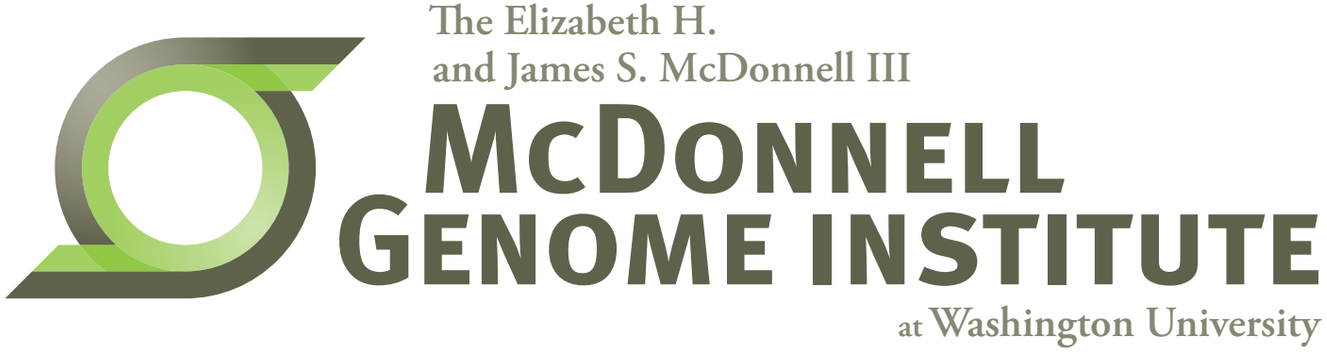
Genome Technology

Ira Hall, Ph.D.
Associate Director, McDonnell Genome Institute
Associate Professor, Department of Medicine
Washington University School of Medicine
ihall@genome.wustl.edu



The scope of the problem: basic genome facts

The human genome is big

- 3 billion nucleotides, 2 copies.

Mutations are arising constantly

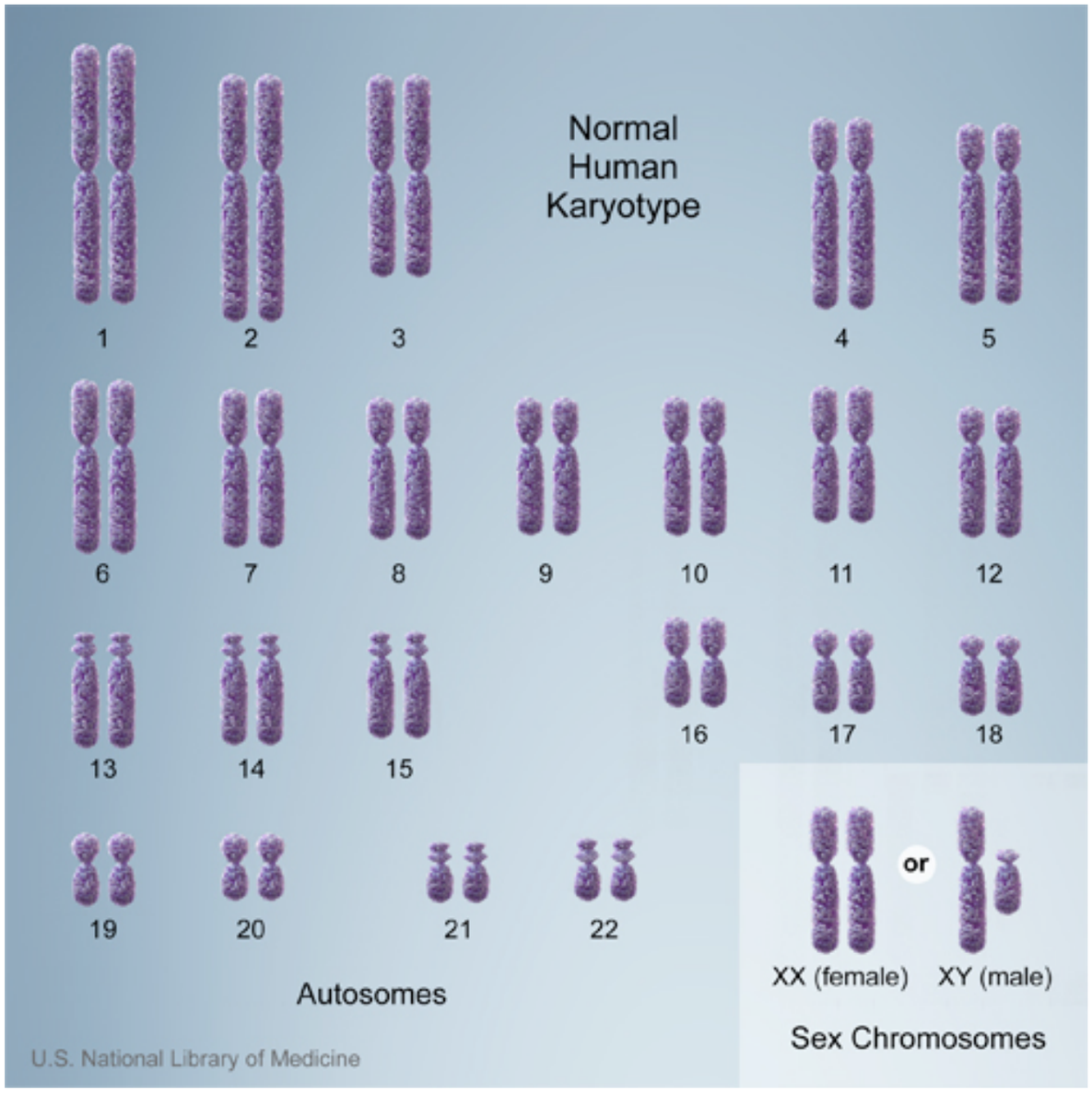
- Roughly 90 per human generation.
- Roughly 1 per somatic cell division.

Human genomes are diverse

- Germline: ~4 million genetic differences between 2 humans; 1 per ~800 bp.
- Cancer: 10^2 - 10^4 somatic mutations.

Most variants are neutral, or benign

- 1-10% of genome is “functional”.
- But, much buffering and redundancy.
- My guess: 0.1% of variants are “functional” = 4,000 germline variants per person (it could be 10X more).



First generation DNA sequencing technologies

Sanger sequencing (1977)

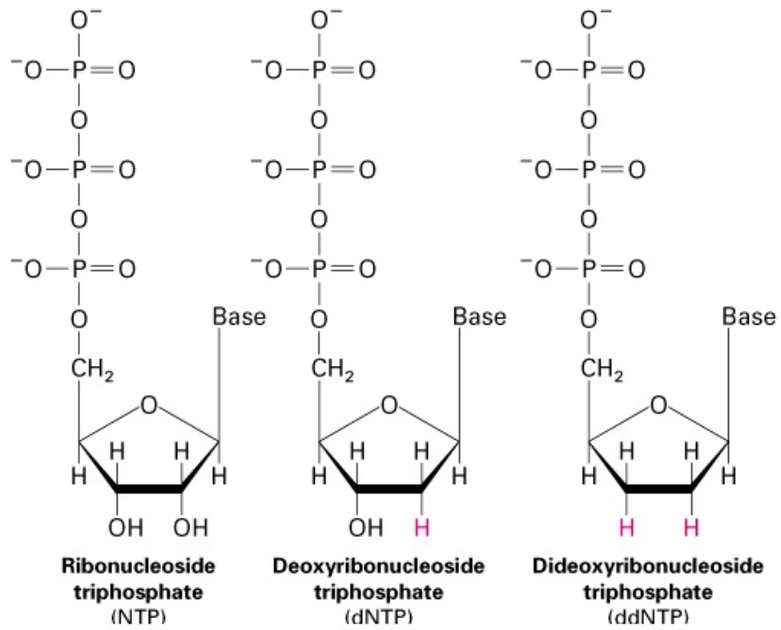
How it works:

