Universität Rostock Traditio et Innovatio

Universitätsmedizin Rostock

Integrative Bioinformatics and Systems Biology

Dr. Mohamed Hamed

LECTURE PLANNING

Lecture: 12 lectures, 3 hrs each.

Total workload: 42 hrs : 36 hrs of lectures and tutorials and 6 hrs of self studies.

Entrance requirements: basic knowledge of biology and computer science.

Literature: Lecture slides, tutorial handouts and problem sets will be provided.

INTEGRATIVE BIOINFORMATICS AND SYSTEMS BIOLOGY

COURSE ABBREVIATION: INT-BIO

LANGUAGE: ENGLISH
USED MEDIA: POWERPOINT PRESENTATION

Module: Lecture and tutorial.

By: Dr. Mohamed Hamed

Head of the Integrative OMICs Analysis Group in Rostock University medical center, Rostock University, Germany.

MOTIVATION AND COURSE OBJECTIVES

The main challenge of modern systems biology is unraveling the holistic picture of the complex molecular interactions that occur on different molecular levels (genomic, transcriptomic, epigenomic, proteomic, etc). Therefore, the needs to integrate/jointly analyze biological data from different high-throughput technologies emerged in order to identify biomarkers for early diagnosis and prognosis of complex diseases and facilitating the development of novel treatment approaches.

This course aims at teaching students how to perform data-specific computational analyses as well as integrative analysis approaches, combining knowledge from different OMICs-based datasets. Both, theoretical and practical aspects will be covered. Students will have the opportunity to work both independently during tutorials and in teams during the research project.

COMPETENCES TO BE DEVELOPED

Students will get practical and extensive hands-on experience on:

-R scripting language and bioconductor packages.

-Data-specific computaional analysis and pipelines for the vast amounts of biological data produced using high-throughput technologies.

-Developing and applying integrative bioinformatics methods that could be utilized in all biology-related areas of interest.

-Basics of machine learning methods as tools for integrating biological features from heterogeneous Omics data.

-Students will be developing their own research projects, interpreting the obtained results, writing a manuscript, scientifically discussing the results.

ASSESSMENT

-Students need to finalize a research project applying all/ most of the learned methods and skills during the course. Novelty and extending the learned methods is highly encouraged and will be well graded.

-The outcomes of each research project should be compiled in a high scientific quality research article that is ready for submission in a peer-review journal.

-All projects will be presented, discussed and scientifically reviewed in the last lecture.

R language mini-course 1	R language mini-course 2	R language mini-course 3	Introduction to integrative bioinformatics
-Introduction to the course -Basics of R language and statistical methods. -R studio IDE	-Advanced R statistics -Bio-conductor packages	-Case study: Microarray analysis using R	-Importance of data integration -Different methods for biological data integration -OMICs data types, and TCGA repository -Databases of diseases-related genes and miRNAs
Transcriptomic analysis	**Non-coding RNAs**	**Network- based integrative methods**	**Epigenetics**
-From microarrays to RNA-seq -RNA-seq analysis -Linking to Ontologies and pathways -Drug signature databases: CMaP and LINKS	-Small and long non-coding RNAs -miRNA sequencing analysis -miRNA databases -lncRNAs analysis	-TFmiR analysis -Network motif analysis -Central hubs identifications - Network visualization (Cytoscape)	-Introduction to the epigenetic landscapes of normal and tumor cells -DNA methylation, Co-methylation analysis -DMRs identifications
Chip-seq experiments and/or GWAS	**Integrative analysis based on machine learning.**	**Supervised machine learning methods**	**PROJECTS DISCUSSION AND CLOSURE**
-Computational analysis of Chip-Seq data and/or -Downstream analysis of genetic variants	-Introduction to machine learning in bioinformatics. -Unsupervised methods: Clustering biological data -PCA analysis	-Classification and regression analysis -Model selection and evaluation of learning methods -Outlook at deep neural networks applications in bioinformatics	-Projects presentation. -Reviews of the potential manuscripts

Lecture 6
Practical RNA-Seq using R

Genome Sequencing

- Sequencing refers to getting access to the sequence or primary structure of any biopolymer
- Examples include nucleic acid sequencing (DNA, mRNA, miRNA,..etc)
- high-throughput and highly parallel sequencing approaches have been denoted as Next-Generation Sequencing (NGS)

NGS overview

- Next Generation Sequencing (NGS) = high throughput sequencing, massive parallel s., second generations
- Population studies (1000 Genomes project)
- Medical applications (TCGA, ICGC)
  - ✓ sequence patient DNA to find variants / mutations that cause diseases
  - ✓ sequence pathogenic organisms
- Fast and comparably cheap => many labs can afford this for their samples
  - ✓ sequencing companies (Illumina, LifeTechnologies, ...)
  - ✓ cooperation with sequencing center (BROAD, Sanger, BGI, ...)
  - ✓ buy their own sequencer
- Billions of short reads with comparably low quality => traditional sequence analysis methods do not work (or do not even exist)
- Bioinformaticians / computational biologists needed to analyze the data
- special research questions => combine / develop tools

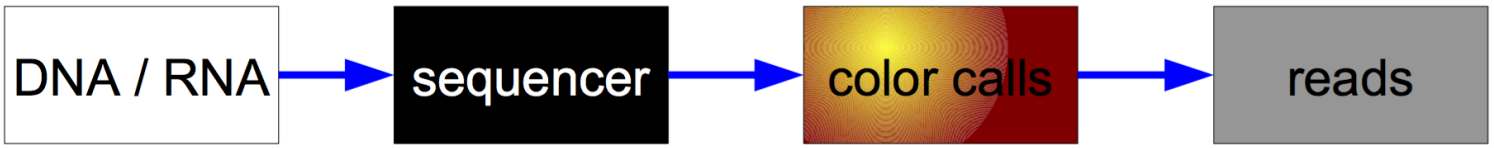
Sequencing technologies

- Sanger
- 454
- Illumina (Solexa)
  - Genome Analyzer II
  - HighSeq 2000
- SOLiD (ABI, LifeTechnologies)
  - SOLiD 4
  - 5500(xl)
- Complete Genomics
- Ion Torrent
- Pacific Biosciences SMRT
- Oxford Nanopore

first generation

second generation

third generation



More details….

ELSEVIER

Genomics

Volume 107, Issue 1, January 2016, Pages 1–8

Review

The sequence of sequencers: The history of sequencing DNA

James M. Heather, Benjamin Chain

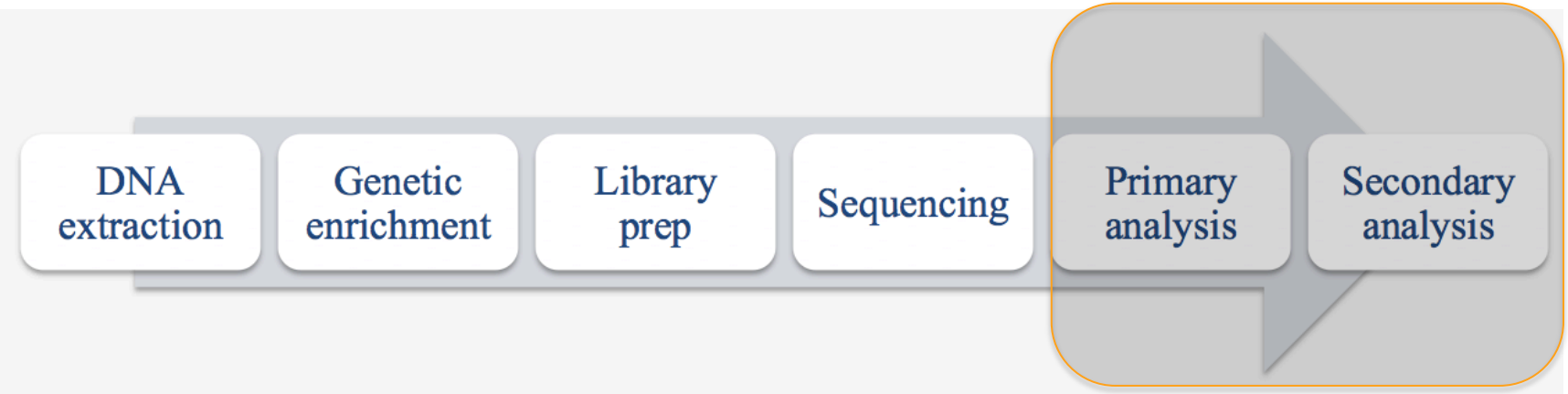
Show more

https://doi.org/10.1016/j.ygeno.2015.11.003

Highlights

- We review the drastic changes to DNA sequencing technology over the last 50 years.
- First-generation methods enabled sequencing of clonal DNA populations.
- The second-generation massively increased throughput by parallelizing many reactions.
- Third-generation methods allow direct sequencing of single DNA molecules.

