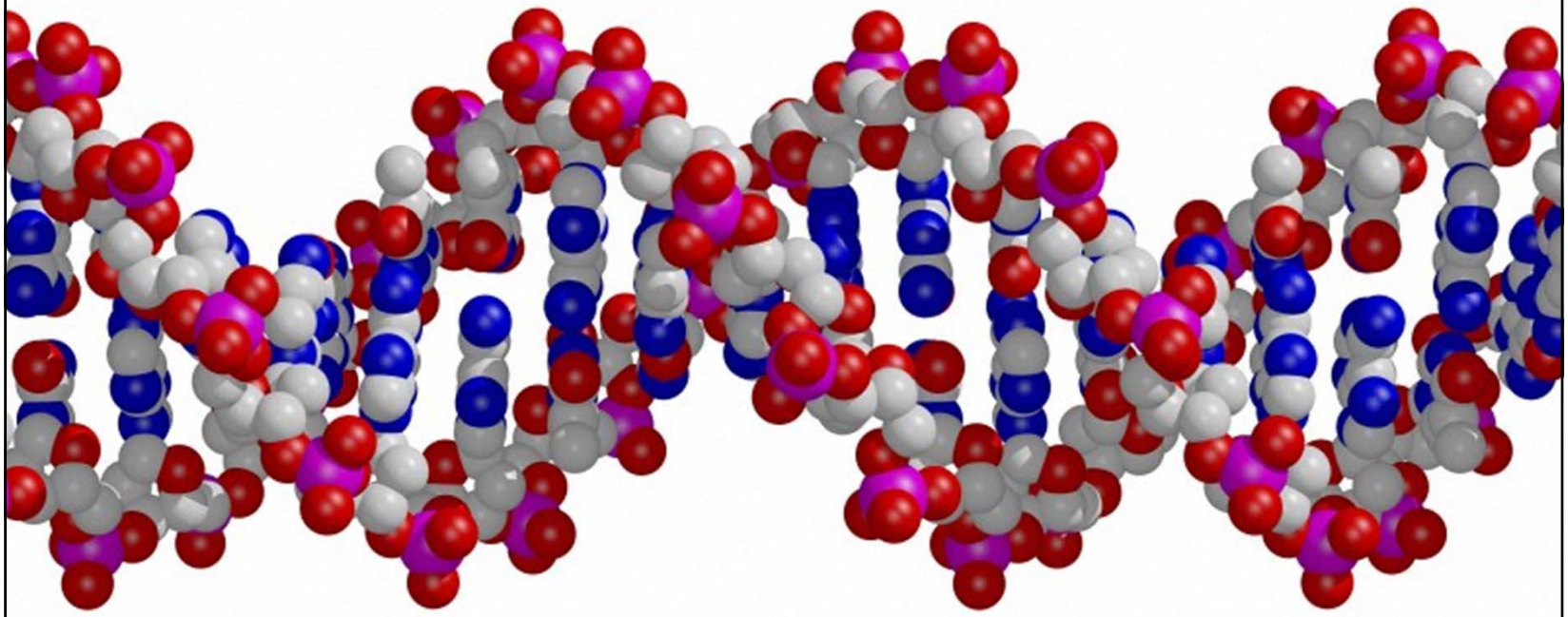
Evolutionary Genomics



Antonis Rokas
Department of Biological Sciences
Vanderbilt University

http://people.vanderbilt.edu/~antonis.rokas/

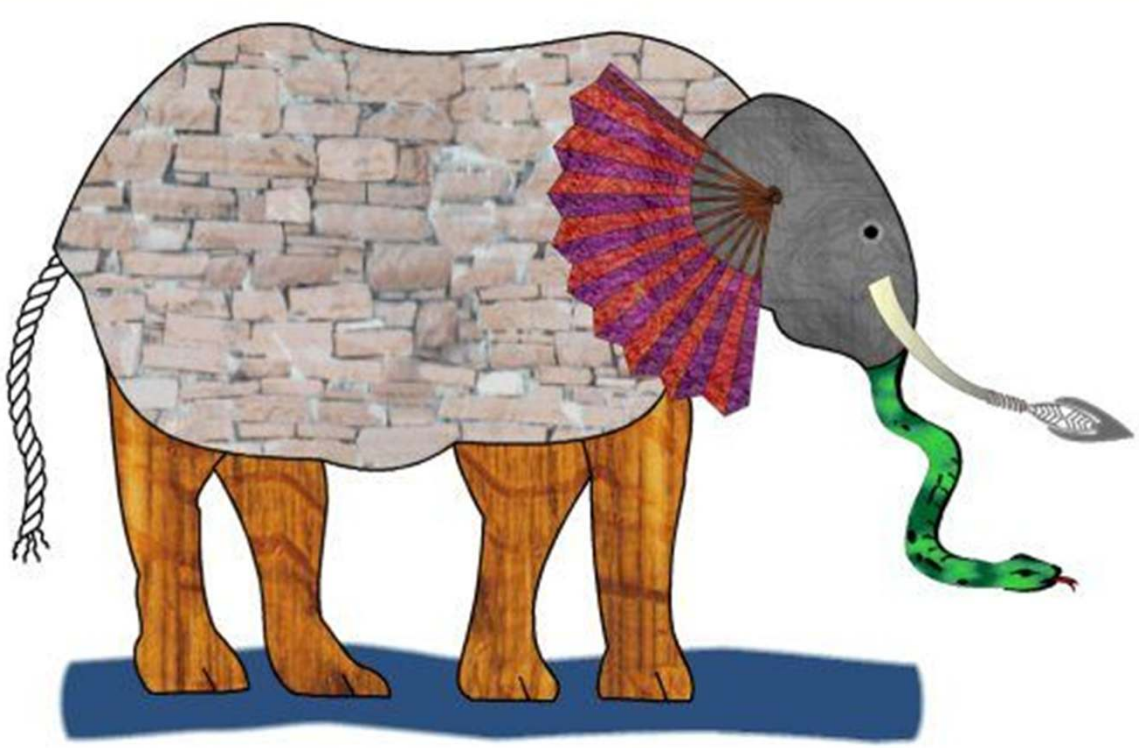
Lecture Outline

- **Introduction to evolutionary genomics**
- **Brief review of next-generation DNA sequencing & its applications**
- **Benchmarking transcriptome sequencing for functional and evolutionary genomics**
- **Transcriptome sequencing for molecular phylogenetics**

----------Coffee Break----------

- **Genomics within species (using next-generation DNA sequencing)**
- **Genomics between species (using publicly available genome data)**

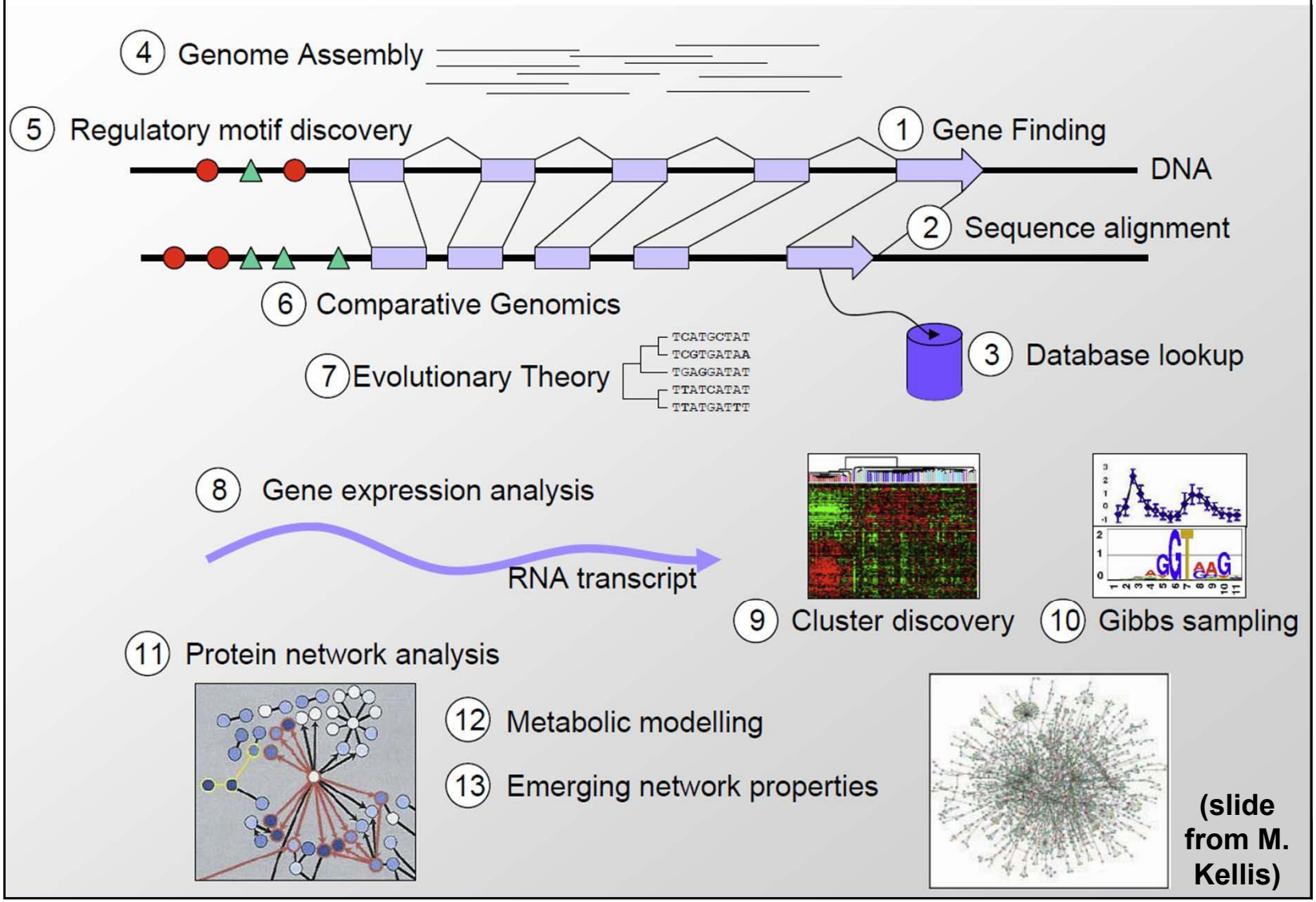
What is an Elephant Like?



What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

Understanding Sequences Requires Tools and Evolution

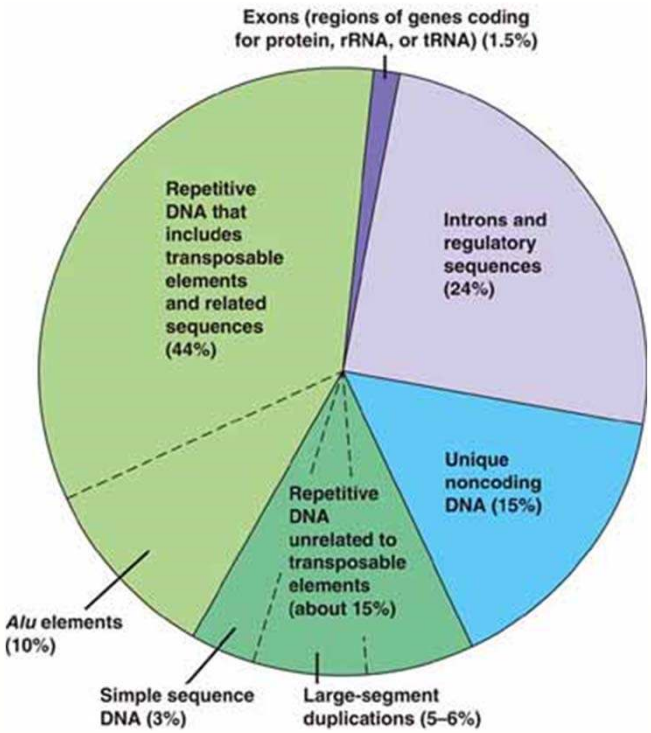


What is a Genome Like?

ACAACCCCTCCACCTCATGTACCTGCGGACTCTCCTCCAGTCACAGCTCAGGCAGTCCACTTTGCAACCCCTAAACCTCAAAACCGGTTT
GACGTTCTGTTAGACGAACAACTATGATATATCGACCCCGCCTAAGAACGGAGCCTCTGTCAGTGCTCCAGCTGAACGTAGGCCGCGGG
CCAGCCACTCATGAAATCGCCCTTTCATTAGCATACTCTAGCGGCATTGATATCATCCTTATACAGGAGCCATACATATATACTGACCTCA
GCCGGCAAATCACAAAAAGGCACCCATCATACGAGTGCTTCTCCCCAACAGACAGCTGGCTTGTAAGCGGTCGACCCCGGGTCCTCACC
TATGTCCGGAAAAAGATGGGCATTCGGGCCTCTCAGCTCCGCCCTCAGCCAATAGATCAAGATGTTCTCTCAGACCTTCTTCTACTACAG
ATCCTCTCCCGCTCTGGACAATCTGCATTGATAATCAACATCTATAATGCTCCAATCGGCTCAATCAGGTCAGGTGAGGCTGCAAAAGCG
CTTACACTCCTGCCTGACTCCTACTTTTCCCAGCCTACCGTGCTTGCCGGCGACTTCAACCTACTACATAGCAGGTGGCAGCCATCACTG
CATTGCAGCCCTACCACCTTTGCTGAGCCATTTGTTGACTGGCTTGATCGCCTAGGGCTGGTTCTTATCTCCGAGATAGACCAGCCTACAC
ACGATAGAGGCAACGTTCTTGACCTCACTTTCGCCTCCAGCTCCCTAGCACTGGCAGGGTCGAGTACCAGGATAGCAAGTCATTTAGAGT
CAACATCAGATCATCGGCCACTCCTCACCACCATGCCATGGAGCCAGAGATTCACAGAGGCAGCTCAGAAACTGAGATTTGATACATTA
GACCACCCTCGCTTCCTCTCACTACTCAGTTCCCACCTTGCTGTCATTGAATGCTCAGCTACAACAGAAGAGGGCCTGGACAGTCTAGCT
CATGGGTTAACCTTAGCAACTGCTAGTGCGTATAAAGGCTCTGCTAGGAGCTCCTTGGCGCAGGGAATAGGTCAGCCATGGTGGAATATT
GACTGCAGAAAAGCGTTGCAAGACTTCCGCTTAGGTCTCTGTTCAAGAAACGACTTCCGTCGGATAACTAGACGGTCTAAATAGCAGTTC
TGGCGAGATAAACTTACCGCAGTGACACAGATCAAAGATGTCTTTGACATAAGCAAGTGACATAAGTTTACAGGATCTTATCGAAACCCT
CCACTAAACGACCCTTTAAGGCCAAACAGCCCTCCAGCAGGGGCTCTGAATGAGAAACAAGACGTATTAGTCCGTAATCTTCTTCAGAAT
ACTGCTGAAGCGGGTGATATTGTCATAGGCTATGGCCTGGGCTGTGGTTGTCAGCCATGCCCTCAACCATAGAACATTCTAGAAGAACCA
TCGGGAAGAGGTTGGAACCCAGTGGAAGTTTGGGAACATGTATATAAGAAGGAGAGGGAGATGTATCTGCCTATTTCTCTCTCCAAGTCT
GCGATATTCGTTAATACATTATACAGGATTGCCAGTTGAAAACAATACTGCCTACGCCCGTCACAGGTACTGCAGTTTCCAACAAGAATC
AACGCTCGACCCGGCAATTATGGCTCAAGGTTAGACTACGTCCTGTGTAGCCTTGATATGCAAGATTAGTTCTGCGATTTGAATATCTAAG
AGGATCTAATGGTAAGCCCCAAGGCTGCCATGGCTTTATTGTAGATTGATTTTCTAGCTGACAATATGCAATTTGGGACAGGGATCTGATG
ATTGTCCGGTTTATGCTGTCTTCAAAAATGTTATACGCCTCGGCGAAGAAGAGGTCAACATTAAATGAGCCCTCCTGGGATGTTTAAAGAT
GGCGAGCGTCAGCAGGAATACTCTACTAAATATCTTCTGCCTACATCAGGGCGCTTAATACCAGAATTTAACAAGCGGAGGAGGATCAA
GGACATGTTCTTGCGTAAACCATCAGCCAACGTATAGAGACCGACGACGAACATCCTGACATTGAGATATTTTACCTCTAGTCAGGAAAA
GGGAACAGCACCCGCTATTTTGGAGAGTGCTGCCAGCGTCATAGCTACCTGCCAGCCTGTAGTAGCTGCTGACAGCACTCAAATGAAAG
AAGTTATTCGTAAGAGCTCTCAGAAATATGAGACAGGTTCCCCTGTCTCAGTCCAGTATTTGACATCGGGTTCAGCCCAATCATCAACAC
CCCCCACTGCTGGACAGAGGACTCTAAAGGGGTTCTTCAAACTTAAAAGTGGTCTAGCCAGCCAAATGGCCATAGCCCAGGATCCTGCA
ACAGTGTCTACTATGCCAACGAAACAACCAGCCGCATCCCCTACAAAATCTACCCCAGTTACAGAACCTCCTGCACTGGAAGCATTACTG
ACAGCTCCCGCTGGTGAAGCTTCTCCAGGAGAACAGCCAAATTCCGCGACTCCTACAGCTCCCGCTTCACCCCAAAGCAATGATACTATT
ATCGATCCCATTGTCAGCAAGGAAGATTGGTCAAAGCTCTTCACTAAAAAGCCCATTCCCAAGTGCGAGGGCCACCAGGAACCATGTTT
CAGTCTGACAACTAAGAAGCCTGGCATCAACTGCGGAAGATCGTTCTGGATCTGTTTGAGACCCCTTGGGCCCAGCGGAAACAAGGAAA
AGGGGATACAGTGGCGATTTCCTACATTCATATGGGCCAGCGATTGGAACCCTTCCGCTCCGTAGATTTTCTGTCTGGGGCAACTTCTTTT
TGCGATAGTGTAACGATACCCGGTTTTATACTTAGAAGGCTACGAATGGTATGATGTATCATGGTTTCAATGATAAGACATTTCGTCAAGT

