Introduction to R for Bioinformatics

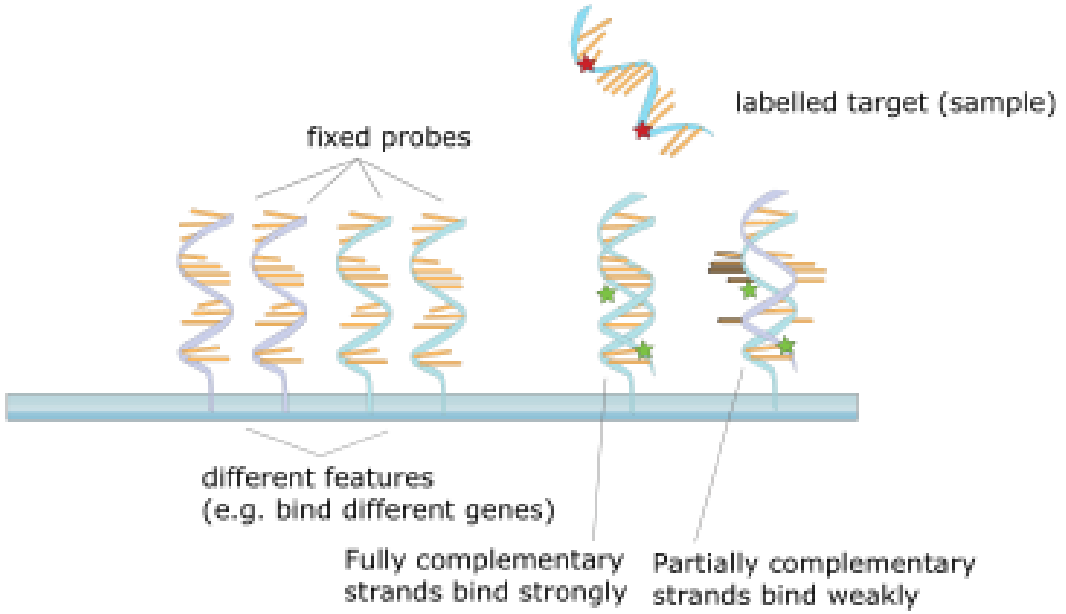
VINCENZO LAGANI

04 MAY 2017

Outline

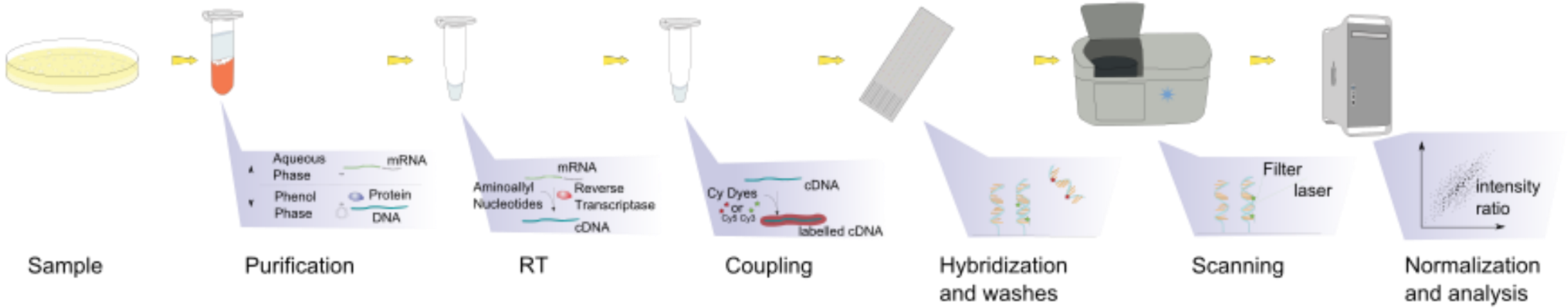
- Microarray technology
- Preprocessing microarray data: the Robust Multy-Array Average (RMA) method
- Microarry in R: the affy and oligo packages
- Introduction to Bioconductor
- Annotation packages for microarray
  - CDF files (altcdfenvs package)
  - DB packages
  - Differences between Bioconductor and microarray

