Intro to Bioinformatics

Marylyn D Ritchie, PhD
Professor, Biochemistry and Molecular Biology
Director, Center for Systems Genomics
The Pennsylvania State University

Sarah A Pendergrass, PhD
Research Associate
Biochemistry and Molecular Biology
The Pennsylvania State University

Center for Systems Genomics
Pennsylvania State University

Ritchie Lab
Pennsylvania State University

Molecular biologist

Epidemiologist

Bioinformatician

Biologist

Virologist

Statistician

Clinical geneticist

Human Geneticist

Biostatistician

Immunologist

Physician/clinician

Other

What is bioinformatics?

- Conceptualizing biology in terms of molecular and applying informatics techniques to understand and organize the information associated with these molecules on a large scale. –Luscombe et al. 2001
- Application of computational techniques to analyze information associated with biomolecules on a large scale
- Established as a discipline in molecular biology
- Wide range of subject areas
  - Structural biology, genomics, transcriptomics, etc.

Luscombe et al., Yearbook of Medical Informatics 2001

Aims of bioinformatics

1. Organize data in a way to allow researchers to access existing information and submit new information as it is produced
2. Develop tools and resources to aid in the analyses of data
3. Use these tools to analyze the data and interpret the results in a biologically meaningful manner

Luscombe et al., Yearbook of Medical Informatics 2001

Data source	Data size	Bioinformatics topics
Raw DNA sequence	11.5 million sequences (12.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	400,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	15,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions)
Genomes	300 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes in yeast	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

Luscombe et al., Yearbook of Medical Informatics 2001

That was then….

What is Bioinformatics? A Proposed Definition and Overview of the Field

N. M. Luscombe, D. Greenbaum, M. Gerstein
Department of Molecular Biophysics and Biochemistry
Yale University, New Haven, USA

Summary
Background: The recent flood of data from genome sequences and functional genomics has given rise to new field, bioinformatics, which combines elements of biology and computer science.

1. Introduction

Biological data are being produced at a phenomenal rate [1]. For example as

This is now …

"I think you'll find that mine is bigger"

What is Big Data?

- Big data is a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. – Wikipedia

- Data sets that are too large and complex to manipulate or interrogate with standard methods or tools. – Oxford Dictionary

- Computers, data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process. – Dictionary.com

Where do we see Big Data?

To put the data explosion in context, consider this. Every minute of every day we create:

- More than 204 million email messages
- Over 2 million Google search queries
- 48 hours of new YouTube videos
- 684,000 bits of content shared on Facebook
- More than 100,000 tweets




http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html - March, 2014

Where do we see Big Data?

