R for immunologists

Anna Lorenc

DIIID - Crick

anna.lorenc@kcl.ac.uk

2017

Why bother with programming?

- **Big** data
- Re-running analysis: change of parameters, added data etc.
- Documenting all steps of analysis
- Sharing your analysis
- Repetitive analysis, plotting etc. straightforward thanks to loops

Why not a spreadsheet?

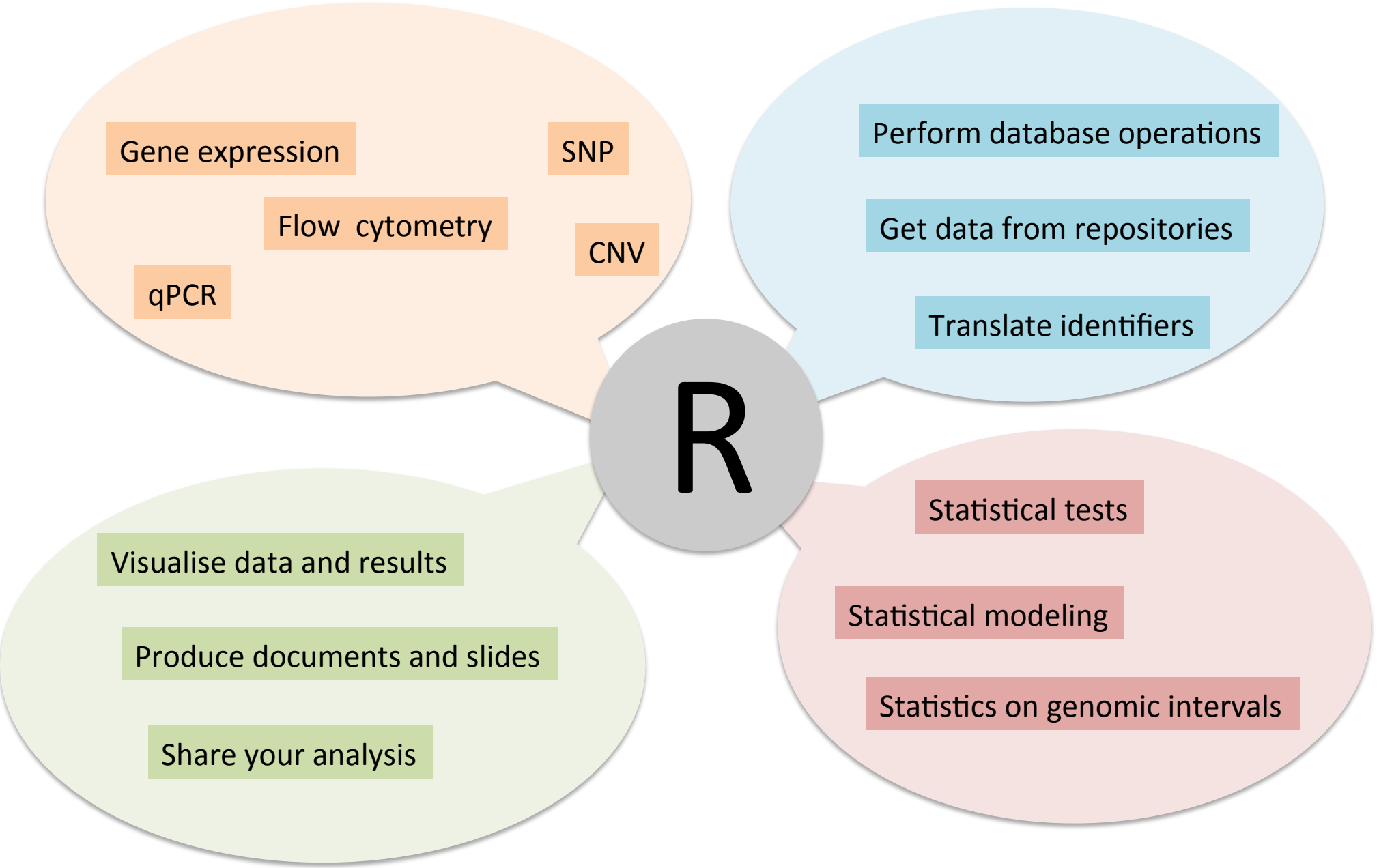
- Not enough rows/columns
- Human/spreadsheet error: it is easy to change data

