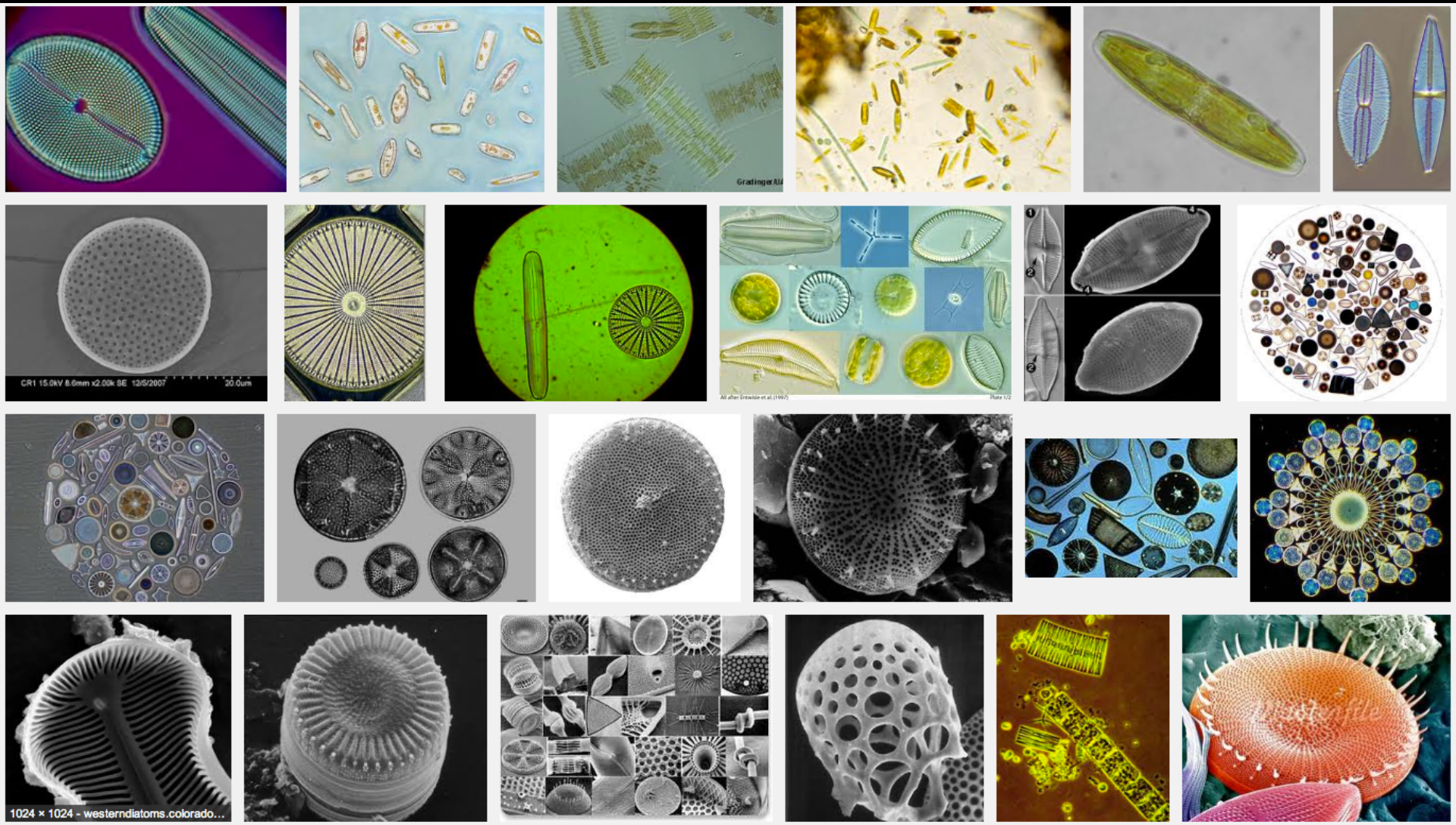
Exascale, Data and Biology

Rick Stevens



Argonne and University of Chicago

362363
1-Jl-45
BECKENHAM, KENT
RAILWAY STATION
OXFINGTON S.C.M.

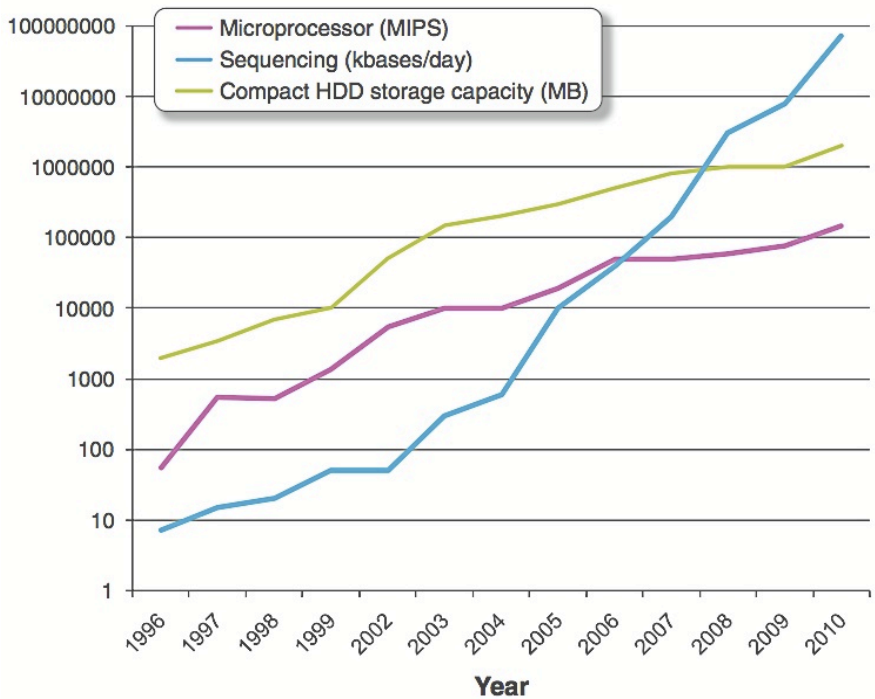
seeing the flowers & experimentising on them. But if you intend to experimentise on them, of course you will not send me the seeds, as I sh. be very unwilling to interfere in any way with your work. I sh. also rather like to look at the flowers of Cassia chamaecrista.

Many years ago I tried some experiments in a remotely

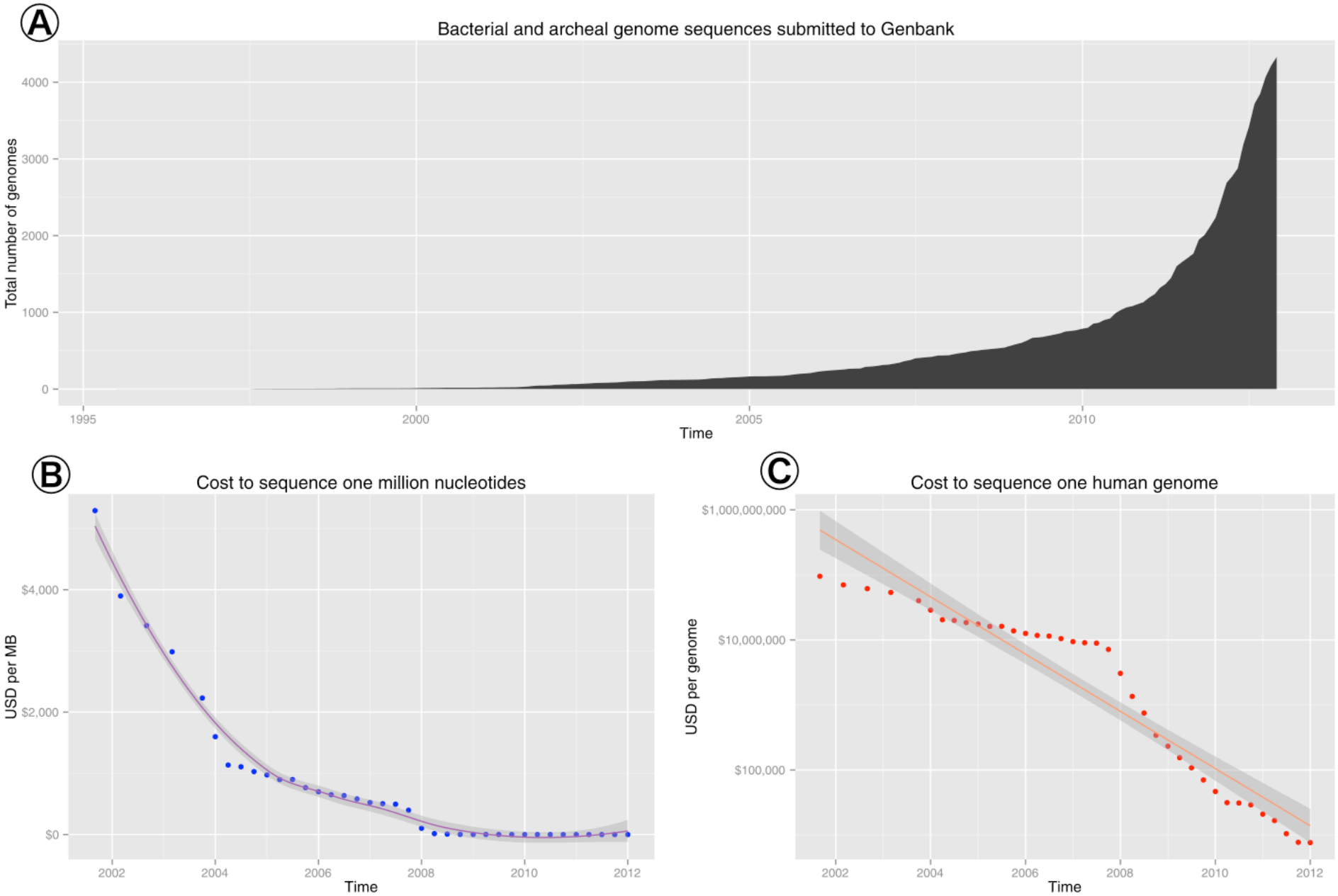
PPS 11/7/45

analogous case & this year am trying others. I described what I was doing to Dr. Fritz Müller (Blumenau, Sta. Catharina, Brazil) & he has told me that he believes that in certain plants producing 2 sets of anthers of a different colour, the bees ~~sat~~ collect the pollen from one of the sets alone. He w. therefore be much interested by your paper, if you have a spare copy that you could send him. I think, but my memory

