Core Facility Genomics

versatile genome or transcriptome analyses

based on quantifiable highthroughput

data ascertainment

Topics

Collaboration with Harald Binder and Clemens Kreutz

Project:

Microarray Validation of Cardiovascular Risk Factors (NGFN)

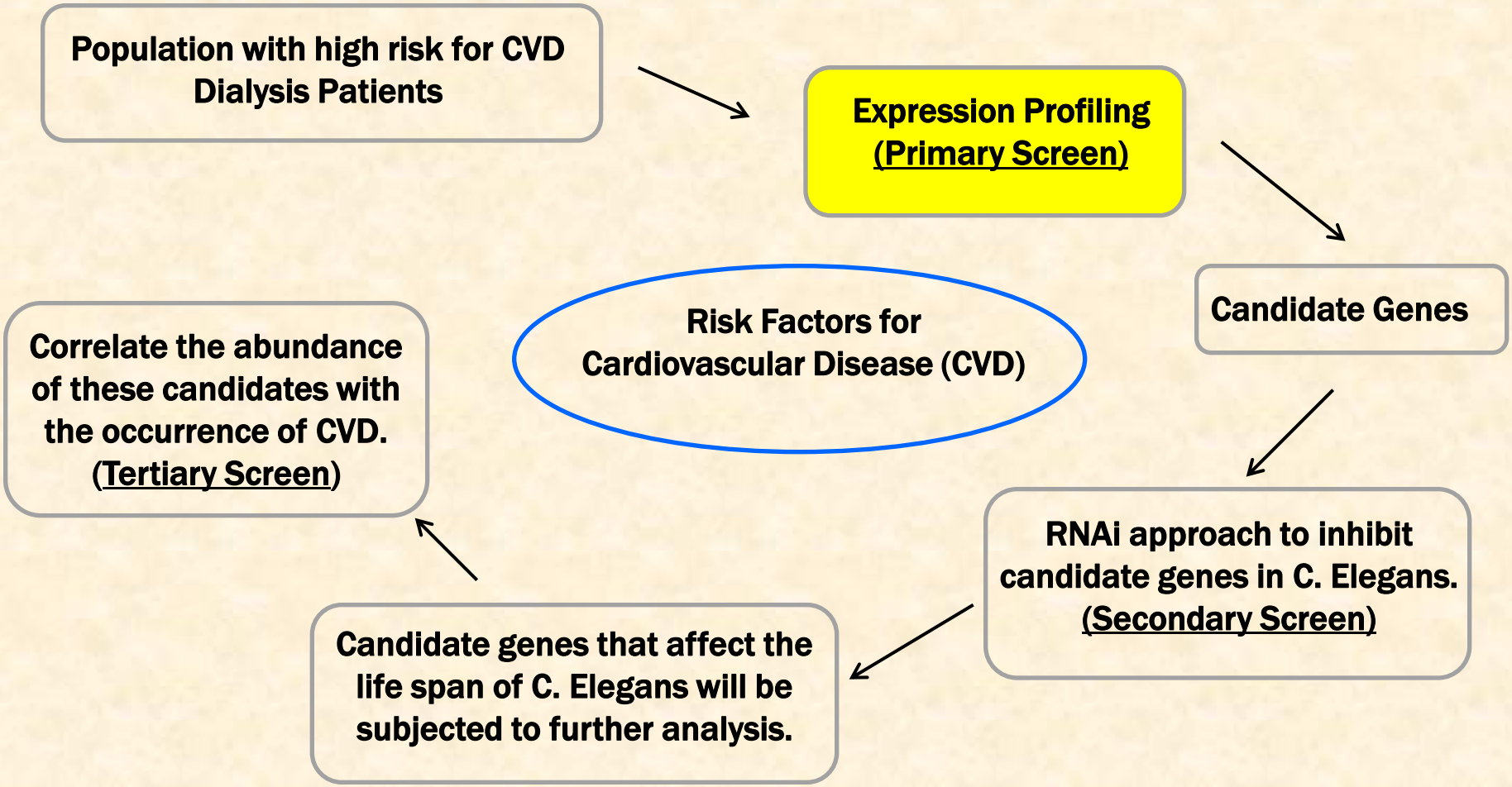
Principal Investigator: Prof. Dr. G. Walz

Collaborator: Prof. Dr. R. Baumeister / Prof. Dr. J. Timmer

Technologies and application spectra of the Core Facility Genomics

(Sample processing / Data analysis (QC) / Statistical analysis)

Workflow


Population with high risk for CVD
Dialysis Patients
Expression Profiling
(Primary Screen)
Candidate Genes
Risk Factors for
Cardiovascular Disease (CVD)
RNAi approach to inhibit
candidate genes in C. Elegans.
(Secondary Screen)
Candidate genes that affect the
life span of C. Elegans will be
subjected to further analysis.
Correlate the abundance
of these candidates with
the occurrence of CVD.
(Tertiary Screen)

Study Population

Collecting blood samples from 324 patients

RNA Isolation

Follow up study
Data acquisition of all relevant clinical data

Printing and Hybridization

Printing of the human Unigene Set - RZPD 3.1 (RZPD-Nr. 983)
37 358 human cDNA clones → two slides

Amplification of the RNA samples using the
Ambion Kit (aRNA)

Sample RNA labelling = Cy3 (green)
Universal Reference (Human) = Cy5 (red)
↓
log-ratio of measured Cy3/Cy5 intensities

Stringency Criteria

Clemens Kreutz, Institute of Physics

Starting with ≈ 12 103 992 data points

Normalization and flagging based on the Genedata Software (Refiner)

Filter criteria (stringent):

- **Rejection of saturated data points / genes (Intensity > 65 000)**
- **Rejection of genes with more than 50% missing data (samples)**
- **Signal / Noise Ratio > 2 (Foreground > Background)**
- **Maximum relative Error < 0.2 (standard deviation / mean x number of pixels)**

2 882 888 data points do not fulfil the stringency criteria

Grouping based on clinical covariates

Group-1

living (221) versus died (100)

Group-2

no coronary heart disease (117)
versus
coronary heart disease (160)

Group-3

living with no coronary heart disease (97)
versus
died by coronary heart disease (30)

Number of regulated Genes

Group-1 (living versus died)

fold > 1.5 = 20	fold < 1.5 = 15
p-value < 0.05 and fold > 1.5 = 14	p-value < 0.05 and fold < 1.5 = 9

Group-2 (no CHD versus CHD)

fold > 1.5 = 9	fold < 1.5 = 11
p-value < 0.05 and fold > 1.5 = 2	p-value < 0.05 and fold < 1.5 = 5

Group-3 (living with no CHD versus died by CHD)

fold > 1.5 = 96	fold < 1.5 = 17
p-value < 0.05 and fold > 1.5 = 78	p-value < 0.05 and fold < 1.5 = 13

Comparing different kinds of Cases!!!

Cardiovascular Disease (CVD)
Polygenic Multifactorial Disease

Dialysis Patients
Population with high risk for CVD

Epigenetics

Genetics

Environment

One and the same clinical phenotype in two patients (e.g. CHD) but different history !!!

Different history could lead to inhomogeneous expression profile within two patients with the same clinical phenotype

Predictive Modeling / Harald Binder, FDM

Subdivision of the study population based on risk strata into low risk (25%), middle risk (50%) and high risk (25%) individuals

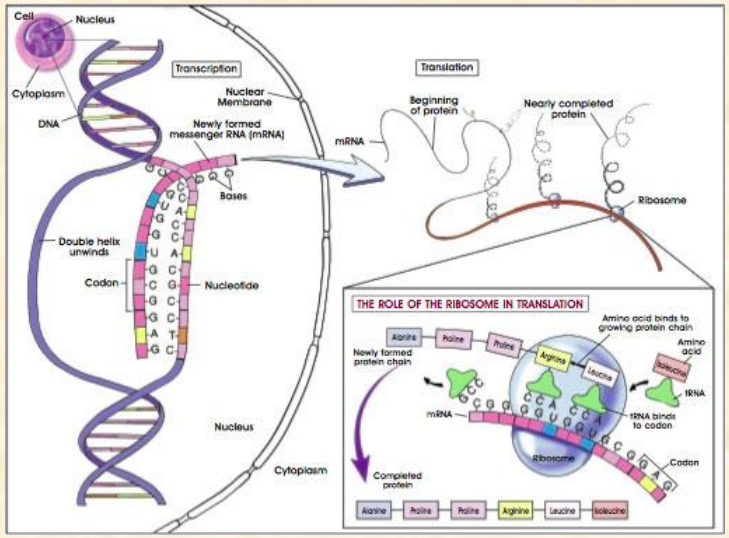
<u>low risk (81) versus high risk (81)</u>

