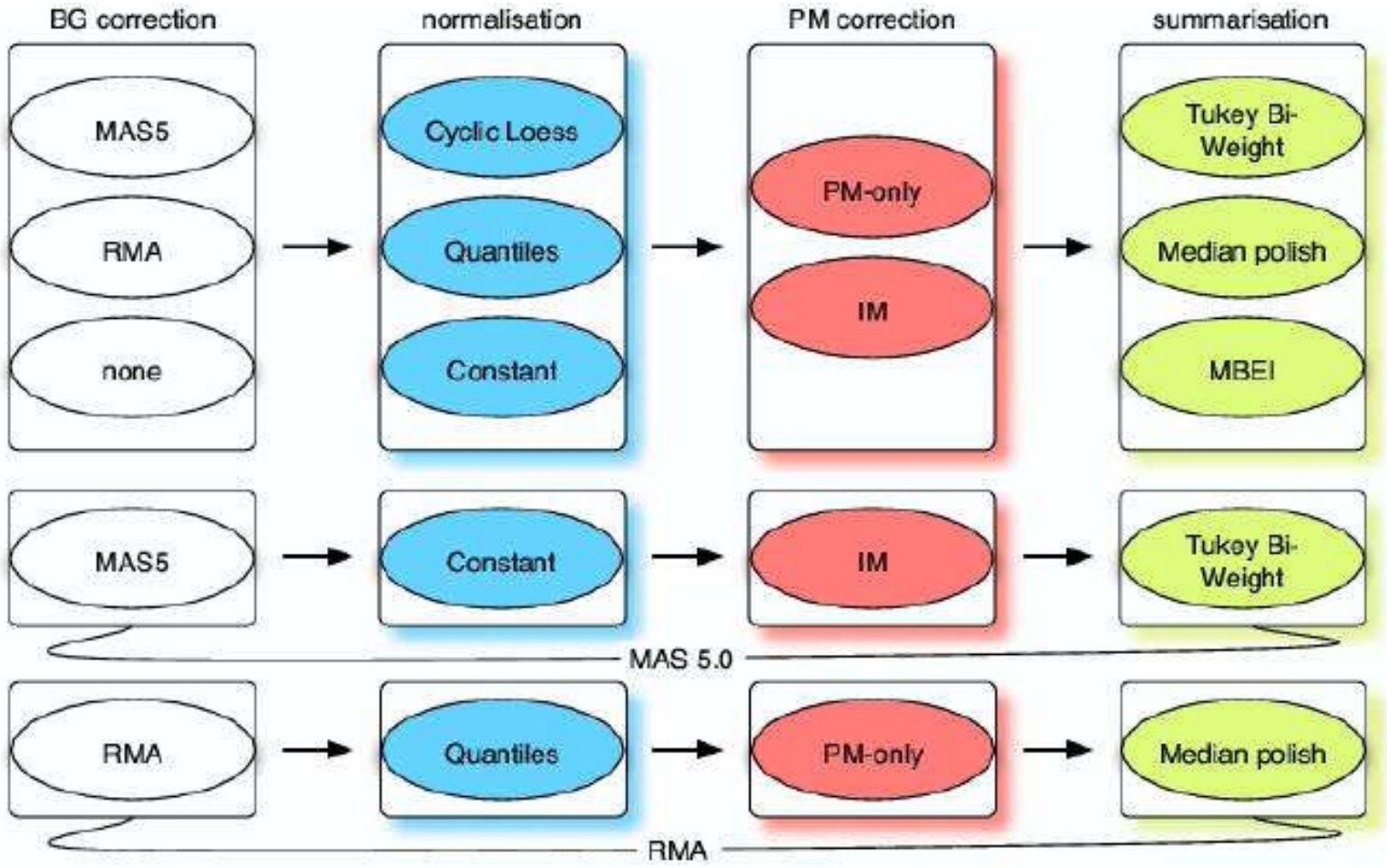
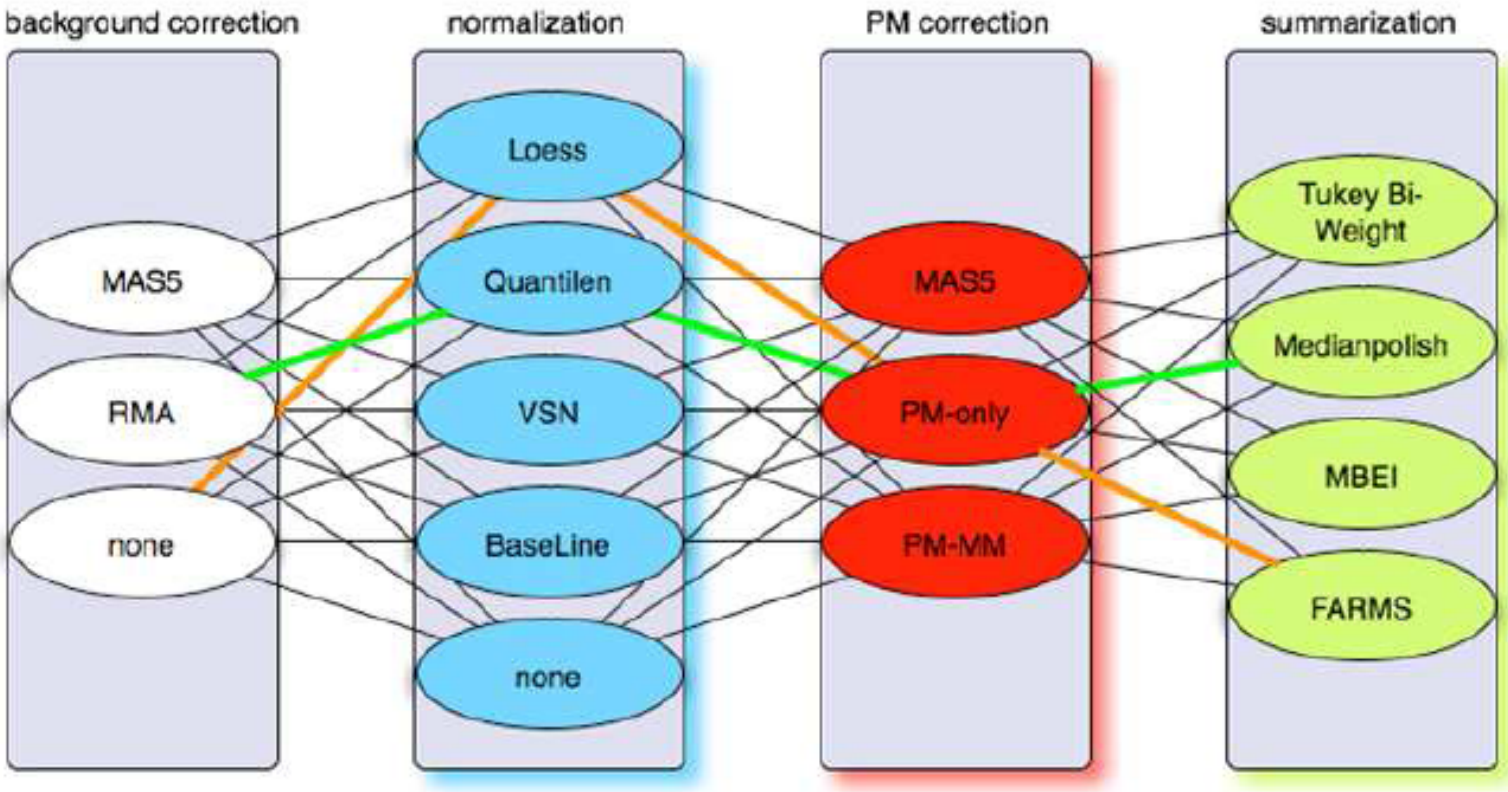
Chapter 7. DNA Microarrays

7. DNA Microarrays
7.7 Different combination of the processing steps



RMA and MBEI are shown



RMA and FARMS are shown

Statistics. To analyze the differential regulation of genes under alternative treatments, to identify "co-regulated" genes and expression patterns → **Expression Arrays**

Data generated used to pose a number of biological or medical questions

For statistical model:

How reliable are the data:

measuring the quantity of mRNA from one gene

replicates to generate the same ratio (F)

Linear regression model with independent homoscedastic errors, f(x) and x linearly dependents and assumed fixed variance

t-test

Statistical hypothesis for two groups with student's t distribution as the test statistic

Estimation of the mean μ

Two samples to be compared

The σ (sd) is unknown and has to be estimated from the data

7. DNA Microarrays
Example

"Preferred analysis method for Affymetrix GeneChips revealed by a wholly defined control dataset"

Choe et al. 2005

Question: the signal intensities are due to DEGs or due to different noise-sources ???

Control data set cRNA spike-in to check the effectiveness

F= Spike/Control --------→Number of fold changes

Desired method: prediction of DEGs by low ↓F + ↑accuracy + ↑Sensitivity

Accuracy Nowadays: 95% with F≥2x ,, sample with DEGs twice concentrated

30% with F<2x

Currently: Decreasing of F due to signal saturation and not to DEGs

3860 cRNAs of known sequence into 2 samples

1309 Spiked ↑**[S]** estimation of false-negative and false-positive

2551 Control **[C]** = **[S]** normalization purposes

Idea: Observed F <<< Theory or known F

BG correction → Normalization → PM correction → Summarization

CONCLUSION

- Subtracting non specific signal from the PM probe intensities
- Performing an intensity-dependent normalization at the probe set level
- Incorporating a signal intensity dependent standard deviation in the test statistic

7. DNA Microarrays
Applications

- R.Redon et al. 2006- **Global variation in copy number in the human genome**

  IDEA: to find and study CNV of DNA sequences by SNPs genotyping arrays screening the International HapMap Project. Genes functionality and diseases related

- M-L Chen et al. 2002- **Identification of a single nucleotide polymorphism at the 5' promoter region of human reelin gene and association study with schizophrenia**

  IDEA: transversion 888G>C (SNPs) as one possible cause of the illness

- Ed S.Lein and col. –**Genome-wide atlas of gene expression in the adult mouse brain**

  IDEA: Comprehension between relationships genes, brain and behavior

  CNS experiment on mouse brain

7. DNA Microarrays
Applications

- Charles G. Mulligan and col. 2007- Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia

  IDEA: PAX5 genes responsible of alteration of B-cell development that contributes to ALL pathogenesis. Detection by SNP arrays

- Florian Markowetz, J.Bloch and R.Spang – Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. 2005

  IDEA: reconstruction of cellular signaling pathways by iRNAs

- Redon R et al - Global variation of copy number in the human genome_COMM. 2006

  IDEA Genome variation profiling by SNP array SNP genotyping by SNP array

- Richard S Sandstrom et al. - DNaseI sensitivity/hypersensitivity Part II. 2007

  IDEA Genome binding/occupancy profiling by genome tiling array

  Focal alteration in chromatin structure in vivo, detectable through hypersensitivity to DNaseI and other nucleases (diverse transcriptional regulatory elements including enhancers, promoters, insulators and locus control regions)

7.9 Gene selection

7.9.1 Protocol description

- Expression values vs. log-ratios
- Normalization and summarization
- Present call
- Standardization
- Gene ranking and gene selection

7.9.2 Comments on the protocol and on Gen selection

- Normalization and Summarization of new arrays
- Correcting the outliers
- Computational costs
- Correlations by chance
- Redundancy

7.9.3 Classification of samples

Gene selection

Extraction of meaningful genes from a set of expression values

↓

↓

Labels prediction of the sample classes

Data preprocessing

Normalization (“Outliers clean-up”)

Feature construction: F.Extraction + F. Selection

Ranking steps

Predictor construction

Expression values vs. log-ratios

Appropriate representation of the data is chosen

$$T_j = \frac{R_j}{G_j} \qquad L_j = \log_2(T_j)$$

R_j red in experiment sample

G_j green in the control sample

} Gene $_j$

Affymetrix→ Expression level [mRNA] in the sample and Log expression value

Normalization and summarization

Measurements from different arrays to be compatible

Quantile normalization and Cyclic Loess

Affymetrix→ 11-21 measurements pro gene

↓

probe set **"summarized"** into one signal

"**Cleaning up the "outliers** -- >

"Cleaning up the outliers"

1st step in microarray analysis : Confidence – Noise clean up

Only a relative subset of the many measurements is informative for interpreting the experiment: identification of truly DEGs

False Positive decrease

High-dimensionality of microarray data reduced before the analysis

From the noisy data exclusion of Non-informative genes and selection of the informative ones

Filtering techniques

Feature selection methods to separate signal to noise

Supervised techniques: ranking of genes on fold changes (Overfitting)

Unsupervised techniques: ranking of features on variation (Exclusion)

significance level of final result by multiple test correction

"Cleaning up the outliers" Present call

Error Model constructed from the expression values or their ratios

SPECIFIC NOISE + GENESPECIFIC NOISE

(All exp. Values affected = bg fluctuations) (Different exp. Values affected ≠, binding efficiency of the day)

p-value assignation: probability of observed measurement value due to noise to every measurement of an expression level

P-value < q_1 (1%, 2%, 5%) and q_2 (3-20 samples) expression level reliable and present call

High → Gene expressed in at least one of the samples: Selected

Low → Gene NOT expressed : excluded

Not included in FARMS

“Cleaning up the outliers“ INI call as Feature Filtering technique

Multiple probes measurement to quantify the Signal to Noise ratio of the probe set under consideration

Bayesian factor analysis with a chosen prior to model the probe level information

Range of exclusion rates from 70-99%

Solution to the high-dimensionality problem

Function implemented in FARMS (R package)

The log-observations of X (PMs) depend on the log Z (mRNA)

$$x = \lambda z + \varepsilon$$

"Cleaning up the outliers" INI call as Feature Filtering technique

[mRNA] Normalized in the mixture → Loading Factor Z

BUT

"*Probes within the same probe set my have different behavior and response to the same target sequence*" *Li and Wong*

Shape parameter : λ_j for each PM_j as individual parameter and specific binding characteristics of each probe

$$\lambda_j = \sigma + \tau_j$$ **Signal strength**

Contributes to the final signal and Large signals leads to large σ which scales up the λ_j

"Cleaning up the outliers" INI call as Feature Filtering technique

The variance of factor Z given the data X: How much variation in the probe set level is explained by the factor Z

$$P(z|x) = N(\mu, \sigma^2)$$
$$Var(z|x) = \sigma^2$$
$$\sigma^2 = (1 + \lambda^T \Psi^{-1}\lambda)^{-1}$$

High λ leads to small sigma: when σ^2 close to zero GOOD RESULT, HIGH CONCORDANCE

Signal to noise ratio $\lambda\lambda^T / \Psi$ as a bimodal distribution with a threshold 0.5

Var(z|x) = o.5 Z and ε contribute equally to the total variation

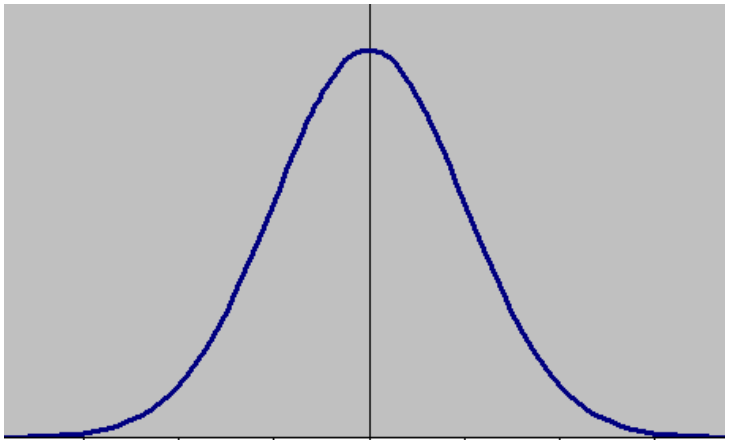
Var(z|x) < 0.5 larger contribution of Z to explain the variation in than ε

Probes Sets are selected

“Cleaning up the outliers“ INI call as Feature Filtering technique

Estimation for λ, Ψ, ε by Bayesian approach: $p(\lambda, \Psi|\{x\}) \alpha p(\{x\}|\lambda, \Psi)p(\lambda, \Psi)$

The prior for the λ from a $N(\mu_\lambda, \sigma_\lambda)$



$\mu = 0 \quad \sigma = \lambda\lambda^T + \Psi$

$x \approx \lambda z$

Noise neglected

Highly correlated probes lead to high values of λ and low values of Ψ

Strong correlation $Var(z|x) \approx 0$

Weak correlation $Var(z|x) \approx 1$

Standardization

Expression values ~N (o,1) all training samples and every gene

No necessary step when the summarization method accounts for $\mu = 0$ and $\sigma^2 = 1$

Expression values differ by orders of magnitude between genes

Goal

↓

↓

Genes with small expression values to be assessed : SIGNIFICANCE

1 Gene Ranking and gene selection

Feature selection method has been chosen where the size of the set of selected genes is controlled with hyperpameter ε

1. **Inner loop**

- Features are ranked (when a subset method is available →P-SVM)
- Features are selected
- Ranking is obtained

2. **Outer loop** ("leave-one-out loop")

- Generalization Model on the test sets
- Optimization: selected genes and hyperparameters values for the unknown examples
- Predictor is obtained (High Ranking Genes): constructed and optimize
- *L* sets of samples with *L*-1 size constructed leaving 1 sample for validation

Subset selection method: GENE SELECTION & RANKING

“Outer Loop”

Inner loop results combined across L different sets of samples

Final ranking → how often genes are selected in the L “leave-one-out” runs of the inner loop

High ranked gene when is selected in many runs otherwise low

Advantage

A high correlation between expression values and class labels induced by a single sample is scaled down if the according sample is removed: more robust against outliers

Other purposes of the Outer-loop

Optimal Gene Number selection
Hyperparameter selection

} v-SVMs trained for each of the *l* sets of samples with ≠ values of the hyperparameter v and the number F

"Leave-one-out" error = function of F of selected genes is NOISY

↓

Replaced by the average of the "leave-one-out" error for F → F+ a and F-a

Average error calculated

↓

F and v with lowest error value obtained SELECTED

FEATURE SELECTION PROCEDURE COMPELETED !!!!!

2 Comments on the protocol and on Gene Selection

Normalization and summarization of new arrays

All known arrays + new array normalized and used for summarization

ML to the training set and new array classified

Correction to the "Outer Loop"

Samples removed for gene ranking should NOT be considered for present call and normalization parameters determination

Feature or hyperparameter selection NOT optimal

Computational costs

Feature selection protocol → L x n_ε feature selection runs

n_ε = Different values of the ε parameter

Computational effort compensation by the increased robustness against correlation by chance

Correlations by chance

IF the N° selected genes is small compared to N° of probes on the chip

High noise induces spurious correlations between genes and class labels

Problem → Spurious Correlations and large negative effects

Solution → N° of selected genes not too small

→ Increasing q_2 (minimum number of reliable expression values for making a present call)

Avoids the selection of genes for which too few samples contribute to the correlation measure

Redundancy

IF the N° selected genes is large

Low generalization performance due to selection of too many genes

Redundant genes indicate the same cause

Problem → Redundant sets of genes with correlated expression patterns

→ Not all causes could be recognized

Solution → Avoiding redundancy

3 Classification of samples

Goal→ Predictor Model construction for the class labels of new samples

Cross-validation procedure to control the performance

Performance of the full gene selection procedure

+

All preprocessing steps separately on all L cross-validation subsets

Before applying the classifier to the new data !!!!

→ The expression values for the new sample must be scaled according to the parameters derived from the training set: Optimizing the performance

→ When expression values exceed the maximal value in the training set, all are set to this maximal value: Underestimation of certain expression levels BUT robustness increased against unexpected deviations

CLASSIFIER APLIED TO THE NEW DATA

7. DNA Microarrays
7.10 Next Generation Sequencing

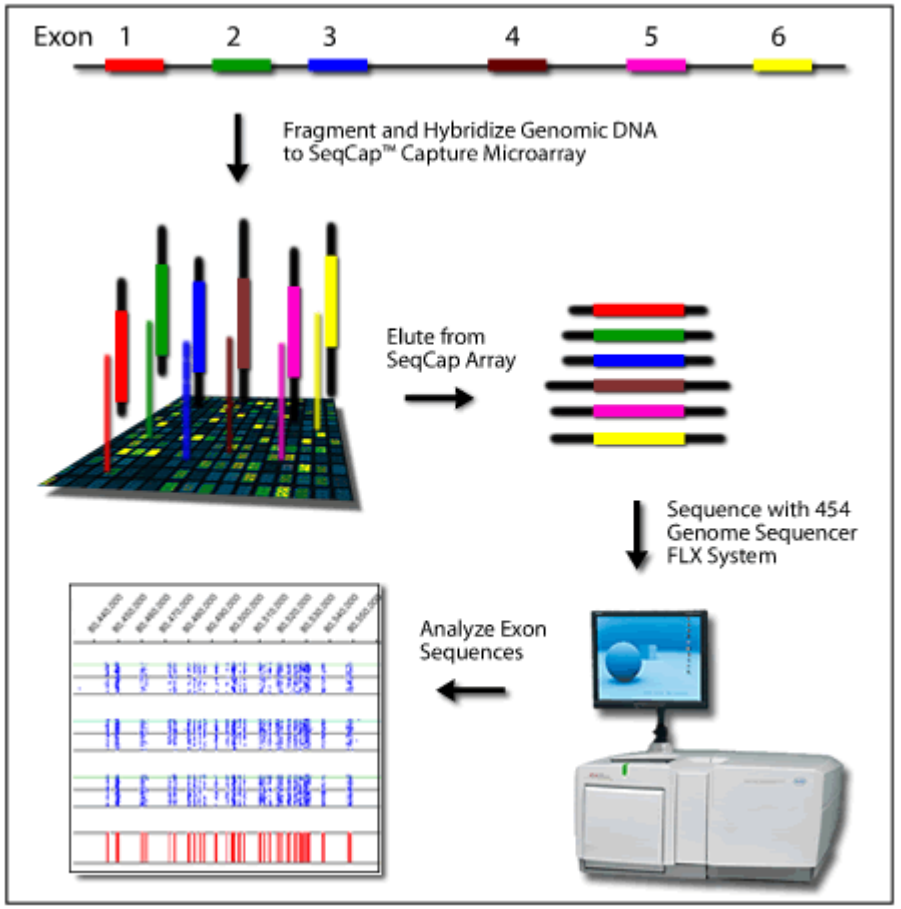
Next Generation Sequencing

Community of genomics and transcriptomics
As an alternative to array based methods

Sequences of the GENOME
DNA Suffix and prefix

BIOINFORMATICS CHALLENGE
Data Infrastructure
+
Data Analysis

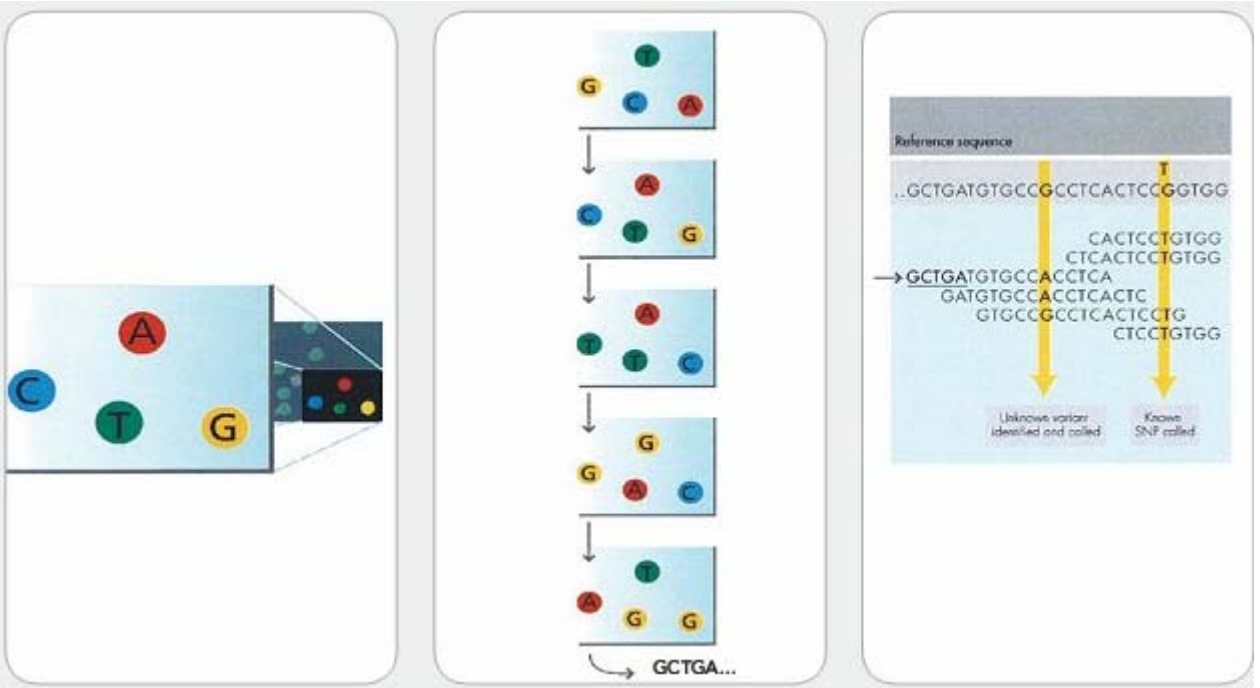
Parallel and GRID computing
http://www.austriangrid.at/



7. DNA Microarrays
7.10 Next Generation Sequencing

ILLUMINA Sequencing

- ✓ Reads
- ✓ Storing



✓Back-mapping

Algorithms main memory and Need of parallel on multiprocessors and/or Run on computers GRID

✓Analysis

Assemble a genome
Transcripts determination
Transcripts concentration
Nucleosome position detection
SNPs detection, CNV estimation

ILLUMINA Sequencing

Produces more than 50 million reads

- One read: 30 – 72 long prefix or suffix sequences of DNA fragments with length 100 to 500 base pairs
- Lane: 150 Gb image data per run/experiment
- First: Reads divided into 8 lanes
  - 1 Experiment 1.2 Tb of image data→ 100 Experiments 120 Tb
- Second: reads mapped back to the reference genome
- Third: analysis on the reference genome

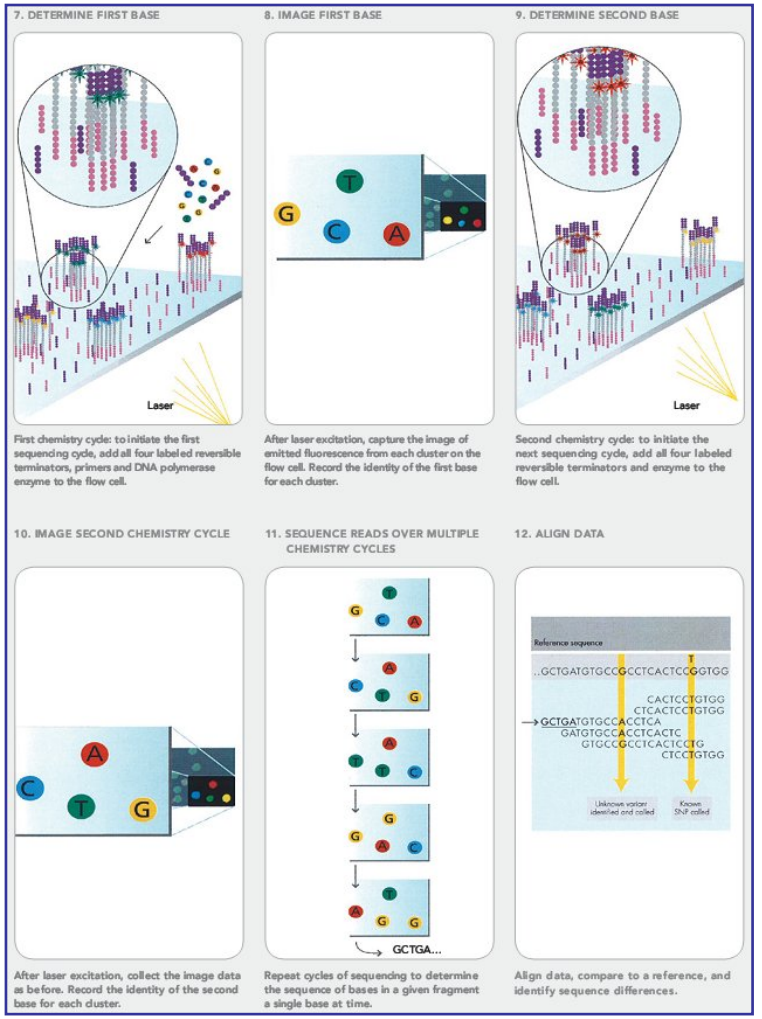
Cluster Generation by Bridge Amplification

The flow cell surface is coated with single stranded oligonucleotides that correspond to the sequences of the adapters ligated during the sample preparation stage

Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polyermase-based extension

Priming occurs as the free/distal end of a ligated fragment "bridges" to a complementary oligo on the surface

Repeated denaturation and extension results in localized amplification of single molecules in millions of unique locations across the flow cell surface



Illumina Summary

Massive parallel sequencing of millions of fragments by reversible terminator-based chemistry

Templates sequenced using four color DNA sequencing-by-synthesis

Removable terminators with removable fluorescent dyes

Randomly fragmented DNA attached to a planar and transparent surface

Attached DNA fragments are extended and bridge amplified

Cell flow created with > 50 million clusters (each approx. 1000 copies same template)

High sensitivity fluorescence detection by laser excitation

“Paired ends”

Both prefix and suffix (algorithm has to include the constraint of matching pairs of reads)

Second > 36bp reads from the opposite end of the fragment and a cluster formation by a bridge (Total > 3Gb paired-end data)

454 Roche

Reads: 400.000 per run BUT 400-800 bases length

Better suited for genome assembly by larger overlaps

Coverage of the genome 100 times smaller than Solexa

SOLID™ AppliedBiosystem

Sequencing of clonally amplified DNA fragments linked to beads

The sequencing methodology is based on sequential ligation with dye-labeled oligonucleotides

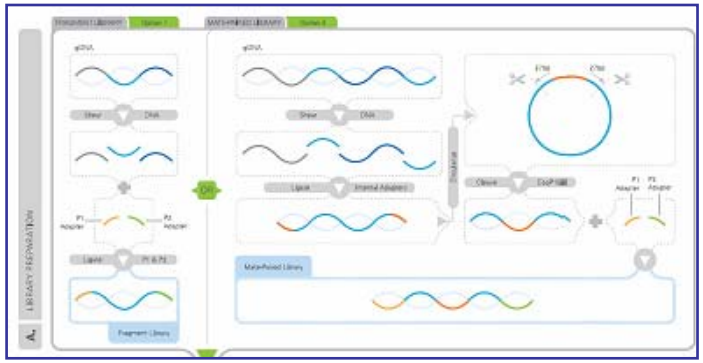
Comparable to Solexa

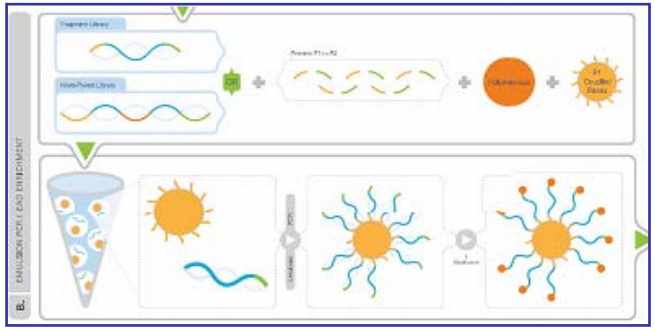
Reads: 10-20 millions per run

SOLID™ AppliedBiosystem

1. Library preparation

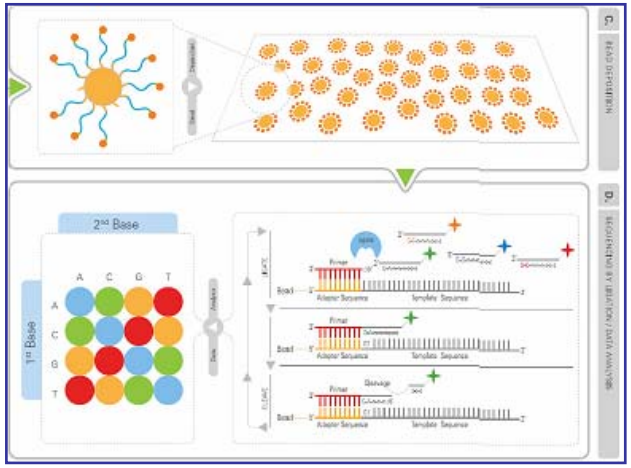
   Two types of libraries can be used





2. Emulsion PCR/ Beads enrichment

   Cloning fragments are prepared



3. Bead deposition and sequencing by ligation data analysis

Solexa Illumina
http://www.illumina.com

s454 Sequencing
http://www.roche.com/

SOLID™ System
http://www3.appliedbiosystems.com/AB_Home/index.htm