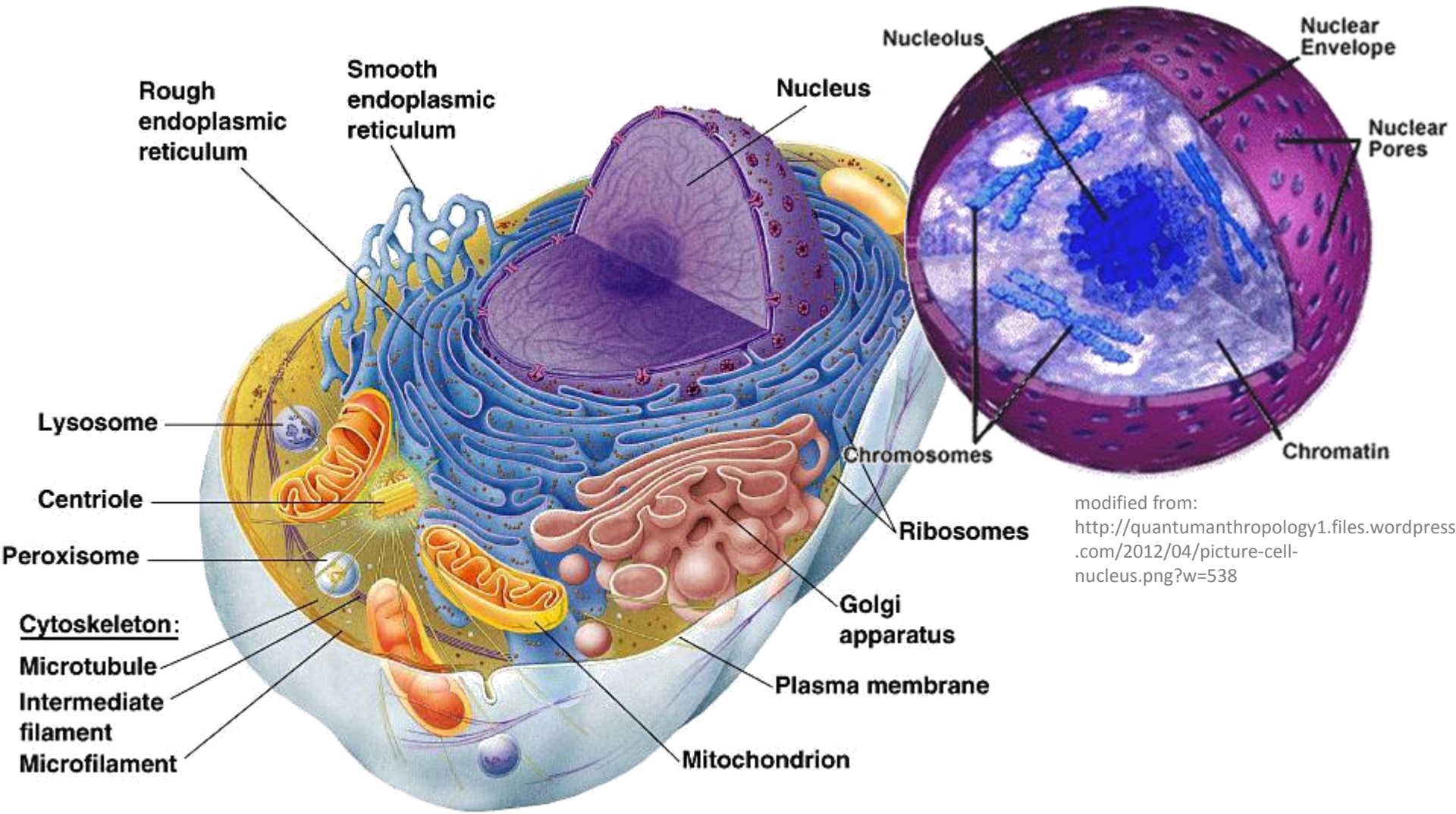
Introduction to Next-Generation Sequencing (NGS)