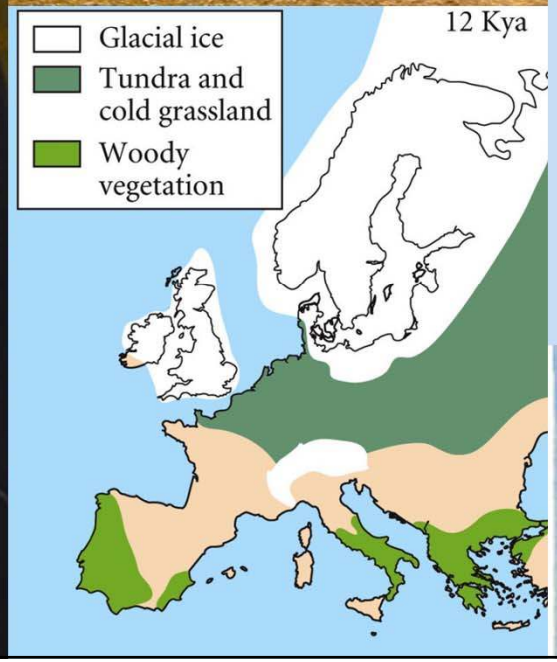
Transposon Protein Binding Site Exon Intron

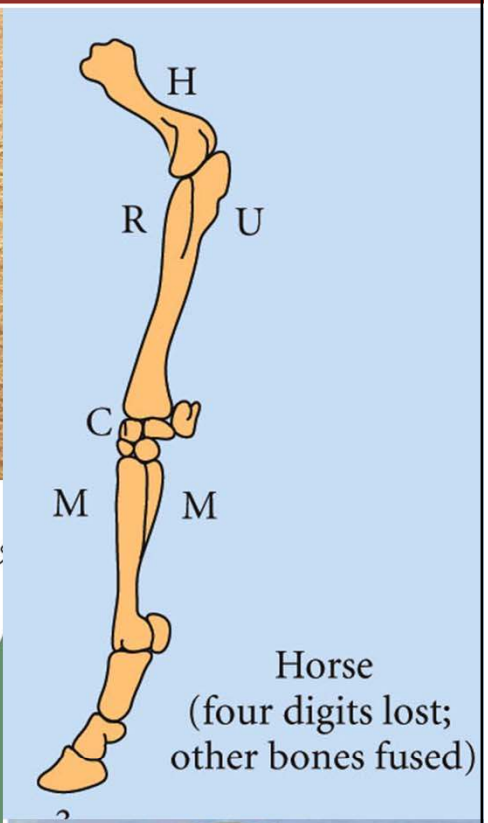
Organization of the Human Genome

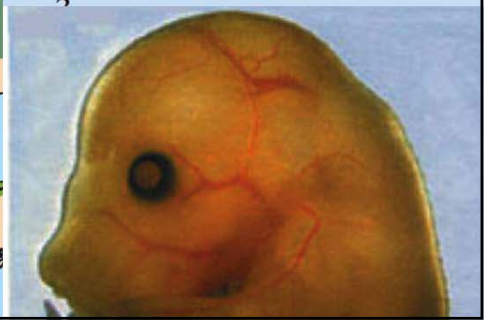


http://www.bio.miami.edu/~cmallery/150/gene/c7.19.14.human.genome.jpg

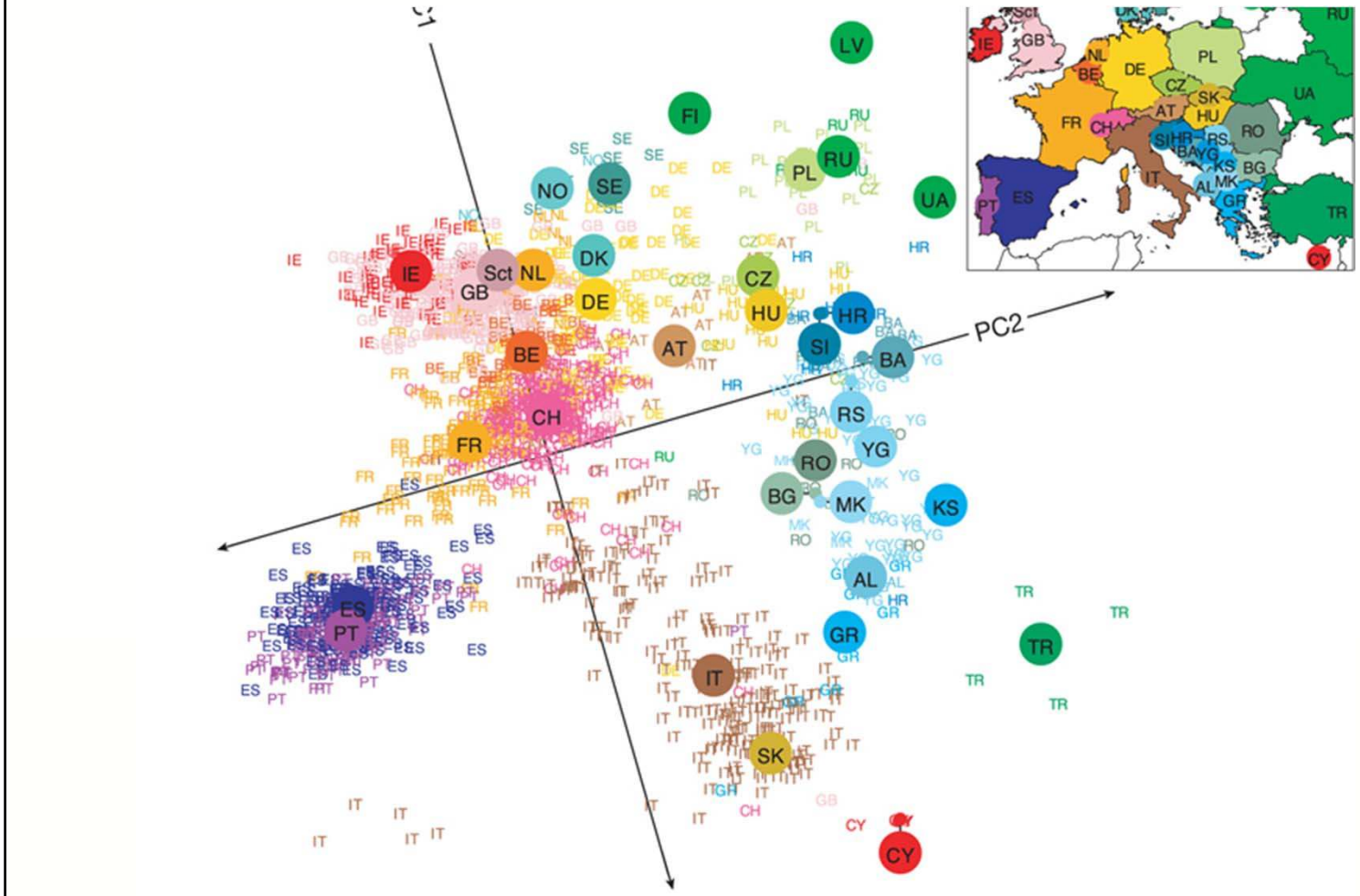
The DNA Record







Human Genes Mirror Geography



Novembre et al. (2008) Nature

Recent Positive Selection in Human Populations

in the Asian Population, involved in hair follicle development

The twenty-two strongest candidates for natural selection

Chr:position (MB, HG17)	Selected population	Long Haplotype Test	Size (Mb)	Total SNPs with Long Haplotype Signal	Subset of SNPs that fulfil criteria 1	Subset of SNPs that fulfil criteria 1 and 2	Subset of SNPs that fulfil criteria 1, 2 and 3	Genes at or near SNPs that fulfil all three criteria
chr1:166	CHB + JPT	LRH, iHS	0.4	92	39	30	2	*BLZF1, SLC19A2*
chr2:72.6	CHB + JPT	XP-EHH	0.8	732	250	0	0	
chr2:108.7	CHB + JPT	LRH, iHS, XP-EHH	1.0	972	265	7	1	*EDAR*
chr2:136.1	CEU	LRH, iHS, XP-EHH	2.4	1,213	282	24	3	*RAB3GAP1, R3HDM1, LCT*
chr2:177.9	CEU,CHB + JPT	LRH, iHS, XP-EHH	1.2	1,388	399	79	9	*PDE11A*
chr4:33.9	CEU,YRI, CHB + JPT	LRH, iHS	1.7	413	161	33	0	
chr4:42	CHB + JPT	LRH, iHS, XP-EHH	0.3	249	94	65	6	*SLC30A9*
chr4:159	CHB + JPT	LRH, iHS, XP-EHH	0.3	233	67	34	1	
chr10:3	CEU	LRH, iHS, XP-EHH	0.3	179	63	16	1	
chr10:22.7	CEU, CHB + JPT	XP-EHH	0.3	254	93	0	0	
chr10:55.7	CHB+JPT	LRH, iHS, XP-EHH	0.4	735	221	5	2	*PCDH15*
chr12:78.3	YRI	LRH, iHS	0.8	151	91	25	0	
chr15:46.4	CEU	XP-EHH	0.6	867	233	5	1	*SLC24A5*
chr15:61.8	CHB + JPT	XP-EHH	0.2	252	73	40	6	*HERC1*
chr16:64.3	CHB + JPT	XP-EHH	0.4	484	137	2	0	
chr16:74.3	CHB + JPT, YRI	LRH, iHS	0.6	55	35	28	3	*CHST5, ADAT1, KARS*
chr17:53.3	CHB + JPT	XP-EHH	0.2	143	41	0	0	
chr17:56.4	CEU	XP-EHH	0.4	290	98	26	3	*BCAS3*
chr19:43.5	YRI	LRH, iHS, XP-EHH	0.3	83	30	0	0	
chr22:32.5	YRI	LRH	0.4	318	188	35	3	*LARGE*
chr23:35.1	YRI	LRH, iHS	0.6	50	35	25	0	
chr23:63.5	YRI	LRH, iHS	3.5	13	3	1	0	
	Total SNPs		16.74	9,166	2,898	480	41	

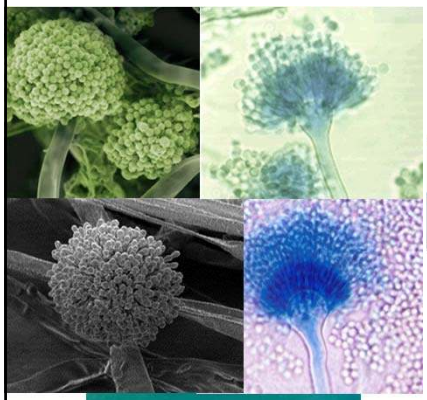
In the European population, involved in skin pigmentation

In the West African population, related to Lassa virus infection

Sabeti et al. (2007) Nature

Genomes Provide a Common Yardstick for Comparison

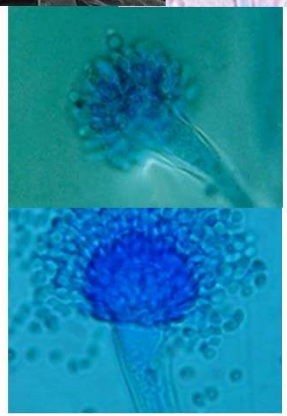
Morphologically very similar!



A. nidulans/niger

A. oryzae/terreus

A. clavatus



N. fischeri

A. fumigatus

Average protein sequence identity

60 — 70 — 80 — 90 — 100

Kluyveromyces lactis

Candida glabrata

fish

birds

S. uvarum

They Interbreed!

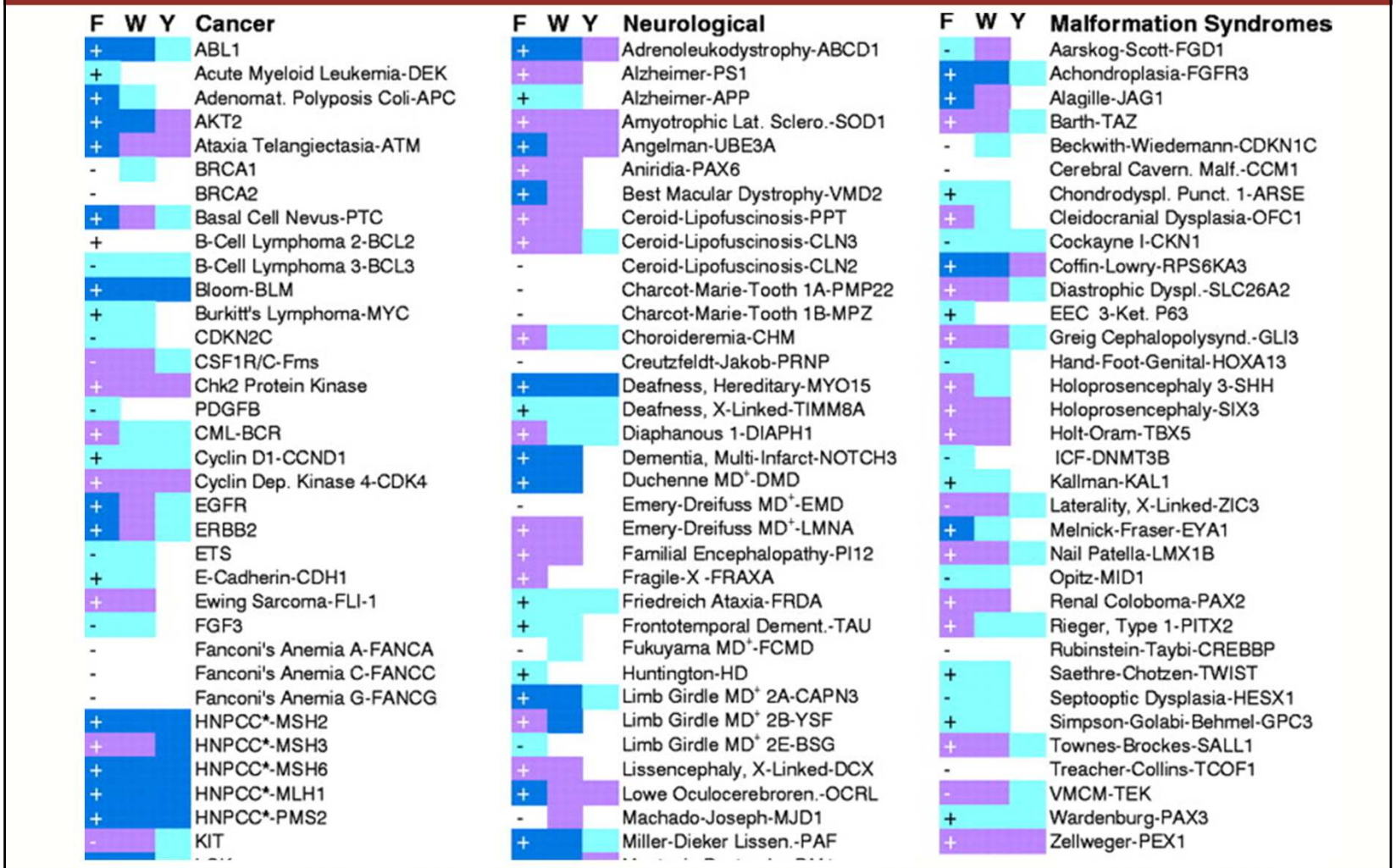
S. paradoxus ↔ *S. cerevisiae*

mice

humans

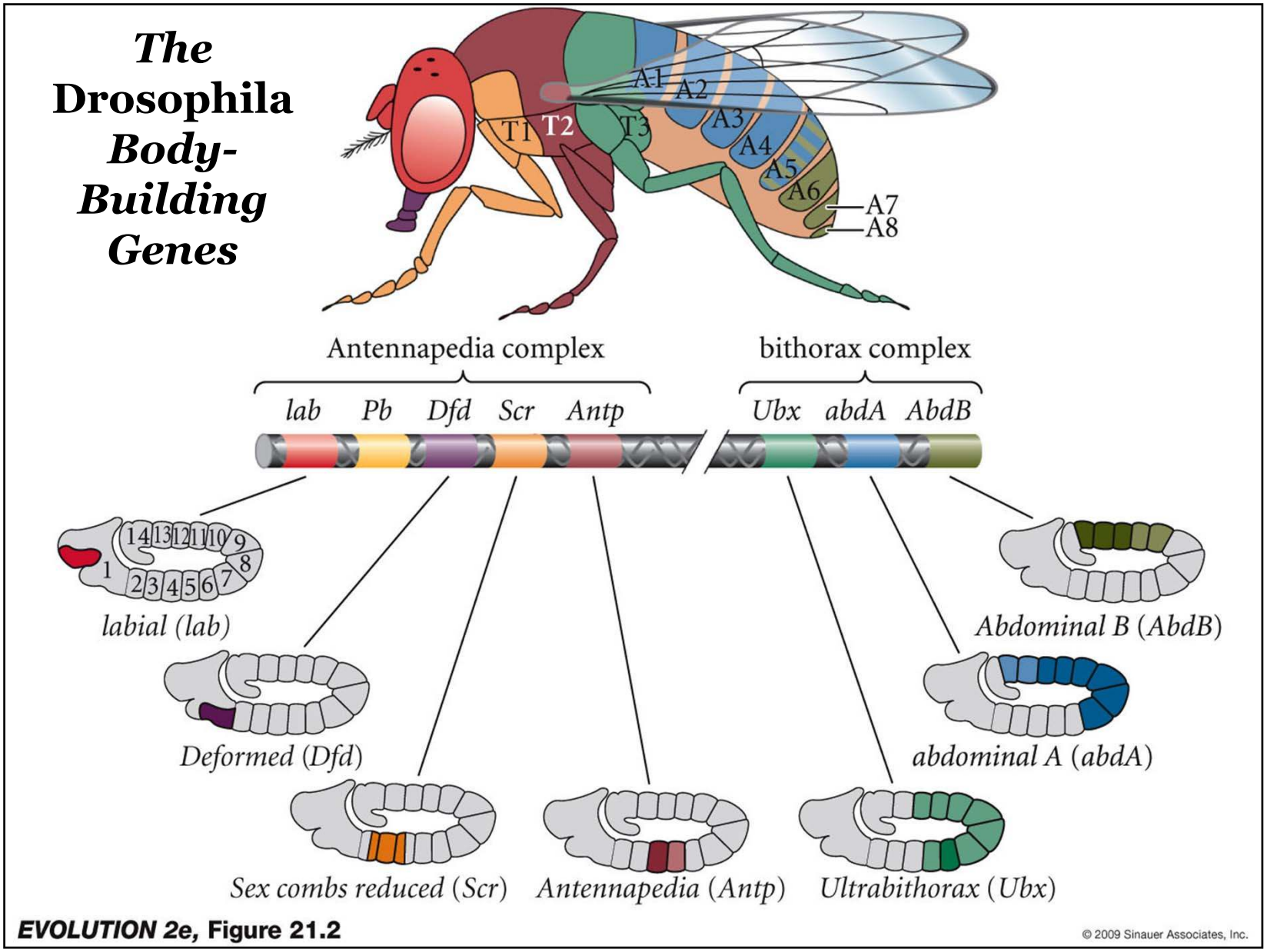
Fedorova et al. (2008) PLoS Genet.

