The genome sequencer, the flow cell, and related sequence capacity

The Sequencer:

The MPS core is equipped with the Illumina HiSeq 1000. It can be used for genome wide DNA sequencing, RNAseq, ChipSeq, methylation sequencing, metagenomics, exome sequencing, and small RNA and DNA sequencing to name a few.

It is not designed to replace a Sanger (capillary) sequencer because each run generates over 300 billion bases of sequence and requires intensive bioinformatics support.

Experimental design should be approached carefully and with the input of both the MPS core and Bioinformatics core personal.

The Flow Cell:

The Flow Cell is where the sequencing takes place and is the dimensions of a microscope slide with 8 lanes. Each lane is capable of sequencing up to 46 billion bases pairs and can multiplex up to 24 samples.
Schematic view of single lane:

The sequencing happens as a cluster on both top and bottom surfaces of inner walls of flow cell. Each lane has apx 230,000,000 clusters \((800,000 \text{ clusters/mm}^2 \times 288\text{mm}^2)\)

Each cluster can also be called a "read". Each read may contain up to 200 bases of sequence. This calculates to apx 46 billion bases of actual sequence \((7 \text{ diploid human geneomes with } 1x \text{ coverage})\)

A single flow cell can sequence over 350 billion bases of DNA. This is equate to the human genome with >50 fold coverage.

\[
\text{Total BP} = (8 \text{ lanes/flow cells}) \times (800,000 \text{ reads/mm}) \times (288\text{mm/lane}) \times (200 \text{ base pairs/read}) = 3.5 \times 10^{11}
\]

Generating the Data:

The sequencer can do reads in various lengths; 50, 100, 200 bp reads. For example, applications of microRNA, a 50 bp read is more than adequate because the target is only 22bp. However, for whole genome sequencing where you need maximum read length and coverage, a 200 bp read is needed. The difference is the cost and time. A 50 bp read takes 2-3 days, a 100 bp reads takes 4.5 days, and a 200 bp reads take 8-9 days.

The Data:

The resulting data is collected as bcl files which need to be converted to fastq format through a special Linux software called Casava. This is performed by the bioinformatics core. The resulting file size for all raw data per flow cell is 3 terabytes and would take about 500 DVD's. However, the processed or converted data is smaller at 330 gigabytes or 40 gigabytes per each lane.