Zeeberg, B. et al. Mistaken identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics 5, 80 (2004) **Sep10**

- Human error: it is not easy to follow how the cell values were derived
- Re-running analysis: change of parameters, added data etc. very tedious
- Does not allow for more sophisticated analyses.
- Difficult to share

Why R?

- High level language
- Programming language and **statistical** environment
- Easy to read, analyze and write “big tables” (hundreds of columns and millions of rows) – efficient functions and data structures
- On top of basic installation - thousands of packages – stats, graphics, domain-specific
- Big and growing community
- *Bioconductor* – biology-high throughput branch – most widely used language in bioinformatics.


Gene expression
SNP
Flow cytometry
CNV
qPCR
Perform database operations
Get data from repositories
Translate identifiers
R
Visualise data and results
Produce documents and slides
Share your analysis
Statistical tests
Statistical modeling
Statistics on genomic intervals

When you analyse your data…

- Clear idea about what do you want to do (Visualize measurements? Statistical testing? Derive a measure from many parameters?)
- Do only things which are relevant, not (all) things which are possible
- If deploying your own analysis on a big datasets, try it out on a small subsets of simulated (fully understood) and real-life data

When you analyse your data…

- Raw data is sacrosanct. Do not touch it.
- If it requires any transformations/name changes - do it in a derived file. Do not touch the raw data.
- Keep tab on what version of software, databases etc. you use.
- Comment your code AND analysis.

Confusing terminology

- R – programming language
- Interactive environment ("read data into R")

R code might be run in the interactive environment or from the commandline:

```
R CMD myscript.R
```

Both offer some advantages. We'll use IDE.

atient X weights 43 kg and is 1m 45cm tall.

