Enabling Large-scale Next-generation Sequence Assembly With Blacklight

Philip Blood, Pittsburgh Supercomputing Center
Brian Couger, Oklahoma State University
Lenore Pipes, Cornell University

Acknowledgements

- PSC
  - Deb Nigra and Rick Costa
  - Sergiu Sanieleivici
  - Derek Simmel
  - J Ray Scott and Brian Johanson
- Oklahoma State
  - Brian Couger
- Cornell
  - Lenore Pipes and Chris Mason

GENOMICS?

METAGENOMICS?

TRANSCRIPTOMICS?

METATRANSCRIPTOMICS?

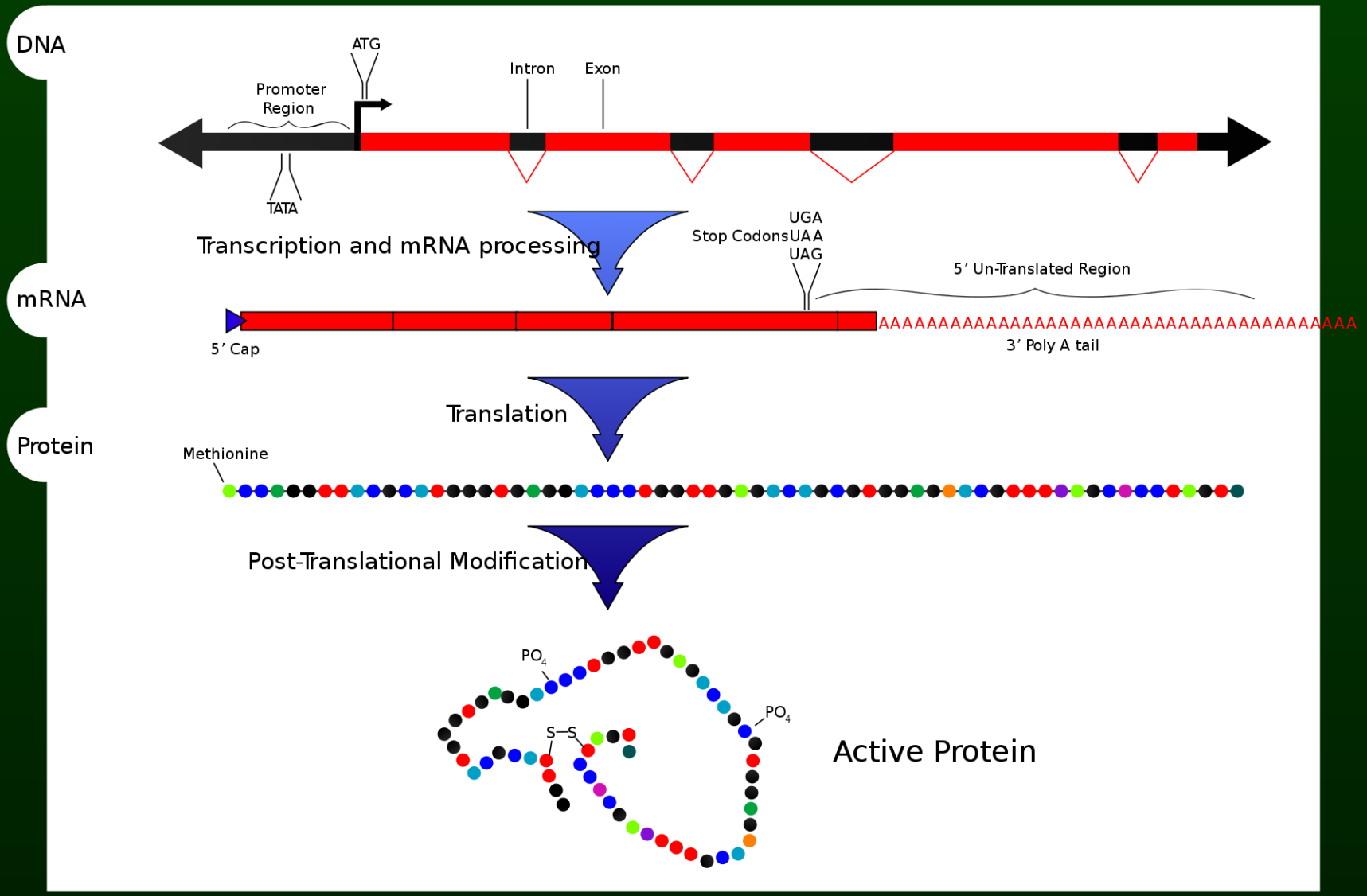
http://www.ark-genomics.org/badomics-generator

PHARMACOMICROBIOMICS?

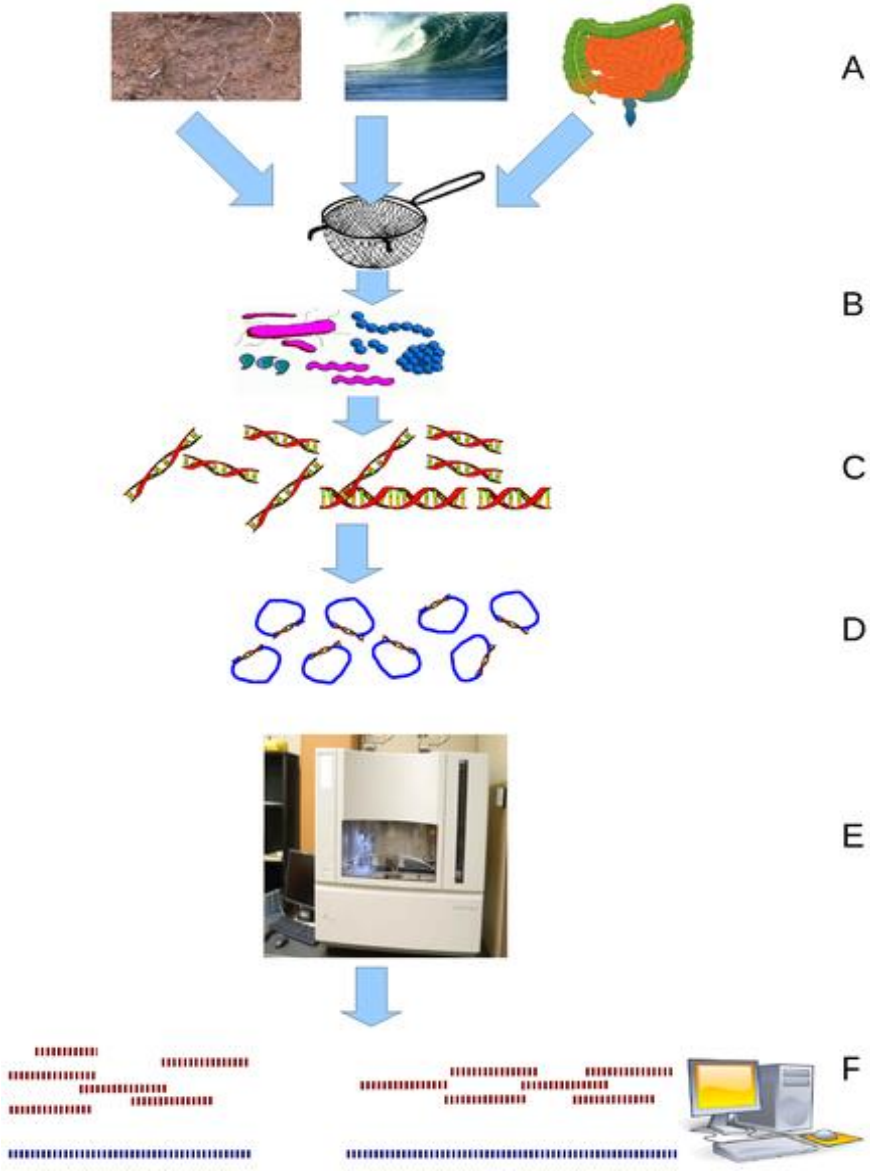
*Blinov ML, Moraru II. Logic modeling and the ridiculome under the rug. BMC Biol. 2012 Nov 14;10:92. doi: 10.1186/1741-7007-10-92.

RIDICULOMICS*?

Central Dogma of Molecular Biology : Eukaryotic Mode



Metagenomics: Environmental Shotgun Sequencing



A **Sampling from habitat**

B **filtering particles, typically by size**

C **DNA extraction and lysis**

D **cloning and library**

E **sequence the clones**

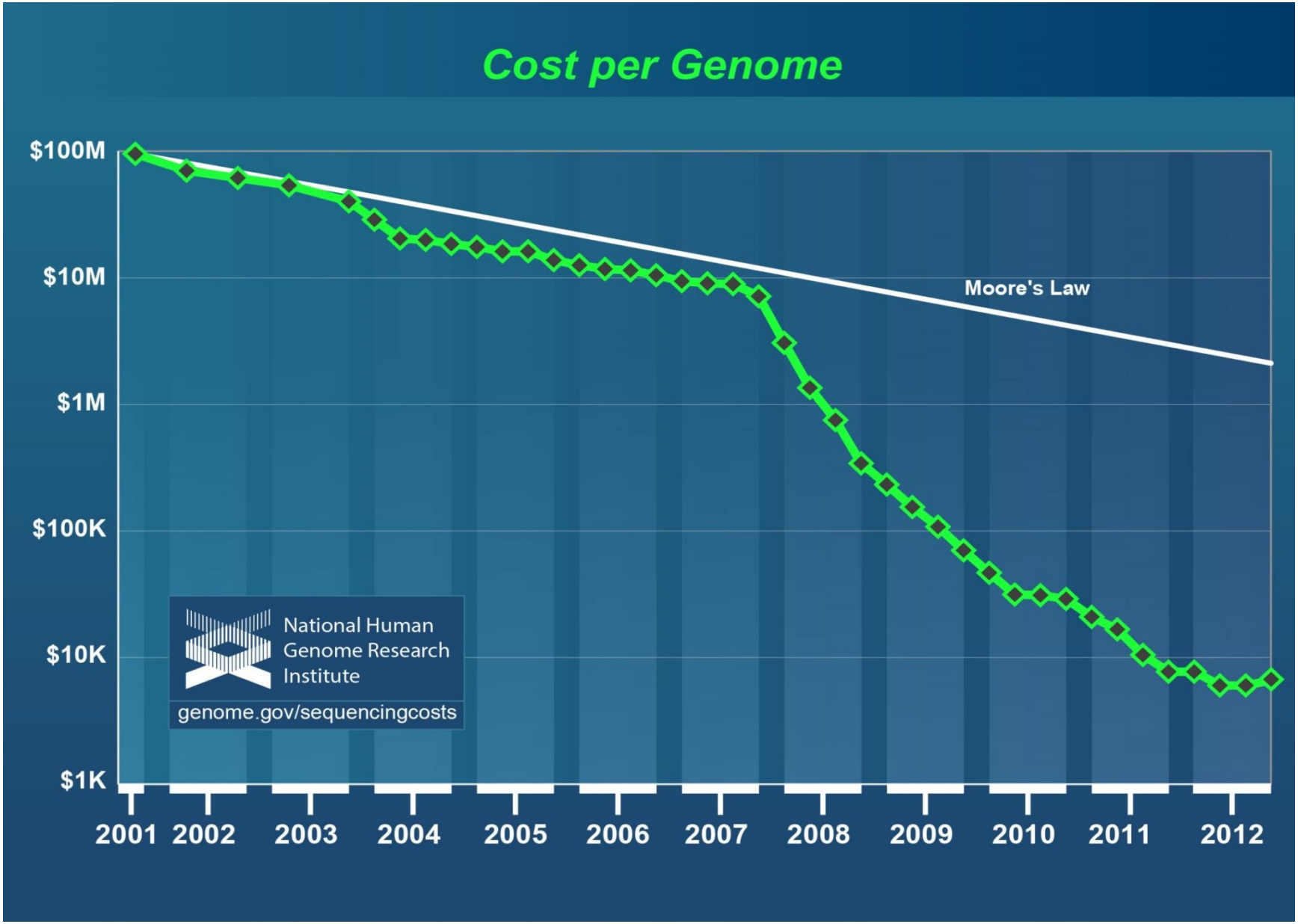
F **sequence assembly**

Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. doi:10.1371/journal.pcbi.1000667
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000667

PLOS | COMPUTATIONAL BIOLOGY

Why do you care?

- Metagenomics
  - Create new sources of energy
    - Biofuels
  - Human health
    - Evidence that diabetes and Crohn's disease related to the bacteria in your gut (microbiome)
    - Transplanting bacteria from one person to another cures "incurable" hospital bacterial infection (C. difficile)
- Genomics and Transcriptomics
  - Human health
    - Find genes/factors controlling disease/health
  - Agriculture
    - Drought resistant crops – dealing with climate change


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

643 HiSeqs = 6.5 Pb/year

http://omicsmaps.com/

http://www.illumina.com/systems/hiseq_2000.ilmn

Genomics analysis: two basic flavors

- Loosely-coupled problems

  Sequence alignment and Variant Calling: Read many short DNA sequences from disk and map to a reference genome

  – Lots of disk I/O

  – Fits well with MapReduce framework

- Tightly-coupled problems

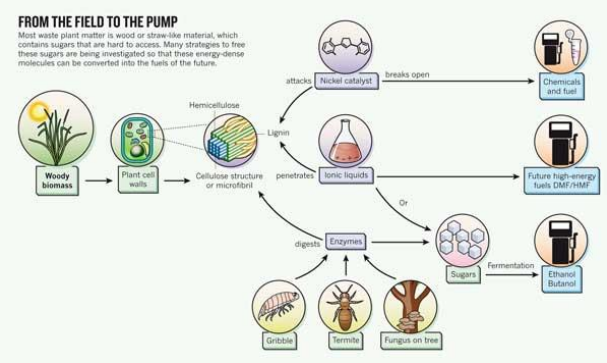
  De novo assembly: Assemble a complete genome from short genome fragments generated by sequencers

