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, Spring 2017

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

Why are my children such pigs?

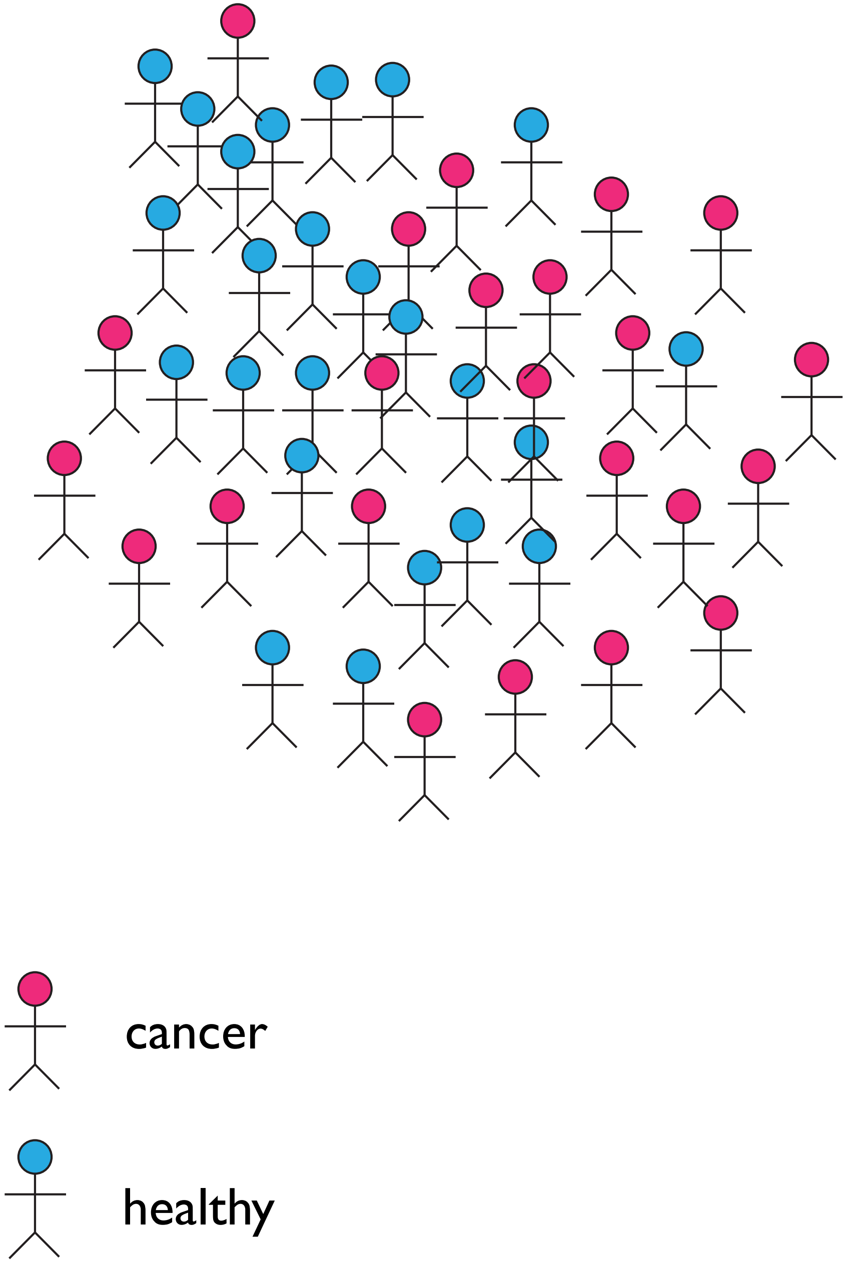
What is Genomics?

- 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.

What is Genomics?



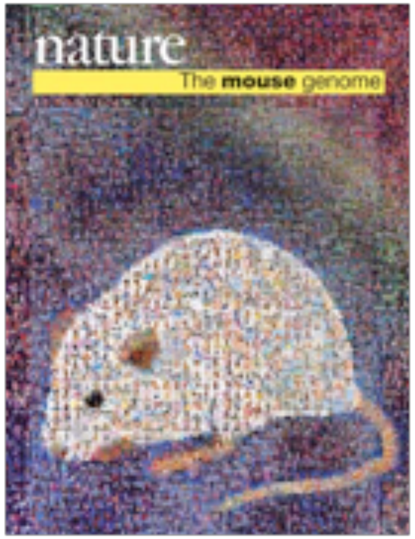
- Study the **molecular** basis of *variation* in development and disease
- Using **high-throughput** experimental methods
  - algorithms
  - ML
  - data management
  - modeling

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

Genome

```
TCAGTTGGAGCTGCTCCCCCACGGCCTCTCCTCACATTCCACGTCCTGTAGCTCTATGACCTCCACCTTTGAGTCCCTCCTCTCACACCTGAC
ATGAAAAGGCACATGAGGATCCTCAAATACCCCGTGATCAGTCTCAGGGTAGCTCTCATAGCCTGGACAGGGCCCCCCTCGGGGGTTGCGCCC
AGGTCCAGGCGGGGGATGCACAGCAACAGTCACCGAAGCAGAAGCCGTCACAGTGGTGATGGGCTGGCAGTAGCTGGGCACAGAGCTGCCCAT
GGCGGTGGACGTTGGGTTCCGAGGGTTGTGAGAACGGGCCCCACGGGGCCCTGAGCGGTCCCTATTGCTAGGGCCAGAATGCCCTTCAGTAGA
AATTTCAAAAGCGTCTCTGCGCGGTCTGTAGGGGGGTGGCCGCAAGCCTTCTCTAGGGGGATCCCTTCGAGGCTGCTGGCCTTGCCGTCCAGG
GGACAAGGAGCCAGAGTCCAGGTGGGGCTGTTGCCGAGGGGTCAAGGGAGGCTGATGTCTGGAGTCCGGATGGACCACCTGCAGAGGAGAGAC
ATAGGTCAACACAGGGAGGTAGGATGGTGGTGATGTTCCACCCACAAAAGAAAACCTATTCCTTTAGAAACCTCCAGGATGTGAATCCTGCCT
GCACCTGCACAGCTGGCTGGAGGCATATAGCCACTGCCCATAGATCTCAACTTACCCTCACAACCAACTGCCCCCAGGCCTAAGTTCTCTGCC
TCAAAACTGCCAAGGCCTGGATAGCCAAGAGCCTGGGTGTCTTGGAAATATGCAACCATAAATAGTAGCTTTTAGAAGTATAAGGCTCCTGTT
TCTGGGTCATATTAGTGTTGTTTTCACCTGTCCCCAGCCCTAAGCCAGGTGTGGCCAGAAGCAAATGTACTGTAAGAGCAGAGCAAAAACTTC
CACACAGATAGTTCTGTTAGGCAATACATCTCTGCCTGACTATTAGGAATCTGGTTTCTGGGTCCTCTGTACAAAGCTCGGAGCAACACAGTG
GCCACATCAATCAAAAGGACCGTGACCAACTTCAAAGTCGGTGAGCTTGTACCTATTTTTAGGCTCCTGCTGAACAGAACCAGATTCACACTA
CAGCTCAGCAGGGCATCGTCACGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGGGGGGGGGGGGTGGACAGAGGACGGGGAC
ACAATTCACTGGCCAGCCCTTCTCTCCTTCAAGGAAGGCTGCTCTAGCCTGGGACTGGAATACACATTTCCTGTAAACATGGTGGGGGCCTCA
GGCAAGCCAGAGTTTTGGAGCCTTCCTTAACTCTTCAAGGTGAGCATCTTGACTTGGAGGGTGGGGGTGCGGGTAAGGAAGGAACCTGTGGAC
TCCTCCCTACAAGACAGAAAAGGAATAAGCCACGAAGACAATAACGATTTTTGTATCAAGCGTCCTCTCCCATTTCAGCTTACCTGACAATGA
AATCAAATTCGGACCCTGCAAGCATCAGTACACCCAGCAGAGTGGACACAGCACCGTCCAGAACGGGAGCAAACATGTGCTCCAGAGCGAGCA
TAGCCCTGTGGTTCTTGTCCCCAATGGCTGTCAGAAAGGCCTGAACAAAGGAGAAAATTGACACGGTCACATTCTGGGTGTGGTAAAGTGCTC
AGCTGTGTCTATACTTGGGTTTTGTAT…
```

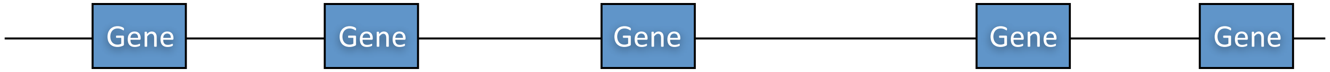
Total amount of DNA in human genome:
3 * 10^9 base pairs (bp)

Why are these two different?

