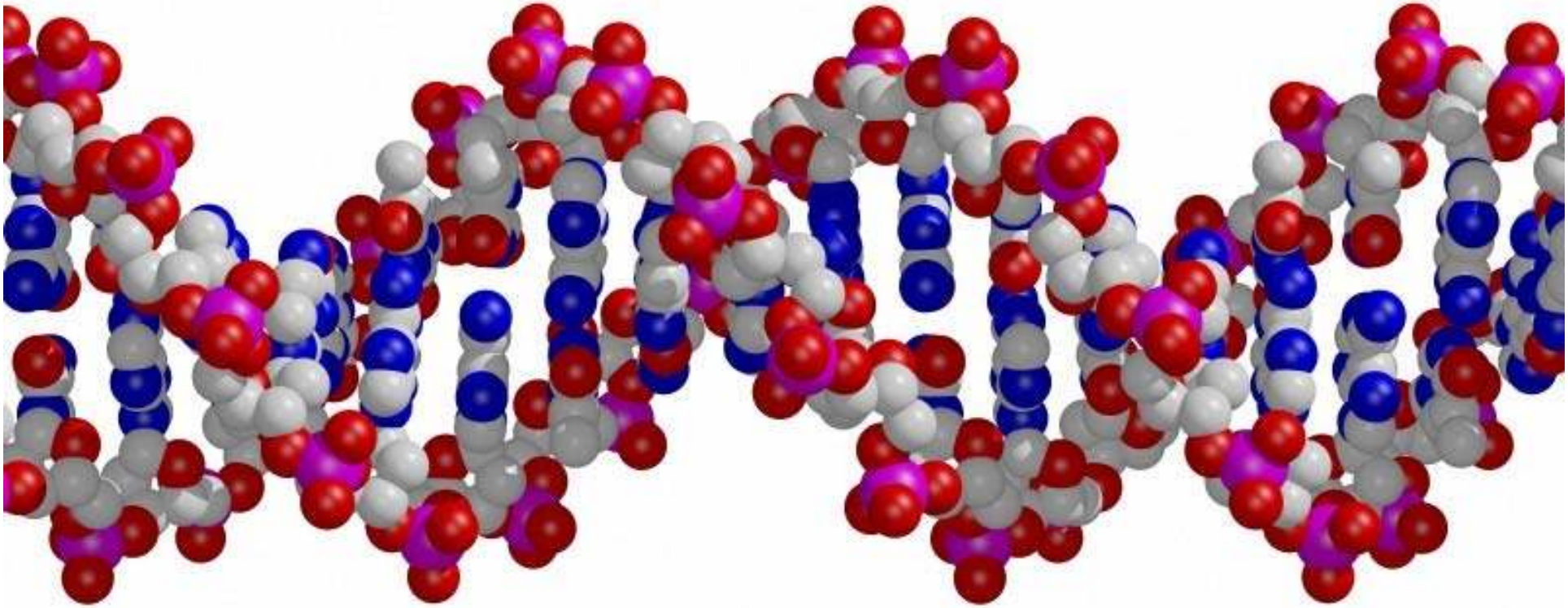
Phylogenomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University

http://as.vanderbilt.edu/rokaslab

High-Throughput DNA Sequencing Technologies

454 / Roche	Illumina	Helicos
450 bp 1.5 Gbp / day	150 bp 35 Gbp / day	55 bp 4.5 Gb / day

“Traditional” Sanger / Capillary Sequencing	
650 bp	2 Mbp / day

SOLiD ABI	PacBio	Ion PGM
75 bp 22 Gbp / day	1000 bp 70 Gbp / day	100 bp 120 Gbp / day

High-Throughput DNA Sequencing Technologies

The principle:

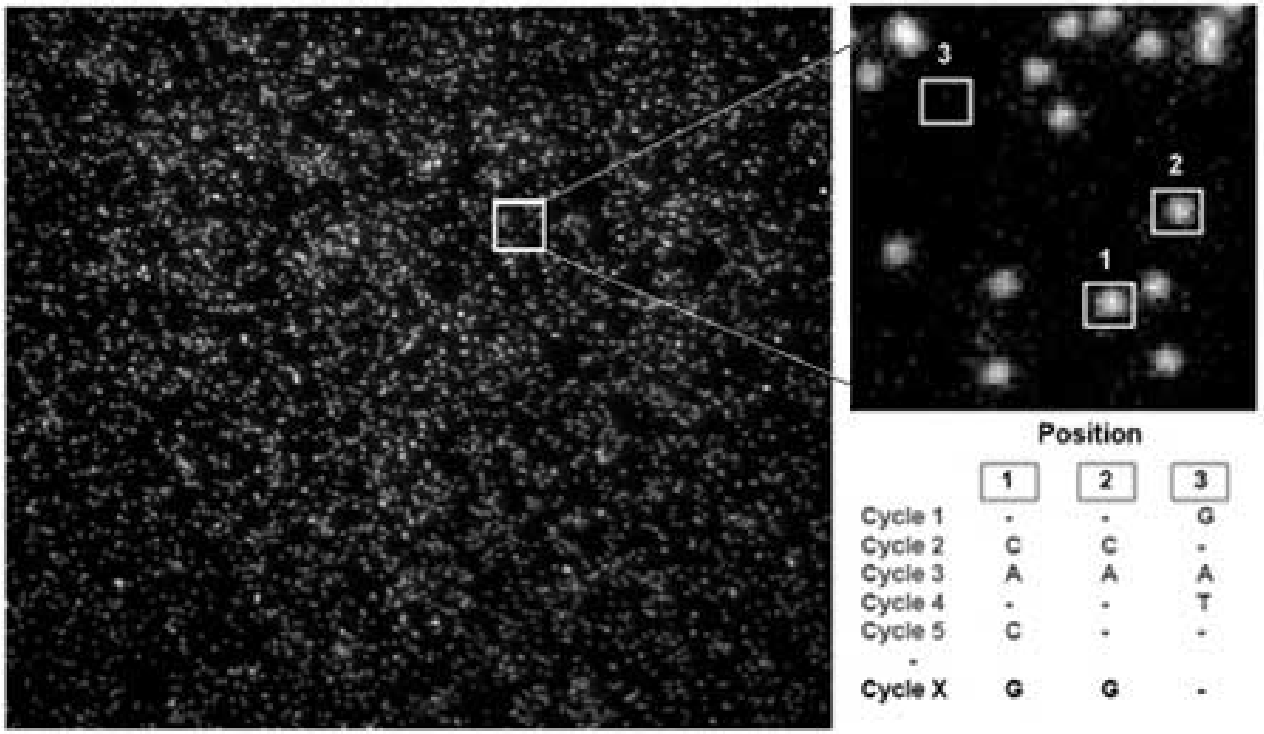
Detecting base incorporation during DNA strand synthesis at a massively parallel scale. Base detection and strand synthesis are key differences between competing technologies

How it works:

1. Physically separate a large number of ssDNA fragments

2. Fix the fragments' location on a substrate (& amplify)

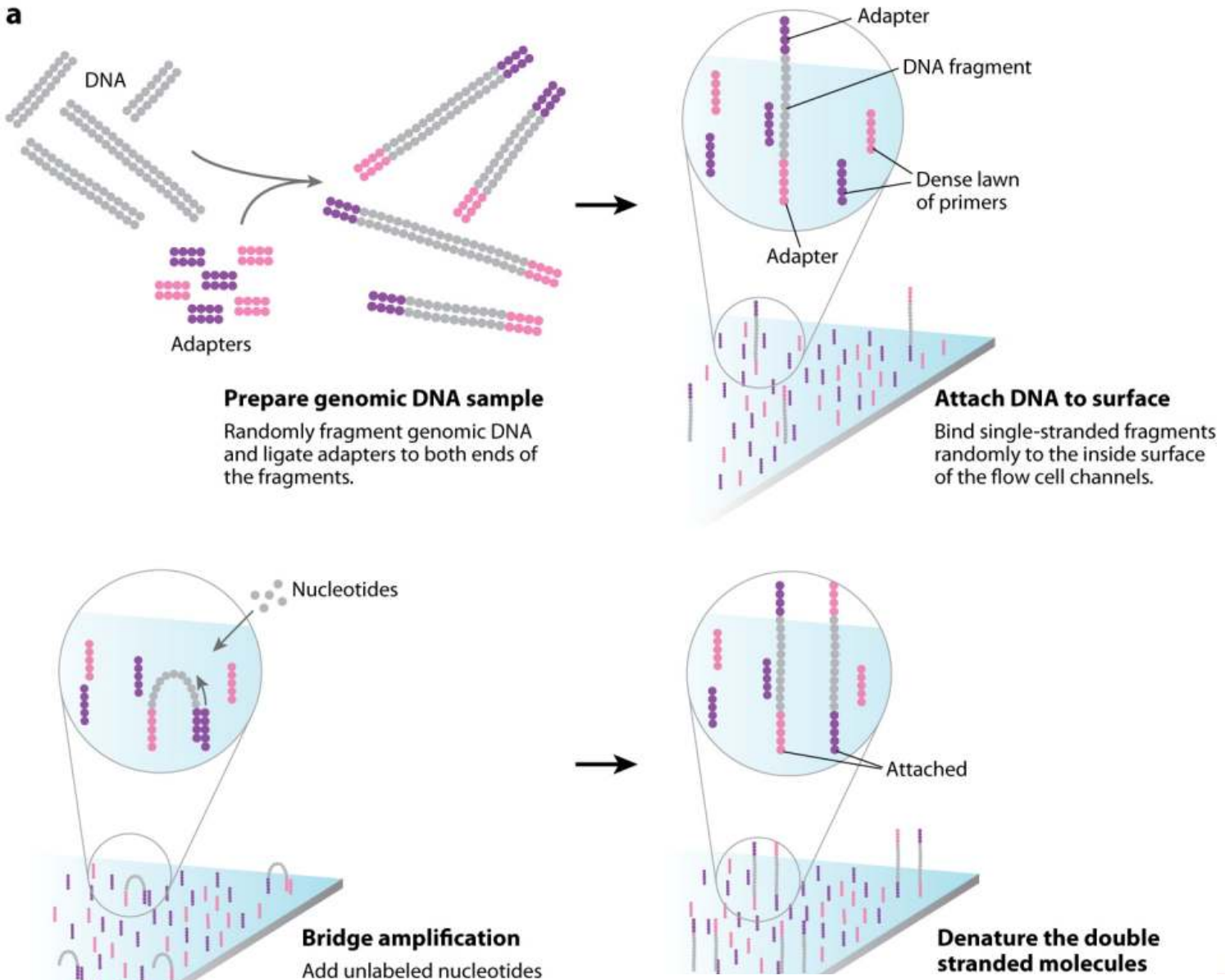
3. Detect the incorporation of each base during strand synthesis in each location through a light signal