Standard NGS process



1. DNA extraction: from biological samples such as blood to DNA material.
2. Genetic enrichment: focusing on certain regions from the full genome.
3. Library prep: the DNA has to be prepared for sequencing.
4. Sequencing: the real readout of the biological material.
5. Primary analysis: Alignment / Assembly and SNP calling
6. Secondary analysis: phenotyping properties such as virulence, integrative analys

NGS applications

DNA

Whole human genome sequencing
Exome sequencing
Gene panel sequencing
Pathogene sequencing

SNP calling
INDEL calling
Copy Number Variations
Genomic Rearrangements

RNA

Transcriptome sequencing
miRNOme sequencing

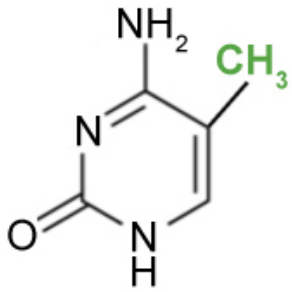
Expression level
Alternative splicing events
Fusion genes

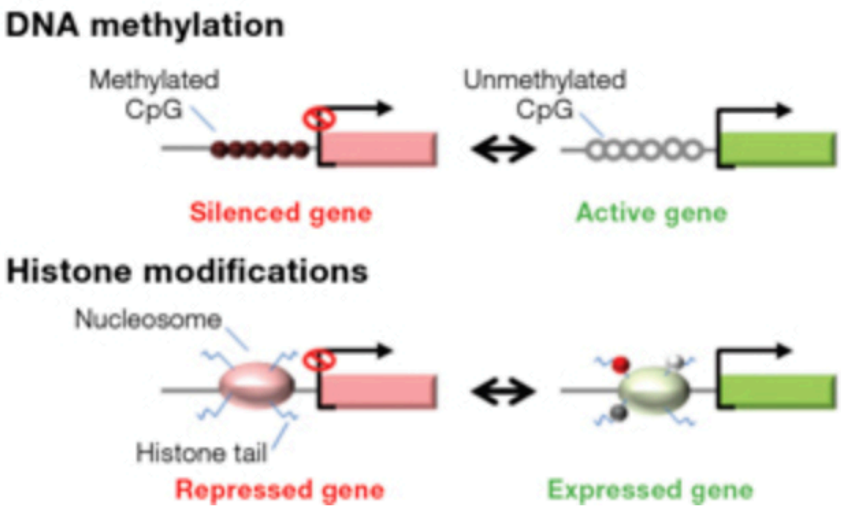
Others

interactions of proteins with DNA (ChIP seq)
interactions of proteins with RNA (PAR-CLIP)
Bisulfite sequencing (methylation)

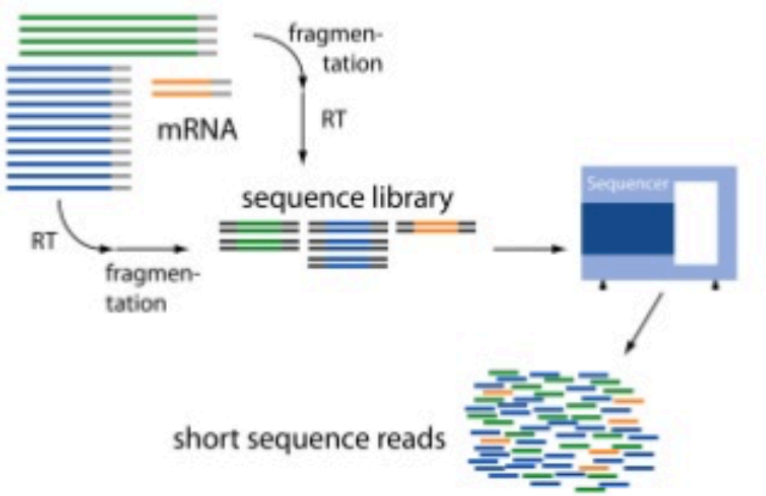
NGS applications: Epigenomics

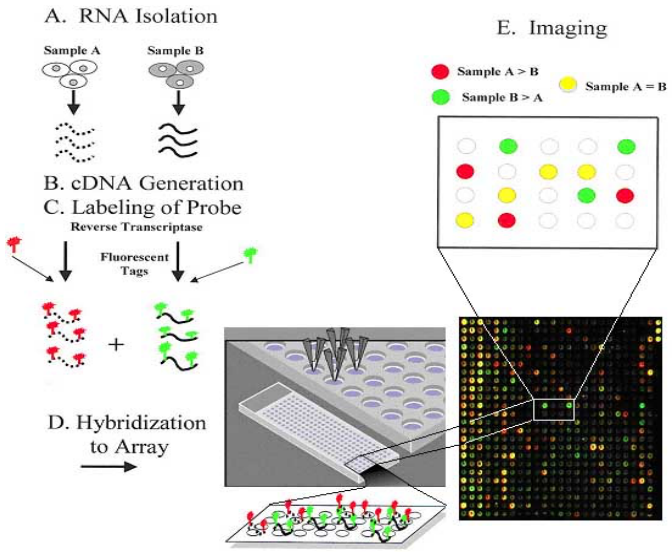
- Histone modifications
  - ChIP-Seq
- Methylome
  - 5-methylcytosine
  - whole genome or reduced representation bisulphite sequencing (WG-BS, RRBS)
  - meDIP-Seq
- Global methylation changes
- Differentially methylated regions
- Integration with gene expression



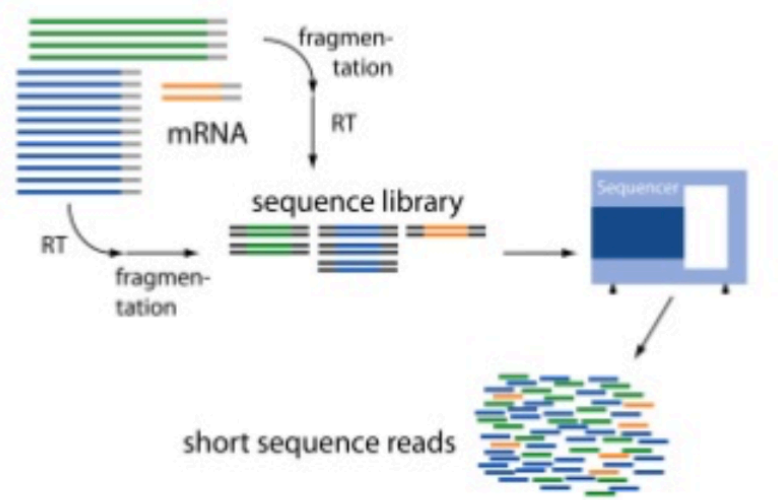


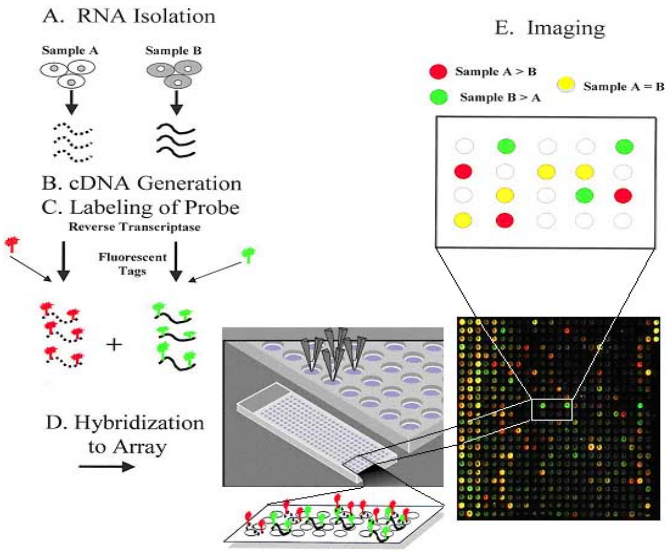
Ex1: RNA-Seq and Microarrays





Ex1: RNA-Seq and Microarrays





Today -alternative to expression microarrays
- "digital gene expression"

Single base resolution
identification of:
- differential expression (clustering)
- novel transcripts, exons, isoforms, TSS, TES
- alternative splicing events

Gene level resolution
identification of:
- differential expression (clustering)
- Cheap and fast

But many artifacts

File formats

- Image data
  Usually discarded after base calling

- FASTA/FASTQ
  -Identifier
  -Sequence
  -Quality scores
  (FASTQ only)

*.fastq

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- SAM/BAM
  -File format for aligned reads
  -However due to good compression and annotation, also often used for storing unaligned reads
  -More in the alignment