What Makes Us Sick Is the Stuff of Life



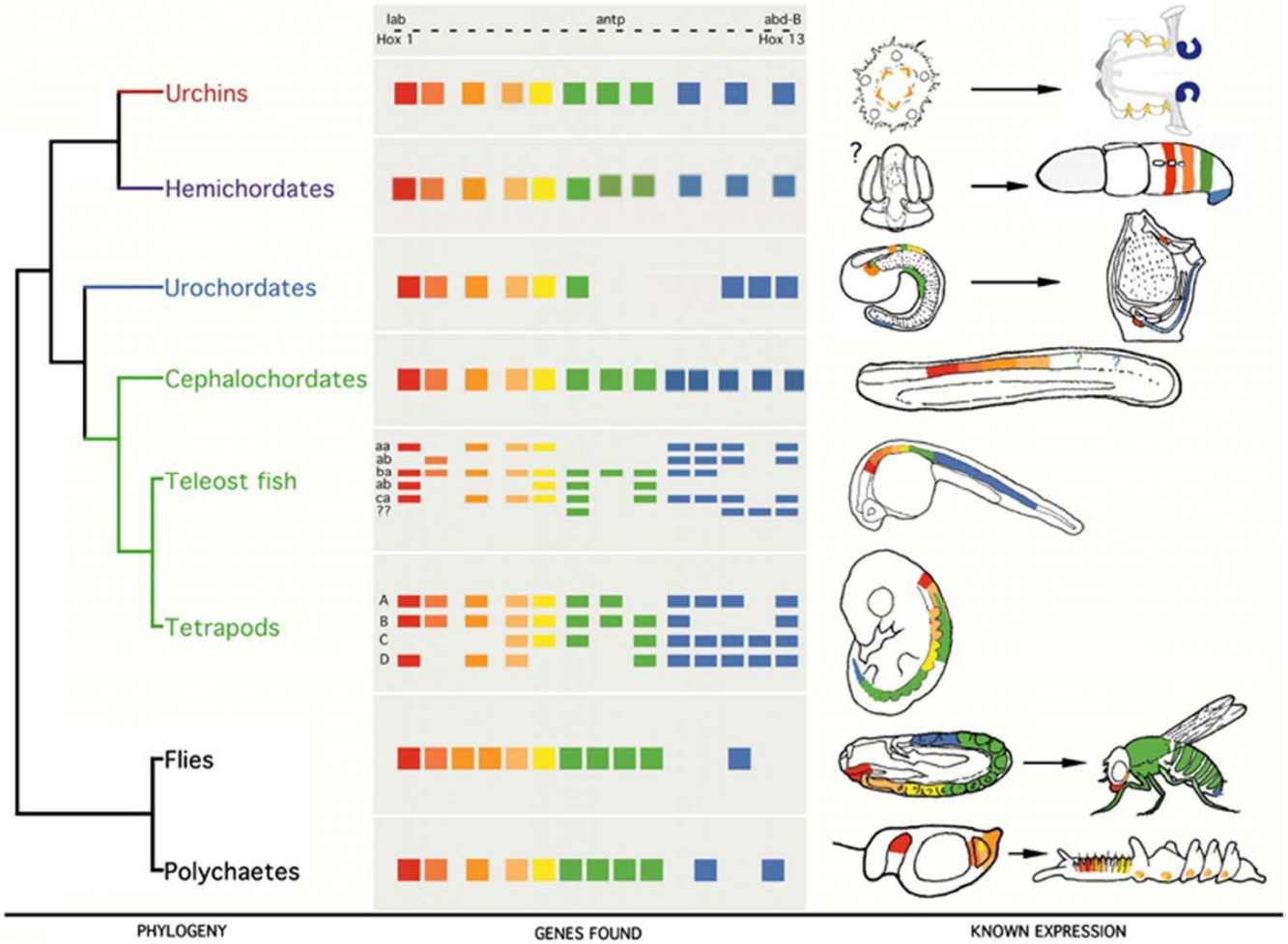
Human disease-associated genes shared with flies (F), worms (W), and Yeast (Y); from Rubin et al. (2000) Science

The Drosophila *Body-Building Genes*



EVOLUTION 2e, Figure 21.2

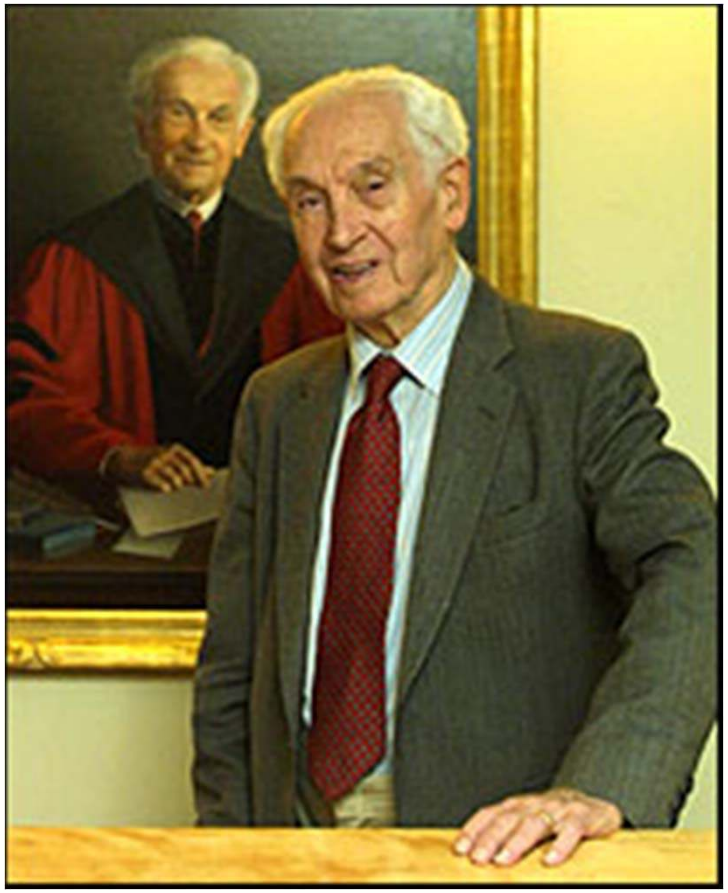
Animal Bodies are Built from the Same Genetic Toolkit

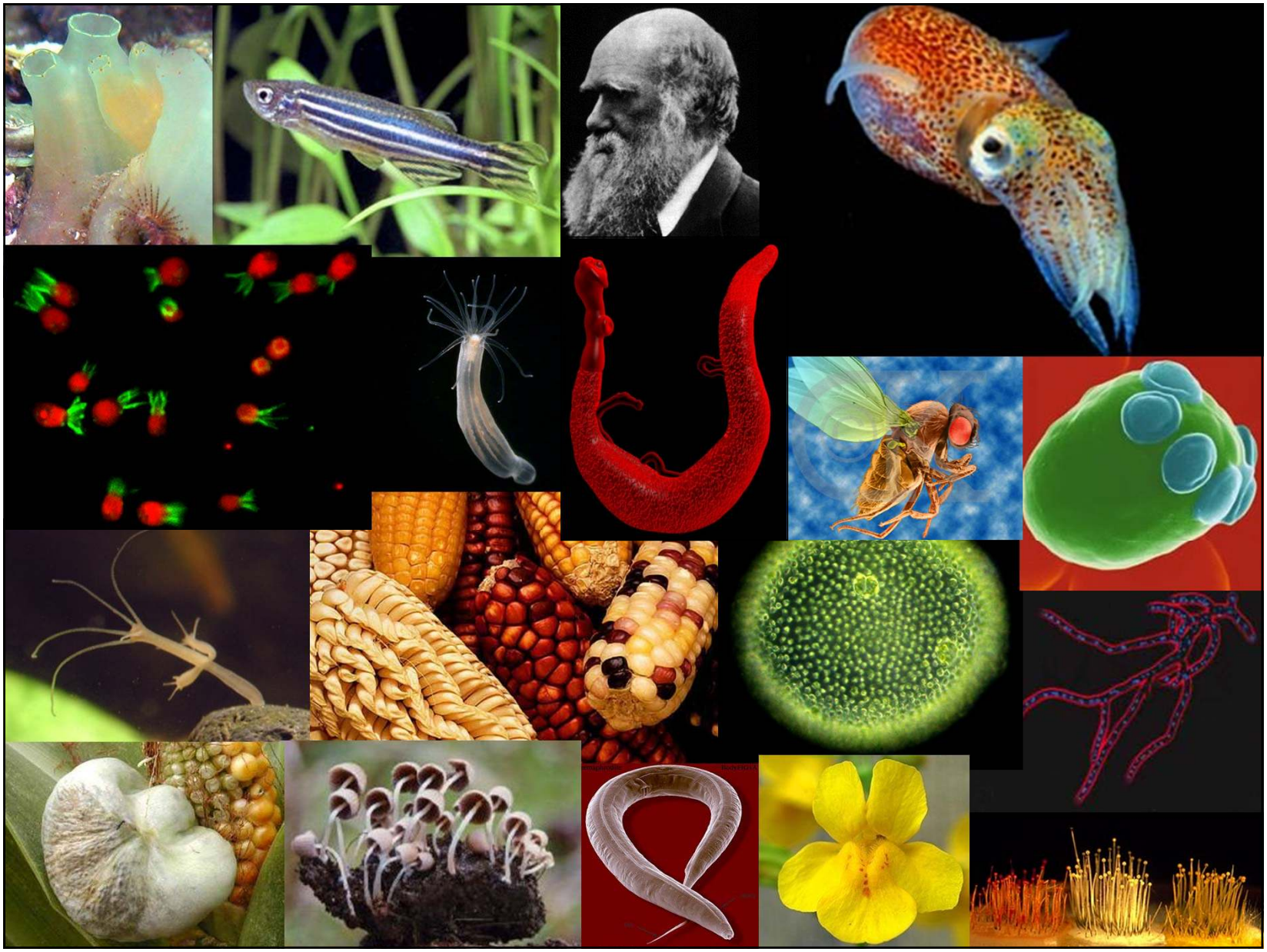


Swalla (2006) Heredity

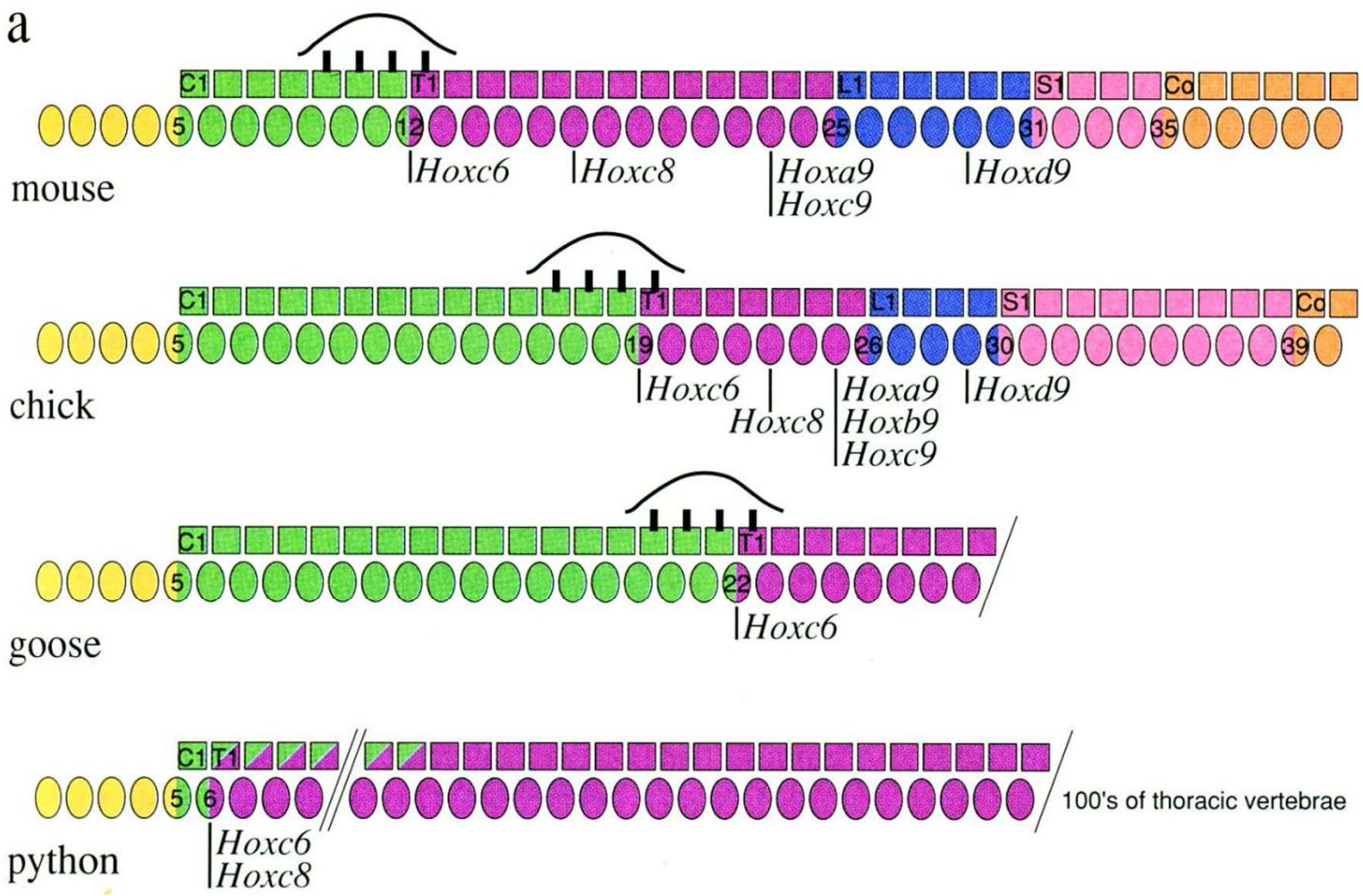
"...the search for homologous genes is quite futile except in very close relatives. If there is only one efficient solution for a certain functional demand, very different gene complexes will come up with the same solution, no matter how different the pathway by which it is achieved. The saying 'Many roads lead to Rome' is as true in evolution as in daily affairs"

