JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Computing for Genomics

Dan Rokhsar,
DOE JGI/LBNL & UC Berkeley

NERSC User Meeting
4 February 2014

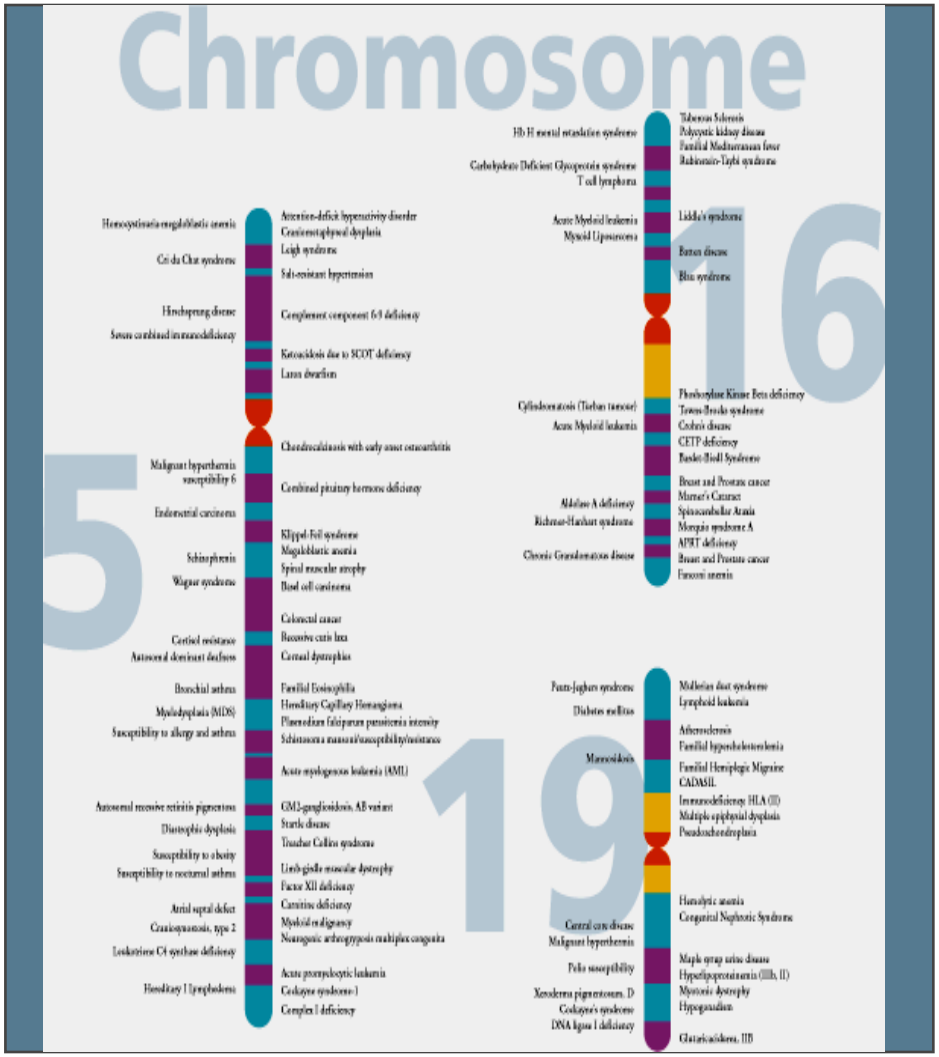
Computing for Genomics

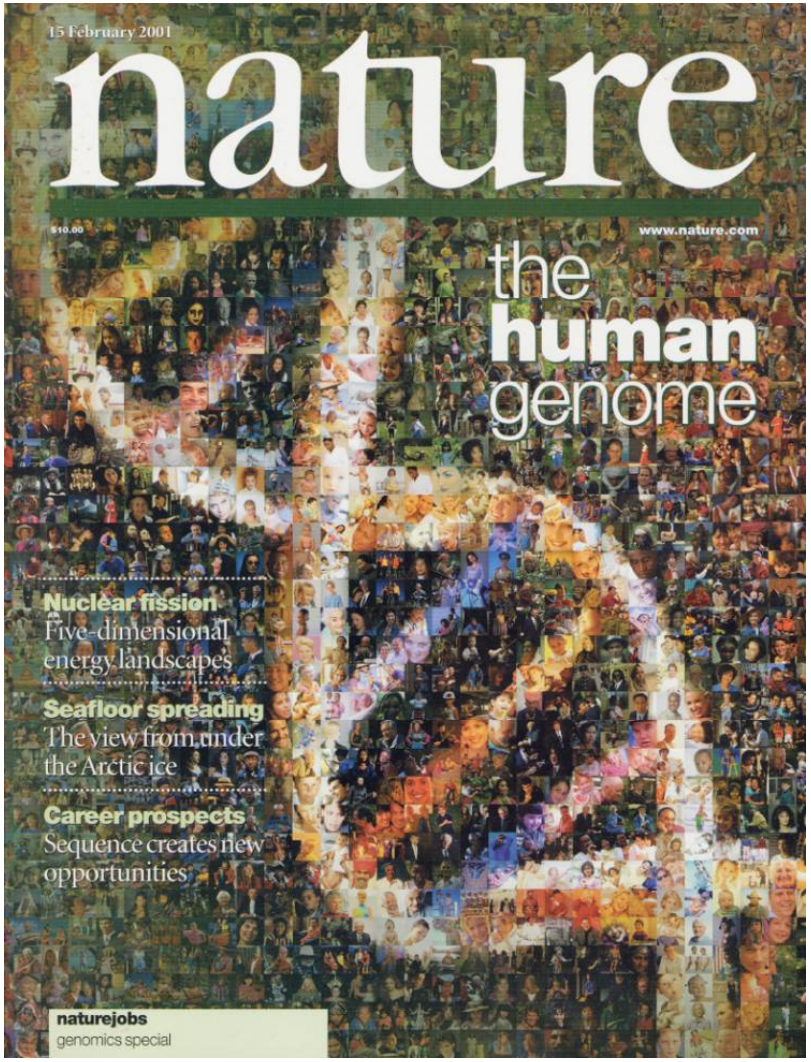
- **The DOE Joint Genome Institute**
  - Who we are, what we do, and why
- **Two computational problems in genomics**
  - "Assembling" genomes from "shotgun" sequence
  - "Annotating" "metagenomic" data

DOE JGI- *A Genomics User Facility*

- In 1997, as part of the ramp-up for the human genome project, the DOE Office of Biological and Environmental Research created a virtual "Joint" Genome Institute
- DOE relevance: baseline for study of DNA damage from radiation.
- Goal: sequence 3 of the 23 human chromosomes (~11% of genome).
- "Joint" = LBNL, LLNL, LANL; in 1999 efforts were centralized at one location in Walnut Creek, CA

By 2003, the “mission” was completed!





DOE JGI- *A Genomics User Facility*

- **In the meantime: sequencing had become more efficient and cost-effective.**

- **And genomic information was becoming more and more central to studying all biological systems.**

- **As a continuation of the human genome project over a dozen animal genomes were sequenced by JGI for comparative purposes.**

DOE JGI- *A Genomics User Facility*

- **In 2004, JGI became a DOE User Facility for Large Scale Genomics to Enable Bioenergy and Environmental Research**
- **Not "just" sequencing:**
  - New technologies & strategies
  - Novel computational methods and pipelines
  - Access to JGI scientific staff
  - DNA synthesis
  - Building user communities
- **Over 1,200 users …**
- **In 2010, established partnership with NERSC to manage JGI computing infrastructure and collaborate on common scientific interests**

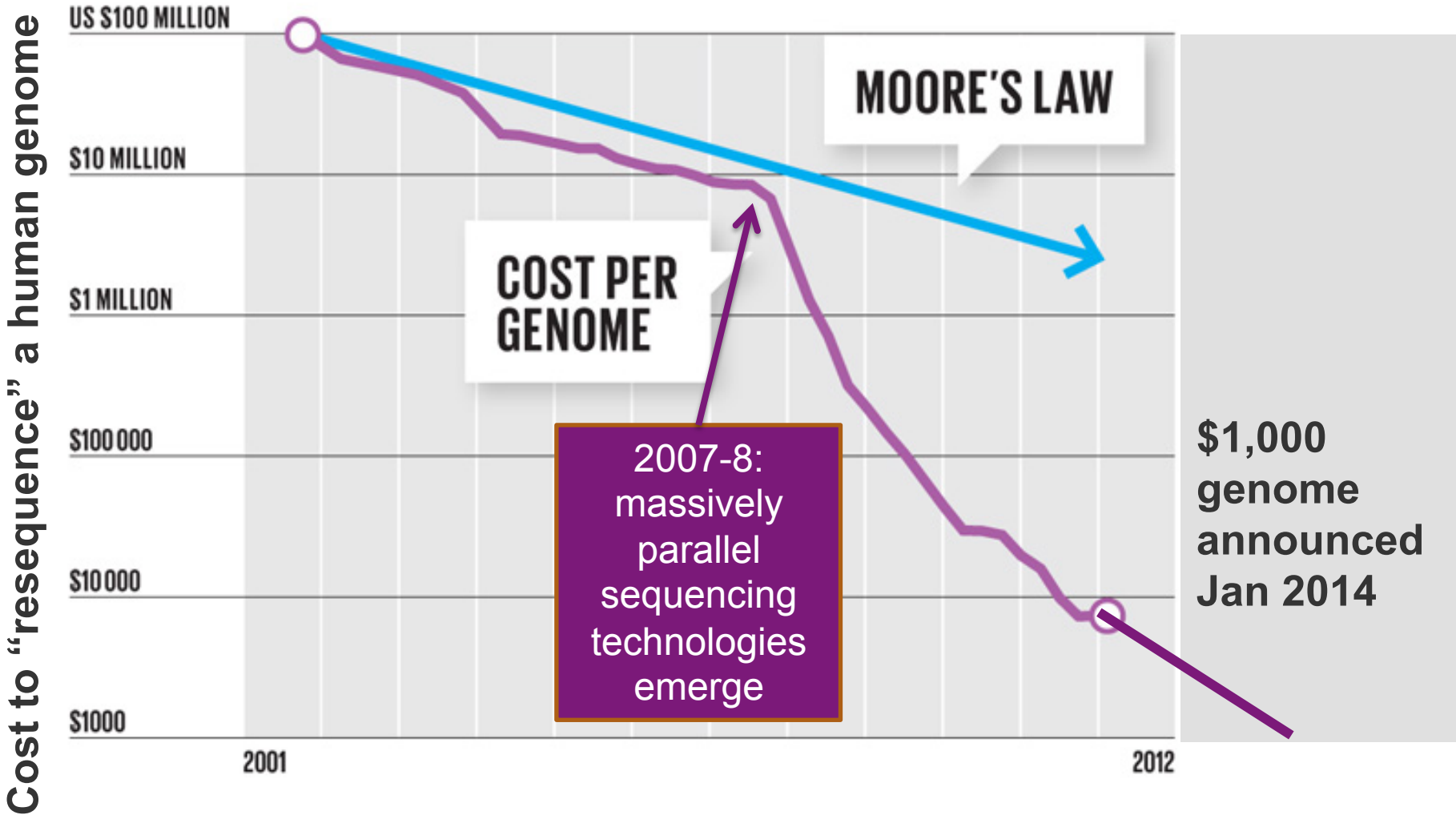

SECOND
“META-
GENOME”


US DEPARTMENT
OF ENERGY
JOINT GENOME
INSTITUTE (JGI)
THIRD ANNUAL
USER MEETING
WALNUT CREEK, CA
MARCH 26-28, 2008
FOCUS:
Genomics of renewable
energy strategies, biomass
conversion to biofuels, envi-
ronmental gene discovery,
engineering of fuel-producing
organisms. Featuring infor-
matics workshops & tutorials.
JGI
GENOMICS
OF ENERGY &
ENVIRONMENT
KEYNOTE:
Steven Chu
Nobel Laureate, Director,
Lawrence Berkeley National
Laboratory
TO REGISTER:
WWW.JGI.DOE.GOV


Science
FIRST
TREE
GENOME
AAAS

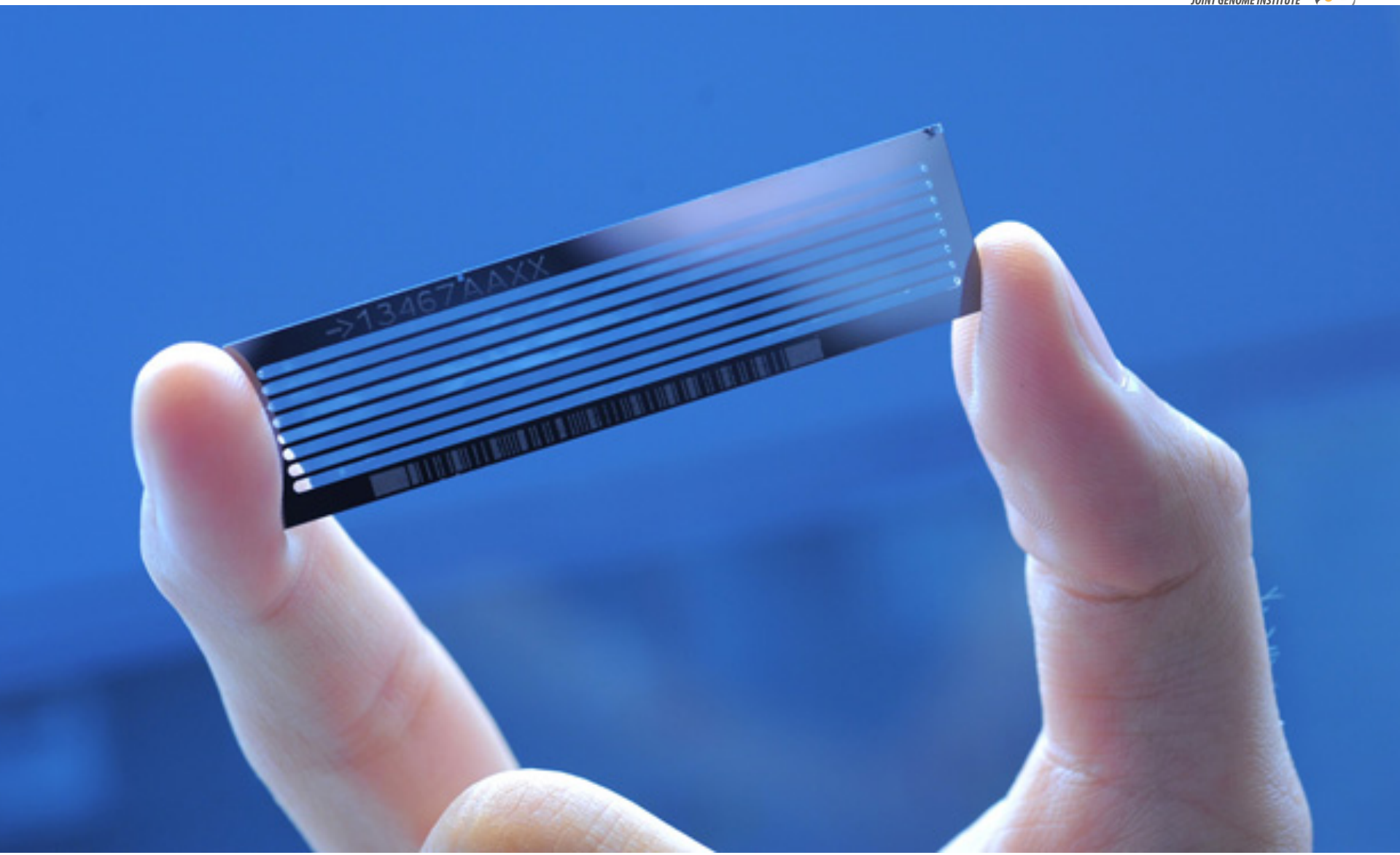

nature
biotechnology
VOLUME 22 NUMBER 6 JUNE 2004
FIRST
FILAMENTOUS
FUNGUS
GENOME
Genome of a white-rot fungus
Boosting essential oils in plants
Imaging cancer drug effects on HER2

Sequencing capacity is outstripping Moore's Law

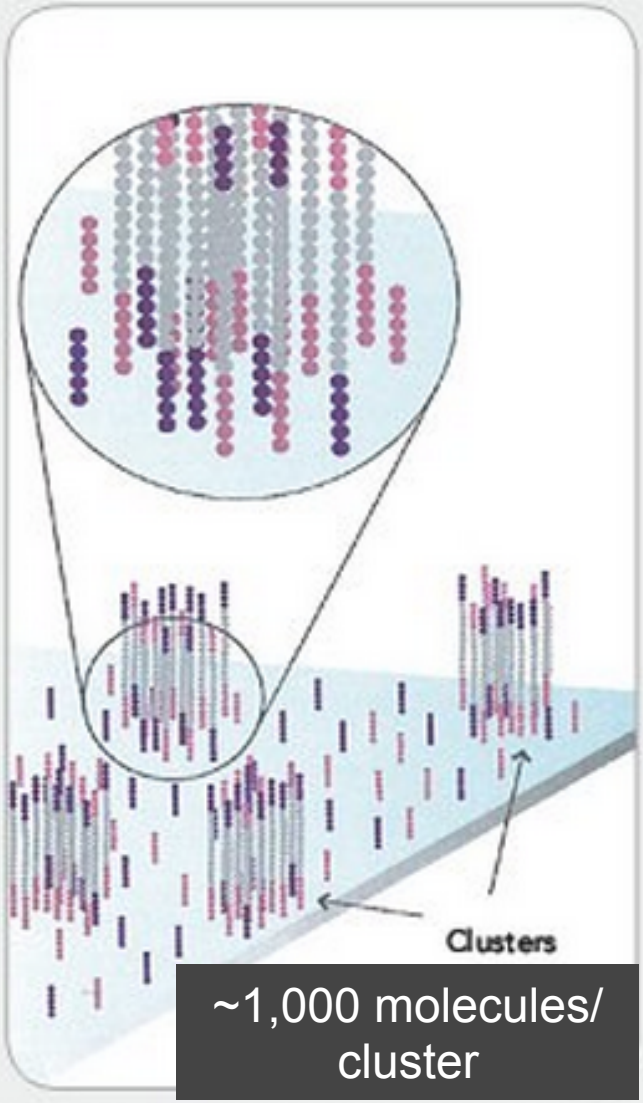


Source: National Human Genome Research Institute

Massively parallel sequencing

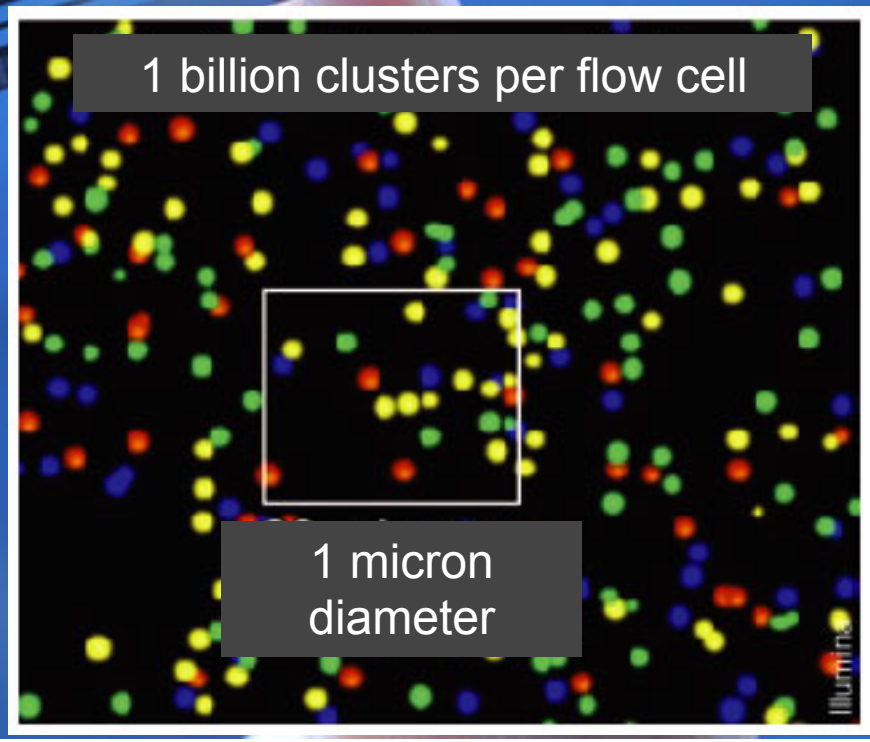


Massively parallel sequencing



~1,000 molecules/ cluster

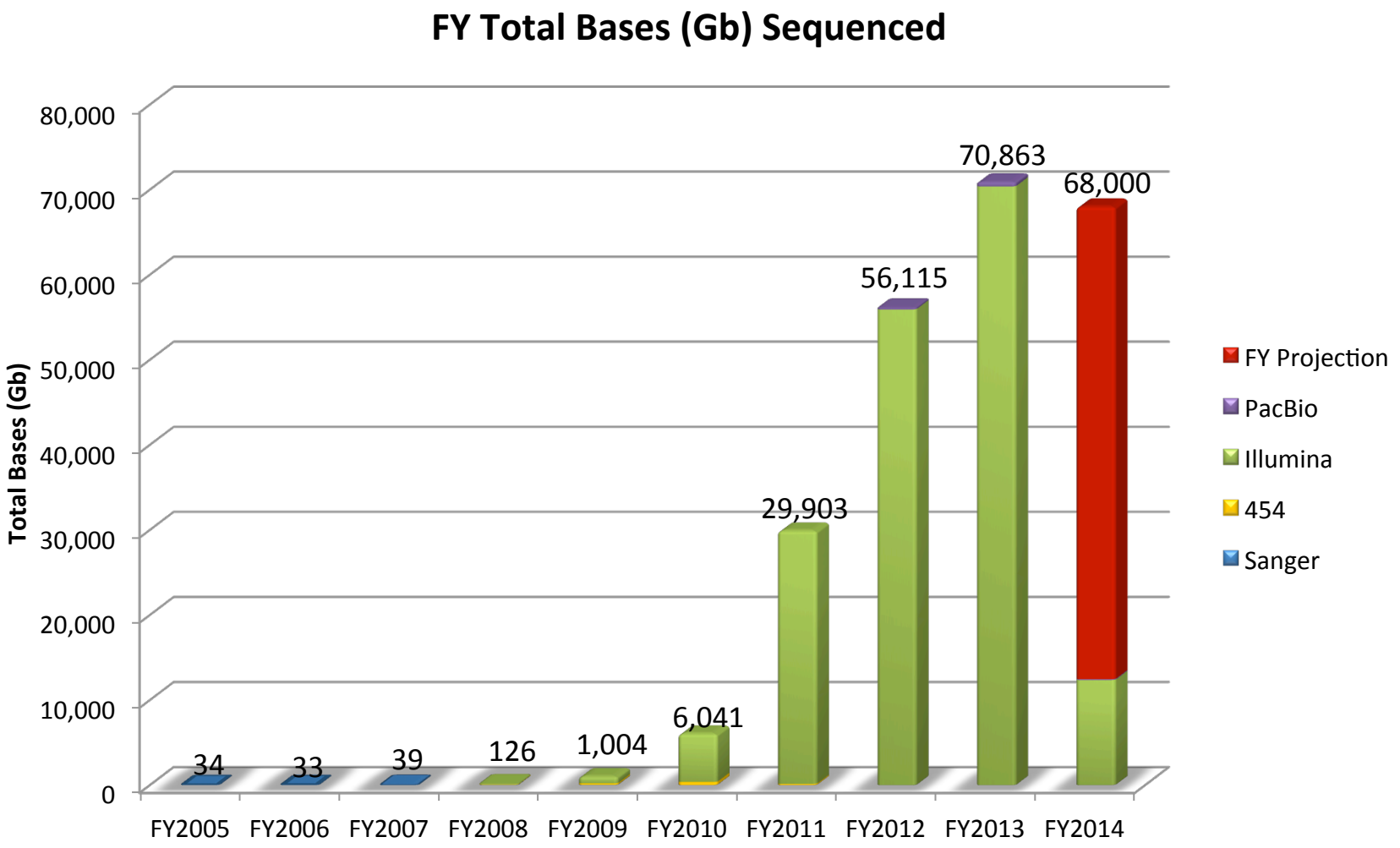
These meethods produce billions of short sequences, each ~100-200 base pairs (bp) long



1 billion clusters per flow cell

1 micron diameter

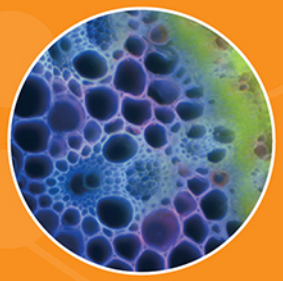
Yearly JGI Sequencing Output



JGI Sequencing Output is Flat
Reagents Costs are Increasing

JGI's Science Programs

DOE Mission Areas

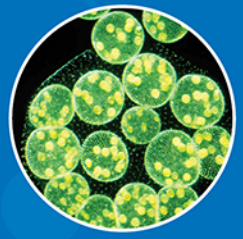

Bioenergy


Carbon Cycling

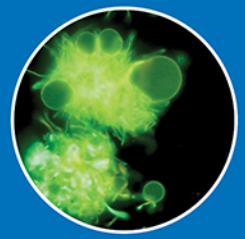

Biogeochemistry

JGI Programs


Metagenomes


Plants


Fungi


Microbes


DNA Synthesis Science

JGI Infrastructure

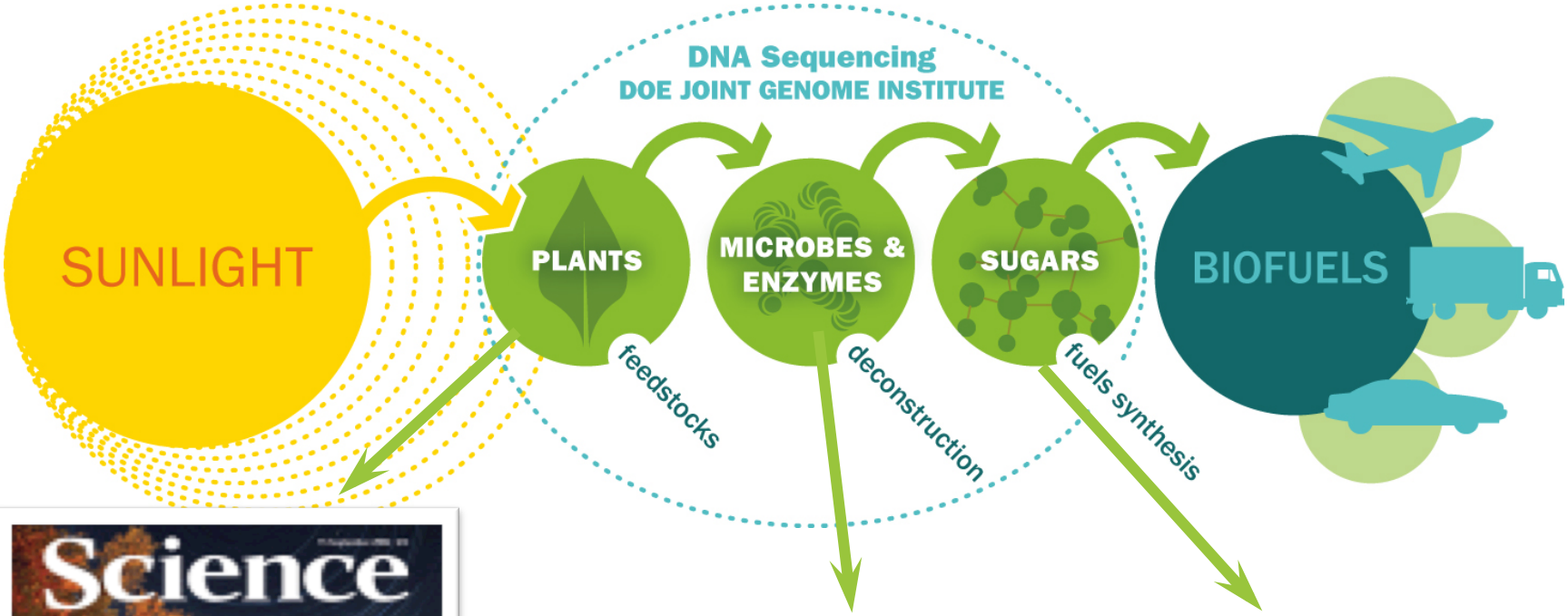

DNA Sequencing

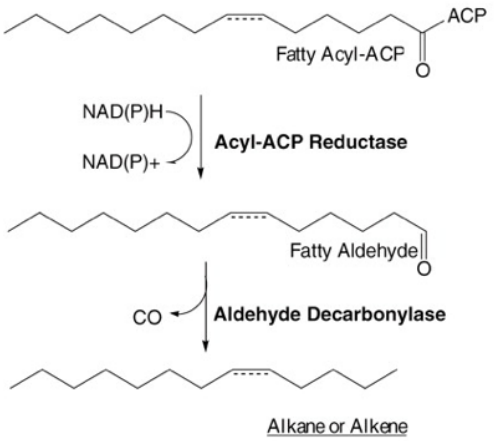

Advanced Genomic Technologies


Computational Analysis

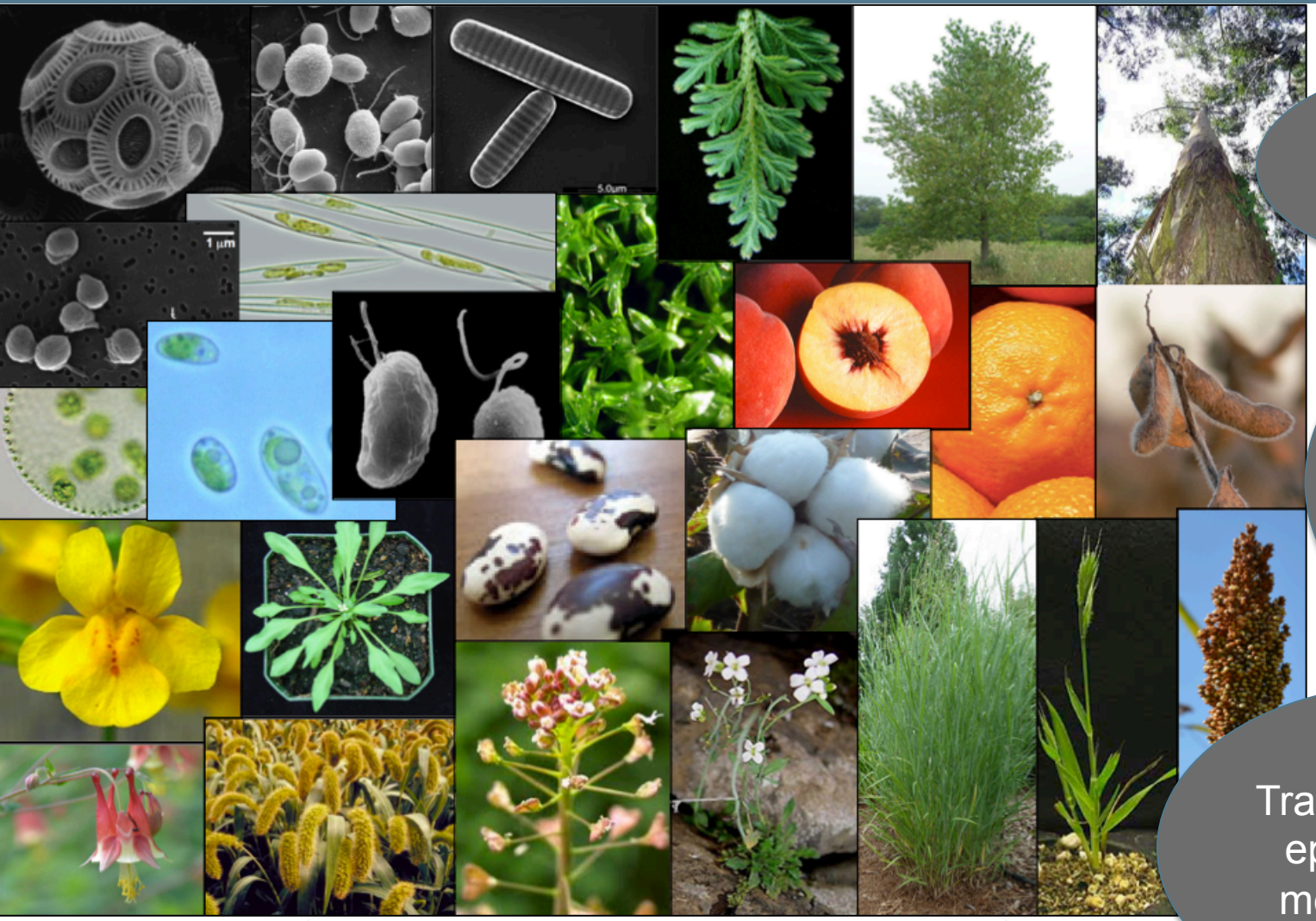

DNA Synthesis

Genomics & Cellulosic Biofuels





JGI Plant Program Focus



Comparative genomics

Genetic diversity

Transcriptomics, epigenomics, metabolomics

Reference Plant Genomes

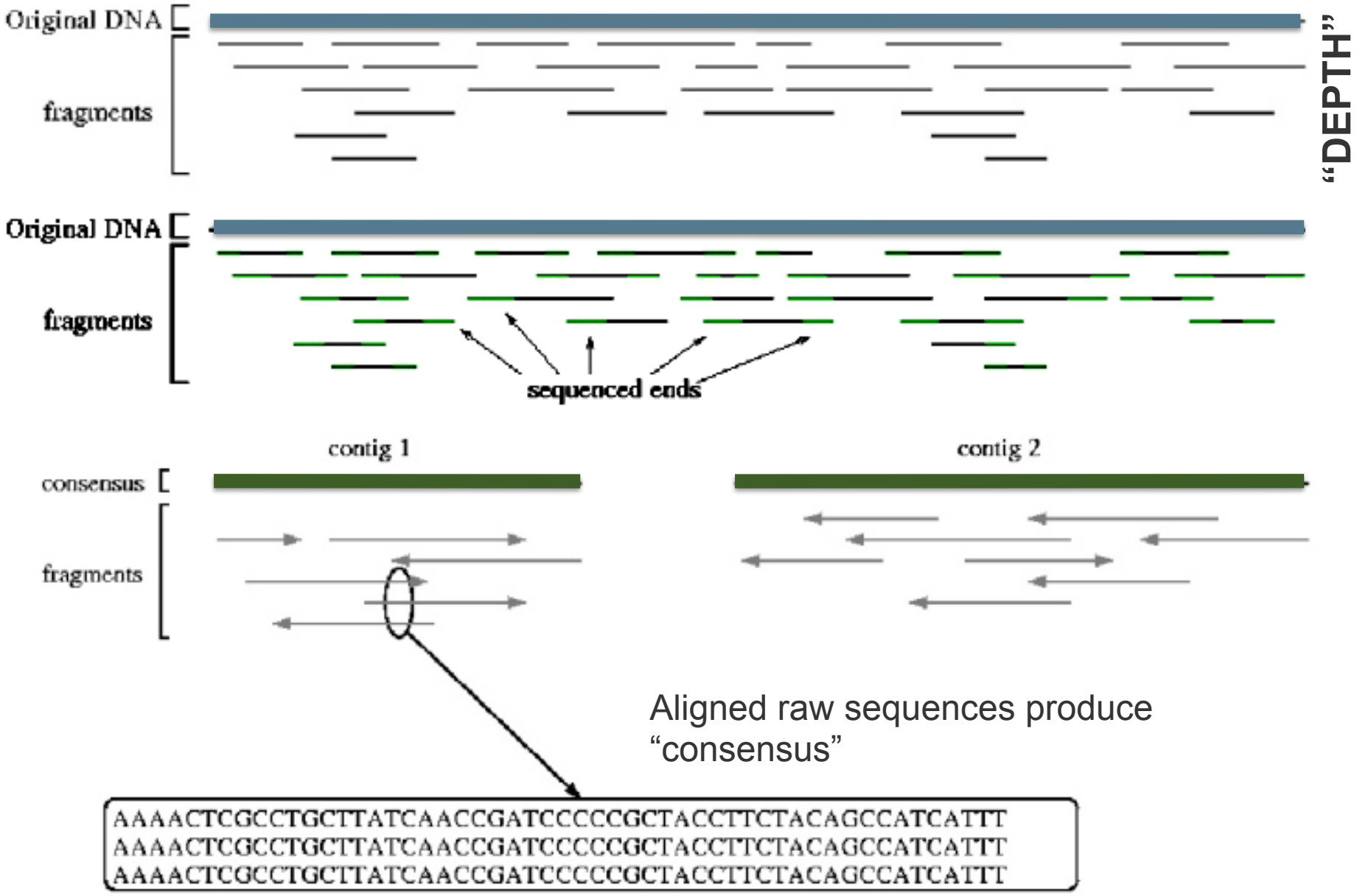
Large-scale genome assembly and mapping

- **JGI: Jarrod Chapman**, Isaac Ho, Eugene Goltsman, Martin Mascher, Dan Rokhsar
- **NERSC/Berkeley/UCSB:** Evangelos Georganas, Veronika Strnadova, Aydin Buluc, Lenny Oliker, Joey Gonzales, John Gilbert, Stefanie Jegelka, Kathy Yelick

- **The "assembly" problem:**
  - We want the DNA sequence of chromosomes
  - Each chromosome is a single DNA molecule that can be represented as a string of the chemical "letters" A,G,C, and T
  - Chromosomes can be hundreds of thousands of "letters" (base pairs, nucleotides) long

- **But the data produced by modern sequencing instruments are billions of short (~hundred letters), redundant sequence fragments ("reads")**

“Shotgun” sequencing: break up genome into short pieces, determine sequence, and reassemble.



Impact of nearly identical repetitive sequence

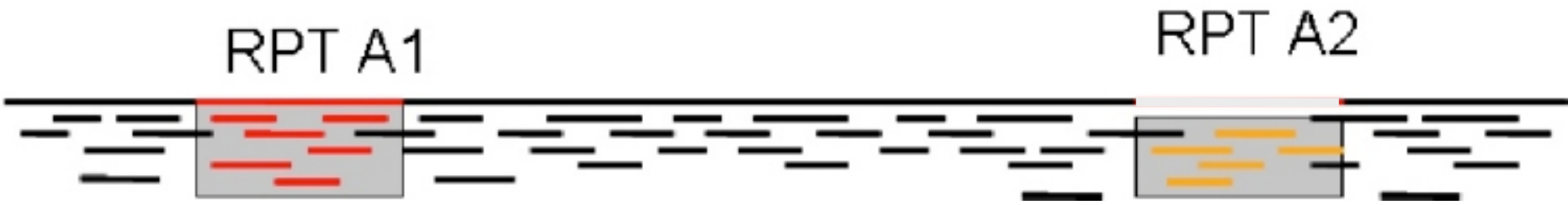


Figure 5. Two copies of a repeat along a genome. The reads colored in red and those colored in yellow appear identical to the assembly program.

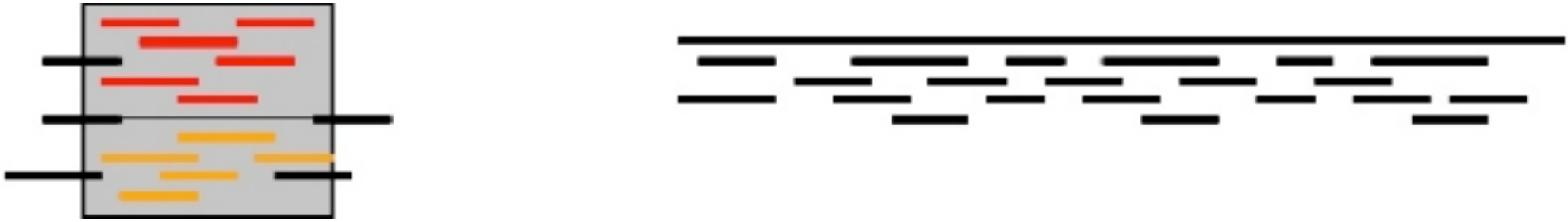
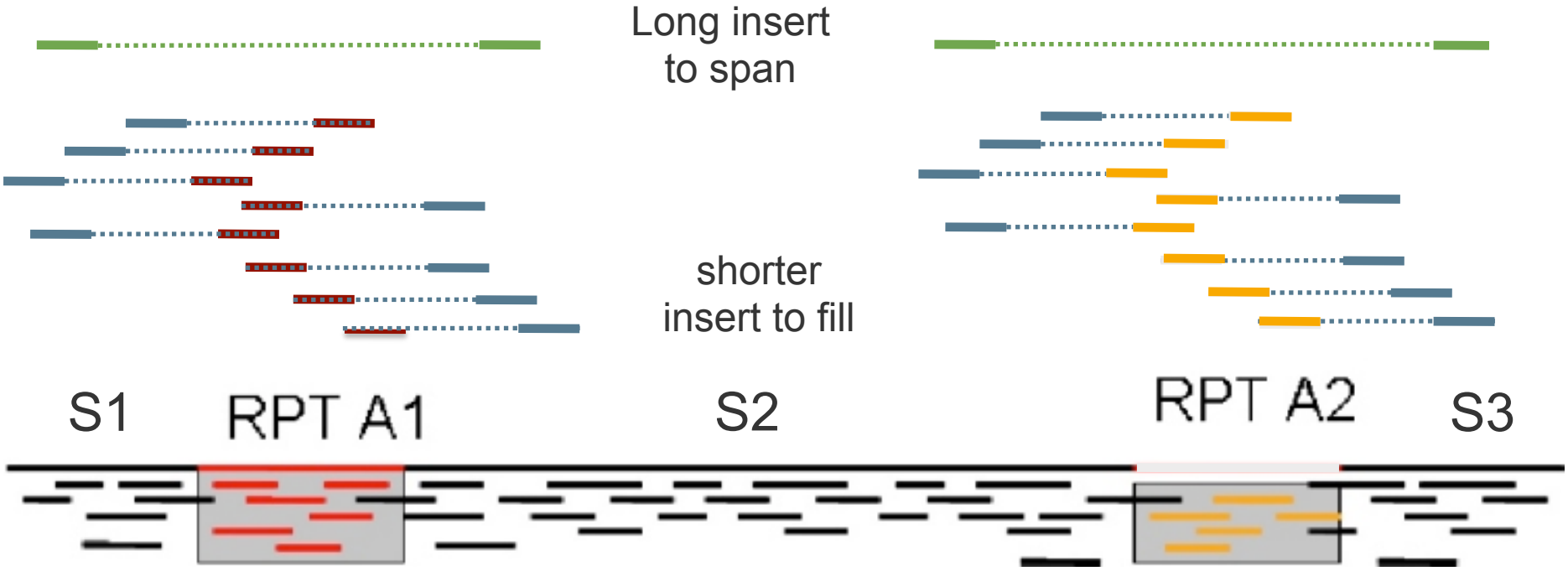