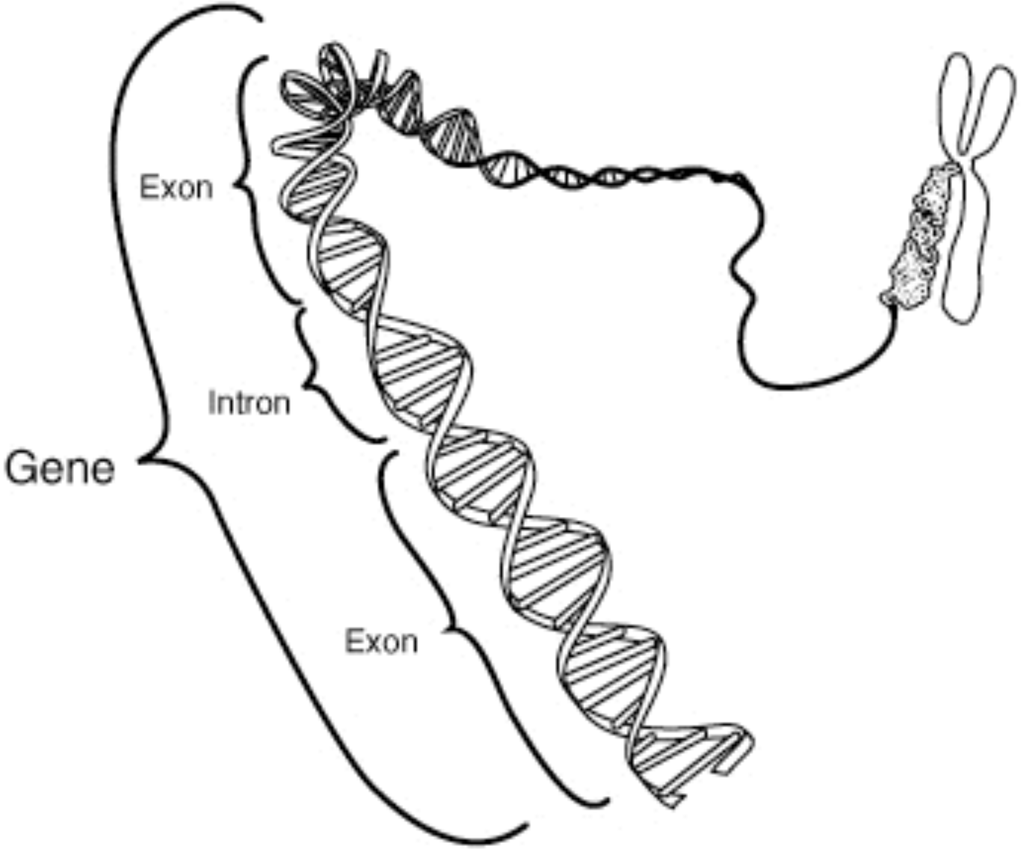
Differences explained by 1-10% difference in genome