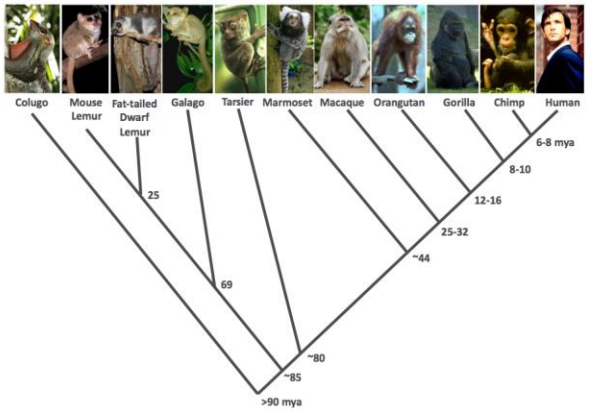
  – Primarily a large graph problem

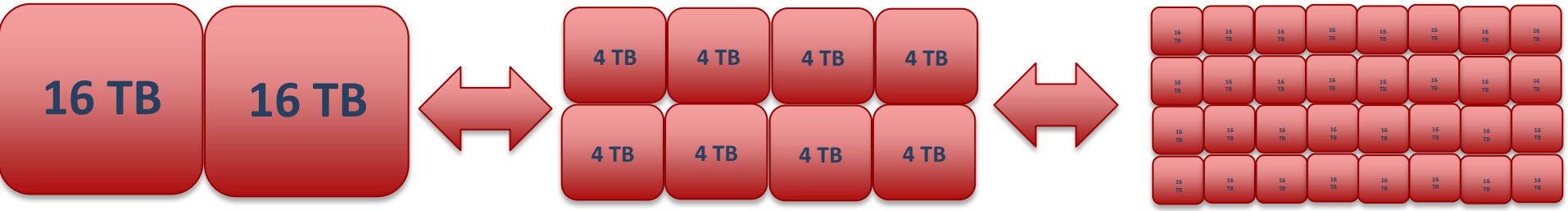
  – Works best with a lot of shared memory

Blacklight: SGI Altix UV 1000
2×16 TB of cache-coherent shared memory
4096 cores

Tackling some of the most challenging problems in genomics with Blacklight







2 x 16 TB RAM:

Capability

8 x 4 TB RAM

Capability + Throughput

32 x 1 TB RAM

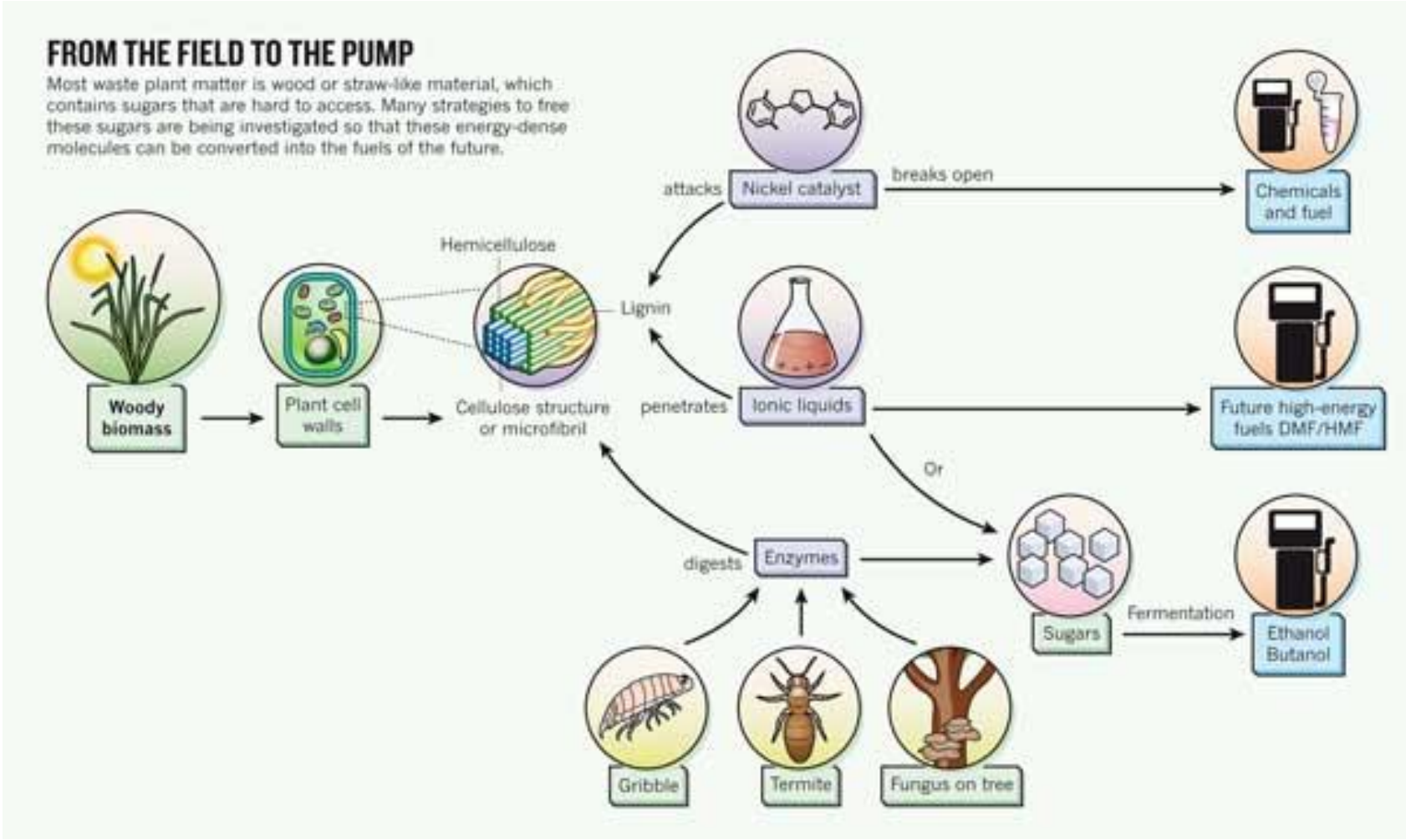
Large Memory + High Throughput

nature

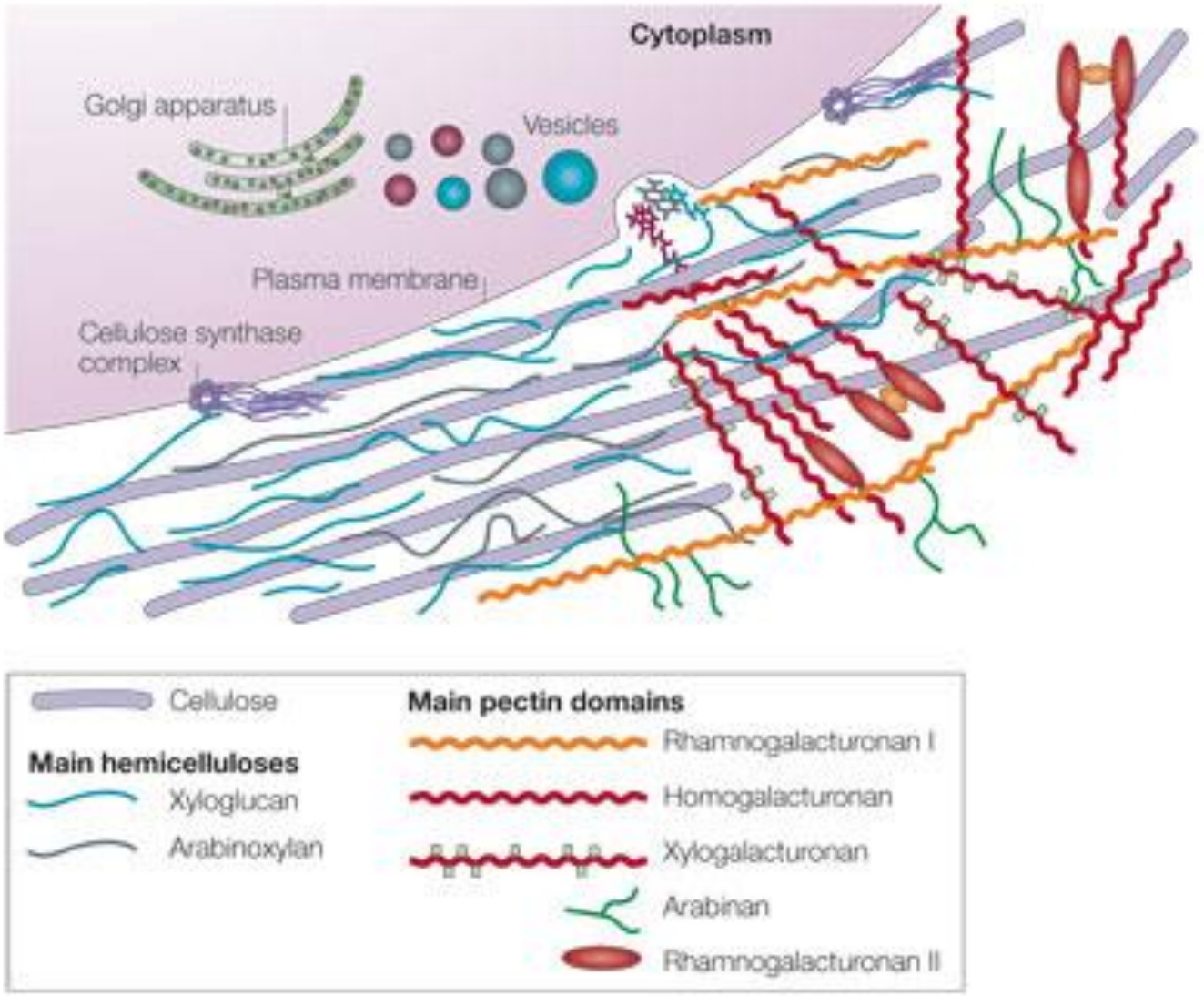
Lignocellulose: A chewy problem

Katharine Sanderson

Nature **474**, S12–S14 (23 June 2011) | doi:10.1038/474S012a
Published online 22 June 2011

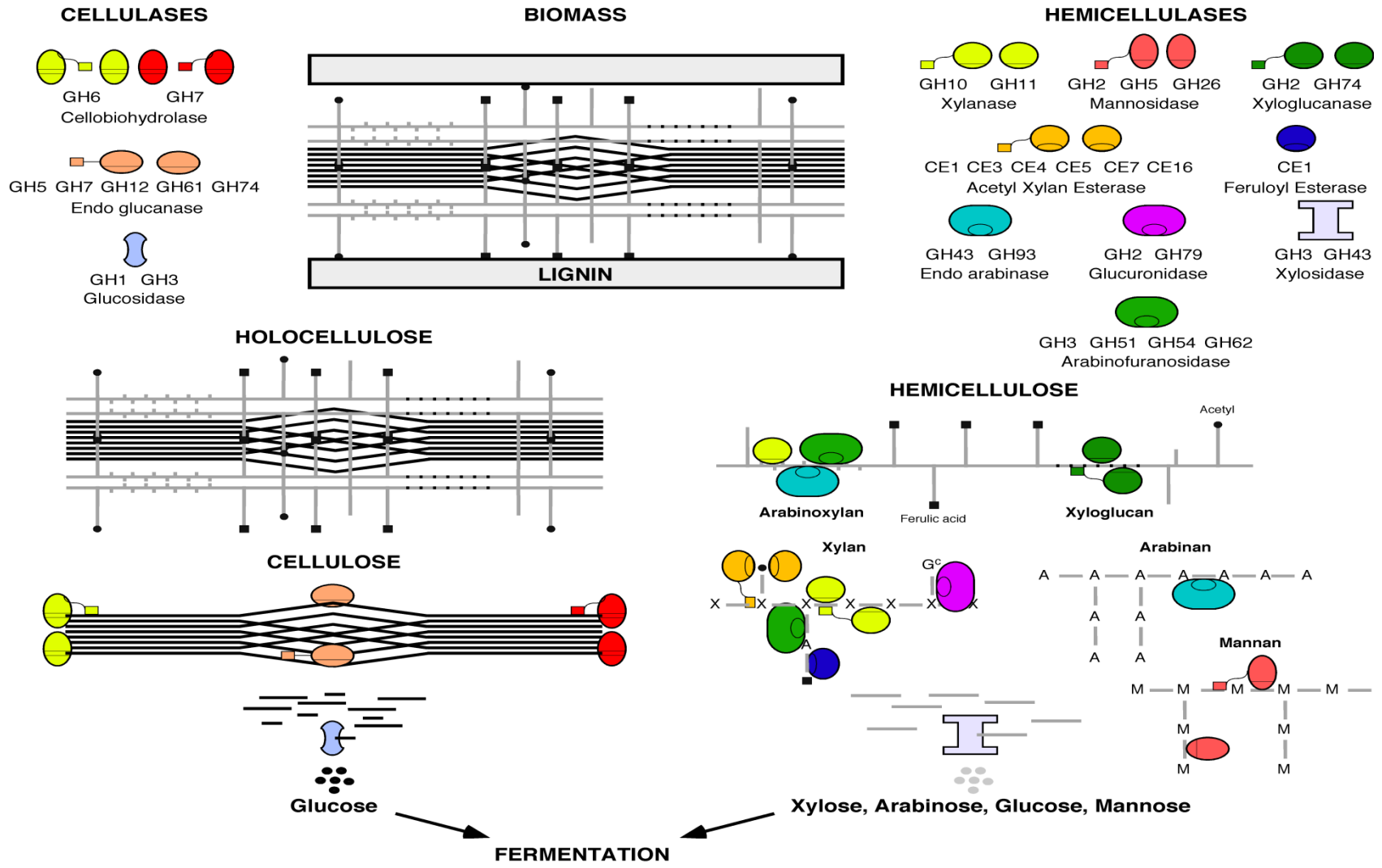


Source: Google Image



Daniel J. Cosgrove

Nature Reviews Molecular Cell Biology **6**, 850-861



nature REVIEWS MICROBIOLOGY

Nature Reviews Microbiology **7**, 756 (October 2009) | doi:10.1038/nrmicro2119-c1

Soil genomics

Brajesh K. Singh[1], Colin D. Campbell[1], Soren J. Sorenson[2] & Jizhong Zhou[3]

constitute 60% of the Earth's biomass. A current global estimate suggests that soil contains 4–5 × 10^{30} microbial cells (excluding viruses), 10 times more than the seas. In addition to reservoirs of industrial products worth UK£100s of billions, microorganisms play vital parts in biogeochemical cycling and sustainability[2]. Therefore, understanding microbial community structure, diversity, functions and stability is essential to our understanding of evolution, community formation and sustainability of life on the Earth. However, obtaining this information has been difficult, owing to our inability to grow microorganisms in laboratory conditions. It is estimated that >99% of microorganisms are currently unculturable under laboratory conditions[3]. Recent developments in technologies such as metagenomics offer a real opportunity for discoveries in the fundamental science of evolution and community formation.

nature REVIEWS MICROBIOLOGY

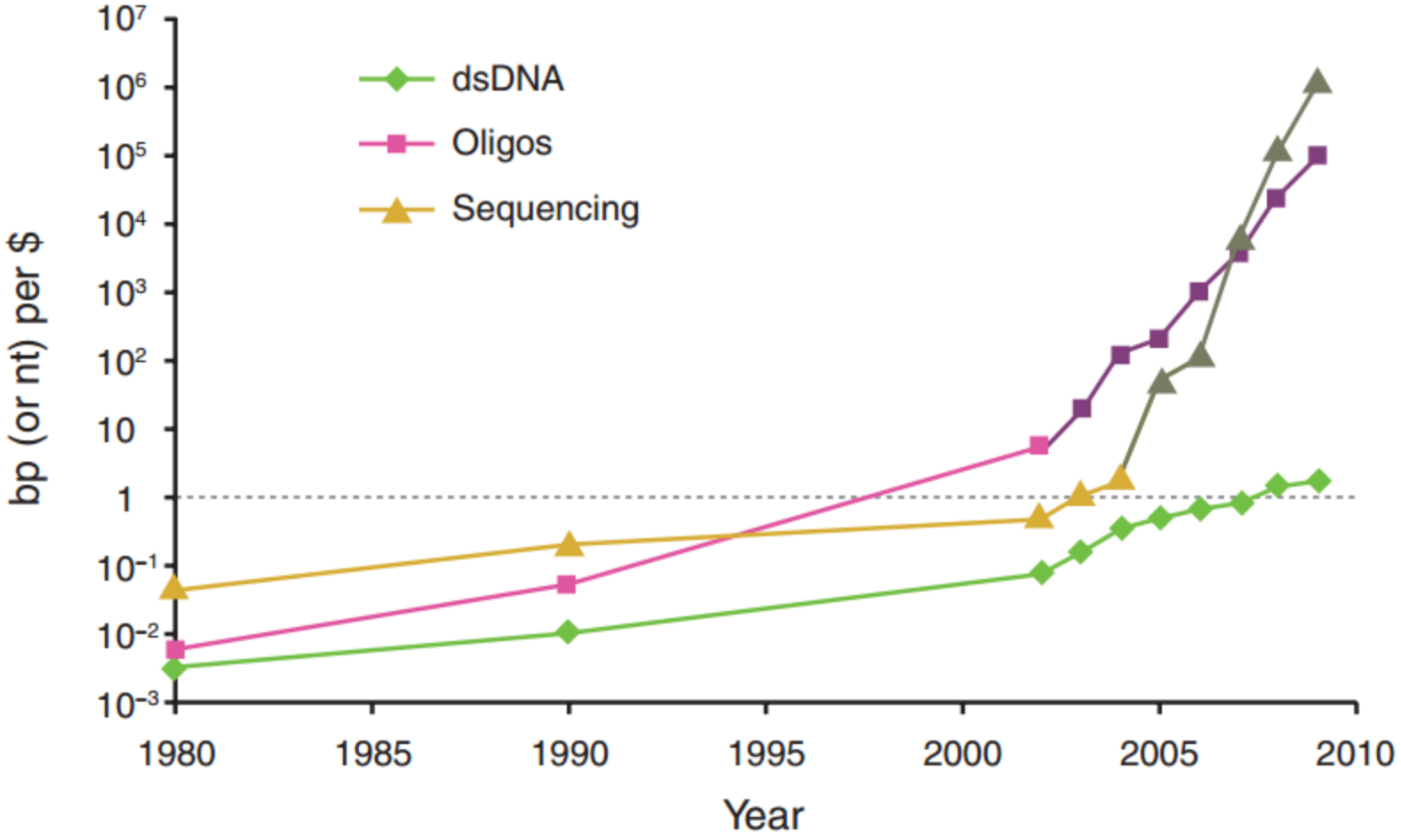
Nature Reviews Microbiology **7**, 252 (April 2009) | doi:10.1038/nrmicro2119

TerraGenome: a consortium for the sequencing of a soil metagenome

See also: Correspondence by Brajesh K. Singh *et.al* | Correspondence by Philippe C. Baveye | Author's reply by Timothy M. Vogel *et.al*

Timothy M. Vogel[1], Pascal Simonet[1], Janet K. Jansson[2], Penny R. Hirsch[3], James M. Tiedje[4], Jan Dirk van Elsas[5], Mark J. Bailey[6], Renaud Nalin[7] & Laurent Philippot[8]

Soil is the most biodiverse environment on the Earth: it is estimated to contain approximately 1,000 Gbp of microbial genome sequences per gram of soil! Compared with the Human Genome project (in which 3 Gbp were sequenced)[1] and sequencing projects that target microbial habitats, such as the Sargasso Sea (for which6 Gbp were sequenced)[2], metagenomic sequencing of soil remains rudimentary and constitutes a new and ambitious challenge. We propose that soil should be our next global metagenomic sequencing initiative.



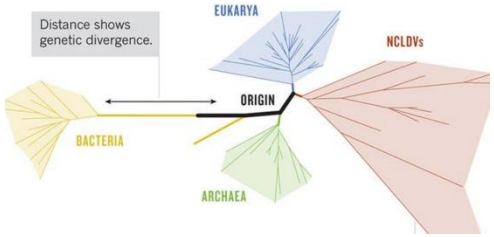
Genome engineering

Peter A Carr[1] & George M Church[2]

Nature Biotechnology **27**, 1151 - 1162 (2009)
Published online: 9 December 2009 | doi:10.1038/nbt.1590

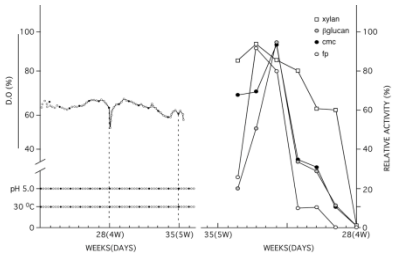
Synthetic Ecosystem Evolution

Complex Microbial Founder Community



Nature **476**, 20-21 (2011)

Experimenter Selected for Metabolic Potential



Evolutionary Pressure Applied for Enrichment of Desired Metabolic Function

Community Sequencing and Assembly

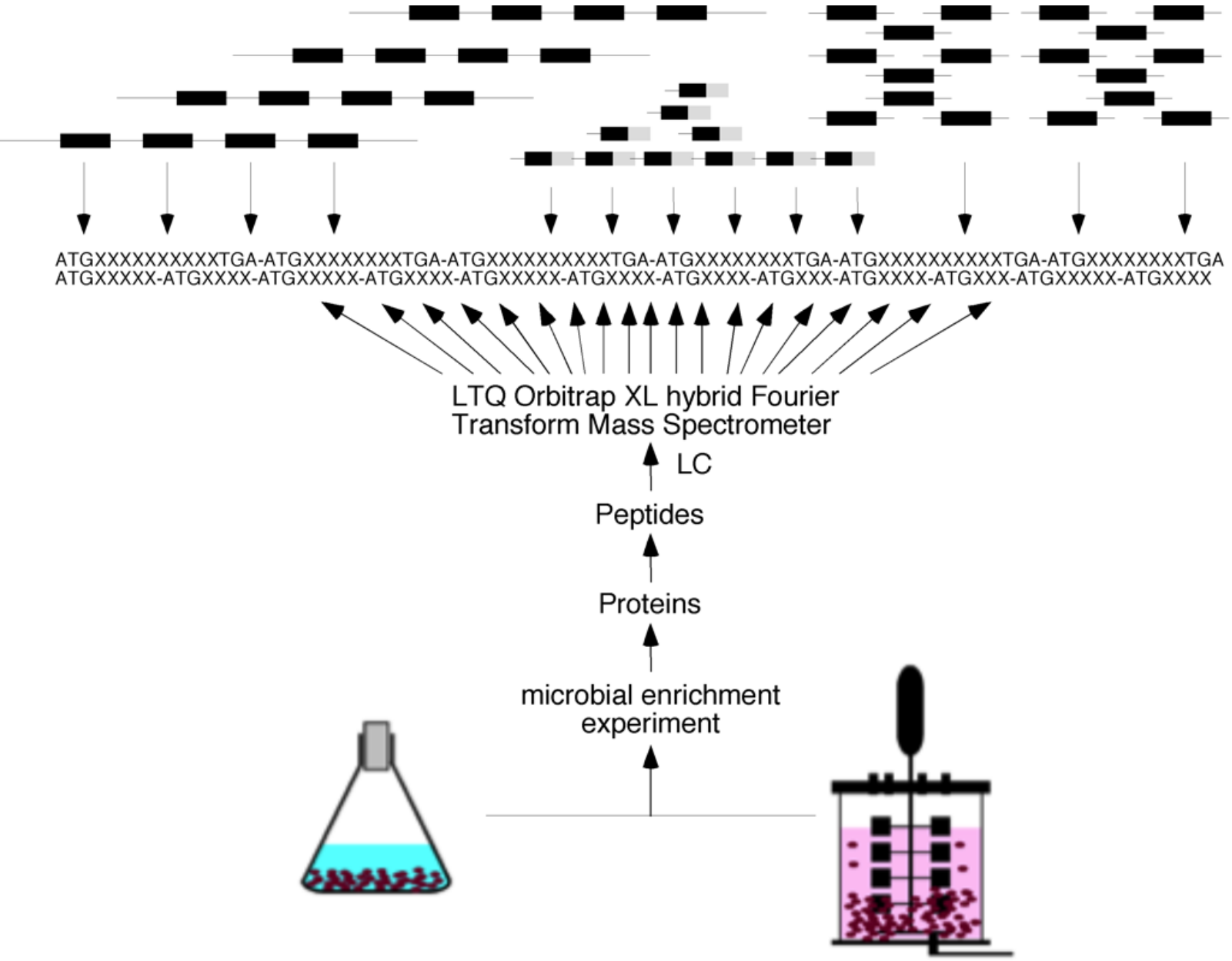
illumina

PITTSBURGH SUPERCOMPUTING CENTER

XSEDE

Velvet

the HOARD memory allocator


ATGXXXXXXXXXXTGA-ATGXXXXXXXXTGA-ATGXXXXXXXXXXTGA-ATGXXXXXXXXTGA-ATGXXXXXXXXXXTGA-ATGXXXXXXXXTGA
ATGXXXXX-ATGXXXX-ATGXXXXX-ATGXXXX-ATGXXXXX-ATGXXXX-ATGXXXX-ATGXXX-ATGXXXX-ATGXXX-ATGXXXXX-ATGXXXX
LTQ Orbitrap XL hybrid Fourier Transform Mass Spectrometer
LC
Peptides
Proteins
microbial enrichment experiment

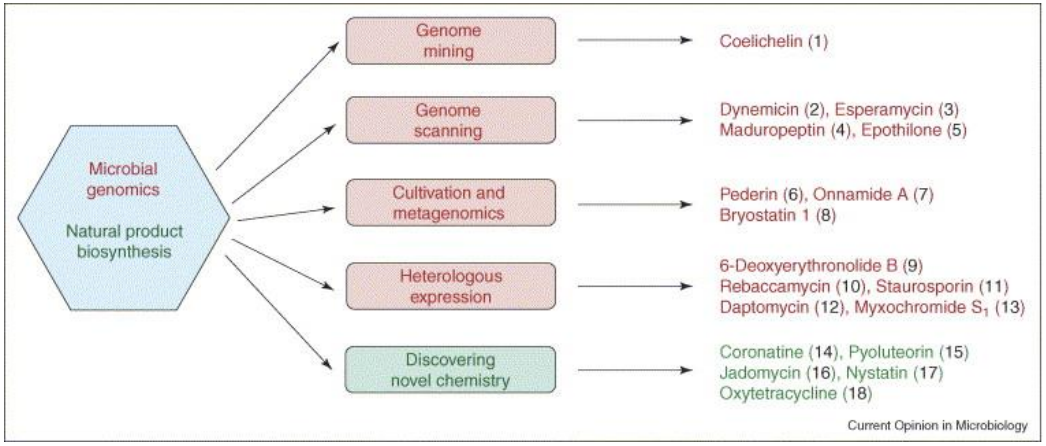
Current Opinion in Microbiology

Volume 9, Issue 3, June 2006, Pages 252–260

Microbial genomics for the improvement of natural product discovery

Steven G Van Lanen, Ben Shen

School of Pharmacy, University of Wisconsin-Madison, 777 Highland Avenue, Madison, WI 53705, USA



“Natural products remain a consistent source of drug leads with more than 40% of new chemical entities reported since 1981 being derived from microbial natural products. Perhaps more astonishing is that more than 60% of the anticancer and 70% of the anti-infective antibiotics currently in clinical use are natural products or natural product-based. ”

Current Opinion in Biotechnology

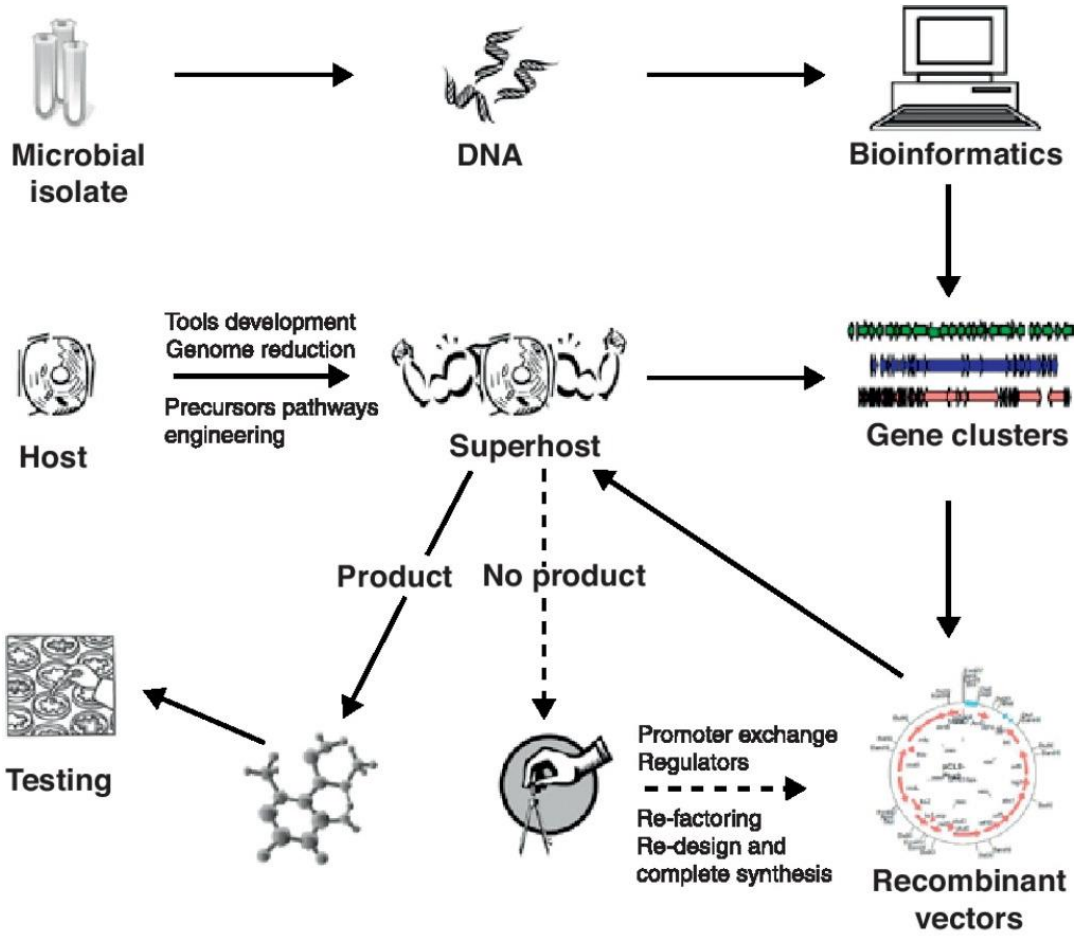
Available online 2 May 2012

Genome-based bioprospecting of microbes for new therapeutics

Sergey B Zotchev[1], Olga N Sekurova[1], Leonard Katz[2]

