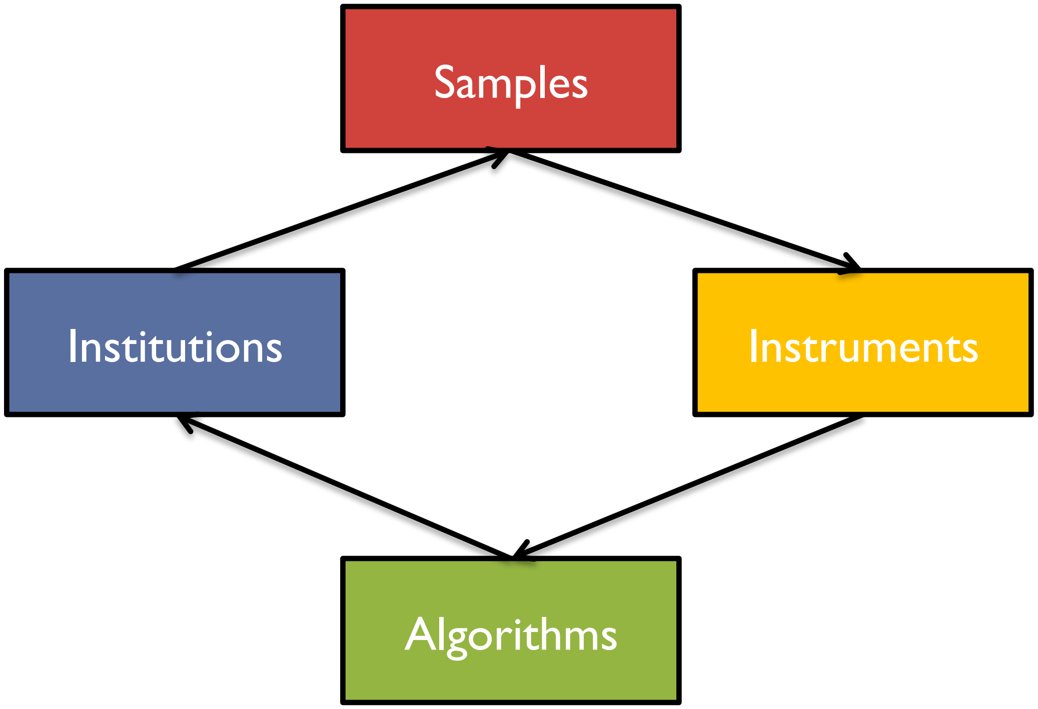
IT Considerations: Hurdles and Solutions