Similarities explained by similar genes

Genes

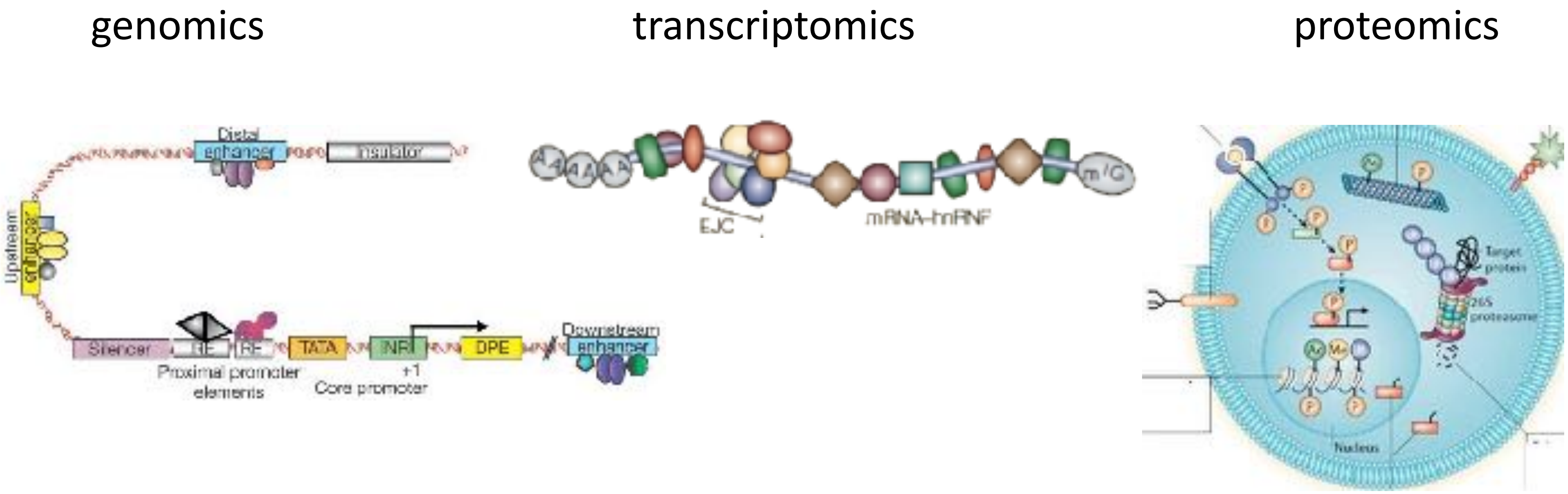


Gene Gene Gene Gene Gene



Computational Biology

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

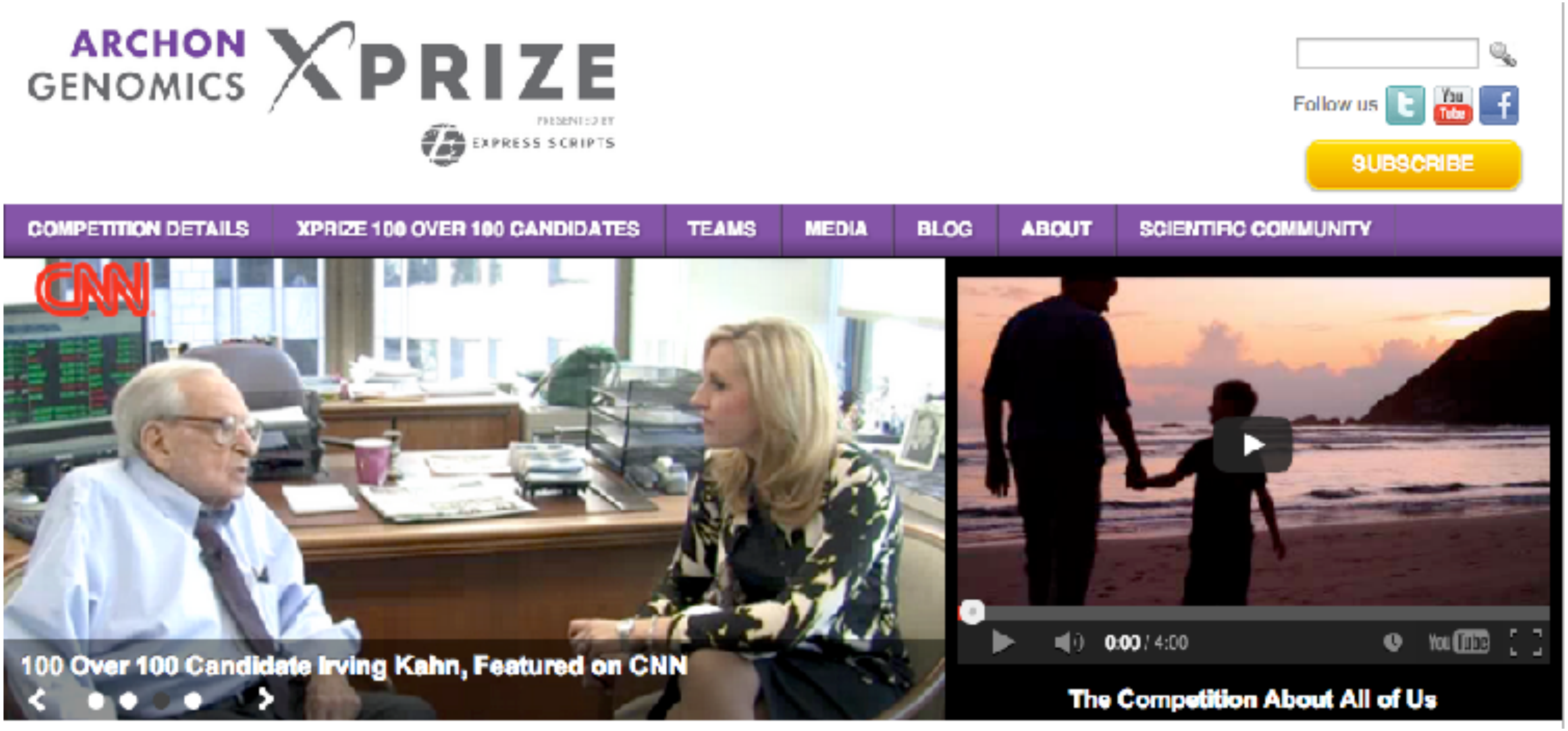


genomics

transcriptomics

proteomics

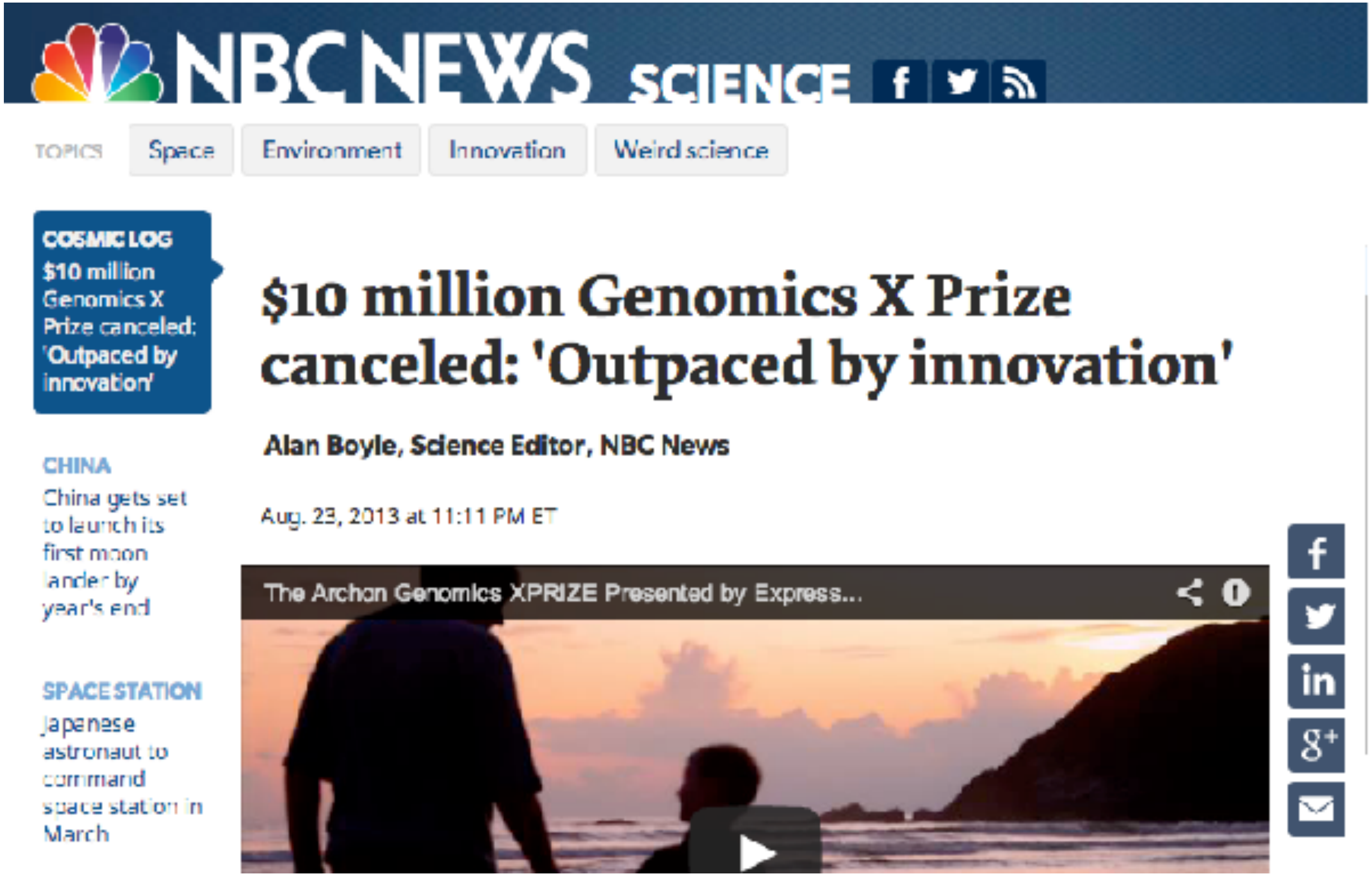
Major technological advances allow **unprecedented** data acquisition



100 Over 100 Candidate Irving Kahn, Featured on CNN

The Competition About All of Us

build a **whole human genome** sequencing device and use it to sequence **100 human genomes** within **30 days or less**, with an accuracy of no more than one error in every 1,000,000 bases sequenced, with an accuracy rate of at least 98% of the genome, and at a recurring cost of **no more than $1,000 (US) per genome**.



NBC NEWS SCIENCE

COSMIC LOG
$10 million Genomics X Prize canceled: 'Outpaced by innovation'

$10 million Genomics X Prize canceled: 'Outpaced by innovation'

Alan Boyle, Science Editor, NBC News

Aug. 23, 2013 at 11:11 PM ET

The Archon Genomics XPRIZE Presented by Express...

CHINA
China gets set to launch its first moon lander by year's end

SPACE STATION
Japanese astronaut to command space station in March

“genome sequencing technology is plummeting in cost and increasing in speed independent of our competition”

> “companies can do this for less than $5,000 per genome, in a few days or less — and are moving quickly towards the goals we set for the prize.”

What makes them different?

Much human variation is due to difference in ~ 6 million base pairs (0.1 % of genome) referred to as SNPs

Single Nucleotide Polymorphism (SNP)

SNP

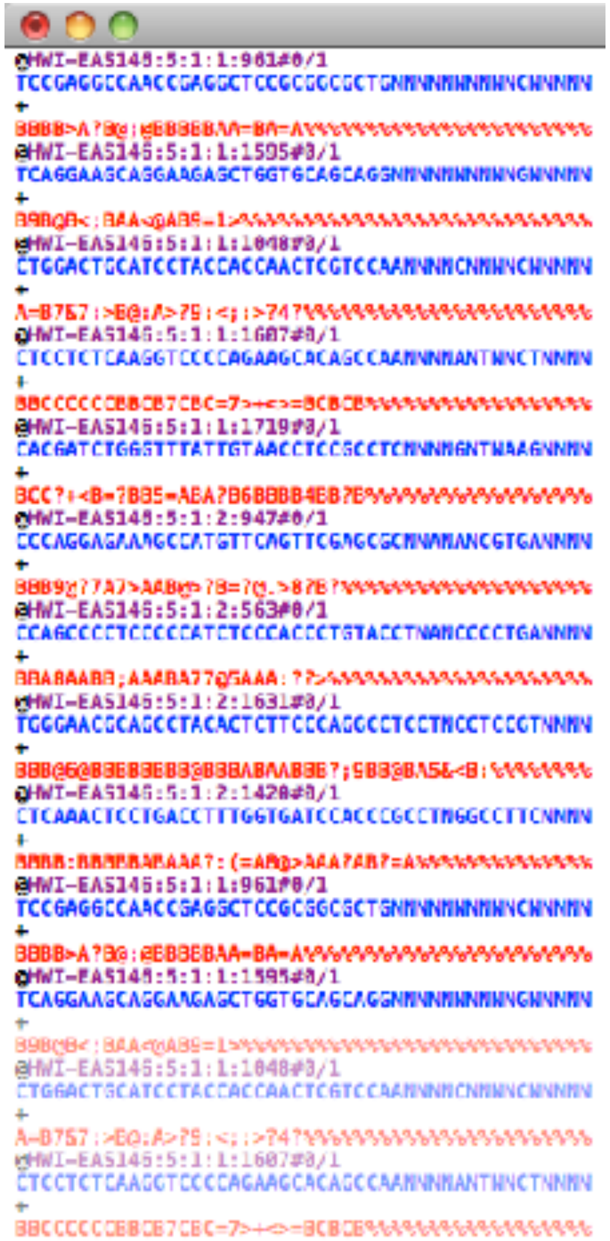
Genomic DNA:

A

TACATAGCCATCGGTANGTACTCAATGATGATA

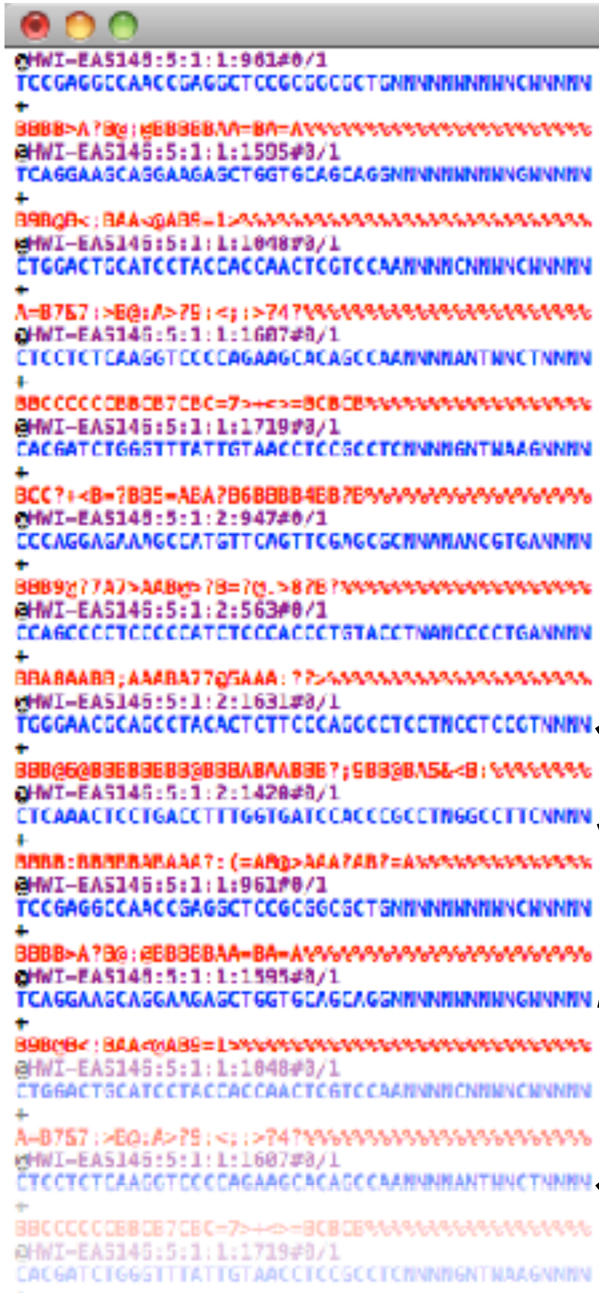
G

From reads to evidence



```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNN
+
BBBB>A?B@;@BBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNN
+
A=B7E7;>B@:A>?9;<;;>?4?%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAGTTCGAGCGCNNANANCGTGANNNN
+
BBB9@?7A7>AAB@>?B=?@.>87B?%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCTCCCCCATCTCCCACCCTGTACCTNANCCCCTGANNNN
+
BBABAABB;AAABA77@5AAA:??>%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTCTTCCCAGGCCTCCTNCCTCCGTNNNN
+
BBB@6@BBEBBEBB@BBBABAABBE?;5BB@BA5&<B:%%%%%%%%
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTCNNNN
+
BBBB:BBBBBABAAA?:(=AB@>AAA?AB?=A%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNN
+
BBBB>A?B@;@BBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNN
+
A=B7E7;>B@:A>?9;<;;>?4?%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%
```

From reads to evidence



```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNNN
+
A=B7B7;>B@;A>?9;<;;>?4?%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
BCC?;<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAANGCCATGTTCAGTTCGAGCGCNNANNANCGTGANNNN
+
BBB9@?7A?>AAB@>?B=?@.>8?B?%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCTCCCCCATCTCCCACCCTGTACCTNANCCCCTGANNNN
+
BBABAABB;AAABA77@5AAA:??>%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTCTTCCCAGGCCTCCTNCCTCCGTNNNN
+
BBB@6@B3EB3EB3@B3BABAABBE?;5BB@BA5&<B:%%%%%%%%
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTCNNNN
+
BBBB:BBBBBABAAA?:(=AB@>AAA?AB?=A%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@BBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNNN
+
A=B7B7;>B@;A>?9;<;;>?4?%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTNAAGNNNN
+
```

1. Comparative

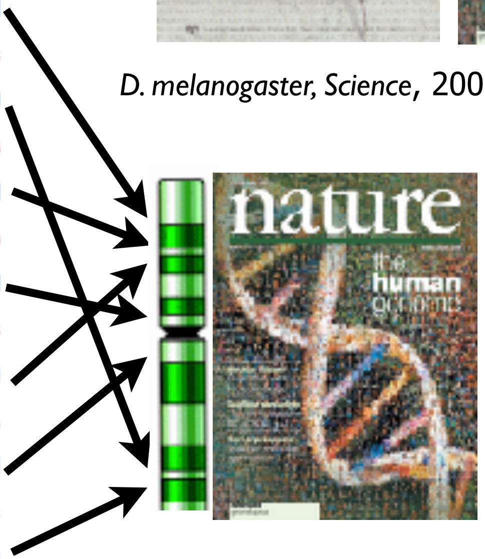
Sequence-wise, individuals of a species are nearly identical

Well curated, annotated "reference" genomes exist

D. melanogaster, Science, 2000

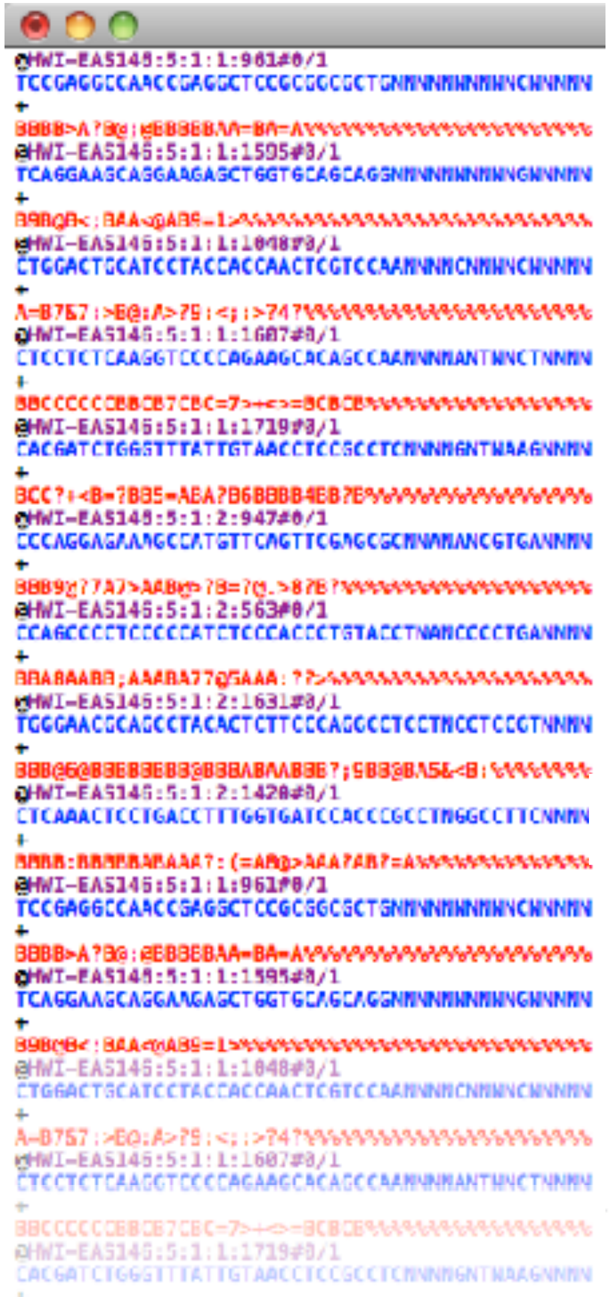
H. sapiens, Nature, 2000 and *Science*, 2000

M. musculus, Nature, 2002



Idea: "Map" reads to their point of origin with respect to a reference, then study differences

From reads to evidence



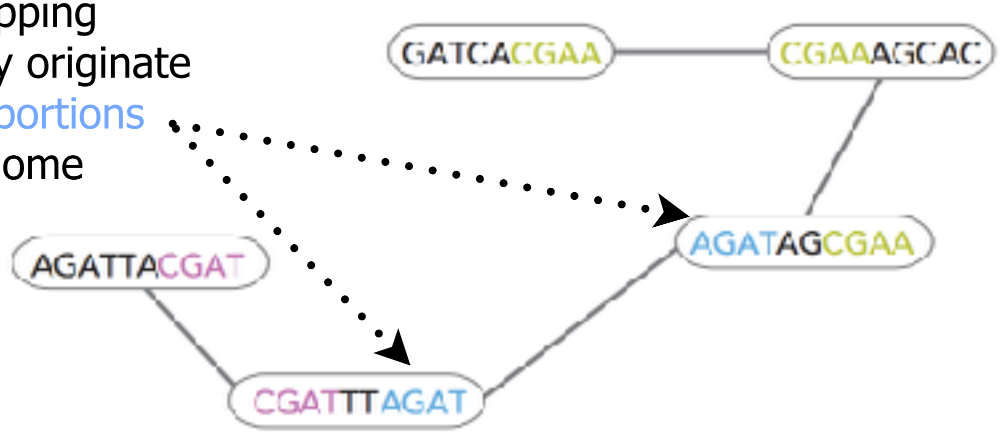
```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@EBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNNN
+
A=B7B7;>B@:A>?9;<;;>74?%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAANGCCATGTTCAGTTCGAGCGCNNANNANCGTGANNNN
+
BBB9@??A?>AAB@>?B=?@.>8?B?%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCTCCCCCATCTCCCACCCTGTACCTNANCCCCTGANNNN
+
BBABAABB;AAABA77@5AAA:??>%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTCTTCCCAGGCCTCCTNCCTCCGTNNNN
+
BBB@6@BBBBBBBB@BBBABAABBB?;5BB@BA5&<B:%%%%%%%%%
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACTCCTGACCTTTGGTGATCCACCCCGCCTNGGCCTTCNNNN
+
BBBB:BBBBBABAAA?:(=AB@>AAA?AB?=A%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNCNNNNN
+
BBBB>A?B@;@EBBEBAA=BA=A%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNCNNNNCNNNNN
+
A=B7B7;>B@:A>?9;<;;>74?%%%%%%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCCAGAAGCACAGCCAANNNNNANTNNCTNNNN
+
BBCCCCCCBBCB7CBC=7>+<>=BCBCB%%%%%%%%%%%%%%%%%%%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNNGNTNAAGNNNN
+
```

2. *de novo*

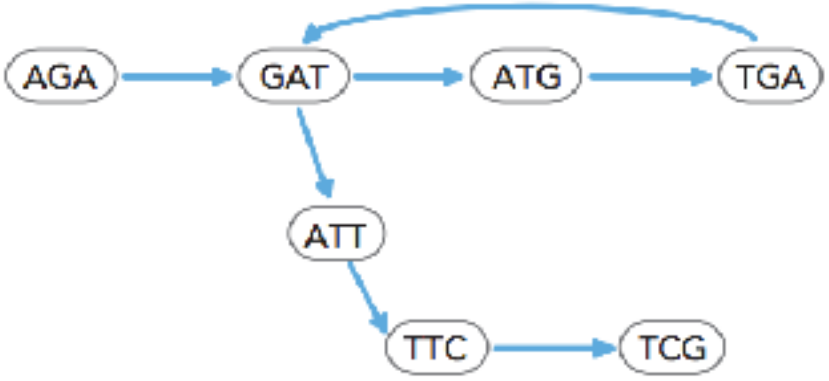
Assume nothing! - let reads tell us everything

Reads with overlapping sequence probably originate from overlapping portions of the subject genome

Encode overlap relationships as a graph



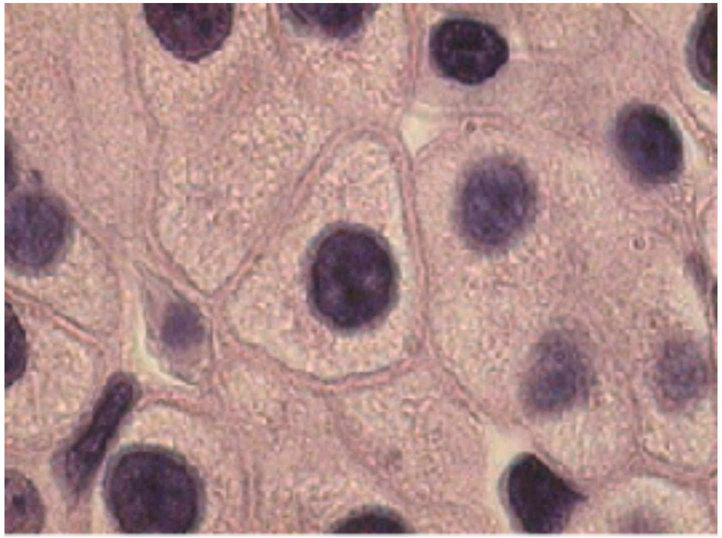
Source: De Novo Assembly Using Illumina Reads. Illumina. 2010

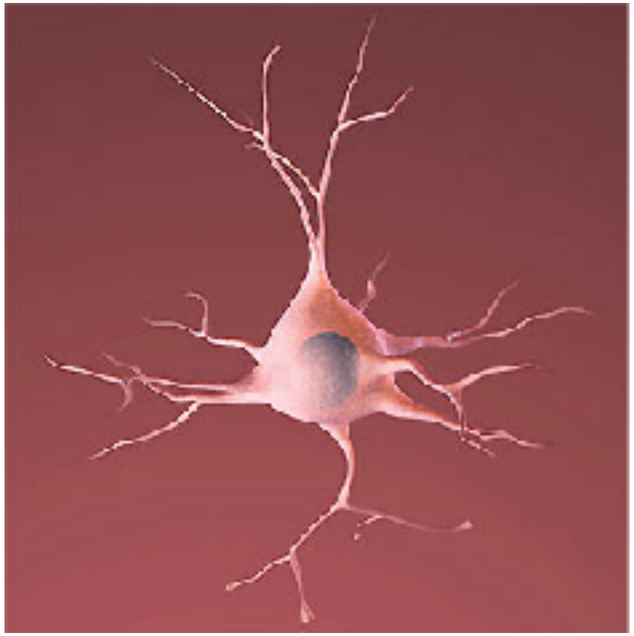


The full genome sequence is a "tour" of the graph

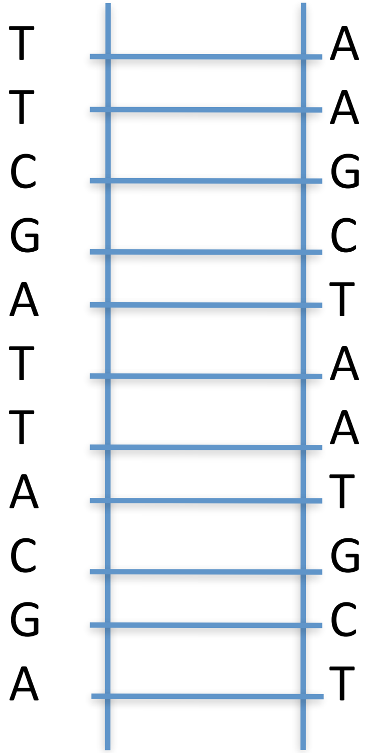
Source: De Novo Assembly Using Illumina Reads. Illumina. 2010
http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf

How many basepair differences?

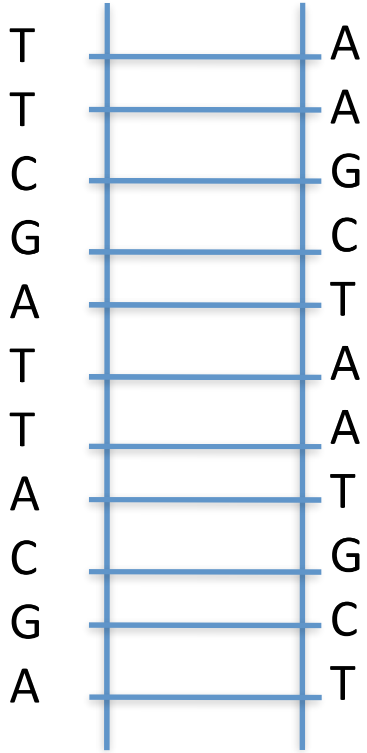




Liver



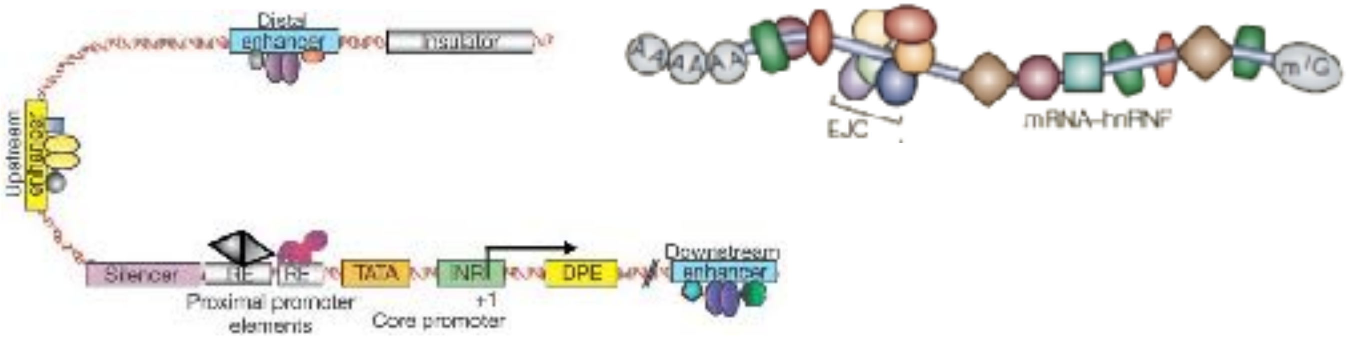
Brain

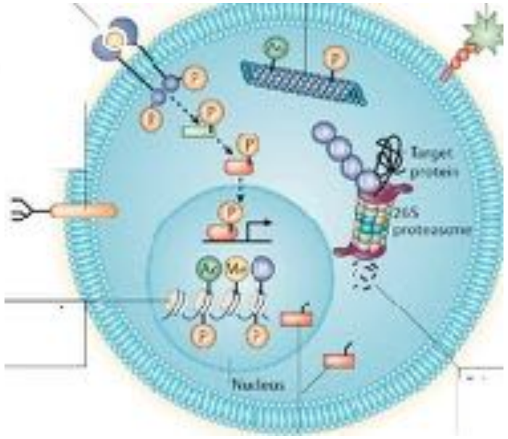


Computational Biology

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

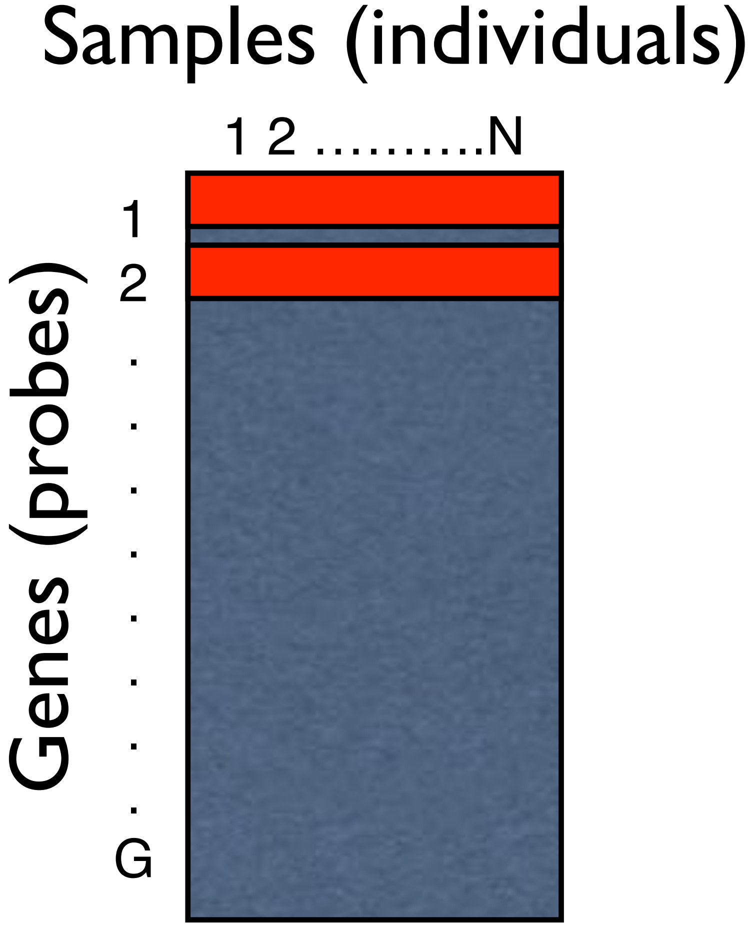
genomics transcriptomics proteomics



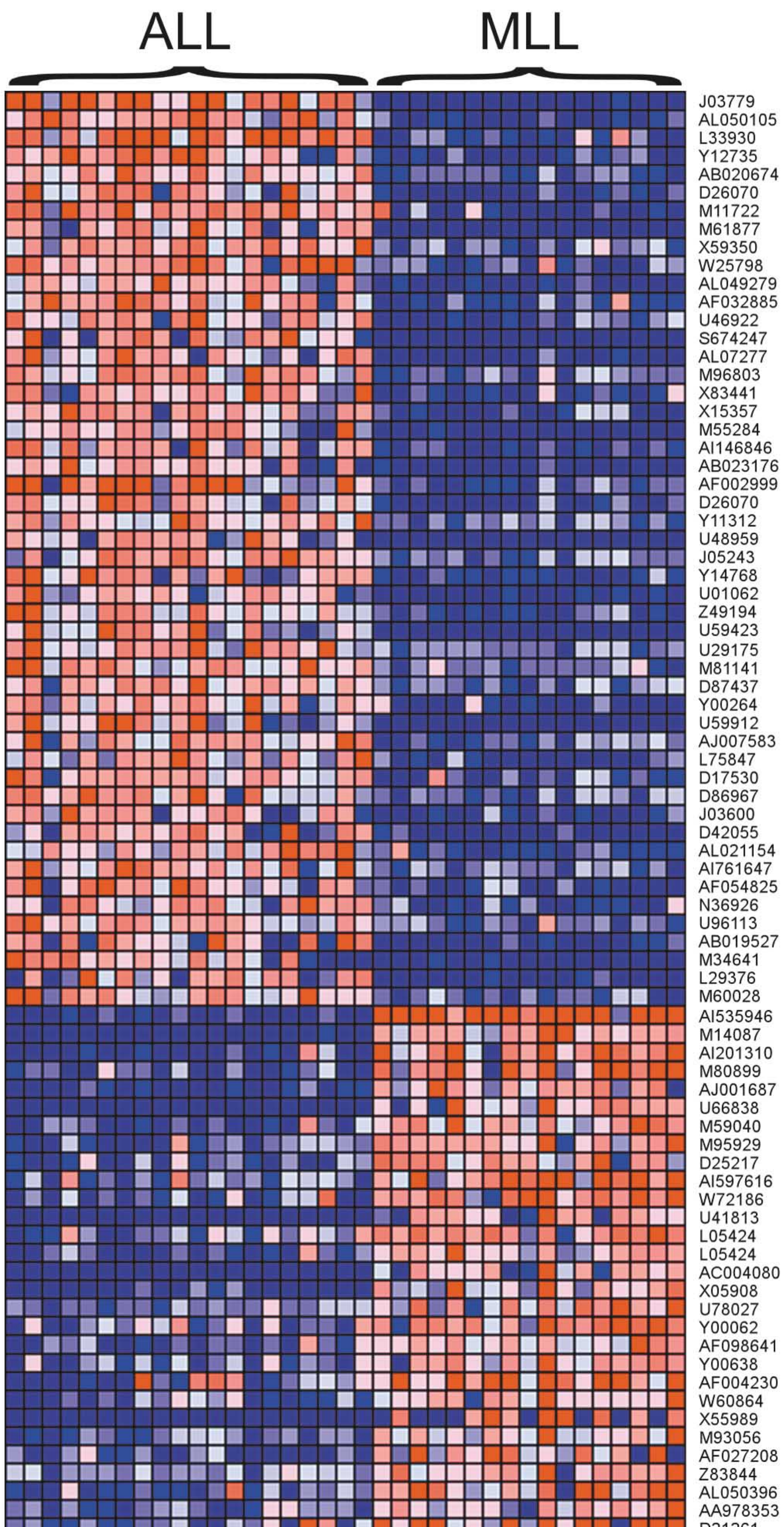


Major technological advances allow **unprecedented** data acquisition

Measurements



DATA MATRIX



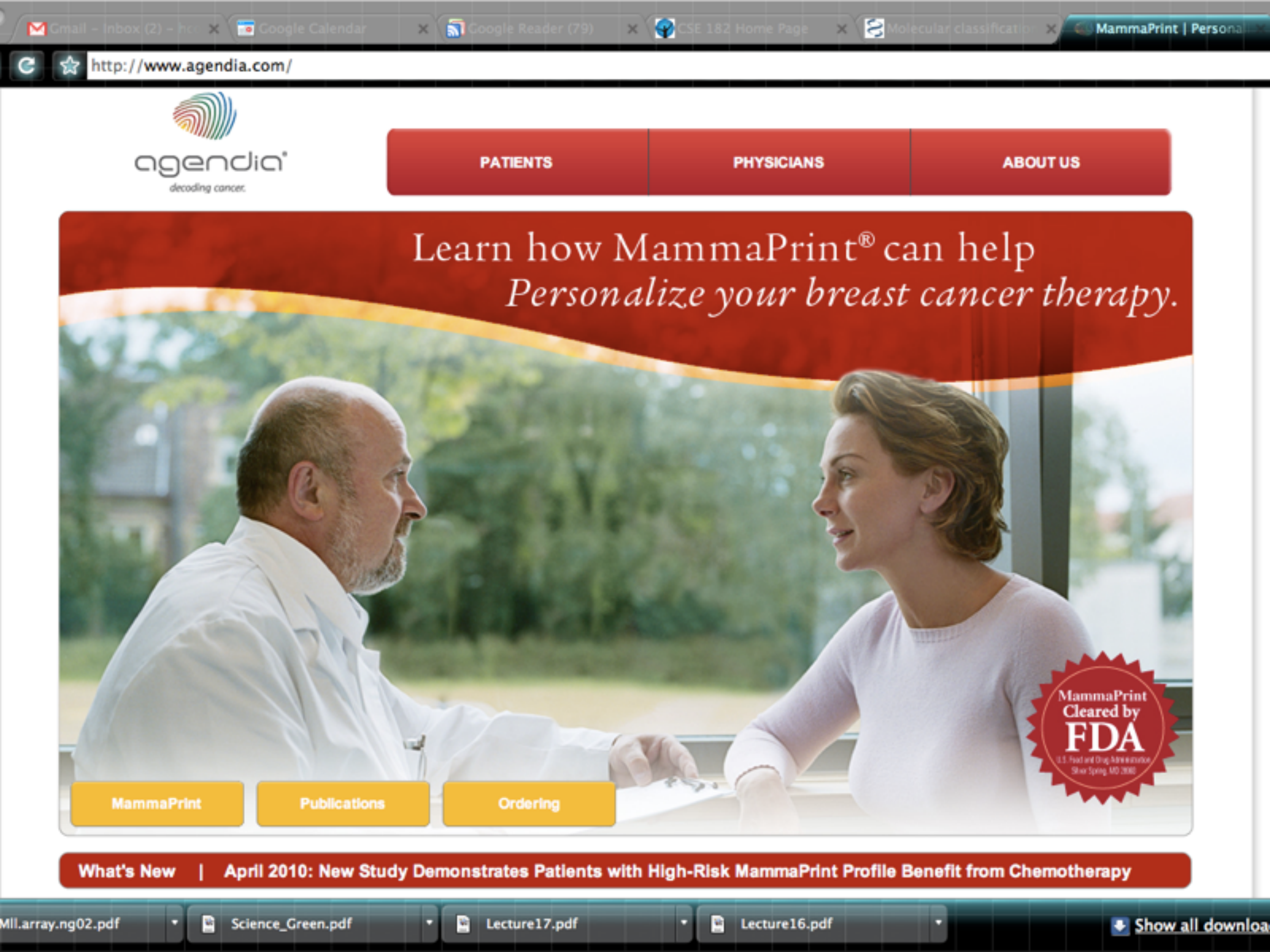
article

MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia

Scott A. Armstrong[1–4], Jane E. Staunton[5], Lewis B. Silverman[1,3,4], Rob Pieters[6], Monique L. den Boer[6], Mark D. Minden[7], Stephen E. Sallan[1,3,4], Eric S. Lander[5], Todd R. Golub[1,3,4,5]* & Stanley J. Korsmeyer[2,4,8]*

**These authors contributed equally to this work.*

Published online: 3 December 2001, DOI: 10.1038/ng765



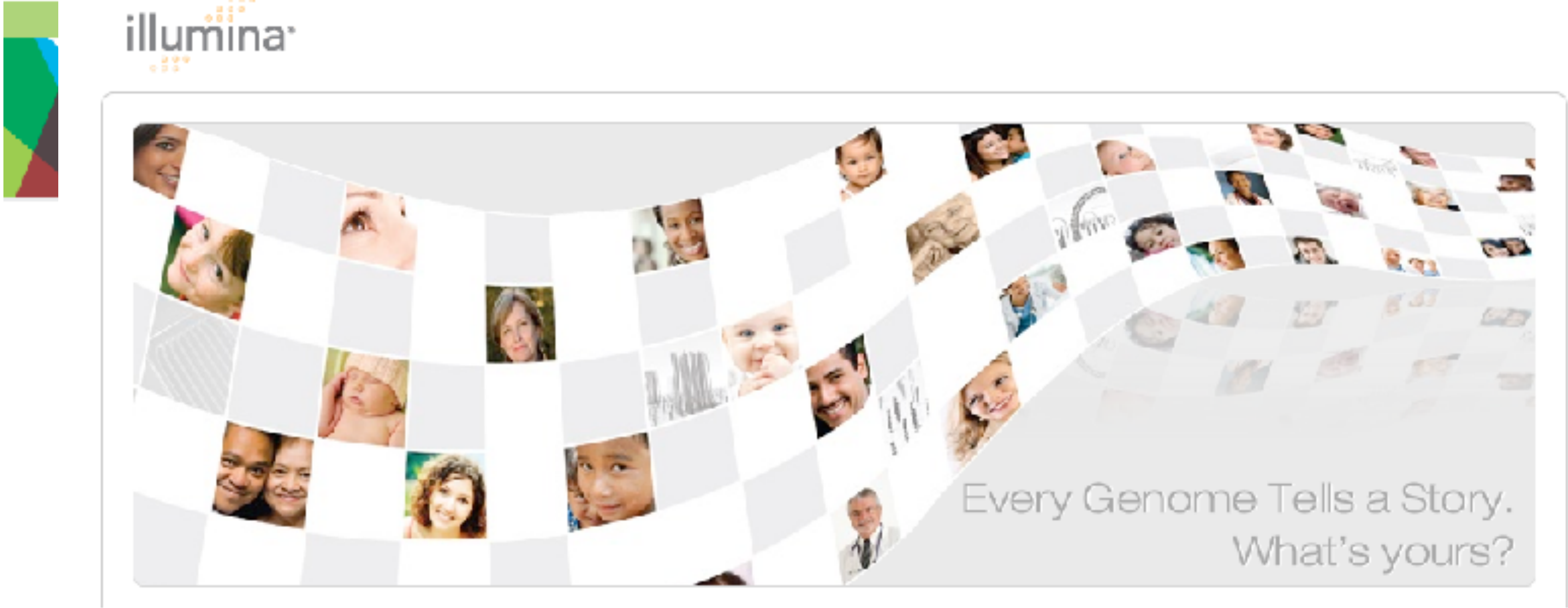
Learn how MammaPrint® can help
Personalize your breast cancer therapy.

What's New | April 2010: New Study Demonstrates Patients with High-Risk MammaPrint Profile Benefit from Chemotherapy

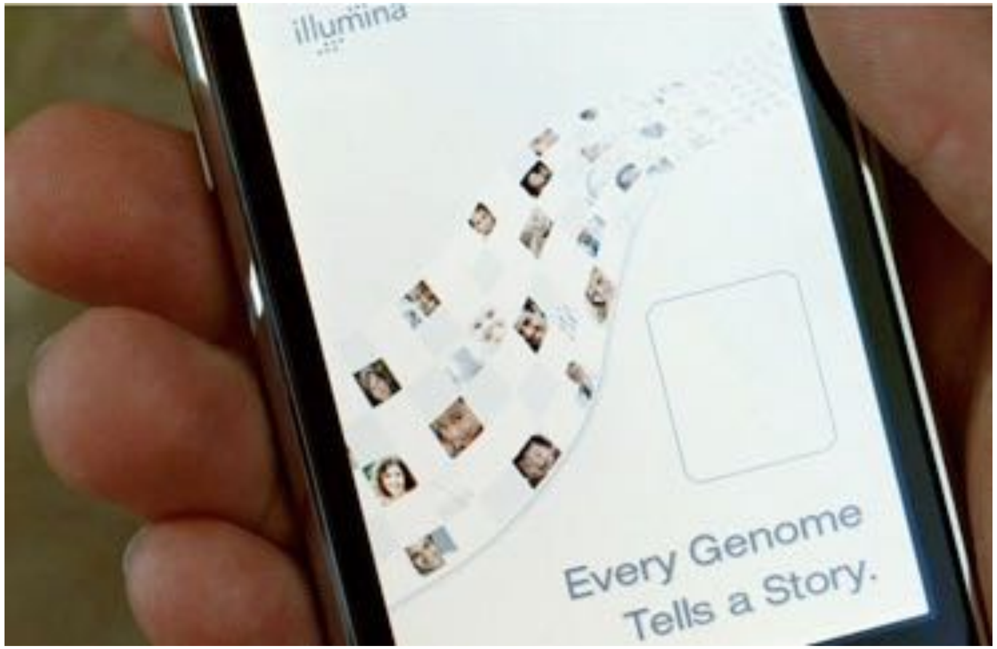
Personal Genomics

Search 23andMe
Go
log in | claim codes | blog | help | your cart
23andMe
genetics just got personal.

Get the latest on your DNA with $399 and a tube of saliva



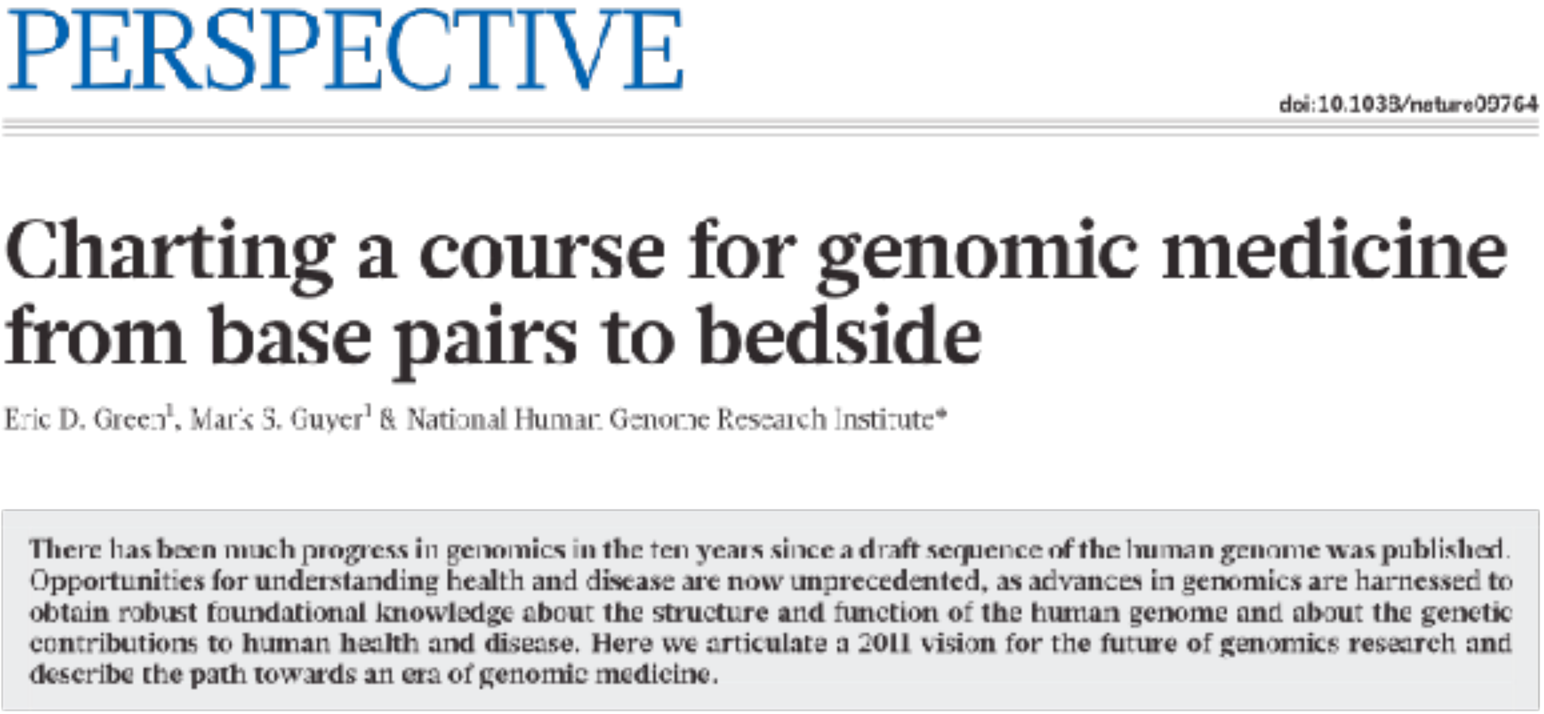
Personal Genomics



- 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 developed105.

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.*

What else is the class about?

- Gives you an example of end-to-end use of what you've learned as CS as a practice
  - We discuss the design and analysis of algorithms (e.g., string algorithms, dynamic programming, iterative optimization methods)
  - We implement algorithms (python)
  - We analyze data (also in python)
- We also learn about biology, medicine and why government shutdowns are really awful

Administrative Details

Class webpage:

1. http://www.hcbravo.org/cmsc423

Everything you want to know is there.

Todo after class today

1) Enroll in Piazza class
2) Enroll in Rosalind pre-lecture and final submission pages (links posted in Piazza)
3) Complete course survey on ELMS

For next class

1) Reading
2) Pre-lecture reading quiz on ELMS
3) Pre-lecture rosalind exercises