eScience and Post-Genome Biomedical Research

Thomas L. Casavant, Adam P. DeLuca

Departments of Biomedical Engineering, Electrical Engineering and Ophthalmology

Coordinated Laboratory for Computational Genomics (CLCG)

Center for Bioinformatics and Computational Biology (CBCB)

http://genome.uiowa.edu

Outline

1. General Observations about Post-genomic eScience

1. Specific Case Study where Traditional Publication and Archiving Practices are Lacking

1. Impact of Emerging Technology on Scale of the eScience Problems

Post-Genomic *eScience*

- The "Post-Genome" Era

3 Primary Types of Investigation

1. Generation of New High-Throughput Data (new "Genome Projects")
2. Generating New Data in the Context of Existing Results (Published or in Databases)
3. No New Data – Exclusive re-use of "Published" Data

Case Study: Genomic Rearrangements or Deletions/Duplications

(Dr. Krishna Rani Kalari, Mayo Clinic)

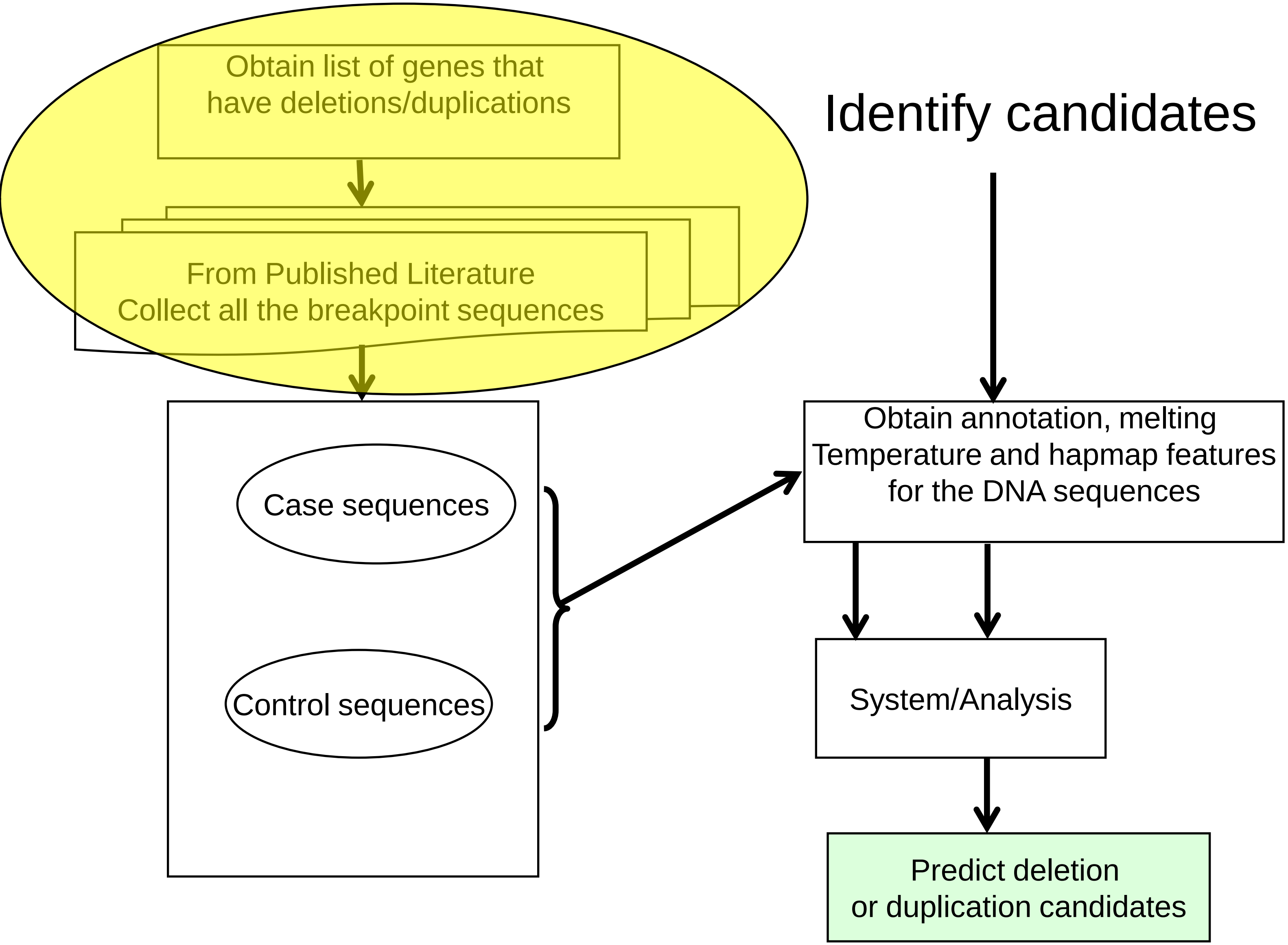
1. Goal: Identification of Human disease causing mutations
2. Observation: Assays exist to identify deletions and duplications
   - time consuming
   - laborious
   - expensive

