CMSC423: Bioinformatic databases, algorithms and tools

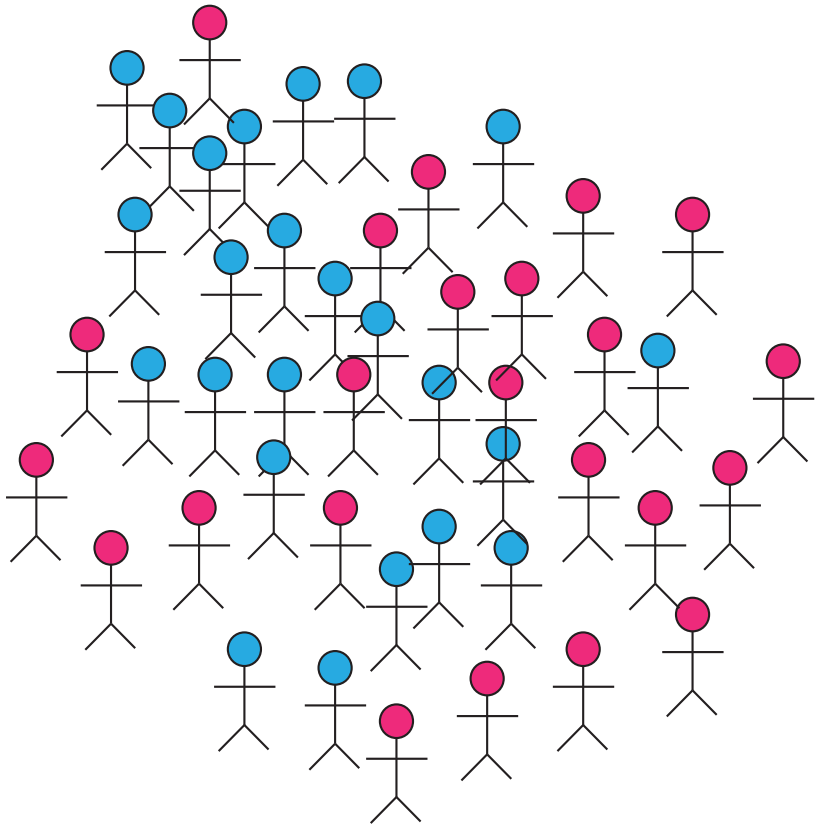
Héctor Corrada Bravo
Dept. of Computer Science
Center for Bioinformatics and Computational Biology
University of Maryland

University of Maryland, Fall 2013

Advances in Biology and Medicine needed, need, and will continue to need computational and statistical thinking (and their tools)

Héctor Corrada Bravo
Dept. of Computer Science
Center for Bioinformatics and Computational Biology
University of Maryland

What is Genomics?

