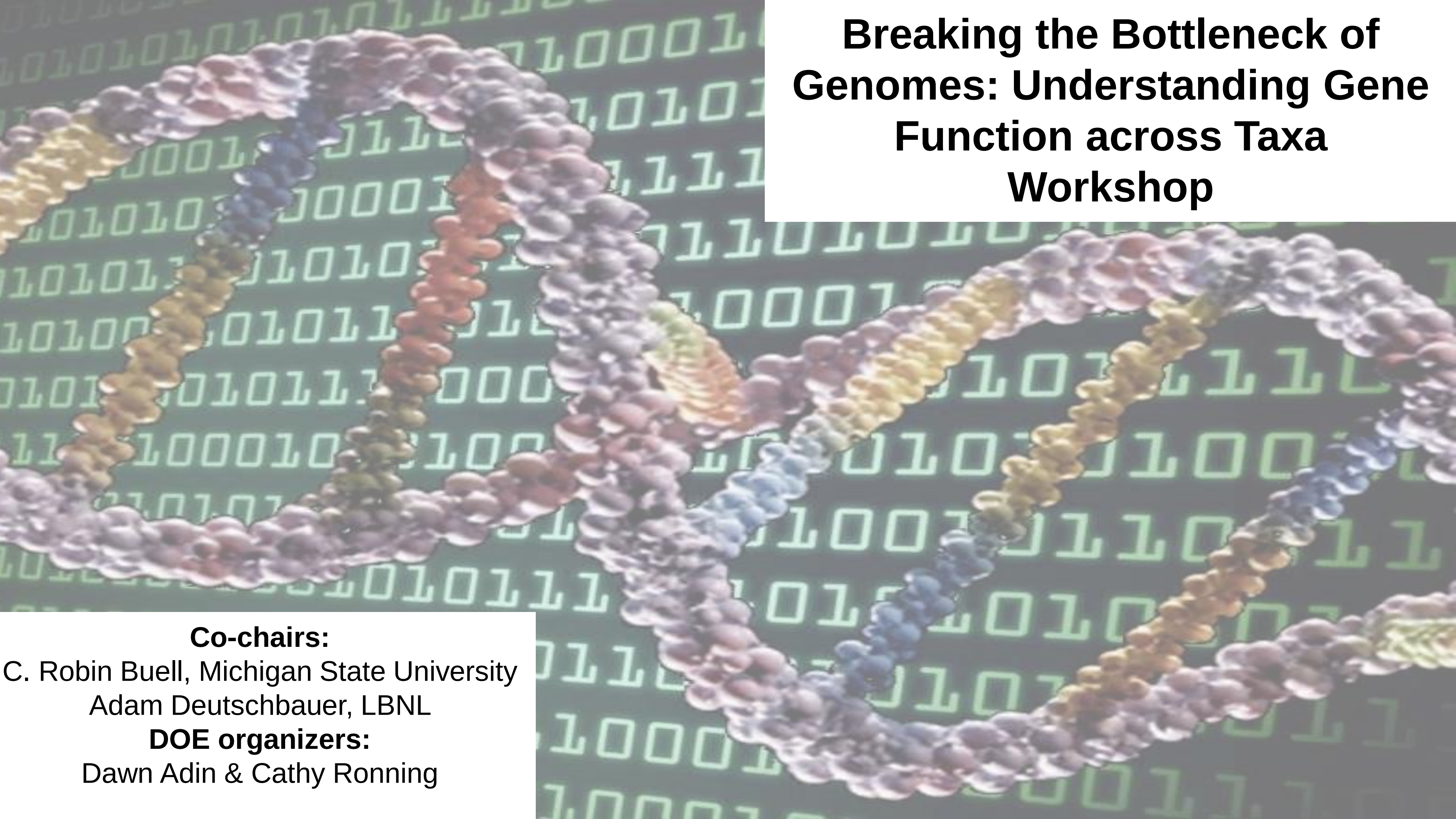
Breaking the Bottleneck of Genomes: Understanding Gene Function across Taxa Workshop



Co-chairs:

C. Robin Buell, Michigan State University

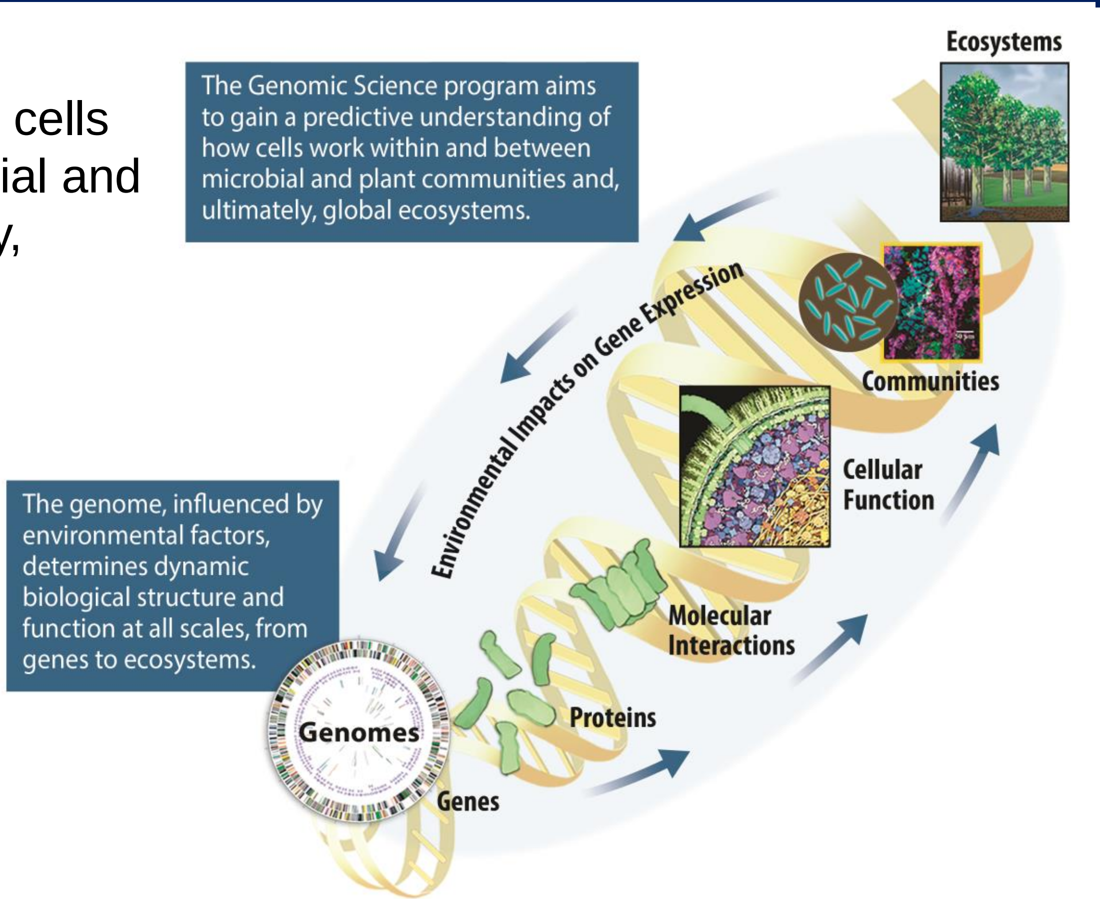
Adam Deutschbauer, LBNL

DOE organizers:

Dawn Adin & Cathy Ronning

Genomics Sciences Program

Goal:
Predictive understanding of how cells work within and between microbial and plant communities and ultimately, global ecosystem level

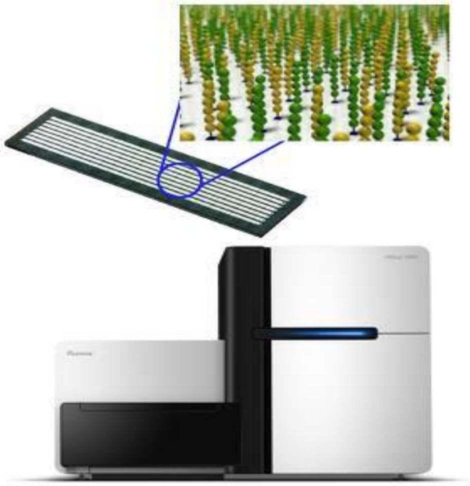


The Genomic Science program aims to gain a predictive understanding of how cells work within and between microbial and plant communities and, ultimately, global ecosystems.