3. Approach: Develop *In-silico* procedures to identify and prioritize candidate deletion/duplication sites and accelerate the finding of disease mutation discovery

Approach Details

1. Construct case and control data sets for all known cases of disease causing unequal recombinations
2. Identify and obtain informative sequence-based features to create a training set
3. Evaluate machine learning methods on the training set
4. Design and develop a computational system to identify and prioritize candidate intragene deletions and duplications

Approach - System Level



Obtain list of genes that have deletions/duplications

From Published Literature
Collect all the breakpoint sequences

Identify candidates

Case sequences

Control sequences

Obtain annotation, melting Temperature and hapmap features for the DNA sequences

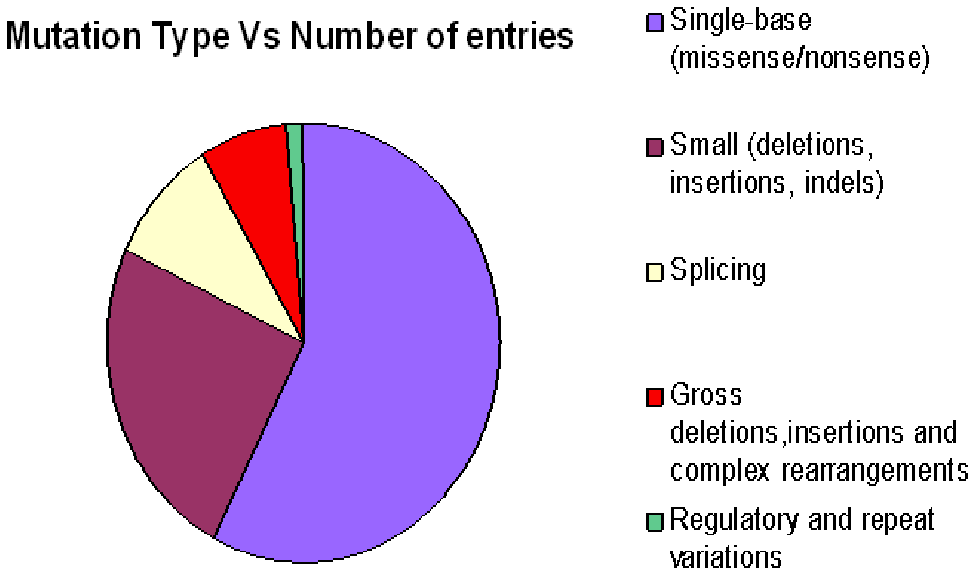
System/Analysis

Predict deletion or duplication candidates

HGMD statistics

- **2362 genes have 64251 mutations**

- 7 % (4,500) of the mutations in HGMD are caused by gross deletions and duplications.