Microarray technology



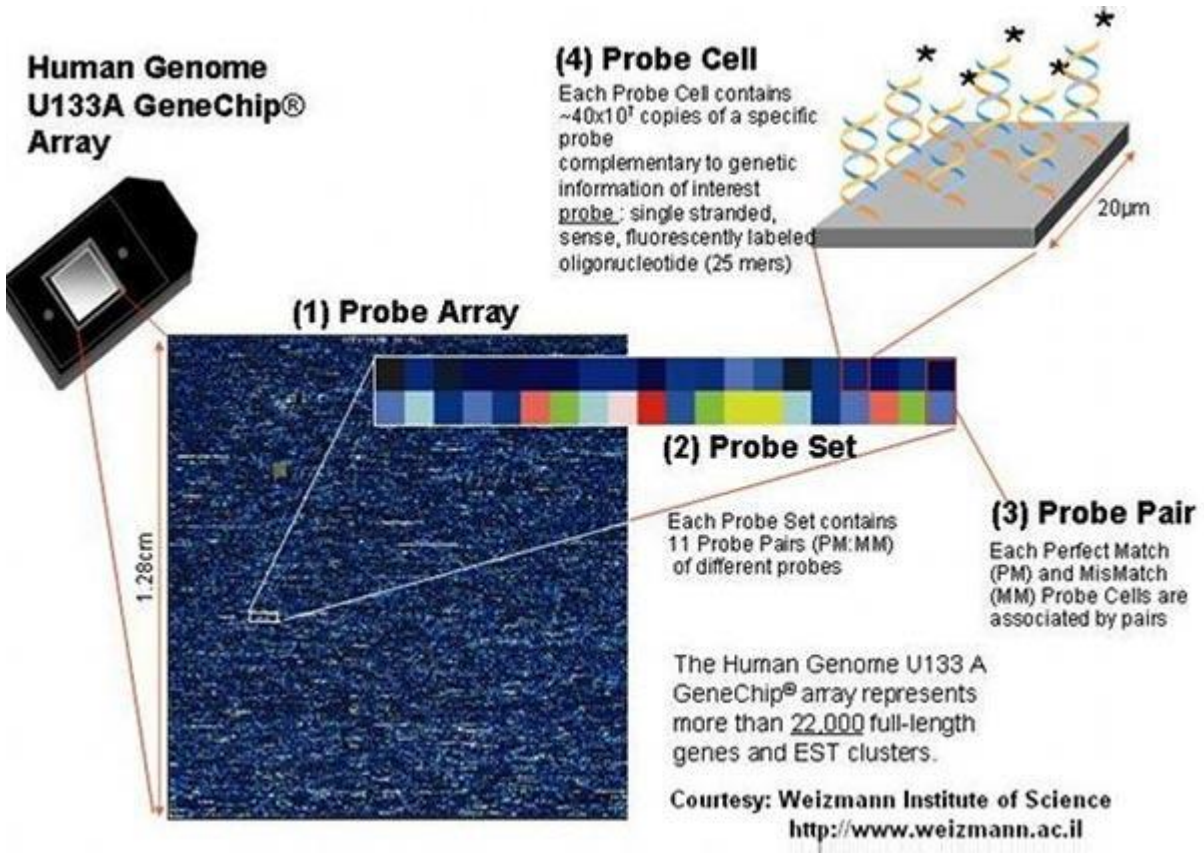
Source: https://en.wikipedia.org/

Sample preparation for microarray

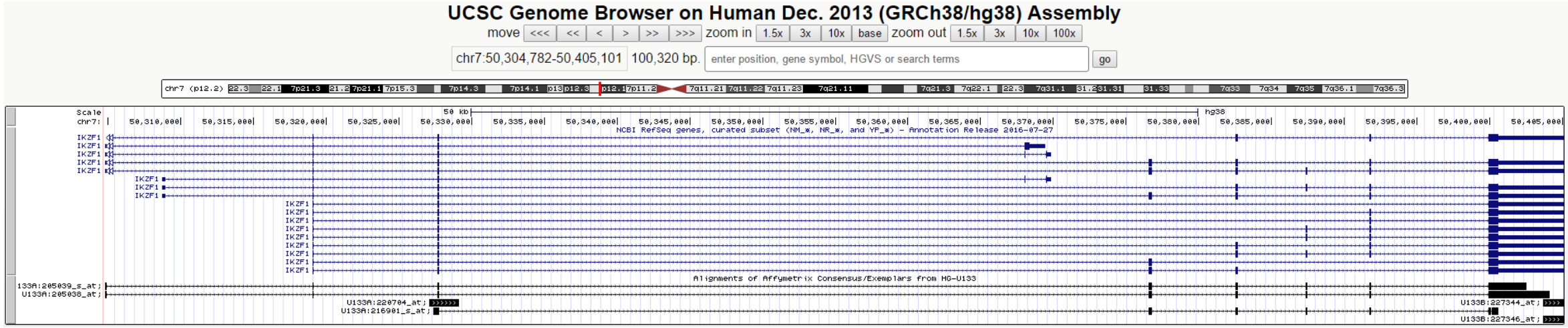


Source: https://en.wikipedia.org/

Probes and probesets



Probesets mapping on genome



Probesets – genes associations

- Most of the probesets measure a single gene
  - Difficult to distinguish transcripts
- Some probesets measure multiple or no genes
- The same gene can be measured by multiple probesets
  - Not in brainarray

Preprocessing microarray data

- We need to quantify the level of expression based on fluorescence intensity

- The Robust Multy-Array Average (RMA) algorithm is one of the most used methods
  - Irizarry, RA; Hobbs, B; Collin, F; Beazer-Barclay, YD; Antonellis, KJ; Scherf, U; Speed, TP (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data.". *Biostatistics*. **4** (2): 249–64

The Robust Multy-Array Average (RMA) method

- Background correction
- Log2 transform
- Quantile normalization

- Performed at probe level
- Perfect Match (PM) probes only

- Summarization (median polish)

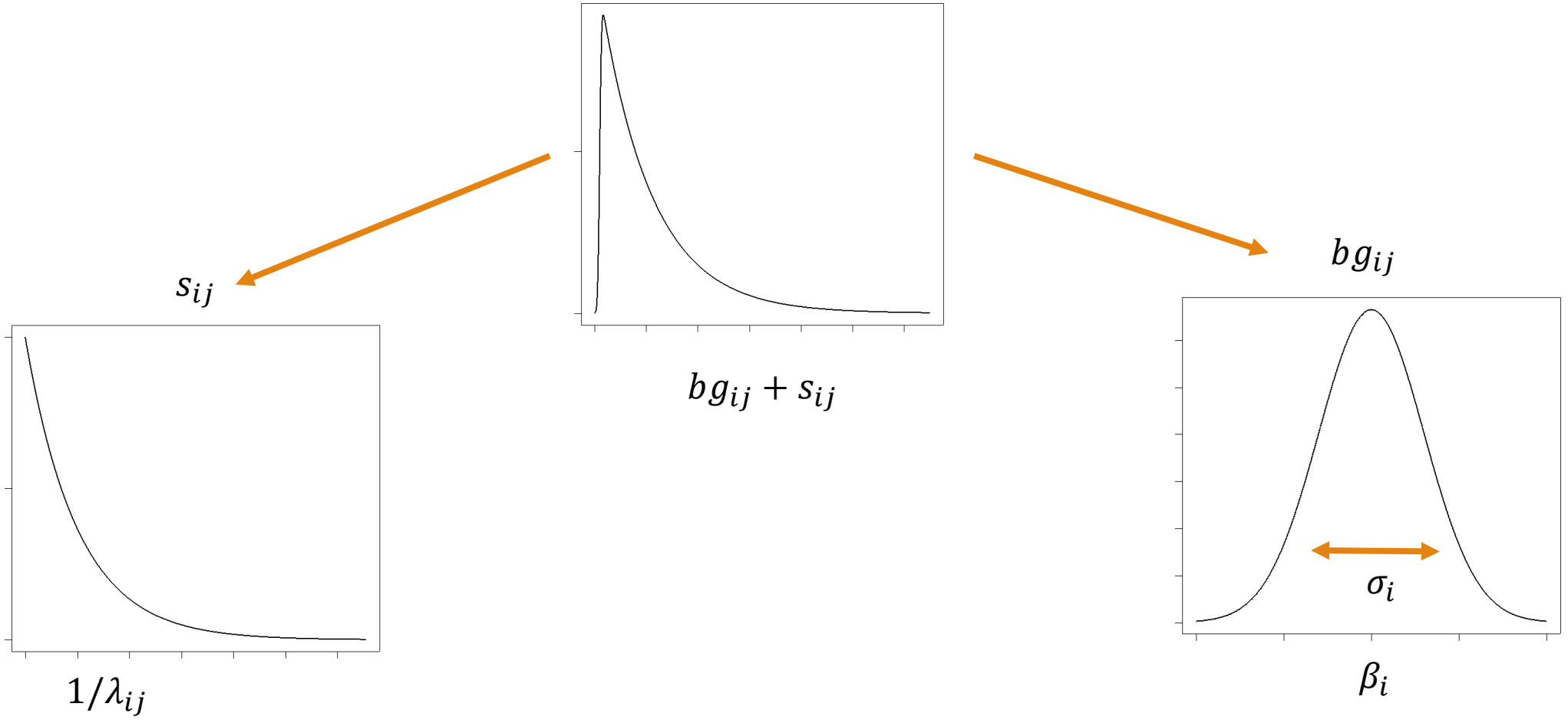
Summarizing probes into probesets

RMA background correction

- $PM_{ij} = \underbrace{bg_{ij}}_{\text{Background noise for probe j in sample i}} + s_{ij}$ } Signal for probe j in sample i

- We assume that $s_{ij} \sim \exp(\lambda_{ij})$ and $bg_{ij} \sim N(\beta_i, \sigma_i)$

Signal and background distributions



Probe expected signal

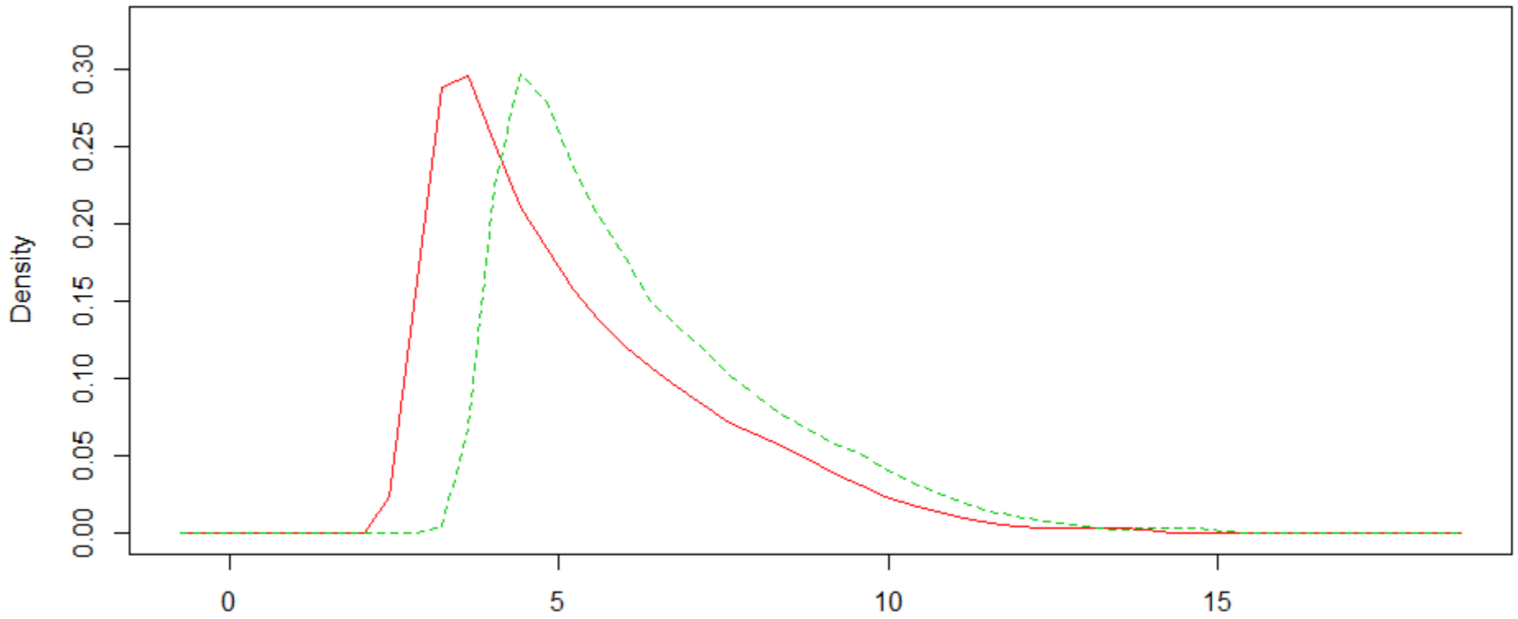
- $E(s_{ij}|PM_{ij}) = PM_{ij} - \beta_i - \lambda_{ij}\sigma^2 + \sigma \cdot \frac{\phi\left(\frac{PM_{ij}-\beta_i-\lambda_{ij}\sigma^2}{\sigma}\right)-\phi\left(\frac{\beta_i+\lambda_{ij}\sigma^2}{\sigma}\right)}{\Phi\left(\frac{PM_{ij}-\beta_i-\lambda_{ij}\sigma^2}{\sigma}\right)-\Phi\left(\frac{\beta_i+\lambda_{ij}\sigma^2}{\sigma}\right)}$

- Where $\phi()$ and $\Phi()$ are the normal density and distribution functions, respectively

- Complex formula, however easy to apply once $\beta_i, \lambda_{ij}, \sigma^2$ are estimated
  - The estimation is based on an empirical procedure

Quantile normalization

- Separate arrays can produce different results even if the samples are identical
  - This may be due to differences in the sample preparation or during hybridization



Quantile normalization example

- Table of log2 transformed probes values

	A1	A2	A3
P1	4	5	2
P2	7	4	8
P3	3	8	5
P4	9	7	6
P5	6	3	4

Quantile normalization example

- Probes are sorted within each sample

	A1	A2	A3
Q1	P3:3	P5:3	P1:2
Q2	P1:4	P2:4	P5:4
Q3	P2:7	P1:5	P3:5
Q4	P5:6	P4:7	P4:6
Q5	P4:9	P3:8	P2:8

Quantile normalization example

- Computing the average value for each quantile

	A1	A2	A3	Average
Q1	P3:3	P5:3	P1:2	2.7
Q2	P1:4	P2:4	P5:4	4
Q3	P2:7	P1:5	P3:5	5.7
Q4	P5:6	P4:7	P4:6	6.3
Q5	P4:9	P3:8	P2:8	8.3

Quantile normalization example

- Ensuring equal values across quantiles

	A1	A2	A3	Average
Q1	P3:2.7	P5:2.7	P1:2.7	2.7
Q2	P1:4	P2:4	P5:4	4
Q3	P2:5.7	P1:5.7	P3:5.7	5.7
Q4	P5:6.3	P4:6.3	P4:6.3	6.3
Q5	P4:8.3	P3:8.3	P2:8.3	8.3

Quantile normalization example

- Reordering probes values

	A1	A2	A3
P1	4	5.7	2.7
P2	5.7	4	8.3
P3	2.7	8.3	5.7
P4	8.3	6.3	6.3
P5	6.3	2.7	4

- Maximum, minimum, median values across arrays are now equal (as well as all other quantiles)

Summarization: median polish

- We now want to obtain a single value for each single probeset
  - Simple solution: taking the mean or median value across probes

- Unfortunately, normalized probes values still suffer of “probe affinity effect”:
- $PM_{ij}^{*} = \mu_i + \alpha_j + \epsilon_{ij}$, where $\sum a_j = 0$ within each probeset

- We must correct for this bias with the median polish algorithm

Median polish example

- Computing and subtracting row medians

	P1	P2	P3	P4	P5	Median
A1	3.8	7.5	6.7	4.3	4.5	4.5
A2	7	4	5.3	6.3	3.8	5.3
A3	2.4	8.2	5.7	9.3	4.2	5.7

	P1	P2	P3	P4	P5
A1	-0.7	3	2.2	-0.2	-0.7
A2	1.7	-1.3	0	1	1.7
A3	-3.3	2.5	0	3.6	-3.3

Median polish example

- Computing and subtracting column medians

	P1	P2	P3	P4	P5
A1	-0.7	3	2.2	-0.2	-0.7
A2	1.7	-1.3	0	1	1.7
A3	-3.3	2.5	0	3.6	-3.3
Median	-0.7	2.5	0	1	-1.5

	P1	P2	P3	P4	P5	Median
A1	0	0.5	2.2	-1.2	1.5	0.5
A2	2.4	-3.8	0	0	0	0
A3	-2.6	0	0	2.6	0	0
median	0	0	0	0	0	

- The algorithm converges when both column and row medians are zero.

Median polish example

- Computing the final values

	P1	P2	P3	P4	P5
A1	3.8	7.5	6.7	4.3	4.5
A2	7	4	5.3	6.3	3.8
A3	2.4	8.2	5.7	9.3	4.2

−

	P1	P2	P3	P4	P5
A1	0	0	1.7	-1.7	1
A2	2.9	-3.8	0	0	0
A3	-2.1	0	0	2.6	0

=

	P1	P2	P3	P4	P5	Average
A1	3.8	7.5	5	6	3.5	5.16
A2	4.1	7.8	5.3	6.3	3.8	5.46
A3	4.5	8.2	5.7	6.7	4.2	5.86

Bioconductor.org

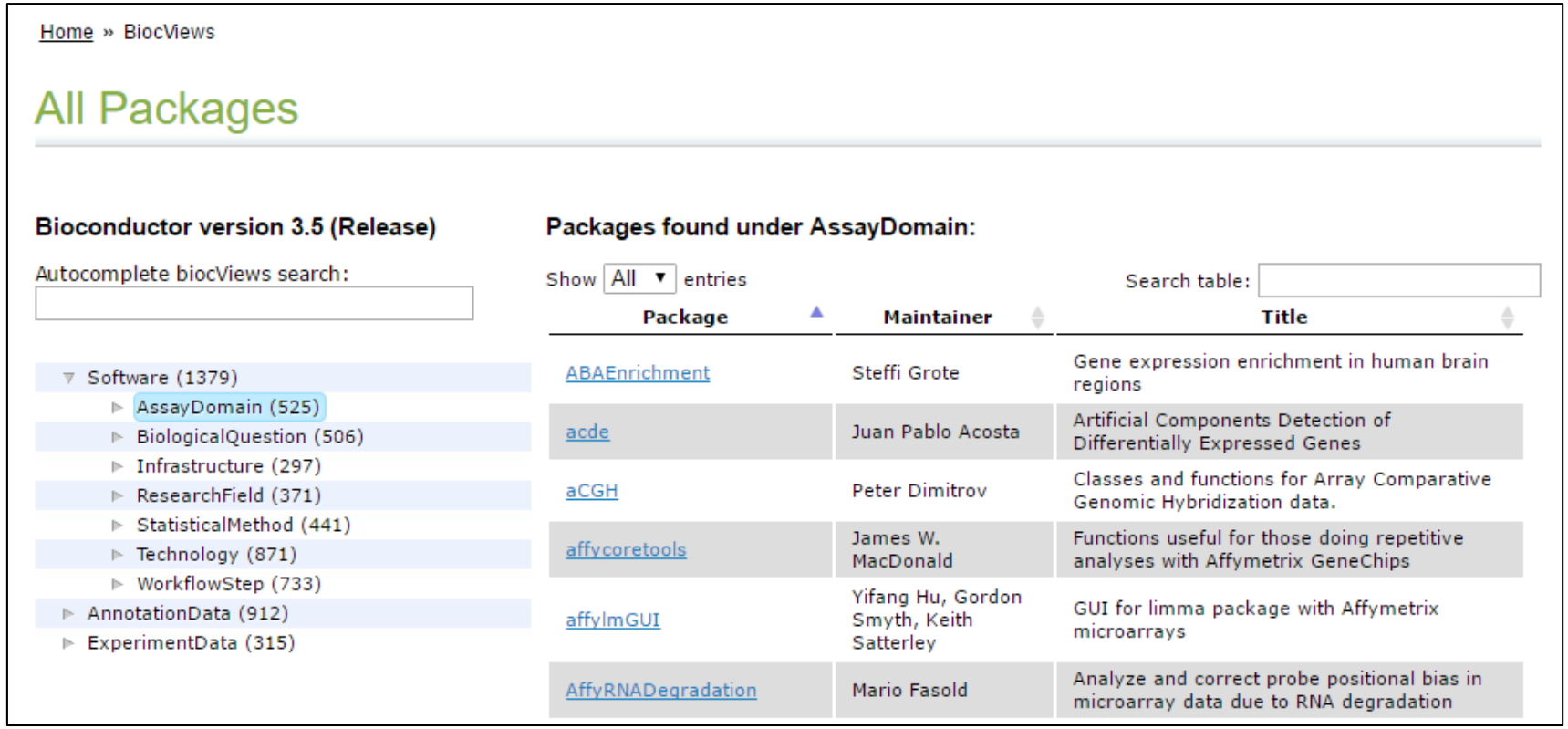
Bioconductor characteristics

- Collection of 2500+ packages specific for bioinformatics

- Three categories of packages:
  - Software
  - Annotation
  - Experimental data

- Each package goes through a rigorous assessment for technical requirements and ensuring minimal overlapping among packages

Bioconductor packages



Home » BiocViews

All Packages

Bioconductor version 3.5 (Release)

Autocomplete biocViews search:

- Software (1379)
  - AssayDomain (525)
  - BiologicalQuestion (506)
  - Infrastructure (297)
  - ResearchField (371)
  - StatisticalMethod (441)
  - Technology (871)
  - WorkflowStep (733)
- AnnotationData (912)
- ExperimentData (315)

Packages found under AssayDomain:

Show All entries Search table:

Package	Maintainer	Title
ABAEnrichment	Steffi Grote	Gene expression enrichment in human brain regions
acde	Juan Pablo Acosta	Artificial Components Detection of Differentially Expressed Genes
aCGH	Peter Dimitrov	Classes and functions for Array Comparative Genomic Hybridization data.
affycoretools	James W. MacDonald	Functions useful for those doing repetitive analyses with Affymetrix GeneChips
affylmGUI	Yifang Hu, Gordon Smyth, Keith Satterley	GUI for limma package with Affymetrix microarrays
AffyRNADegradation	Mario Fasold	Analyze and correct probe positional bias in microarray data due to RNA degradation

Bioconductor other resources

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Home » Help » Courses and Conferences

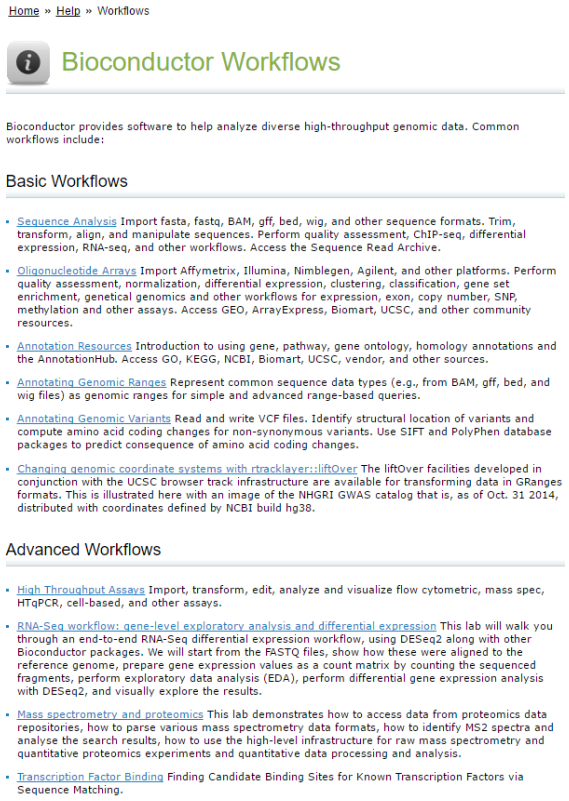
Courses & Conferences

Bioconductor provides training in computational and statistical methods for the analysis of genomic data. You are welcome to use material from previous courses. However, you may not include these in separately published works (articles, books, websites). When using all or parts of the Bioconductor course materials (slides, vignettes, scripts) please cite the authors and refer your audience to the Bioconductor website.

Upcoming events are advertised 6 to 8 weeks in advance.

Show 25 entries Search:

Keyword	Title	Course	Materials	Date	Bioc/R Version
Talk	Good software: simple, tidy, rich, Martin Morgan	Meetup	pdf	2017-04-20	3.5/3.4
Introduction	Introduction to *R* and *Bioconductor*, Martin Morgan, Lori Shepherd	Moffitt	Install, Intro to R, Data input and manipulation, Statistics, Graphics, Intro to Bioconductor, Key classes and methods, An RNA-seq work flow, Next steps	2017-03-02	3.4/3.3
Packages	Software development in *R* and *Bioconductor*, Martin Morgan	UIdaho	Packages & version control	2017-01-31	3.4/3.3
Overview	R / *Bioconductor* for 'Omics Analysis (U. Idaho), Martin Morgan		pdf (slides)	2017-01-30	3.4/3.3
Introduction	Introduction to *R*, Martin Morgan	RPCI RIntro	Installation, Using R, Input and manipulation, Statistics, Workflows and visualization	2017-01-09	3.4/3.3
Overview	*Bioconductor* for 'Omics Analysis (University of Rochester Medical Center), Martin Morgan		pdf (slides)	2016-12-01	3.4/3.3
Annotation	Annotation, Communication, and Performance, Martin Morgan	Technion-BKU	html, R, Rmd	2016-11-20	3.4/3.3
RNAseq	An RNA-seq work flow, Martin Morgan	Technion-BKU	html, R, Rmd	2016-11-20	3.4/3.3
Introduction	Introduction to *Bioconductor*, Martin Morgan	Technion-BKU	html, R, Rmd	2016-11-20	3.4/3.3
Status	The *Bioconductor* Project: Current Status, Martin Morgan	BioC Asia	pdf (slides)	2016-11-04	3.4/3.3
Introduction	*Bioconductor* for Genomic Analysis, Martin Morgan	BioC Asia	html, R, Rmd, github	2016-11-03	3.4/3.3



Home » Help » Workflows

Bioconductor Workflows

Bioconductor provides software to help analyze diverse high-throughput genomic data. Common workflows include:

Basic Workflows

- Sequence Analysis Import fasta, fastq, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.
- Oligonucleotide Arrays Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.
- Annotation Resources Introduction to using gene, pathway, gene ontology, homology annotations and the AnnotationHub. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.
- Annotating Genomic Ranges Represent common sequence data types (e.g., from BAM, gff, bed, and wig files) as genomic ranges for simple and advanced range-based queries.
- Annotating Genomic Variants Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding changes.
- Changing genomic coordinate systems with rtracklayer::liftOver The liftOver facilities developed in conjunction with the UCSC browser track infrastructure are available for transforming data in GRanges formats. This is illustrated here with an image of the NHGRI GWAS catalog that is, as of Oct. 31 2014, distributed with coordinates defined by NCBI build hg38.

Advanced Workflows

- High Throughput Assays Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.
- RNA-Seq workflow: gene-level exploratory analysis and differential expression This lab will walk you through an end-to-end RNA-Seq differential expression workflow, using DESeq2 along with other Bioconductor packages. We will start from the FASTQ files, show how these were aligned to the reference genome, prepare gene expression values as a count matrix by counting the sequenced fragments, perform exploratory data analysis (EDA), perform differential gene expression analysis with DESeq2, and visually explore the results.
- Mass spectrometry and proteomics This lab demonstrates how to access data from proteomics data repositories, how to parse various mass spectrometry data formats, how to identify MS2 spectra and analyse the search results, how to use the high-level infrastructure for raw mass spectrometry and quantitative proteomics experiments and quantitative data processing and analysis.
- Transcription Factor Binding Finding Candidate Binding Sites for Known Transcription Factors via Sequence Matching.