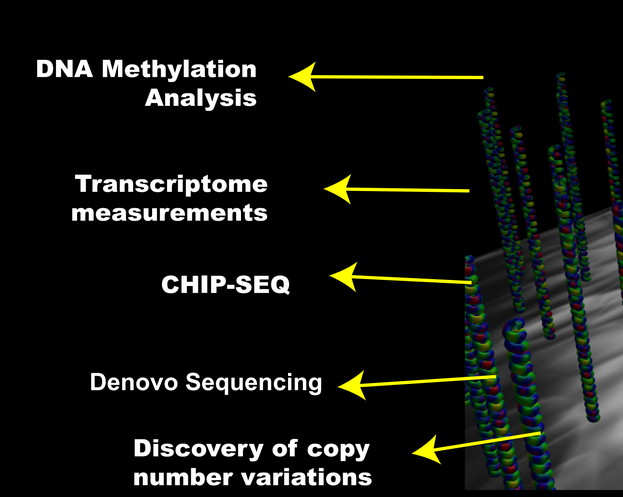
Deep Sequencing

Werner Van Belle, Ina Nissen, Christian Beisel
Laboratory for Quantitative Genomics
Department Biosystems Science and Engineering - ETH Zurich

D-BSSE
Department of Biosystems Science and Engineering

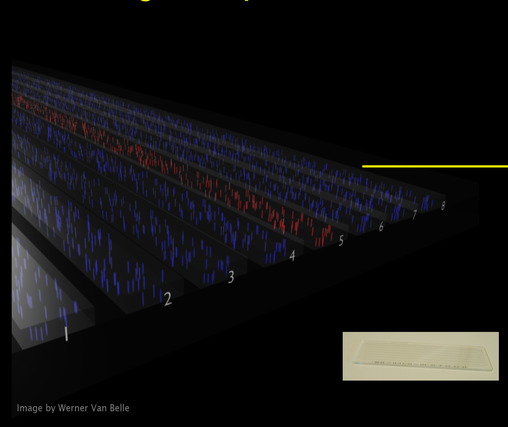
FMI
UNI BASEL

1. Sequencing By Synthesis

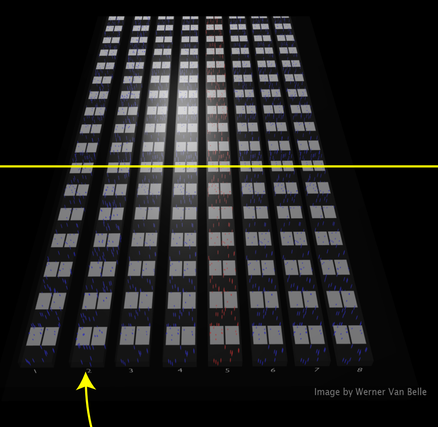


Around 100'000'000 short fragments (200-300bp) are grafted to a special plate. One can sequence the first 36 bases of each fragment by incorporating bases with fluorophores.

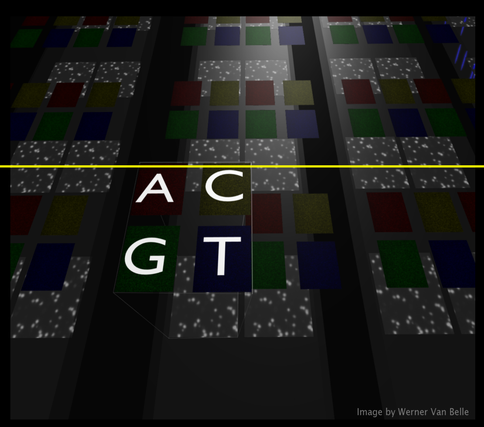
2. Image Acquisition



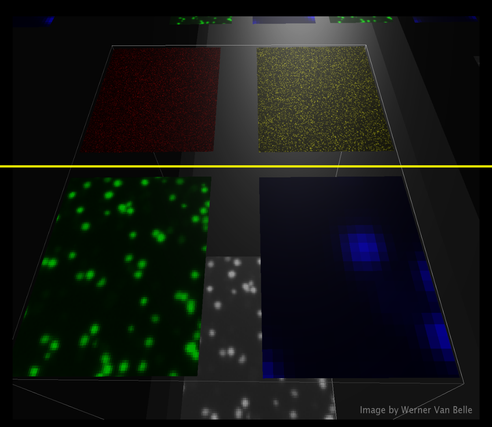
A flowcell consists of 8 lanes. Each lane can contain one biological sample. Lane 5 is always reserved for a control sample (PhiX).



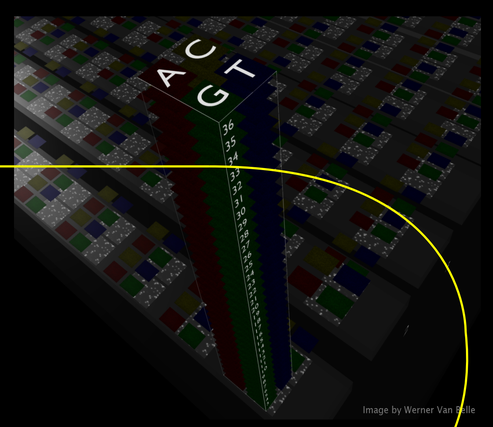
Each lane has 100 imaging positions, called tiles, spaced in two columns



Each tile is imaged 4 times. Once for each fluorophore.



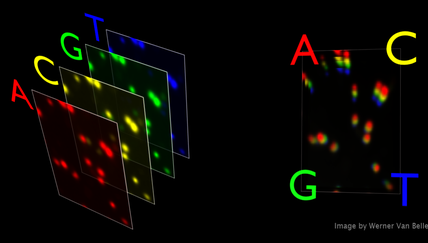
4 tile images. Topleft: normal size. Topright: 2x Bottomleft: 50x. Bottomright 100x zoom. Each spot is a short fragment (cluster)



The imaging cycle is repeated 36 times. Between two imaging cycles we incorporate a new base.

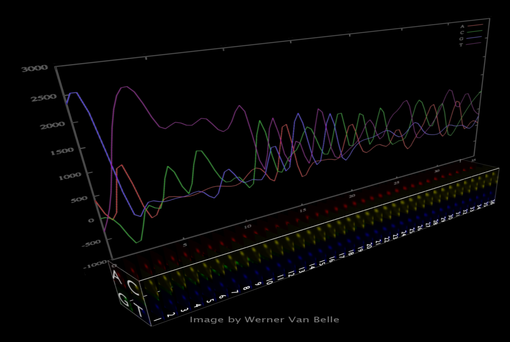
36 x

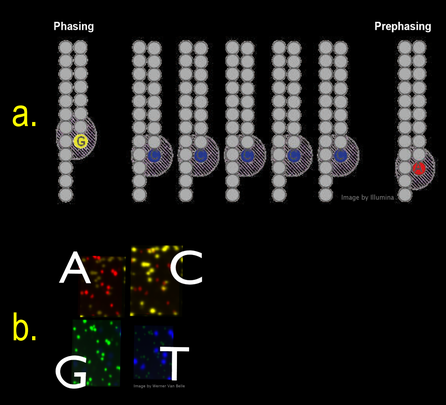
3. Image Alignment



Because the lightpath for each of the filters is slightly different we must first align all images. Afterwards clusters are identified and listed. This is done with Firecrest.

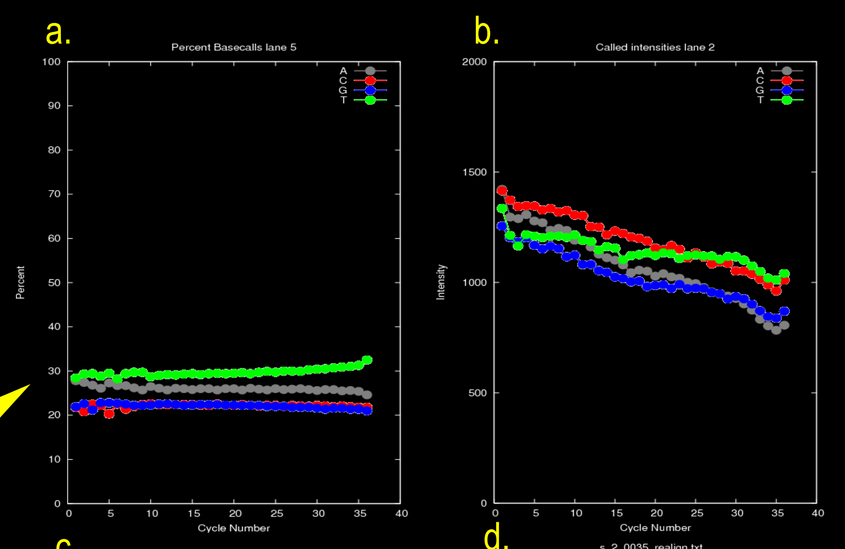
4. Basecalling





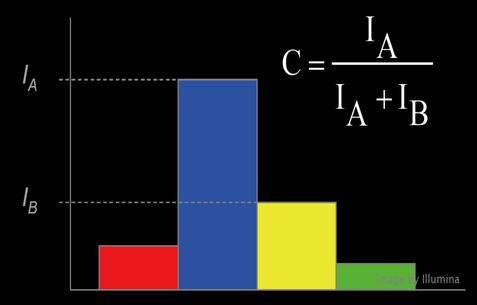
Each cluster has 36 intensities for A, C, G and T. These intensities require further normalization because
a) certain molecules will run ahead or lag behind (phasing and pre-phasing) and
b) there is crosstalk between the A, C, G and T channels.
Afterwards, the basecalling can be performed. This is done with Bustard.

7. Reports



Gerald generates various error reports
a) % basecalls per base per cycle
b) average intensities for each base per cycle
c) statistics on base mismatches per cycle
d) % of read errors per base per cycle

5. Data Filtering

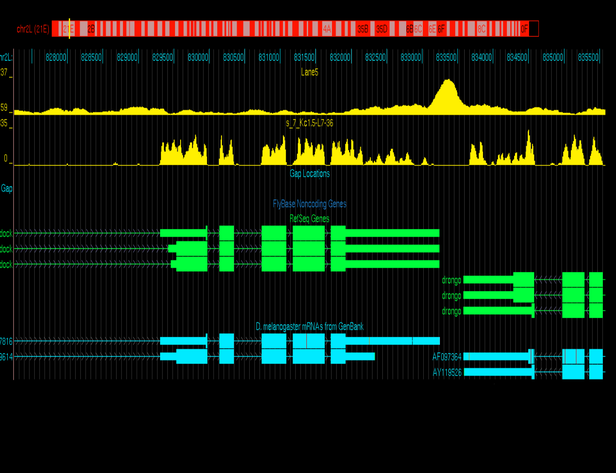


Further data filtering is based on the ratio between the highest intensity and the second highest intensity. If the ratio C is smaller than 0.6 then the base can be considered unclean.

6. Short Fragment Alignment

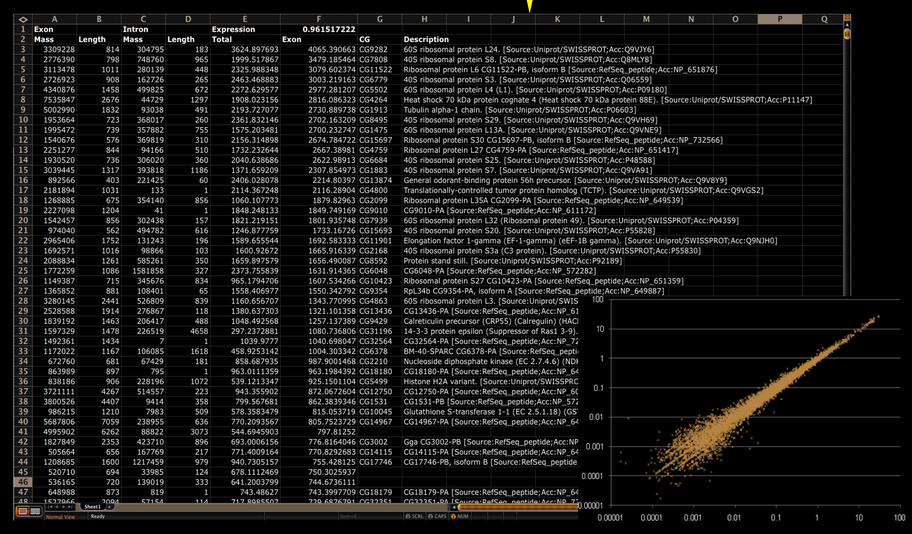
The short fragments can be aligned to a reference genome (yellow). The first program (Eland) allows for at most two mismatches per fragment. The second program (PhageAlign) will find the best match for each fragment.

8. Genome Browsing



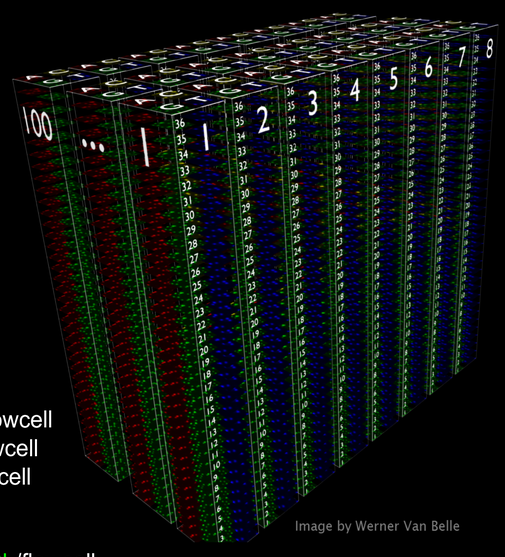
After converting the alignment files to WIG files one can visualize them in the UCSC genome browser.
Track 1) output from a CHIP-SEQ experiment.
Track 2) output from an RNA experiment. Notice the sharp exon boundaries aligning perfectly to the genome tracks.

9. Expression Reports



By counting the number of bases matched at a certain genome position, we can accurately report the expression level of each gene. This includes exon and intron expressions. The correlation between two technical replicates is 0.9903, outperforming existing techniques.

10. Data Delivery



Images: 100 Gb per lane - 800 Gb per flowcell
IPAR Output: 10.4 Gb/lane - 83.2 Gb/flowcell
Intensity files: 8.9 Gb/lane - 71.2 Gb/flowcell
Basecalls: 22 Gb/lane - 176 Gb/flowcell
SRF: 7.53 Gb/lane - 60.42 Gb/flowcell
Filtered Sequences: 1.6 Gb/lane - 12.8 Gb/flowcell
Alignment exports: 1.23 Gb/lane - 14.76 Gb/flowcell
Error reports: 6.47 Gb/lane - 51.76 Gb/flowcell
Minimal Dataset: 8.76 Gb/lane - 70 Gb/flowcell
Everything without images: 66.89 Gb/lane - 535 Gb/flowcell
Everything including images: 166.7 Gb/lane - 1.3 Tb/flowcell