Michael Schatz

April 29, 2013

Developing a Neuroscience Consortium

Outline



Samples

Institutions

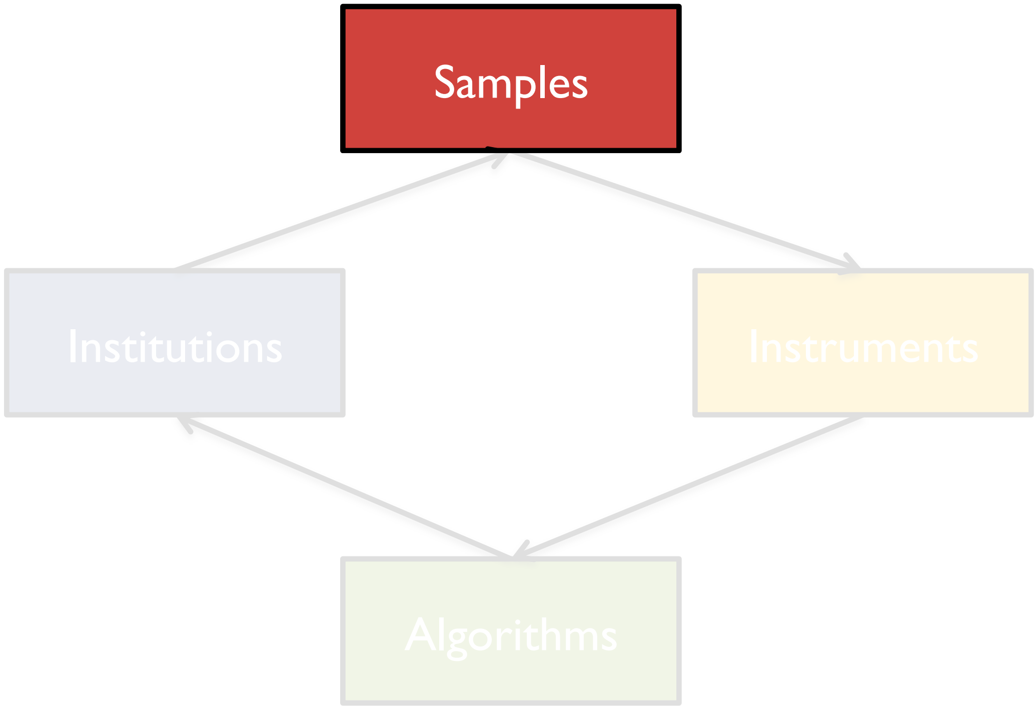
Instruments

Algorithms

The biggest IT challenge is managing diversity

Corollary: There is no single magic bullet

Outline



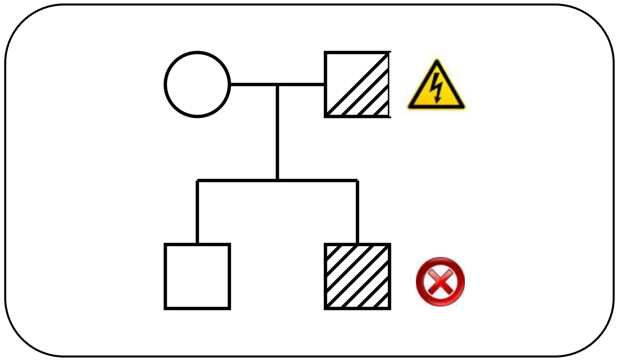
Samples

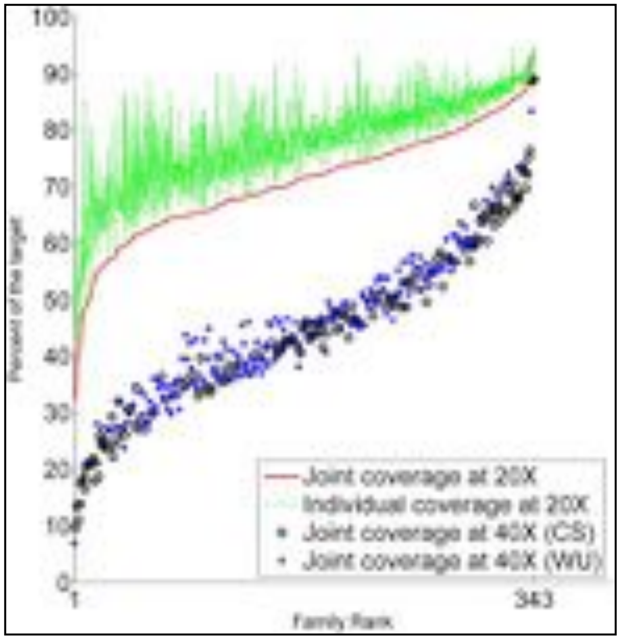
Institutions

Instruments

Algorithms

De novo genetics of autism





Sequencing of 343 families from the Simons Simplex Collection

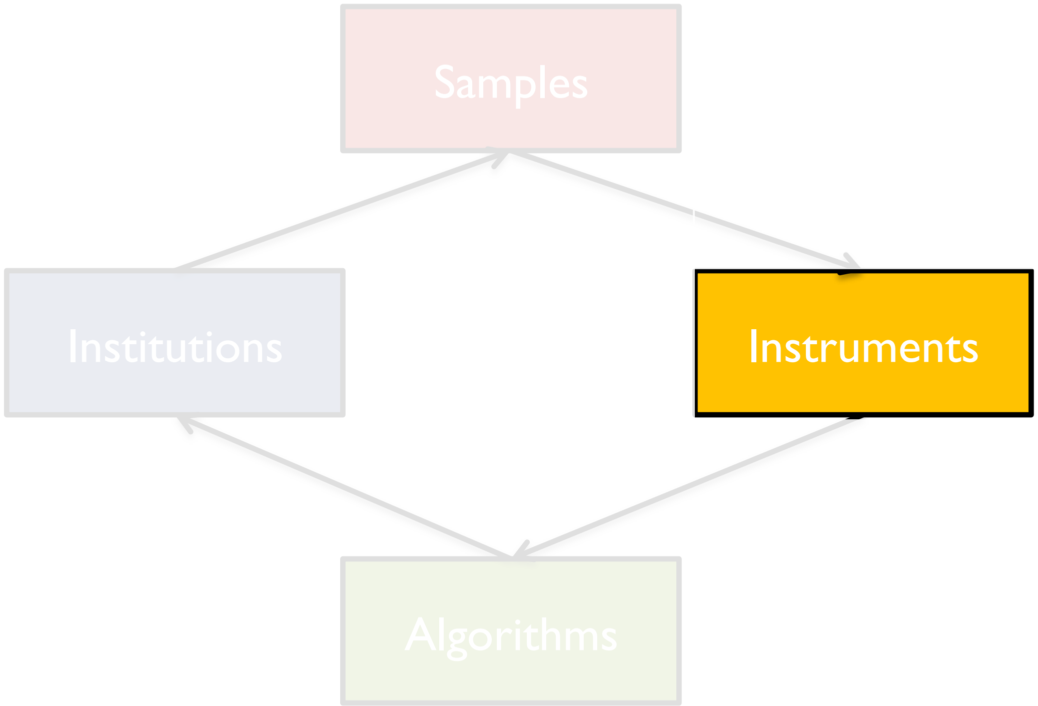
- Parents plus one child with autism and one non-autistic sibling
- Chose to do whole exome sequencing to balance costs with genome coverage
- Discovered significant enrichment in de novo likely gene killing mutations