Ernst Mayr, 1963



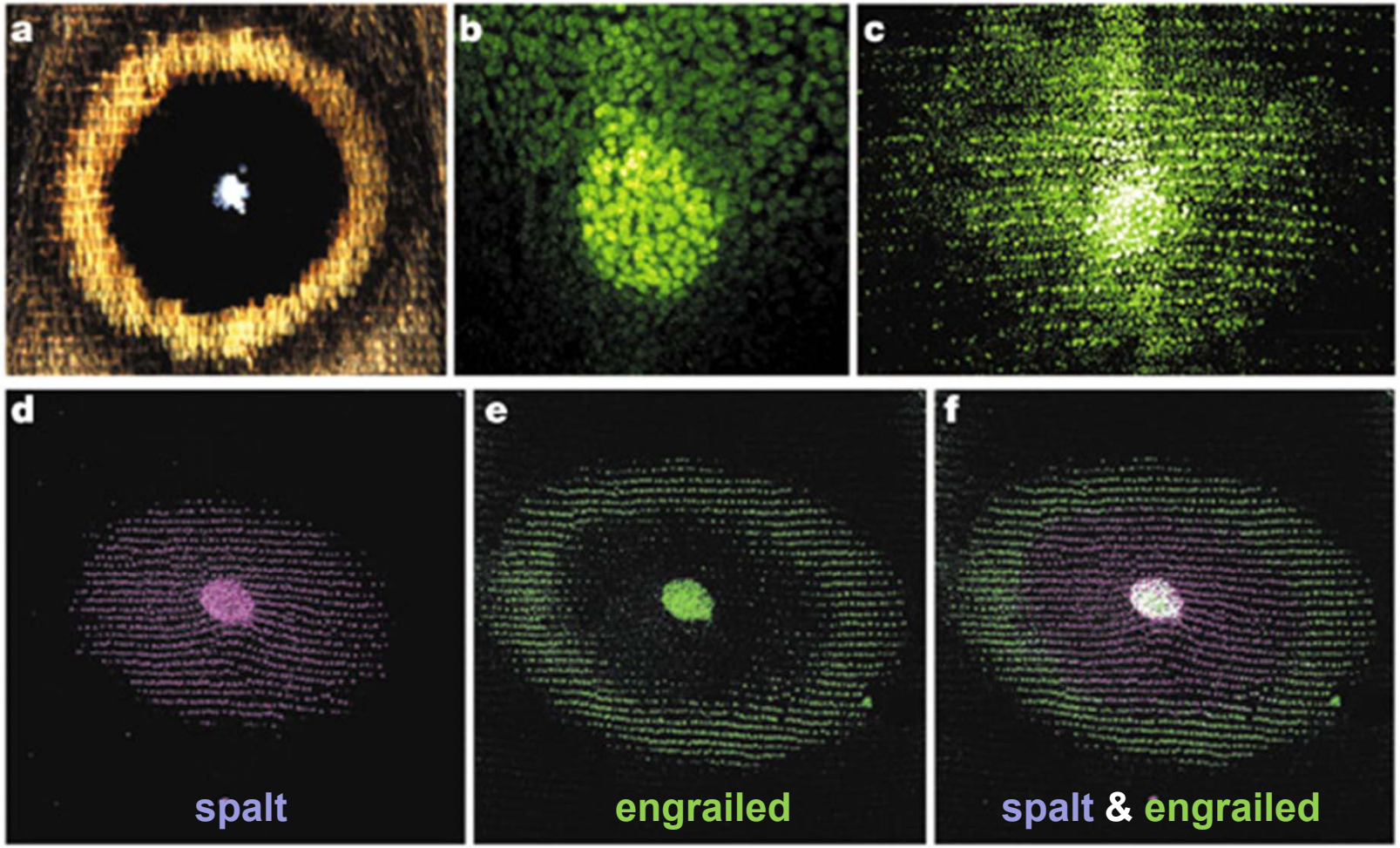


Altering the Expression of Old Genes



Carroll et al. (2005) From DNA to Diversity, 2nd ed., Fig. 5.6

Teaching Old Genes New Tricks (Gene Co-Option)



Beldade & Brakefield (2002) Nature Rev. Genet.

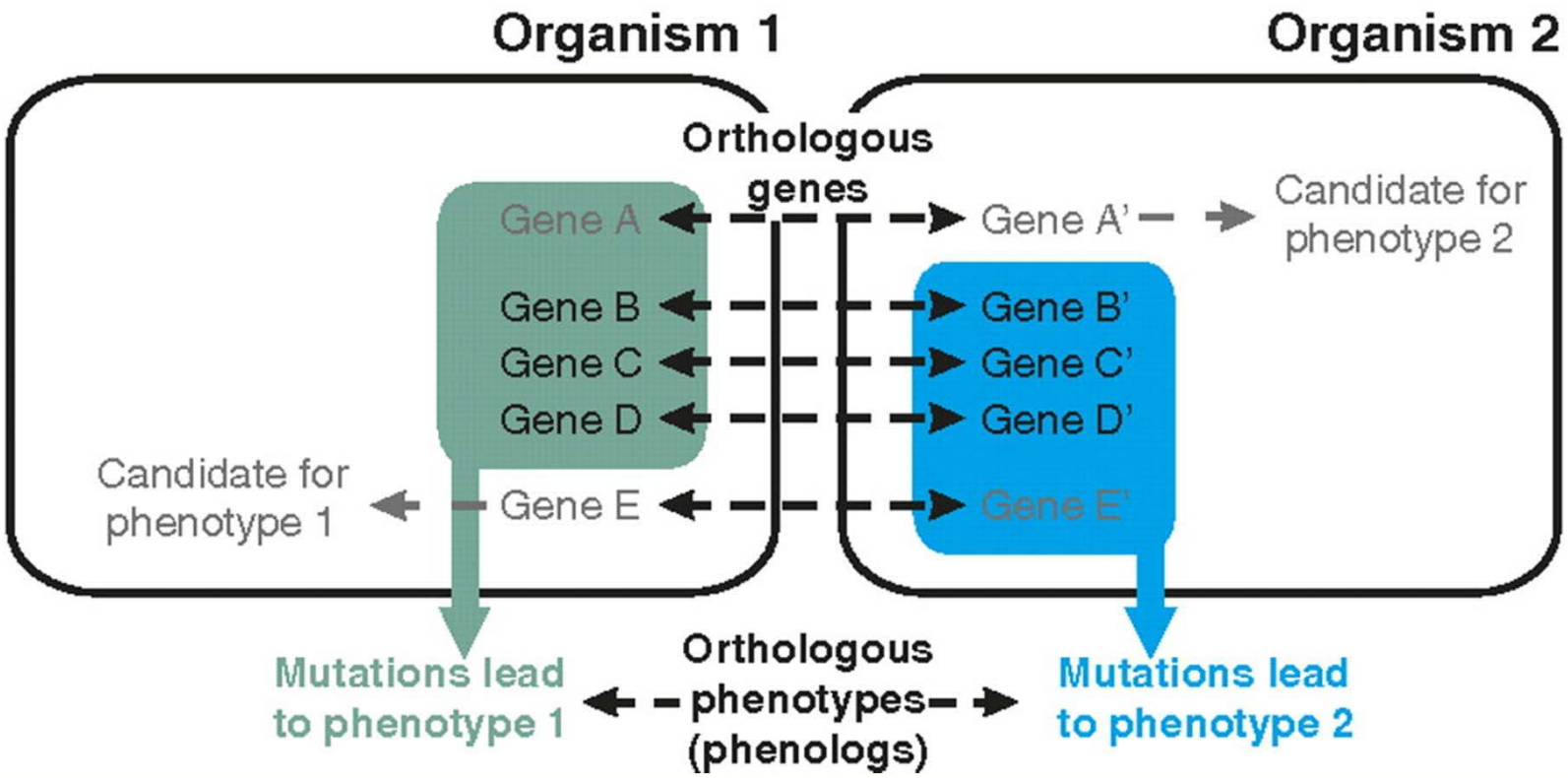
"Our method suggests a yeast model for angiogenesis defects, a worm model for breast cancer, mouse models of autism, and a plant model for the neural crest defects associated with Waardenburg syndrome, among others"

McGary et al. (2010) PNAS

"Modelling neurodegeneration in *Saccharomyces cerevisiae*"

Khurana & Lindquist (2010) *Nat. Rev. Neurosci.*

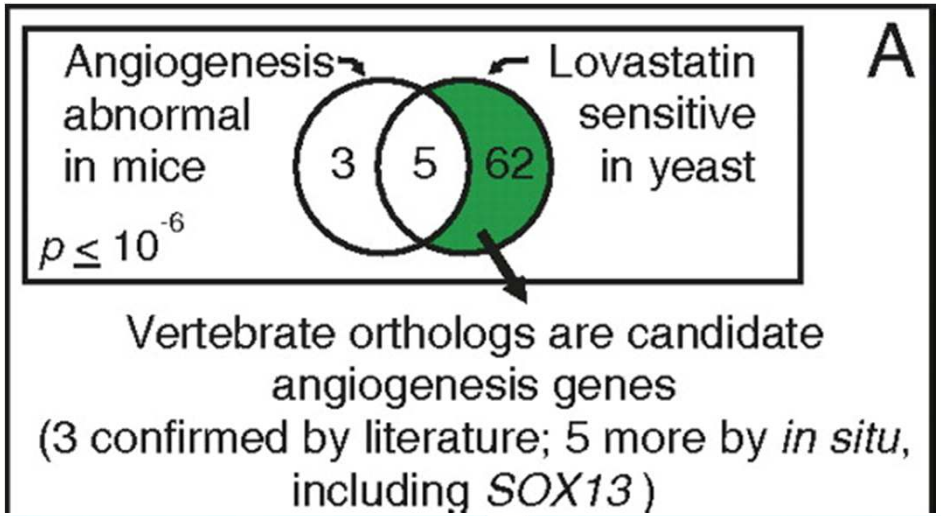
Deep Homology for Nonobvious Human Disease Gene Discovery

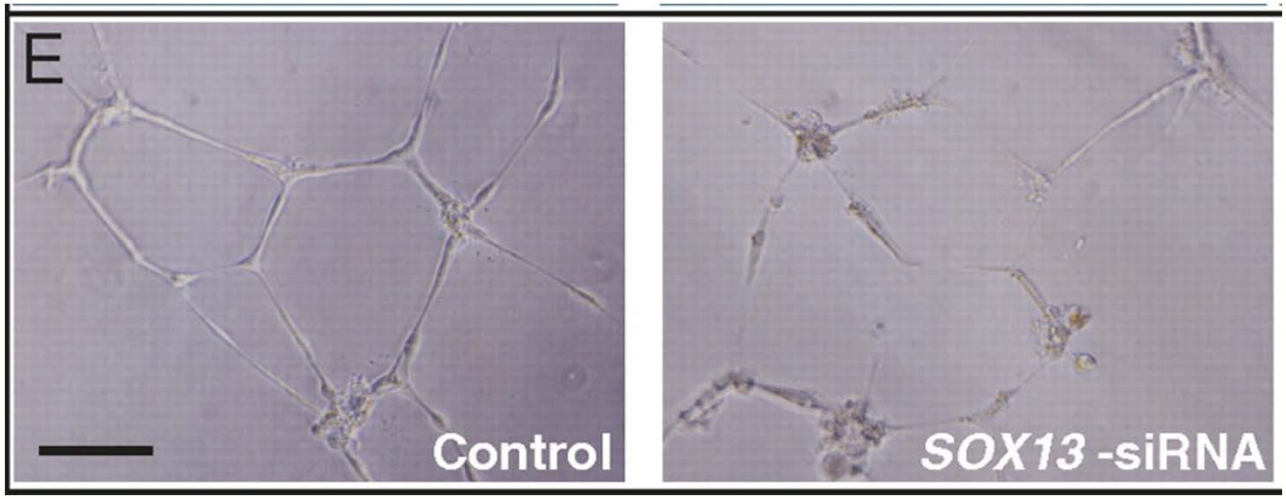


- **A worm model for breast cancer**
- **A yeast model for angiogenesis defects**
- **A plant model for the neural crest defects**

McGary et al. (2010) PNAS

A Yeast Model for Angiogenesis





McGary et al. (2010) PNAS

Lecture Outline

- Introduction to evolutionary genomics
- **Brief review of next-generation DNA sequencing & its applications**
- **Benchmarking transcriptome sequencing for functional and evolutionary genomics**
- **Transcriptome sequencing for molecular phylogenetics**

----------Coffee Break----------

- **Genomics within species (using next-generation DNA sequencing)**
- **Genomics between species (using publicly available genome data)**