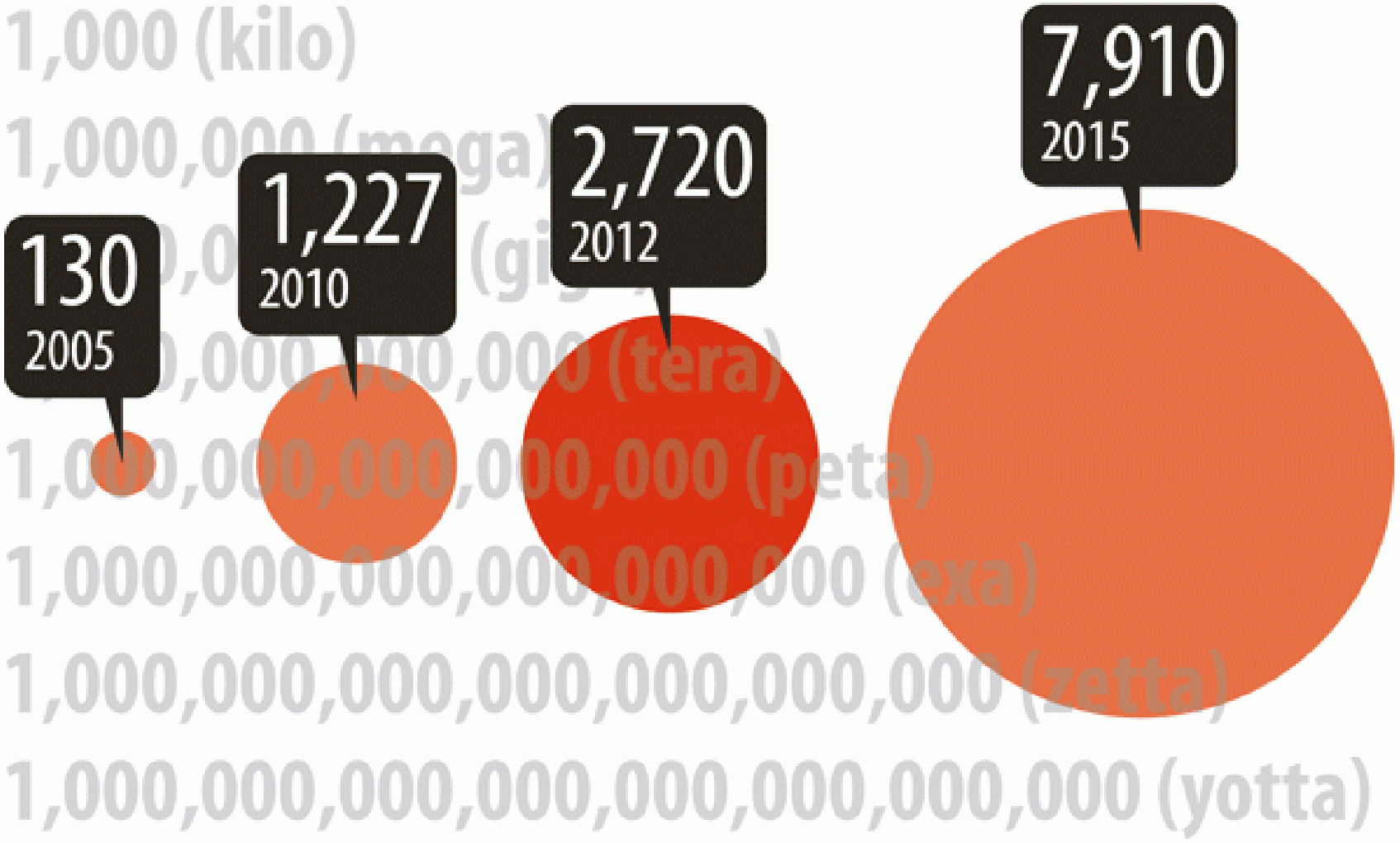
90 PB (pedabytes)
May, 2013

300 PB (pedabytes)
April, 2014

Exabytes of data

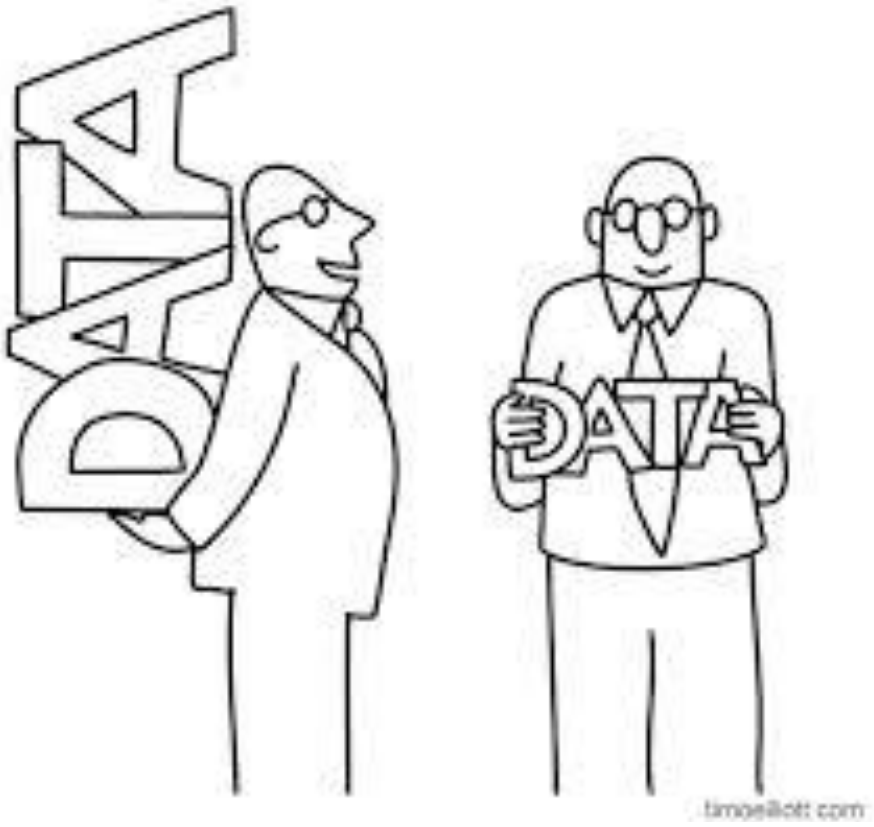
Exaponential

Quantity of global digital data, exabytes

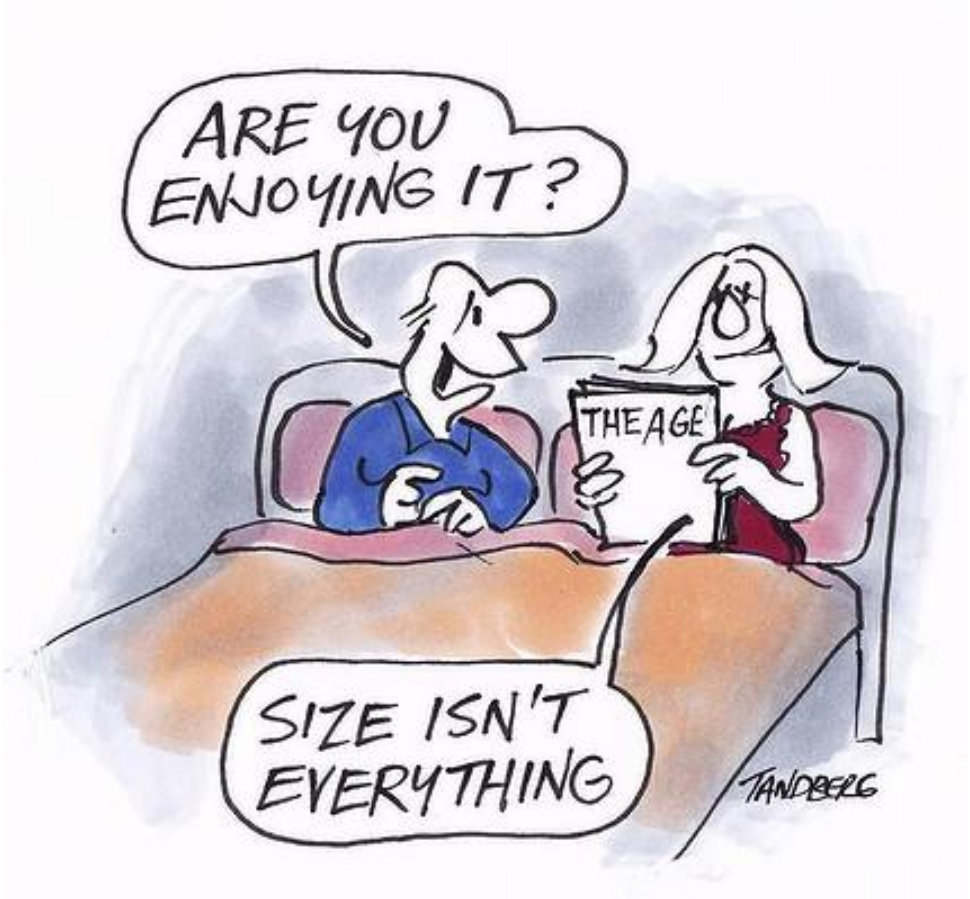


Source: EMC/IDC Digital Universe Study, 2011

Big Data: Is it all about size?

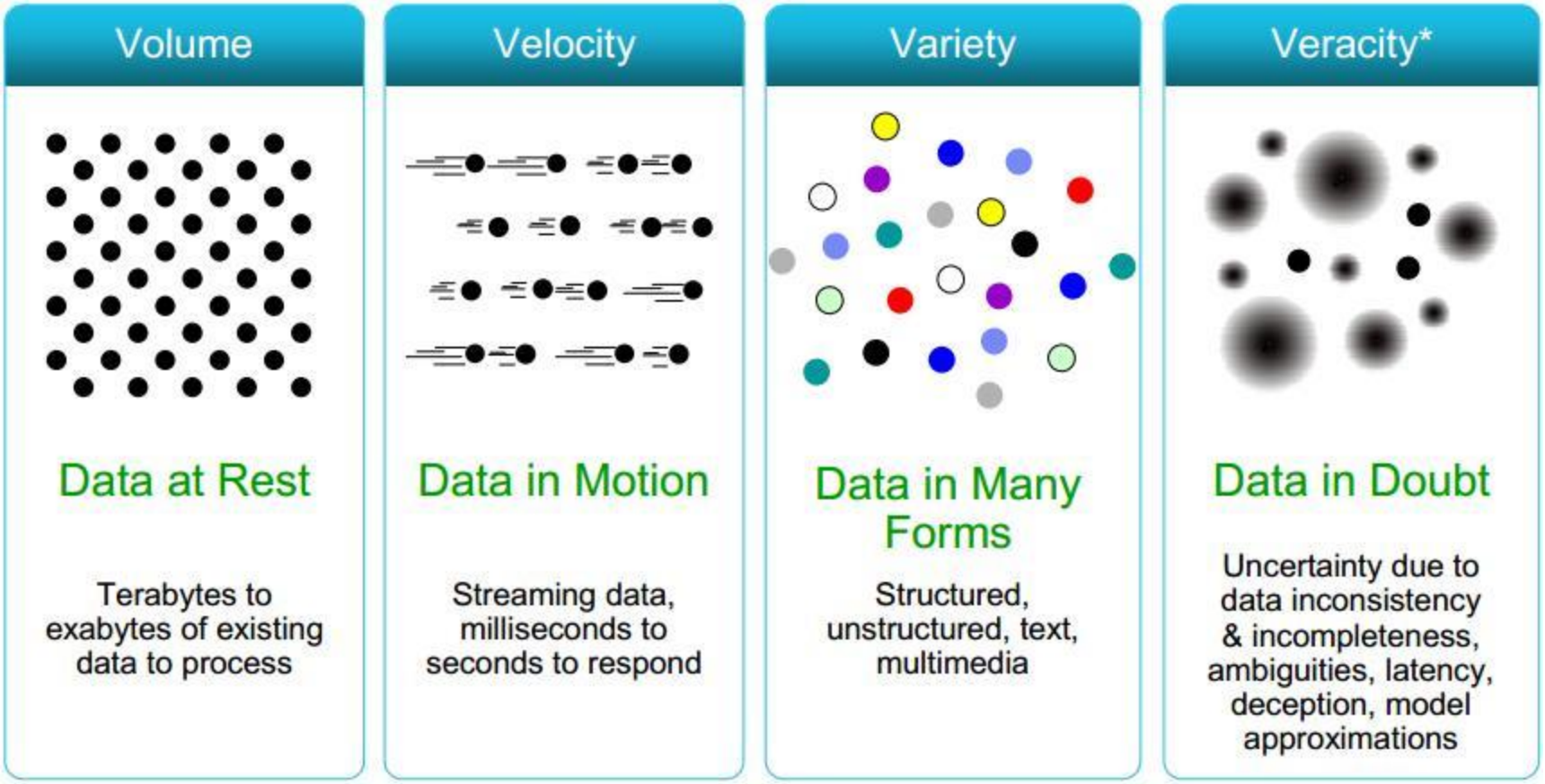


"I think you'll find that mine is bigger"



Depends on your frame of reference

The ~~three~~ four V's of Big Data



http://www.datasciencecentral.com/profiles/blogs/data-veracity

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Effective Use of Big Data

- The use of big data effectively is critical to reap benefit from the massive resources
- Online retail – compile history of every click to recommend additional purchases

amazon.com

- Traffic data – needs to be real time
  - No need for 5 minute old traffic data
- Industry buzzwords:
  - “streaming data”
  - “complex event processing”

Google Flu

- In 2009, Google published in Nature
- Google search queries to track influenza-like illness
- Relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms

- Accurately estimate the current level of weekly influenza activity in each region of the United States
- Reporting lag of about one day
- Better than the Centers for Disease Control
  - More than a week

Google Flu

- Few months after announcing Google Flu, the world was hit with the 2009 swine flu pandemic
- Caused by a novel strain of H1N1 influenza
- Google Flu missed it
- A bigger problem with Google Flu, though, is that most people who think they have “the flu” do not
- The vast majority of doctors’ office visits for flu-like symptoms turn out to be other viruses

What happened?

- Unpredictability
- Complexity

- Not trying to determine what caused flu
- Correlation not causation

Challenges

- Big data is big
  - Seeing an inversion in priorities
  - Rather than moving data, we are moving programs to where the data are
- Big data is messy
  - “80% of the effort involved in dealing with data is cleaning it up” – Pete Warden
- Emergence of a new field
  - Data science
  - Combines math, computer science, and scientific instinct

Challenges

- Capture
- Storage
- Curation
- Search
- Sharing
- Transfer
- Analysis
- Visualization

Big Data

- Data alone does not answer all questions
- More data alone cannot solve all problems

- Hypothesis generating strategies
- Paradigm shift from hypothesis testing science

Big Data

Formula one race car

Nascar headquarters

200 data feeds

5GB per lap

Joel Dudley, Mt. Sinai School of Medicine

Big Data

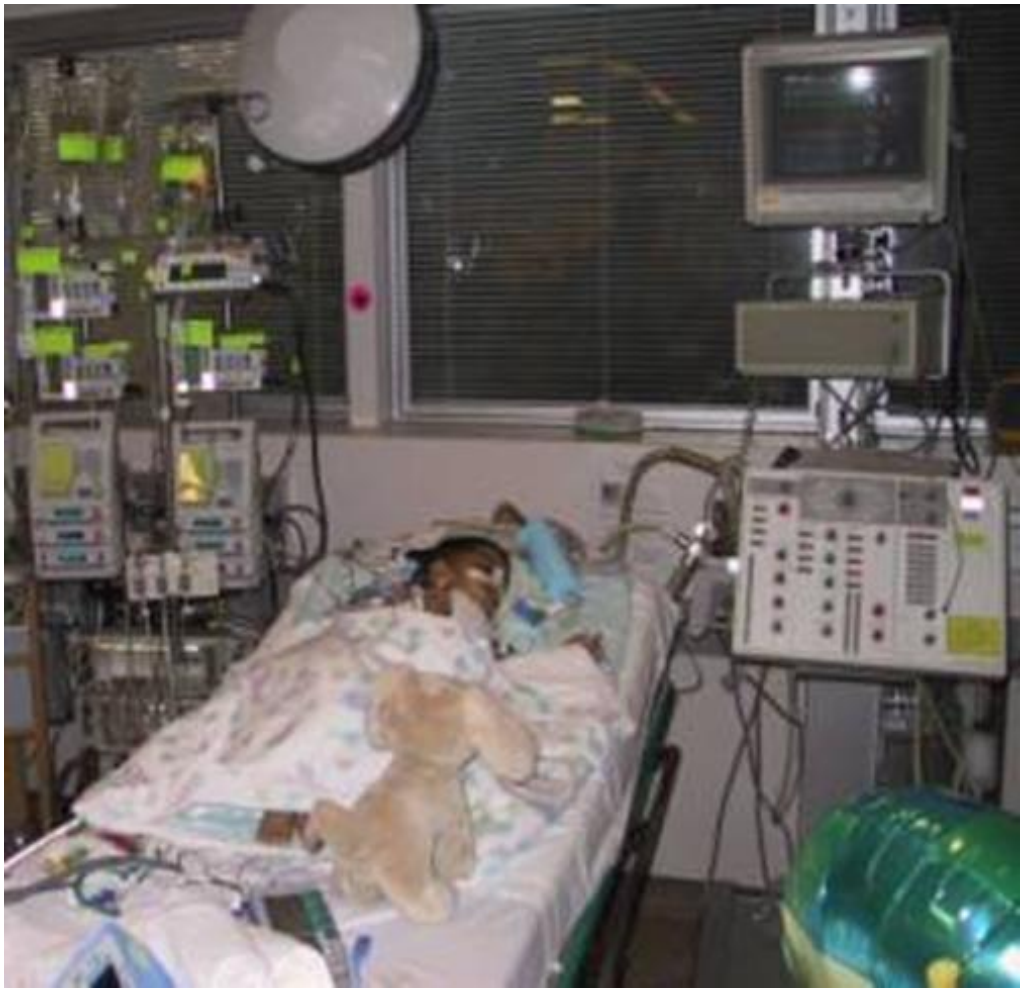
Blood samples for genetic screening are collected from a heel prick before the newborn is discharged from the hospital.

Photo: Courtesy of The Museum of disABILITY HISTORY

SIGN	0	1	2	1 min	5 min
Heart Rate	Absent	Less Than 100	Over 100	2	2
Respiratory Effort	Absent	Slow, Irregular	Good Cry	1	2
Muscle Tone	Limp	Some Flexion	Active Motion	1	2
Reflex Irritability	No Response	Grimace	Cry	1	2
Color	Pale	Body Pink, Extr. Blue	All Pink	1	2
TOTAL SCORE				6	10

Joel Dudley, Mt. Sinai School of Medicine

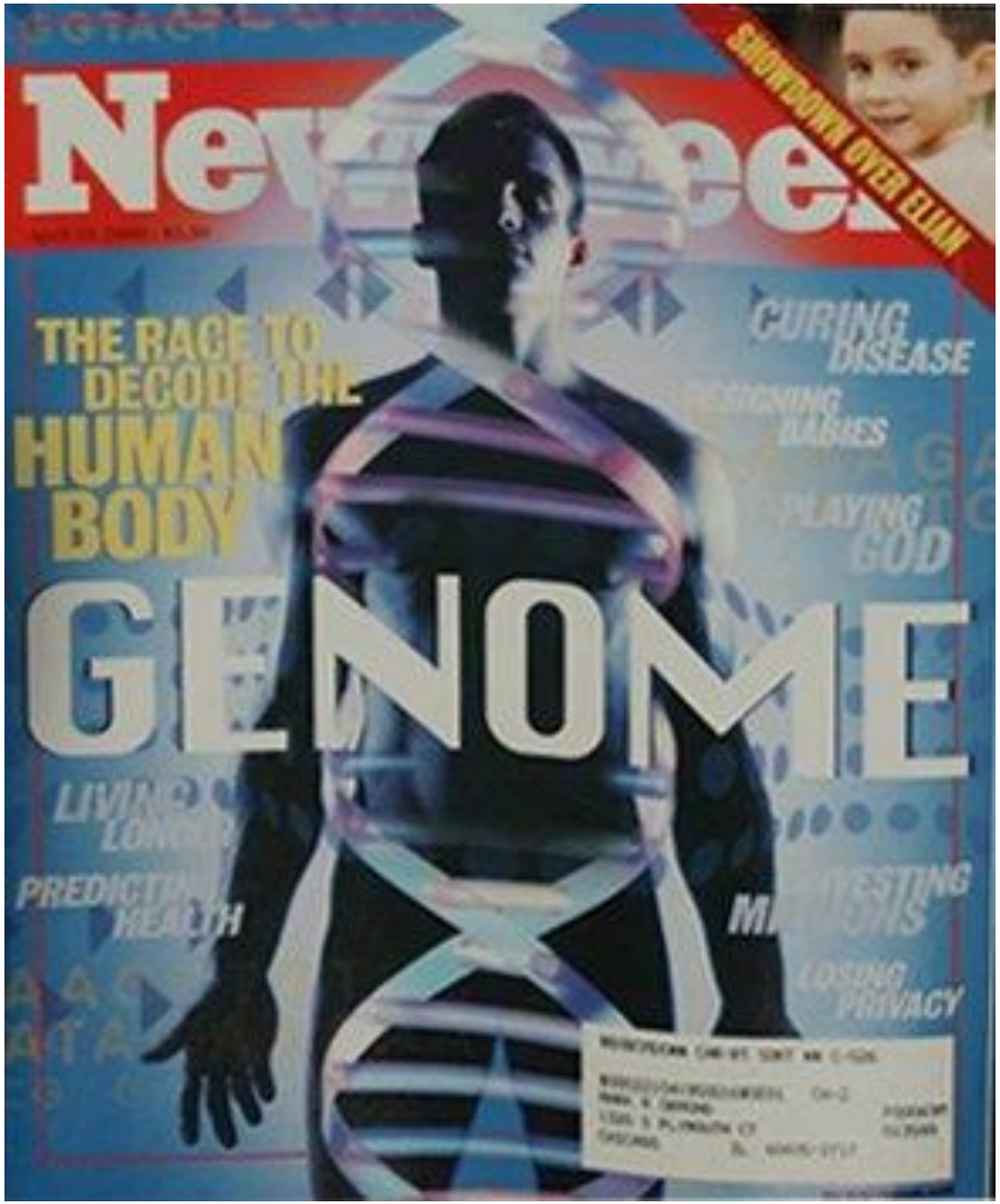
Big Data



Bioinformatics is big data

- Ability to generate data is at an unprecedented rate
- High-throughput technologies have moved scientific disciplines leaps and bounds
- Bioinformatics as a discipline is emerging, expanding, running to keep up with the data

How did we get here?


SHOWDOWN OVER ELIAN
THE RACE TO DECODE THE HUMAN BODY
GENOME
CURING DISEASE
DESIGNING BABIES
PLAYING GOD
LIVING LONGER
PREDICTING HEALTH
INVESTING
LOSING PRIVACY


15 February 2001
nature
$10.00
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

Goal of the Human Genome Project (HGP)

- To obtain a highly accurate sequence of the vast majority of the euchromatic portion of the human genome
- Launched in 1990
- International Human Genome Sequencing Consortium (IHGSC) formed
- 20 centers
- 6 countries
- Manuscript contained 14 pages of authors (in supplemental material)

Human Genome Project (HGP)

- 3 Phases in HGP
  1) Preliminary phase that developed and refined approaches
  2) Draft phase that yielded 90% of the information
  3) Finishing phase that yielded 99% of the information
- 1% of the euchromatic genome remains

Human Genome Project (HGP)

- Key challenges
  1) Systematic identification of all genetic polymorphisms carried in human populations

- Started in October, 2002
- First Haplotype Map published in October, 2005

Vol 437|27 October 2005|doi:10.1038/nature04226

nature

ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

Human Genome Project (HGP)

- Key challenges
  2) Systematic identification of all functional elements in the human genome including genes, proteins, regulatory controls, and structure elements
  3) Systematic identification of all the “modules” in which genes and proteins function together
     - Requires the study of expression, localization and interaction in a spatial and temporal context

- Launched in September, 2003
- Pilot project published in June, 2007

Vol 447|14 June 2007|doi:10.1038/nature05874

nature

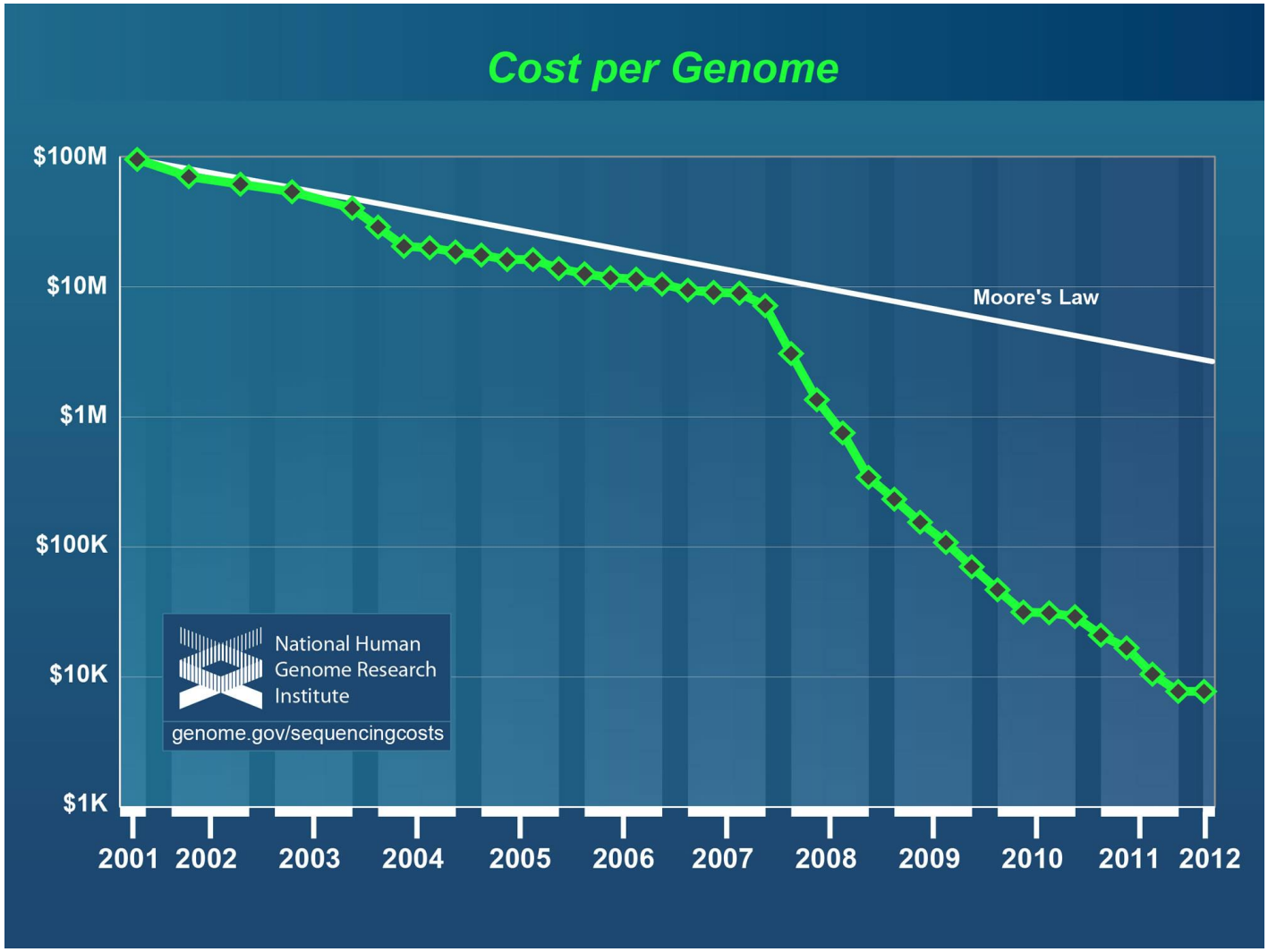
ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

We report the generation and analysis of functional data from multiple, diverse experiments performed on a targeted 1% of the human genome as part of the pilot phase of the ENCODE Project. These data have been further integrated and augmented by a number of evolutionary and computational analyses. Together, our results advance the collective knowledge about human genome function in several major areas. First, our studies provide convincing evidence that the genome is pervasively transcribed, such that the majority of its bases can be found in primary transcripts, including non-protein-coding transcripts, and those that extensively overlap one another. Second, systematic examination of transcriptional regulation has yielded new understanding about transcription start sites, including their relationship to specific regulatory sequences and features of chromatin accessibility and histone modification. Third, a more sophisticated view of chromatin structure has emerged, including its inter-relationship with DNA replication and transcriptional regulation. Finally, integration of these new sources of information, in particular with respect to mammalian evolution based on inter- and intra-species sequence comparisons, has yielded new mechanistic and evolutionary insights concerning the functional landscape of the human genome. Together, these studies are defining a path for pursuit of a more comprehensive characterization of human genome function.

Perspective



Perspective

THEN

The Whitehead/MIT Center for Genome Research

NOW

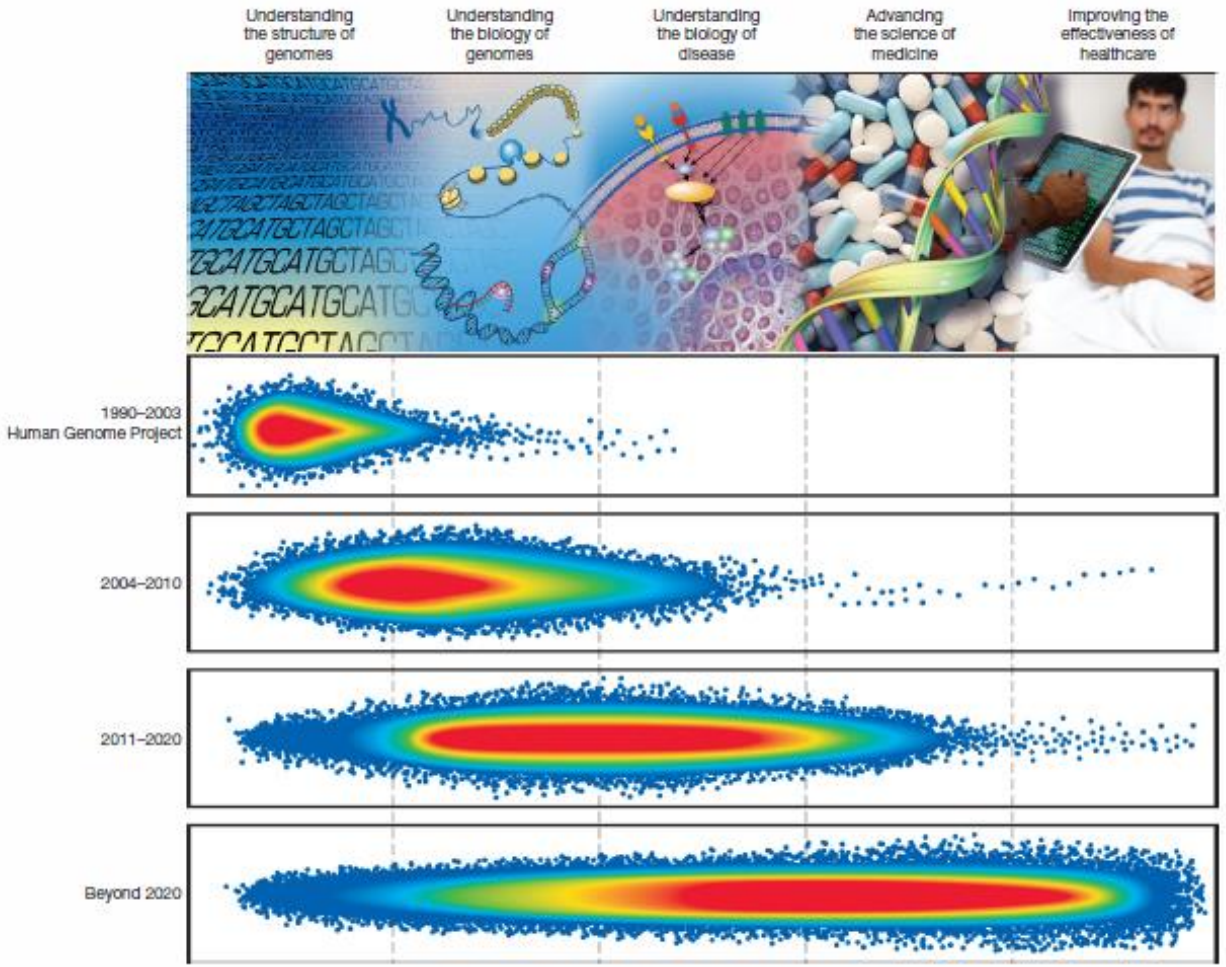
Oxford Nanopore

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*



DNA sequencing changed biology in many ways…

1. Evolutionary and comparative genomics

Adapted from Elliot Marguiles' 2/9/2010 NHGRI talk (http://www.genome.gov/12514288)

DNA sequencing changed biology in many ways…

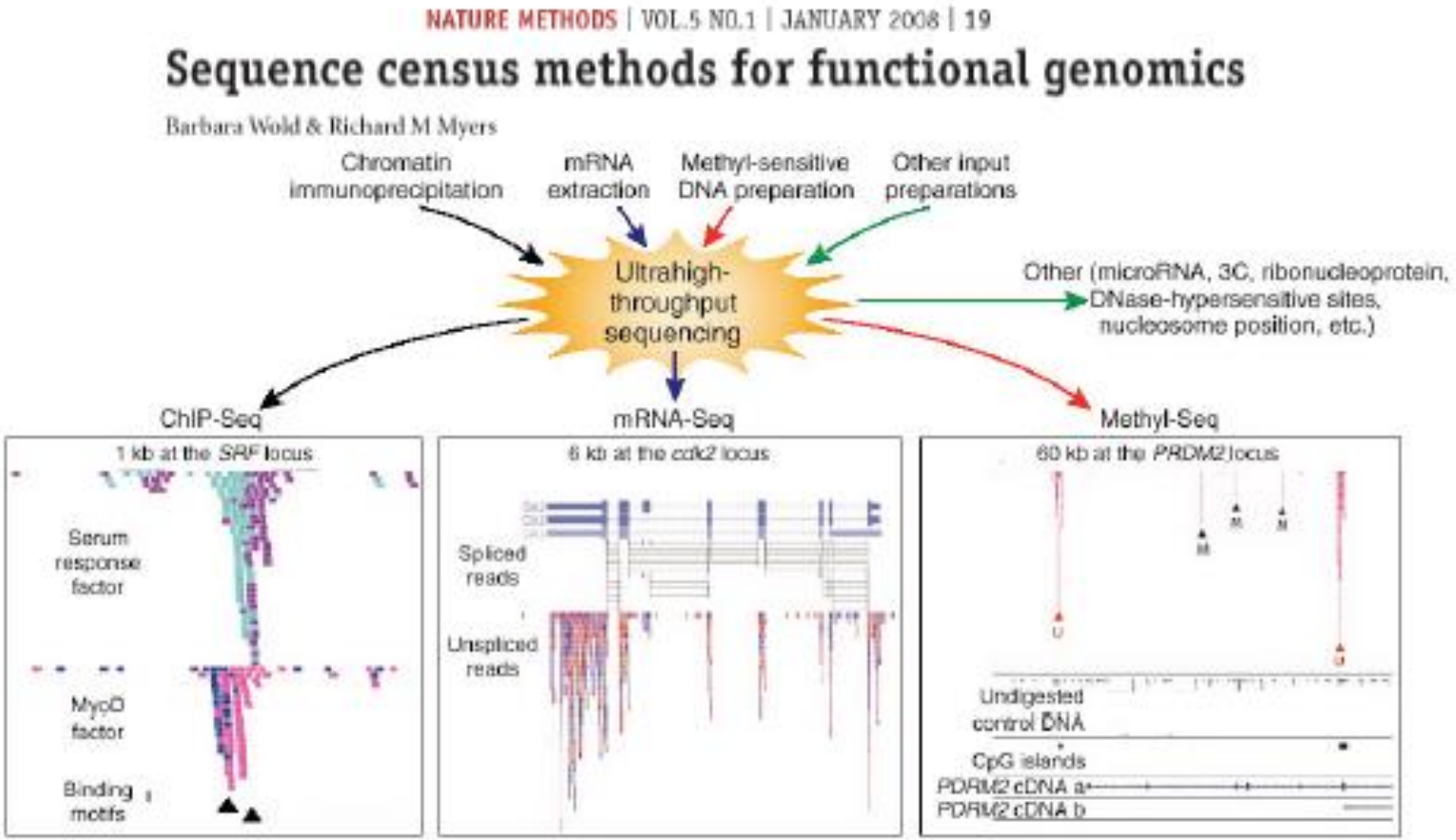
2. Understanding health and disease

Table 1 | **Potential frequencies of causal variants in complex traits**

Variant class	Minor allele frequency	Implications for analysis
Very common	Between 5 and 50%	Amenable to association analysis using current genome-wide association methods
Less common	Between 1 and 5%	Amenable to association analysis using variants catalogued in the 1000 Genomes Project
Rare (but not private)	Less than 1% but still polymorphic in one or more major human populations	Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families
Private	Restricted to probands and immediate relatives	Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants

DNA sequencing changed biology in many ways…

3. Identifying and quantifying rare transcripts, splicing, etc.



DNA sequencing changed biology in many ways…

Table 2 Applications of next-generation sequencing

Category	Examples of applications	Refs
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes	44
Reduced representation sequencing	Large-scale polymorphism discovery	45
Targeted genomic resequencing	Targeted polymorphism and mutation discovery	46–52
Paired end sequencing	Discovery of inherited and acquired structural variation	53,54
Metagenomic sequencing	Discovery of infectious and commensal flora	55
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	56–63
Small RNA sequencing	microRNA profiling	64
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA	60,65,66
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions	67–70
Nuclease fragmentation and sequencing	Nucleosome positioning	69
Molecular barcoding	Multiplex sequencing of samples from multiple individuals	61,71

Shendure and Ji, Nat. Biotech, 2008

DNA sequencing changed biology in many ways…

4. Identifying or classifying species (viruses, bacteria, etc).

Int. J. Mol. Sci. **2011**, *12*, 7861-7884; doi:10.3390/ijms12117861

OPEN ACCESS

Viruses **2011**, *3*, 1849-1869; doi:10.3390/v3101849

OPEN ACCESS

viruses

ISSN 1999-4915

www.mdpi.com/journal/viruses

Review

Applicati
Diagnosti

Luisa Barzon

Review

Next Generation Sequencing Technologies for Insect Virus
Dis

Siju

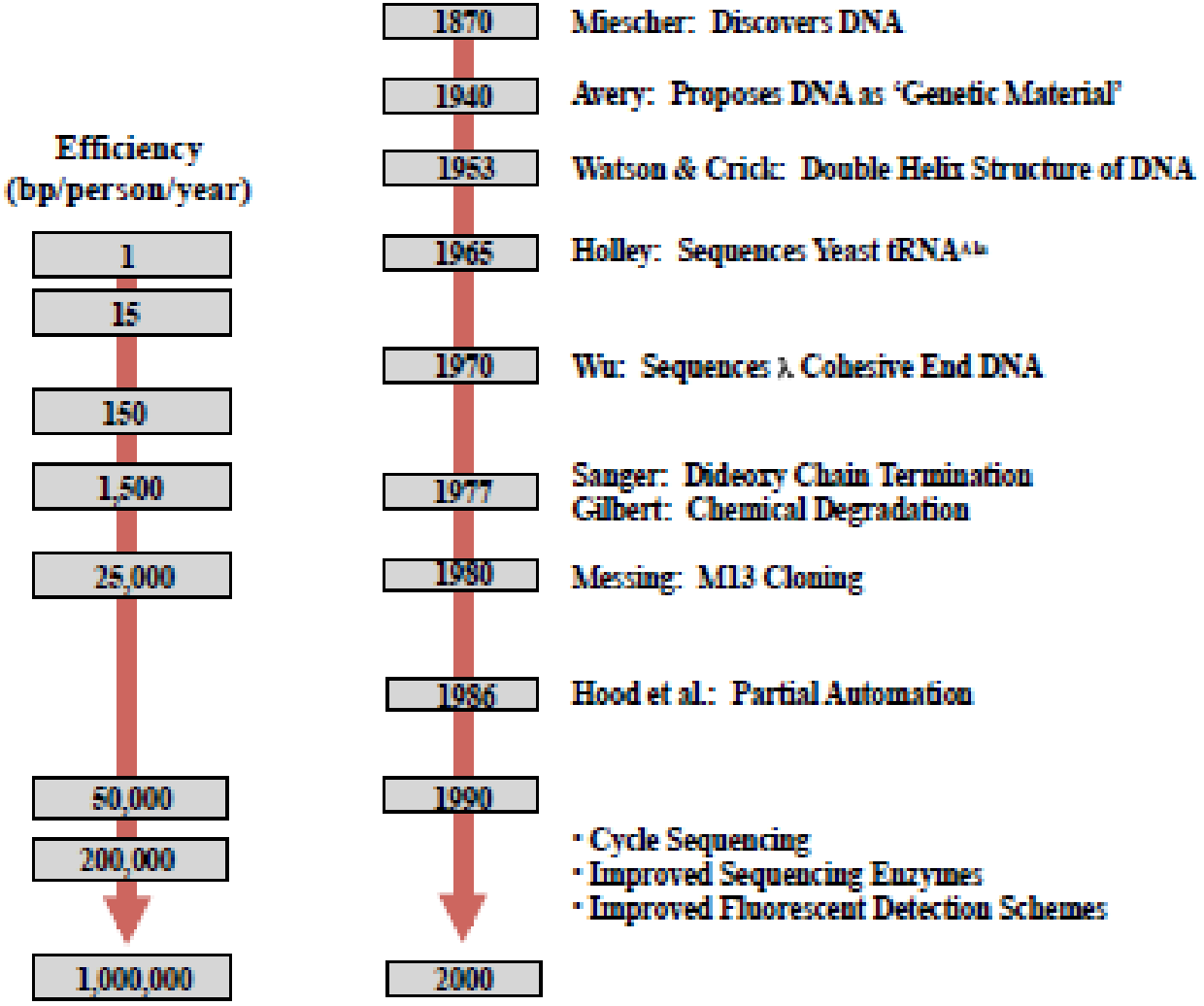
OPEN ACCESS Freely available online

PLoS one

Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach

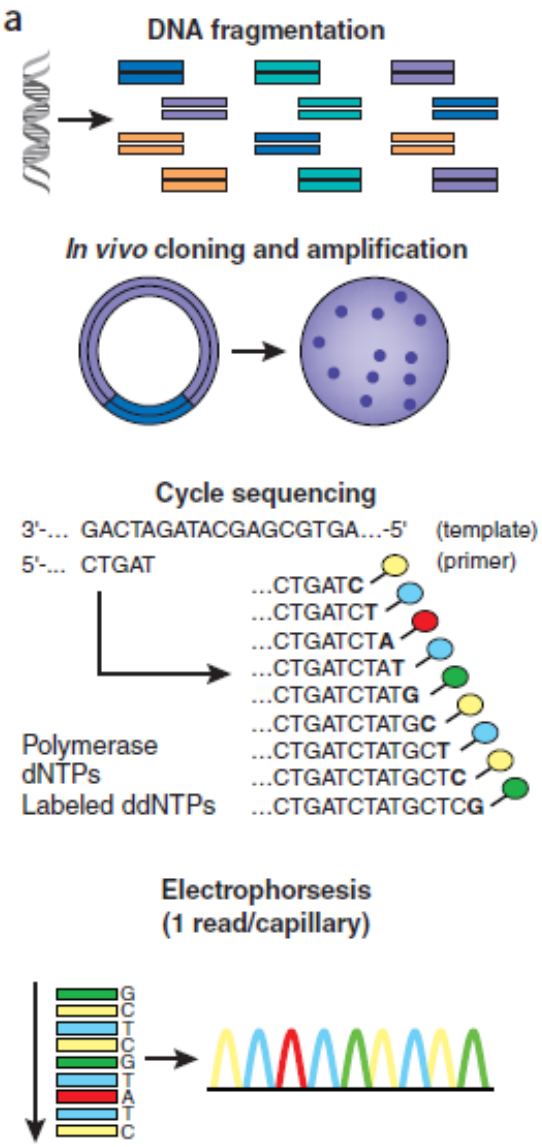
Shota Nakamura[1,9], Cheng-Song Yang[2,3,9], Naomi Sakon[4], Mayo Ueda[2,3], Takahiro Tougan[5], Akifumi Yamashita[1], Naohisa Goto[1], Kazuo Takahashi[4], Teruo Yasunaga[1], Kazuyoshi Ikuta[3], Tetsuya Mizutani[6], Yoshiko Okamoto[7], Michihira Tagami[8], Ryoji Morita[8], Norihiro Maeda[8], Jun Kawai[8], Yoshihide Hayashizaki[8], Yoshiyuki Nagai[7], Toshihiro Horii[2,5], Tetsuya Iida[2], Takaaki Nakaya[2]*

History of Nucleic Acid Sequencing

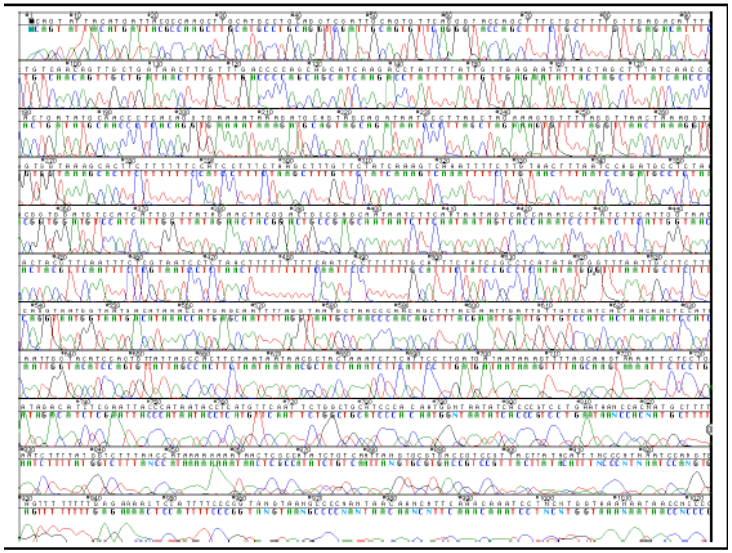


Adapted from Elliot Marguiles' 2/9/2010 NHGRI talk (http://www.genome.gov/12514288)

1st Generation: Sanger Sequencing



AB 3730 xl



Shendure and Ji, Nat. Biotech, 2008

2nd Generation: "Next Generation" Sequencing

Clonally amplified single molecules for sequencing

454

Pyrosequencing

Illumina HiSeq 2000

Reversible Terminator Chemistry

SOLiD

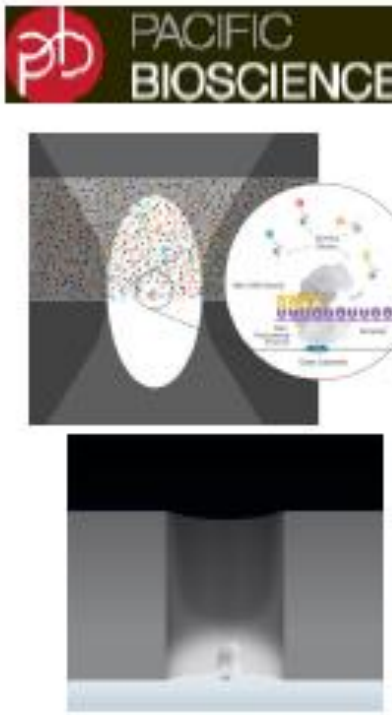
Ligation-based extension

Adapted from Elliot Marguiles' 2/9/2010 NHGRI talk (http://www.genome.gov/12514288)

3rd Generation: Next-Next Generation Sequencing

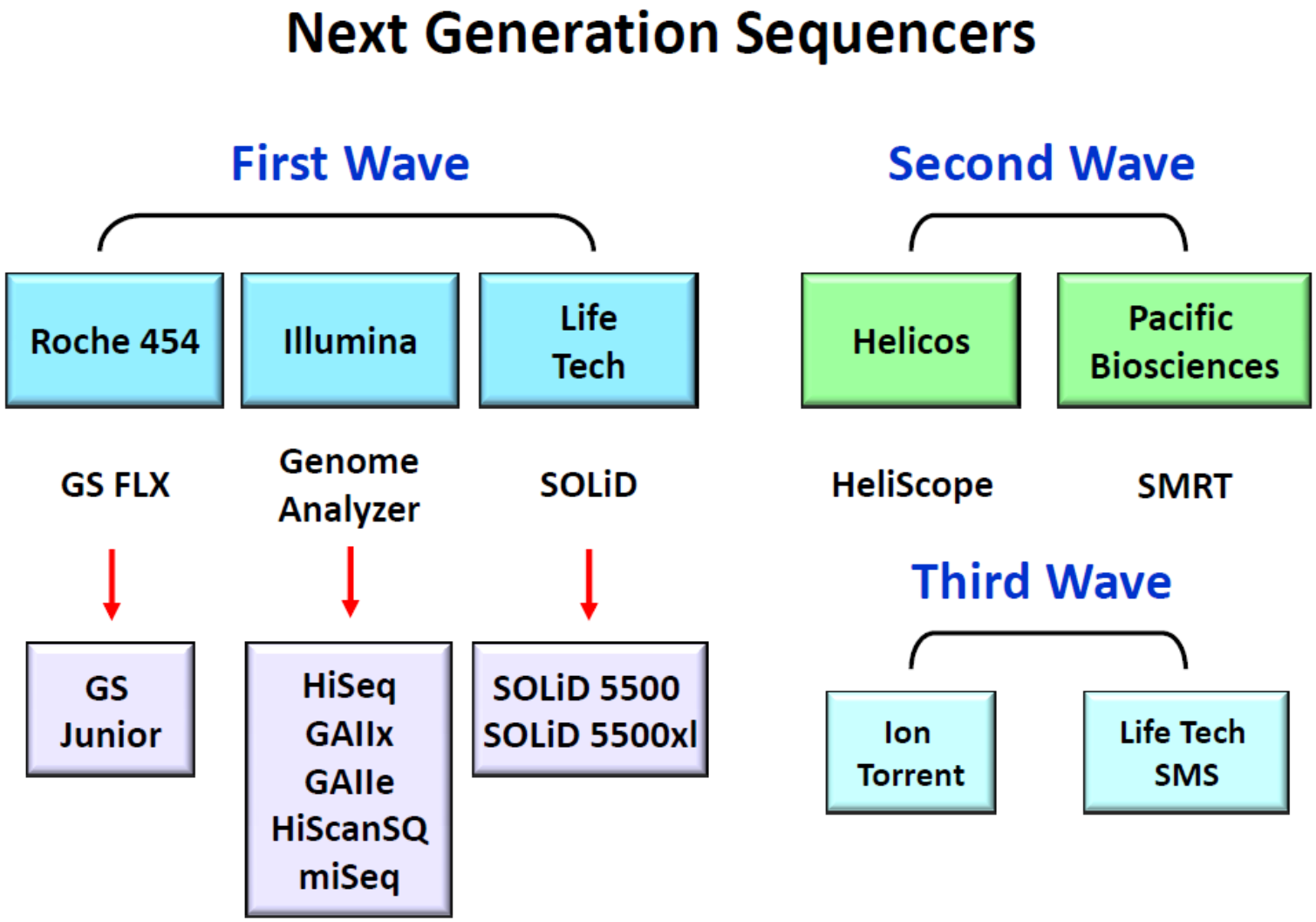
True Single Molecule Sequencing

HeliScope



SMRT Technology

The Waves of Next-Gen Sequencing



The Rate of DNA Sequencing Continues to Accelerate…

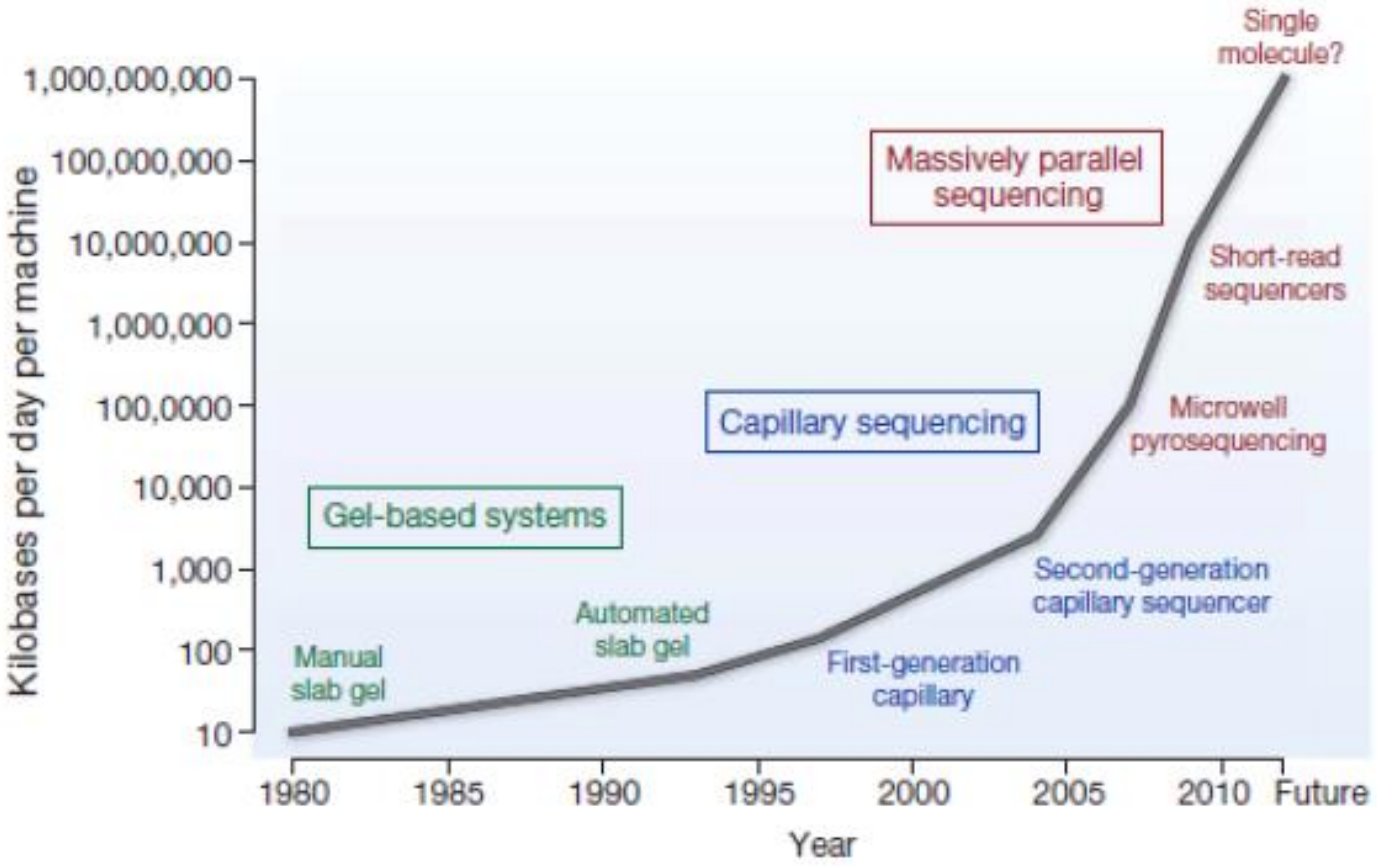
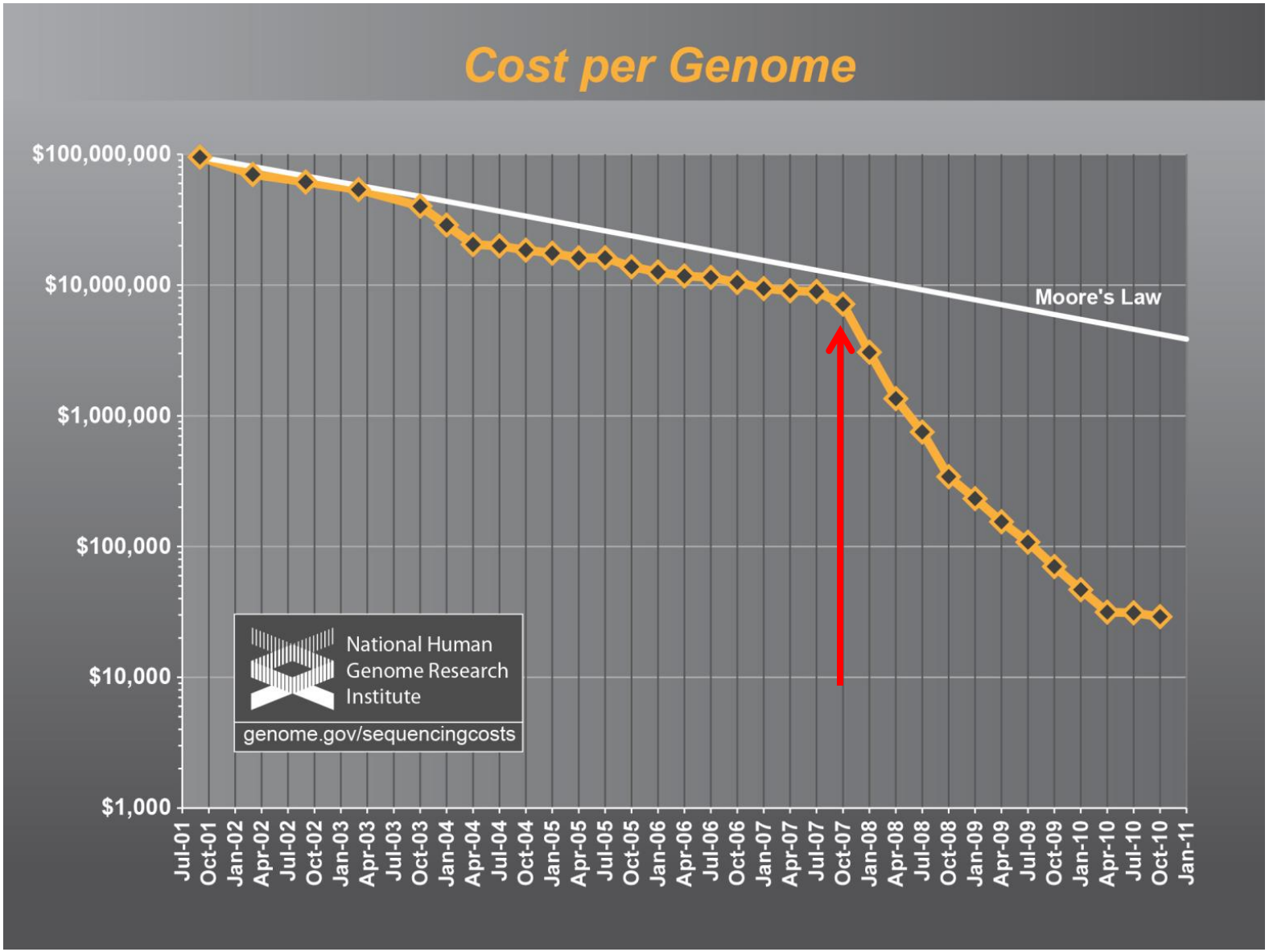


Figure 3 | Improvements in the rate of DNA sequencing over the past 30 years and into the future. From slab gels to capillary sequencing and second-generation sequencing technologies, there has been a more than a million-fold improvement in the rate of sequence generation over this time scale.

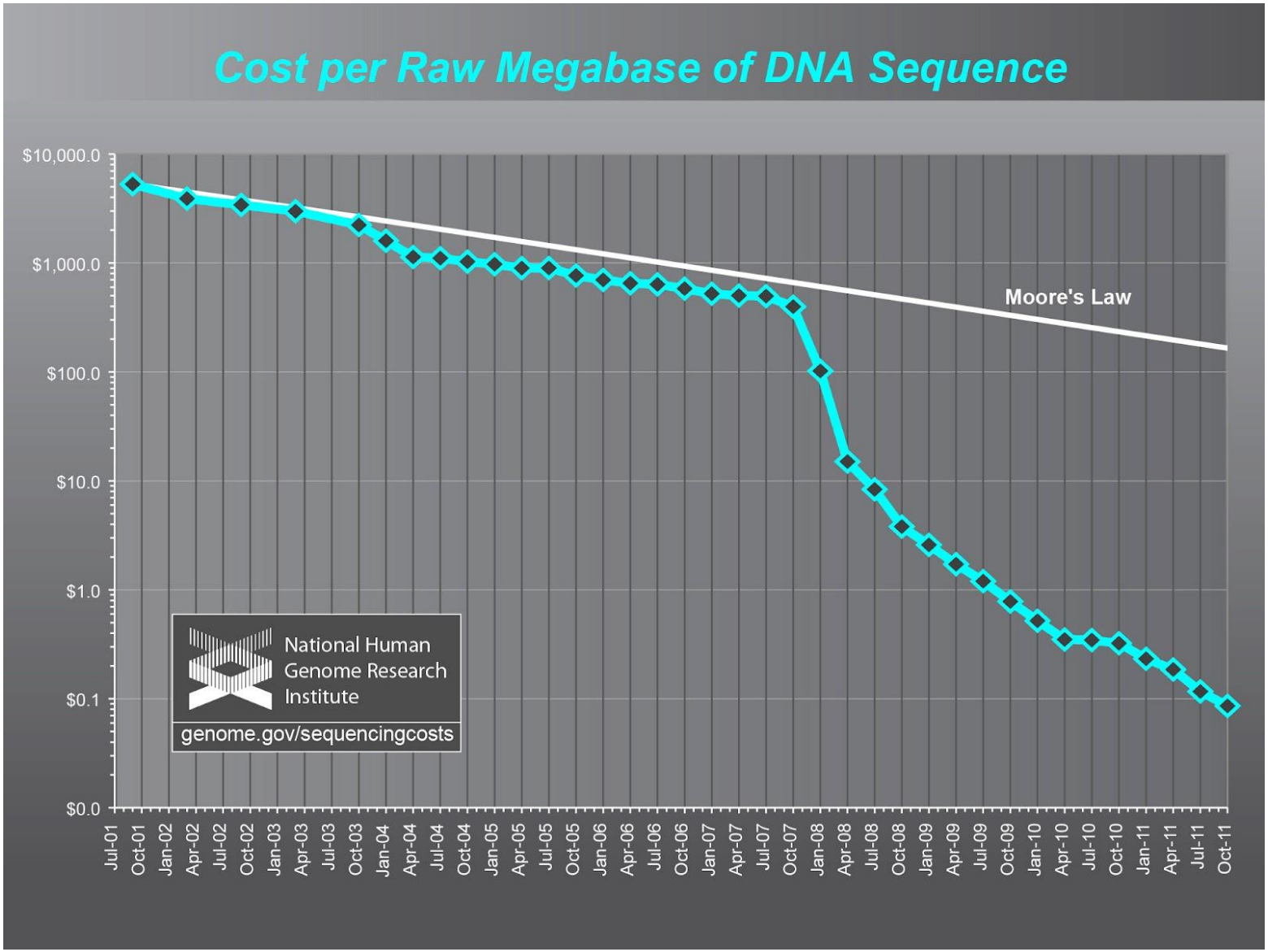
MR Stratton *et al*. *Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

… While Sequencing Costs Decline



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program

… While Sequencing Costs Decline



NATURE REVIEWS | GENETICS VOLUME 11 | JANUARY 2010 | 31

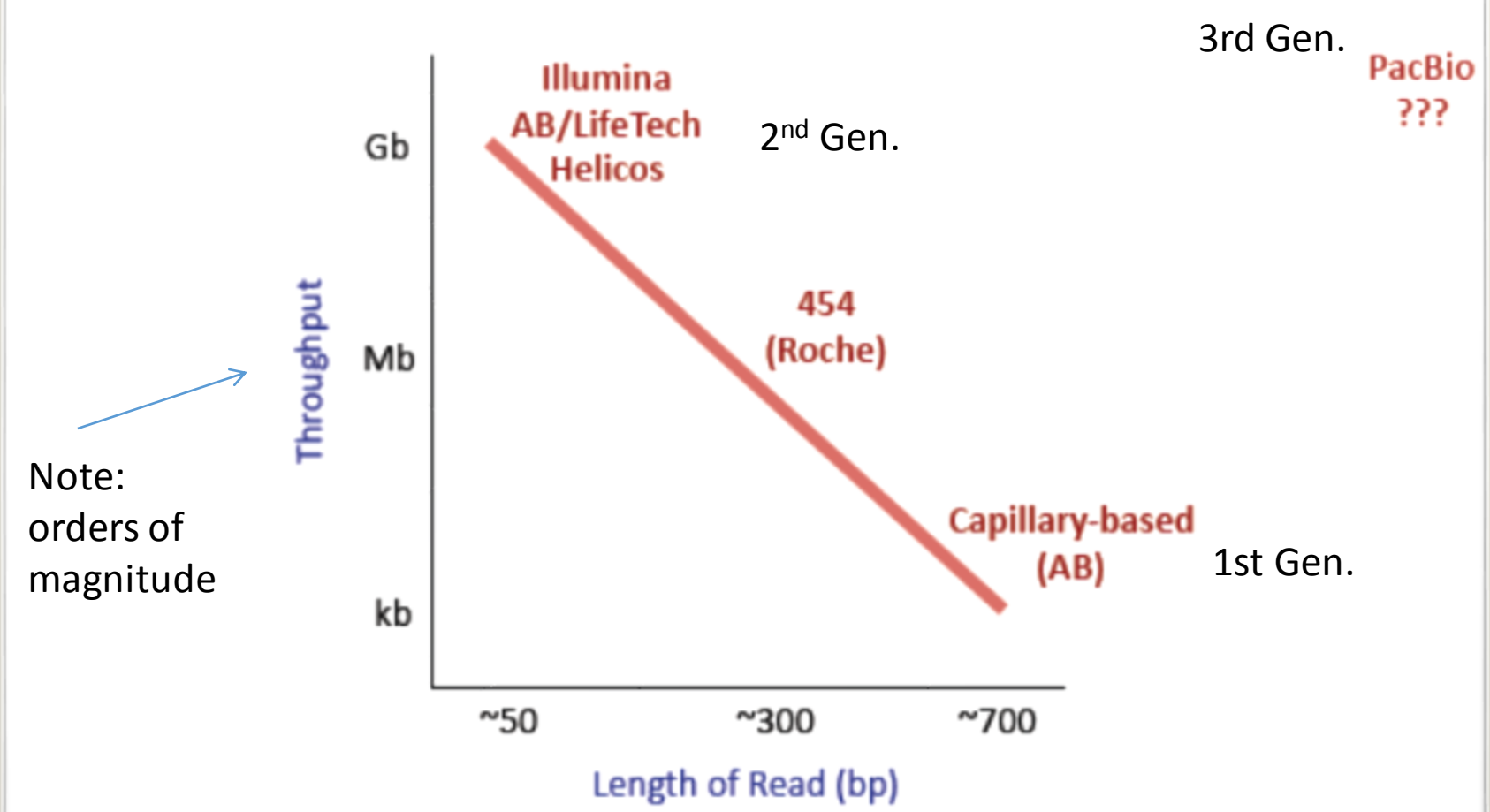
APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

Michael L. Metzker * ‡

- Akin to early days of PCR
- Enormous volumes of data cheaply
  - Different scale, however
  - 1 billion reads per run

Trade-offs


3rd Gen.
PacBio
???
Illumina
AB/LifeTech
Helicos
2nd Gen.
454
(Roche)
Capillary-based
(AB)
1st Gen.
Throughput
Gb
Mb
kb
~50
~300
~700
Length of Read (bp)
Note:
orders of
magnitude
Throughput = Amount of Sequence Generated / Unit of Time or Cost

Comparison of methods

Table 1: (a) Advantage and mechanism of sequencers. (b) Components and cost of sequencers. (c) Application of sequencers.

(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

Liu et al. Journal of Biomedicine and Biotechnology. Volume 2012 (2012), Article ID 251364, 11 pages

Comparison of methods

(b)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

Liu et al. Journal of Biomedicine and Biotechnology. Volume 2012 (2012), Article ID 251364, 11 pages

Comparison of methods

(c)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Resequencing		Yes	Yes	
De novo	Yes	Yes		Yes
Cancer	Yes	Yes	Yes	
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes		
Mutation detection	Yes	Yes	Yes	Yes

Liu et al. Journal of Biomedicine and Biotechnology. Volume 2012 (2012), Article ID 251364, 11 pages

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel

required for it to occur. I therefore offer the following as food for thought.

One source of difficulty in using resequencing approaches for diagnosis centers on the need to improve the quality and completeness of the human reference genome. In terms of quality, it is clear that the clone-

There is job security in bioinformatics…

Data source	Data size	Bioinformatics topics
Raw DNA sequence	11.5 million sequences (12.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	400,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	15,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations, molecular movements, docking predictions)
Genomes	300 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes in yeast	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

Luscombe et al., Yearbook of Medical Informatics 2001

Updated numbers, 2014

Intro to Bioinformatics

4 day short course

Tuesday, September 30, 2014

- Retrieving information on genes and proteins from biological and genomic databases
- Predicting genes from DNA sequences
- Identifying promoters and regulatory elements in DNA sequences

Intro to Bioinformatics

4 day short course

Wednesday, October 1, 2014

- Analyzing protein sequences
- Comparing protein and DNA sequences
- Visualizing and analyzing protein structures

- Functional annotations and predictions
- Predict function
- Compare/contrast functional prediction tools

Intro to Bioinformatics

4 day short course

Thursday, October 2, 2014

- Place function in the context of biological pathways
- Pulling information from multiple sources together
- Methods and Applications
  - Genome-phenome analysis

Intro to Bioinformatics

4 day short course

Friday, October 3, 2014

- Bioinformatics pipelines and workflows
- Understanding the problems associated with analyzing large datasets
- Resources to go to in the future → the field moves fast

Questions???