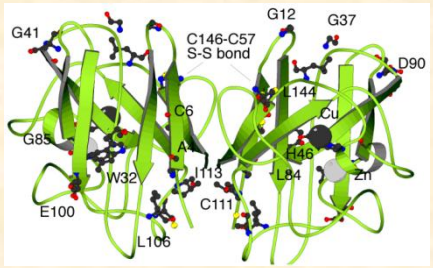
fold > 1.5 = 602	**fold < 1.5 = 281**
p-value < 0.05 and fold > 1.5 = 579	**p-value < 0.05 and fold < 1.5 = 276**
p-FDR < 0.05 and fold > 1.5 = 567	**p-FDR < 0.05 and fold < 1.5 = 276**

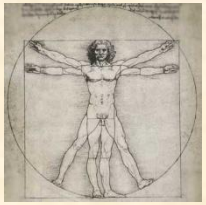
Grouping based on survival analyses lead to more homogenous groups !!!

Technologies and application spectra of the Core Facility Genomics

Genome → ***Transkriptome*** → ***Proteome*** → ***Organism***

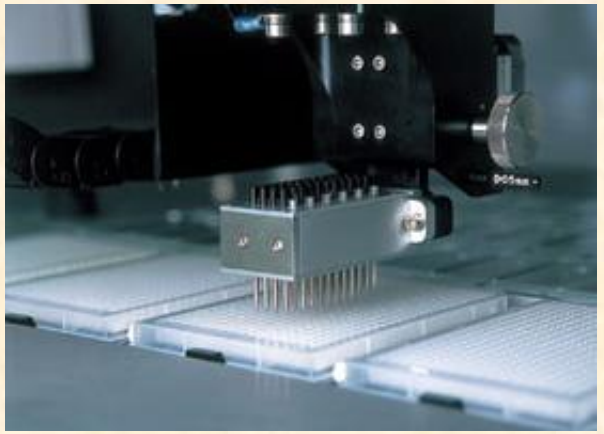


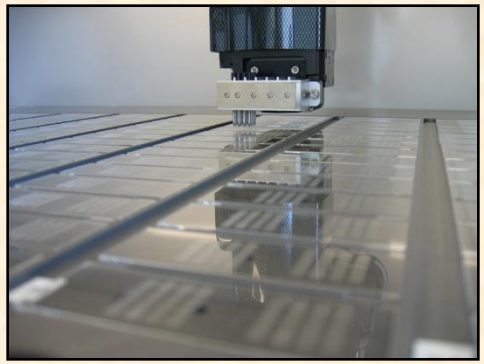


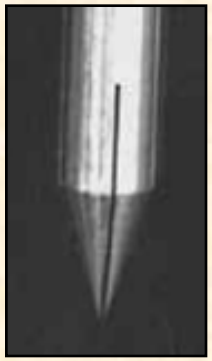


Printing microarrays

cDNA Library

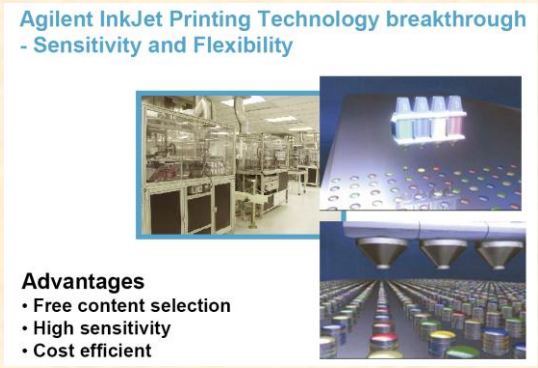




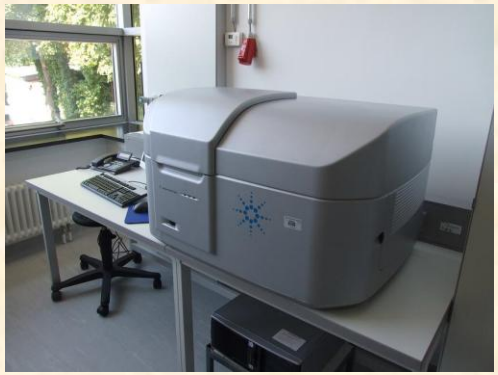


Printing

75µm/250nl



Agilent Platform

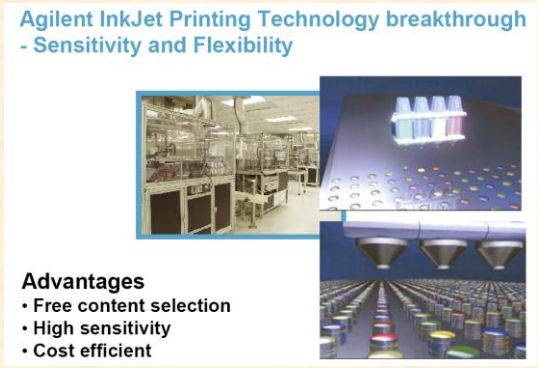


Genome Analysis

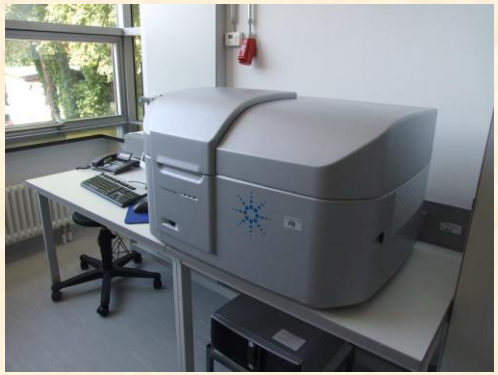
- Tiling Arrays: 1 x 244K, 2 x 105K, 4 x 44K, 8 x 15K
- Comparative Genomic Hybridization (CGH)
- Copy Number Variation (CNV)
- ChIP-on-chip (e.g. DNA methylation, transcription factor binding sites)
- Custom Arrays

Trancriptome Analyses

- Whole Genome Expression for every known or partly known genome
- Micro RNAs
- Splice Variants
- Custom Arrays



Agilent Platform



Scanner upgrade

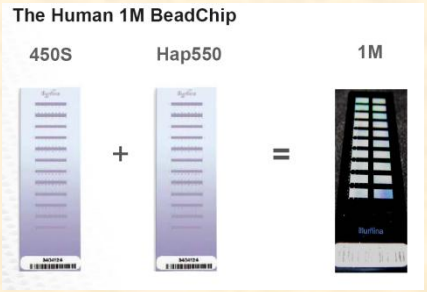
Scanner resolution after upgrade: 2µm, 3µm + 5µm, 10µm

Available tiling arrays formats after upgrade:

1 x 1M, 2 x 400K, 4 x 180K, 8 x 60K

+

1 x 244K, 2 x 105K, 4 x 44K, 8 x 15K



iScan (BeadStation)

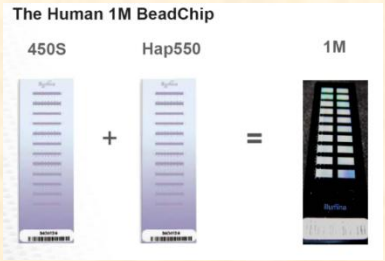
Illumina

Genome Analysis

- Comparative Genomic Hybridization (CGH)
- Copy Number Variation (CNV)
- ChIP-on-chip (e.g. DNA methylation, transcription factor binding sites)
- Custom Arrays

Trancriptome Analyses

- Whole Genome Expression
- Micro RNAs
- Splice Variants
- Custom Arrays



iScan (BeadStation)

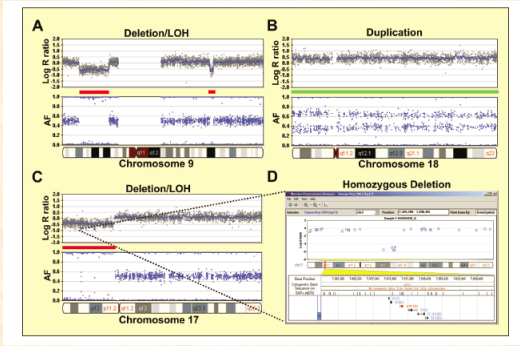
Illumina

SNP Genotyping

SNP-CGH:

Loss of Heterozygosity (LOH)

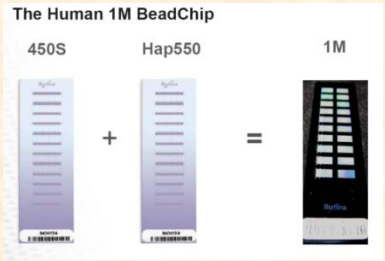
Copy neutral LOH (e.g uniparental disomy)



+

Flexible custom genotyping for fine mapping studies or diagnostics:

96, 384, 768, 1536, 3072 ..... 60.000 SNPs per sample

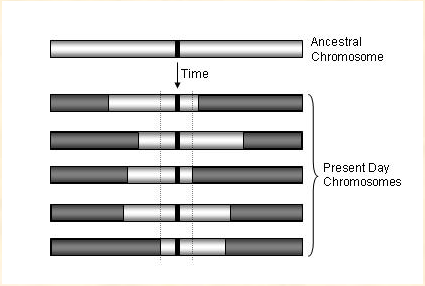


iScan (BeadStation)

Illumina

Whole Genome Association Studies with up to 1 Million SNPs

Haplotype tag (tag) SNPs capture most of the haplotype structure / variation within an haplotype block

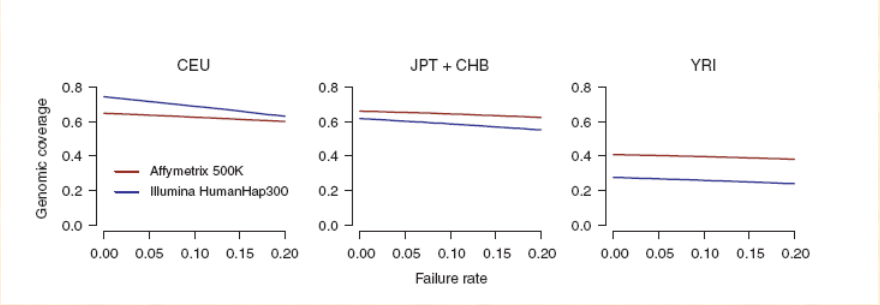


New England Journal of Medicine; Medical Uses of Statistics; 3'"Edition; Identifying Disease Genes in Association Studies (Dan L. Nicolae, **Thorsten Kurz**, and Carole Ober).

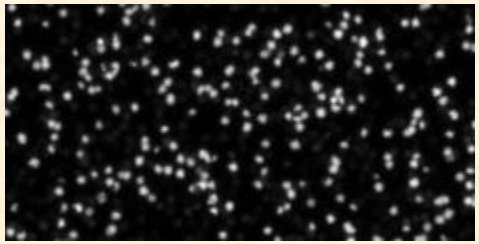
International HAPMAP Project:

tag SNPs that uniquely identify common haplotypes

INCREASE STATISTICAL POWER



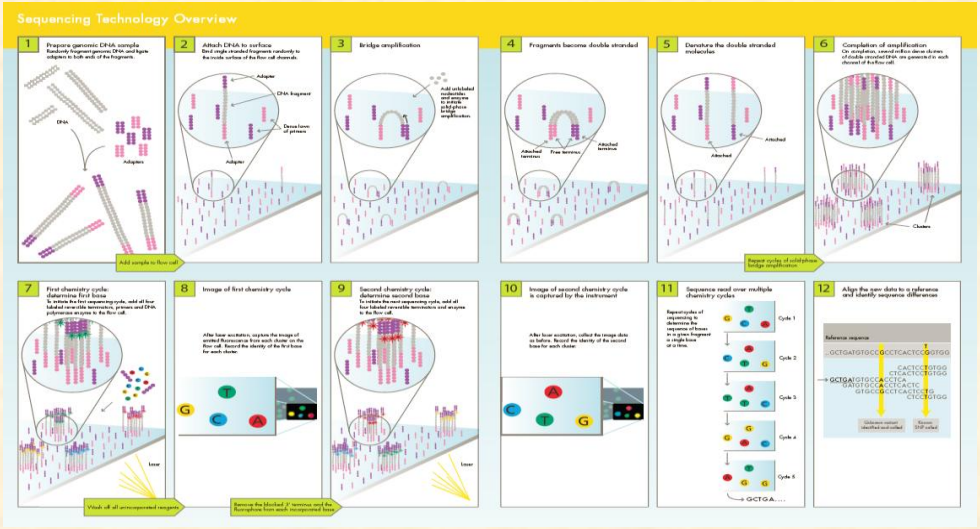
Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659-62