Genomics: "Big Science" Driven by a Few Centers

wellcome trust sanger institute

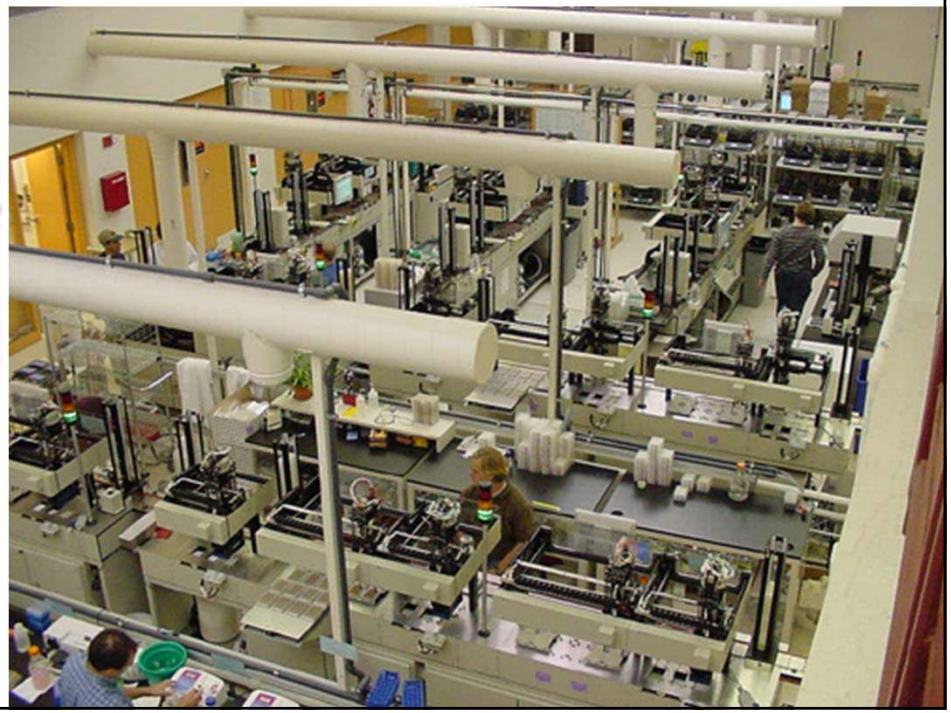
BCM
Baylor College of Medicine

BROAD INSTITUTE

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

J. Craig Venter INSTITUTE

THE Genome CENTER
AT WASHINGTON UNIVERSITY



Next-Generation DNA Sequencing Technologies

Helicos		
2 x 55 bp	~8 days	35 Gb

Illumina		
2 x 100 bp	~8 days	200 Gb

454 / Roche		
2 x 400 bp	~0.4 days	0.6 Gb

SOLiD ABI		
2 x 75 bp	~14 days	300 Gb

The Speaker Declares No Competing Financial Interests

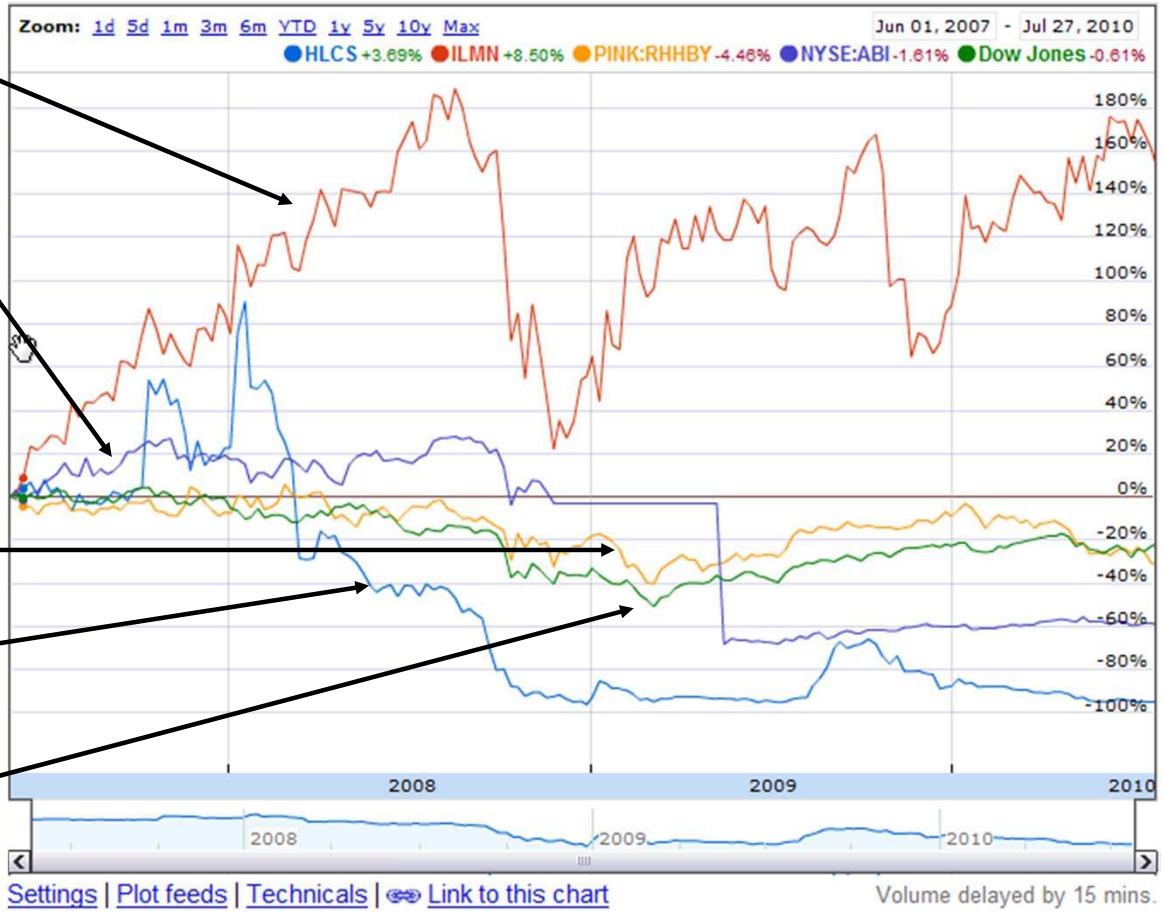
Illumina

ABI SOLiD

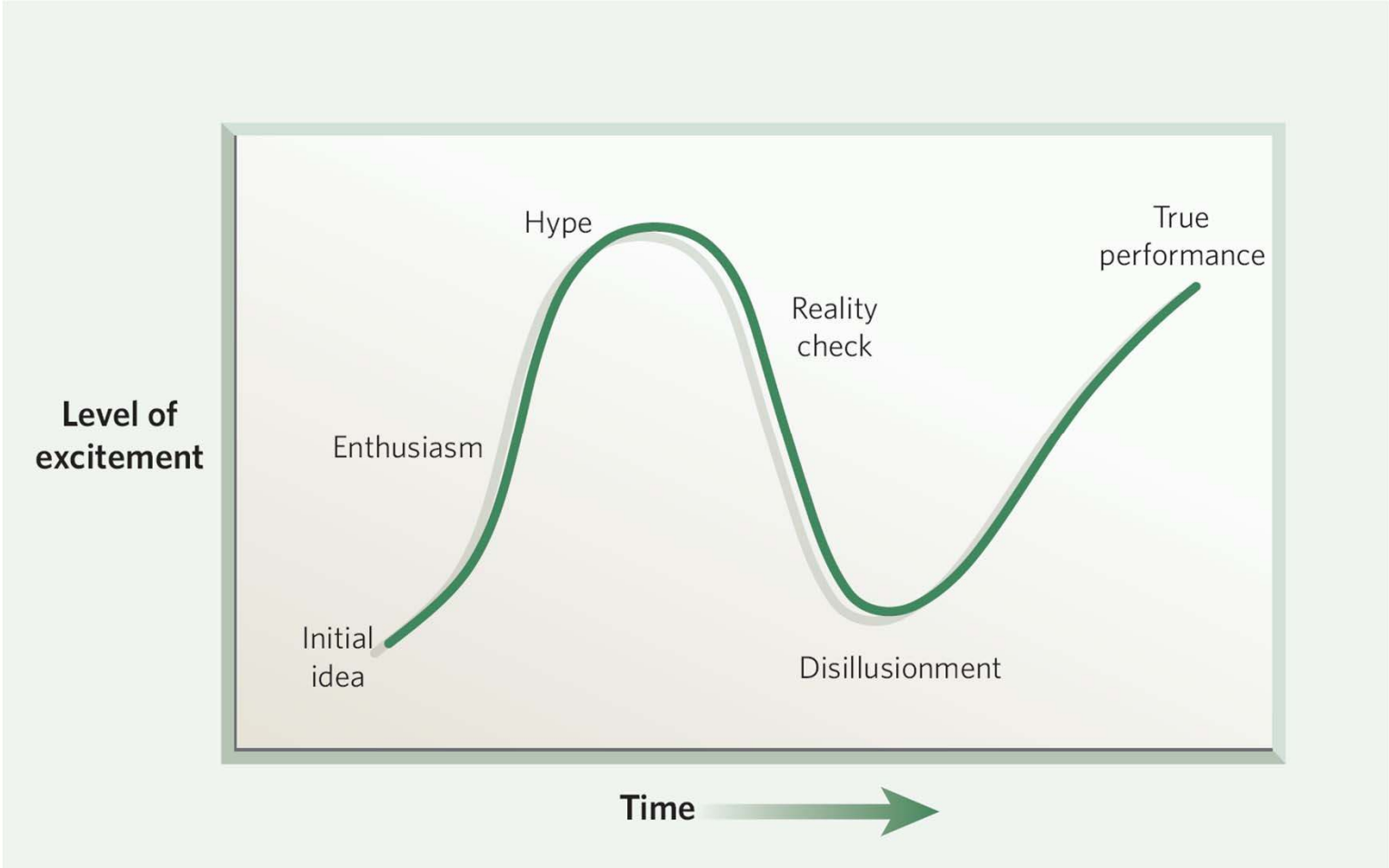
Roche/454

Helicos

My Retirement



"New Technology Level of Enthusiasm" Curve



Collins (2006) Nature

Meeting report

Genome informatics: taming the avalanche of genomic data

Erik LL Sonnhammer

Address: Center for Genomics and Bioinformatics, Karolinska Institutet, 171 77 Stockholm, Sweden. E-mail: Erik.Sonnhammer@cgb.ki.se

Comment

Tsunami

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

PERSPECTIVES

Preparing for the flood: evolutionary biology in the age of genomics

James R. Brown

Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data

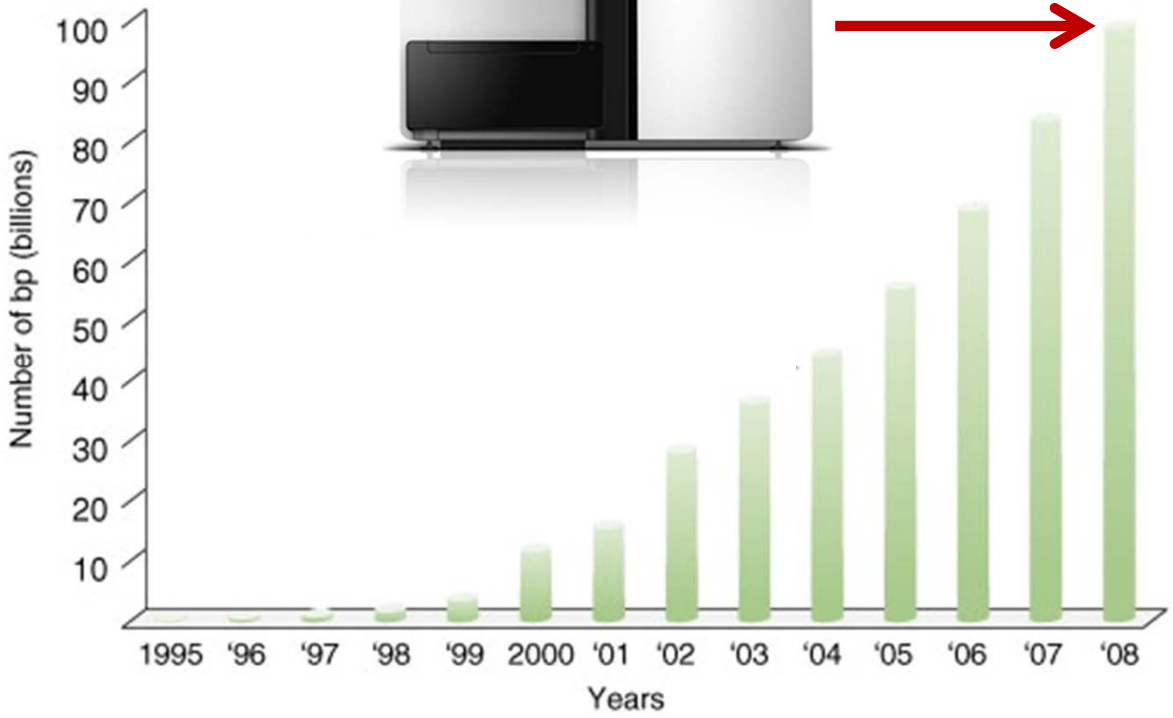
CHRISTOPHER SOUTHAN

GRAHAM CAMERON

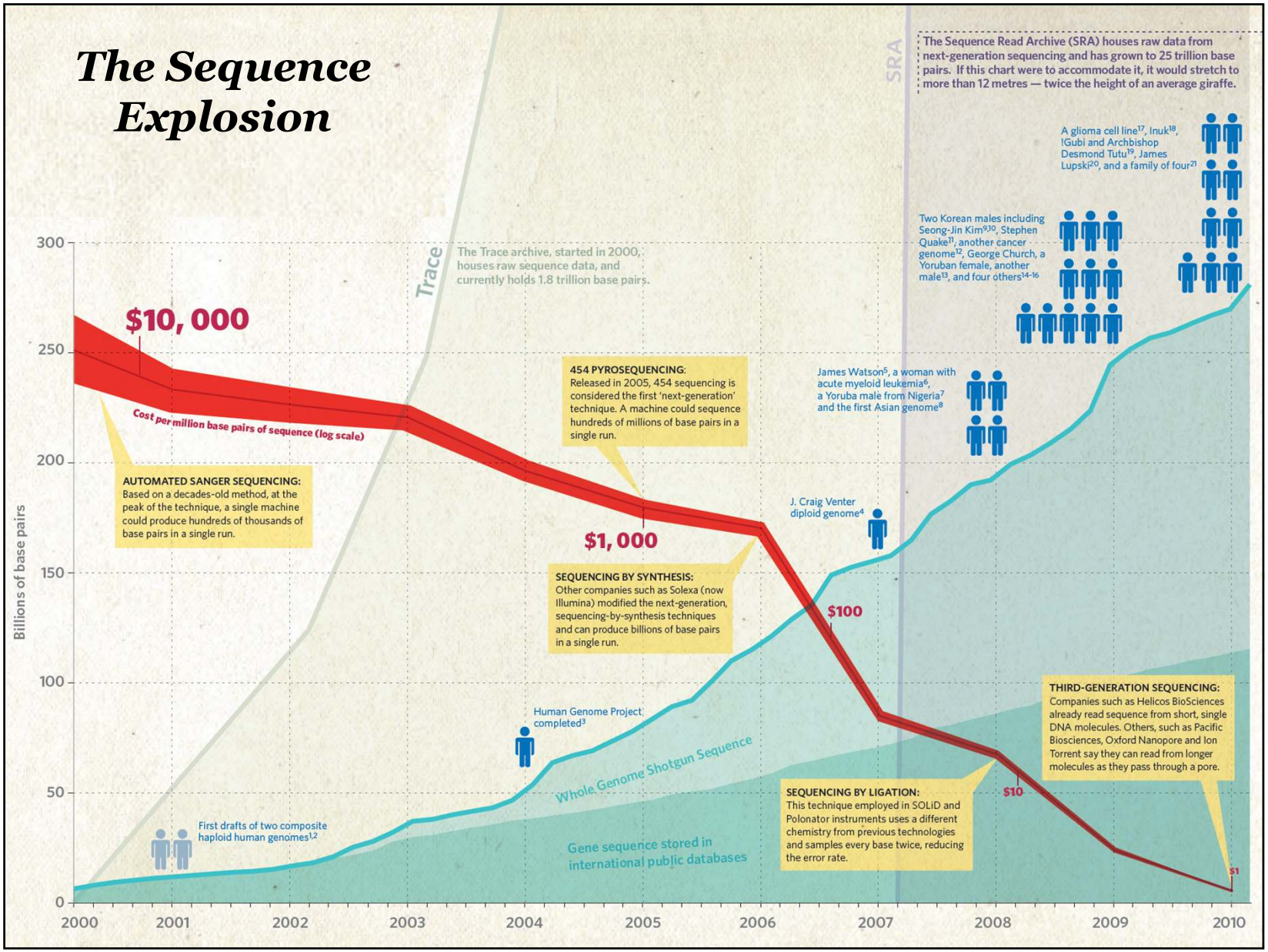
EMBL-European Bioinformatics Institute

Why Is Next-Generation DNA Sequencing So Exciting?

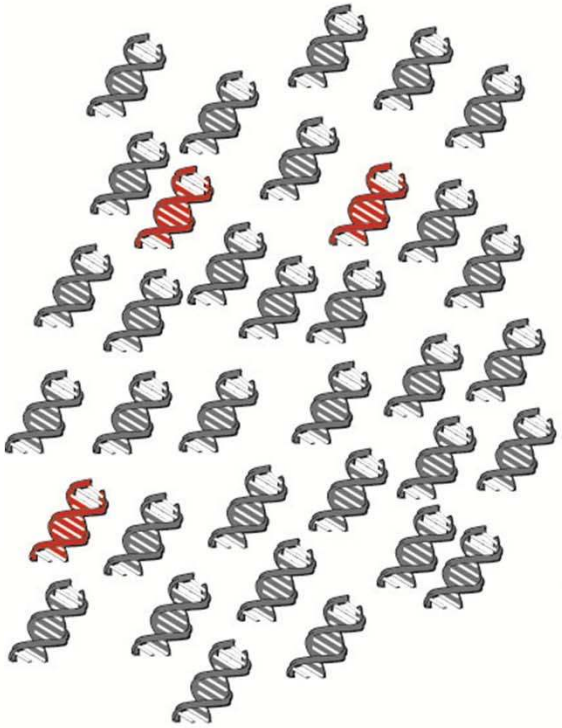
4 days' output



Gilad et al. (2009) Trends Genet.


The Sequence Explosion
The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.
SRA
Trace
The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.
A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]
Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14-16]
James Watson[5], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]
J. Craig Venter diploid genome[4]
Human Genome Project completed[3]
First drafts of two composite haploid human genomes[1,2]
$10, 000
Cost per million base pairs of sequence (log scale)
$1, 000
$100
$10
$1
AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.
454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.
SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.
SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.
THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.
Whole Genome Shotgun Sequence
Gene sequence stored in international public databases
Billions of base pairs
0
50
100
150
200
250
300
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

Next-Gen Sequencing is Qualitative and Quantitative



NGSTs

Each DNA template is sequenced directly

```
Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red trascript
Grey transcript
Grey transcript
Grey transcript
Grey transcript
Red trascript
Grey transcript
Grey transcript
Red transcript
Grey transcript
Grey transcript
```

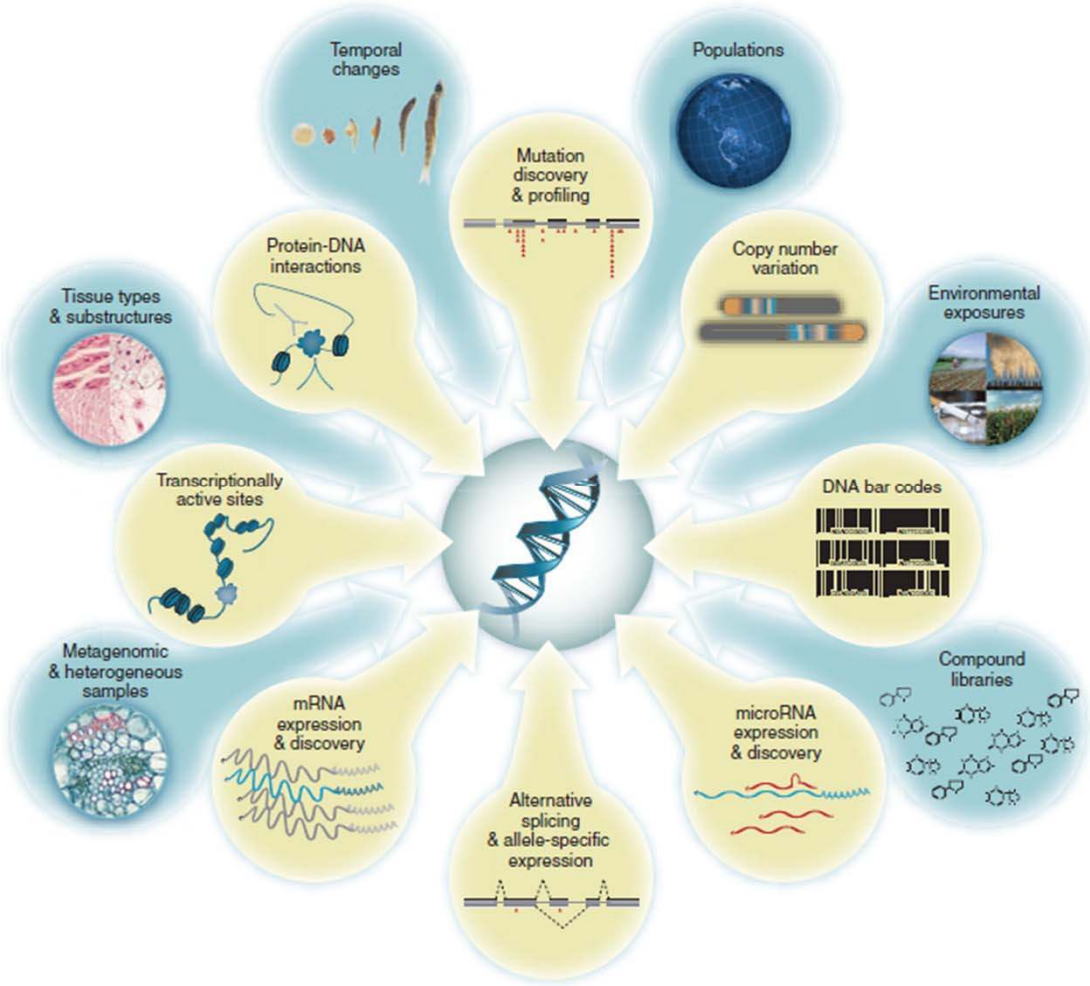
Capillary Sequencing

All DNA templates are sequenced together to create a single consensus sequence

```
Grey transcript
```

Rokas & Abbot (2009) Trends Ecol. Evol.

Sequence Census Assays



Kahvejian et al . (2008) Nature Biotech.

The Genomes of Non-Model Organisms are the New Frontiers

Epidemiology	What is the aetiology of emerging, uncharacterized infectious diseases? What are the core genomes of different microbes or pathogens occupying different niches or hosts?
Social Evolution	Are maternal care and sibling care behaviors regulated by similar patterns of gene expression?
Biodiversity	What is the extent of undocumented microbial diversity in different ecosystems? What is the microbial flora of the human gut?
Phylogenetics	What is the phylogeny of all vertebrate species or angiosperms using organelle or nuclear genome data?
Population Genetics	What is the variation in the mutation rate across organisms?
Experimental evolution	What is the genetic basis of phenotypes emerging during laboratory evolution?
Palaeontology	What are the evolutionary relationships of ancient organisms to extant taxa?
Evolution of development	How do regulatory networks evolve and rewire? What are the *cis*-regulatory targets of all transcription factors in a species and how do they evolve?