CGEpi Winter school 2017

13.02.2017

Aarif Mohamed Nazeer Batcha & Guokun Zhang

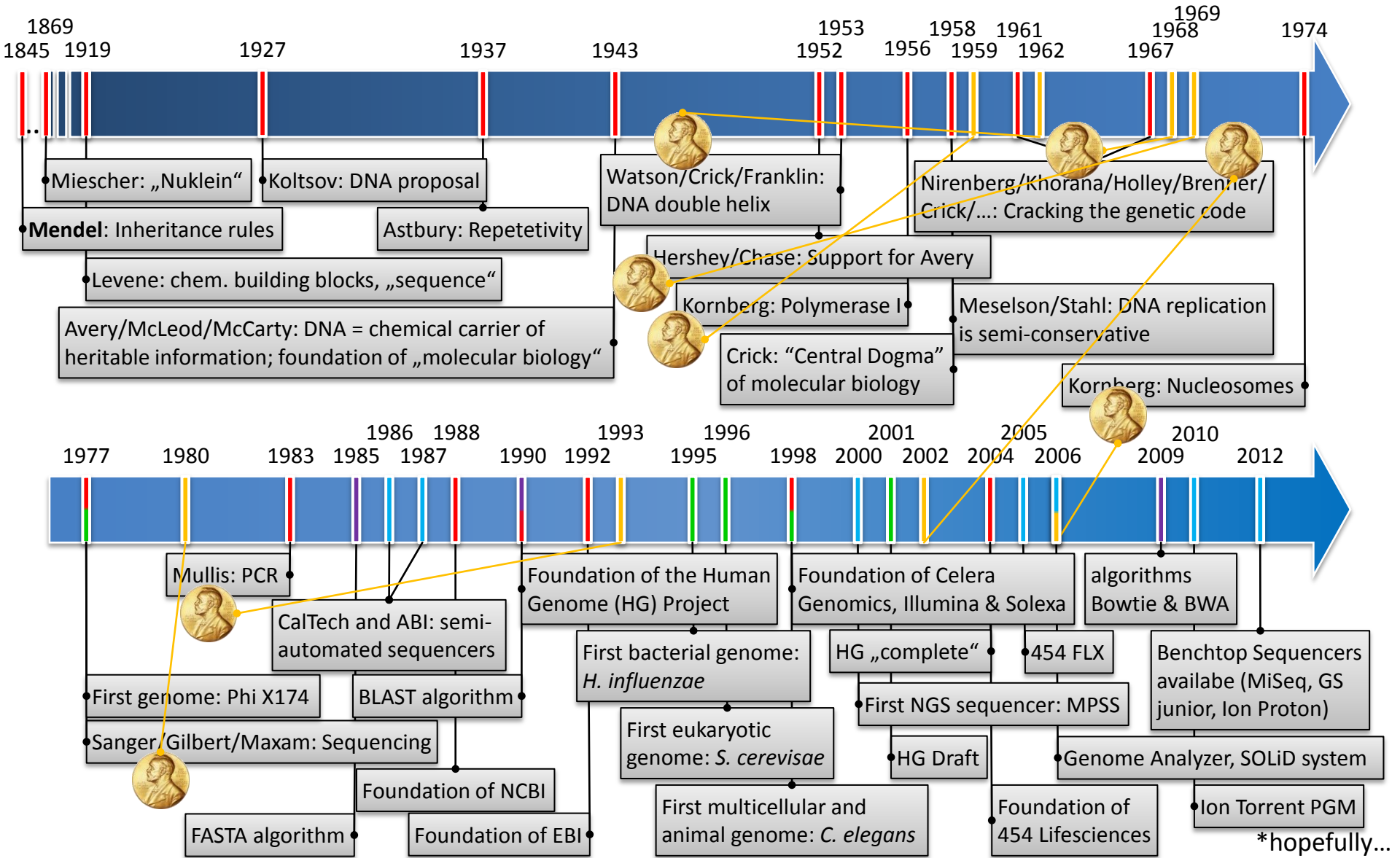
Biology: Cells, Genes and Molecules



modified from: http://quantumanthropology1.files.wordpress.com/2012/04/picture-cell-nucleus.png?w=538

from: http://www.biologyjunction.com/images/04-05A-AnimalCell-L.jpg

From Mendel to nowadays: in short* the history of molecular genetics

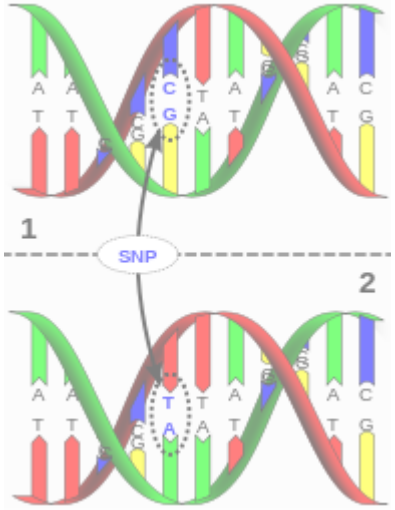


Wanted: DNA variations

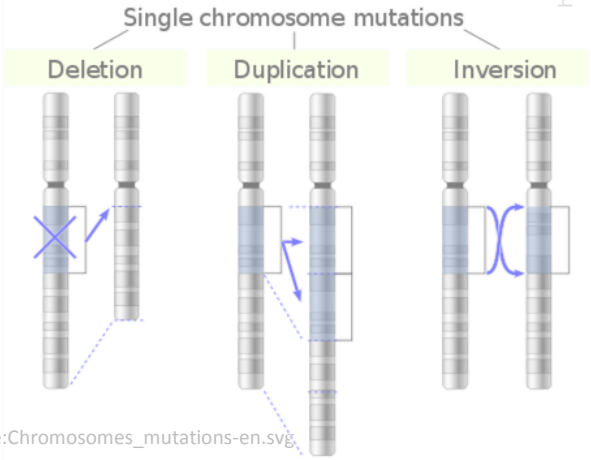
A difference compared to a defined ‚wildtype' reference

Nomenclature(s) - by effect on:

- structure
  - small scale: single nucleotide polymorphisms (SNPs) up to some hundred basepairs; substitutions, insertions, deletions, …)
  - large scale: chromosome level; amplifications, deletions, translocations, gene fusions, inversions, loss of heterozygosity, …
- protein sequence: frameshift, nonsense, missense, neutral, silent, …
- function: loss/gain, dominant negative, lethal
- fitness: harmful/beneficial, (nearly) neutral
- … (http://en.wikipedia.org/wiki/Mutation)



http://commons.wikimedia.org/wiki/File:Dna-SNP.svg



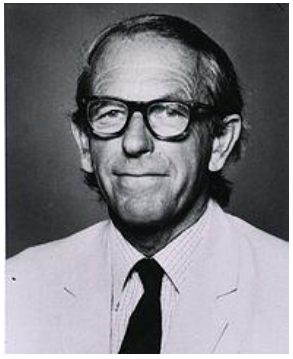
http://en.wikipedia.org/wiki/File:Chromosomes_mutations-en.svg

→ We need a reference sequence

(and samples…)

„Next-Generation“ Sequencing?

- Sanger technique published in 1977, based on the polymerase chain reaction (PCR)
  - → high quality, long reads
  - → slow, expensive
  - ➔ „First Generation“

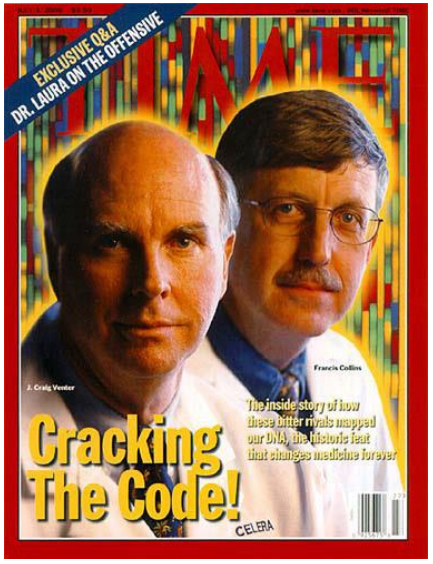


http://upload.wikimedia.org/wikipedia/commons/thumb/f/f8/Frederick_Sanger2.jpg/195px-Frederick_Sanger2.jpg

http://www.lifesciencesfoundation.org/content/media/2011/08/17/2001_genome_covers.png

- Human Genome Project: 1990 – 2004
- Investments >3,000,000,000 $ (~1$/bp)

- Development of cheaper „NGS“ technologies
- Compete with microarray technology*



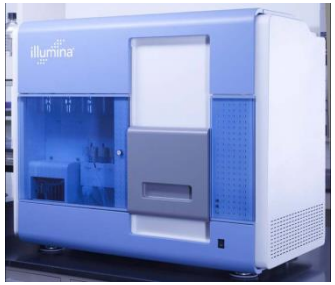
http://singularityhub.com/wp-content/uploads/2010/05/Venter_Collins_Genome.jpg

* http://en.wikipedia.org/wiki/DNA_microarray

„Next-Generation“ Sequencing!

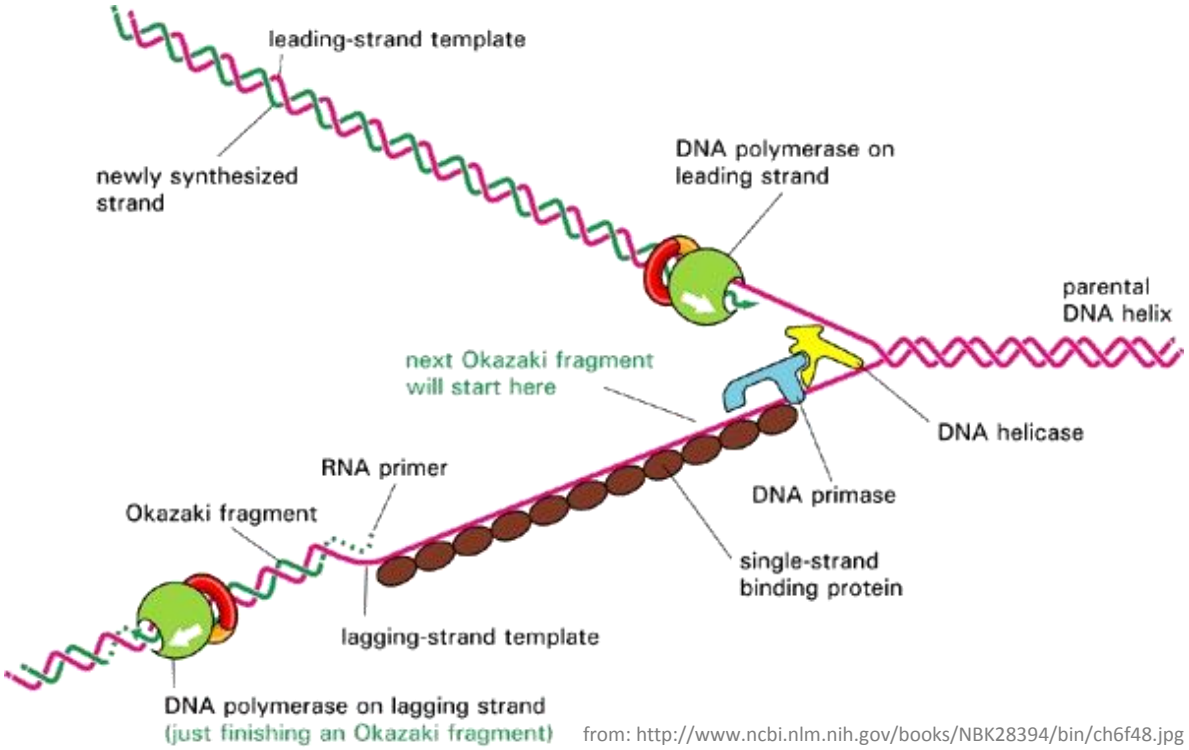
NGS depends on massive, molecular parallelization

- Four major platform types (and further ones):
  - Solexa/Illumina (Genome Analyzer, MiSeq, HiSeq)
  - Applied Biosystems (ABi; SOLiD)
  - Roche/454 (GS FLEX, GS Junior)
  - Ion (Torrent PGM, Proton)
- Advantages: fast (massively parallel) & cheap (low cost per bp)
- Disadvantages: error-prone, extensive needs in computation time and storage

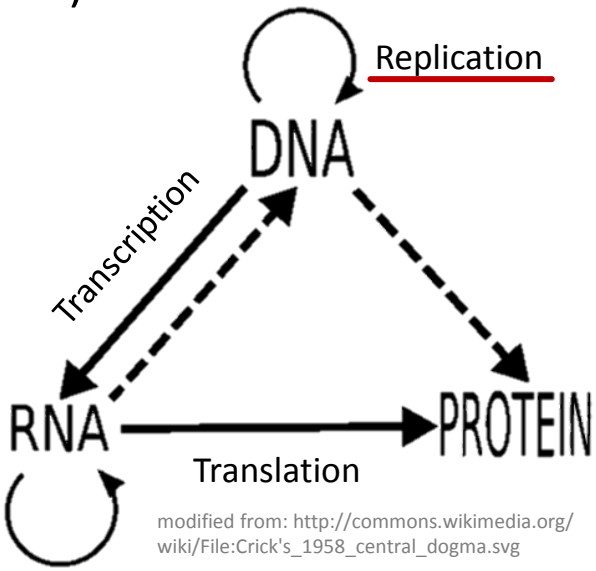


Principles of Sequencing (I)

- Central „Dogma“ of molecular biology (Crick 1958/70)
- DNA polymerases (Kornberg 1956 and others)
  - replicate DNA (template + nucleotides + energy)
  - act semi-conservative (Meselson & Stahl 1958)
  - work also *in vitro*



from: http://www.ncbi.nlm.nih.gov/books/NBK28394/bin/ch6f48.jpg

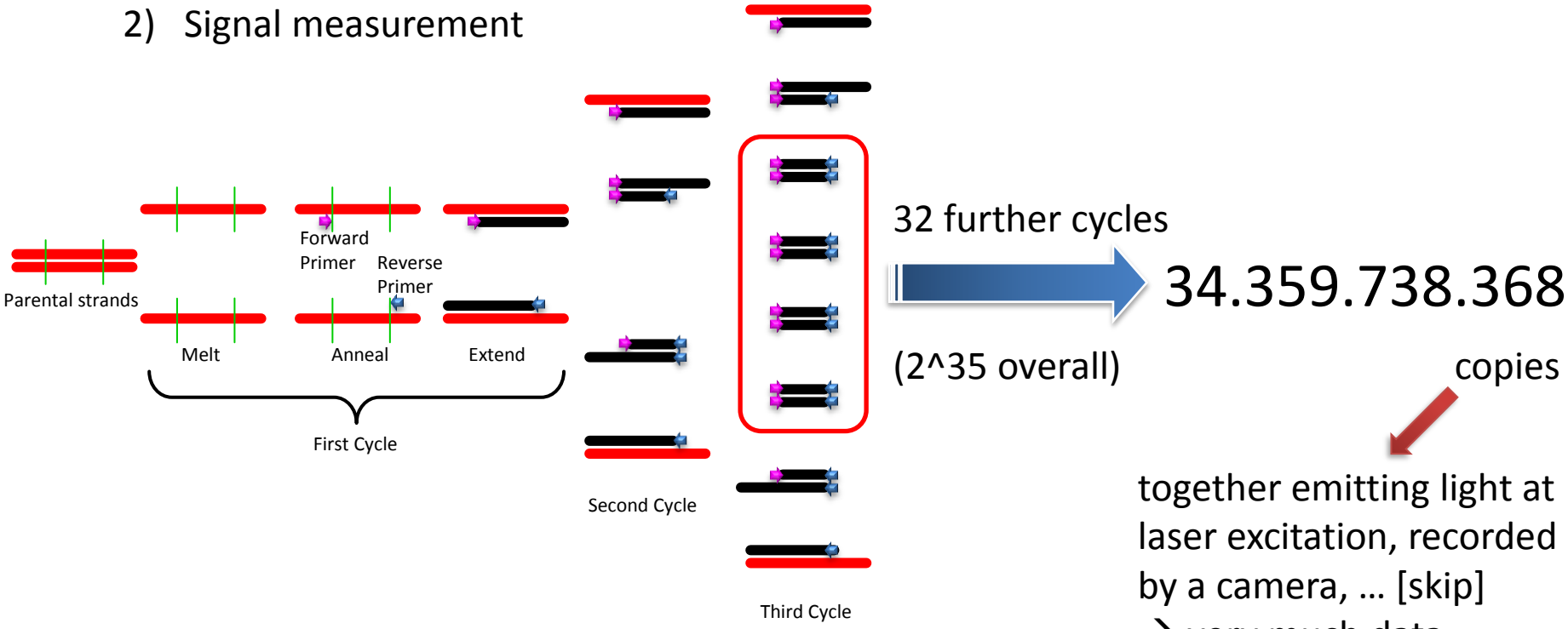


modified from: http://commons.wikimedia.org/wiki/File:Crick's_1958_central_dogma.svg

record what a polymerase is doing in order to get the template sequence

Principles of Sequencing (II)

- How to „observe“? E.g. measure light emission during a biochemical reaction, tracing the underlying base sequence
- Requires a pre-step, due to limited signal strength:
  1) DNA amplification
  2) Signal measurement



made from: http://www.summerschool.at/static/andreas.zoller/images/cycle1.jpg

a bit of… Theory

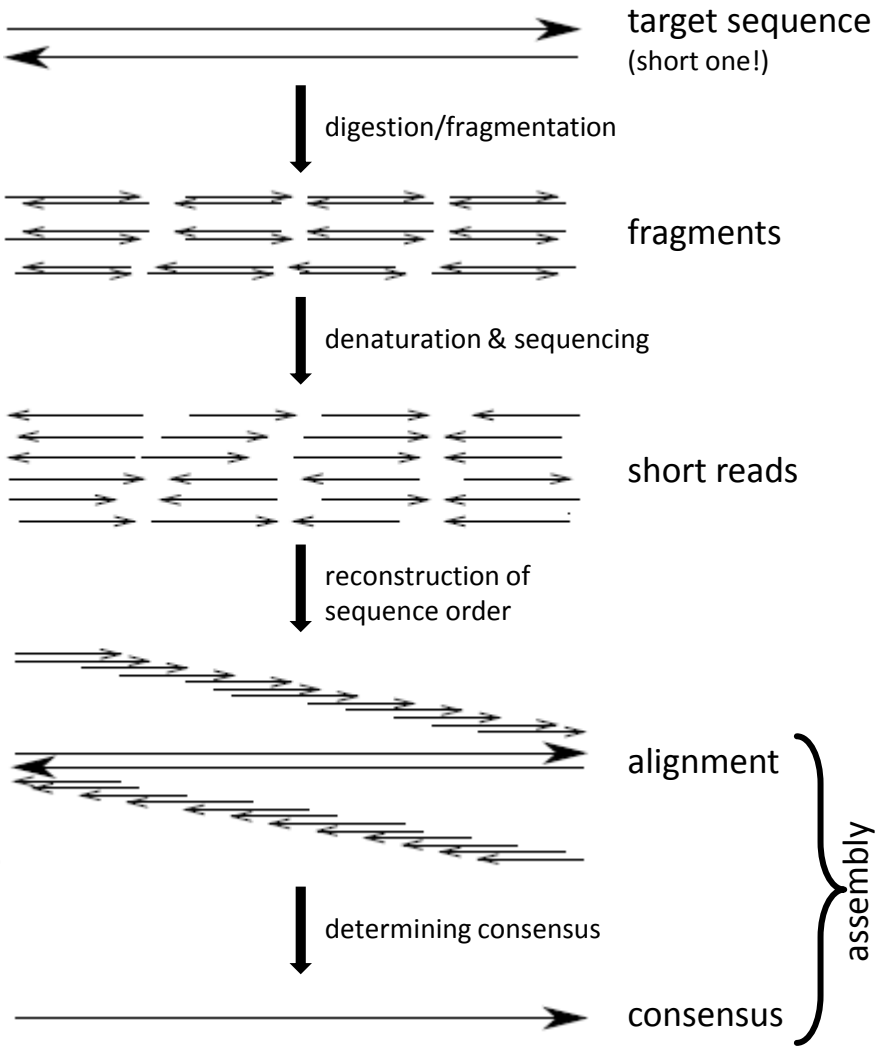
- per-base quality (sequencing errors)
- short length of reads:
  - more gaps
  - more comparisons
  - more uncertain positions
  - may not bridge repetetive regions
  - …
- coverage problem in genomics (Lander-Waterman 1988):

  How much read data must be generated in order to ensure a certain statistical correctness of the consensus sequence?
  - Poisson distribution of residing gaps
  - „Coverage“ (e.g. 10x) applies to read length

→ NGS technology requires a multiple of data to be generated compared to the original sequence

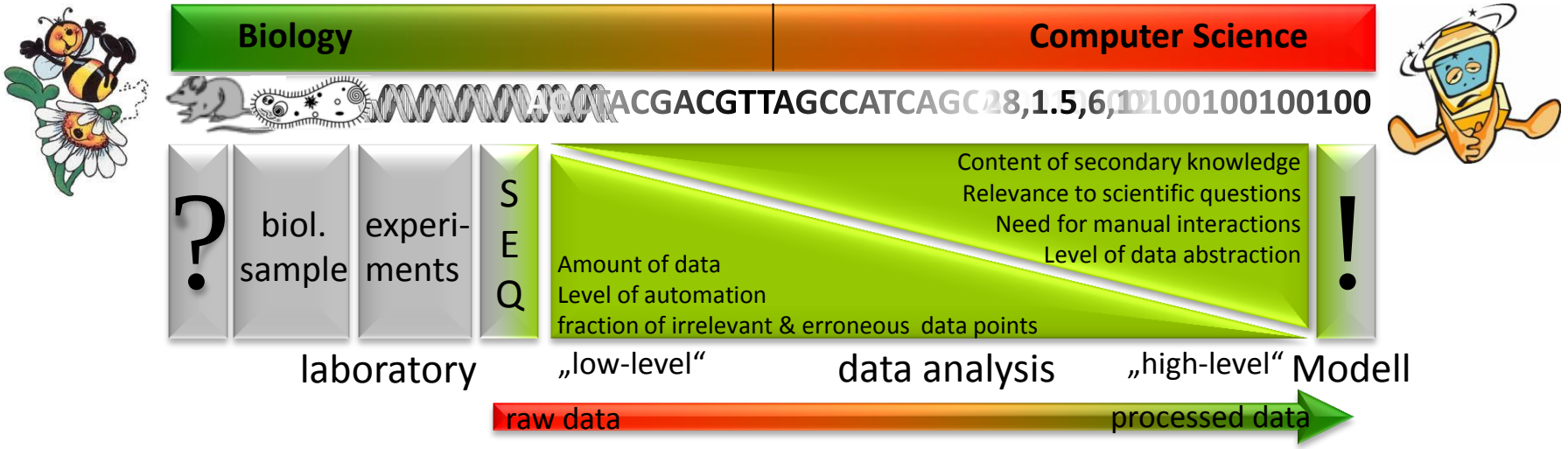
→ but: massive cost reduction (per base)

➔efforts shift from lab to computing



modified from: S. Kurtz, ZBH Hamburg (lecture winter semester 2008)