De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron.* 74:2 285-299

Samples

- **Organisms**
  - Humans, Animal models, Others?
  - Different scales, complexity, genome structures
- **Genetics**
  - Genome, Exome, Transcriptome, Methylome, etc
  - LIMS, Metadata of sample treatment
- **Phenotypes and Environments**
  - Behavior, growth, response to treatments, etc
  - Ontologies, Qualitative/Quantitative scoring
- **Populations**
  - Large numbers, different conditions, timeseries
  - Database of individuals, privacy, access control
- **Sample types**
  - Gross tissue, single tissue, single cell, single molecule
  - Sample tracking errors, QA/QC

Outline



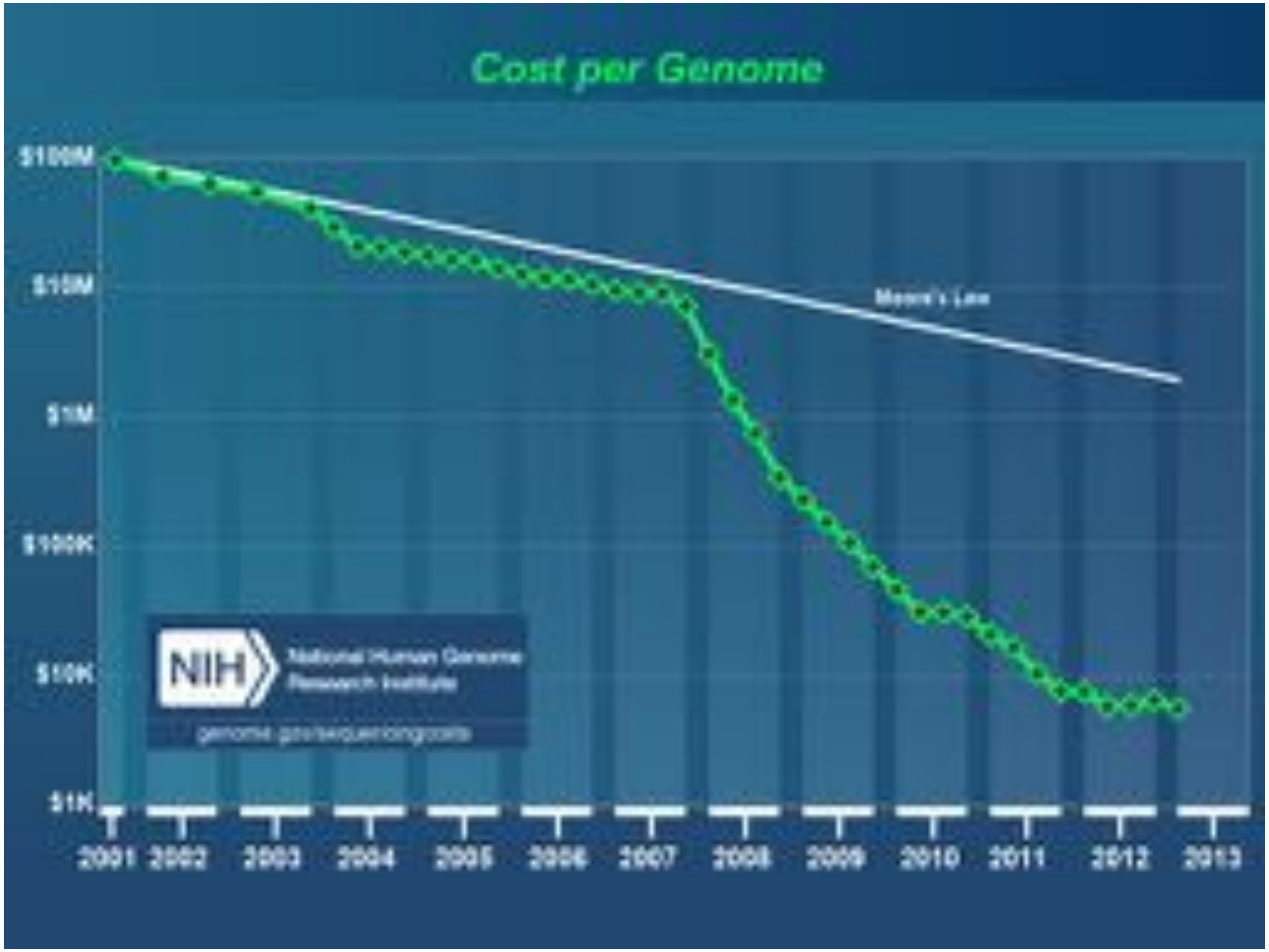
Samples

Institutions

Instruments

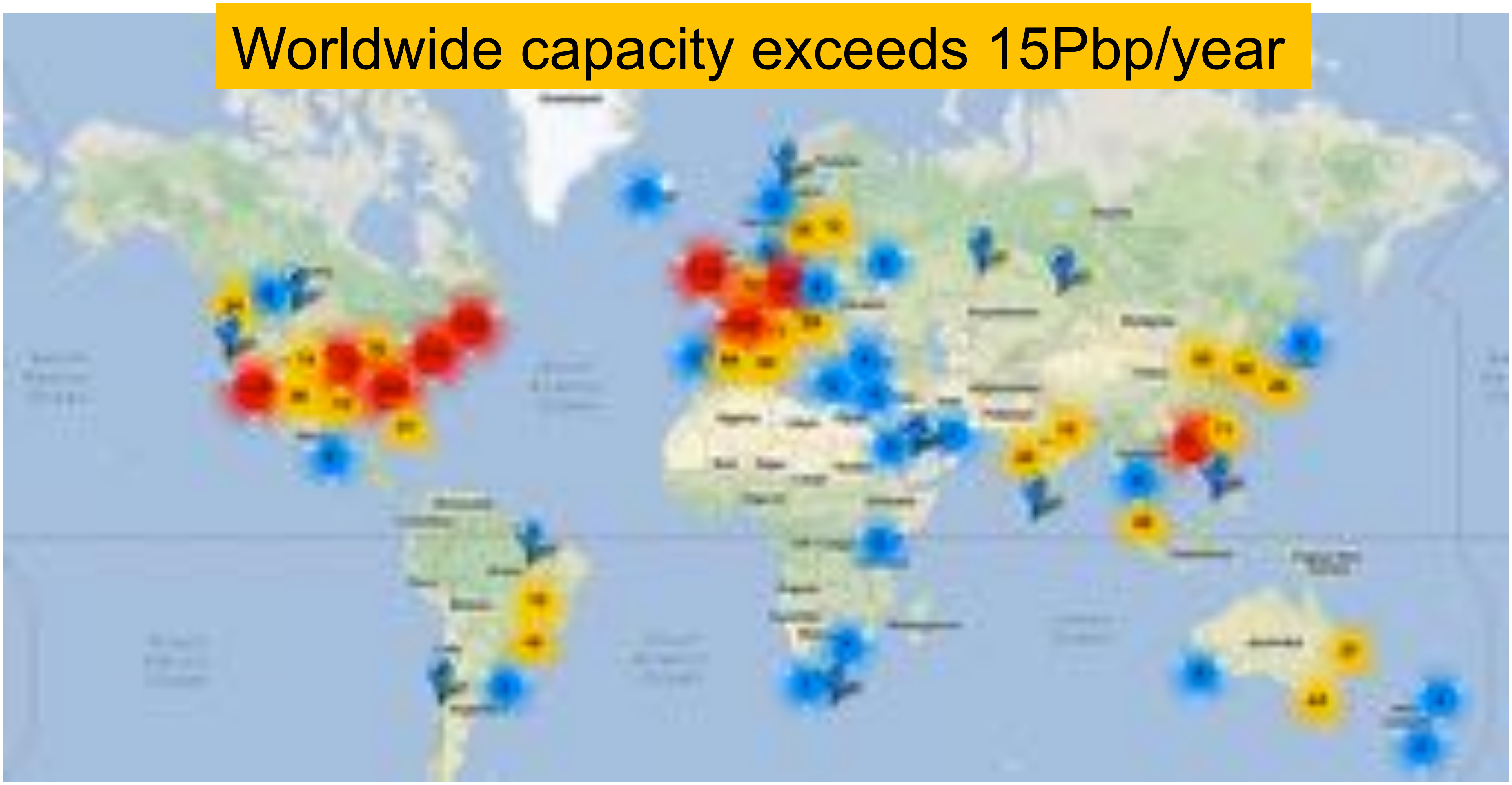
Algorithms

Cost of Sequencing



NHGRI: DNA Sequencing Costs
http://www.genome.gov/sequencingcosts/

Sequencing Centers

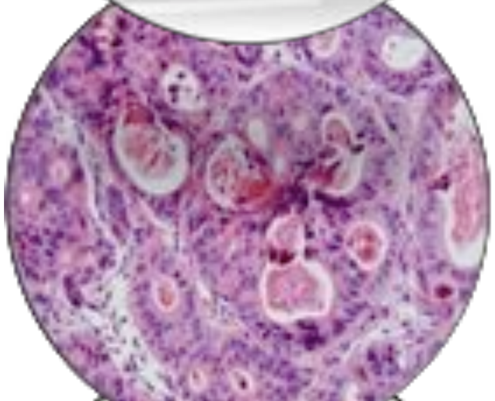


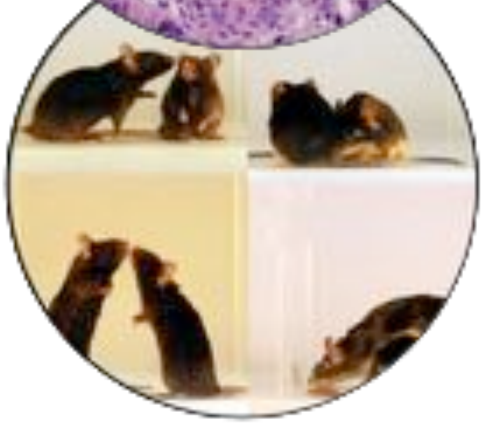
Worldwide capacity exceeds 15Pbp/year

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com/

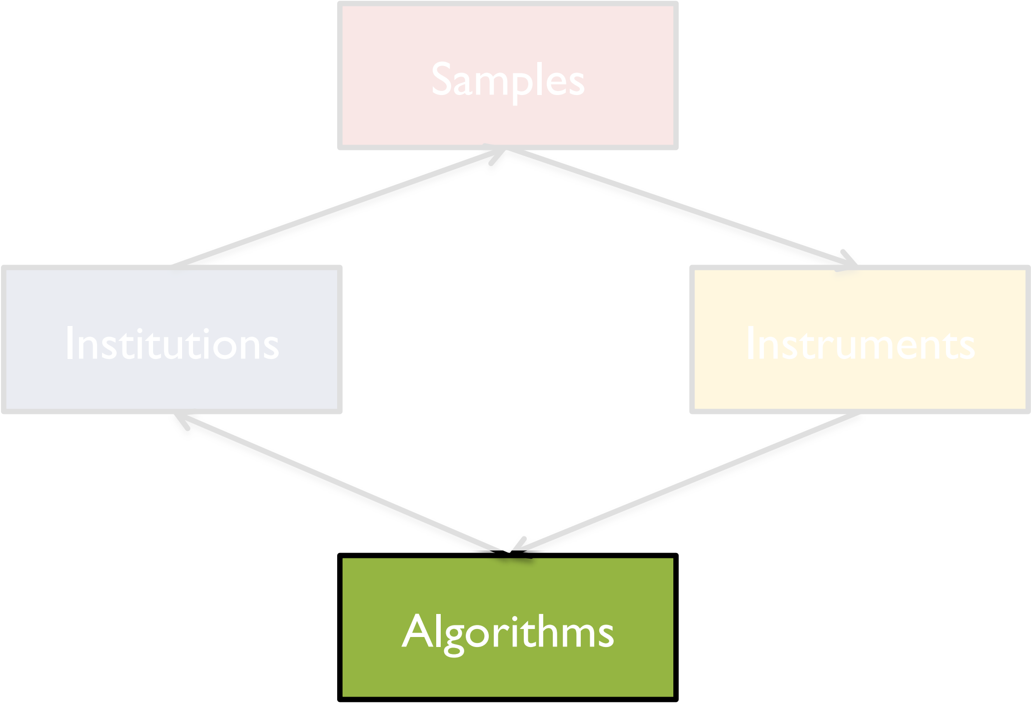
Instruments

- **Sequencing Platforms**
    - Illumina/Life/Ion/PacBio/Moleculo/Oxford Nanopore
- **Phenotyping Platforms**
    - Animal Tracking, Growth Tracking, Cell Tracking
- **Scale**
    - 1 instrument ~100Gbp / day;
    - Institute: 1Tb/day;
    - Worldwide: 15Pb/year
    - Compression: Precious samples to routine analysis
- **Dispersed Resources**
    - Not organized around a few large collectors
    - Variable quality
- **Rapidly changing landscape**
    - Each instrument has different characteristics and error models that need to be modeled and corrected