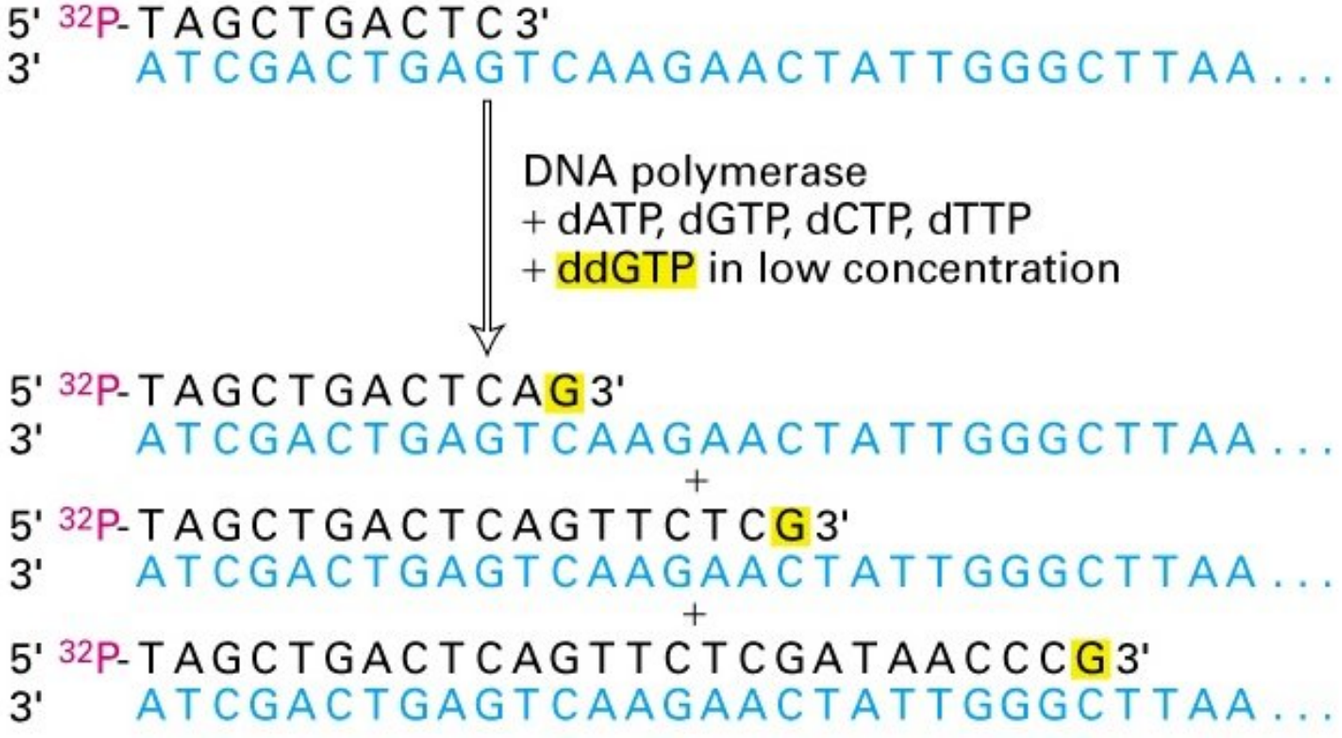
(1) Isolate a pure DNA fragment in a tube.

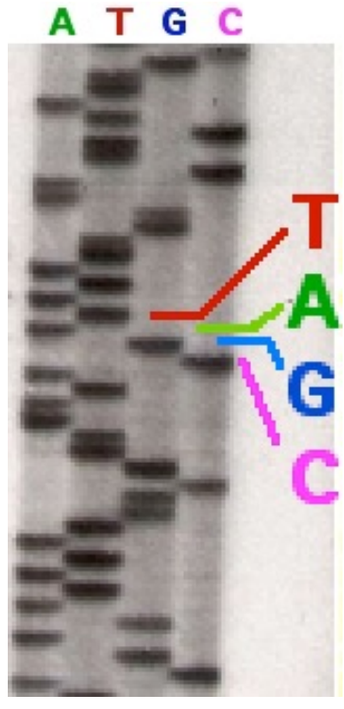
(2) Anneal radioactive primers to single stranded DNA.

(3) Add polymerase + dNTPs + ddNTP terminators.
DNA synthesis proceeds until a ddNTP is incorporated.

(4) Run gel: 1 lane per base, visually interpret ladder.



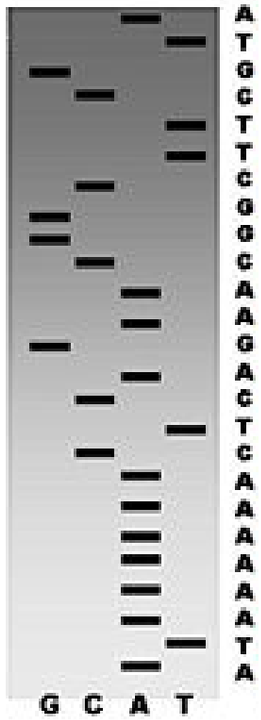




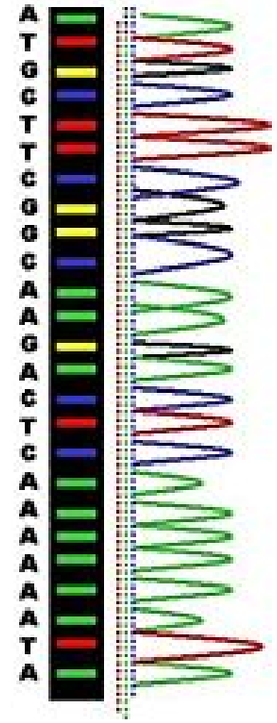
- A competing technology based on chemical DNA degradation was developed by Maxam and Gilbert (1977)
- Fred Sanger & Walter Gilbert shared the Nobel Prize in 1980 for DNA sequencing methods.

Automated Sanger sequencing with dye-labelled primers (Hood, 1985)

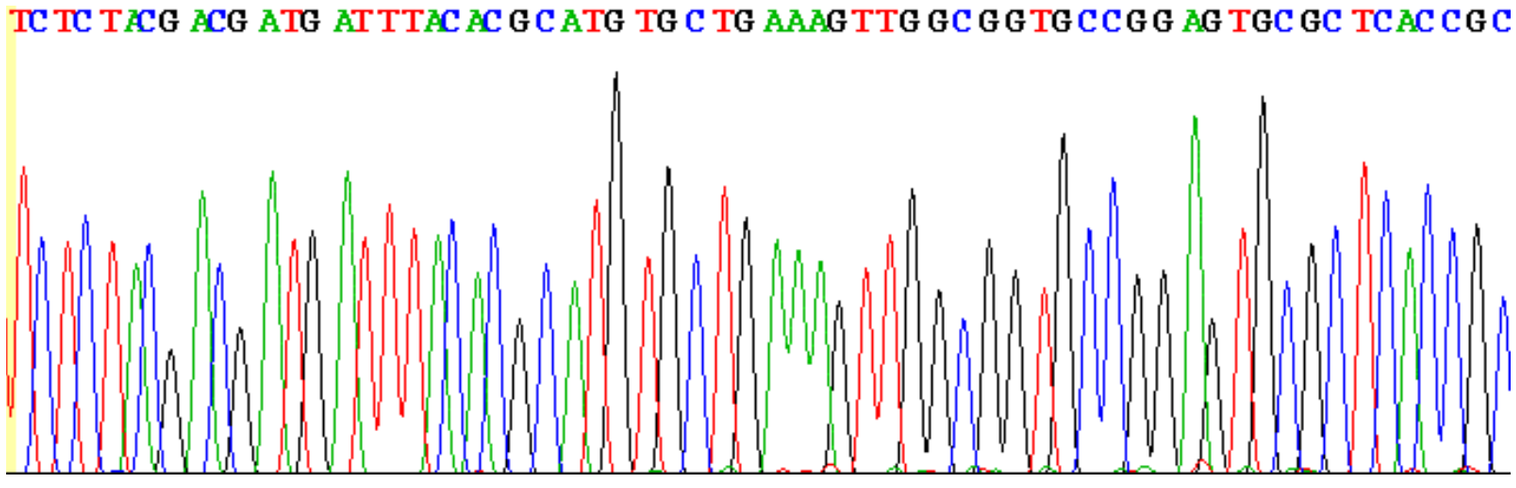
before



after



* By the 1990's we could obtain ~500 bp reads with a ~1% error rate

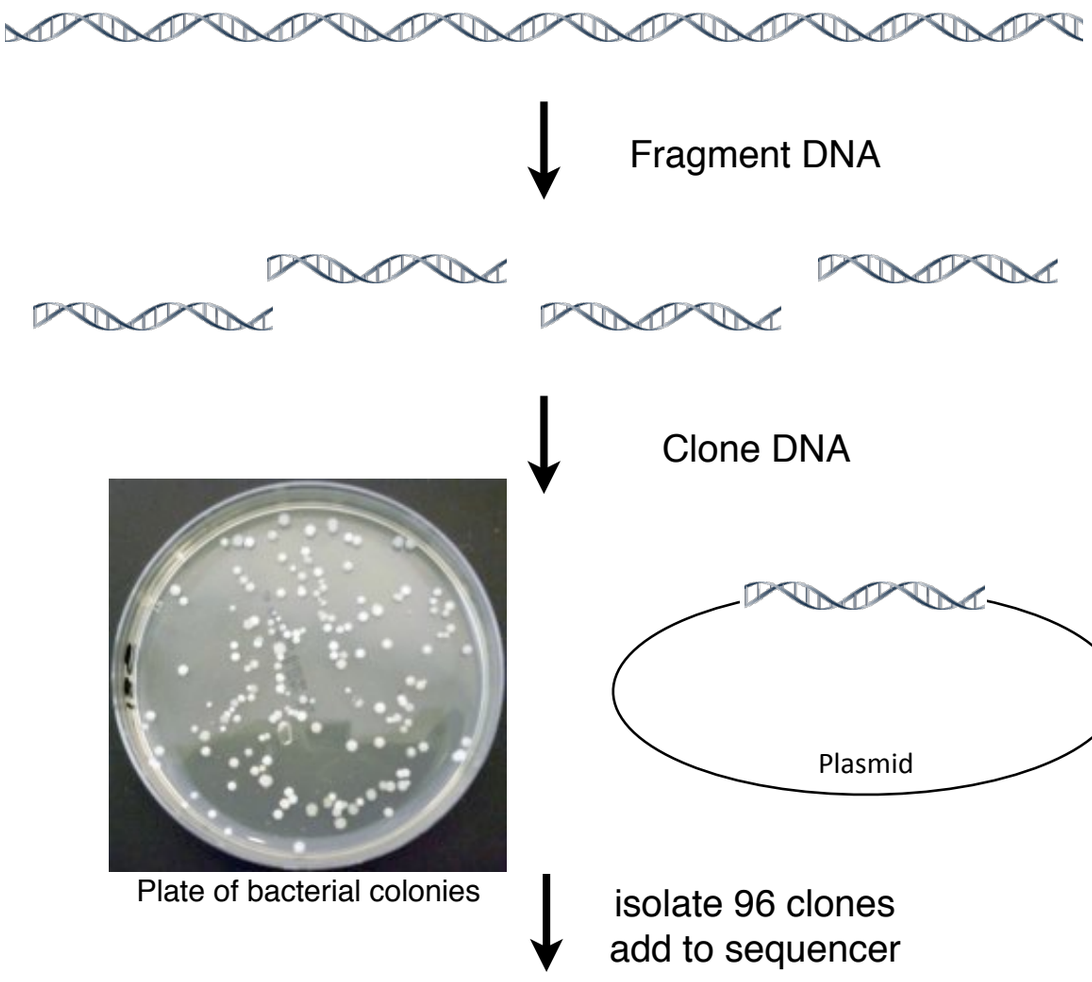


Read: The string of contiguous bases generated by a single sequencing reaction.

Read Length: The number of contiguous bases generated by a single sequencing reaction.

Error Rate: The fraction of bases that are read incorrectly by the sequencer.

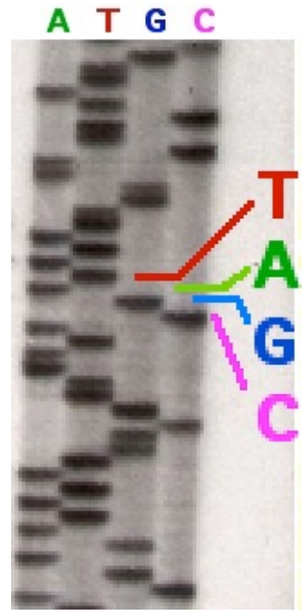
With first generation sequencing methods, each DNA molecule must be isolated and sequenced separately



- Need to isolate and amplify each individual DNA molecule by cloning it into a “vector” (e.g., a plasmid).
- 1 DNA molecule –> 1 bacterial colony –> 1 tube –> 1 physically distinct sequencing reaction.
- This makes the technique extremely slow, laborious and expensive for organisms with large genomes.
- Thus, DNA sequencing was not a very useful tool for mapping human disease genes in the 20th century.

Sanger sequencing throughput timeline

1977: Fred Sanger
- 1 hardworking technician + 1 sequencer = 700 bases per day



1985: ABI 370 (first automated sequencer)
- 5,000 bases per day

1995: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers)
- 19,000 bases per day

1999: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots)
- 400,000 bases per day
- 205 years to sequence human genome
- Or, 205 sequencers ($61.5M) + 205 technicians = 1 year

ABI 3700 ($300K)

<u>Sequencing Throughput:</u> The number of bases that can be interrogated per unit time/labor

2000-2003: Two “complete” human genomes

Public Project: (NIH + Universities)

