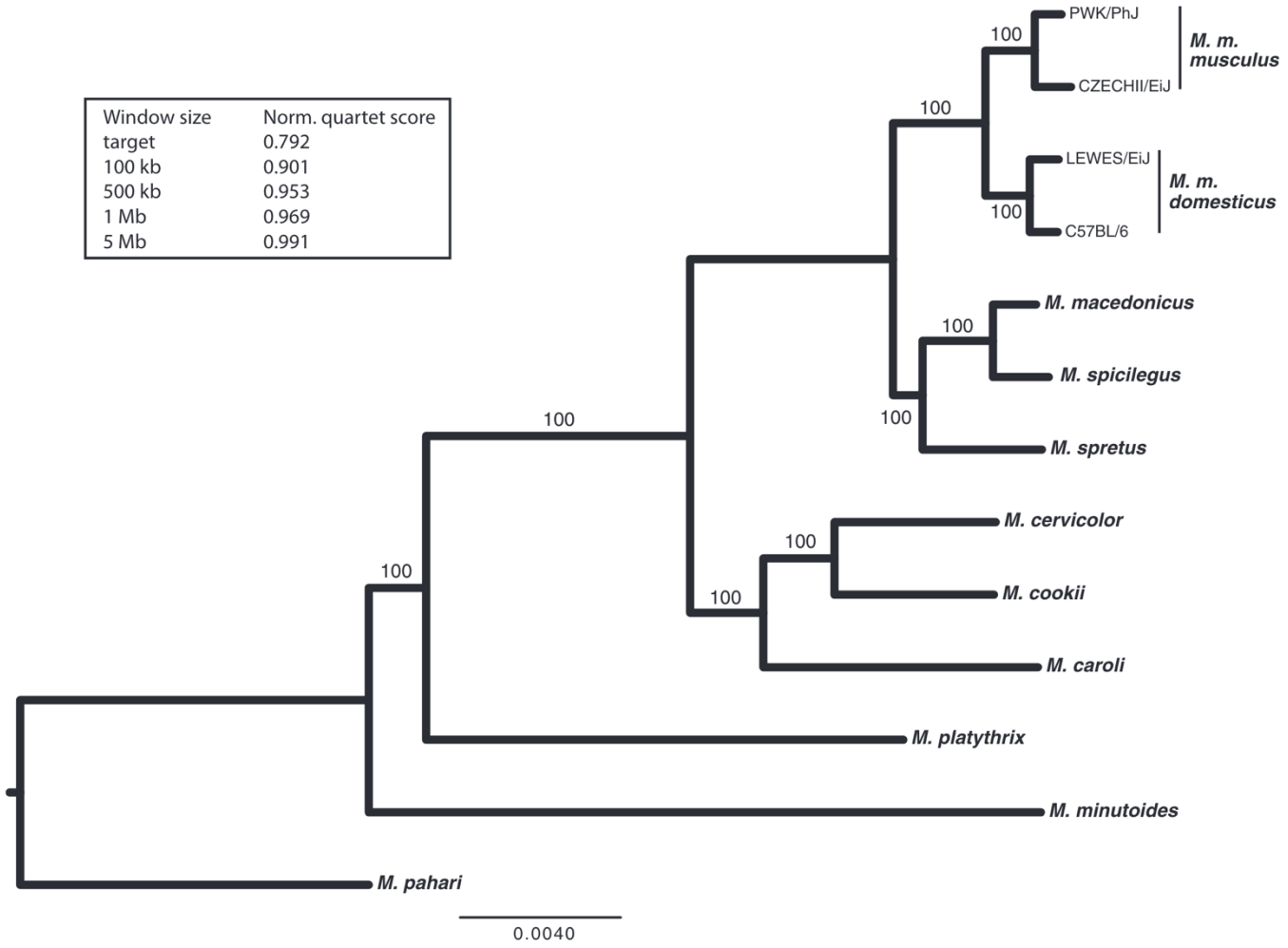


FIG. 3.—*Mus* phylogeny, rooted on *M. pahari*, estimated using all extended targets from Chromosome 1. ML bootstrap support values are listed above branches. There was no discordance between this tree and trees estimated from other chromosomes. The inset provides the normalized quartet scores calculated with ASTRAL from local genealogies estimated at five genomic scales.