Outline



Samples

Institutions

Instruments

Algorithms

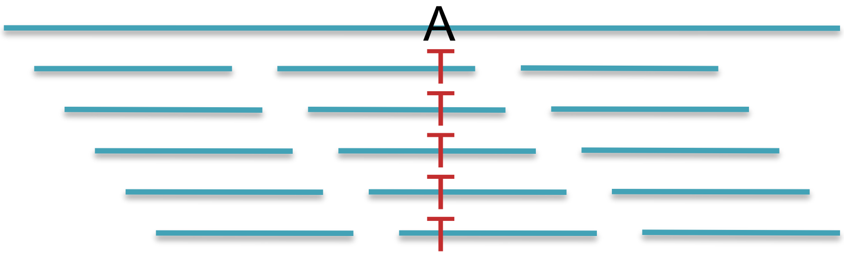
Genomics Applications

GENOME 10K

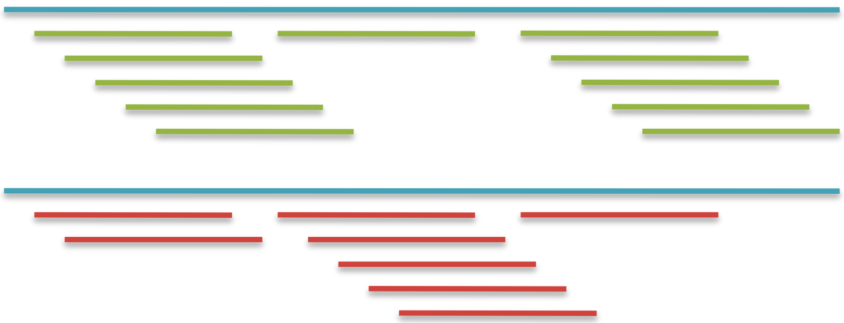
ENCODE

HMP

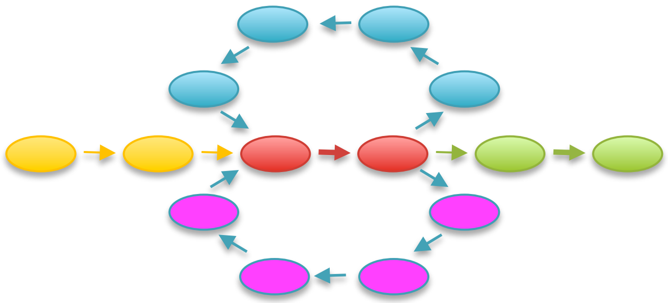
Alignment & Variations



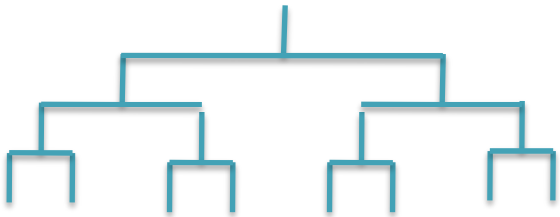
Differential Analysis



De novo Assembly

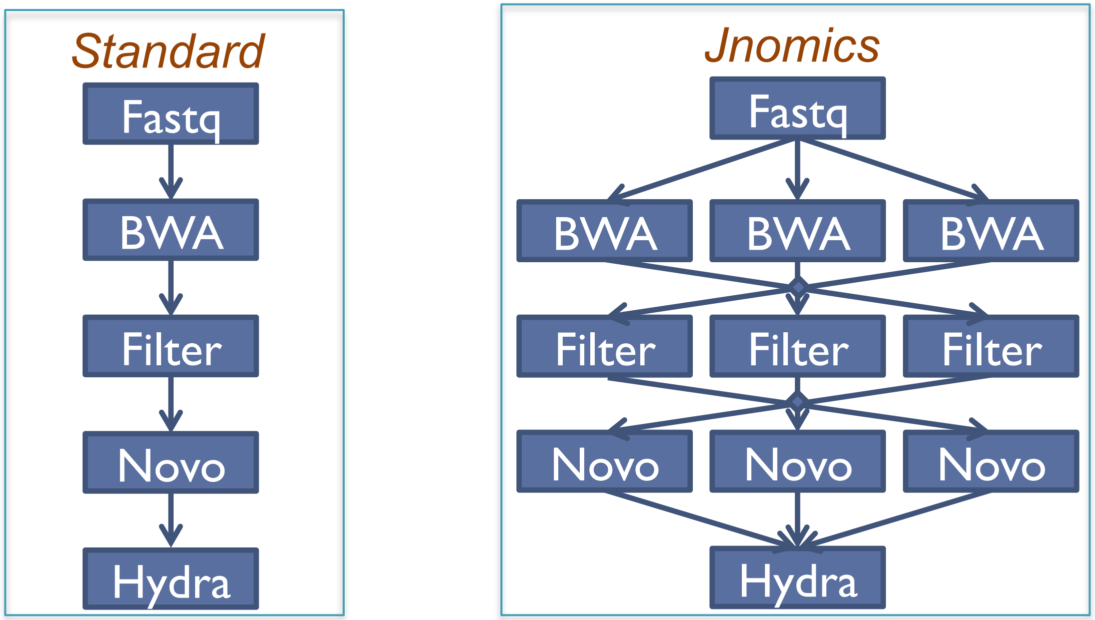


Phylogeny & Modeling



Jnomics: Cloud-scale genomics

James Gurtowski, Matt Titmus, Michael Schatz



- Rapid parallel execution of NGS analysis pipelines