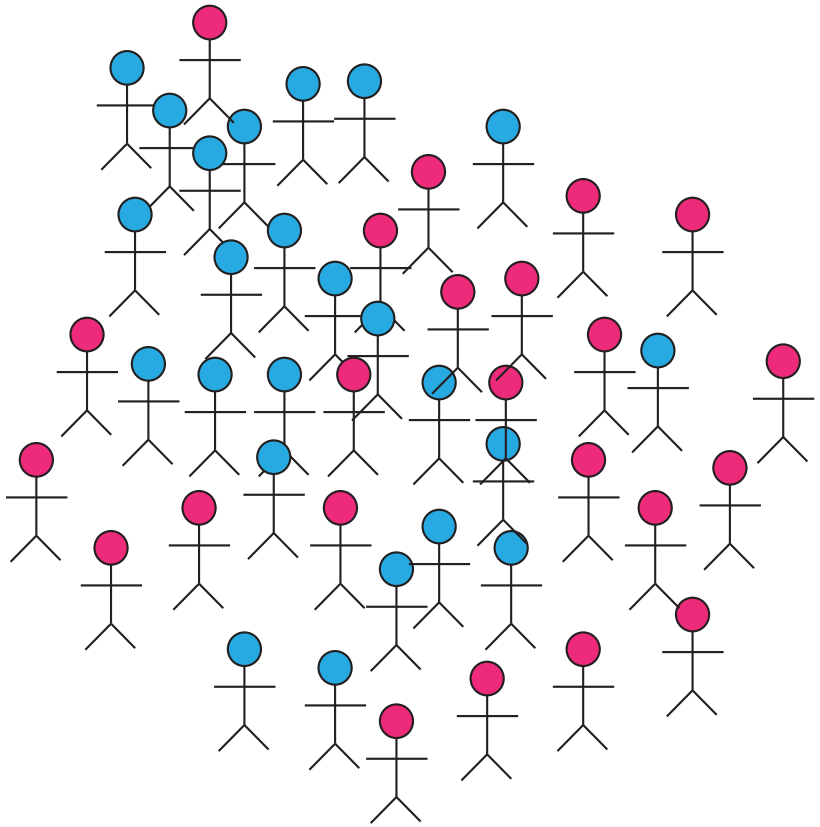


cancer

healthy

- Study the **molecular** basis of *variation* in development and disease
- Using **high-throughput** experimental methods
  - algorithms
  - ML
  - data management
  - modeling

What is Genomics?

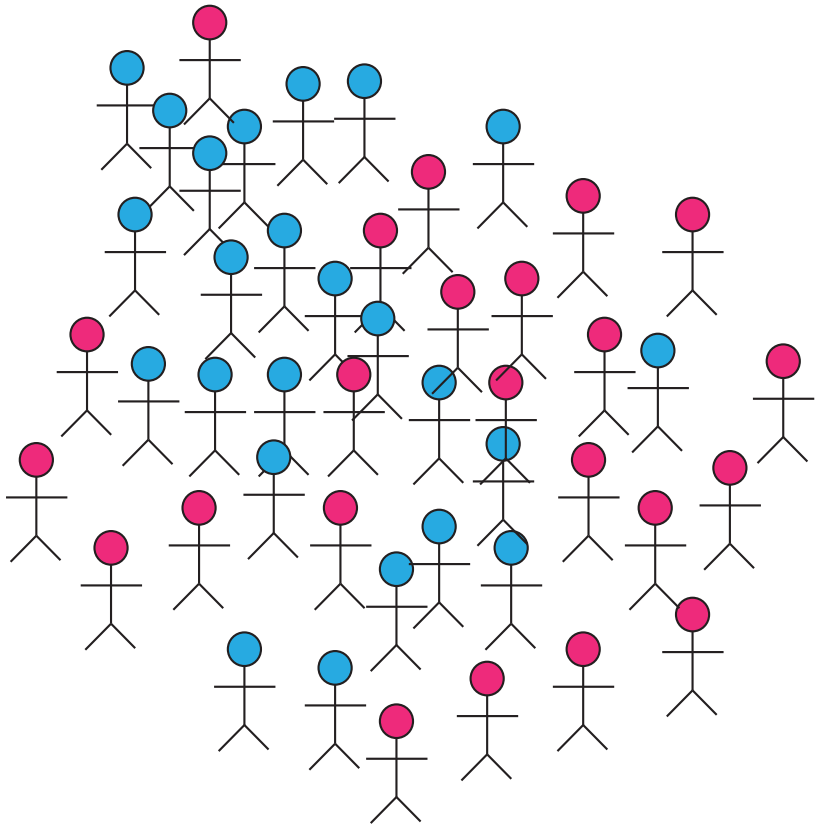


cancer

healthy

- Study the **molecular** basis of *variation* in development and disease
- Using **high-throughput** experimental methods
  - algorithms
    - String Algorithms
    - DP for string matching
    - Pattern-Finding

What is Genomics?



cancer

healthy

- Study the **molecular** basis of *variation* in development and disease
- Using **high-throughput** experimental methods
  - ML
    - clustering
    - classification
    - probabilistic methods


Why are my children such pigs?

What is Genomics?

- Each cell contains a complete copy of an organism's genome, or blueprint for all cellular structures and activities.

- The genome is distributed along chromosomes, which are made of compressed and entwined DNA.

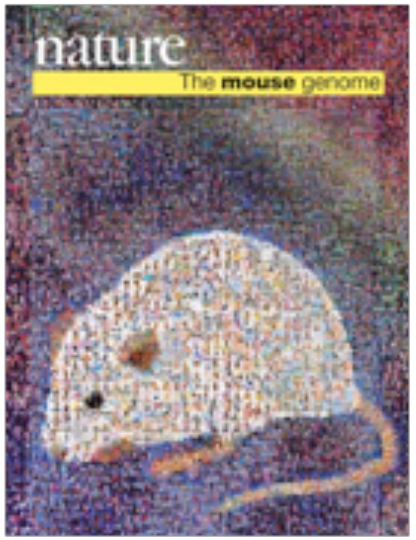
- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

Measurement

- For a small enough piece, we can measure the sequence of bases, referred to as *sequencing*
- Human Genome Project