10 years, $3 billion

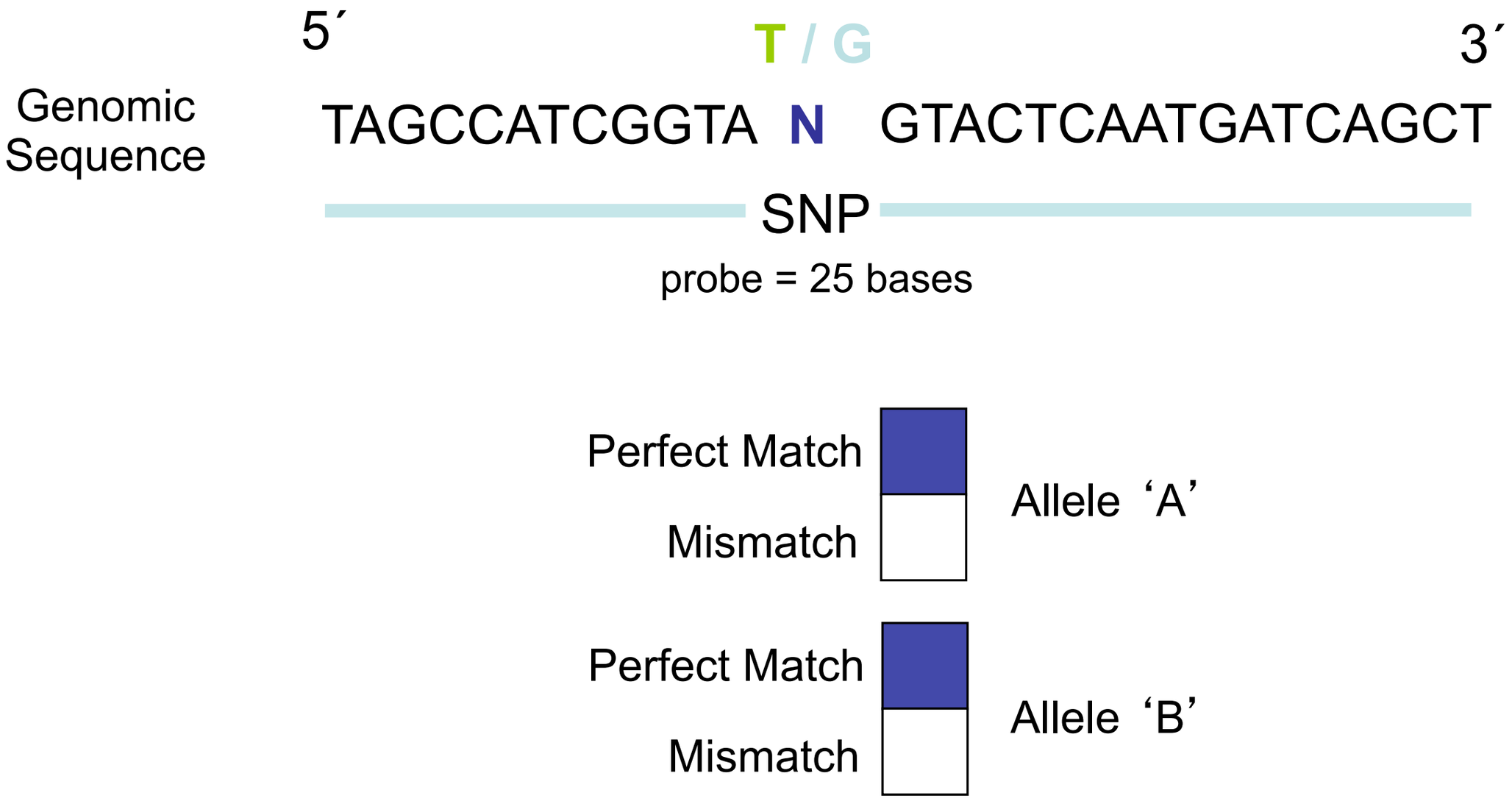
Private Project (Celera Corp.)

2 years, $300 million

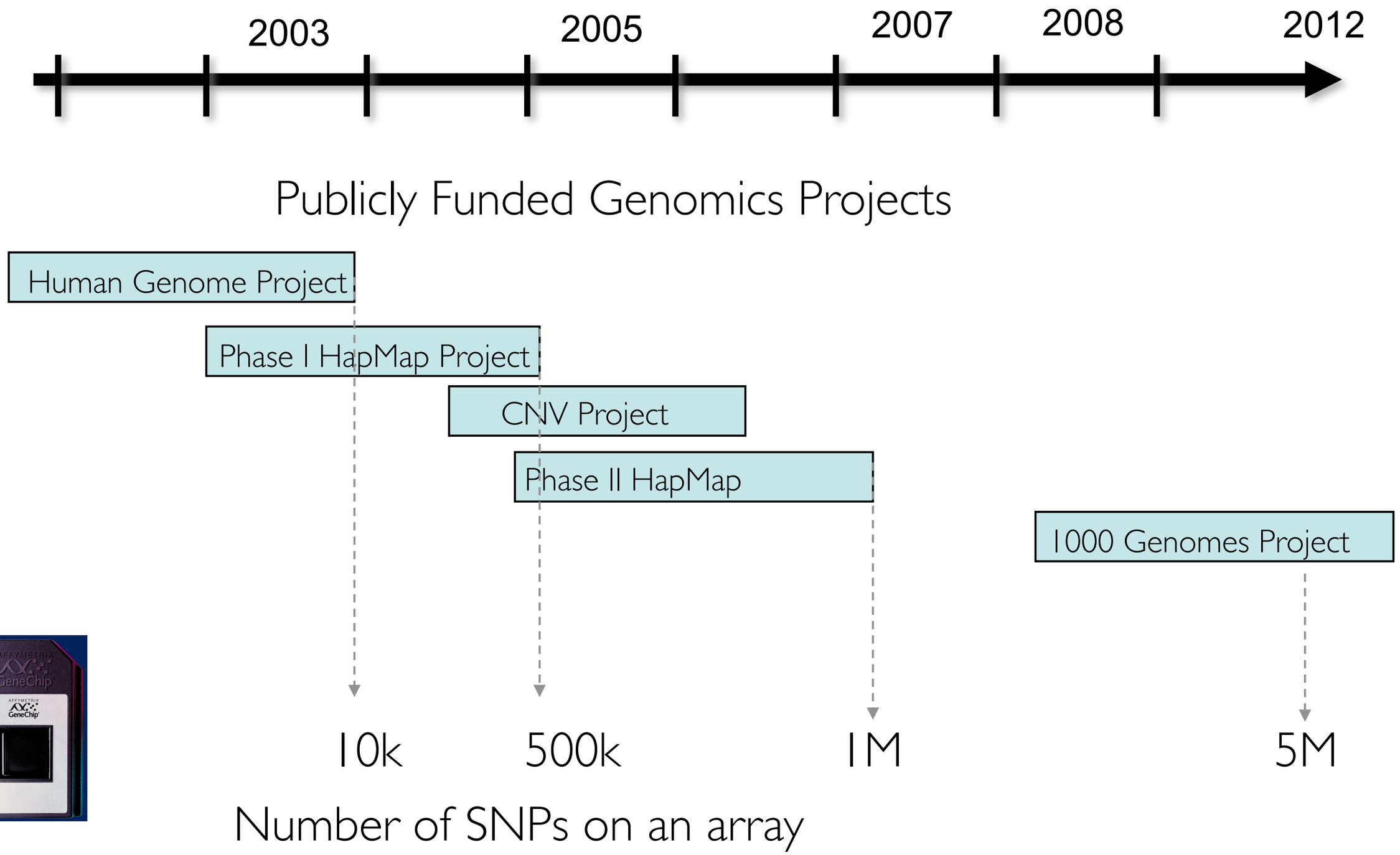
IMPORTANT: We refer to the genome sequence generated by the public project as the “reference genome”. The reference genome has been invaluable for medical research.

Affymetrix SNP arrays

SNP probe design



Where does the content come from?