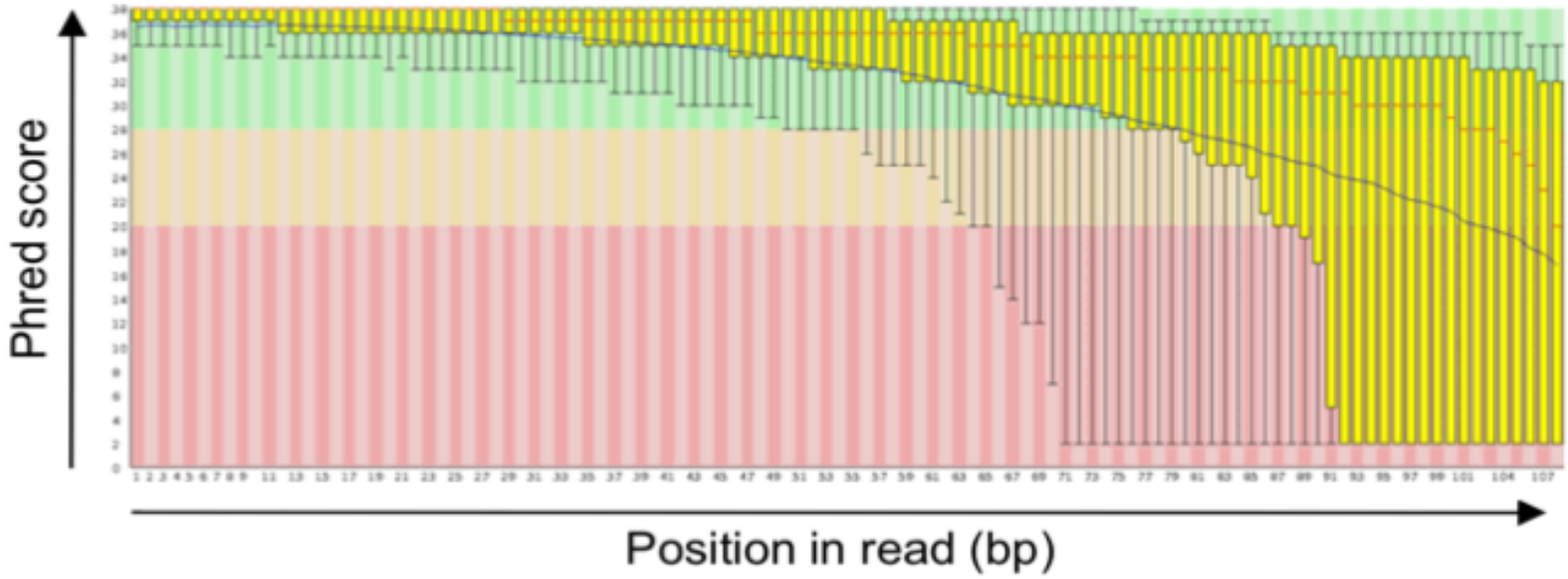
Quality (Phred) score

- **Phred Score (*Q*):**

$$Q = -10 \log_{10} P$$

  - Here *P* denotes the estimated base calling error probability

- Base quality scores tend to decline towards the end of the read
  - Reads are often trimmed before or in the alignment step



Data Analysis Pipeline

Quality Control

Mapping

Quantifying

Differential expression

Presenting the results

Quality control- FastQC

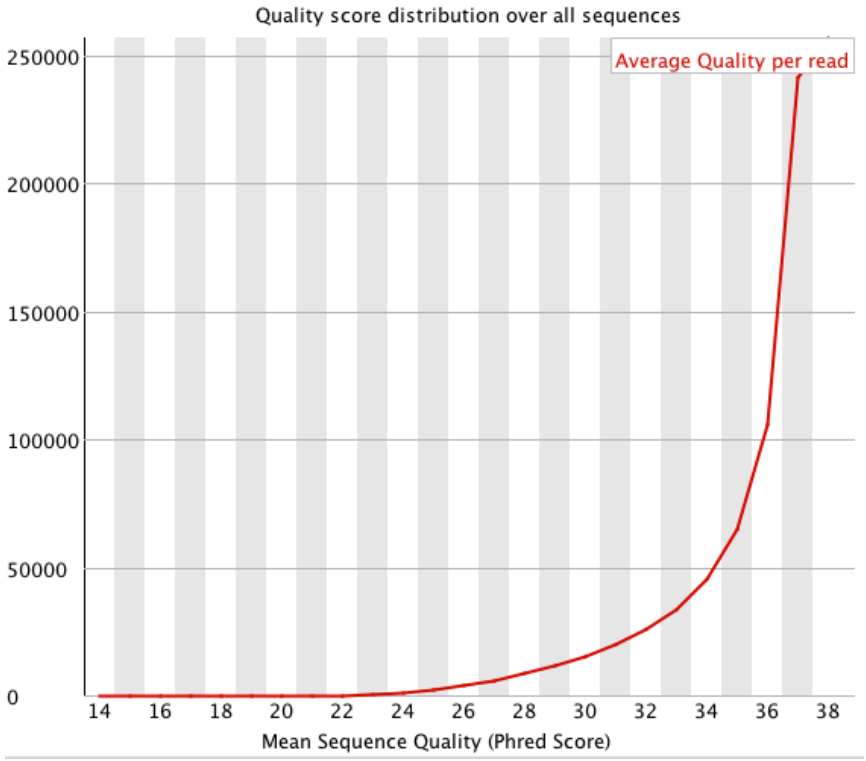
Quality control of short reads

1- Import data from BAM, SAM or FastQ files (any variant)

2-Providing a quick overview to tell you in which areas there may be problems

3-Summary graphs and tables to quickly assess your data

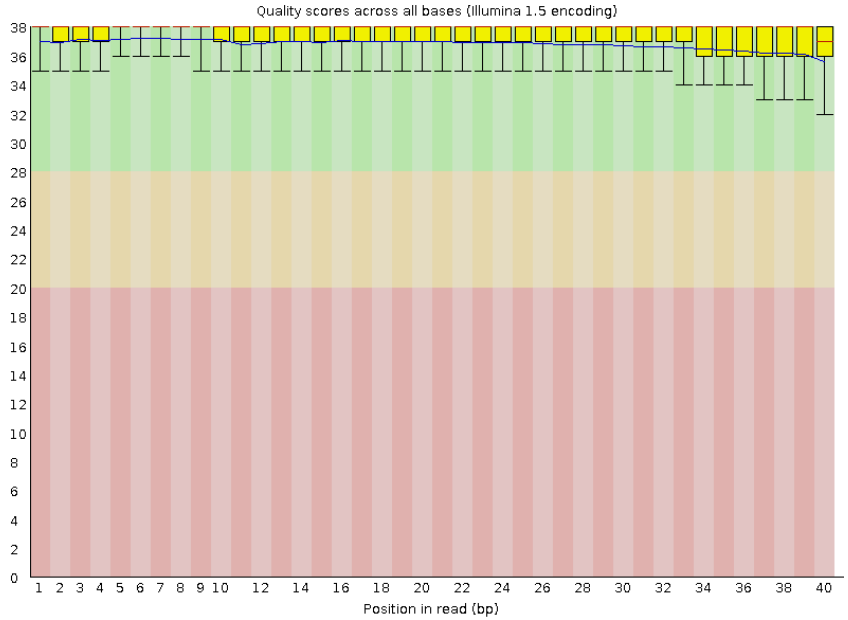
4-Export of results to an HTML based permanent report



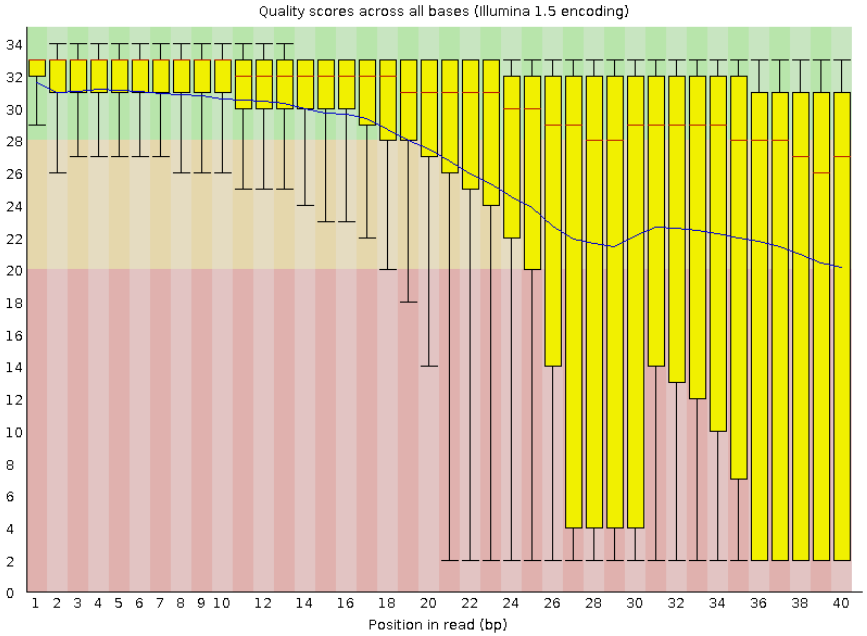
Per sequence quality scores

Quality control- FASTQC

Per base sequence quality



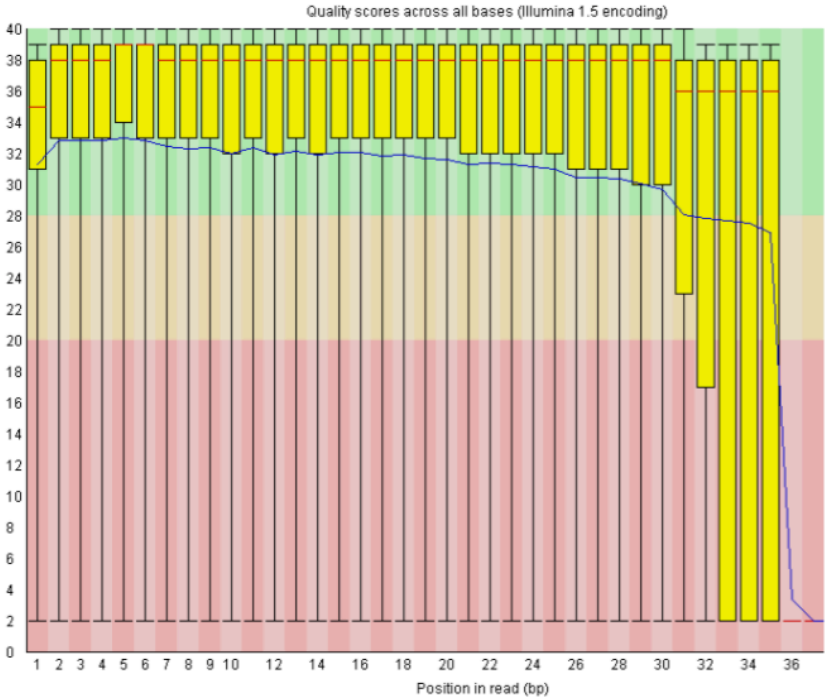
Per base sequence quality



Per base sequence quality

Tackling quality problems (**Trimming**)

- Method 1:
  - Keep all reads as is
  - Map as many as possible
- Method 2:
  - Drop all poor-quality reads
  - Trim poor-quality bases
  - Map only good-quality bases



- Which makes more sense for your experiment?