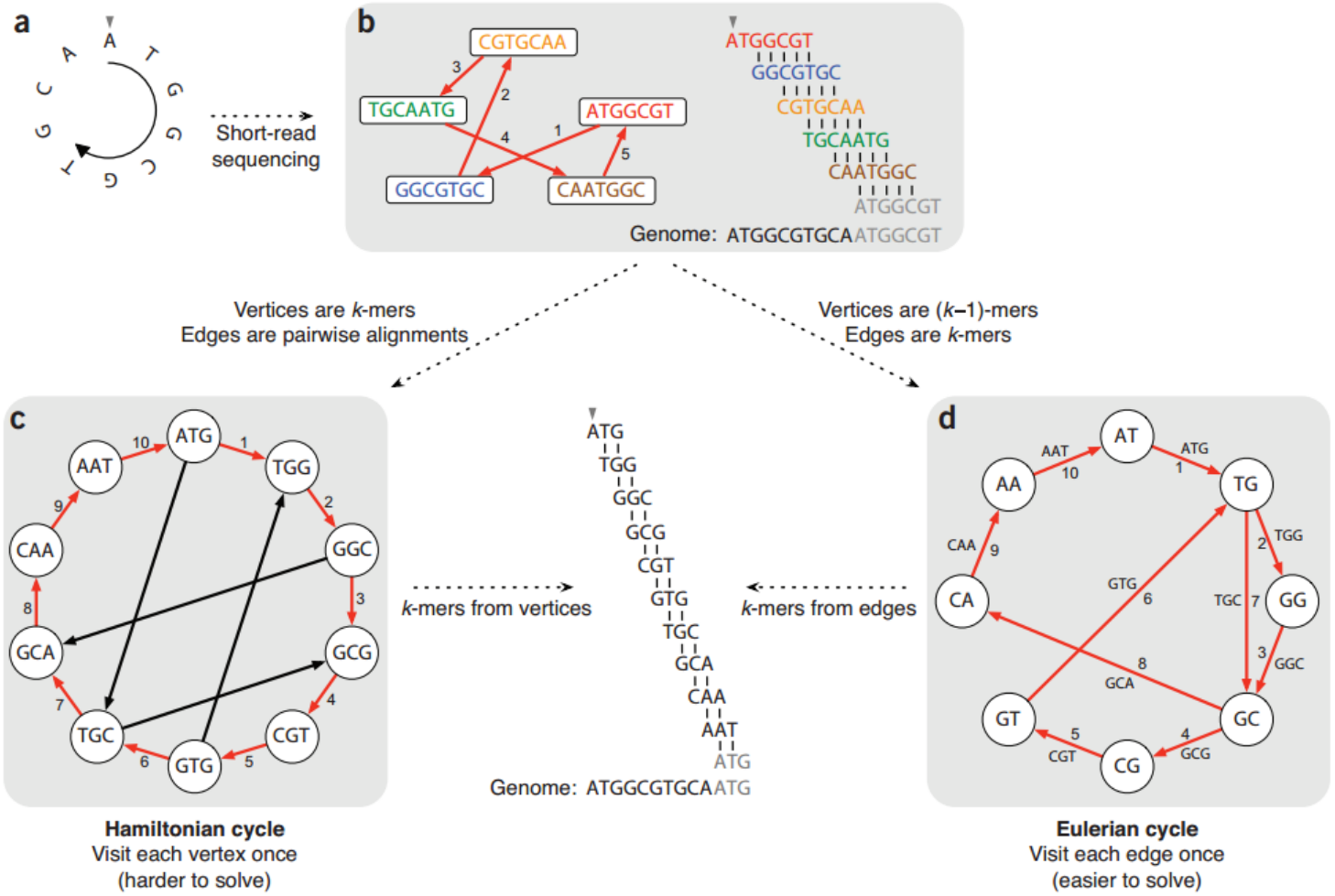
[1] Department of Biotechnology, Norwegian University of Science and Technology, Trondeim, Norway

[2] Synthetic Biology Engineering Research Center, University of California-Berkeley, Berkeley, USA



How to apply de Bruijn graphs to genome assembly

Phillip E C Compeau, Pavel A Pevzner & Glenn Tesler



PITTSBURGH SUPERCOMPUTING CENTER

XSEDE

Metagenomic Read Data

Time point	Number of Paired End Reads	Total Sequence
Initial (0-Weeks)	481,578,176	96GB
Intital (0-Weeks)	1.1 Millon (FLX Titanuim)	0.5GB
Shaker(6-Weeks)	487,079,477	97GB
Fermenter (8-Weeks)	542,583,705	108GB
TOTAL	1,512,341,358	~300GB

Computational Run-Time Statistics

Assembly Characteristic	Value
Velveth peak RAM Usage	3.6 TB (3,600GB)
Velveth CPU Hours	1,200 (50 days)
Velveth Wall Time Hours	60 (2.5 days)
Velvetg peak RAM Usage (w/RAMDISK)	1.0 TB (1,000GB)
Velvetg CPU Hours	1,600 (66 days)
Velveth Wall Time Hours	83 (3.5 days)

Velvet

EMBL-EBI

[135452.458854] No more memory for memory chunk!

Could not find 500kb of continuous RAM in 6000GB (6TB)!

the HOARD memory allocator

Emery Berger

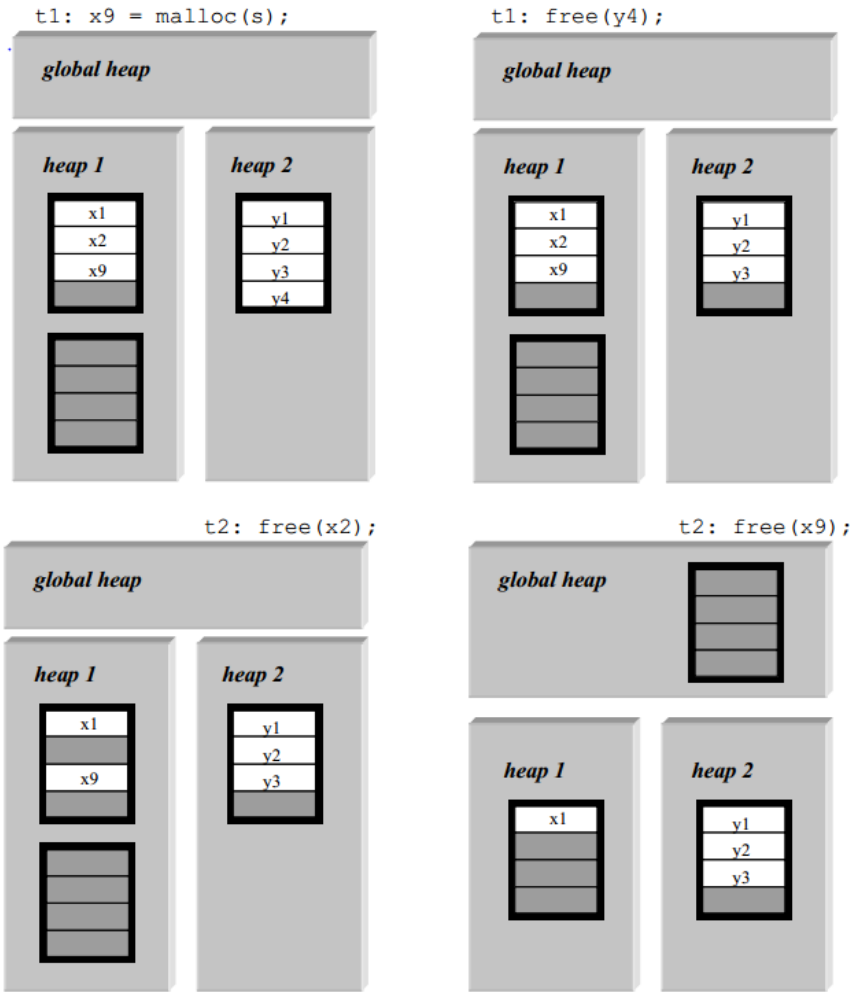
Associate Professor
Department of Computer Science
University of Massachusetts, Amherst

Hoard: A Scalable Memory Allocator for Multithreaded Applications

Emery D. Berger* Kathryn S. McKinley† Robert D. Blumofe* Paul R. Wilson*

*Department of Computer Sciences
The University of Texas at Austin
Austin, Texas 78712
{emery, rdb, wilson}@cs.utexas.edu

†Department of Computer Science
University of Massachusetts
Amherst, Massachusetts 01003
mckinley@cs.umass.edu



Velvet

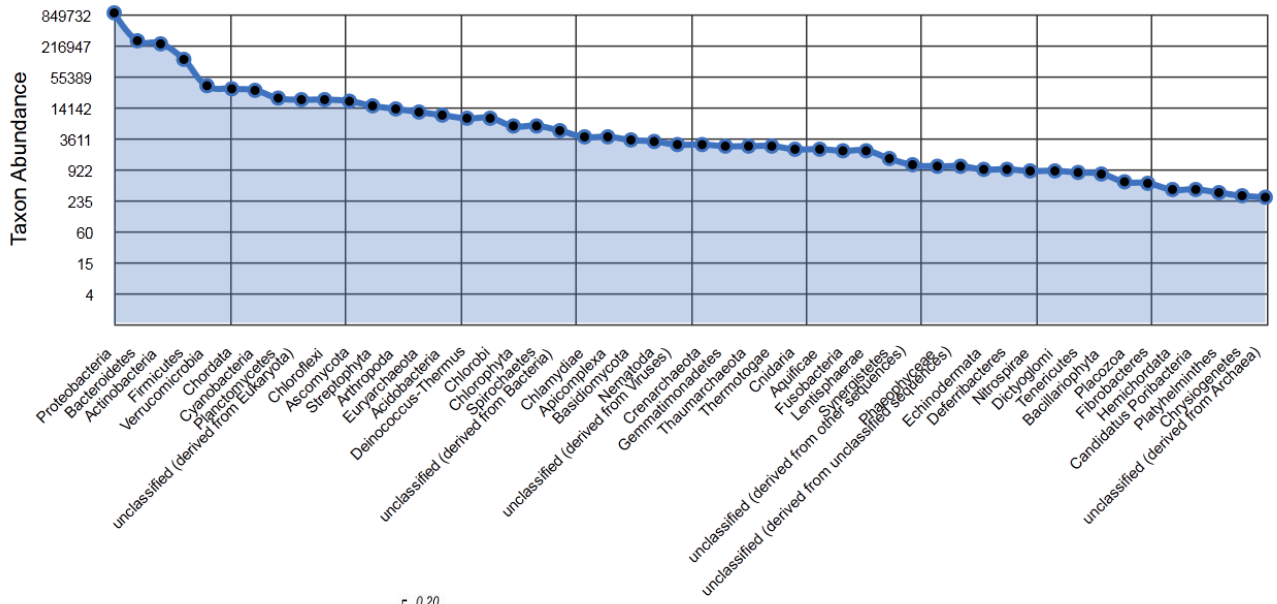
EMBL-EBI 

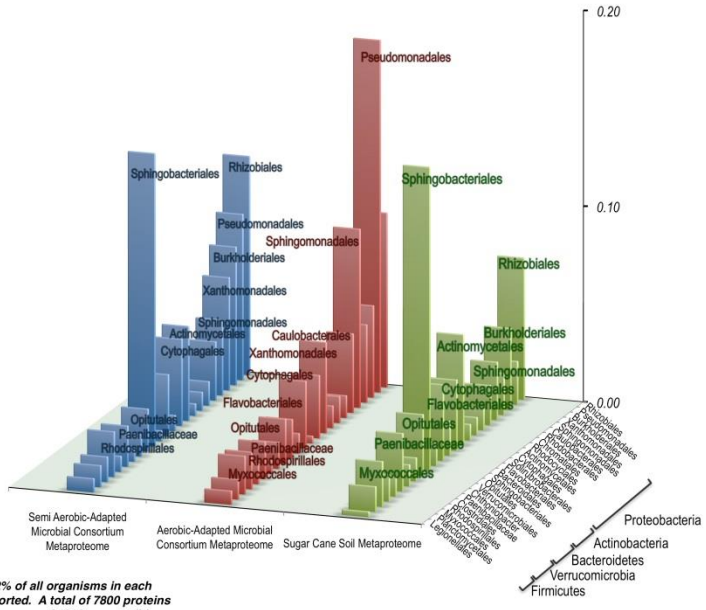
[232433.185224] Inputting sequence 1511000000 / 3022482716
[233050.932981] === Sequences loaded in 205756.407926 s
[233050.943863] Done inputting sequences
[233050.943874] Destroying splay table
[244658.550910] Splay table destroyed

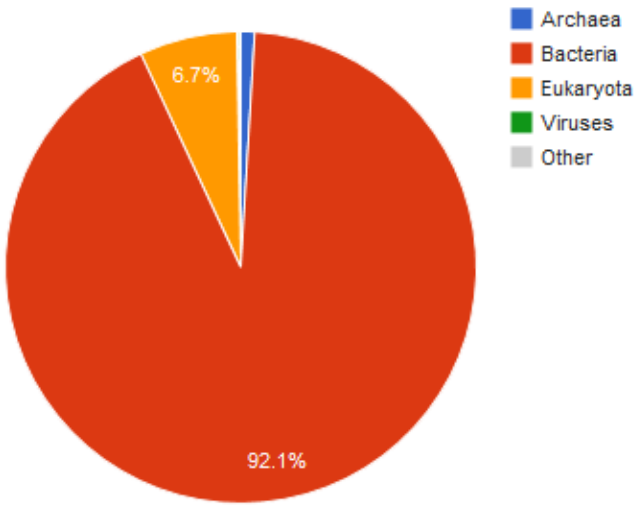
the HOARD memory allocator

Metagenomic Assembly	Velvet Assembly Size (300 BP Contigs) /n50
MSCA Initial Soil Sample	112 MB/1.3kb
MSCF Fermenter Enrichment	626 MB/2.8kb
MSCS Shaker Enrichment	577 MB /2.6kb
Total Combined Assembly	1.2 GB /2.6kb

