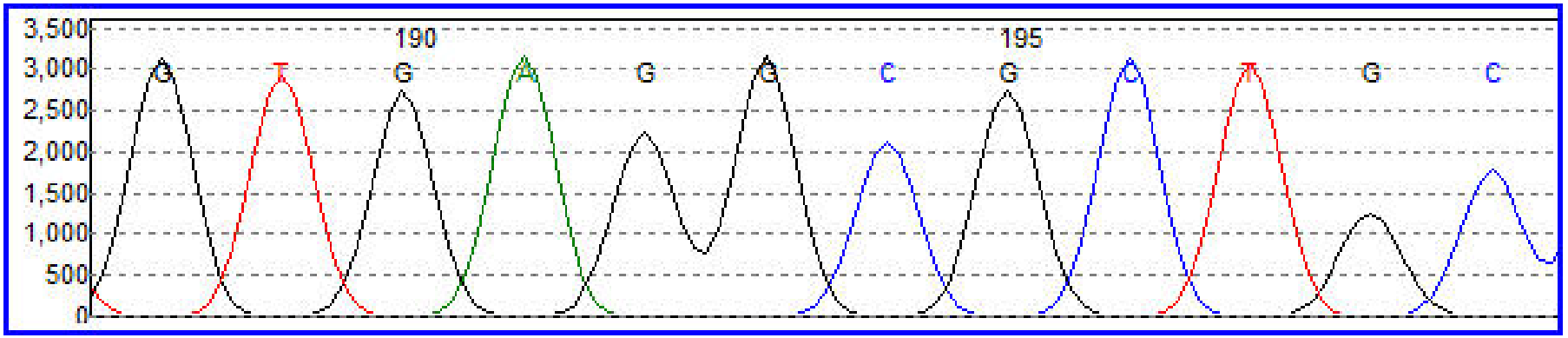
GS01 0163
Analysis of Microarray Data

Keith Baggerly and Bradley Broom
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
bmbroom@mdanderson.org

1 December 2009

Lecture 26: Next-Generation Sequencing



- Sanger Sequencing
- Exponential Growth of Sequencing Capability

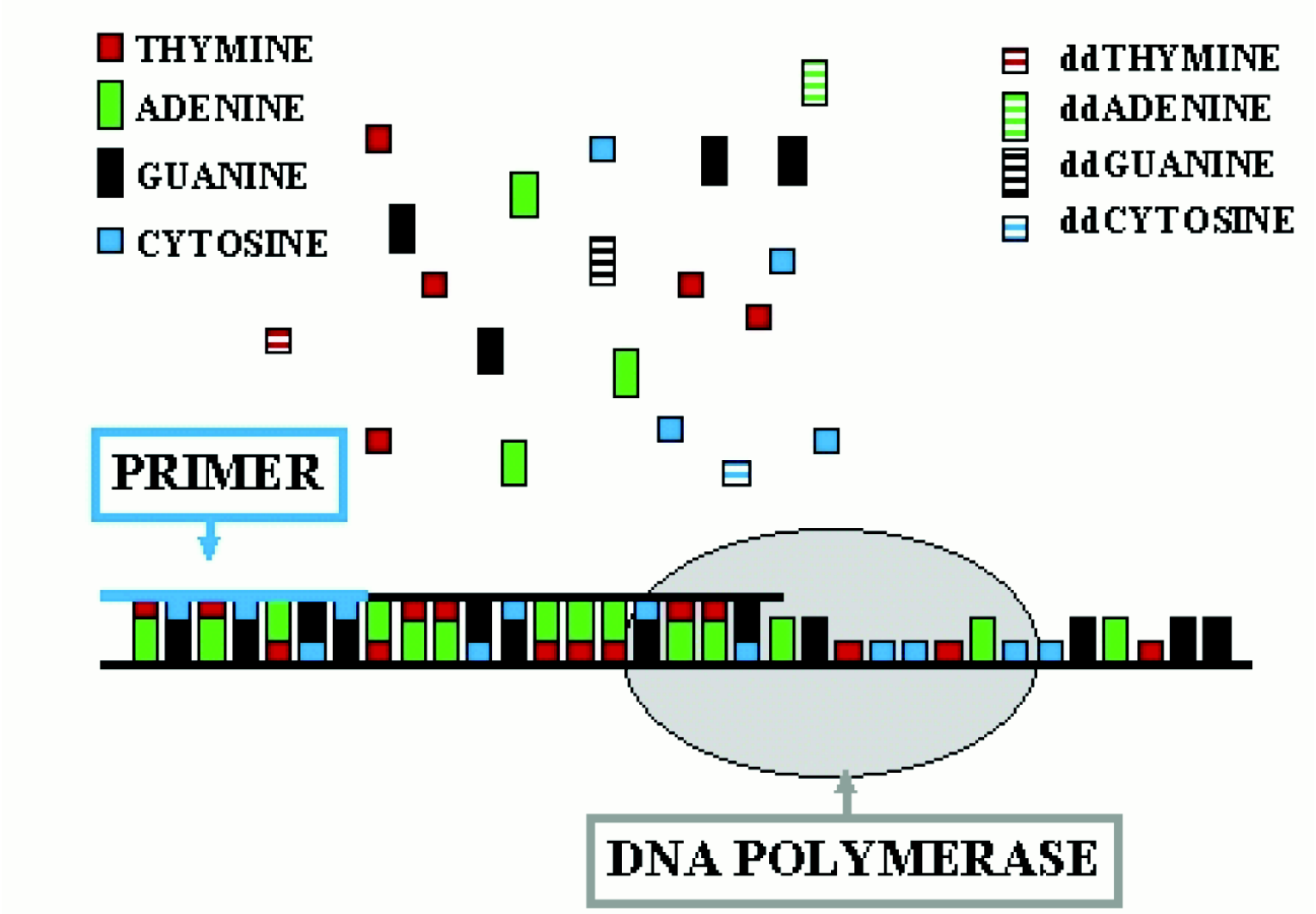
Sanger Sequencing

Sequencing method described by Sanger and Coulson in 1977 that was dramatically better than earlier methods.