42

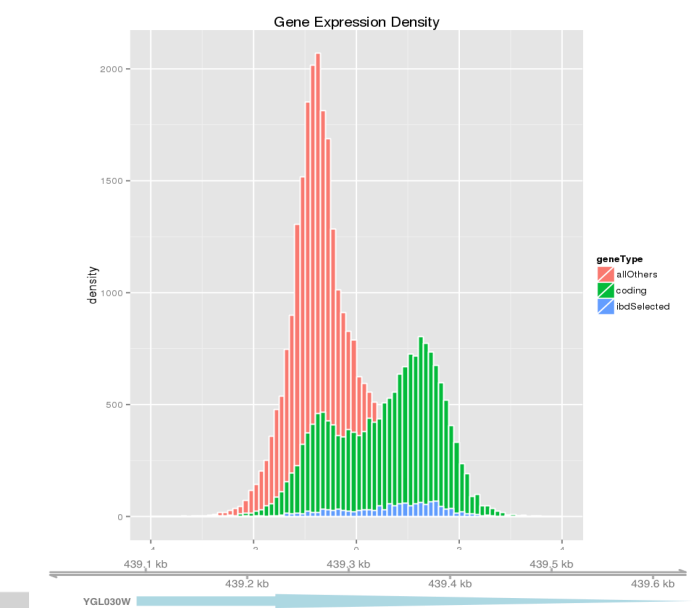
	Mouse	Gender	Assay_Date	Total_T_ce	Ly6C+_CD5+	Total_ab_T	Total_CD4+	CD4+_T_hel	Effector_C	Resting_CD	KLRG1+_CD
1:	M02406315	Female	2016-08-01	31.9520	29.226	95.733	57.574	89.815	29.853	69.474	2.43
2:	M02406314	Female	2016-08-01	27.9560	28.073	94.850	58.243	90.221	33.597	65.886	3.45
3:	M02406310	Male	2016-08-01	23.8100	35.395	95.153	57.030	89.100	36.818	62.258	2.81
4:	M02406309	Male	2016-08-01	27.5430	37.309	96.011	57.613	89.404	31.150	68.238	2.42
5:	M02376651	Female	2016-06-29	30.6930	30.210	96.291	58.341	90.456	27.096	72.701	1.34
6:	M02376650	Female	2016-06-29	32.8250	30.003	96.260	58.993	92.070	20.428	79.552	1.30
7:	M02416303	Male	2016-08-22	30.5890	35.361	95.987	57.893	91.356	20.971	78.872	2.12
8:	M01553164	Male	2013-11-13	27.5492	NA	94.535	56.712	89.604	25.563	71.400	0.97
9:	M01976527	Female	2015-02-25	30.1690	37.362	94.555	57.463	88.412	29.868	69.857	0.89
10:	M01976518	Female	2015-02-25	28.6680	41.351	92.734	61.558	91.124	23.544	76.449	2.65
11:	M01837260	Female	2014-09-22	29.5530	39.064	94.093	56.925	88.429	23.552	76.356	0.73
12:	M02415261	Male	2016-08-17	28.3930	36.821	96.390	59.999	89.651	32.268	67.070	1.76
13:	M02425263	Female	2016-09-01	34.1290	36.077	96.650	60.045	92.714	33.390	65.006	1.09
14:	M02387647	Male	2016-07-13	33.4260	37.617	97.011	64.220	92.602	43.982	53.256	1.53
15:	M01542979	Female	2013-10-30	26.6000	NA	94.314	51.716	92.185	23.725	74.751	1.07
16:	M01913572	Female	2014-12-15	29.4000	39.322	95.180	58.313	88.922	32.104	67.528	1.12
17:	M01913571	Female	2014-12-15	26.1120	39.142	93.777	55.684	89.755	23.620	76.296	1.11
18:	M01832867	Female	2014-09-10	31.4590	11.267	94.133	58.003	89.858	20.406	79.570	1.06
19:	M01800538	Male	2014-08-06	38.3630	41.041	94.766	48.640	90.553	26.102	73.250	0.73
20:	M01800536	Male	2014-08-06	37.1640	42.554	94.095	40.551	90.460	25.722	73.515	1.67
21:	M01792855	Male	2014-07-30	24.6790	45.733	94.234	51.548	87.990	34.294	65.539	0.92
22:	M01792854	Male	2014-07-30	25.1940	39.466	93.906	58.188	88.311	46.149	53.366	0.83
23:	M01769859	Male	2014-07-08	25.6710	49.390	90.521	48.257	89.691	24.395	75.661	1.39
24:	M01729423	Male	2014-05-27	30.0640	NA	93.989	51.315	91.960	34.888	62.304	1.26
25:	M01729422	Male	2014-05-27	25.6490	NA	92.968	50.845	91.574	40.651	55.961	1.88
26:	M01723957	Female	2014-05-19	29.5480	NA	91.916	53.831	90.351	48.148	49.027	1.93
27:	M01723955	Female	2014-05-19	30.0840	NA	90.546	51.412	92.614	27.419	71.678	2.50
28:	M01676584	Male	2014-03-19	28.1000	NA	94.558	54.579	90.155	25.480	72.718	1.06
29:	M01676583	Male	2014-03-19	28.8200	NA	92.636	53.334	90.895	25.192	73.044	1.06
30:	M01676575	Female	2014-03-24	35.5640	NA	94.869	54.798	92.320	19.861	79.354	0.95
31:	M01676573	Female	2014-03-24	31.9730	NA	94.749	54.061	93.267	25.028	74.051	1.96

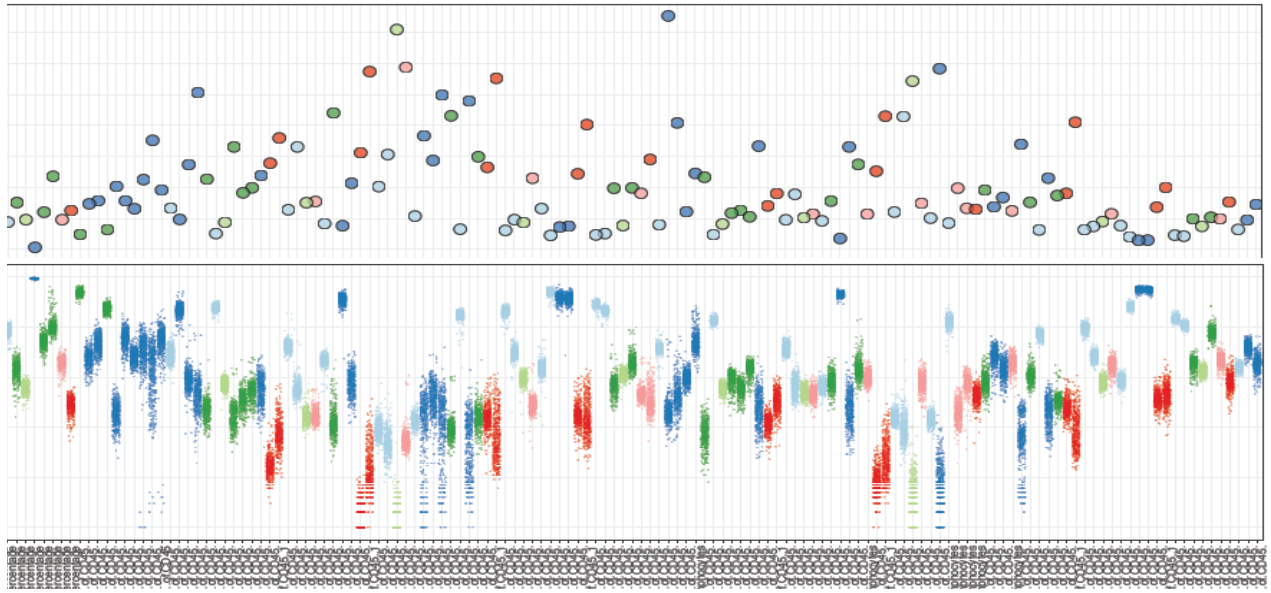
Data

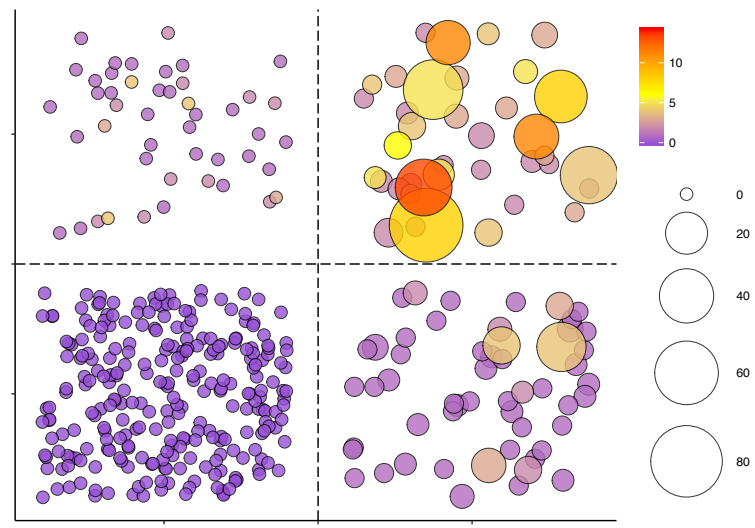
R

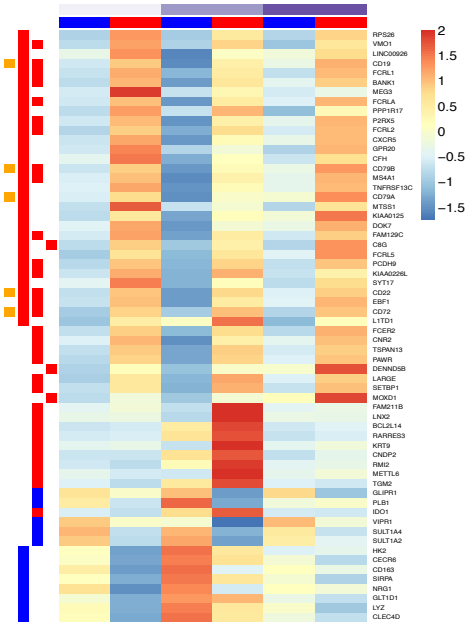
functions
/verbs/

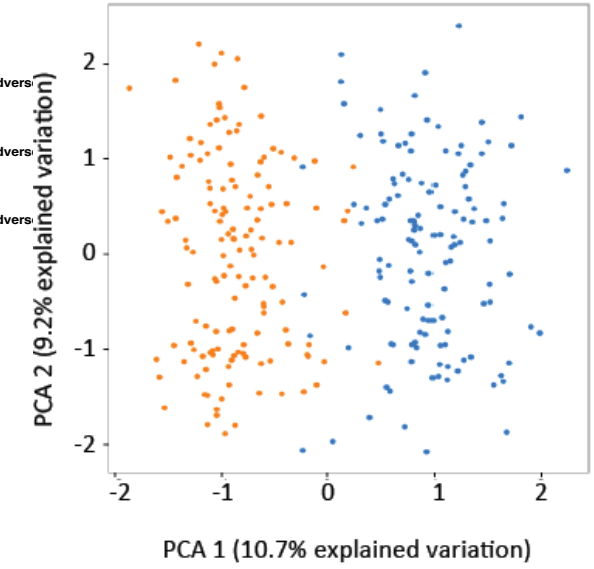
data containers
/nouns/

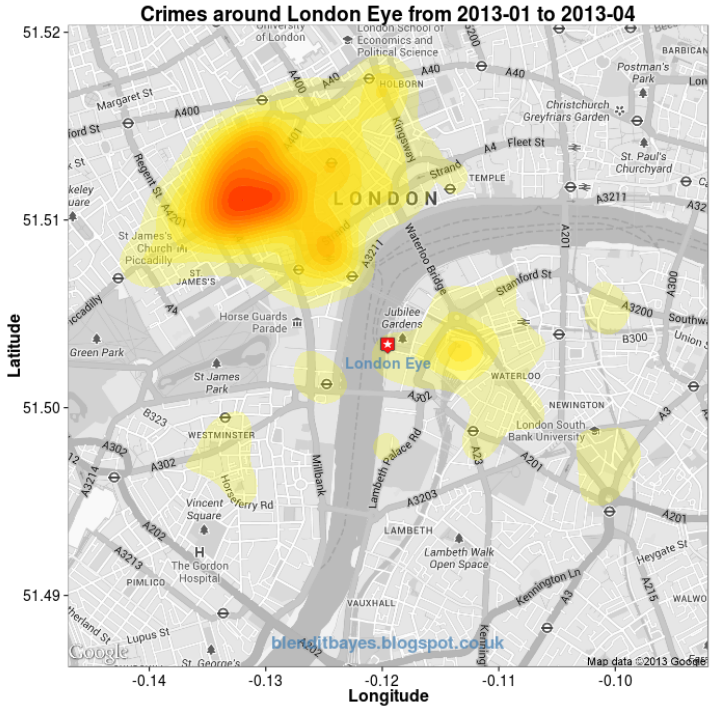

Gene Expression Density
density
2000
1500
1000
500
0
geneType
allOthers
coding
ibdSelected
439.1 kb
439.2 kb
439.3 kb
439.4 kb
439.5 kb
439.6 kb
YGL030W
YGL030W

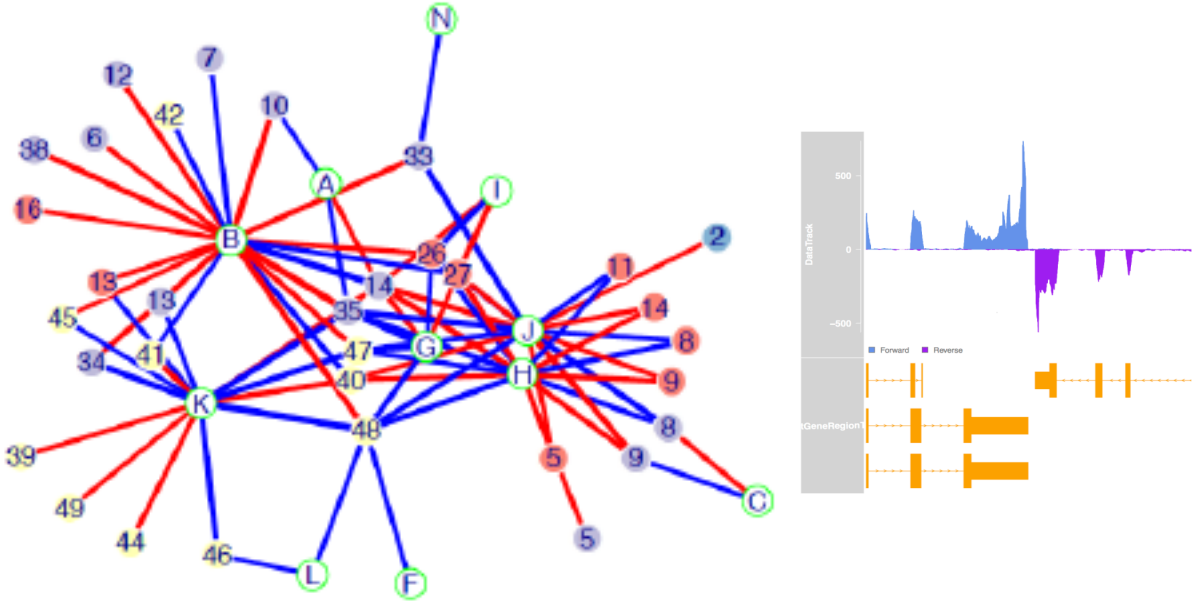


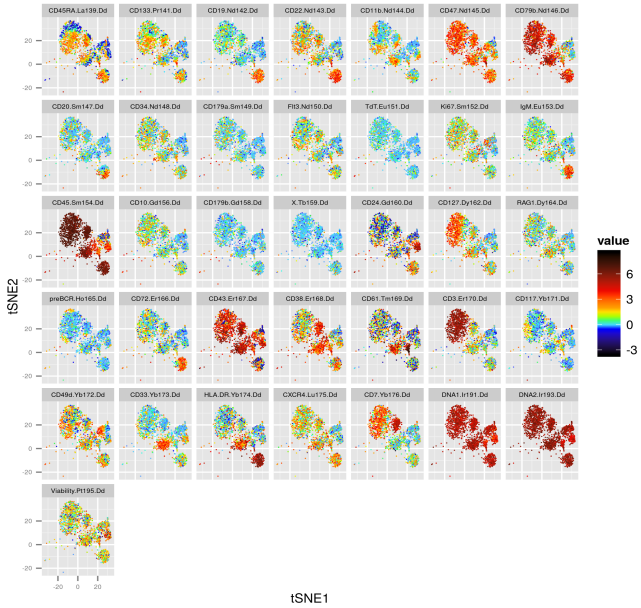

10
5
0
0
20
40
60
80

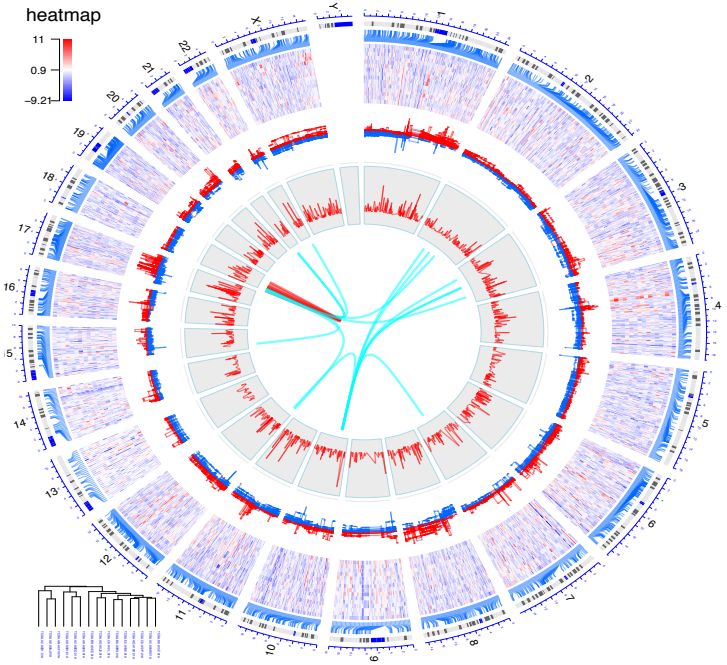

2
1.5
1
0.5
0
−0.5
−1
−1.5
X1
T1
T2
T3
adverse
no
yes
bcell
0
1
adverseVSnonadvers
−1
0
1
adverseVSnonadvers
−1
0
1
adverseVSnonadvers
0
1


PCA 2 (9.2% explained variation)
PCA 1 (10.7% explained variation)
2
1
0
-1
-2
-2
-1
0
1
2


Crimes around London Eye from 2013-01 to 2013-04
Latitude
Longitude
51.52
51.51
51.50
51.49
-0.14
-0.13
-0.12
-0.11
-0.10
LONDON
London Eye
WESTMINSTER
LAMBETH
VAUXHALL
WATERLOO
NEWINGTON
HOLBORN
TEMPLE
Horse Guards Parade
Jubilee Gardens
St James's Park
Green Park
Vincent Square
The Gordon Hospital
Lambeth Walk Open Space
London South Bank University
Christchurch Greyfriars Garden
St. Paul's Churchyard
Postman's Park
St James's Church in Piccadilly
Waterloo Bridge
Lambeth Palace Rd
Millbank
Horseferry Rd
Kennington Ln
Stamford St
Strand
Fleet St
Margaret St
Regent St
Lupus St
PIMLICO
BARBICAN
Economics and Political Science
blenditbayes.blogspot.co.uk
Map data ©2013 Google


N
A
B
C
F
G
H
I
J
K
L
DataTrack
GeneRegion
Forward
Reverse


tSNE1
tSNE2
value
6
3
0
-3


heatmap
11
0.9
-9.21

Bioconductor

http://www.bioconductor.org/

tools for the analysis of high-throughput genomic data in R
open source and open development