Fully resolved concatenated phylogeny with no discordance among chromosomes.

Concatenation can inflate confidence in the overall tree and obscure incongruence.

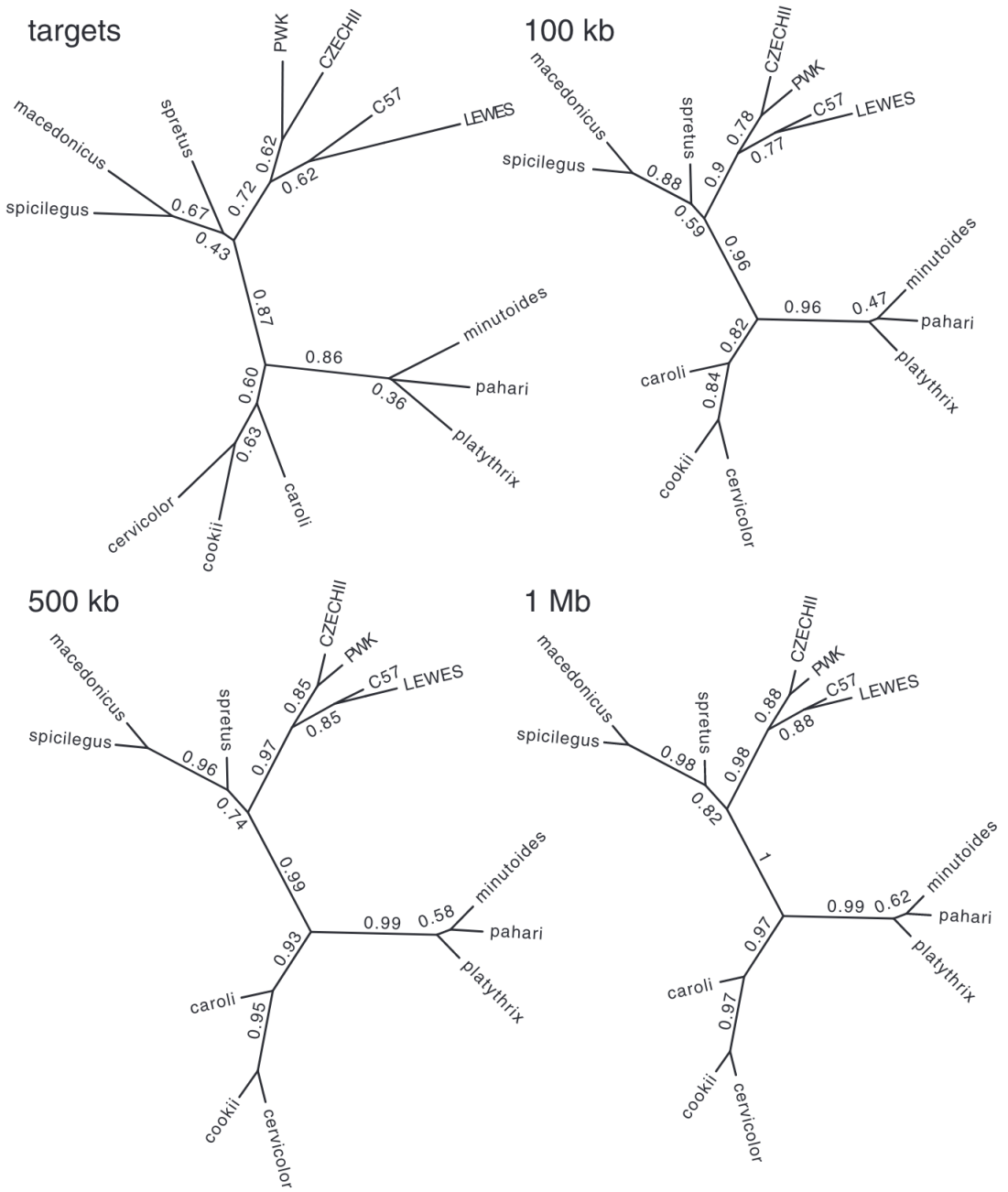


FIG. 4.—Unrooted species tree estimates from ASTRAL across four different window sizes (5 Mbp not shown). Branches are annotated with their local quartet scores.

No discordance in the species tree

Variation in quartet support among different window sizes

Targets and lower size windows have lower supports