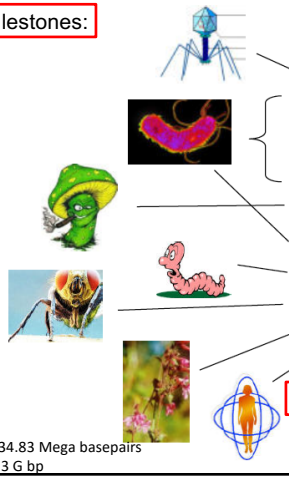
Molecular and Cellular Biology

5. Basic Molecular Genetic Techniques

Prof. Dr. Klaus Heese

Sequencing Methods & Applications

Milestones:



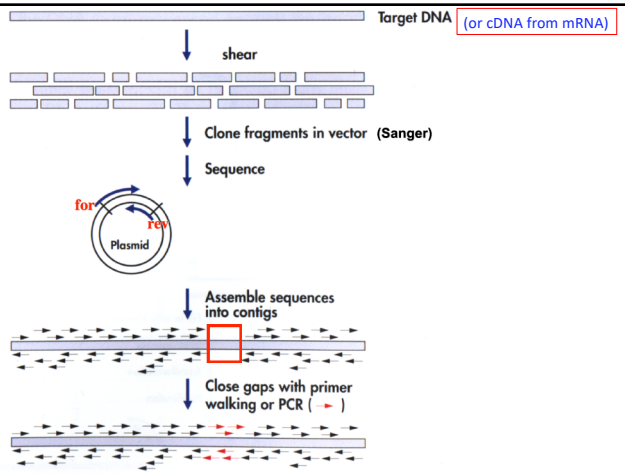
Phi X 174	1977	5.386 bp
λ– Phage	1982	48.502 bp
M. genitalium	1995	580.000 bp
H. influenzae	1995	1.830.000 bp
M. jannaschii	1996	1.660.000 bp
S. cerevisae	1997	12.500.000 bp
E. coli	1997	4.654.000 bp
C. elegans	1998	97.000.000 bp
D. melanog.	1999	116.000.000 bp
A. thaliana	2000	115.000.000 bp
H. sapiens	2001	2.693.000.000 bp*
bread wheat	2018	17 G bp !!

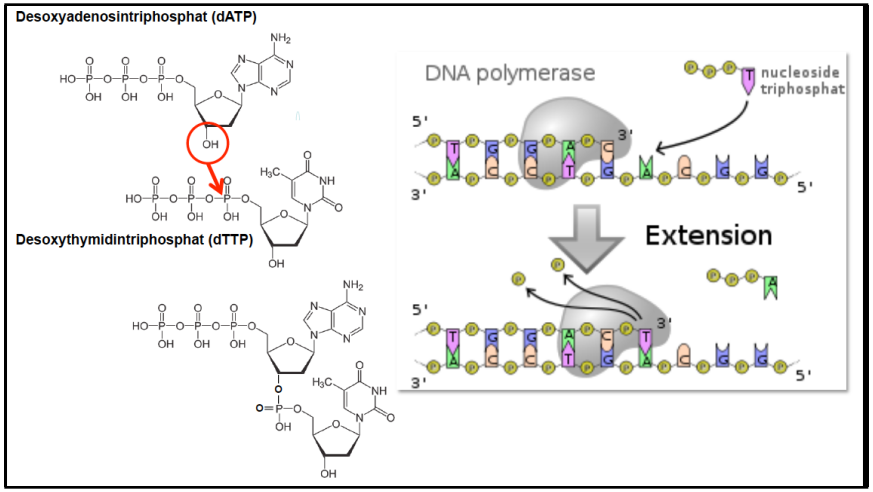
2012 > 1000genomes project (human)

soon > Genome 10K (vertebrates)

* 3,234.83 Mega basepairs
~ 3.3 G bp

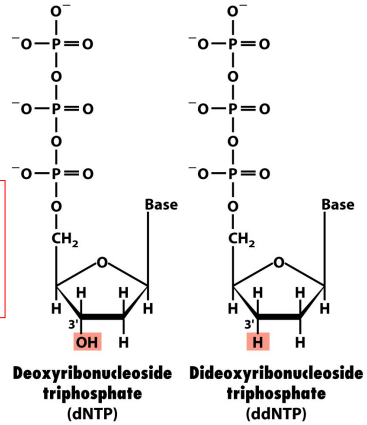
classical 'Sanger' sequencing method:

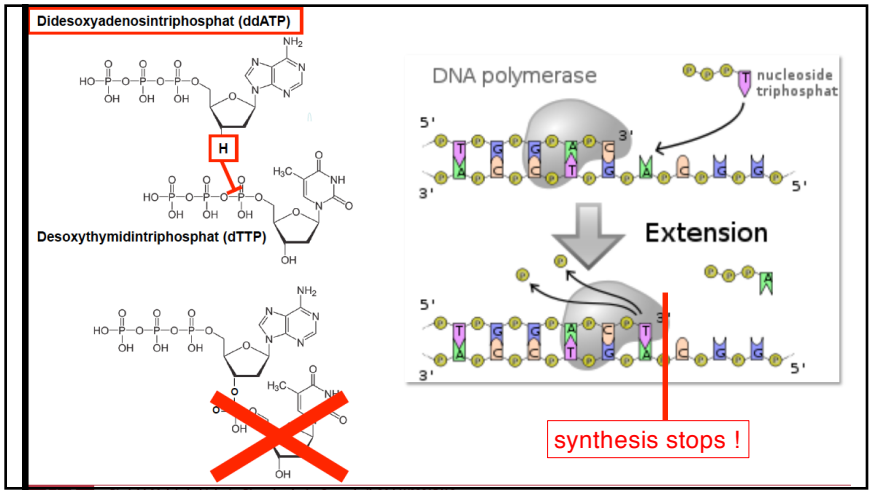




Sequencing of DNA

characterization of DNA quality by specific gene (pDNA + gene of interest) sequencing.

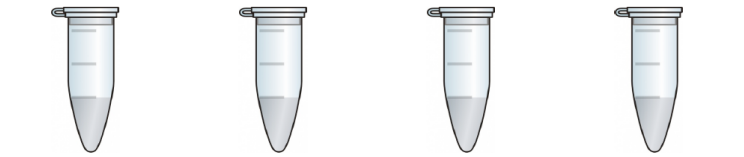




DNA sequencing is based on enzyme activity of DNA polymerase:

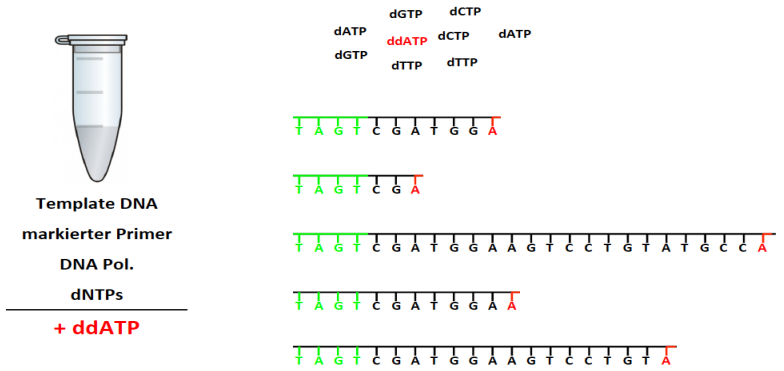
How does DNA sequencing work?

4 reactions take place:



Template DNA markierter Primer DNA Pol. dNTPs	Template DNA markierter Primer DNA Pol. dNTPs	Template DNA markierter Primer DNA Pol. dNTPs	Template DNA markierter Primer DNA Pol. dNTPs
+ ddATP	+ ddGTP	+ ddTTP	+ ddCTP

DNA sequencing is based on enzyme activity of DNA polymerase:
4 reactions take place:



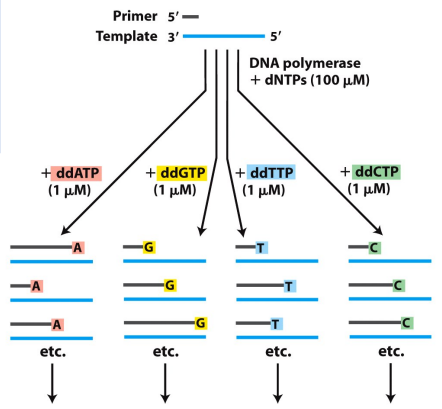
DNA sequencing is based on enzyme activity of DNA polymerase:
4 reactions take place:

5′ T A G C T G A C T C 3′
3′ A T C G A C T G A G T C A A G A A C T A T T G G G C T T A A ...

DNA polymerase
+ dATP, dGTP, dCTP, dTTP
+ ddGTP in low concentration

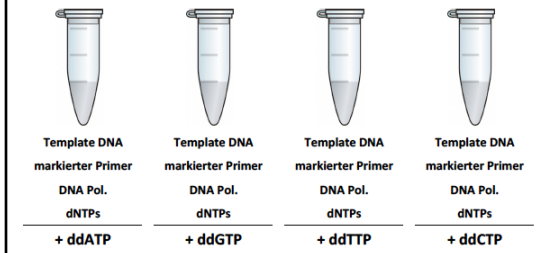
5′ T A G C T G A C T C A G 3′
3′ A T C G A C T G A G T C A A G A A C T A T T G G G C T T A A ...
+
5′ T A G C T G A C T C A G T T C T T G 3′
3′ A T C G A C T G A G T C A A G A A C T A T T G G G C T T A A ...
+
5′ T A G C T G A C T C A G T T C T T G A T A A C C C G 3′
3′ A T C G A C T G A G T C A A G A A C T A T T G G G C T T A A ...

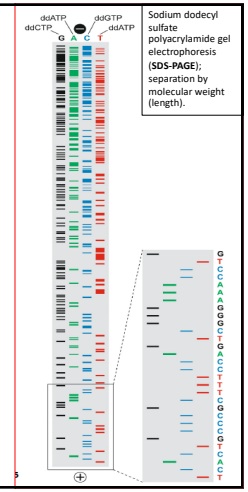
DNA sequencing is based on enzyme activity of DNA polymerase:
4 reactions take place:



DNA sequencing is based on enzyme activity of DNA polymerase:

4 reactions take place:





DNA sequencing process

① A mixture of DNA go into sequencing machine.

② Move past a laser that causes the dye molecules to fluoresce.

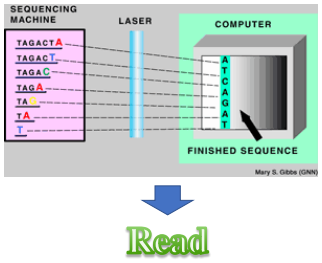
③ A detector reads the color of the fluorescence.

1: blue-T, blue-TAGACT

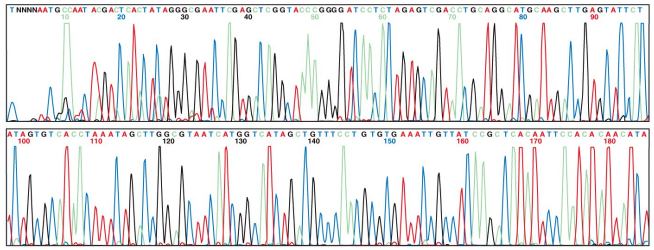
2: red-TA, red-TAGA

3: yellow-TAG

4: green-TAGAC



DNA sequencing is based on enzyme activity of DNA polymerase:
4 reactions take place:



Separation method is similar to SDS-PAGE (see slides further below)

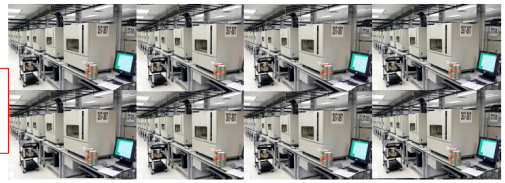
1. DNA polymerases reading about 80 bp per reaction; today about >> 1,000 bp

2. Improved detection

40 probes/samples ---> 10 sequences

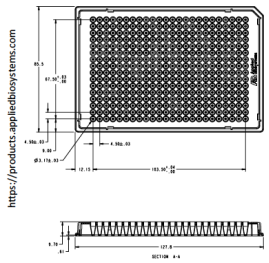
384 probes/samples ---> 384 sequences

3. Automation
digitization / digitalization
massive Investments



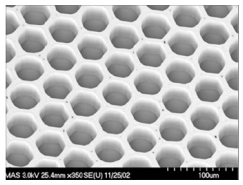
since about 2005 development to massive parallel sequencing, '454 pyrosequencing', 'Illumina sequencing', …

Sanger:

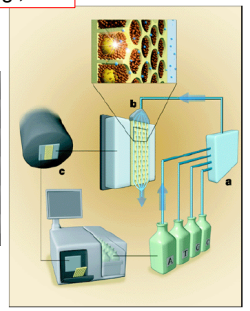


- microtiter plates
- 384 wells
- ca. 300 000 bases per run

pyrosequencing:



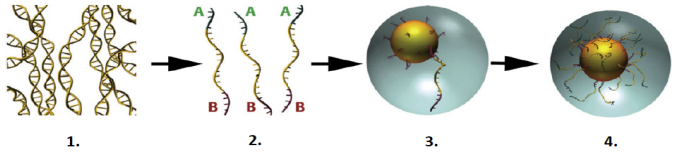
http://www.rocheinstitute.com/



- microtiter plates made by polydimethylsiloxane
- 1 000 000 wells
- improved enzymes and detection systems
- ca. 400 000 000 bases per run

How does Next Generation Sequencing (NGS) work ?

aus: Tucker et al., The American Journal of Human Genetics 2009

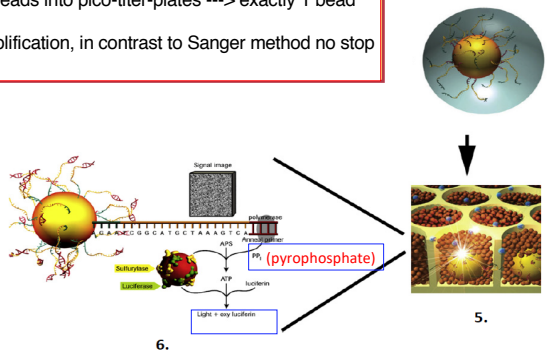


- Isolation of DNA, make fragments
- Ligation of adaptor linkers
- Ligation of 1 fragment onto 1 carrier-bead
- Clonal amplification by emulsion-PCR ---> many copies of the same DNA molecule attached to the 1 carrier-bead

How does Next Generation Sequencing (NGS) work ?

from: Tucker et al., The American Journal of Human Genetics 2009

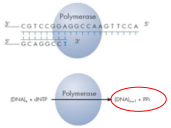
- Distribution of the beads into pico-titer-plates ---> exactly 1 bead per well
- Sequencing by amplification, in contrast to Sanger method no stop by ddNTPs



How does Next Generation Sequencing (NGS) work ?

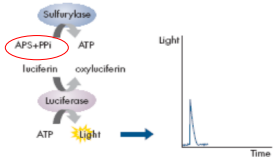
Chemistry of this sequencing reaction method

1. Klenow-Fragment



$DNA_n + dNTP \rightarrow DNA_{n+1} + PP_i$

2. ATP-Sulphurylase + Luciferase



Adenosinphosphosulphat + PP → ATP

Luciferin + ATP → Oxyluciferin + light

3. Apyrase

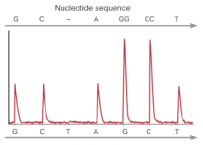
dNTP —Apyrase→ dNDP + dNMP + phosphate

ATP —Apyrase→ ADP + AMP + phosphate

abreaction of unused dNTP (and ATP)

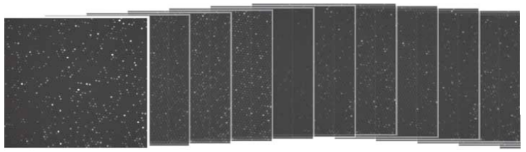
How does Next Generation Sequencing (NGS) work ?

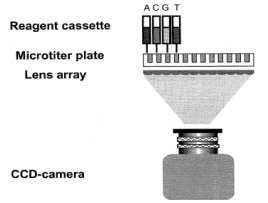
- Intensity of emitted light is proportional to the amount of ATP = incorporated NTPs
- After each cycle fresh dNTP is added
- Alternating adding dATP, dTTP, dCTP and dGTP



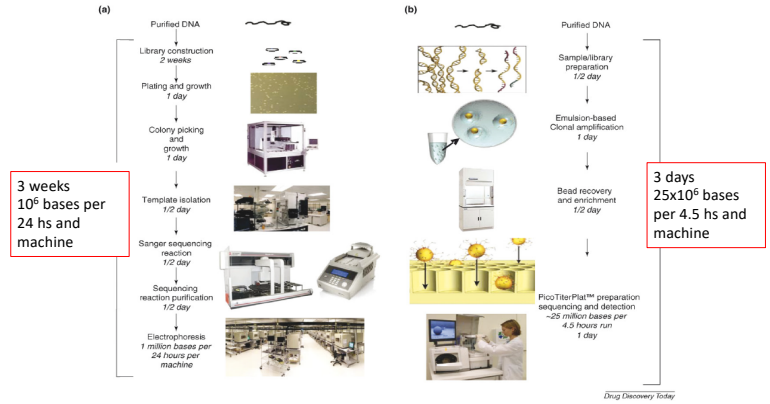
Detection

CCD camera, each well is captured separately / individually





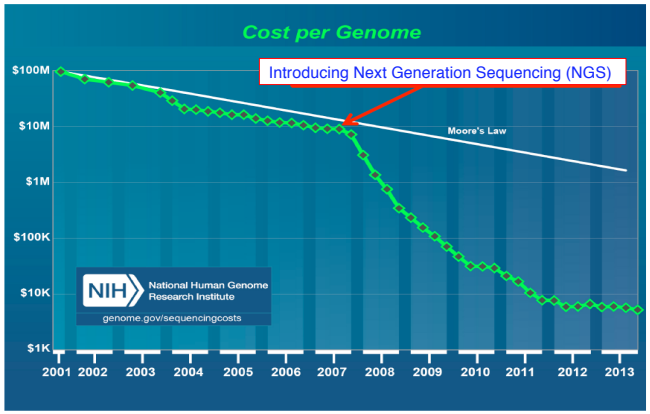
Comparison of conventional high-throughput sequencing with Next Generation Sequencing (NGS)



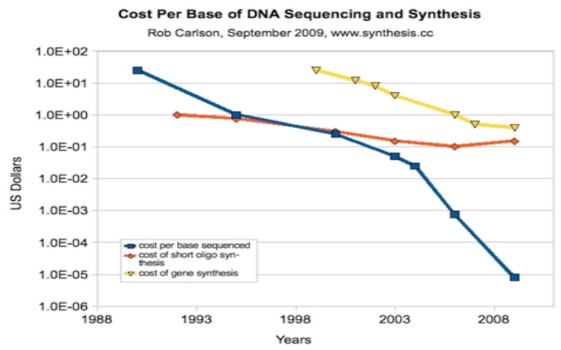
3 weeks
10^6 bases per 24 hs and machine

3 days
$25x10^6$ bases per 4.5 hs and machine

Comparison of conventional high-throughput sequencing with Next Generation Sequencing (NGS)



Comparison of conventional high-throughput sequencing with Next Generation Sequencing (NGS)



2003:

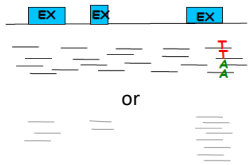
first genome:
several Mrd $

2009:

ca. 10$ per 10^6 bases
= 33 000 $ per genome

Genome Sequencing: **Re**ference- or ***de novo*** method

Re:



reference genome
(from database)

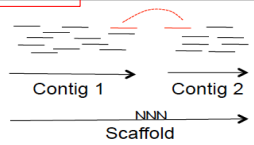
genomic DNA reads

cDNA reads

Exomes, RNA (transcripts: mRNA-->cDNA), …

- **variant detection**
- differential gene expression (RNA-Seq)

de novo:



genome draft

Chromatin (ChIP), methylations (Methyl-Seq, Bisulfite-Seq), …

- **de novo genome**
- **gene discovery**
- **genome evolution**

Re- vs. de novo-sequencing

...require different bioinformatics algorithms

- Re-Seq

> short reads ok
> very high redundancy (SNP detection!)
> no assembly, but **alignment to existing reference sequence („*Mapping*“)**
> possible on normal PCs (64bit, quadcore, 16 GB RAM)

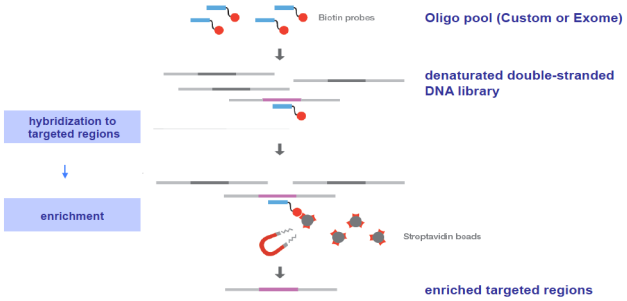
- de novo

> longer reads better (due to repeats)
> classic **assembly** (de Bruijn graph)
> extremely RAM intensive (512 GB RAM)

Martin JA, Wang Z, Nature Reviews Genetics 2011

Exome-Seq

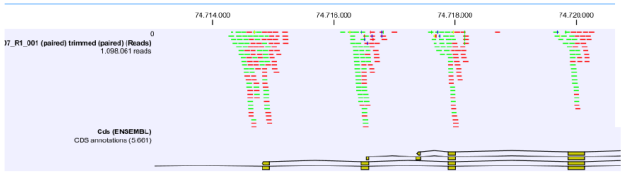
- **requires *a priori* knowledge of target to be sequenced** (e.g. exons)
- **target enrichment by hybridization**



Exome-Seq

Exome Enrichment:

Target region size	62 Mb
Number of target genes	20,794
Number of target exons	201,121
Number of probes	340,427



Exome-Seq

Result of exome-Seq is a list of nt-differences observed in the sequenced DNA and compared with the reference-genome

SNVs : single nucleotide variants

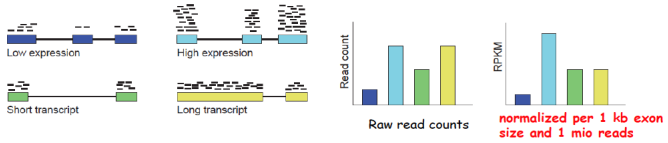
SNPs : single nucleotide polymorphisms*

cSNPs : coding SNPs

* at least 1% as Allele in human population

Transcriptome analysis (RNA-Seq)

- **RNA isolation > cDNA production & fragmentation**
- **high-throughput sequencing**
- **mapping of fragments to reference genome**
- **read-counting as measure of gene transcription**



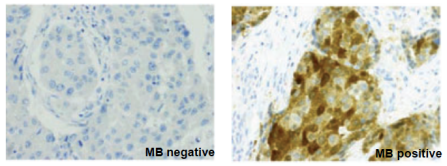
A typical measure is "RPKM"
= reads per Kb of exon sequence and 1 mio reads in the dataset

MB immunostaining on breast tumors

Example:

Cancer analysis:

Myoglobin (MB) in breast cancer

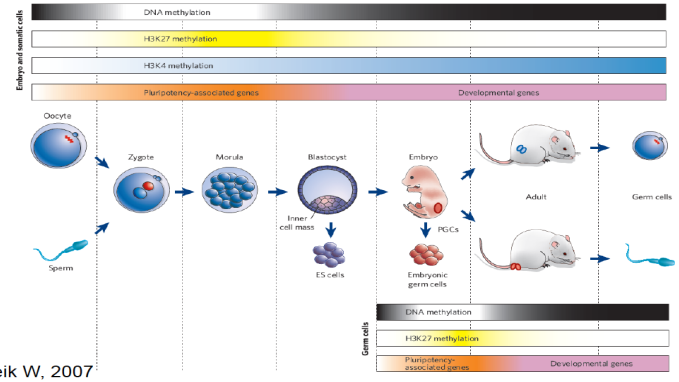


RNA-Seq

differential expressed genes
hint at molecular changes causing symptoms, signs and cell / tissue organ-functions observed

Result of the RNA-Seq is a list of genes that are statistically significantly differentially regulated between the considered datasets.
Sometimes these lists contain hundreds or thousands of genes.
This makes it necessary to functionally categorize the genes (GeneOntology Vocabulary, KEGG Pathways).
Are certain functional categories overrepresented or underrepresented in the differentially regulated genes?
What does that say about the "biology" of the comparative samples?

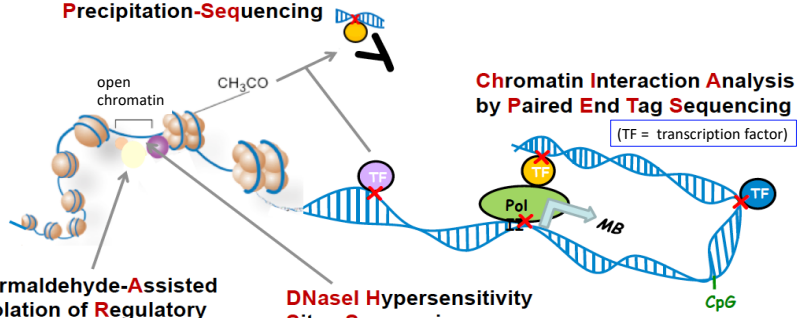
Epigenetics – Genome modifications by e.g. methylation during the development (differentiation) of an organism



Reik W, 2007

Genome-wide detection of chromatin modifications by next generation sequencing (NGS) methods

Chromatin Immuno-Precipitation-Sequencing



Chromatin Interaction Analysis by Paired End Tag Sequencing

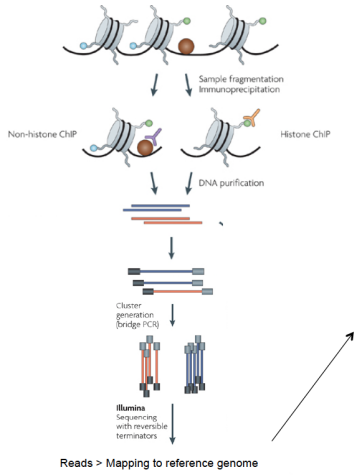
(TF = transcription factor)

Formaldehyde-Assisted Isolation of Regulatory Elements-Sequencing

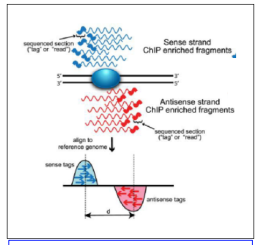
DNaseI Hypersensitivity Sites-Sequencing

Bisulfite-Sequencing

ChIP-Seq

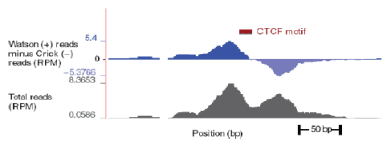


„peak calling“

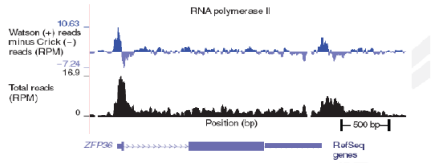


Peaks show sites where TFs have bound

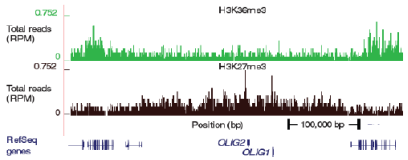
ChIP results examples:



Trancriptional regulation protein (CTCF)



RNA polymerase II



Epigenetics: Repressive mark for H3K27me

ChIP Seq peak types from different experiments

Computation for ChIP-seq and RNA-seq studies
Shirley Pepke, Barbara Wold & Ali Mortazavi
Nature Methods 6, S22 - S32 (2009)

Next Generation Sequencing (NGS) methods

2nd Next Generation Sequencing (NGS) methods

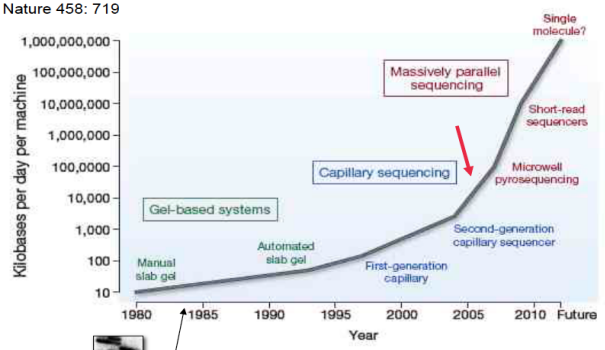
Common feature:

based on enzyme (DNA polymerase)

	454 (Roche)	Ion Torrent	Illumina	Complete Genomics	Sanger
DNA matrix	Emulsion PCR	Emulsion PCR	Bridge PCR	amplification: DNA nano balls	Plasmids Clones PCR
Sequencing Method	seq-by-synth: Pyrosequencing	seq-by-synth: Proton release	seq-by-synth: reversible Dye-Terminators	Seq-by-ligation	seq-by-synth: Dye Terminator 96 capillaris
Read length	av. 600 bp (up to 1000)	Up to 700	2 x 100 bp (up to 2 x 300)	70 bp	1000 bp
Data	600 Mbp	1 Gbp	Up to 1,5 Tbp	20-60 Gbp	0,1 Mbp
Runtime	10 hrs	90 min	2-10 Days	?	2 hrs

Gene/genome sequencing technologies – a million-fold improvement

Nature 458: 719



my diploma thesis: 1kb (chicken TrkC) , 4 weeks (day & night in the isotope lab)

NGS technology: How to...

tedious cloning
high chemical costs
slow electrophoresis

PCR or even single molecules
extreme miniaturisation
massively-parallel read-out

2nd Next Generation Sequencing (NGS) methods

Common feature:

based on enzyme (DNA polymerase)

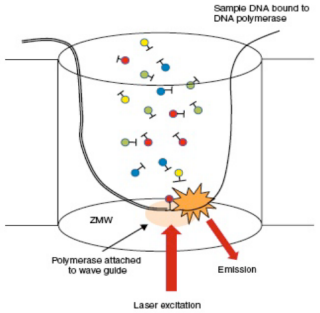
	454 (Roche)	Ion Torrent	Illumina	Complete Genomics	Sanger
DNA matrix	Emulsion PCR	Emulsion PCR	Bridge PCR	amplification: DNA nano balls	Plasmids Clones PCR
Sequencing Method	seq-by-synth: Pyrosequencing	seq-by-synth: Proton release	seq-by-synth: reversible Dye-Terminators	Seq-by-ligation	seq-by-synth: Dye Terminator 96 capillaris
Read length	av. 600 bp (up to 1000)	Up to 700	2 x 100 bp (up to 2 x 300)	70 bp	1000 bp
Data	600 Mbp	1 Gbp	Up to 1,5 Tbp	20-60 Gbp	0,1 Mbp
Runtime	10 hrs	90 min	2-10 Days	?	2 hrs

3rd Next Generation Sequencing (NGS) methods

	Pacific Bioscience	Genia (Roche) Not released	Oxford Nanopore
based on: enzyme (DNA polymerase)	√	√	No enzyme !
DNA matrix	Single-mol	?	Single-mol
Sequencing Method	seq-by-synth: labeled hexaphosphate Nt's	seq-by-synth: PEG tagged Nt´s + Nanopore	direct sequencing
		Combination of Nanopore and enzyme	
Read length	20 kbp (mean) 90 kbp max	?	20 kbp (mean) Up to 1.2 Mbp
Data	5-8 Gbp /SMRT Cell (16 Cells)	?	10-20 Gbp / Flow Cell
Runtime	30 min – 10 hrs	?	Realtime (48 hrs max)

Pacific Biosciences – PacBio

single molecule long-range sequencing

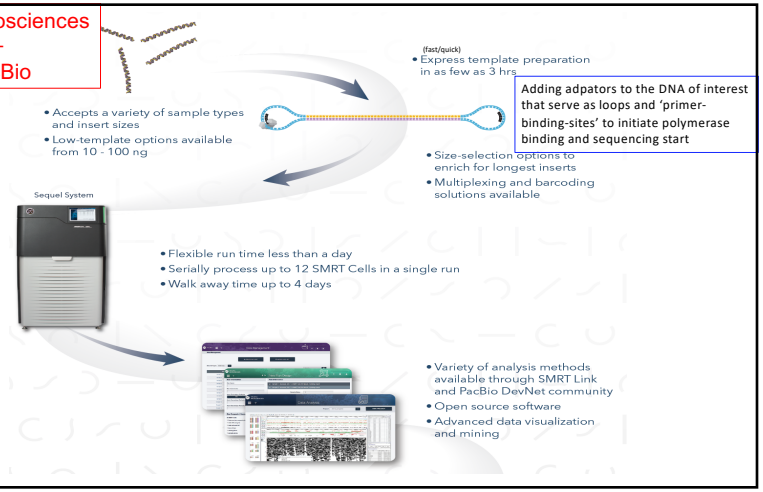


- **read length of > 5000 Bp !!**
- **long-range sequence information !!**

but

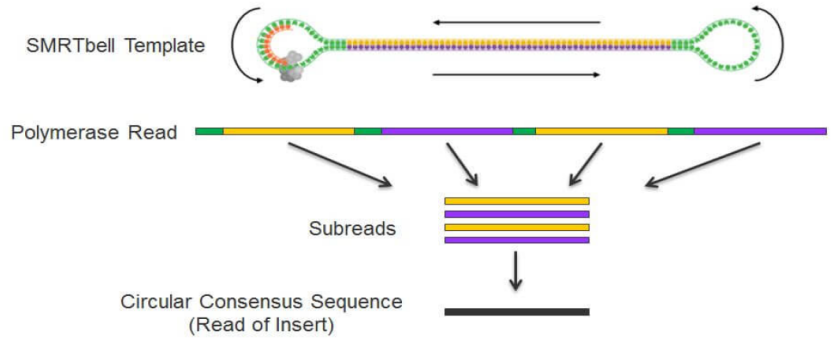
- **expensive machine**
- **low to medium throughput**
- **extreme error rate (up to 20%)**
- error-rate decreasing year by year by improved chemistry etc

Pacific Biosciences – PacBio



Adding adpators to the DNA of interest that serve as loops and 'primer-binding-sites' to initiate polymerase binding and sequencing start

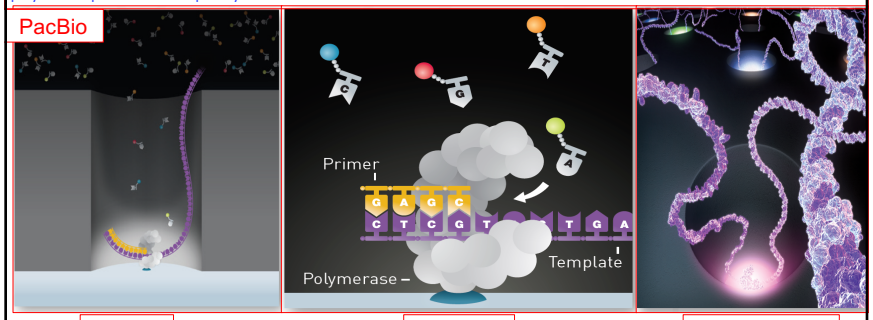
PacBio circular consensus sequencing provides high-accuracy long reads at the single-molecule level



The circular nature of the SMRTbell DNA template allows polymerase to sequence the same DNA molecule multiple times with multiple passes. This produces high intramolecular consensus accuracy.

Single Molecule, Real-Time (SMRT®) technology is built upon two key innovations that overcome major challenges in the field of sequencing. Zero-Mode Waveguides (ZMWs) allow light to illuminate only the bottom of a well in which a DNA polymerase / template complex is immobilized. Phospholinked nucleotides allow observation of the immobilized complex as the DNA polymerase produces a completely natural DNA strand.