> 1100 packages

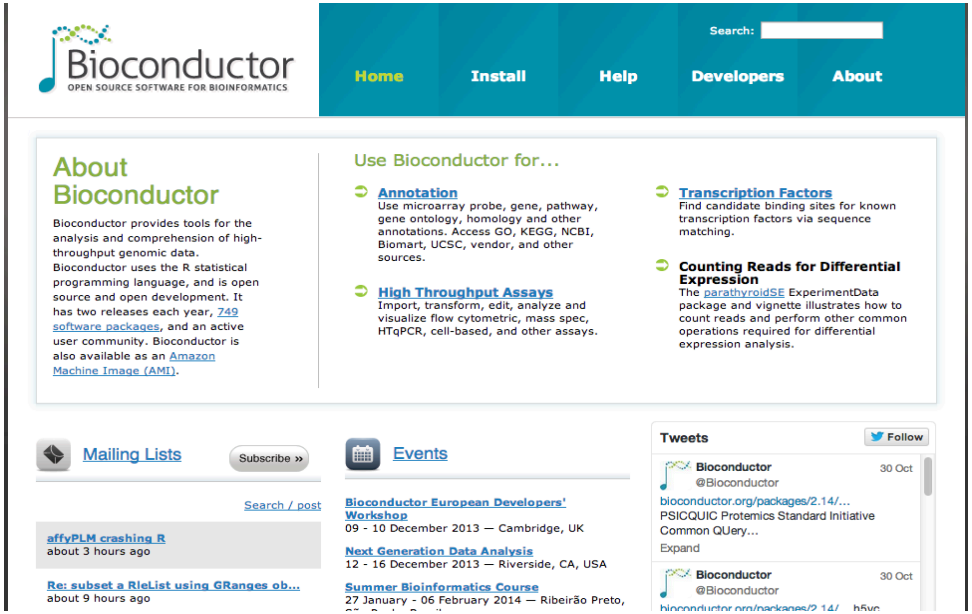
Course materials

workflows

mailing list with archive

Community

www.bioconductor.org



In R:

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
```

NATURE METHODS | PERSPECTIVE

Orchestrating high-throughput genomic analysis with Bioconductor

Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron + *et al.*

Affiliations | Corresponding author

Nature Methods **12**, 115–121 (2015) | doi:10.1038/nmeth.3252

Received 30 July 2014 | Accepted 09 December 2014 | Published online 29 January 2015

Abstract

BiologicalQuestion (260)
- AlternativeSplicing (7)
- Coverage (13)
- DifferentialExpression (164)
- DifferentialMethylation (7)
- DifferentialPeakCalling (1)
- DifferentialSplicing (6)
- FunctionalPrediction (1)
- GeneRegulation (21)
- GeneSetEnrichment (41)
- GeneTarget (1)
- GenomeAnnotation (10)
- GenomicVariation (6)
- LinkageDisequilibrium (1)
- MotifAnnotation (3)
- MotifDiscovery (4)
- NetworkEnrichment (13)
- NetworkInference (17)
- SequenceMatching (17)
- SomaticMutation (4)
- VariantDetection (1)

Sequencing based (213)
- ChIPSeq (40)
- DNASeq (6)
- ExomeSeq (3)
- MethylSeq (10)
- Microbiome (3)
- miRNA (4)
- RIPSeq (1)
- RNASeq (77)
- TargetedResequencing (2)
- WholeGenome (3)

Infrastructure (185)
- DataImport (81)
- DataRepresentation (34)
- GUI (19)
- ThirdPartyClient (9)
- ResearchField (193)
- StatisticalMethod (262)
- Technology (591)
- FlowCytometry (36)
- MassSpectrometry (44)
- Microarray (328)
- MicrotitrePlateAssay (13)
- qPCR (9)
- SAGE (9)

AssayDomain (299)
- aCGH (12)
- CellBasedAssays (38)
- ChIPchip (7)
- CopyNumberVariation (43)
- CpGIsland (6)
- DNAMethylation (38)
- ExonArray (6)
- GeneExpression (131)
- GeneticVariability (23)
- SNP (40)
- Transcription (47)

WorkflowStep (477)
- Alignment (8)
- Annotation (67)
- BatchEffect (5)
- ExperimentalDesign (2)
- MultipleComparison (76)
- Normalization (15)
- Pathways (69)
- Preprocessing (128)
- QualityControl (95)
- ReportWriting (25)
- Visualization (218)

BioC workflows

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,

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Packages »

Bioconductor's stable, semi-annual release:

- Analysis software packages.
- Annotation packages.
- Illustrative experiment data packages.
- Latest release announcement.

Bioconductor is also available via Docker and Amazon Machine Images.

Prepare a folder Rcourse

and inside folders:

- Rcourse/data
- Rcourse/scripts
- Rcourse/analysis