Rokas & Abbot (2009) Trends Ecol. Evol.

Lecture Outline

- Introduction to evolutionary genomics
- Brief review of next-generation DNA sequencing & its applications
- **Benchmarking transcriptome sequencing for functional and evolutionary genomics**
- Transcriptome sequencing for molecular phylogenetics

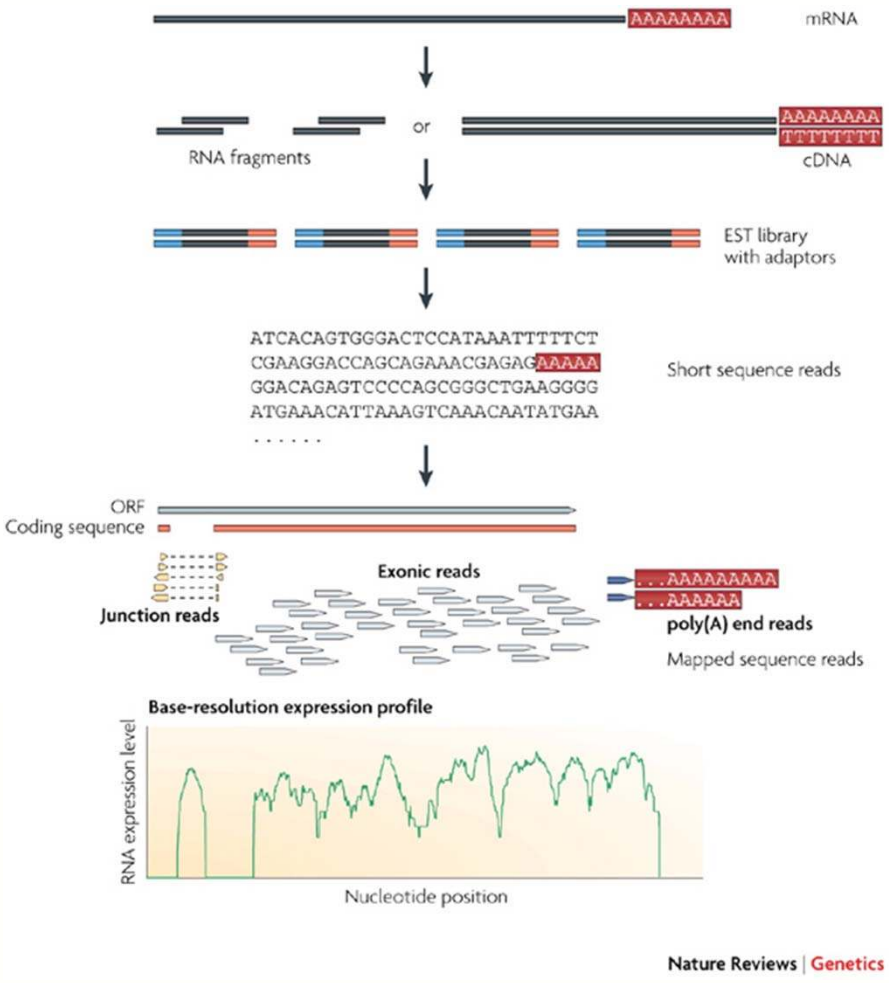
----------Coffee Break----------

- Genomics within species (using next-generation DNA sequencing)
- Genomics between species (using publicly available genome data)

RNA-Seq for Non-Model Organisms: What's in it for You?

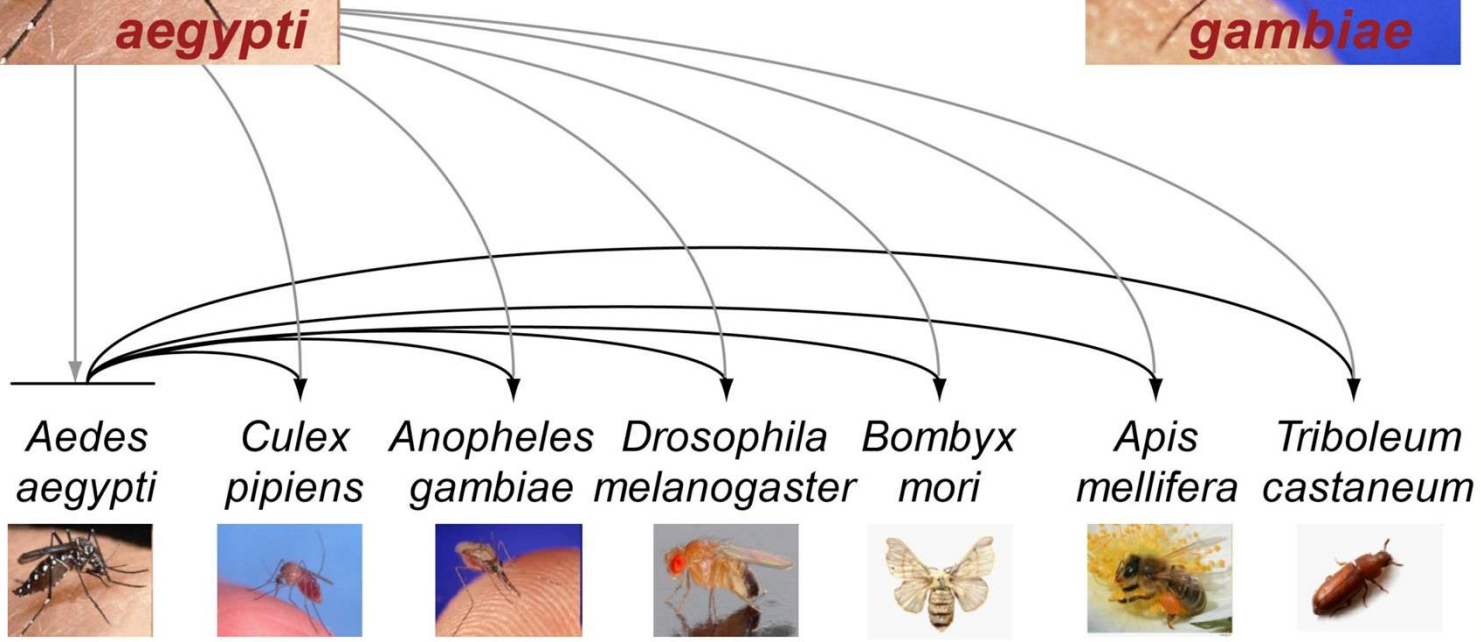
John Gibbons

A Typical RNA-Seq Experiment



Wang et al. (2009) Nature Rev. Genet.

The Next-Gen Revolution: What's in it for You?



How much? How good? How useful for functional studies?
How useful for evolutionary studies?

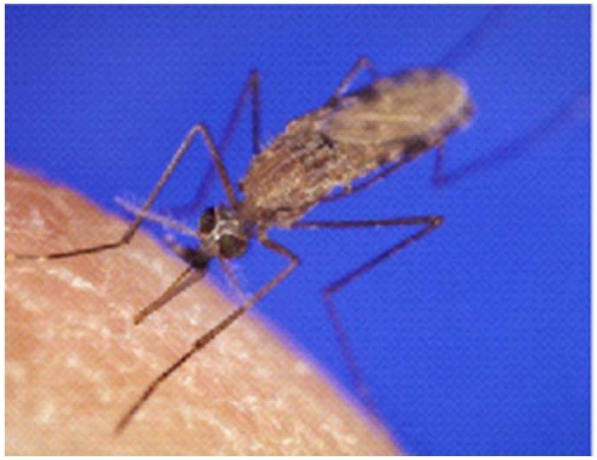
Gibbons et al. (2009) Mol. Biol. Evol.

How Much?

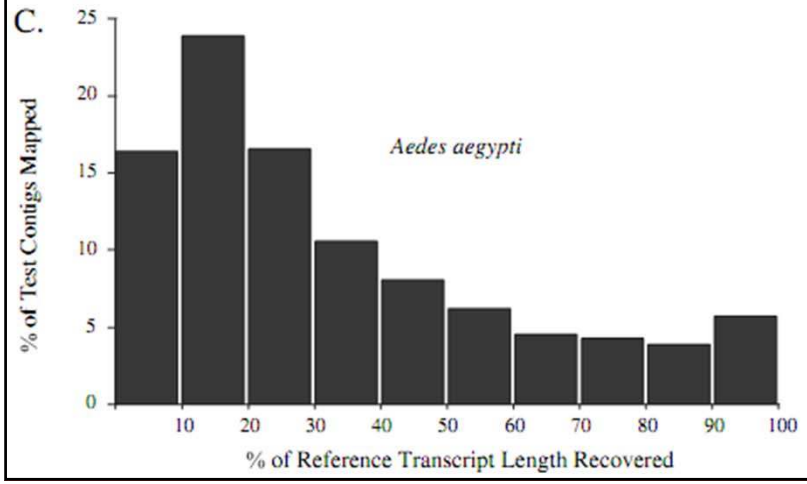
Aedes aegypti

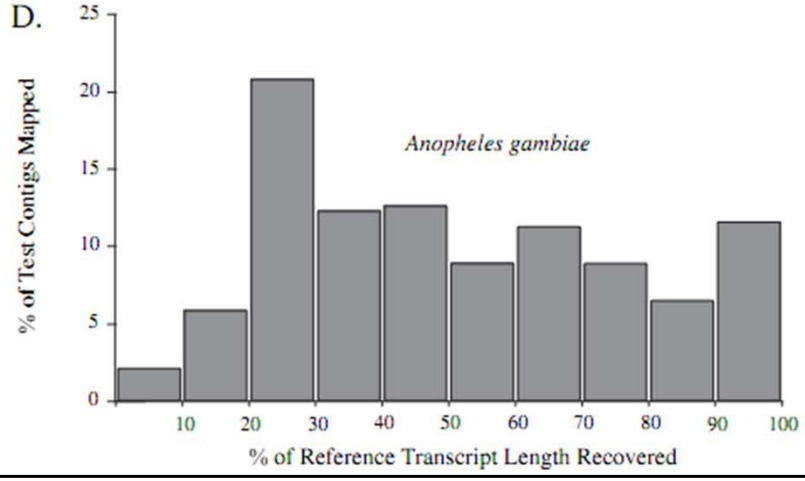
27% of the transcriptome
34% of each transcript

Anopheles gambiae



20% of the transcriptome
50% of each transcript





RNA-Seq Data is Qualitatively Accurate

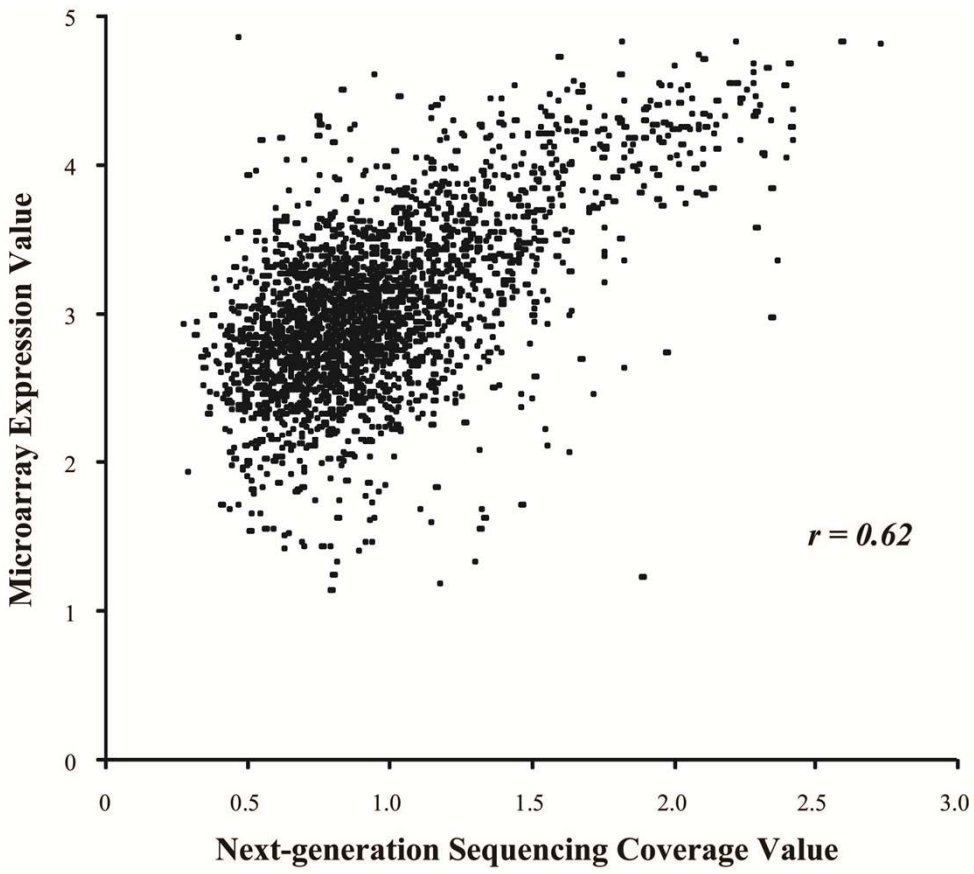
```
                Reference CAG CCA ATG AGG TCT CAG TCG
Substitutions                ||| ||| ||  ||| ||  ||| |||   3x10^-3
                     Test CAG CCA ATC AGG TCA CAG TCG

                Reference CAG CCA ATG --- --- CAG TCG
Indels                       ||| ||| |||       ||| |||   4x10^-5
                     Test CAG CCA ATG AGG TCA CAG TCG

                Reference CAG CCA ATG AGG TCT CAG TCG
Assembly Breaks                |       | ||| ||| ||| |||  2x10^-4
                     Test ATG GGG TAC AGG TCT CAG TCG
```

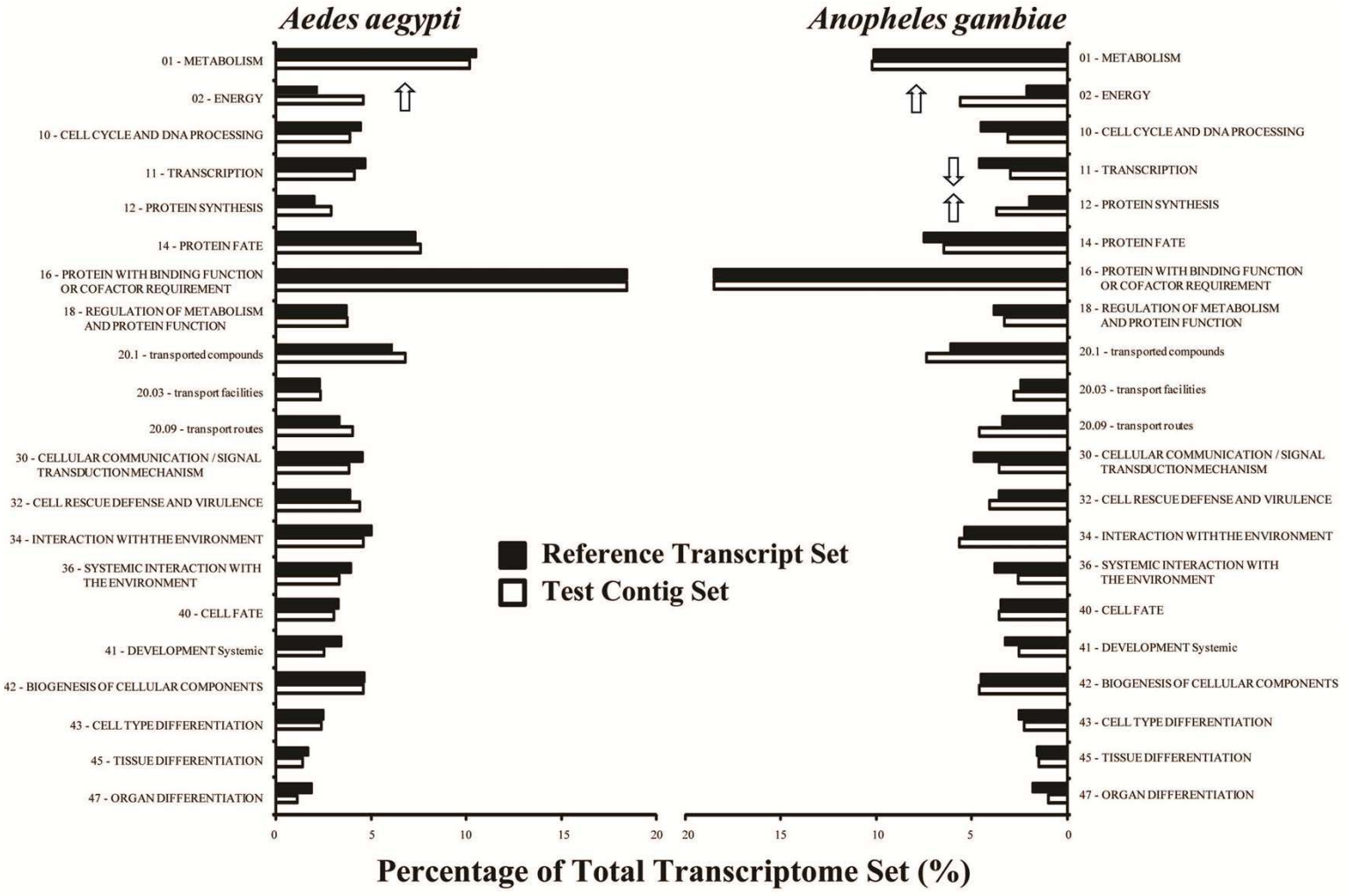
Gibbons et al. (2009) Mol. Biol. Evol.