Genomic sequencing output **x2** every **9 month**
>300 public centers



GENOMICS

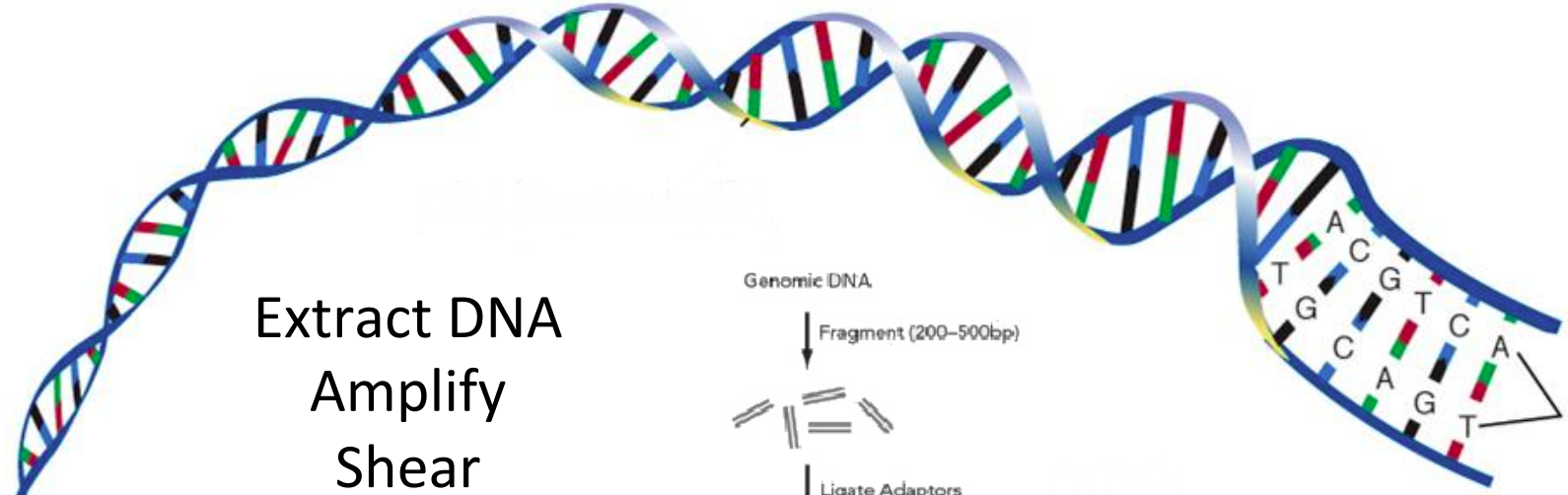

A
Bacterial and archeal genome sequences submitted to Genbank
Total number of genomes
4000
3000
2000
1000
0
1995
2000
2005
2010
Time
B
Cost to sequence one million nucleotides
USD per MB
$4,000
$2,000
$0
2002
2004
2006
2008
2010
2012
Time
C
Cost to sequence one human genome
USD per genome
$1,000,000,000
$10,000,000
$100,000
2002
2004
2006
2008
2010
2012
Time

Million Plant & Animal Genomes Project

Million Human Genomes Project

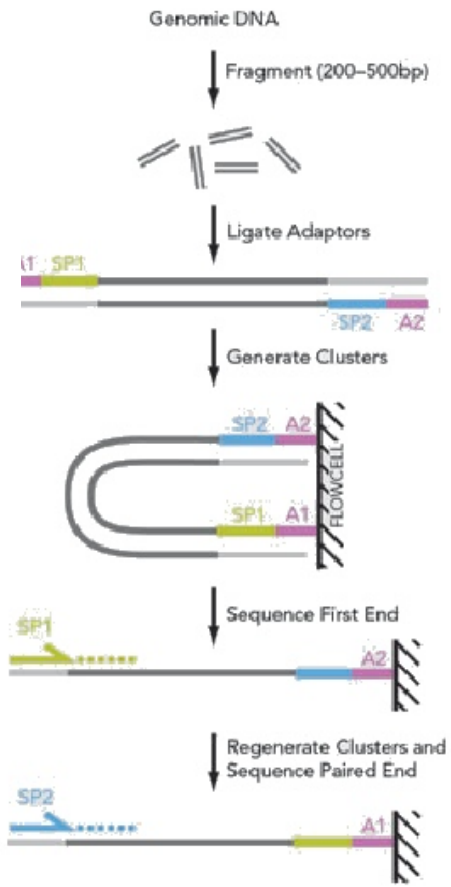
Million Microecosystem Genomes Project

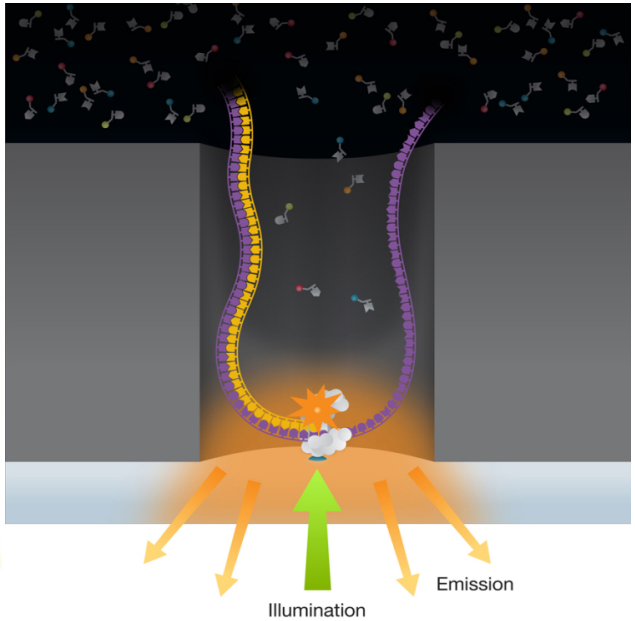
Next Generation Sequencing



Extract DNA
Amplify
Shear
Sequence

Base Pairs





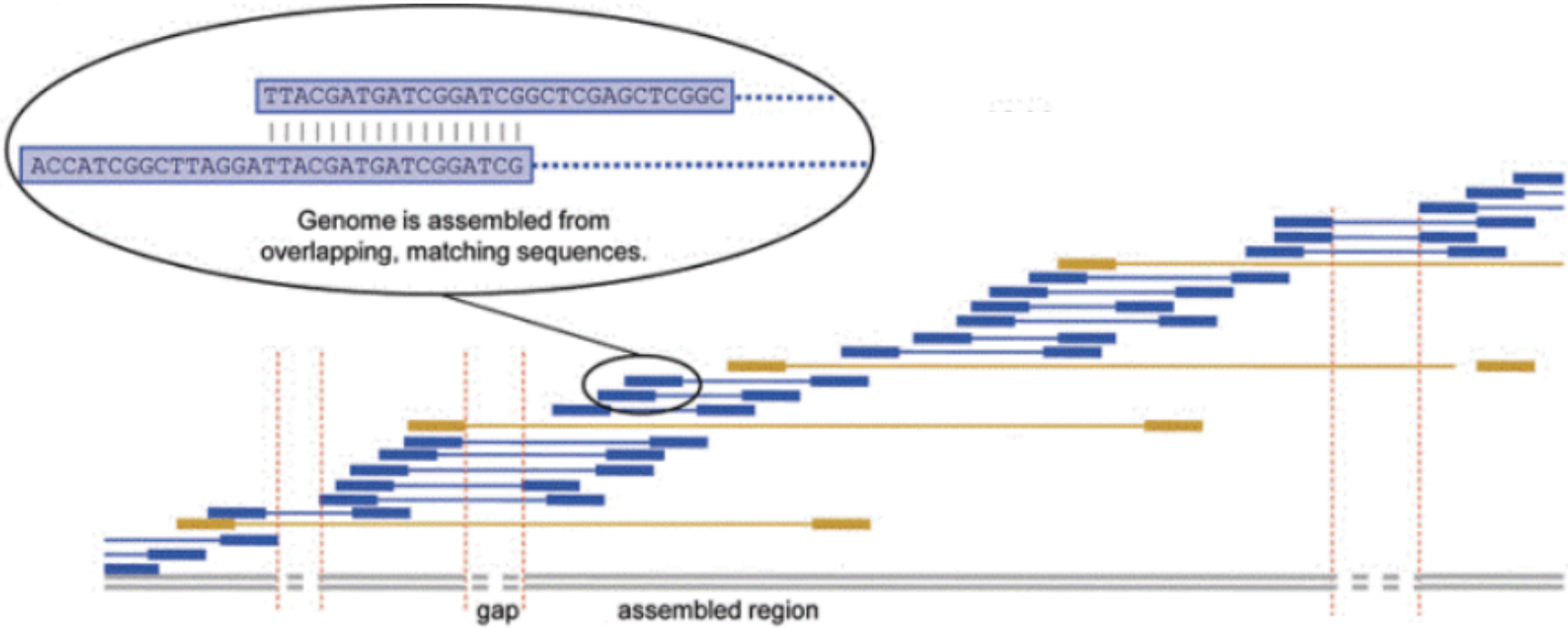
Analogy

- How we assemble
  - Pattern
  - Grammar

- DNA sequence
  - AGATTCATAG - ???
  - We don't fully understand the language yet

