Next Generation Sequencing
for
DUMMIES

(Version March 14th 2018)

Looking at a presentation without the explanation from the author is sometimes difficult to understand. This document contains extra information for some slides that need it. Making a presentation, showing it once and keeping it for myself is a waste of time. That is why the presentation is available on the internet. I hope you find it informative.
The presentation can be found on my webpage:
http://users.ugent.be/~avierstr/
andy.vierstraete@ugent.be
Enjoy, Andy

Slide 1:

Choice of the title: I do not mean that you are dummies when we are talking about Next Generation Sequencing. Maybe you are, maybe you are not. What I mean is that I still find myself a dummy when talking about Next Generation Sequencing. Even after 4 years trying to keep up with the latest developments, I still have the feeling that I only know the "tip of the iceberg".
With this presentation, I'm trying to give you the basic principles and work flow, so if you know what is happening, you can understand what can go wrong. Once you know the basics, it is easier to read more detailed publications.
Presentation available at this link.

Slide 2:

The topics I will talk about. The *Workflow* and *Different platforms* will be explained simultaneously.

Slide 3-4:

(One of) the first publication(s) of a 11 bp sequence from Bacteriophage MS2. Revolutionary in that time, background noise in the sequencing world of today. A reference in that publication from 1961 identified the last base of the sequence to be "A".

Slide 5-6:

Some more milestones in the DNA research history.

Slide 7:

A list of the most commonly used platforms. The ones in gray are not yet available, but according to the manufacturers, they will be available within the next year(s).
Purple: Next Generation Systems: need an amplification step.
Red: Third Generation Systems: single molecule sequencing.

Slide 8:

First explain the **Next Generation Sequencers** (purple list on slide 6):

All systems require a library preparation.

Slide 9:

Library preparation:
- gDNA (or RNA) is fragmented randomly (e.g. by sonication, enzymatic, ...).
- Adapter and primer are ligated to the fragments (I use the term “primer” for the part that will be used when sequencing is performed, and the term “adapter” for the part that will anneal to the bead or the slide. Actually they are both primer sequences). Notice that the ligation is random, some fragments will contain adapters on both sides, some will contain primers on both sides, the good ones are those with an adapter on one side, and a primer on the other side. A barcode (or MID (=Multiplex IDentifier)) can be included.
- Some platforms do a size selection, so that only fragments with the ideal length are used.

Slide 10:

There are two types of amplification: Emulsion PCR or “Polony” PCR on a slide

Slide 11 -15:

Emulsion PCR is created by mixing oil and water. The result are little droplets (micro reactors). This is also a random reaction, so not all droplets contain all needed material: one bead, one DNA strand, primers and PCR mix. The bead contains the complementary sequence of the adapter on the DNA strand.

Slide 12:

Denaturation of the DNA: double strand becomes single strand.

Slide 13:

One (reverse) strand anneals to the adapter on the bead and the polymerase amplifies the strand starting from the bead (light blue) towards the primer side (green) (the adapter on the bead has the same function as a primer).
(At the same time, it is possible that the complement (forward) strand is amplified in the other direction: starting from the primer side (green) and going to the dark blue side of the strand. (not shown in this picture, shown in slide 15))

Slide 14:

Denaturation: double strand becomes single strand. The reverse strand that annealed in slide 13 detaches from the bead, but the amplified complement (forward strand) is physically attached to the adapter by the sugar phosphate backbone of DNA.
Annealing: the reverse strand anneals to an other complementary adapter on the bead. On the forward strand (which is attached to the bead), the green primer anneals to the primer side of the strand.

Slide 15:

Extension: polymerase amplifies the forward strand starting from the bead to the primer side, the reverse strand is amplified starting from the primer towards the bead.
The process 4 – 5 – 6 is repeated for 30-60 cycles.

Slide 16:

Only 15 % of the droplets are ideal for emulsion PCR. Left two columns: before PCR, right two columns: after PCR. After PCR, some droplets will only contain a bead (top left), some will be ideal: one bead, one DNA strand (top right), some will only contain DNA (middle left), some will contain two beads and one DNA strand: these will result in the exact same sequence (middle right), some are empty (bottom left), some will contain two or more DNA strands and one bead: these are polyclonal and will result in double signal when sequenced (bottom right).

Slide 17:

the other amplification method: “Polony” PCR on a slide for Illumina and SOLiD.

Slide 18 -19:

Bridge amplification for Illumina.
- The slide contains the complementary sequence of the primer and adapter.
- ssDNA anneals to a primer attached to the surface, the polymerase amplifies the strand. The amplified strand is attached to the surface.
- The original ssDNA strand is washed away.
- The other end of the DNA anneals to the complementary primer on the surface.
- Extension starts and a double stranded bridge is formed.
- After denaturation: each strand is attached to the surface (one strand with adapter side, other strand with primer side).
- Amplification goes on until clusters or polonies (PCR colonies) are formed.

Slide 20:

Second form of “polony” PCR for SOLiD

Slide 21:

Wildfire amplification for SOLiD (older SOLiD systems still use Emulsion PCR)

- ssDNA anneals to the complementary primer site on the surface.
- Extension of the template. The formed fragment is attached to the primer on the surface.
- Partial denaturation: strand “walks” to nearby primersite.
- Extension and strand displacement: two strands become four strands...

Slide 22 - 23:

SOLiD sequencer: sequencing by ligation: table with information about the systems.

Slide 24:

Sequencing by ligation:

SOLiD uses probes based on 2-base encoding. This means that 2 bases have to fit to the template to be attached. Four colors are used, and per color four dual bases are possible. Left bottom picture: e.g. a green color can be CA (1st base C and 2nd base A) or AC or TG or GT.

Slide 25-47: Sequencing by ligation step by step.

Slide 25:

All the probes are present (4 colors x 4 combinations), ligase, primer and the template attached to the bead.

Slide 26:

The primer anneals to the adapter. This is the 1st run.

Slide 27:

The probe with the complementary dual bases anneals and the ligase attach the probe to the primer.

Slide 28:

A laser excites the fluorescent label, and fluorescence is detected by the camera.

Slide 29:

There are four different combinations for the first two bases as shown above the two green XX. We can not know which one it is right now.

Slide 30:

The last three bases with the fluorescent label is cleaved off.

Slide 31 - 35:

The ligation - excitation - detection and cleavage goes on until 75 bp are reached.

Slide 36:

The extended sequence is melted off, and an new primer anneals one position to the left in comparison with the first primer. This is the 2nd run.

Slide 37:

The ligation - excitation - detection and cleavage starts again.

Slide 38:

The first of the dual bases (red XX) of the probe anneals to the last base of the P1 adapter (green) on the bead. Since this first base is known (e.g. T), the first (RED) base is the complement: A.

Slide 39:

Now we know the first base of the dual base, and there is only one possibility for the second base: T. If we look at the four green possibilities from the first run just above, we know that the first base from the green probe was a T (corresponds with the last base of the red probe just below). There is only one possibility for the second green base: G.

Slide 40:

The ligation - excitation - detection and cleavage goes on until 75 bp are reached. The rest of the bases are still unknown.

Slide 41:

The extended sequence is melted off, and an new primer anneals two positions to the left in comparison with the first primer. This is the 3th run.

Side 42 - 43:

The ligation - excitation - detection and cleavage goes on until 75 bp are reached. The rest of the bases are still unknown.

Slide 44:

The extended sequence is melted off, and an new primer anneals three positions to the left in comparison with the first primer. This is the 4th run.

Side 45:

The ligation - excitation - detection and cleavage goes on until 75 bp are reached. The rest of the bases are still unknown.

Slide 46:

The extended sequence is melted off, and an new primer anneals four positions to the left in comparison with the first primer. This is the 5th run.

Slide 47:

All the bases have been sequenced twice now, and we can decode the complete sequence in colorspace.
We know the sequence from slide 39 (last base from the adapter and 2 bases from the unknown sequence: ATG.
That G was the last base from the green encoding and corresponds to the first base of the red encoding (follow the purple circles column after column, you must find the same base twice in each column).

Slide 48:

For higher accuracy, a next run with 3-base encoding probes is possible.
The table shows which base is read in which run.

Slide 49:

Next system: Illumina with reversible terminator sequencing.

Slide 50 – 52:

The sequencers from Illumina with information about the systems. Only the longest read lengths per system are shown.

Slide 53:

Left drawing: after the bridge amplification, the template is denatured (becomes single stranded), and a primer is annealed .

Right drawing: the four nucleotides, each with a different dye are present, and the one which is complementary to the template is incorporated. A wash step removes the unincorporated nucleotides, and a color image is taken. The terminator group and color dye is cleaved off and are washed away. A next incorporation step is started.
The advantage of “reversible terminator sequencing” is that homopolymer regions can be read exactly. E.g. if there are 12 T's in a row, it will take 12 steps to detect the T's (NOT 1 step to incorporate 12T's and have a signal that is 12 times higher than normal).

Left bottom image: each “polony” has a fixed place on the slide and is imaged after each incorporation.

In 4-channel SBS (sequencing by synthesis) , the acquisition of 4 distinct images enables a cycle-by-cycle observation of which color dye is incorporated into an individual cluster. Cluster detection software algorithms then process the images to determine the individual base calls for each unique cluster. With 4-channel sequencing, all 4 images are required to build up the DNA sequence. The MiSeq and HiSeq Series Systems currently use 4-channel SBS.

Two-Channel SBS (sequencing by synthesis) Imaging. Accelerated detection of all 4 DNA bases is performed on the MiniSeq and NextSeq Series Systems using only 2 images to capture red and green filter wavelength bands. A bases will be present in both images (yellow cluster), C bases in red only, T bases in green only, and G bases in neither.

Slide 54:

movie that explains Illumina sequencing:
https://youtu.be/fCd6B5HRaZ8

Slide 55:

Next system: Ion Torrent with semiconductor sequencing.

Slide 56:

The Ion Torrent sequencers and information about them.

Slide 57:

Sequencing on IonTorrent: The enriched beads with DNA are in a tube in buffer with primer. DNA on the bead is single stranded. Denaturation (95°C) opens possible secundary structures.

At 37°C the primer anneals to the DNA. Polymerase is added and attaches to the position where the primer is annealed (polymerase needs a double strand to start from. The primer annealed to the DNA forms a short piece of dsDNA). The beads are loaded on the sequencing chip and a well can only contain one bead. Sequencing happens from the primer towards the bead.

Slide 58:

The PGM measures a pH-change when a base is incorporated. On the left picture, when the T is attached to the C, a diphosphate and a hydrogen is released. This hydrogen is causing the pH-change.

Right top picture: the chip contains wells which can hold one bead with DNA. The layers below are responsible for detecting the pH-change.

Right bottom picture: the time frame in which a pH-change is measured (0,5 sec) and going back to baseline (1,5 sec).

Slide 59:

Picture of the wells in a 318 chip

Slide 60:

The four nucleotides (not labeled) flow sequentially over the chip. When the base is complementary to the template, it is incorporated and a hydrogen is released and detected.

Slide 61:

When there is no match in nucleotide, ho hydrogen is released and no pH change detected.

Slide 62:

When two or more bases are incorporated, the pH-change is two or more times higher and is detected as such.

Slide 63:

An example of an Ionogram. The key sequence (first 4 artificial bases just after the primer) is used to normalize the signal. The instrument knows that only one base is incorporated, so the pH-change corresponds to one base. If the signal intensity is higher after the first four bases, than this means that more bases are incorporated at the same time. The colored bases just above the x-axis are the nucleotide flows. The bases below the figure is the detected sequence.

Slide 64:

For maximizing sequencing yield, a size selection is done. This way, fragments with the ideal length are selected for sequencing (in this case: fragments of 200 bp are for the 100 bp sequencing because they also contain the adapter and primer sequence on the sides.) Shorter fragments have to less information, longer fragments have the change not getting completely amplified in the emulsion PCR.

Slide 65:

The sheared and ligated DNA runs over a gel with open wells halfway the gel. When the fragments of the desired length are passing through that well, it is pipetted out and used in the emulsion PCR.

Slide 66:

Left picture: the emulsion PCR machine from Ion Torrent: Ion OneTouch. The left tube contains emulsion oil, the right tube contains detergent (SDS) to breakup the emulsion after the PCR. Below the disc on the right are two tubes in a small centrifuge. The emulsion PCR mix is recovered in those tubes after amplification.

Bottom middle picture: the tube contains the emulsion PCR mix (lower half) and emulsion oil (top half). On top is a “filter” with three openings. This is slowly inverted and placed on the emulsion PCR machine. When the mix is pressed through the filter, emulsion droplets are formed.

Right top picture: the PCR plate. On the back top right side there are two tubes (in- and outlet). This plate is positioned between two heating plates in the machine (groove behind the iron bar). The inside of the plate is one long fine tube. First, the emulsion mix goes from right to left in the top zone of the plate (denaturation). When it reaches the left side, it is going from left to the middle, back to the left, and so on until it reaches the bottom of the plate. Then it is going all the way to the right, back to the middle and goes up that way until it reaches the exit point (yellow tube). down and up (95° - 64°) 60 times until it reaches the right top side outlet of the plate. When migrating trough the plate, it passes zones at 95°C (denaturation) and zones at 65°C (annealing and extension).

Right bottom picture: The Ion OneTouch ES: enrichment system to select the beads that contain DNA (see slide 76).

Slide 67:

To maximize the sequencing yield, an enrichment is done for the beads that contain DNA. This is done with magnetic beads.
The primer contains a biotin label. Magnetic Streptavidin beads are added tot the pcr-beads from the emulsion PCR, and they bind to the biotin. This is immobilized with a magnet. Only beads that contain DNA are immobilized by the magnet. A wash step removes the beads without DNA. Denaturation with NaOH makes the DNA on the bead single stranded and those beads are collected for sequencing.

Slide 68:

Initialization of the sequencer. Since the systems is detecting pH changes, it has to be prepared for that. In front of the machine: 4 tubes with the different dNTP's. On the side, a W1 bottle with 350 µl 100mM NaOH, a W2 bottle with 2 liter W2 solution ((freshly prepared milliQ water) contains a pcr solution), and a W3 bottle with 50 ml buffer with stable pH.

1. The system measures the ph of the buffer W3. This is the reference pH.
2. It measures the pH of W2 solution. This is not the correct pH. It pumps an amount of W2 in the W1 bottle and mixes it by bubbling argon through the solution.
3. It pumps a volume of W1 solution in the W2 bottle and mixes with argon. It measures the pH again. If the pH is not correct, the system repeats step 3 until the pH is correct.

4. When the pH of the W2 solution is ok, the 4 tubes with the dNTP's are filled with W2 solution and mixed with argon gas. The pH of the four tubes is measured and should be the same as the W2 solution.
5. The system is ready to load a chip to be sequenced within 10 hours time (after that time, the pH can change by CO2 dissolving in the solutions).

Slide 69:

Movie time
https://youtu.be/WYBzbxIfuKs

Slide 70:

Just to give you an idea about the amount of GB that is produced during sequencing. These are just small sequencing chips with limited capacity… Imagine how much intermediate data is produced on high capacity sequencers.

Slide 71:

Systems marked with red color: Single Molecule Sequencers

Slide 72:

Advantages of Single Molecule Sequencing.

Slide 73:

The Pacific Biosciences systems with information.

Slide 74:

The accuracy is low for the raw reads. Because the errors are random, when a sequence is read several times, the consensus sequence has a high accuracy (high QV value).

High consensus reads can be made by:
- Circular Consensus Sequencing (CCS): one sequence read several times
- Standard sequencing: several reads of the same gene that are sequenced once

Bottom of page: assembly (not from PacBio reads) to explain the consensus sequence with high accuracy. The black sequence is the consensus. Green sequences are forward reads, red sequences are reverse reads. At some positions in the assembly, you can see bases with a colored background. These are sequencing errors and can be seen as random errors. If you look at the first 'wrong' A in the first red sequence, you see that the sequence has an A on that position, but the 6 other sequences above and below don't have an A in that position. So the consensus sequence on top will not have an A in that position and as a result, the consensus will have a high accuracy because the errors in the raw reads are filtered out.

Slide 75:

Library prep for Pac Bio. After fragmentation of the DNA, a loop adapter is ligated to the ends of the DNA fragments so the fragment becomes circular.

Slide 76:

A primer and polymerase is attached in the loop. This way, the froward and reverse strand are read after each other.

Slide 77:

The polymerase is attached to the bottom of a transparent well. The four nucleotides with different labels are present in the solution. When a nucleotide is incorporated it is illuminated and the emission light is detected.

Slide 78:

Movie time
https://www.youtube.com/watch?v=v8p4ph2MAvI

Slide 79:

Oxford Nanopore: sytems and information.

From the latest information (March 2017), it seems that they can reach 24 GB of data on the MinIon by a software update. They found out that the adapter complex could jam the prore, preventing the pore to be used again. By reversing the voltage in that pore, the adapter complex is ejected and the pore is available for more sequencing.

The ONT MinION permits the simultaneous quantification of short and long transcripts. On the contrary, the design of the PacBio RS II sequencing flow cell (SMRT Cells) has the disadvantage of biasing towards sequencing short cDNA molecules with the "diffusion loading method" or towards sequencing cDNA molecules longer than 700 bp with the "MagBead loading method". This can considerably affect the absolute quantification of genes and isoforms as the abundance of short transcripts relative to the long ones cannot be accurately determined.
(Spyros Oikonomopoulos et al. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations Scientific Reports 6, Article number: 31602 (2016))

Slide 80-81:

A current is measured in a nanopore (small hole). When something goes through that hole, the current changes. Each of the four bases cause a different change in current when passing through the hole. A protein (to read RNA, DNA, protein, ...) can be attached on top of that nanopore.

Slide 82:

The 2D kit is discontinued. The replacement kit is a $1D^2$ kit. There is no longer a loop at one end of the double strand, but both strands contain its own motor-adapter (helicase is the motor). When one strand is sequenced, the other strand can enter the pore and be sequenced.

Slide 83:

Movie time:
https://youtu.be/3UHw22hBpAk

Slide 84:

SeqLL (Sequence the Lower Limit): system and information.

tSMS sequencer (True Single Molecule Sequencing) will probably become available in the second half of 2017. The applications will be mostly for clinical research. The technology is based on the Helicos system: Sequencing by Synthesis.

Slide 85:

Library prep for the tSMS sequencer

Slide 86:

The four nucleotides (dye labeled) flow sequentially over the system. When a base is complementary to the template, it is incorporated. Unincorporated nucleotides are washed away, an image is taken. Dye and inhibiting group are cleaved off and washed away. Next cycle with other dNTP is performed.
Right top picture : when a base is incorporated, light is detected on that spot. When no incorporation is done, no light on that spot.

Slide 87:

Movie time:
https://youtu.be/s4UXK8vFhAY

Slide 88:

What sequencer to choose for your project ?

Don't let your choice depend on what system is available in your facility. All systems have pro's and con's.

1. Do you have long homopolymer (more than 7-8 times the same base) regions that needs to be sequenced exactly ?
2. Do you need your results fast ? (e.g. patient infected, need fast results for medication).
3. What amount of data are you expecting ? e.g. if you want to sequence the genome of a bird (1Gb) with a coverage of 15 x . You are expecting 15Gb of data. You can barcode samples and run them together on a system.
4. For de novo sequencing, it is better to have longer reads.
5. For museum samples, or forensic material, better to use systems that don't use amplification.
6. If you have enough money, you can combine systems.
7. ...

Compare prices of sequencing providers. Some are listed on:
https://genohub.com
https://www.scienceexchange.com/browse?category=ngs

Slide 89-90:

Homopolymer (several repeats of the same base in a row) problems: Ion Torrent has problems with that. Consensus sequence can solve the problem (within limits).

Slide 91-92:

Homopolymer on a Illumina HiSeq: no problems with T, seems to have problems with G. Consensus sequence can solve the problem.

Slide 93-94:

Homopolymer on a Illumina MiSeq: sometimes perfect, sometimes not. Consensus sequence can solve the problem.

Slide 95:

Quality scores in sequencing. Q-values or Phred scores are used to indicate the probability that the called base is correct. Each base in a sequence has a quality score. Examples:
Q10: 90,0% chance that the base is correct
Q30: 99,9% chance that the base is correct

Q20 and above are generally accepted as reliable.

If fragments are sequenced several times (more coverage), the consensus sequence will have a reduced number of errors, and as a result it will have a higher Q-value than the raw sequences.

For each sequencing run the average sequencing value can be calculated (slide 96-102).

Slide 96:

Ion Torrent run: first 300 bases have an average quality of 28-30. Between 300-400 the average Q is around 28. The error bars show that there are also positions in the sequences with higher and lower quality (between 14 and 34).

Maximum readlength around 400 bp

Slide 97:

MiSeq forward run: higher quality scores than IonTorrent, smaller error bars.

Maximum readlength around 250 bp

Slide 98:

MiSeq reverse run: higher quality scores than IonTorrent, but dropping fast around 200 bp !

Maximum readlength around 250 bp, probably a lot of reads will be trimmed by quality check.

Slide 99:

HiSeq run: higher quality scores than IonTorrent, smaller error bars.

Maximum readlength around 100 bp (in this example)

Slide 100:

SOLiD run: (raw data downloaded from internet) large errror bars, high and low quality ? I have no confirmation if this is a 'bad run' example or typical for SOLiD sequencing !!

Maximum readlength around 50 bp (in this example)

Slide 101:

PacBio run: much lower quality than IonTorrent and Illumina on the raw reads.
Maximum readlength around 4400 bp (in this example)

Slide 102:

MinIon run: much lower quality than IonTorrent and Illumina on the raw reads. Comparable with PacBio.

Maximum readlength around 32.000 bp (in this example)

Slide 103-104:

Quality scores in Sanger Sequencing.

Slide 104:

Bases with a red line below have a Q-value lower than 20 and this is generally accepted as less reliable. With 15 years of experience in Sanger sequencing, I don't agree with that. But Sanger sequencing has the advantage that you still can manually check the sequences.

Slide 105:

Examples of raw reads with the quality scores. On some positions in the reads, I have placed the Q-values.
Ion Torrent goes up and down.
Illumina is quite stable but has some low drops in qualtiy on some positions.
SOLiD is also jumping up and down ?? (from the same dataset as in slide 100 ! Is this normal for SOLiD ?)

Slide 106:

Other examples from Ion Torrent and Illumina. These are more 'stable' reads. Also compare the difference in read length and Quality scores for that length. Illumina HiSeq is reading 100 bases with high quality scores, Ion Torrent has longer reads with high scores.

Slide 107:

A 'stable' read from Illumina MiSeq.
A read from MinIon: lower Q-values, and they are jumping up and down.

Slide 108 – 109:

How important are quality scores ?
It is generally accepted that high quality scores guarantee correct basecalling and that lower

quality scores have higher chances of calling an incorrect base. Everything above Q20 (99%) is accepted as 'good'.

This is my personal opinion:
I don't agree with the fact that high Q-values give you the guarantee that the bases are correct. By taking a closer look at some reads, I noticed that there are random errors in assemblies. That could have 2 reasons: assembling contamination to my sample, or having errors in the reads.

Slide 108:

An Ion Torrent assembly.
The sample contains amplicons from Caenorhabiditis elegans. I downloaded the 18S gene from NCBI and mapped the reads to that reference gene. In a certain zone, I had a coverage of 1700 reads. I selected the first 10 reads and blasted those. A few returned contaminants, the rest returned C. elegans 18S gene. I selected 5 that were from C. elegans and assembled them. The result is shown in this slide.

Top half: assembly from base 40 till 140. First error is a G at position 40. When I look at the raw data, the G has a Q34 which means the basecaller was 99,96% sure that it was a G. But that G should be an A !!! At positon 41 in the second sequence, there is an A with Quality score of 32 that should be a G !!! Same problem at position 137.

Bottom half: assembly from approximately base 270 until 325 (quality normally decreases to the end of the read). A G with a Q15: acceptable that it is wrong. An A with a Q30 that is wrong !! A C with Q26 that is wrong !! The quality drops after bases 303, so that is acceptable that some of those bases are wrong.

My point here is that in a conserved gene like 18S, in the first 5 random reads, there are already more that 5 errors in the first 300 bases that have a high Q-value and are wrong.

What is the point than of the high quality scores ????
Maybe some will say: Ion Torrent is less reliable… That is why I tried it also on MiSeq reads from Illumina.

Slide 109:

A MiSeq assembly:
I used the same procedure. Species is Halomonhystera sp. Downloaded 18S from NCBI, mapped the reads to that reference, blasted the first 10 sequences, removed contaminant sequences and assembled the good ones. Result is shown in this slide.

The 2 first errors A and C have a Q40 (99,99% sure it is correct) and still they are wrong !!! The following T and G have Q38 and Q39 and are also wrong. At position 134, a T with Q38 is wrong. Also in the bottom part of the slide : G, T, G with Q36, Q36 and Q25 are wrong…

So Illumina is even worse than Ion Torrent. It has higher Quality scores, and some the bases are also wrong ;-)

So far my personal opinion about quality scores.

Anyway, in the end, the consensus sequence has a high quality score and is reliable.

Slide 110:

How are sequencing errors introduced ?

Slide 111:

After sequencing, the reads have to be trimmed on quality. Q-values of 20 and above are accepted as good. If trimming is done on a way that everything is removed after the first base that has a Q-value below 20, it is possible that a lot of reads will be trimmed drastically.
In this example, the read would be trimmed at base 28 (G) and everything after it would be removed.
For that reason, trimming is often done with a sliding window (there are also other possibilities). For example a “sliding window 30” and “cutoff 15”. This means that the average quality of the first 30 bases is calculated. As long as the average is above 15, the window slides to the next base and the average is calculated again. This sliding goes on until the average score drops below 15. When that happens, everything after that window will be removed. This way, you will retain more and longer reads.

Slide 113:

This is a limited list of possible applications in NGS. (for some, it can be discussed if they are in the correct ‘group’ or not)

De novo sequencing: sequence the genome of an organism that not has been sequenced yet.

Resequencing: sequence the genome of a species that has been sequenced before to compare them (it has to be the same species).

Slide 114:

Targeted (re)sequencing: after random amplification, target genes are isolated with probes and sequenced.

Slide 115:

Mutation detection: individuals are sequenced, and compared (to a reference) to find mutations.

Slide 116:

Mutation detection: this can not been seen as ‘random errors’ because it is in a specific position and there are 2 possible bases in that position.

Slide 117:

(Moorcraft et all., Understanding next generation sequencing in oncology: A guide for oncologists, 2015, Critical Reviews in Oncology/Hematology 96)

Single Nucleotide Variation (SNV): mutations in exomes (regions that code for genes) can cause problems in proteins. A codon is a sequence of 3 bases that corresponds with an amino acid. There are 3 types of mutations:
- *silent mutation*: a base in the codon is changed but does not cause a change in the translated amino acid. A normal protein is formed.

- *missense mutation*: a base in the codon is changed and does cause a change in the translated amino acid. A faulty protein is formed. Some changes have no or little effect on the protein's function. Other changes can have a huge effect (folding or binding to receptor capability) on the protein's function.
- *nonsense mutation*: a base in the codon is changed and codes for a stop codon. An incomplete protein is formed.

Missense and nonsense mutations can be the reason for developing cancer.

Insertions and deletions (indels): one or more bases are inserted or deleted from the sequence. If this change is a multiple of 3 (size of codon), one or more amino acids will be added to the protein. If the change is not a multiple of 3, this will result in a frameshift. Each codon after the change will start in a new place and the resulting amino acid will change. Mostly, frameshifts will result in a loss of function of the protein. "In frame" mutations can reduce, enhance or transform the function of the protein which can lead to the development of cancer.

Slide 118:

Amplicon sequencing: this can be used for heterozygote genes. In Sanger sequencing, this gives double peaks which are useless. A fusion PCR method is possible: the normal primers for the pcr contain the adapter and primer sequence for the system you will use. The ligation step in the library preparation is no longer necessary.

Slide 119:

In amplicon sequencing it is also possible to sequence (parts of) multiple genes at once. One option is to do that in separate pcr's for each amplicon and each species. That is a lot of work. Afterwards, the samples are pooled and adapters are ligated in the final library preparation. When this is followed by a PCR step of a few cycles, there is a chance that a low percentage of chimera are formed. (chimeras are sequences formed from two or more biological sequences joined together. Amplicons with chimeric sequences can form during PCR)

Slide 120:

An other option in amplicon sequencing is to use multiplex pcr. Different primersets are added to one PCR tube per species. Pro is that it is less work to perform. Con is that is is difficult to optimize the primer concentrations. Some genes will be amplified a lot, other genes will be amplified a few times (or not at all). Also the possibility for chimera formation in the pcr after the final library preparation.

Slide 121:

Amplicon Cancer Panels: a few examples of many kits that are available to screen for mutated genes that increase the chance of developing cancer.

Slide 122:

Phylogenomics and phylogenetics

Slide 123:

RNA sequencing: one option is to see if different genes are up or down regulated in different

conditions.
An other option in RNA sequencing is to look for the genes in the transcriptome (eg to design primers, look for differences with other species, …)

The transcriptome also contains long non coding RNA (lncRNA). These do not code for proteins but have other functions:
- regulation of gene transcription
- post transcriptional regulation
- epigenetic regulation
- translation

A problem with NGS is the short read length. Assembling the reads can be problematic with isoforms of genes.

Slide 124:

RNA Cancer Panels: a few examples of many kits that are available to screen for mutated genes that can increase the chance of developing cancer.

Slide 125:

Isoforms in a publication: 68 isoforms in library. 63 discovered by Illumina and 27 false positive. Oxford Nanopore found all 68. PacBio missed 1 (reason: size selection in library prep (missed gene is 219 bp long)).

Slide 126:

2D transcriptomics: www.spatialtranscriptomics.com

Now it is also possible to combine histological information with genome wide transcriptome data. High-resolution imaging and RNA capture are used for this.

Principle:
- top left image: a slide contains 1007 spots.
- middle left image: each spot has its own barcoded poly-T tailed probes.
- bottom left image: tissue is place on top of the slide, the tissue is fixated and imaged. The tissue is permeabilized with reagent. RNA exit the cells and bind to the poly-T tailed probes on the chip. cDNA synthesis is performed on the chip. The cDNA is cleaved off the chip, used for library prep and sequenced. Based on the barcode, the reads can be mapped back to the position on the tissue.

Some figures from a publication:
Visualization and analysis of gene expression in tissue sections by spatial transcriptomics
Patrik L. Ståhl et all. Science 1 July 2016

Comparative analyses of tissue domains
(B) The features placed back onto the two tissue images. (C and D) Histological section of a breast cancer biopsy (C) containing invasive ductal cancer (INV) and six separate areas of ductal cancer in situ (1 to 6), with analyzed spatial transcriptomics features in (D). INV areas without, or with minimal, stromal infiltration were selected. (E) Gene expression heat map over the different areas in four adjacent sections (D)

Slide 127:

microRNA sequencing:

Mature microRNAs (miRNAs) are a class of naturally occurring, small non-coding RNA molecules, about 21–25 nucleotides in length. MicroRNAs are partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to down regulate gene expression in a variety of manners, including translational repression, mRNA cleavage, and deadenylation.

Slide 128:

SAGE and CAGE sequencing: SAGE is the 3' side of the RNA, CAGE is the 5' side. 17 bp of the start or stop site are sequenced and mapped to the genome. New start and stop sites can be discovered, antisense transcription can be found.

Slide 129:

ChIP sequencing: Chromatin Immuno Precipitation.

Some proteins are linked to specific places in the DNA and have a function there. These complexes are isolated, the proteins are removed, the DNA fragments are sequenced and mapped to the genome.

Slide 130:

Structural variation: the search for insertions, deletions, duplications, inversions, translocations, ... in the genome. These changes can cause errors in genes, change the expression of the genes, can lead to fusion of (parts of) genes that produce proteins with abnormal functions. These changes can cause cancer.

Slide 131:

Metagenomics: isolate DNA form an environmental sample, sequence everything and identify the species that were present in the sample (BLAST).
Metagenetics: isolate DNA form an environmental sample, amplify a few genes, sequence those genes and identify the species that were present in the sample (BLAST).
Metatranscriptomics: isolate RNA form an environmental sample, sequence everything and identify the species that were present in the sample (BLAST).

Slide 132 - 135:

Microsatellite sequencing: (Wikipedia) a microsatellite is a tract of repetitive DNA in which certain DNA motifs (ranging in length from 2–5 base pairs) are repeated, typically 5–50 times. Microsatellites occur at thousands of locations within an organism's genome; additionally, they have a higher mutation rate than other areas of DNA leading to high genetic diversity.

Side 133:

Old school microsatellite sequencing: fragment analyses. Orange peaks are the size standard (= ladder) What does a small peak before or after a big one mean ? How to interpret this result ?

Slide 134:

This is an example of microsatellites sequenced on an Ion Torrent PGM. The microsatellite is the repeat of “aag” for 11 or 12 times.

The length of the pcr fragment (including the microsatellite) is 111 bases for most of the reads in the top half of the alignment.
Five reads have a length of 108 bases, which means one repeat of “aag”is missing. This is what you can expect: some have 11 repeats, other have 12.

The bottom part of the alignment contains fragments of 112 bases. So there is one base extra. In the conventional ‘sizing’ of the fragment, this will only give you a number: 112 bases. If we look closer: we can see that the extra base is not always in the same position. So even the length of the fragment is the same, the sequence is different. Sequencing gives you more information than the sizing of the fragment.

Slide 135:

For some of the reads, I checked the quality scores of the bases. To my frustration, the quality scores are going down to the end of the repeat. But that does not mean that the bases are called wrong for the reads with a length of 111 bases. (see my personal opinion of slides 104-105) I checked several of them, and they all have the same pattern. The quality scores are going up again after the repeat.

I also checked a read with length of 108 bases, and for the last part of the repeat, the second A always has a lower quality score (17).

For the reads with a length of 112 bp, the quality pattern is similar. Three A’s after each other are a polymer for Ion Torrent and that is the reason why the quality score is lower for the 3th A. But because several reads have the same sequence (see slide 130), this probably means that the sequence is correct, and this that there are differences in the sequence of this microsatelitte, although this is not visible with fragment analyses.

It is also known that sequencing repeats is more difficult (in Sanger and in NGS) than reading random sequences.

Slide 136-137:

Genetic marker discovery: the purpose is to reduce the complexity of comparing whole genomes between different population. Polymorfisms (little changes in genes) are heritable so they can occur in a big part of a population. If you can identify such a genes, then you don’t have to sequence whole genomes to compare populations, but you can use several genetic markers for that.

Microsatellite development: sequencing (parts) of the genome, and search for microsatellites in the sequences to develop primers.

RADSeq (Restriction-site-Associated Sequencing), RRLs (Reduced-Representation libraries) and GBS (Genotyping By Sequencing).

Figure explanation:
Methods for high-throughput marker discovery.
a | An example genomic region containing restriction sites (red). A sample of DNA from each of

two individuals (sample 1 is dark blue and sample 2 is light blue) is to be sequenced. Sample 2 has a variation in the cut site at 1,300 bases (grey arrow) and so this site will not be cut.
b | Protocols for developing sequenced markers. All methods begin with a restriction enzyme digestion.
Reduced-representation library (RRL; left panels): fragments from all samples are pooled and size selected to 300–700 bp. Fragments are ligated to standard sequencing adaptors (grey squares) and sequenced. In this simple case, only the ends of fragments will be sequenced, but the protocol can be modified to sequence entire fragments.
Restriction-site-associated DNA sequencing (RAD-seq; middle panels): fragments are ligated to P1 adaptors (yellow for sample 1, purple for sample 2), pooled, randomly sheared and size selected to 300–700 bp. P2 adaptors with divergent ends (grey, Y-shaped) are ligated to the fragments with and without P1 adaptors. The fragments are PCR amplified with P1- and P2-specific primers. The P2 adaptor is completed when fragments containing P1 adaptors are bound by a P1 primer and copied, and the P2 primer only binds to completed P2 adaptors (grey squares). This means that only fragments with P1 and P2 adaptors (the fragments containing restriction sites) are amplified.
Genotyping by sequencing (GBS; right panels): barcoded adaptors (yellow) and common adaptors (grey) are ligated to digested fragments, producing fragments with barcode+common, barcode+barcode and common+common adaptor combinations. Samples are pooled and amplified on the Illumina Genome Analyzer flowcell. Only short samples featuring a barcode+common adaptor combination are amplified for sequencing.
c | Sequenced markers are aligned to the original reference genome.
RRL: either fragment ends (thick lines) or entire fragments (thin lines) between 300 and 700 bp are sequenced. Because the site at 1,300 bases is not present in sample 2, the long fragment between 700 and 2,000 bases is filtered by size selection.
RAD-seq: downstream regions of all fragments above 300 bases are sequenced, but not the fragment between 150 and 350 bases. Thin lines indicate the sequence that would be covered using paired-end sequencing.
GBS: dashed lines represent regions that would be filtered during amplification, but could be imputed using (for example) the multiplexed shotgun genotyping hidden Markov model.
The short fragment between 150 and 350 bases will be sequenced.

Slide 138:

Run types: Single-end sequencing: sequencing is done in one direction (forward OR reverse).

Slide 139:

Paired-end sequencing: sequencing is done in forward and reverse direction. If the fragment is short, there will be overlap between the forward and reverse read. If the insert is too long, there will be no overlap. Fragment size can be between 200-1200 bp

Slide 140:

Mate-pair sequencing: the outside of long fragments is sequenced, and since the length of the insert is known, e.g. gaps in an assembly can be closed. Long inserts can be used (0,6 – 25 kb).

Slide 141:

Library prep principle for mate-pair sequencing.

Slide 142:

Barcoding samples: a barcode or MID (Multiplex IDentifier) is a unique sequence (10-12 bp) is added to the DNA (in the ligation step, or in the PCR primers). This barcode is sequenced first, and can be used to identify a sample in a mixture of samples. Barcode can be on one or on both sides of the sequence.

Slide 143-148:

Explanation of the complexity of the data analyses.

Slide 143:

Data analyses: lets go back to Sanger sequencing (capillary sequencing) where you sequence a single pcr product with 6 primers. You assemble the fragments and check the assembly for errors. This will take 5-10 minutes for a 2000 bp assembly.

Slide 144:

Lets simplify the idea from slide 109: you have 8 sequences (8 dots), you connect them (assemble) and you get your result that you can compare to the database on the internet (BLAST).

Slide 145:

In this example we sequenced one gene of a fish, and when we BLAST that gene, the BLAST result is a fish. We are happy.

Slide 146:

Simplified next generation sequencing: each dot is a sequence. Connect the dots...

As you can see, it is impossible to connect the dots manually, and correcting them for errors would take a few years. The only option is to use a computer program to assemble that.

Slide 147:

The computer program has assembled the sequences, and this is the result. As you can see, there are still dots that are not used in the assembly. These sequences can be contamination, can be parts of the genome that does not have enough coverage to be assembled.

So do not expect that the genome will be covered completely or that every read will be used in the assembly. You will have gaps, and sequences that don't fit in the assembly (contamination, reads with much sequencing errors).

Slide 148:

The result of previous slide was nice, but the problem is that I sequenced a human genome. When I use the same dataset with other parameters in the computer program, I can get (slightly) different results.

You are never sure that your assembly is the correct one. There is no “best” assembler, you should try different ones and compare the results. You should run multiple assemblies multiple

times. Determining which assembly is the best is not an easy question. (Monya Baker. Nature Methods Volume 9 No.4, 333-337 (2012))

Slide 149:

It is impossible to check the assemblies manually. The top view is a contig of 7.500.000 bp. When zoomed in to a part of 23.581 bp, you can see parts with high coverage and parts with no coverage.
The bottom picture is a assembly zoomed in to the base level. As you can see, every few bases, a new sequence starts. There is one column with a mutation (SNP), or is it from a heterozygote gene.

Slide 150 -153:

Example of the assembly of the first sequence I have done as training on the Ion Torrent: a bacteria Clavibacter sp.

Slide 150:

This is an example of a bacteria sequenced on a 314 chip with an output of 11,97 Mb Q20. There are 4206 contigs, but no complete genome. The longest contig is 5882 bp and contains 4028 reads.

Slide 151:

The largest contig of the assembly. You can see sequencing errors in some places.

Slide 152:

The same contig, but here the forward sequences are shown in light blue, the reverse sequences in dark blue. This proofs that the ligation of the primer and adapter in the library preparation (slide 9) is random, and sometimes the adapter is on the 3' side, and sometimes on the 5' side.

Slide 153:

When selecting a part of the assembly in the top area, you can see the coverage of that part. Here 115 x coverage, in slide 143 a different part is selected, and has 42 x coverage.

Slide 154:

What to expect from a NGS run ?
Result from a MiSeq run, 2 x 250 bp. Left figure the forward reads. Most reach around 250 bp, but as you can see there are also reads that are only 50, 100 , 150,... bases long. The figure on the right side are the reverse reads. Most of them only reach 160 bp. (is this typical for Illumina, or is this a bad example ?)

So in this example, suppose you want to sequence amplicons and you want 50 bp overlap between the forward and reverse read, you can only sequence fragments of 200+110=310 bp because most of the reverse reads do not reach 250 bp.

Slide 155:

Result from an Ion Torrent PGM run, 400 bp. A size selection has been done in the library prep, and as you can see in the figure on the right side, most of the reads are around 400 bp, but there are also reads that are between 20 and 400 bp. The reasons can be: high GC or AT content in the reads, low bascalling quality scores, …

In the left window you can see that there was 81% loading of the wells, producing 990 Mbases of data.

The middle window is showing the 81% loading, from that 81% there is 100% enrichment (all beads contain DNA), from that 100% enrichment there is 71% clonal (29% of the reads give double peaks) and from that 71% there is 93% usable. So from the 6,2 M wells, 66% (or 3,360,432) gives usable reads.

Slide 156:

It is important for your project that you know the genome size, and how much coverage you want. 15 times coverage is normal, but If you are are looking for rare genes, more than 100x coverage can be necessary to find that gene.

Remember that mitochondrial DNA is 100 times more present than genomic DNA. This will also show in your results.

And sequencing is the easy part. Analyzing your data afterwords is a huge task for big datasets.

Slide 157:

Pro's and con's of commercial and open source programs for Next Generation Data processing.

Slide 158 -161:

A not up-to-date list of programs for Next Generation analysis on the website/forum SEQanswers.

Slide 162:

Galaxy: a free online server for Next Generation analysis. Can be installed locally.

Slide 163:

The FastQ file format. Results are mostly delivered in this format. For the Quality score the ascii table is used. Different platforms use different parts of the ascii table.

Depending on the application, different steps have to be done to get your results.

Slide 164:

Think before you start with Next Generation Sequencing…

(example of wrong result: patient had a mutation. The program did not find the mutation. When taking a closer look, the program had to find (for example) 1000 reads with that sequence, and in 60% of those reads the mutation had to be there. The patient was male (XY) and because the mutation was on the X-chromosome, it has less reads than a sample from a female (XX)

patient. So the program found (for example) only 600 reads with that sequence, and that was not enough to be reported.)

Slide 165:

link with the latest version of this presentation:

http://users.ugent.be/~avierstr/

andy.vierstraete@ugent.be