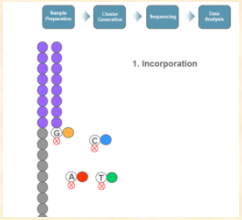
Genome Analyzer IIx

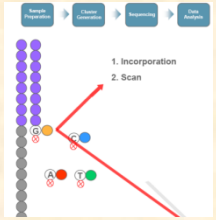
Illumina (Solexa)

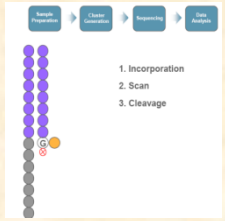
Deep Sequencing



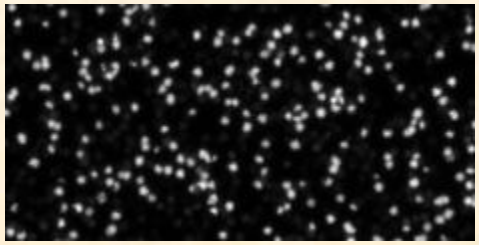
Principle







Dr. Thorsten Kurz
Core Facility Genomics (ZBSA)



Genome Analyzer IIx

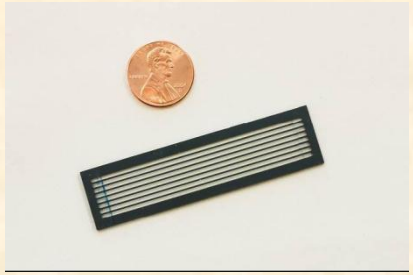
Illumina (Solexa)

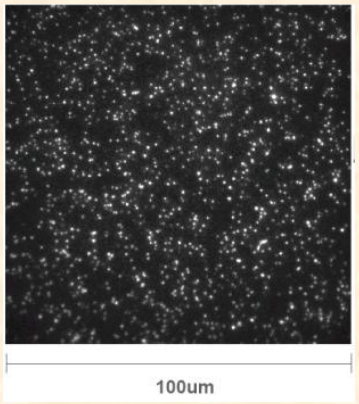
Deep Sequencing / Dimensions

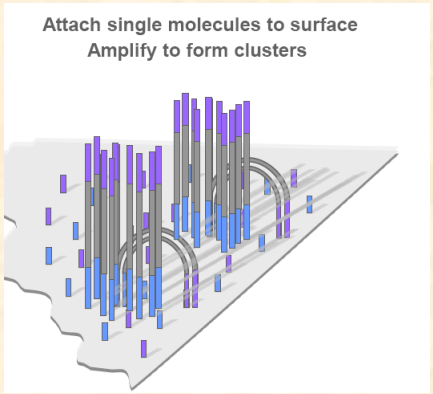
Flow Cell

17mm x 66mm

1.4mm wide channels



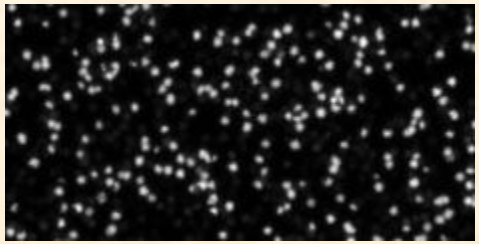




1000 molecules per 1µm cluster = up to 150.000 clusters per tile = 120 tiles per lane

Per Flow Cell:

8 lanes = 960 tiles = 1 50.000.000 clusters = 120.000.000 reads = 14 - 18 GB high-quality output (2x75bp) = 5TB raw data = 9.5 days run time



Genome Analyzer IIx

Illumina (Solexa)

Deep Sequencing / Applications

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation– sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

The Core of the Core Facility Genomics