  - FASTX, BWA, Bowtie, Novoalign, SAMTools, Hydra
  - Population analysis: Clustering, GWAS, Trait Inference
  - Integrate compute and storage resources together
- 200-fold performance gains analyzing 1TB genetic data

Answering the demands of digital genomics
Titmus, M.A.., Gurtowski, J, Schatz, M.C.. (2012) *Concurrency & Computation*

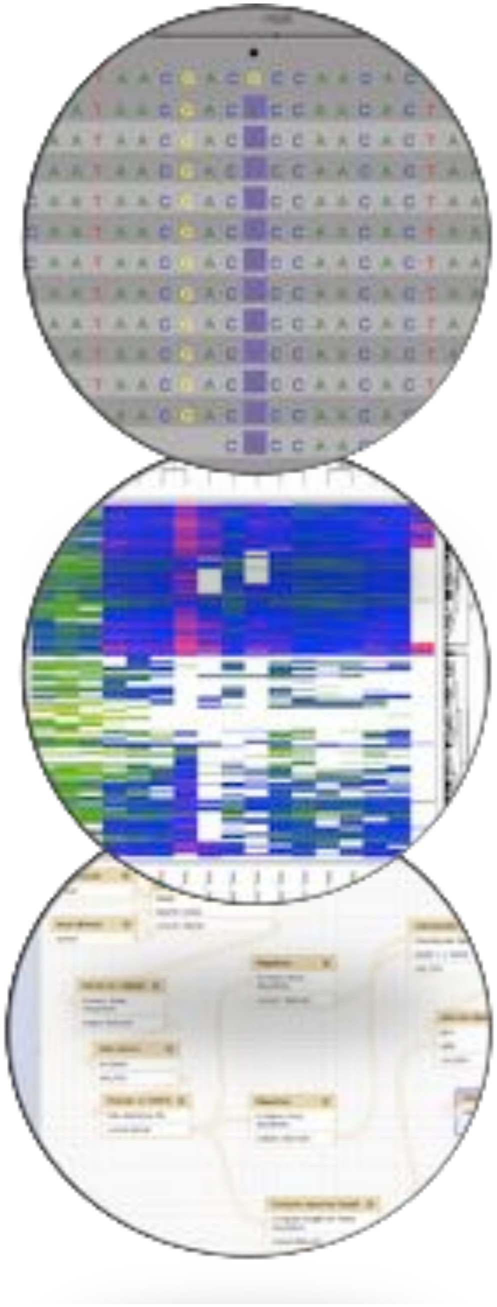
Algorithms

- **Applications**
  - De novo assembly, Variant Detection, Phylogenies
  - Differential Expression, Correlations
  - Integration, Modeling, Machine Learning