Mapping versus alignment

- **Mapping**

  The process of finding the region where a read is placed in the context of a ref genome

- **Alignment**

  The exact positioning of each base in a read in a target region.

Short read alignment

- Map each read to the reference genome

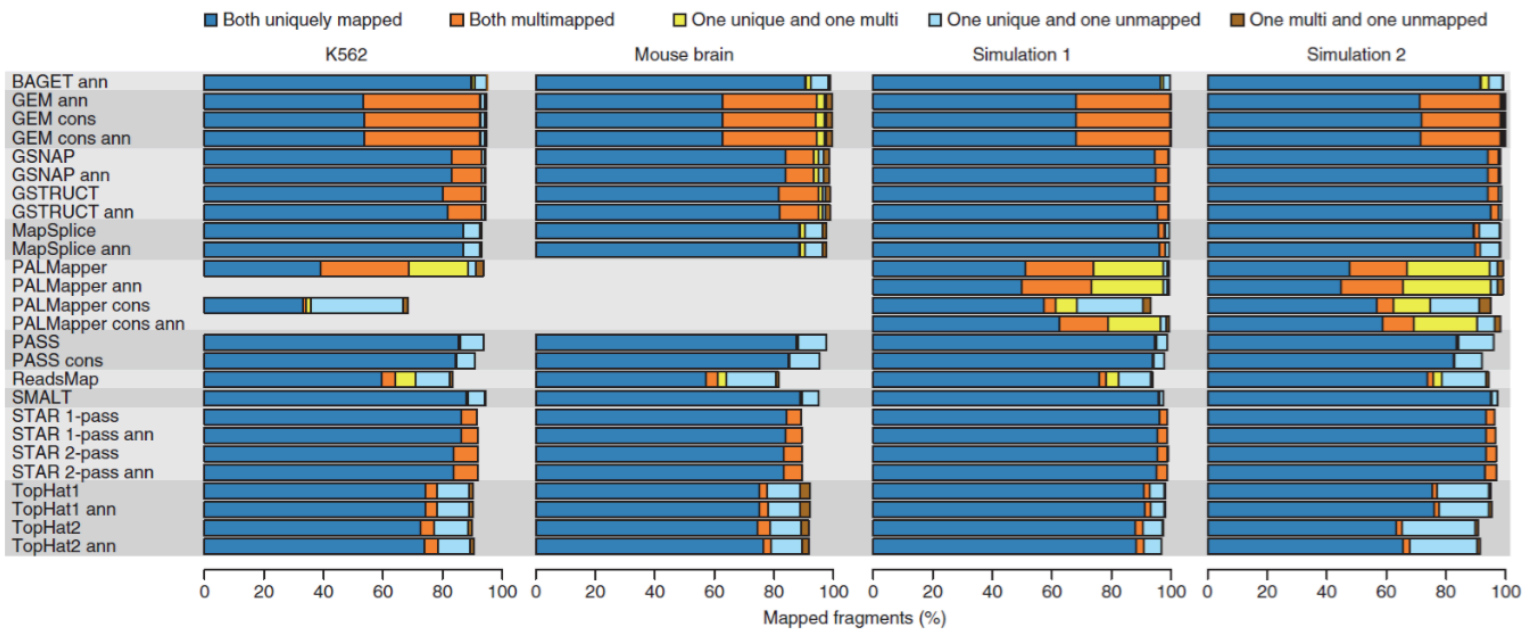
- Why not BLAST (or BLAT)?
  - optimized for longer reads : early NGS reads were <= 36 bp
  - not base quality aware
  - no means to judge uniqueness / non-uniqueness of alignments such as repetitive regions
  - too slow

Conclusion => Need for specialized short read mappers with highly efficient algorithms

Mapping short Reads

Various mappers for spliced alignment:

TopHat (recommended), STAR, GSNAP, MapSplice, PALMapper,ReadsMap, GEM, PASS, GSTRUCT, BAGET…



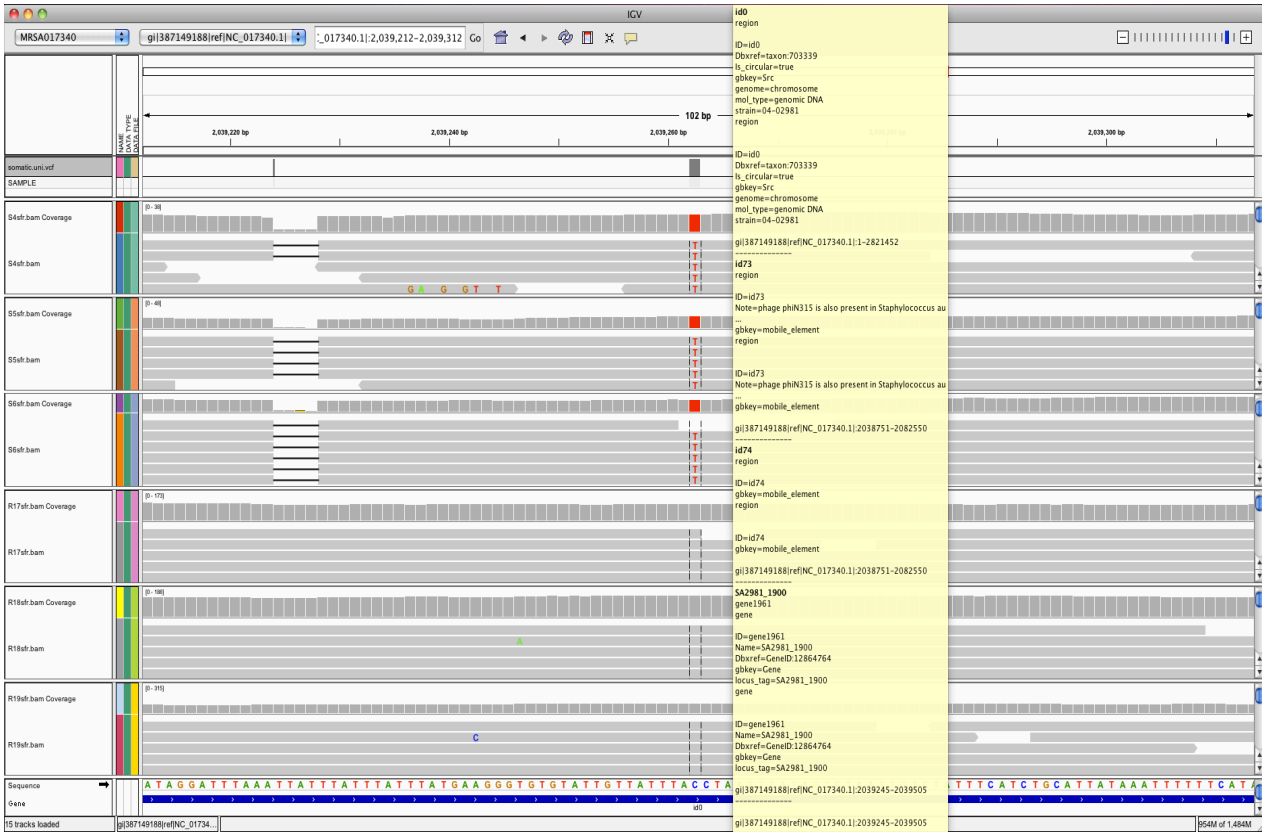
Pär G Engström *et al.* **Nature Methods** | VOL.10 NO.12 | DECEMBER 2013