Community Characteristics







Average of 82% of all organisms in each condition reported. A total of 7800 proteins considered, average of 2600 per condition

2.2 Million Peptides
32,143 Glycoside Hydrolase

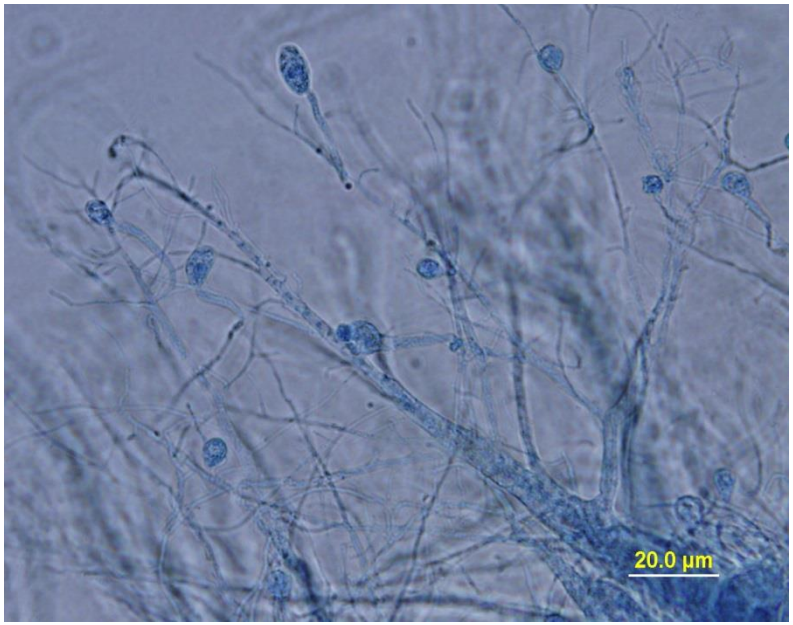
PFAM Domain	Number of Genes Models with Evalue < -4 PFAM	Physiological role
Polyketide Cyclase	1,597	Antibiotic Synthesis
NRPS	16	Peptide Secondary Metabolites
SnoaL Polyketide Cyclase	1,597	Antibiotic Synthesis
Skimate Kinase	350	Aromatic Compound Precursor Synthesis
Thioesterase	1,567	Terminal Step in Cyclic Antibiotic Synthesis
Antibiotic Biosytheis Monoxygenase	83	Antibiotic Synthesis

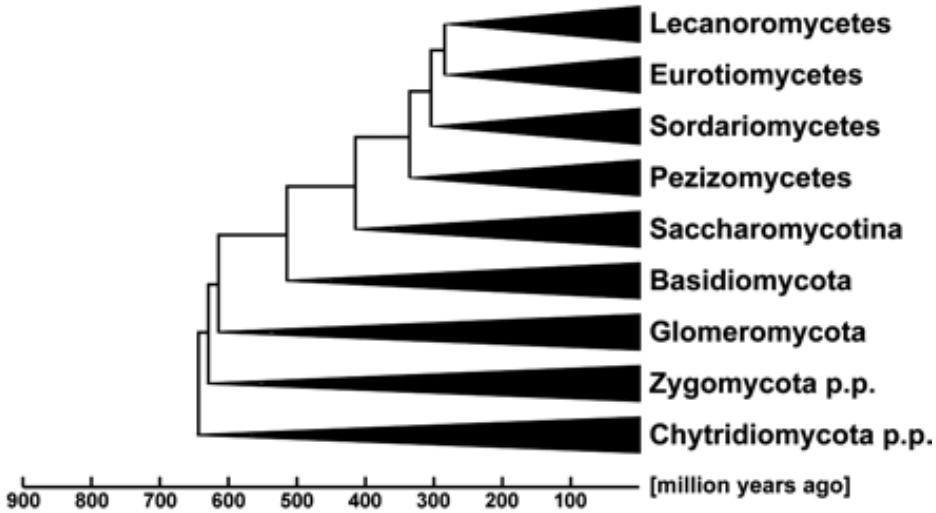
AEM
JournalsASM.org

The Genome of the Anaerobic Fungus *Orpinomyces* sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader

Noha H. Youssef,[a] M. B. Couger,[a] Christopher G. Struchtemeyer,[a] Audra S. Liggenstoffer,[a] Rolf A. Prade,[a] Fares Z. Najar,[b] Hasan K. Atiyeh,[c] Mark R. Wilkins,[c] Mostafa S. Elshahed[a]

Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, USA[a]; Department of Chemistry and Biochemistry, University of Oklahoma, Stillwater, Oklahoma, USA[b]; Department of Biosystems and Agricultural Engineering, Oklahoma State University, Stillwater, Oklahoma, USA[c]





nature biotechnology

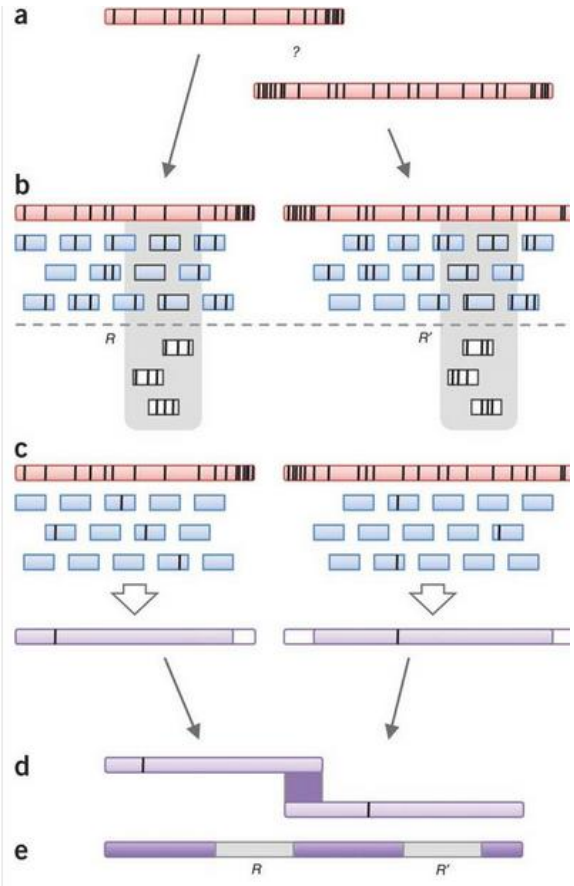
Hybrid error correction and *de novo* assembly of single-molecule sequencing reads

Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis & Adam M Phillippy

Affiliations | Contributions | Corresponding authors

Nature Biotechnology **30**, 693–700 (2012) | doi:10.1038/nbt.2280

Received 26 October 2011 | Accepted 18 May 2012 | Published online 01 July 2012



AEM
Journals.ASM.org

The Genome of the Anaerobic Fungus *Orpinomyces* sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader

Noha H. Youssef,[a] M. B. Couger,[a] Christopher G. Struchtemeyer,[a] Audra S. Liggenstoffer,[a] Rolf A. Prade,[a] Fares Z. Najar,[b] Hasan K. Atiyeh,[c] Mark R. Wilkins,[c] Mostafa S. Elshahed[a]

Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, USA[a]; Department of Chemistry and Biochemistry, University of Oklahoma, Stillwater, Oklahoma, USA[b]; Department of Biosystems and Agricultural Engineering, Oklahoma State University, Stillwater, Oklahoma, USA[c]

1 Table S2. Comparison between genome assembly using Illumina only sequence data and
2 SMRT-Illumina hybrid assembly

	Illumina Only	PacBio/Illumina assembly
Genome assembly size	105.1 MB	100.85 MB
In 1kb+ Contigs	73.52 MB	100.95 MB
Number of ambiguous bp	91,688 bp	0
N50 of 1kb+ Contigs	2,226 bp	3,373 bp
N90 of 1kb contigs	1,072 bp	1,829 bp
Average length of gene model	903 bp	1623 bp
Number of introns	2,458	35,697
Number of gene models	14,594	16,437
GC% Content	15.8	17.0
Total PASA/Trinity Assemblies	10,115	14,009

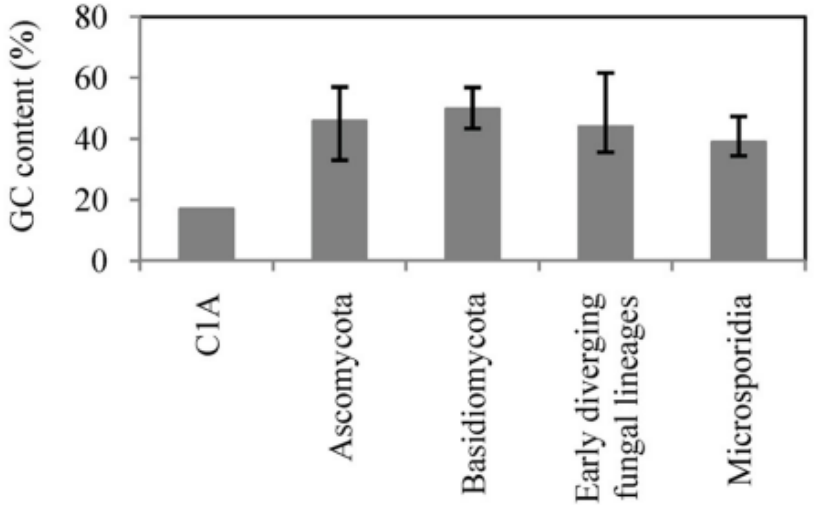
AEM
Journals.ASM.org

The Genome of the Anaerobic Fungus *Orpinomyces* sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader

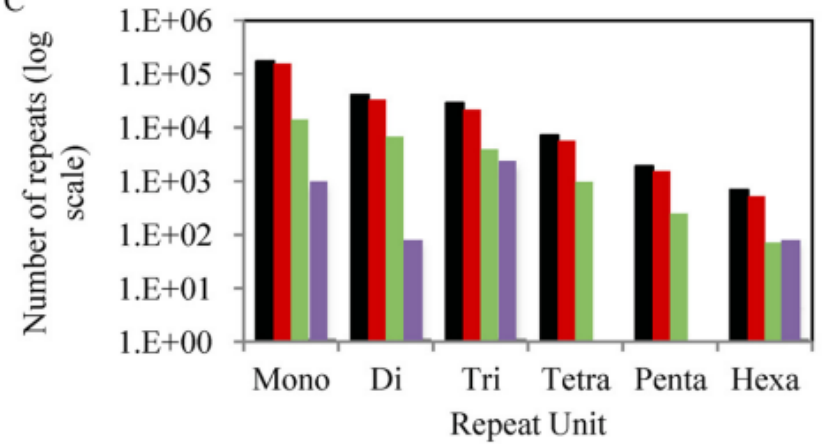
Noha H. Youssef,[a] M. B. Couger,[a] Christopher G. Struchtemeyer,[a] Audra S. Liggenstoffer,[a] Rolf A. Prade,[a] Fares Z. Najar,[b] Hasan K. Atiyeh,[c] Mark R. Wilkins,[c] Mostafa S. Elshahed[a]

Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, USA[a]; Department of Chemistry and Biochemistry, University of Oklahoma, Stillwater, Oklahoma, USA[b]; Department of Biosystems and Agricultural Engineering, Oklahoma State University, Stillwater, Oklahoma, USA[c]