RNA-Seq Data is Quantitatively Accurate



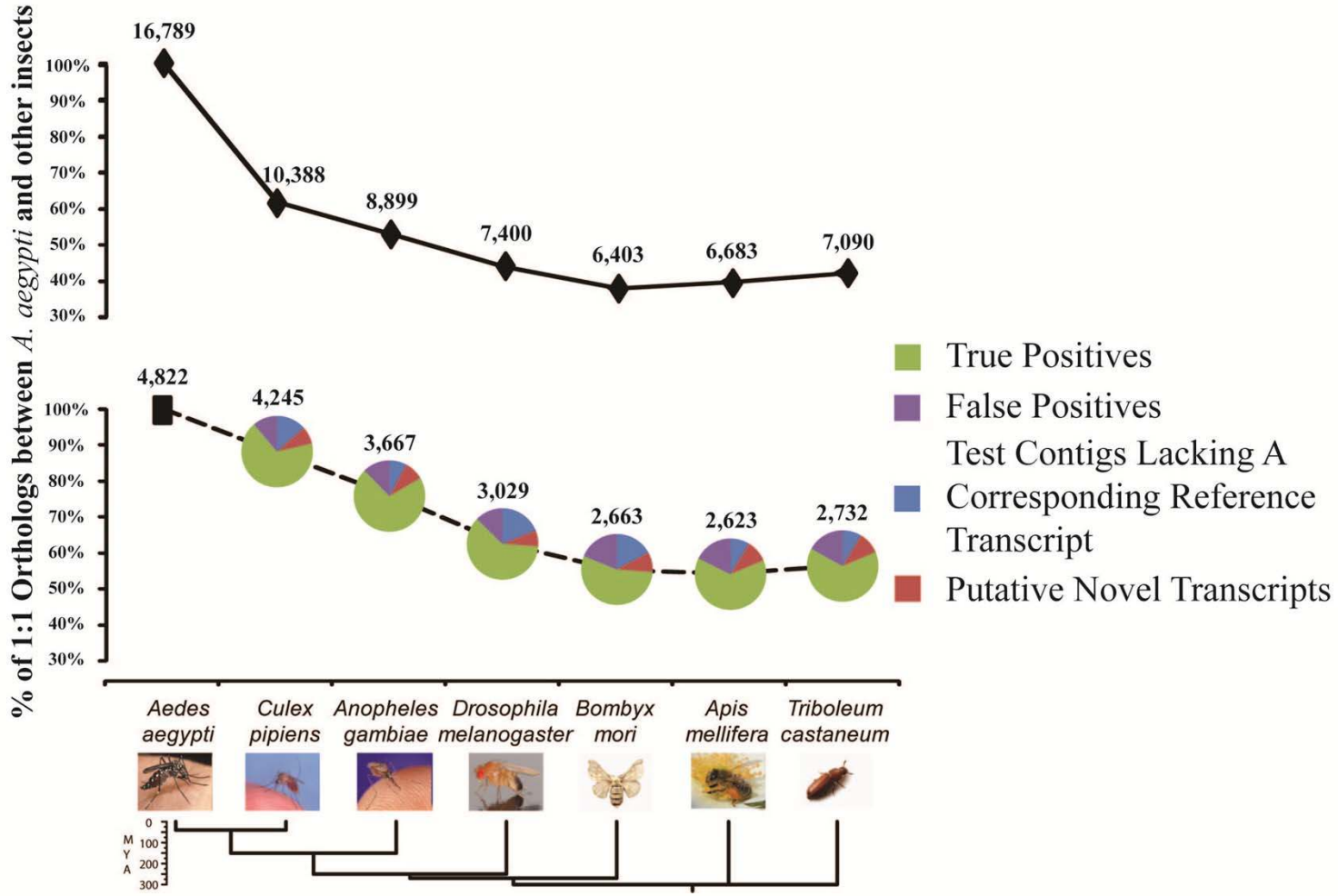
Gibbons et al. (2009) Mol. Biol. Evol.

RNA-Seq Data is Functionally Diverse



Gibbons et al. (2009) Mol. Biol. Evol.

Our RNA-Seq Data is a Goldmine for Molecular Markers



Gibbons et al. (2009) Mol. Biol. Evol.

Lecture Outline

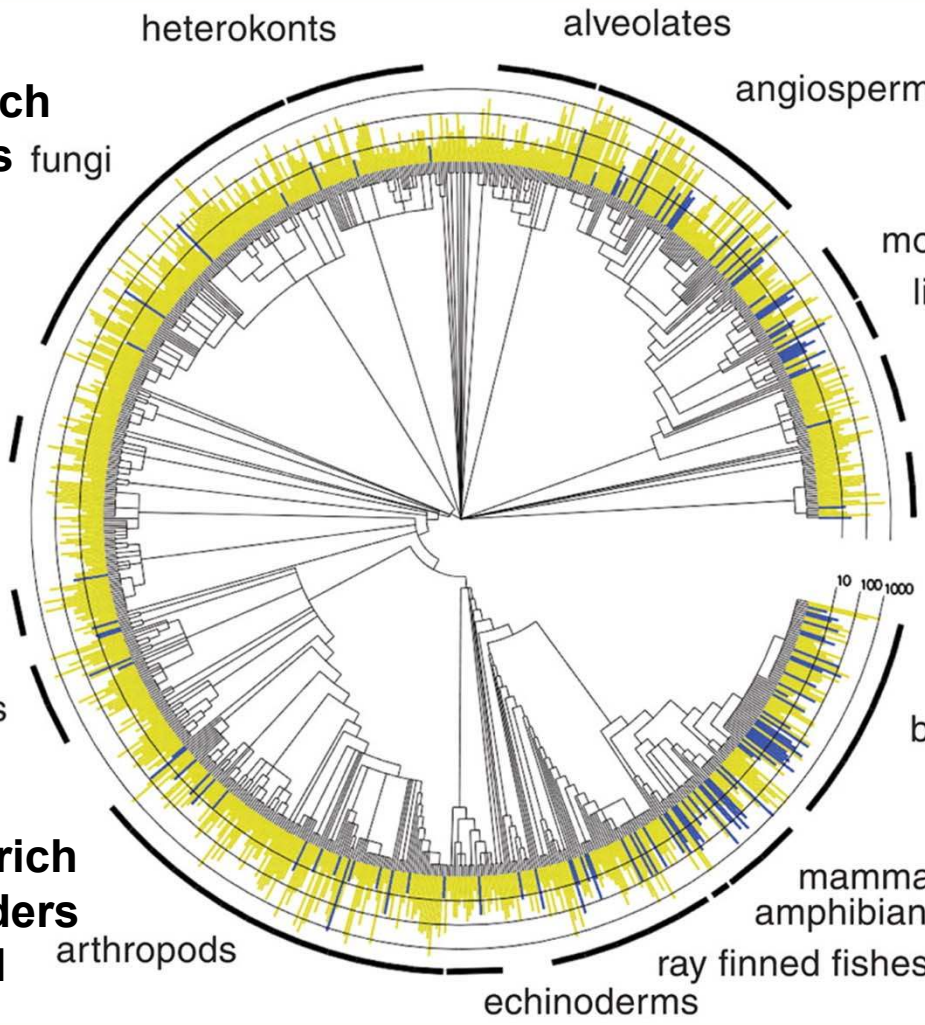
- Introduction to evolutionary genomics
- Brief review of next-generation DNA sequencing & its applications
- Benchmarking transcriptome sequencing for functional and evolutionary genomics
- **Transcriptome sequencing for molecular phylogenetics**

----------Coffee Break----------

- Genomics within species (using next-generation DNA sequencing)
- Genomics between species (using publicly available genome data)

Taxonomic Breadth versus Genomic Depth

1/40 species-rich fungal orders supported



1/45 species-rich arthropod orders supported

16/57 species-rich vertebrate orders supported

Sanderson (2008) Science

Can we Use RNA-Seq to Increase Genomic Depth?

Targeting the transcriptome is a good idea because:

1. **Smaller than genome (coding parts cover 7% of *Anopheles gambiae* genome)**

2. **Fewer repetitive and transposable elements**

3. **Unequal representation (5 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**

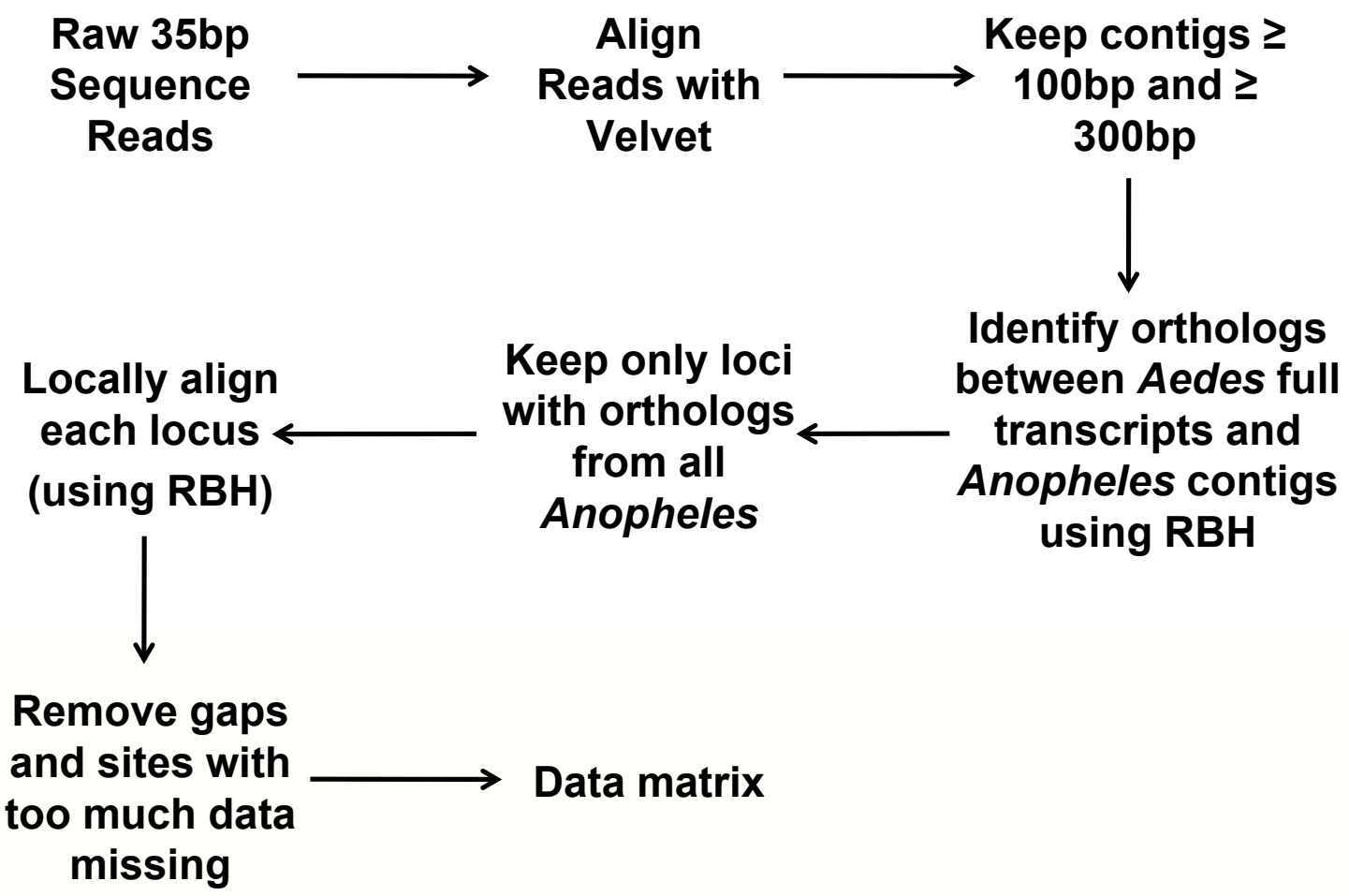
Hittinger et al. (2010) PNAS

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

Hittinger et al. (2010) PNAS

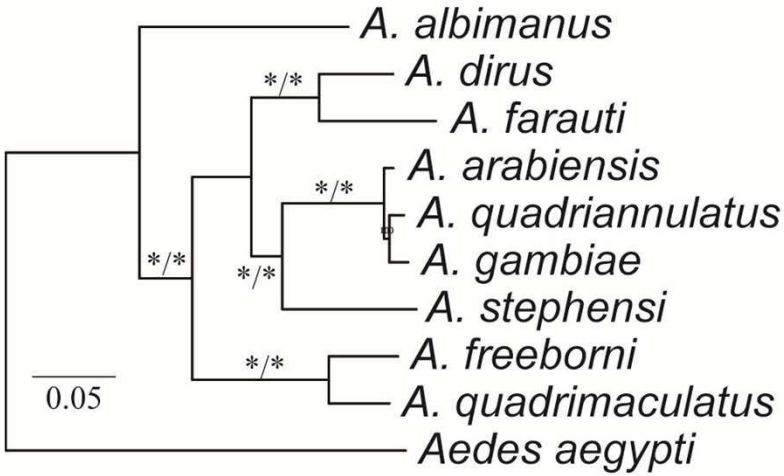
Experimental Design: The "Singlecontig" Strategy



Hittinger et al. (2010) PNAS

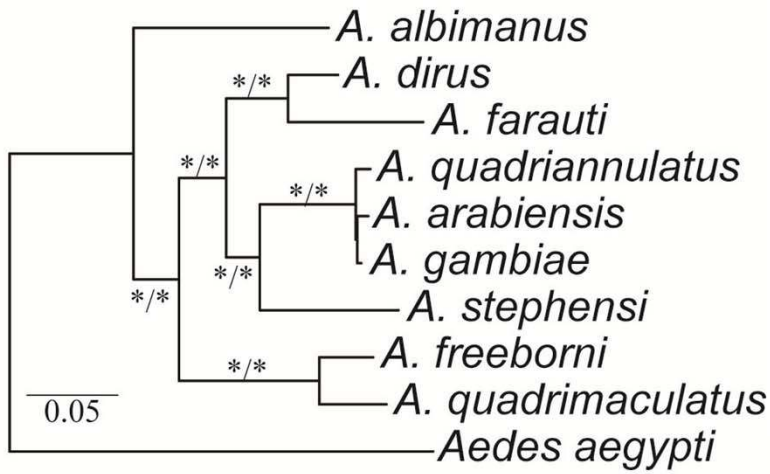
Robust Phylogenetic Inference from RNA-Seq Data

Using ≥ 100bp contigs



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

Using ≥ 300bp contigs

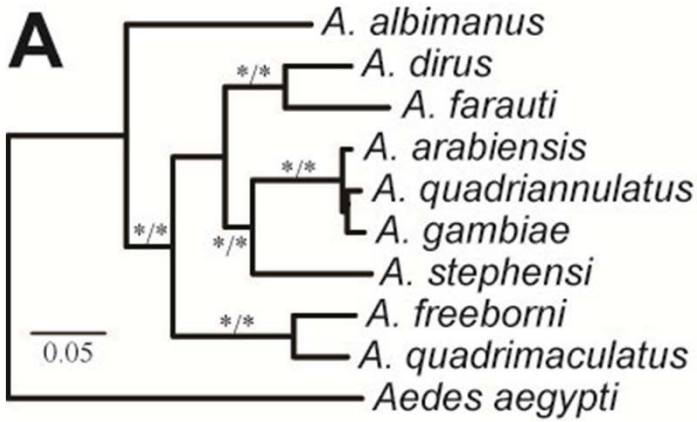


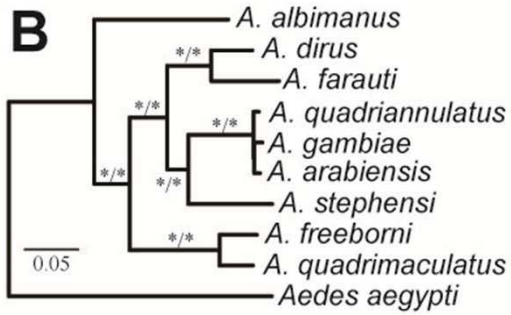
69 loci
Aln L: ~73 Kb
Missing data: 44%

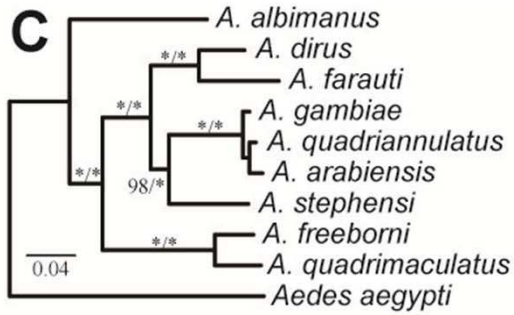
Hittinger et al. (2010) PNAS

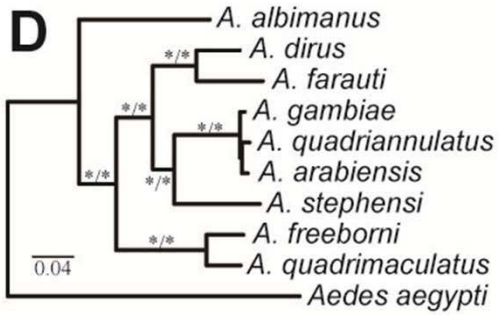
Accurate Phylogenetic Inference From our Data