Visualize /inspect short reads

Visualizing short reads

1-Integrative Genome Viewer (IGV)

http://www.broadinstitute.org/igv/home



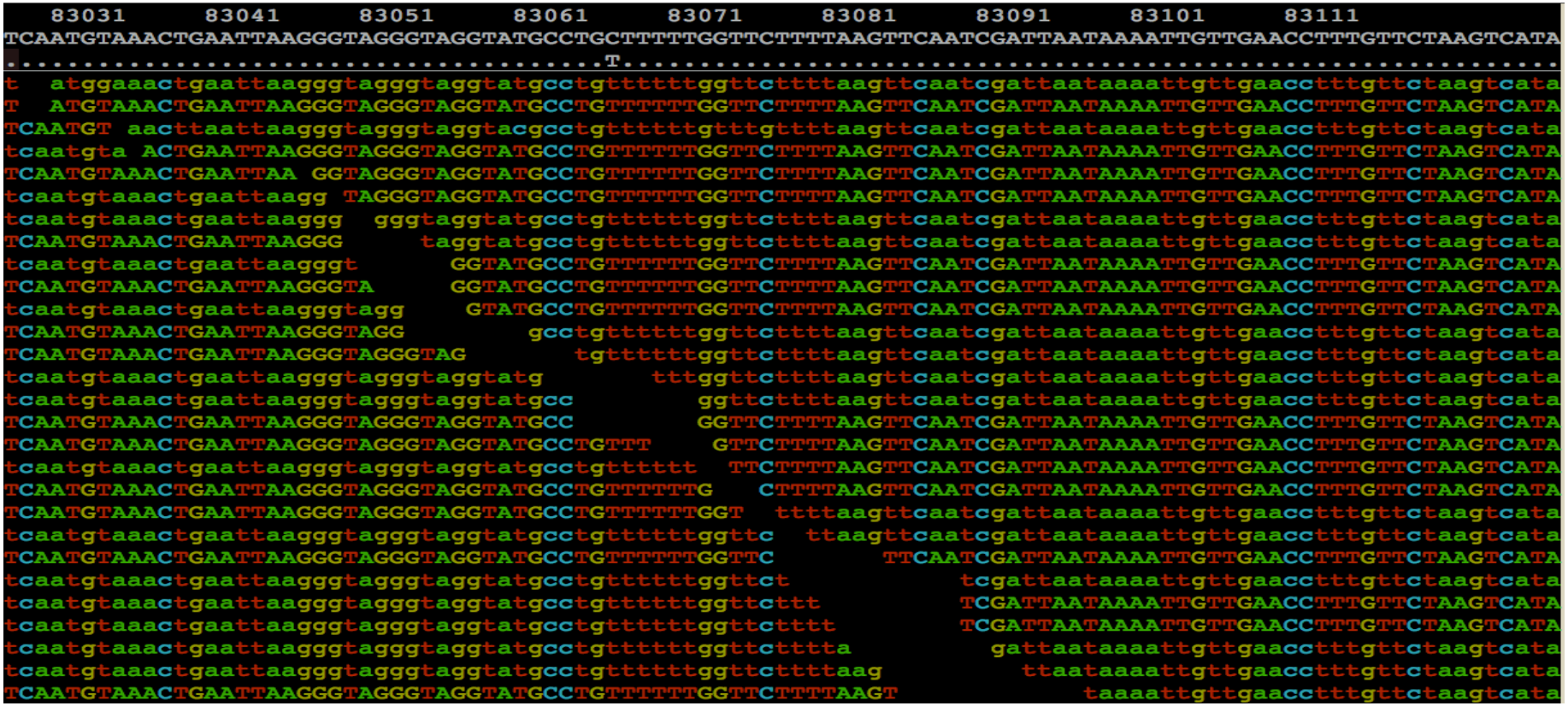
Integrative Genomics Viewer

Remap if necessary

Visualizing short reads

2- Samtools viewer

$ samtools tview myfile.bam reference.fasta



Counting mapped reads

- **htseq-count (recommended)**

  http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

  – Output is raw counts

- Cufflinks

  http://cufflinks.cbcb.umd.edu

  – Output is FPKM and related statistics

- Bedtools (intersectBed; coverageBed)

  http://code.google.com/p/bedtools/

  – Output is raw counts (but may need post-processing)

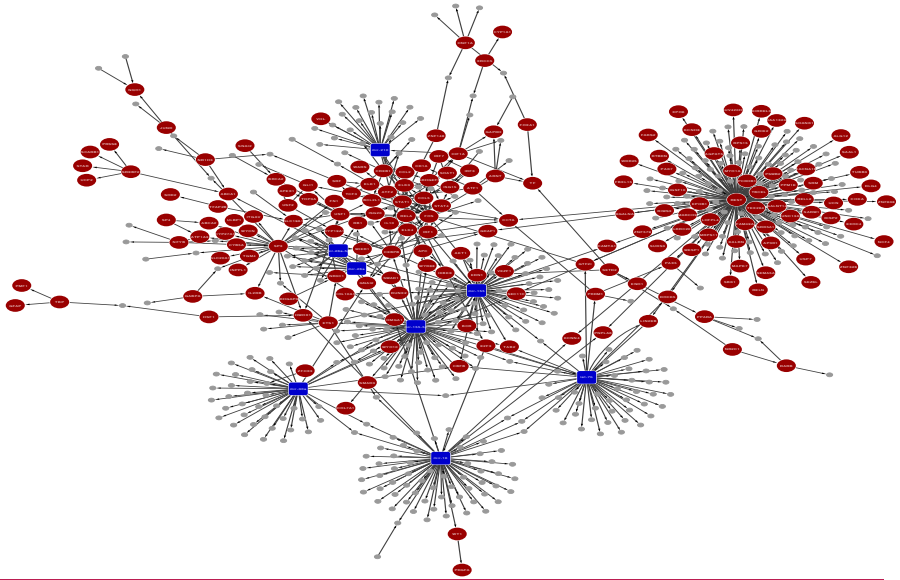
Differential expression analysis

- Normalization
- Log transformation
- DEGs identification

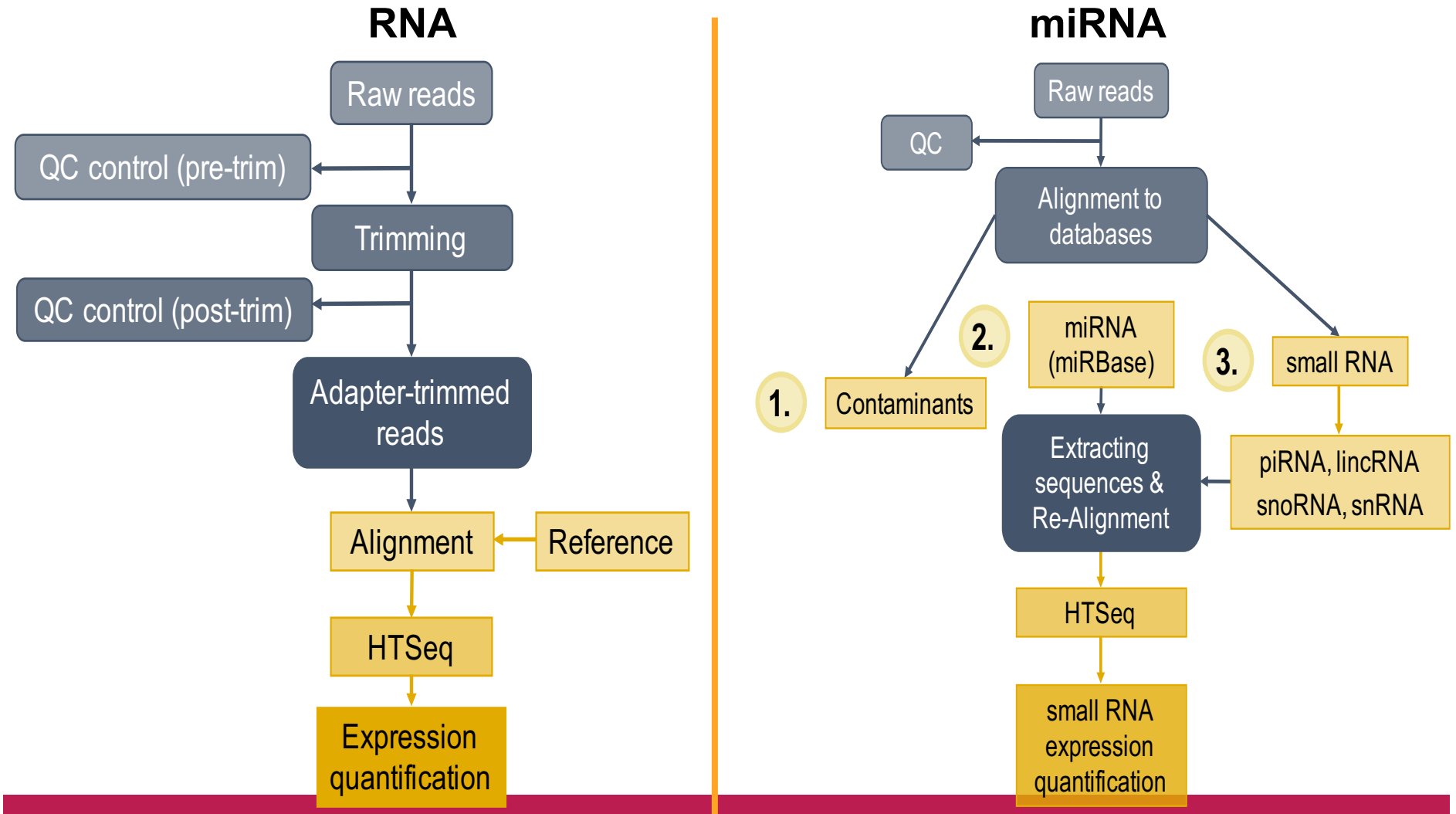
P.S: all the above steps are implemented in DESeq, cuffdiff

Downstream analysis

- Heatmap and clustering
- Functional enrichment analysis
  - GO terms
  - KEGG Pathways
  - miRNA targets
  - TFBS
  - Gene family (MSigDB)
  - …....
- Regulatory network construction



A pre-processing pipelines in practice



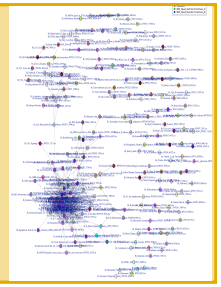
Downstream analysis in practice

Phenotype file

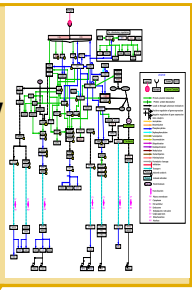
Raw count files

Preprocessing

Pairwise / three-way combined transcriptome analysis



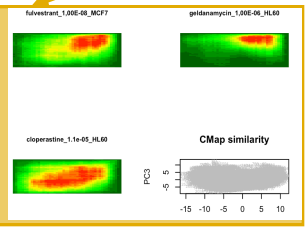
Pathway analysis



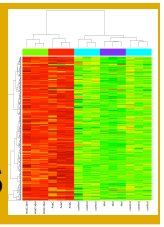
Differential expression analysis

Differential co-expression analysis

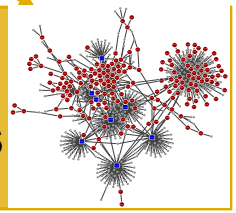
Single drug vs. CMap analysis



Combination effects / Drug synergy analysis



GRN analysis



General datasets

GTEx Portal

GTEx | Datasets | Expression | QTLs & Browsers | Sample Data | Biobank | Documentation | Publications | Contact | FAQs | Search Gene or SNP ID... | Sign In

2018-09-20
Reuse GTEx Visualizations On Your Website
Read More >>

Current Release

Latest Version: V7

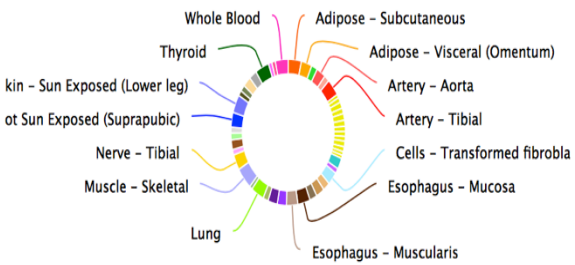
Download | Summary Statistics | How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

The current release is V7 including 11,688 samples, 53 tissues and 714 donors.

Browse eQTL Tissues

Total samples in all eQTL tissues: 10294

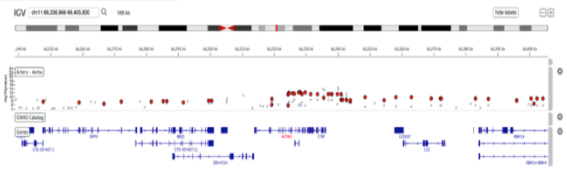


Genetic Association

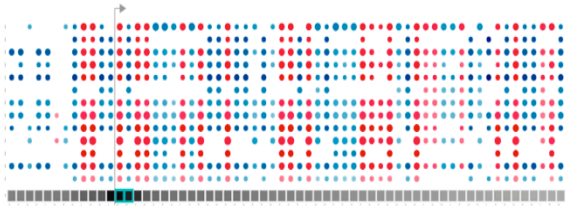
Single-Tissue eQTLs

Search eQTL by gene or SNP ID

GTEx IGV eQTL Browser



GTEx Gene-eQTL Visualizer



View eQTL data of a gene...

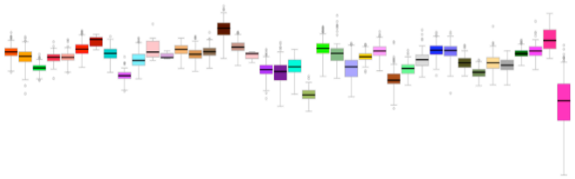
GTEx eQTL Calculator

Transcriptome

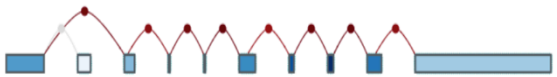
Search expression by gene ID...

Top Expressed Genes in a Tissue

Gene Expression in Tissues



Exon and Transcript Expression



News & Events

Use The GTEx Portal On Your Phone (2018-09-20)
Read More >>

Reuse GTEx Visualizations On Your Website (2018-09-20)
Read More >>

Access GTEx Data Programmatically with the GTEx API (2018-09-20)
Read More >>

NCBI

DATASET BROWSER (CURATED)

GEO Gene Expression Omnibus

Search for [ ] Search Clear Show All **Advanced Search**

Page size 20

4348 DataSet records

Page 1 of 218 > >>

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6248	Diet-induced obesity model: liver	*Mus musculus*	GPL6887	GSE39549	51
GDS6247	Diet-induced obesity model: white adipose tissue	*Mus musculus*	GPL6887	GSE39549	40
GDS6177	Acute alcohol consumption effect on whole blood (control group): time course	*Homo sapiens*	GPL570	GSE20489	25
GDS6176	Caspase-1 deficiency effect on lipid-loaded intestines	*Mus musculus*	GPL11533	GSE32515	18
GDS6100	MicroRNA-135b overexpression effect on prostate cancer cell line: time course	*Homo sapiens*	GPL10558	GSE57820	12
GDS6083	Chronic lymphocytic leukemia cells response to the neutralization of inhibitor of apoptosis proteins	*Homo sapiens*	GPL570	GSE62533	12
GDS6082	Sendai virus infection effect on monocytic cell line: dose response	*Homo sapiens*	GPL10558	GSE67198	11
GDS6064	Arthritic tarsal joints induced by collagen: time course	*Mus musculus*	GPL6246	GSE61140	15
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6016	Transcription factor engrailed-2 loss-of-function model of autism spectrum disorder: hippocampus	*Mus musculus*	GPL7202	GSE51612	6

DataSet Record GDS6248: Expression Profiles | Data Analysis Tools | Sample Subsets

Title: Diet-induced obesity model: liver

Summary: Analysis of livers of C57BL/6J mice fed a high fat diet for up to 24 weeks. Significant body weight gain was observed after 4 weeks. Results provide insight into the effect of high fat diets on metabolism in the liver.

Organism: *Mus musculus*

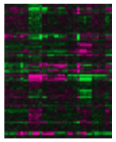
Platform: GPL6887: Illumina MouseWG-6 v2.0 expression beadchip

Citations: Kwon EY, Shin SK, Cho YY, Jung UJ et al. Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity. *BMC Genomics* 2012 Sep 4;13:450. PMID: 22947075

Do GM, Oh HY, Kwon EY, Cho YY et al. Long-term adaptation of global transcription and metabolism in the liver of high-fat diet-fed C57BL/6J mice. *Mol Nutr Food Res* 2011 Sep;55 Suppl 2:S173-85. PMID: 21618427

Reference Series:	GSE39549	**Sample count:**	51
Value type:	transformed count	**Series published:**	2014/03/01

Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

advanced search

ArrayExpress – functional genomics data

ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.

Browse ArrayExpress

Data Content

Updated today at 03:00

- 71502 experiments
- 2315758 assays
- 47.61 TB of archived data

Latest News

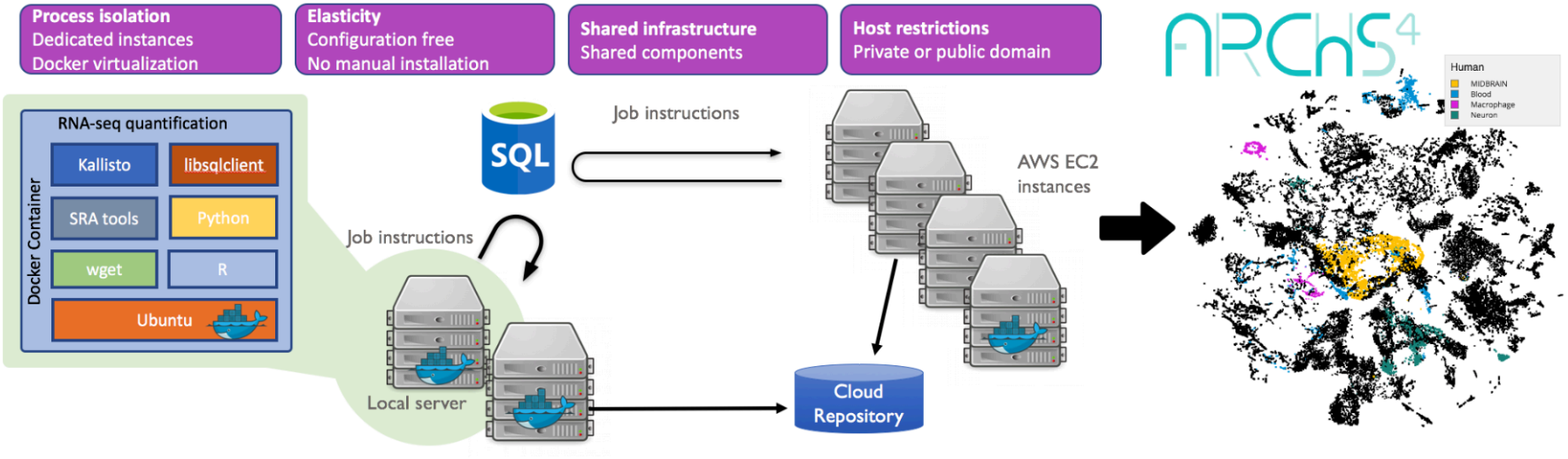
30 October 2018 - **A New and Improved Annotare has been released!**

Recently, we released a new version of Annotare designed to simplify and speed up the submission process by introducing several novel submission templates including a dedicated template for single-cell sequencing experiments. The templates pre-populate your submission with required sample attribute categories thus making it easier for submitters to know what type of information they need to provide with each experiment and sample type. By using the Annotare templates you will reduce the likelihood of being asked for additional metadata by our curation team and thus help us to process your submission more quickly.