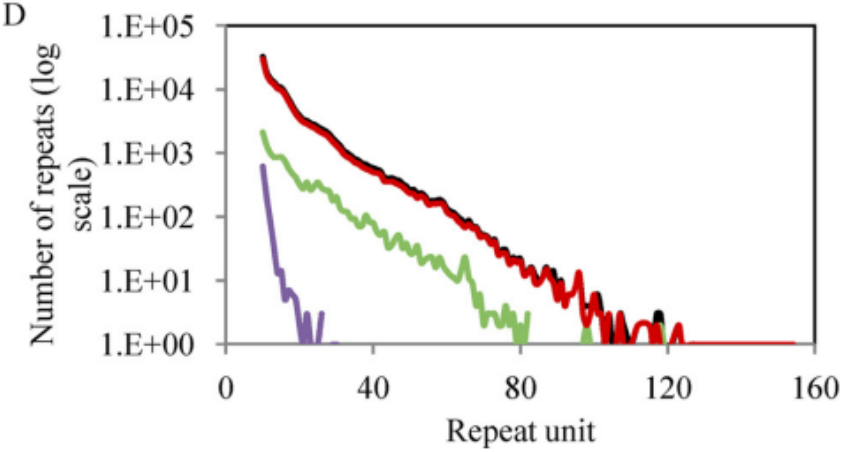
A



C

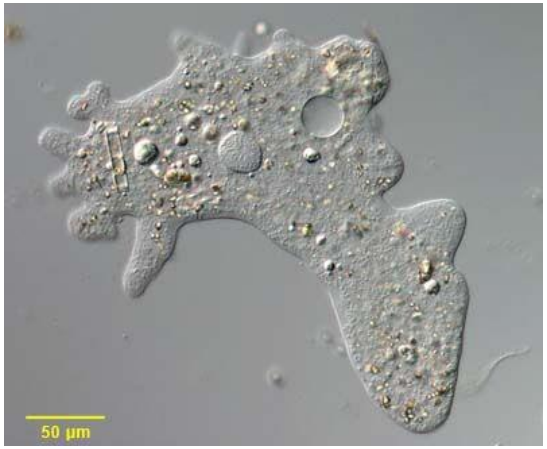


D



Genomic Assembly of Novel Organisms

Protopterus aethiopicus

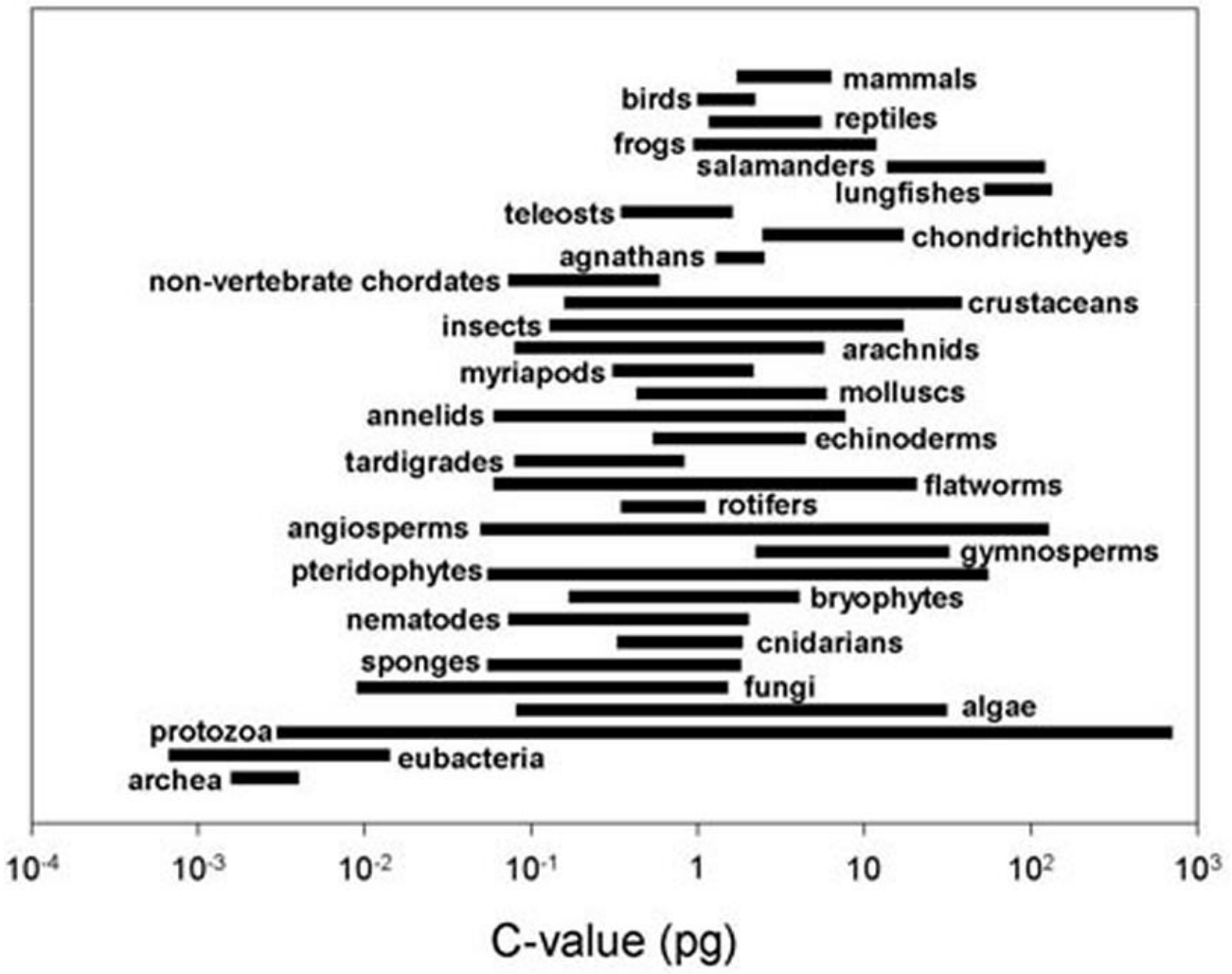


Polychaos dubium

Paris japonica

140+ Giga basepairs

ANIMAL GENOME SIZE DATABASE



Kew
ROYAL BOTANIC GARDENS

	C Mean	C Min.	C Max.	Standard deviation
1C (Mbp)	5883.89	10	148852	9543.79

3040 Plant Genome Listed with 1C value greater than 3GB

Trinotate: Transcriptome Functional Annotation and Analysis

Trinotate

NCBI BLAST

HMMER

Pfam

UniProt

eggNOG version 3.0

SQLite

the Gene Ontology

RNA-Seq ➡ Trinity ➡ Transcripts/Proteins ➡ Functional Data ➡ Discovery

Automated Higher Order Biological Analysis

ParaFly: Simple parallel unix command processing using OpenMP

PITTSBURGH SUPERCOMPUTING CENTER

XSEDE

DNA Sequencing Caught in Deluge of Data

The New York Times

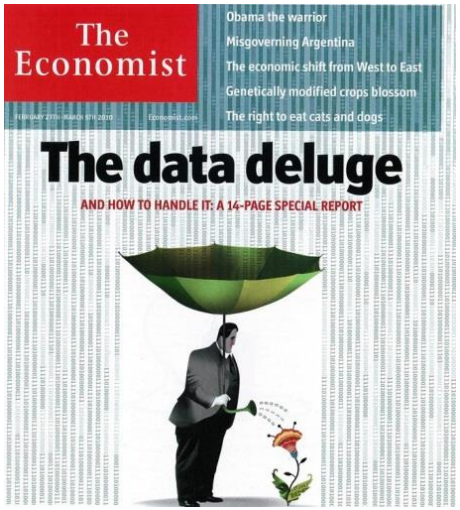
By ANDREW POLLACK
Published: November 30, 2011

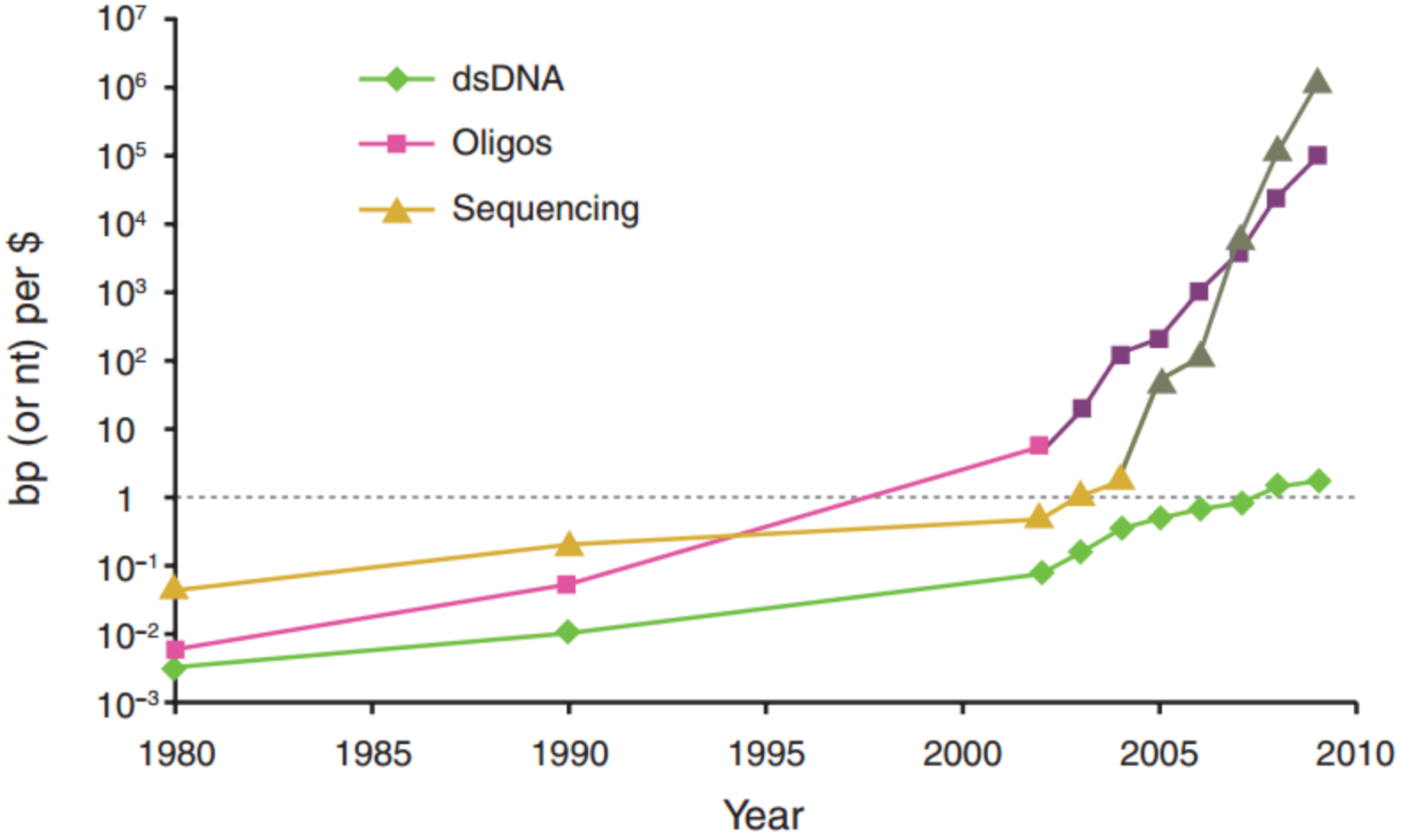
Contracting Sequencing Costs Could Mean Ballooning Informatics Prices

Aabha Khemani, Gauri Jaju

GEN Genetic Engineering & Biotechnology News

Biotechnology from bench to business





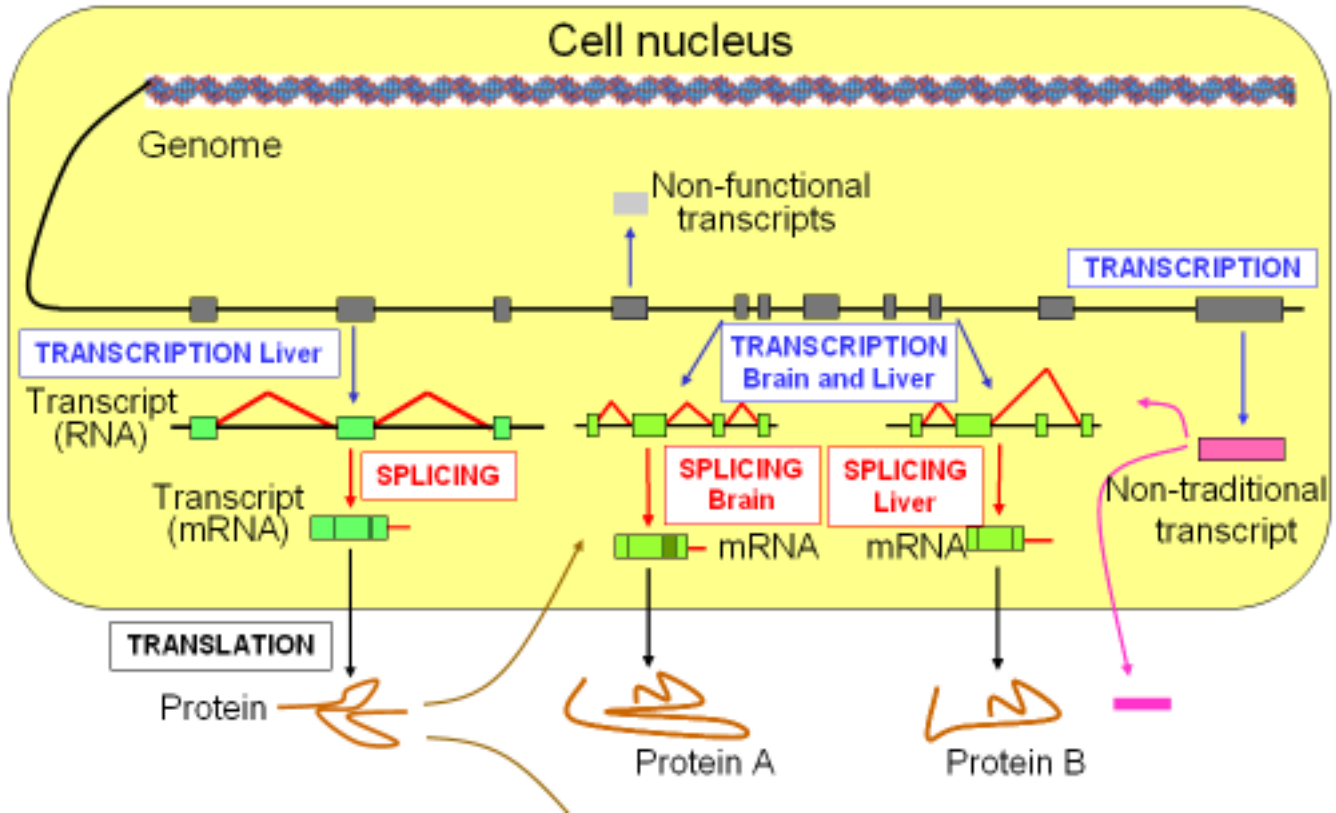
Genome engineering

Peter A Carr[1] & George M Church[2]

Nature Biotechnology **27**, 1151 - 1162 (2009)
Published online: 9 December 2009 | doi:10.1038/nbt.1590

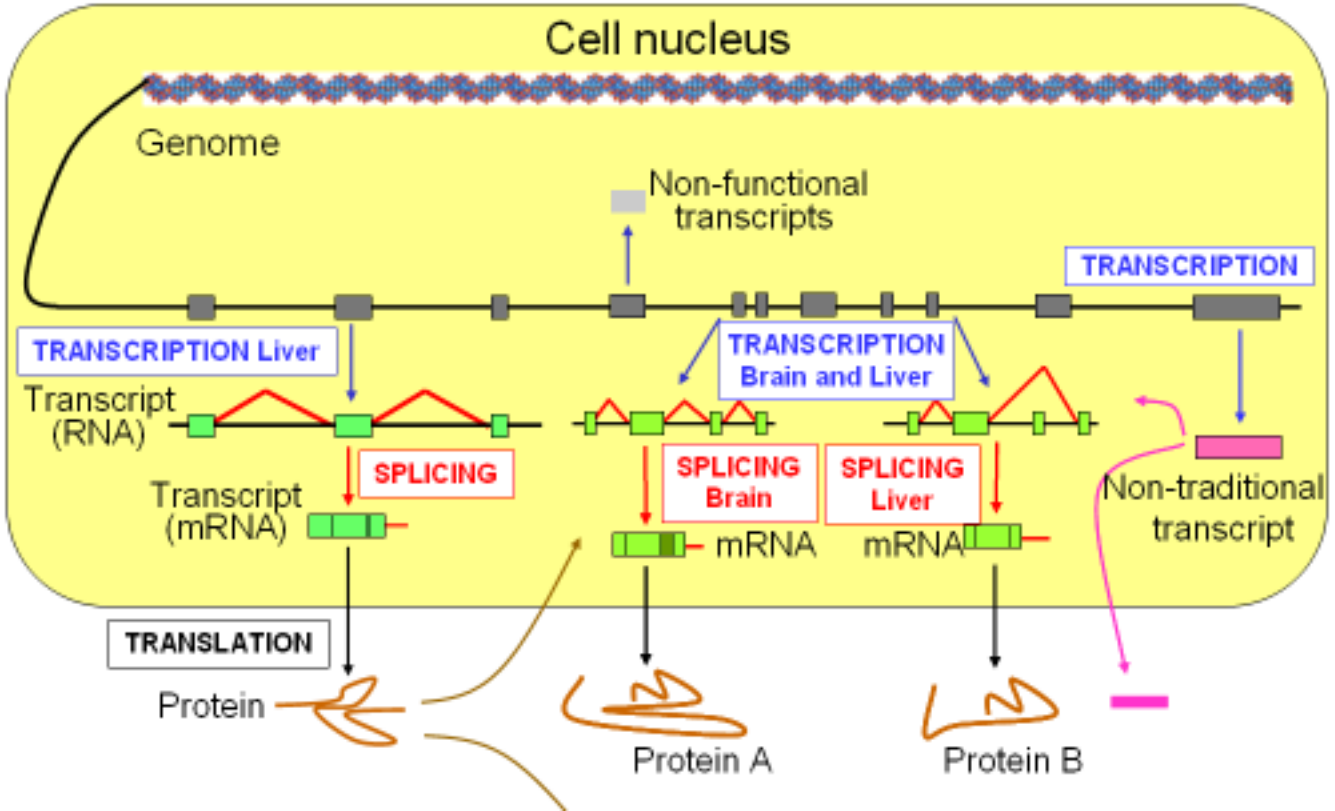
WHAT IS A TRANSCRIPTOME?