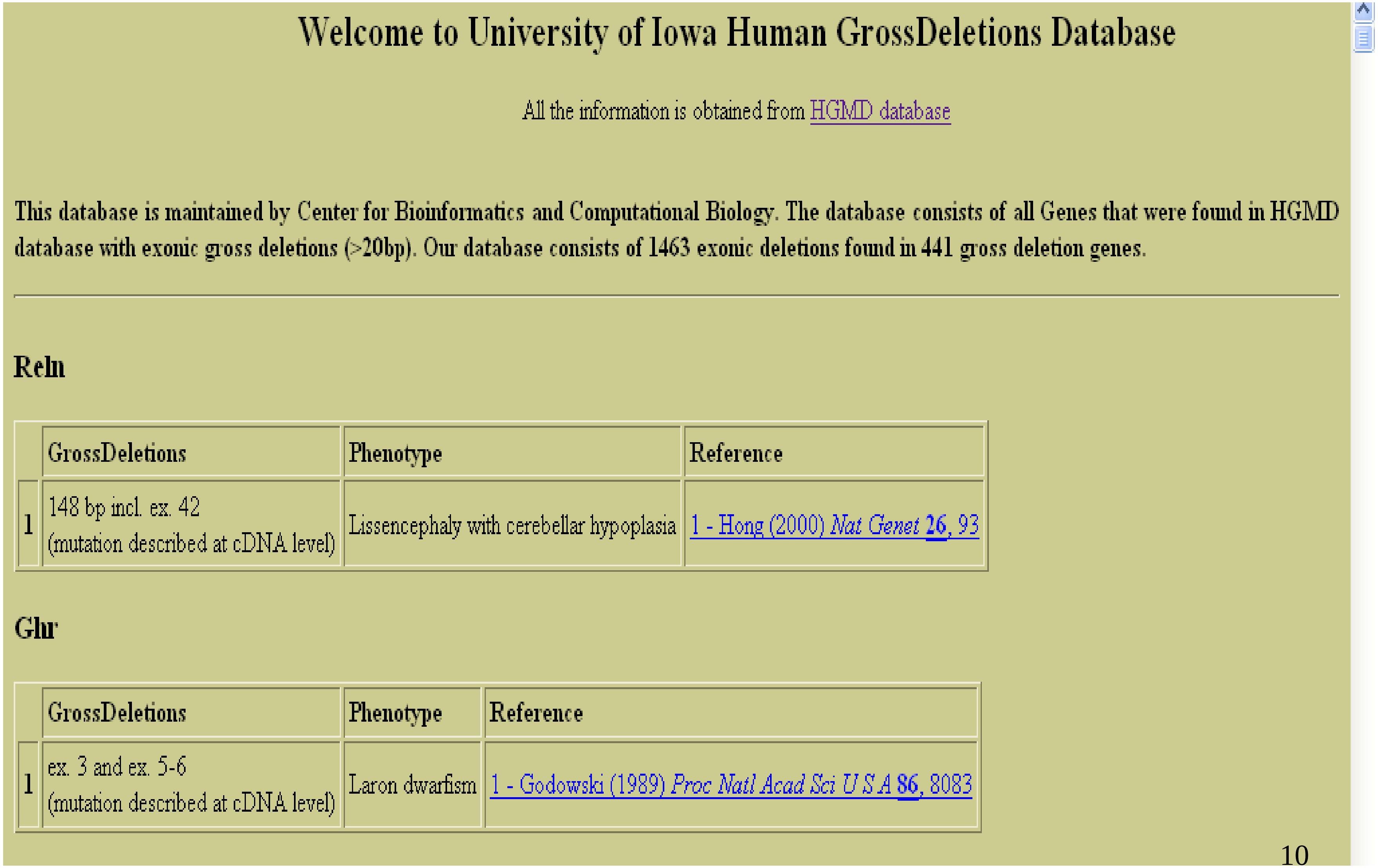
HGMD mutation classification

Mutation type	Total number of mutations
Nucleotide substitutions (missene / nonsense)	294
Nucleotide substitutions (splicing)	46
Nucleotide substitutions (regulatory)	0
Small deletions	52
Small insertions	12
Small indels	1
Gross deletions	2
Gross insertions and duplications	0
Complex rearrangements (inversions)	1
Repeat variations	0

HGMD - Gross deletions

Accession Number	Description	Phenotype	Reference
CG035110	ex. 18 (described at genomic DNA level)	Stargardt disease	Yatsenko (2003) Hum Mutat **21,** 636
CG994802	36 bp nt. 6543 (described at genomic DNA level)	Stargardt disease	Lewis (1999) Am J Hum Genet **64,** 422

Local Deletion Database



Welcome to University of Iowa Human GrossDeletions Database

All the information is obtained from HGMD database

This database is maintained by Center for Bioinformatics and Computational Biology. The database consists of all Genes that were found in HGMD database with exonic gross deletions (>20bp). Our database consists of 1463 exonic deletions found in 441 gross deletion genes.