Figure 6. Genome mis-assembled due to a repeat. The assembly program incorrectly combined the reads from the two copies of the repeat leading to the creation of two separate contigs

Paired-ends allow some shorter repetitive sequences to be skipped over and then back-filled

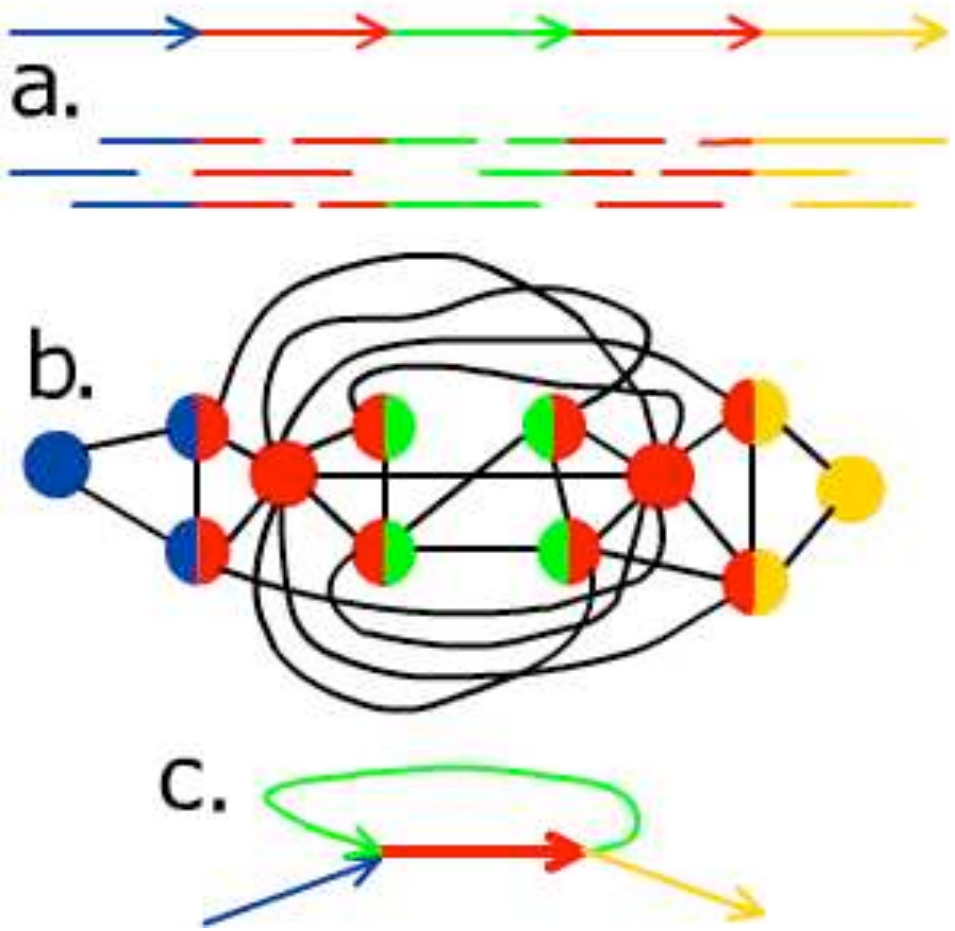


Also can use repeat-boundary spanning reads to define edges

Can represent unresolved assembly as a graph
S1 > A > S2 > A > S3

Complex repetitive sequences

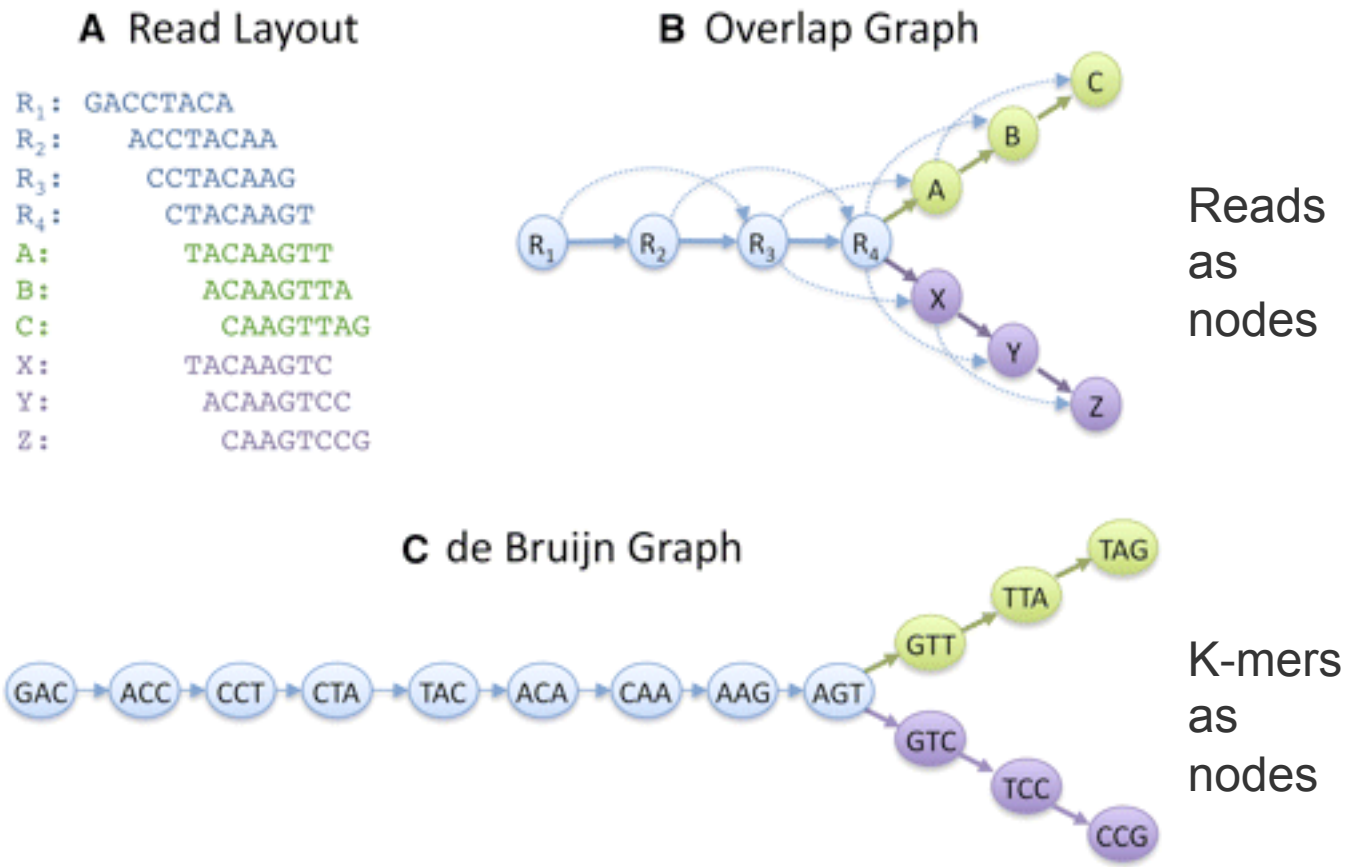
Can sometimes resolve path using flow constraints (repeats are higher apparent coverage) and paired-ends



http://www.ams.org/news/math-in-the-media/mmarc-03-2010-media

DeBruijn graph approach (Waterman, Idury, Pevzner, Myers, et al.)

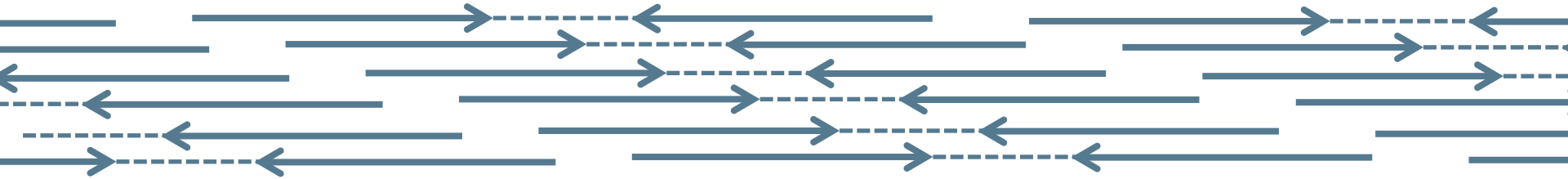
- Represent reads by overlapping k-mers
- Express assembly in terms of connectivity of k-mer graph

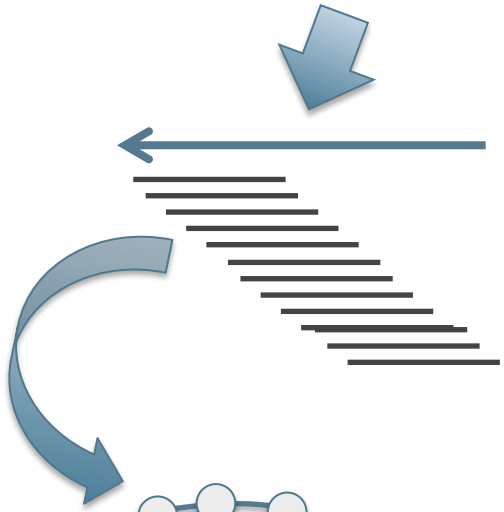


Schatz et al. (2010) Perspective: Assembly of Large Genomes w/2nd-Gen Seq. Genome Res.

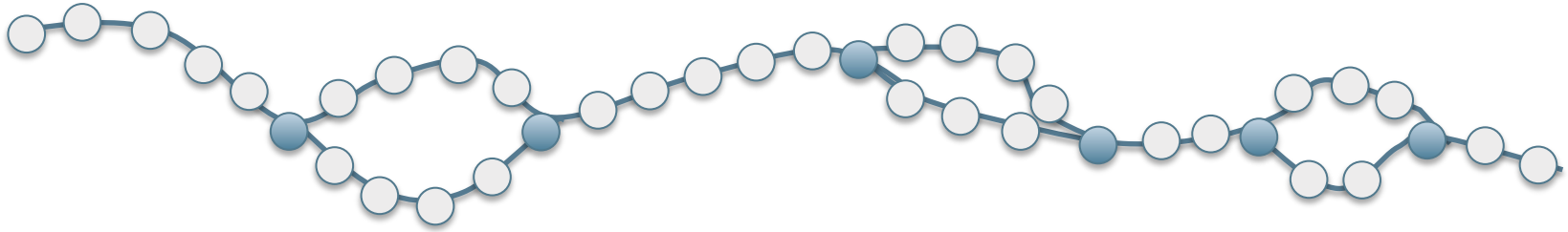
“mer-aculous” genome assembly

“shotgun sequencing” of a diploid genome





Represent each read by a set of overlapping words **(k-mers)** that are long enough to be mostly unique in genome, but short enough to be unlikely to contain a sequencing error

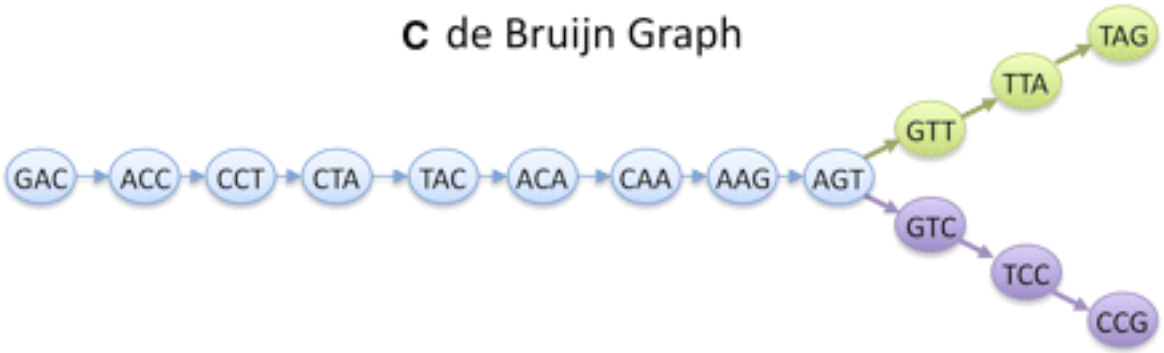


A “stringy” graph with bubbles representing polymorphisms

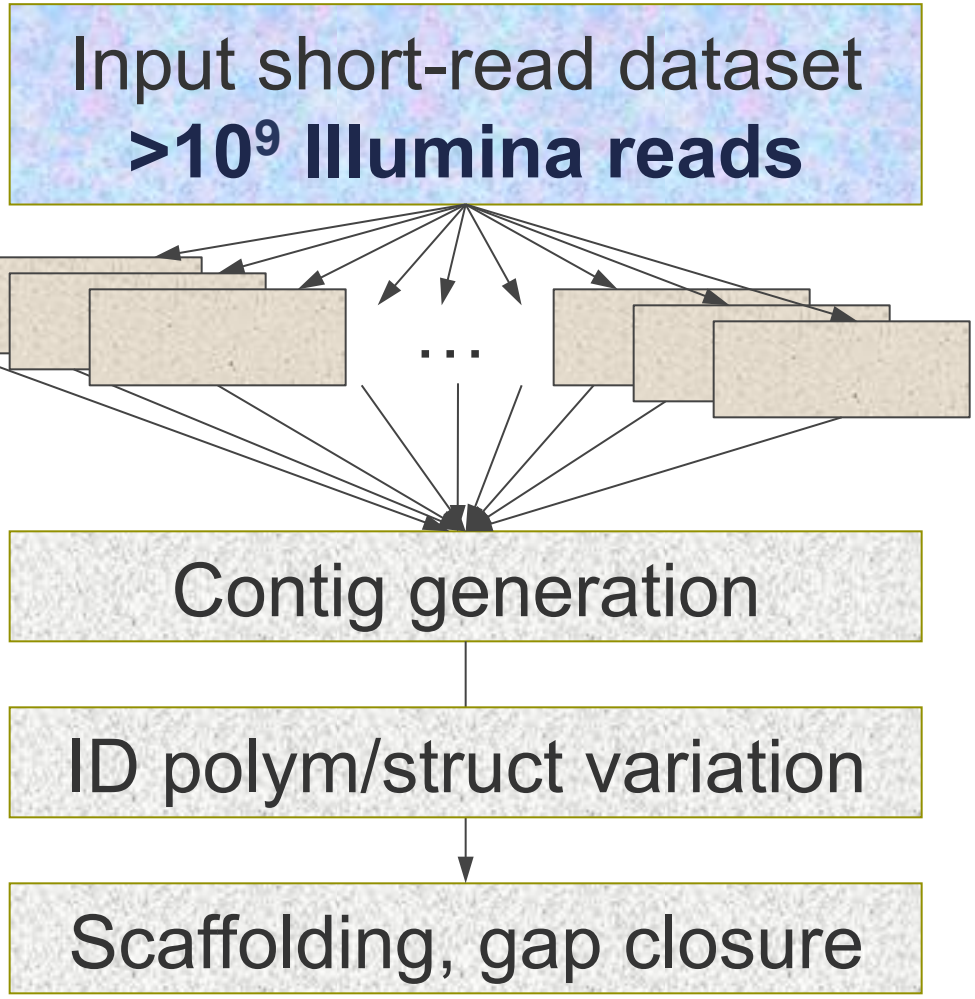
Counting k-mers is computationally simpler than comparing reads to each other to find overlap

Counting k-mers is dependent on total data size, not depth, so it avoids all-vs-all alignments of reads.
In unique genomic sequence, each (error-free) read contains the same k-mers.
So data compression is achieved.

But info is lost in converting reads into strings of adjacent k-mers.
This info must be recovered elsewhere in the algorithm. These are longer-range (beyond nearest neighbor) connections in the graph.



“Meraculous” Chapman, et al.



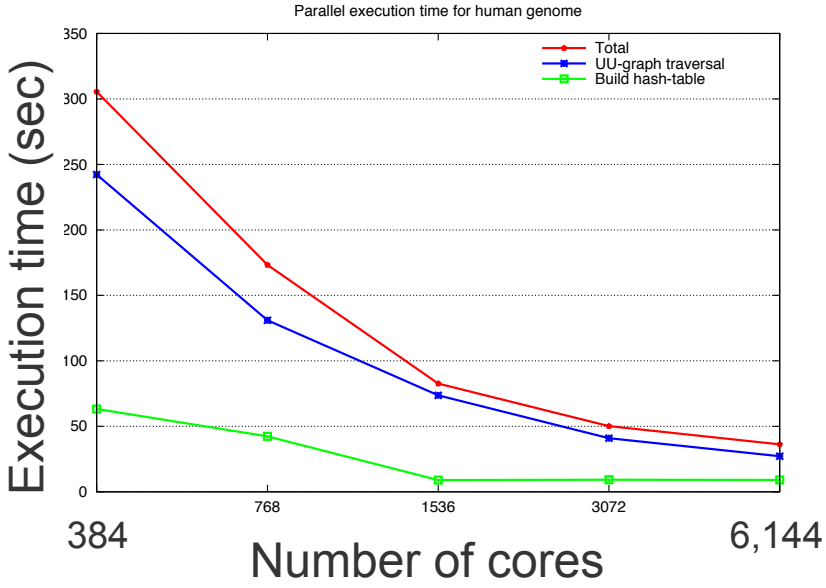
- **Parallelized calculation of mer-graph, traversal, etc.**
- **Load balancing depend on genome, data quality, and uniformity.**
- **led Assemblathon I and II in several key categories.**
- **Ongoing collaboration with NERSC to produce distributed parallel version using UPC/PGAS.**

Special features of metagenomes: sample-specific depth and variation profiles. **Must be learned from the data.**

Towards a highly parallelized mer-aculous assembler

– Jarrod Chapman, Evangelos Georganas, Aydin Buluc, Kathy Yelick, Dan Rokhsar

- **Two critical steps in Meraculous have been translated to UPC** (Unified Parallel C)
- **Realized thousand-fold speedup!**
- **Allows access to arbitrary memory footprint** for large assembly problems
- Remaining steps are also being translated to UPC.



Genomic rearrangements in prostate cancers

With souped-up UPC meraculous, we'll be able to assemble complex genomes like human genomes in minutes ….

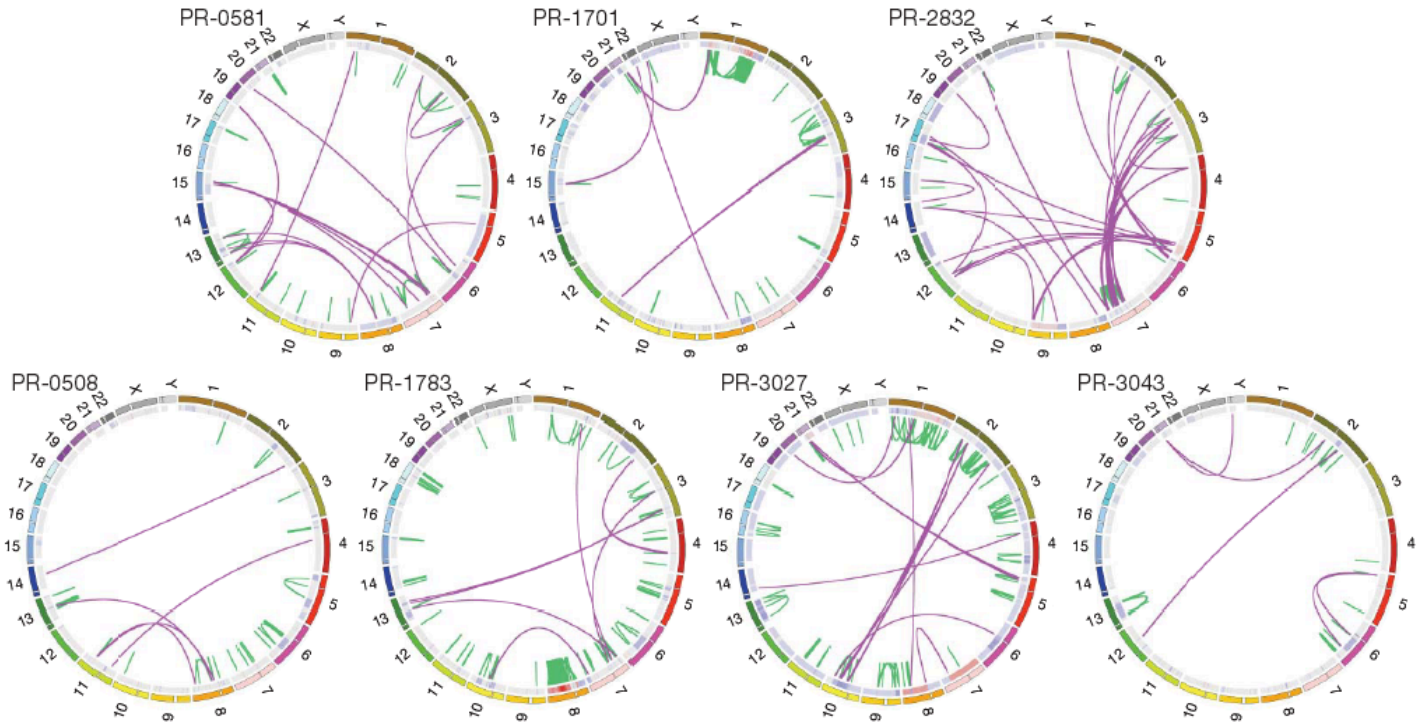


Figure 1 | Graphical representation of seven prostate cancer genomes. Each Circos plot[12] depicts the genomic location in the outer ring and chromosomal copy number in the inner ring (red, copy gain; blue, copy loss). Interchromosomal translocations and intrachromosomal rearrangements are shown in purple and green, respectively. Genomes are organized according to the presence (top row) or absence (bottom row) of the *TMPRSS2–ERG* gene fusion.

Some rearrangements break up genes.
Some of these recur independently across prostate cancers.

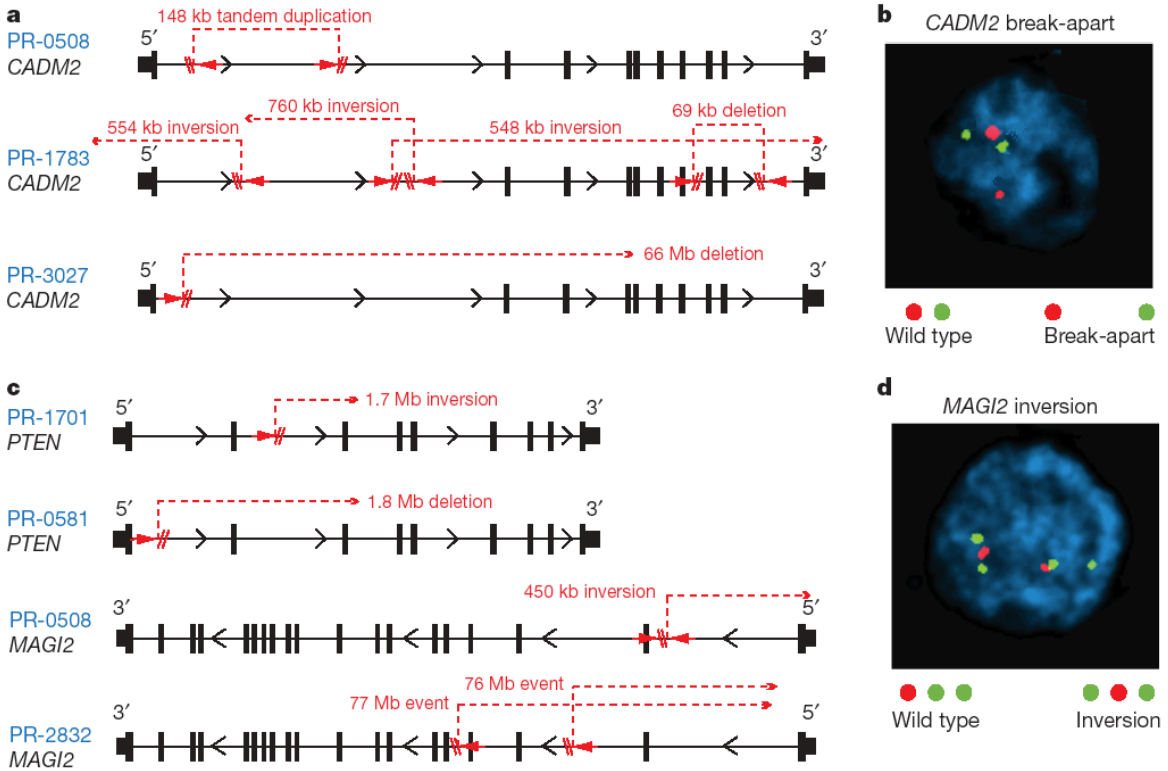
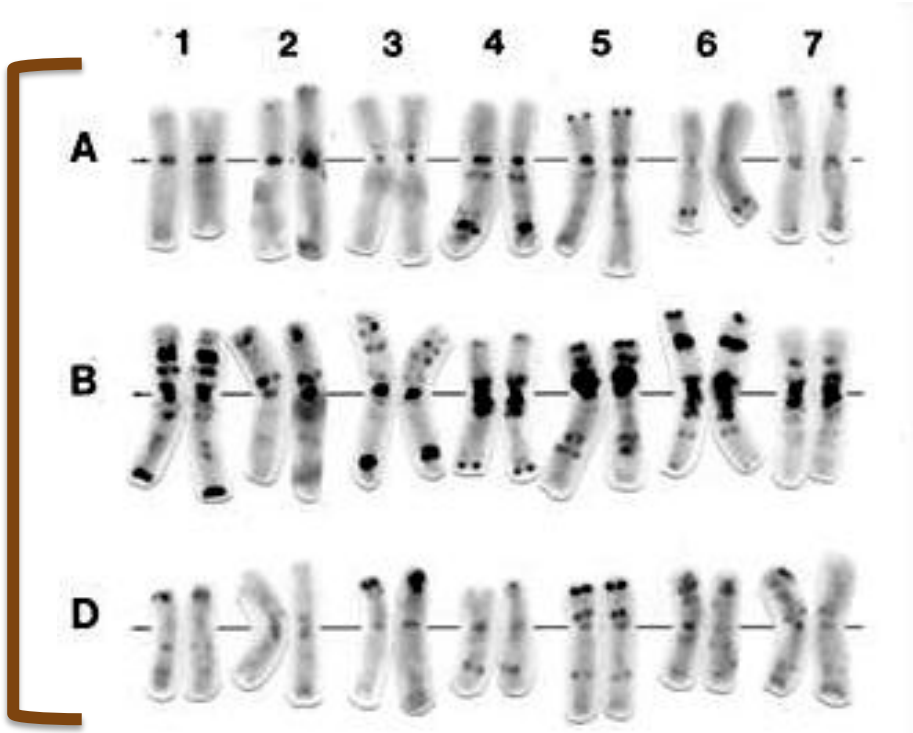


Figure 4 | Disruption of *CADM2* and the *PTEN* pathway by rearrangements. **a**, Location of intragenic breakpoints in *CADM2*. **b**, *CADM2* break-apart demonstrated by FISH in an independent prostate tumour. **c**, Location of intragenic breakpoints in *PTEN* (top) and *MAGI2* (bottom). **d**, *MAGI2* inversion demonstrated by FISH in an independent prostate tumour, using probes flanking *MAGI2* (red and green) and an external reference probe also on chromosome 7q (green). The probes and strategy for detecting novel rearrangements by FISH are shown in diagram form in Supplementary Fig. 8.

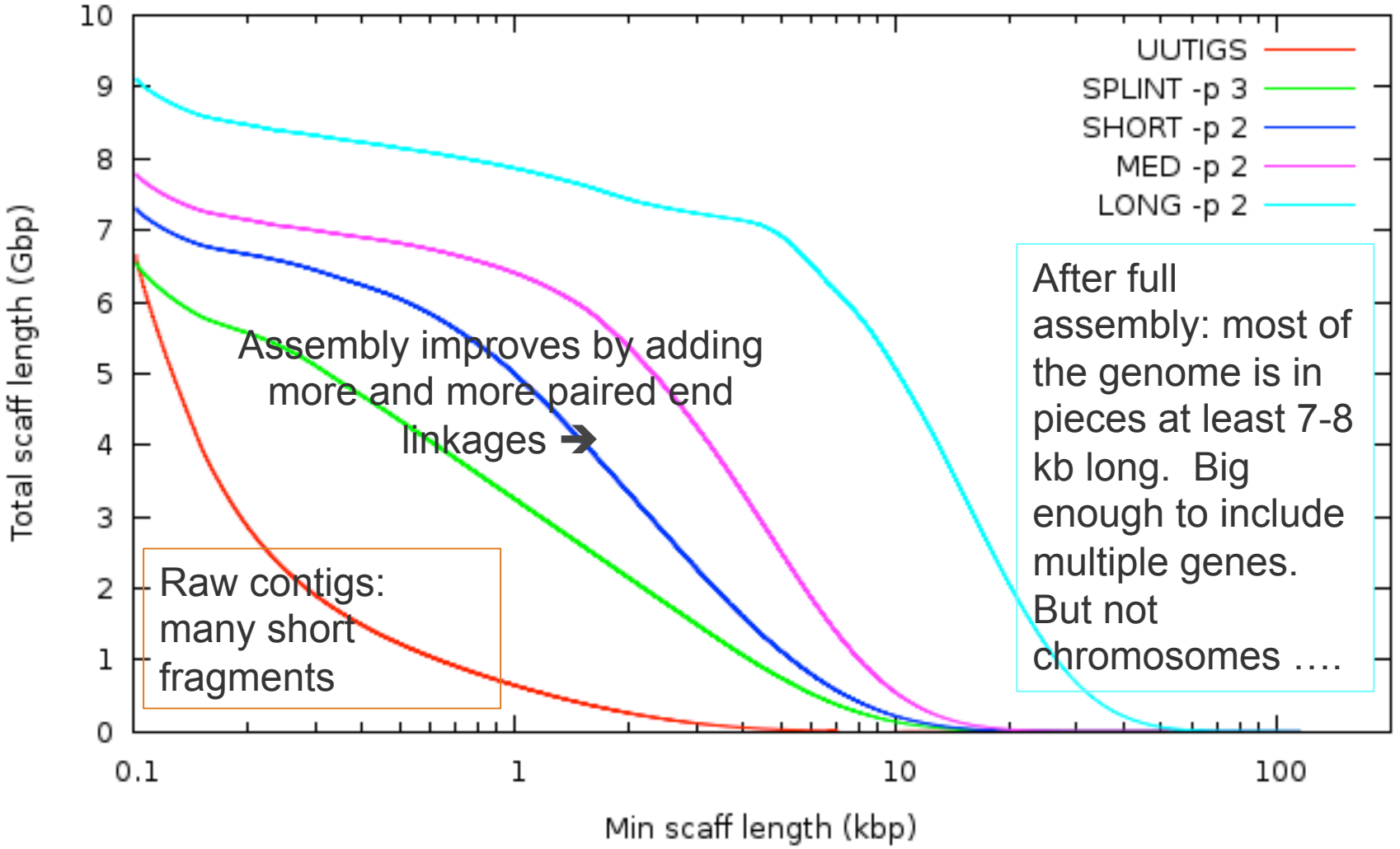
Can we access these kinds of structural changes through a ”diff” that acts on a pair of mer-graphs?

Assembly humongous genomes

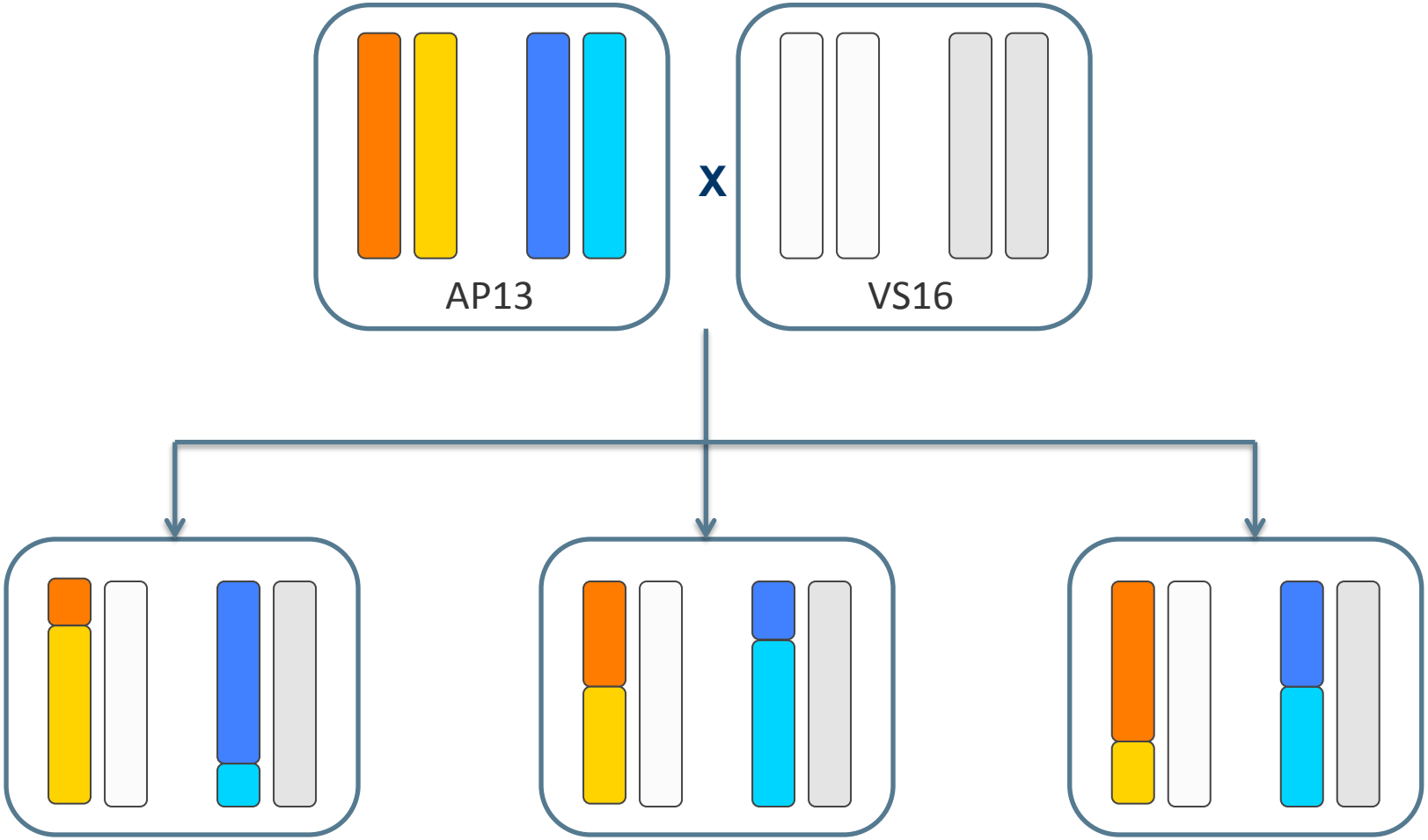
- Human genome: 3 giga-base pairs
- Maize: 2.4 Gbp
- Switchgrass: 1.4 Gb
- Miscanthus: 2.5 Gb
- …
- Barley genome: 7 Gbp
- **Wheat genome: 17 Gbp**
- Pine genome: 20 Gbp
- Salamander: 20-30 Gbp



Progress towards assembling wheat genes (hybrid C/UPC version)



Genetic mapping with millions of markers

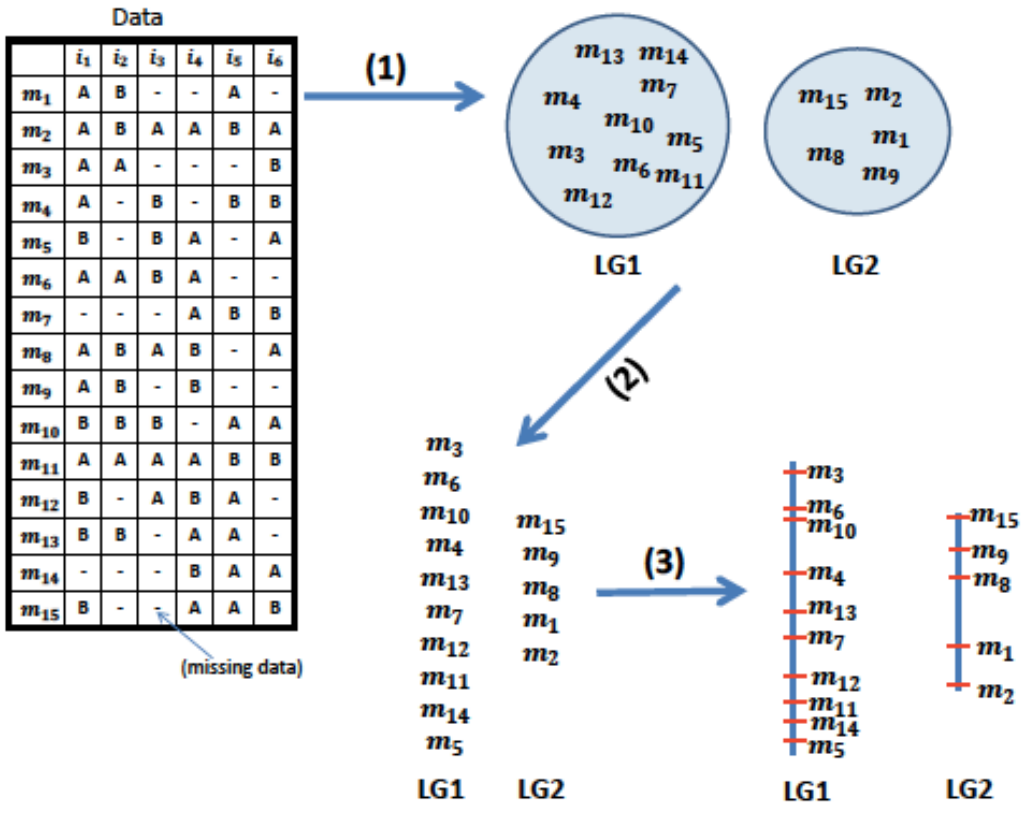


F1 recombinants track “orange” vs “yellow” in offspring

Efficient and Accurate Clustering for Large-Scale Genetic Mapping *

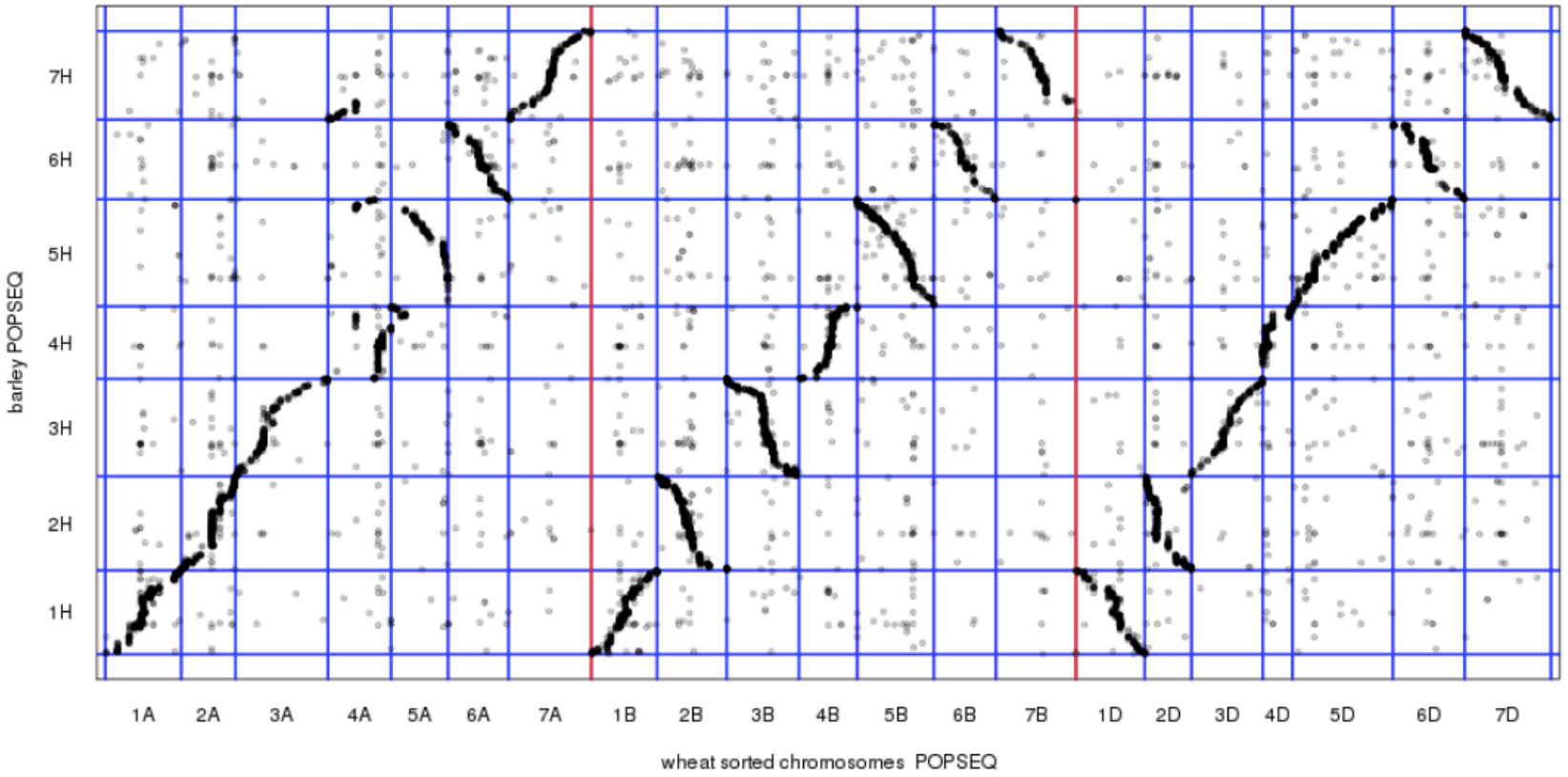
Veronika Strnadova † Aydın Buluç‡ Jarrod Chapman§ John R. Gilbert¶
Joseph Gonzalez‖ Stefanie Jegelka** Daniel Rokhsar†† Leonid Oliker‡‡