Bridging the gap



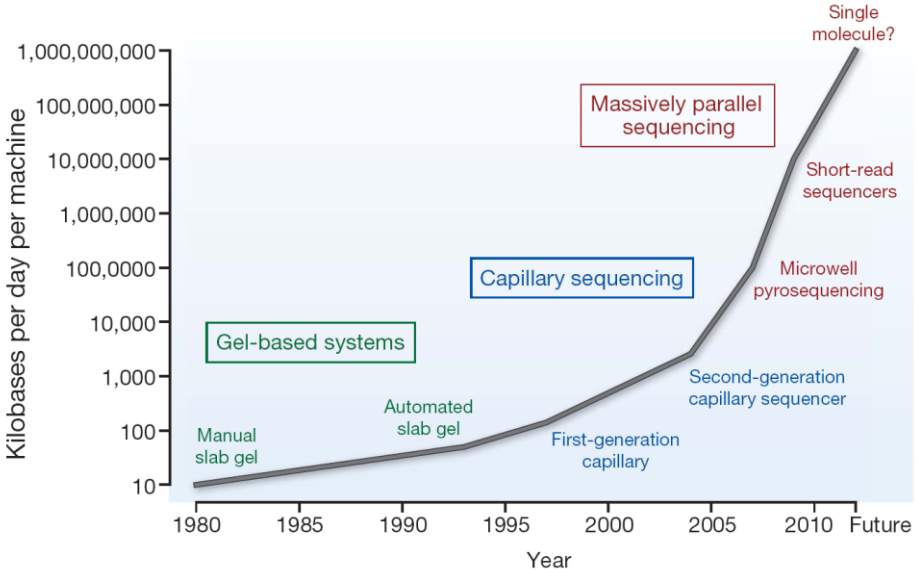
- Computationally intense: often quadratic or even exponential growth of requirements in time and space

  → e.g. double input data, quadruple load

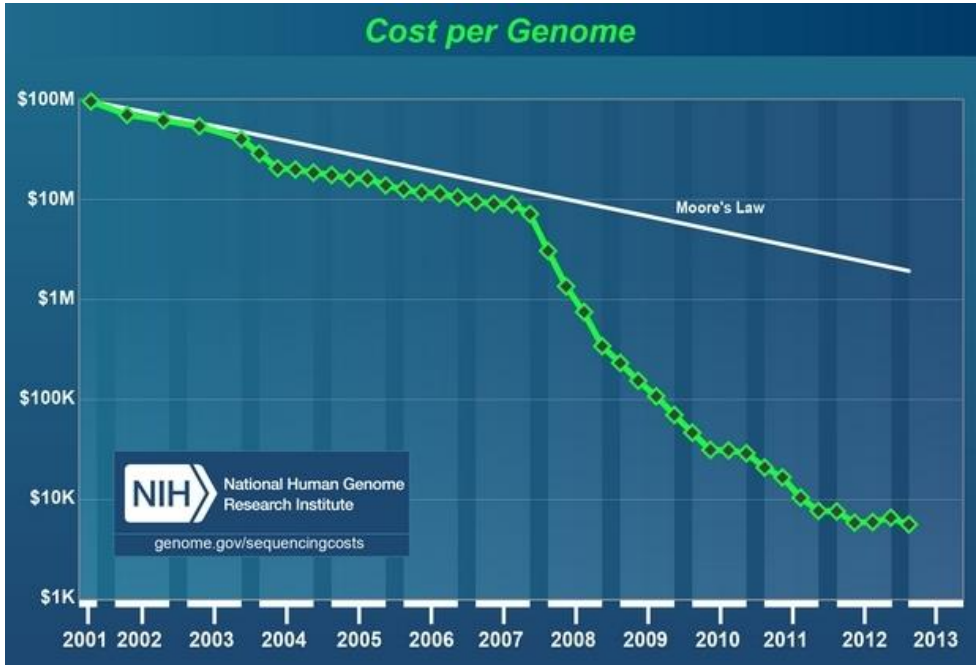
Aim: data reduction and gain of knowledge

Busting storage media

- Current technologies (computing power, storage capacities, …) get overwhelmed by NGS data
- Sequence data generation beats even Moore‘s Law (which postulates exponential growth for hardware)



from: Stratton *et al.* 2009 – The cancer genome

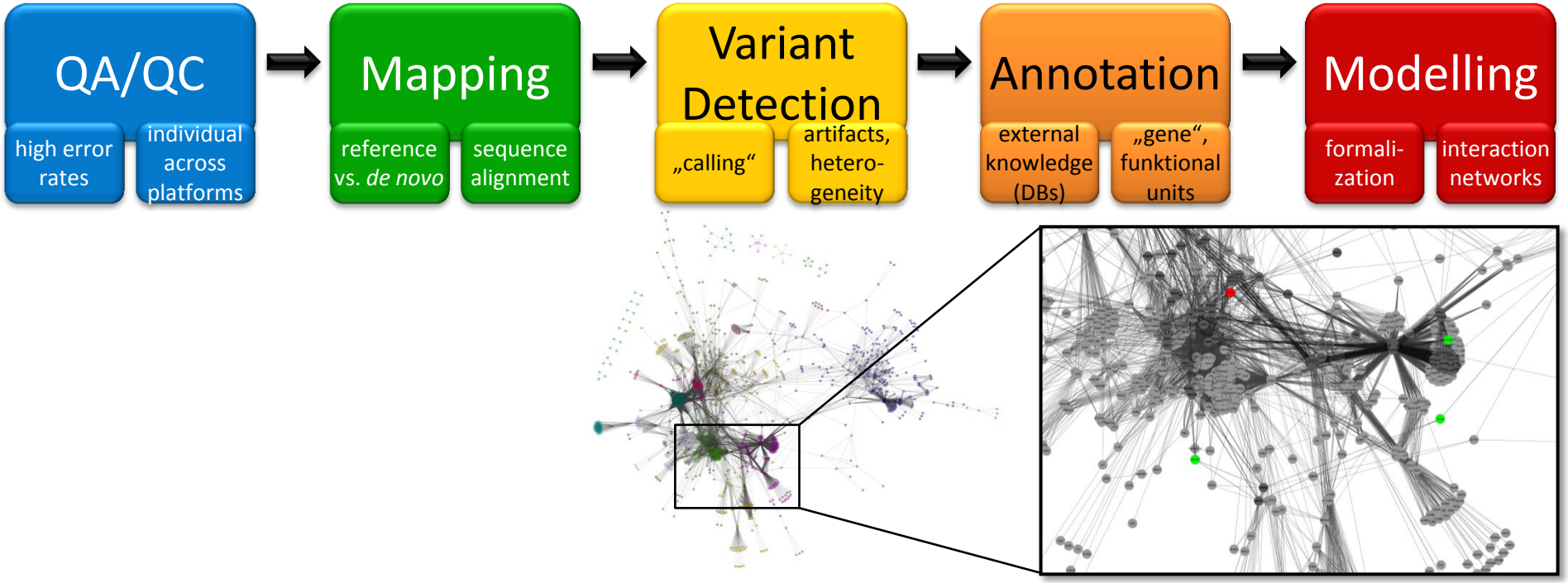


http://genomics.xprize.org/sites/default/files/styles/panopoly_image_original/public/nih-cost-genome.jpg?itok=QPzSRoVY

But: why?

Practice: Using sequence data (I)

- NGS is just a technology generating data
- Scientists need assays in order to get from questions to answers
- Great variety of problems, scientific fields, target molecules, biological mechanism etc. determine the assay and the data analysis
- General scheme:



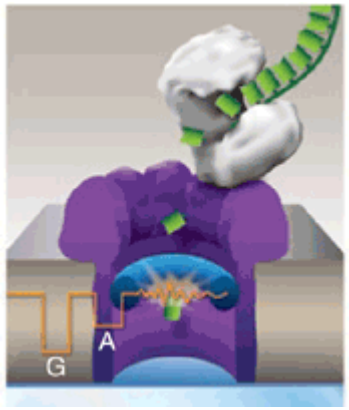
Practice: Using sequence data (II)

- NGS is just a technology generating data
- Scientists need assays in order to get from questions to answers
- Great variety of problems, scientific fields, target molecules, biological mechanism etc. to work on
- General scheme
- find differences („variants“) via comparison („alignments“)
- Examples for sequencing assays:
  - whole genome
  - exome („poor man‘s genome“), transcriptome
  - RNA expression quantification (analogous to microarrays)
  - epigenome („regulatome“)
  - amplicons (deep sequencing on few defined targets)
  - metagenome (genomic content of a biosphere; water, biofilm, gut flora, …)
  - …
- for what?
  - basic research
  - diagnostics (paternity tests, hereditary defects, tumor characterization, …)
  - gain in resolution in existing efforts

Outlook

- faster technologies are under development („third generation sequencing“)
  - Nanopore technology
  - Electron-microscope based
  - Via mass spectroscopy
  - …

  → single molecule approaches: skip the amplification step
- fraction of costs for analysis will rise further (Mardis 2010)
- competition: Archon X Prize (Kedes & Campany 2011; canceled in 2013)
- Ethics!



modified from: http://www.nature.com/nbt/journal/v28/n5/images/nbt0510-426-F1.gif

„Sequencing on a stick“:

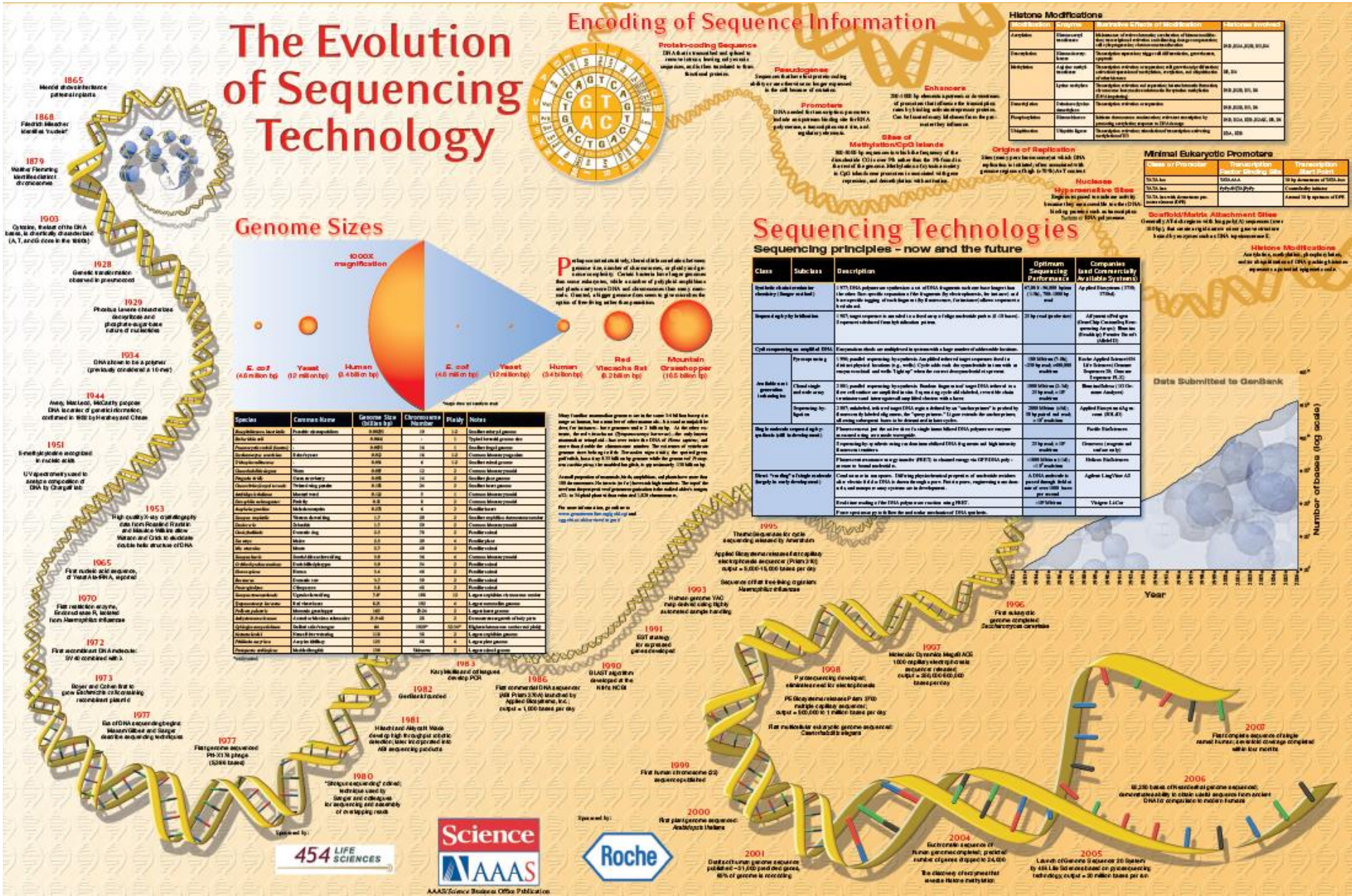
from: http://www.wired.com/wiredenterprise/wp-content/uploads//2012/03/Oxford-Nanopore-MinION.jpeg

“The $10 million grand prize will be awarded to the team(s) able to sequence **100 human genomes** within **30 days** to an accuracy of **1 error per 1,000,000 bases**, with **98% completeness**, identification of insertions, deletions and rearrangements, and a complete haplotype, at an audited total cost of **$1,000** per genome” (→ “outpaced by innovation“)