4. Repeat nucleotide incorporation and detection cycles, thus permitting the simultaneous tracking of thousands to millions of sequencing reactions

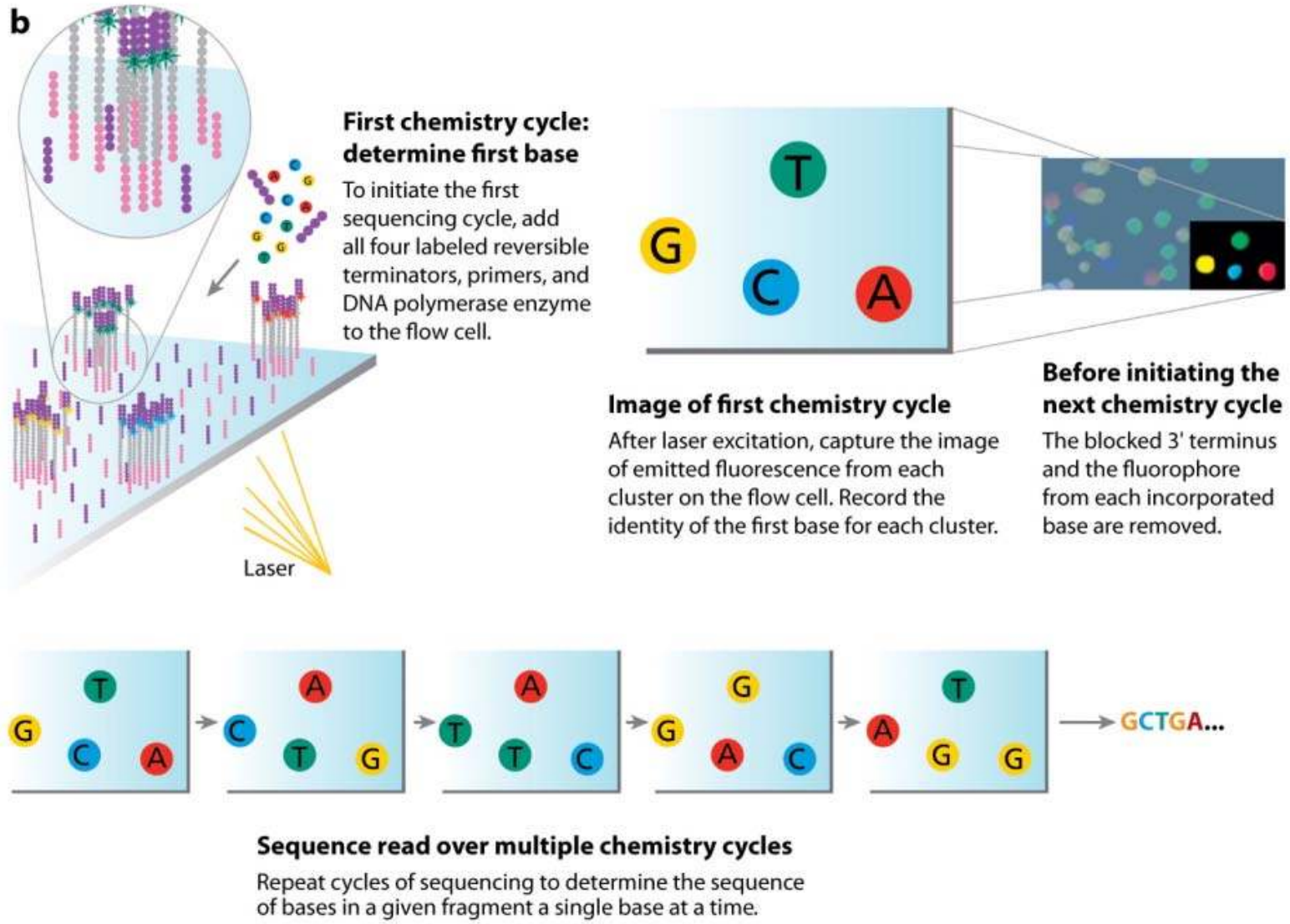
Rokas & Abbot (2009) Trends Ecol. Evol.

The Illumina Sequencing-By-Synthesis Approach



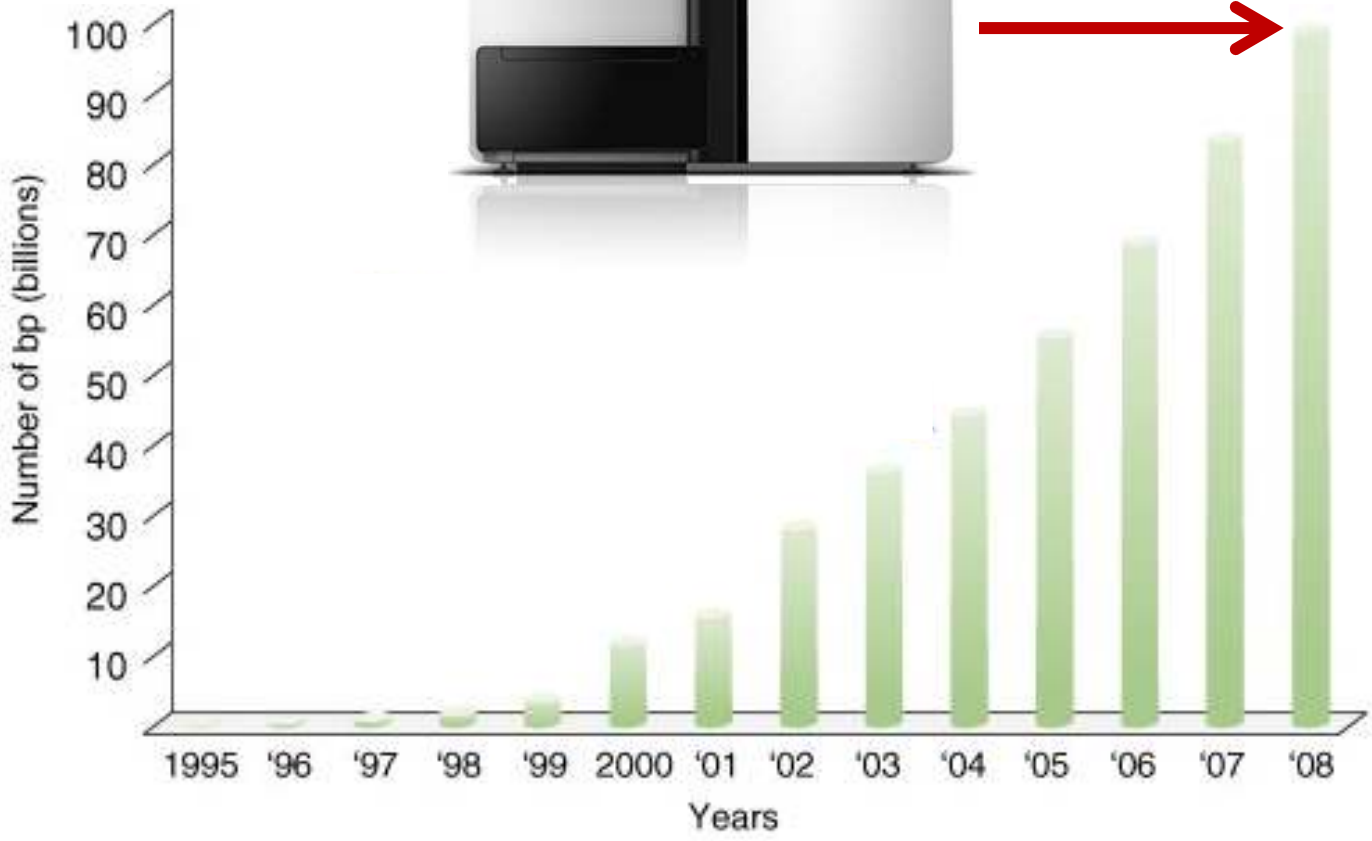
Mardis (2008) Annu. Rev. Genomics Hum. Genet.

The Illumina Sequencing-By-Synthesis Approach



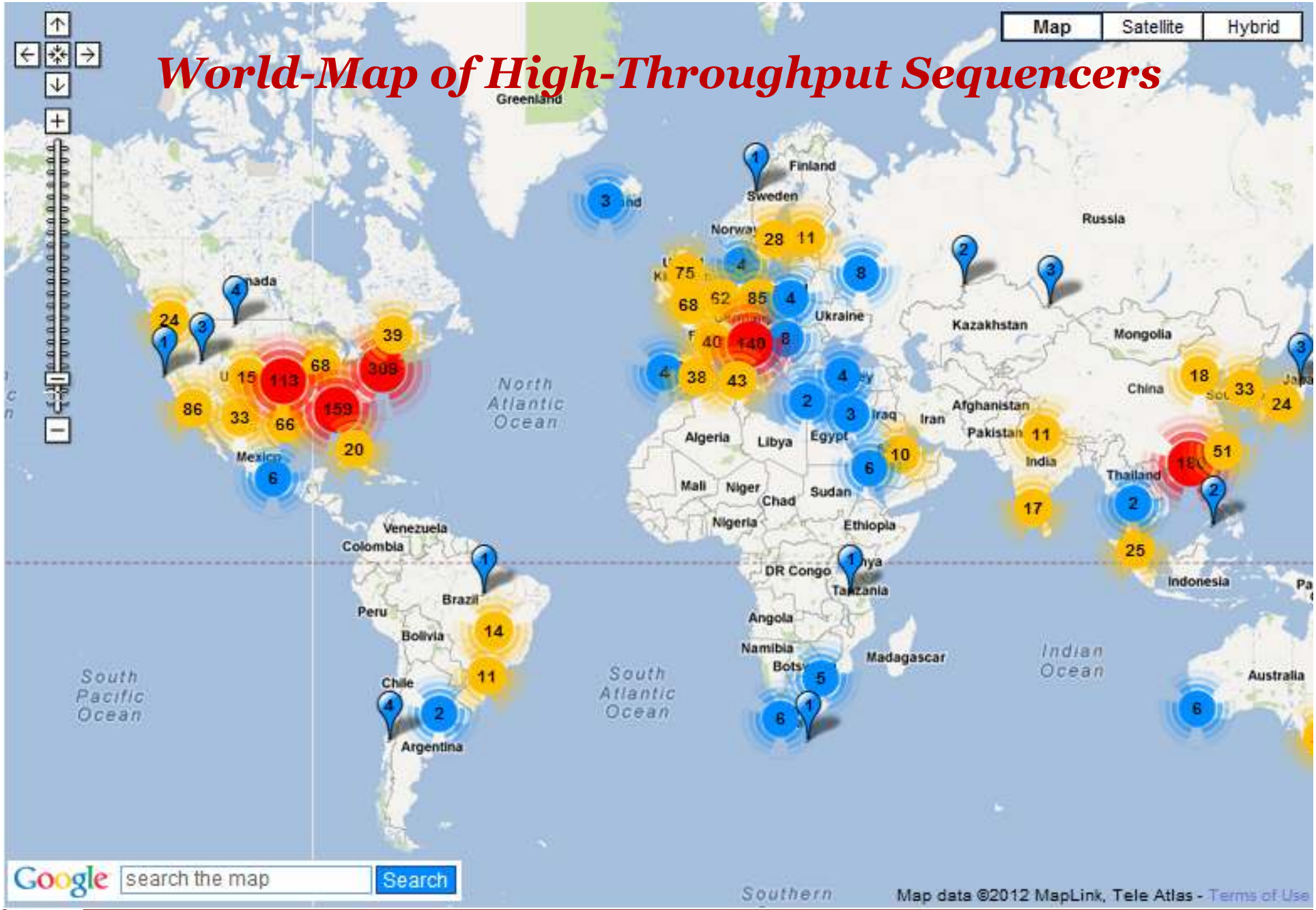
Why Is High-Throughput DNA Sequencing So Exciting?

4 days' output



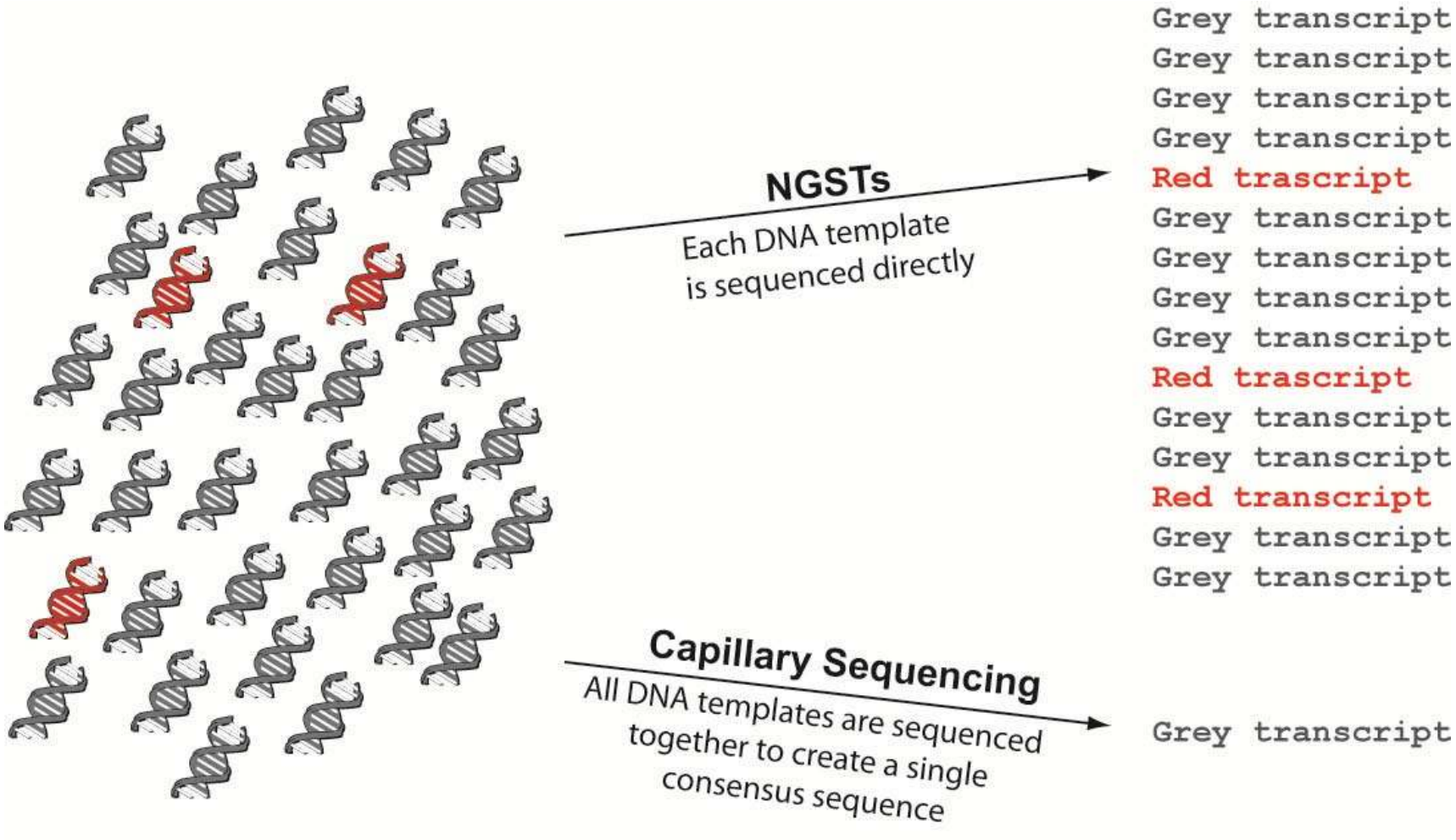
Gilad et al. (2009) Trends Genet.