Reln

	GrossDeletions	Phenotype	Reference
1	148 bp incl. ex. 42 (mutation described at cDNA level)	Lissencephaly with cerebellar hypoplasia	1 - Hong (2000) *Nat Genet* **26**, 93

Ghr

	GrossDeletions	Phenotype	Reference
1	ex. 3 and ex. 5-6 (mutation described at cDNA level)	Laron dwarfism	1 - Godowski (1989) *Proc Natl Acad Sci U S A* **86**, 8083

Of the 4,500 Possible Training Cases, How Many Did We Get???

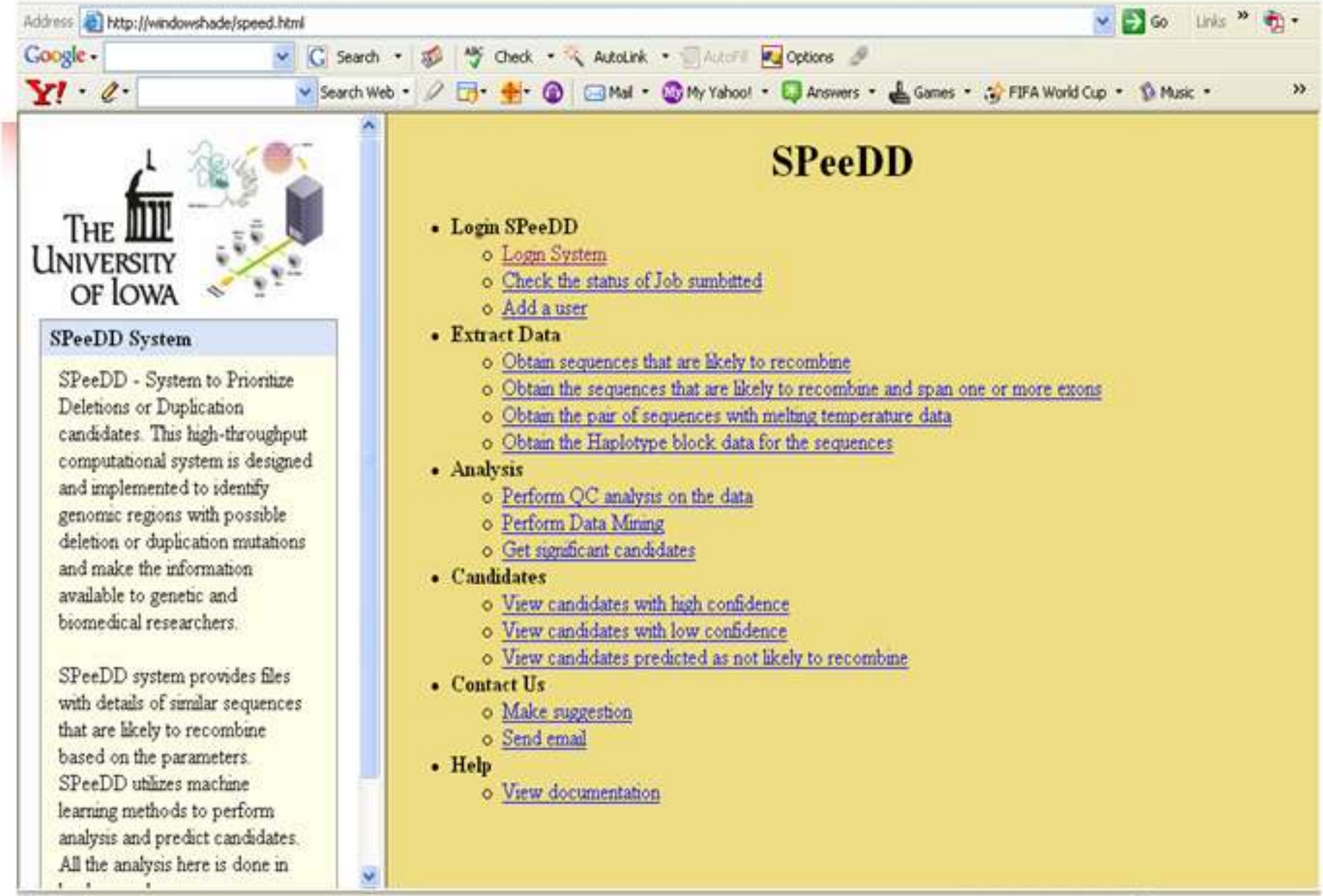
Searched for specific break point information for 1463 IDDs described in HGMD

Identified 102 fully-characterized rearrangement breakpoints (cases)

- know exactly where the breakpoint occurs

Identified 2338 matching set of breakpoints for each of the positives for which IDDs have not been observed (controls)

SPeeDD web-interface



SPeeDD System

SPeeDD - System to Prioritize Deletions or Duplication candidates. This high-throughput computational system is designed and implemented to identify genomic regions with possible deletion or duplication mutations and make the information available to genetic and biomedical researchers.

SPeeDD system provides files with details of similar sequences that are likely to recombine based on the parameters. SPeeDD utilizes machine learning methods to perform analysis and predict candidates. All the analysis here is done in

SPeeDD

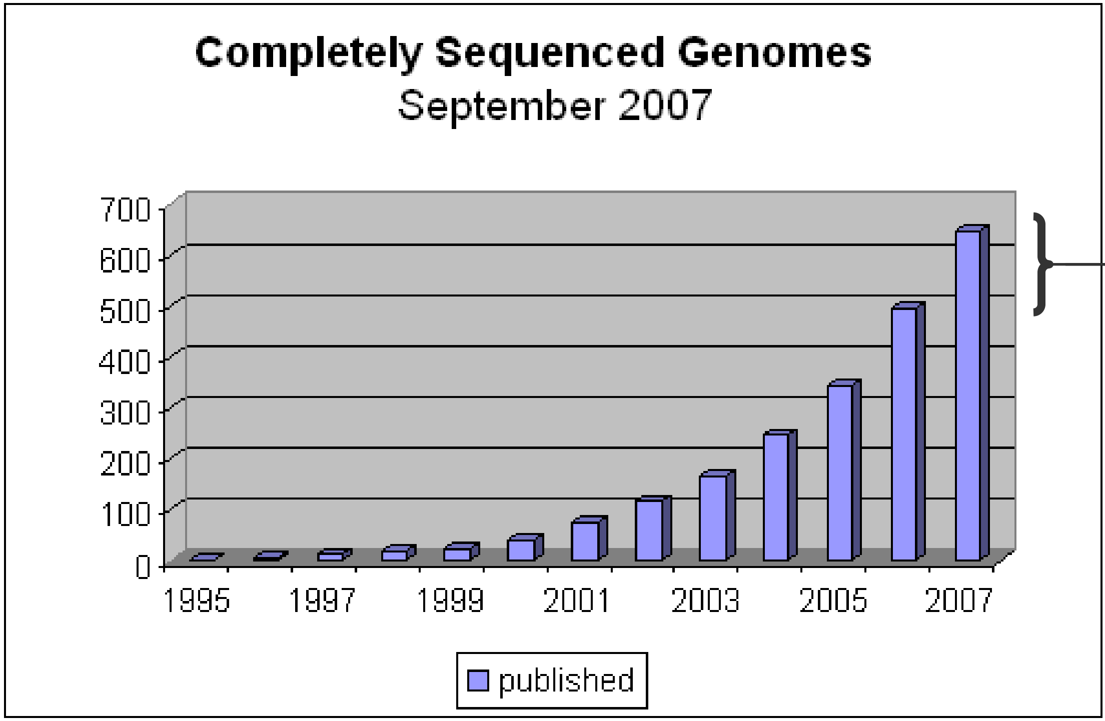
- **Login SPeeDD**
  - Login System
  - Check the status of Job sumbitted
  - Add a user
- **Extract Data**
  - Obtain sequences that are likely to recombine
  - Obtain the sequences that are likely to recombine and span one or more exons
  - Obtain the pair of sequences with melting temperature data
  - Obtain the Haplotype block data for the sequences
- **Analysis**
  - Perform QC analysis on the data
  - Perform Data Mining
  - Get significant candidates
- **Candidates**
  - View candidates with high confidence
  - View candidates with low confidence
  - View candidates predicted as not likely to recombine
- **Contact Us**
  - Make suggestion
  - Send email
- **Help**
  - View documentation

Lessons From Deletion Case Study

- Important results and “data” are buried in traditional forms of scientific publication and dissemination mechanisms (no surprise here).
- Fidelity and throughput of legacy results inadequate
- Necessary data can be requested from investigators
  - In some cases
  - Reference to a changing world of what is assumed to be “known”

- Stay tuned… the problem will only get worse…

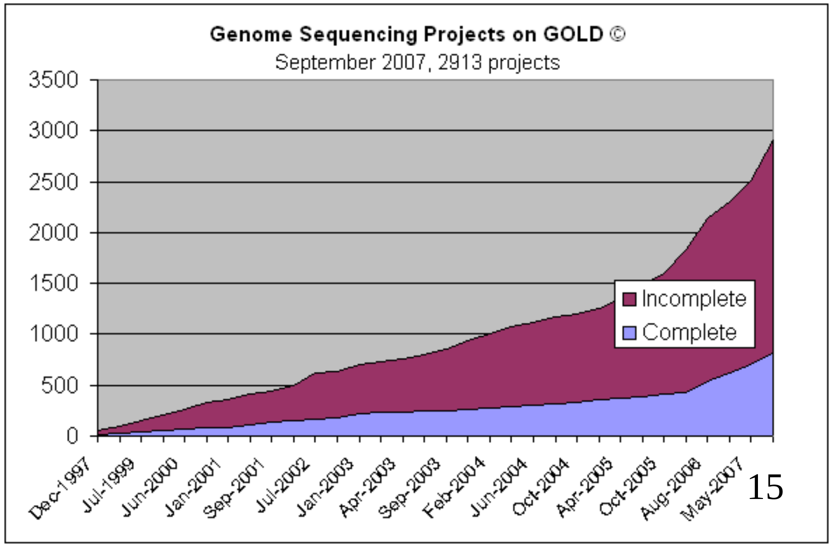
Impact of Emerging Technology on Scale of the eScience Problems



~3 genomes/week

Genomes Online Database v2.0

www.genomesonline.org



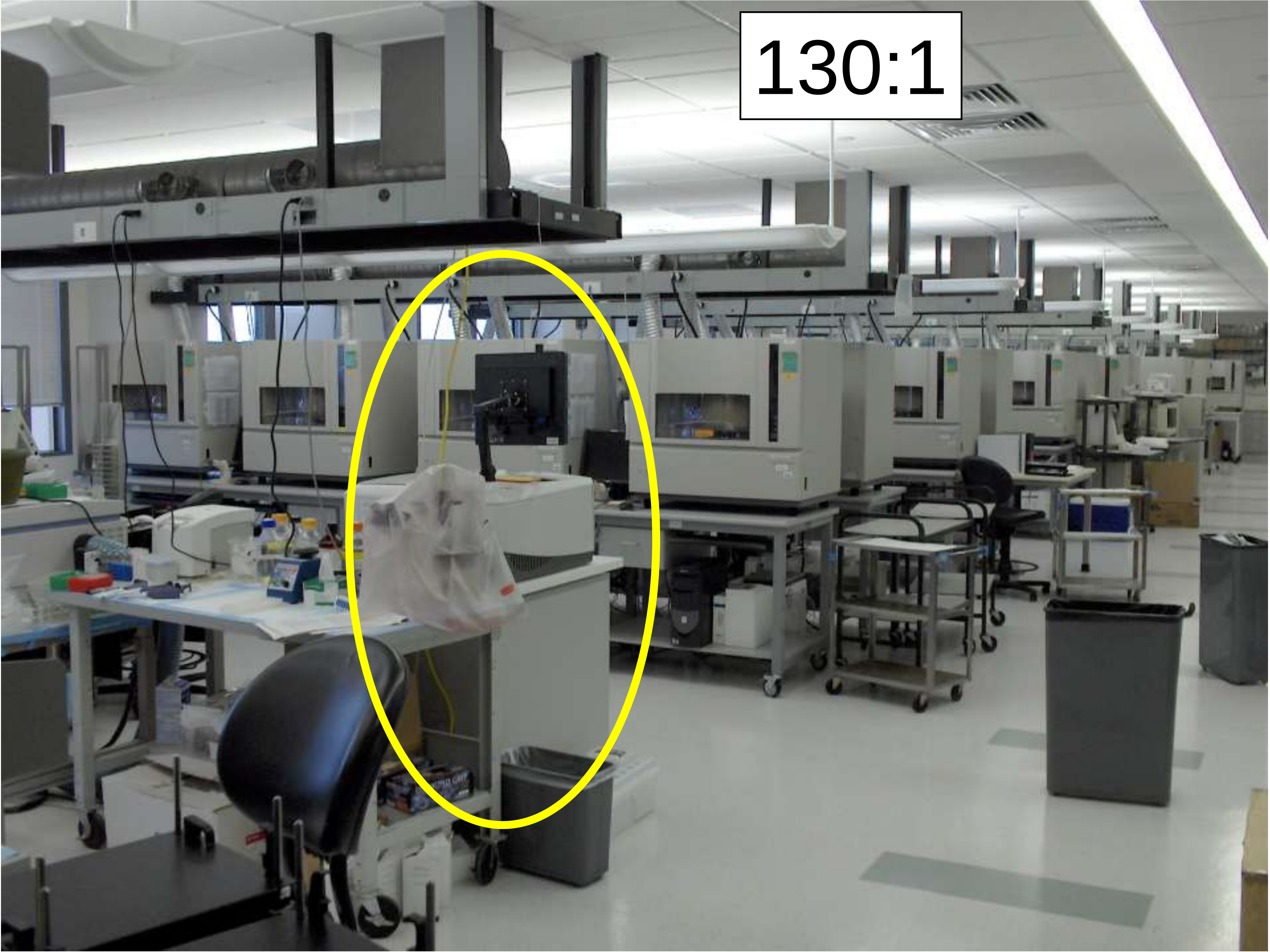
How did we get here?

- Advances in genome sequencing were driven by the Human Genome Project
  - Scale-up started in 1999
  - Resources concentrated in large genome centers
  - Increase in capacity
  - Reduction in cost
    - Economies of scale
    - Improved technology
- Sequencing infrastructure available for non-human projects

Genome Center Perspective

(George Weinstock, Wash-U/Baylor GSCs)

- Research is Data-driven
  - *Produce more data*
  - *Hypothesis generating > hypothesis testing*
  - *Community resource projects*
    - *Rapid data release; prepublication*
    - *Etiquette in use of prepublication data*
    - *No intellectual property contraints*
- Production is Technology-enabled
  - *Develop or acquire new technologies*



130:1

Human disease study

- 500 cases + 500 controls
- 500 genes, 15 exons/targets per gene
- 2 reads/target
- *15 million reads* to screen 1,000 subjects
- 454: 10M rds/d or Solexa: 160M rds/d
- Conclusion: *this is a small experiment*

Project Jim

- Whole human genome “Proof of Principle”
  - What can be learned from a single genome?
  - What biases exist in the data?
  - What analysis issues arise?
    - Not a consensus sequence but need to capture both alleles: 6 GB not 3 GB
    - Data quality vs variation: how do you know a variant base is a mutation and not an error

Conclusions: Post-genomic eScience

1. Generation of New High-Throughput Data (new “Genome Projects”)
2. Generating New Data in the Context of Existing Results (Published or in Databases)
3. No New Data – Exclusive re-use of “Published” Data