SNP genotyping platforms

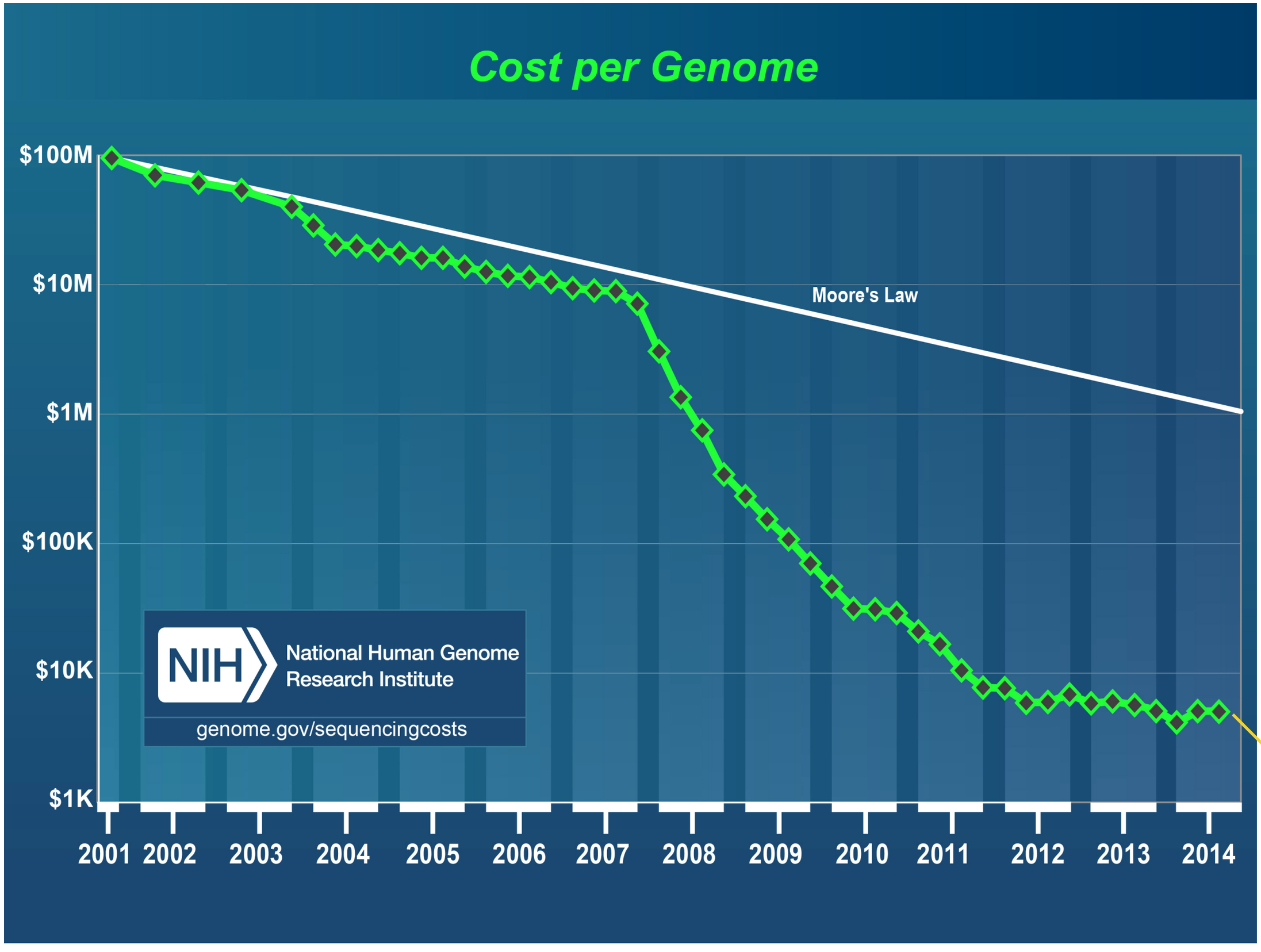
	SNP probes	CNV probes	Cost/sample
Affymetrix 6.0	900K	950K	$450
Affymetrix Axiom	600K	11K	?
Illumina Omni5	5M	0?	$1000
Illumina Omni1	1M	100K	$450
Illumina OmniExp	720K	0	$250

Typical Caucasian is polymorphic at 3M sites, Africans a little more, East Asians a little less

*Prices as of 2012

Second generation DNA sequencing

Genome sequencing costs: 2001-2015



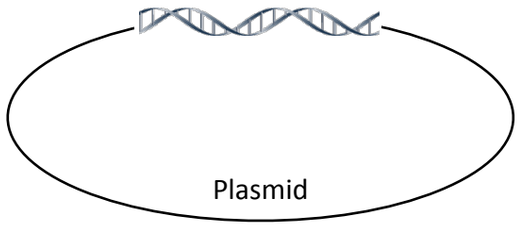
What changed? The level of parallelism. By 7 orders of magnitude!

First Generation:

Fragment DNA

Clone DNA

Plate of bacterial colonies



isolate 96 clones
add to sequencer

ABI 3700

96 well plate

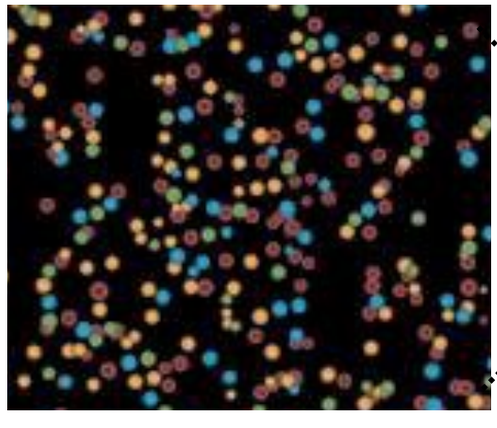
Output: 96 reads, each ~500 bp long

Second Generation:

Fragment DNA

Add adaptors

Add sample directly
to sequencer



~2 billion “clusters” in
a 4 x 1 inch flow-cell

Illumina HiSeq 2500

Output: 2 billion reads, each ~100 bp long

Second generation DNA sequencing technologies

- A.k.a., next-generation sequencing (NGS), high throughput sequencing (HTS), massively parallel sequencing, etc.
- 2005: 454 Technologies (Roche). Obsolete.
- **2006-present: Illumina (formerly Solexa). This is now the dominant technology, with a near monopoly.**
- 2006: SOLiD (Applied Biosystems, Life Technologies). Sequencing by DNA ligation. Obsolete.
- 2009-present: Complete Genomics (BGI). Sequencing as a service. Nearly obsolete.
- 2010-present: Ion Torrent (Life Technologies). This technology uses ion semiconductors to read DNA and holds considerable promise. It is not yet a real competitor to Illumina. We don't have time to talk about it. See wikipedia.

Three popular Illumina sequencers

	MiSeq	HiSeq 2500		HiSeq X*
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	N/A	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-65 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

*The MiSeqDx is the first next-generation sequencing system to receive FDA approval.

* “Rapid Run” mode enables time-sensitive clinical whole genome sequencing

http://www.illumina.com/systems/sequencing.html

An Illumina X10 Cluster (10 “integrated” sequencers)

- The $1000 human genome has finally arrived!
- Caveat: The X10 cluster costs $10 million, and sequencing costs do not include labor, data storage, electricity, etc.
- Throughput: one run: 3 days, 165 genomes = ~20,000 genomes / year
- Parallelism: 3 billion molecules per flowcell, 2 flowcells per machine. 60 billion molecules are sequenced simultaneously.
- We (McDonnell Genome Institute) purchased one last year. A handful of other US institutions also have one.

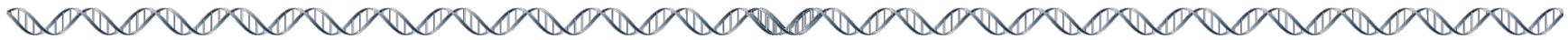
Second generation DNA sequencing summary

- **High throughput and inexpensive. More so each year.**
- **Short reads; initially very short (25-35 bp), but getting longer each year (now 100-300 bp).**
- **Major upgrade relative to microarrays**
  - We are directly sequencing DNA, not indirectly measuring the abundance of molecules that we already know exist.
  - Digital; unlike microarrays, does not rely on fluorescence. This improves dynamic range for molecule counting applications (e.g., RNA expression).
  - Assays are not limited to known genes/isoforms/loci/SNPs where probes can be designed. More comprehensive = less biased = novel discoveries.
- **The impact has been profound**
  - Big projects: 1000 Genomes Project, The Cancer Genomes Project (TCGA), International Cancer Genome Consortium (IGCG), Pediatric Cancer Genome Project (PCGP), UK 10,000 Genomes Project, ENCODE, Human Microbiome Project, etc.
  - Small labs, big data.
  - Many new clinical assays (e.g., genetic testing, cancer)

Application #1: Whole Genome Sequencing

A modern genome sequencing experiment

genomic DNA



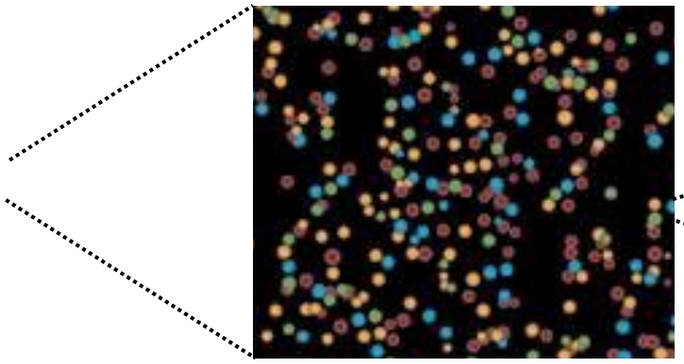
Shear to desired length (~400 bp)

DNA fragments

ligate adapters, size select

sequencing library

Illumina HiSeq



clusters on a flow-cell

500 million paired-end reads (“read-pairs”)

~100 bp ~200 bp ~100 bp

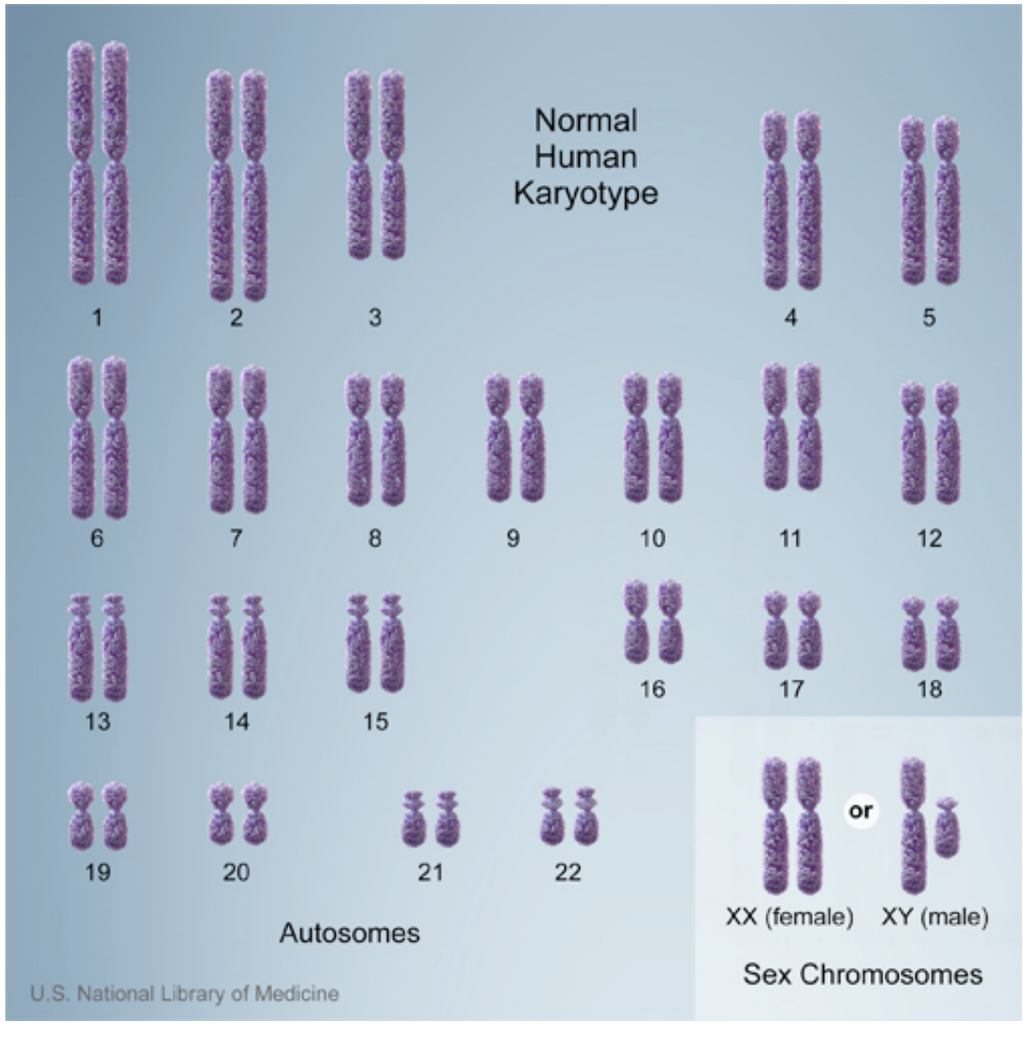
5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

Ideally, we would "stitch" reads together using the process of whole genome assembly to produce the complete diploid genome of that person, or tumor. This is not possible with second generation DNA

The raw data: 500 million read-pairs

5' GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5'

The genome (3.2 billion bp)



Why is genome assembly so hard? Reads are short, the genome is big and complex.

- The human genome is an nasty beast: it is large, complex, and laden with repetitive elements.
- When repeat size exceeds read-length, it is impossible to assemble through them. This causes errors and gaps.
- Long-range sequence information (10-100 kb) is required to resolve complex regions. Short reads (~125 bp) do not suffice.

<u>KEY POINT:</u> We do not sequence human genomes from scratch with second generation technologies. It is too hard.

Instead, we <u>infer</u> genome variation by comparing raw sequencing data from a given individual (or tumor) to the high quality reference genome produced by the public human genome project in ~2003.

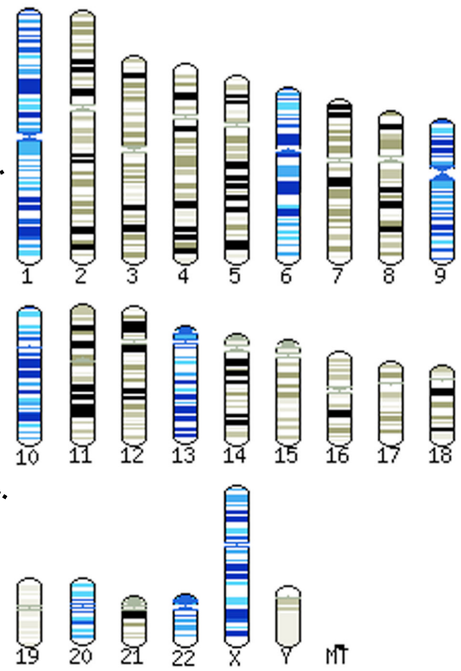
So, we are not really “sequencing genomes”. We are mapping genome variation, indirectly and imperfectly.

Aligning reads to the reference genome

(1) a read-pair (2 x 100 bp)

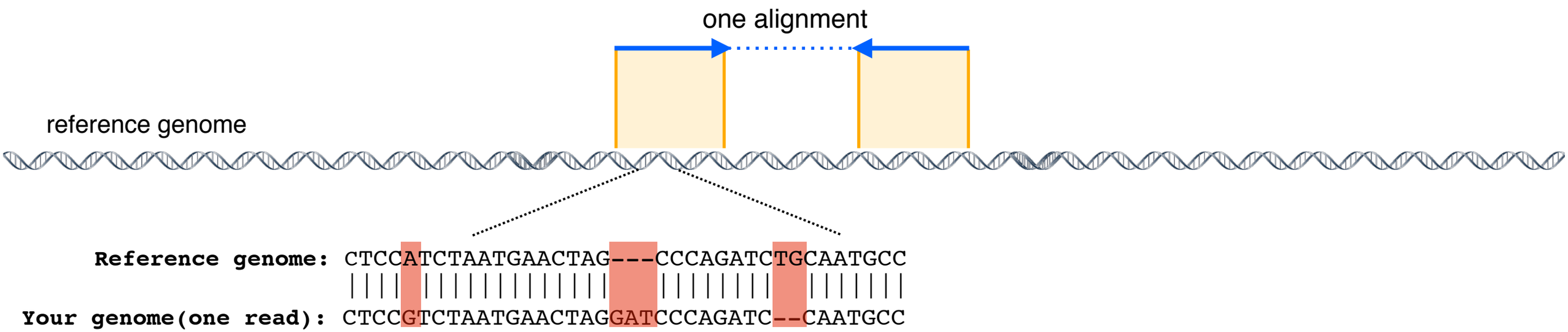
5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

(2) The human reference genome (~3 billion bp)



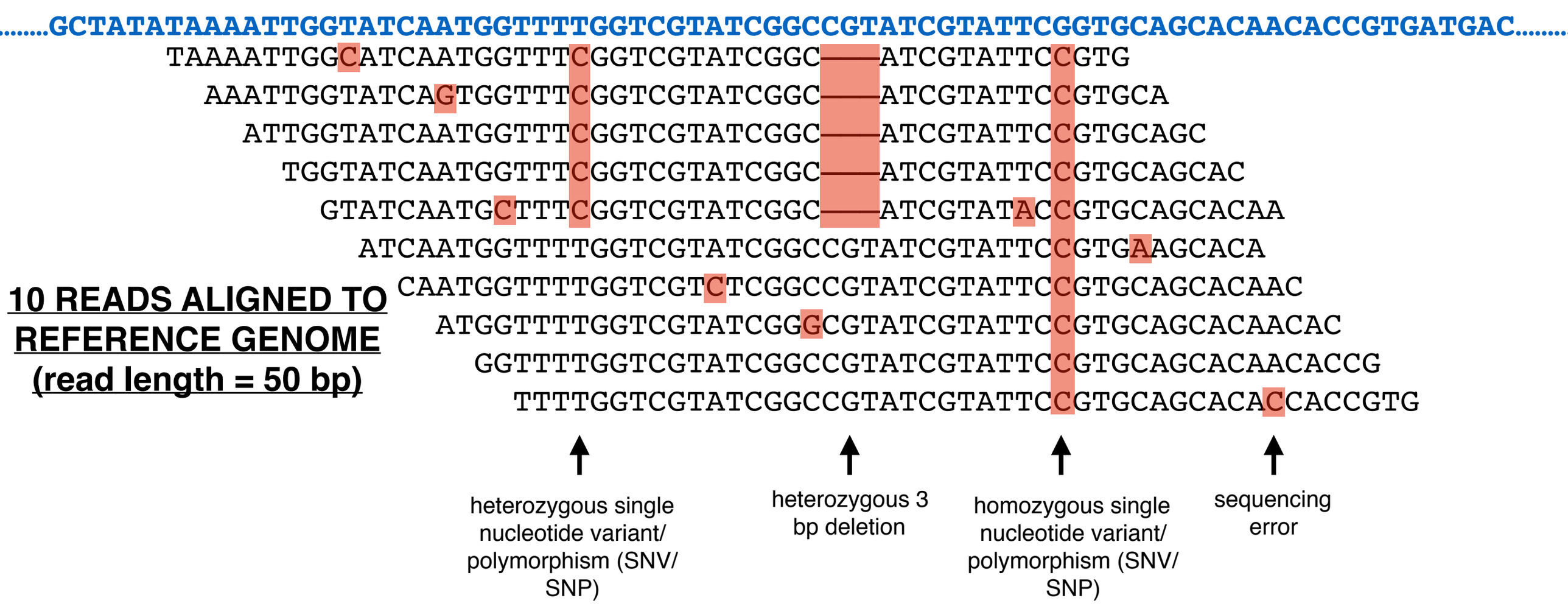
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG
TATGGCAATTAAAATTGGTATCAATGGTTTTGGTCGTATCGGCCGTATCGTATTCCGTGCAGCACAACAC
CGTGATGACATTGAAGTTGTAGGTATTAACGACTTAATCGACGTTGAATACATGGCTTATATGTTGAAAT
ATGATTCAACTCACGGTCGTTTCGACGGCACTGTTGAAGTGAAAGATGGTAACTTAGTGGTTAATGGTAA
AACTATCCGTGTAACTGCAGAACGTGATCCAGCAAACTTAAACTGGGGTGCAATCGGTGTTGATATCGCT
GTTGAAGCGACTGGTTTATTCTTAACTGATGAAACTGCTCGTAAACATATCACTGCAGGCGCAAAAAAAG
TTGTATTAACTGGCCCATCTAAAGATGCAACCCCTATGTTCGTTCGTGGTGTAAACTTCAACGCATACGC
AGGTCAAGATATCGTTTCTAACGCATCTTGTACAACAAACTGTTTAGCTCCTTTAGCACGTGTTGTTCAT
GAAACTTTCGGTATCAAAGATGGTTTAATGACCACTGTTCACGCAACGACTGCAACTCAAAAAACTGTGG
ATGGTCCATCAGCTAAAGACTGGCGCGGCGGCCGCGGTGCATCACAAAACATCATTCCATCTTCAACAGG

(3) Alignment of the “read-pair” to the reference genome gives coordinates describing where in the human genome the read-pair came from, and whether there are any sequence differences.



To distinguish genome variation from sequencing and alignment errors, we weigh evidence from multiple

REFERENCE GENOME (HAPLOID)



10 READS ALIGNED TO REFERENCE GENOME (read length = 50 bp)

IMPORTANT: To do this, we must sequence each base in the genome multiple times. The number of times, on average, that each base in the genome is sequenced is termed *coverage*. Today, the industry standard for Illumina whole genome sequencing is >30X coverage relative to the haploid reference genome. Since our genomes are diploid, this corresponds to >15X coverage for each chromosome.

Well, this doesn’t look so hard.

Why hasn’t whole genome sequencing become a routine assay yet?