- DNA is transcribed into corresponding molecules of RNA
- The transcriptome is all of the transcripts of a particular cell
- The major type of RNA is messenger RNA (mRNA) which a gene may produce various types of

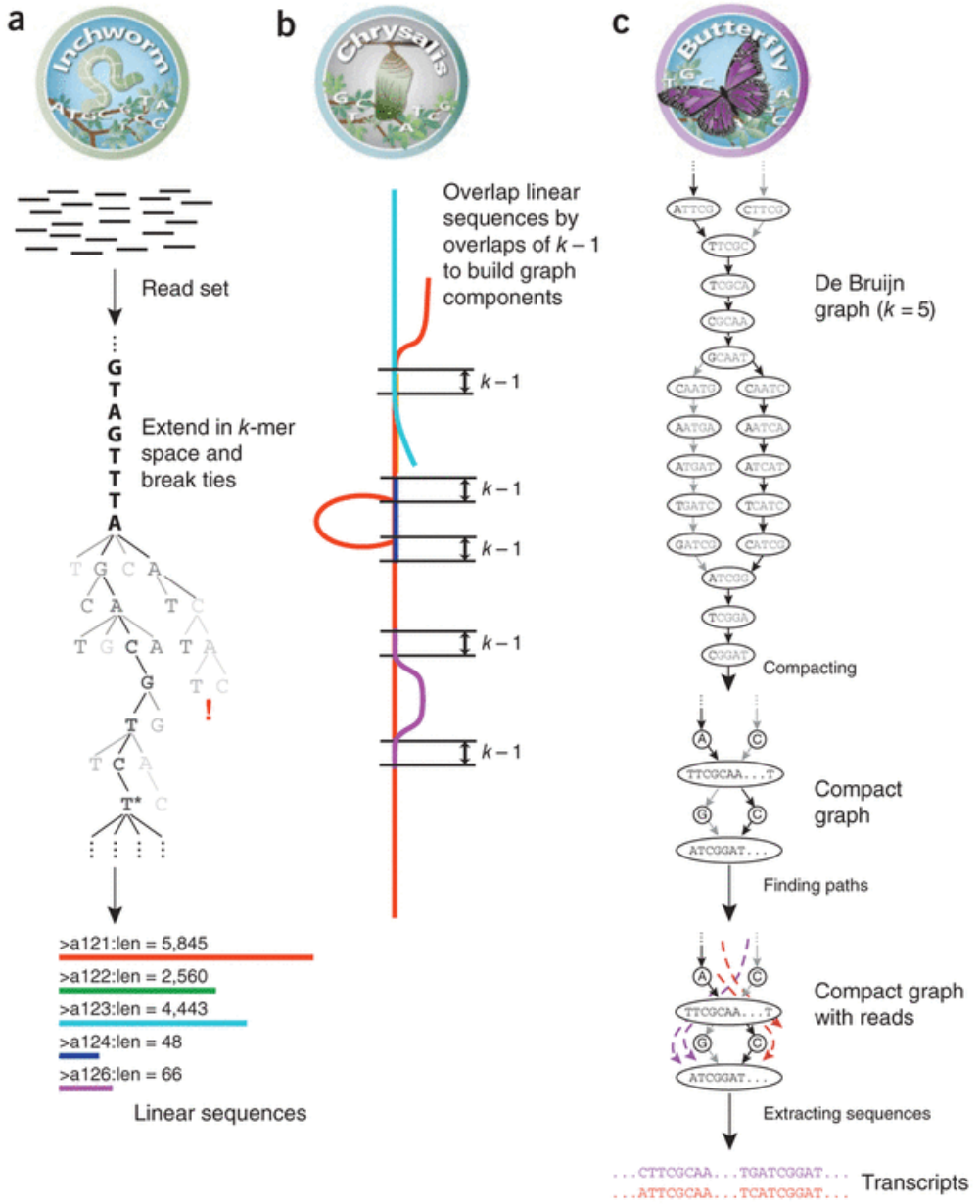


WHAT CAN THE TRANSCRIPTOME TELL US?

- **The transcriptome shows when and where each gene is turned off or on in the cells and tissues of an organism**
- **Counting the number of transcripts for a given gene can determine gene expression**

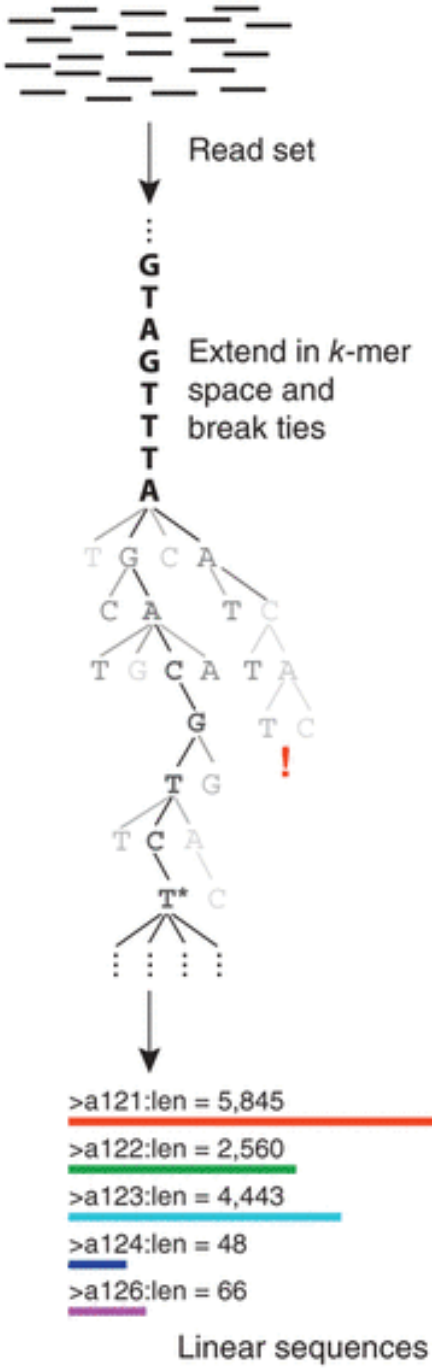


BUILDING THE TRANSCRIPTOME FROM RNA-SEQ READS WITH DE NOVO ASSEMBLY USING TRINITY



Grabherr et al., 2011

BUILDING THE TRANSCRIPTOME FROM RNA-SEQ READS WITH DE NOVO ASSEMBLY USING TRINITY

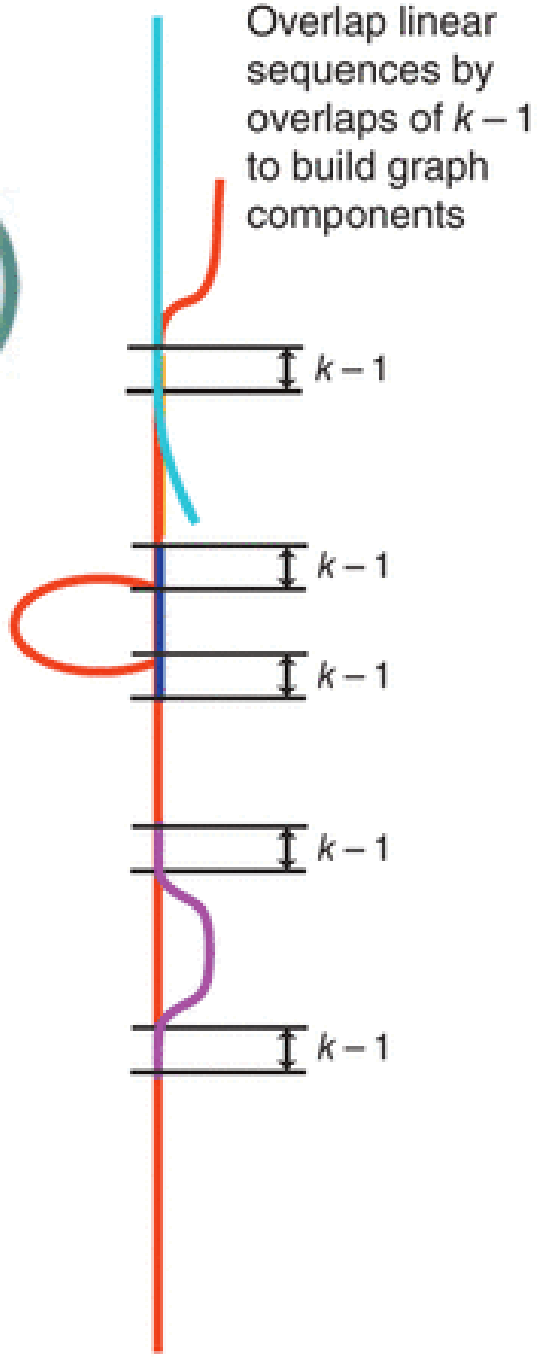


INCHWORM: All overlapping k-mers are extracted from the RNA-seq reads, and each unique k-mer is examined and transcript contigs are generated using a greedy extension based on (k-1)-mer overlaps.

Grabherr et al., 2011

BUILDING THE TRANSCRIPTOME FROM RNA-SEQ READS WITH DE NOVO ASSEMBLY USING TRINITY

CHRYSALIS: related Inchworm contigs are clustered into components using raw reads to group transcripts. A de Bruijn graph for each cluster is built, and then reads are partitioned among the clusters.

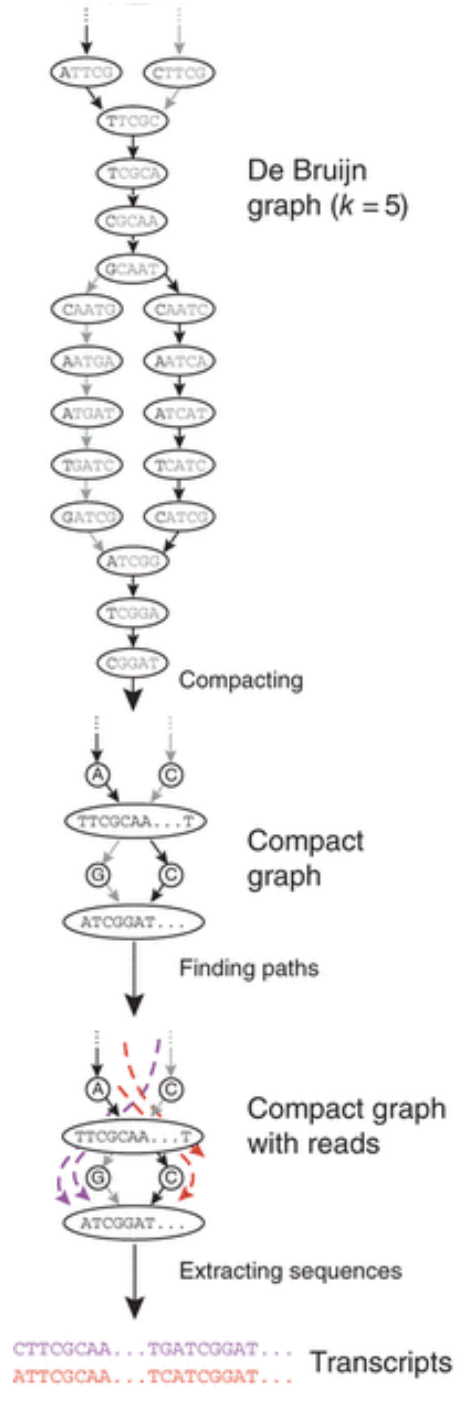


Grabherr et al., 2011

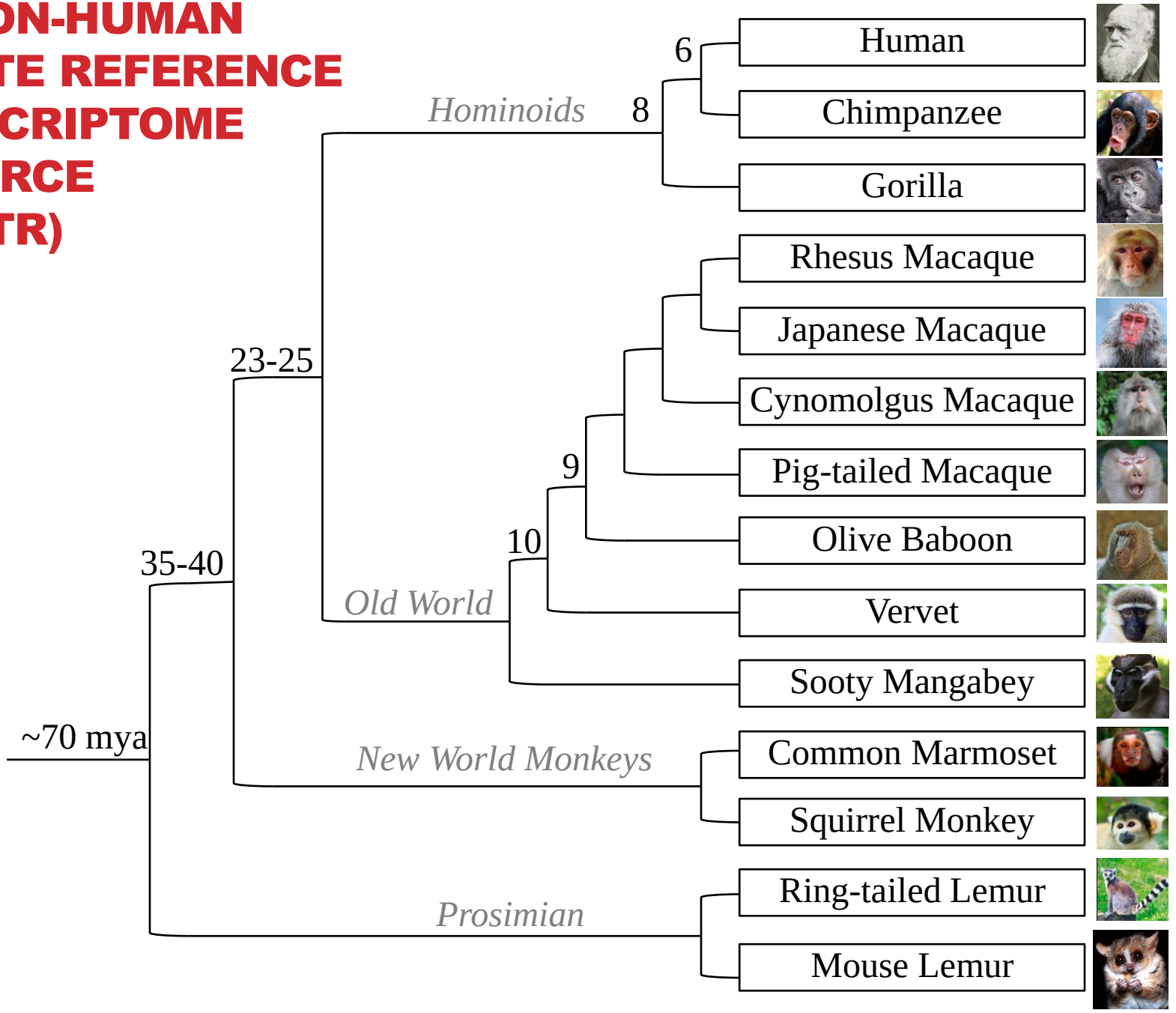
BUILDING THE TRANSCRIPTOME FROM RNA-SEQ READS WITH DE NOVO ASSEMBLY USING TRINITY

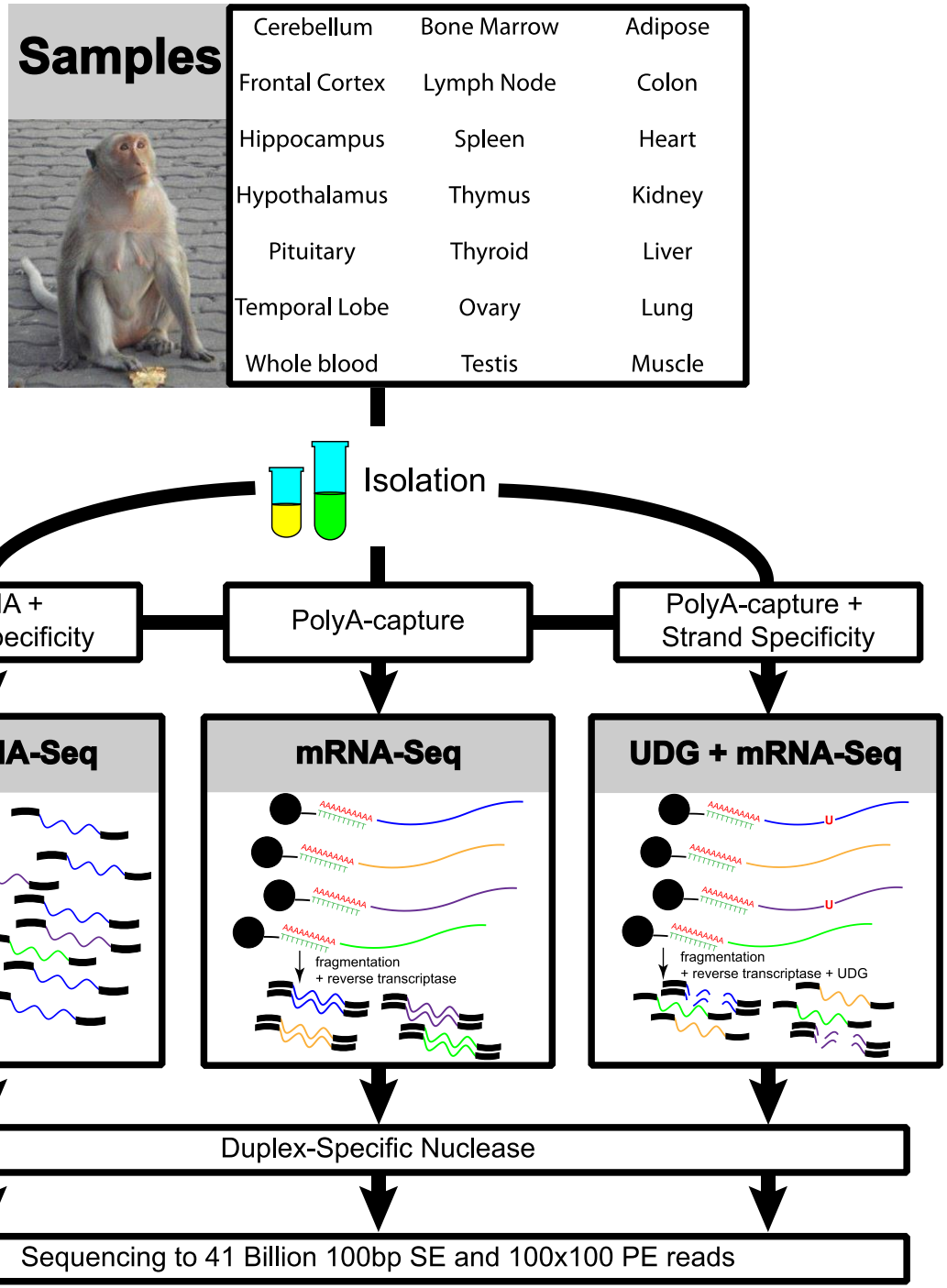
BUTTERFLY: The RNA-seq reads are traced through the graph and connectivity based on the read sequence and on support from any available paired-end data is determined. Butterfly processes each de Bruijn graph in parallel and reports the full-length transcripts for alternatively spliced isoforms.

Grabherr et al., 2011

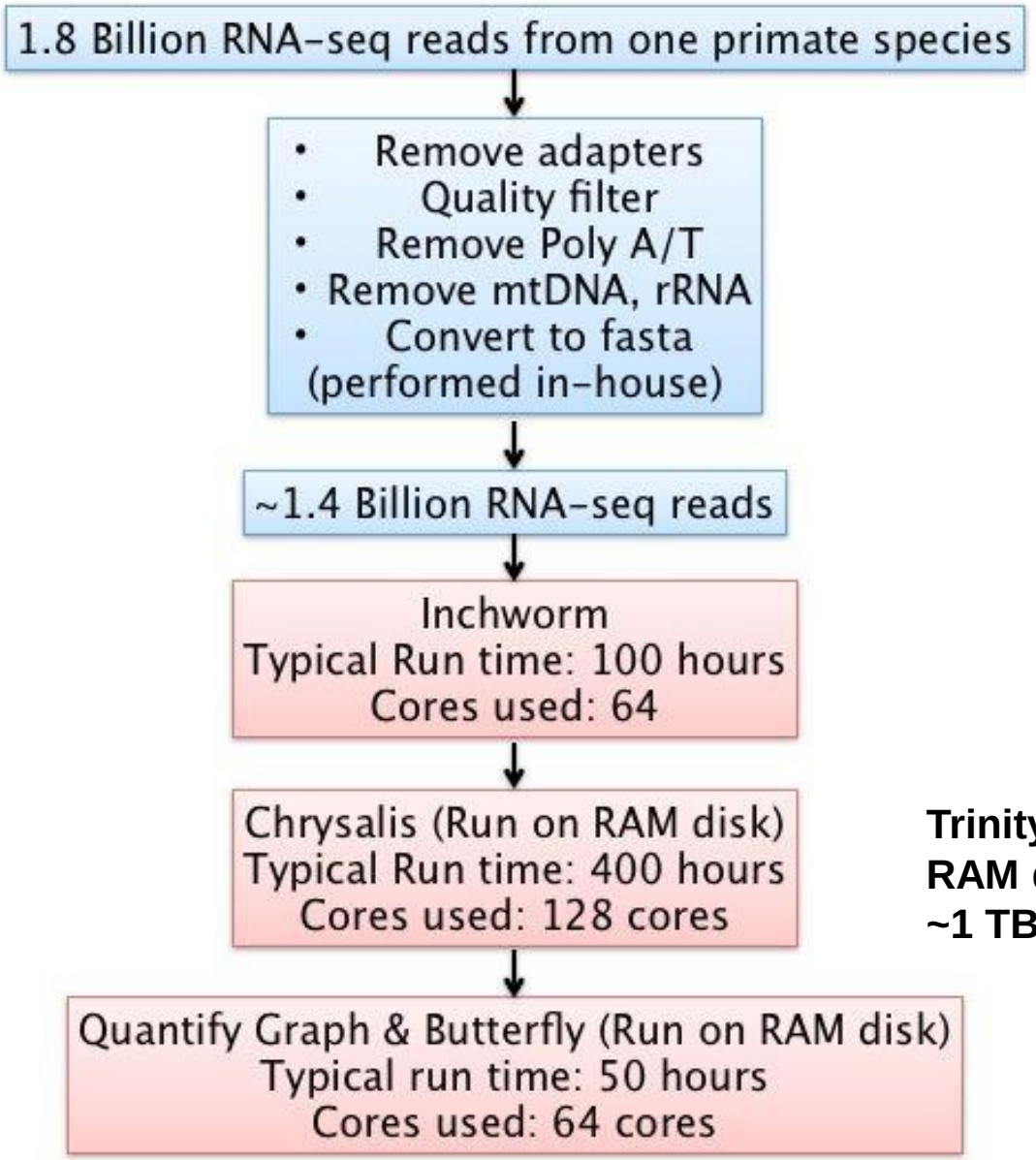


THE NON-HUMAN PRIMATE REFERENCE TRANSCRIPTOME RESOURCE (NHPRTR)




Samples
Cerebellum
Bone Marrow
Adipose
Frontal Cortex
Lymph Node
Colon
Hippocampus
Spleen
Heart
Hypothalamus
Thymus
Kidney
Pituitary
Thyroid
Liver
Temporal Lobe
Ovary
Lung
Whole blood
Testis
Muscle
Isolation
All RNA +
Strand Specificity
PolyA-capture
PolyA-capture +
Strand Specificity
Total RNA-Seq
mRNA-Seq
UDG + mRNA-Seq
fragmentation
+ reverse transcriptase
fragmentation
+ reverse transcriptase + UDG
Duplex-Specific Nuclease
Sequencing to 41 Billion 100bp SE and 100x100 PE reads

TRINITY PIPELINE



Trinity + files on RAM disk required ~1 TB RAM

5x faster on RAM disk

Species	Library	# input sequences (billion)	# contigs	Total length (Mb)	N25 (bp)	N50 (bp)	N75 (bp)	Longest contig (bp)
Baboon	TOT	0.149	658,581	281	1,029	434	276	35,170
	UDG	1.543	1,131,951	844	3,534	1,368	479	131,395
Chimpanzee	UDG	1.465	987,615	1,433	6,356	3,806	1,666	47,873
Cynomologus Macaque Chinese	RNA	1.676	911,282	864	4,769	2,316	706	59,032
	UDG	1.630	990,604	1,055	5,479	2,822	875	122,916
Cynomolgus Macaque Mauritian	RNA	1.078	1,142,531	929	4,368	1,813	525	30,364
	TOT	0.166	526,723	200	657	377	266	36,976
	UDG	1.177	1,015,657	834	4,360	1,927	532	22,239
Gorilla	UDG	1.256	732,336	1,122	5,878	3,680	1,804	33,526
Japanese Macaque	RNA	1.863	703,246	738	5,010	2,687	898	21,620
Marmoset	TOT	0.253	332,782	118	545	357	261	14,561
	UDG	1.659	814,235	476	2,206	785	359	122,518
Pig-tailed Macaque	RNA	1.725	969,993	1,044	5,189	2,807	916	25,056
	UDG	1.573	1,301,087	1,222	5,015	2,265	668	32,637
Rhesus Macaque Chinese	RNA	1.210	923,017	836	4,694	2,230	639	32,796
	UDG	1.310	969,421	850	4,638	2,114	594	34,020
Rhesus Macaque Indian	RNA	3.200	703,246	738	5,010	2,687	898	21,620
	UDG	1.410	1,051,149	833	3,966	1,644	519	68,269
Ring-tailed Lemur	UDG	1.403	611,678	822	6,169	3,630	1,468	33,401
Sooty Mangabey	UDG	1.635	1,188,472	1,466	6,431	3,483	1,116	30,331

CURRENT USES OF THE DE NOVO ASSEMBLED TRANSCRIPTOMES

- Improve existing genome annotation for species with reference genomes
  - Different splice variants
  - Show extensions in 3’ and 5’ UTRs
- Establish genome annotation for species without reference genomes
- Assist in genome annotation for the Baboon genome consortium

Long-term goal is to have an excellent annotation data set for each species to detect species-specific differences that have functional consequences.

BLACKLIGHT AND XSEDE HIGHLIGHTED IN RECENT TRINITY PROTOCOL PAPER

NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

Affiliations | **Contributions** | **Corresponding authors**

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013

ACKNOWLEDGMENTS

Mason Lab (Weill Cornell Medical College)

Christopher E. Mason

Sheng Li

Marjan Bozinoski

XSEDE

Philip Blood

Tri-Institutional PhD Program for Computational Biology and Medicine

Katze Lab (University of Washington)

Michael Katze

Xinxia Peng

Robert Palermo

Illumina

Gary Schroth

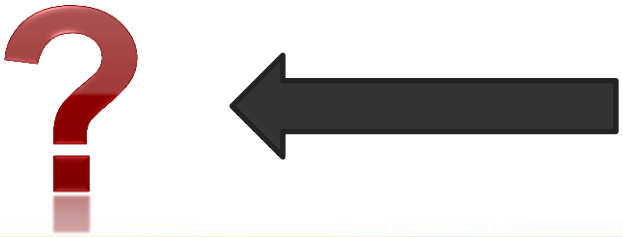
Baylor College of Medicine

Jeff Rogers

Questions?

How can I:

- De novo assemble a 17 Gbp genome?
- Assemble a terabase of metagenomic data?
- De novo assemble 30 transcriptomes for 10 different primate species?



YOUR QUESTION HERE