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









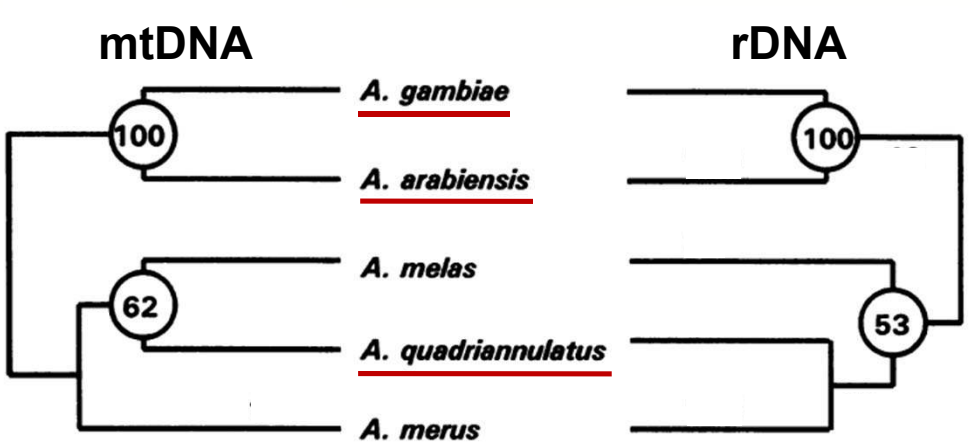
<u>Exclude erroneous loci</u>
491 loci
Aln L: ~329 Kb
Missing data: 50%

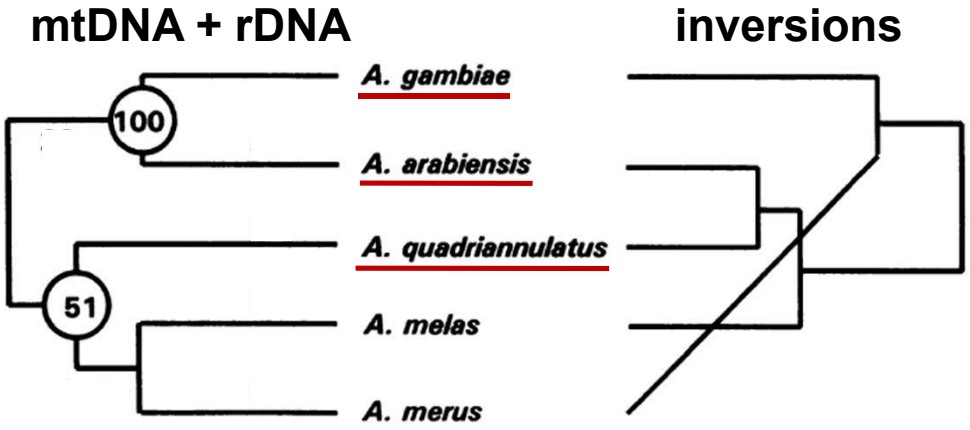
<u>Use only sites without data missing</u>
Aln L: ~15 Kb
Missing data: 0%

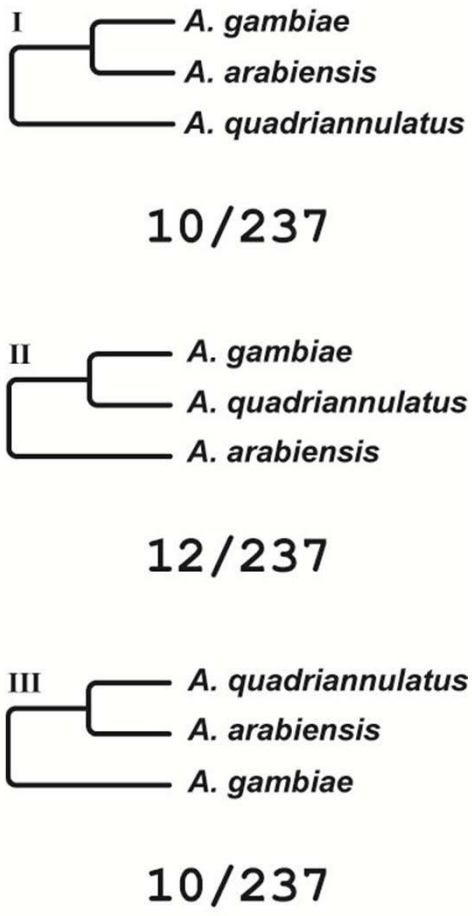
<u>Use *A. gambiae* as ref.</u>
634 loci
Aln L: ~472 Kb
Missing data: 50%

Hittinger et al. (2010) PNAS

Our Data Matrices Can Detect Population-Level Events



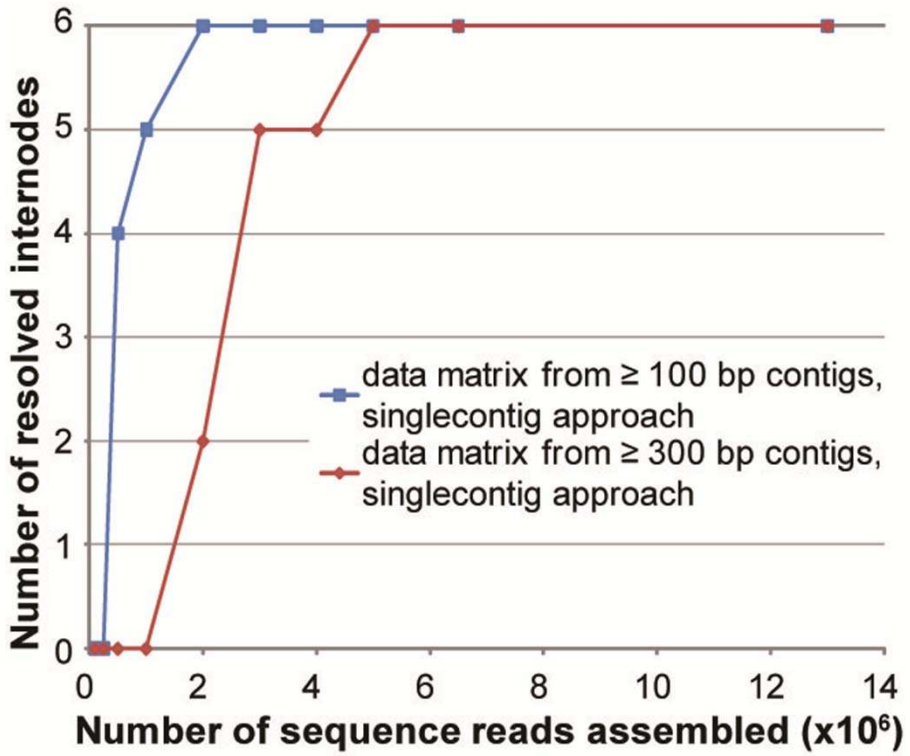




Besansky et al. (1994) PNAS

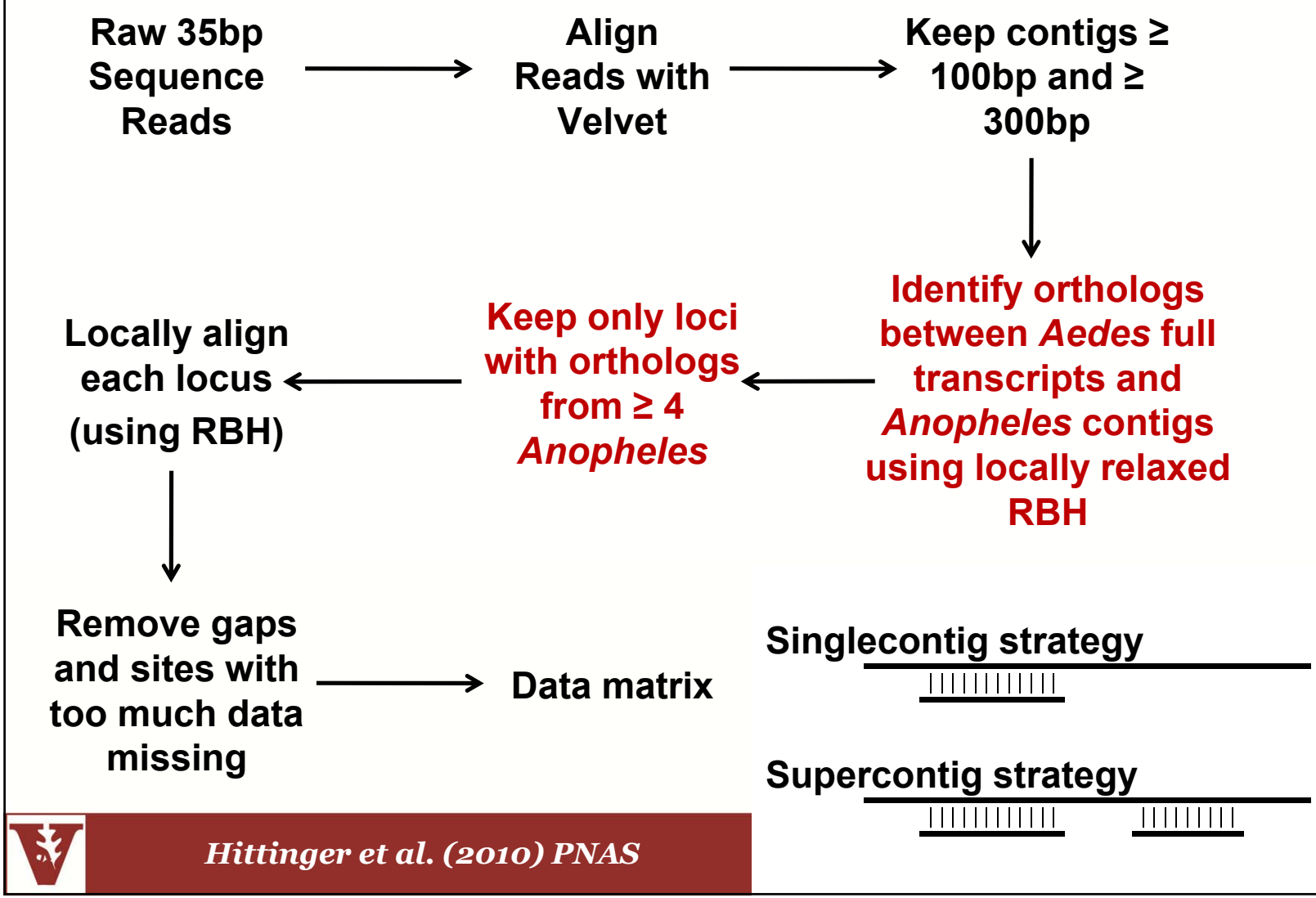
Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Few Sequence Reads



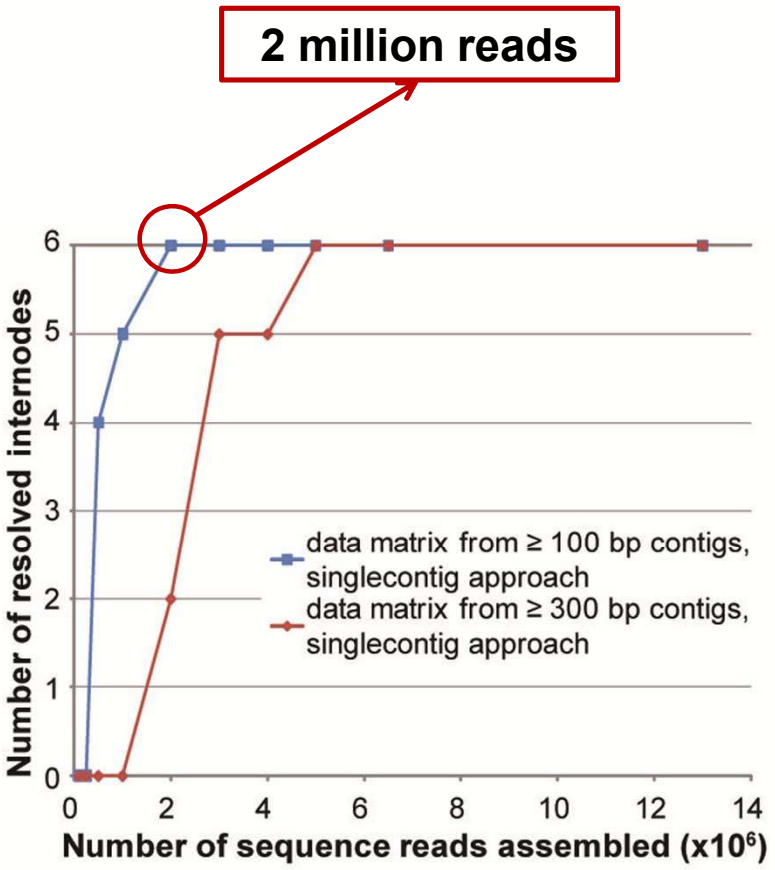
Hittinger et al. (2010) PNAS

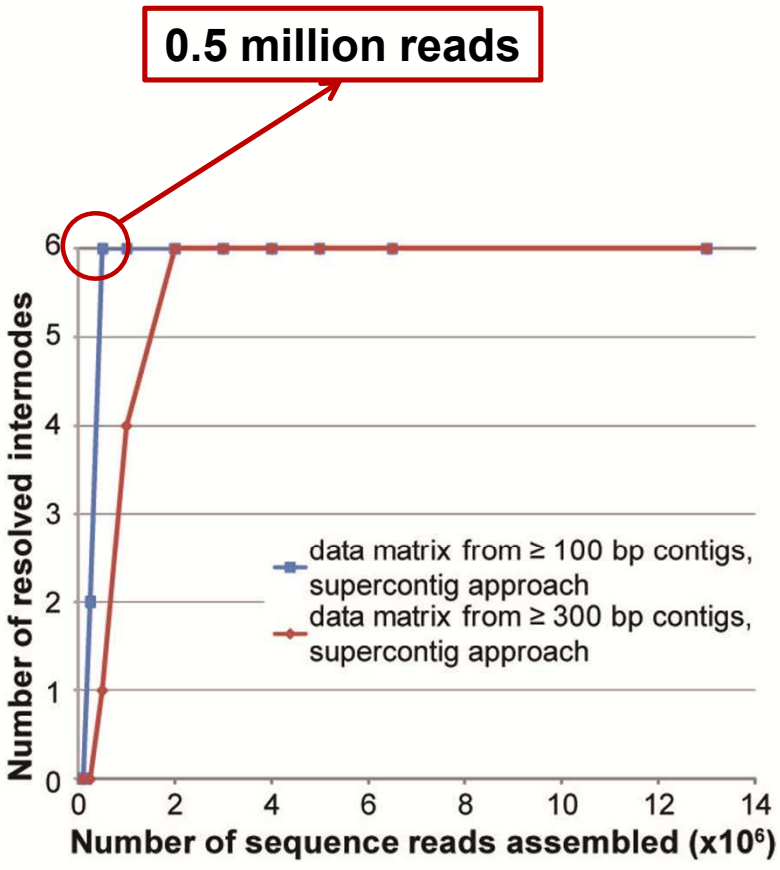
Experimental Design: The "Supercontig" Strategy



Hittinger et al. (2010) PNAS

Robust Phylogenetic Inference From Even Fewer Reads





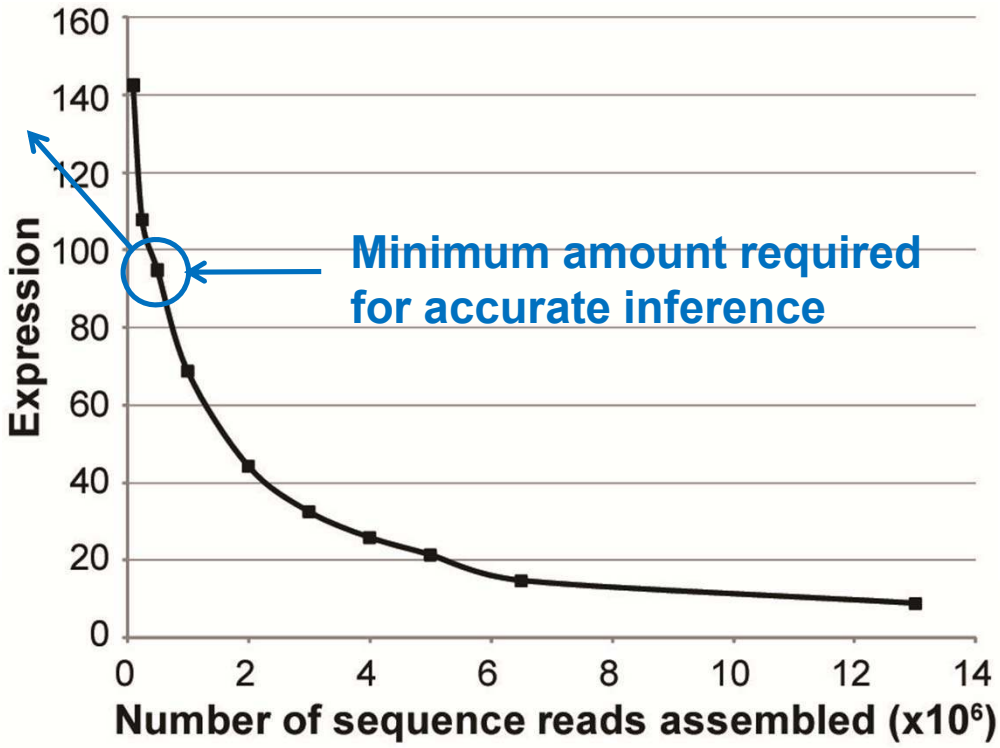
553 loci, AlnL: ~390Kb, %miss: 51

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

Hittinger et al. (2010) PNAS

Our Data Matrices are Derived from Highly-Expressed Transcripts

2008 cost: ~$50
2011 cost: ~$10



Hittinger et al. (2010) PNAS