The genome, influenced by environmental factors, determines dynamic biological structure and function at all scales, from genes to ecosystems.

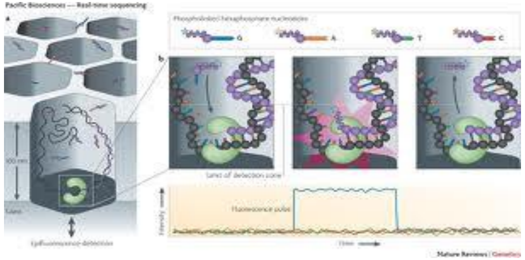
Technology continues to advance throughput coupled with decreased costs

-Ultra,ultra high throughput & very inexpensive



©2010, Illumina Inc. All rights reserved.

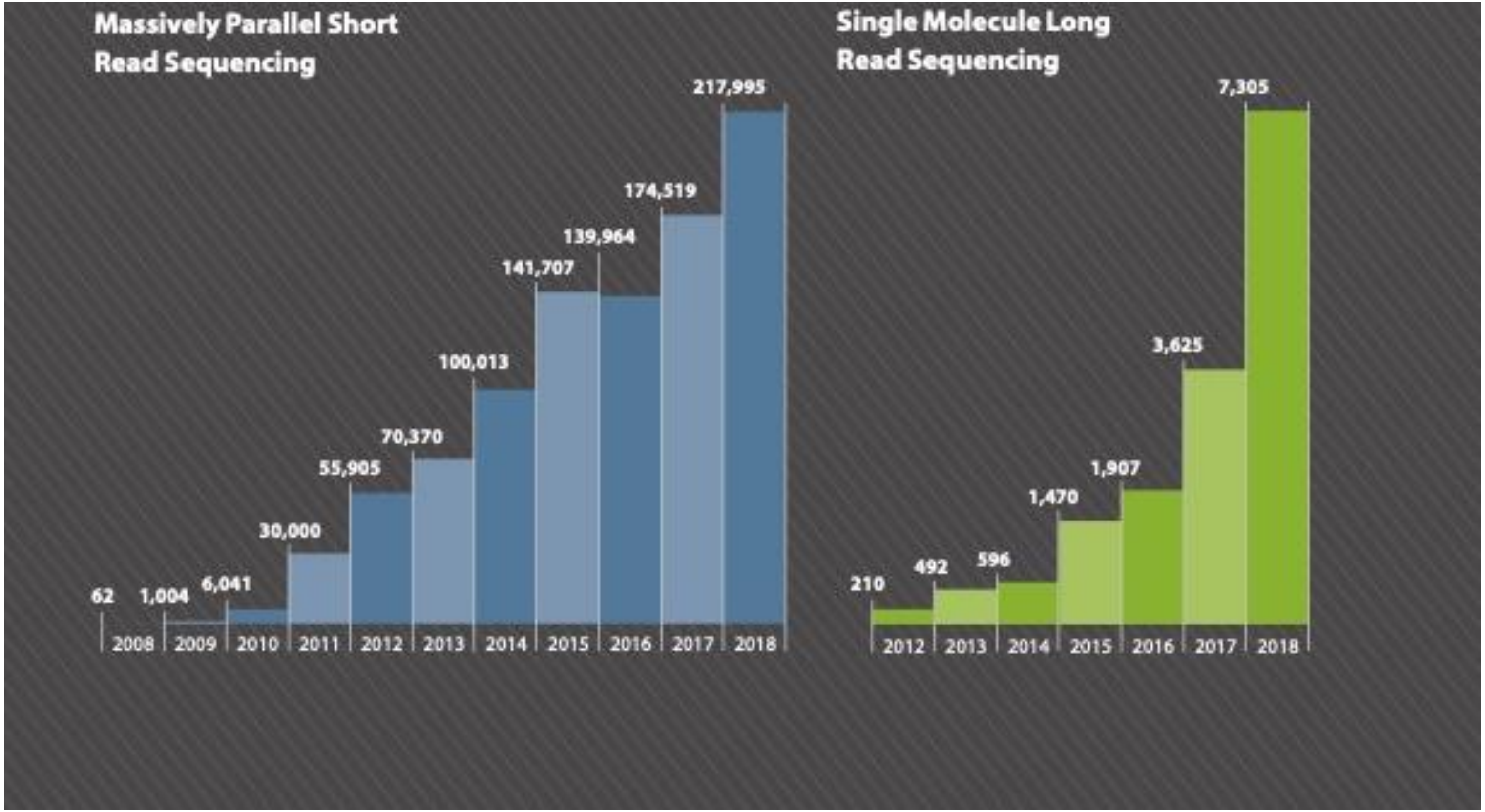
Illumina


Pacific Biosciences — Real-time sequencing
a
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time
Nature Reviews | Genetics

Pac Bio

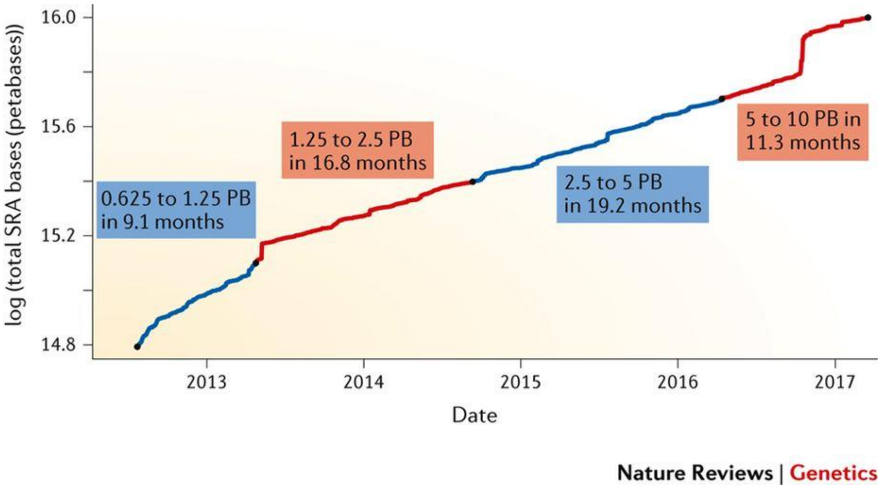
Oxford Nanopore

Significant increases in output from JGI in the last decade



Isn’t this great?

=> Get TONS of data easily, cheaply



From July 2012 to March 2017, the amount of genomic data (total bases) in the Sequence Read Archive (SRA) doubled four times.
Langmead & Nellore Nat Reviews Genetics 2018

This has created a bottleneck of genomes

The bottleneck is now annotation, Specifically quality annotation

Structural gene annotation: Rapid, improved precision in the last decade

Annotation methods have improved substantially in the last 10 years

Still remain highly focused on structural annotation of protein-coding genes

Involves defining gene features such as transcription start/stop, translation start/stop, exon/intron structure, promoters/enhancers

Methods utilize computational algorithms that look for signatures in the DNA sequence along with empirical data (transcript, protein evidence) when available

These can be automated and are fast, cheap

Precision improving

Gene functional annotation: Knowing what genes do

Functional annotation involves determining gene function; can be expanded to understanding the function of other elements in a genome

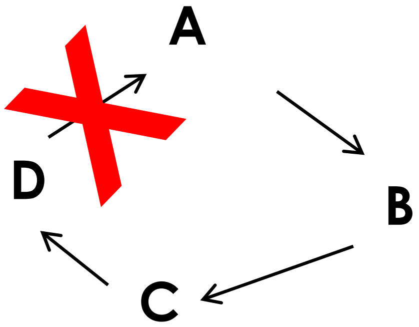
Some genomes such as yeast, fly, human have been manually curated at the structure and function level

Yet outside of key model species where functional, empirical data has been generated for a substantial set of genes, we actually know the “true” function of a fraction of the genes in any one genome

If functional annotation is key to deciphering genomes, why has there been little, if any, improvement in methodology?

Gene function annotation: Highly dependent on transitive, automated methods

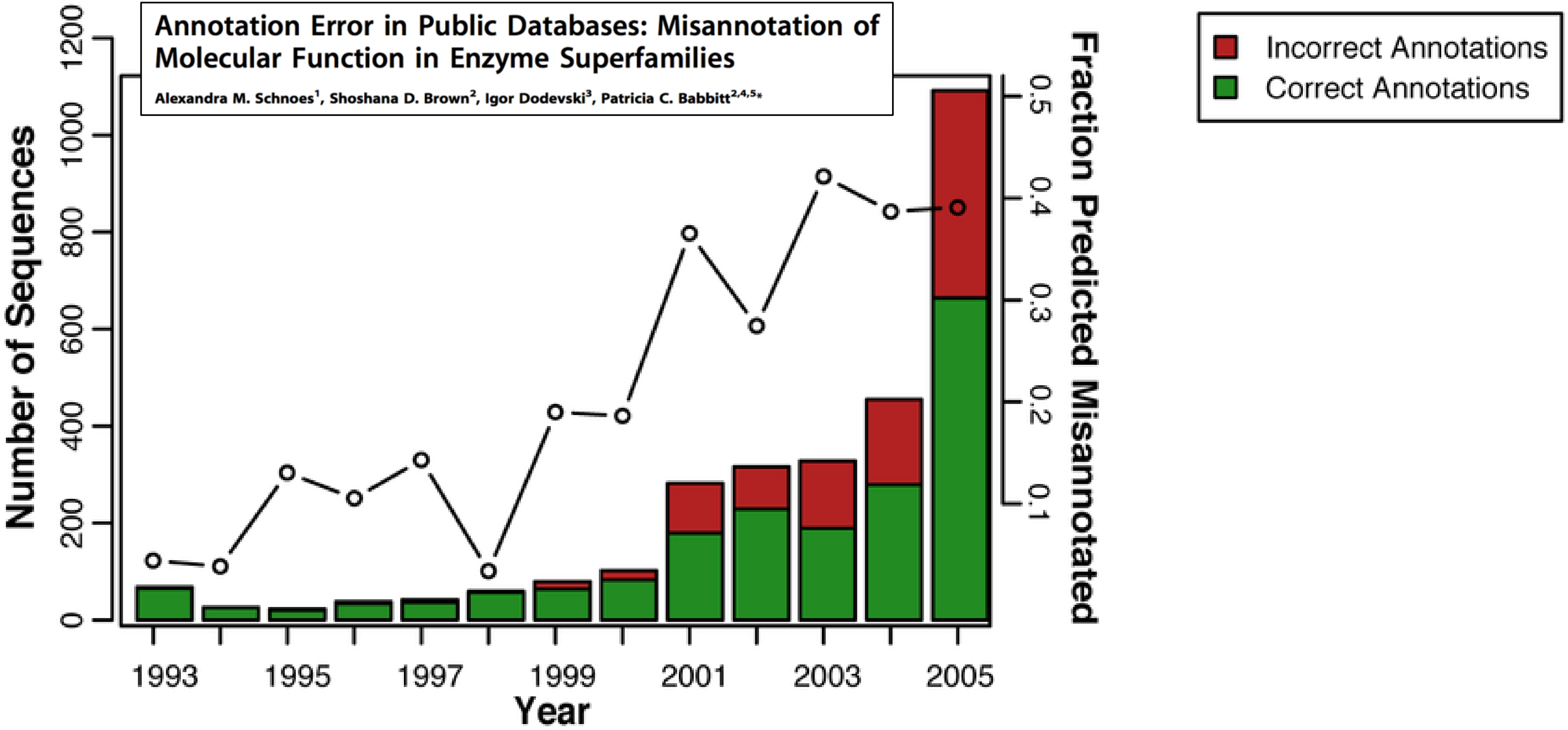
Most annotation is transitive in nature derived through sequence similarity to a sequence available in GenBank and by identifying domains/motifs via algorithms



This yields annotation such as “kinase”, “Cytochrome P450” as well as “hypothetical protein”, “expressed protein”

Not only are these uninformative due to their coarse nature, they are often wrong

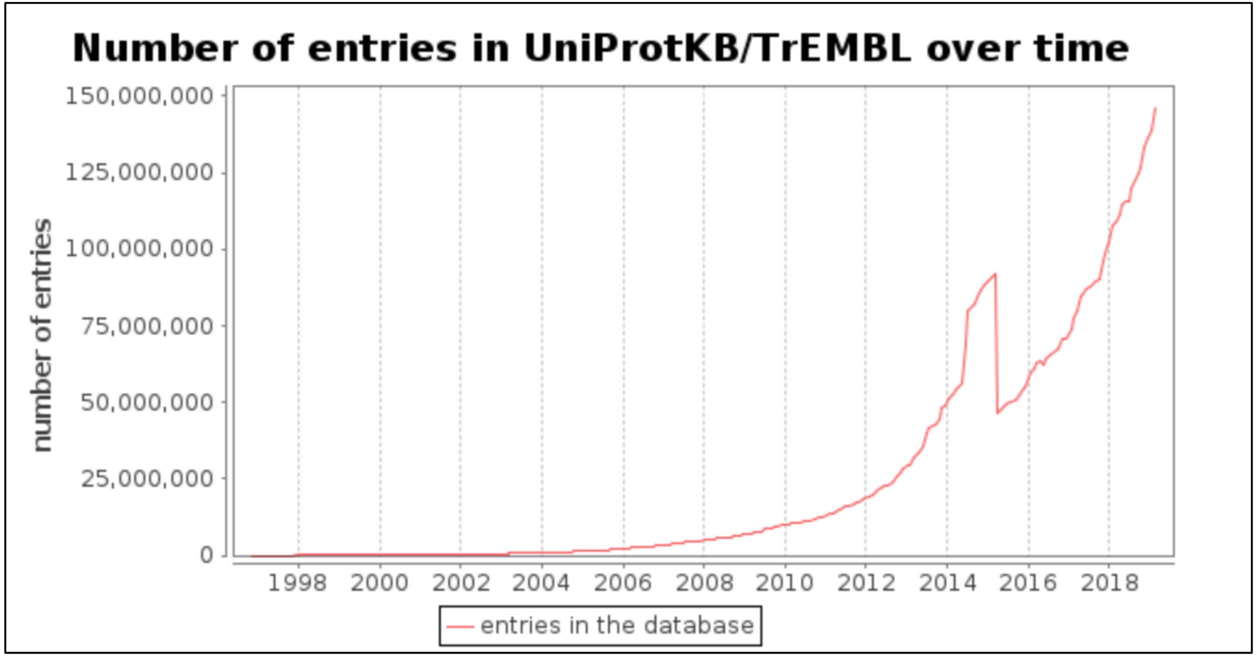
Many automated computational gene annotations are uninformative or (worse) wrong



Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies

Alexandra M. Schnoes[1], Shoshana D. Brown[2], Igor Dodevski[3], Patricia C. Babbitt[2,4,5]*

Gene functional annotation: And it is only going to get worse over time



Automatically annotated and not reviewed

#s from Feb 13, 2019

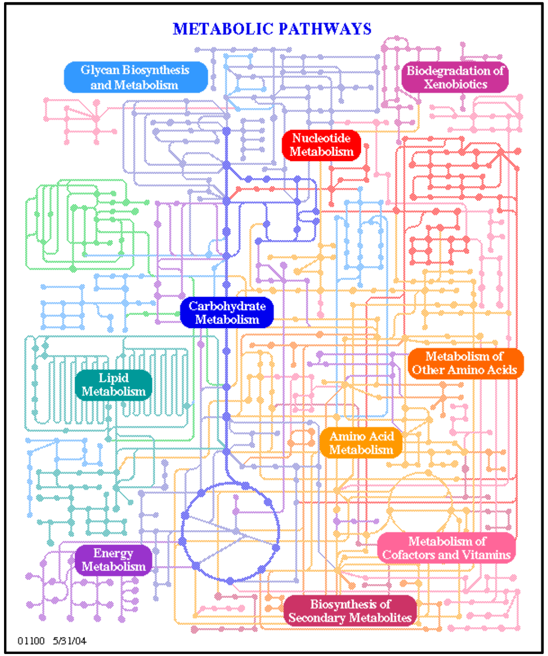
Implications of low quality inaccurate gene function annotation

Error propagation leads to:

- an inordinate waste of researcher’s time
- inability to correctly model and predict biological processes
- low pace of paradigm-changing research to capitalize on the genomics era
- a MAJOR waste of limited research dollars

Lack of gene function understanding negatively impacts all BER-funded research (and biology in general)

Incomplete models and parts-list for engineering


METABOLIC PATHWAYS
Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites
01100 5/31/04

Understanding the impact of allelic variation on plant phenotypes

The BER advisory committee identified improved gene function understanding as a future grand challenge

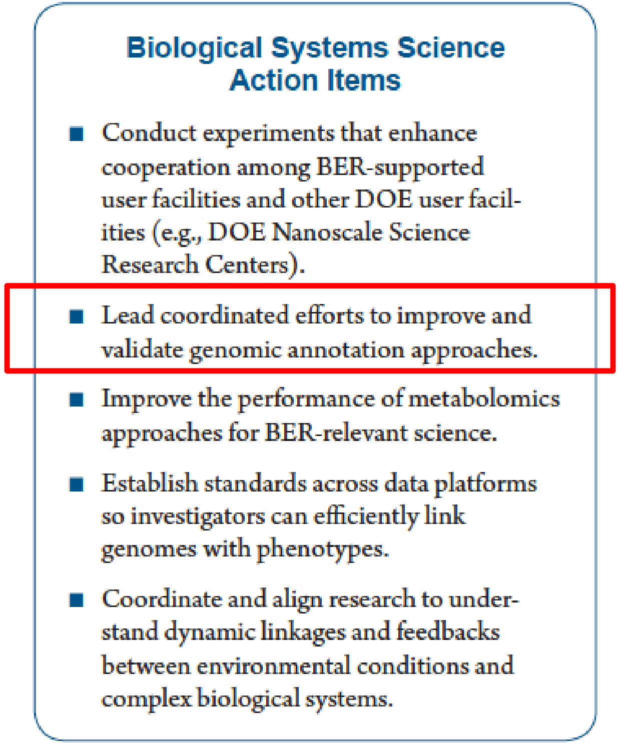
DOE/SC–0190

Grand Challenges for Biological and Environmental Research:

Progress and Future Vision

November 2017

A report from the Biological and Environmental Research Advisory Committee



Biological Systems Science Action Items

- Conduct experiments that enhance cooperation among BER-supported user facilities and other DOE user facilities (e.g., DOE Nanoscale Science Research Centers).
- Lead coordinated efforts to improve and validate genomic annotation approaches.
- Improve the performance of metabolomics approaches for BER-relevant science.
- Establish standards across data platforms so investigators can efficiently link genomes with phenotypes.
- Coordinate and align research to understand dynamic linkages and feedbacks between environmental conditions and complex biological systems.

Breaking the Bottleneck of Genomes: Understanding Gene Function across Taxa Workshop – November 1-2, 2018

- **Provide community input to DOE/BER** on current state and future directions in gene function discovery and annotation

- **Identify challenges, knowledge/technology gaps, and opportunities.**

- **Immediate outcome will be a workshop report** – to be finished in 2019

DOE/SC–0190

Grand Challenges for Biological and Environmental Research:

Progress and Future Vision

November 2017

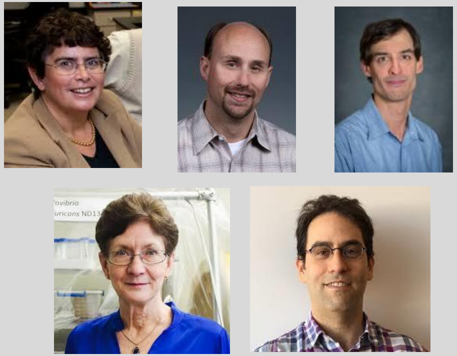
A report from the Biological and Environmental Research Advisory Committee

Workshop participants spanned a range of expertise

Very small workshop
Expertise in target areas
High acceptance rate

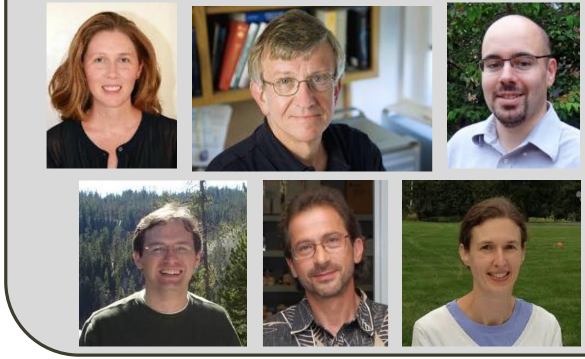
Plants

Microorganisms

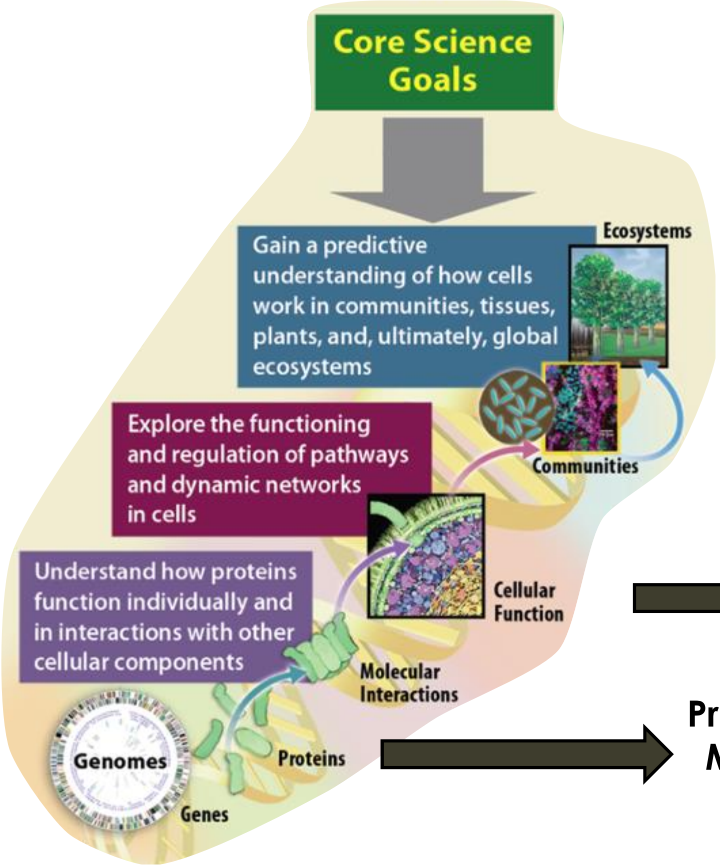


Technology

Computation



First: What is gene function?


Core Science Goals
Gain a predictive understanding of how cells work in communities, tissues, plants, and, ultimately, global ecosystems
Ecosystems
Explore the functioning and regulation of pathways and dynamic networks in cells
Communities
Understand how proteins function individually and in interactions with other cellular components
Cellular Function
Molecular Interactions
Genomes
Proteins
Genes

Genes and proteins function at multiple levels

Importance to Ecosystem process

Biological Process (ex. A phenotype)

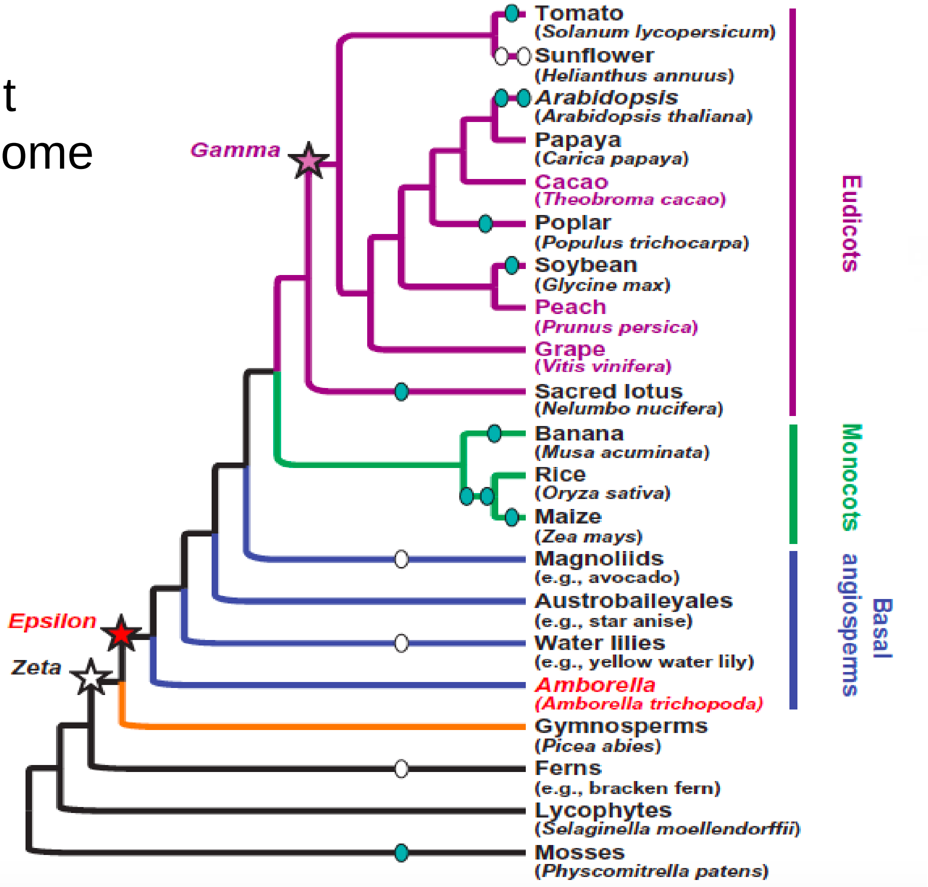
Protein structure, Modifications, Biochemistry

Biology complicates determining gene function determination

Rampant whole genome and segmental duplications present throughout the plant lineage, some which are very recent

Increases gene number!

This also occurs in microbes!

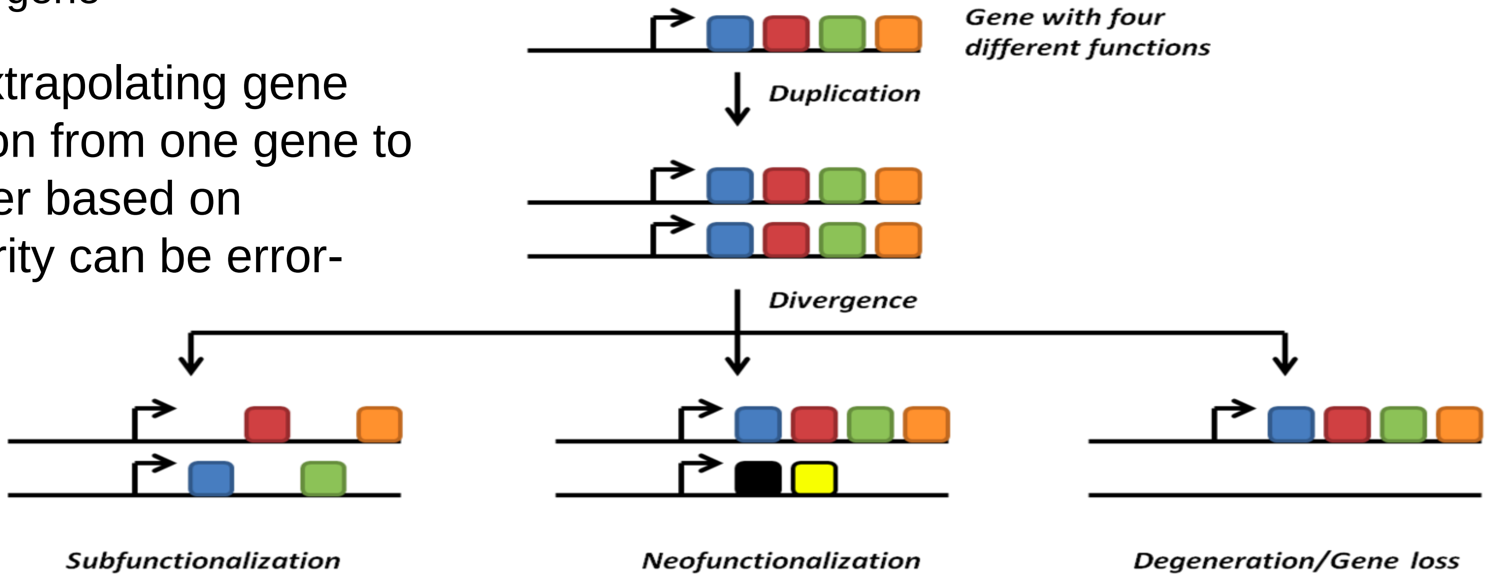


Amborella Genome Project
Science 2012

Consequences of genome/gene duplication

Genome duplication provides additional copies of every gene. This provides a “template” for nature to vary or modify the genes without (potentially) loosing the function of the original gene

So, extrapolating gene function from one gene to another based on similarity can be error-prone



Divide the function among the two copies

Take on NEW functions!

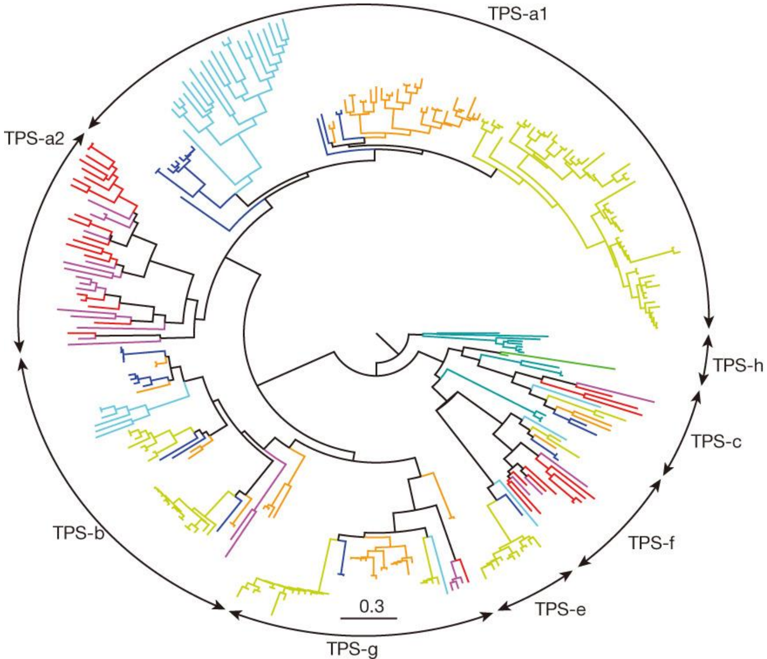
Loose one of the copies

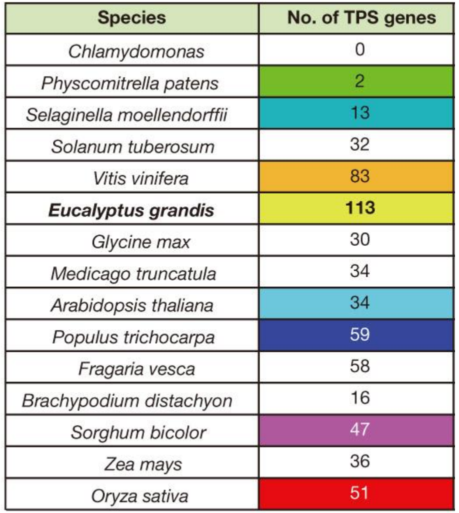
http://en.wikipedia.org/wiki/Gene_duplication

One example: Genes involved in specialized metabolism are replete with gene amplification events

Eucalyptol (terpene-derived) used in mouthwashes, cough suppressants, cosmetics, fragrance, insect repellent

Eucalyptus grandis has had a major expansion of terpene synthase (TPS) genes (n=113) that are hypothesized to function in biotic defense.





Species	No. of TPS genes
Chlamydomonas	0
Physcomitrella patens	2
Selaginella moellendorffii	13
Solanum tuberosum	32
Vitis vinifera	83
Eucalyptus grandis	**113**
Glycine max	30
Medicago truncatula	34
Arabidopsis thaliana	34
Populus trichocarpa	59
Fragaria vesca	58
Brachypodium distachyon	16
Sorghum bicolor	47
Zea mays	36
Oryza sativa	51

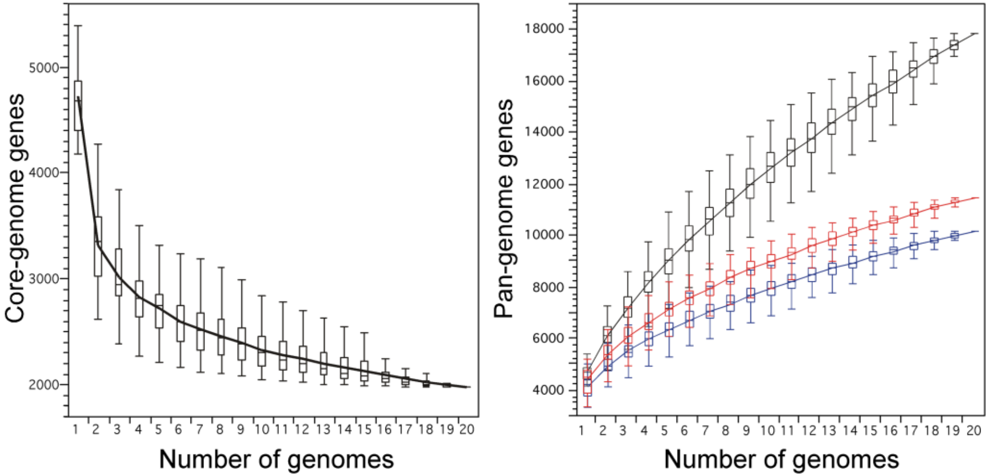
AA Myburg *et al. Nature* (2014)

Species have plastic genomes: One ‘reference’ accession will not reveal all of the genes in a species

Pioneering work in bacteria showed that there was extreme genomic diversity between bacterial isolates of the same species.

Lead to the concept of “pan-genome” composed of “core genes” and “dispensable genes” with the dispensable genes contributing to phenotypic diversity

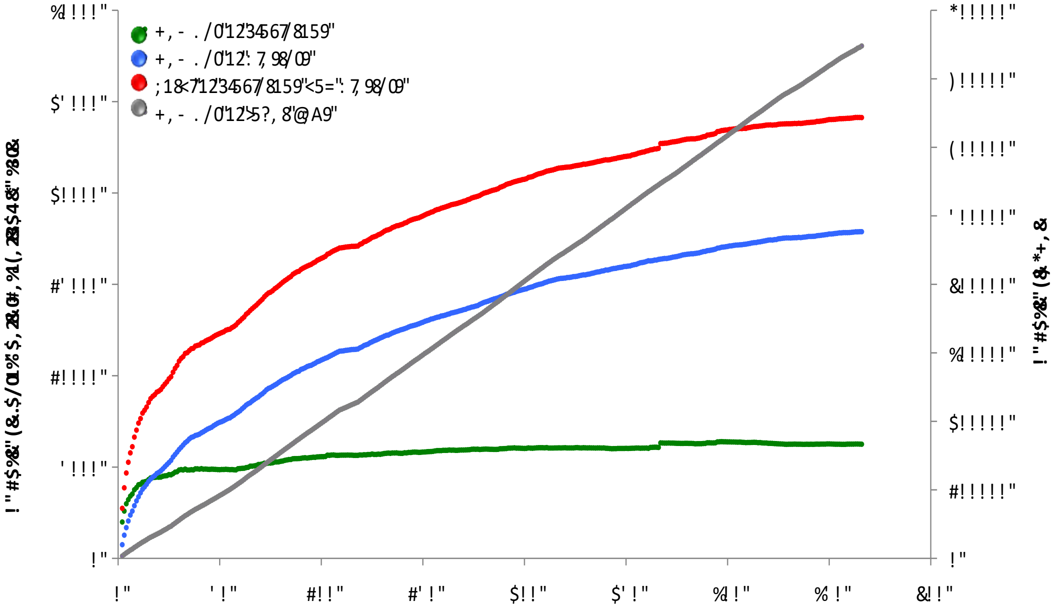
A subset of these dispensable genes function in adaptation to the environment



Touchon et al. PloS Genetics 2009

Pan-Genomes: The Plasticity of Plant Genomes

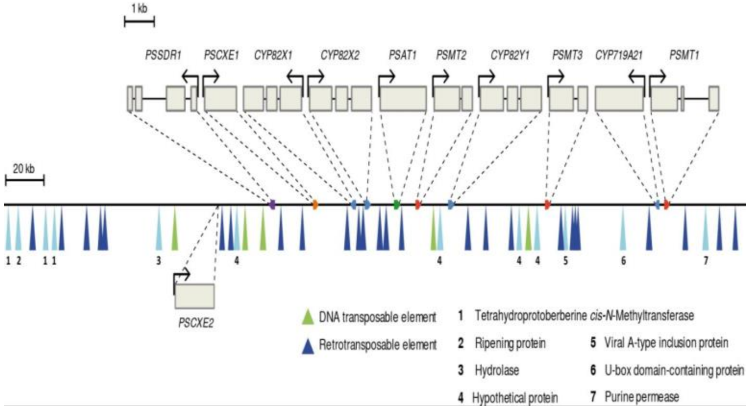
What is the extent of plant pan-genomes? Are they “closed” or “open” like bacteria?



Transcriptome sequencing of 503 maize lines reveals an average of ~2,000 novel transcripts/genes per line

Hirsch et al. Plant Cell 2014

Some dispensable genes encode entire biosynthetic pathways



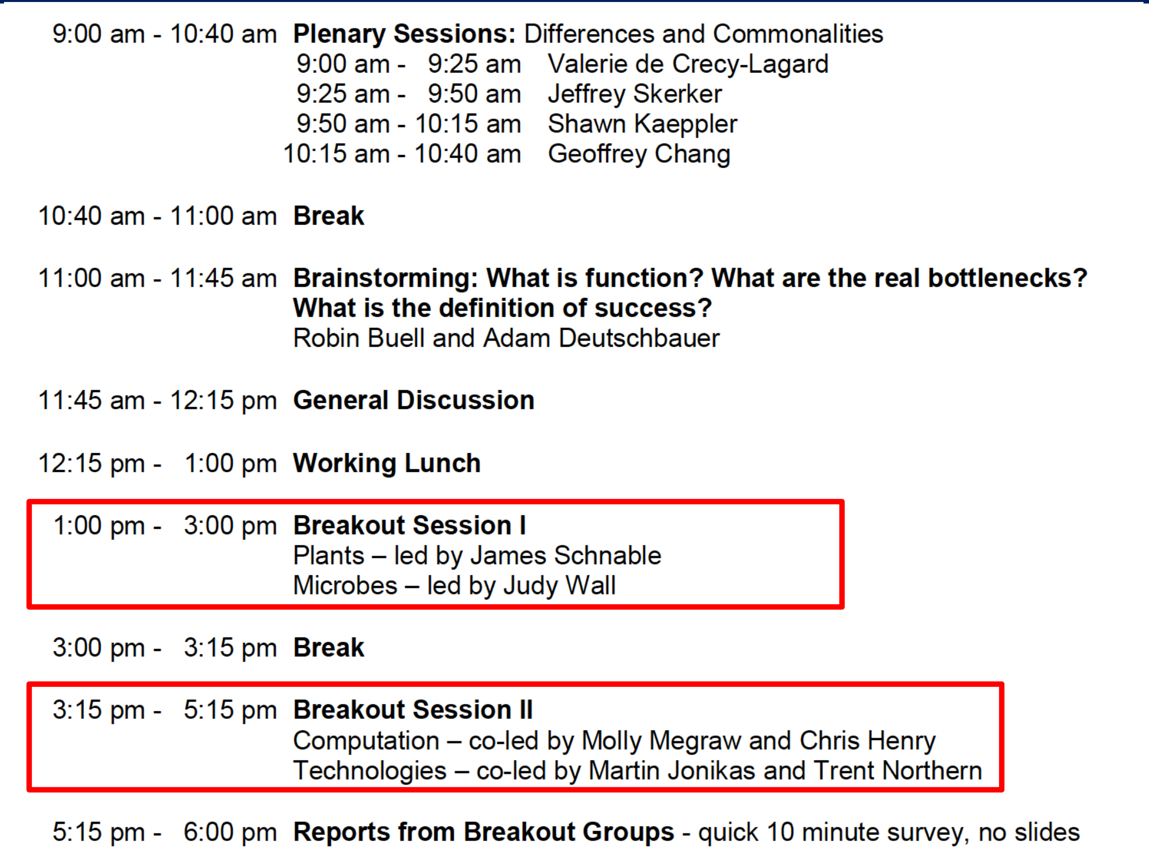
Winzer *et al.*, *Science* 2012

Plants and Microorganisms

- **Some unique challenges**
  - **Microbes** (many are currently uncultivated)
  - **Plants** (transformation, logistics of growing plants for phenotyping)

- **However, many commonalities between plants and microbes**
  - **Computational**: Propagating existing knowledge to new gene annotations
  - **Technology**: Genetic manipulation is often difficult

Breakout sessions for each of these 4 focal areas



9:00 am - 10:40 am	**Plenary Sessions:** Differences and Commonalities 9:00 am - 9:25 am Valerie de Crecy-Lagard 9:25 am - 9:50 am Jeffrey Skerker 9:50 am - 10:15 am Shawn Kaeppler 10:15 am - 10:40 am Geoffrey Chang
10:40 am - 11:00 am	**Break**
11:00 am - 11:45 am	**Brainstorming: What is function? What are the real bottlenecks? What is the definition of success?** Robin Buell and Adam Deutschbauer
11:45 am - 12:15 pm	**General Discussion**
12:15 pm - 1:00 pm	**Working Lunch**
1:00 pm - 3:00 pm	**Breakout Session I** Plants – led by James Schnable Microbes – led by Judy Wall
3:00 pm - 3:15 pm	**Break**
3:15 pm - 5:15 pm	**Breakout Session II** Computation – co-led by Molly Megraw and Chris Henry Technologies – co-led by Martin Jonikas and Trent Northern
5:15 pm - 6:00 pm	**Reports from Breakout Groups** - quick 10 minute survey, no slides

Technology: Challenge, Gaps, Opportunities

- **Challenge**: No one approach is sufficient to determine gene function across scales and taxa

- **Gap**: New, groundbreaking methods for gene function determination in high-throughput are needed.