D. melanogaster, *Science*, 2000

H. sapiens, *Nature*, 2000 and *Science*, 2000

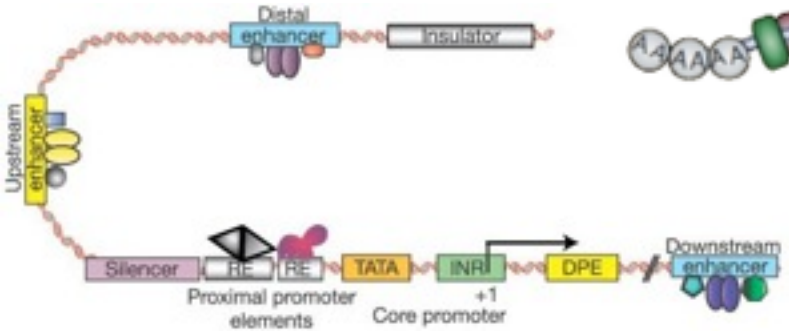


M. musculus, *Nature*, 2002

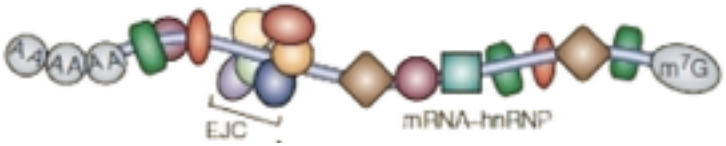
Computational Biology

Genes encode proteins which are transcribed into mRNA and translated into proteins.

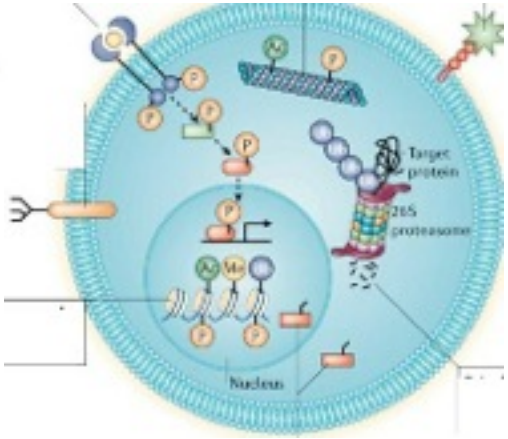
genomics



transcriptomics



proteomics



Major technological advances allow **unprecedented** data acquisition

What you missed

- Scratched surface of population studies
  - gene expression
  - epigenetics
  - genetics

What you missed

- Metagenomics
  - Microbial sampling
  - Clinical association to disease

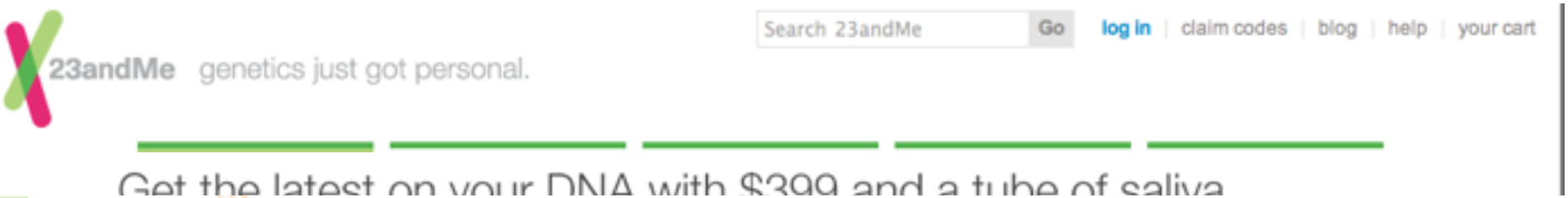
What you missed

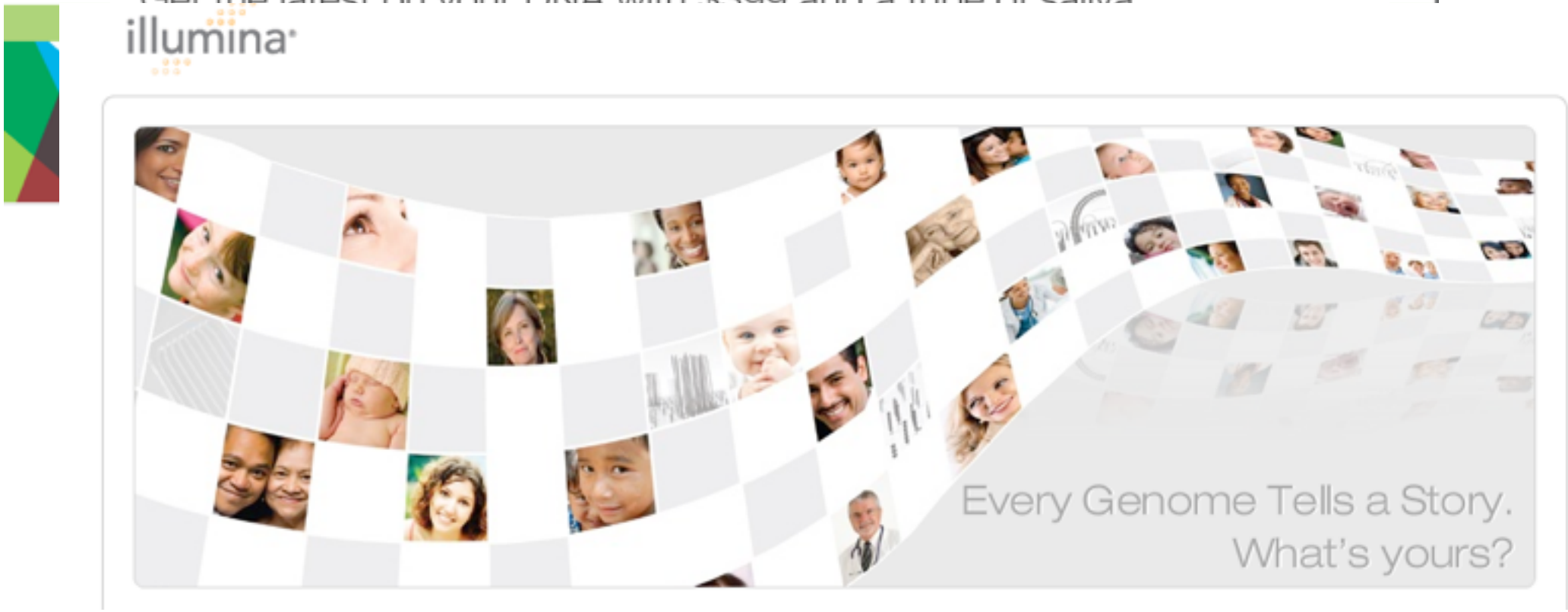
- Networks, Networks, Networks
  - Gene regulatory networks
  - Protein-Protein interaction networks

What you missed

- Proteomics
  - Becoming high-throughput field
  - Awesome new experiments revealing all sorts of new insight
  - Check out Zia's grad seminar next semester

Personal Genomics





Sequence Once
Read Often

Read what?

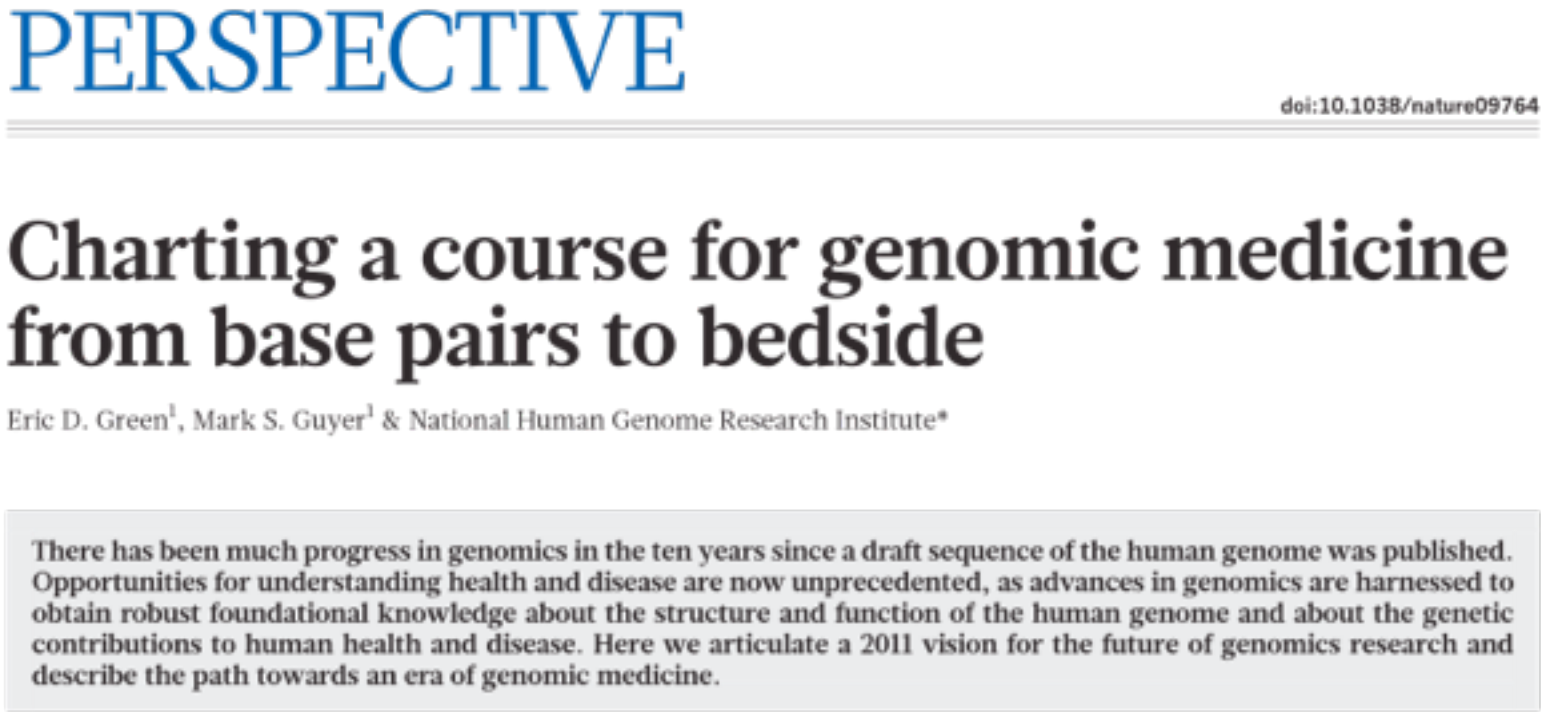
- genome
- variants
- methylation
- expression
- other genome features
- medical literature
- risk models
- population information
- ...

Personal Genomics

- We need to produce reliable genome measurements, but on much bigger scale (Algorithmics, Systems)
- Multiple genome features, decide which are relevant and significant (Information Retrieval, Data Management)
- Population-based science, interpreted individually (Machine Learning/Statistics, Privacy)

NHGRI strategic plan

- What does the NIH think genomics should be for the next 10 years?



PERSPECTIVE

doi:10.1038/nature09764

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green[1], Mark S. Guyer[1] & National Human Genome Research Institute*

There has been much progress in genomics in the ten years since a draft sequence of the human genome was published. Opportunities for understanding health and disease are now unprecedented, as advances in genomics are harnessed to obtain robust foundational knowledge about the structure and function of the human genome and about the genetic contributions to human health and disease. Here we articulate a 2011 vision for the future of genomics research and describe the path towards an era of genomic medicine.

[Nature, Feb. 2011]

Where do we fit in?

- The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges.

- **Data analysis**
  - Computational tools are quickly becoming inadequate for analysing the amount of genomic data that can now be generated, and this mismatch will worsen. Innovative approaches to analysis, involving close coupling with data production, are essential.

- **Data integration**
  - Genomics projects increasingly produce disparate data types (for example, molecular, phenotypic, environmental and clinical), so computational approaches must not only keep pace with the volume of genomic data, but also their complexity. New integrative methods for analysis and for building predictive models are needed.

- **Visualization**
  - In the past, visualizing genomic data involved indexing to the one-dimensional representation of a genome. New visualization tools will need to accommodate the multidimensional data from studies of molecular phenotypes in different cells and tissues, physiological states and developmental time. Such tools must also incorporate non-molecular data, such as phenotypes and environmental exposures. The new tools will need to accommodate the scale of the data to deliver information rapidly and efficiently.

- **Computational tools and infrastructure**
  - Generally applicable tools are needed in the form of robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists. Adequate computational infrastructure is also needed, including sufficient storage and processing capacity to accommodate and analyse large, complex data sets (including metadata) deposited in stable and accessible repositories, and to provide consolidated views of many data types, all within a framework that addresses privacy concerns. Ideally, multiple solutions should be developed[105.]

Where do we fit in?

- Meeting the computational challenges for genomics requires scientists with expertise in biology as well as in informatics, computer science, mathematics, statistics and/or engineering.

- *A new generation of investigators who are proficient in two or more of these fields must be trained and supported.*