

Next-Generation Sequencing

Methods and Computational Analysis

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

2012/10/16

Today

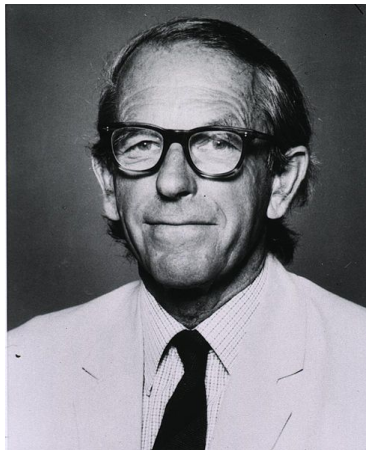
- Sanger Sequencing: Review
- Next-Generation Sequencing
 - ▶ Basic principles
 - ▶ Close look at Illumina's SBS method
 - ▶ Glance at competing methodologies
- Next-next Generation Sequencing
 - ▶ Nanopore sequencing
 - ▶ What's to come
- Computational analysis
 - ▶ Actually, this is the main topic of the course
 - ▶ We will start today with some basic formats and concepts

Outline

- 1 Sanger Sequencing
- 2 The Next Generation
- 3 NGS: File Formats and the Big Picture

Fred Sanger: $1\frac{1}{4}$ Nobel Prizes

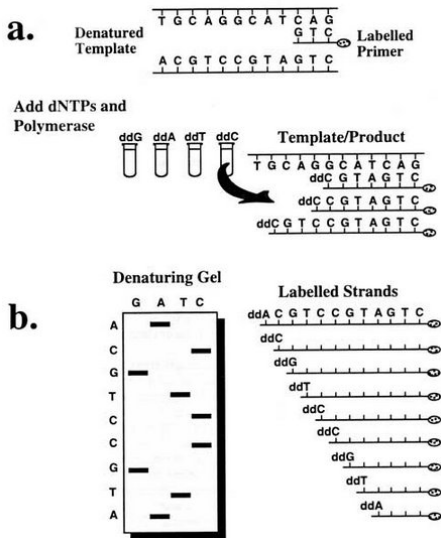
- 1958: Nobel prize in chemistry "for work on structure of proteins, especially that of insulin".
- 1980, Shared half of Nobel prize in chemistry with Walter Gilbert "for determination of base sequences in nucleic acids".
- The major method of sequencing nucleic acids until very recently



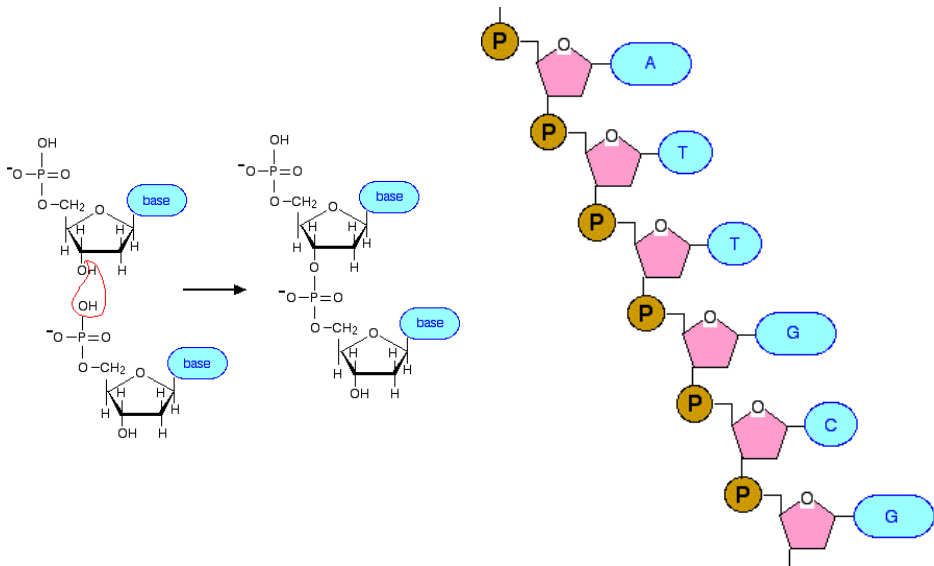
1918–
British biochemist

Sanger Sequencing

- classical chain-termination method
- Ingredients:
 - single-stranded DNA template
 - DNA primer
 - DNA polymerase
 - normal deoxynucleotidetriphosphates (dNTPs)
 - “chain-terminating”, dideoxy NTPs

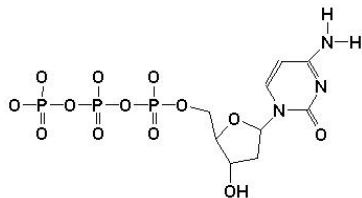


DNA Synthesis: Chain extension

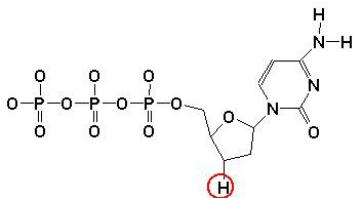


- Recall that DNA is extended from 5' to 3'

Sanger Sequencing: Key idea: dideoxy nucleotides



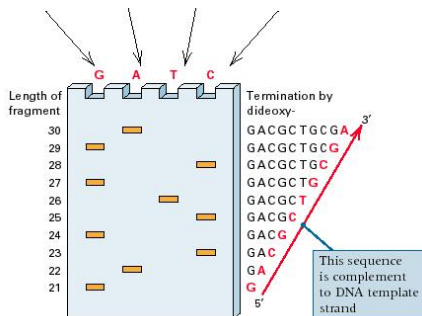
Deoxycytosine (dCTP)



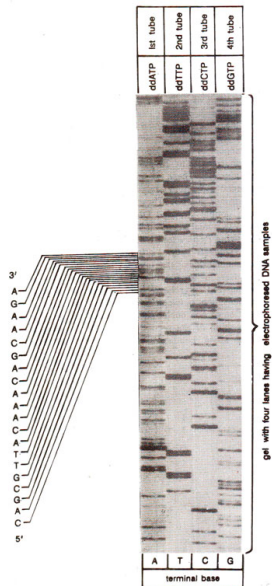
Dideoxycytosine (ddCTP)

- A dideoxy nucleotide (ddNTP) stops DNA synthesis at specific nucleotides.
- For example, if the ddCTP to the right is incorporated into a growing strand of DNA, the lack of a free 3' OH group would prevent the next nucleotide from being added and thereby **terminate chain extension**

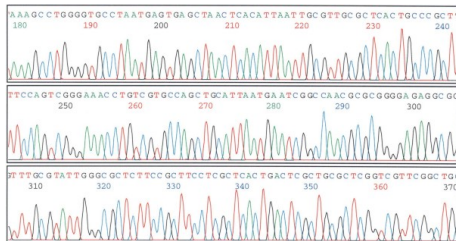
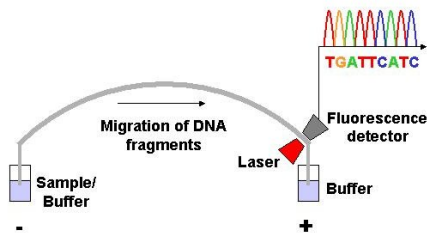
Sanger Sequencing: Radioactive



- Initially done with radioactively marked nucleotides, e.g., dATP- $[\alpha\text{-}^{33}\text{P}]$
- labeled primers, four reactions, each with different ddNTP



Sanger Sequencing: Fluorescent



- Dye-terminator sequencing
- label each of the 4 ddNTPs with different fluorescent dye (different wavelength)
- Allows one reaction rather reaction, rather than four reactions as in the labeled-primer method.
- Plot intensity of each wavelength against electrophoresis time (“chromatogram”)
- Conventional colors: A, T, C, G

Sanger Sequencing: Fluorescent

- Sanger sequencing powered the characterization of the human genome
- But: Still extremely limited in throughput
- At height of Sanger era, 400 kb per machine per day
- about 45 thousand runs needed for one human genome at 6x coverage if everything works perfectly...



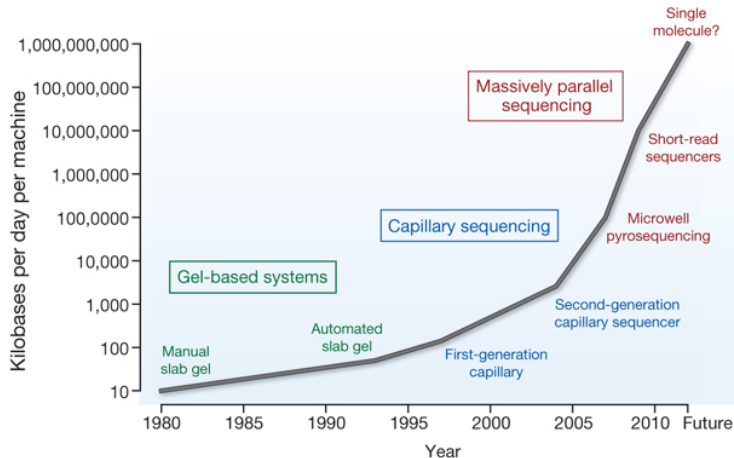
International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome

Nature **409**:860-921

Outline

- 1 Sanger Sequencing
- 2 The Next Generation**
- 3 NGS: File Formats and the Big Picture

Next-Generation Sequencing



MR Stratton et al. *Nature* **458**, 719-724 (2009)

- NGS: refers to one of a number of technologies that enable a massive parallelization of DNA sequencing

Next-Generation Sequencing



- Genbank circa 2005 – 50 Gb sequence data
- Illumina GA at 1000 Genomes project in year 2008 – 2,500 Gb
- “Each week in Sept–Oct of 2008, the 1000 Genomes Project created the equivalent of all the data in GenBank”

Thomas Keane and Jan Aerts. Tutorial 1: Working with next-generation sequencing data - A short primer on QC, alignment, and variation analysis of next-generation sequencing data. 9th European Conference on Computational Biology 26th September, 2010

Illumina Sequencing

- There are several competing NGS technologies
- At present, the technology of Illumina seems to be superior for most applications
- Four Basic Steps:

1. DNA & Library Preparation	Make random DNA fragments, append adapters
2. Chip/flowcell prep	Attach fragments to surface and amplify
3. Sequence	Massively parallel DNA sequencing
4. Analyze	This course!

- Understanding 1.–3. is essential for 4.!

Library Prep

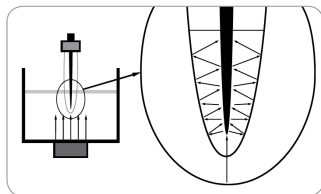
Purpose of Library Prep: Create collection of *random* DNA fragments ready to be sequenced!

- Major difference to Sanger sequencing, which requires some form of targeting the DNA: either by specific PCR primers or by cloning into sequences with universal primer binding sequences.
- Four major substeps for library prep (About 6 hours of lab work)
 - 1 Fragment DNA
 - 2 Repair ends / Add A overhang
 - 3 Ligate adapters
 - 4 Select ligated DNA

Library Prep (1): Fragmentation

- Most Illumina protocols require that DNA is fragmented to less than 800 nt.
- Ideally, fragments have uniform size
- **Sonication** uses ultrasound waves in solution to shear DNA.
- Ultrasound waves pass through the sample, expanding and contracting liquid, creating “bubbles” in a process called *cavitation*.
- Bubbles \Rightarrow focused shearing forces \Rightarrow fragment the DNA

- Sketch of sonication in “Eppi”
- Source: Bioruptor (<http://www.diagenode.com/>)



Library Prep (1): Fragmentation

- **Nebulization**: alternative method
- genomic DNA is physical sheared by repeatedly forcing input DNA as a fine mist through a small hole in the nebulizer
- The size of the fragments can be determined to some extent by
 - ① speed at which the DNA solution passes through the hole
 - ② Pressure of gas blowing through the nebulizer
 - ③ the viscosity of the solution
 - ④ the temperature

- Typical nebulizer
- Source: AIRTM DNA Fragmentation kit
(<http://www.biooscientific.com/>)

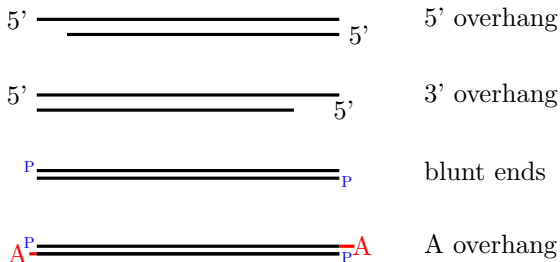


Library Prep (2): End repair

Purpose of End repair: Downstream enzymatic steps won't work unless the DNA fragments are *nice*!

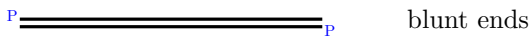
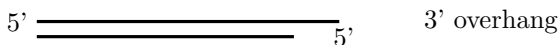
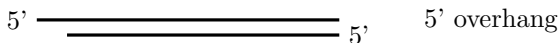
- Following sonication, the ends of the DNA must be polished so that an A-tail can be added
- The A-tail is needed for downstream steps
- DNA can come out of fragmentation procedure with 5' or 3' overhangs

Library Prep (2): End repair



- Phosphorylation of 5'-ends is strictly required for enzymatic ligation of nucleic acids (ligases need a 5'-phosphate and a 3'-OH to link two oligonucleotides by forming a phosphodiester bond).
- *T4 DNA polymerase* and *E. coli DNA polymerase I Klenow fragment*: 3' to 5' exonuclease activity removes 3' overhangs and the polymerase activity fills in the 5' overhangs.

Library Prep (2): End repair



- Add 'A' bases to the 3' End of the DNA Fragments
- Add dATP and Klenow fragment (polymerase activity)
- **purpose**: Keep DNA fragments with ligating with each other & enable ligation to adaptors (which cleverly are designed to have a 'T' -base overhang)

Library Prep (3): Adapter ligation

Purpose of Adapter ligation: Adaptors get attached (ligated) to the end-repaired DNA fragments. This enables three things

- 1 The adaptors will later bind to complementary oligos on the flowcell.
- 2 Allow PCR enrichment of adapter-ligated DNA fragments only
- 3 Allow for indexing or “barcoding” of samples

- Elegant use of primer sequences to achieve these goals.
- It is worth taking the time to understand how it all works

Library Prep (3): Adapter ligation

TruSeq Universal Adapter (P7)

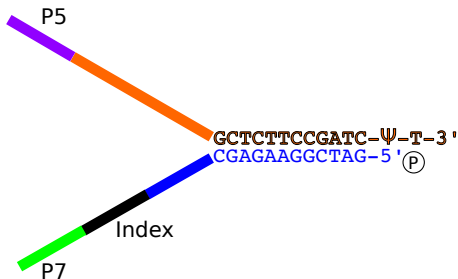
5 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3

TruSeq Indexed Adapter (P5)

5 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC–NNNNNN–ATCTCGTATGCCGTCTTCTGCTTG 3

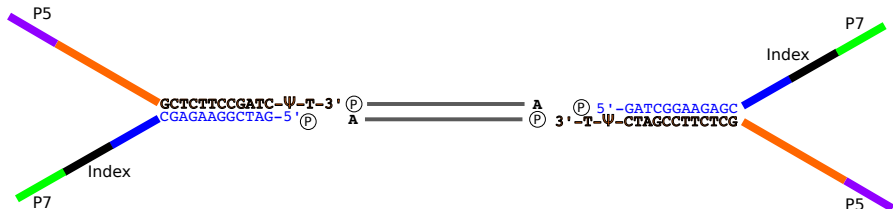
- The indexed adapter (IA) has 6 variable nucleotides (“N”), which are used for the **bar code**.
- The various segments of the primer sequences play several roles in the sequencing process

Library Prep (3): Adapter ligation



- The twelve 3' nucleotides of the Universal Adapter (UA; P7) and the twelve 5' nucleotides of the indexed adapter (IA; P5) are reverse complementary to one another and can thus **anneal**.
- Note the C and T at the very 3' end of the UA are connected by a phosphothiorate (Ψ) bond, which is resistant to exonuclease activity. This T is of course necessary to bind to the 'A' overhang of the ligated fragments

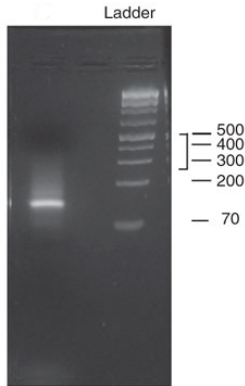
Library Prep (3): Adapter ligation



- DNA ligation: DNA ligase is an enzyme that joins DNA strands together by catalyzing the formation of a phosphodiester bond.
- Following ligation, we purify the products of the ligation reaction on a gel to remove unligated adapters, as well as any adapters that may have ligated to one another, and select a size range of sequencing library appropriate for cluster generation.

Library Prep (3): Adapter ligation

- Gel purification (cut out slice of gel)
- purpose
 - 1 Remove unligated adapters
 - 2 Remove adapters that have ligated to each other (here: band at ca. 100 bp)
 - 3 select size range appropriate for desired sequencing library
- For genomic sequencing, Illumina suggests 300–400 bp (\pm one standard deviation of about 20 bp) insert size for read lengths 2 x 75 bp
- This translates to about 3 mm gel size at 400-500bp to account for length of adapter sequences.

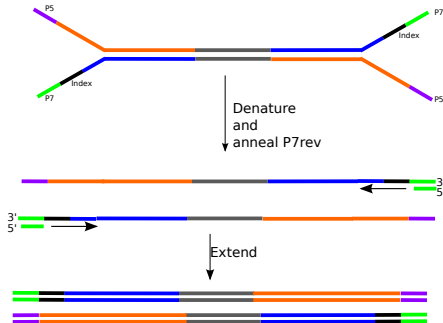


Library Prep (3): Enrichment PCR

Purpose of Enrichment PCR:

- introduce to the adapter-ligated molecules the sequences required for hybridization to flowcell.
 - selectively enrich those DNA fragments that have adapter molecules on both ends and to amplify the amount of DNA in the library
-
- PCR is performed with primers that anneal to the ends of the adapters
 - small number of PCR cycles (e.g., 10) to avoid skewing representation of the library

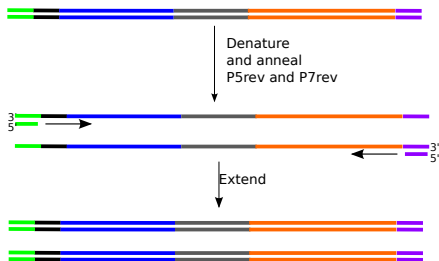
Library Prep (3): Enrichment PCR



- The primer P7 is the reverse complement of the last 24 bases of the indexed adapter (going exactly up to the index)
- 5'-CAAGCAGAAGACGGCATACGAGAT-3'

5' - (. . .) - NNNNNN - ATCTCGTATGCCGTCTTCTGCTTG - 3'
 3' - TAGAGCATACGGCAGAAGACGAAC - 5'

Library Prep (3): Enrichment PCR



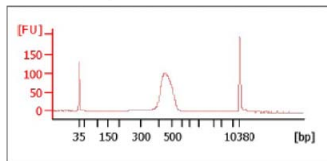
- In the remaining cycles, the primer P5 can also bind
- P5 is identical to the first 44 bp of the universal adapter, and can thus bind to its reverse complement as generated by PCR
- 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3'

5' -AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

5' -AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3'

Library Prep (3): Enrichment PCR

- Validate results to quantify amount and size distribution
- At this point, samples with different bar codes can be pooled.



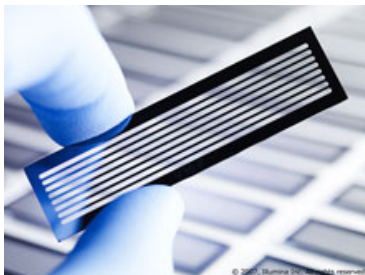
Problems for bioinformatics from this step:

- Duplicate reads (multiple PCR amplicons from same template)
- Allele bias in amplification

Flow-Cell Prep

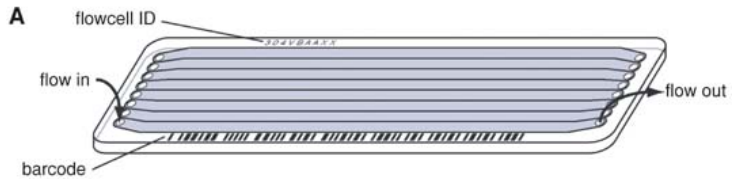
Purpose of flowcell prep:

- Bind ligated DNA fragments to flow cell
- Perform in site amplification of individual molecules to boost signal later during sequencing



Flow-Cell Prep

- A flow cell is essentially a hollow glass slide
- There are eight channels in an Illumina flowcell



Flow-Cell Prep (1): Library deposition

- Specific portions of the adapter sequences bind to oligos that are attached to the flow cell surface

Universal adapter

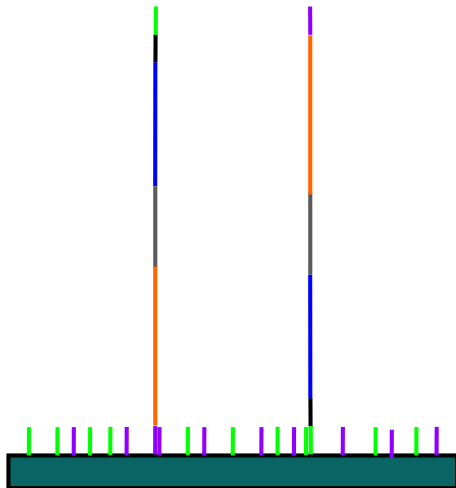
5'-AATGATACGGCGACCACCGA GATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

Indexed adapter:

5'-(...)-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG-3'

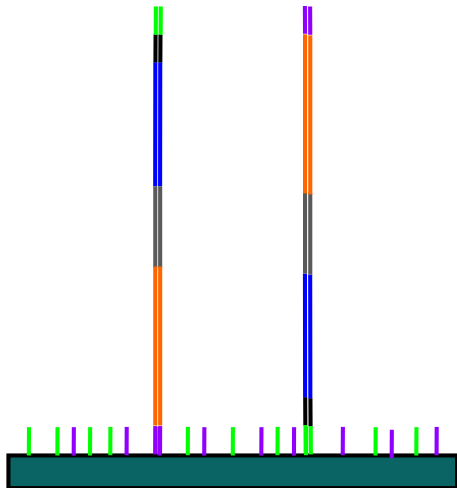
Flow-Cell Prep (1): Library deposition

- The PCR products are then denatured with NaOH
- The single stranded DNA molecules are transferred to the flowcell
- Here, they bind to oligos that are complementary to sequences in the adapters



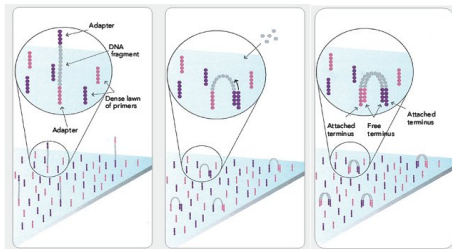
Flow-Cell Prep (1): Library deposition

- We now get a double-stranded molecule
- Formamide is added to denature these products, and the original fragment (which is not covalently attached to the flow cell) is washed away



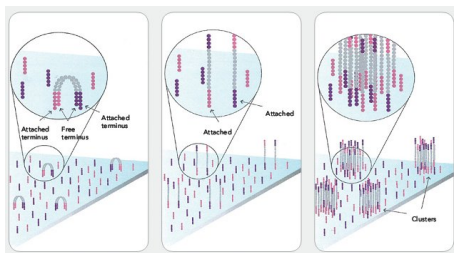
Flow-Cell Prep (2): Bridge amplification

Purpose: amplify single molecules attached to flow cell as described above by factor of ~ 1000 to generate more signal in the following sequencing steps.



- Single-stranded DNA molecule bends and hybridizes to a second oligo on flow cell surface.
- Extension is performed all the way back to the other oligo: a “bridge”

Flow-Cell Prep (2): Bridge amplification



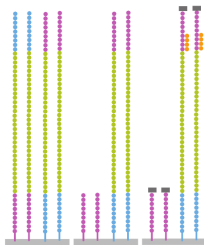
- Surface amplification is performed at constant temperature (60°C) for 35 cycles:
 - 1 formamide at 60°C ⇒ equivalent to “denaturation step” in normal PCR
 - 2 extension buffer ⇒ equivalent to “annealing step” in normal PCR
 - 3 extension mixture ⇒ equivalent to “extension step” in normal PCR

Flow-Cell Prep (3): Flow cell processing

Purpose: We now have colonies of ~ 1000 amplicons. However, they go in both directions

- Make sure only one orientation of sequence is present in each colony
- Make remaining sequences single stranded to prepare for sequencing

- Reverse strands are now cleaved to ensure sequences have only one orientation
- Free 3' ends are blocked to prevent unwanted DNA priming

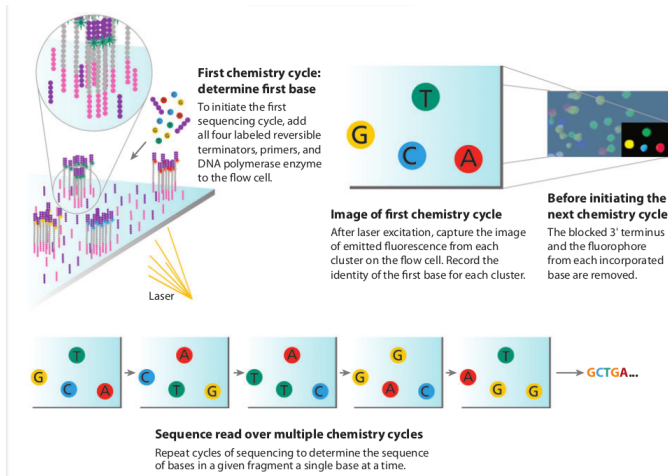


Sequencing by synthesis

- Sequencing by synthesis (SBS) is relatively simple to understand after all of the prep steps
- Basic idea:
 - ① incorporate one base at a time using four differently marked, ddNTPs
 - ② Difference to Sanger sequence: the ddNTPs have *reversible* terminators
 - ③ identify which base was incorporated in each cluster by measuring wavelength of incorporated ddNTP
 - ④ unblock the ddNTP and repeat cycle.

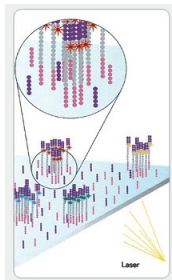
Sequencing by synthesis

- Sequencing by synthesis:



Sequencing by synthesis

- Sequence “down” towards the flow cell
- All four ddNTPs added simultaneously
- Each ddNTP has reversible chemical block of 3' OH group
- Each ddNTP has unique fluorescent label

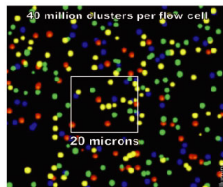


The sequencing primer is essentially the sequence of the UA
(excluding the portion that binds to the flow cell)

UA: 5' -AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
Seq-primer: 5' -TCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

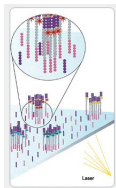
Sequencing by synthesis

- An imaging step follows each base incorporation step
- each flow cell lane is imaged in three 100-tile segments
- cluster density per tile of 30,000 (improving with new machines and chemistry)
- after each imaging step, the 3 blocking group is chemically removed and the cycle is repeated



Sequencing by synthesis: Base calling

- A base-calling algorithm assigns sequences and associated quality values to each read
- quality checking pipeline removes poor-quality sequences.



Important for bioinformatics: Quality is affected by many parameters

- PCR errors in colony amplification
- Phase errors (individual strands do not incorporate nucleotide in some cycle and then lag behind the other molecules)
- Impurities on flow cell

The quality of a base call is reflected in the **PHRED** score (vide infra).

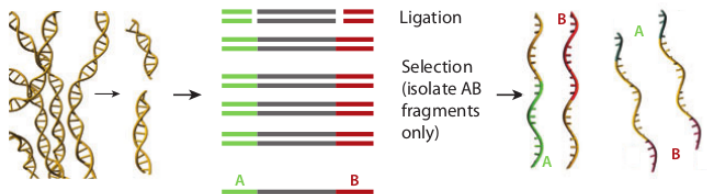
Other NGS Technologies

There are several competing NGS technologies. The main technologies currently on the market are

- Illumina
- Roche 454
- ABI SOLiD
- Life Sciences Ion Torrent
- Pacific Biosciences
- Complete Genomics

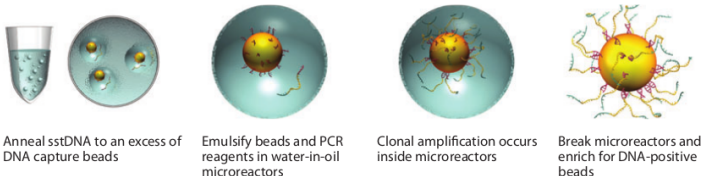
- Each technology has advantages and disadvantages
- Appropriate bioinformatics analysis needs to be adapted for each of the technologies to achieve optimal results
- We will provide a quick glance at some of the competing technologies.

Roche/454 FLX Pyrosequencer



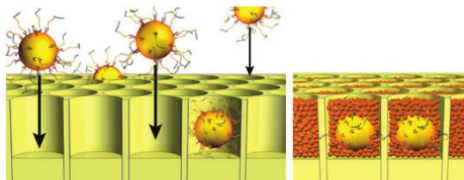
- Library prep roughly similar to that of Illumina

Roche/454 FLX Pyrosequencer



- Fragments are mixed with a population of agarose beads whose surfaces carry oligonucleotides complementary to the 454-specific adapter sequences
- Each bead is associated with (optimally) a single fragment.
- Each of these fragment:bead complexes is isolated into individual oil:water micelles and PCR is performed: “Emulsion PCR” (purpose is the same as with the bridge amplification for Illumina)

Roche/454 FLX Pyrosequencer



- Well diameter: average of 44 μm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

- The beads are arrayed into a picotiter plate
- single beads “land” in each of several hundred thousand single wells
- Sequencing by Pyrosequencing is performed.

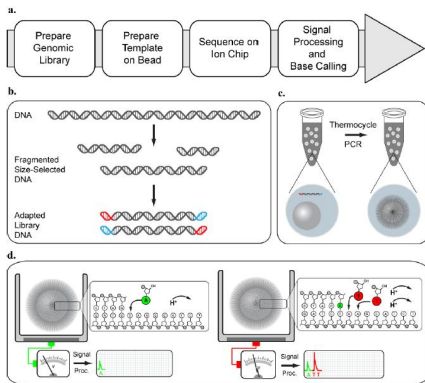
In pyrosequencing, each incorporation of a nucleotide by DNA polymerase results in the release of pyrophosphate, which initiates a series of downstream reactions that ultimately produce light by the firefly enzyme luciferase. The amount of light produced is proportional to the number of nucleotides incorporated

Roche/454 FLX Pyrosequencer

- Major advantages of Roche system
 - ▶ Long sequences (up to 800 nt) can be read
- Major disadvantages of Roche system
 - ▶ Homopolymer tract errors. For instance, the pyrosequencing signal for 7 “A”-bases is not easily distinguishable from the signal for 6 “A” bases.
 - ▶ Price. The most expensive of the major platforms currently on the market

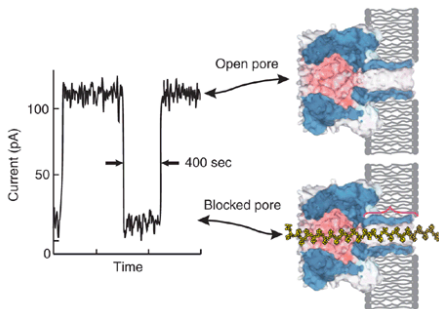
Ion Torrent

- All four (unmarked!) nucleotides cyclically flowed in an automated 2- hour run, releasing H^+ ions
- Bases are called based on H^+ signal



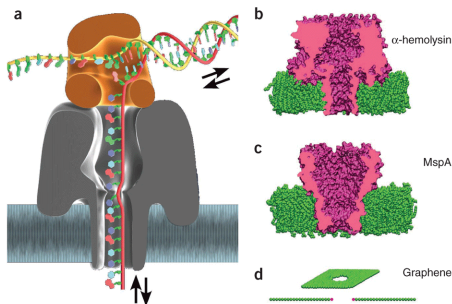
Nanopore Sequencing

- A nanopore: A hole of ~ 1 nm in diameter
- Apply voltage across pore and observe current due to conduction of ions through the nanopore
- The different bases of DNA block the hole to different extents and change the current in a characteristic way



Nanopore Sequencing

- By following changes of current over time, we can determine the sequence
- True single-molecule sequencing
- Advantages: 1) Long reads ($>10,000$ nt) 2) cheap (probably)
- Disadvantages: Probably higher error rates due to signal-to-noise issues inherent in the technology



Outline

- 1 Sanger Sequencing
- 2 The Next Generation
- 3 NGS: File Formats and the Big Picture**

Big Picture

NGS Bioinformatics Pipelines often comprise the following steps

- Q/C of fastq sequence files
- Read mapping against some reference genome
- Some analysis of the mapped reads
 - 1 variant calling (exome, genome, . . .)
 - 2 differential expression (RNA-seq)
 - 3 peak calling (ChIP-seq)
 - 4 etc.
- Visualization
- Biomedical interpretation

The practicum will be a rough and dirty walk through this procedure. You will learn to understand, refine, and extend the various steps in the rest of this course

File Formats

- There are many many different file formats that reflect the various steps of analysis
- We will introduce the major formats here
- Goal of lecture and practicum is to go through the entire process once
- Remaining lectures will fill in details and lead into biomedical analysis of NGS data

FASTQ and PHRED-like Quality Scores

- Illumina sequences are reported in FASTQ format.

```
@My-Illu:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTGTTCAACTCACAGTTT
+
!' '* ((((***+)) %%%++) (%%%) .1***-+' ')) **55CCF>>>>>CCCCCCC65
```

- 1 Read identifier
- 2 sequence reported by the machine
- 3 '+' (can optionally include a sequence description)
- 4 ASCII encoded base quality scores

FASTQ and PHRED-like Quality Scores

The read identifier from the previous slide has the following meaning:

My-Illu	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	x-coordinate of the cluster within the tile
1973	y-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (paired-end/mate-pair only)

(Note the format has changed with recent Illumina software).

PHRED Quality Scores

- The PHRED quality score is defined as

$$Q_{PHRED} = -10 \log_{10} p \quad (1)$$

where p is the probability that the corresponding base call is **wrong**.

- The PHRED quality score is nothing more than a simple transformation.

Q_{PHRED}	p	Accuracy
10	10^{-1}	90%
20	10^{-2}	99%
30	10^{-3}	99.9%
40	10^{-4}	99.99%
50	10^{-5}	99.999%

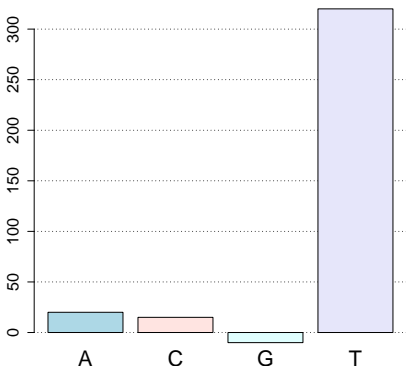
PHRED Quality Scores

- The PHRED quality scores are converted to ASCII characters, by using ASCII character number $Q_{PHRED} + 66$.

Q_{PHRED}	ASCII
2	B: Special indicator: Trim off rest of read
10	L
20	V
30	`
40	j
...	...

PHRED Quality Scores: Illumina

- The PHRED quality scores attempt to quantify
 - 1 Is the signal for the base call much brighter than the others?
 - 2 Does the spot get suspiciously dim compared to the beginning?
 - 3 Does the signal look clean in the next few cycles and the previous few cycles?



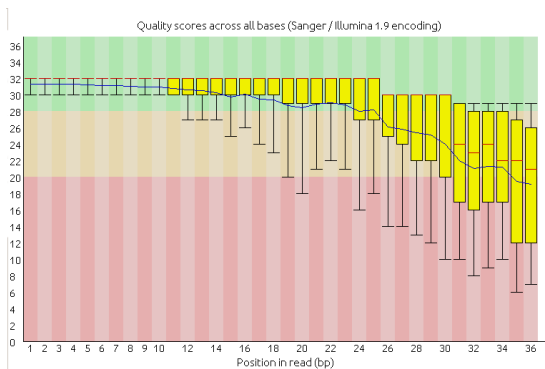
PHRED Quality Scores: Illumina

- The probability is calculated by the Illumina software on the basis of a number of heuristics.
 - 1 Penultimate chastity 25
 - 2 Approximate homopolymer length
 - 3 Intensity decay
 - 4 Signal to noise (SOWB¹)
 - 5 Shifted purity: How much of signal is accounted for by brightest channel?
- Read segment quality: The ends of some reads are unreliable. All bases in such reads are flagged with “B”, corresponding to $Q_{PHRED} = 2$. This part of the read should not be used for downstream analysis.

¹signal overlap with background

PHRED Quality Scores: Illumina

- In the practicum, we will look at some real data
- In general, Illumina data decline in quality along the length of a read (why?)
- Here is a plot showing median (read) and mean (blue) quality scores in Illumina 1G (old!) data for *E.coli*



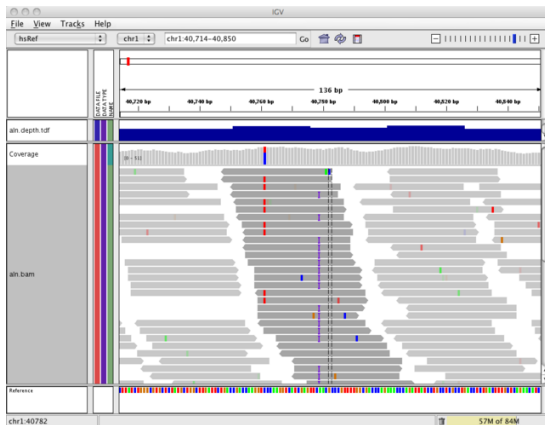
PHRED Quality Scores: Illumina

Types of quality control and procedures

- per-base quality: Trim ends if quality too low
- per sequence quality: Avg quality score should be well above 20, otherwise experiment may need to be repeated
- sequence duplication levels: A large number of identical sequences may indicate duplication by PCR during library prep, this can bias estimates of mRNA expression in RNA-seq and is often filtered out
- Overrepresented sequences: Sometimes adapter sequences are represented in final sequences and these need to be filtered out before downstream analysis

Read alignment

- Read alignment will be covered in detail in the following lectures
- Purpose: find out where in the genome a short read comes from...



- For now we will examine SAM/BAM formats, in the practical you will have a closer look

SAM (Sequence Alignment/Map) format

A generic format for storing large nucleotide sequence alignments.

- One header section
 - ▶ Lines begin with '@'
- One alignment section
 - ▶ Lines do not begin with '@'

A **BAM file** is essentially a binary (\sim gzip-compressed) version of a SAM file.

SAM/BAM files are usually sorted and indexed to streamline data processing

SAM: header

- Each line of header begins with '@' followed by a type code
- Except for @CO (comment) lines, each data field is TAG:VALUE
- The header we will produce in the practical is as follows:

```
@HD      VN:1.0   SO:unsorted
@SQ      SN:gi|254160123|ref|NC_012967.1|      LN:4629812
@PG      ID:Bowtie      VN:0.12.8      CL:" (...) "
```

- VN (version of SAM format): 1.0
- SO: sorting order of alignments
- SQ: reference sequence dictionary (here: the E coli genome)
- LN: reference sequence length (4.6 million nt)
- PG: program with version (VN) and command line (CL) arguments

SAM: alignment

- All remaining lines are alignment lines, one read per line
- Each alignment line has 11 mandatory fields, representing name, position, quality, the actual alignment, and the relationship of an aligned sequence to its paired read (if any)
- Perhaps the most interesting field is the CIGAR field, a way of concisely representing the alignment
- A set of codes describes the operations performed to obtain an alignment
 - ▶ M: match
 - ▶ I: insertion
 - ▶ D: deletion
 - ▶ etc.
- For instance, 36M refers to an alignment in which all 36 positions match

SAM: alignment

Consider the following alignment between a reference (R) and query (Q) sequence

```
AGCATGTTAGATAA--GATAGCTGG   R
      |||||      |||  |||
-----TTAGATAAAGGATA-CTGG   Q
```

- 8 matches
- insertion of 2 nucleotides in the query
- four additional matches
- deletion of one base in the query
- four more matches

We can represent this as

8M2I4M1D4M

Variant calling

The above procedures have basically aligned to reads to a reference genome.

In the practical, we will see that the SAM file is converted to a BAM file, sorted, and indexed

- In essence, this leaves us with a data structure representing the alignments of the short reads with the reference genome

A screenshot of a variant caller interface, likely from a genome browser. The top part shows a reference genome sequence with positions 150, 160, 170, 180, 190, 200, 210, and 220 marked. Below the reference, a read group named "ZAZ0DCYWK" is selected. Multiple short reads are displayed, each with a red bar above it indicating a variant. The variant is located at position 190, where the reference has a 'G' and the reads have a 'C'. The variant is highlighted in red. The interface also includes a "Show Bases" button and a small icon.

Variant calling: VCF File

- Following some further analysis with programs such as samtools, we generate a variant call file
- VCF files contain variants (deviations from the reference genome)
- In general, only the variants are represented
- We also have information about the number of reads supporting the variant call and its overall quality

Variant calling: VCF File

Each line in the VCF file represents a single variant. We will discuss this in more detail in the practical, but here are some of the fields of a typical line

Field	Meaning	Example
CHROM	id from ref genome	gi 254160123 ref NC_012967.1
POS	reference position,	161041
REF	ref sequence	T
ALT	variant	G
QUAL	PHRED quality score	179
FILTER	additional information	see next slide

Thus in this case the variant call is of high quality (PHRED 179)-

Variant calling: VCF File

The **info field** contains some more detailed information about the data leading to the call,

```
DP=60;VDB=0.0839;AF1=1;AC1=2;DP4=0,0,24,32;MQ=20;FQ=-196
```

Field	Meaning	Example
DP	# high-quality bases	60
AF1	max. likelihood estimate of first ALT allele	1.0
DP4	# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases	0,0,24,32
MQ	Root-mean-square mapping quality of covering reads	20

Variant calling: VCF File

Finally, the **Genotype fields** specify the genotype information for each sample

```
GT:PL:GQ 1/1:212,169,0:99
```

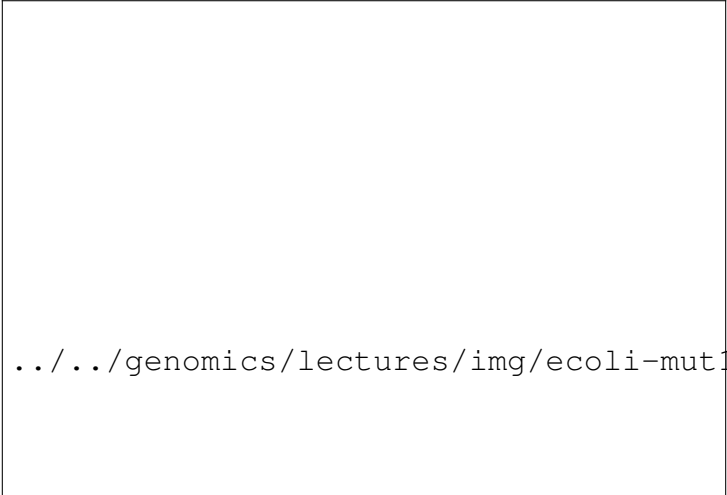
There is first a FORMAT field that specifies the data types and order

- 1 GT genotype. 0: reference, 1: first allele listed in ALT, 2: second allele (if applicable)
- 2 PL: List of Phred-scaled genotype likelihoods for AA,AB,BB genotypes where A=ref and B=alt;
- 3 GQ PHRED coded genotype quality

Variant calling: VCF File

A high-quality variant looks like this

```
gi|254160123|ref|NC_012967.1| 555154 . A C 151 .\  
DP=27;VDB=0.1016;AF1=1;AC1=2;DP4=0,0,7,17;MQ=20;FQ=-99 GT:PL:GQ \  
1/1:184,72,0:99
```



../../genomics/lectures/img/ecoli-mut1.png

Variant calling: VCF File

A lower-quality variant looks like this

```
gi|254160123|ref|NC_012967.1| 555660 . T C 25\  
DP=3;VDB=0.0158;AF1=1;AC1=2;DP4=0,0,1,2;MQ=20;FQ=-36 GT:PL:GQ\  
1/1:57,9,0:15
```

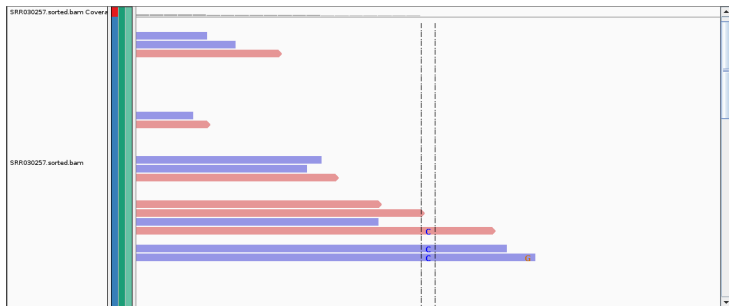


Figure: Visualization of mutation at position 555660, ref=T, alt=C, DP4=0,0,1,2.

Contact Info

Peter N Robinson

Professor of Medical Genomics

Institut für Medizinische Genetik und Humangenetik

Charité - Universitätsmedizin Berlin

Augustenburger Platz 1

13353 Berlin

030 450566006

peter.robinson@charite.de

<http://compbio.charite.de>

<http://www.human-phenotype-ontology.org>

Office hours by appointment (send email or call up)