Sequencing Technology

Lexington Senior Center

December 4, 2013

Allan Kleinman

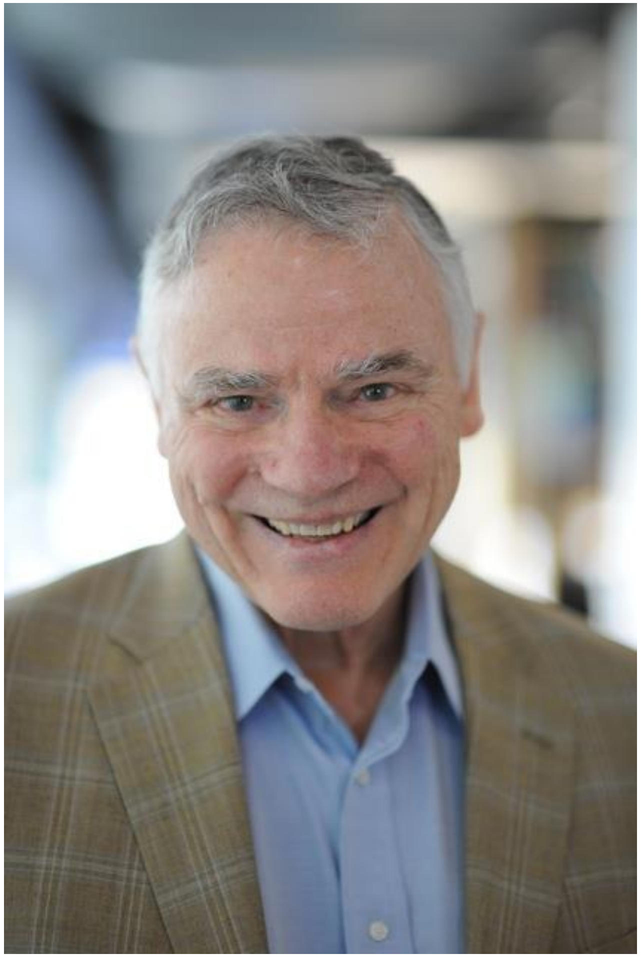
Disclaimer

- I am a retired engineer and volunteer tour guide at Jackson Laboratory in Bar Harbor, ME
- I first got interested in Bioinformatics and Computational Biology 15 years ago
- I am not a Doctor and cannot dispense medical advice – but I can refer you to literature
- I am not a geneticist, nor a biologist – but have arranged for them to answer questions that come up during this talk that I cannot answer

Talk Overview

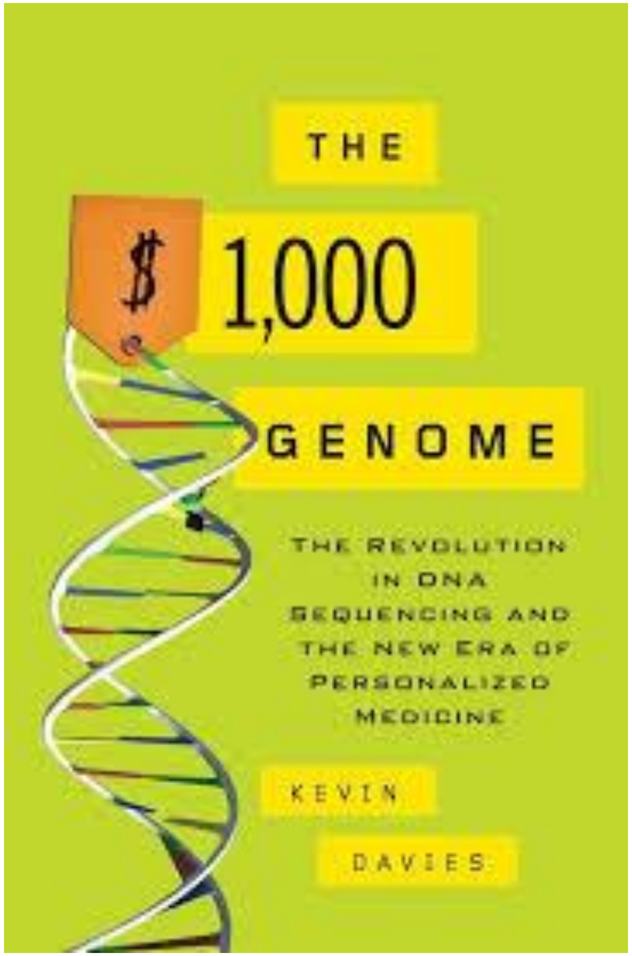
- Future of Medicine and Technology
- Review Cell and DNA Basics
- Sequencing Technology - Current and Future
- Computer Challenges and Response
- Personalized Medicine in the News
- No Synthetic Biology Today

P4 Medicine - Leroy Hood



- Predictive
  - Know what’s coming
- Personalized
  - Specific to your genome
- Preventive
  - Avoid getting sick
- Participatory
  - Advocate for your health

The $1000 Genome


THE
$ 1,000
GENOME
THE REVOLUTION IN DNA SEQUENCING AND THE NEW ERA OF PERSONALIZED MEDICINE
KEVIN
DAVIES

- $1000 is a Tipping Point
- Makes Personalized Medicine Possible
- Will Lead to Ubiquitous Genomic Sequencing Replacing Single Gene Tests
- Kevin Davies Lives in Lexington

Creative Destruction of Medicine by Eric Topol

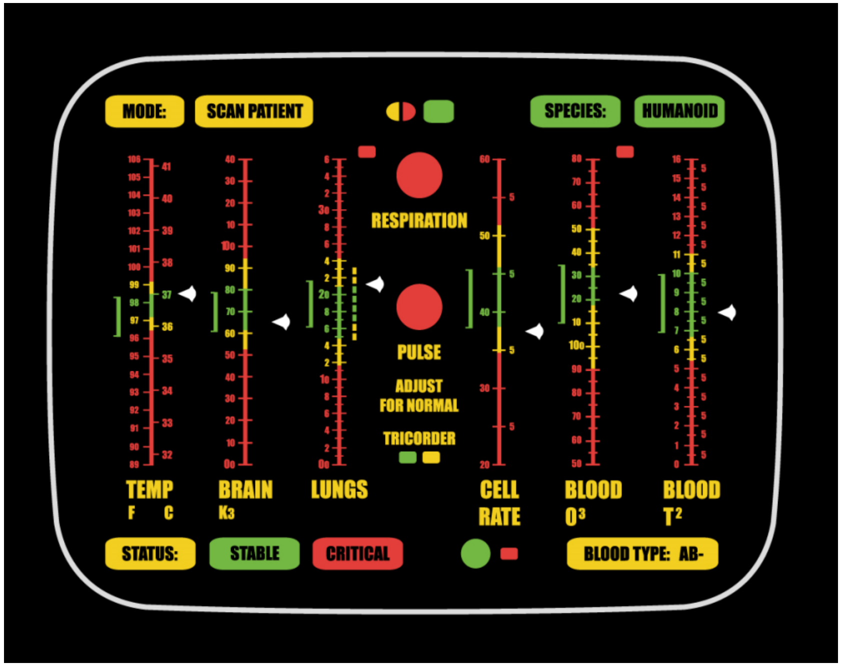
- Phoebe Reads Topol
- Convergence Coming
  - Medical sensors
  - Smart phones
  - Wireless communication
  - Genome sequencing
  - Electronic Health Records in the “Cloud”
- Need Consumer Advocates to Push to Realize Medical Benefits

Future of Medicine - Star Trek ?

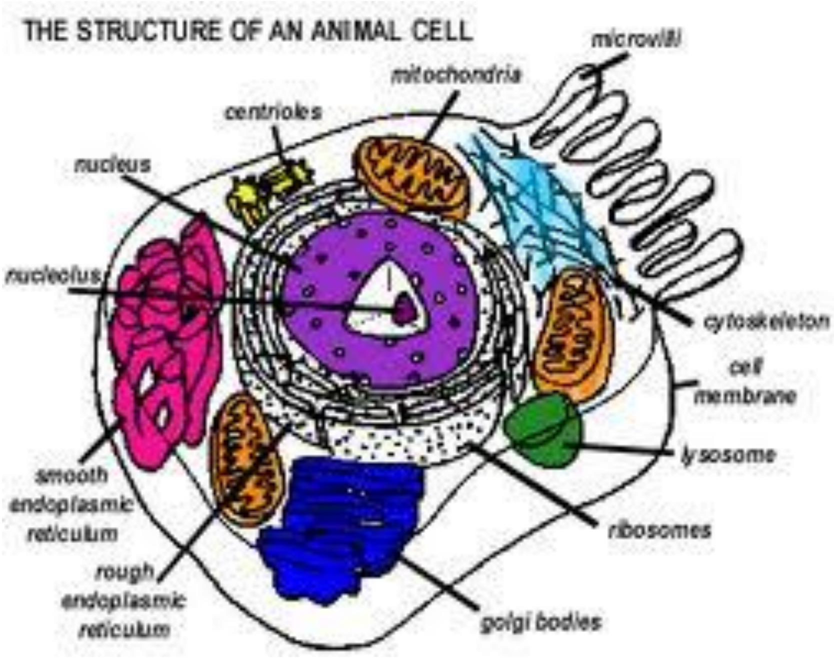
Handheld DNA Analyzers

Handheld Protein Analyzers

Is the Tricorder coming?

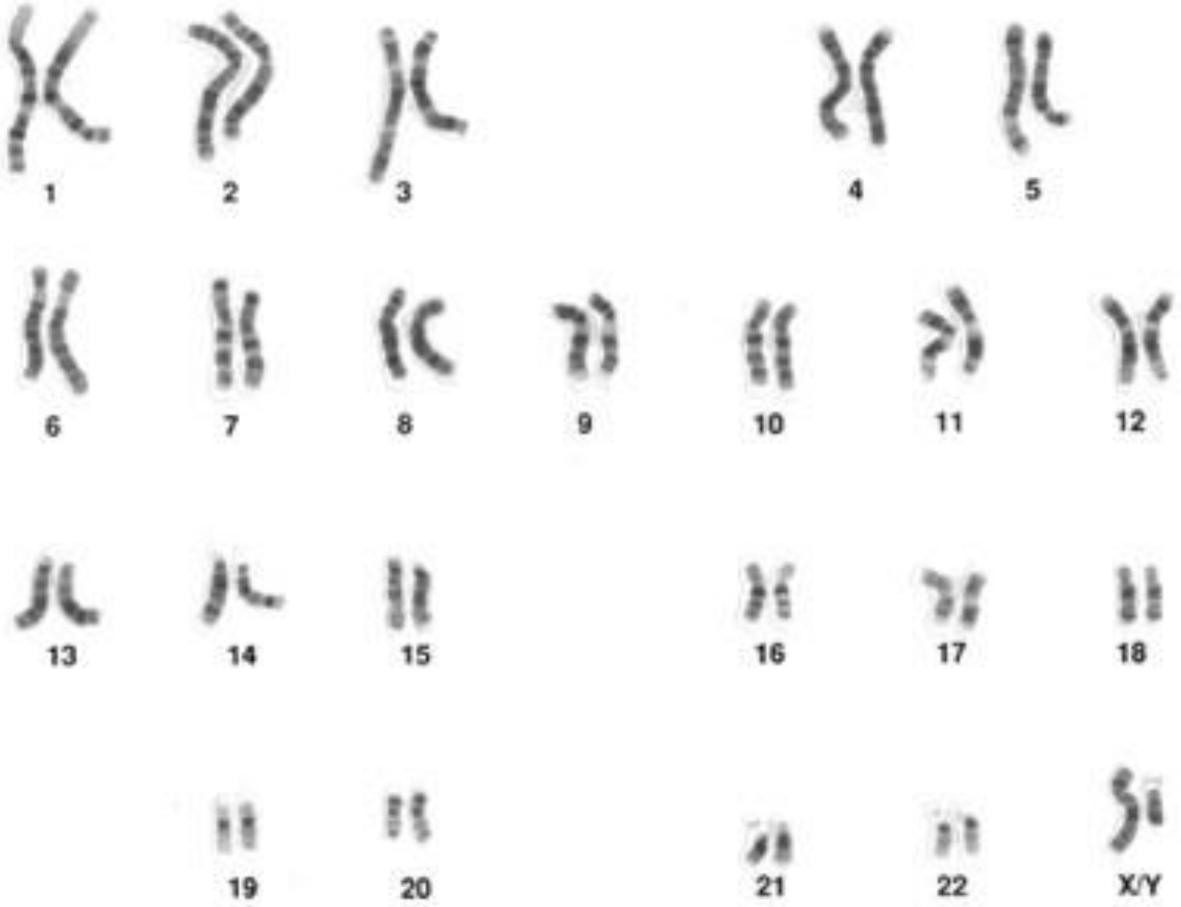


Cell Basics

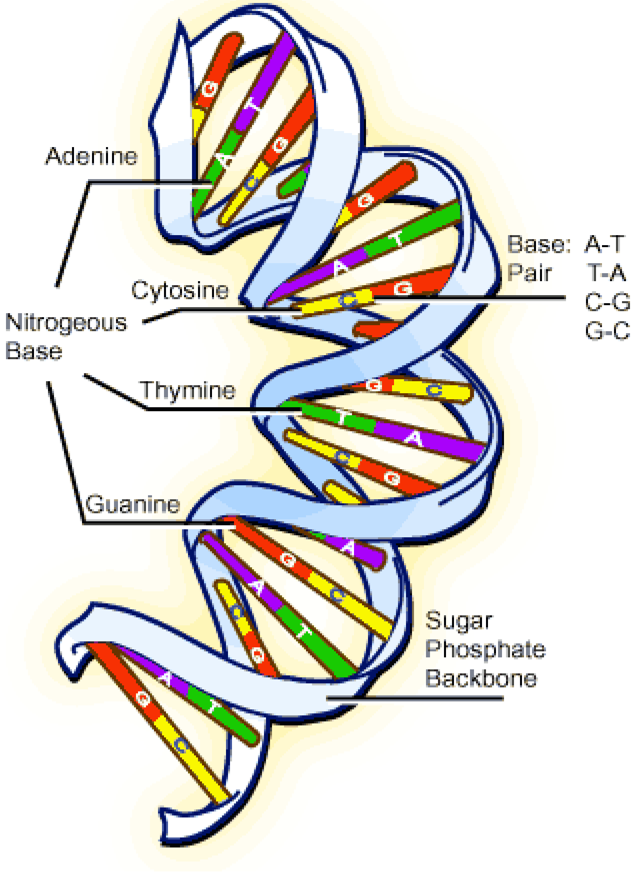


- Our bodies contain 10 trillion cells of 200 types
- Eukaryotic cells have a nucleus, Prokaryotics (bacteria) do not
- Our DNA resides in the nucleus on 46 chromosomes
- Proteins = “workhorses”
  - Structural elements
  - Muscles
  - Enzymes – speed reactions
  - Signals – turn on/off DNA

Chromosome Map

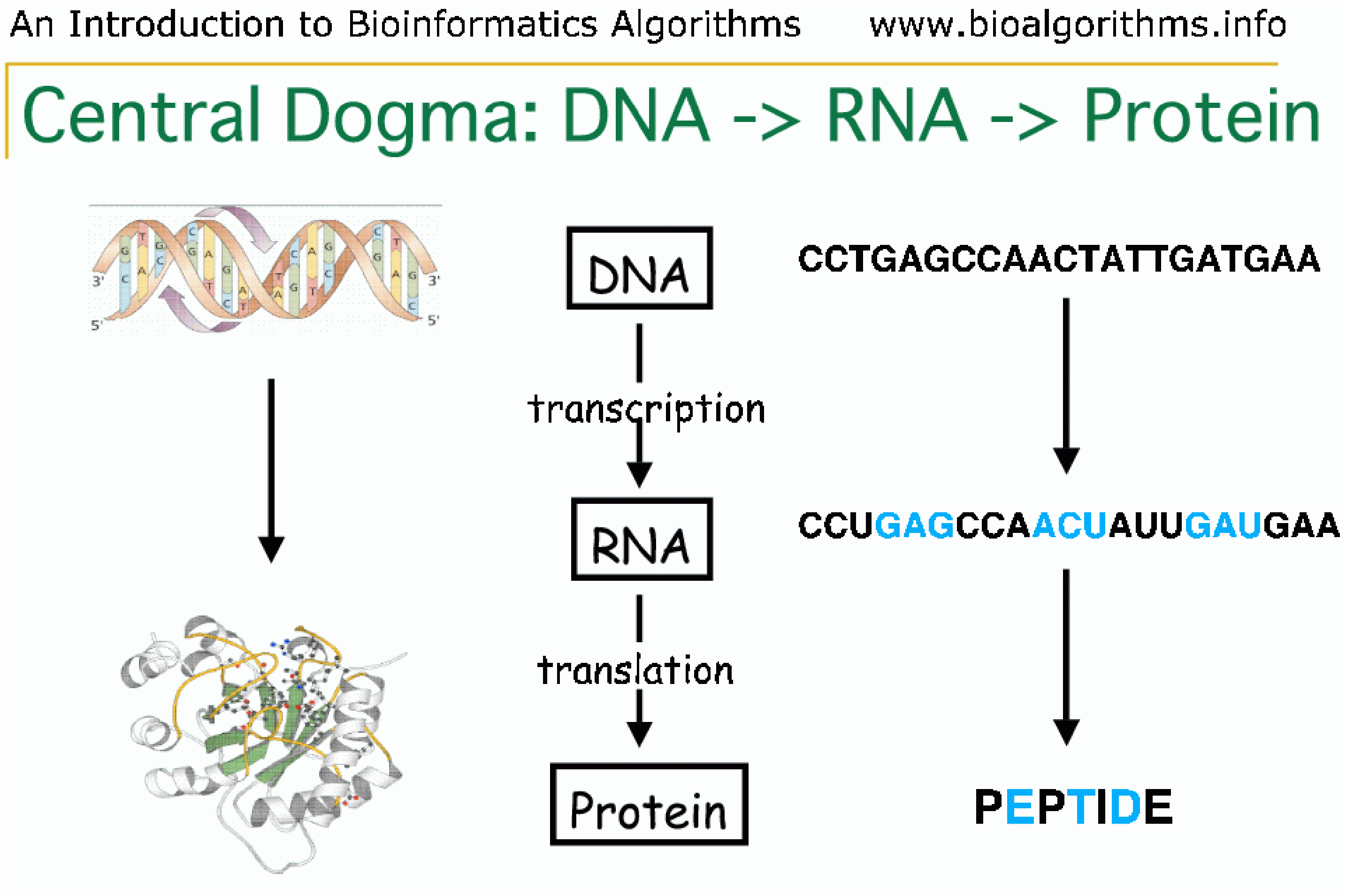


DNA Basics



- Human DNA has 3 billion base pairs
- Arranged in roughly 23,000 genes
- We also have about 20,000 non-coding DNA control elements and large areas of “junk DNA”
- Humans have about 3 million Single Nucleotide Polymorphisms (SNPs), a difference in one base
- Genes have related promotor controllers and exons that get copied to make mRNA

Central Dogma of Molecular Biology



An Introduction to Bioinformatics Algorithms www.bioalgorithms.info

Central Dogma: DNA -> RNA -> Protein

DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

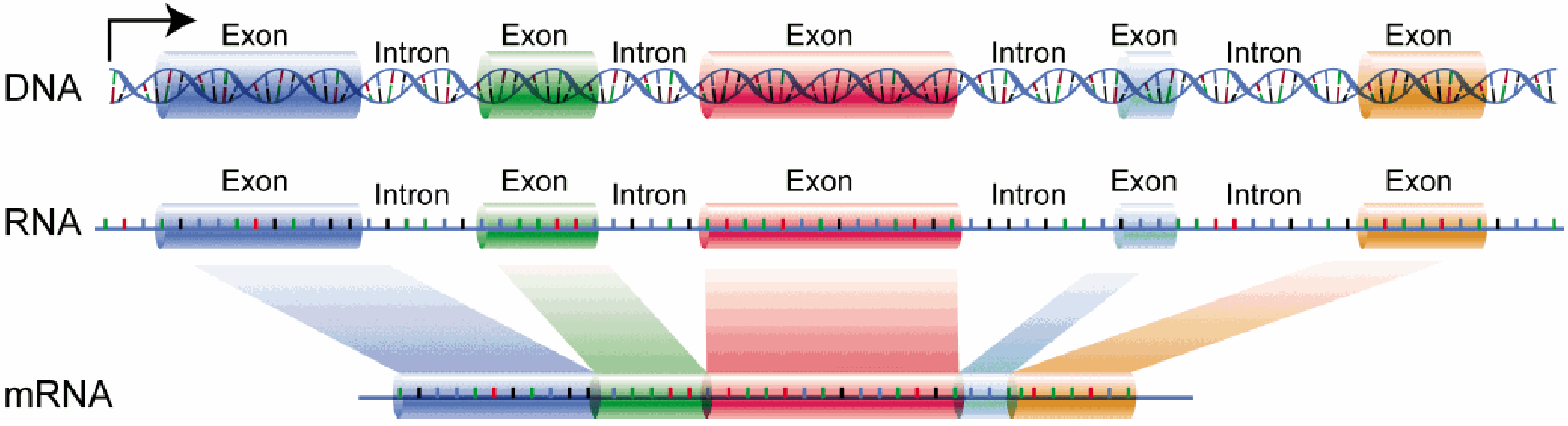
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Findings of the Human Genome Project

- Basic Facts
  - 23,500 Genes
  - 270,000 Exons
  - 20,000+ non-coding regulatory sections
  - 200,000+ proteins
- “Vestigial” DNA Blocks from Ancient Viruses
- Mobile DNA Segments During Cell Division
- Copy Number Variations (CNVs)

Anatomy of a Gene

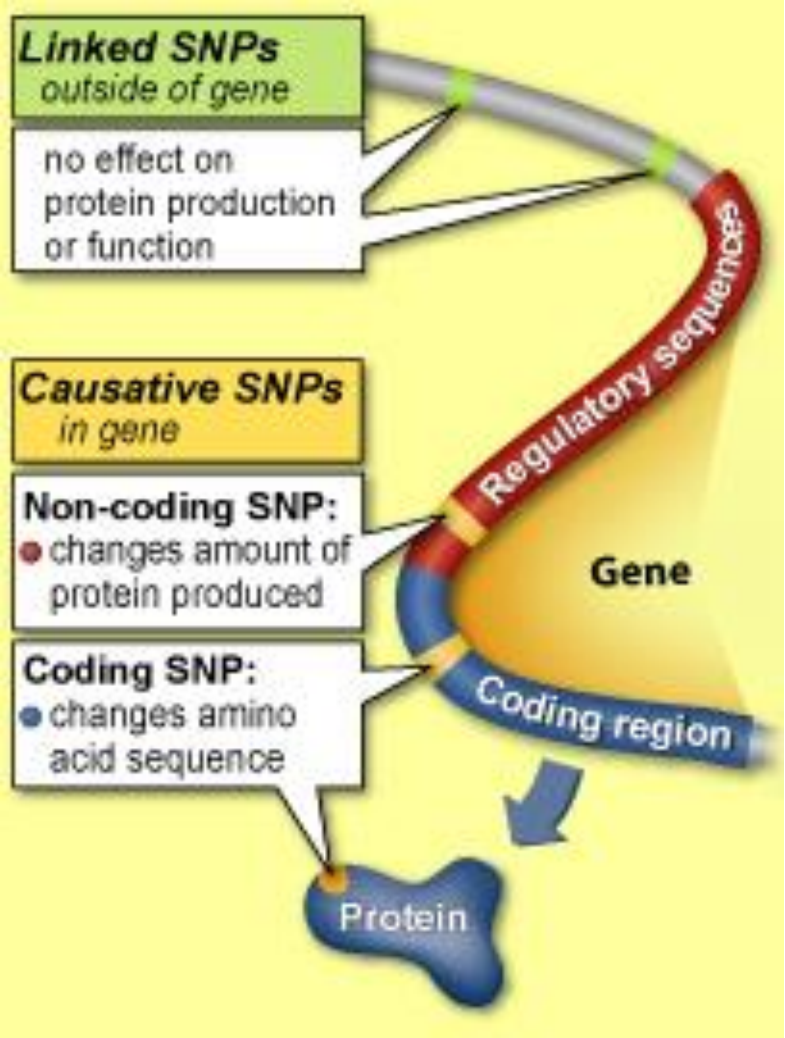


Promoter/Regulatory Section Before Gene

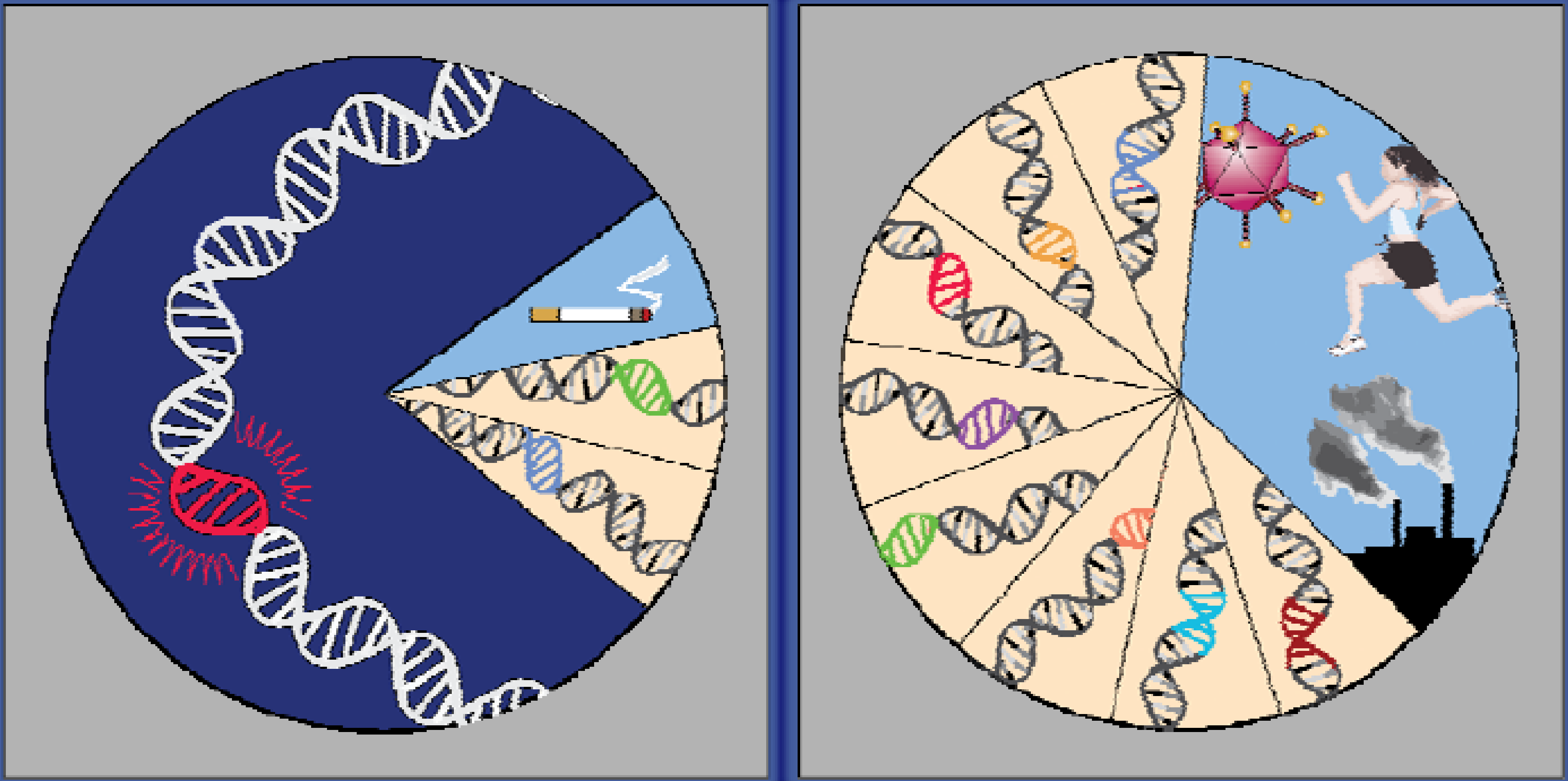
- Selects Alternative Exon Splicing
- Controls Amount of Protein Produced

RNA Interference (RNAi) = Gene Silencing

Non-Coding DNA Can Change Protein Production



Genomic Architecture of Genetic Diseases



Rare, Simple, Monogenic, Mendelian…

Common, Complex, Multigenic, Non-Mendelian…

Mostly *Coding* Mutations

Mostly *Non-Coding* Mutations

feature

A vision for the future of genomics research

A blueprint for the genomic era

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Guyer on behalf of the US National Human Genome Research Institute*

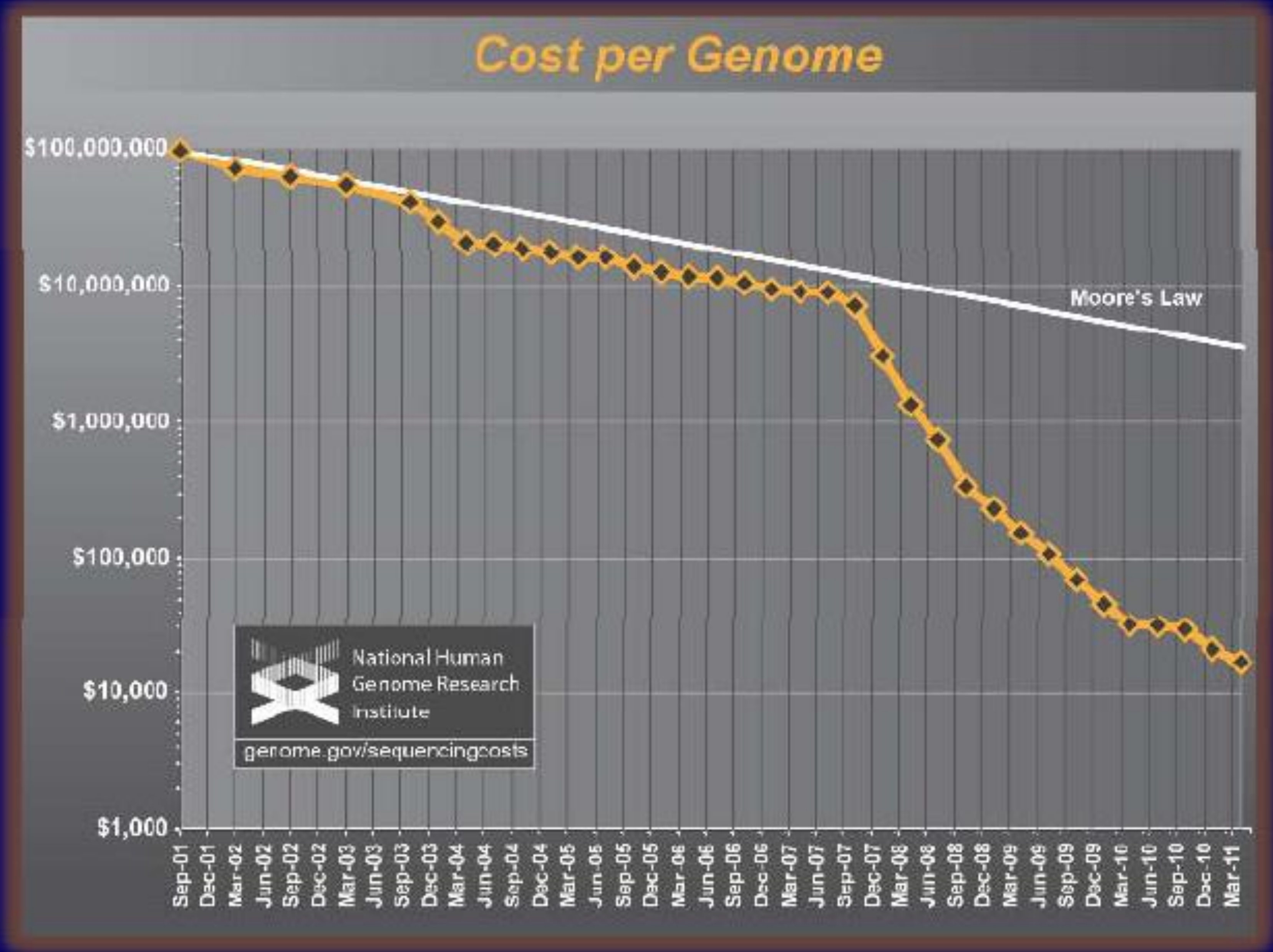
in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler

"…'technological leaps' that seem so far off as to be almost fictional but which, if they could be achieved, would revolutionize biomedical research and clinical practice.

[For example,]… the ability to sequence DNA at costs that are lower by four to five orders of magnitude than the current cost, allowing a human genome to be sequenced for $1,000 or less."

Nature, April 2003

Cost per Sequenced Human Genome



Ten Years On — The Human Genome and Medicine

Harold Varmus, M.D.

On a June day nearly 10 years ago, the leaders of the United States and the United Kingdom, accompanied by the leaders of the public and private teams deciphering the human genome, announced that a draft sequence had been completed. That occasion was rich with promises of new and more powerful ways to understand, diagnose, prevent,

some strong genetic markers for assessing drug responsiveness, risk of disease, or risk of disease progression — have entered routine medical practice. And most of these can be traced to discoveries that preceded the unveiling of the human genome. As Francis Collins, formerly the leader of the publicly funded sequencing efforts, recently commented: "the consequences for clinical medicine . . . have thus far been modest . . . the Human Genome Project has not yet directly affected the health care of most individuals."[2]

In this issue, the *Journal* begins another series of articles on genomic medicine.[3] Is it appropriate for the *Journal* to be taking stock so soon? It is, and for the following reasons.

First, readers will want to know the state of

influential haplotypes, and in general, other implicated susceptibility haplotypes collectively account for only a small fraction of the apparent heritable risk. Clearly, more than one decade of genomics will be required to understand the inborn risks of most common disorders, such as diabetes and hypertension.

Second, readers will enjoy learning from these articles how rapidly the engines of genomics and

Physicians are still a long way from submitting their patients' full genomes for sequencing, not because the price is high, but because the data are difficult to interpret.

N ENGL J MED 362;21 NEJM.ORG MAY 27, 2010

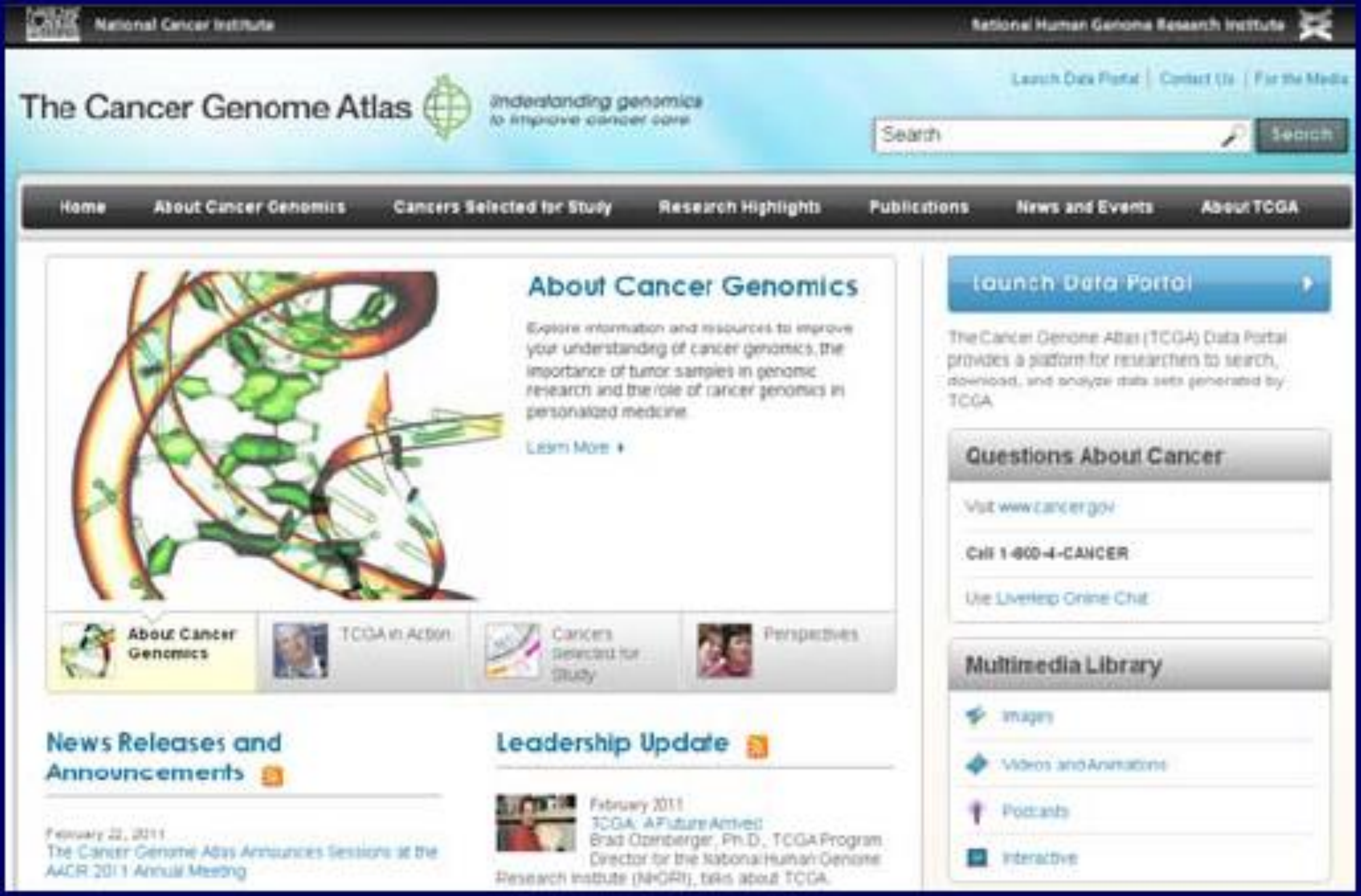
***NEJM* (2010)**

The Informational Bottleneck

```
TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGAT                                                                   GTCTGGCGGACCCTGA
CGCGAAGGAGGGTCT                                                                   TGGACCTAAAGAGAGG
CCGCGACTGTCGCCC                                                                   AGGGAGGCTGGGAGTC
AGAATCGGGAAAGGG                                                                   GTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGC                                                                   CAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGG                                                                   GCACCCAGAGTAGTAG
GTCTTTGGCATTAGG                                                                   TGGAAAAGGCCAGCGT
TGTCTCCAAACTTTT                                                                   GTGTATGGGTTGGGTT
TGGGGTAAAGGAATA                                                                   AAAACAGAAAGCATTA
AGAAGAGATGGAAGA                                                                   ACTCAAGTACGCTACT
ATGCACTTGTTTTAT                                                                   CCCCTTCATGCCTTGG
ACACTTGATTGTCTT                                                                   TCAGCCAACAAAAATT
TTGGGGTAGGTAGAA                                                                   GATTCTTCAAAAATTG
AAAGCAAATTTGTTG                                                                   CCGAAGTTATATCCAA
CTGACATTTAATAAA                                                                   TAGCATCTAAGTTCGG
AATCTTAGGCAAAGT                                                                   TATTATACTGGTGTGA
ATGAATGAATAGGTA                                                                   AAAAAGTCAAATATGT
TATAAATAGCTCATA                                                                   CAGTTAATCCTGGAAC
TCCGGTGCTAAGGAG                                                                   AATTATCTTTTTGTGT
TGATGTTATCCACCT                                                                   AAATGTTAATTGGCAT
AAATTATAGACTTTT                                                                   GAATATTTCATGGATA
GTTCTAAATACTAAT                                                                   ATCACCTGACACATTT
AATATAGGTTAAAAA                                                                   CTCATTTCTGTTCTCC
AAAATATTTCATAAG                                                                   CCTAAAATACCAATGA
TTACAAACTTCCTTC                                                                   TTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGA                                                                   CCTTAACATCTCTGTG
TGTGACTTGACCTTT                                                                   GTT------CTATTATT
ATGGATTACCATATT                                                                   TTTTGTGACTCTCAAT
CTGGATATGCAATGA                                                                   GGAAACACGTCACATG
TTTCTATTGTATGTT                                                                   AAAATTATTATGGTAT
TTACAAACTTCCTTC                                                                   TTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGA                                                                   CCTTAACATCTCTGTG
```

GTACT
CAAGT
TTAC
CTA
ATCC
TAC
CCGA
TAGG
ATCC
TGG
ATCC
GATC
CTC
A
S. Harris

The Future: Genome Sequencing



National Cancer Institute

National Human Genome Research Institute

Launch Data Portal | Contact Us | For the Media

The Cancer Genome Atlas

understanding genomics to improve cancer care

Search

Search

Home | About Cancer Genomics | Cancers Selected for Study | Research Highlights | Publications | News and Events | About TCGA

About Cancer Genomics

Explore information and resources to improve your understanding of cancer genomics, the importance of tumor samples in genomic research and the role of cancer genomics in personalized medicine.

Learn More ›

About Cancer Genomics | TCGA in Action | Cancers Selected for Study | Perspectives

News Releases and Announcements

February 22, 2011
The Cancer Genome Atlas Announces Sessions at the AACR 2011 Annual Meeting

Leadership Update

February 2011
TCGA: A Future Arrived
Brad Ozenberger, Ph.D., TCGA Program Director for the National Human Genome Research Institute (NHGRI), talks about TCGA.

Launch Data Portal ›

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use LiveHelp Online Chat

Multimedia Library

Images

Videos and Animations

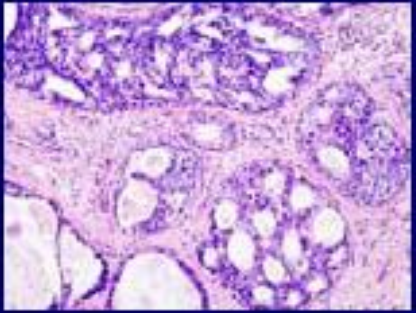
Podcasts

Interactive

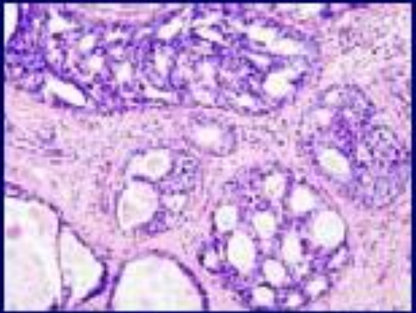
Cancer Genomics

Genomic Medicine: Cancer Diagnostics

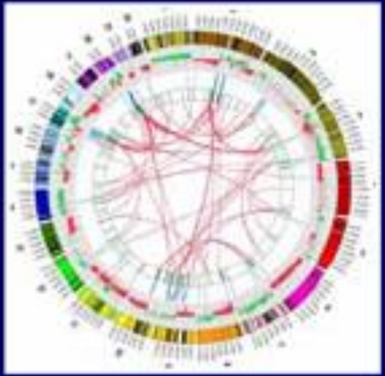
Now



Future



+



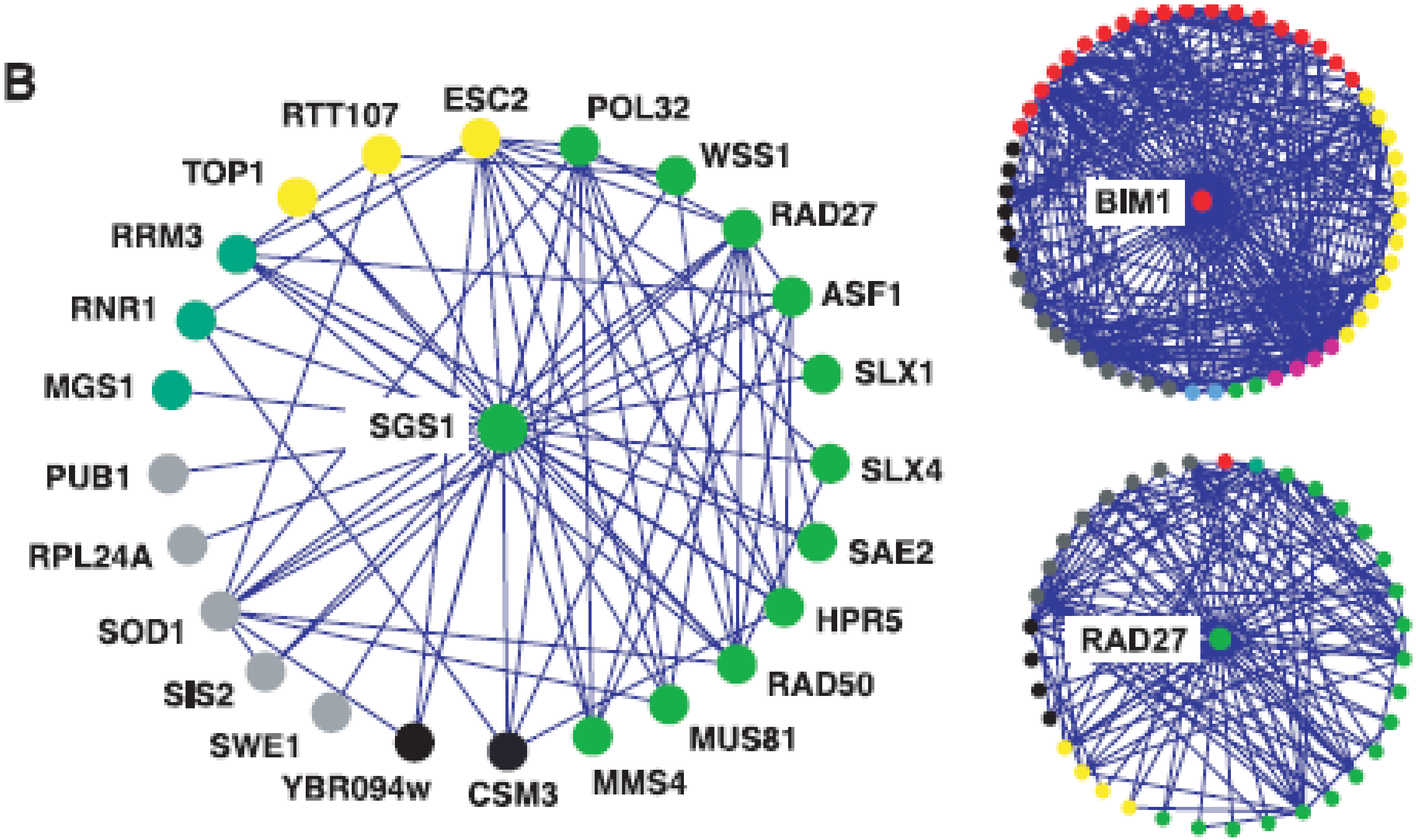
Computer Models and Databases

- Needed to Explain Mechanism of Disease
  - Define/Model disease mechanisms and simulate
  - Guides strategy for determining therapies/drugs
  - Required for FDA drug approval
  - Validate in animal models and clinical trials
- Example Databases
  - GENMAP, ENCODE, MGD, ENTREZ, …..
  - Cancer Genome Atlas

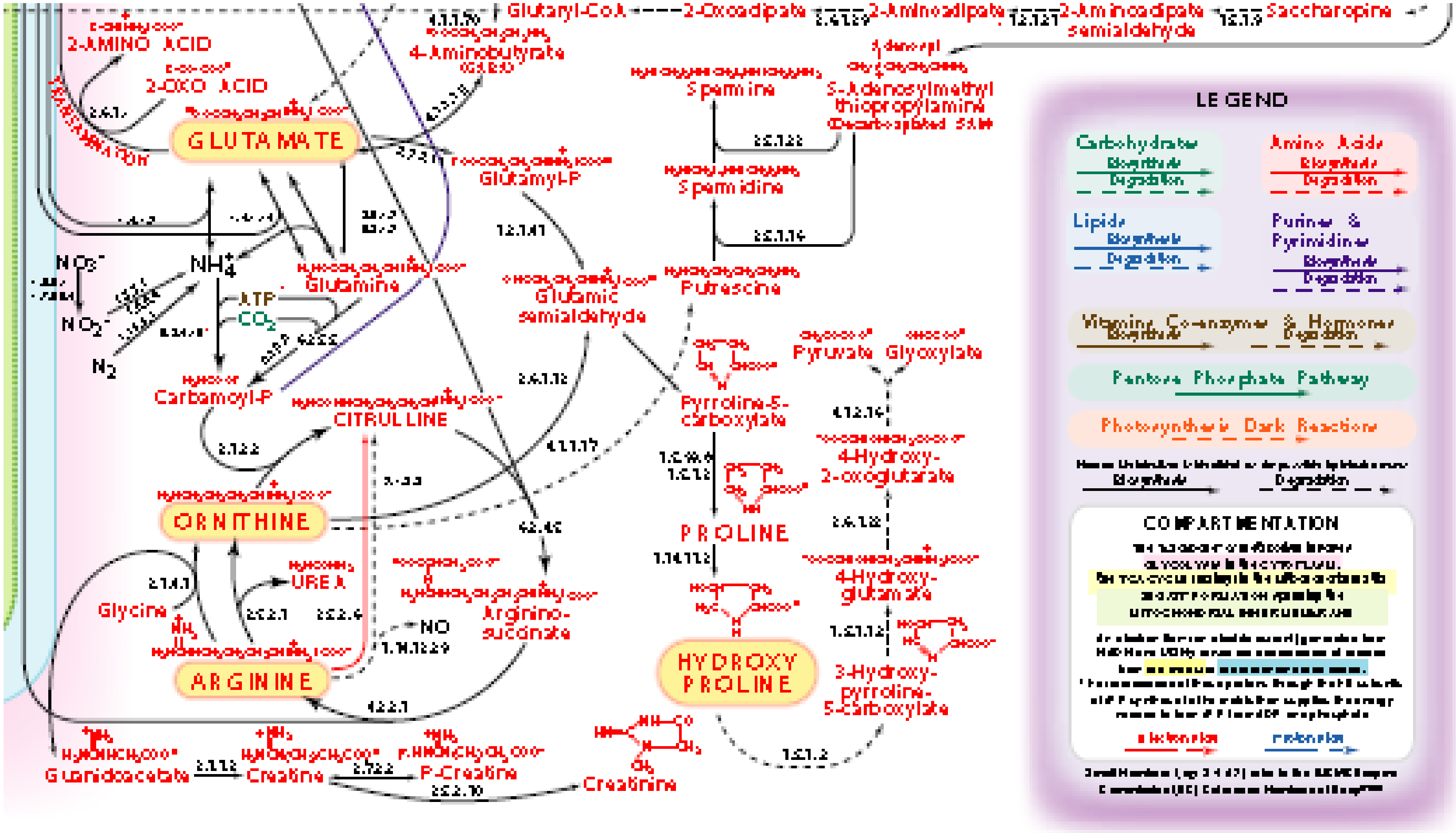
Computational Gene Finding

- Using Bioinformatics to Identify Genes:
  - Identifying common phenomena in known genes
  - Building a computational framework/model that can accurately describe the common phenomena
  - Using the model (Hidden Markov Models and neural networks) to scan uncharacterized sequence to identify regions that match the model, which become putative genes
  - Using Bayesian statistics to make predictions
  - Test and validate the predictions

Gene Interaction Networks - Yeast



Excerpt - Metabolic Pathways

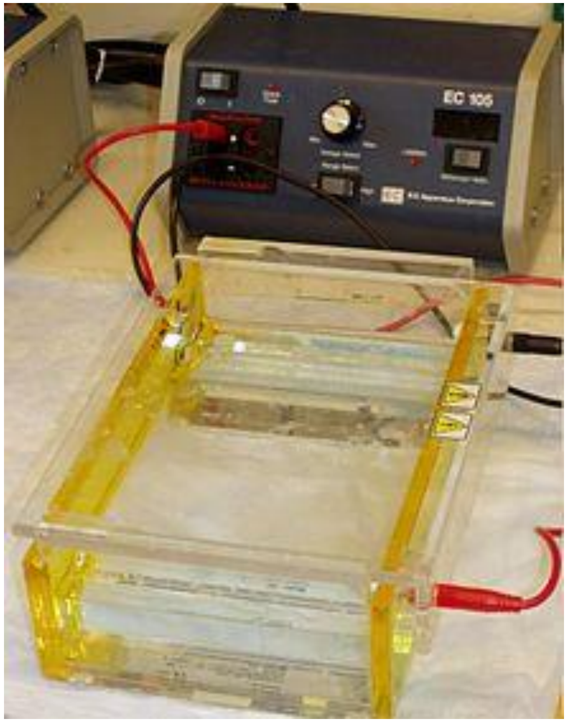


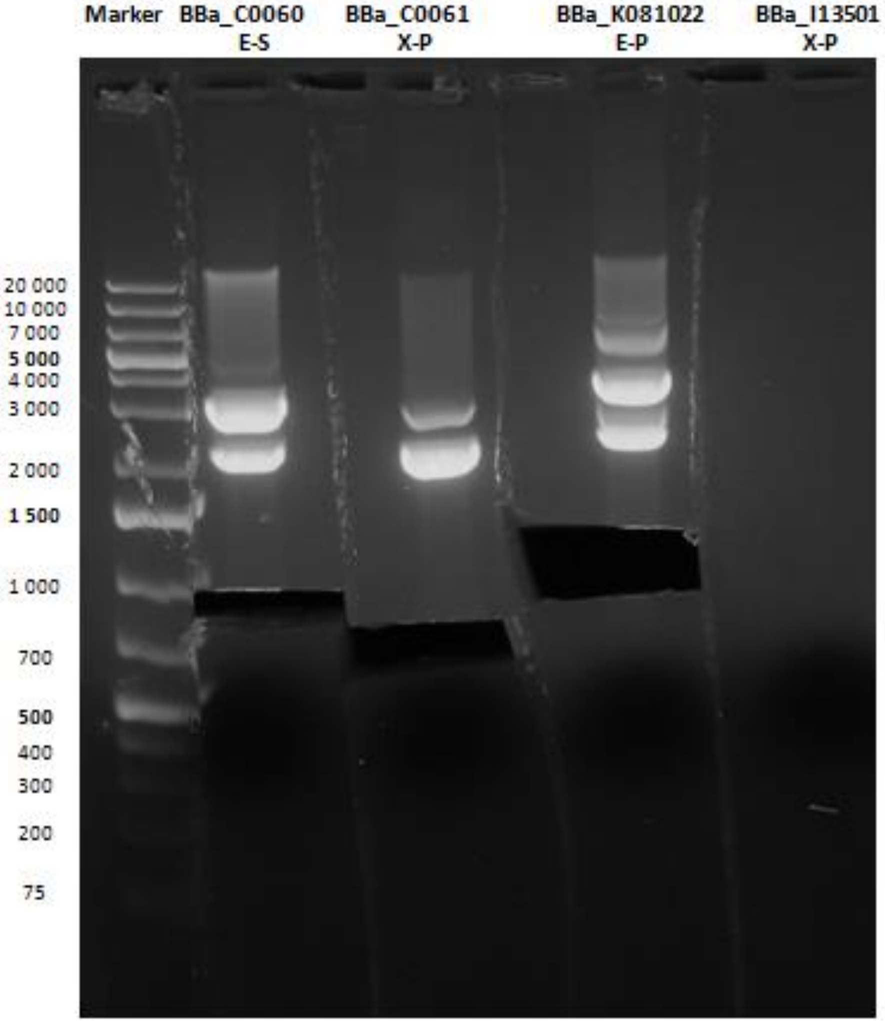
Genome Sequencing Overview

- Gel Electrophoresis and RFLP
- Gene Chip Overview
- Gene Sequencing Basics and Trends
- Illumina Sequencing Machine
- Ion Torrent Sequencing Machine
- Novel Sequencing Concepts in Research Stage

Gel Electrophoresis

- Apparatus and Results





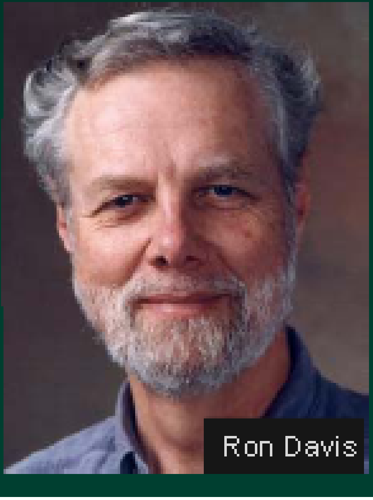
RFLPs provide abundant markers (1980)

AJHG 32: 314, 1980

Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms

DAVID BOTSTEIN,[1] RAYMOND L. WHITE,[2] MARK SKOLNICK,[3] AND RONALD W. DAVIS[4]

David Botstein


Ron Davis

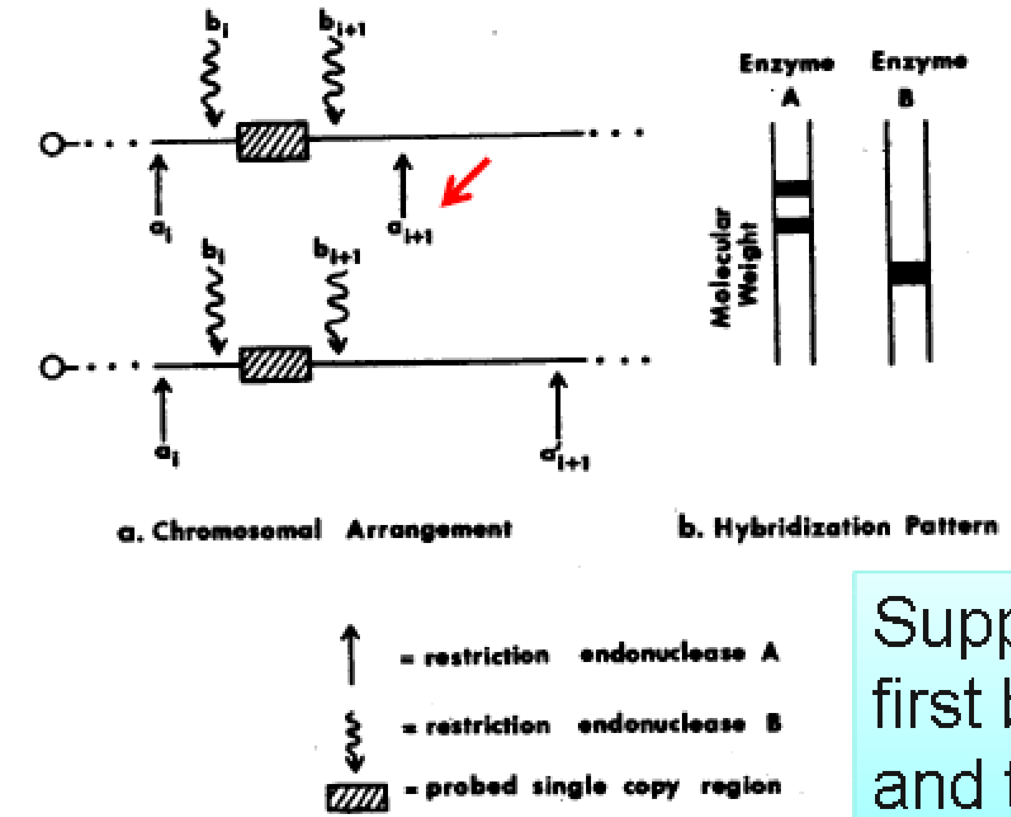
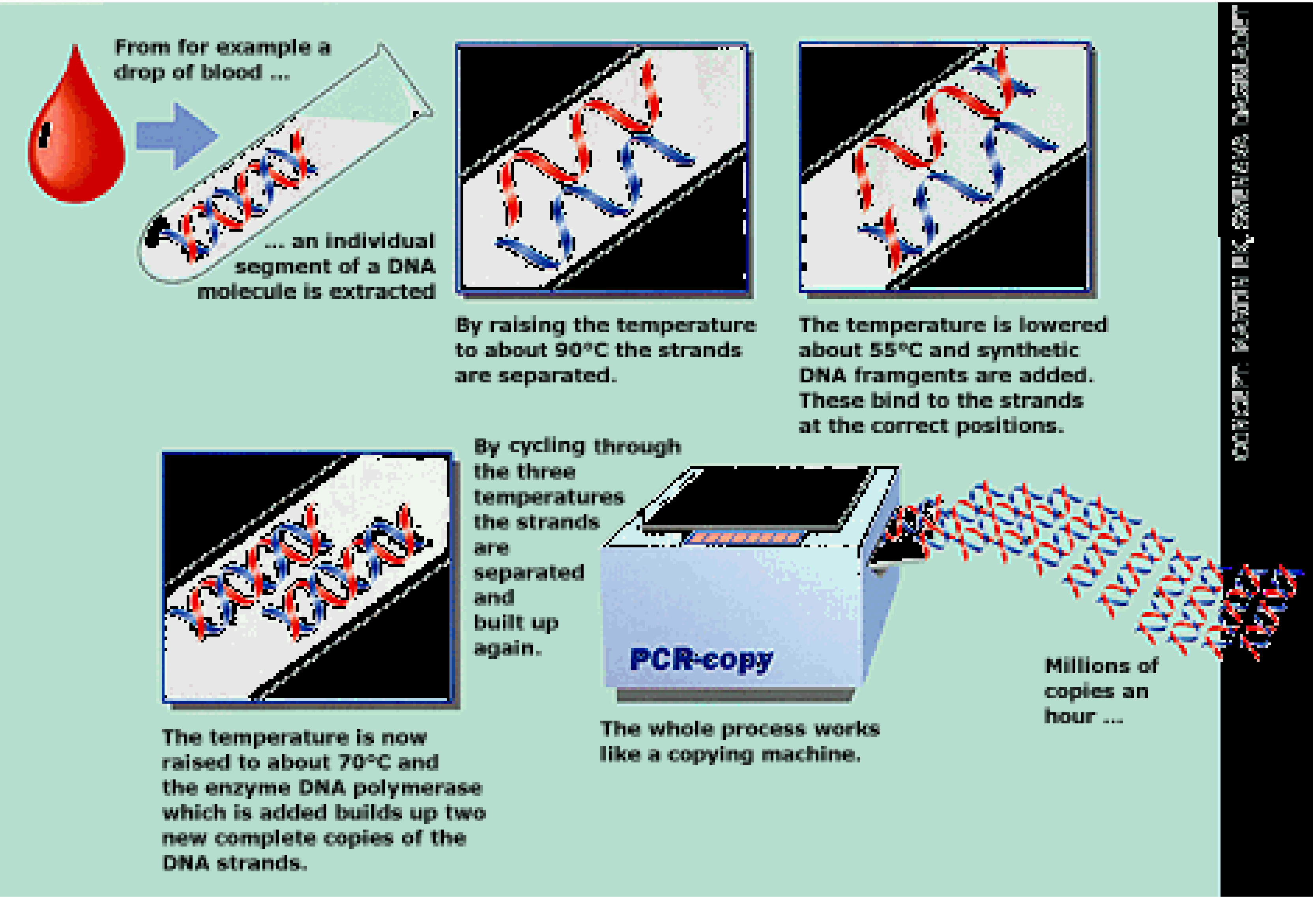


FIG. 1.—*a*, Cuts made in pair of homologous chromosomes by enzyme A and enzyme B; *b*, hybridiza attern of enzymes A and B given cuts of *a*.

Supplanted, first by STRp's and then by SNPs

Polymerase Chain Reaction (PCR)



From for example a drop of blood ...

... an individual segment of a DNA molecule is extracted

By raising the temperature to about 90°C the strands are separated.

The temperature is lowered about 55°C and synthetic DNA framgents are added. These bind to the strands at the correct positions.

The temperature is now raised to about 70°C and the enzyme DNA polymerase which is added builds up two new complete copies of the DNA strands.

By cycling through the three temperatures the strands are separated and built up again.

The whole process works like a copying machine.

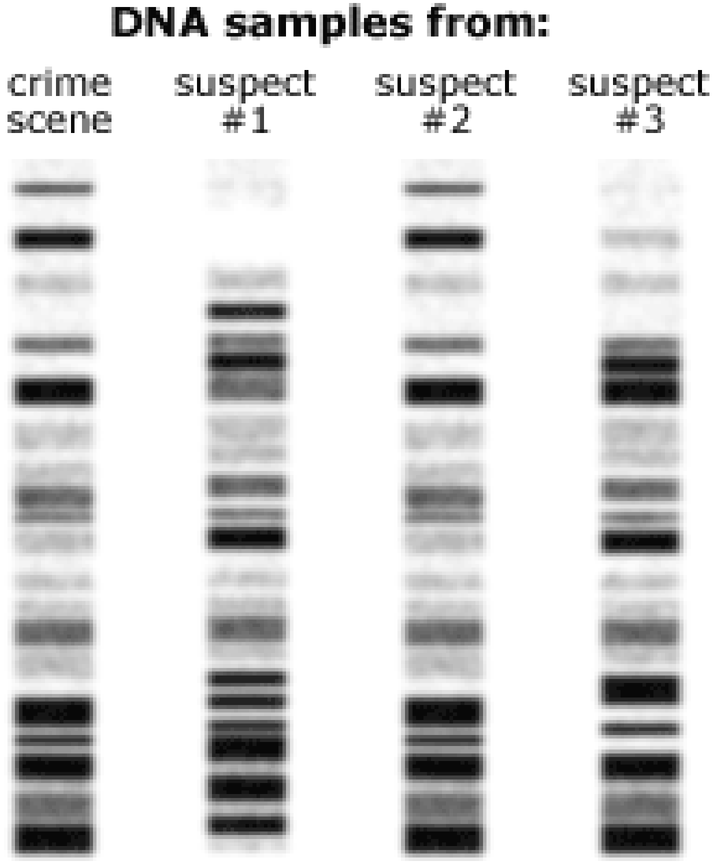
Millions of copies an hour ...

CONCEPT: MARTIN EK, GRAPHICS: DAGBLADET

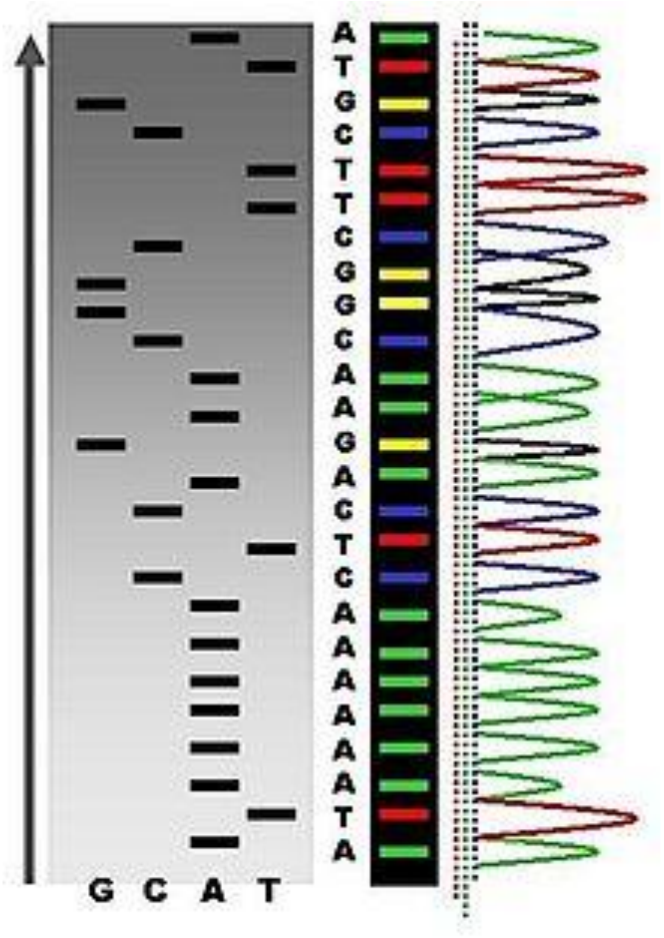
Gel Electrophoresis Applications

DNA Fingerprinting (RFLP)

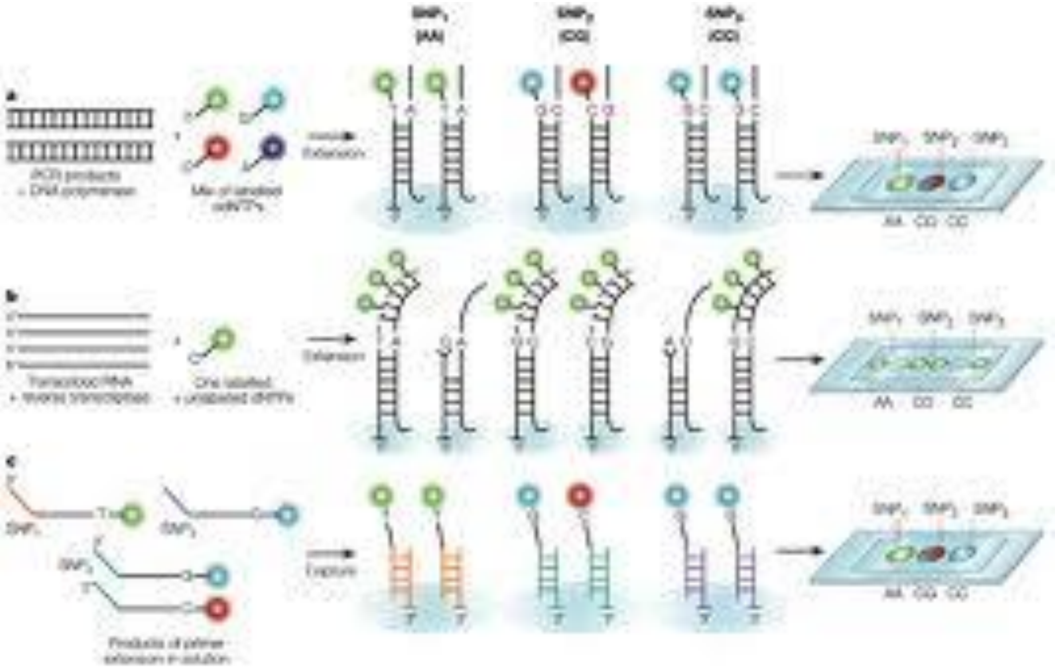
- Restriction Enzymes Cut DNA



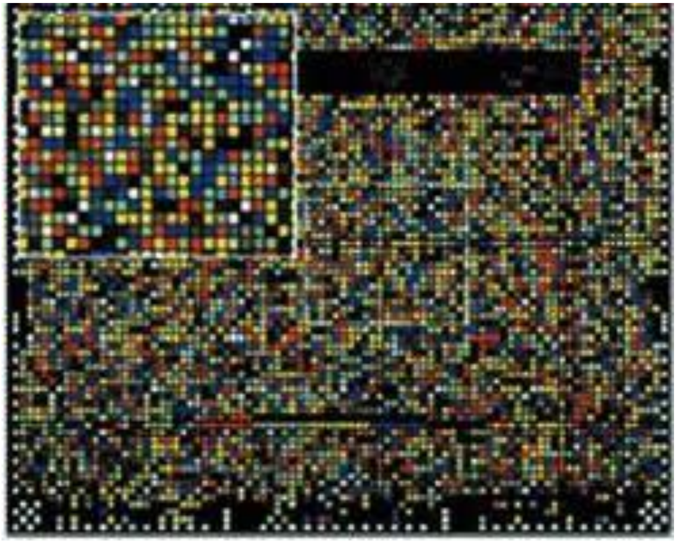
DNA Sequencing - Sanger Method



Gene Chip Overview



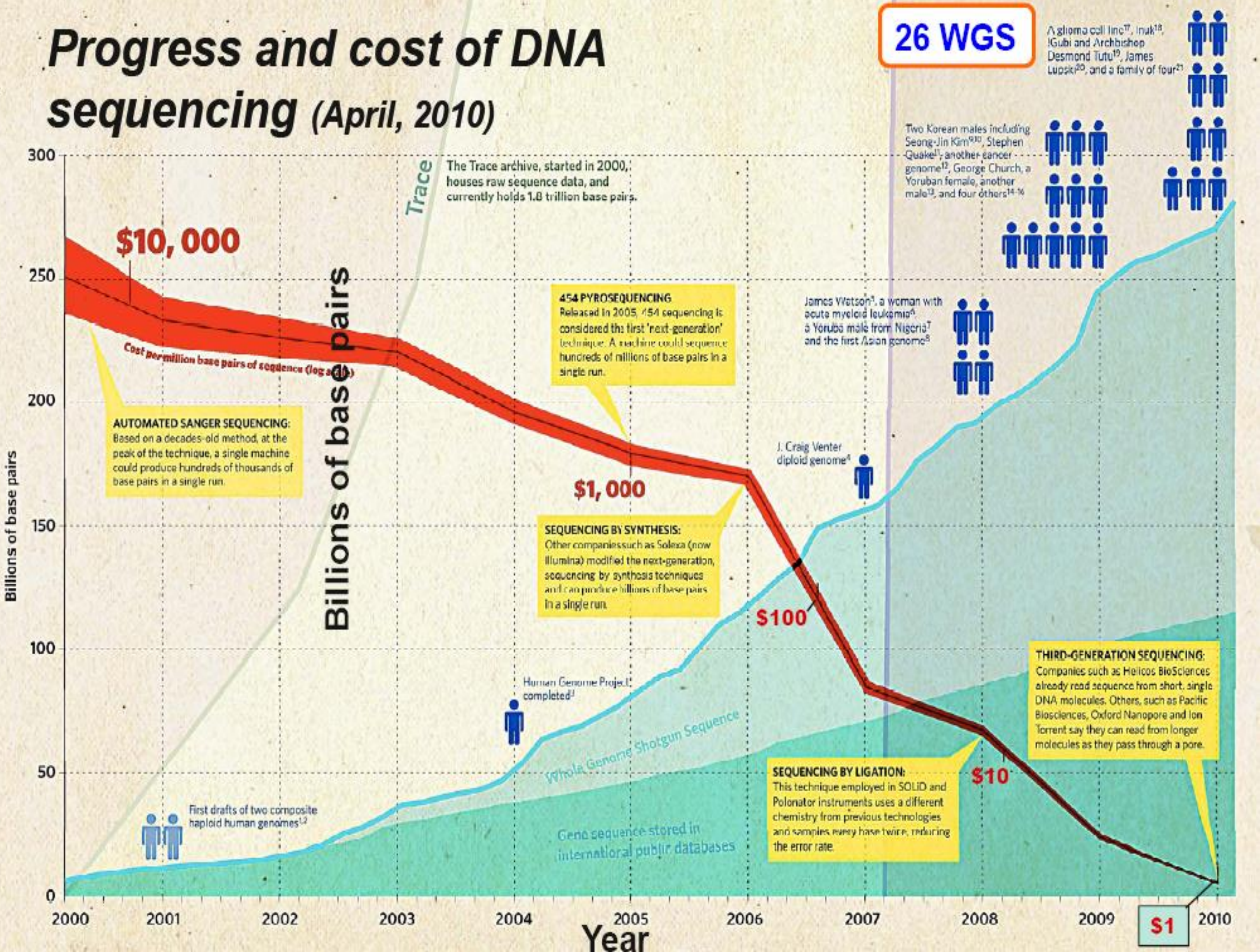
Nature Reviews | Genetics



Gene Chip Applications

Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.[8]
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.[9][10]
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO [1]), mycoplasms in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome.
DamID	Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations.[11] Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
Alternative splicing detection	An '*exon junction array* design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene).
Fusion genes microarray	A Fusion gene microarray can detect fusion transcripts, *e.g.* from cancer specimens. The principle behind this is building on the alternative splicing microarrays.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

Progress and cost of DNA sequencing (April, 2010)



26 WGS

A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]

Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14-16]

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companiessuch as Solexa (now Illumina) modified the next-generation, sequencing by synthesis techniques and can produce billions of base pairs in a single run.

James Watson[3], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]

J. Craig Venter diploid genome[4]

Human Genome Project completed[3]

First drafts of two composite haploid human genomes[1,2]

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

Comparative costs: sequencing a human genome

Capillary technology

Applied Biosystems 3730xl (2004)

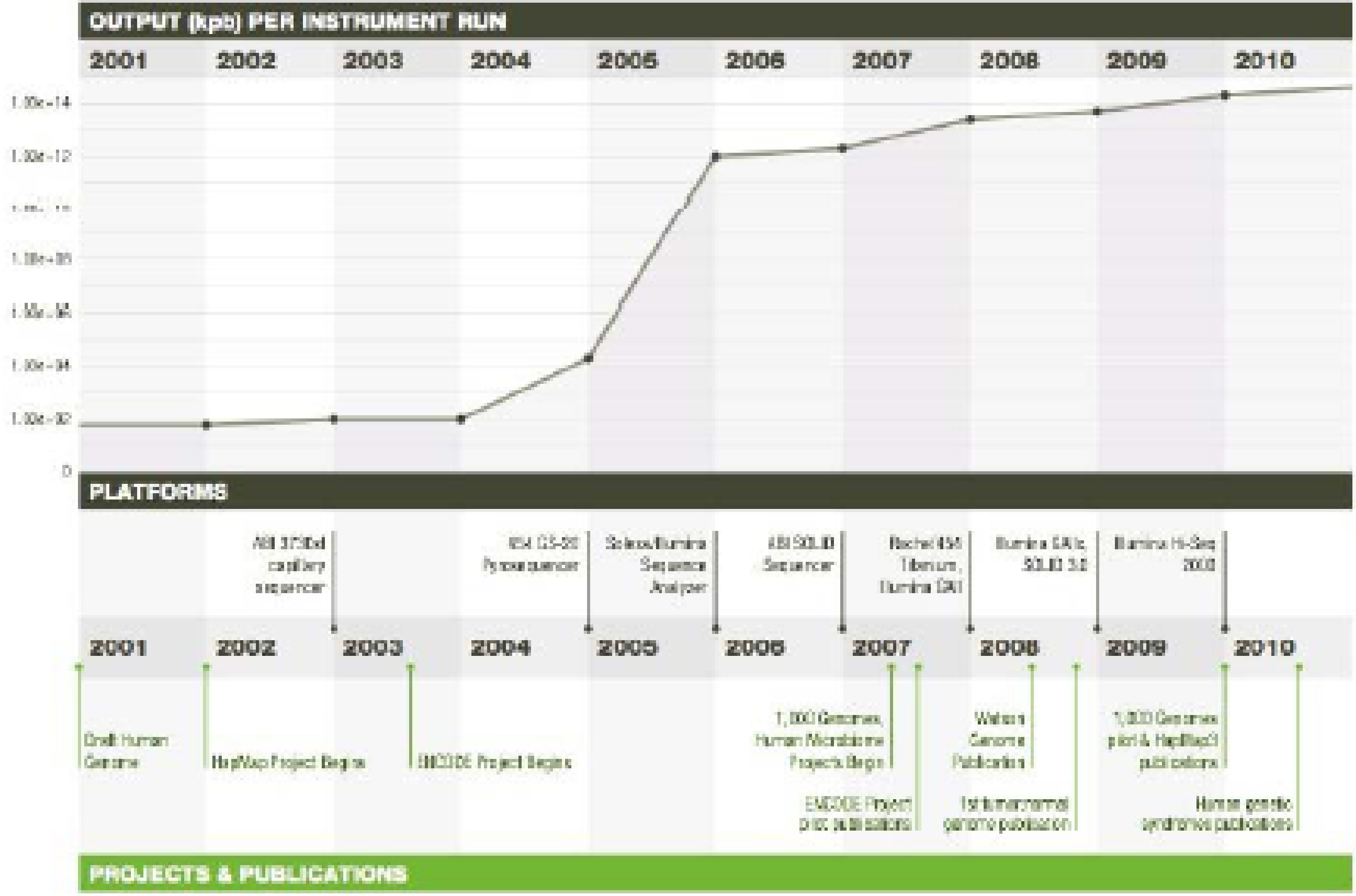
$15,000,000

Next-gen technology

Illumina HiSeq (2011)

$10,000

The Trajectory of Throughput: 10 years



E.R. *Mardis*, Nature (2011) 470: 198-203

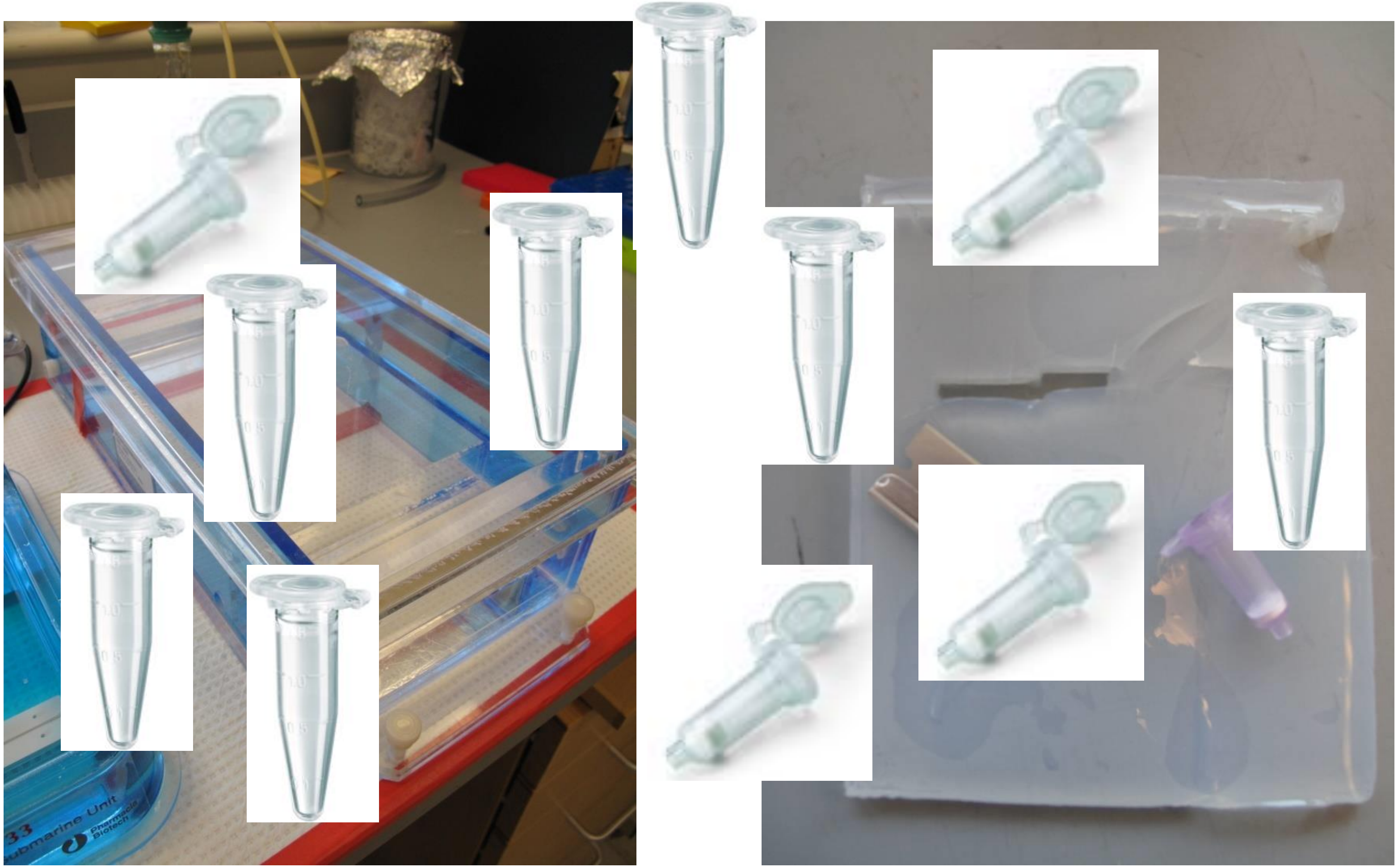
Next-generation DNA sequencing instruments

- **All commercially-available sequencers have the following shared attributes:**
  - Random fragmentation of starting DNA, ligation with custom linkers = "a library"
  - Library amplification on a solid surface (either bead or glass)
  - Direct step-by-step detection of each nucleotide base incorporated during the sequencing reaction
  - Hundreds of thousands to hundreds of millions of reactions imaged per instrument run = "massively parallel sequencing"
  - Shorter read lengths than capillary sequencers
  - A "digital" read type that enables direct quantitative comparisons
  - A sequencing mechanism that samples both ends of every fragment sequenced ("paired end" reads)

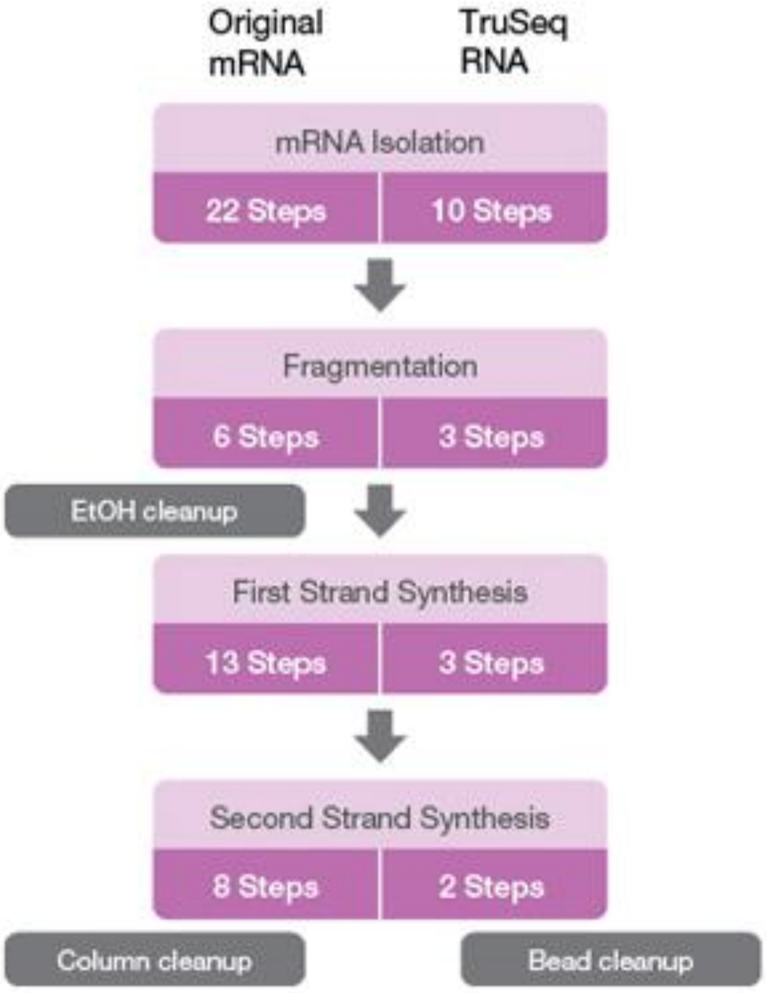
Next Generation Sequencing Platforms

Company	Platform Name	Sequencing	Amplification	Run Time
Roche	454 Ti	DNA Polymerase “Pyrosequencing”	emPCR	10 hours
Illumina	Hi-Seq/MiSeq	DNA Polymerase	Bridge amplification	10 days/24 hours
Life	SOLiD/5550	DNA Ligase	emPCR	12 days
Ion Torrent	PGM	Synthesis H^+ detection	emPCR	2 hours
Pacific Biosciences	RS	Synthesis	NONE	45 min

mRNAseq library prep: PAST

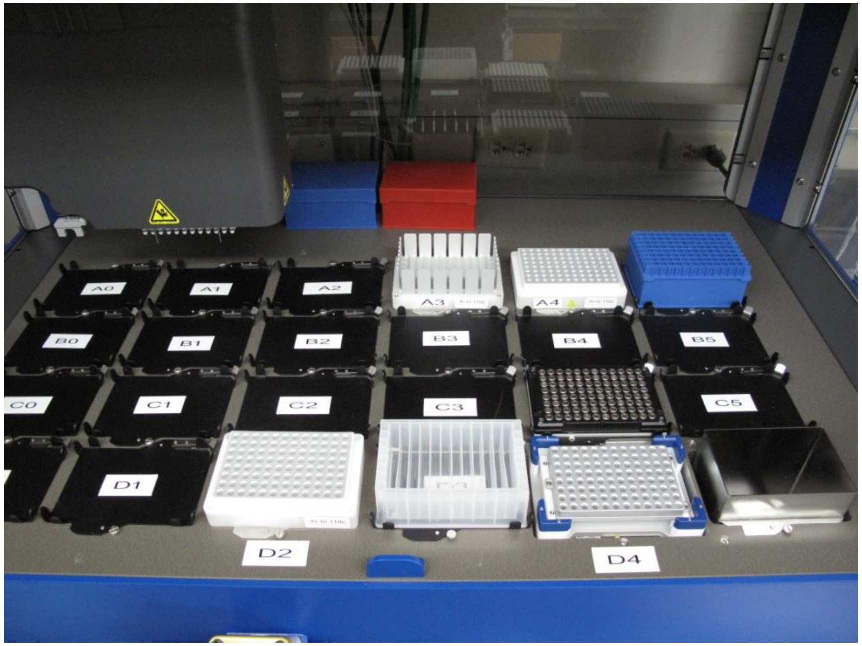


mRNAseq library prep: PRESENT

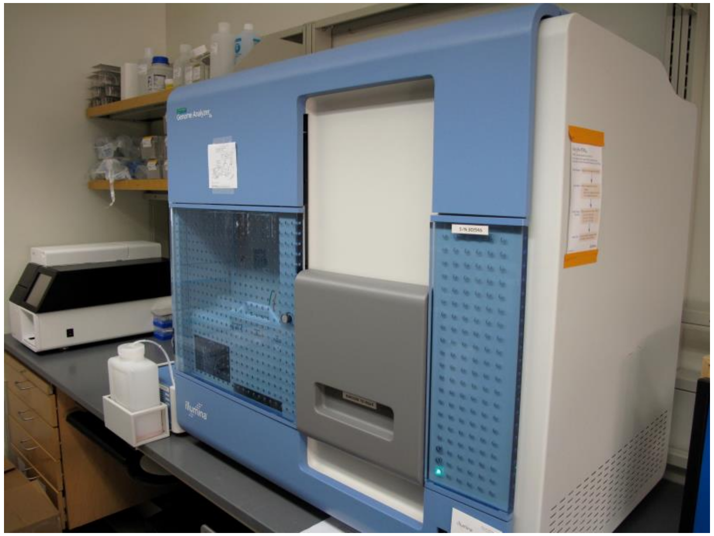


	Original mRNA	TruSeq RNA
mRNA Isolation	22 Steps	10 Steps
Fragmentation	6 Steps	3 Steps
	EtOH cleanup	
First Strand Synthesis	13 Steps	3 Steps
Second Strand Synthesis	8 Steps	2 Steps
	Column cleanup	Bead cleanup

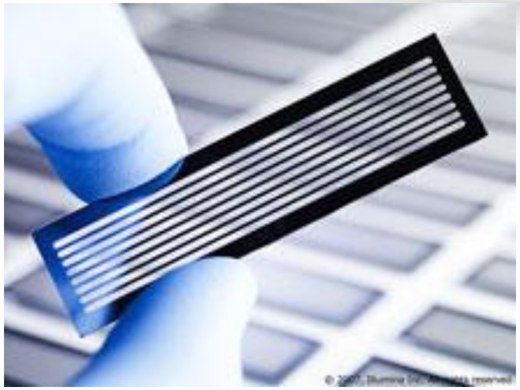
Caliper Sciclone NGS robotic liquid handler



Illumina Sequencing

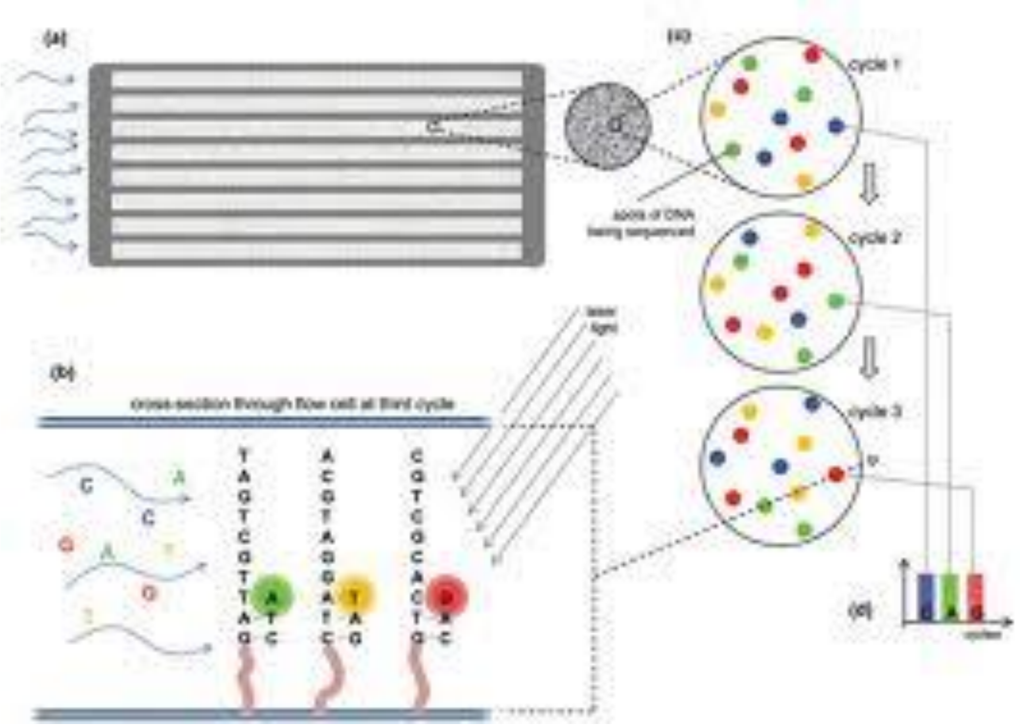


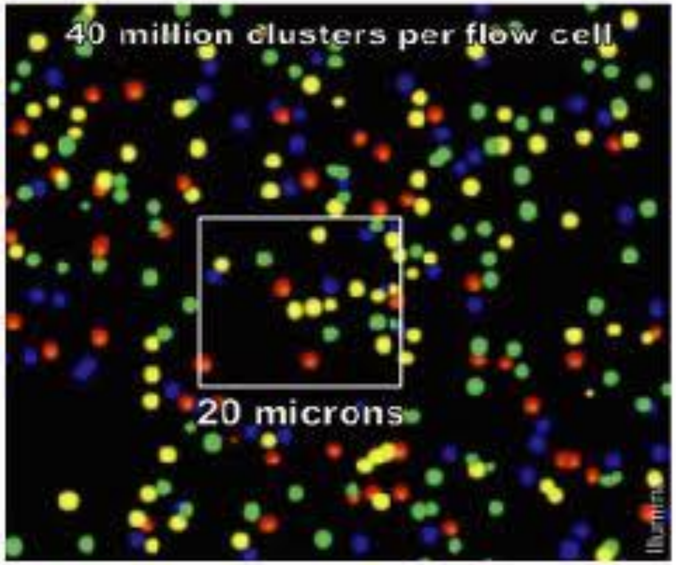
40 gb/run



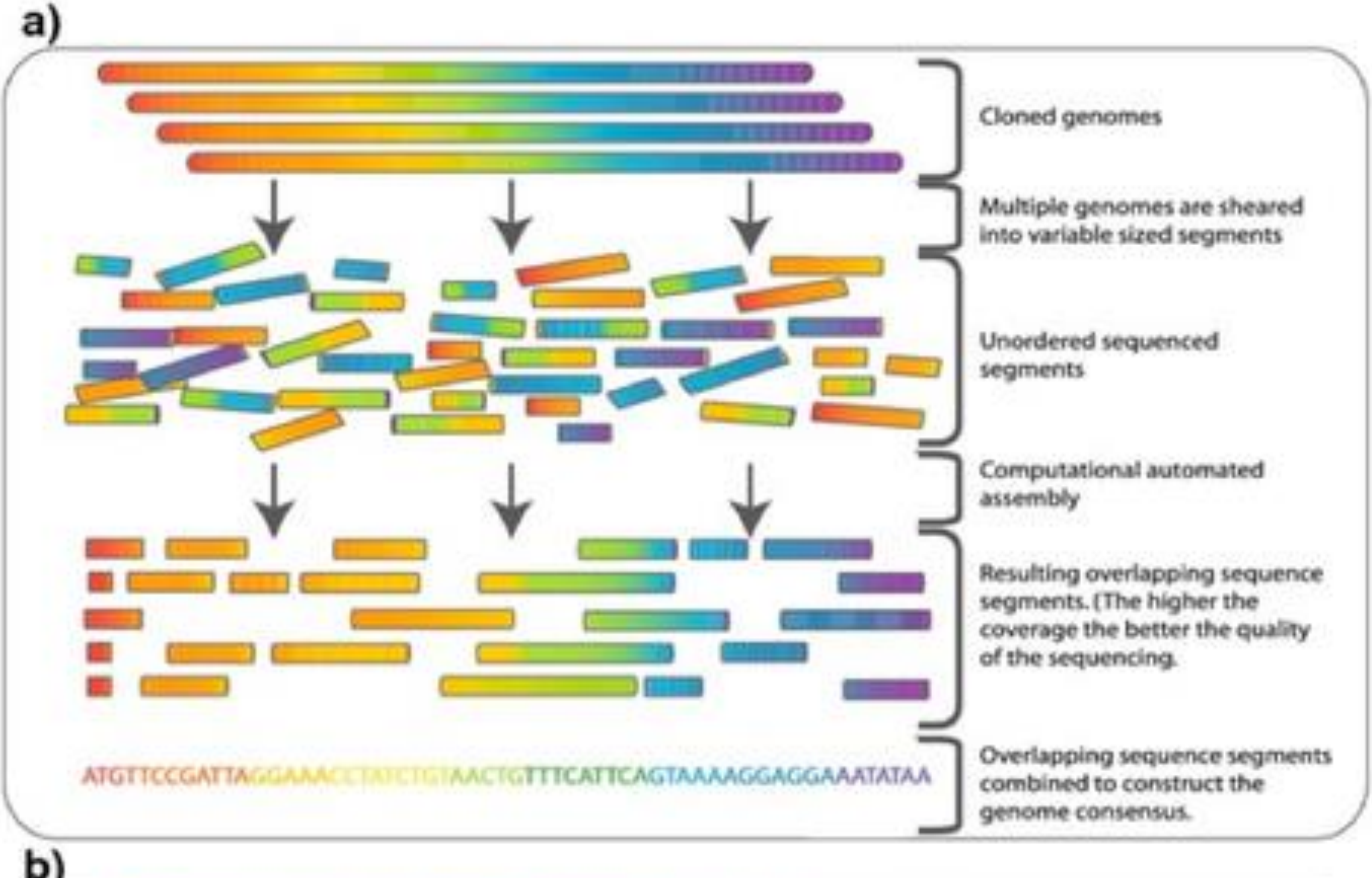
300gb/run
600gb/run soon

Illumina Sequencing

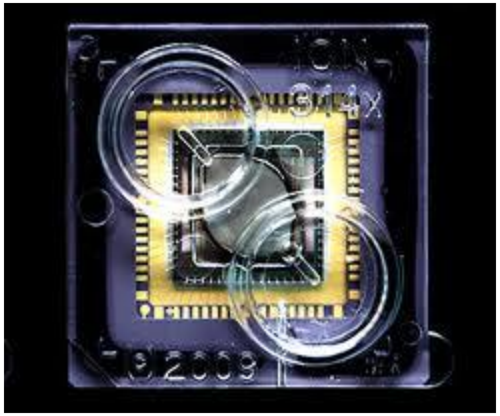

(a)
(c)
cycle 1
cycle 2
cycle 3
spots of DNA being sequenced
laser light
(b)
cross-section through flow cell at third cycle
(d)
cycles



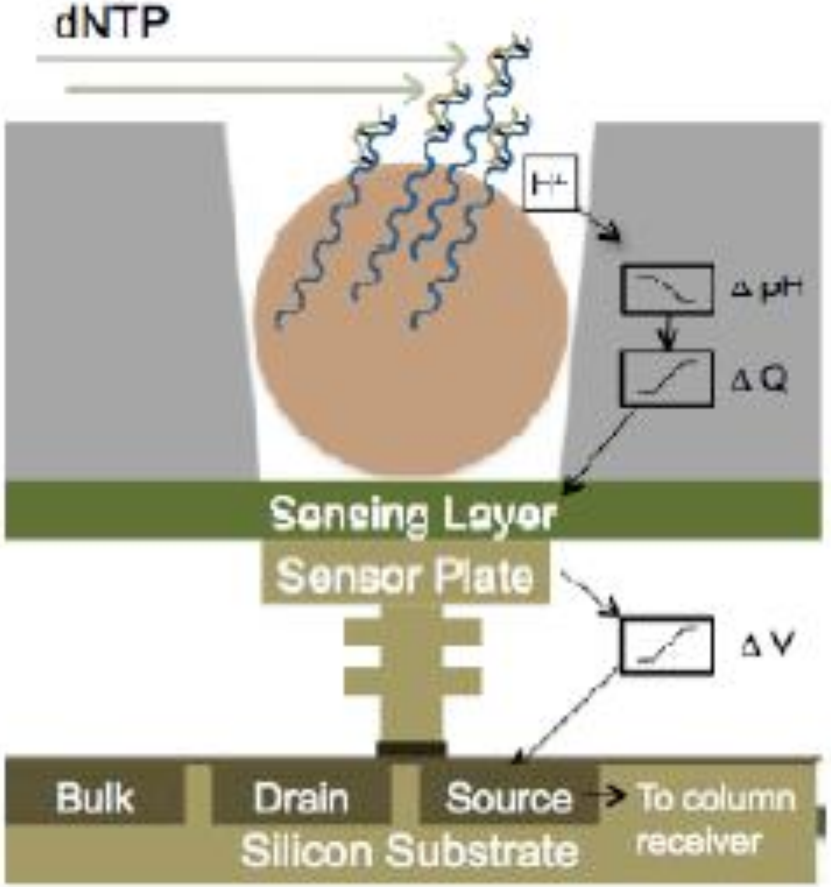
Sequence Alignment - Shotgun Method

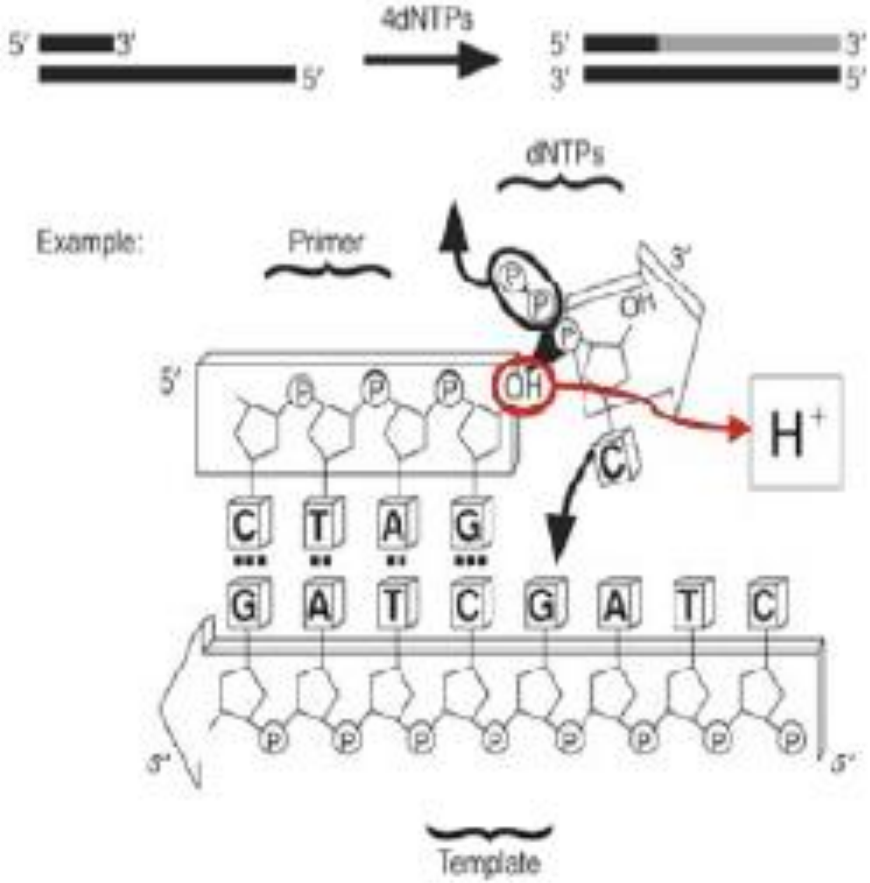

a)
Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.
b)

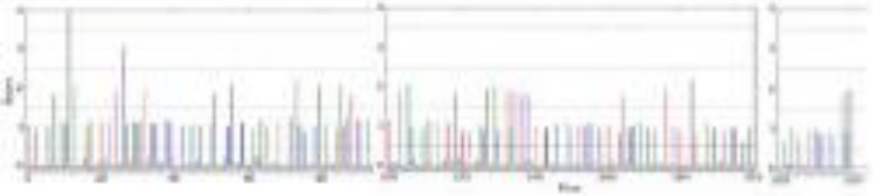
Ion Torrent Personal Genome Machine



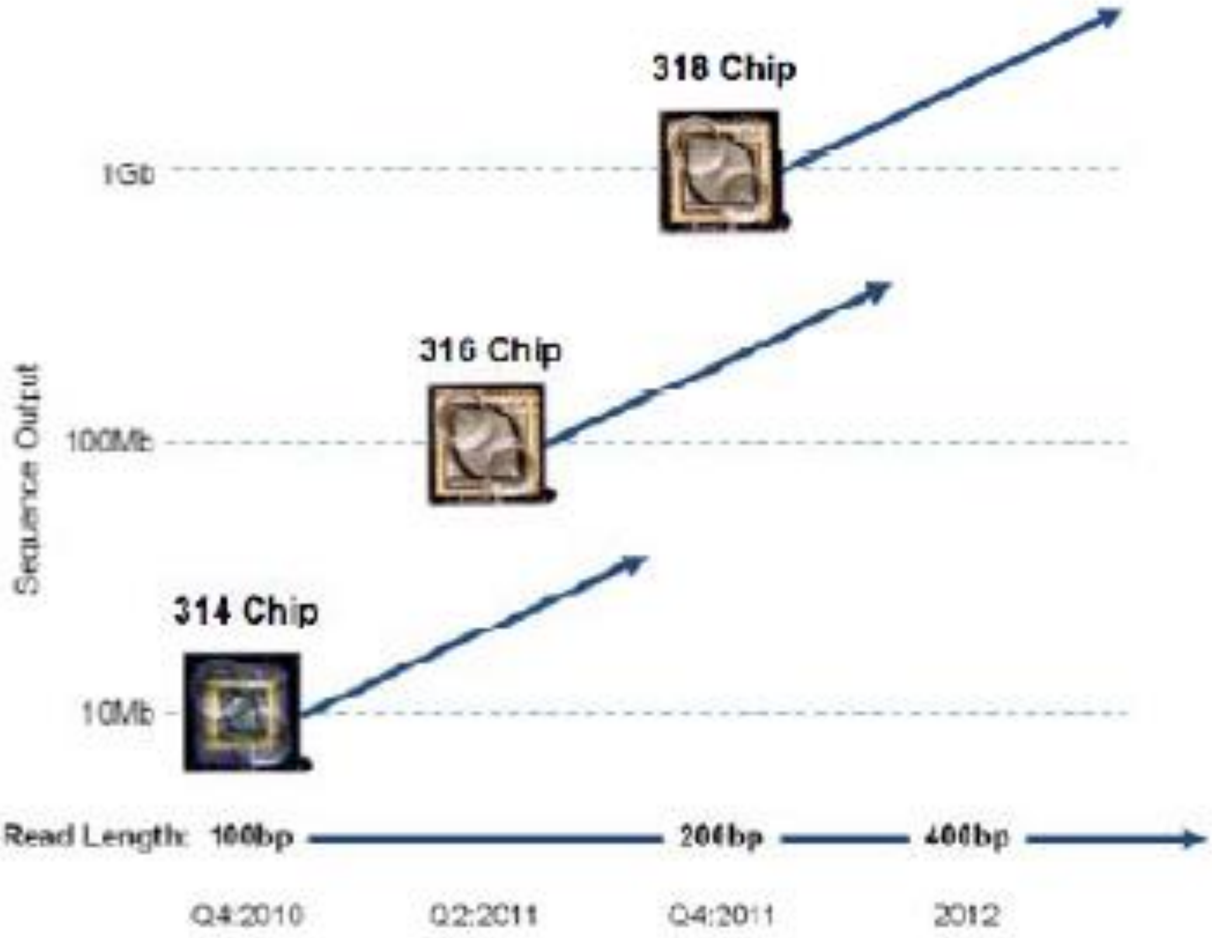
ION Torrent


dNTP
H+
Δ pH
Δ Q
Sensing Layer
Sensor Plate
Δ V
Bulk
Drain
Source
To column receiver
Silicon Substrate

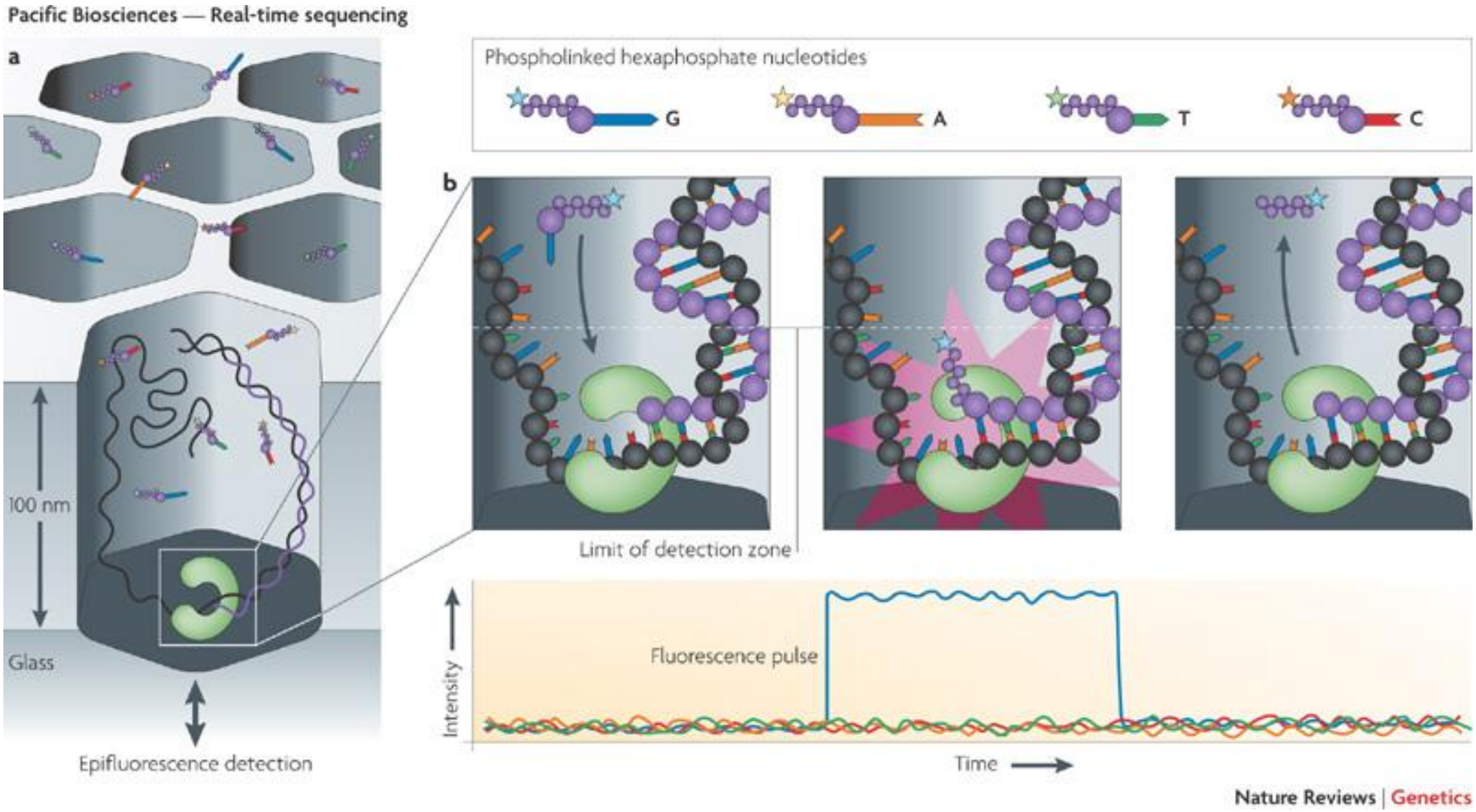

5'
3'
5'
4dNTPs
5'
3'
3'
5'
dNTPs
Example:
Primer
3'
OH
5'
P
P
P
P
P
OH
H+
C
C
T
A
G
G
A
T
C
G
A
T
C
3'
5'
Template



Ion Torrent Yield Trajectory

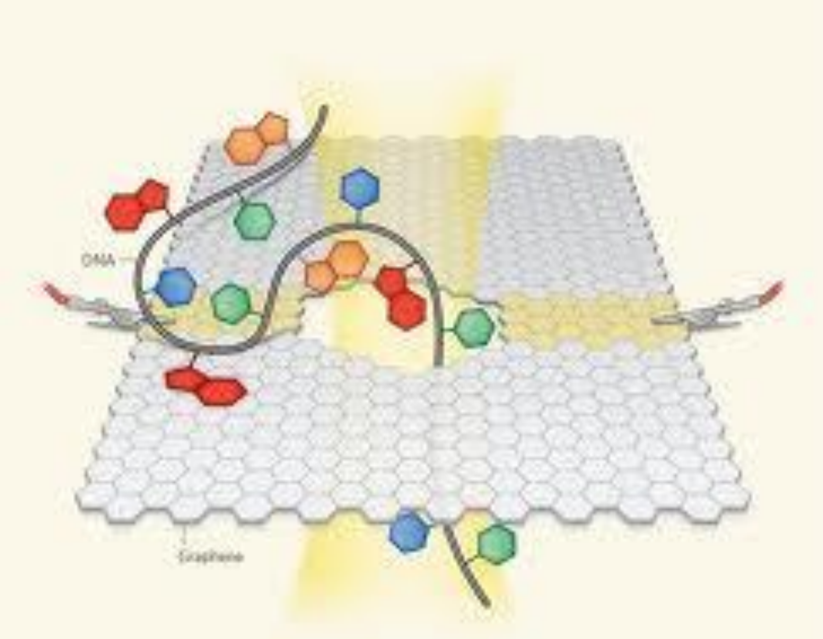


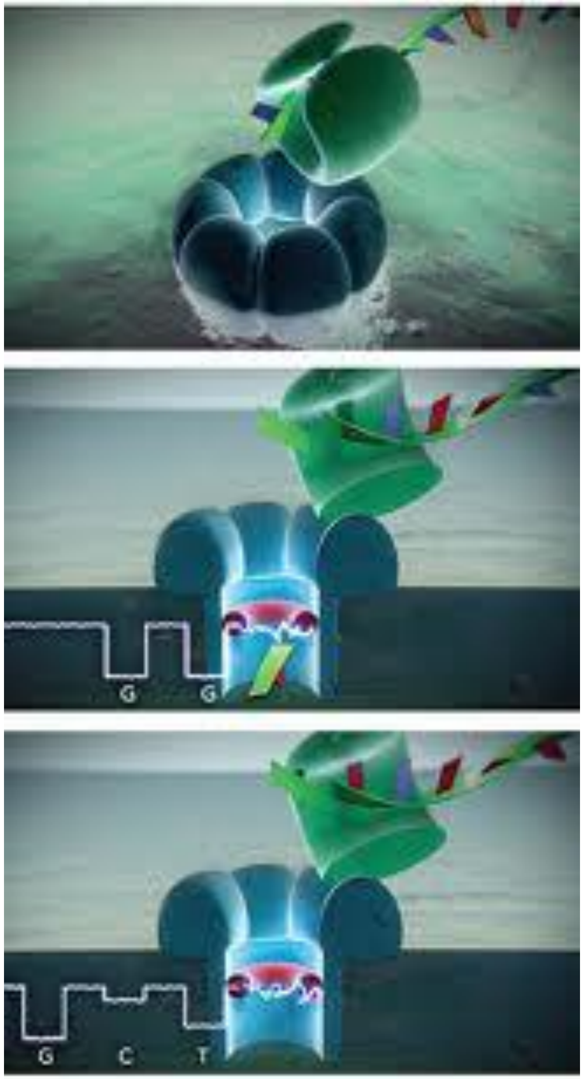
Pacific Biosciences - Real-time Single Molecule Sequencing



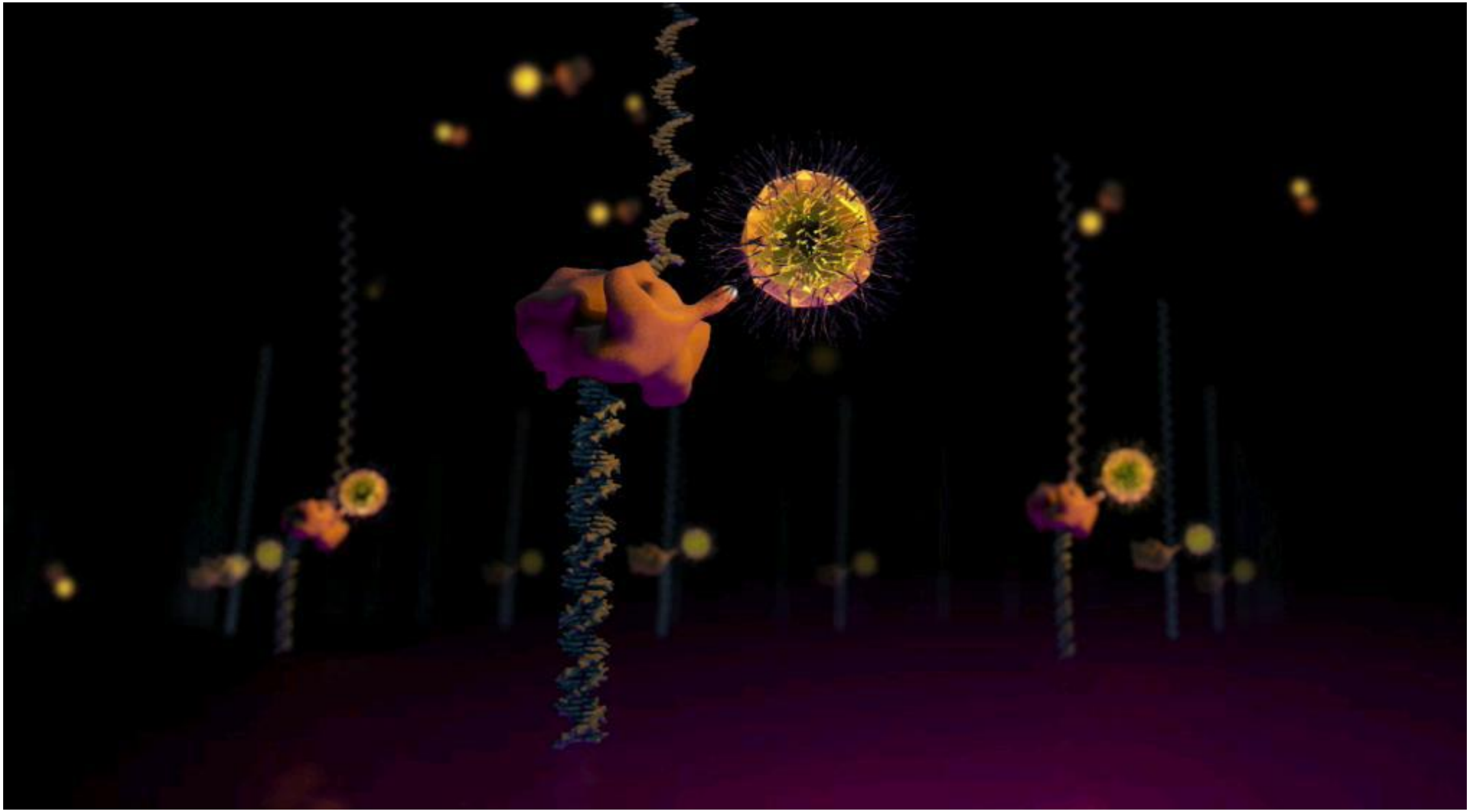
Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Fluorescence pulse
Intensity
Time
Nature Reviews | Genetics

Oxford Nanopore - Single Molecule Sequencing

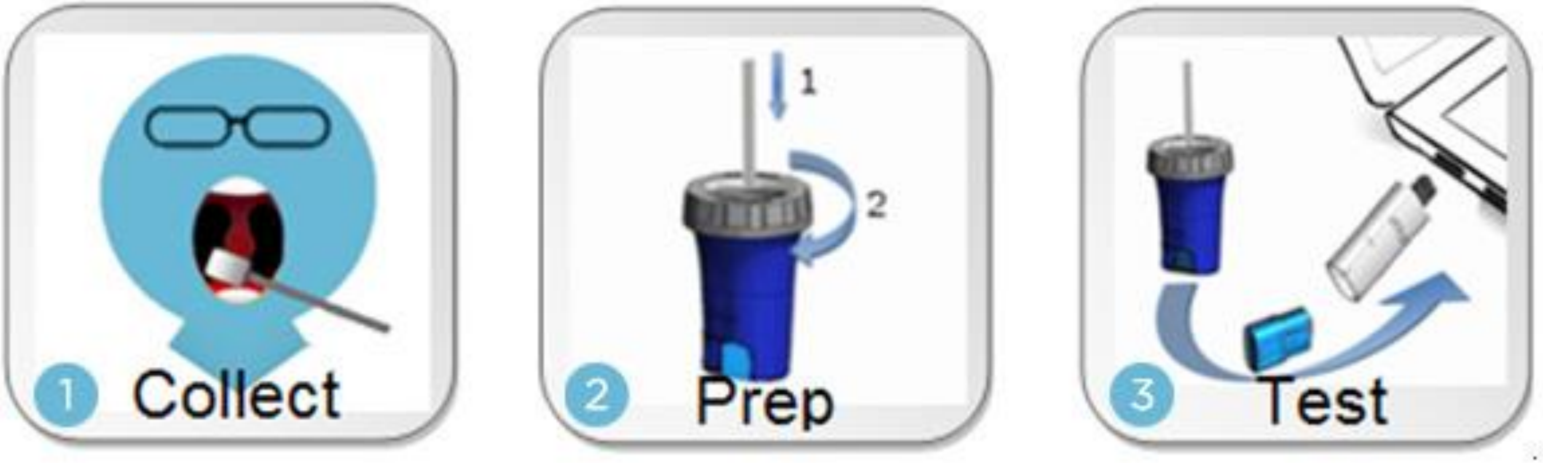




Starlight - Life Technologies



DNAe - Lab-free DNA Testing



Benefits

30 Minute Sample Preparation

No Special Skills Required

Cartridge Connects to PC via USB Port

Applications

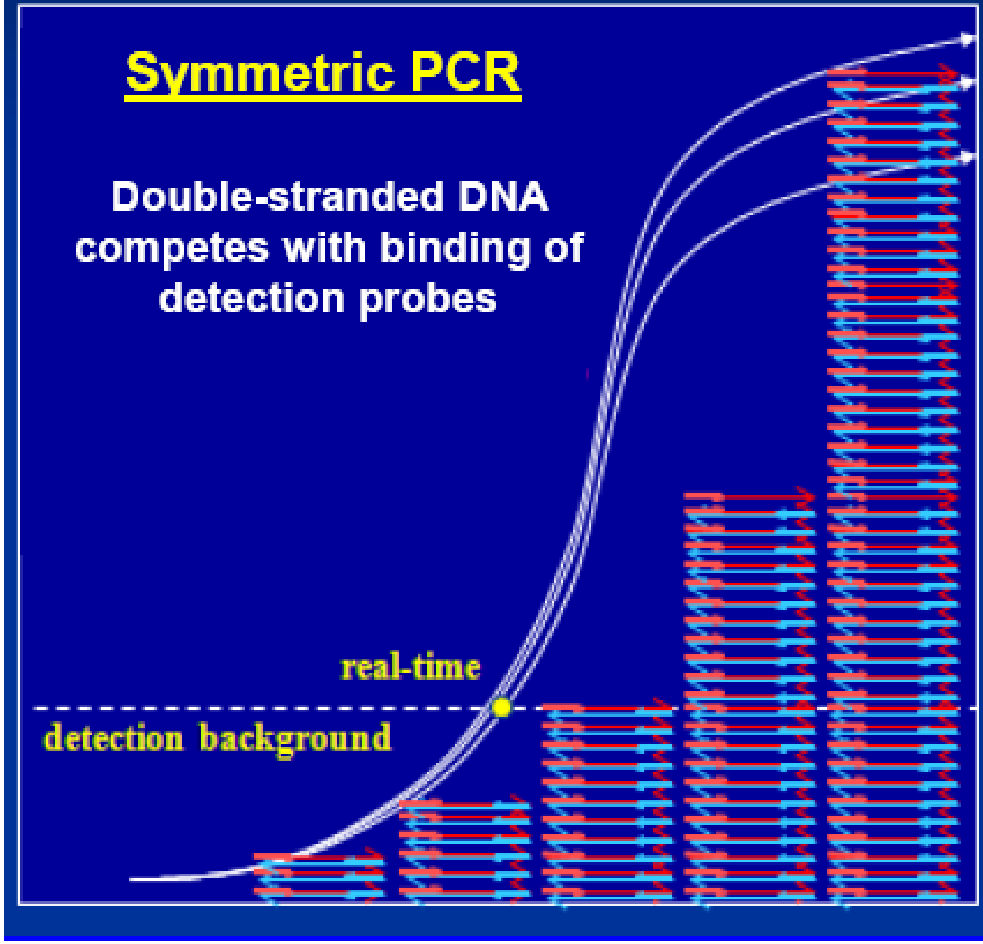
Get results while you wait in Doctor/Dentist Office

Check for infections - bacterial, viral, fungal

Faster PCR

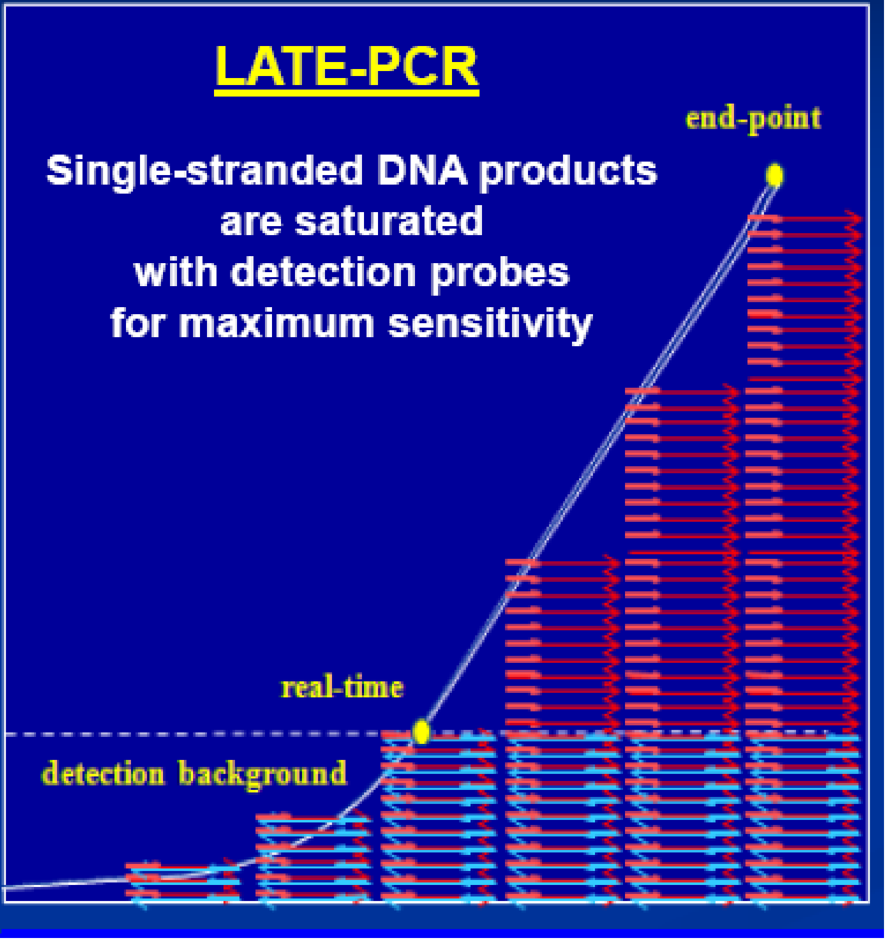
- Brandeis L&L Talk about Viruses and Bacteria
  - Arthur Reis, Brandeis Chem Prof about Wangh Lab
  - LATE-PCR Technique to Assay Pathogens
  - Use to Detect Flu Variants, Resistant Bacteria
  - Cheaper than Gene Chips ($5 instead of $100)
  - Takes about 1 Hour in Field Use
  - Brandeis Holds Patents

LATE-PCR Provides Increased Detection Sensitivity



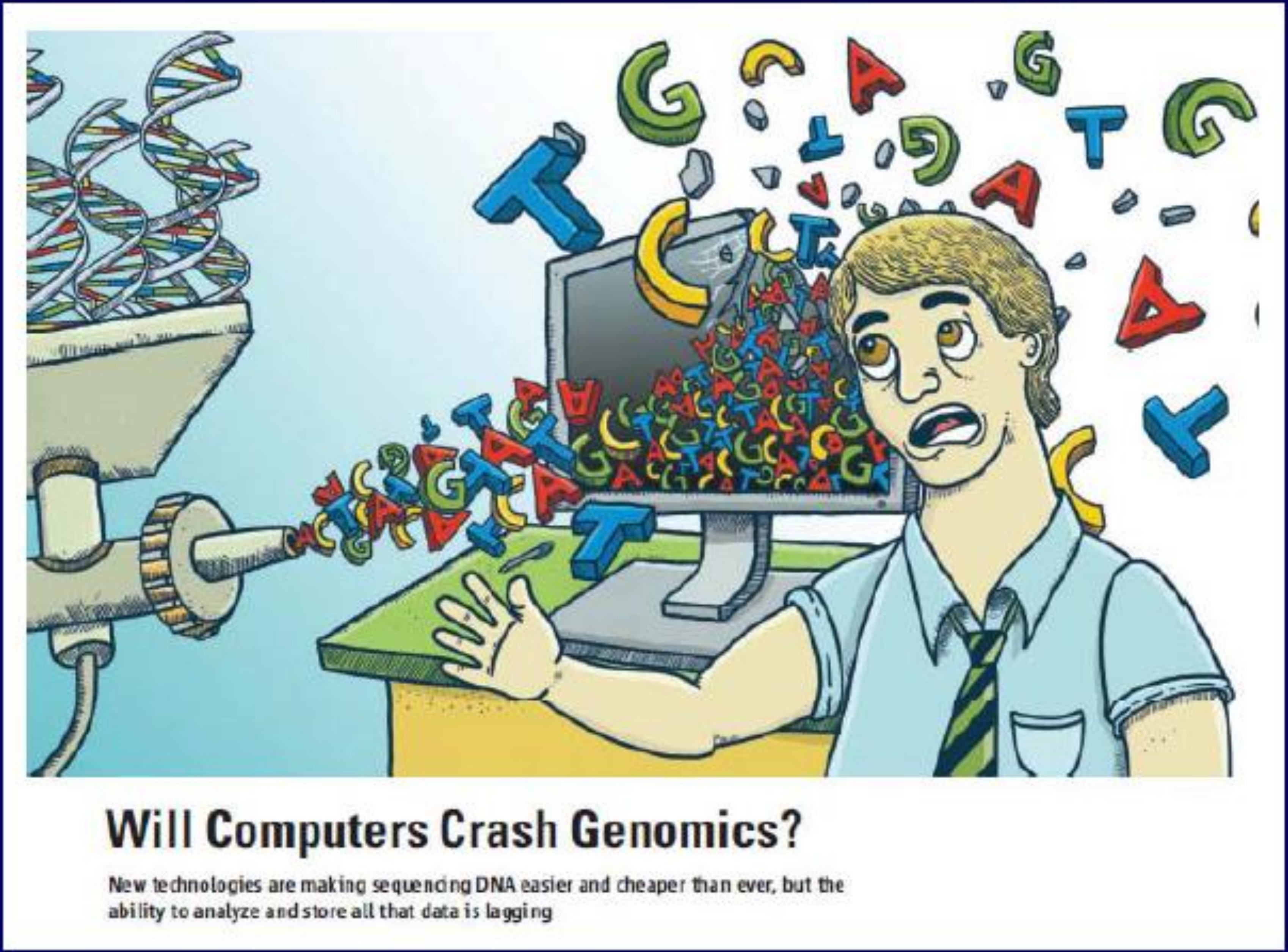
Symmetric PCR

Double-stranded DNA competes with binding of detection probes



LATE-PCR

Single-stranded DNA products are saturated with detection probes for maximum sensitivity



Will Computers Crash Genomics?

New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging

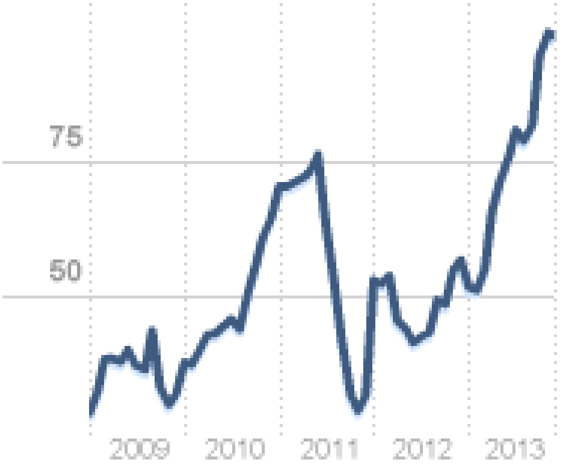
***Science* (2011)**

Response to Data Tsunami

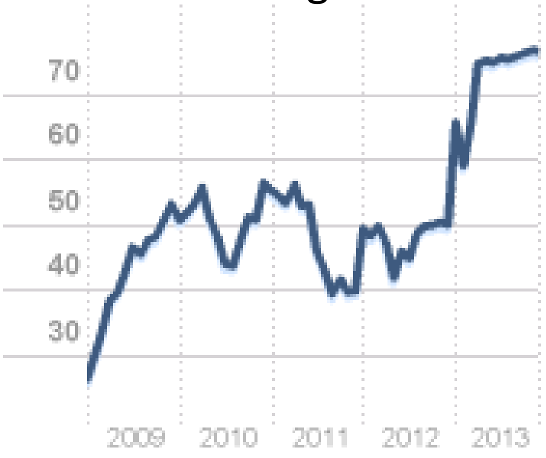
- Cloud Computing and Storage
- Computer Architectures
  - Massively Parallel Graphical Processing Units
  - Quantum Computing?
- Novel Compression Algorithms
  - Store only the 3 million SNP differences from reference genome - 3 Mb instead of 100 Gb
- Decision Support Systems for DNA Diagnostics
  - Avoid the $1000 Genome with $100K analysis
  - AI - IBM Watson applied to genome sample
  - Pattern matching to known variants for $400

Sequencing Company Stock Prices

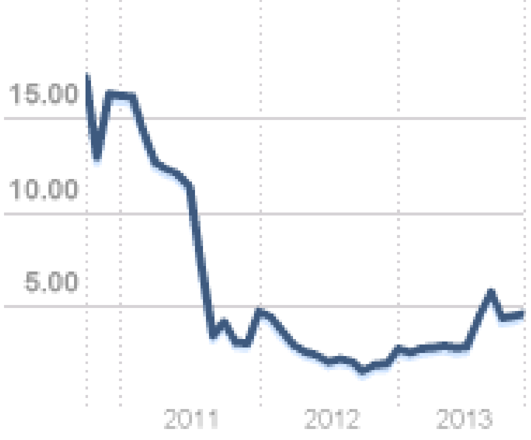
Illumina - 5 Years



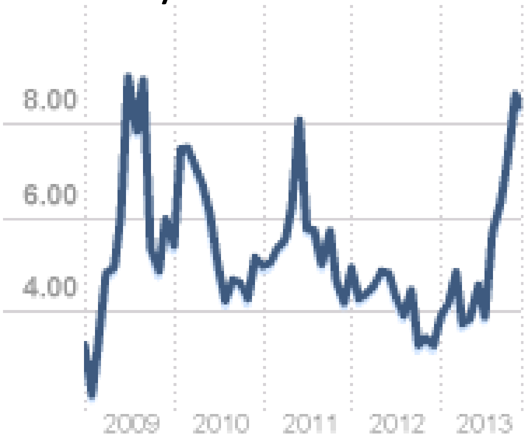
Life Technologies - 5 Years



Pacific Biosciences - 3 Years



Affymetrix - 5 Years



Individualized Medicine in the News

- NY Times, December 3, 2013
  - Learning to Defuse the Aorta
- NY Times, November 25, 2013
  - In Israel, a Push to Screen for Cancer Gene Leaves Many Conflicted
- NY Times, November 25, 2013
  - F.D.A. Orders Genetic Testing Firm to Stop Selling DNA Analysis Service
- NY Times, November 25, 2013
  - Microbes May Add Special Something to Wines

Closing Thoughts

- Thanks for paying attention and asking good questions
- If interested I can post the slides and a list of videos to the group website for those who want to learn more via Al Sherman
  - Email me at: allankleinman@rcn.com with comments and questions