References

- **Metzker 2010** - Sequencing technologies - the next generation; Nat Rev Genet. 2010 Jan;11(1):31-46
- **Stratton *et al.* 2009** - The Cancer Genome; Nature. 2009 April 9; 458(7239): 719–724
- **Mardis 2010** - The $1,000 genome, the $100,000 analysis?; Genome Medicine 2010, 2:84
- **Liu *et al.* 2012** - Comparison of Next-Generation Sequencing Systems; J Biomed Biotechnol. 2012;2012:251364
- **Kedes & Campany 2011** - The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition; Nat Genet. 2011 Oct 27;43(11):1055-8
- Science Poster „The Evolution of Sequencing Technology“

For those who are curious…



Comparison of next-generation sequencing methods

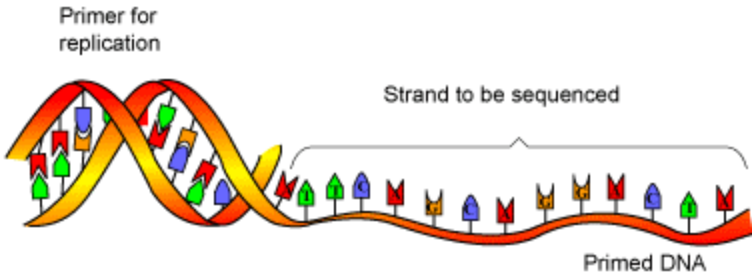
Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	2900 bp average[38]	200 bp	700 bp	50 to 250 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	87% (read length mode), 99% (accuracy mode)	98%	99.90%	98%	99.90%	99.90%
Reads per run	35–75 thousand [39]	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours [40]	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length[41]	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$2	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[42]	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Low yield at high accuracy. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.

Quail, Michael; Smith, Miriam E; Coupland, Paul; Otto, Thomas D; Harris, Simon R; Connor, Thomas R; Bertoni, Anna; Swerdlow, Harold P; Gu, Yong (1 January **2012**). "A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers". BMC Genomics 13 (1): 341

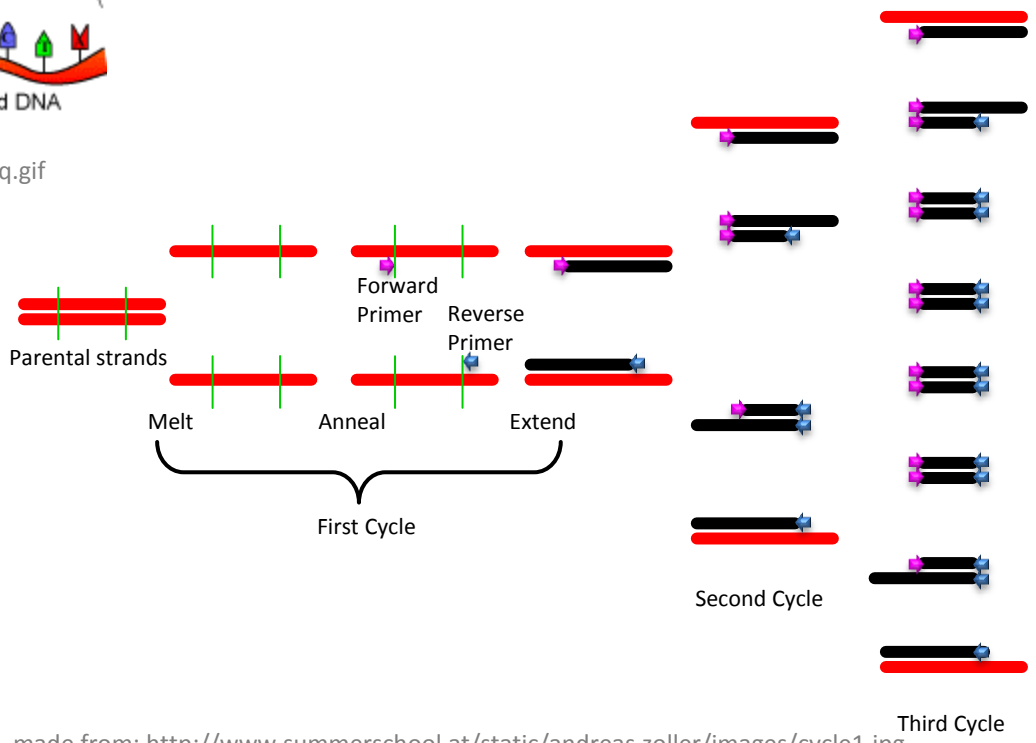
Liu, Lin; Li, Yinhu; Li, Siliang; Hu, Ni; He, Yimin; Pong, Ray; Lin, Danni; Lu, Lihua; Law, Maggie (1 January 2012). "Comparison of Next-Generation Sequencing Systems". Journal of Biomedicine and Biotechnology **2012**: 1–11

Sanger Method (I)

- Sequence of interest is targeted via designed primers
- Massive amplification of target (e.g. by ca. 35 rounds of PCR amplification)



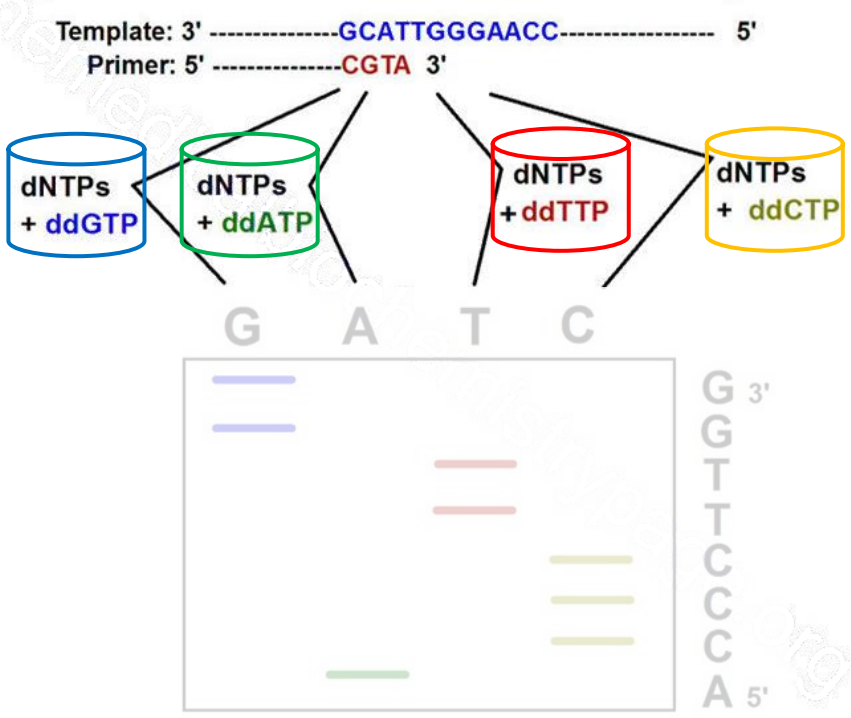
modified from: http://www.eisenlab.org/FunFly/wp-content/uploads/2012/06/science-creative-quarterly-seq.gif



made from: http://www.summerschool.at/static/andreas.zoller/images/cycle1.jpg

Sanger Method (II)

- Elongation of DNA sequences by polymerase
- Enzyme stops at a random position per copy (by **d**dNTP)
- Terminated copies are separated within a gel (smaller ones run further)
- Sequence can be read directly



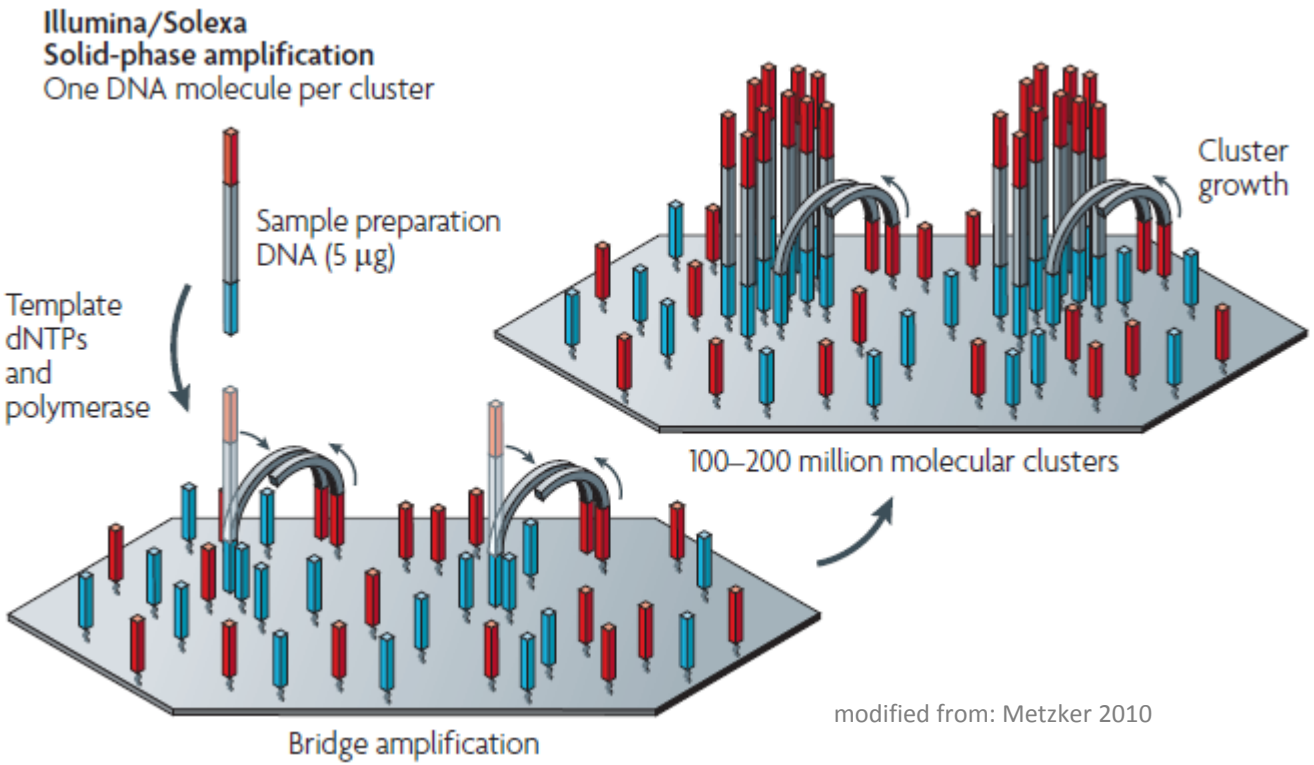
→ Extremely accurate: „gold standard“ (error rate ~1:100,000)

→ Slow: poorly parallelizable (60x max.)

modified from: http://themedicalbiochemistrypage.org/images/sangersequencing.jpg

Illumina Technology (I)

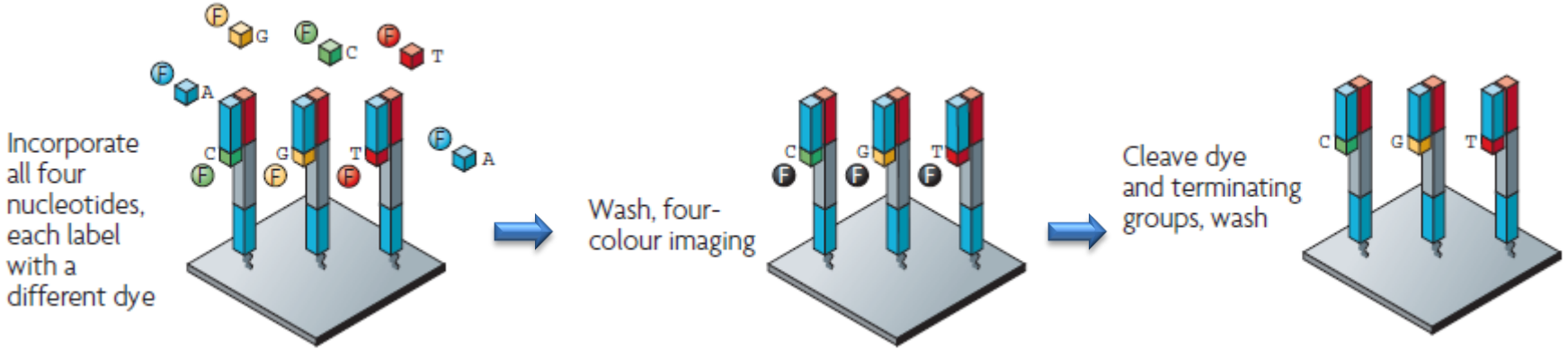
- Ca. 80 % of the NGS market
- Preparation: solid-phase bridge amplification (one cluster = one target sequence)



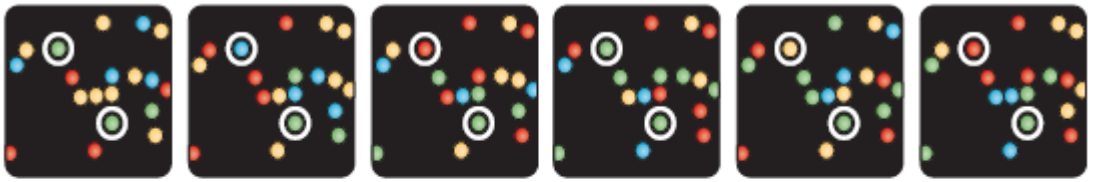
modified from: Metzker 2010

Illumina Technology (II)

- Ca. 80 % of the NGS market
- Preparation: solid-phase bridge amplification (one cluster = one target sequence)
- Sequencing: Reversible terminators, fluorescent dyes



modified from: Metzker 2010



Top: CATCGT
Bottom: CCCCCC

→ Extremely fast: speedup by massive parallelization on the moleculare level (~100 Mio. x)

→ Inaccurate: poor per-base quality (error rate ~1:500)