Genomics at the Broad's Microbial Sequencing Center

Matthew Henn, Ph.D.
Genome Biology Program

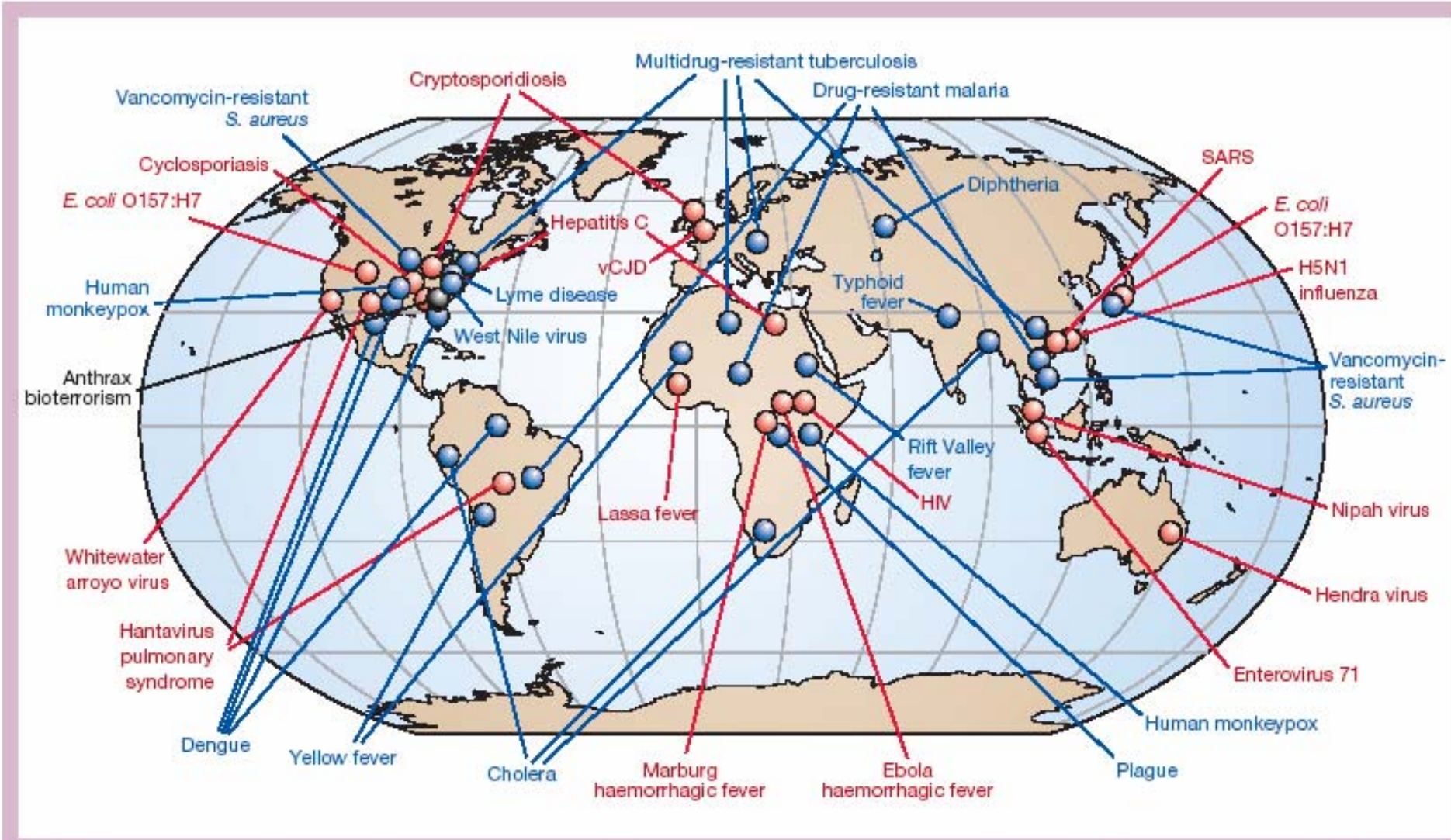
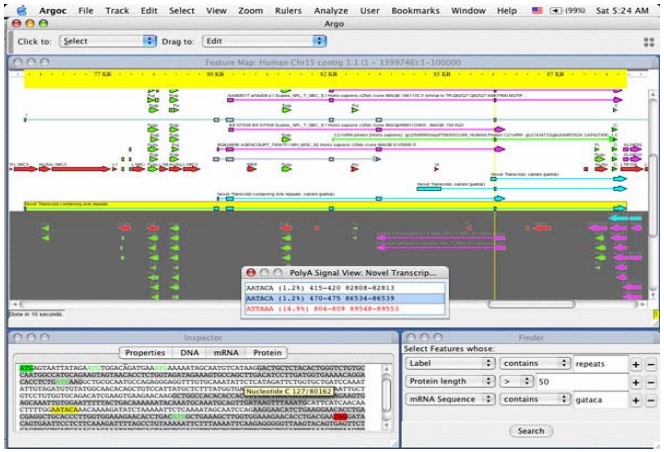


Figure 1 Global examples of emerging and re-emerging infectious diseases, some of which are discussed in the main text. Red represents newly emerging diseases; blue, re-emerging/resurging diseases; black, a 'deliberately emerging' disease. Adapted, with permission, from ref. 23.

Morens, Folkers, & Fauci (2004) Nature

Broad Institute's Microbial Sequencing Center

High-Throughput Genomic Technologies



Computational Genomic Analysis

Collaboration with Research & Clinical Communities

ABI Sequencing Capacity

105M reads/year

>60 Billion Q20 bases/yr

whole genome or directed sequencing at scale

Next Generation Sequencing Technologies

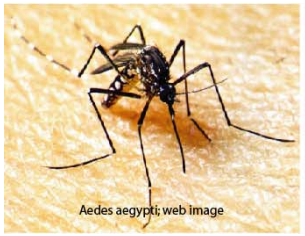
454

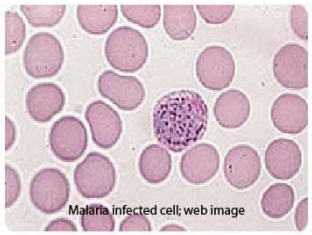
Read length: 100-200b
Capacity: >20-60Mb/run

