GENOMIC DATA ANALYSIS

Irene Tiemann-Boege
Irene.tiemann@jku.at

LVA-Nr. 320.301 and 320.304

JOHANNES KEPLER UNIVERSITÄT LINZ

OUTLINE FOR TODAY

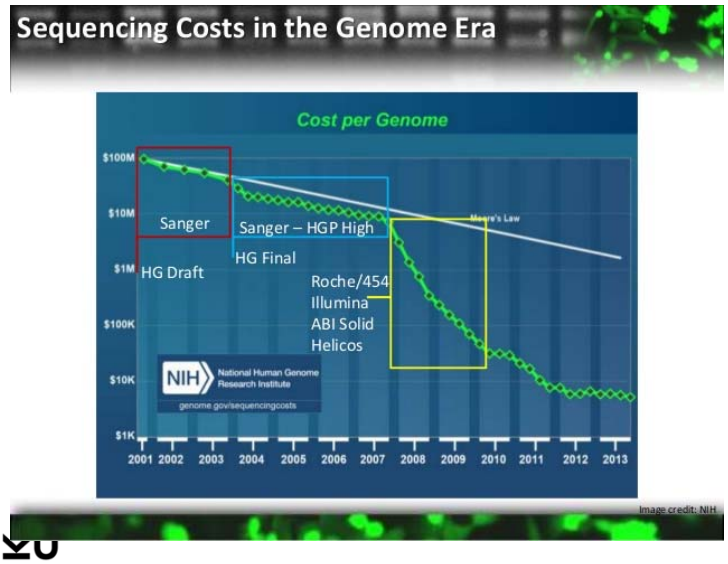
- New sequencing technologies (NGS)
- Principles of next generation sequencing technologies
- Commercial platforms
- Uses of NGS
- Individual genomes

NEXT GENERATION SEQUENCING (NGS)

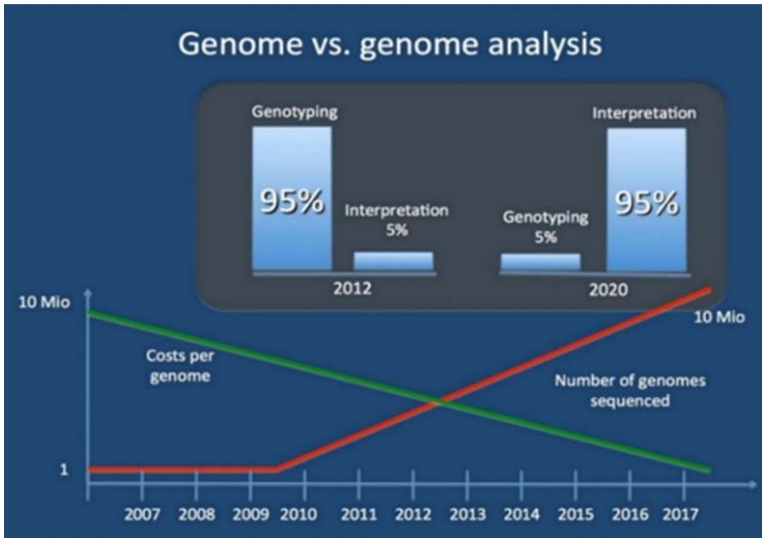
- Generation of 100Mb in one day…a whole human genome in a few days!
- Very high throughput and a very low cost
- Aim: $1000 per human genome per day
- In 2001: Human genome Cost: $0.3-3 billion
- 3Gb (one human genome) sequenced in ~8 years by many centers

Today, sequencing a human genome costs ~$1000-a takes 3-5 days!

SEQUENCING FASTER AND CHEAPER



THE ANALYSIS IS GETTING MORE EXPENSIVE



JKU

NEW STRATEGIES FOR SEQUENCING WERE DEVELOPED BASED ON:

- Capillary Sanger sequencing is already optimized to its utmost potential
- Short-read sequencing is possible with the availability of whole genomes as a reference
- Progress in technology:
  - ☐ microscopy
  - ☐ surface chemistry
  - ☐ nucleotide and enzyme biochemistry
  - ☐ computation
  - ☐ data storage and handling
  - ☐ data analysis

JKU

OUTLINE FOR TODAY

- New sequencing technologies (NGS)
- Principles of next generation sequencing technologies
- Commercial platforms
- Uses of NGS
- Individual genomes

JKU

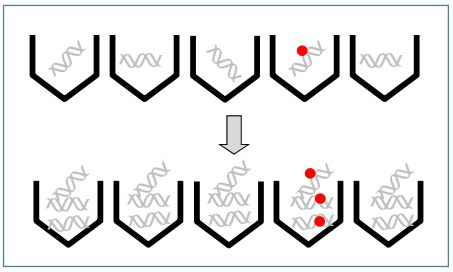
SEQUENCING TECHNOLOGIES

- Next generation sequencing (2nd generation):
  - amplification of single DNA molecules
  - PCR in microscopic compartments
    - Emulsion PCR or bridge amplification
  - Sequencing of millions of PCR products in parallel
- Third generation sequencing (Nanopore/ PacBio)
  - No amplification step
  - Directly sequencing single DNA molecules

JKU

WHAT IS SINGLE MOLECULE AMPLIFICATION?

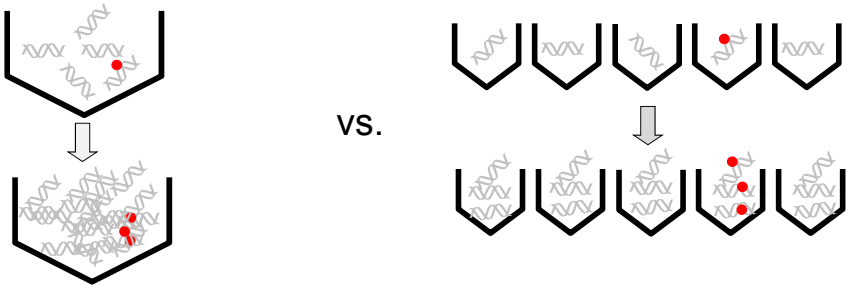
- Single molecules are amplified in separate compartments
- Large amounts of clonal products are produced
- Easy analysis of the amplified material



JKU

ADVANTAGES OF SINGLE MOLECULE ANALYSIS

- In pooled samples sequences are diluted out
- A single molecule approach represents all DNA variants-no mixing of variants or sequences



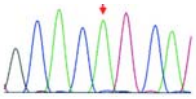
Multi-template PCR

Single template PCR

JKU

CLONING IS ALSO SINGLE MOLECULE AMPIFICATION

Each colony is derived from one initial bacterium

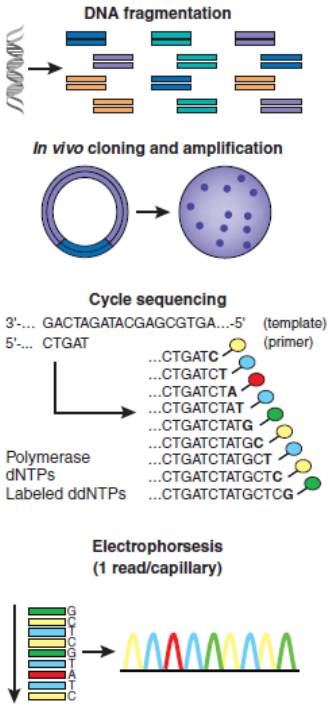


JKU

OLD SEQUENCING TECHNOLOGY

- Old: Sanger capillary sequencing
- Genomic DNA is fragmented, then cloned into E. coli
- A single bacterial colony is picked, the plasmid DNA is isolated.
- The sequencing reaction generates a ladder of ddNTP-terminated, dye-labeled products
- Fragments are separated with high-resolution
- electrophoresis in one of 96 or 384 capillaries
- As fluorescently labeled fragments of discrete sizes pass a detector, the four-channel emission spectrum is used to generate a sequencing trace.

Taken from Shendure & Ji.2008. NatBiotech.



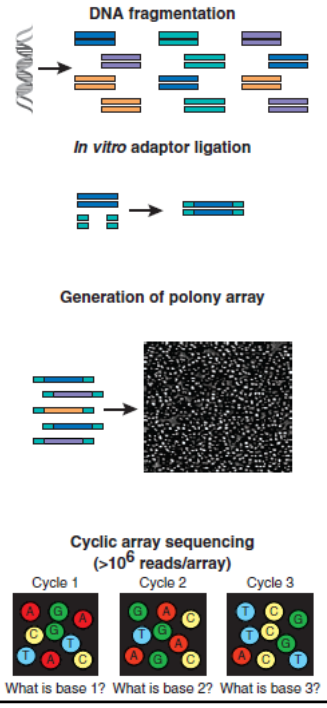
JKU

OLD VS. NEW SEQUENCING TECHNOLOGIES

- New: Cyclic array methods
- Genomic DNA is fragmented, then common adaptors are ligated
- Single molecules are amplified via diferent methods in picoliter volumes that results in an array of millions of spatially immobilized PCR colonies
- All PCR colonies are tethered to a planar array and all colonies can be treated and analyzed **in parallel**
- Imaging and analysis of the fluorescently labelled colonies are used to acquire sequencing data **in parallel**.
- Successive iterations of enzymatic treatment and imaging are used to build a sequencing read for each array feature

Taken from Shendure & Ji.2008. NatBiotech.

JKU



OLD VS. NEW SEQUENCING TECHNOLOGIES

- Advantages of second generation technologies:
  - Throughput and cost—possible via miniaturization

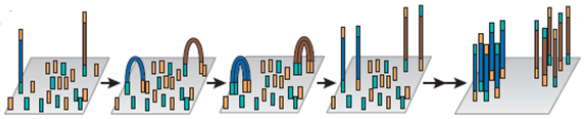
- Disadvantages:
  - Read-length (read-lengths are currently much shorter than conventional sequencing); longest sequence is ~700bp on average per read
  - Raw accuracy (base-calls generated by the new platforms are at least tenfold less accurate than base-calls generated by Sanger sequencing).

Taken from Shendure & Ji.2008. NatBiotech.

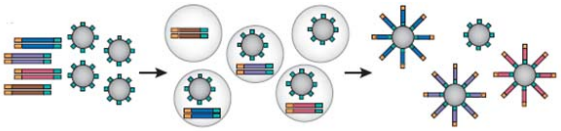
JKU

A QUANTUM LEAP IN SEQUENCING

- very efficient single molecule analysis platforms based on miniaturization and parallel processing of reactions
- Single molecules amplified on a glass surface (bridge amplification)



- Single molecules amplified on beads (emulsion PCR)



JKU

OUTLINE FOR TODAY

- New sequencing technologies (NGS)
- Principles of next generation sequencing technologies
- Commercial platforms
- Uses of NGS
- Individual genomes

JKU

Next generation sequencing are very efficient single molecule analysis platforms

Illumina/MySeq

PacBio

Ion Torrent

Nanopore

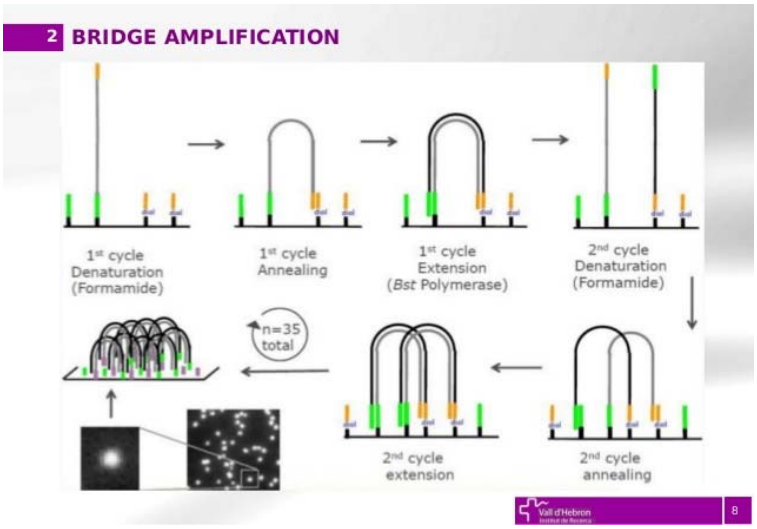
JKU

ILLUMINA-BRIDGE AMPLIFICATION

- Single templates are attached on a glass surface covered by the forward and the reverse primer
- By bridge amplification clusters of PCR products or PCR-colonies are formed
- Sequencing chemistry is based on reversible dye-terminators (3'-OH is protected and each nucleotide emits a different color)
- The color of the cluster (spot) is determined by the incorporated nucleotide and translated into a base

JKU

BRIDGE AMPLIFICATION



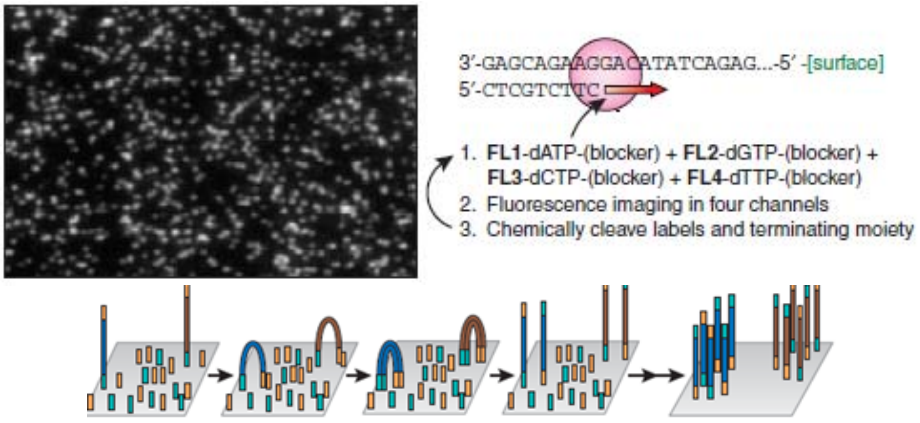
JKU

BRIDGE AMPLIFICATION

- Both primers are covalently attached to a solid surface (glass); distributed randomly and at a high density
- one DNA template is attached to the primer array
- PCR components (polymerase, dNTPs, buffers, etc) are added to the array
- Cycling is started
- The end of the template can anneal to the complementary primer and get extended
- PCR products accumulate around the area that can be reached by the template

JKU

ILLUMINA GENOME ANALYZER (SOLEXA)



JKU

ILLUMINA GENOME ANALYZER

- A dense area of DNA molecules is generated directly on a surface by bridge PCR (cluster PCR) forming immobilized PCR colonies (polonies)
- Each sequencing cycle includes the simultaneous addition of four modified deoxynucleotides
- Deoxynucleotides have the following modification:
  - a specific fluorescent label
  - a reversibly terminating moiety at the 3’ hydroxyl position.
- A DNA polymerase incorporates the modified deoxynucleotides
- The polonies are imaged in four channels
- Remove fluorescent labels and the terminating moiety
- Turcatti et al. 2008. Nucleic Acid Research.
- https://www.youtube.com/watch?v=womKfikWlxM

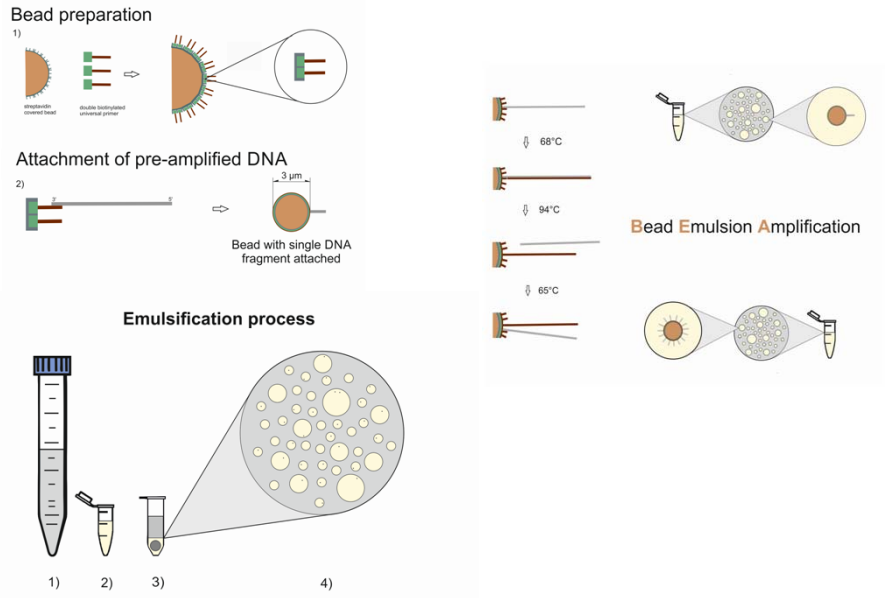
JKU

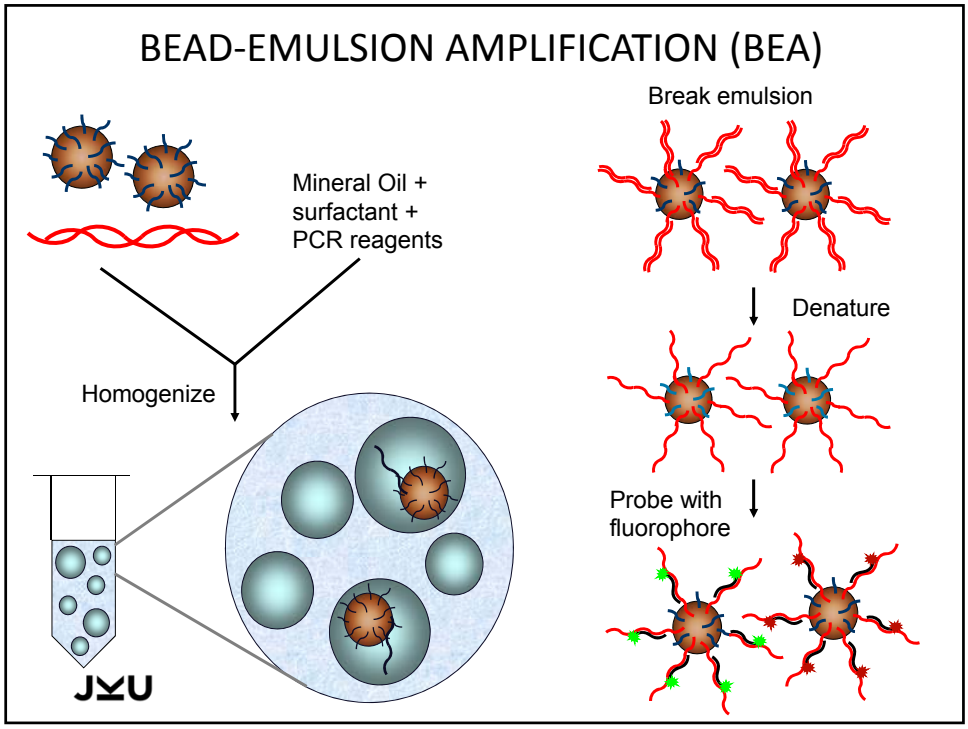
EMULSION PCR

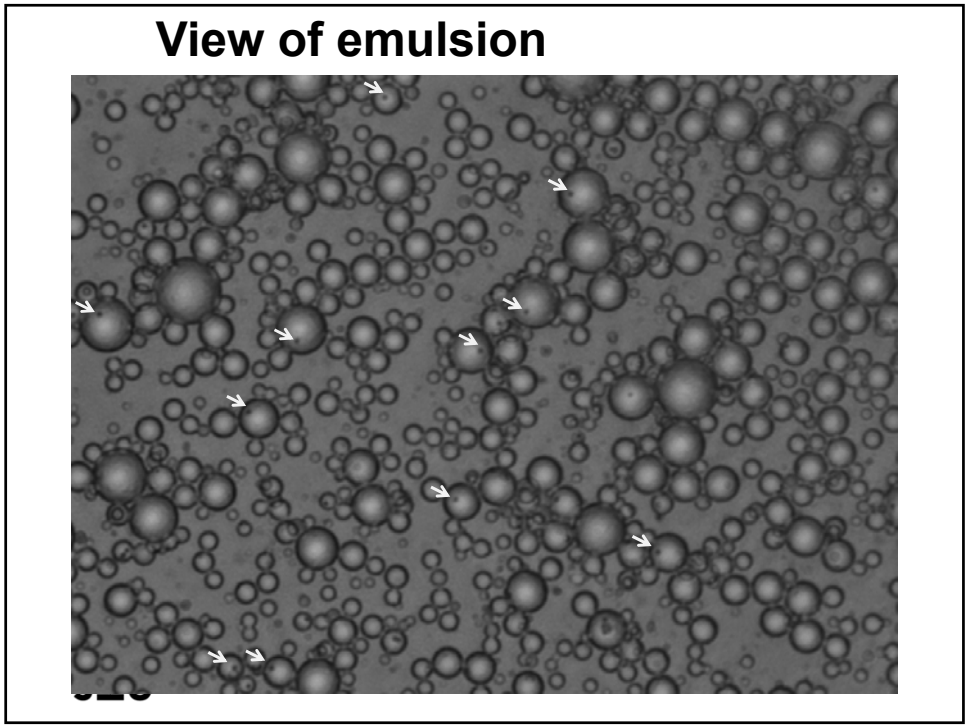
- Only one primer is attached to a microscopic bead
- PCR is carried out within an emulsion formed by an water-oil-phase
- Amplified DNA stays bound to the bead
- Only one complementary template can move, but is limited by the compartment formed by the emulsion

JKU

EMULSION PCR
EMULSION PCR



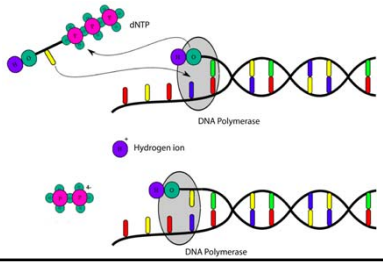

BEAD-EMULSION AMPLIFICATION (BEA)
Mineral Oil +
surfactant +
PCR reagents
Homogenize
Break emulsion
Denature
Probe with
fluorophore
JKU


View of emulsion

EM-PCR BASED SEQUENCING

- Ion-Torrent
  - Single molecules are amplified on beads in an emulsion
  - Principle: smallest pH meter
  - Flushing of different dNTPs (one at a time) per cycle
  - The incorporation of a dNTP releases one proton
  - The change in pH is measured and translated into a sequence

- Video:
- https://www.youtube.com/watch?v=KzdWZ5ryBIA
- http://www.youtube.com/watch?v=MxkYa9XCvBQ



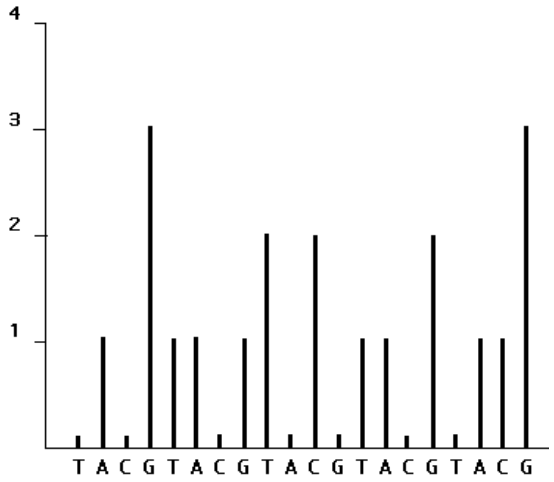
JKU

ION TORRENT

- Change in pH detected by Ion-Sensitive Field Effect Transistors
- Direct measurement of nucleotide incorporation during DNA synthesis by changes in pH
- No need for modified DNA bases; non-optical sequencing
- Read length ~400bp
- Potential difficulties reading through highly repetitive or homopolymer regions

JKU

OUTPUT IN PH DIFFERENCES?



JKU

3RD GENERATION SEQUENCING— SEQUENCING SINGLE MOLECULES

- PacBio
- Nanopore technologies

JKU

REAL TIME SEQUENCING

- Oxford Nanopore sequencing
- Ultra-fast DNA sequencing in which nucleic acids are driven through a nanopore (either a biological membrane protein such as alpha-hemolysin or a synthetic pore).
- Fluctuations in DNA conductance through the pore, or, potentially, the detection of interactions of individual bases with the pore, are used to infer the nucleotide sequence.

Video:
https://www.youtube.com/channel/UC5yMlYjHSgFfZ37LYq-dzig

JKU

OXFORD NANOPORE SEQUENCING

- Portable DNA Sequencer Can ID Bacteria and Viruses
- A palm-sized, nanopore-based USB device can recognize E. coli, cowpox, and vaccinia
- Needs preamplified material (PCR)

BIOMED CENTRAL, ANDREW KILIANSKI

http://www.the-scientist.com//?articles.view/articleNo/42542/title/Portable-DNA-Sequencer-Can-ID-Bacteria-and-Viruses/

JKU

ADVANTAGES AND DISADVANTAGES OF NEW SEQUENCING TECHNOLOGIES

- Advantages of second generation technologies:
  - Throughput and cost
- Disadvantages:
  - Read-length (except PacBio)
  - Accuracy (error rates)

JKU

PACBIO

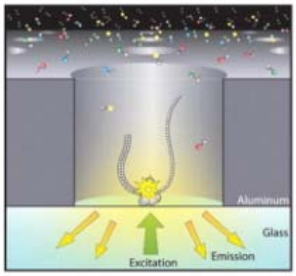
- Single molecule, real-time sequencing (SMRT)
- Longest sequencing reads (~30 kb; average 10,000-15,000bp)
- Hairpin adaptors are ligated to the DNA ends (circularize DNA; several rounds of sequencing)

- Sequence on zero-mode waveguides (polymerase attached to a surface) via light pulses
- Each of the four nucleotides is labeled with a different fluorescent dye. As a nucleotide is held by the polymerase, a light pulse is produced that identifies the base
- Can detect epigenetic modifications such as methylated DNA

JKU

PACBIO

- Each SMRT cell contains 150,000 ZMWs. Approximately 35,000–75,000 of these wells produce a read in a run lasting 0.5–4 h, resulting in 0.5–1 Gb of sequence



Rhoads et al. 2015. Genomics, Proteomics & Bioinformatics

JKU

PACBIO

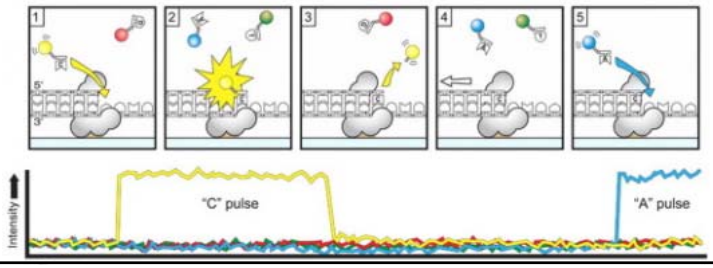
(1) A fluorescently-labeled nucleotide associates with the template in the active site of the polymerase.

(2) The fluorescence output of the color corresponding to the incorporated base (e.g. yellow)

(3) The dye-linker-pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW, ending the fluorescence pulse.

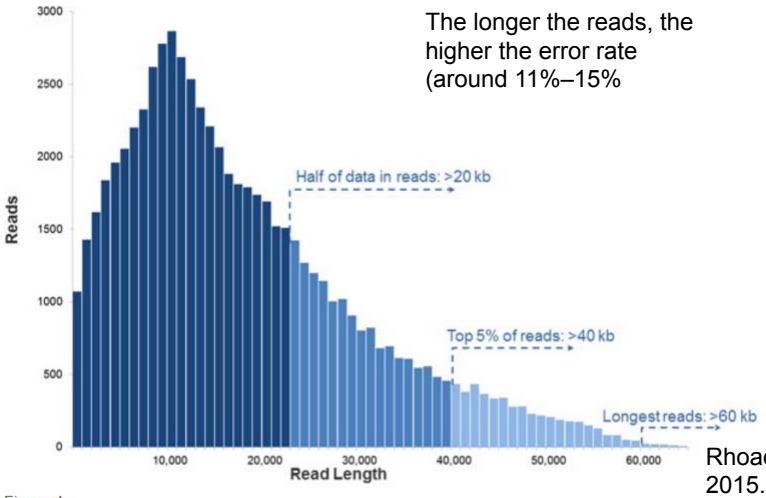
(4) The polymerase translocates to the next position.

(5) The next nucleotide associates with the template in the active site of the polymerase, initiating the next fluorescence pulse, which corresponds to base A here.



Rhoads et al. 2015. Genomics, Proteomics & Bioinformatics

PACBIO



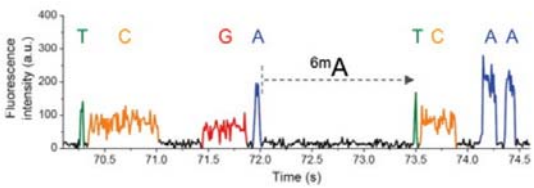
The longer the reads, the higher the error rate (around 11%–15%

Figure 4.
PacBio RS II read length distribution using P6-C4 chemistry

Rhoads et al. 2015. Genomics, Proteomics & Bioinformatics

PACBIO

- Uses of PACBio:
- *De novo* genome assemblies
- Can detect epigenetic modifications such as methylated DNA (6mA)



- Pac f sequencing highly-repetitive genomic regions
- https://www.youtube.com/watch?v=v8p4ph2MAvI

JKU

COMPARISON OF DIFFERENT NGS PLATFORMS

Table 1.
Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)
Sanger ABI 3730×l	1st	600–1000	0.001	96	0.5–3 h	500
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57
Illumina HiSeq 2500 (High Output)	2nd	2 × 125	0.1	8×10^9 (paired)	7–60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	2nd	2 × 250	0.1	1.2×10^9 (paired)	1–6 days	0.04
SOLiD 5500×l	2nd	2 × 60	5	8×10^8	6 days	0.11
PacBio RS II: P6-C4	3rd	1.0–1.5 × 10^4 on average	13	3.5–7.5 × 10^4	0.5–4 h	0.40–0.80
Oxford Nanopore MinION	3rd	2–5 × 10^3 on average	38	1.1–4.7 × 10^4	50 h	6.44–17.90

JKU

OUTLINE FOR TODAY

- New sequencing technologies (NGS)
- Principles of next generation sequencing technologies
- Commercial platforms
- Uses of NGS
- Individual genomes

JKU

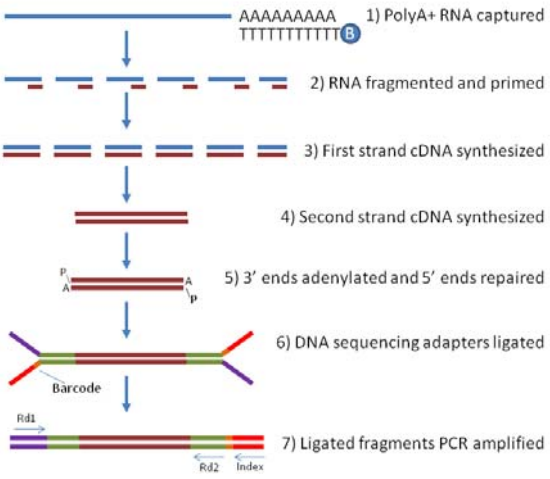
NEW SEQUENCING TECHNOLOGIES: NEXT GENERATION SEQUENCING

- Uses:
  - Re-sequencing
    - Discovery of new polymorphisms
    - Personalized medicine (discover mutations)
    - Sequence tissues or cancers
  - De novo sequencing of unsequenced genomes
  - Exome sequencing
  - RNA Seq
  - Protein-DNA interactions
  - Chromosome conformation
    - Epigenetics

JKU

STUDYING THE TRANSCRIPTOME: RNA-SEQ ANALYSIS

- Study of the transcriptome (mRNA in a tissue)
- Library preparation from mRNA



JKU

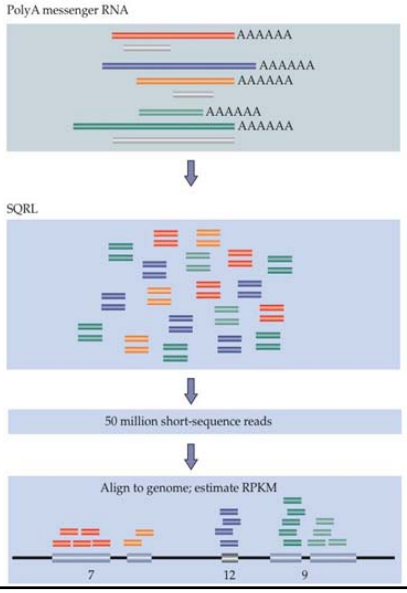
STUDYING THE TRANSCRIPTOME: RNA-SEQ ANALYSIS

- The polyAfraction of cellular RNAis isolated and fragmented into 200 base sequences
- Result: short quantitative random RNA libraries (SQRLs)
- These libraries become the template for one of the next-generation sequencing platforms
- 10 million short sequence reads for each RNA sample are generated
- Short sequences are then mapped back to the reference genome of the species

JKU

STUDYING THE TRANSCRIPTOME: RNA-SEQ ANALYSIS

- Produce short random RNA libraries (SQRLs)
- SQRLs are sequenced by next-generation sequencing platform
- Sequences are mapped to the RefSeq
- Digital estimate of the abundance of each transcript = quantitative determination of transcript abundance (expression)



JKU

USES OF RNA-SEQ

Coverage as measure of expression

- 4x coverage = expressed gene
- No. of reads in a particular exon are normalized by length and total reads (RPKM= reads per kilobase of predicted exon per million total reads)
- Units of RPKM = expression levels and correlate with qPCR results

JKU

STUDYING THE TRANSCRIPTOME: RNA-SEQ ANALYSIS

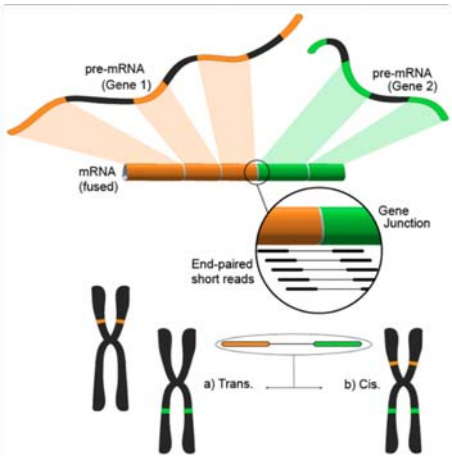
- Advantage over microarrays (based on hybridization of mRNA on known probes)
- RNA-Seq provides:
  - Accurate representation of splicing variants (exon-exon junctions represented)
  - New types of mRNAs: all mRNAs are represented even unknown RNA
  - New mutations
  - The number of reads is proportional to the number of initial mRNA (quantify expression)

JKU https://www.youtube.com/watch?v=PisqV9cyfoA

USES OF RNA-SEQ

Post-transcriptional variation

- Comparison of the genomic with the transcriptomic sequence can help discovering post-transcriptional edits
- Fusion genes: exon-exon junction where the exons come from different genes



JKU

TARGETED NGS SEQUENCING

- Select specific regions in the genome before sequencing
- This occurs during library preparation
- Advantage: unwanted regions are not sequenced; cheaper sequencing costs and larger throughput

JKU

TARGET-ENRICHMENT STRATEGIES

- PCR- Based
- Hybrid- or in solution capture:
  - pool of custom oligonucleotides (probes)
  - Probes attached to beads or to solid surface
  - Selective hybridization to genomic regions of interest
  - Hybridizided probes can be separated and cleared from unwanted DNA
  - ~3.5 Mb can be captured
- Molecular inversion probes (MIP)

JKU

TARGET-ENRICHMENT STRATEGIES

- PCR- Based
  - Amplifying segments of interest prior to sequencing
  - Amplify 480 kb in 124 PCRs,
  - E.g. candidate genes involved in altitude adaptation
  - This approach is laborious and time-consuming, and limited in the number of kb to be sequenced

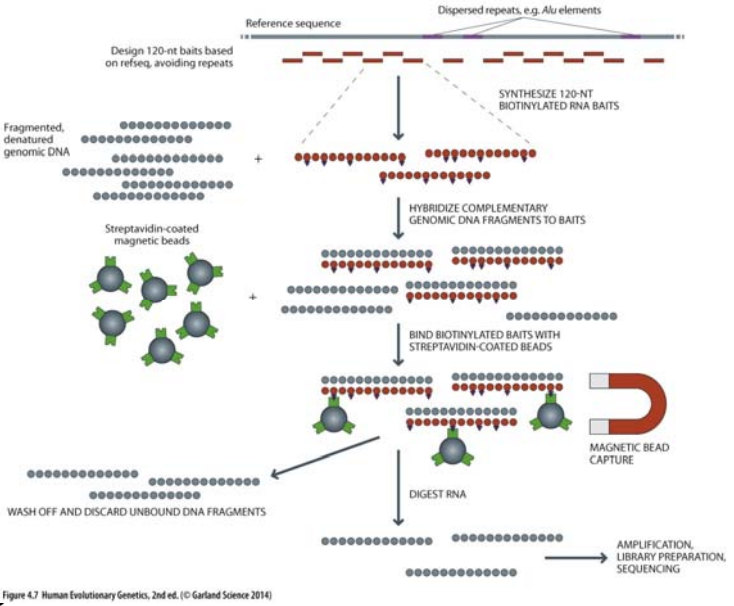
JKU

TARGET-ENRICHMENT STRATEGIES

- Hybrid- or in solution capture
  - pool of custom oligonucleotides (probes 50-120nt)
  - Probes attached to beads or to a solid surface
  - Selective hybridization to complementary single-stranded genomic DNA fragments
  - Hybridizided products can be separated and cleared from unwanted DNA
  - Clearing of capture probes
  - Library preparation from captured DNA
  - Capture from a few hundred kilobases up to tens of megabases)
  - ~3.5 Mb can be captured

JKU

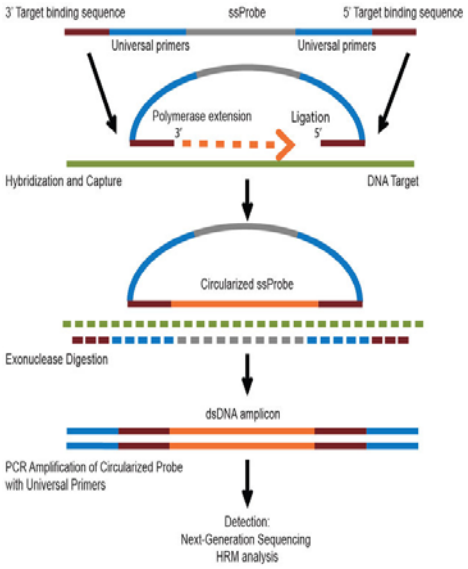
HYBRID- OR IN SOLUTION CAPTURE



JKU Figure 4.7 Human Evolutionary Genetics, 2nd ed. (© Garland Science 2014)

MOLECULAR INVERSION PROBES

- Single stranded probe with specific target binding sequences
- Binding sequence is extended and ligated
- Exonuclease digest (Circularized probes cannot be digested)
- Circularized probes are amplified
- PCR products are sequenced



JKU

EXOME SEQUENCING (WES)

- Whole exome sequencing (WES)
  - ☐ only sequence the exons of our genome (protein coding information)-
  - ☐ 1% of the genome ($30x10^6$nt)
  - ☐ initially ~30 Mb, but increased to ~50 Mb in later designs
  - ☐ Captures the exons of all annotated protein-coding genes (plus a small amount of flanking sequence including splice sites),

JKU

OUTLINE FOR TODAY

- New sequencing technologies (NGS)
- Principles of next generation sequencing technologies
- Commercial platforms
- Uses of NGS
- Individual genomes

JKU

SEQUENCE INDIVIDUAL GENOMES

- Individual human genomes: James Watson (2008) followed by the genome of Craig Venter-completed in 2 months; 1/100 of the cost of the human genome project
- Personalized medicine
- Sequencing of large cohorts (e.g. 1000 Genomes Project)

JKU

1000 GENOMES PROJECT

- 1000 genome project—sequencing 1000 genomes in 2 years..in reality it is less individuals!
- Launched in Jan 2008 finished in 2010
- International research effort
- Objective: detailed catalogue of human genetic variation

JKU

COHORT SEQUENCING

- What information is obtained by sequencing individual genomes?
- Find a gene variant that reduces cholesterol in the blood

- Inhibition of PCSK 9 reduces levels of cholesterol in the blood dramatically
- Use information on gene function for new drug development

JKU

HOW WAS PCSK9 DISCOVERED?

- Look for a extreme phenotypes with a large effect
- Cohort 1: 1000 individuals with high blood cholesterol
- Cohort 2: 1000 individuals with low blood cholesterol
- Sequence the whole genome of the 2000 individuals to find differences.
- 7 individuals with two distinct nonsense mutations in PCSK9 had almost 10% of the normal levels of cholesterol
- Identification of an individual with two dysfunctional copies of PCSK9—equivalent to a human knock-out model which was completely healthy—
- For the pharmaceutical industry this means that PCSK9 can be controlled without any side effects!

JKU

SEQUENCING INDIVIDUAL GENOMES: THE PERSONAL GENOME PROJECT

- Personal Genome Project (PGP)
- launched by Dr. George Church in 2005
- The Personal Genome Project hopes to enroll 100,000 participants from the general public who are willing to have their genomes sequenced and allow the results to be published in a massive database along with extensive information about their traits and medical history.
- It is hoped that the information provided will help scientists test hypotheses about the relationships among genes, traits, and environment.

JKU

SEQUENCING INDIVIDUAL GENOMES

- Perhaps less well known is what it takes to become a volunteer for this project.
- In order to enroll as a volunteer, potential participants must take an entrance exam that tests:
- basic genetics literacy
- informed consent expertise
- knowledge about the rights and responsibilities of human research subjects.

- That's right…you must take a test and score 100% in order to qualify for participation in the study!

JKU

JOHANNES KEPLER
UNIVERSITÄT LINZ

QUESTIONS?

JOHANNES KEPLER
UNIVERSITÄT LINZ
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at