World-Map of High-Throughput Sequencers



http://pathogenomics.bham.ac.uk/hts/

Next-Gen Sequencing is Qualitative and Quantitative



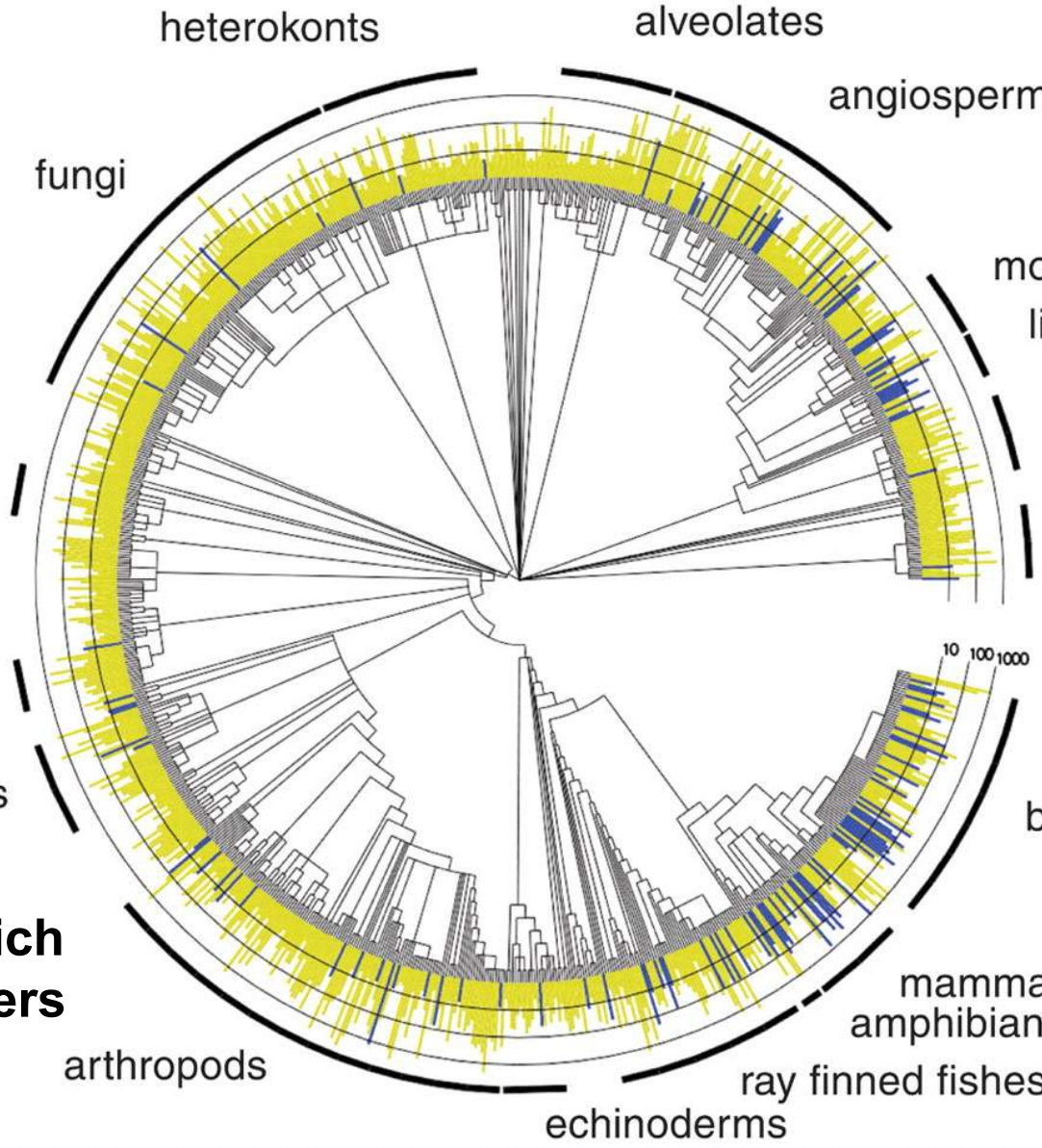
Rokas & Abbot (2009) Trends Ecol. Evol.

Estimating the Taxonomic Breadth of the Tree of Life

2002: Cracraft "guess-timates" that 0.4% of all known species have been included in at least one published phylogeny

2008: Sanderson reports that molecular sequence data have been sampled from 10% of all known species

The Genomic Depth of the Tree of Life



1/40 fungal orders

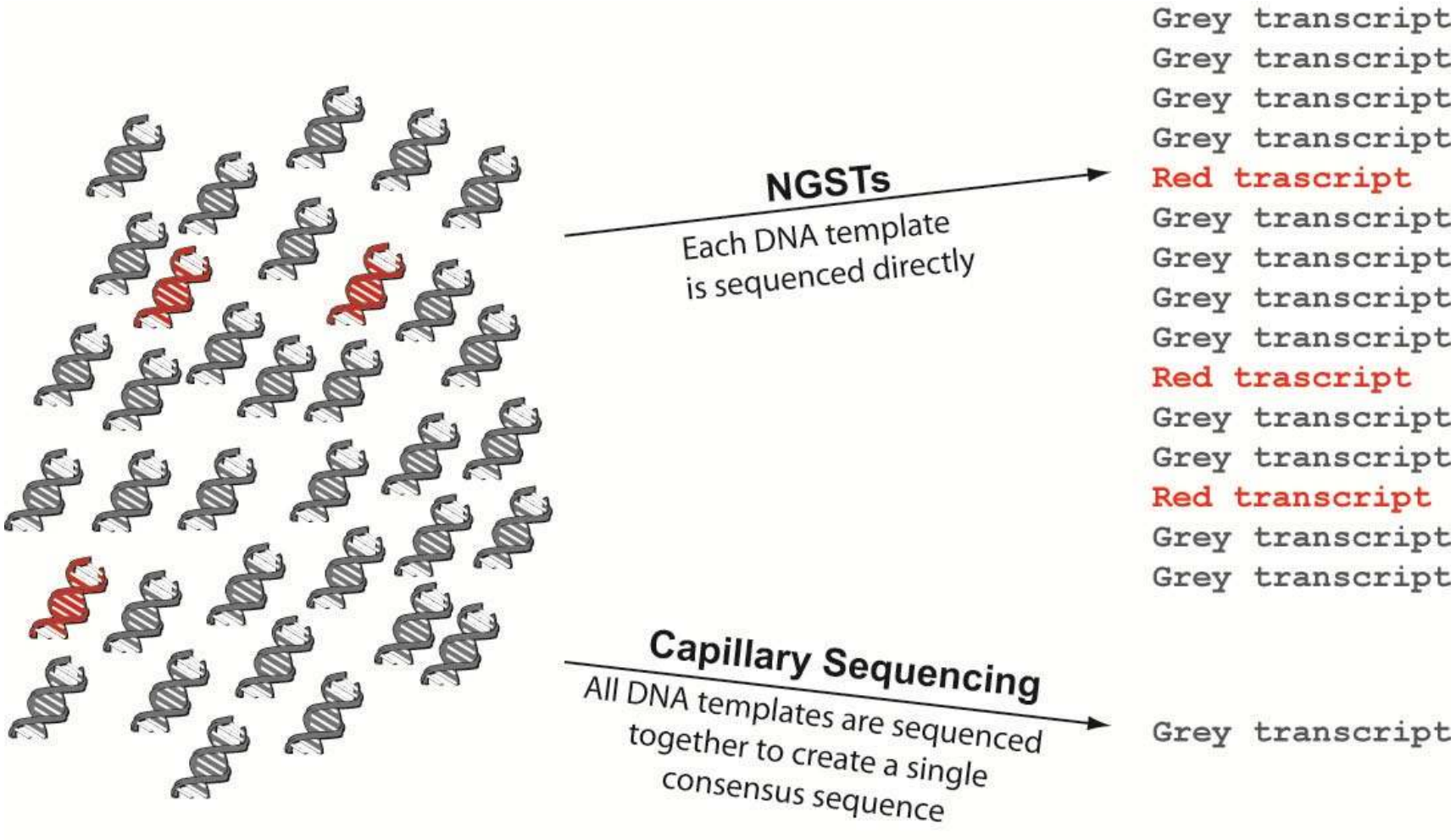
3/45 angiosperm orders

1/45 species-rich arthropod orders supported

16/57 vertebrate orders

Sanderson (2008) Science

Next-Gen Sequencing is Qualitative and Quantitative



Rokas & Abbot (2009) Trends Ecol. Evol.

Transcriptome Sequencing to Increase Genomic Depth

1. Smaller than genome (in *Anopheles gambiae* the transcriptome makes up 7% of the genome)

2. Fewer repetitive and transposable elements

3. Unequal representation (up to 6-7 orders of magnitude)

    Enriched for housekeeping and energy genes (they tend to be conserved)

4. The overwhelming majority of sequence evolution models have been developed for and tested in coding sequences

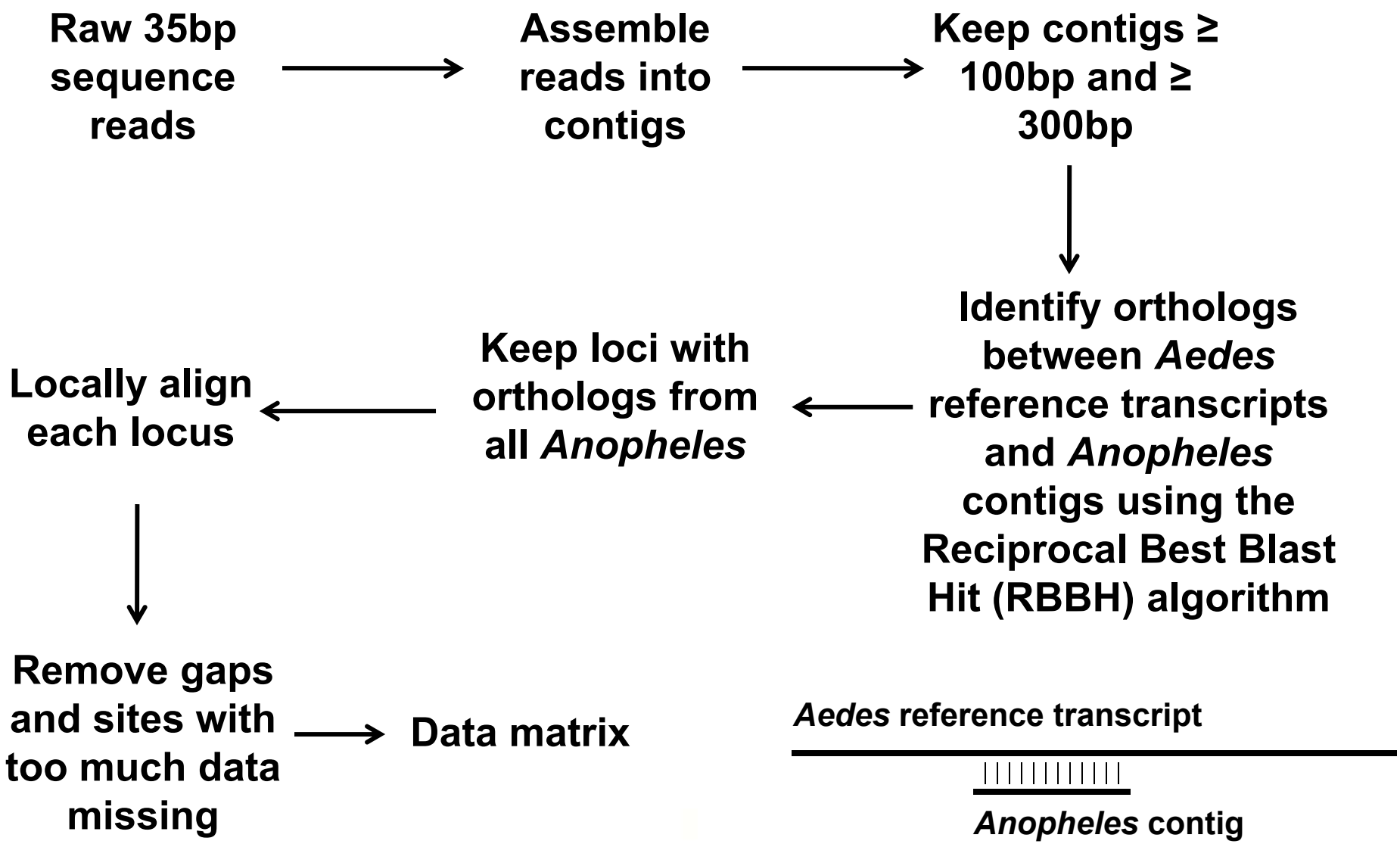
Hittinger et al. (2010) PNAS

Can we Use RNA-Seq to Increase Genomic Depth?

Species	Stock No.	Collection Location
Anopheles albimanus(Nyssorhynchus)	MRA-126	El Salvador
Anopheles arabiensis Cellia)	MRA-339	Zimbabwe
Anopheles dirus (Cellia)	MRA-700	Thailand
Anopheles farauti (Cellia)	MRA-489	Papua New Guinea
Anopheles freeborni (Anopheles)	MRA-130	USA
Anopheles gambiae (Cellia)	MRA-765	Liberia
Anopheles quadriannulatus (Cellia)	MRA-761	South Africa
Anopheles quadrimaculatus (Anopheles)	MRA-139	USA
Anopheles stephensi (Cellia)	MRA-128	India
Aedes aegypti (Stegomyia)	MRA-735	West Africa

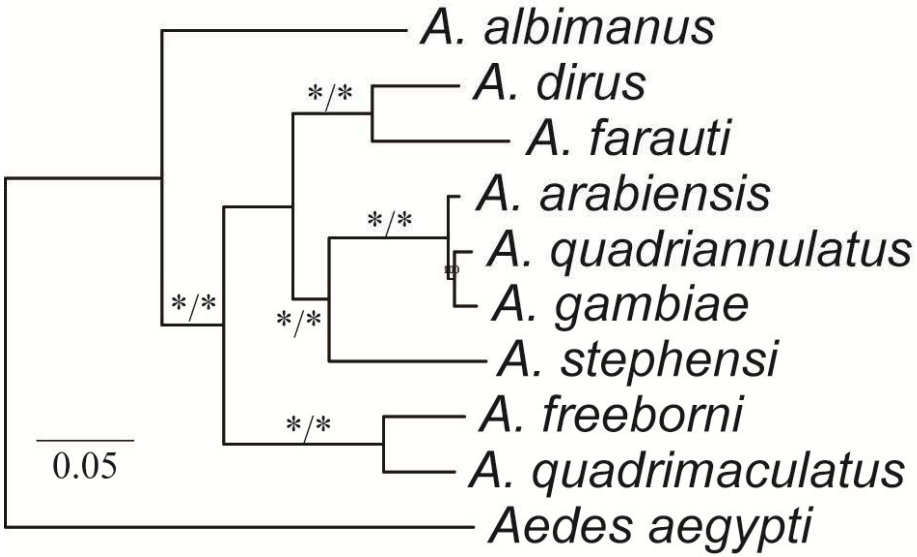
Illumina RNA-Seq → **~150,000,000 reads 5,250,000,000 bp**

Data Matrix Construction: The "Singlecontig" Strategy



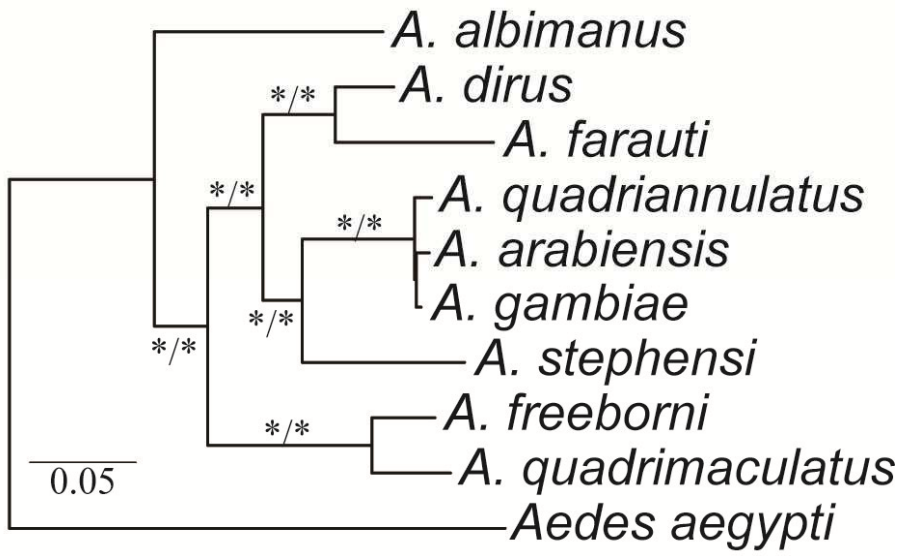
Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100bp contigs



 # Loci = 553
Aln Length = ~390 Kb
% Missing data = 51

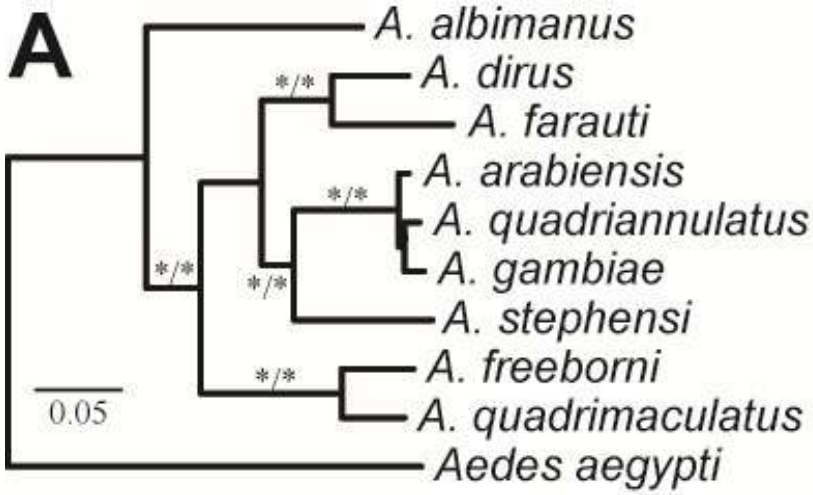
Using ≥ 300bp contigs

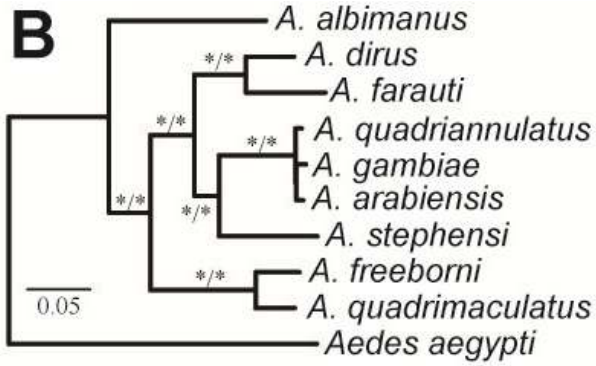


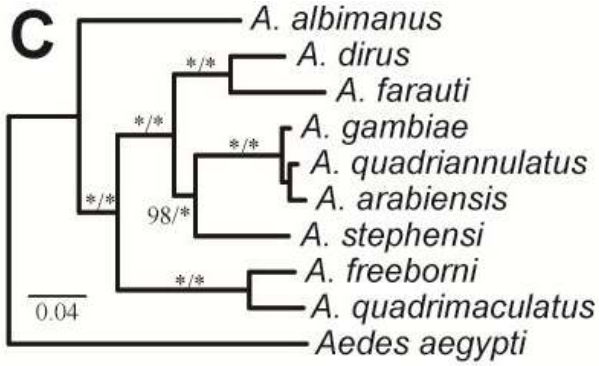
 # Loci = 69
Aln Length = ~73 Kb
% Missing data = 44