Solexa

Read length: 25-35b
Capacity: 40Mb-1Gb/run

The Starting Point: Genome Sequences

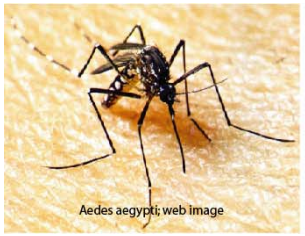


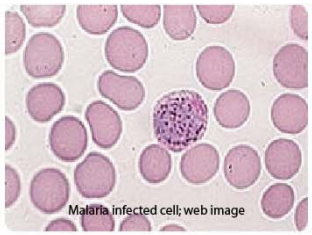


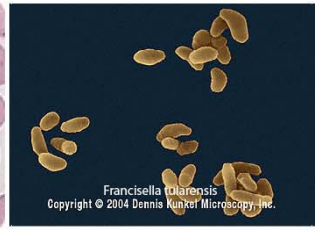
Bacteria

	In Sequencing	Assembled	Annotated
Burkholderia spp.		2	
Enterococcus spp.	~24		
Francisella spp.	6	1	
Listeria monocytogenes	2	16	
Mycobacterium tuberculosis	7	3	2
Pseudomonas aeruginosa			2

The Starting Point: Genome Sequences





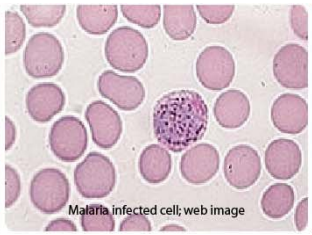


Fungi

	In Sequencing	Assembled	Annotated
*Aspergillus terreus**		1	1
Candida spp.[#]		5	5
Coccidioides spp.*	16	3	1
Dimorphic fungal pathogens*	11	2	2
Cryptococcus neoformans[#]		2	1
Dermatophyte spp.[#]	5		

*NIAID
[#]NHGRI

The Starting Point: Genome Sequences



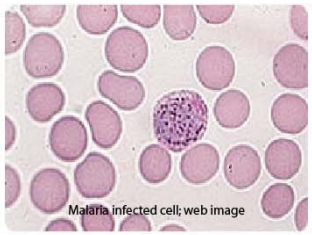
Vectors

	In Sequencing	Assembled	Annotated
Aedes aegypti		1	1
Culex pipiens quinquefasciatus	1		
Ixodes scapularis	1		

Protists

	In Sequencing	Assembled	Annotated
Plasmodium falciparum		11	

The Starting Point: Genome Sequences



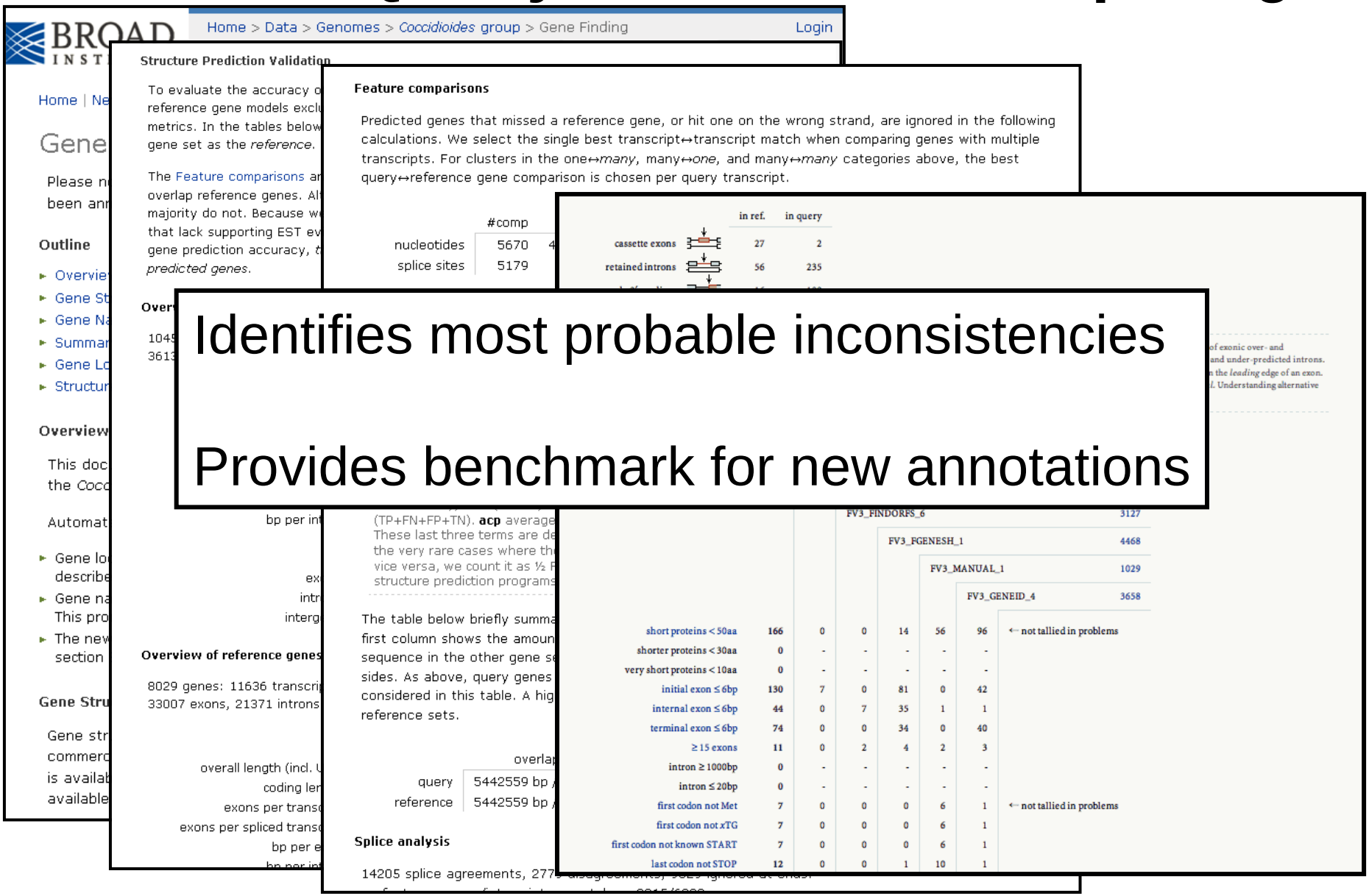
Viruses

	In Sequencing	Assembled	Annotated
Dengue Virus	3300		
Hepatitis C Virus	535		
HIV	2500		
West Nile Virus	TBD		