- **Given a collection of "markers" (short sequences that can be either "yellow" or "orange")**
- **Compare markers with every other marker, across all family members.**
- **Look for pairs of markers such that when one is yellow, the other is almost always yellow; when one is orange, the other is almost always orange.**
  - "Color" is unknown at the start. Like a (trivial) Ising gauge factor.
  - Amount of mis-correlation is related to distance along the chromosome.
- **Need clever ways to organize the calculation, taking advantage of the inherent linearity of chromosomes.**
- **Can handle 10's of millions of markers, far more than other genetic mapping codes.**



	i_1	i_2	i_3	i_4	i_5	i_6
m_1	A	B	-	-	A	-
m_2	A	B	A	A	B	A
m_3	A	A	-	-	-	B
m_4	A	-	B	-	B	B
m_5	B	-	B	A	-	A
m_6	A	A	B	A	-	-
m_7	-	-	-	A	B	B
m_8	A	B	A	B	-	A
m_9	A	B	-	B	-	-
m_{10}	B	B	B	-	A	A
m_{11}	A	A	A	A	B	B
m_{12}	B	-	A	B	A	-
m_{13}	B	B	-	A	A	-
m_{14}	-	-	-	B	A	A
m_{15}	B	-	-	A	A	B

Progress in mapping the wheat genome



Within a week or two we will have a de novo assembly and map of the wheat genome, after the algorithmic dust settles!

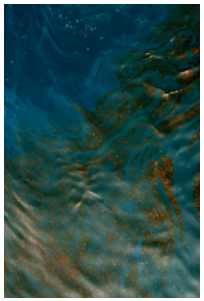
“Annotating” DNA sequence

- **JGI: Amrita Pati,** Marcel Huntemann, Nikos Kyrpides
- **NERSC:** Seung-Jin Sul, Kjiersten Fagnan, Shane Canon

- **The “annotation” problem:**
  - When sequencing a “metagenome” from a microbial community, we may sample a billion gene fragments, derived from the constituent microbes.
  - Q. How do we computationally infer the function of these fragments?
  - A. By grouping related gene sequences together, and detecting similarity with genes of known function.

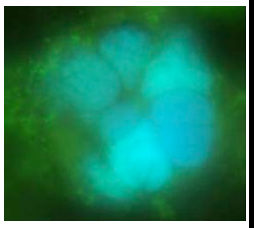
Microbial communities ("meta" genomes)

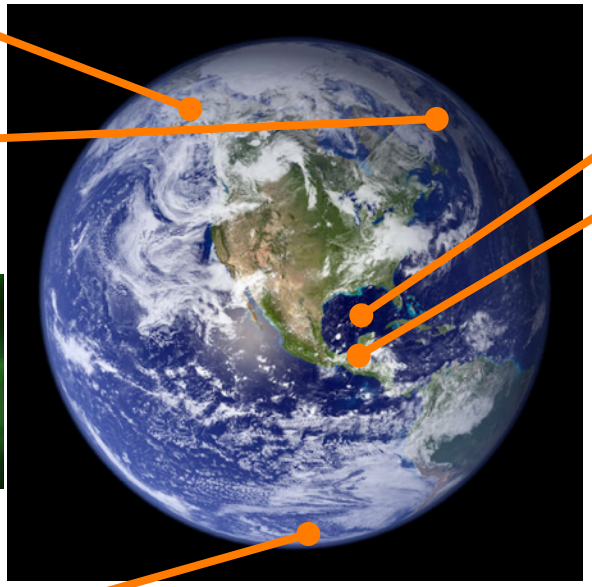
Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*



Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J*

The metagenome of a marine anammox bacterium illustrates role in the global nitrogen cycle. *Environmental Microbiology*





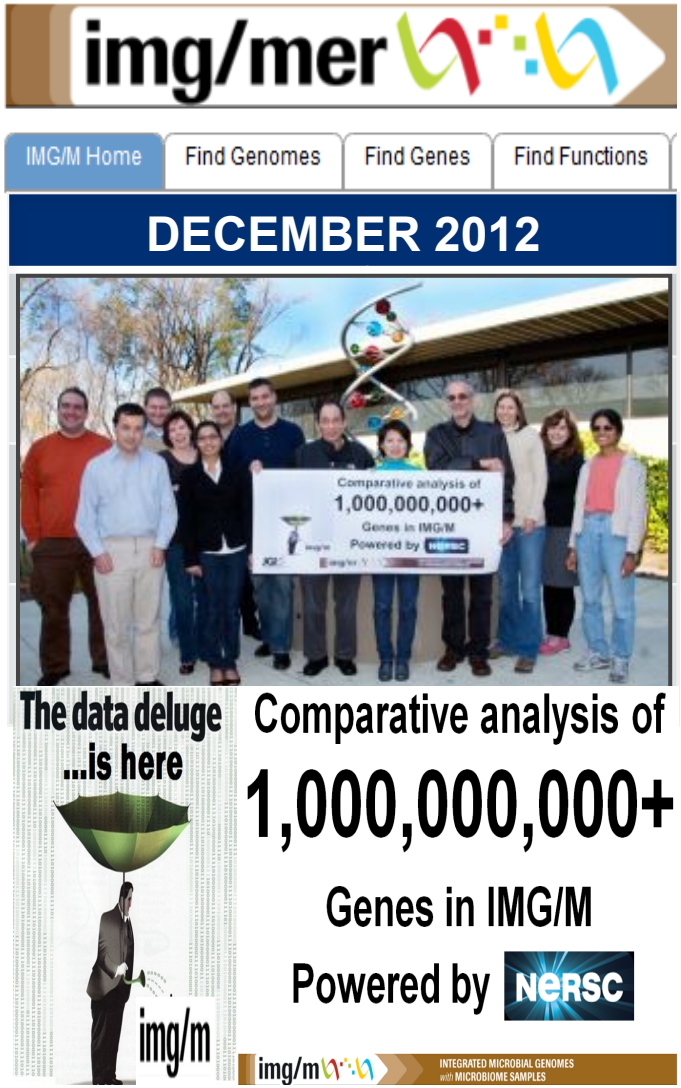
Global transcriptome response to ionic liquid by a tropical rain forest soil bacterium, *Enterobacter lignolyticus*. *PNAS*

The genome of the polar eukaryotic microalga coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biology*

Growth of metagenomic data

JGI JOINT GENOME INSTITUTE

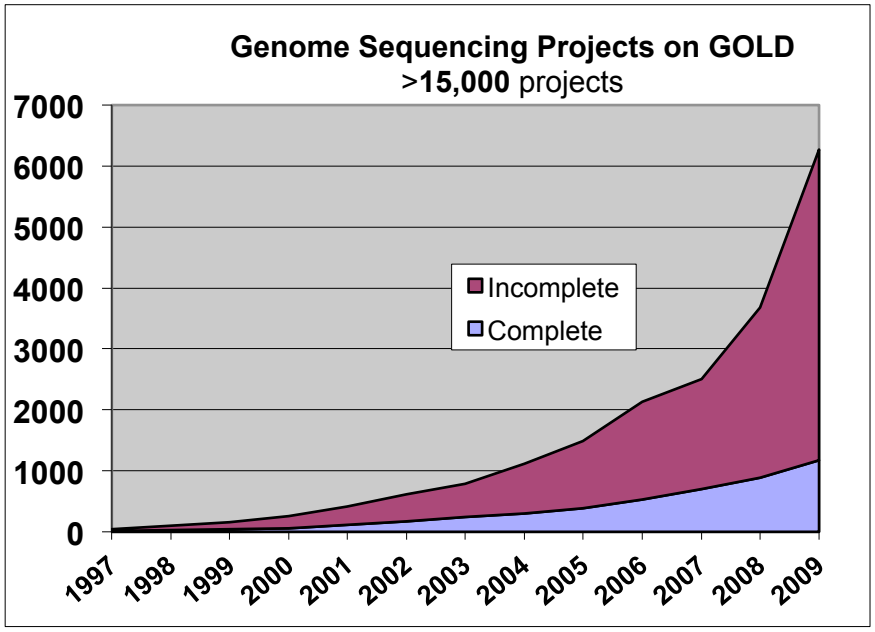
POWERED BY NERSC



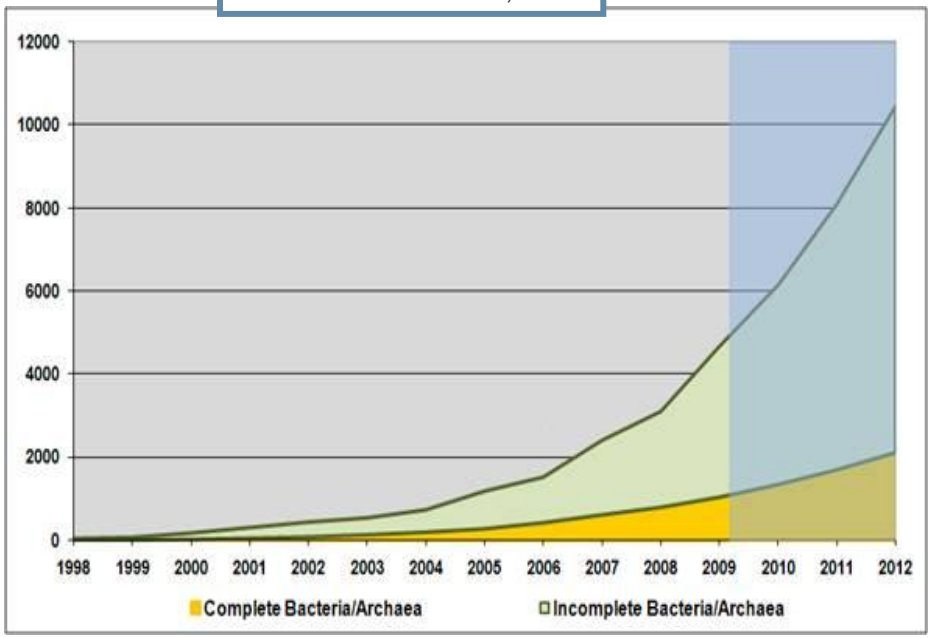
February 2014	
Samples	3,854
DNA (bps)	3,9 Tb
Genes	23.4 Billion

CLUSTER CAPACITY	
minimum	250 M genes /wk
maximum	1.3 Billion genes /wk

Microbial Genome Projects



P. Chain *et al.* Science, 2009



	1995-2009	2010-2014
Finished	1000	3,000
Draft	1000	10,000
Genes	6 Million	39 Million

now
2,975
12,379
52 Million

Metagenomics analysis challenges

JGI JOINT GENOME INSTITUTE

POWERED BY NERSC

- **How do you compare 15 Billion genes (all vs all)?**

1. **Cluster all genes from "isolate" genomes**
   - 22 Million Genes

2. **Pledge all genes of the metagenomes to the clusters of the isolate genomes**
   - 15 Billion genes

3. **Cluster all the metagenomes genes that did not pledge and create additional (purely metagenomics) clusters.**
   - These may have interesting unknown functions