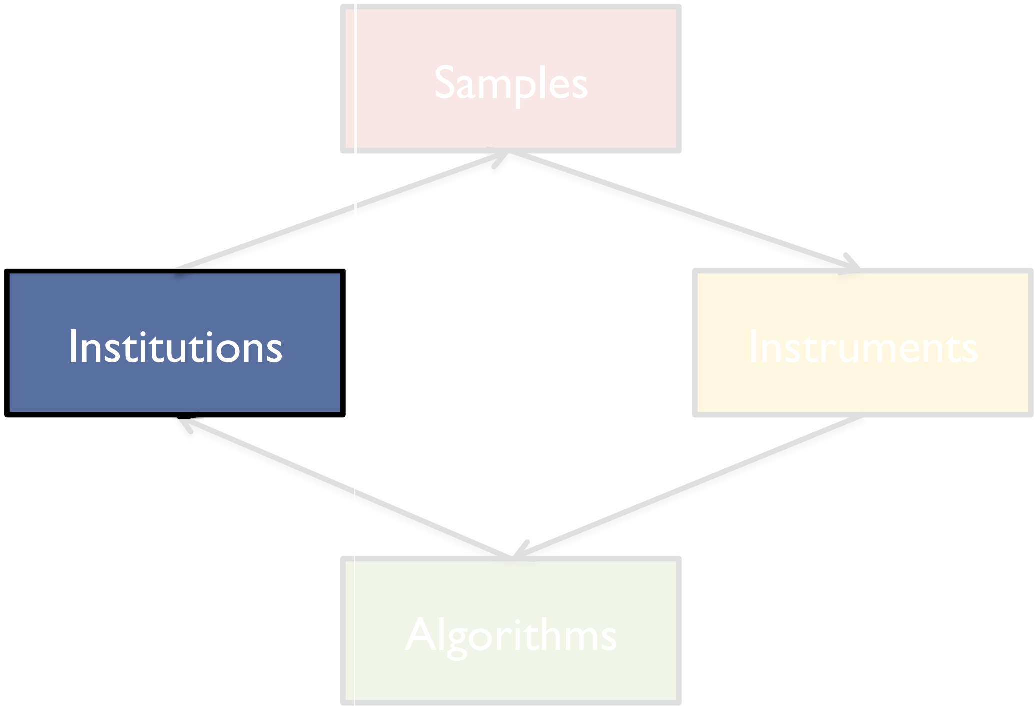
- **Requirements**
  - High memory/High CPU/High IO/High throughput
  - Visualization & user friendliness

- **Integration**
  - Federation or mirroring of data
  - Monitoring data quality; IDR
  - Rich resources in some species/diseases, less so in others

- **Workflows**
  - Provenance, reproducible workflows
  - Rapidly changing best practices