PacBio



Zero-Mode Waveguides

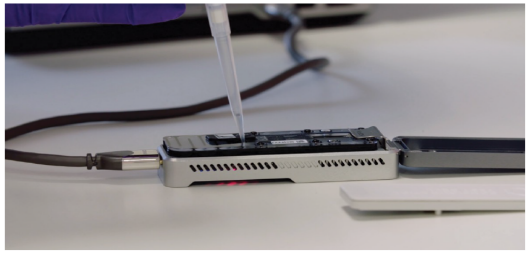
Phospholinked Nucleotides

Up to a million ZMWs per SMRT Cell

SMRT Cells containing up to a million ZMWs are processed on PacBio® Systems which simultaneously monitor each of the waveguides in real time.

Oxford Nanopore Technologies (ONT)

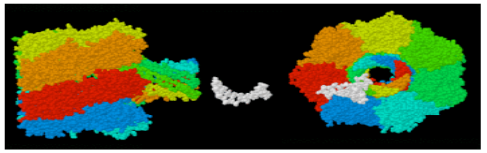
Nanopore sequencing: towards single molecule detection

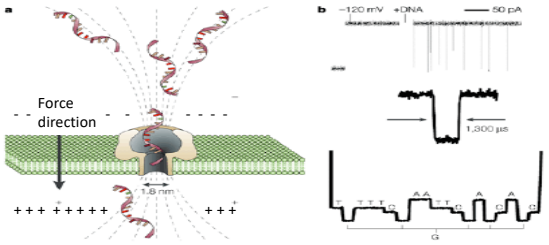


Oxford Nanopore
„MinION“

Nanopore sequencing

www.wikipedia.org

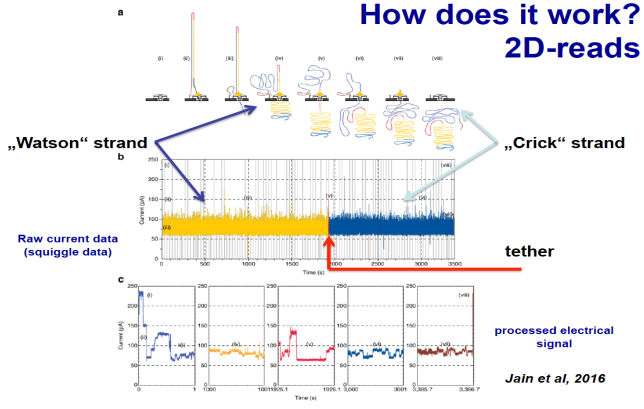




Nature Reviews | Drug Discovery

Oxford Nanopore Technologies (ONT)

How does it work? 2D-reads



Jain et al, 2016

Oxford Nanopore Technologies (ONT)

The future...

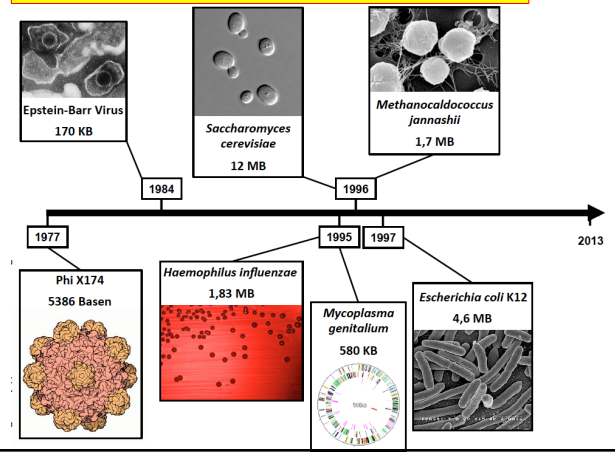
- Output = ?
- Mobile DNA analysis for everyone and anywhere

PromethION

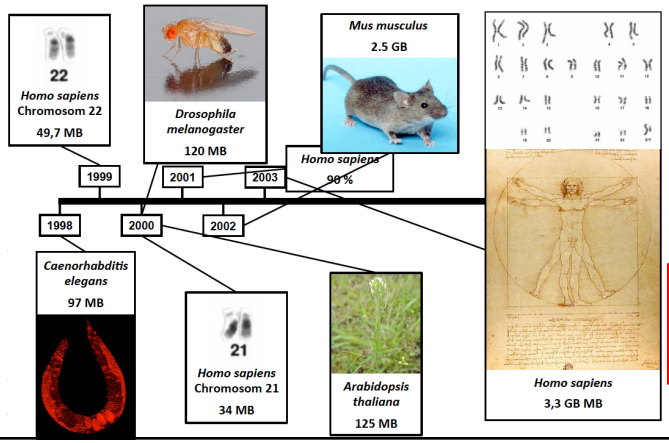
SmidgION

- 48 Flowcells with 2048 Channels (= 192 MinIONs)
- 6-11 Tbp Output / 24h
- Direct RNA sequencing?
- Direct 5-mC sequencing

Milestone in Genome Sequencing of various organisms



Milestone in Genome Sequencing of various organisms



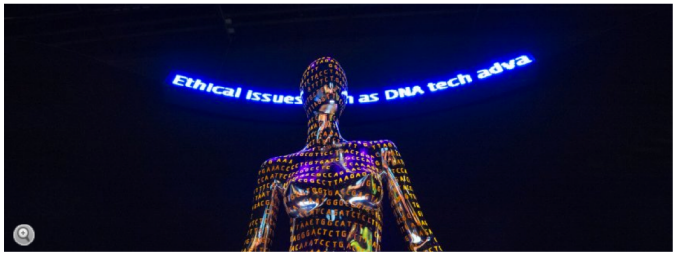
bread wheat
2018
17 G bp !!

Milestone in Genome Sequencing of various organisms

The 1000 Genomes Project 2008-2011:

1000 Genomes
A Deep Catalog of Human Genetic Variation

1000-Genomes-Projekt: Genome of ~ 2500 humans decoded



Example: Designer Baby (bio system)

- **Definition** (Wikipedia):
  A designer baby is a human embryo that has been genetically modified, usually following guidelines set by the parent or scientist, to produce desirable traits.

Genome Sequencing: practical applications:

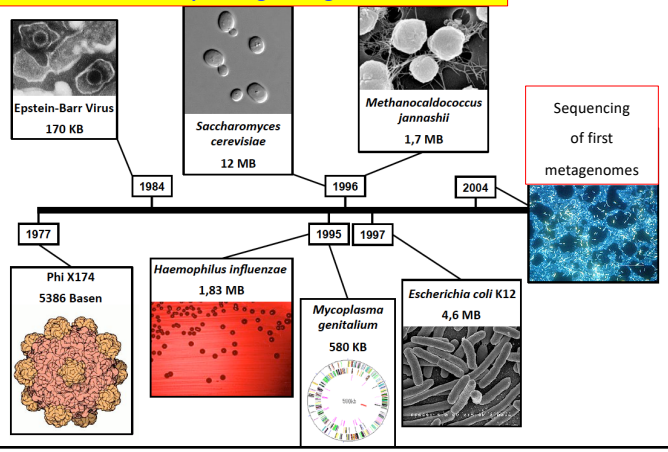
***homo sapiens*:**
- mapping/annotation of gene defects (genome-wide association study (GWAS))
- correlation between human genes and model-organisms (e.g. mouse) genes

***Mus musculus*:**
- International knock-out (ko) mouse consortium (IKMC):
- Establishing ko mice (>9000 genes)
- ko in embryonic stem cells (ESCs)

***outlook*:**
- Diagnosis of gene defects by gene sequencing (---> Oxford Nanopore Technologies (ONT))
- Gene repair and treatment by CRISPR|Cas and iPSC technology

International knock-out (ko) mouse consortium (IKMC)

Genome Sequencing: metagenomics:



Genome Sequencing: metagenomics:

Definition of metagenome:

All genetic information of a defined habitat (genome / genetic information of all residential organisms of that defined habitat)

Flora of intestine:

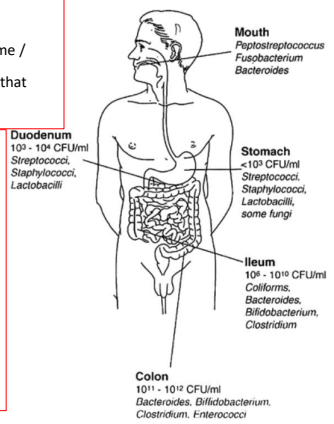
about 10^{11} – 10^{12} microorganism per gram colon !

Total: 10^{14} microorganism = 10x more than body cells

about 1000 different species

about 2000 genes per species

total genetic information (metagenome): 100x more genes than human genome (about 20,000 genes)

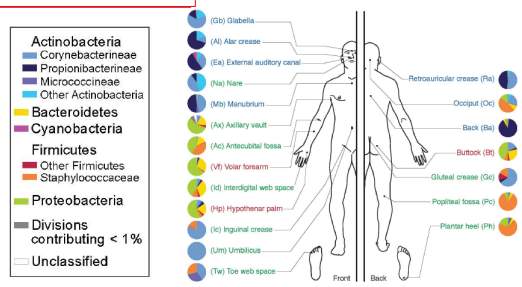


Genome Sequencing: metagenomics:

Problem:

- Which organisms are present within a defined habitat?
- ---> 16S rDNA/rRNA sequencing

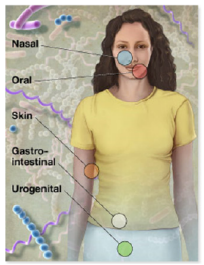
Metagenome of the human skin flora



Genome Sequencing: metagenomics:

Human Microbiome Project (2008-2013):

- Sequencing of the genome of 600 - 1000 bacteria species from different habitats of the human body
- Metagenome of the microflora of different parts of the body



MetaHIT (2008-2012):

Metagenome of the flora of the intestine of 124 humans

Genome Sequencing: metagenomics:

Problem:

- Which organisms are present within a defined habitat?
- ---> 16S rDNA/rRNA sequencing

Metagenome of your working place

Metagenome of your office working place: ~ 30, 000 bacteria

Genome Sequencing: metagenomics:

Problem:

- Which organisms are present within a defined habitat?
- ---> 16S rDNA/rRNA sequencing

Metagenome of dust sample from your working place

January 2014 | Volume 9 | Issue 1 | e87093

OPEN ACCESS Freely available online

PLOS ONE

Architectural Design Drives the Biogeography of Indoor Bacterial Communities

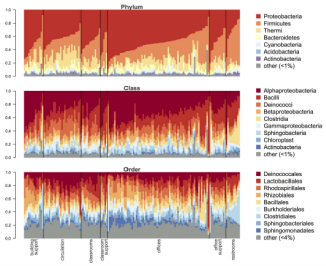
Steven W. Kembel[1,2,3☯], James F. Meadow[2,3,4☯], Timothy K. O'Connor[2,3,4], Gwynne Mhuireach[2,5], Dale Northcutt[2,5], Jeff Kline[2,5], Maxwell Moriyama[2,5], G. Z. Brown[2,5,6], Brendan J. M. Bohannan[2,3], Jessica L. Green[2,3,7]

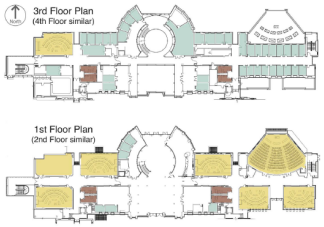
Genome Sequencing: metagenomics:

Problem:

- Which organisms are present within a defined habitat?
- ---> 16S rDNA/rRNA sequencing

Metagenome of dust sample from your working place



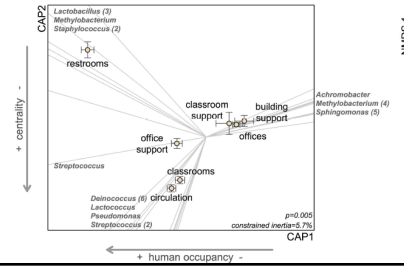


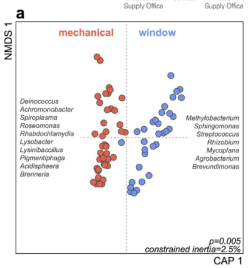
Genome Sequencing: metagenomics:

Problem:

- Which organisms are present within a defined habitat?
- ---> 16S rDNA/rRNA sequencing

Metagenome of dust sample from your working place





Genome Sequencing: concluding remarks

Sequencing becomes faster and easier – still, classical molecular biology methods, applying e.g. target-mutagenesis, are needed to identify NT-variation-induced phenotypes.