1. Generation of isolate protein clusters

IMG-nr

↓ BLAT

All vs. all similarities, retain highly similar pairs

70% alignment and 70% identity

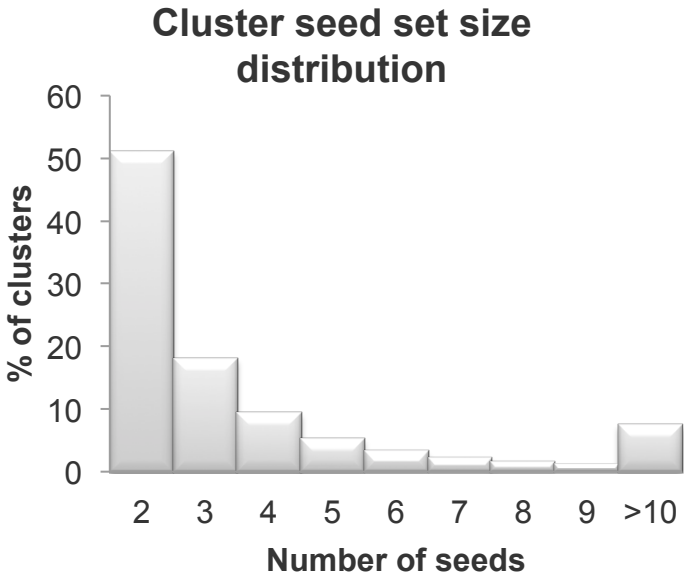
Single-linkage clustering

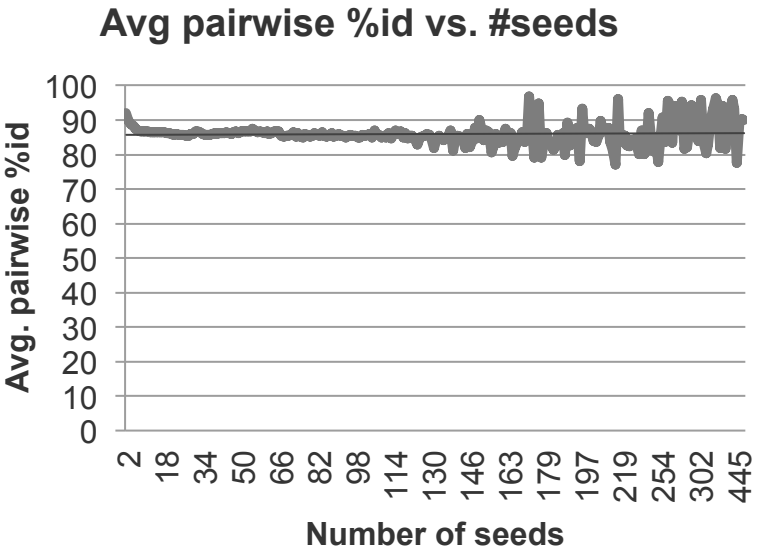
UClust

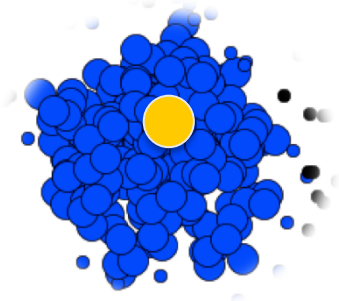
Refine clusters obtained above

70% alignment and 70% identity

Isolate protein clusters



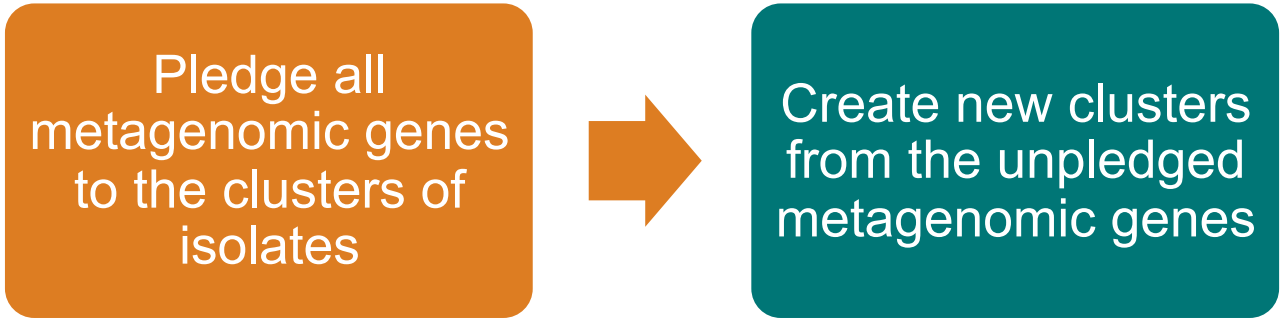




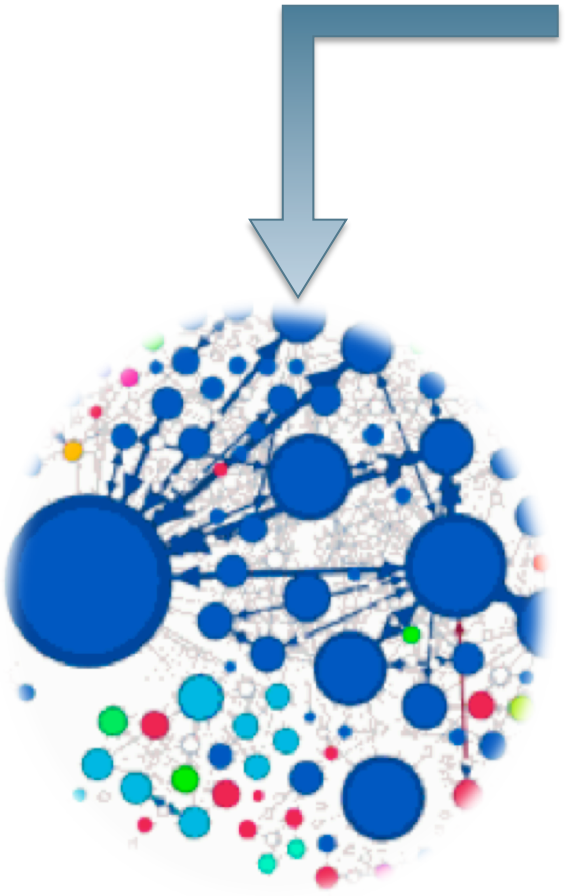
Isolate Genomes	
Total Genes	22 Million
NR Genes	16 Million
Clusters	1.8 Million
Genes in clusters	8 Million
Singletons	8 Million

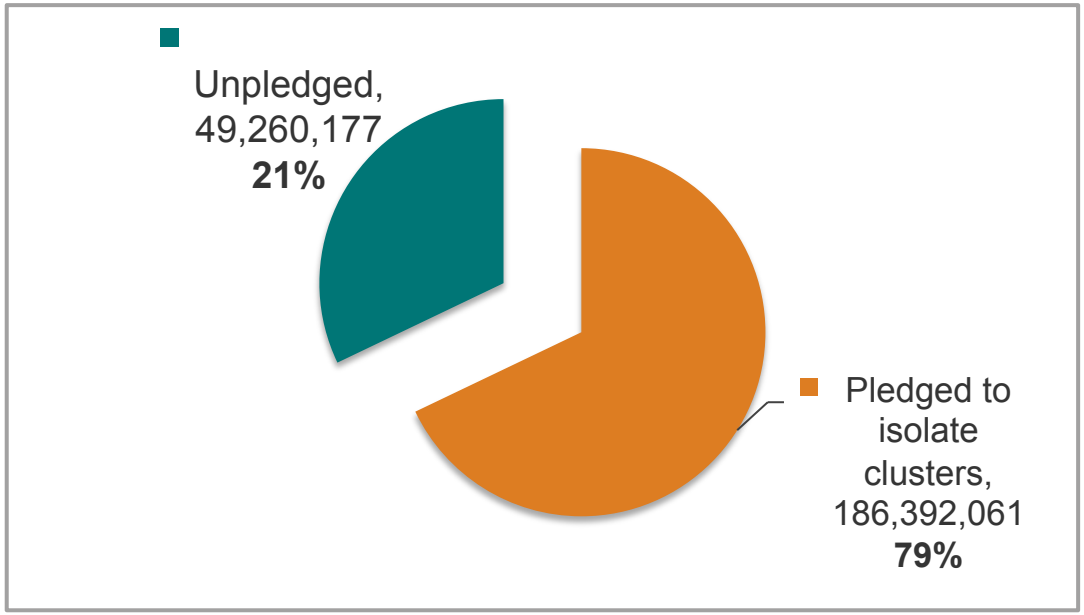
Next steps

Create new clusters from metagenomic genes that don't have any hits to gene clusters from isolates



Metagenomic Clustering from assembled metagenomes

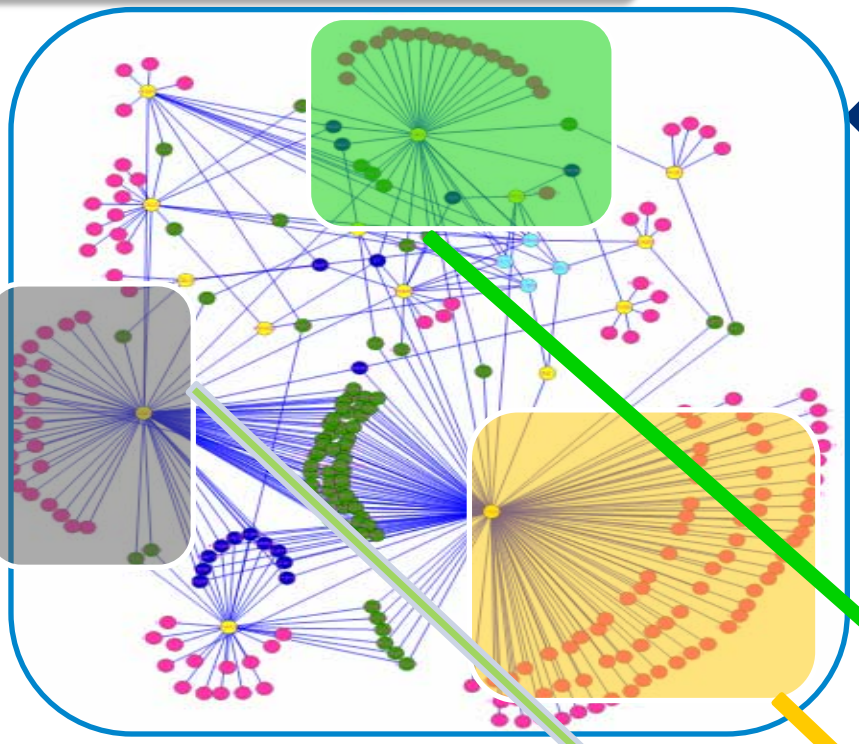




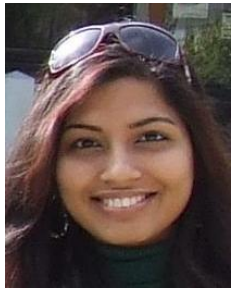
Metagenome Genes	
Total Genes	49 Million
Clusters	1.3 Million
Genes in clusters	25 Million
Singletons	24 Million

Metagenome clusters
1.27M clusters
Within a cluster linkages have 70%id
and 50% alignment length across both members

IMG/M: 23 Billions Genes

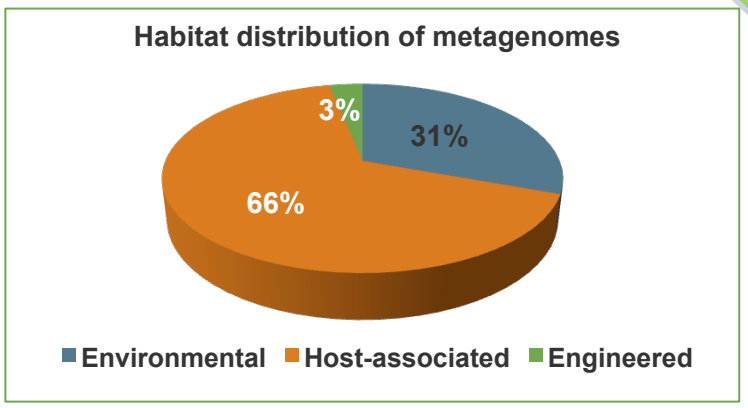


Gene Clustering



Metagenome Classification

- Metagenomic Samples (1930)
  - Environmental (711)
    - Aquatic (580)
      - Non-marine Saline and Alkaline (70)
      - Marine (264)
      - Thermal springs (0)
      - Freshwater (152)
    - Terrestrial (119)
      - Rock-dwelling (subaerial biofilms) (5)
      - Soil (113)
      - Deep subsurface (1)
    - Air (7)
      - Indoor Air (2)
      - Outdoor Air (4)
  - Host-associated (1122)
  - Engineered (97)



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Marcel Huntemann
Amrita Pati (PI)

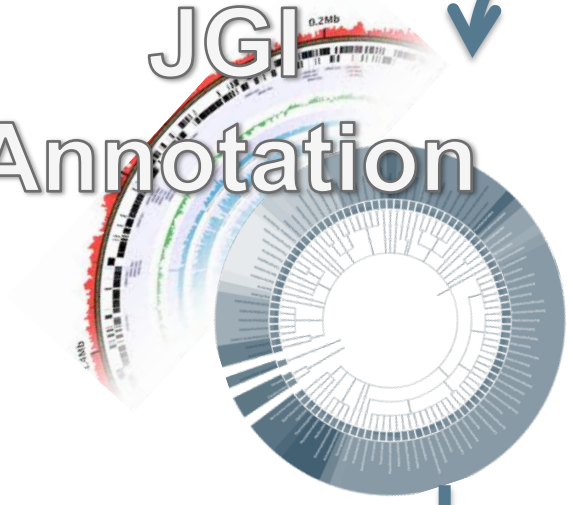
- Terabytes of sequence data
- JGI engineers
- Distributed computational paradigms
- Tools for bioinformatics sequence analysis

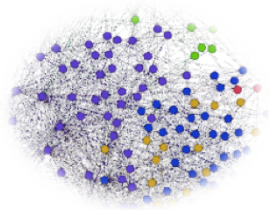
Seung-Jin Sul
Shane Canon (PI)

- Massive computational systems
- Consultants to help build interfaces with custom hardware

NERSC

JGI Annotation





Functional and phylogenetic annotation for 12 Billion metagenome genes in 12 months

Massive backlogs cleared

R&D towards generating scalable computational frameworks for big sequence data

Computing for Genomics

- **The DOE Joint Genome Institute**
  - Who we are, what we do, and why
- **Two computational problems in genomics**
  - "Assembling" genomes from "shotgun" sequence
  - "Annotating" "metagenomic" data
- **We have a User Group meeting too!**

GENOMICS OF ENERGY AND ENVIRONMENT

Meeting

March 18 - 20, 2014

Walnut Creek, CA

- **DNA Synthesis & Synthetic Biology**
- **Single-Cell Genomics for Bioprospecting**
- **Biofuel Traits in Biomass Feedstocks**
- **HPC for Next-Gen Sequencing Applications**
- **Functional Metagenomics**

Invited presentations, workshops and tutorials on sequence-based bioinformatics, and data management systems.

Thanks!

Assembly

- **JGI: Jarrod Chapman**, Isaac Ho, Eugene Goltsman, Martin Mascher, Dan Rokhsar
- **NERSC/Berkeley/UCSB:** Evangelos Georganas, Veronika Strnadova, Aydin Buluc, Lenny Oliker, Joey Gonzales, John Gilbert, Stefanie Jegelka, Kathy Yelick

Metagenome annotation

- **JGI: Amrita Pati,** Marcel Huntemann, Nikos Kyrpides
- **NERSC:** Seung-Jin Sul, Kjiersten Fagnan, Shane Canon

Thanks to DOE OBER, ASCR, and LBNL for support

Also: Thanks to NERSC for JGI computing partnership!