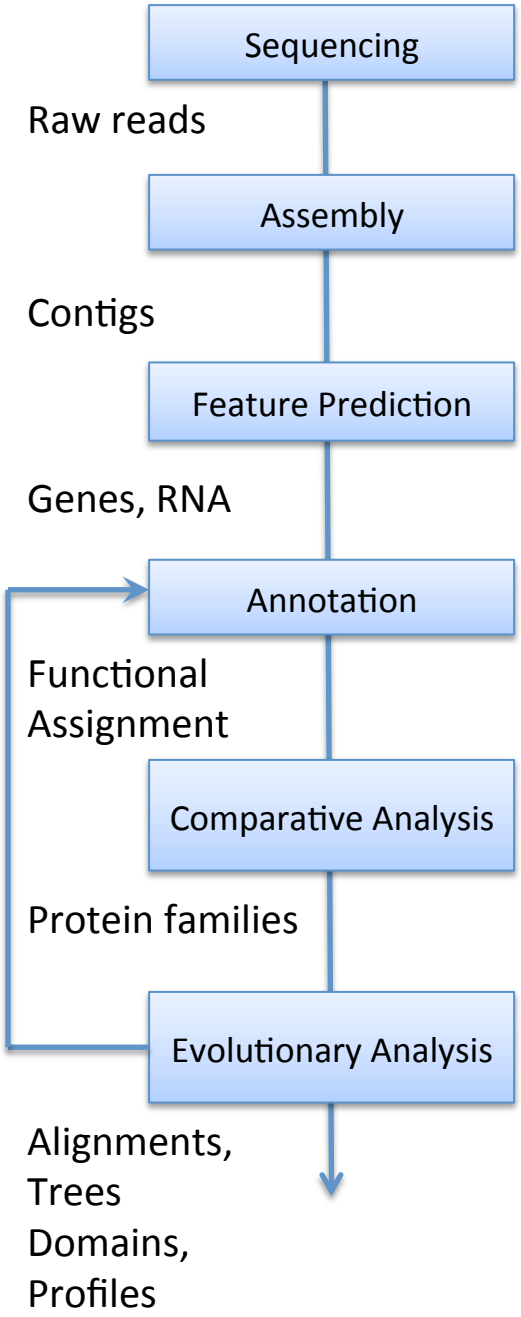
The Sequence Assembly problem

- Reconstructing contiguous DNA regions (contigs) from a set of short sequences (reads, kmers)

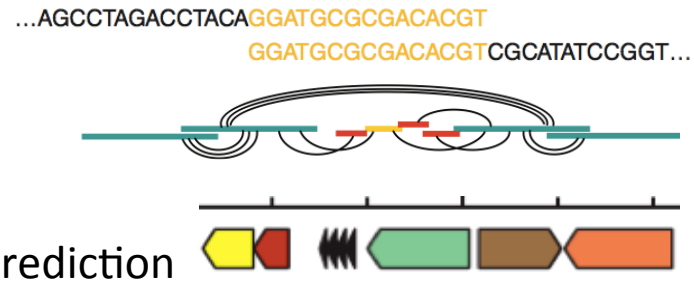


(Gilchrist, 2010)

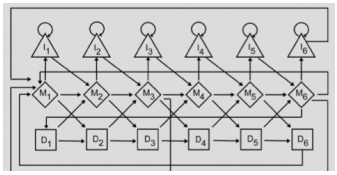
Microbial Genome Analysis Pipeline



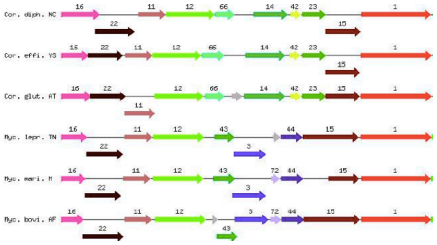
Read cleaning
De novo assembly
Assembly validation



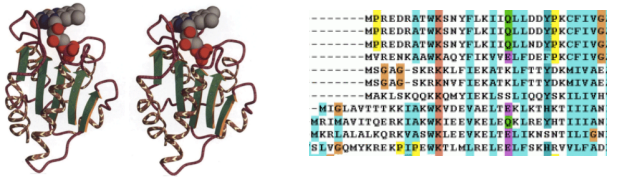
ORF prediction, rRNA & tRNA prediction
Gene model building



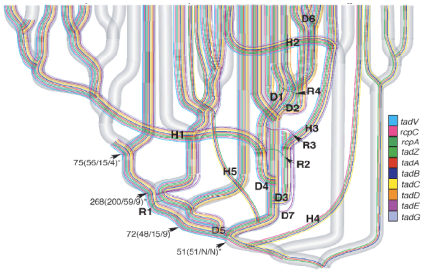
Universal genes: reverse profile search
Homology search: by signature or alignment
Chromosomal clustering



Protein families: existing and new
Structural analysis

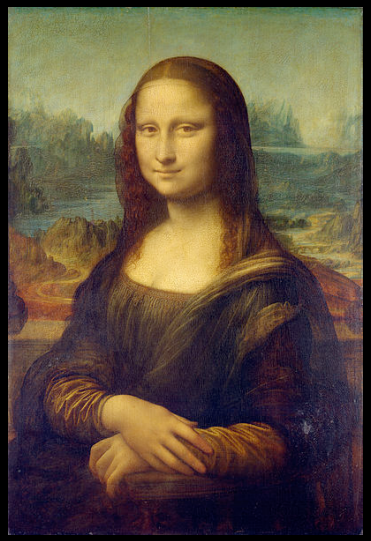


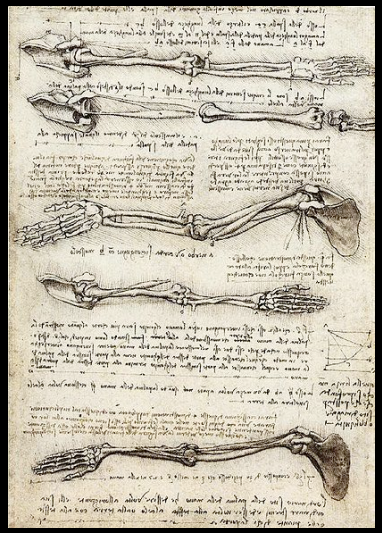
Multiple Sequence Alignment
Phylogenetic trees
Profiles for protein families/domains
Orthorlogs and paralogs
Horizontal Gene Transfer



“We know more about the movement of celestial bodies than about the soil underfoot.”

- Leonardo DaVinci, circa 1500





Microbial Universe

1 kilogram of soil

10^{13} microbial cells (10^{30} for the entire Earth)

Most (>99%) have never been cultivated and their properties are unknown

Stars in the Galaxy

10^{11}

Stars in the Universe

10^{24}

There are a million times as many microbes on earth as stars in the known universe

Microbes are responsible for fundamental life processes on a global scale, they under pin the Earth's ecosystems and control cycling of C, N, P and other nutrients.

How many kinds?

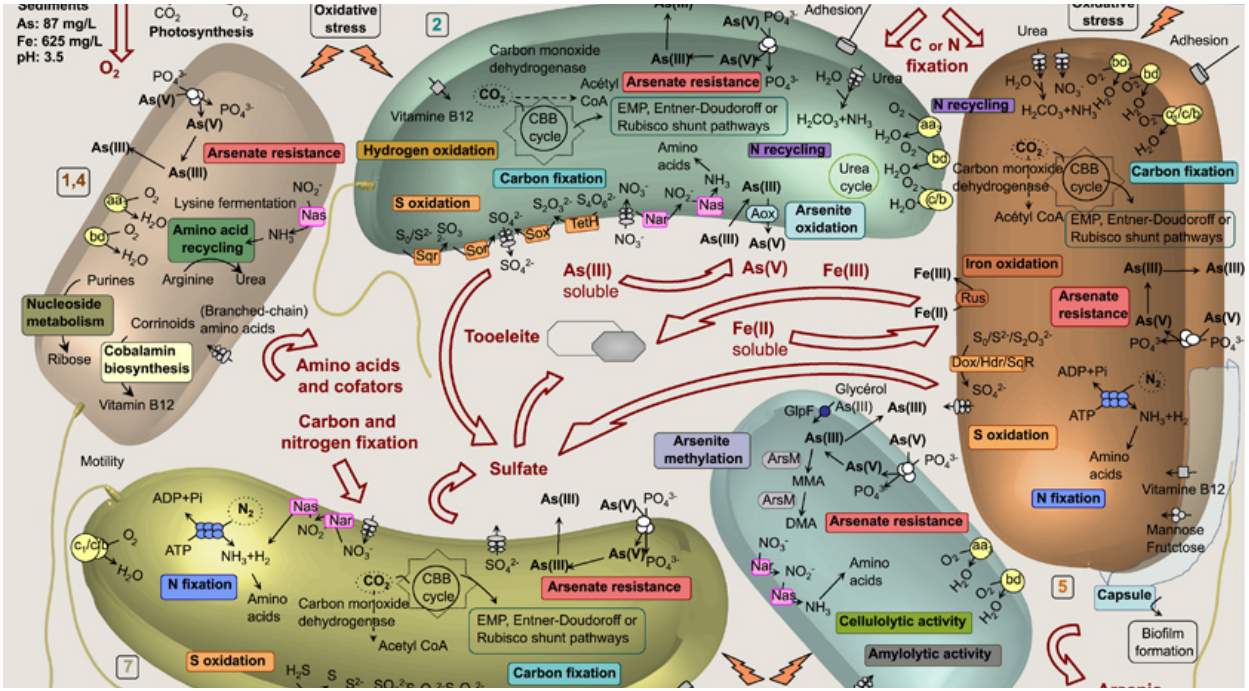
Where are they?