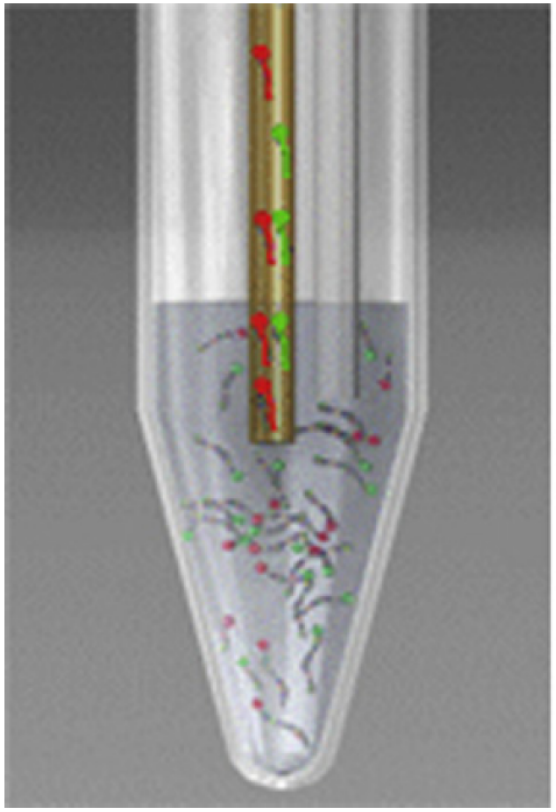
For the next 30 years, 'Sanger sequencing' was practically the only DNA sequencing method used.

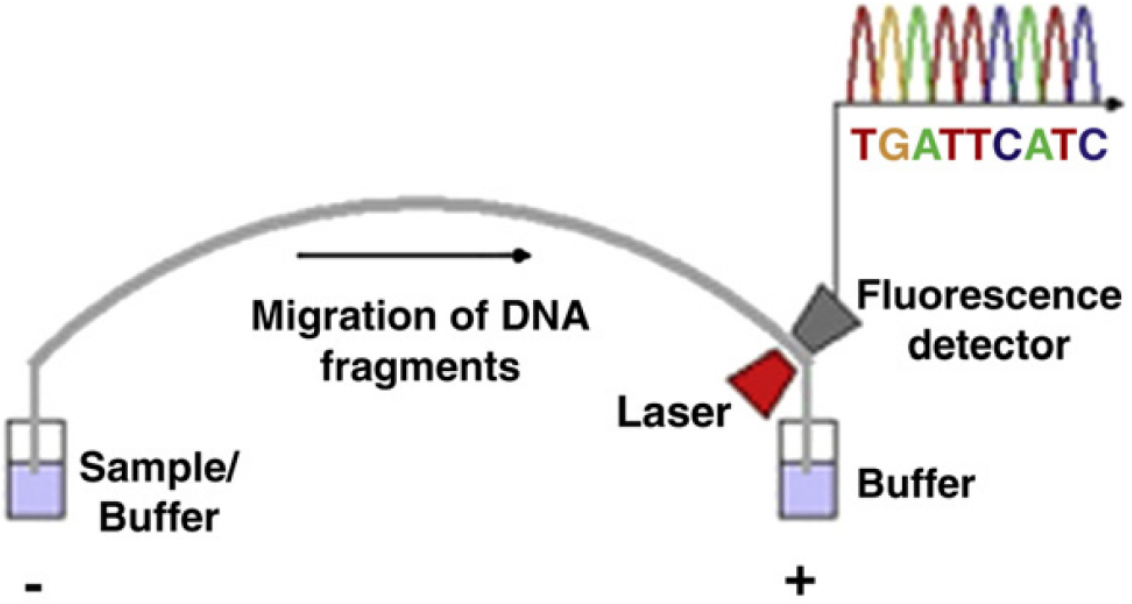
Need for sequencing created many sequencing centers containing hundreds of automated DNA sequencing instruments operated by a large staff.