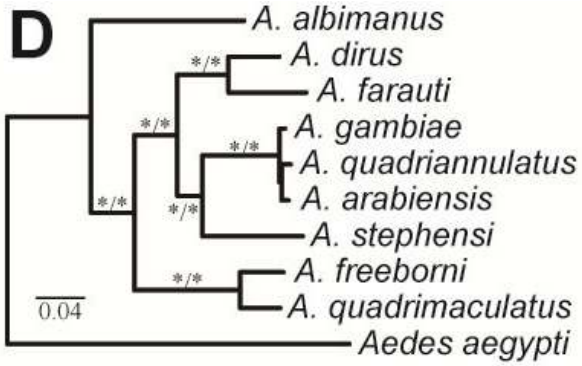
Accurate Phylogenetic Inference From our Data

553 loci
Aln L: ~390 Kb
Missing data: 51%









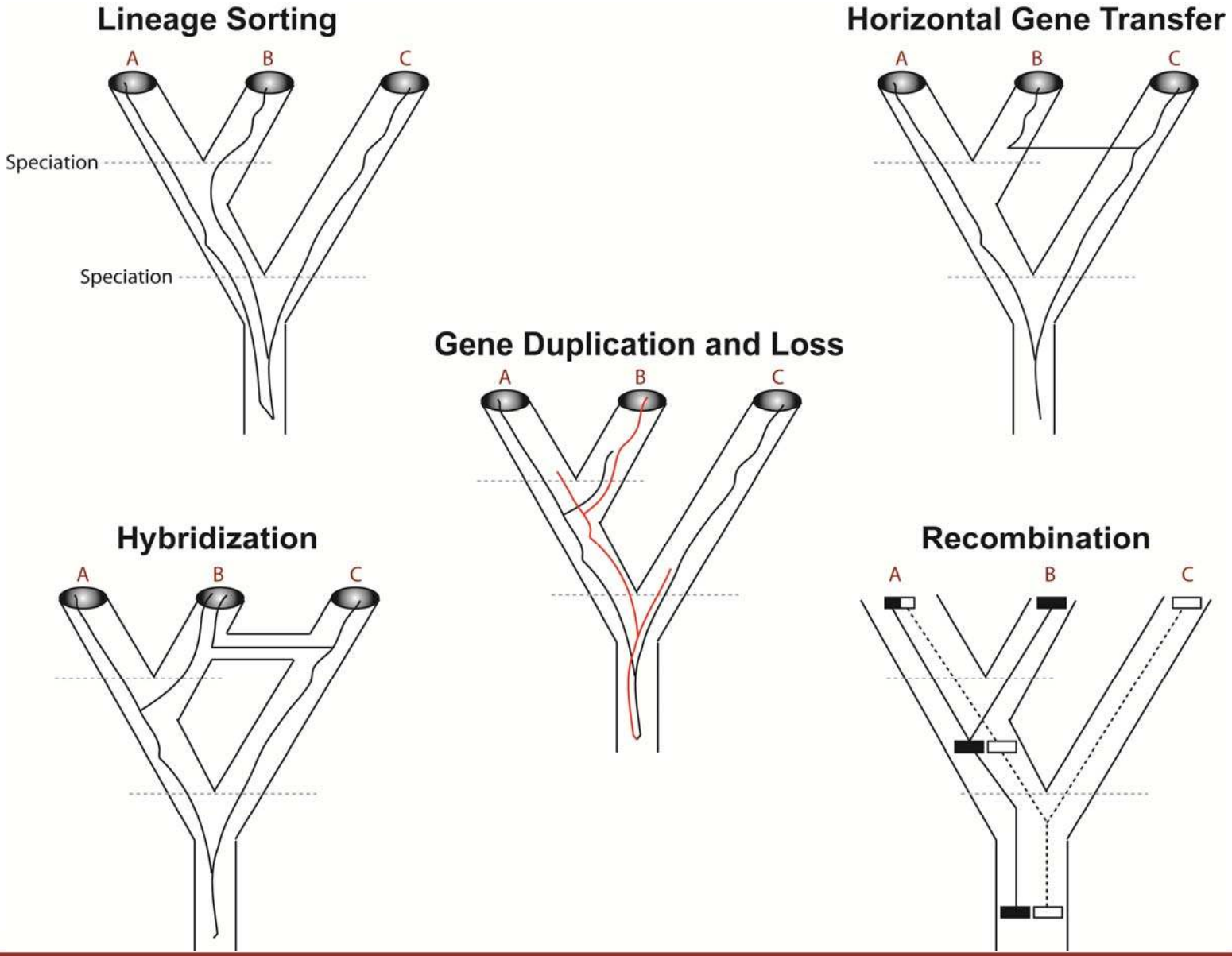
<u>Exclude erroneous loci</u>
491 loci
Aln L: ~329 Kb
Missing data: 50%

<u>Use only sites without data missing</u>
Aln L: ~15 Kb
Missing data: 0%

<u>Use *A. gambiae* as ref.</u>
634 loci
Aln L: ~472 Kb
Missing data: 50%

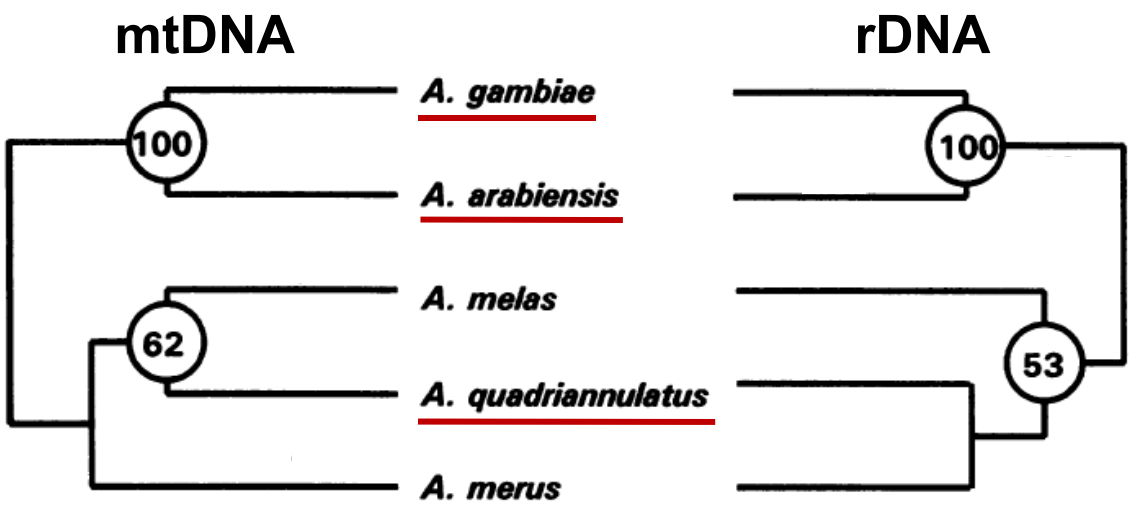
Hittinger et al. (2010) PNAS

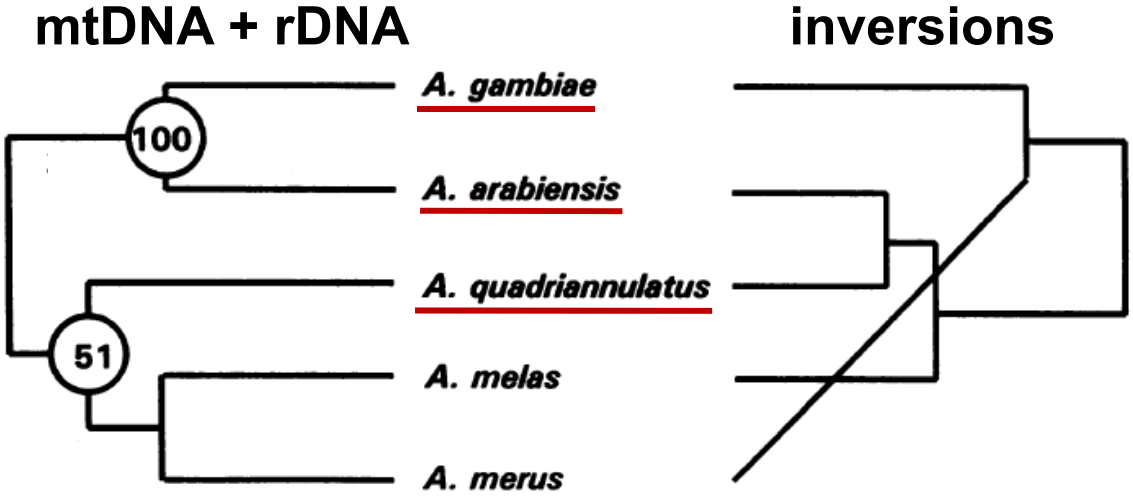
Gene Trees Can Differ from Species Trees

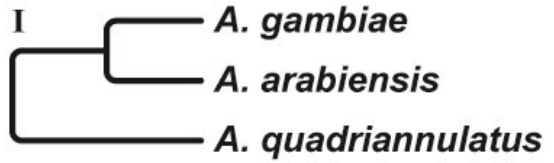


Degnan & Rosenberg (2009) Trends Ecol. Evol.

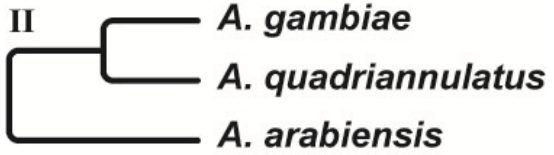
Our Data Matrices Can Detect Population-Level Events



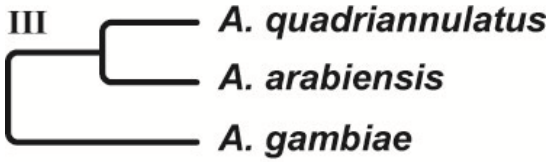




10/237



12/237

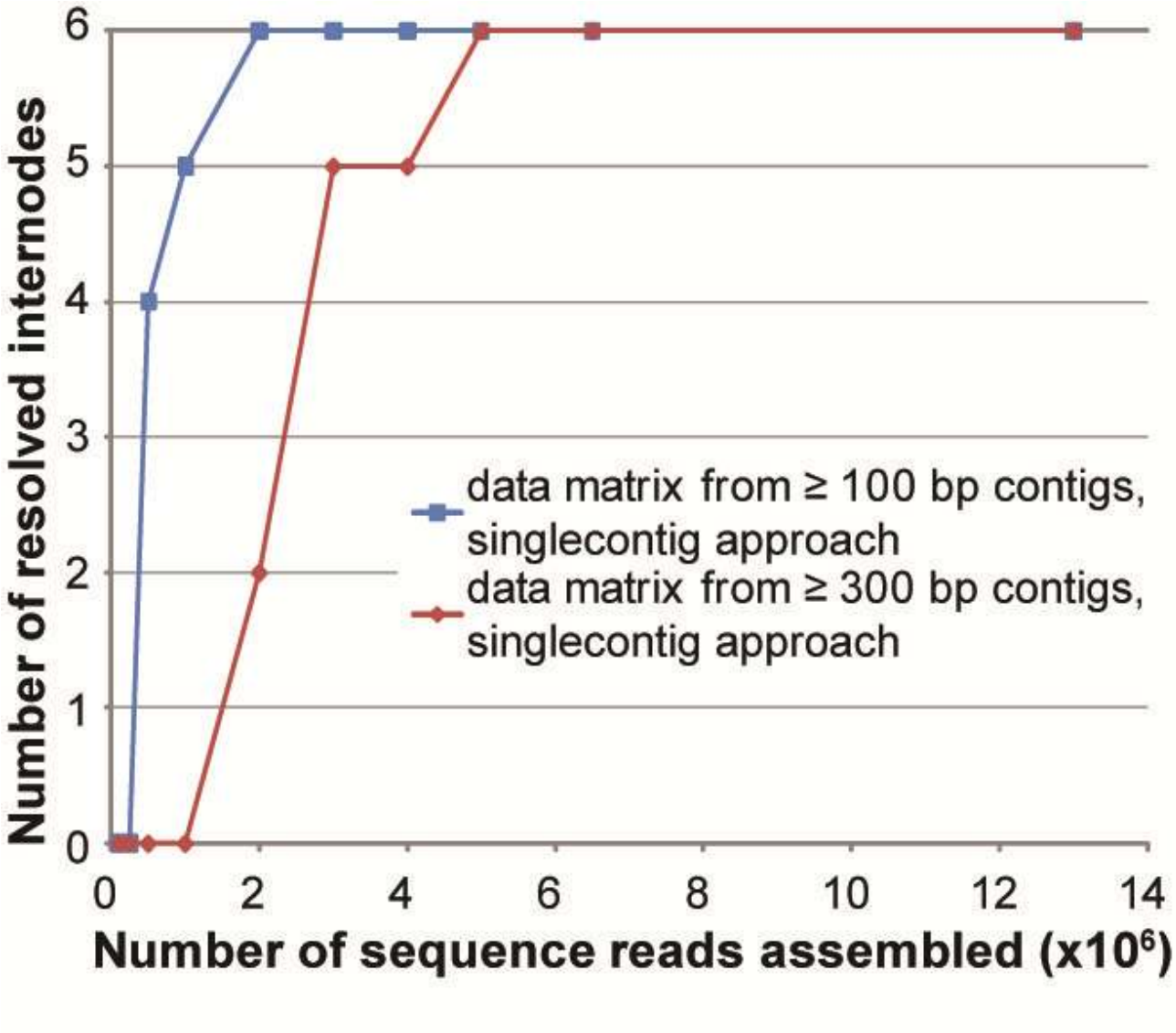


10/237

Besansky et al. (1994) PNAS

Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Few Sequence Reads



Hittinger et al. (2010) PNAS

Experimental Design: The "Supercontig" Strategy

Raw 35bp sequence reads → Assemble reads into contigs → Keep contigs ≥ 100bp and ≥ 300bp

↓

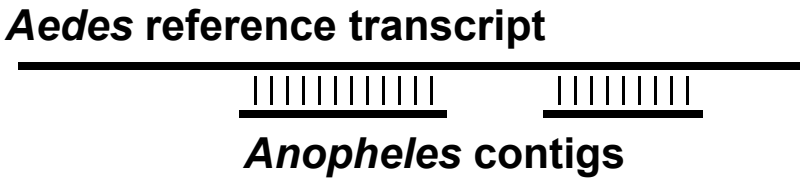
Identify orthologs between *Aedes* reference transcripts and *Anopheles* contigs using a locally relaxed RBBH algorithm

← Keep only loci with orthologs from ≥ 4 *Anopheles*

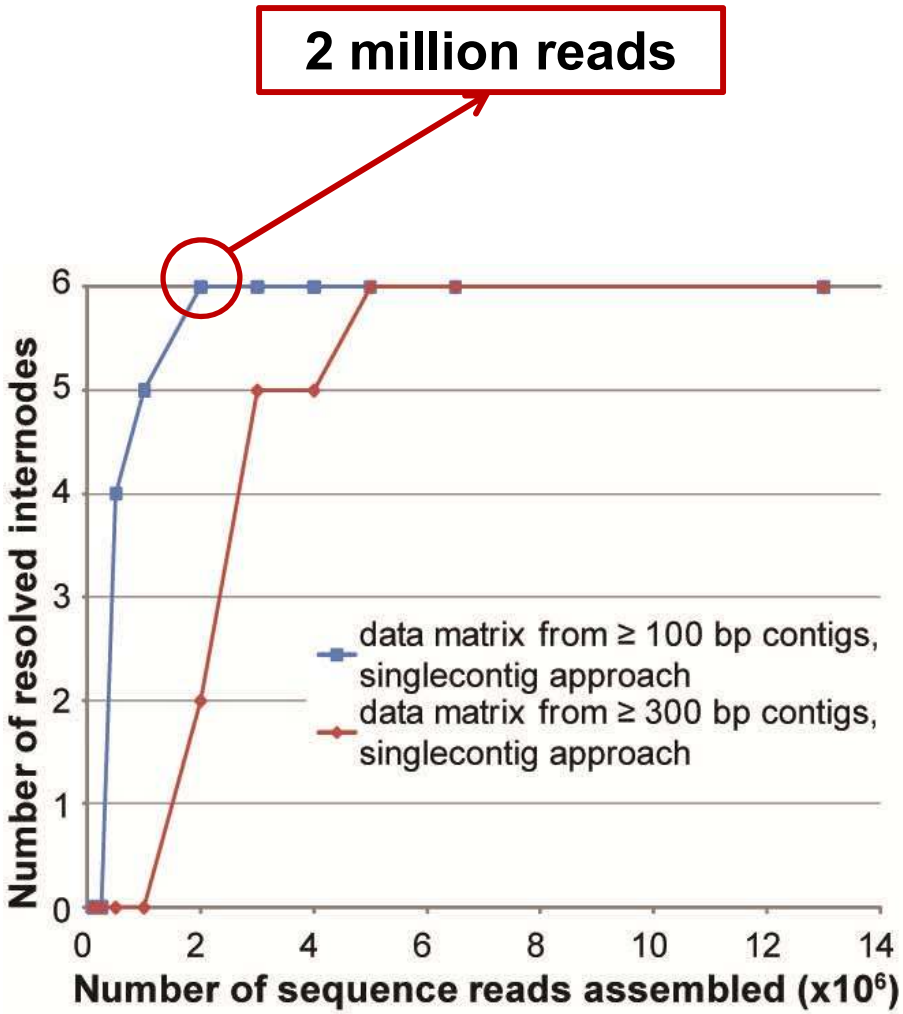
← Locally align each locus

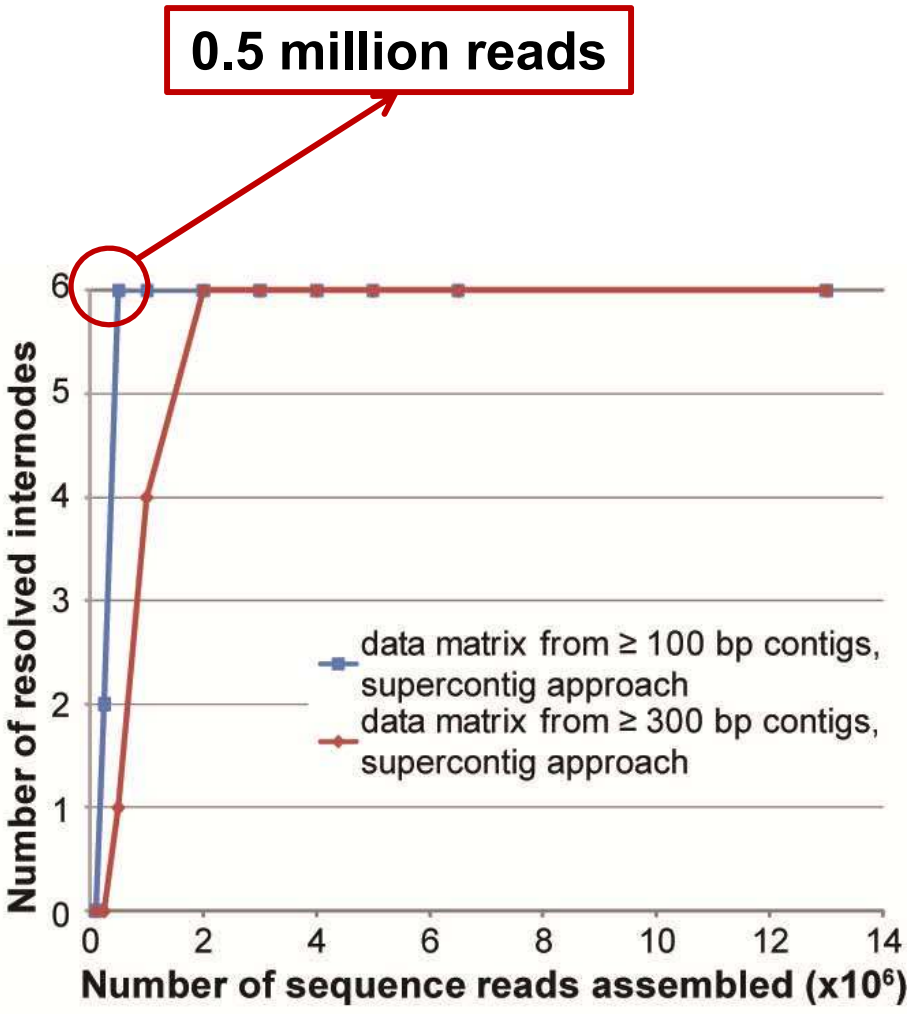
↓

Remove gaps and sites with too much data missing → Data matrix



Robust Phylogenetic Inference From Very Few Reads





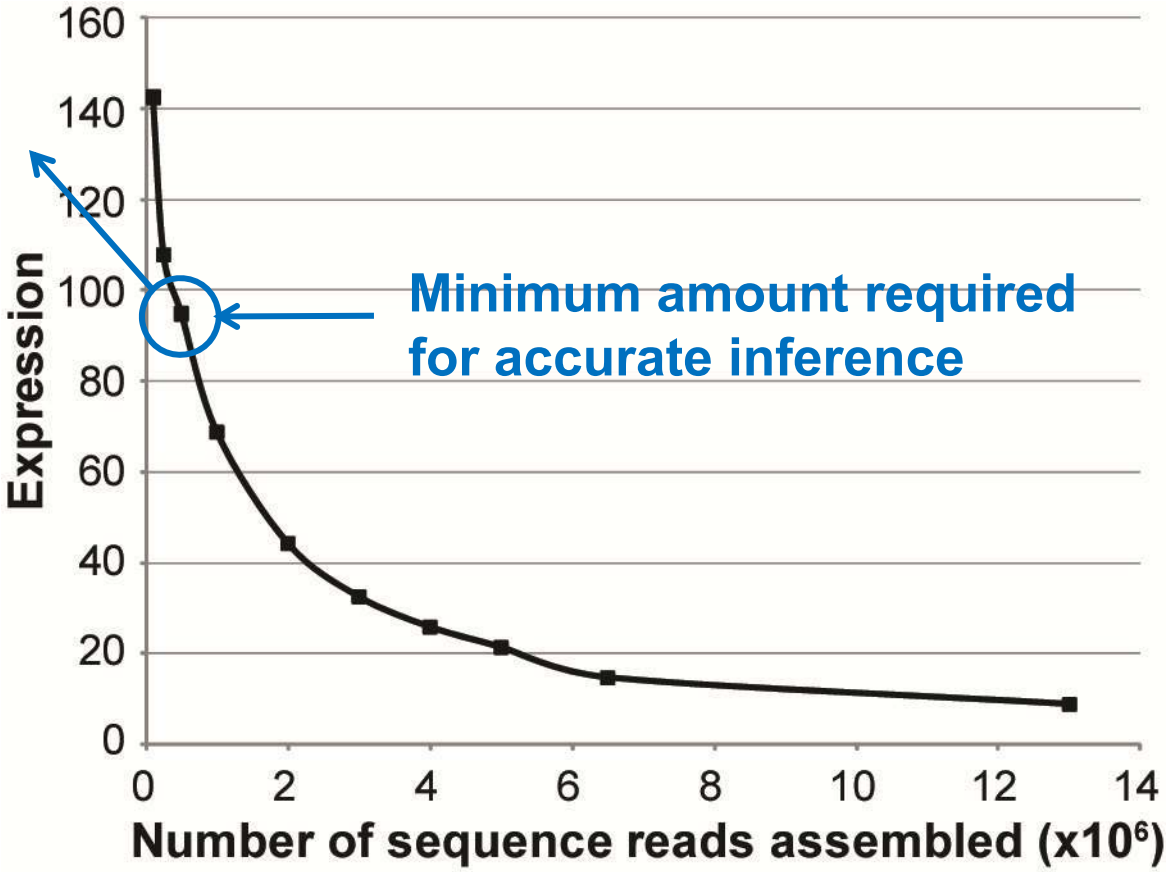
553 loci, AlnL: ~390Kb, %miss: 51

2,661 loci, AlnL: ~971Kb, %miss: 62

Hittinger et al. (2010) PNAS

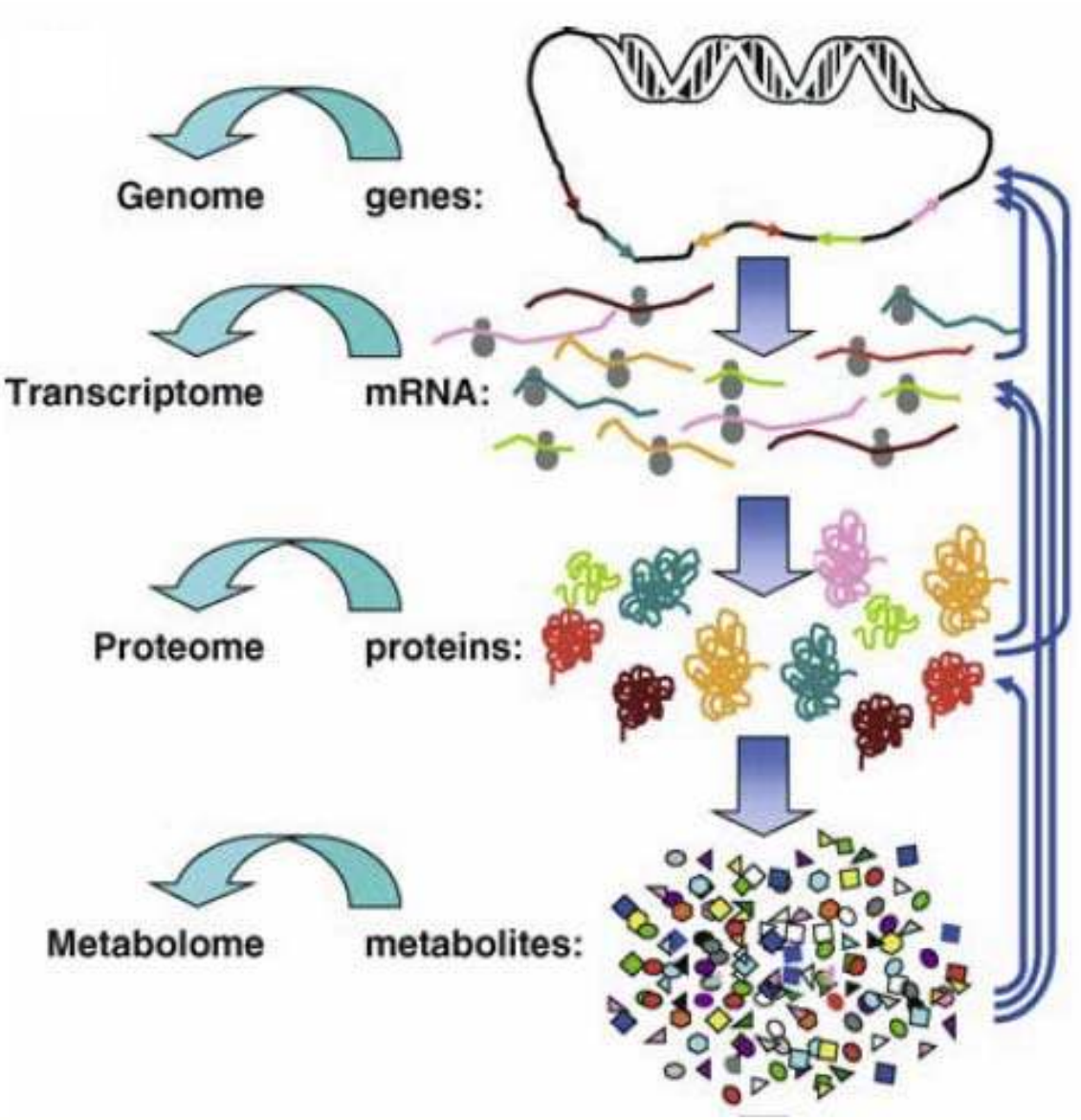
Our Sequences are from Highly-Expressed Transcripts

2008 cost: ~$50
2013 cost: ~$5



Hittinger et al. (2010) PNAS

The Age of High Throughput Technologies



Goodacre (2005) Metabolomics

The Genomes of Non-Model Organisms are the New Frontiers

Rokas & Abbot (2009) Trends Ecol. Evol.