Where do they come from?

What are they doing?

Metagenomics

- DNA sequence information is extracted from entire samples in situ
- 10^{30} microbial cells on earth, 99% unculturable
- Ecology



(Bertin et al, 2011)

Sequencing the Environment

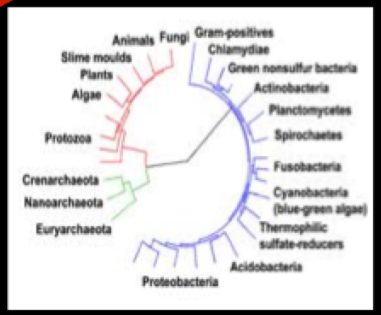
Metagenomic data collection

Collecting samples

Sequencing

Sequence fragments

Associating fragments to taxonomical groups

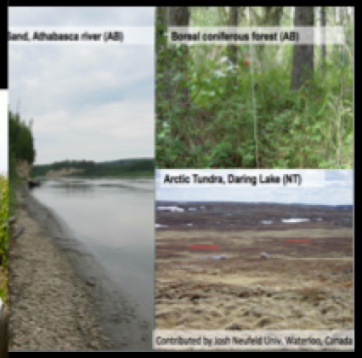


Assembly of most abundant microbes into complete genomes

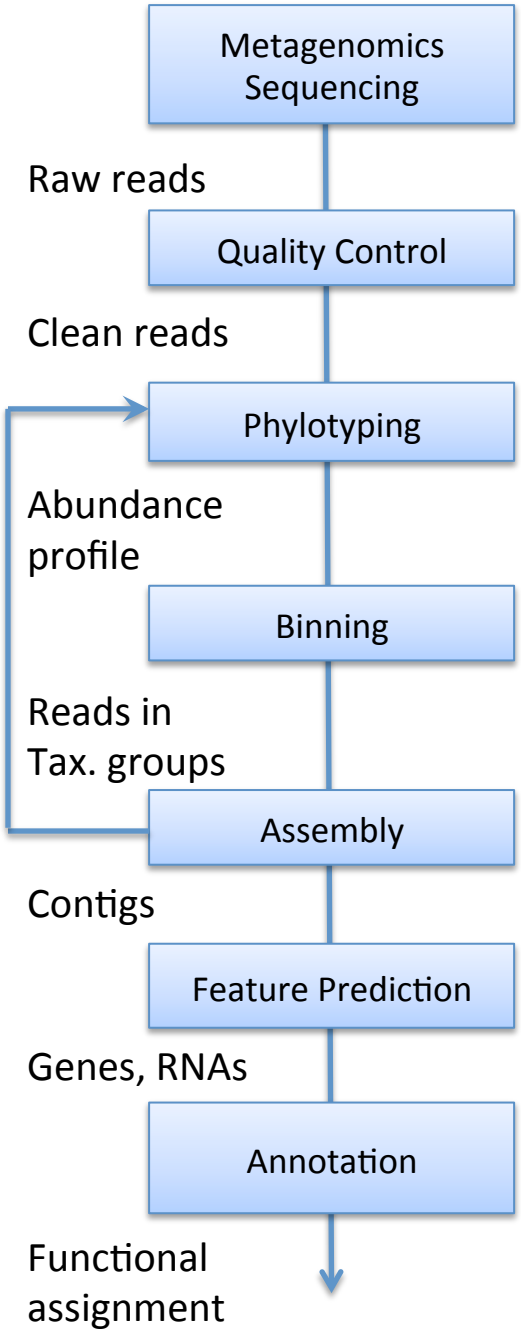
Extreme environments:
Acidic hot springs, Yellowstone–
contributed by Greg Caparoso

Merlot Microbiome:
High school volunteers
Long Island

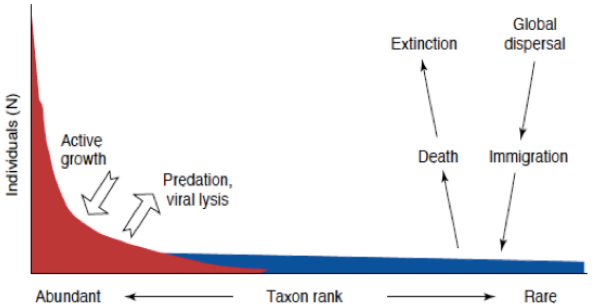
Beck Wehrle, The Iguana Microbiome



Metagenomic Pipeline

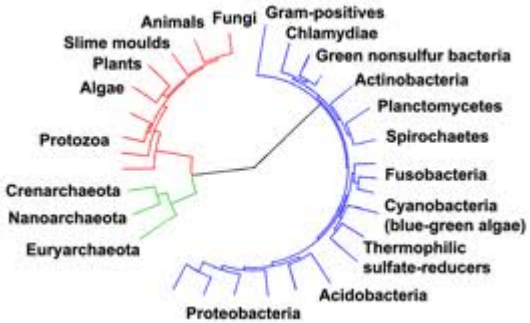


Filtering, trimming, dereplication
Adapter removal
Model organism screening



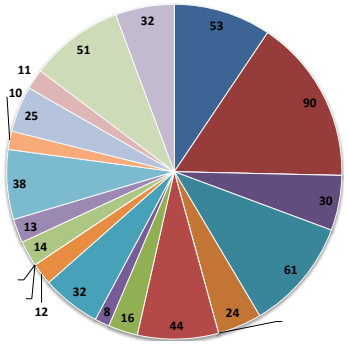
K-mer analysis
Species diversity estimation

Probabilistic inference
Homology search
Clustering



De novo metagenome assembly

ORF prediction, rRNA & tRNA prediction
Gene model building



Universal genes: reverse profile search
Homology search: by signature or alignment
Chromosomal clustering

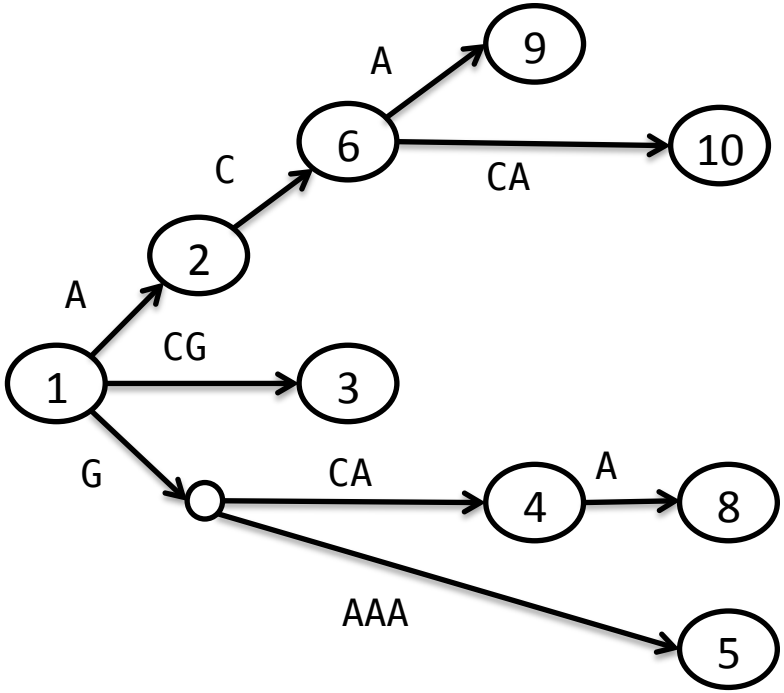
Assemblying Metagenomes

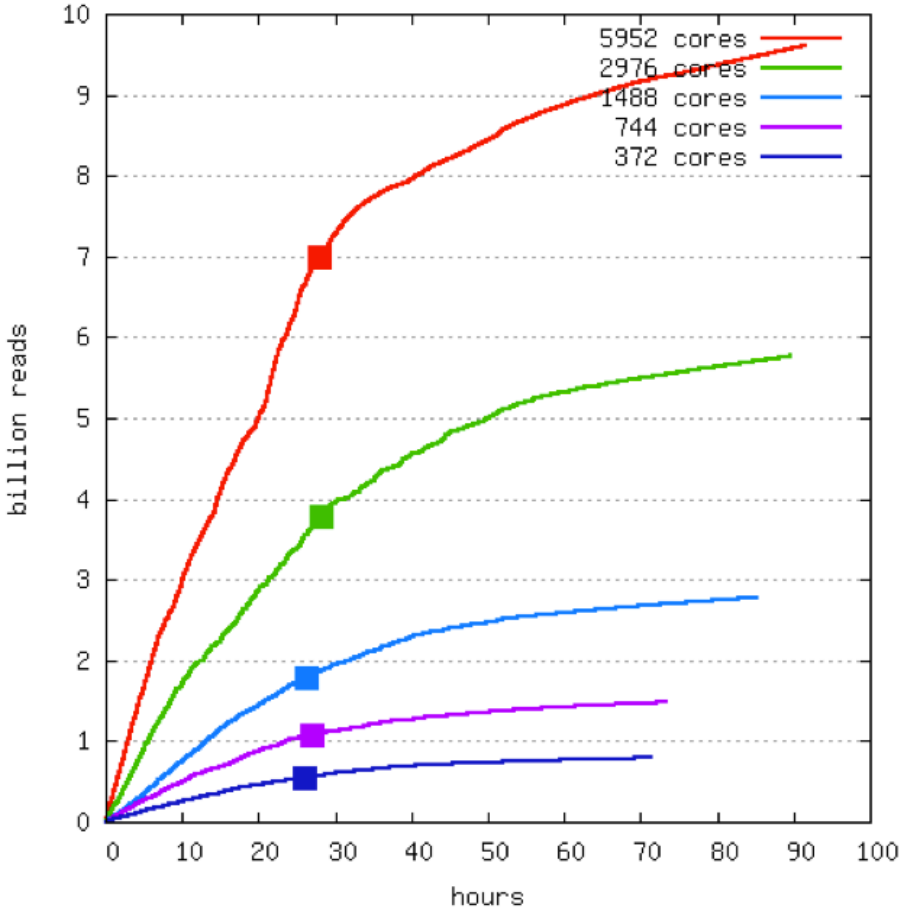
Kiki: a Parallel *k*-mer Indexing Engine for Metagenome Assembly

- Parallel library implemented in C with MPI
- Majority of nodes used for indexing sequences
- Efficient hash table allows for fast queries
- Advantages of Kiki assembly
  - Greedy algorithm provides most important results first
  - avoids $O(n^2)$ computation
  - uses distributed memory
  - supports job persistence

Kiki

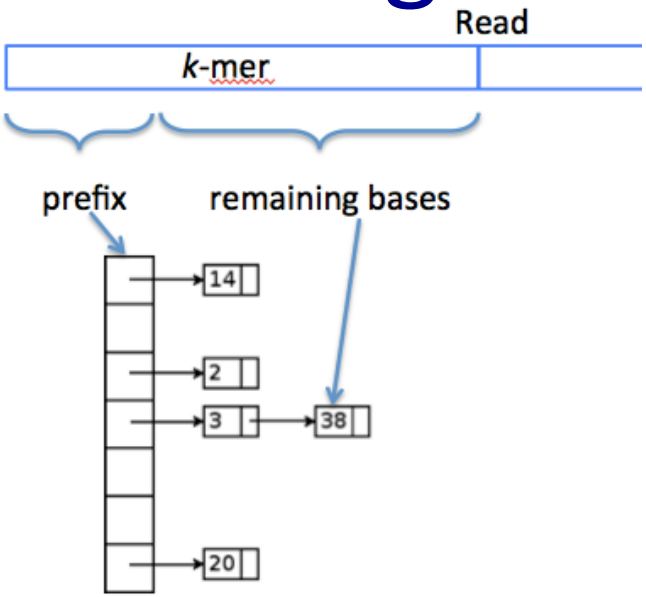
```
[1]   GTAATTGCCATCGGTTGTACGGGTGG    (SEED seq)
[2]    TAATTGCCATCGGTTGTACGGGTGGA
[3]     AATTGCCATCGGTTGTACGGGTGGCG
[6]     AATTGCCATCGGTTGTACGGGTGGAC
[4]      ATTGCCATCGGTTGTACGGGTGGGCA
[9]      ATTGCCATCGGTTGTACGGGTGGACA

[5]        TTGCCATCGGTTGTACGGGTGGGAAA
[8]        TTGCCATCGGTTGTACGGGTGGGCAA
[10]       TTGCCATCGGTTGTACGGGTGGACCA
```





Hashing

```
[1]  GTAATTGCCATCGGTTGTACGGGTGG          (Seed)

      TAATTGCCATCGGTTGTACGGGTGGA
       AATTGCCATCGGTTGTACGGATGGAC
       AATTGCCATCGGTTGTACGGGTGGCG

            ...  ...

                  GTACGGGTGGACTGCAGCTAGCGTGA

Old consensus
------------------------------------------------------------

      GTAATTGCCATCGGTTGTACGGGTGGACTGCAGCTAGCGTGA

                  TACGGGTGGA................
                   ACGGGTGGAC................
                    CGGGTGGACT................

new queries                                          (moving prefix)

                          ... ...

                               GCTAGCGTGA................
```

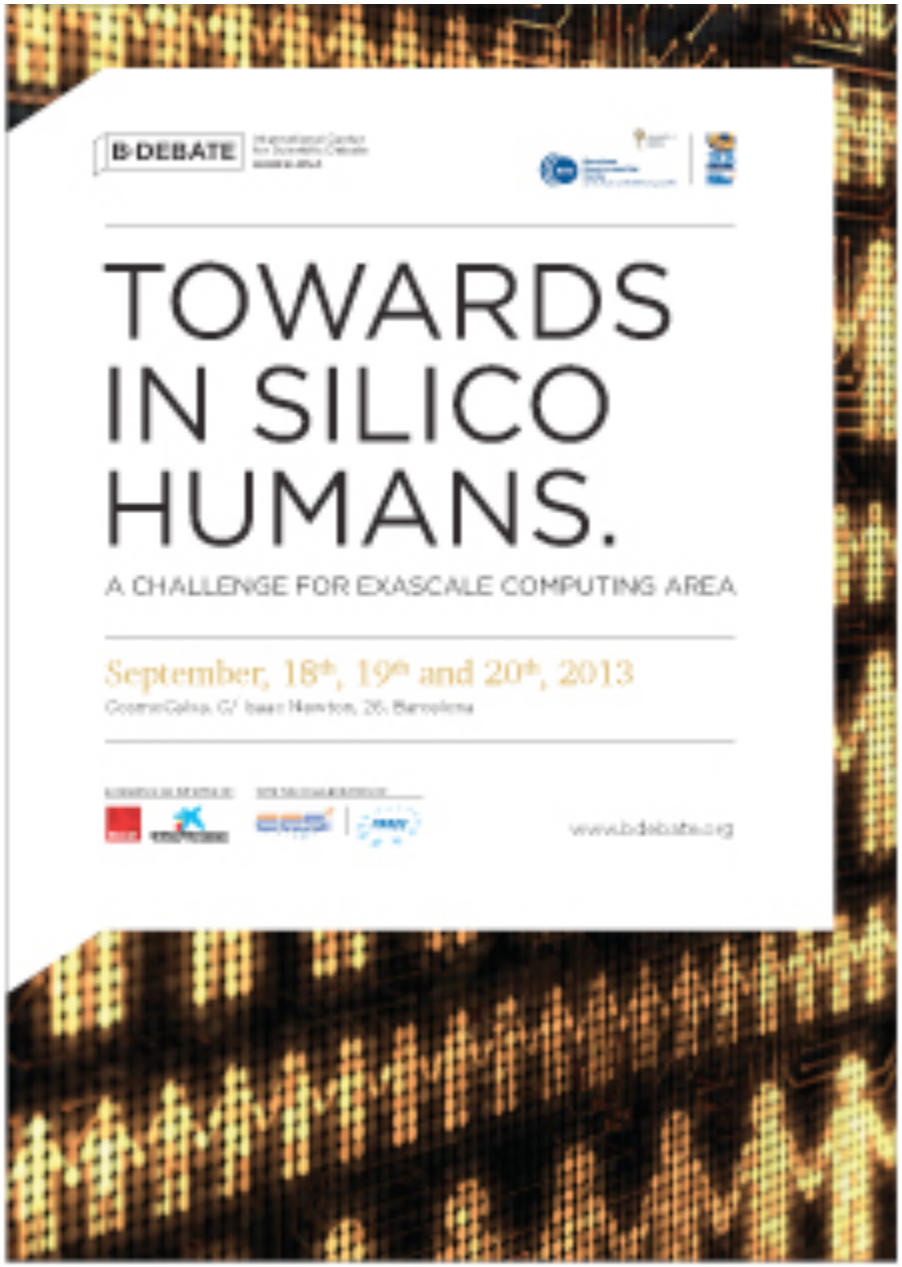


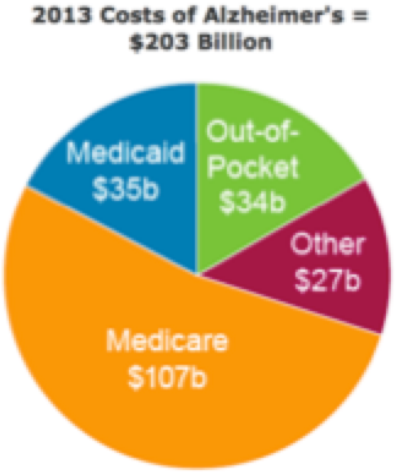

```
getPrefixSeqs( query       = TACGGGTGGACTGCAGCTAGCGTGA,
               minOverlap  = 10,
               maxMismatch = 0.05 )
```

1330 molec. bio databases

Nucleic Acids Research (96 in Jan 2001)

HUMAN GENETIC VARIATION





It is the only cause of death among the top 10 in America without a way to prevent it, cure it or even slow its progression.

1 in 3 seniors dies with Alzheimer's or another dementia.

On the allelic spectrum of human disease

David E. Reich and Eric S. Lander

Human disease genes show enormous variation in their allelic spectra; that is, in the number and population frequency of the disease-predisposing alleles at the loci. For some genes, there are a few predominant disease alleles. For others, there is a wide range of disease alleles, each relatively rare.

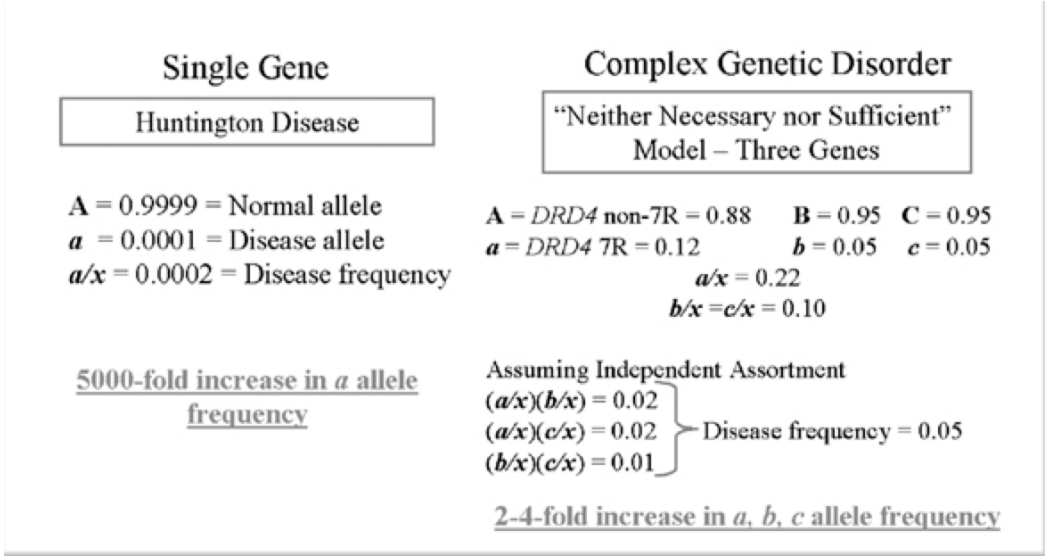
Reich and Lander's 2001 paper on the possibility of Genome Wide Association Studies or GWAS

By 2007 there were many studies underway

By 2013 over 1,700 published GWAS studies

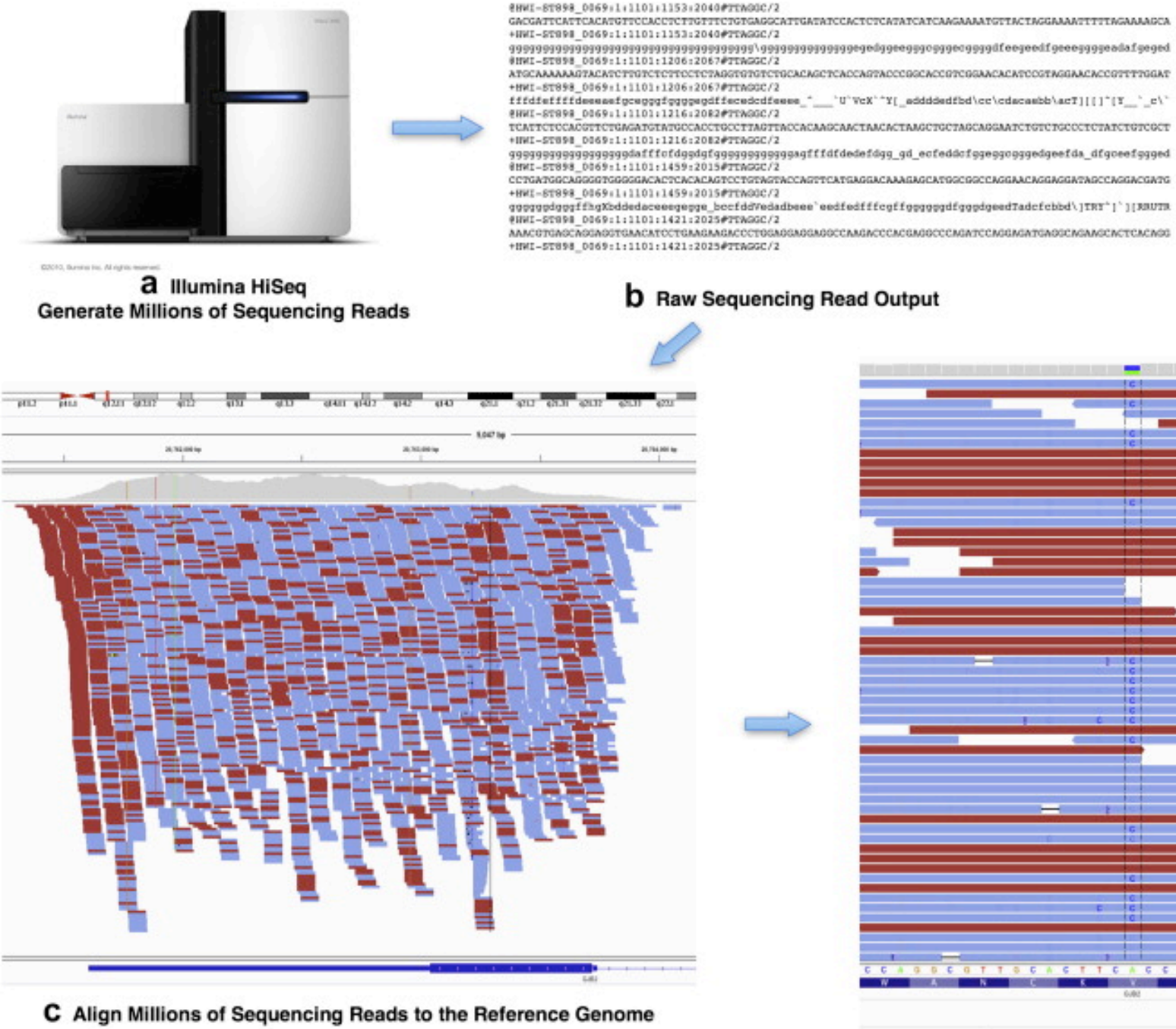
Complex Genetic Diseases

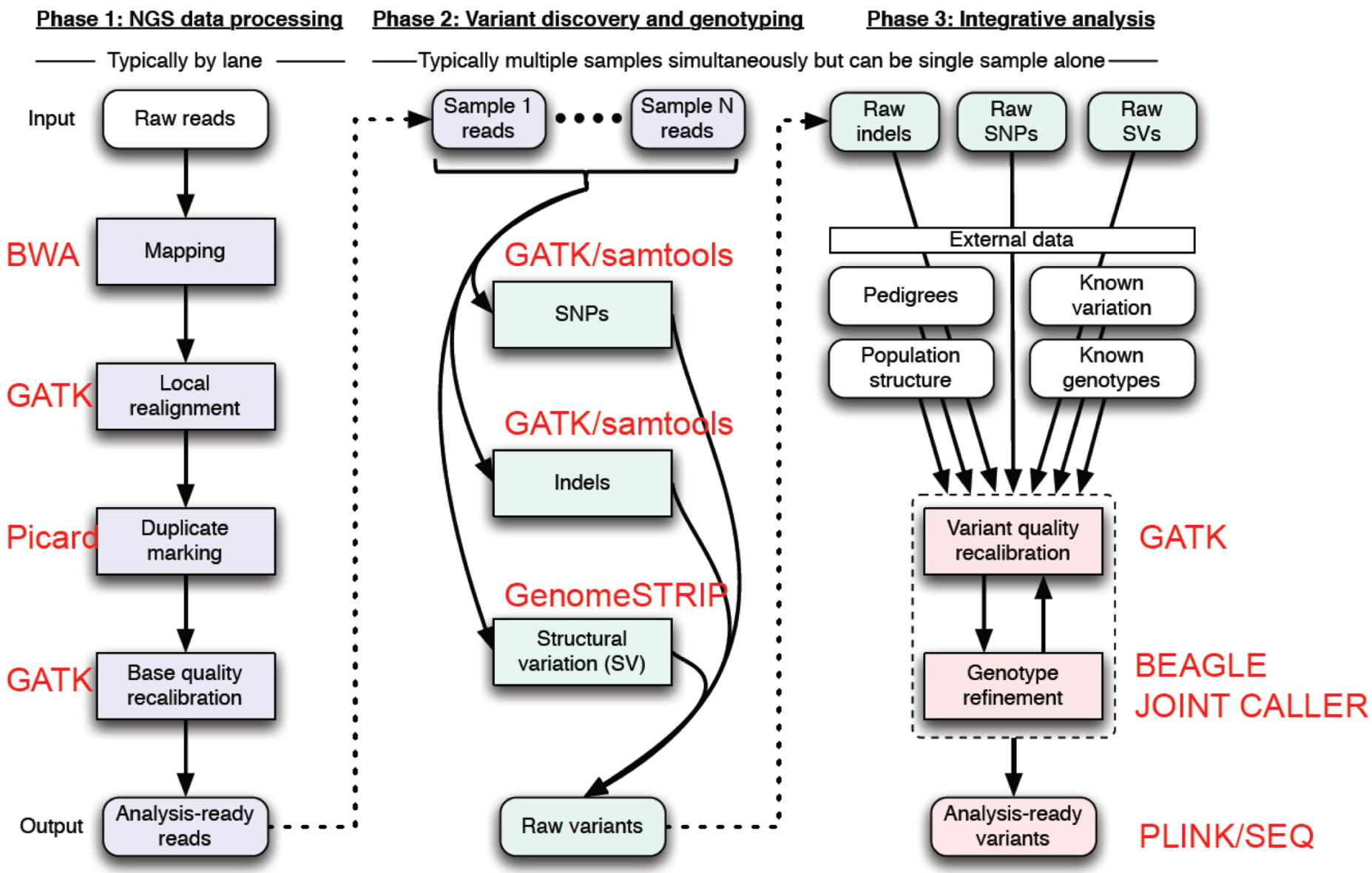
- **Alzheimer's 39**
  - 5M $200B
- **Parkinson's 19**
  - 1M $25B
- **Diabetes 62**
  - 27M $174B
- **Autism 8**
  - 2M $60B



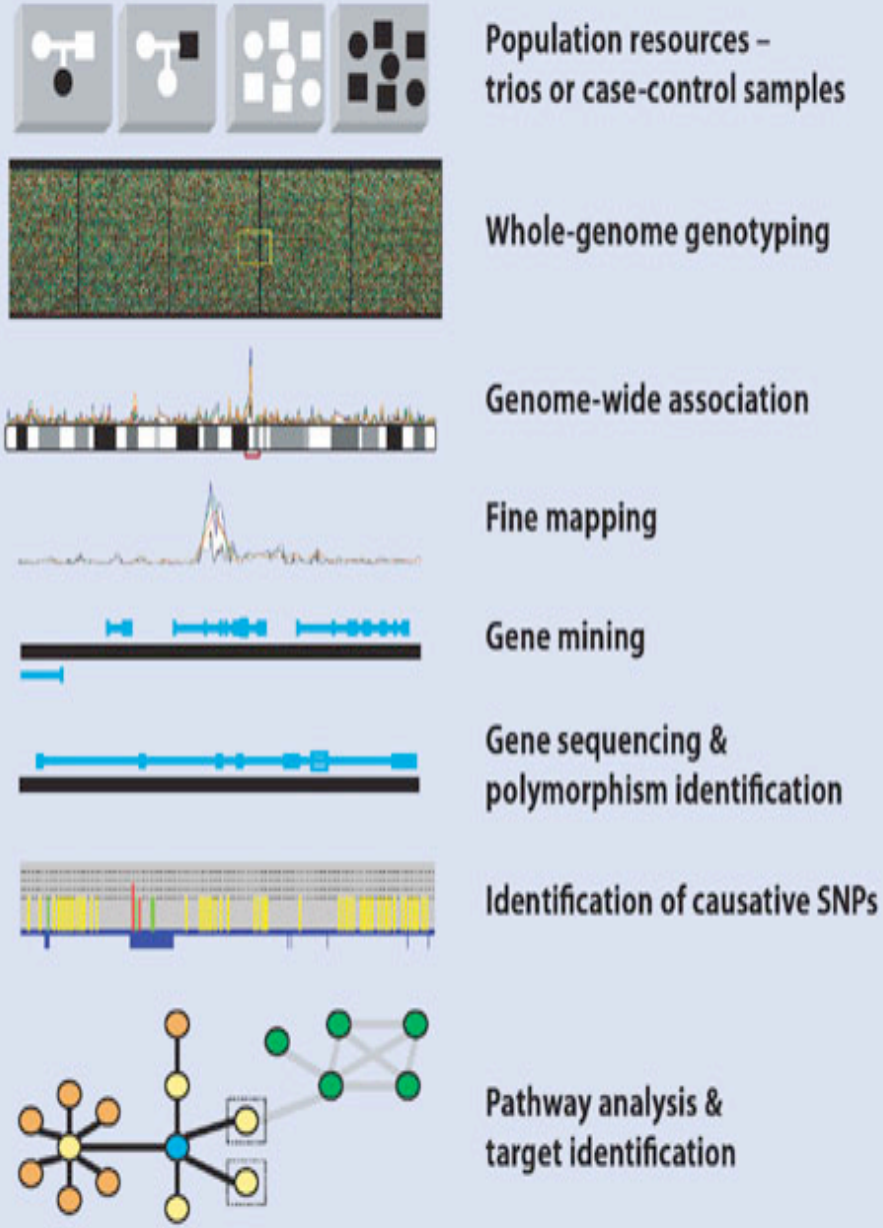
Current GWAS methods consider single SNPs in a linear Additive fashion. Limits due to method, study design, Assay cost and and computing power.

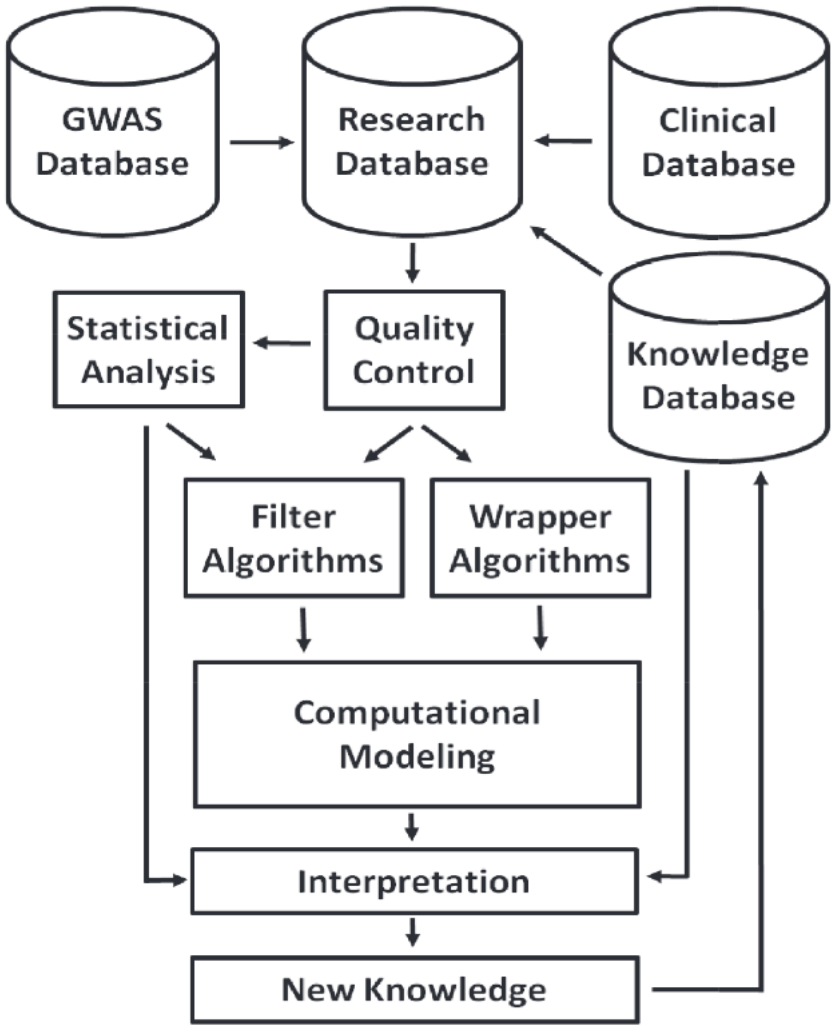
With increases in sequencing capability and computing capability it will be possible to dramatically increase the power of GWAS studies for complex genetic diseases


a Illumina HiSeq
Generate Millions of Sequencing Reads
b Raw Sequencing Read Output
c Align Millions of Sequencing Reads to the Reference Genome
d Identify and Annotate Variant:
GJB2 Val167Met carrier

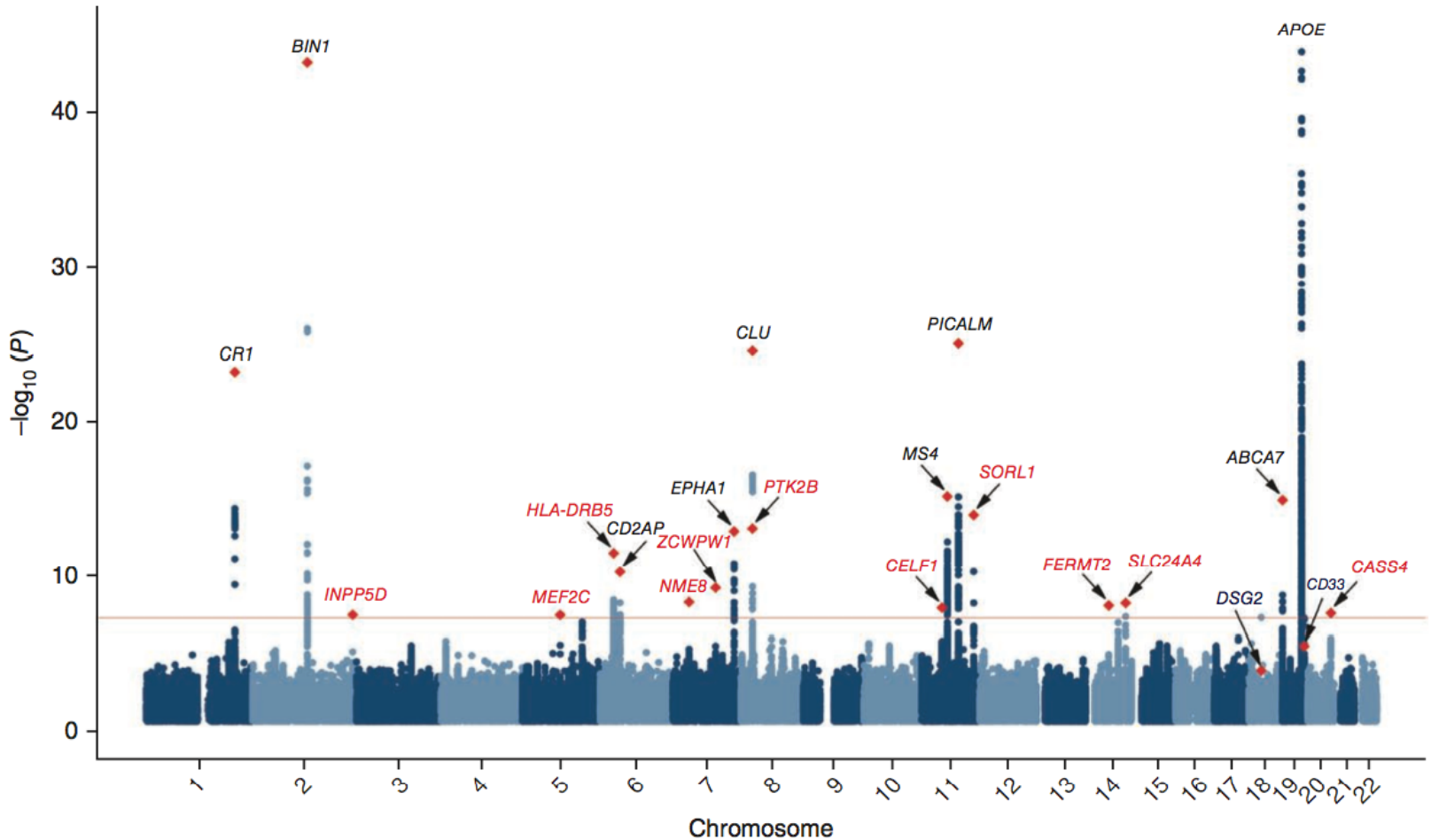


http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v2


Population resources –
trios or case-control samples
Whole-genome genotyping
Genome-wide association
Fine mapping
Gene mining
Gene sequencing &
polymorphism identification
Identification of causative SNPs
Pathway analysis &
target identification