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT

Annotation needs to scale with sequence data production

Annotation Quality Evaluation and Reporting



Identifies most probable inconsistencies

Provides benchmark for new annotations

Annotation is Dynamic

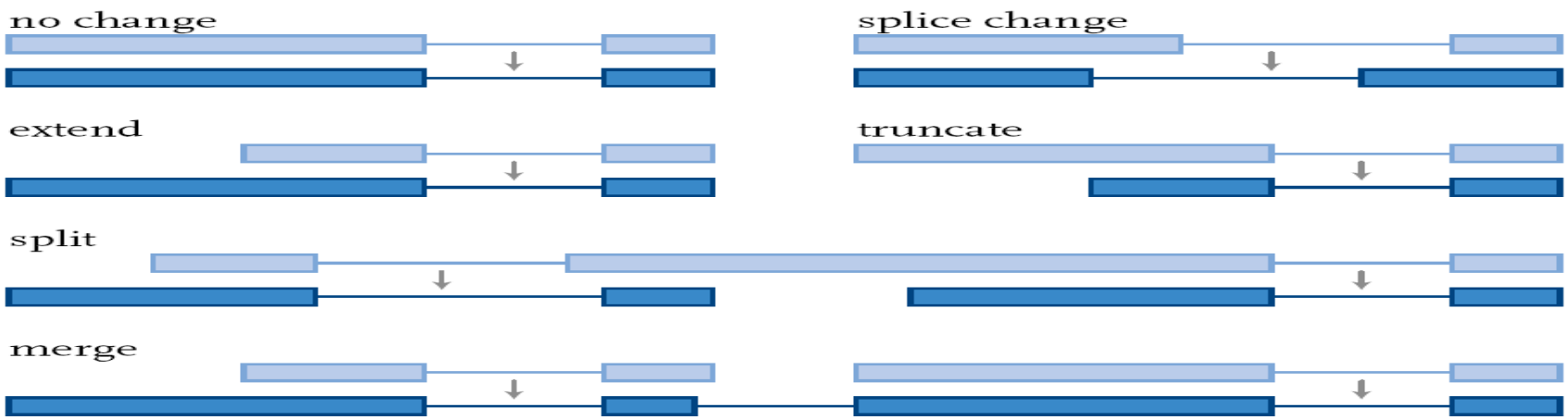
New Evidence Added

Gene Prediction Algorithms Improved

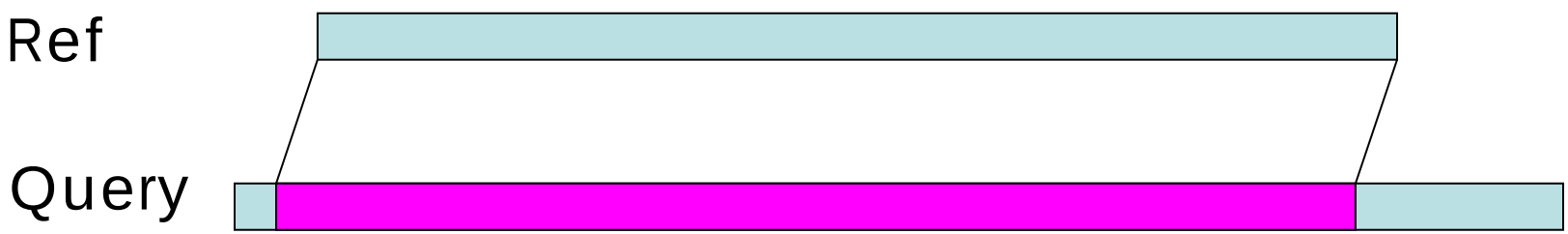
New Sequence Data Generated

Consequences of a Dynamic Annotation World

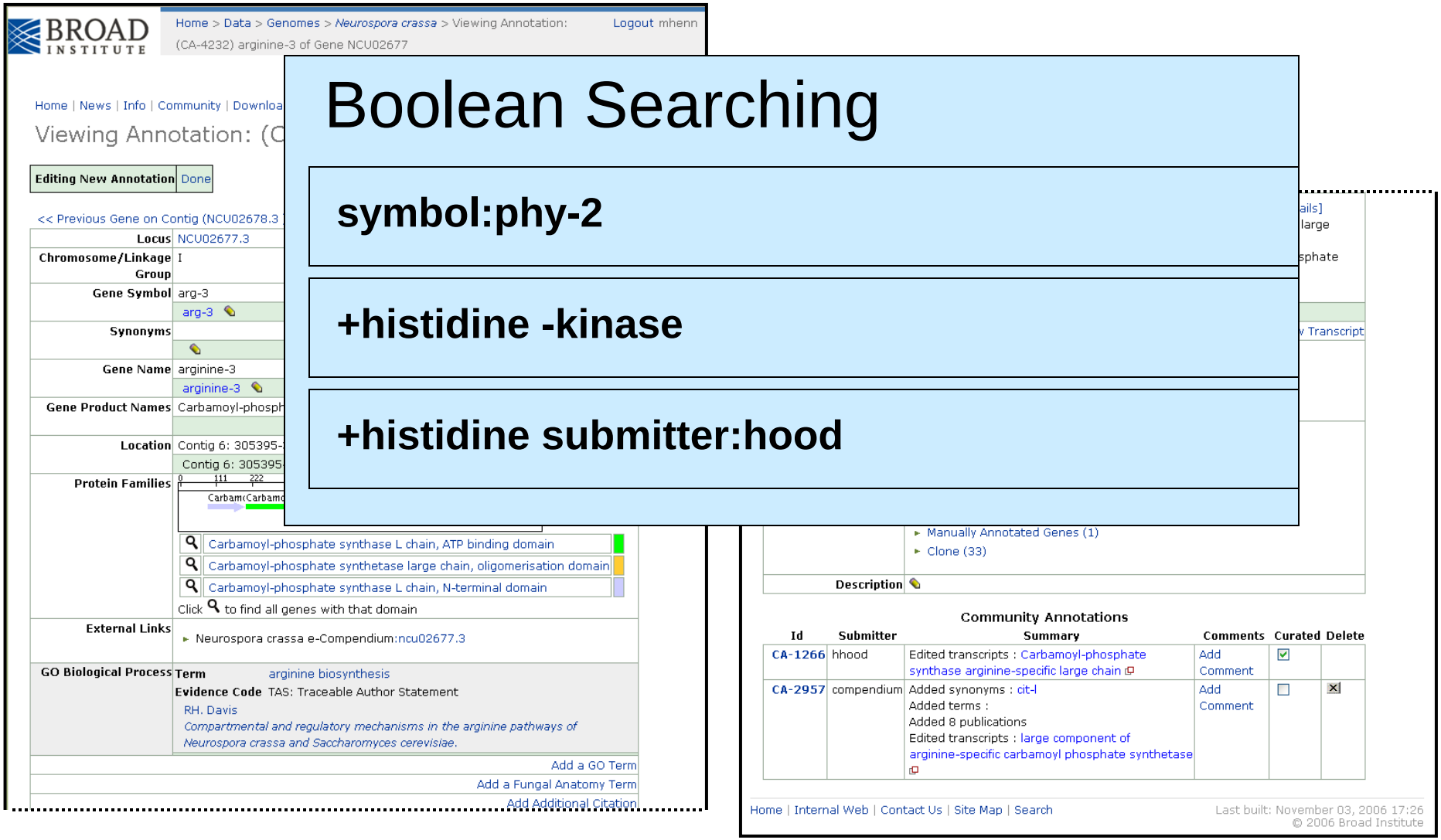
Mapping New Versions



Transfer of Reference Annotations



Improving annotations with the wealth of expert community knowledge



Boolean Searching

symbol:phy-2

+histidine -kinase