Whole genome sequencing challenges

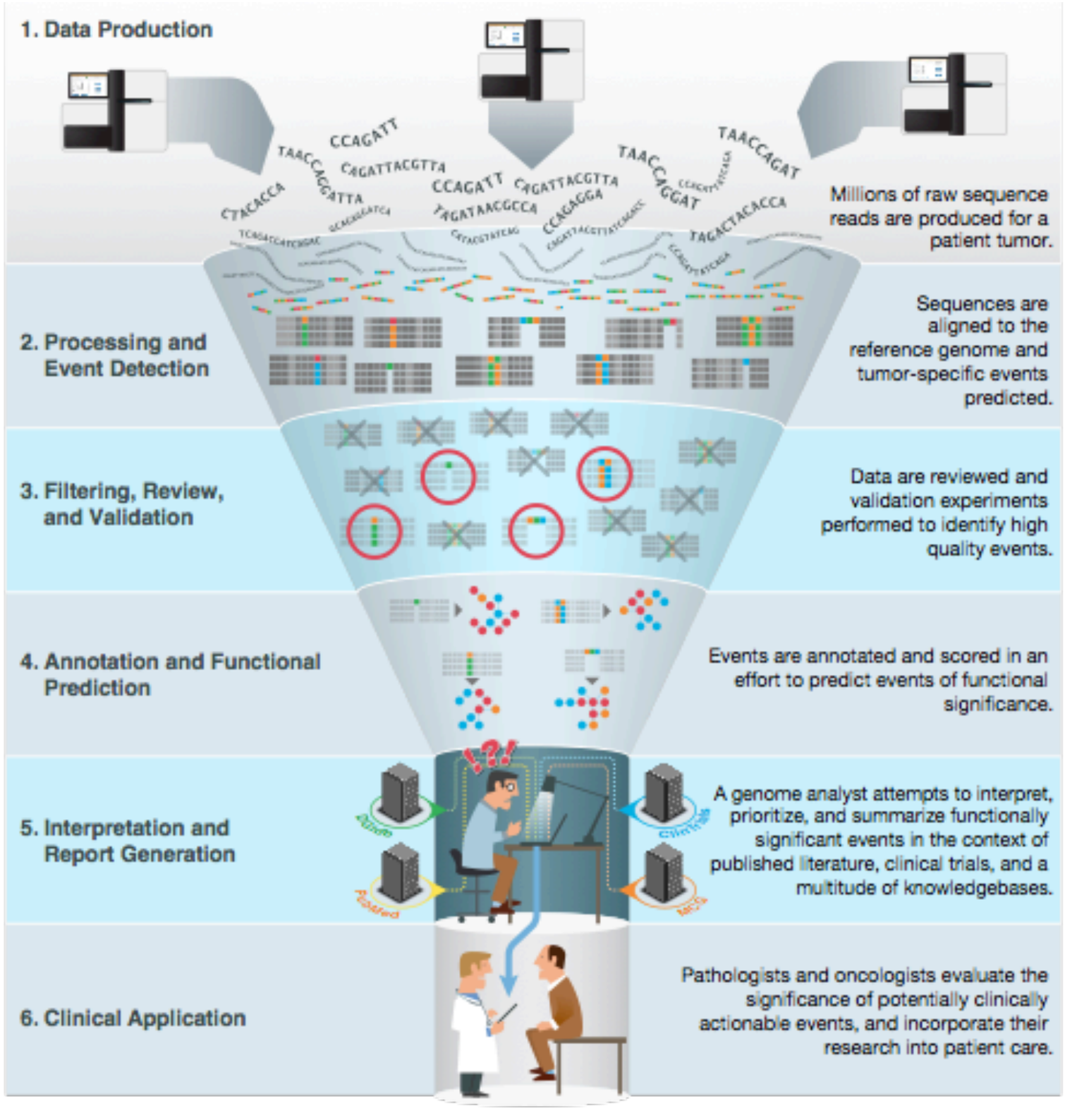
(1) Cost. Genome sequencing was too expensive for routine use until this year. And, although sequencing costs have decreased quickly, data analysis and storage costs have not.

- Infrastructure: a HiSeq X10 cluster can generate 1.5 Petabytes (1,500 Terabytes) of raw data per year. Data analysis requires a large compute cluster and another Petabyte of storage.
- People: skilled data analysts are not cheap, or easy to find.

(2) Mapping genome variation is harder than you might think.

- False positives caused by alignment errors and reference genome effects.
- False negatives (missed variants) in difficult-to-sequence regions.
- INDELs and structural variants are especially prone to both types of error.

Interpretation of clinical relevance of genomic events represents a serious bottleneck - manual labor!



Obi Griffith and Malachi Griffith <ogriffit@genome.wustl.edu and mgriffit@genome.wustl.edu>

Exome sequencing summary

STRENGTHS:

- Cheap relative to genome sequencing (but less so each year).
- Relatively deep coverage (100X) is possible at relatively low cost ($350), enabling detection of low frequency mutations in tumor samples.

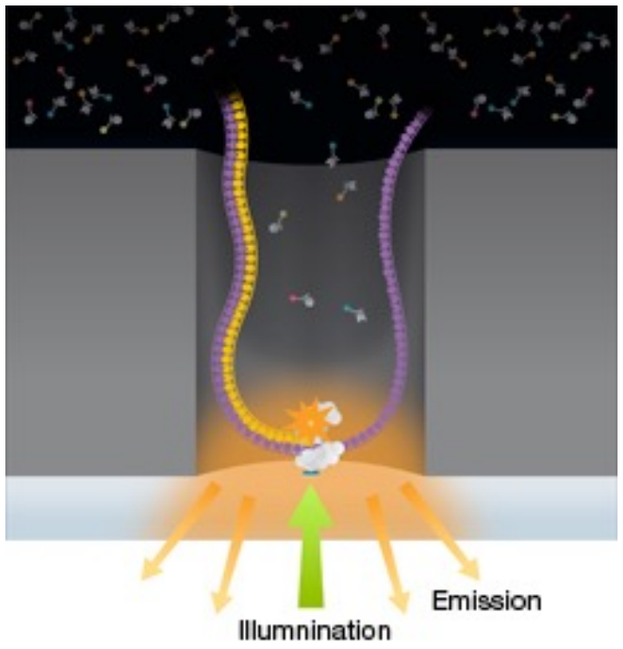
WEAKNESSES:

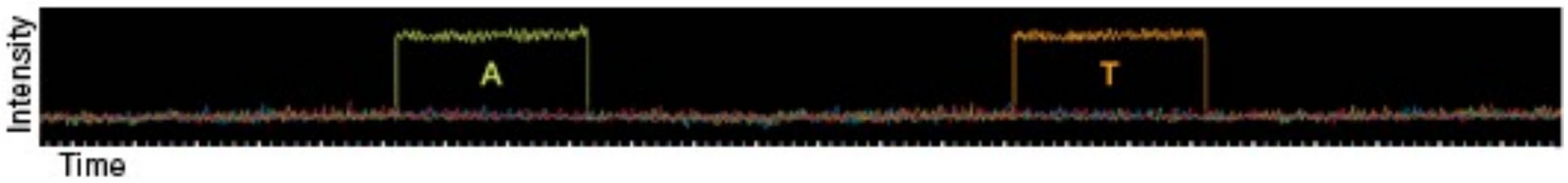
- Only assays a small portion of the genome (<2%); much can be missed (e.g., structural variation breakpoints rarely lie within exons).
- Capture techniques result in non-uniform sequence coverage.
- Non-uniform coverage results in variable power to detect mutations at different captured regions. This complicates statistical analyses.
- Non-uniform coverage makes it very difficult to detect copy number variants (CNVs) via read-depth analysis.

The Future: third generation DNA sequencing

Defining Characteristics: long reads (10-100 kb) from single molecules

Pacific Biosciences: watching a polymerase synthesize DNA in real time





<u>The promise:</u> Long reads will allow us to accurately sequence and assemble whole human genomes, from scratch, without using the reference genome.

<u>Status</u>: Currently can achieve long reads (10-30 kb), but utility is limited by low throughput, high error rate (10-15%) & high cost (> $100,000 per human genome). Not widely used in human genetics.

One ongoing application is high accuracy whole genome assembly, to improve reference genome resources (e.g., MGI Reference Improvement Project)