Genomic data analysis with R

Macrogen Inc
김세환

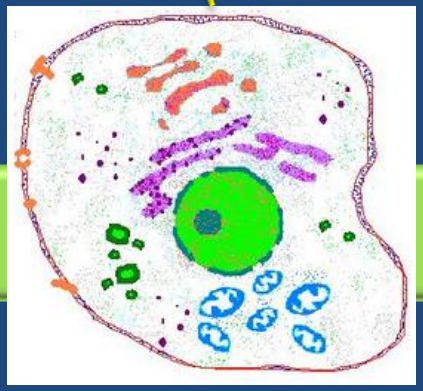
Contents

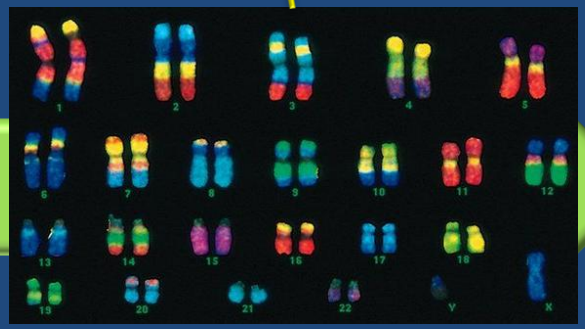
- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

Where We Are

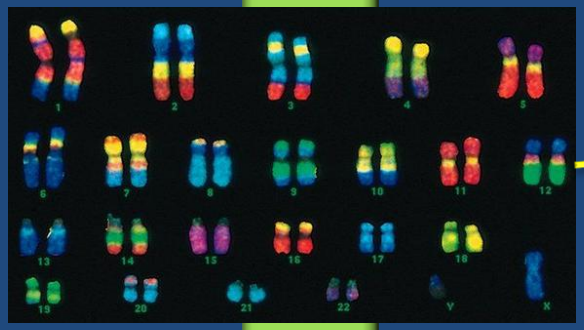
- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

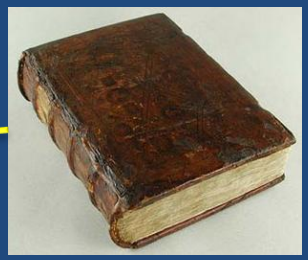
Biological Background for Genomics



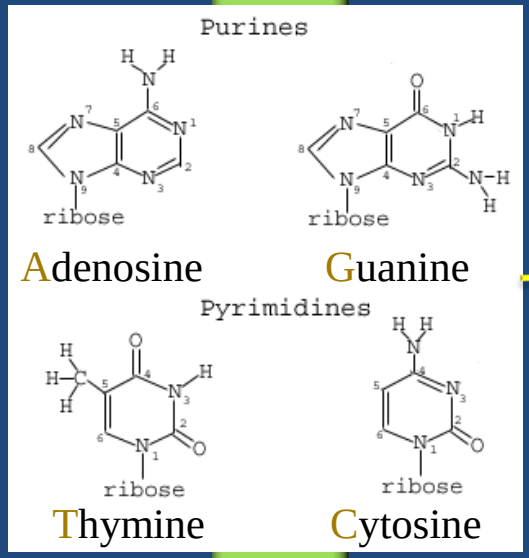


Biological Background for Genomics





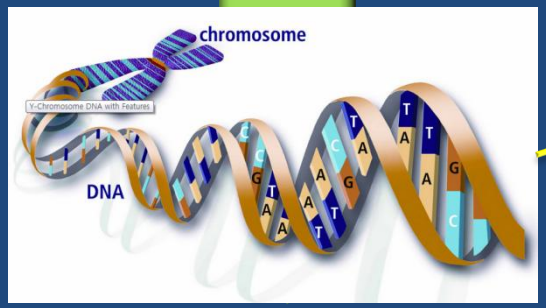
46 chromosomes = 22 * 2 + (1+1)



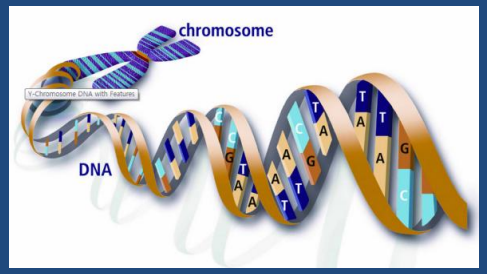
AGTC

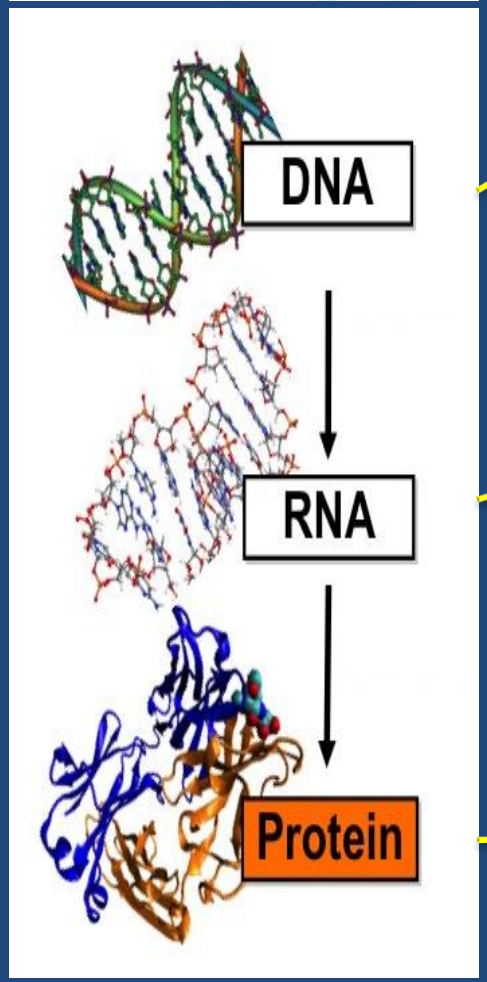
4 bases

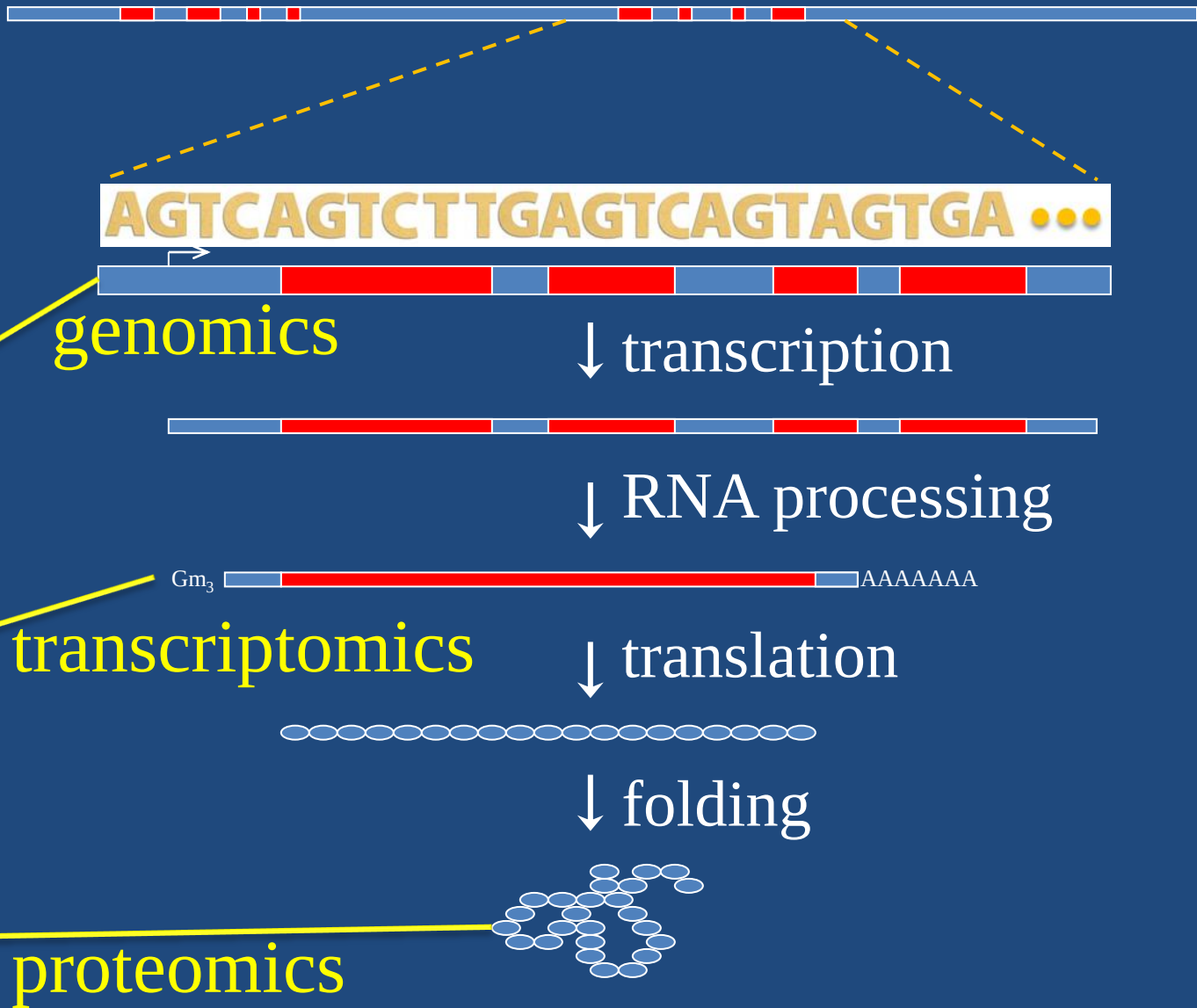
$3 * 10^9$ bp



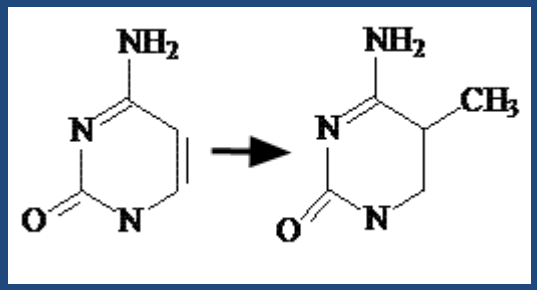
Central Dogma

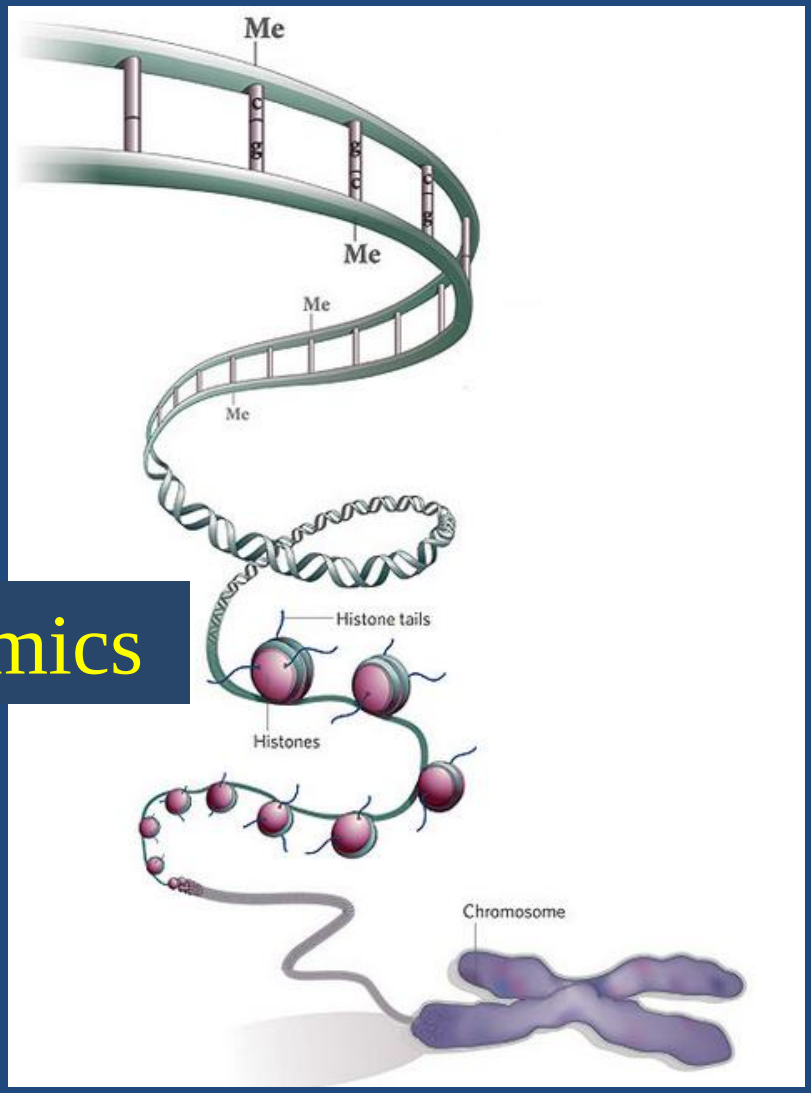




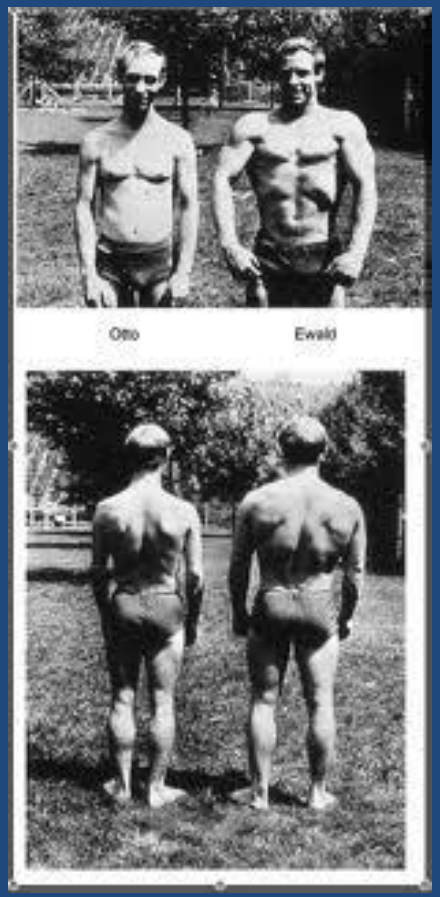


Methylated 5’ C ; Fifth base in DNA





epigenomics

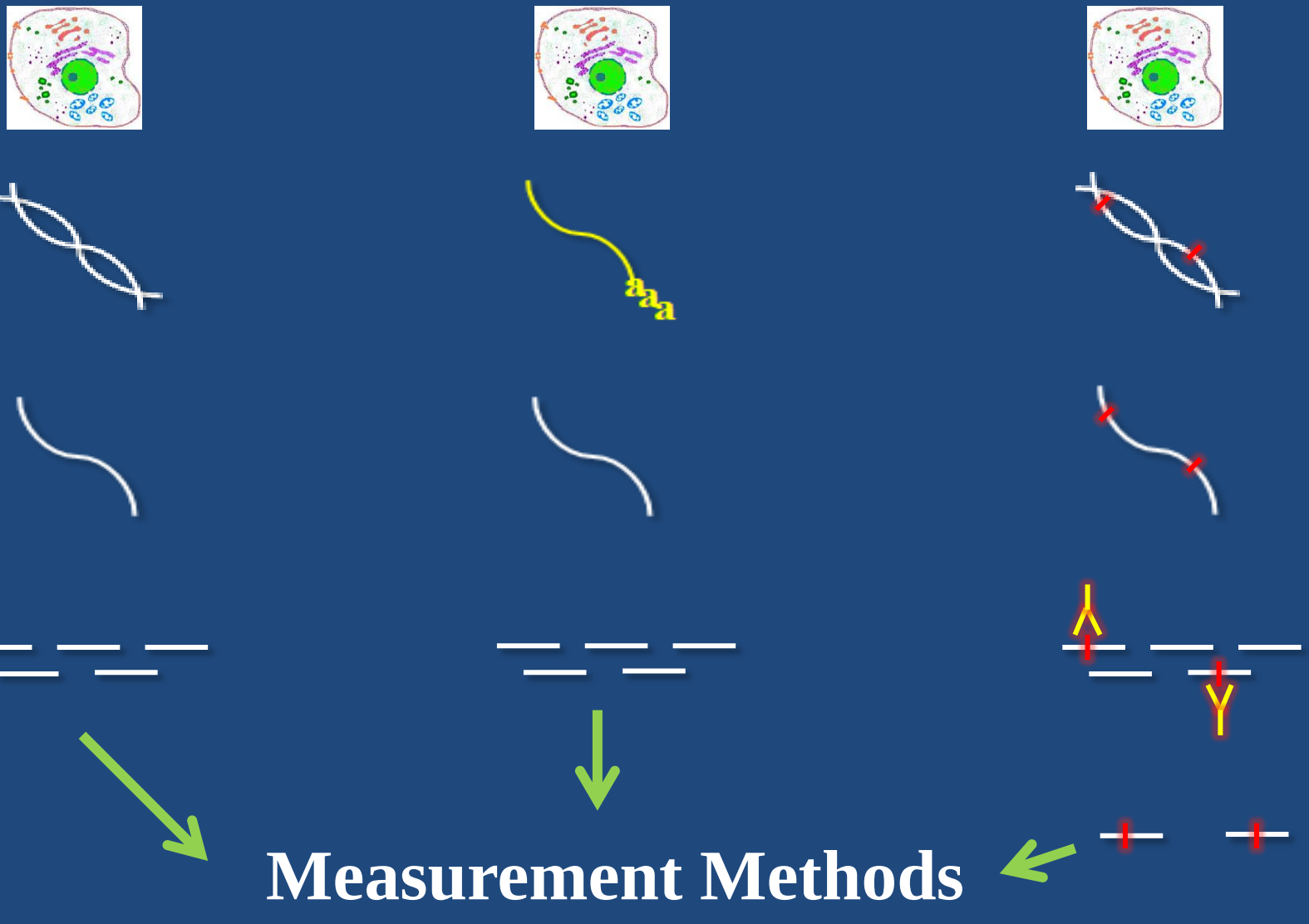


Where We Are

- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

How to get these biological data?

genomics transcriptomics epigenomics



Measurement Methods

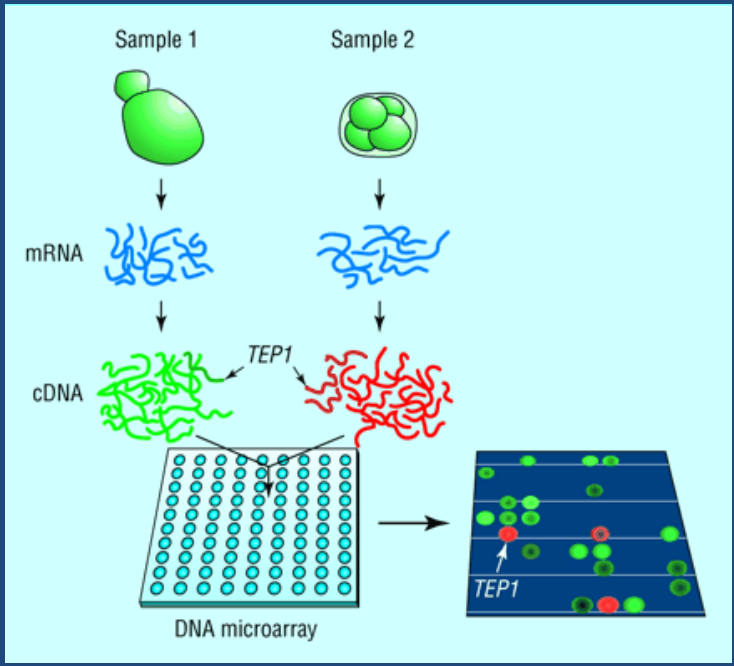
Measurement Methods

- Microarray

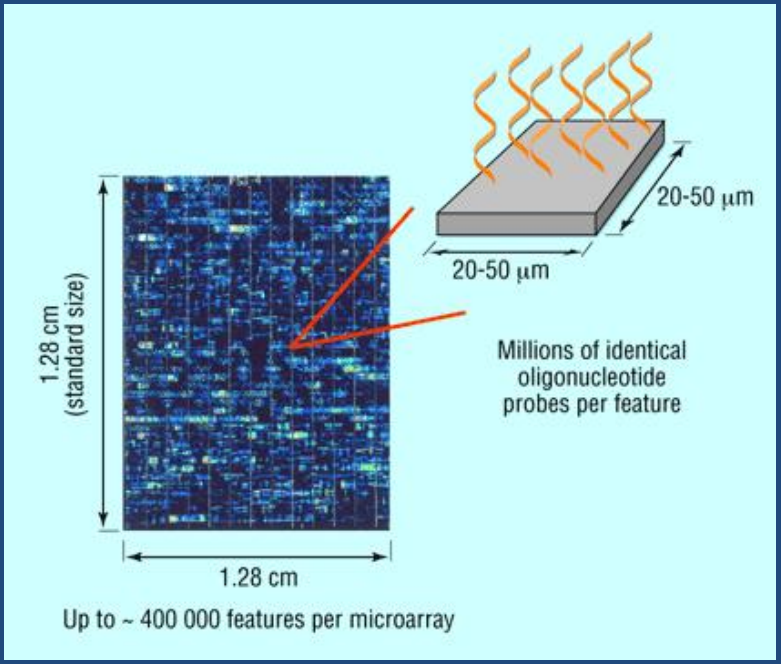
- DNA sequencer

Microarray

- Probe => Measurement
- Analogue signal
- oligonucleotide array VS two-color spotted array

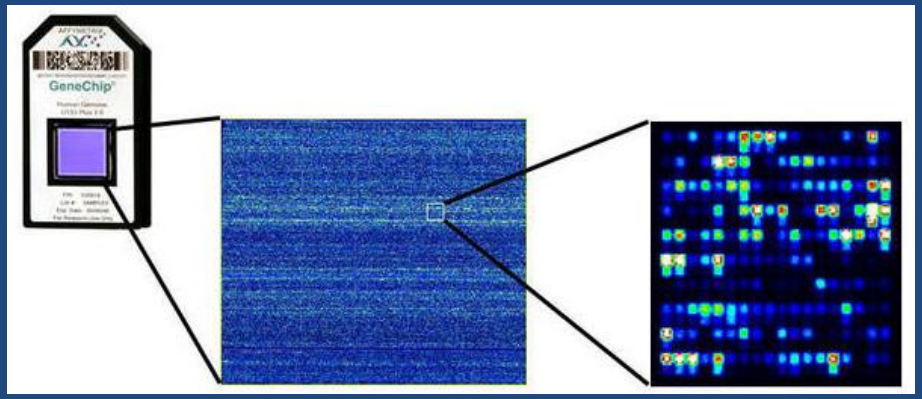


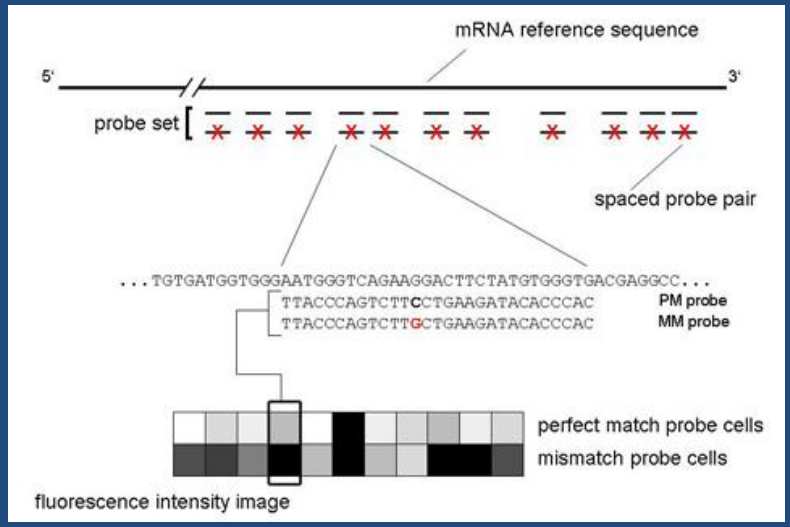
<two-color spotted array>



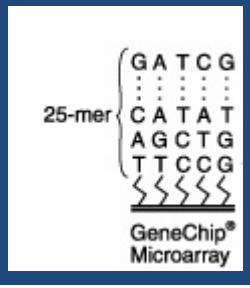
<oligonucleotide array>

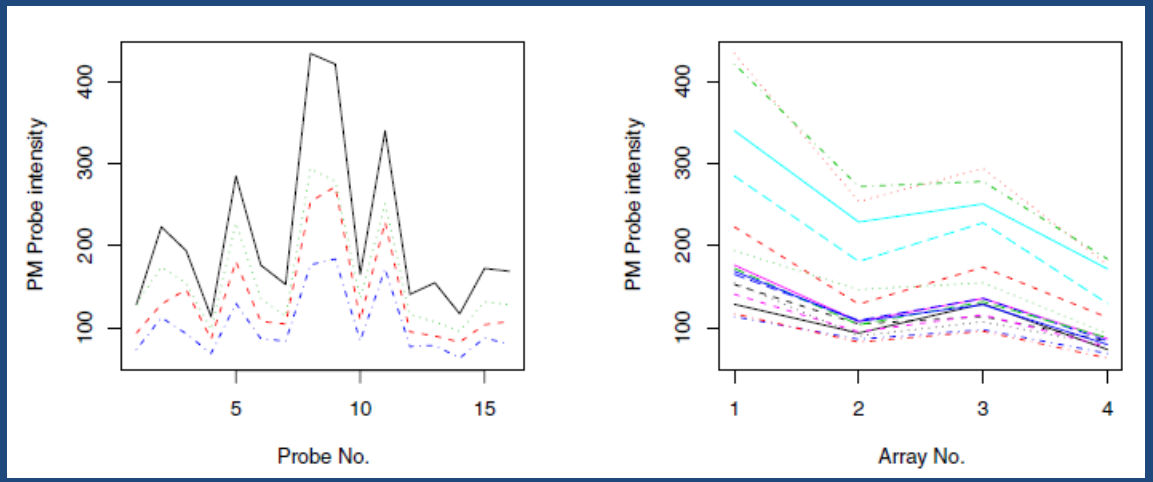
Preprocessing of Microarray Data

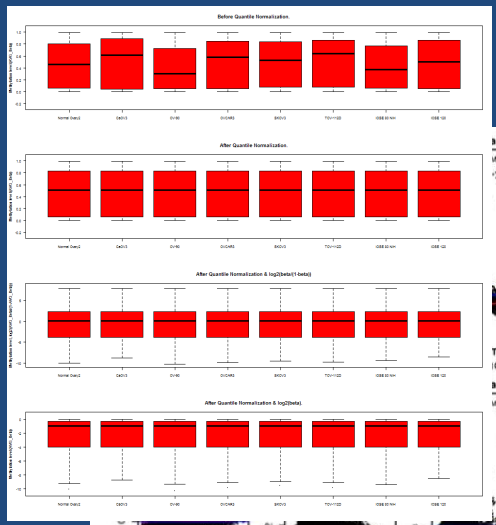




affy, affyPLM, affycomp, gcrma, affypdnn

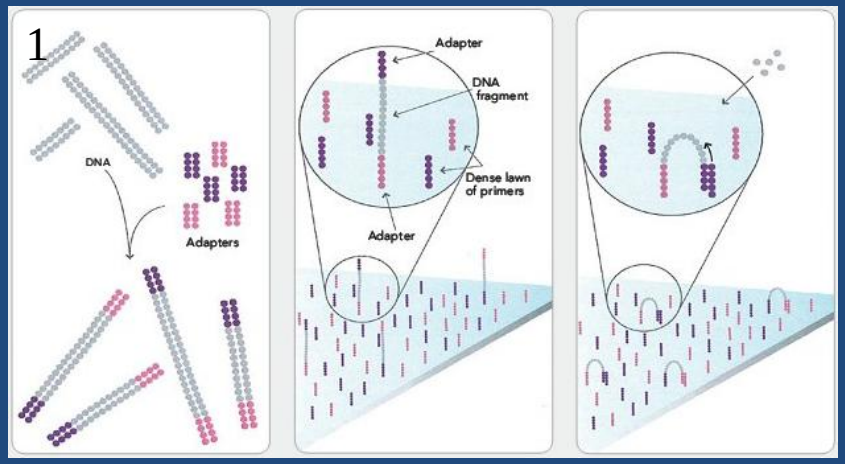


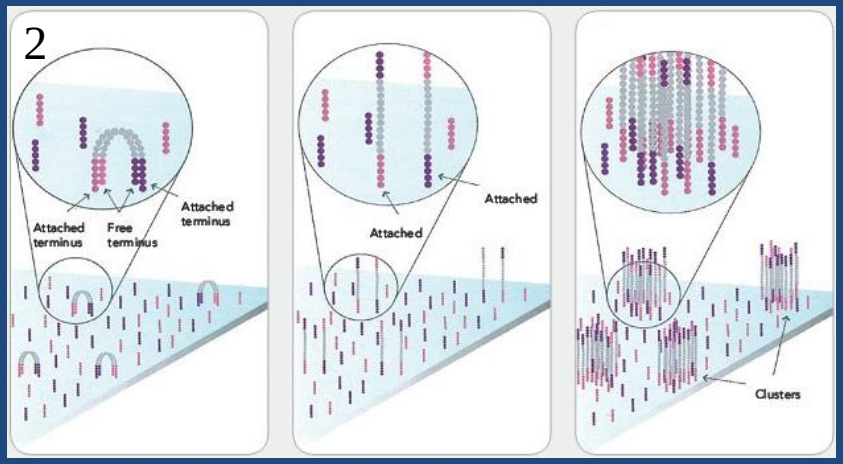


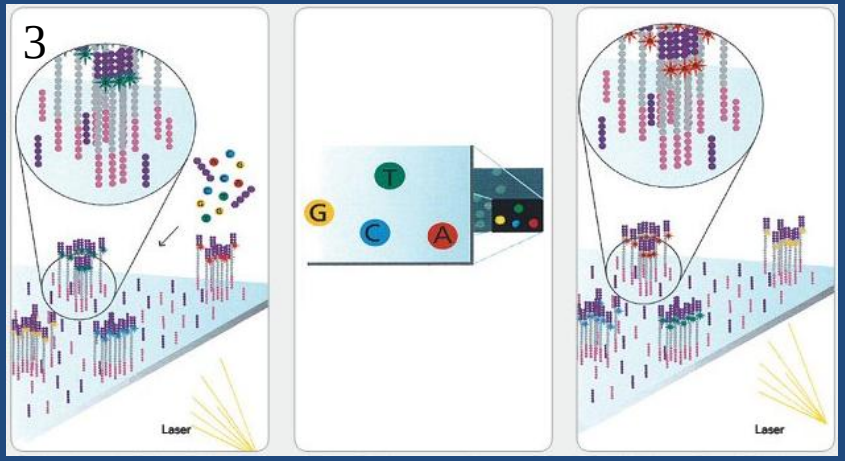


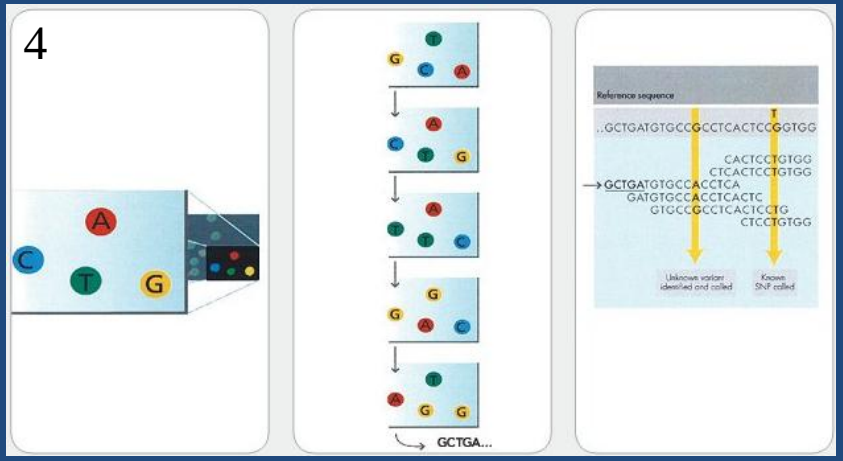
DNA sequencer

- Measurement => Probe
- Digital signal









Preprocessing of Sequence Data

@HWUSI-EAS052R_0029:3:1:4952:1104#1/1
CTATCATGATCACCAACATCACCATCACC
+HWUSI-EAS052R_0029:3:1:4952:1104#1/1
gggggfffefba_caB``abbbabaeeee

@HWUSI-EAS052R_0029:3:1:4952:1104#1/1
CTATCATGATCACCAACATCACCATCACCANGAACN
+HWUSI-EAS052R_0029:3:1:4952:1104#1/1
gggggfffefba_caB``abbbabaeeeecB_\ZaB

Fastq file format

Reference genome

```
@HD VN:1.3 SO:coordinate        } Header section
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

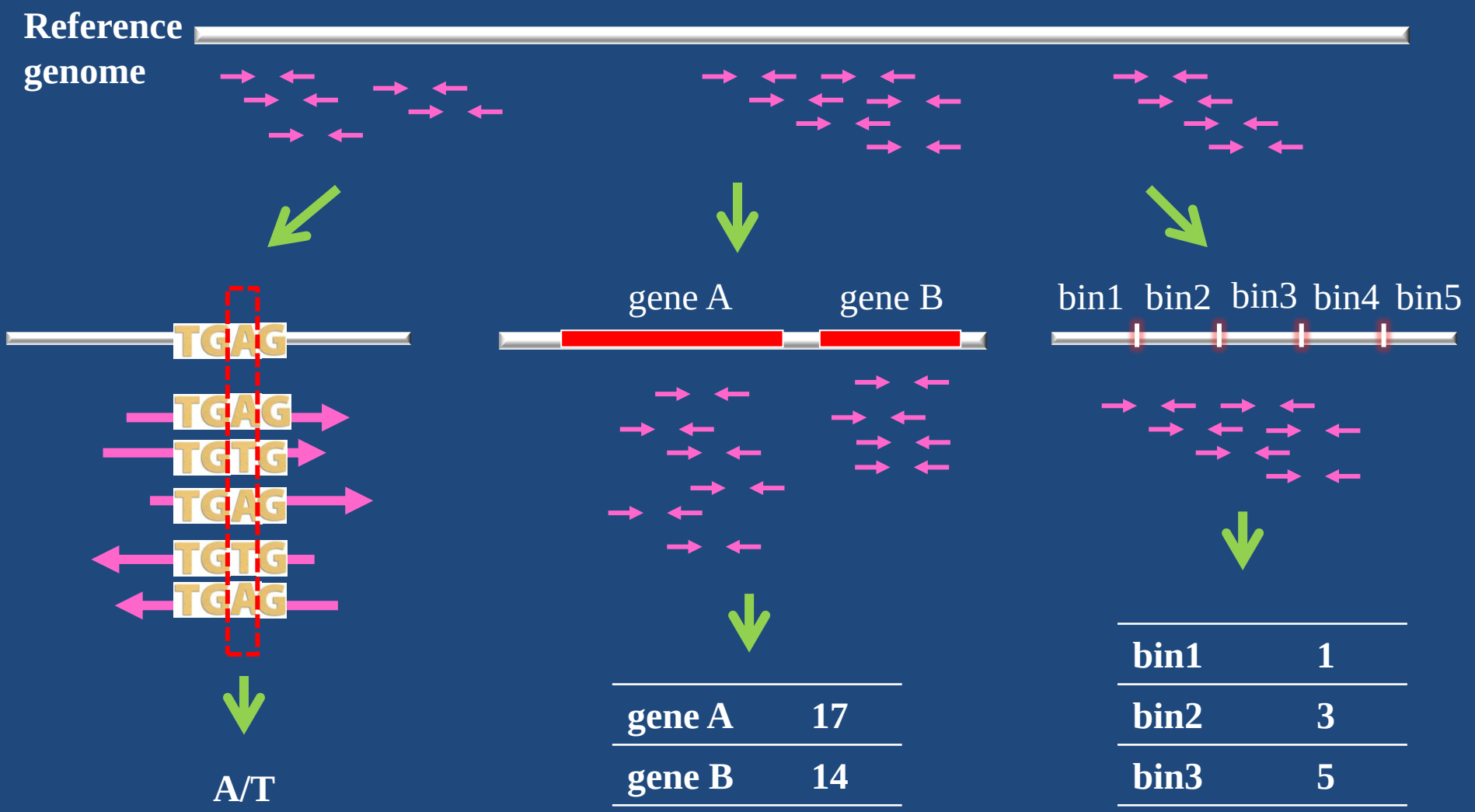
Alignment section

QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL

sam format

shortRead, Rsamtools

Data from Sequencer



Data Formats

- SNP (A, C, G, T) : categorical
  - Genotype (AA, AB, BB)
  - Sample x Genotype for each SNP locus
  - Genetic Model(dominant, recessive, co-dominant, additive, allele models)
- Expression
  - Signal to noise for each sample and probe (continuous)
- Methylation : categorical or continuous
  - Avg_beta, methylation level for each CpG site(continuous)
  - methylated/unmethylated (categorical) or hypo methylated, medium methylated, hyper methylated (categorical)

Where We Are

- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

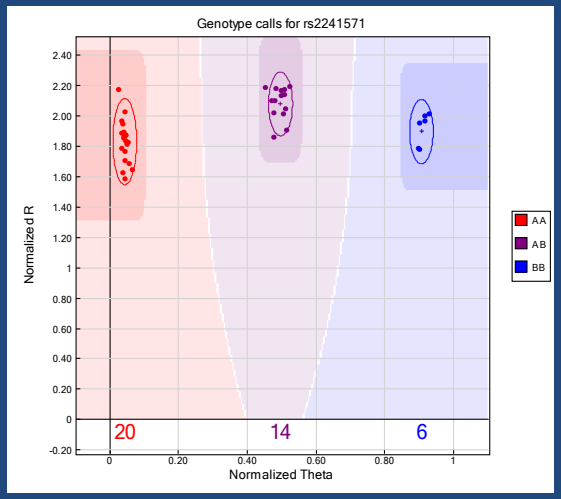
Genomics Data Analysis

A	A	A	SNP1	A	A	A
T	T	T	SNP2	T	T	T
G	G	G	SNP3	G	G	G
C	C	C	SNP4	A	A	A
G	G	G	SNP5	G	G	G
G	G	G	SNP6	G	G	G
C	C	C	SNP7	C	C	C
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
C	C	C	SNP8	C	C	C
T	T	T	SNP9	T	T	T
T	T	T	SNP10	T	T	T

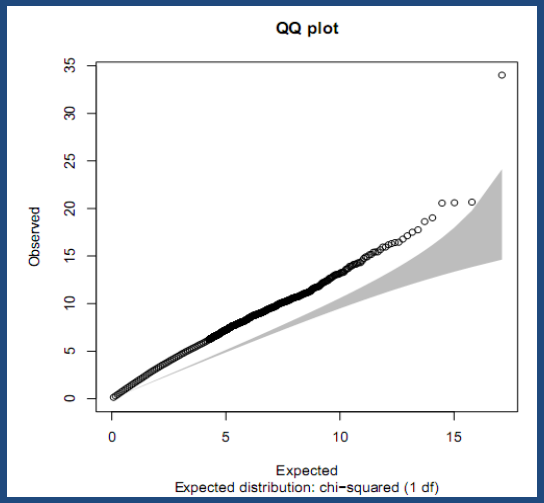
Genomics Data Analysis

→ SNPasoc, GenAbel

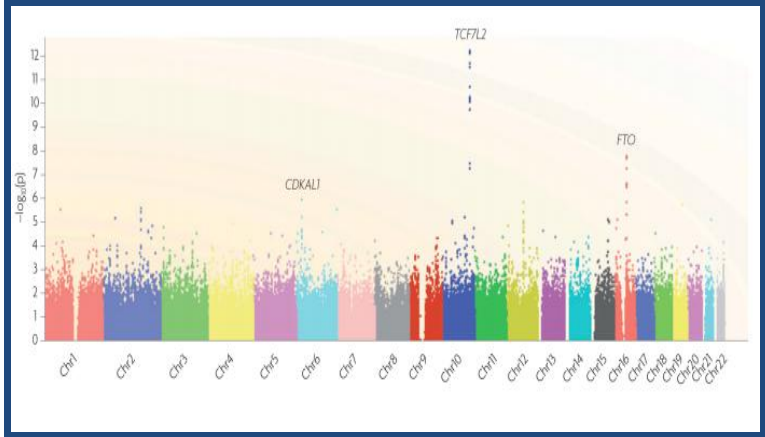
- Genotype clustering



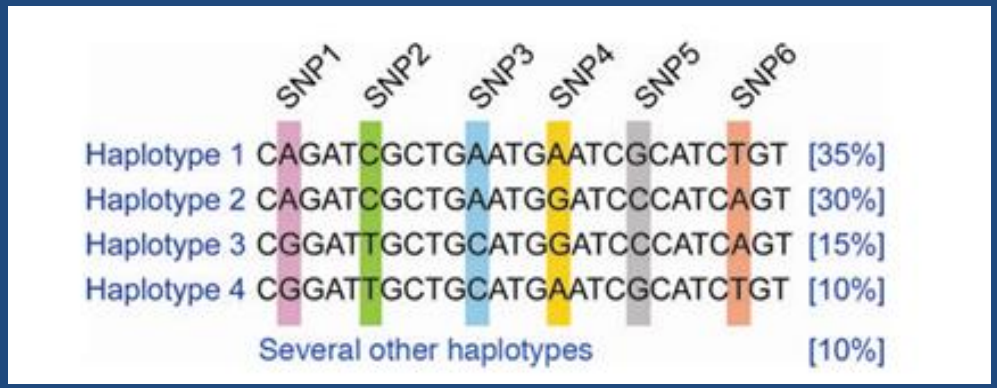
- HWE test

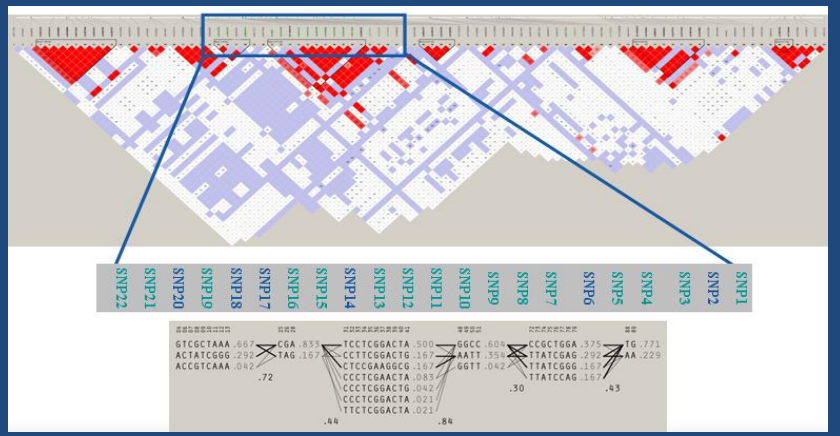


- Manhattan Plot

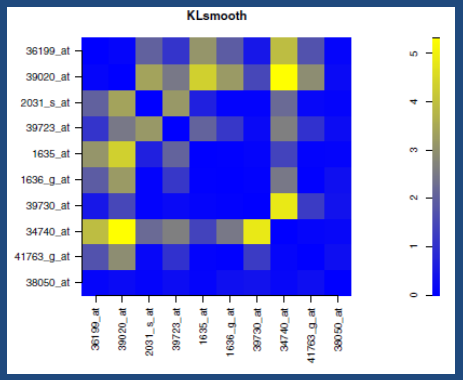


- LD & Haplotype

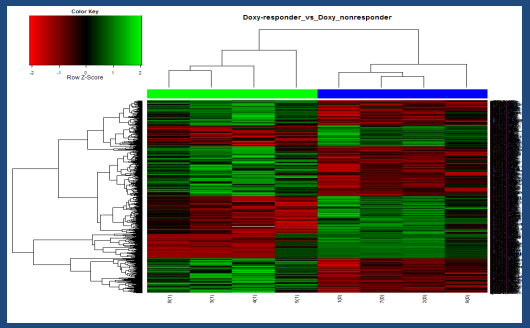




Transcriptomics/Epigenomic Data Analysis

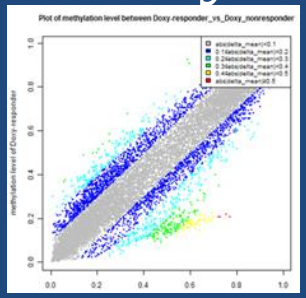


- Distance between Gene or Samples
  - Minkowski metric distance
  - Correlation-based distance
  - Distance between distributions
- Cluster analysis → cluster, hopach
  - Partitioning
  - Hierachical
  - Hybrid



- Analysis of Differential Gene Expression/methylation
  - parametric test
  - non-parametric test
  - multiple test adjustment

→ limma, multtest



Transcriptomics Data Analysis

normal

	Sample1	Sample2	Sample3
Gene A	13	14	15
Gene B			
.			
.			
Gene N			

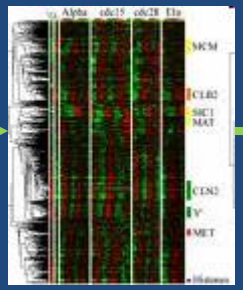
patient

	Sample1	Sample2	Sample3
Gene A	4	5	3
Gene B			
.			
.			
Gene N			

Non-specific filtering

	Sample 1	Sample 2	Sample 3
Gene A	13	14	15
Gene C			
.			
.			
Gene M			

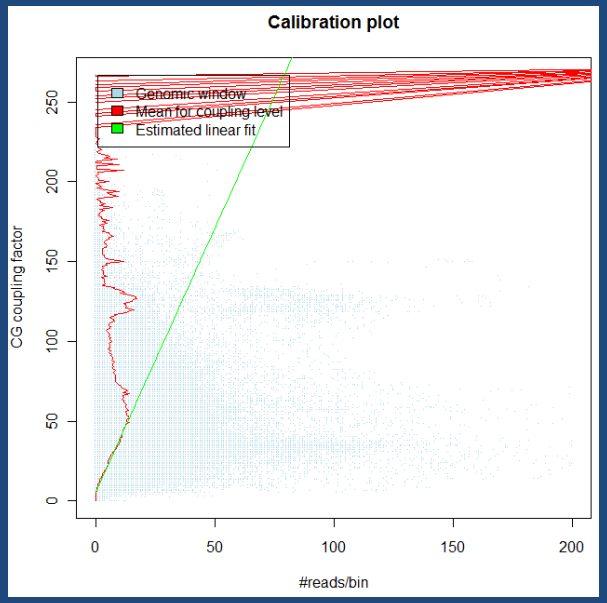
clustering

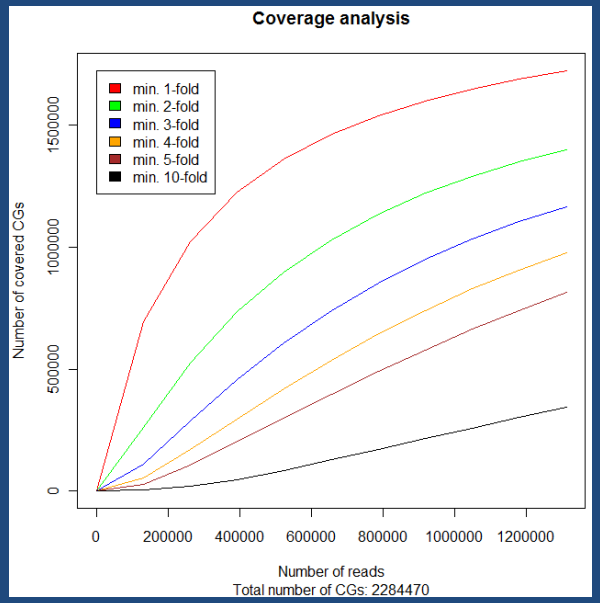


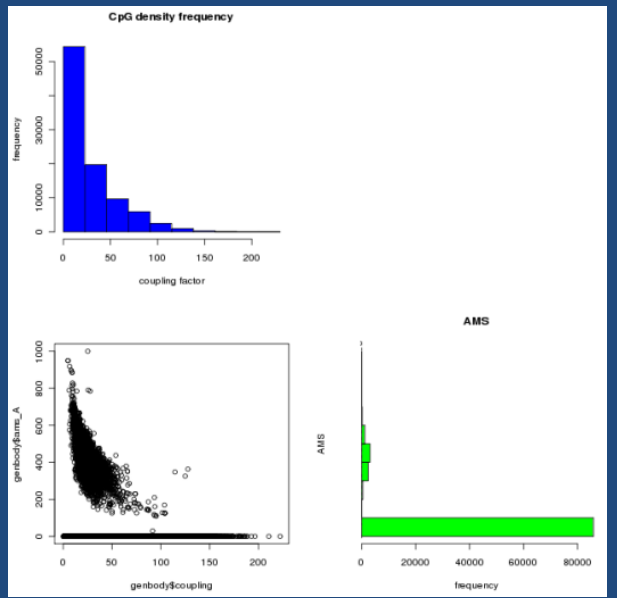
Finding DEG

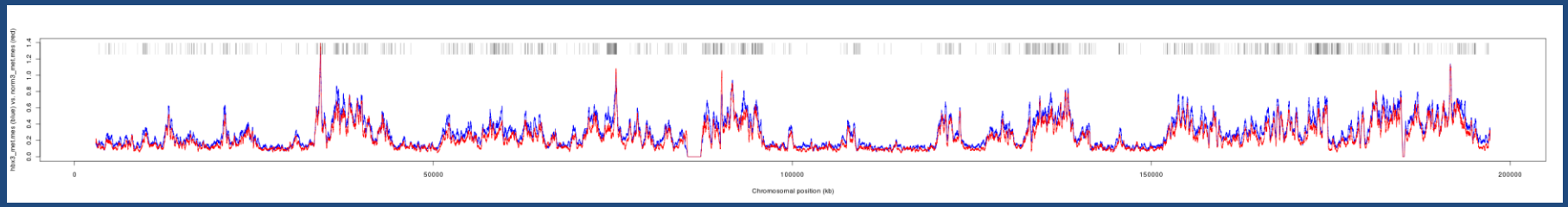
Gene K
Gene I
Gene M
.
.

Epigenomics Data Analysis



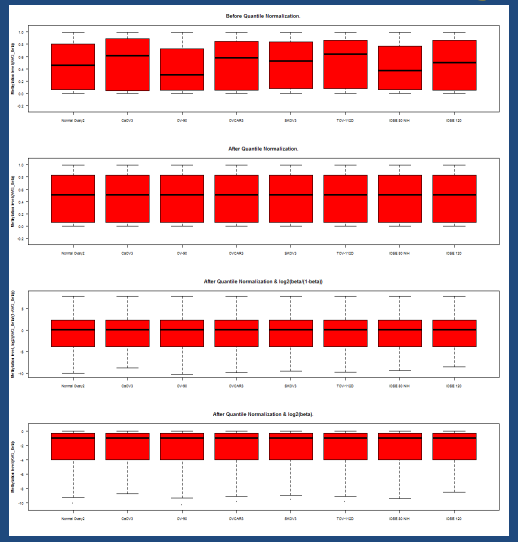


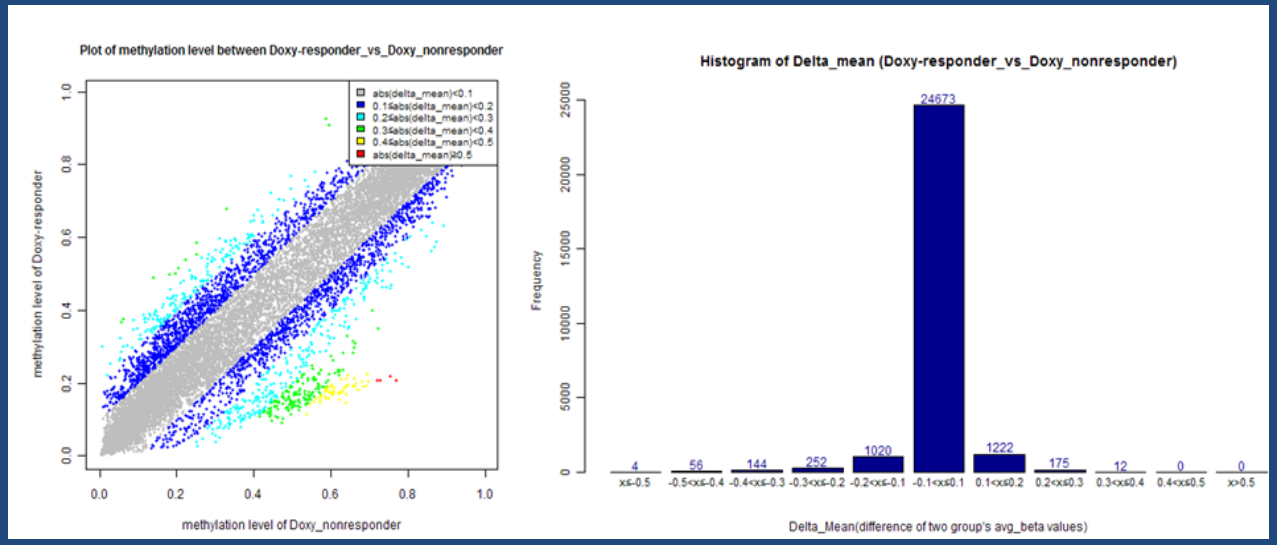


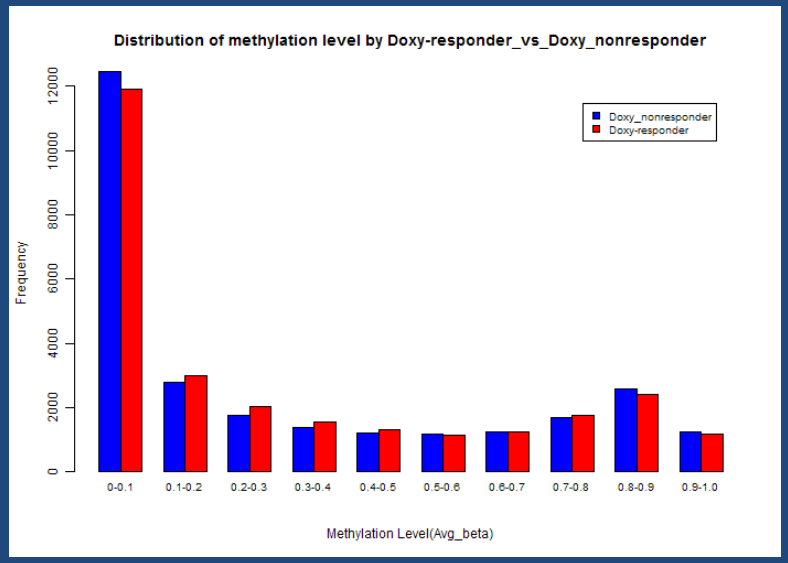


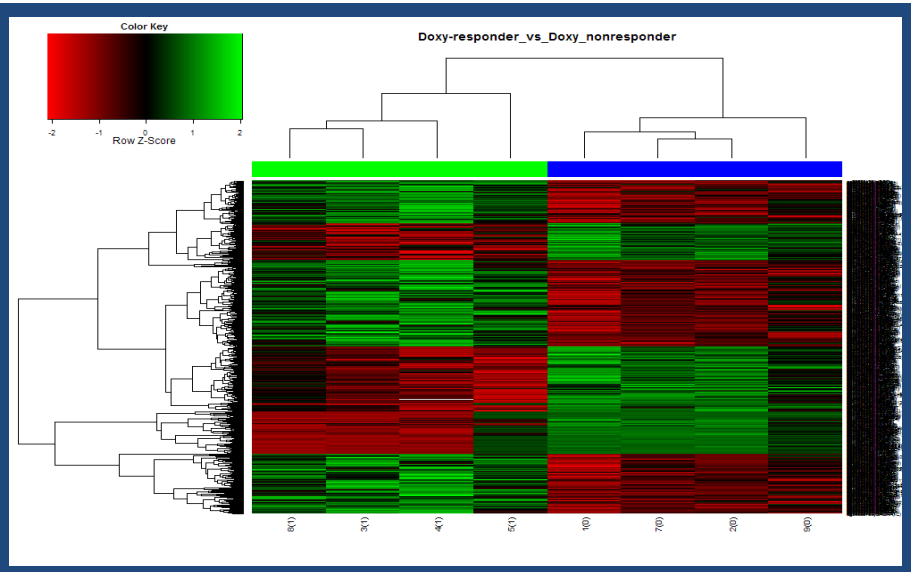
MEDIPS

Transcriptomics/Epigenomic Data Analysis

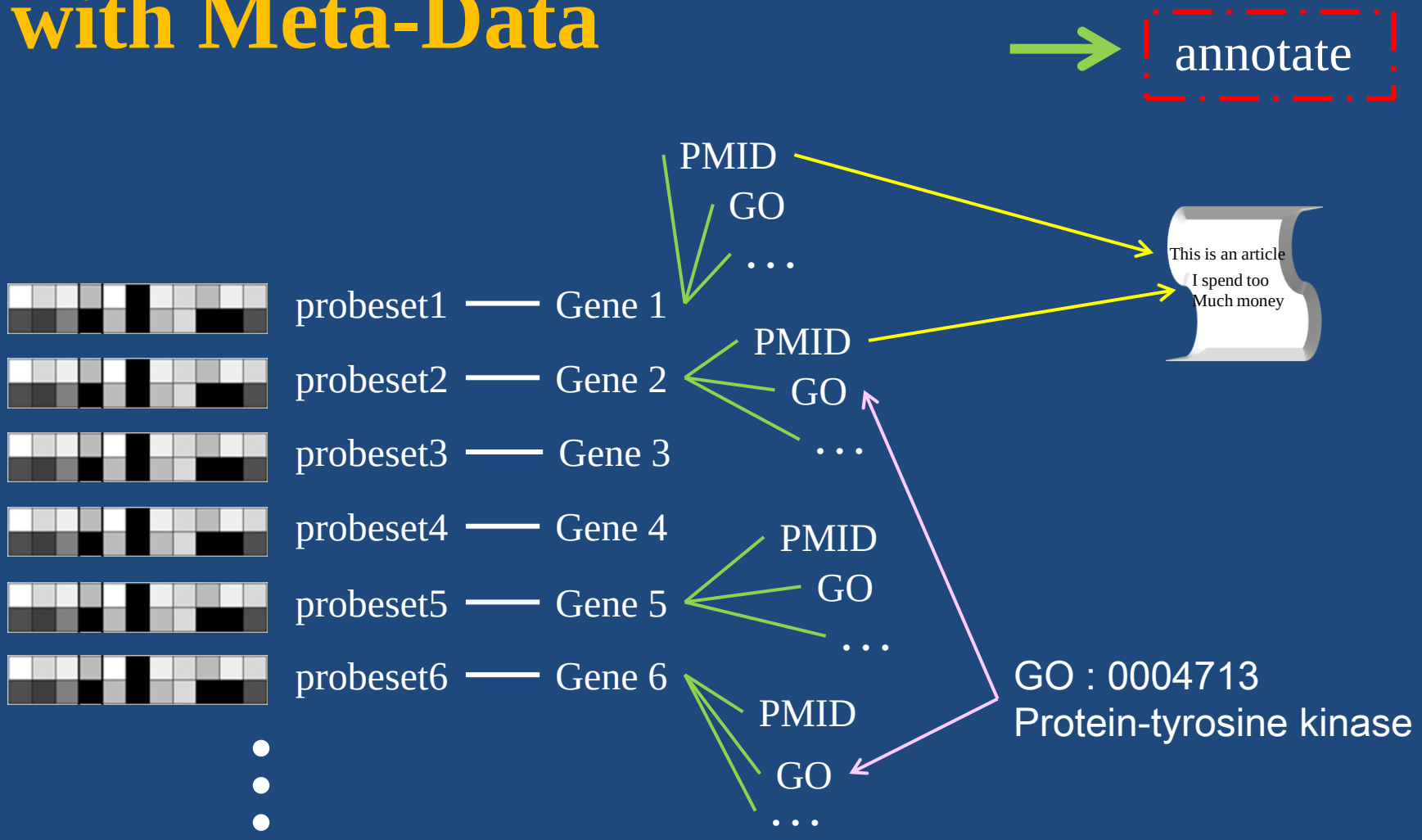








Association Genomic Data with Meta-Data



Where We Are

- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

Bioconductor

Method

Open Access

Bioconductor: open software development for computational biology and bioinformatics

Robert C Gentleman[1], Vincent J Carey[2], Douglas M Bates[3], Ben Bolstad[4], Marcel Dettling[5], Sandrine Dudoit[4], Byron Ellis[6], Laurent Gautier[7], Yongchao Ge[8], Jeff Gentry[1], Kurt Hornik[9], Torsten Hothorn[10], Wolfgang Huber[11], Stefano Iacus[12], Rafael Irizarry[13], Friedrich Leisch[9], Cheng Li[1], Martin Maechler[5], Anthony J Rossini[14], Gunther Sawitzki[15], Colin Smith[16], Gordon Smyth[17], Luke Tierney[18], Jean YH Yang[19] and Jianhua Zhang[1]

Addresses: [1]Department of Biostatistical Science, Dana-Farber Cancer Institute, 44 Binney St, Boston, MA 02115, USA. [2]Channing Laboratory, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA. [3]Department of Statistics, University of Wisconsin-Madison, 1210 W Dayton St, Madison, WI 53706, USA. [4]Division of Biostatistics, University of California, Berkeley, 140 Warren Hall, Berkeley, CA 94720-7360, USA. [5]Seminar for Statistics LEO C16, ETH Zentrum, Zürich CH-8092, Switzerl. [6]Department of Statistics, Harvard University, 1 Oxford St, Cambridge, MA 02138, USA. [7]Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, Lyngby 2800, Denmark. [8]Department of Biomathematical Sciences, Mount Sinai School of Medicine, 1 Gustave Levy Place, Box 1023, New York, NY 10029, USA. [9]Institut für Statistik und Wahrscheinlichkeitstheorie, TU Wien, Wiedner Hauptstrasse 8-10/1071, Wien 1040, Austria. [10]Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Waldstraße6, D-91054 Erlangen, Germany. [11]Division of Molecular Genome Analysis, DKFZ (German Cancer Research Center), 69120 Heidelberg, German. [12]Department of Economics, University of Milan, 23 Via Mercalli, I-20123 Milan, Italy. [13]Department of Biostatistics, Johns Hopkins University, 615 N Wolfe St E3035, Baltimore, MD 21205, USA. [14]Department of Medical Education and Biomedical Informatics, University of Washington, Box 357240, 1959 NE Pacific, Seattle, WA 98195, USA. [15]Statistisches Labor, Institut für Angewandte Mathematik, Im Neuenheimer Feld 294, D 69120, Heidelberg, Germany. [16]Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, TPC-28, La Jolla, CA 92037, USA. [17]Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia. [18]Department of Statistics and Actuarial Science, University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242, USA. [19]Center for Bioinformatics and Molecular Biostatistics, Univerisity of California, San Francisco, 500 Parnassus Ave, San Francisco 94143-0560, USA.

Correspondence: Robert C Gentleman. E-mail: rgentlem@jimmy.harvard.edu

Published: 15 September 2004

Genome ***Biology*** 2004, **5**:R80

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/10/R80

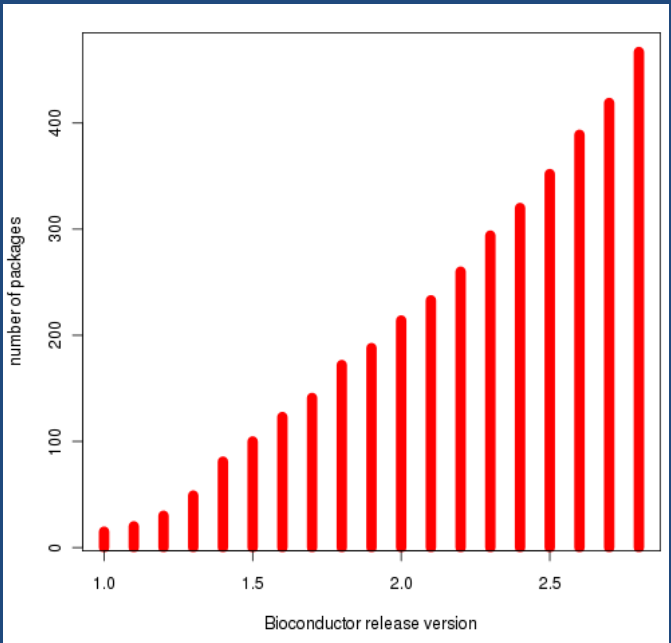
Received: 19 April 2004
Revised: 1 July 2004
Accepted: 3 August 2004

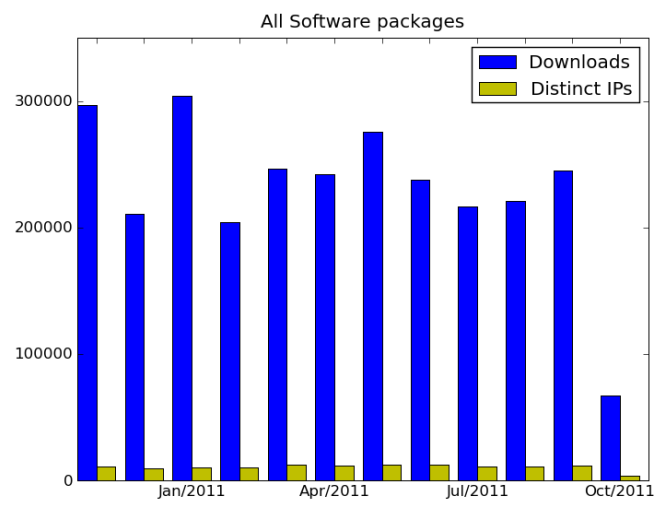
Purpose

Merit

expreSet class

Bioconductor





Bioconductor version 2.8 (Release)

- AnnotationData (593)
  - ChipManufacturer (356)
  - ChipName (190)
  - CustomArray (2)
  - CustomCDF (16)
  - CustomDBSchema (10)
  - FunctionalAnnotation (10)
  - Organism (424)
  - SequenceAnnotation (2)
- ExperimentData (82)
  - Cancer (18)
  - ChIPchipData (1)
  - ChIPseqData (3)
  - EColiData (1)
  - FlowCytData (1)
  - HapMap (7)
  - HighThroughputSequencingData (1)
  - HIV (1)
  - MassSpectrometryData (1)
  - NormalTissue (1)
  - RNAExpressionData (1)
  - RNAseqData (1)
  - StemCells (1)
  - Yeast (9)
- Software (467)
  - Annotation (61)
  - AssayDomains (182)
  - AssayTechnologies (290)
  - Bioinformatics (273)
  - BiologicalDomains (46)
  - Infrastructure (194)

Where We Are

- Biological background
- Methods for biological data measurement
- Example of biological data analysis
- Bioconductor
- Trend in Genomics

Why is genomics highlighted?

2001

Human Genome Project

2007

Next Generation Sequencer

1st generation sequencer

next generation sequencer

2011 ..

And 3rd …

Genomics in Business

Thank you

GATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTCTAGAG
CCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTGCTT
TCCACGACGGTGACACGCTTCCCTGGATTGGgtaagctcctgactgaacttgatgagtcctctctgagtcacgggctctcggctccgt
gtattttcagctcgggaaaatcgctggggctgggggtggggcagtggggacttagcgagtttgggggtgagtgggatggaagcttggctagagggatcatc
ataggagttgcattgttgggagacctgggtgtagatgatgggatgttaggaccatccgaactcaaagttgaacgcctaggcagaggagtggagctttggg
gaaccttgagccggcctaaagcgtacttctttgcacatccacccgtgctgggcgtaggaatccctgaaataaaagatgcacaaagcattgaggtctga
gacttttggatctcgaaacattgagaactcatagctgtatattttagagcccatggcatcctagtgaaaactgggctccattccgaaatgatcatttgggggtg
atccggggagcccaagctgctaaggtcccacaacttccggaccttgtccttcctggagcgatctttccaggcagccccggctccgctagatggagaaaa
tccaattgaaggctgtcagtcgtggaagtgagaagtgctaaaccaggggtttgcccgccaggccgaggaggaccgtcgcaatctgagaggcccggcag
ccctgttattgtttggctccacatttacatttctgcctcttgcagcagcatttccggtttctttttgccggagcagctcactattcacccgatgagaggggaggaga
gagagagaaaatgtcctttaggccggttcctcttacttggcagagggaggctgctattctccgcctgcatttctttttctggattacttagttatggcctttgcaaag
gcaggggtatttgttttgatgcaaacctcaatccctccccttctttgaatggtgtgccccaccccgcgggtcgcctgcaacctaggcggacgctaccatggcg
tgagacagggagggaaagaagtgtgcagaaggcaagcccggaggtattttcaagaatgagtatatctcatcttcccggaggaaaaaaaaaaaagaatg
ggtacgtctgagaatcaaattttgaaagagtgcaatgatgggtcgtttgataatttgtcggaaaaacaatctacctgttatctagctttgggctaggccattcca
gttccagacgcaggctgaacgtcgtgaagcggaaggggcgggcccgcaggcgtccgtgtggtcctccgtgcagcccctccggcccgagccggttcttcct
ggtaggaggcggaactcgaattcatttctcccgctgccccatctcttagctcgcggttgtttcattccgcagtttcttcccatgcacctgccgcgtaccggccact
ttgtgccgtacttacgtcatctttttcctaaatcgaggtggcatttacacacagcgccagtgcacacagcaagtgcacaggaagatgagtttttggcccctaac
cgctccgtgatgcctaccaagtcacagacccttttcatcgtcccagaaacgtttcatcacgtctcttcccagtcgattcccgacccccacctttattttgatctccat
aaccattttgcctgttggagaacttcatatagaatggaatcaggctgggcgctgtggctcacgcctgcactttgggaggccgaggcgggcggattacttgag
gataggagttccagaccagcgtggccaacgtggtgaatccccgtctctactaaaaaatacaaaaattagctgggcgtggtgggtgcctgtaatcccagcta
ttcgggagggtgaggcaggagaatcgcttgaacccgggaggcagaggttgcagtgagccaagatcgtgccactacactccagcctgggcgacaaga
acgaaactccgtctcaaaaaaaaggggggaatcatacattatgtgctcatttttgtcgggcttctgtccttcaatgtactgtctgacattcgttcatgttgtatatat
cagtattttgctccttttcatttagtatagtccatcgattgtatatccgtccttttgatggccttttgagttgtttcccatttgcggttatgaaataaagctgctataaacatt
cttgtacaattcttttgtgatcatatgttttcgtgtttcttggagaaatacttaggaggggaattgcgagtttggaagtaaaaagtagctgtattttgaacttttcaga
a