+histidine submitter:hood

Moving beyond a single sequence view to comparative genomics…

Use evolution to filter the signal from the noise

CONSERVATION ⇔ FUNCTION ⇔ SELECTION

Project Design to Enable Comparative Genomics

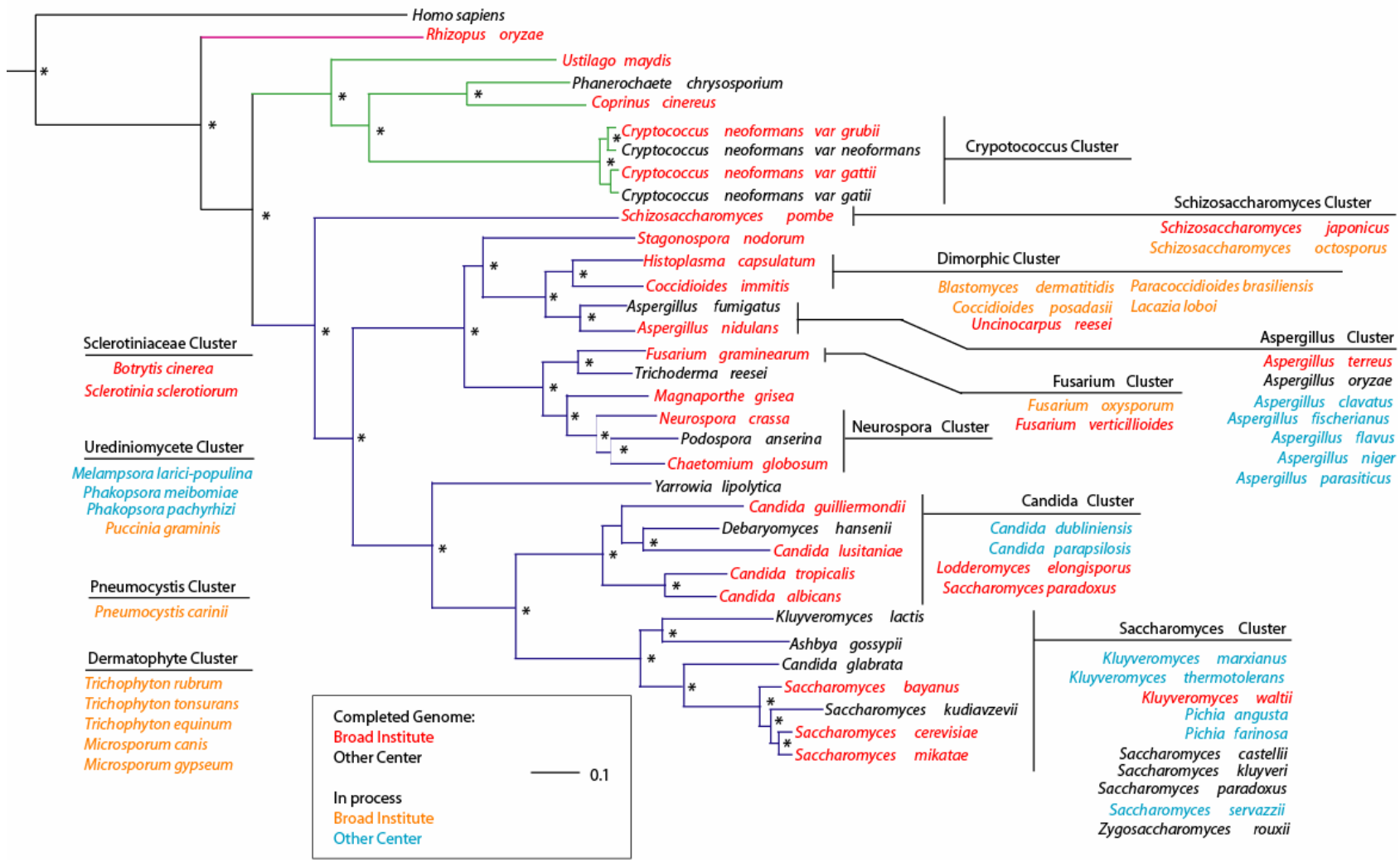


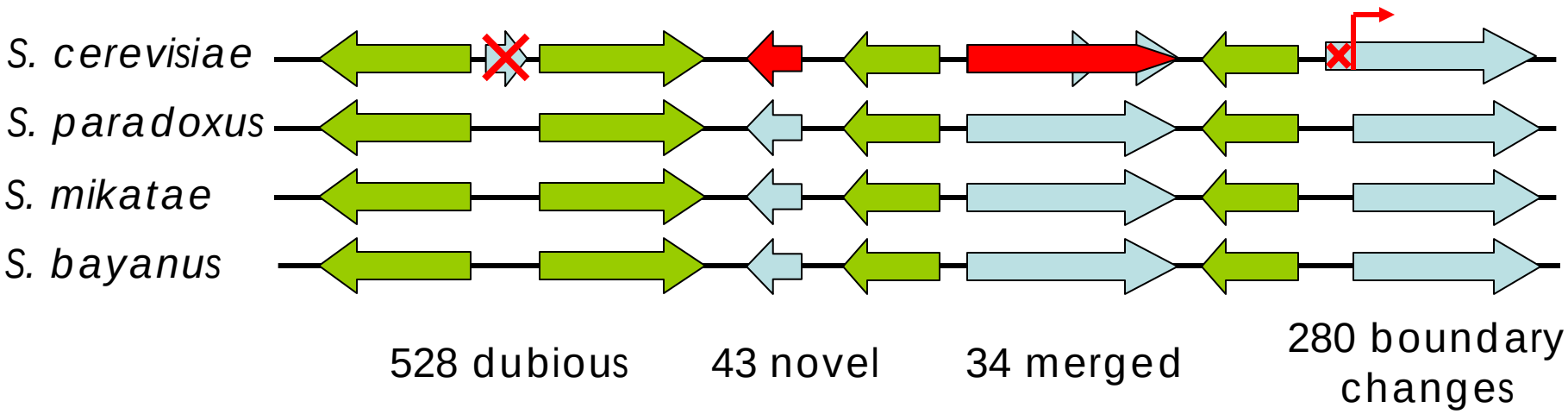
Figure modified from Galagan, Henn, et al. (2005) Genome Research

Comparative approach greatly improves Gene Annotation

Saccharomyces cerevisiae

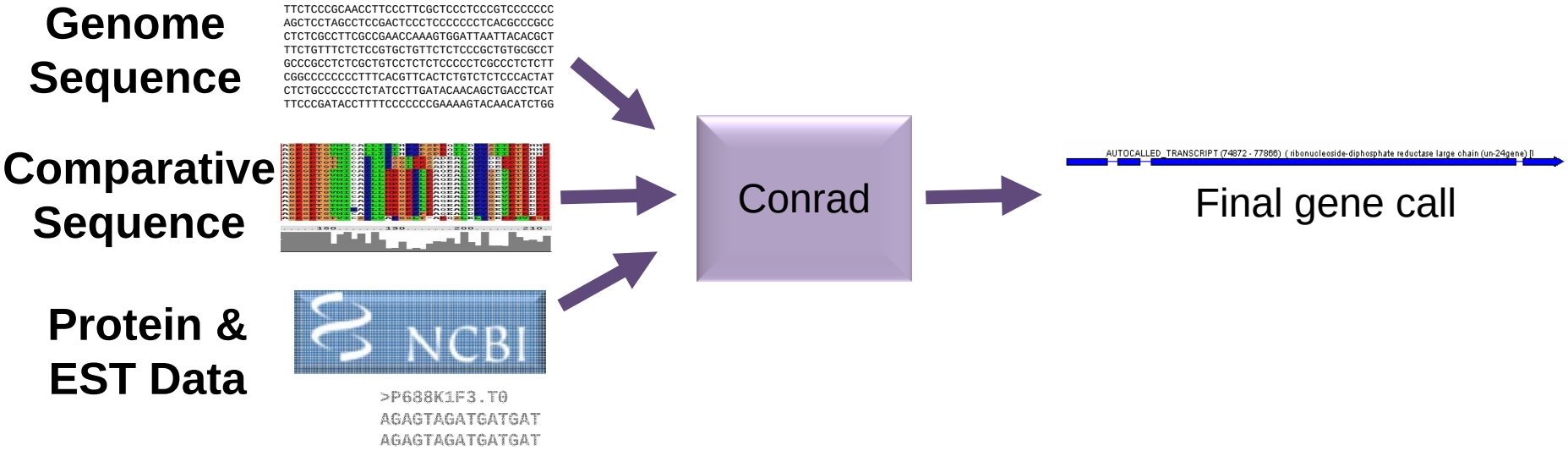
Revisions affect 15% of all genes

6235 previously annotated genes → 5695 'verified' genes



Kellis et al. (2003) Nature

Conrad: Comparative Gene Prediction

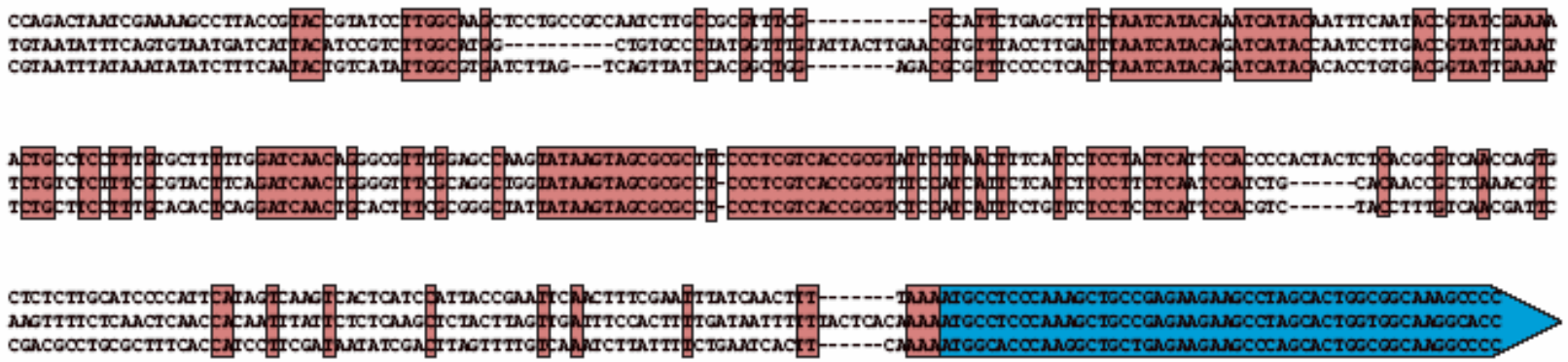


- Uses probabilistic models and non-probabilistic data
- Easily trainable for new data types and genomes

Vinson, DeCaprio, Galagan et al. (2006) NIPS

Regulatory Motif Discovery Using Comparative Genomics

69 Conserved Patterns in Aspergillus



Conpat consensus	Similar sequence	No. 5' sites	No. 3' sites	5' /3' bias?	Aligned (%)	Strand bias?	Enriched in COG categories (no. genes in category)
TCACGTG	bHLH	115	8	up***	35	Yes***	J: Translation, ribosomal structure, biogenesis (16)*
ATCTTATC	*areA/gln3*	25	3	up*	60		
GTTTG... TGTTTG		46	11	up**	35		A: RNA processing, modification (8)
TCACATG	bHLH	36	3	up***	47	Yes*	I: Lipid transport, metabolism (10)**
TGTATA	Puf	5	20	down*	30	Yes***	