GWAS Database
Research Database
Clinical Database
Knowledge Database
Statistical Analysis
Quality Control
Filter Algorithms
Wrapper Algorithms
Computational Modeling
Interpretation
New Knowledge

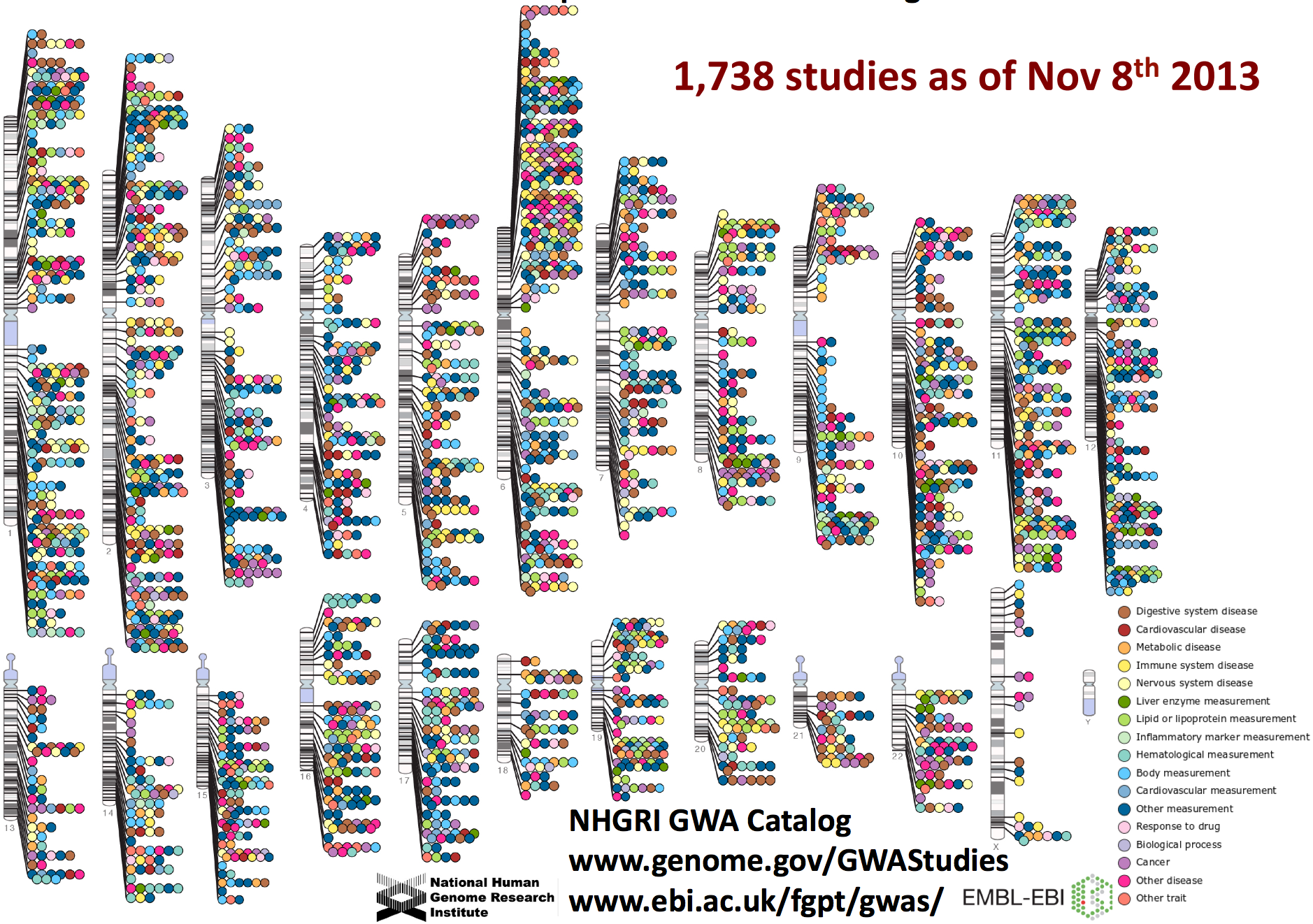
11 new loci Implicated in Alzheimer's from a meta-analysis of 54,162 individuals



Nature Genetics published online 27 October 2013; doi:10.1038/ng.2802

Published Genome-Wide Associations through 12/2012
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

1,738 studies as of Nov 8th 2013



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Pair-HMM is the biggest culprit for the low performance

Stage	Time	Runtime %
Assembly	2,598s	13%
Pair-HMM	14,225s	70%
Traversal + Genotyping	3,379s	17%

NA12878 80xWGS chromosome 20 haplotype caller run
Chr20 time: 5.6 hours
WGS time: 7.6 days

Sandbox Performance comparison

Data: NA12878 80xWGS chromosome 20

TECH	Hardware	Runtime (seconds)	Improvement (fold)
AVX	Intel Xeon 24-core*	15	720x
GPU	NVidia Tesla K40	160	67x
GPU	NVidia GeForce GTX Titan	161	67x
GPU	NVidia GeForce GTX 480	190	56x
GPU	NVidia GeForce GTX 680	274	40x
GPU	NVidia GeForce GTX 670	288	38x
AVX	Intel Xeon 1-core*	309	35x
FPGA	Convey Computers HC2	834	13x
-	C++ (baseline)	1,267	9x
-	Java (gatk)	10,800	-

* Reported by Intel - unreleased hardware

Sandbox Performance comparison

Data: NA12878 80xWGS chromosome 20

TECH	Hardware	Runtime (seconds)	Improvement (fold)
AVX	Intel Xeon 24-core*	15	720x
GPU	NVidia Tesla K40	160	67x
GPU	NVidia GeForce GTX Titan	161	67x
GPU	NVidia GeForce GTX 480	190	56x
GPU	NVidia GeForce GTX 680	274	40x
GPU	NVidia GeForce GTX 670	288	38x
AVX	Intel Xeon 1-core*	309	35x
FPGA	Convey Computers HC2	834	13x
-	C++ (baseline)	1,267	9x
-	Java (gatk)	10,800	-

Computational Characteristics of Bioinformatics Analyses

- Compute-intensive, small to moderate data
  - Similarity search, alignment, trees, metabolic modeling, protein folding, molecular docking, GWAS
- Data-intensive, big data
  - Short read mapping, NGS error correction, kmer counting, phylogenetic binning, variation analysis
- Compute-intensive, big data
  - Assembly, network analysis, all-to-all similarity, inferencing over multiple data types
    - Tradeoffs: speed, memory, quality

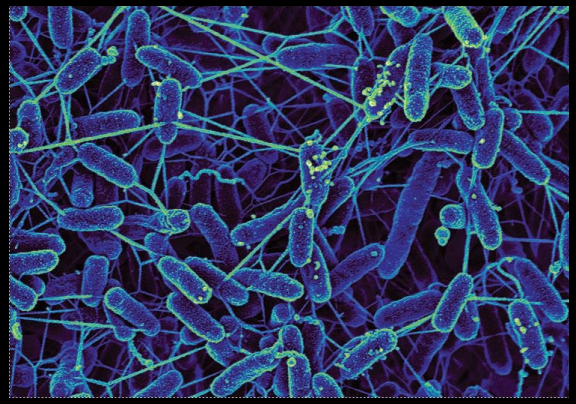
Big Data Challenges for Bioinformatics

- New types of methods and new algorithms
  - From $O(N^2) \Rightarrow O(N \log N) \Rightarrow O(N) \Rightarrow O(K)$
  - Non-alignment methods and streaming
- New types of Infrastructure bringing biological data and computing together
  - Users need to have an environment where they don't need to move the data to work
- Ability to share methods, protocols, tools and insights leveraging social networks
  - Enable the best methods to win regardless of where they come from

DOE KBase is One Such Effort

Knowledgebase enabling ***predictive*** systems biology.

- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Resource to enable **experimental design** and **interpretation** of results.



Microbes

Communities

Plants

Acknowledgements

- Many many people are working to build the things I've talked about today.. I can't mention them all but you should know about a few
- Ross Overbeek, Chris Henry, Fangfang Xia, Folker Meyer, Tom Brettin, Bob Olson, Terry Disz, Andreas Wilke, Sam Seaver, Veronika Vonstein, Scott Devoid, Gordon Pusch, Bruce Parello, Jared Wilkens, Adam Arkin, Daniel Quest, Chris Bun, Jennifer Salazar, Elizabeth Glass, Shane Canon, Narayan Desai, Matt Dejongh, Aaron Best, Tobias Paczian, Peter Larson and many more

Acknowledgements

- Many thanks to DOE, NSF, NIH, DOD, ANL, UC, Moore Foundation, Sloan Foundation, Apple, Microsoft, Cray, Intel and IBM for supporting our research groups over the years

Thank You for Listening