Outline



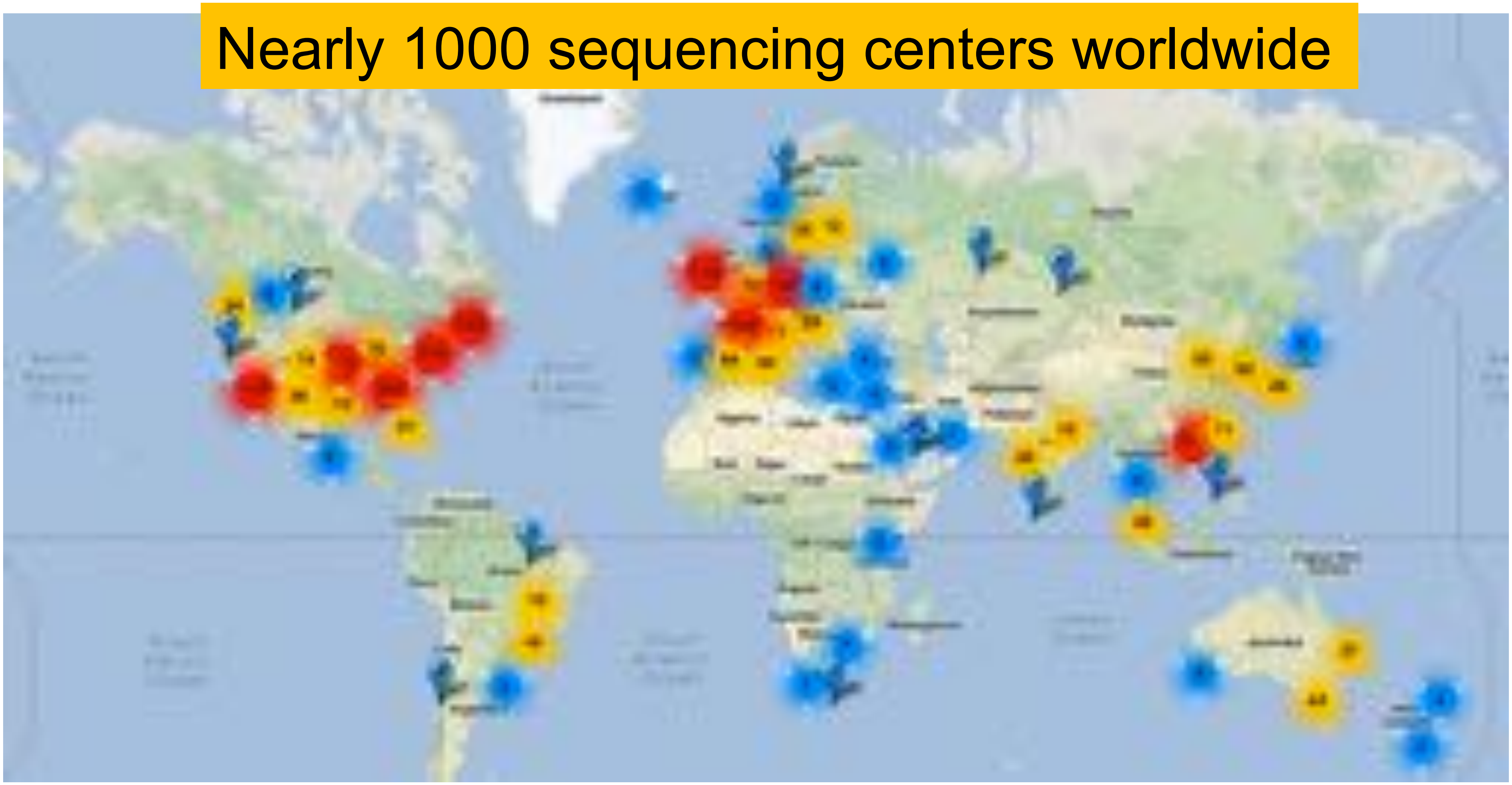
Samples

Institutions

Instruments

Algorithms

Sequencing Centers

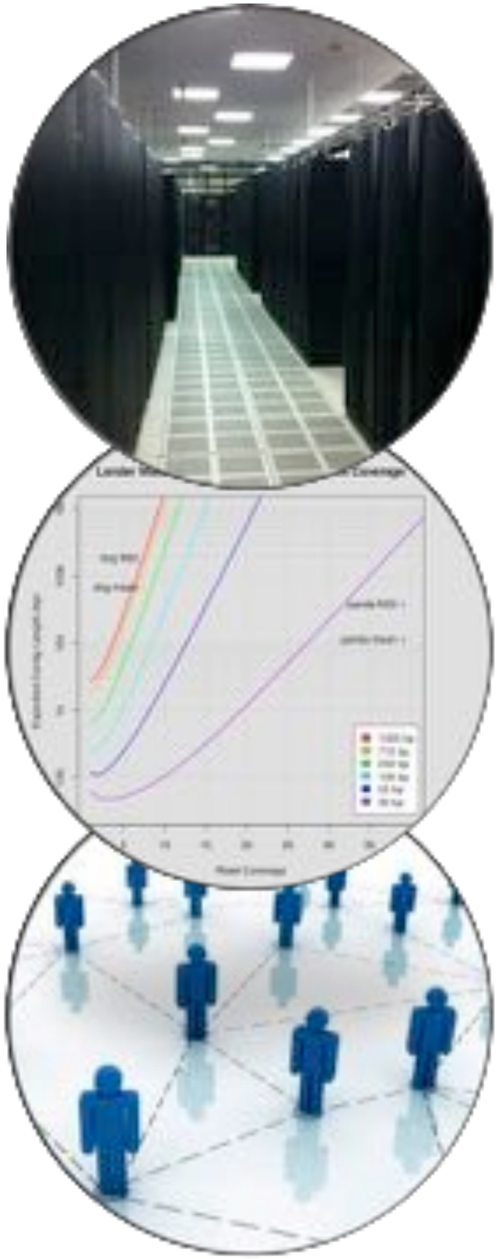


Nearly 1000 sequencing centers worldwide

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com/

Institutional

- **IT Resources**
  - Network; Storage; Cores; HVAC; Power
  - Parallel Computing is hard
- **Expertise**
  - Alg Developers/Expert users/End users
  - Quantitative Education is hard
- **Data Reuse**
  - Moving large amounts of data is hard
  - Data quality becomes essential
- **Collaborative projects**
  - How do we coordinate/communicate resources
  - Cross-institution access and privacy requirements
- **Data are complex, requiring deep understanding**
  - Reinventing the wheel is (generally) okay, because we all need slightly different wheels



Summary

- Potential scale of data is enormous
  - Parallel computing aka distributed computing aka cloud computing may be our only hope for keeping up with the pace of advance
  - Move code to data whenever possible
- Managing the diversity of projects is the biggest challenge
  - Certain applications are common, but a long tail of important, but lesser used ones
  - Landscape is extremely dynamic with new instruments and algorithms released every day
- Key to success is a focused vision.
  - Integrate resources into the existing ecosystem
  - What are the incentives and enforcements available?

Thank You

http://schatzlab.cshl.edu/

@mike_schatz