Galagan et al. (2005) Nature

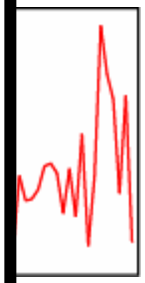
Annotating Regions of Rapid & Slow Evolution

Fusarium graminearum strains:
8x assembly of PH-1 (Broad)

- ID therapeutic targets
- Determine stability of disease loci

SNPs/
NQS kb

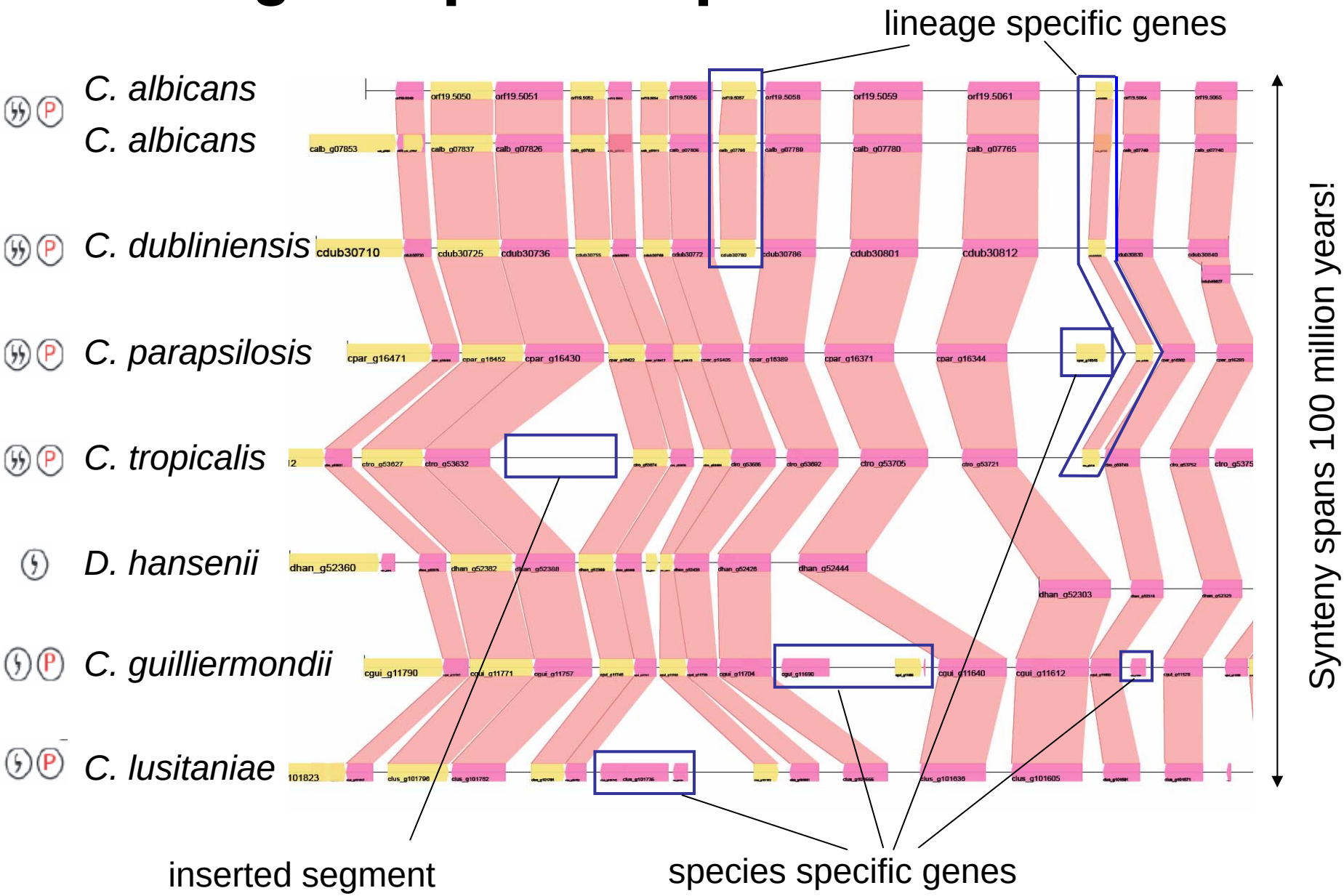
FG specif
genes



chromosome 1 position

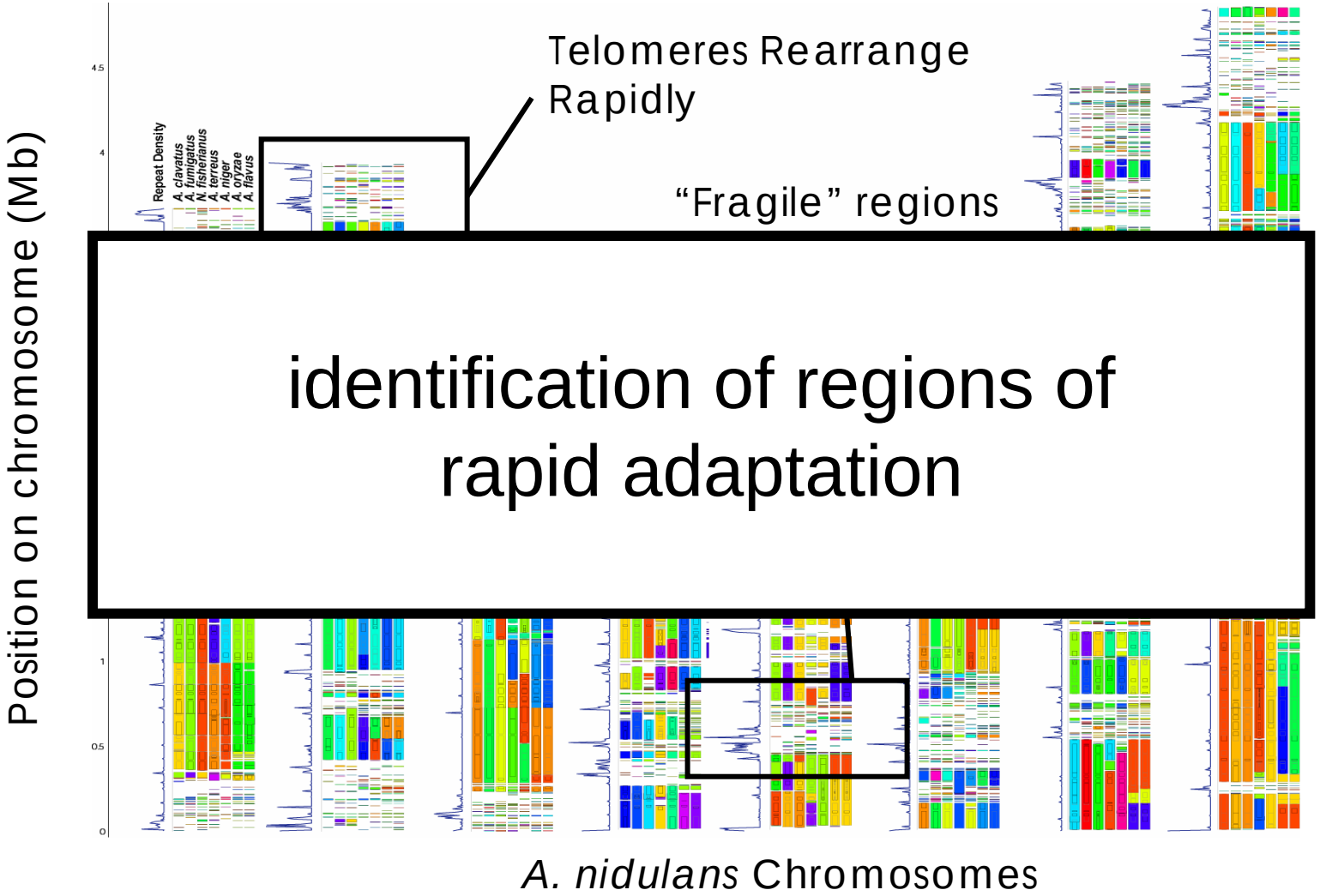
C. Cuomo (unpublished)

Lineage & Species Specific Innovation



Cuomo, Kellis, & Rasmussen (unpublished)

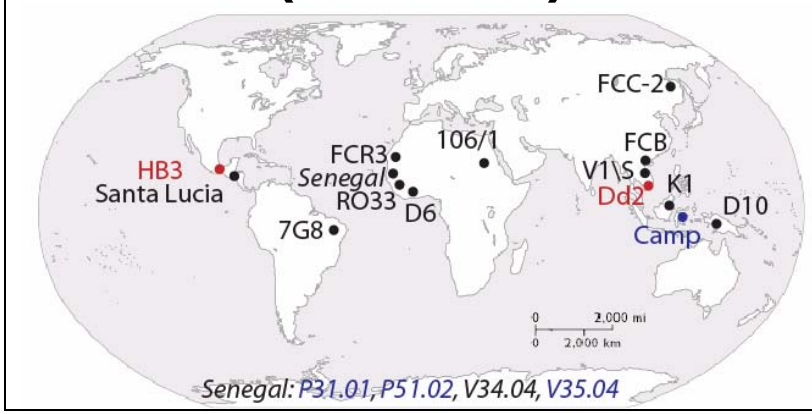
Identifying Regions of Rapid Rearrangement: whole genome alignment of 8 Aspergillus spp.