Chain termination using dideoxynucleotide triphosphates



Capillary electrophoresis





Sequencing by synthesis

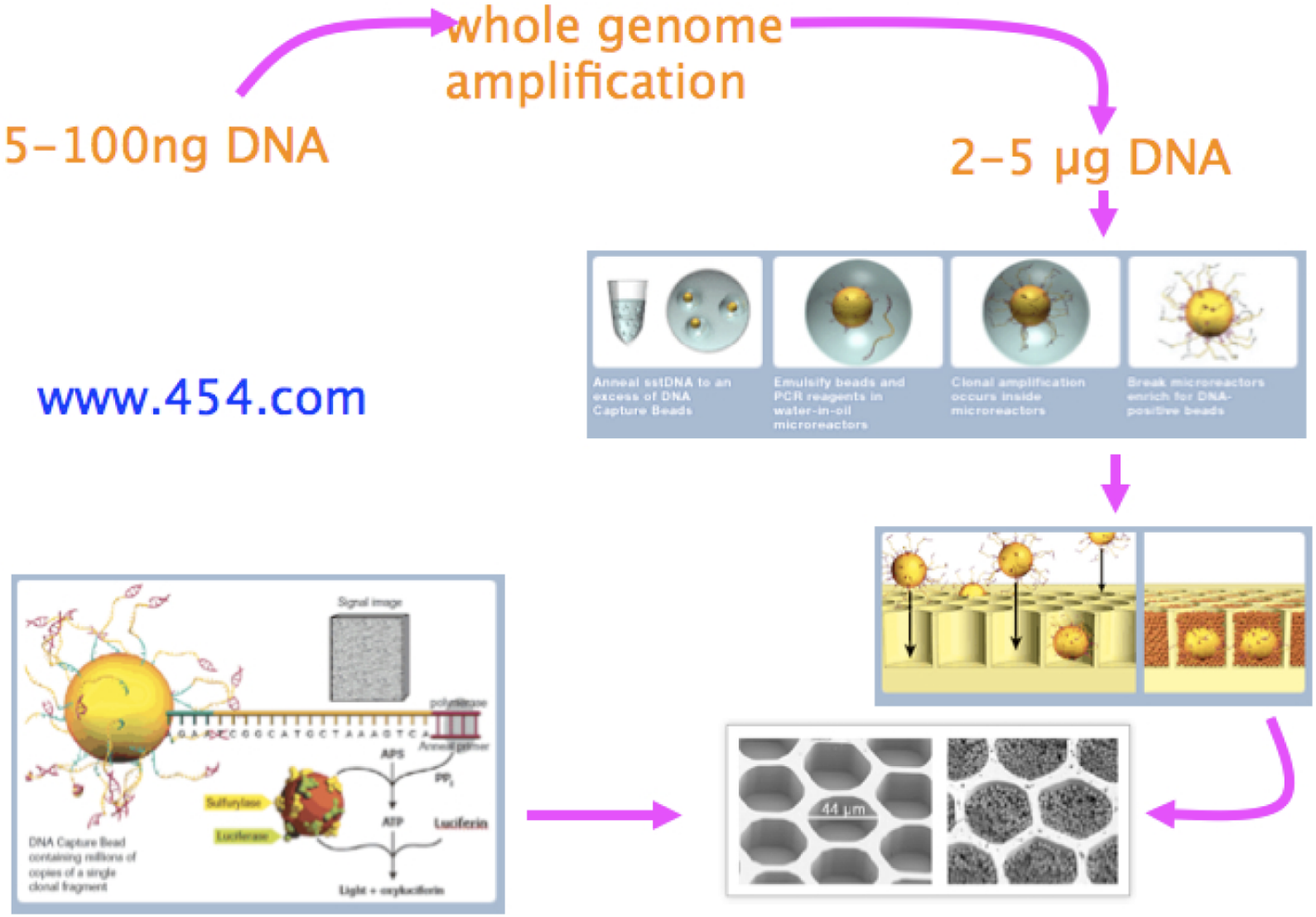
Sequencing revolution began in 2005 with the introduction of the sequencing-by-synthesis technology developed by 454 Life Sciences and the multiplex polony sequencing protocol developed by George Church's lab.

Pyrosequencing involves constructing a complementary DNA strand one base at a time and detecting the actual base that was incorporated.

The first 454 instrument was about 300 times cheaper than previous Sanger technology.

A single next-generation sequencing instrument can generate as much data as several hundred Sanger-type sequencers, but can be operated by a single person.

Sequencing by synthesis: pyrosequencing



Initial concerns about NGS

Concerns raised about:

- sequencing fidelity,
- read length,
- infrastructure cost (and obsolescence of existing Sanger machines),
- large data volume.

Read length

Sanger sequencing:

- Initially read length of Sanger sequencing rarely $>$ 25 bp.
- Read length is now approximately 800 to 1000 bp.
- First 15 to 40 bases have poor quality, and it deteriorates again after 700-900 bases.

Pyrosequencing (sequencing-by-synthesis):

- Initial read length of 100 bp,
- Read length is now approximately 300 to 500 bp.

Re-sequencing

Shorter read length is a problem for de novo sequencing. It's less of a problem for resequencing.

Sequencing closely related organisms can be done using resequencing, in which the assembly is guided by a reference sequence.

Resequencing requires only about 10 times coverage which is much less than for assembling genomes de novo (which needs 25 to 70 times).

In cancer genetics, next-generation sequencing's short read length is a minor disadvantage compared to the much larger number of sequence reads attainable.

Future Improvements in Sequencing Technology

- Further cost reduction: a reduction of 12 orders of magnitude is needed to deliver on the promise of personal genomics, which targets a cost of $1,000 for the resequencing of a human genome.
- Reduced sequencing error rate.
- Skyrocketing quantities of data, creating a bioinformatics analysis bottleneck.

Exponential Growth of Sequencing Capability

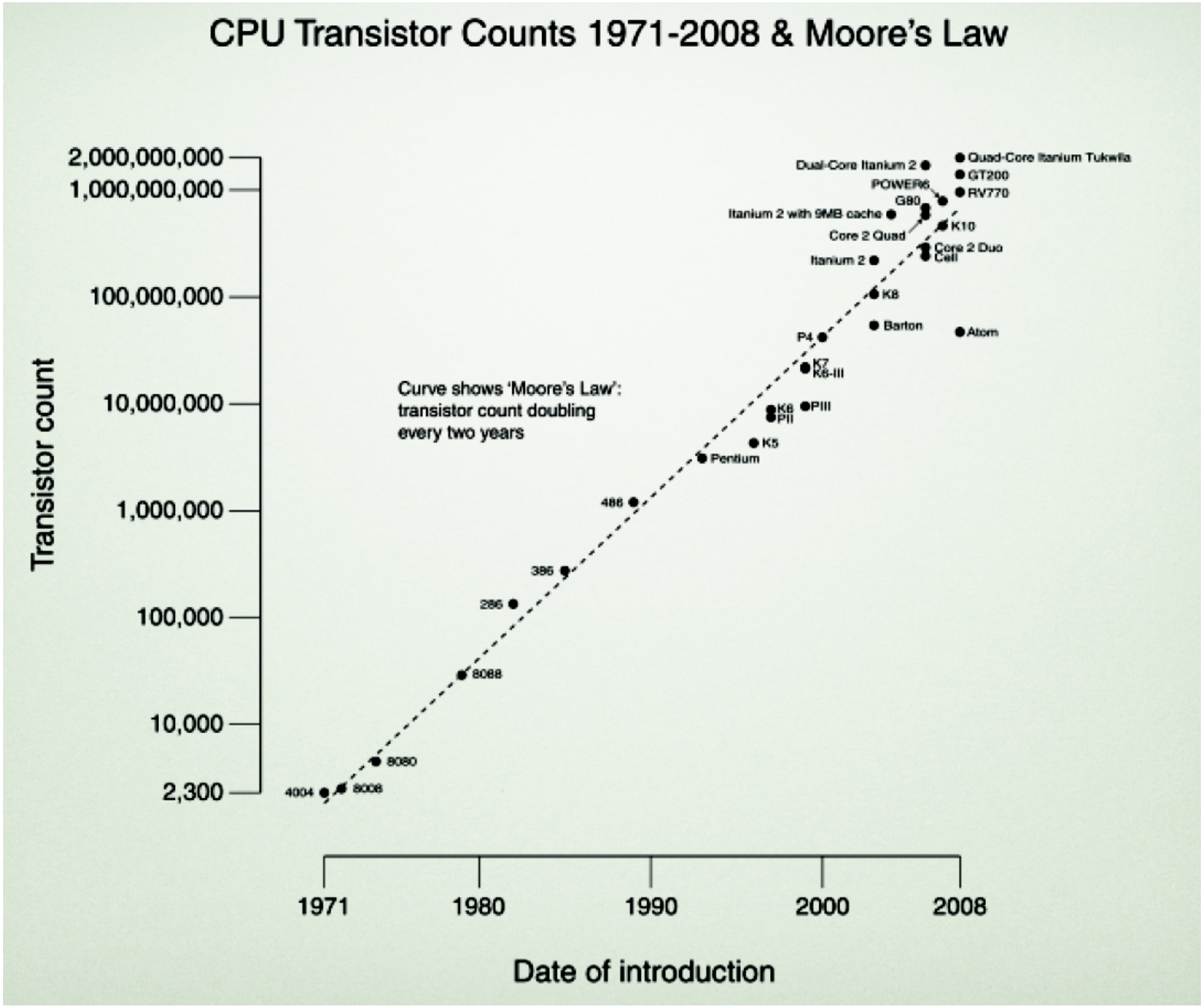
First human genome sequenced over ten years at $3 billion.

In 2007, Watson's genome was sequenced in two months by 454 at $2 million.

In 2008, the cost (list price of reagent) of human genome re-sequencing using Solexa is $250,000.

ABI SOLiD claim to be able to re-sequence at $10,000 this year (2009).

From 2005 to 2009, the cost of DNA sequencing has dropped ten times per year.


CPU Transistor Counts 1971-2008 & Moore's Law
2,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
Transistor count
Curve shows 'Moore's Law': transistor count doubling every two years
Dual-Core Itanium 2
Quad-Core Itanium Tukwila
POWER6
GT200
G80
RV770
Itanium 2 with 9MB cache
K10
Core 2 Quad
Core 2 Duo
Cell
Itanium 2
K8
Barton
Atom
P4
K7
K8-III
K6
PIII
PII
K5
Pentium
486
386
286
8088
8080
4004
8008
1971
1980
1990
2000
2008
Date of introduction

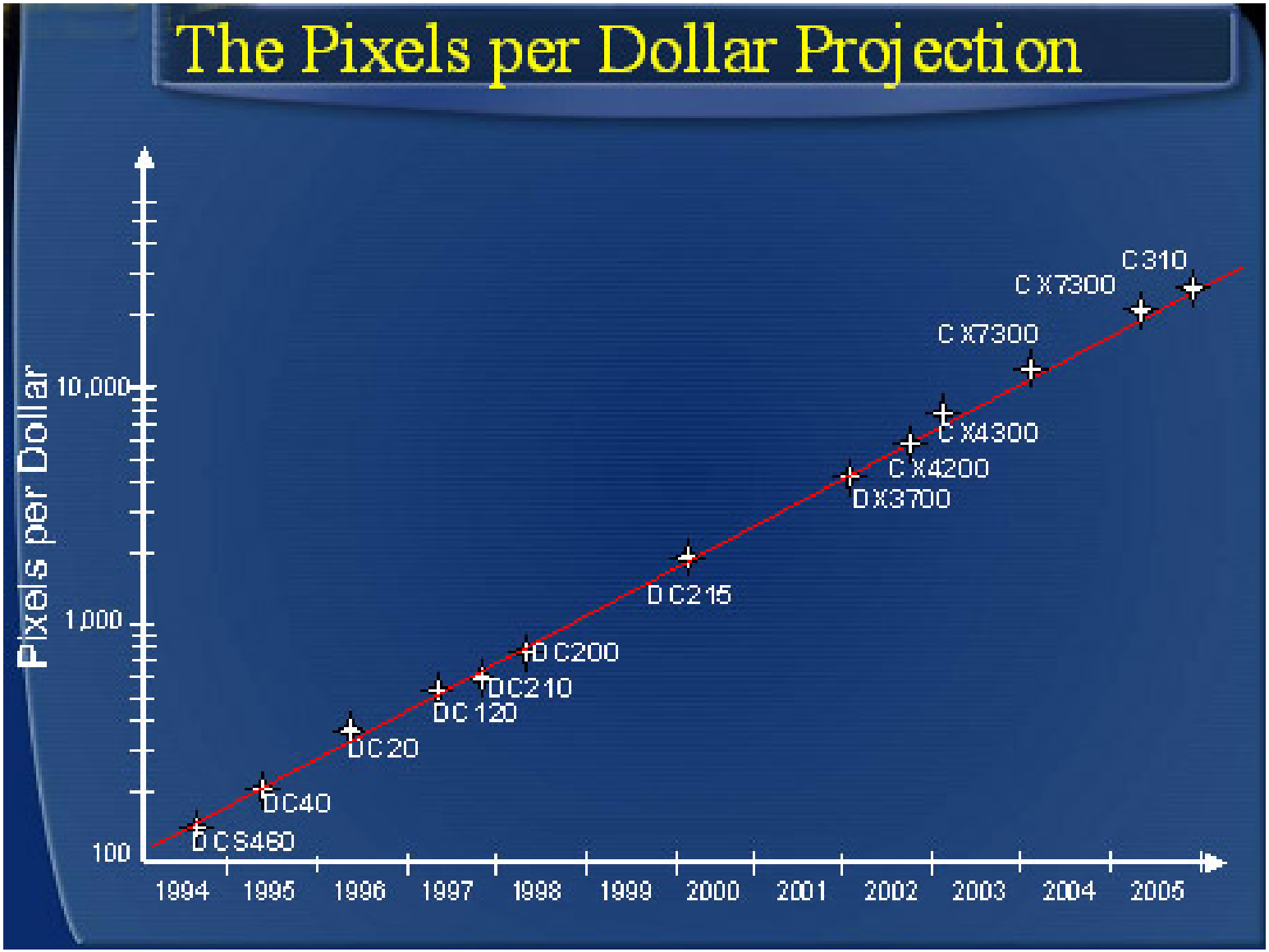
Sequencing expected to follow similar law

Moore's law: computing power a dollar can buy doubles every 18 months

The rate limiting step in next generation sequencing is imaging.

The sequencer's digital camera will increase in capacity following Moore's law.

Pixels per A$ of Kodak Digital Cameras



Next generation sequencing: hardware

Sanger

454

ABI SOLiD

Illumina Solexa

Complete Genomics

Pacific Biosciences

Solexa

8 lanes per run

200 pictures per lane

4x36 pictures for 36-mer

1/4 million pictures per 3-day run $\rightarrow$ 0.5TB of data

Complete Genomics

Reagent cost is 1/1000 of Solexa

Demonstrated 8.8 Gb per machine run per day.

A completed genome sequence on company's web site

June 2009, launch of commercial run: 200Gb per machine run lasting 8 days.

Data center: 60,000 processors and 30 petabytes storage.

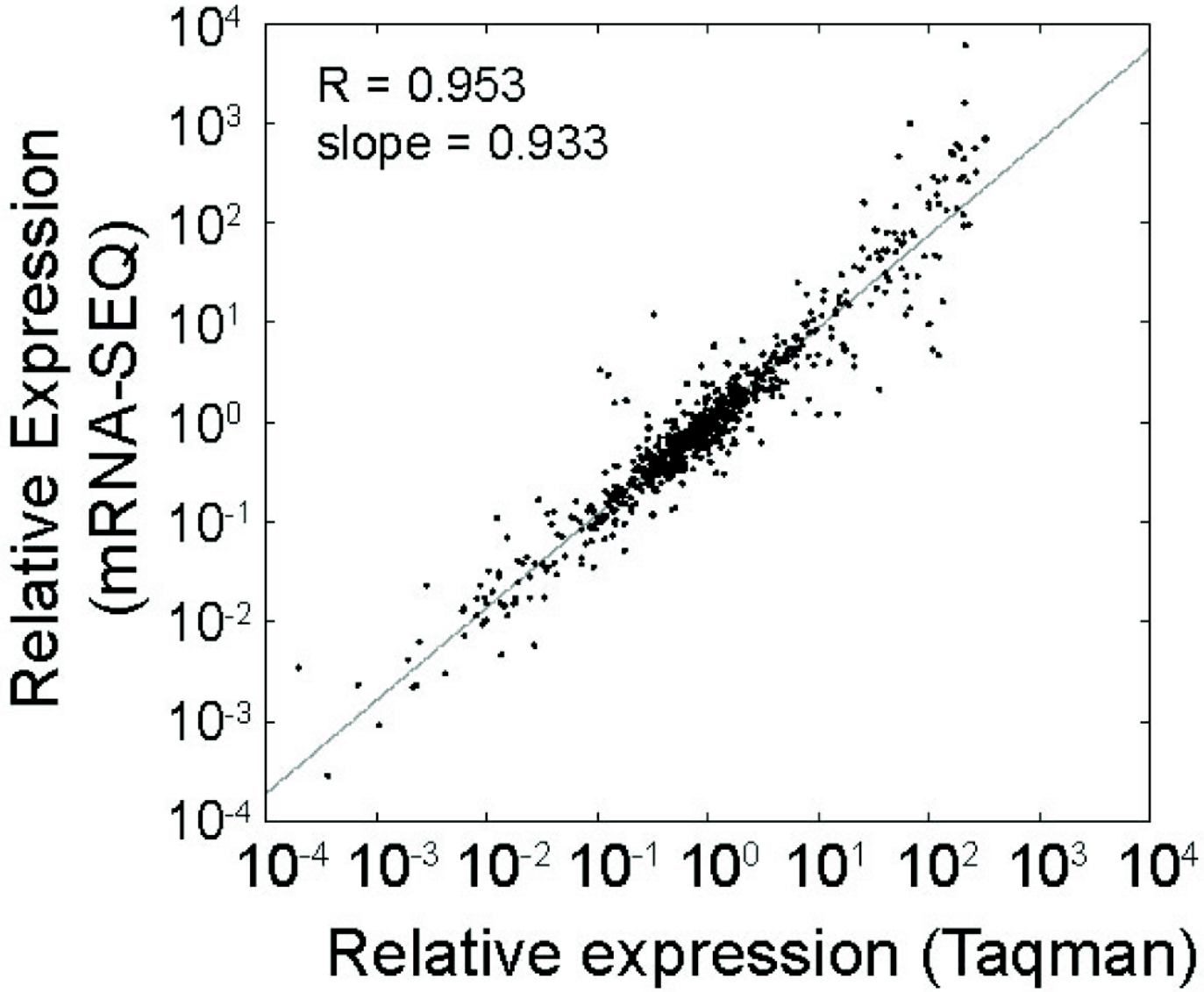
Next Generation Sequencing Methods

	Feature generation	Sequencing by synthesis	Cost per mega base	Cost per instrument	most common error	Read-length
454	Emulsion PCR	Polymerase (pyrosequencing)	$60 ($20)	$500,000	Indel	400 bp
Solexa	Bridge PCR	Polymerase (reversible terminators)	$2	$430,000	Subst.	36 bp
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	$2	$591,000	Subst.	35 bp
Polonator	Emulsion PCR	Ligase (nonamers)	$1	$155,000	Subst.	13 bp
HeliScope	Single molecule	Polymerase (asynchronous extensions)	$1	$1,350,000	Del	30 bp

Jay Shendure & Hanlee Ji, *Nature Biotechnology* **26,** 1135 - 1145 (2008)

1

Digital gene expression: Solexa vs Taqman

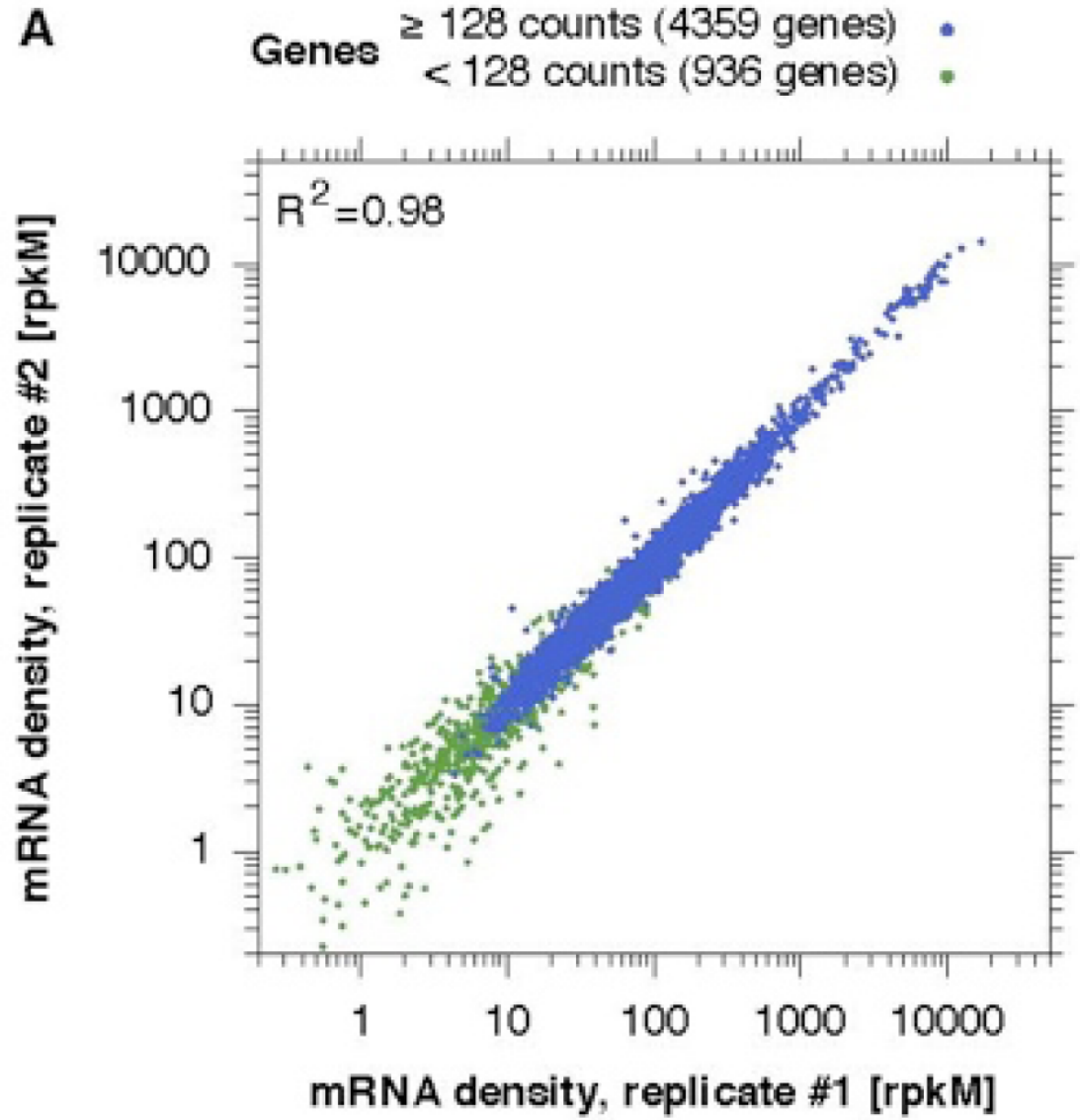


787 RefSeq human transcripts in brain and UHR.

TaqMan is considered a gold standard.

Note that unlike Sanger sequencing, which averages over many molecules, PCR errors in next generation sequencing do not average away.

Digital gene expression: mRNA reproducibility



Clinical Applications

Sanger sequencing is used clinically to identify mutations in selected Mendelian disease genes.

Wider application of sequencing to clinical testing has been limited by cost and throughput.

Array genomic hybridization is used to detect copy number changes that cause intellectual disability and other birth defects.

How can NGS be used to perform these and similar clinical tests more effectively?

Clinical Applications

Detecting Mutations in Disease Influencing Genes

- Ability to sequence entirety of even very large genes at very high coverage
- To increase utilization of machine, pool bar-coded DNA from multiple individuals

Simultaneous Genetic Disease Screening

- Test thousands of candidate disease influencing loci simultaneously

Clinical Applications

Discovering Disease Influencing Genes

- Get information on all genetic changes, not just common tag SNPs

Personalized Adverse Reaction Sequencing

- Screen all for genome variants that influence adverse drug reactions
- Particularly valuable for patients taking many drugs concurrently (e.g. elderly)

Clinical Applications

Improved Cancer Diagnosis and Treatment

- Screen for LOH, CNV across entire genome
- Detect gene mutations and fusions
- Detect accurate detection of mutations in a rare subpopulation of cells
- Estimate fraction of tumor cells in tumor sample

Epigenetics

- Genome-wide tests of epigenetic changes known to cause disease states

Clinical Applications

Identification of Structural Variants

- Detect inversions and other chromosomal rearrangements that do not affect copy number
- CGH can detect unbalanced SVs but not balanced SVs
- Requires long reads or paired-end reads

Clinical Applications

Pathogen Detection and Screening

- Sequence all DNA
- 'Subtract' all human sequences
- 'Subtract' sequences of normal commensal organisms
- Use remaining paired-end reads to assemble denovo sequences and identify pathogens

Ethical Issues

Raised before by genetic testing but NGS amplifies concerns:

- A specific clinical application needs only a fraction of the data collected.
- Other data may be important later.
- We cannot interpret most of the data.
- We can use it to uniquely identify an individual.

Consent

- Does next generation sequencing require a different level or kind of consent than other medical assessments?
- Should next generation sequencing be done when the same question can be answered by a more limited test?
- Is it appropriate to generate whole-genome data that may or may not be of clinical significance but analyze only a small portion of it to answer a specific clinical question?
- Should next generation sequencing be done in children or incompetent adults?
- Is informed consent for next generation sequencing possible?

Releasing Findings from sequence data

Should patients be informed of results that:

- have uncertain clinical significance?
- predict serious disease that cannot be prevented or treated?
- do not have direct implications for them, but do for other family members?

If a genetic analysis reveals findings that have direct implications for the patient's relatives, should they be informed?

Incidental Findings

Should patients be informed of incidental findings that

- Unequivocally predict serious disease that can be prevented or ameliorated by early detection? What if the disease cannot be prevented or ameliorated?

- Indicate an increased (or reduced) risk for disease that can be prevented or ameliorated by early detection? What if the disease cannot be prevented or ameliorated?

Should physicians or clinical laboratories recontact patients if previously collected sequence data is later found to have serious medical implications?

Findings of no medical importance

Should physicians or clinical laboratories provide genomic information that has no medical importance but is of

- Social or personal consequence to the patient (e.g., ancestry or paternity)?
- General interest to the patient (e.g., SNPs associated with athletic or musical ability)?

Storing Sequence Data

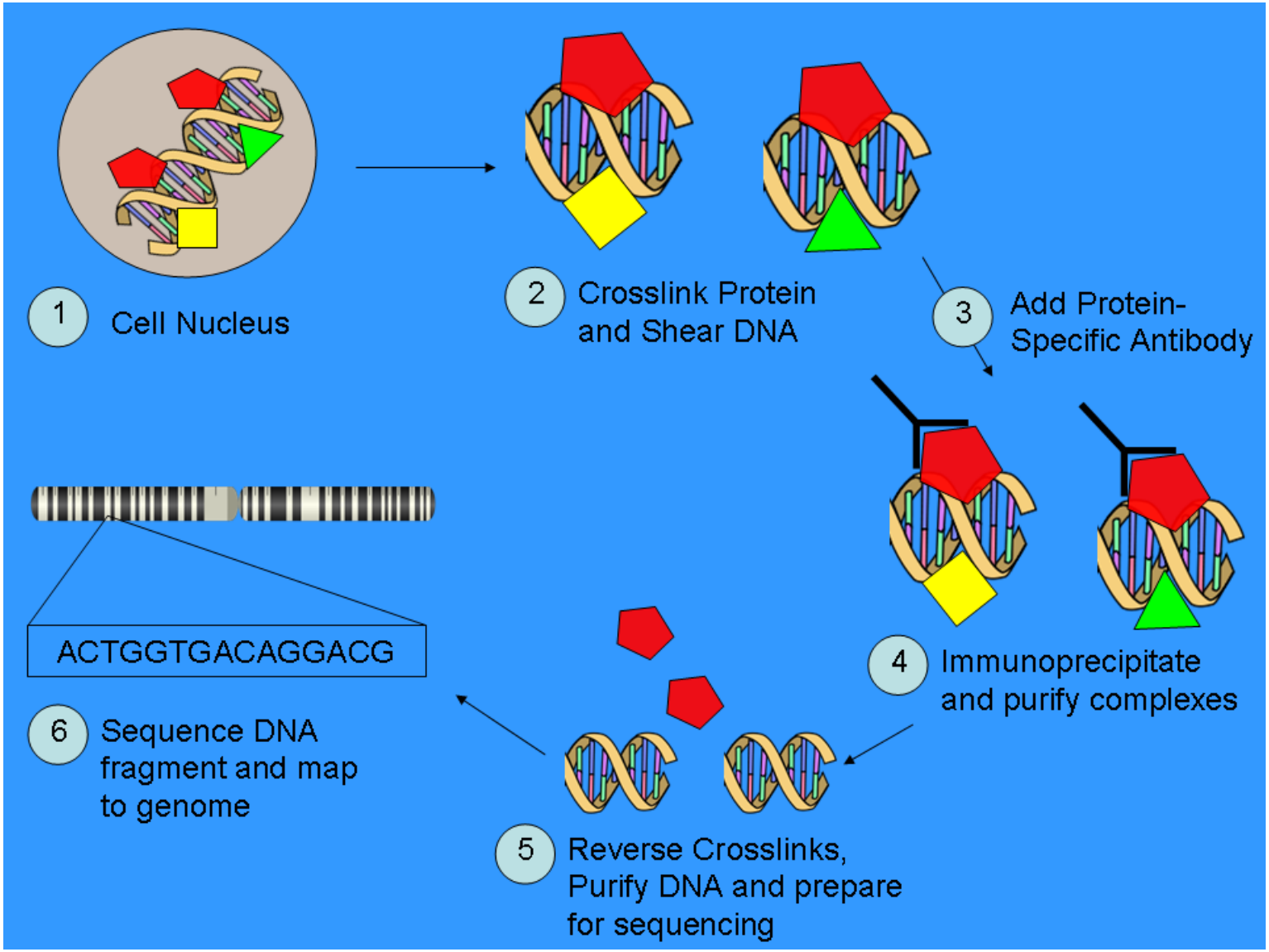
- Where should individual sequence data be stored, and who should be responsible for the stored data?
- Should the individual sequence data be retained for long periods of time (or throughout life or longer) in case future reanalysis is necessary?
- Who should be able to obtain access to an individual's complete genomic sequence? The individual? Any treating physician? Insurance companies? Police?
- Under what circumstances should stored genomic data be used for purposes of identification (e.g., for identification of disaster victims or confirmation of citizenship)?

Applications in Genome Research

A DNA sequence is a bar-code, and therefore an addressing system of a genome.

Share similarities with microarray in measuring amount of DNA by genome locations.

chIP-seq: analyze protein-DNA interactions



RNA-Seq

Use NGS sequencing to measure genome wide mRNA levels:

- Purify RNA
- Bind polyA fraction (mRNA)
- Fragment RNA (200 bp)
- Convert to cDNA by random priming
- Apply adaptors and sequence
- Analyze millions of 25 bp reads

Statistical Issues in mRNA Seq Analyses

Gene length can bias differential expression results.

Not using the standard phi X control lane does not negatively impact detection of different expression, but leads to more balanced and cost-effective designs.

Different lanes have total different read counts (sequencing depth).

Quantile-based normalization methods work better than scaling by total lane counts (e.g. RPKM model of Mortazavi).

Reading for next time

Read the TCGA paper *Comprehensive genomic characterization defines human glioblastoma genes and core pathways*, Nature 455(23), October 2008, 1061–1068, and its supplementary methods and data (available online at http://www.nature.com/nature/journal/v455/n7216/abs/nature07385.html).