- **Opportunities**:
  - Rapid and low-cost transfer of existing tools across diverse taxa (for example CRISPR-Cas9 editing).
  - Application of mature omics approaches systematically across species-microbes and plants
  - Nexus of DNA synthesis, cell-free biochemistry, and microfluidics

Technology: Specific opportunities

Specific Opportunities:

- The need for scalable experimental technologies
- Reduction in technology barriers
- Improving gene manipulation efficiencies and phenotyping
- Capturing molecular processes at the level of single cells
- Targeting classes of proteins
- Advancing molecular measurements of proteins
- Extension of high-throughput genetic approaches to relevant ecological contexts
- Integrating technologies to scale gene function determination

Computation: Challenge, Gaps, Opportunities

- **Challenge**: Automated approaches to infer gene function (from prior knowledge and diverse omics data).

- **Gap**: Appropriate algorithms for accurate inference; benchmarking datasets; versioning of annotations and evidence scores

- **Opportunities**:
  - Existing resources can be leveraged
  - Machine learning is well developed
  - Inference through models
  - Community engagement

Computation: Specific opportunities

Specific Opportunities:

- Computationally-driven gene function discovery
- Databases and knowledgebases of gene annotations
- A computational framework for discovery of new gene functions and accurate annotation
- Infrastructure requirements for integrating diverse omics data
- Gaps in experimental data
- Strategies and data sources for evaluating confidence in the functional annotation of a gene
- Community engagement
- Potential for high-performance computing and new algorithms to discover gene functions

Microorganisms: Challenges, Gaps, Opportunities

- **Challenges**: Amazing diversity (bacteria, fungi, archaea); many uncultivated including microbiomes
- **Gaps**:
  - Moving tools to non-model species
  - Secondary metabolite characterization
- **Opportunities**:
  - Many omics assays are rapid and cheap (pooled assays with genetically modified strains).
  - Examine microbes within more natural ecosystems (interactions between microbes, hosts, and the environment)

Microorganisms: Specific opportunities

- **Opportunities**:
  - Target microorganisms for intensive study
  - Move experimental tools from model to non-model microorganisms
  - Move experimental tools from model to non-model microorganisms
  - Determine gene function in natural contexts (microbiomes, biofilms)
  - Genetic redundancy and functionally distinguishing paralogs

Plants: Challenges, Gaps, Opportunities

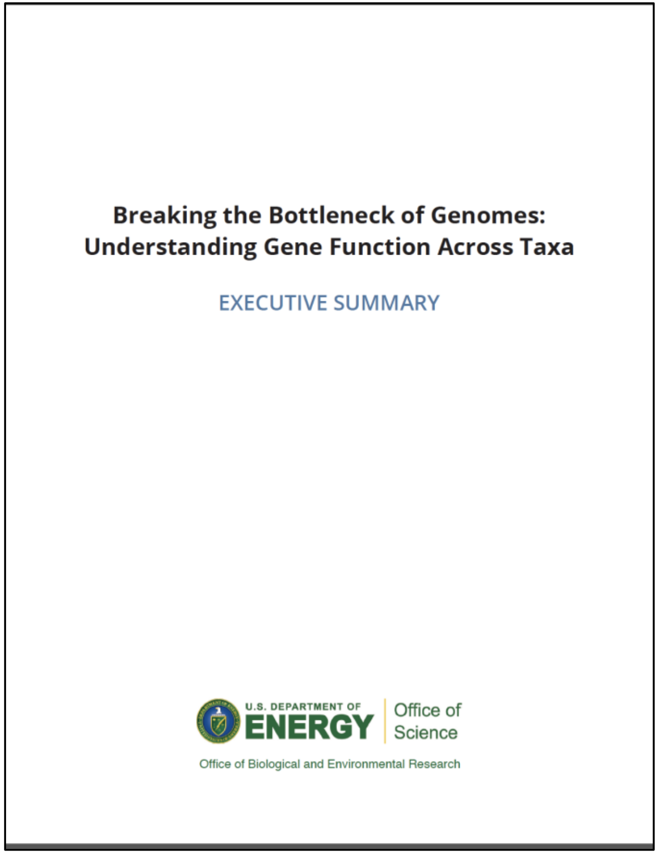
- **Challenge**: Genome size and complexity; environmental heterogeneity; ploidy
- **Gaps**: Barriers to genetic manipulation, large-scale phenotyping
- **Opportunities**:
  - Some BER-relevant species have immense datasets and resources in place that can be leveraged
  - Comparative methods permit leveraging knowledge across related taxa

Plants: Specific opportunities

- **Opportunities**
  - Focal species to accelerate gene function discoveries
  - Well annotated genomes and associated datasets
  - Prioritizing gene sets for functional experimentation
  - Perturbation of genes via gene editing
  - Modeling of relevant plant processes
  - GxE: Role of environment

- A transformative platform: A minimal plant genome as a chassis for gene function discovery

Summarizing the outcomes of the workshop



Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa

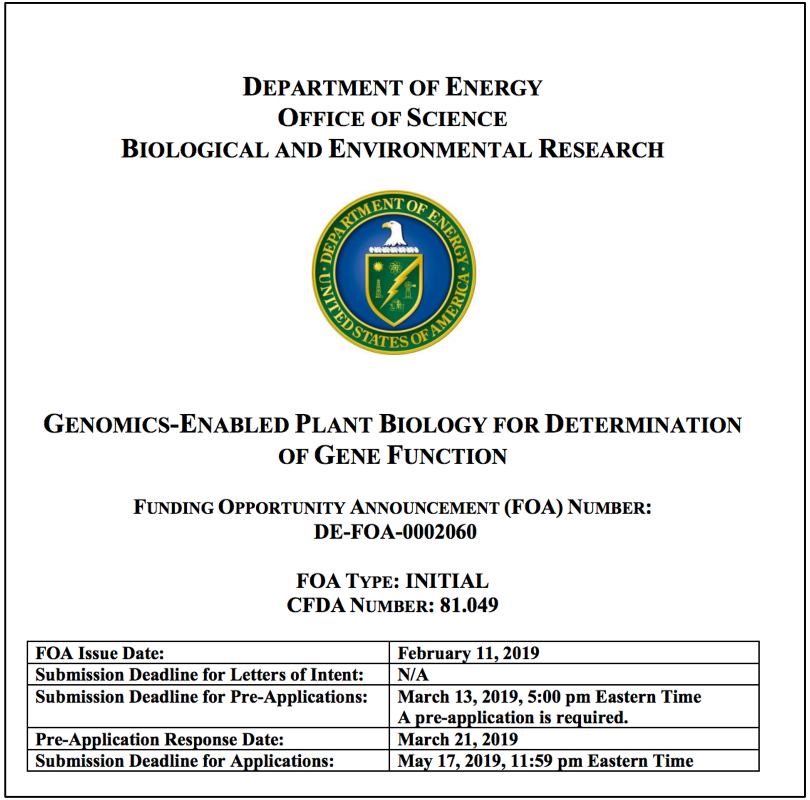
EXECUTIVE SUMMARY

U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Biological and Environmental Research

genomicscience.energy.gov/genefunction/

Developing opportunities



Department of Energy
Office of Science
Biological and Environmental Research

Genomics-Enabled Plant Biology for Determination of Gene Function

Funding Opportunity Announcement (FOA) Number:
DE-FOA-0002060

FOA Type: INITIAL
CFDA Number: 81.049

FOA Issue Date:	February 11, 2019
Submission Deadline for Letters of Intent:	N/A
Submission Deadline for Pre-Applications:	March 13, 2019, 5:00 pm Eastern Time A pre-application is required.
Pre-Application Response Date:	March 21, 2019
Submission Deadline for Applications:	May 17, 2019, 11:59 pm Eastern Time