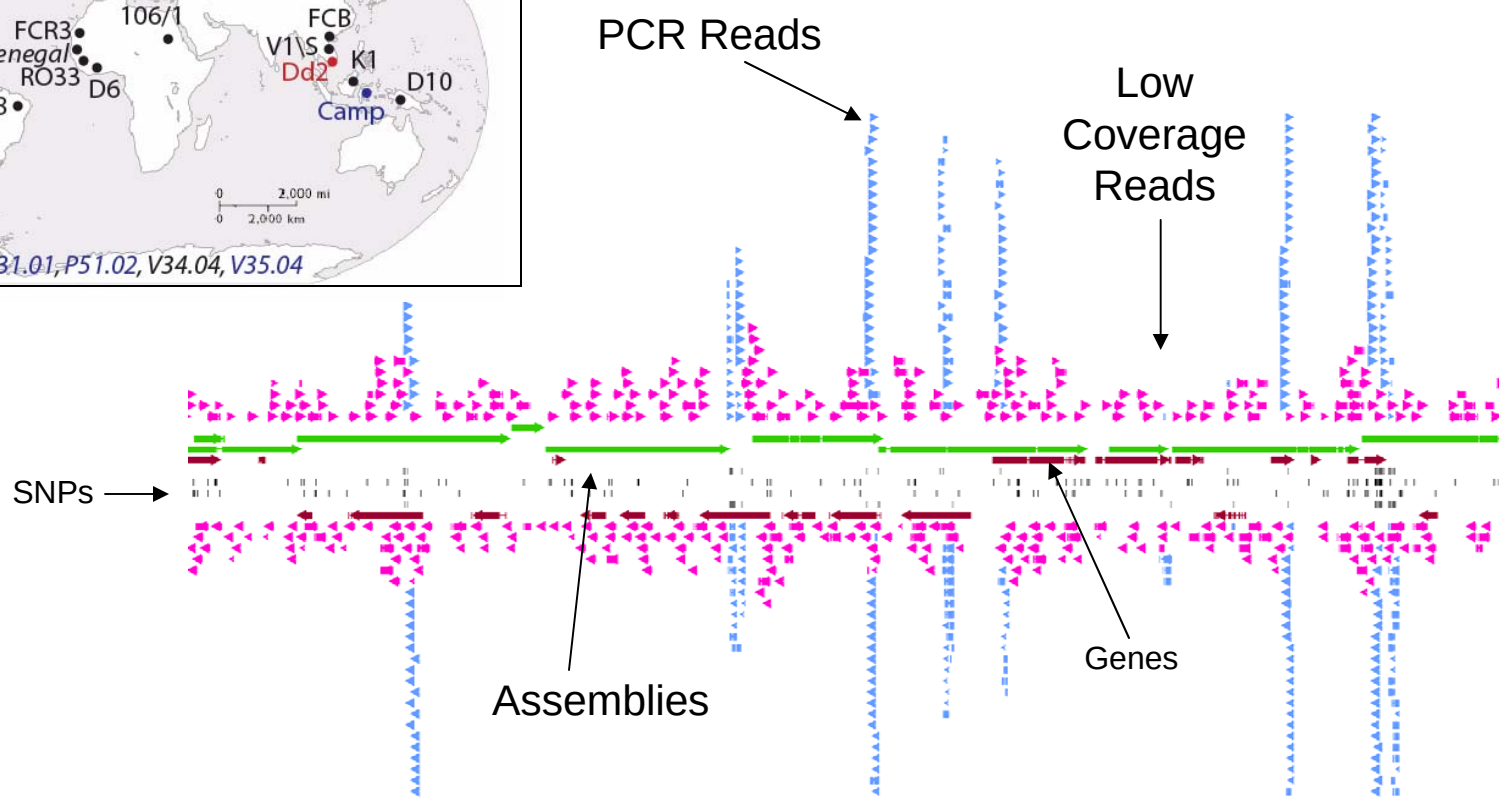


Galagan et al. (2005) Nature

Comparative Genomics at the Population Scale: *Plasmodium falciparum*

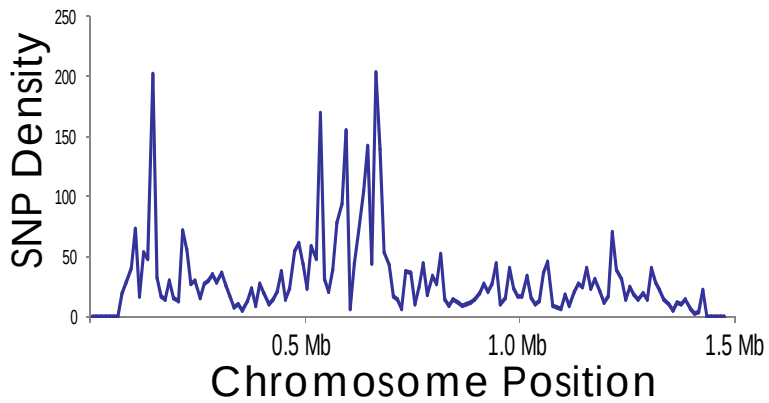
Generate Genomic Sequences (14 Strains)



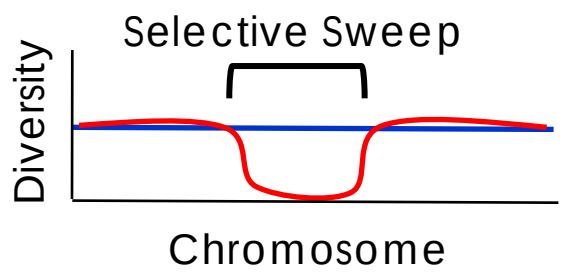


Comparative Genomics at the Population Scale: *Plasmodium falciparum*

Chromosome-Scale Variation



Diversity to Identify Regions of Selection

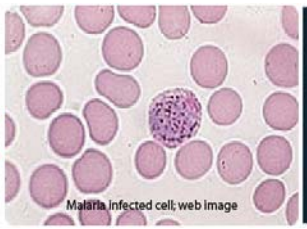


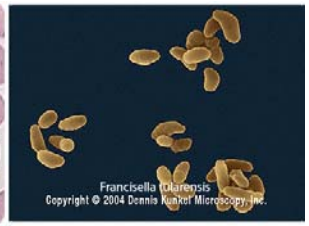
Interrogate Phenotypes

Resistant Sensitive

Volkman, Sabeti, DeCaprio, Neafsey, Schaffner, Galagan, Wirth (2007) Nature Genetics

Conclusions - I





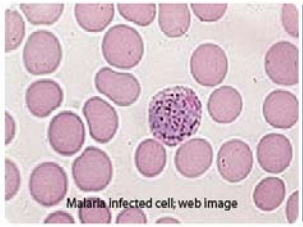
Sequencing

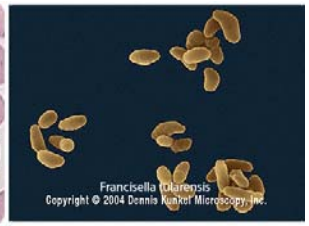
- Large sequencing capacity
- New data types
- Previously infeasible projects are possible

Annotation

- Needs to scale with sequencing
- Dynamic
- More than just genes

Conclusions - II





Comparative Approaches

- Conservation ⇔ Function ⇔ Selection
- Improve & Enable Annotation
- Species and Lineage Specific Innovation
- Identify Regions of Rapid and Slow Evolution
- Bring Genetics to Genomics

Collaborative Projects

- Scientific Planning and Project Development

MSC & FGI Director
Bruce Birren

Assoc. Dir. Microbial Analysis & Annotation
James Galagan

Broad Institute Director
Eric Lander

Genome Biology Program
Bruce Birren (co-Director)
Chad Nusbaum (co-Director)
Keri Devon (Assistant Director)
Christina Cuomo
Carsten Russ
Li-jun Ma
Manolis Kellis
Pardis Sabeti
Manuel Garber
Steve Shaffner
Matt Rasmussen

Assembly Analysis
Sara Young
Kurt LaButti
Sean Sykes

Annotation Informatics
Dave Decaprio (Gr. Leader)
Reinhardt Engels
Philip Montgomery
Matt Crawford
Matt Pearson
Tim Ledlie
Michael Koehrsen
Clint Howarth
Lisa Larson
Jared White
Danny Park

Microbial Analysis
Mark Borowsky
Antonis Rokas
Dan Neafsey
Desmund Lun
Aaron Brandes
Brian Weiner

Microbial Annotation
Chinnappa Kodira (Gr. Leader)
Qiandong Zeng
Chandri Yandava
Sinead O'leary
Lucia Alvarado

Whole Genome Assembly
David Jaffe (Director)
Jonathan Butler
Pablo Alvarez
Sante Gnerre
Manfred Grabherr
Michael Kleber
Evan Mauceli
William Brockman
Iain McCallum

Genome Sequencing Platform
Jen Baldwin (Director)
Rob Nicol (Dir. Operations)
Pat Cahill
Nicole Stange-Thomann
Niall Lennon
Georgia Giannoukos
Will Lee
Mike Weiand

Infectious Disease Initiative
Roger Weigand (Director)
Dyann Wirth
Sarah Volkmann

BROAD INSTITUTE

National Institute of Allergy and Infectious Diseases
National Institutes of Health

BILL & MELINDA GATES foundation