What are Genomics and Computational Genomics?

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

Genomics in the news

- **Fearing Punishment for Bad Genes by Kira Peikoff. New York Times, April 7, 2014.**
- **Out of Siberian Ice, a Virus Revived by Carl Zimmer. New York Times, March 3, 2014.**
- **A Powerful New Way to Edit DNA by Andrew Pollack. New York Times, March 3, 2014.**
- **The Mammoth Cometh by Nathaniel Rich. New York Times, Feb 27, 2014.**
- **Tracing Ancestry, Researchers Produce a Genetic Atlas of Human Mixing Events by Nicholas Wade. New York Times, Feb 13, 2014.**
- **A Catalog of Cancer Genes That's Done, or Just a Start by Carl Zimmer . New York Times, Feb 6, 2014.**
- **The $1,000 Genome Arrives - For Real, This Time by Matthew Herper. Forbes, January 14, 2014.**
- **Aiming to Push Genomics Forward in New Study by Andrew Pollack. New York Times, January 13, 2014.**
- **I Had My DNA Picture Taken, With Varying Results by Kira Peikoff. New York Times, December 30, 2013.**
- **Baffling 400,000-Year-Old Clue to Human Origins by Carl Zimmer. New York Times, December 4, 2013.**
- **FDA Approves New Gene-Sequencing Devices by Thomas M. Burton. The Wall Street Journal, November 20, 2013.**
- **Frederick Sanger, 95, Two-Time Winner of Nobel and Pioneer in Genetics, Dies by Denise Gellene. New York Times, November 20, 2013.**
- **Same Gene Mutations Tied to 12 Cancers by Ron Winslow. The Wall Street Journal, October 16, 2013.**
- **Biology's Big Problem: There's Too Much Data to Handle by Emily Singer. Wired, October 11, 2013.**
- **DNA Double Take by Carl Zimmer. New York Times, September 16, 2013.**
- **A Family Consents to a Medical Gift, 62 Years Later by Carl Zimmer. New York Times, August 7, 2013.**

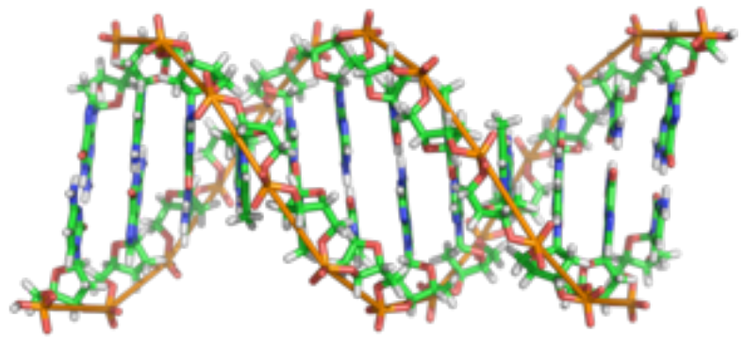
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Genome

"The complete set of genes or genetic material present in a cell or organism."

Oxford dictionaries

"Blueprint" or "recipe" of life



Self-copying store of read-only information about how to develop and maintain an organism

TAGCCCGACTTG

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Genomics

Oxford dictionaries

"The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of **genomes**."

where are the genes and other interesting bits?

how do sequences *change* over evolutionary time?

what does all the DNA *do*?

what are the physical shapes of the genome and its products?

Collins English Dictionary

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Genomics: contrast with biology & genetics*

* Everything on this slide is a gross generalization

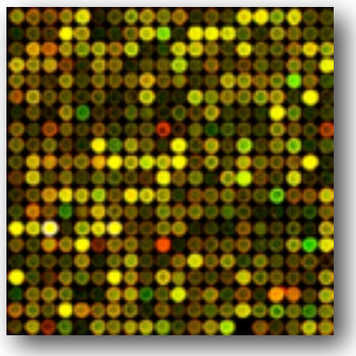
Biology & Genetics		**Genomics**
Targeted studies of one or a few genes	scope	Studies considering all genes in a genome
Targeted, low-throughput experiments	technology	Global, high-throughput experiments
Clever experimental design, painstaking experimentation	hard part	Tons of data, uncertainty, computation

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Genomics: shaped by technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

These provide very high-resolution snapshots of the world of nucleic acids (not just DNA)

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Genomics: tool for basic science

"The branch of molecular biology concerned with the **structure**, **function**, **evolution**, and **mapping** of genomes." Oxford dictionaries

Structure / mapping

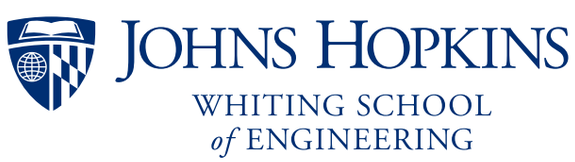
What is the DNA sequence of the genome?
Where are the genes?
What is the genome's three dimensional shape in the cell?

Function

What does all the DNA in the genome do?
What genes interact with what other genes?
How does the cell know what DNA is on/off?

Evolution

How did history shape our ethnicities and populations?
What big events shaped our current genetics?
Which portions of the genome are conserved by evolution?



Genomics: tool for medicine

"The branch of molecular genetics concerned with the study of genomes, specifically the identification and sequencing of their constituent genes and the **application** of this knowledge in **medicine**, **pharmacy**, **agriculture**, etc."

Collins English Dictionary

How is genotype related to health phenotypes?

What's the difference between DNA in a tumor vs DNA in healthy tissue?

Can genomic data help predict what drugs might be appropriate for:

- a particular cancer patient?
- a particular genetic disorder?

Can genomic data reveal weaknesses in the defenses of pathogens?

Can genomic data help us predict what flu strains will prevail next year?

Computational Genomics

Addresses crucial problems at the intersection of genomics and computer science

The intersection:

Key biological models are straight out of computer science: **circuits** and **networks** for molecular interactions, **trees** for evolution and pedigrees, **strings** for DNA, RNA and proteins

Thanks to sequencers and microarrays, research bottlenecks increasingly hinge on computational issues: **speed**, **scalability**, **energy**, **cost**

With large, noisy, biased high-throughput datasets comes a critical need for **machine learning** and **statistical reasoning**

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Computational Genomics: computation

How to efficiently analyze the huge quantities of fragmentary evidence that come from DNA sequencers

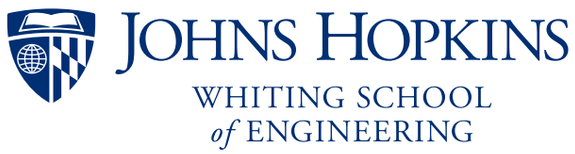
How to model biological phenomena and make predictions

How to combine data from disparate datasets to reach new conclusions in the presence of error and systematic bias

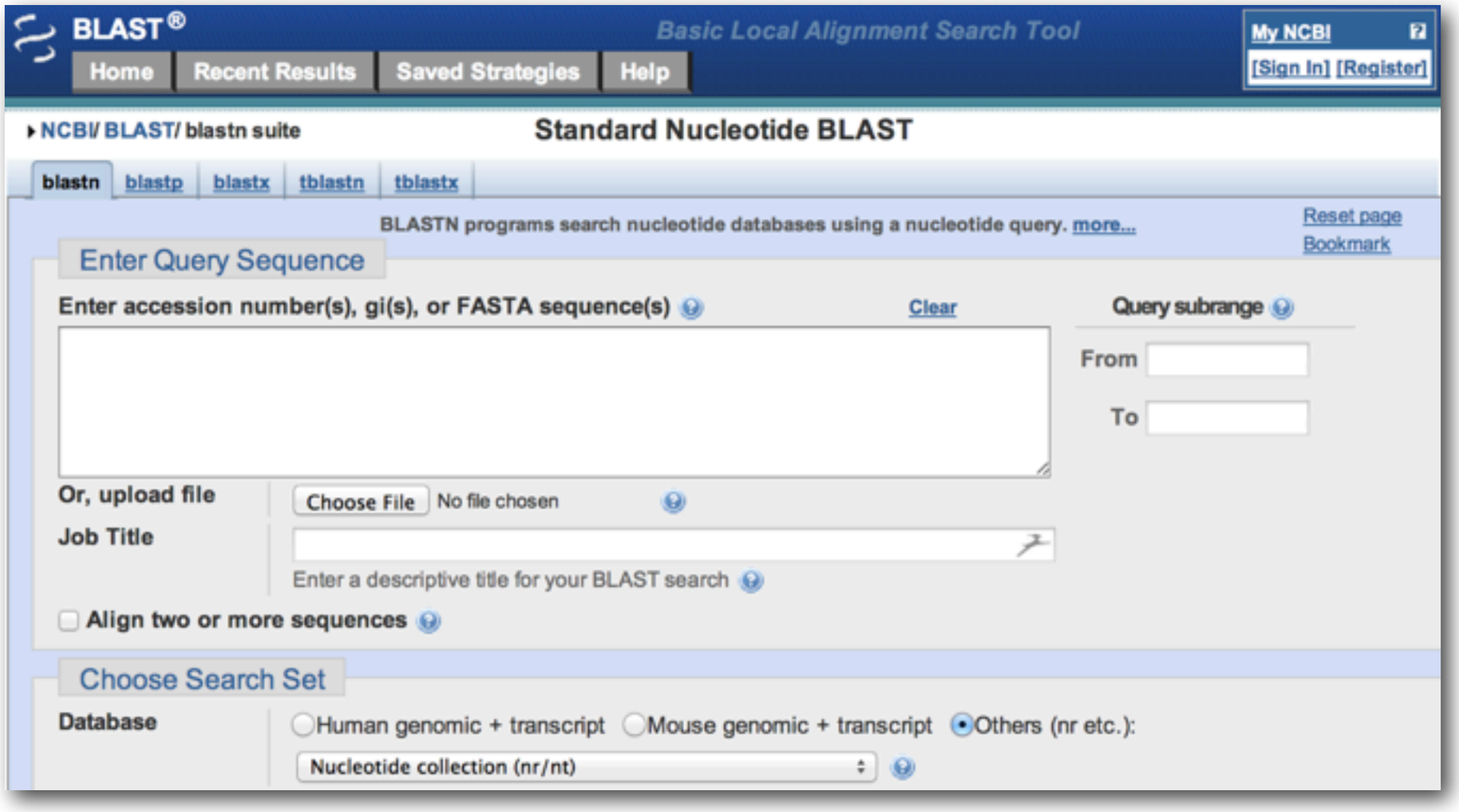
How to store huge quantities of data economically and securely while also allowing it to be queried

How to visualize large, complicated datasets

Draws on: Algorithms, data structures, pattern matching, indexing, compression, information retrieval, distributed and parallel computing, cloud computing, machine learning, ...



Computational Genomics: success stories



http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

The BLAST sequence alignment program is a hugely successful tool, a fixture of biological analysis and cited over 50,000 times

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Computational Genomics: success stories

The Human Genome Project depended crucially on contributions by computer scientists, especially new methods for assembling DNA fragments into chromosomes.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Computational Genomics: success stories

The idea of using high-throughput DNA sequencing in medical settings is only possible because of novel, extremely efficient software developed in the years after second-generation sequencers arrived.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Links

Past winners of the (Computational Biology) Overton Prize:
www.iscb.org/iscb-awards/overton-prize

Genomics in the popular press:
www.cs.jhu.edu/~langmea/poppress.shtml

The DNA Data Deluge (*behind paywall*):
ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6545119

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING