

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 2015

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

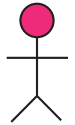
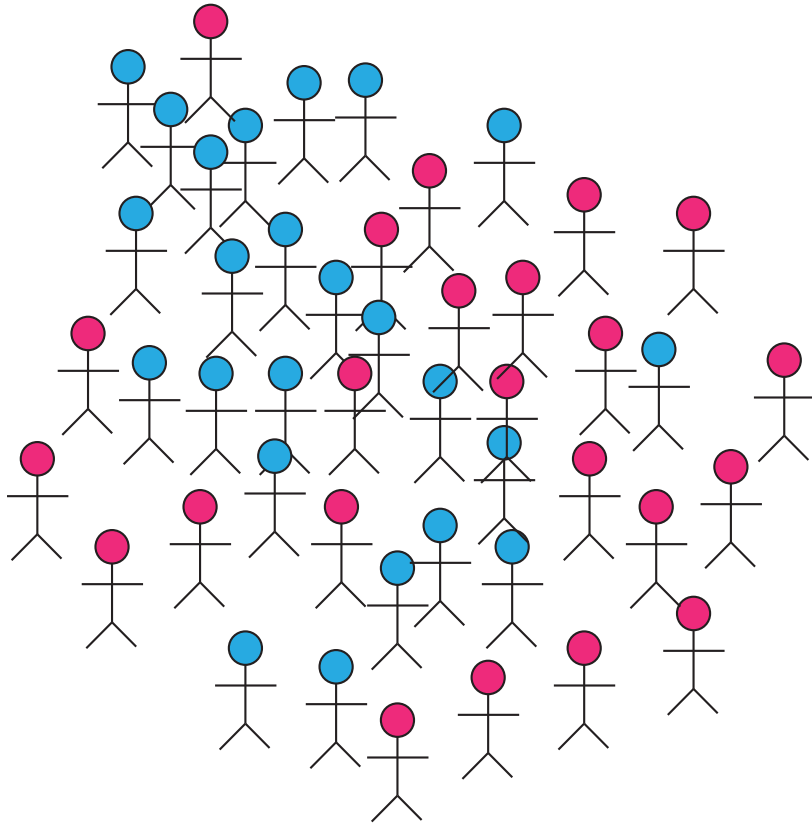
Why are my children
such pigs?



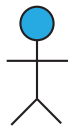
What is Genomics?

- 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?



cancer



healthy

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

Measurement

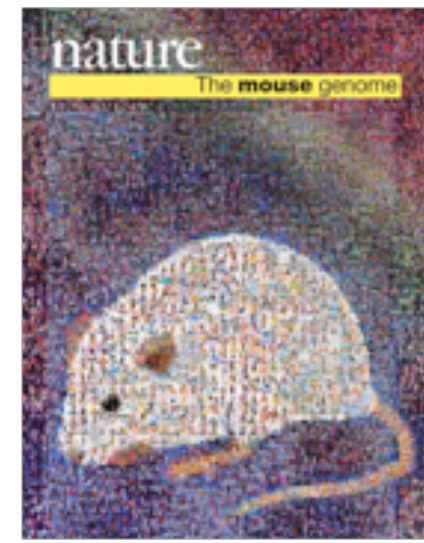
- For a small enough piece, we can measure the sequence of bases, referred to as *sequencing*
- Human Genome Project



D. melanogaster, Science, 2000



H. sapiens, Nature, 2000
and Science, 2000



M. musculus, Nature, 2002

Genome

[illegible]

**Total amount of DNA in human genome:
3 * 10⁹ base pairs (bp)**

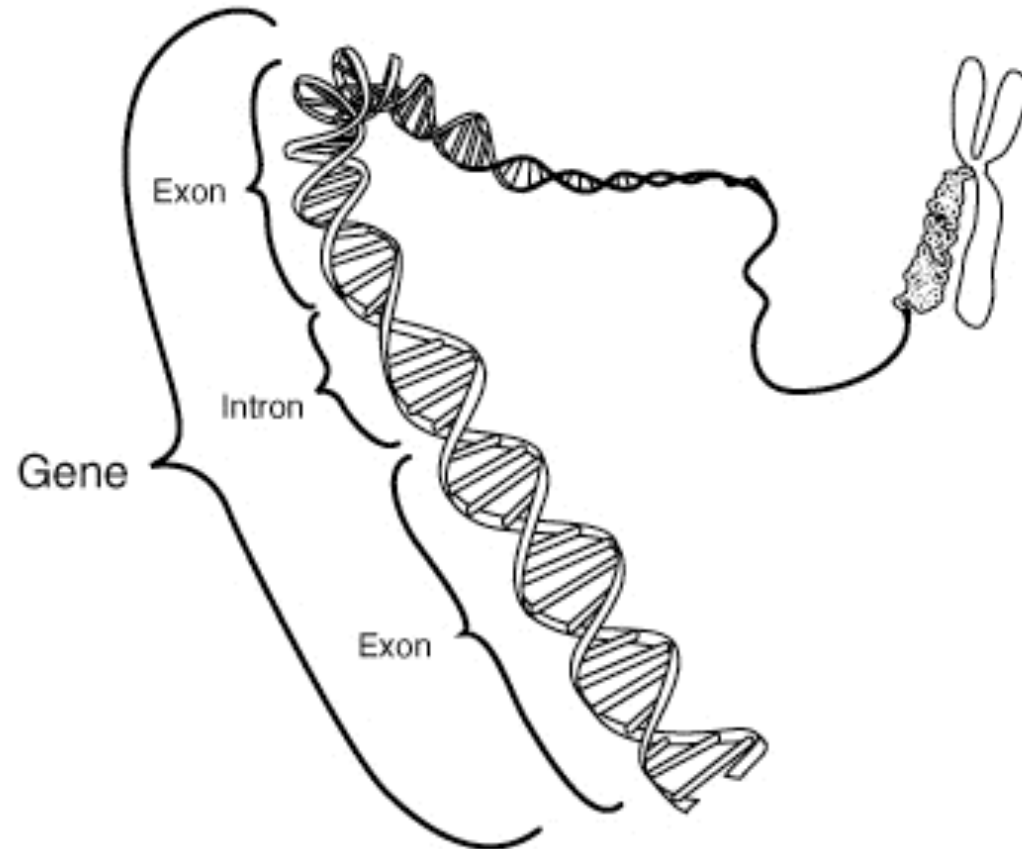
Why are these two different?



Differences explained by 1-10% difference in genome

Similarities explained by similar genes

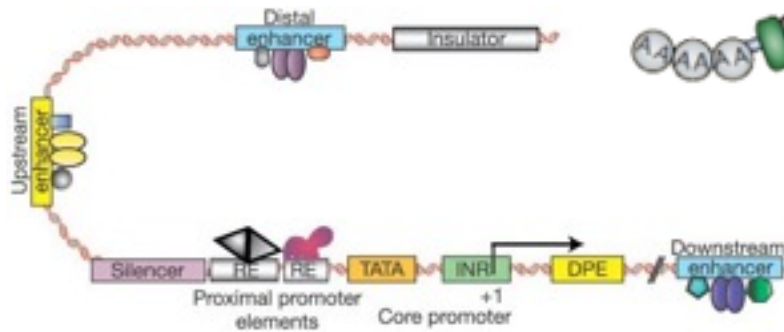
Genes



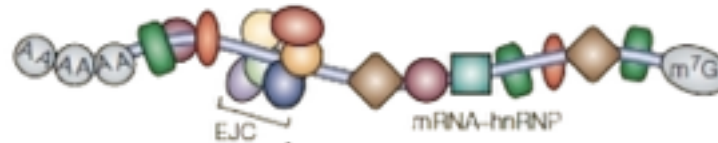
Computational Biology

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

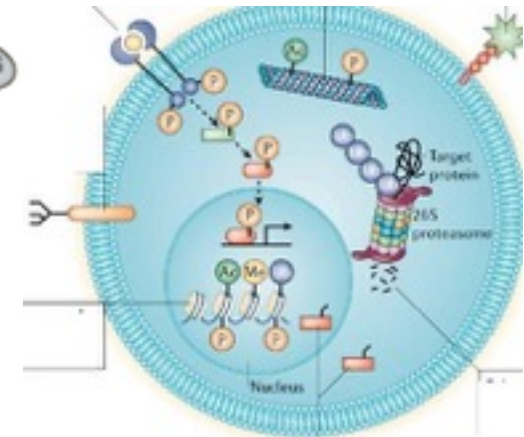
genomics



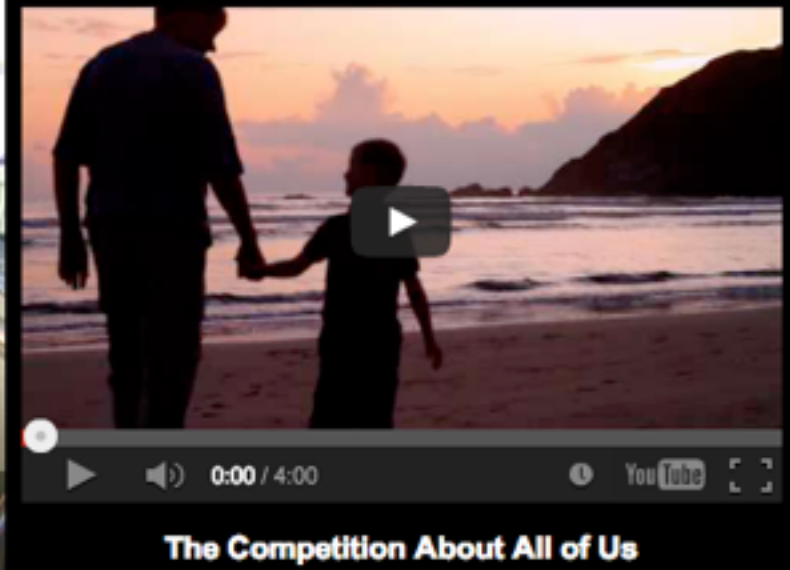
transcriptomics



proteomics



Major technological advances allow **unprecedented** data acquisition



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   

TOPICS [Space](#) [Environment](#) [Innovation](#) [Weird 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

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

SPACE STATION
Japanese astronaut to command space station in March

The Archon Genomics XPRIZE Presented by Express...



“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.”

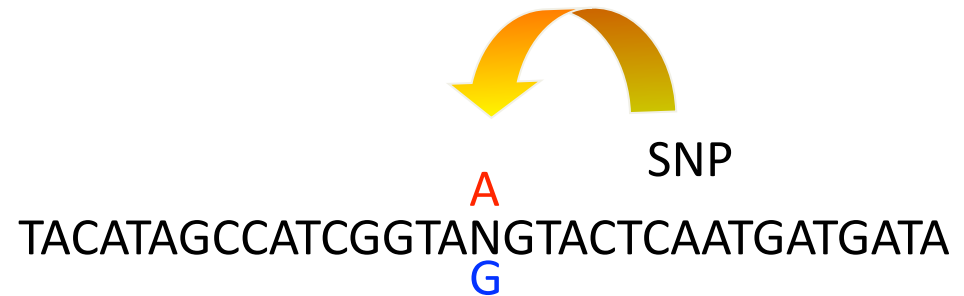
What makes them different?



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

Single Nucleotide Polymorphism (SNP)

Genomic DNA:



From reads to evidence

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNNNNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A%NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%NNNNNNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNNNNNNNNNNNN
+
A=B7&7:>B@:A>79:<;:>747%NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAAGCACAGCCAANNNNNANTNNTNNTNNN
+
BBCCCCCBB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACTCCGCCTCNNNNNGNTNAAAGNNNN
+
BCC7+<B=7BB5=ABA7B6BBB84BB7B%NNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:947#0/1
CCAGGAGAAAGCCATGTTTCAGTTCGAGCGCNANANCGTGANNNN
+
BBB9@?7A7>AAB@>7B=?@.>87B7%NNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCCTCCCCATCTCCACCCCTGTACCTNANCCCTGANNNN
+
BBABAABB;AAABA77@5AAA:??>%NNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTTTCCAGGCCTCCTNCCTCCGTNNNN
+
BBB@5@BBB8BBB8@BBB8BAA8BB7;9BB@BA5&<B:%NNNNNN
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACTCCTGACCTTTGGTGATCCACCGCCTNGGCCTTCNNNN
+
BBBB:BBBBBAAAA?: (=AB@>AAA7AB7=A%NNNNNNNNNNNN
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNNNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A%NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%NNNNNNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNNNNNNNNNNNN
+
A=B7&7:>B@:A>79:<;:>747%NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAAGCACAGCCAANNNNNANTNNTNNTNNN
+
BBCCCCCBB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7CB7
```

From reads to evidence

I. Comparative

Sequence-wise, individuals of a species are nearly identical

Well curated, annotated “reference” genomes exist

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A\NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>\NNNNNNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B767:>B@:A>79:<;>747\NNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAGCACAGCCAANNNNNANTNNTNNNN
+
BBCCCCCBBCB7CBC=7>+<=BCBCB\NNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCINNNGNTNAAGNNNN
+
BCC?+<B=7BB5=ABA7B6BBB4BB7B\NNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCAAGTTCAGTCAGCGCANNANANCGTGANNNN
+
BBB9@77A7>AAB@>7B=?@.>B7B?NNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCTCCCCCATCTCCACCCTGTACCTNANCCCTGANNNN
+
BBA8AABB;AAABA77@5AAA:??>\NNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTCTTCCAGGCCTCCTNCCTCCGTNNNN
+
BBB@6@BBB8888@BBBABAABBB?;9BB@BA5<B:\NNNNNNNN
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACCTCTGACCTTTGGTGATCCACCCGCTNGGCCTTCNNNN
+
BBBB:BBBBBAAAA?:(=AB@>AAA?AB?>A\NNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A\NNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTCAGCAGGNNNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>\NNNNNNNNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B767:>B@:A>79:<;>747\NNNNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAGCACAGCCAANNNNNANTNNTNNNN
+
BBCCCCCBBCB7CBC=7>+<=BCBCB\NNNNNNNNNNNNNNNNNN
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCINNNGNTNAAGNNNN
+

```



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

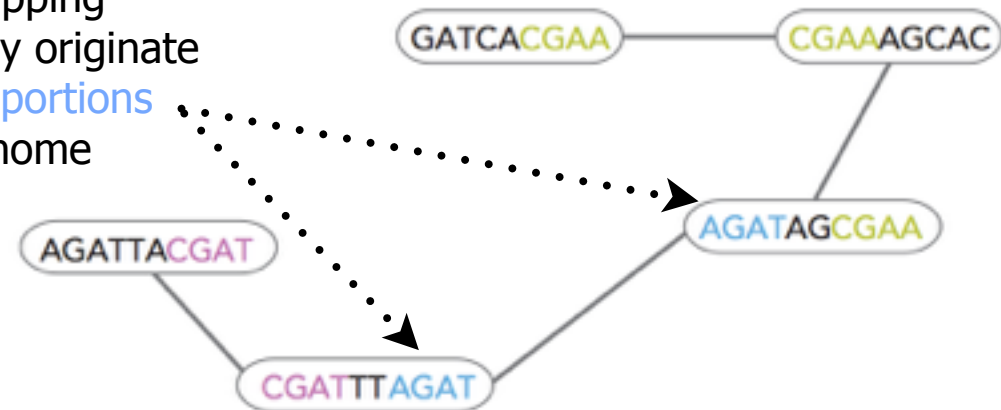
From reads to evidence

2. *de novo*

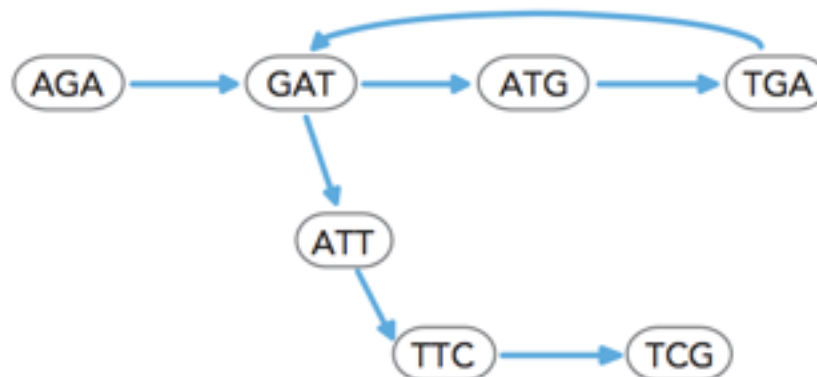
Assume nothing! - let reads tell us everything

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



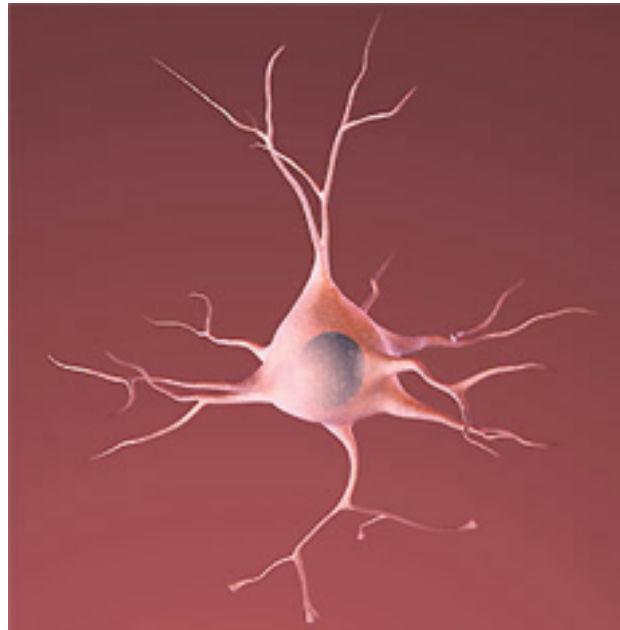
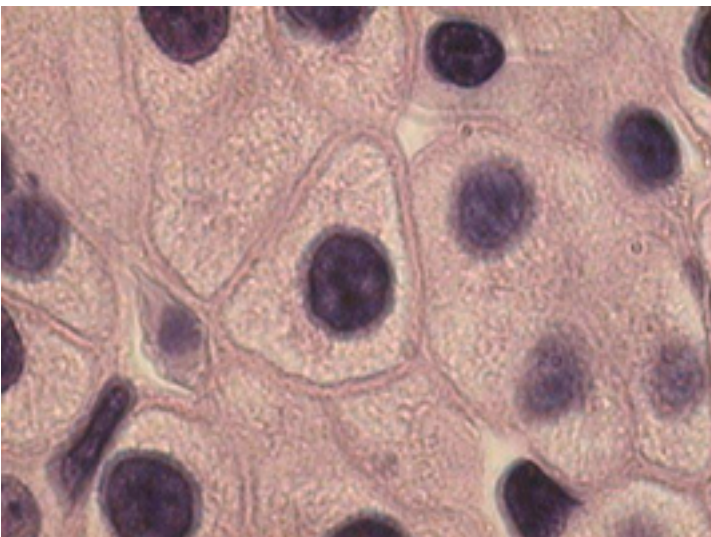
The full genome sequence is a "tour" of the graph

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

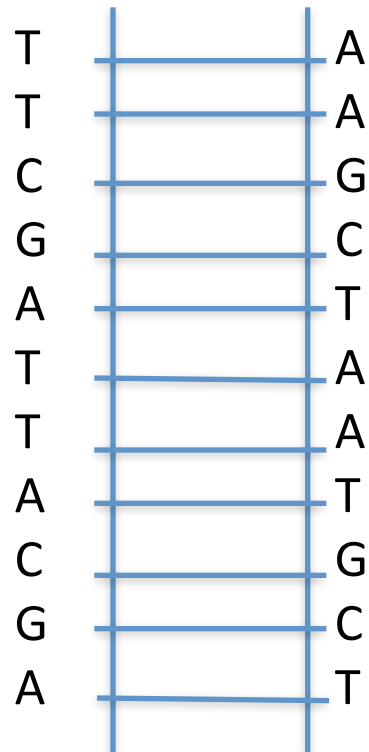
http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGGNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B767:>B@:A>79:<;>747%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAAGACAGCCAANNNNNANTNCTNNNN
+
BBCCCCCBBCB7CBC=7>+<=BCBCB%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCINNNGNTNAAGNNNN
+
BCC?+<B=7BB5=ABA7B6BBB4BB7B%
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTCACTTCGAGCGCNNANANCGTGANNNN
+
BBB9@77A7>AAB@>7B=?@.>87B?%
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCCTCCCCATCTCCACCCTGTACCTNANCCCCCTGANNNN
+
BBA8AABB;AAABA77@5AAA:??>%
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCCTACACTCTTCCAGGCCCTCCTNCCTCCGTNNNN
+
BBB@6@BBB8888@BBBABAABBB?;9BB@BA5<B:%
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACCTCTGACCTTTGGTGATCCACCCGCTTNGGCTTCNNNN
+
BBBB:BBBBBAAAA?:(=AB@>AAA?AB?>A%
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB>A7B@;@BBBBBAA=BA=A%
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGGNNNNNNNNNGNNNNN
+
B9B@B<;BAA<@AB9=1>%
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNCNNNNCNNNNN
+
A=B767:>B@:A>79:<;>747%
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAAGACAGCCAANNNNNANTNCTNNNN
+
BBCCCCCBBCB7CBC=7>+<=BCBCB%
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCINNNGNTNAAGNNNN
```

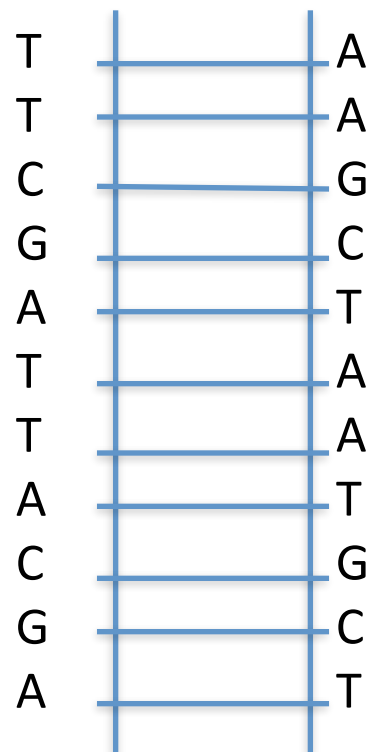
How many basepair differences?



Liver



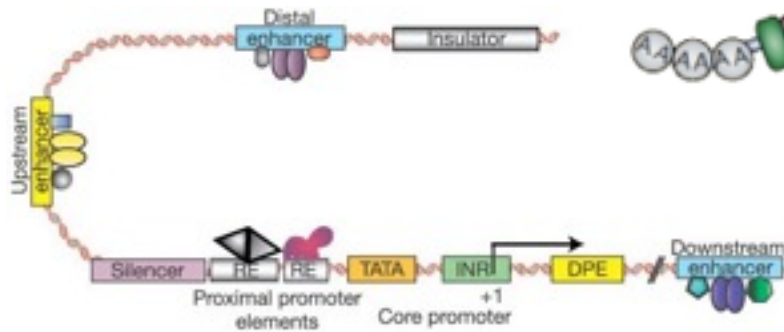
Brain



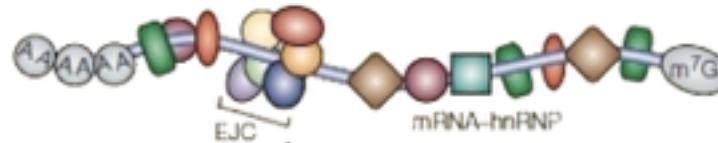
Computational Biology

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

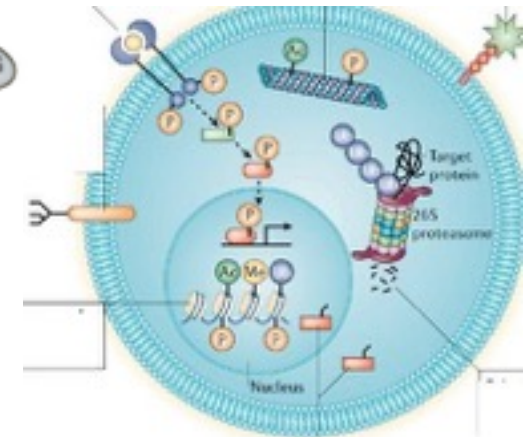
genomics



transcriptomics

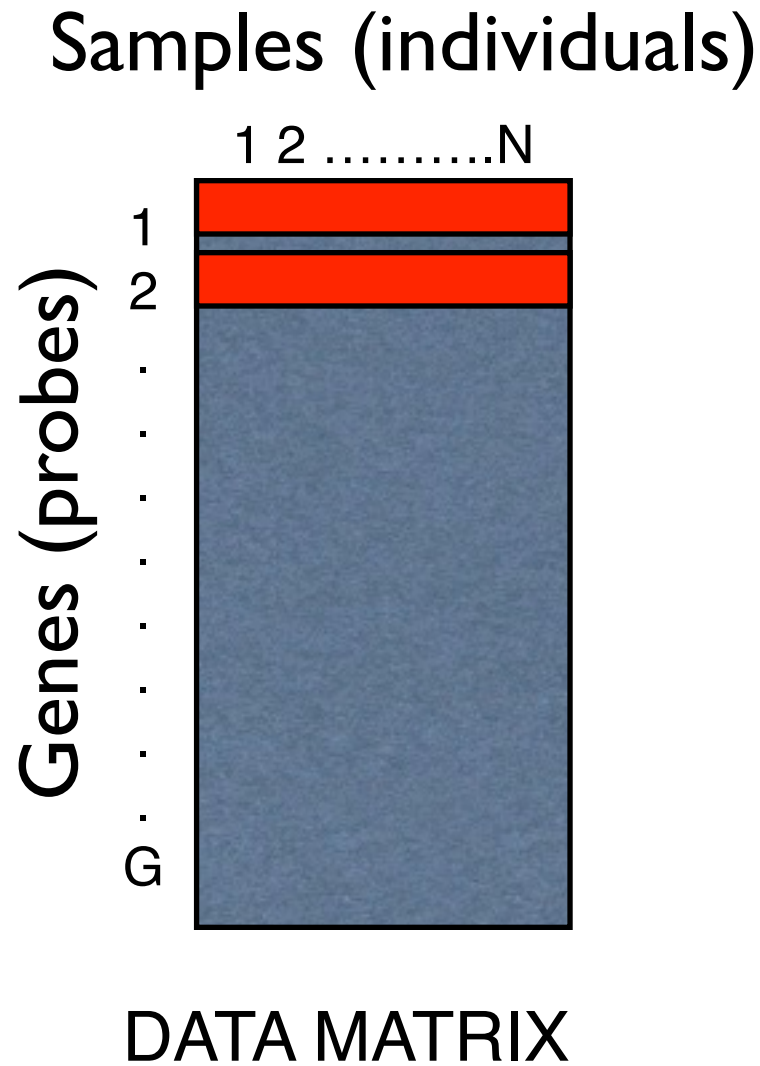


proteomics



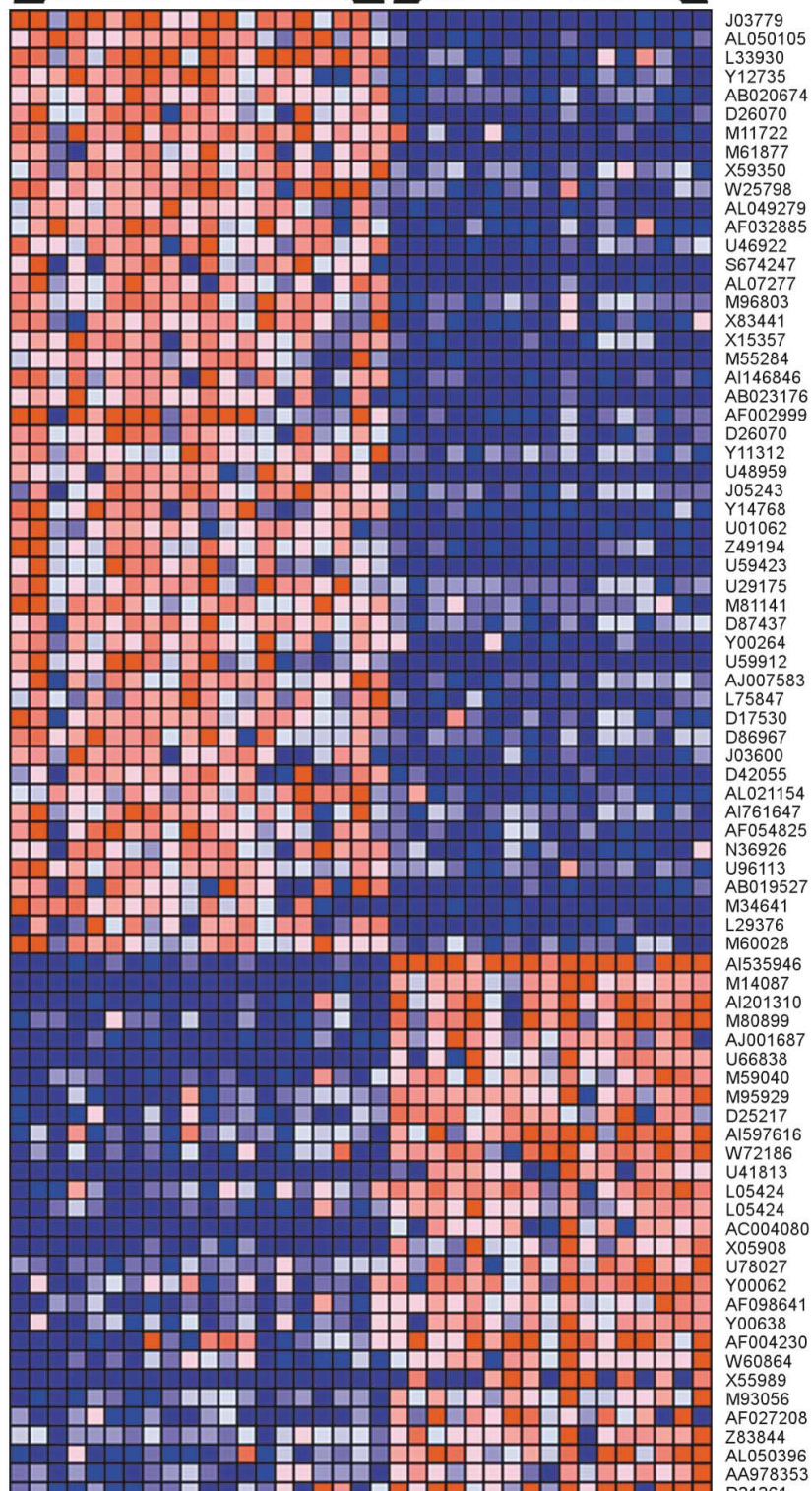
Major technological advances allow **unprecedented** data acquisition

Measurements



ALL

MLL



J03779
AL050105
L33930
Y12735
AB020674
D26070
M11722
M61877
X59350
W25798
AL049279
AF032885
U46922
S674247
AL07277
M96803
X83441
X15357
M55284
AI146846
AB023176
AF002999
D26070
Y11312
U48959
J05243
Y14768
U01062
Z49194
U59423
U29175
M81141
D87437
Y00264
U59912
AJ007583
L75847
D17530
D86967
J03600
D42055
AL021154
AI761647
AF054825
N36926
U96113
AB019527
M34641
L29376
M60028
AI535946
M14087
AI201310
M80899
AJ001687
U66838
M59040
M95929
D25217
AI597616
W72186
U41813
L05424
L05424
AC004080
X05908
U78027
Y00062
AF098641
Y00638
AF004230
W60864
X55989
M93056
AF027208
Z83844
AL050396
AA978353
D21264

article

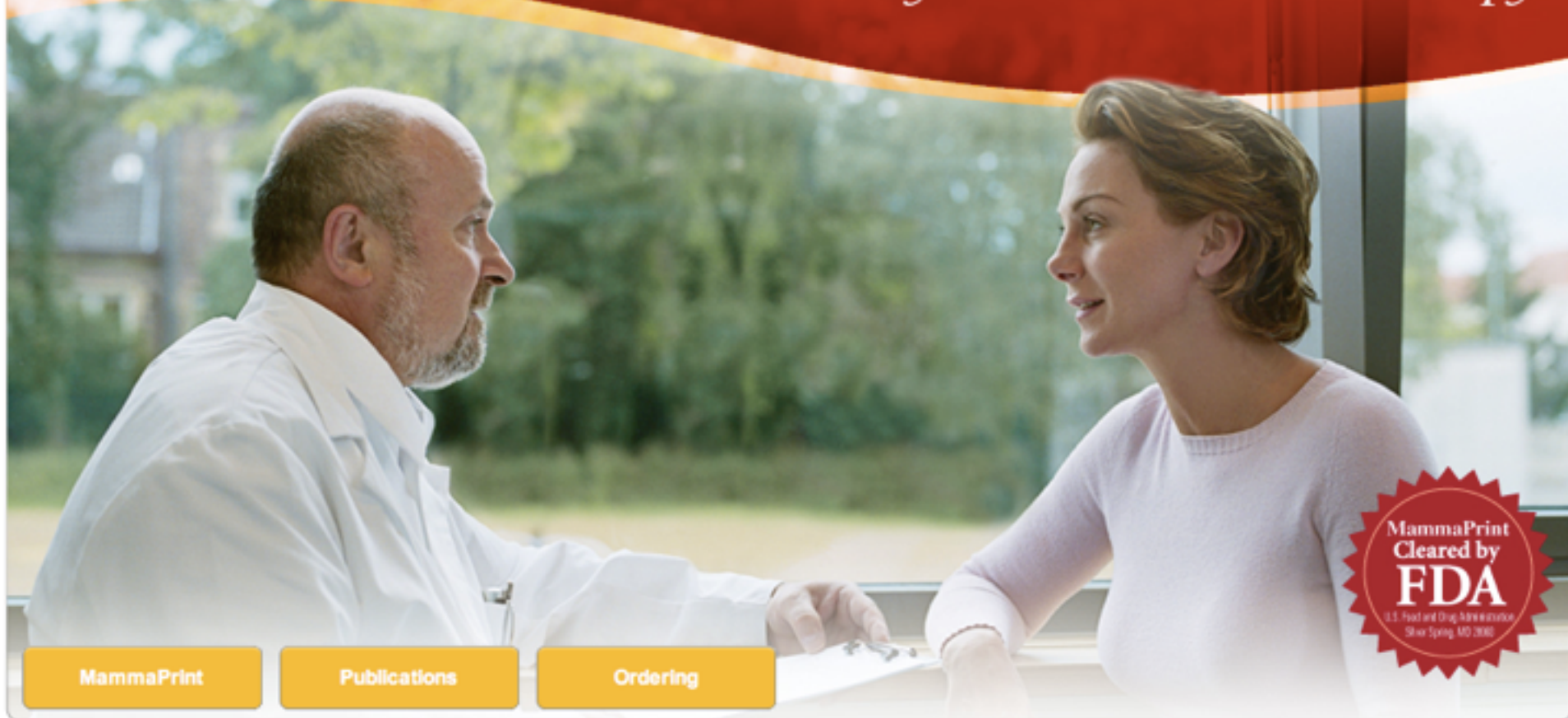
MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia

Scott A. Armstrong¹⁻⁴, Jane E. Staunton⁵, Lewis B. Silverman^{1,3,4}, Rob Pieters⁶, Monique L. den Boer⁶, Mark D. Minden⁷, Stephen E. Sallan^{1,3,4}, Eric S. Lander⁵, 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.



MammaPrint

Publications

Ordering

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

PERSONAL GENOMICS



23andMe genetics just got personal.

Search 23andMe

Go

[log in](#)

[claim codes](#)

[blog](#)

[help](#)

[your cart](#)

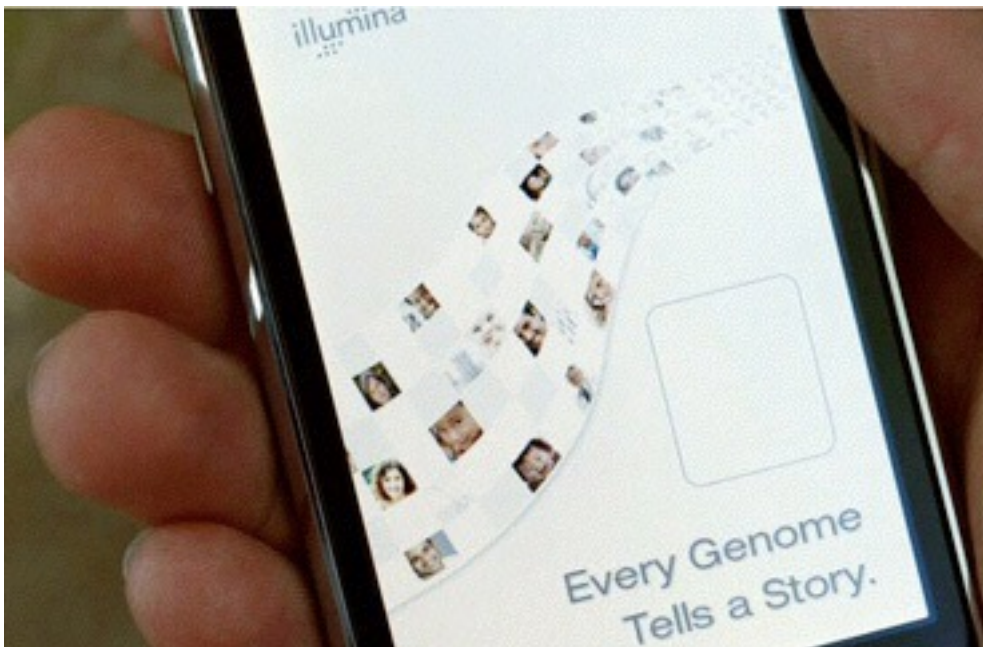
Get the latest on your DNA with \$399 and a tube of saliva

illumina[®]



Every Genome Tells a Story.
What's yours?

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. Green¹, Mark S. Guyer¹ & 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.*

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.cbcb.umd.edu/~hcorrada/CMSC423>

Everything you want to know is there.