You can find more details about the updated Annotare and the new templates here.

We're always looking to improve our service and very much appreciate any feedback from our users – please let us know what you think about the new Annotare and how we can further improve it to make the submission process as smooth as possible. Contact us at annotare@ebi.ac.uk

ARCHS4: Massive Mining of Publicly Available RNA-seq Data from Human and Mouse



Get Started

Species

Human | Mouse | Sample | Gene

Search

Metadata | Signature | Enrichment

Metadata Search

Human examples:
GSE81547, GSM2679484, Macrophage, Brain

MG63 Search

Tissue Types

- Cardiovascular System
- Connective Tissue
- Digestive System
- Immune System
- Integumentary System
- Muscular System
- Nervous System
- Respiratory System
- Urogenital System

Cell Lines

- Bone
- Brain
- Breast / Mammary
- Crevix
- Colon
- Connective

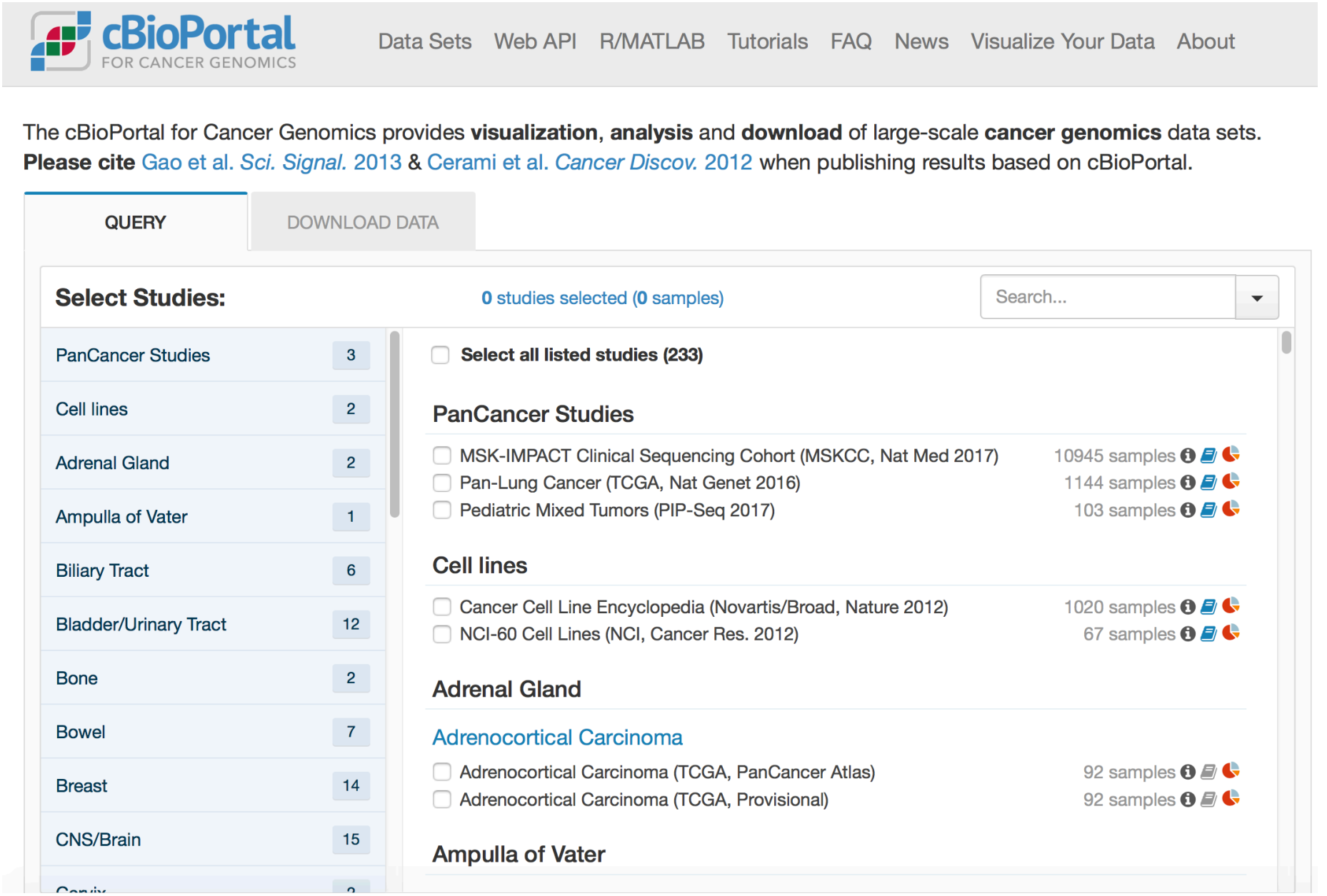
t-SNE view

Search Result

Samples

	Description	Organism	Samples	Series	Download	Delete
	MG63	Human	15	3	R	✖

Oncology-specific datasets


cBioPortal
FOR CANCER GENOMICS
Data Sets
Web API
R/MATLAB
Tutorials
FAQ
News
Visualize Your Data
About
The cBioPortal for Cancer Genomics provides **visualization**, **analysis** and **download** of large-scale **cancer genomics** data sets.
Please cite Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012 when publishing results based on cBioPortal.
QUERY
DOWNLOAD DATA
Select Studies:
0 studies selected (0 samples)
Search...
PanCancer Studies 3
Cell lines 2
Adrenal Gland 2
Ampulla of Vater 1
Biliary Tract 6
Bladder/Urinary Tract 12
Bone 2
Bowel 7
Breast 14
CNS/Brain 15
Select all listed studies (233)
PanCancer Studies
MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017) 10945 samples
Pan-Lung Cancer (TCGA, Nat Genet 2016) 1144 samples
Pediatric Mixed Tumors (PIP-Seq 2017) 103 samples
Cell lines
Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012) 1020 samples
NCI-60 Cell Lines (NCI, Cancer Res. 2012) 67 samples
Adrenal Gland
Adrenocortical Carcinoma
Adrenocortical Carcinoma (TCGA, PanCancer Atlas) 92 samples
Adrenocortical Carcinoma (TCGA, Provisional) 92 samples
Ampulla of Vater

R/Matlab

CGDS-R Package

Description

The CGDS-R package provides a basic set of functions for querying the Cancer Genomic Data Server (CGDS) via
Maintained by Anders Jacobsen at the Computational Biology Center, MSKCC.

Documentation

- CGDS-R Package on CRAN
- The CGDS-R reference manual
- The CGDS-R documentation vignette

Installation

1. The CDGS-R package currently **only works with R Version 2.12 or higher**.
2. Then install the cgds-R package from within R: `install.packages('cgdsr')`

ICGC Data Portal

Cancer Projects | Advanced Search | Data Analysis | DCC Data Releases | Data Repositories

Cancer genomics data sets visualization, analysis and download.

Quick Search

Search

e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

Advanced Search

By donors | By genes | By mutations

Data Release 27

April 30th, 2018

Cancer projects	84
Cancer primary sites	22
Donor with molecular data in DCC	20,487
Total Donors	24,077
Simple somatic mutations	77,462,290

Download Release

ICGC Data Portal

Quick Search

DATA ANALYSIS

aunch Analysis | Saved Sets

Enrichment Analysis

Find over-represented groups of gene sets (e.g. Reactome pathways) that are of statstical significance when comparing with your gene set.

Select | Demo

Demo showing top 50 genes in Cancer Gene Census.

Set Operations

Display Venn diagram and find intersection or union, etc. of your sets of the same type.

Select | Demo

Cohort Comparison

Display the survival analysis of your donor sets and compare characteristics such as gender, vital status and age at diagnosis between your donor sets.

Select | Demo

OncoGrid

Visualize genetic alterations affecting a set of donors.

Select | Demo

Jupyter Notebooks | Powered by Cancer Genome Collaboratory

The Jupyter Notebook sandbox is provided by ICGC as a way of programmatically interacting with, analyzing, and visualizing data. Users with DACO approval will be able to log in and use the sandbox as a place to experiment and explore. We recommend users download any notebooks of value and interest as the sandbox will be periodically refreshed to ensure continued operation for all users.

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 472 | GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

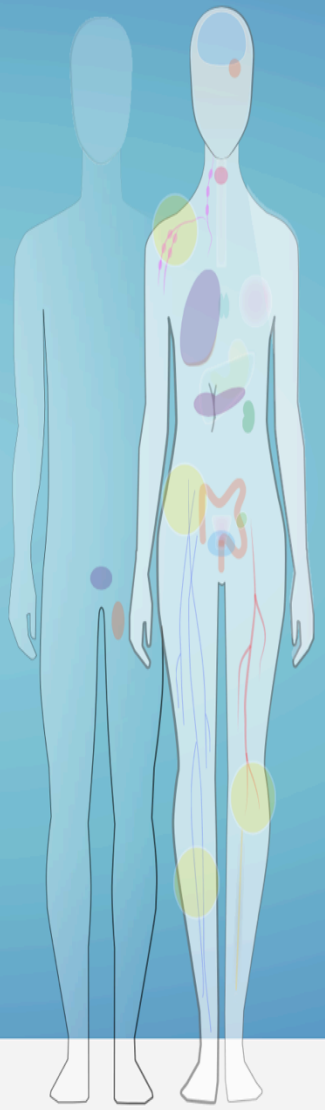
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

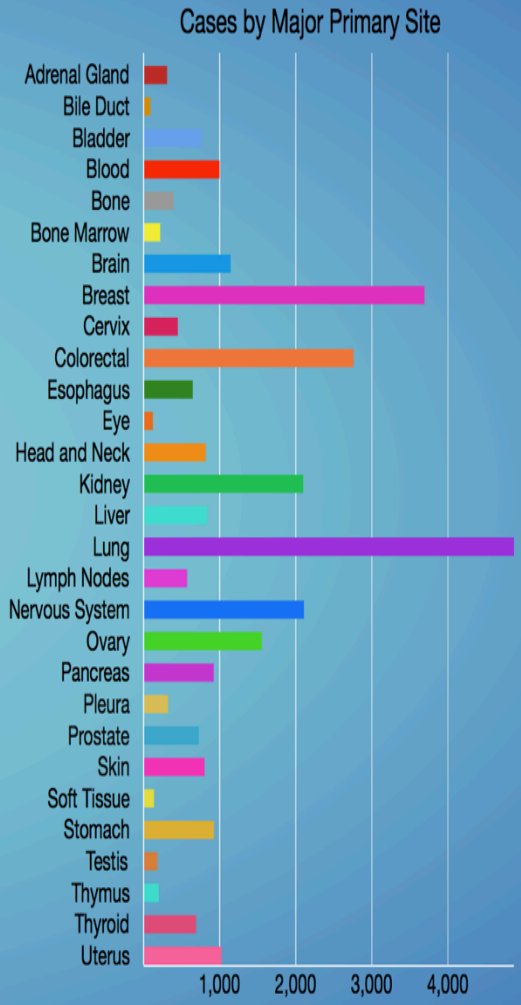
Data Portal Summary

Data Release 13.0 - September 27, 2018

PROJECTS	PRIMARY SITES	CASES
43	69	33,096

FILES	GENES	MUTATIONS
358,092	22,872	3,142,246





TCGAbiolinks

platforms all | rank 95 / 1649 | posts 18 / 0.6 / 2 / 3 | in Bioc 3 years
build error | updated < 1 week

DOI: 10.18129/B9.bioc.TCGAbiolinks

TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

Bioconductor version: Release (3.8)

The aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses and iv) to easily reproduce earlier research results. In more detail, the package provides multiple methods for analysis (e.g., differential expression analysis, identifying differentially methylated regions) and methods for visualization (e.g., survival plots, volcano plots, starburst plots) in order to easily develop complete analysis pipelines.

Author: Antonio Colaprico, Tiago Chedraoui Silva, Catharina Olsen, Luciano Garofano, Davide Garolini, Claudia Cava, Thais Sabedot, Tathiane Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr

Maintainer: Antonio Colaprico <antonio.colaprico at ulb.ac.be>, Tiago Chedraoui Silva <tiagochst at usp.br>

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

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Online Analysis Tools
For Full Analysis Path

BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help Sign In

BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface

Get Started

To acknowledge BioJupies in your publications, please use the following reference: Torre, D., Lachmann, A., and Ma'ayan, A. (2018). BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. Cell Systems.

BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help Sign In

BioJupies Automatically Generates RNA-seq Data Analysis Notebooks

With BioJupies you can produce in seconds a customized, reusable, and interactive report from your own raw or processed RNA-seq data through a simple user interface

Get Started

To acknowledge BioJupies in your publications, please use the following reference: Torre, D., Lachmann, A., and Ma'ayan, A. (2018). BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud. Cell Systems.

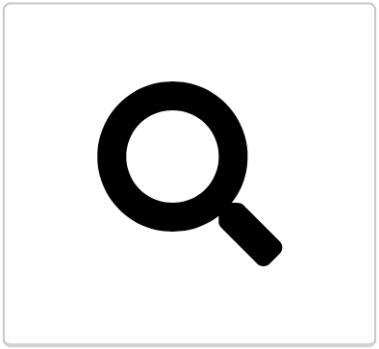
BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help

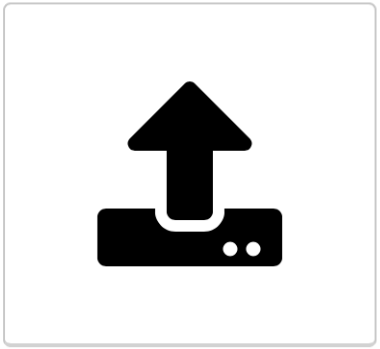
Sign In

What data would you like to analyze?



Published Data

Search thousands of published, publicly available datasets



Your Data

Upload your own gene expression data for analysis

Example Data

Learn to generate notebooks with an example dataset

BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help Sign In

Select an RNA-seq Data Repository

Gene Expression Omnibus

Search thousands of RNA-seq datasets published on GEO

GTEx

Analyze RNA-seq samples from the publicly available GTEx Portal

BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help

Sign In

Which dataset would you like to analyze?

Use the form below to **search 9,145 publicly available datasets** published in the Gene Expression Omnibus database and processed by ARCHS4.

cancer

Displaying 1-10 of 789 results | Organism: All | Sort by: Newest | Samples: 6 35 70+

Mutant p53R270H drives altered metabolism and increased invasion in pancreatic ductal adenocarcinoma

GSE106853 *12 samples* Published *September 2018*

Analyze > More Info

Replicated transcriptome profiling of Normal and Cancerous Prostate Cells [RNA-Seq]

GSE70466 *6 samples* Published *September 2018*

Analyze > More Info

RNA profiling of 6xAOM induced colon tumors from wt, miR-34a IEC deficient, p53 IEC deficient and miR-34a/p53 IEC deficient mice

GSE99452 *12 samples* Published *August 2018*

Analyze > More Info

BioJupies

Search GEO datasets (e.g. cancer, p53)

Chrome Extension Help Sign In

Which analyses would you like to perform?

Use the form below to **add or remove data analysis and visualization tools** to your notebook. These tools will analyze the selected dataset and embed interactive results in your notebook. Once you have selected the desired tools, click **Continue** to proceed.

Back Continue

Exploratory Data Analysis

These tools assist in visually exploring global patterns within the dataset.

PCA

Linear dimensionality reduction technique to visualize similarity between samples

Remove More Info

Clustergrammer

Interactive hierarchical clustering heatmap visualization

Remove More Info

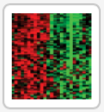
Library Size Analysis

Analysis of readcount distribution for the samples within the dataset

Remove More Info

Differential Expression Analysis

These tools allow for calculation and exploration of differential gene expression between two groups of samples.



Differential Expression Table

Differential expression analysis between two groups of samples

Add +

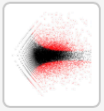
More Info

Volcano Plot

Plot the logFC and logP values resulting from a differential expression analysis

Add +

More Info



MA Plot

Plot the logFC and average expression values resulting from a differential expression analysis

Add +

More Info

Enrichment Analysis

These tools analyze the results of a differential expression analysis by placing results in context of prior knowledge.

Enrichr Links

Links to enrichment analysis results of the differentially expressed genes via Enrichr

Add +

More Info

Gene Ontology Enrichment Analysis

Identifies Gene Ontology terms which are enriched in the differentially expressed genes

Add +

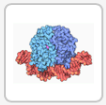
More Info

Pathway Enrichment Analysis

Identifies biological pathways which are enriched in the differentially expressed genes

Add +

More Info



Transcription Factor Enrichment Analysis

Identifies transcription factors whose targets are enriched in the differentially expressed genes

Add +

More Info

Kinase Enrichment Analysis

Identifies protein kinases whose substrates are enriched in the differentially expressed genes

Add +

More Info

miRNA Enrichment Analysis

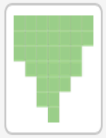
Identifies miRNAs whose targets are enriched in the differentially expressed genes

Add +

More Info

Small Molecule Queries

These tools allow for identification of small molecules to mimic or reverse differential gene expression signatures.



L1000CDS^2 Query

Identifies small molecules which mimic or reverse a given differential gene expression signature

Add +

More Info

L1000FWD Query

Projects signatures on a 2-dimensional visualization of the L1000 signature database

Add +

More Info

In [1]:
```
# Initialize Notebook
%run ../library/v1.0/init.ipy
HTML('''<script> code_show=true;  function code_toggle() {  if (code_show){  $('div.input').hide();  } else {  $('div.input').show();  }  code_show = !code_show }  $( document ).ready(code_toggle); </script> <form action="javascript:code_toggle()"><input type="submit" value="Toggle Code"></form>''')
```

Out[1]: Toggle Code

Control vs Perturbation Analysis Notebook | BioJupies

Introduction

This notebook contains an analyis of GEO dataset GSE106853 (https://www.ncbi.nlm.nih.gov/gds/?term=GSE106853) created using BioJupies. For more information on BioJupies, please visit http://biojupies.cloud. If the notebook is not correctly displayed on your browser, please visit our Notebook Troubleshooting Guide.

Table of Contents

The notebook is divided into the following sections: