Lecture Notes

Big Data in Medical Informatics

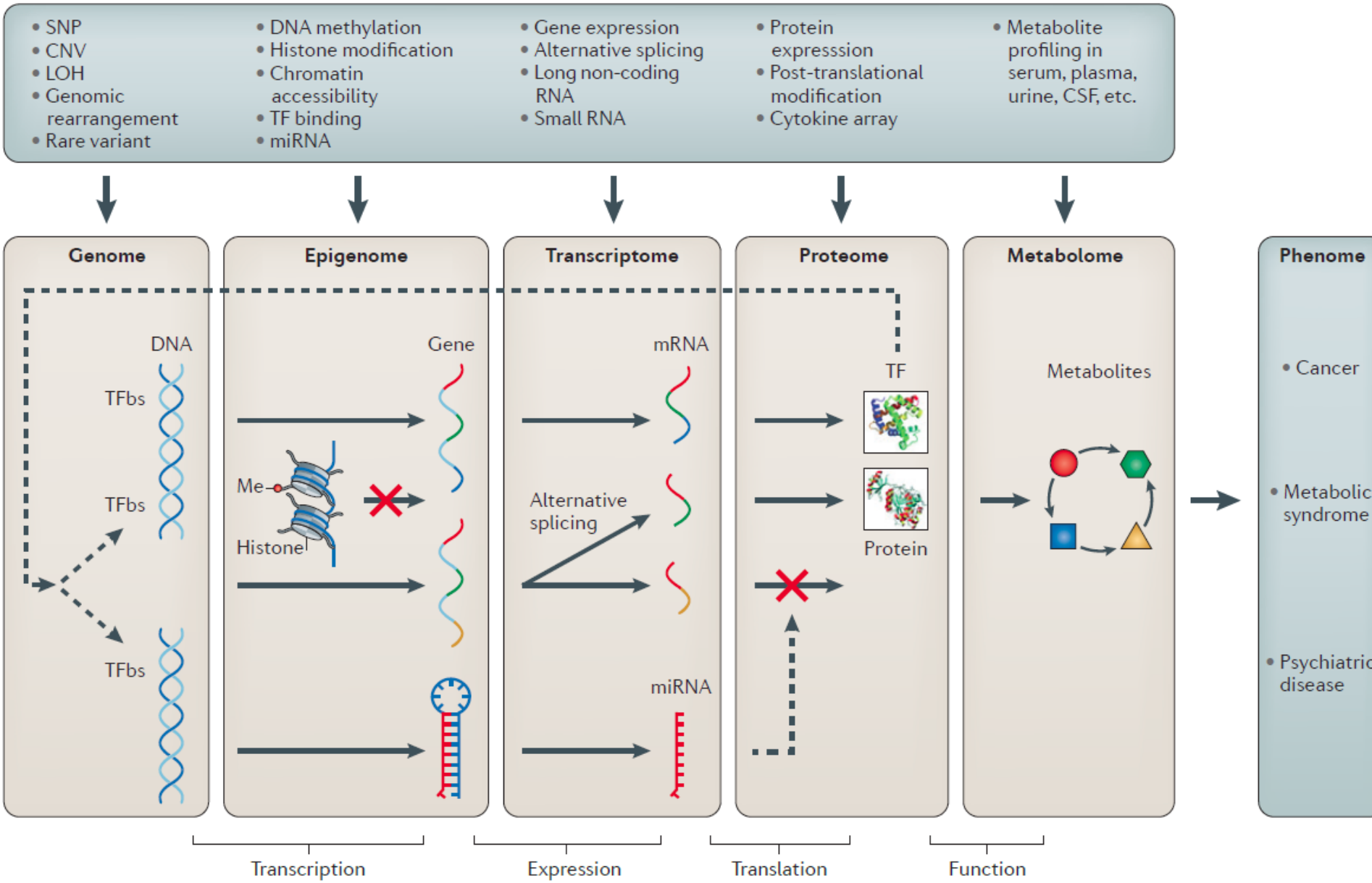
Week 13:
Genomic Data Analysis – Part 2

RWTH AACHEN UNIVERSITY

Genomic Data Analysis

- Why do we analyze genomic data?
  - to generate new insights into the biology of human disease
  - to predict the individual response to treatment
  - to enhance the understanding of the underlying mechanisms
  - to promote the knowledge exchange between doctors and patients,
  - To facilitate clinical decision making

Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome to the phenome



Ritchie, Marylyn D., et al. "Methods of integrating data to uncover genotype-phenotype interactions." *Nature Reviews Genetics* 16.2 (2015): 85-97.

Genomic Data Analysis

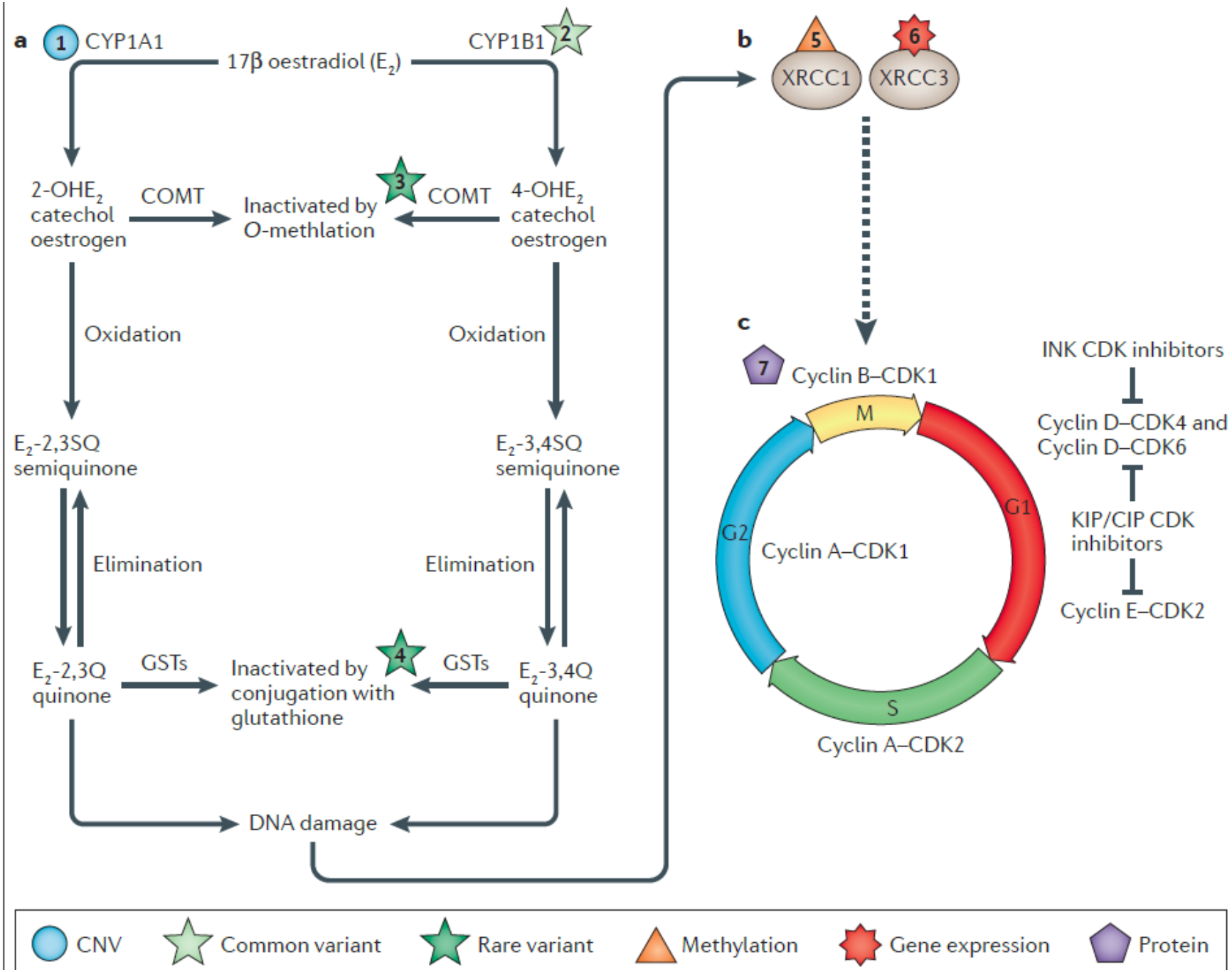
How data-driven approaches facilitate the generation of new discoveries and insights into biology?

- The genomic landscapes in complex diseases (e.g. cancers) are overwhelmingly complicated, and reveals a high order of heterogeneity among different individuals

- Questions:
  - if any of mutations are indeed responsible for the development of the diseases?
  - if yes, how to identify these real contributors
  - when multiple mutations are involved, can we infer the evolutionary relation they may have against each other?

Genomic Data Analysis

- very simple and straightforward approach:
  - catalog all the genetic changes in many samples so that one can identify the common changes across individuals with the same or different cancers.
  - Group the common mutations, and identify the genetic changes linking to onset or progress of the disease
  - Determine if the changes reflect genomic regions that are associated with clinical responses or can be targeted by a specific drug.
  - The evolutionary patterns among these changes can be studied based on the diverging lineages among different genetic populations.



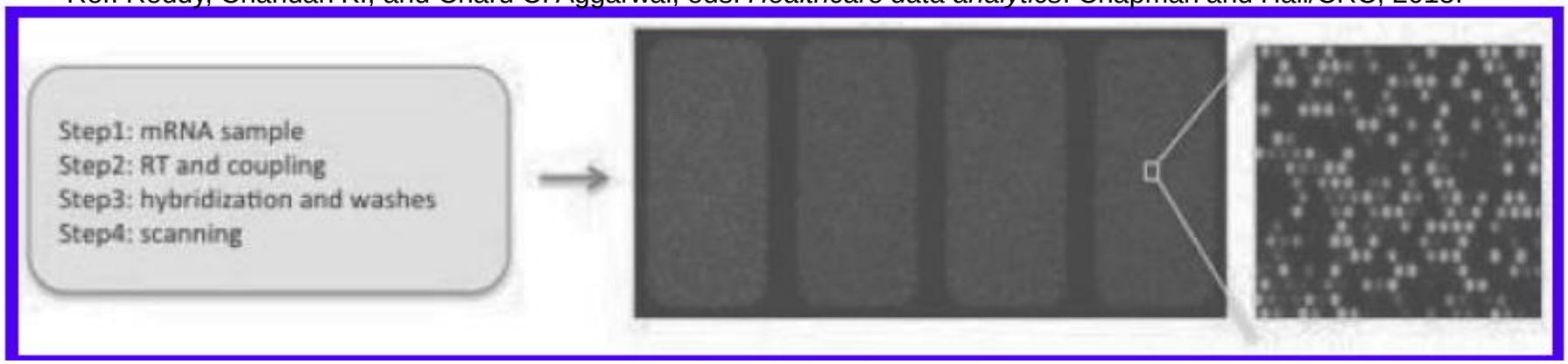
Ritchie, Marylyn D., et al. "Methods of integrating data to uncover genotype-phenotype interactions." *Nature Reviews Genetics* 16.2 (2015): 85-97.

Genomic Data Generation

- Different types of omics data including genomics, epigenetics, proteomics, and metabolomics data are generated by the state-of-the-art high throughput technologies as well as conventional biological experiments.

- Microarray Data
  - microarray (also known as gene/protein-chips) and mass spectrometry MS) are widely used to determine the presence and abundance of genes, proteins, and metabolites in biological samples including tissues, cells, blood, and urine.

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

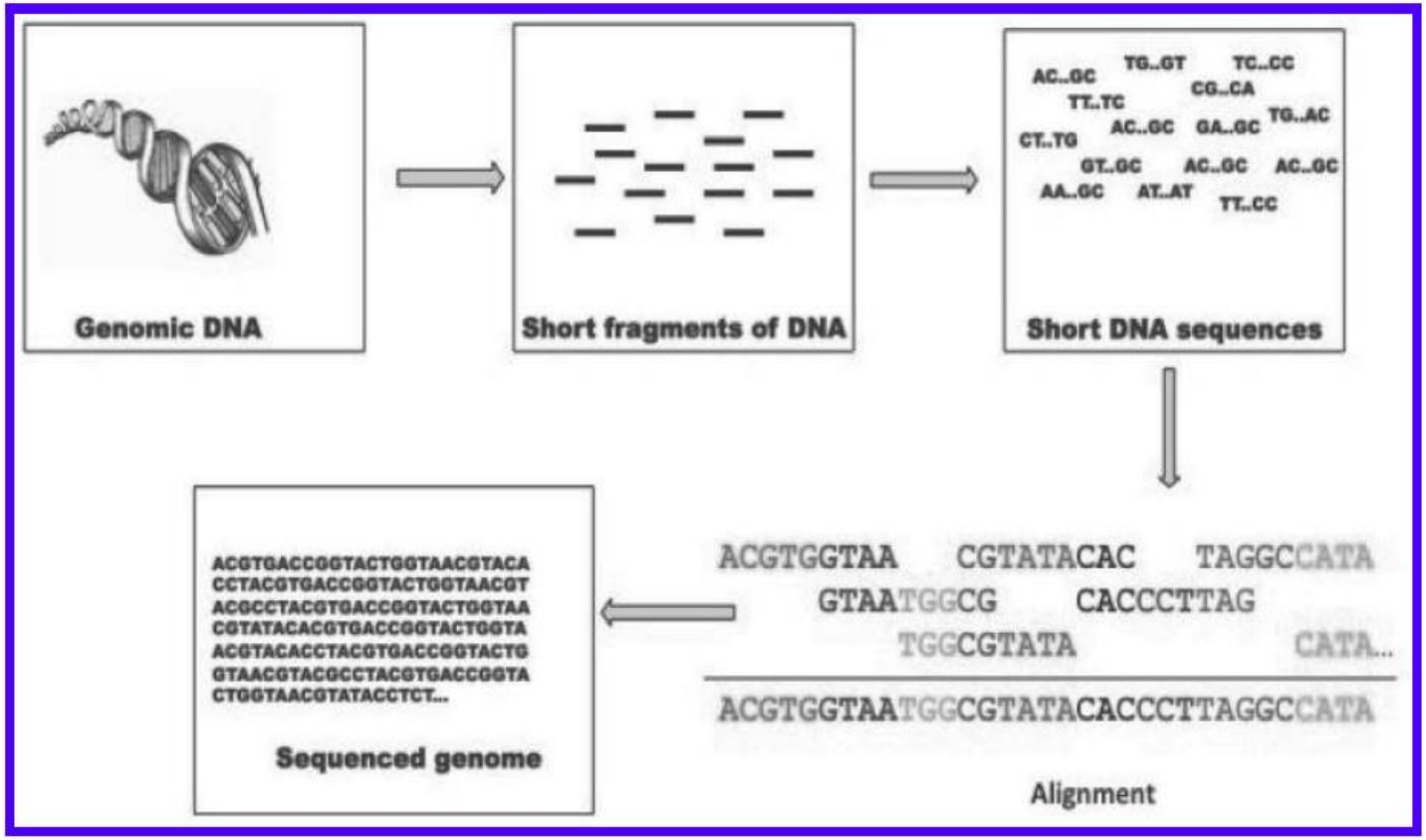


Scanned image data generated from standard DNA microarray protocols, e.g., gene array platform from Affymetrix, Agilent and ALMAC, where the signals extracted from the scanned array image reflect the gene abundance

Genomic Data Generation

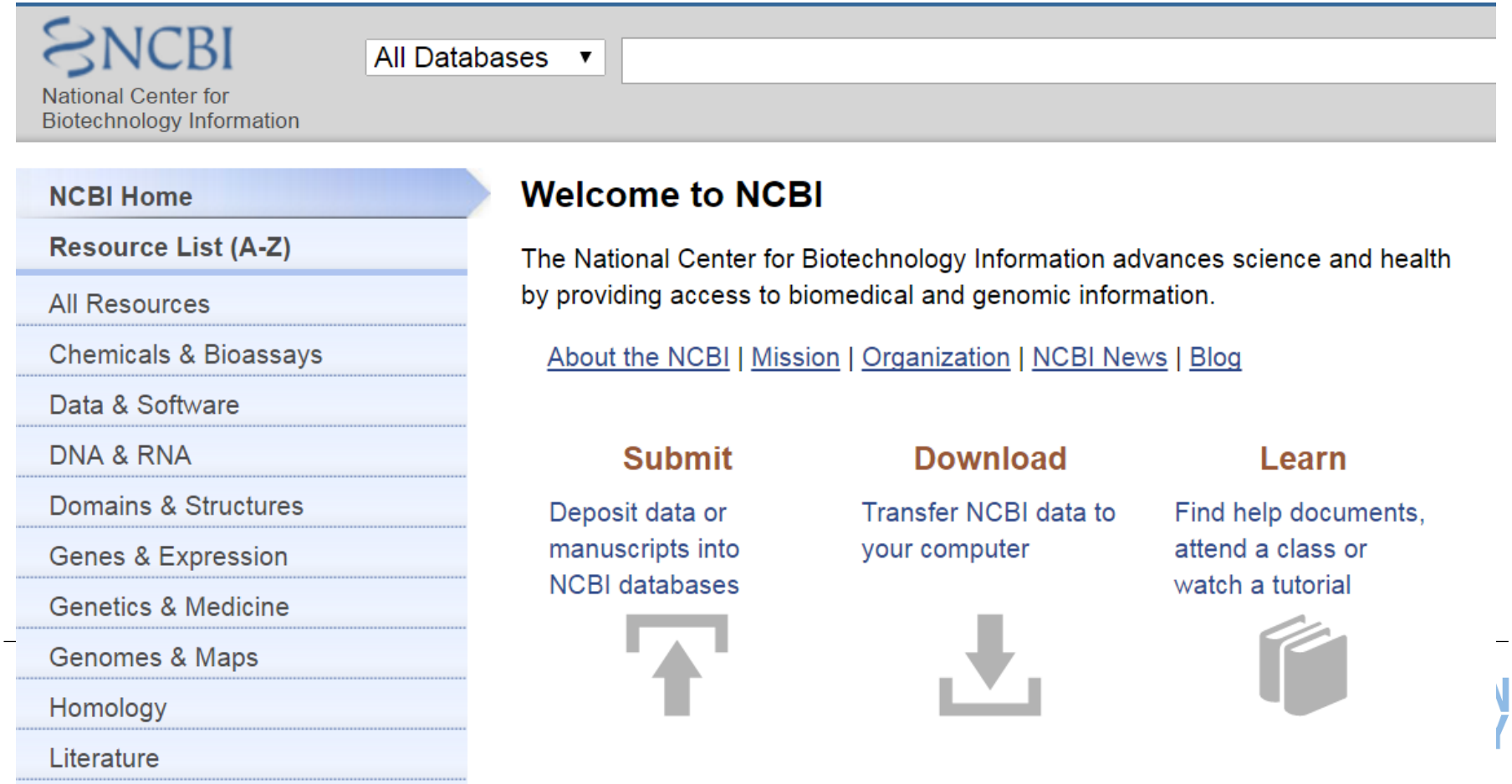
- Next-Generation Sequencing
  - Any given single genomics DNA is first fragmented into a library of small segments that can be uniformly and accurately sequenced in millions of parallel reactions.
  - The identified strings of bases, called reads,
  - Reads are then assembled through aligning to a known reference genome (resequencing), or in the absence of a reference genome (de novo sequencing).

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.



Public Repositories for Genomic Data

- Repositories of biological information are so essential for biomedical or bioinformatics studies as they organize a large variety of biological data and enable researchers to get access to the structured information and utilize them in their respective researches
- NCBI database (http://www.ncbi.nlm.nih.gov/



Public Repositories for Genomic Data

- **Human genomes, mutations and epigenome databases**:
  - HGMD (Human Genome Mutation Database): contains 141,161 germline mutations associated with human inheritable diseases
  - dbSNP database (Single Nucleotide Polymorphism Database): archives comprehensive genetic variation data across different species.
  - TCGA (The Cancer Genome Atlas) and ICGC (International Cancer
  - Genome Consortium) [74] : are two of the largest cancer genome projects to sequence thousands of whole genomes, along with other types of omic data, for many cancer types.
  - COSMIC (Catalog of Somatic Mutations In human Cancer) : large cancer genomic database which contains 1,592,109 gene mutations identified on 947,213 tumor samples.

Public Repositories for Genomic Data: dbGAP- PheGENI

NCBI Resources How To

PheGenI
Phenotype-Genotype Integrator

All Databases

Search Summary

Search Criteria
Phenotype Selection
Trait: Celiac Disease

Modify Search

Search Results

Association Results	1 - 50 of 50	Searched by phenotype trait.
Genes	1 - 50 of 62	Searched by gene IDs retrieved from association results.
SNPs	1 - 43 of 43	Searched by SNP rs numbers retrieved from association results.
eQTL Data	1 - 3 of 3	Searched by SNP rs numbers retrieved from association results.
dbGaP Studies	1 - 1 of 1	Searched by traits retrieved from association results.
Genome View	43 SNPs and 50 of 62 genes over 18 chromosomes.	

Modify Search Show All Hide All

Search Criteria

Association Results

1 - 50 of 50 Download Modify Search

#	Trait	rs #	Context	Gene	Location	P-value	Source	Study	PubMed
1	Celiac Disease	rs2187668	intron	HLA-DQA1	**6**: 32,605,884	1.000×10^{-50}	NHGRI		20190752
2	Celiac Disease	rs1464510	intron	LPP	**3**: 188,112,554	3.000×10^{-40}	NHGRI		20190752
3	Celiac Disease	rs17810546	intergenic	RPS2P19, IL12A	**3**: 159,665,050	4.000×10^{-28}	NHGRI		20190752
4	Celiac Disease	rs13151961	intron	KIAA1109	**4**: 123,115,502	2.000×10^{-27}	NHGRI		20190752
5	Celiac Disease	rs653178	intron	ATXN2	**12**: 112,007,756	7.000×10^{-21}	NHGRI		20190752

Public Repositories for Genomic Data: dbGAP- PheGENI

Genome View

Ideogram Setup

Orientation: Select

Include: Genes SNPs Location

Display: Current Subset

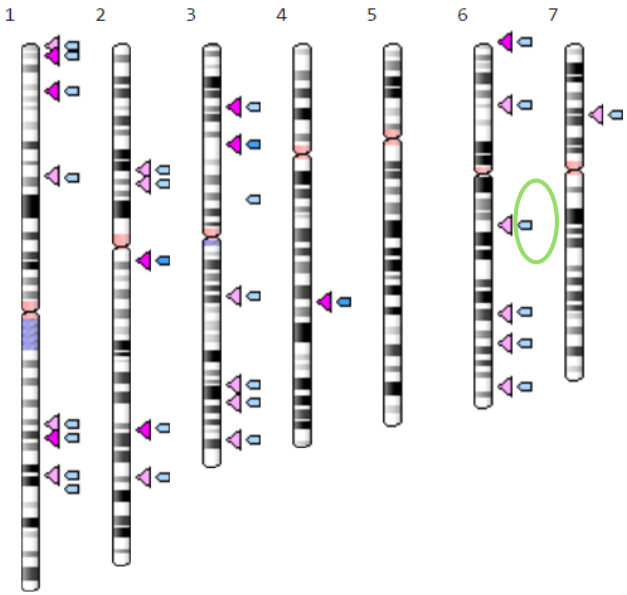
Chromosomes: All

Update Download

Summary

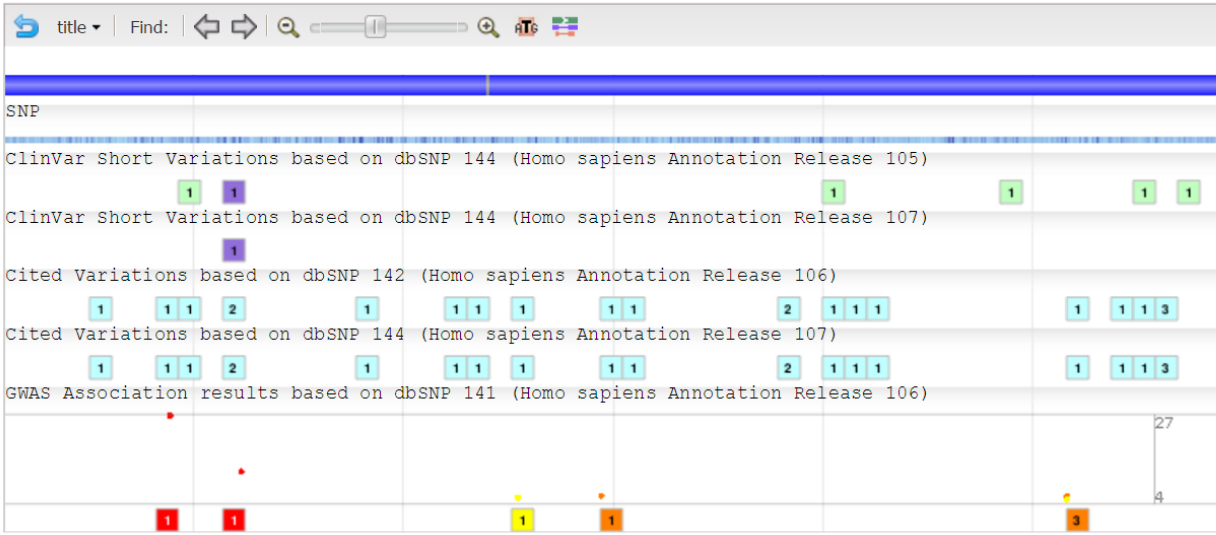
43 SNPs searched by SNP rs numbers retrieved from association results and 50 of 62 genes searched b
chromosomes.

SNP	Gene	Count
		1 SNP or gene
		2 - 10 SNPs or genes
		11 - 20 SNPs or genes
		more than 20 SNPs or genes



Click on ideogram annotation to show sequence display

Sequence Display
2 SNPs



Genes

1 - 50 of 62 < Previous Next > Page 1 Go Download Modify Search

Open All Close All

#	Symbol	Description	Location	OMIM
1	MMEL1	membrane metalloendopeptidase like 1	1: 2,633,042 - 2,590,639	

RWTH Informatik 5 | Ahornstr. 55 D-5205
Tel +49/241/8021501 | Fax +49/241/8022

Public Repositories for Genomic Data: ClinVar

NCBI Resources How To Sign in to NCBI

Variation Viewer

Homo sapiens: GRCh38.p7 (GCF_000001405.33) Chr 1 (NC_000001.11): 196.6M - 196.8M

Reset All Share this page FAQ Help Version 1.5.3

New to Variation Viewer? Read our quick overview! X

Pick Assembly

Search

Location, gene or phenotype

Enter a location, gene name or phenotype

Region CFH (Gene) NM_000186.3 (Transcript)

Exons: click an exon above to zoom in

Your Data

History

Region Details

Features of Interest

Other sequence representations - *None*

1 GRC genome issue in this view. Add Track

NC_000001.11: 197M..197M (115Kbp) Tools Tracks

Genes, NCBI Homo sapiens Annotation Release 108, 2016-06-07

CFH

ClinVar Short Variations based on dbSNP Build 149 (Homo sapiens ...

dbVar ClinVar Large Variations

dbSNP Build 149 (Homo sapiens Annotation Release 108) all

196690048..196690194

Variation ID: rs34815383, with benign allele
Location: 196,690,048
Variation Viewer: CFH
ClinVar: rs34815383

Variation ID: rs1061170, with pathogenic allele
Location: 196,690,107
Variation Viewer: CFH
ClinVar: rs1061170

Variation ID: rs121913061, with pathogenic allele
Location: 196,690,125
Variation Viewer: CFH
ClinVar: rs121913061

Variation ID: rs121913056, with pathogenic allele
Location: 196,690,194
Variation Viewer: CFH
ClinVar: rs121913056

Variation Data

Filter by

Source database
- dbSNP (6,827)
- dbVar (609)

In ClinVar
- Yes (30)
- No (7,406)

Most severe clinical significance
- Pathogenic (18)
- Likely pathogenic (0)
- drug response (0)

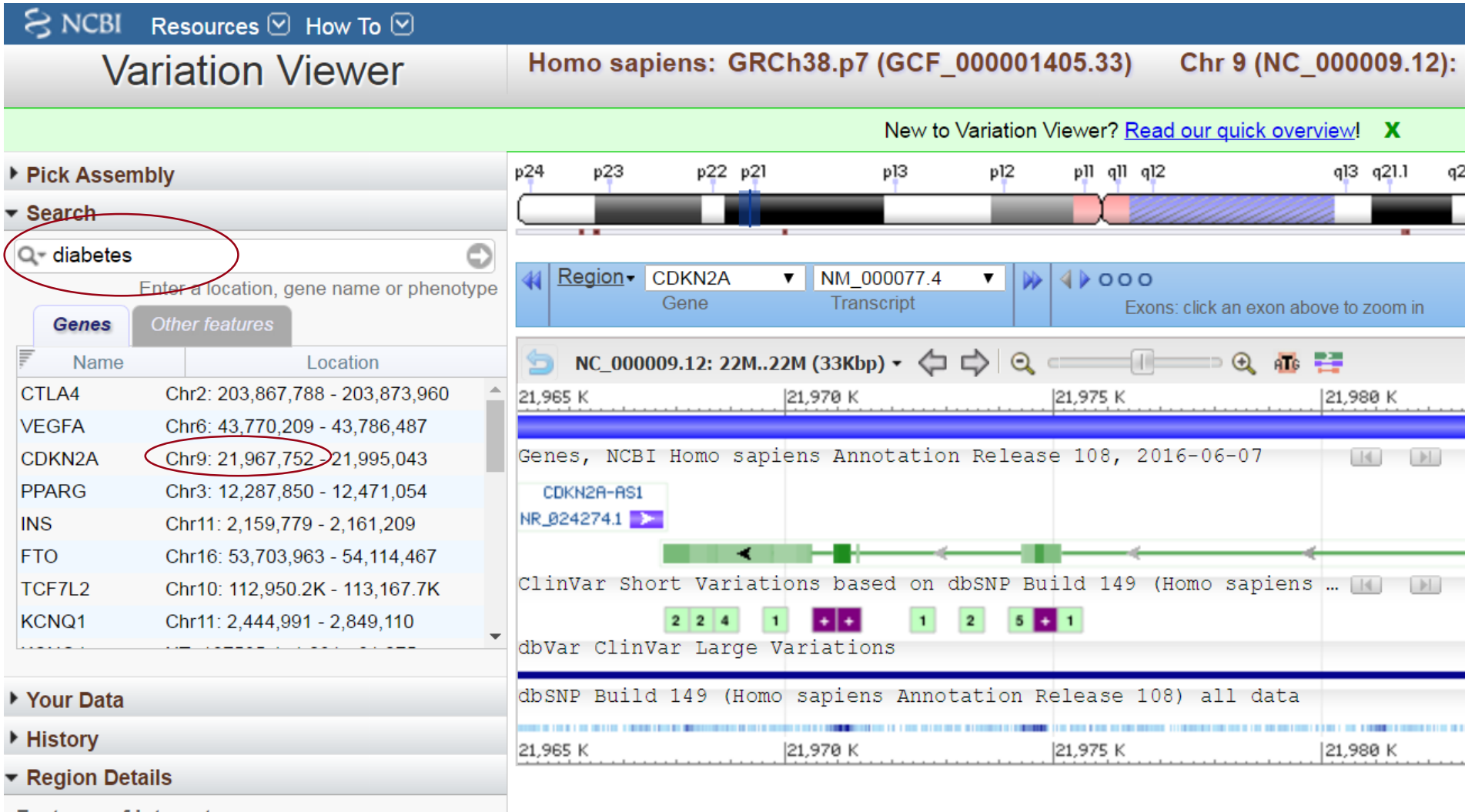
Download Edit columns

< Prev Page 1 of 248 Next > Last >>

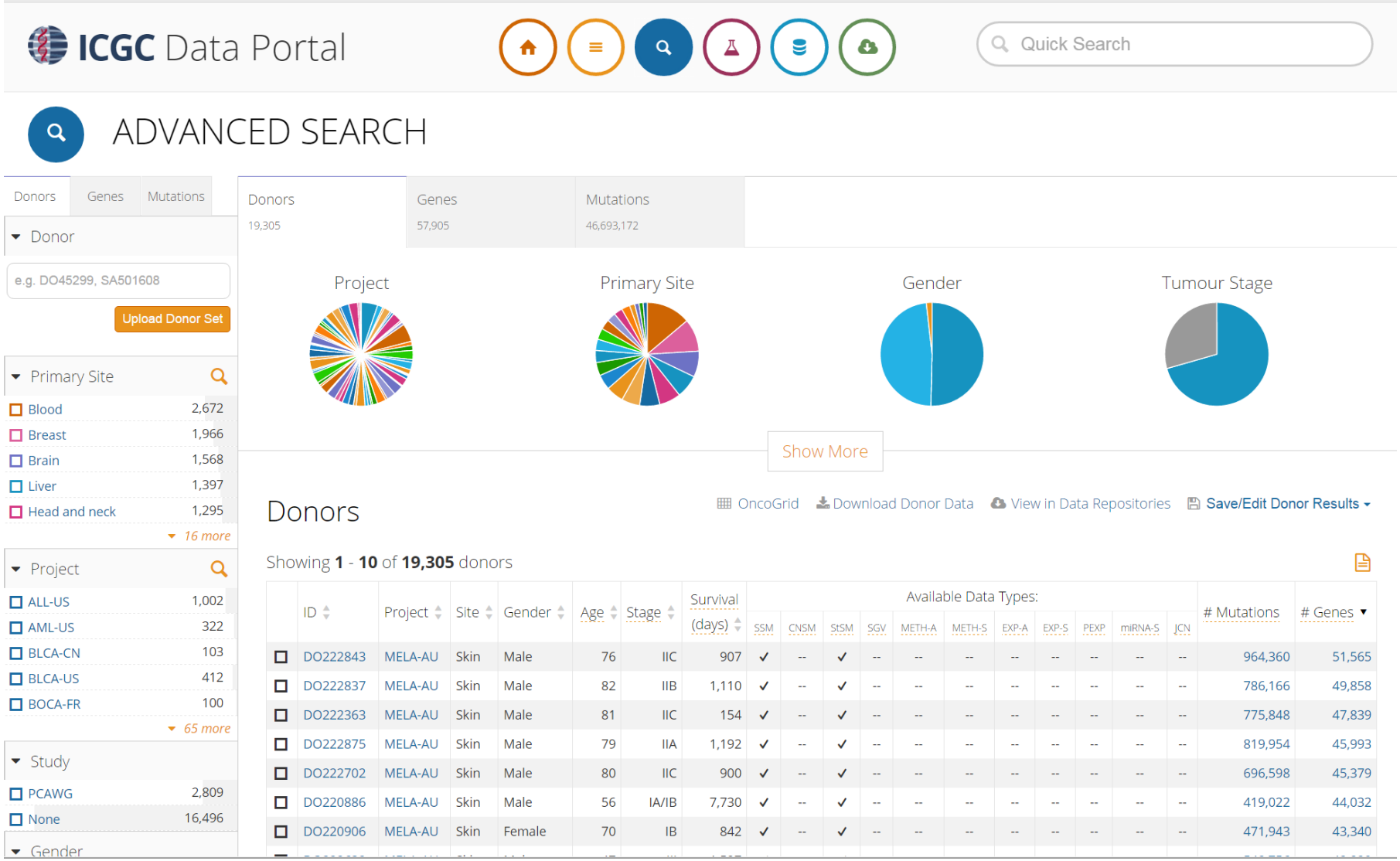
Variant ID	Location	Variant type	Gene	1000G MAF	GO-ESP MAF	ExAC MAF	Publications
nsv984836	61,735 - 248,930,189	copy number variation	AQP10 and 2				1
nsv436033	794,863 - 224,014,422	insertion	AQP10 and 2				1
nsv436505	795,801 - 224,010,171	copy number variation	AQP10 and 2				1
nsv2772868	914,087 - 248,930,485	copy number variation	AQP10 and 2				1
nsv1146931	6,798,591 - 214,701,964	inversion	AQP10 and 2				1
nsv1132997	16,516,919 - 234,817,464	inversion	AQP10 and 2182 more				1
nsv436816	16,594,222 - 206,307,186	insertion	AQP10 and 1926 more				1

Public Repositories for Genomic Data: ClinVar

- Genomic variations related with diabetes
- https://www.ncbi.nlm.nih.gov/variation/view/



Public Repositories for Genomic Data: ICGC



Public Repositories for Genomic Data: TCGA

001cef41-ff86-4d3f-a140-a647ac4b10a1

Add all files to the Cart

Summary

Case UUID	001cef41-ff86-4d3f-a140-a647ac4b10a1
Case Submitter ID	TCGA-E2-A1IU
Project ID	TCGA-BRCA
Project Name	Breast Invasive Carcinoma
Disease Type	Breast Invasive Carcinoma
Program	TCGA
Primary Site	Breast

FILES
32

ANNOTATIONS
0

File Counts by Experimental Strategy

Experimental Strategy	Files
Genotyping Array	4
Methylation Array	1
WXS	18
RNA-Seq	4
miRNA-Seq	3

File Counts by Data Category

Data Category	Files
Raw Sequencing Data	4
Transcriptome Profiling	5
Simple Nucleotide Variation	16
Copy Number Variation	4
Clinical	1
Biospecimen	1

Clinical

Export

Demographic | Diagnoses / Treatment (1) | Family Histories (0) | Exposures (1)

ID	d14426b2-e0a0-519a-bea6-4fe07d11ce95
Ethnicity	Not Hispanic Or Latino
Gender	Female

Public Repositories for Genomic Data:

Human Genome, Mutation, and Epigenome Databases

Database	Content	URL
HGMD	A database for germline mutations that are associated with heritable diseases	www.hgmd.org/
dbSNP	A catalog for genome variations	www.ncbi.nlm.nih.gov/projects/SNP/
TCGA	A cancer *omic* data resource containing genomic, epigenomic, and transcriptomic data sponsored by NIH	https://tcga-data.nci.nih.gov/tcga/
ICGC	A cancer *omic* data resource containing genomic, epigenomic and transcriptomic data sponsored by ICGC	http://icgc.org/
COSMIC	A catalog of somatic mutations in human cancers containing $>$ 50,000 mutations	http://www.sanger.ac.uk/perl/genetics/CGP/cosmic
Cancer gene census	A catalog of mutations in more than 400 cancer-related genes	www.sanger.ac.uk/genetics/CGP/Census/
CanProVar	A database for single amino-acid alterations including both germline and somatic variations	http://bioinfo.vanderbilt.edu/canprovar/
IARC TP53	A database for sequence-level variations in P53 identified in human population and tumor samples	http://p53.iarc.fr
CDKN2A	A database for variants of CDKN2A identified in human disease samples	https://biodesktop.uvm.edu/perl/p16
Androgen receptor gene mutations	A dataset of 374 mutations identified in patients with androgen insensitivity syndrome	http://androgendb.mcgill.ca
NIH roadmap epigenomics program	A database for human epigemomes now covering at least 23 cell types	http://www.roadmapepigenomics.org/data
Human epigenome project	A database for genome-wide DNA methylation patterns of all human genes in all major tissues	http://www.epigenome.org/
MethyCancer	A database for DNA methylation information in cancer-related genes, collected from public resource	http://methycancer.genomics.org.cn

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

RWTH AACHEN UNIVERSITY

Public Repositories for Genomic Data

- **Gene expression databases:**
  - Compared to other omics databases, there is a much larger collection of transcriptomic data on the Internet.
  - Two of the most popular ones are GEO (Gene Expression Omnibus) at the NCBI that has more than 32,000 sets of gene-expression data collected from 800,000 samples of 1,600 organisms and Arrayexpress at the EBI that consists of 1,245,005 sets of gene-expression data collected through 43,947 experiments using microarray and RNA sequencing.

Database	Content	URL
NCBI GEO	A comprehensive collection of gene expression data	http://www.ncbi.nlm.nih.gov/gds
Arrayexpress	A database of functional genomics including gene expression data in both microarray and RNA-seq forms	http://www.ebi.ac.uk/arrayexpress/
SMD	Stanford microarray database for gene expression data covering multiple organisms	http://smd.stanford.edu/
Oncomine (research edition)	A commercial database for cancer transcriptomic and genomic data, with a free edition to academic and nonprofit organizations	https://www.oncomine.org/resource/login.html
ASTD	A database for human gene-expression data and derived alternatively spliced isoforms of human genes	http://drcat.sourceforge.net/astd.html

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

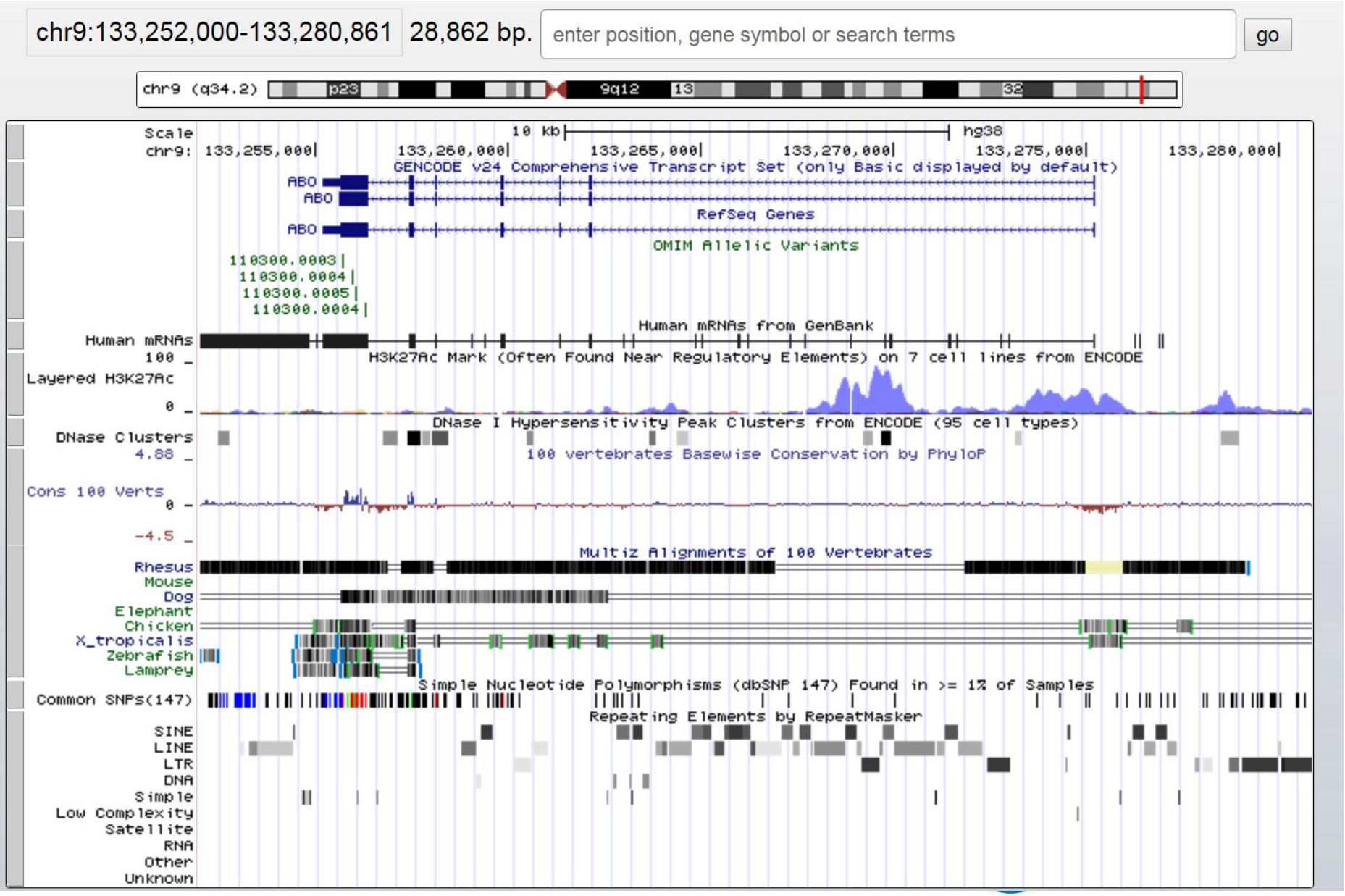
Public Repositories for Genomic Data

- **MicroRNAs and target databases:**
- Interactions of human mRNAs, microRNAs have roles in regulating many major cellular processes such as cell growth, differentiation, and apoptosis, as well as disease development
- Many earlier researches in this field are focused on microRNA identification and targets prediction.
- Major databases archiving validated microRNAs with sequence, structure, and interaction information : MiRecords and miRBase

Database	Content	URL
miRecords	A database for animal microRNA-target interactions	http://mirecords.biolead.org
miRBase	A database for published microRNA sequences and annotations covering numerous species	http://www.mirbase.org
TargetScan	A database for microRNA targets	http://www.targetscan.org
MiRanda	A databases for predicted microRNA targets	http://www.microrna.org/microrna/home.do
MirTarBase	A database for experimentally validated microRNA-target interactions	http://mirtarbase.mbc.nctu.edu.tw

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

Genome Browsing: UCSC Genome Browser

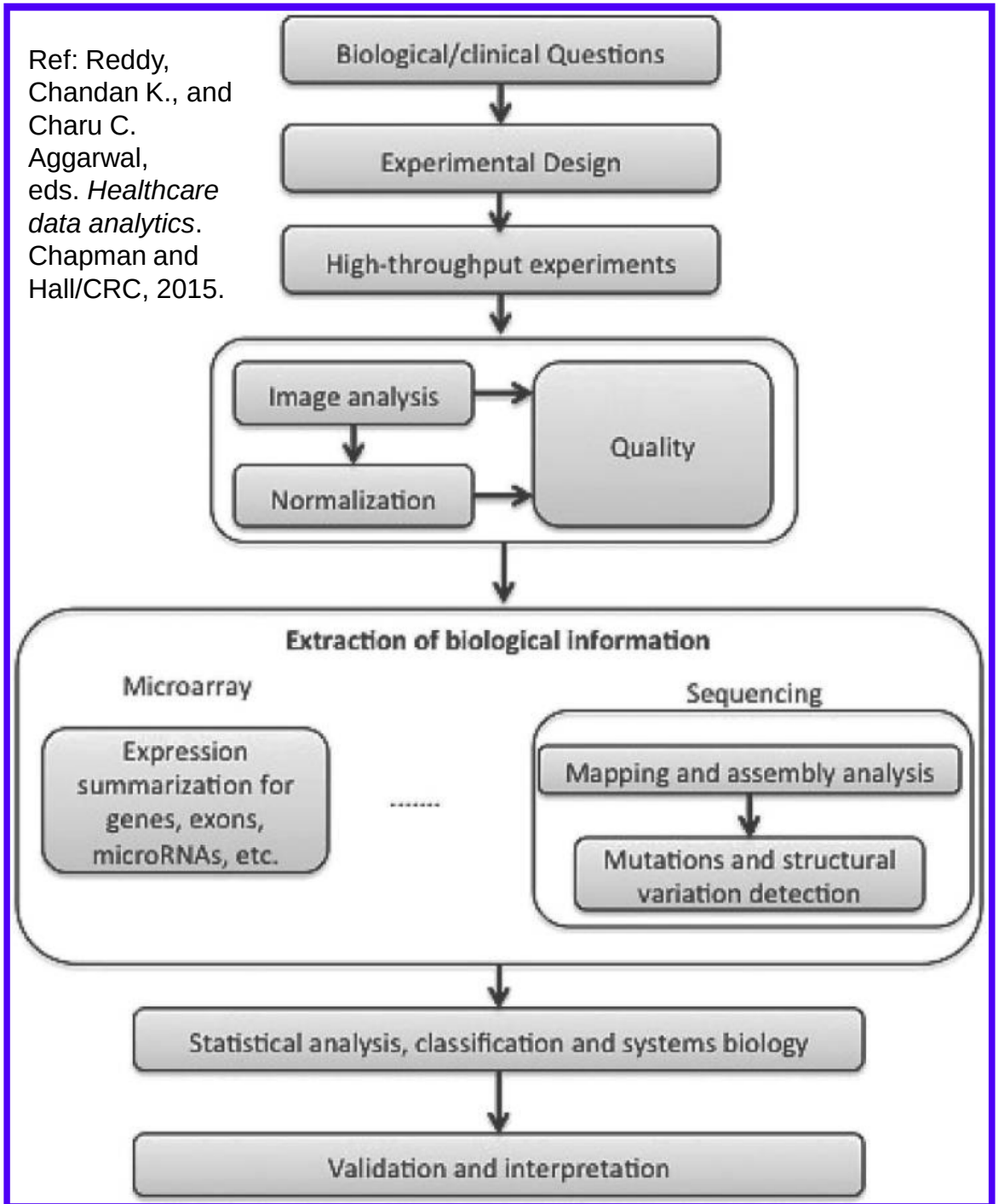


Methods for Genomic Data Analysis

- A large collection of different methods and algorithms have been developed for genomic data analysis, each serving a specific analytic step within the standard bioinformatics workflow

- They are generally categorized into three groups
  - data preprocess,
  - data analysis,
  - result interpretation

Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.



The standard bioinformatics workflow to analyze the genomic data

- Example:
- microarray, sequencing slides, or phenotyping screening will have to be analyzed through the scanner using appropriate algorithms to quantify the raw signal, followed by data normalization to improve the signal-to-noise ratio.
- The quality of the data is checked at the level of both the image analysis and the normalization steps.
- After the preprocess, meaningful biological information will be extracted from the data and then subjected to further analysis using clinical statistics, classification or the systems biology approach, followed by the validation and interpretation of the results.

Methods for Genomic Data Analysis: Clustering and Classification

- To identify meaningful expression patterns , **clustering** methods can be applied to identify if some genes shows correlated expression across the given set of biological groups or if some samples share similar gene expression profiles

- Like the clustering strategy for identifying gene expression patterns, **classification** methods can be used to identify gene signatures, which represent a set of genes that can differentiate different biological groups based on the gene expression.

Hierarchical clustering:

- produce a gene/condition tree where the most similar expression profiles are joined together
- Strategies generally fall into two types:
  - 1. agglomerative approach: where each observation (expression profile for one gene or one sample) starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy and
  - 2. divisive approach: where all observations start in one cluster and splits are performed recursively as one moves down the hierarchy.
- In general, the merge and splits are determined in a greedy manner.
- The measure of dissimilarity of observations can be calculated based on various distance functions including Euclidean distance, Manhattan distance, maximum distance, etc.
- Different strategies are used to calculate the distance between clusters including complete linkage, single linkage, average linkage, and centroid linkage.

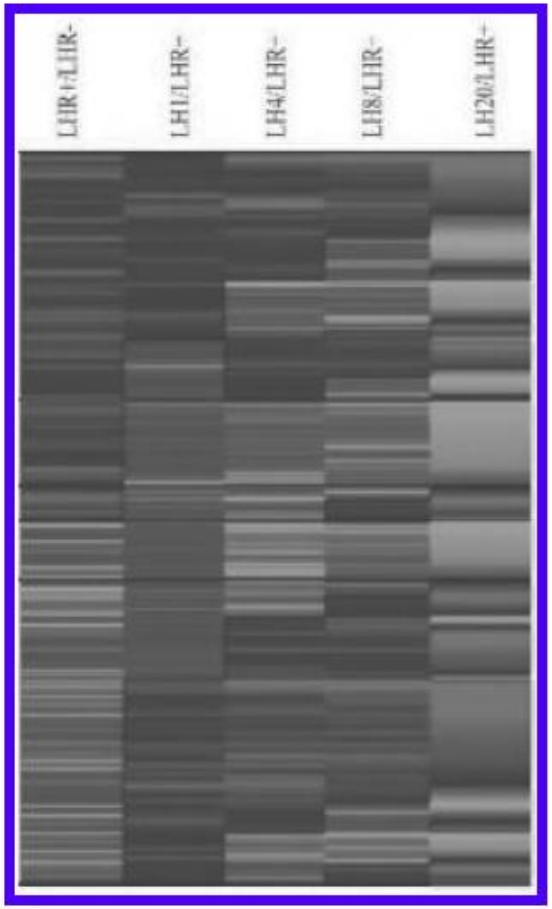
Methods for Genomic Data Analysis: Clustering and Classification

- K-mean clustering:
  - a representative partitioning method that needs to define k, the number of clusters in which to partition selected genes or conditions.
  - The algorithm attempts to minimize the mean-squared distance from each data point to its nearest center, the intracluster variability, and maximized intercluster variability.

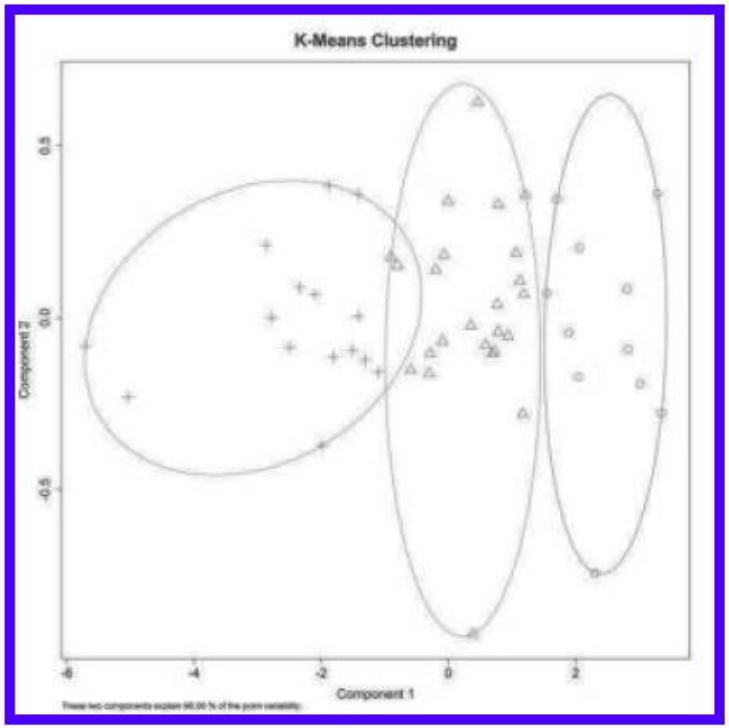
- SOM (Self-Organizing Map):
  - artificial neural network-based.
  - The goal is to find a set of centroids and to assign each object in the dataset to the centroid that provides the best approximation of that object,
  - produces information about the similarity between the clusters

Methods for Genomic Data Analysis:
Clustering and Classification

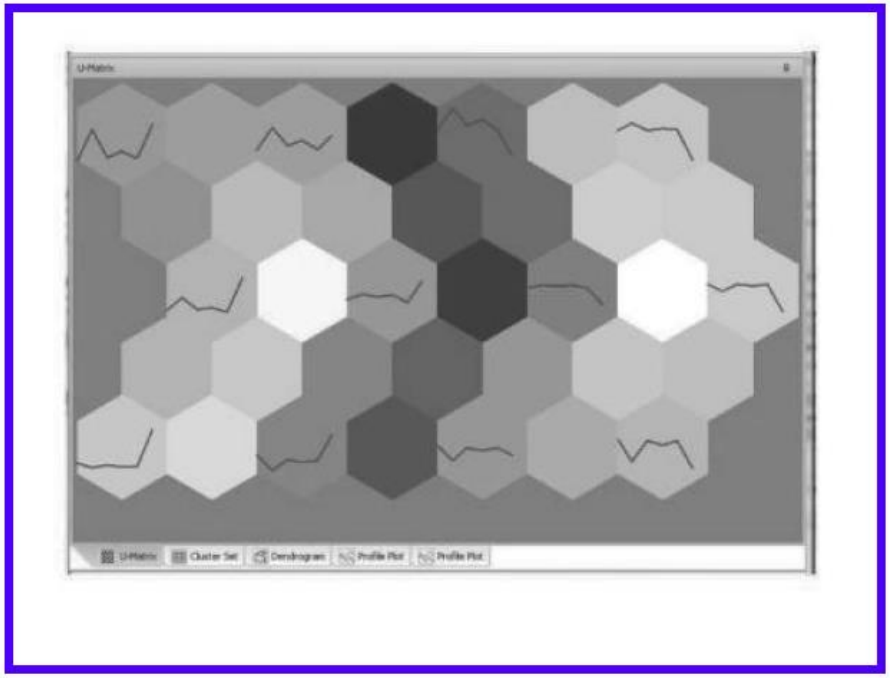
Results from three different clustering analyses including hierarchical clustering (a), K-mean clustering (b) SOM (c), based on the same data from



(a)



(b)



Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

Public Tools for Genomic Data Analysis

- A variety of computational analysis and data mining tools have been published and deployed on the Internet, which can be used to analyze the databases
- Genome analysis tools:
  - ABSOLUTE for computing absolute copy number and mutation multiplicities in the genomes,
  - MuTect for identifying point mutations,
  - Breakpointer for pinpointing the breakpoints of genomic rearrangements
  - dRanger for identifying genomic rearrangements,
  - Oncotator for annotations of the point mutations
  - INDELs in the sequenced genome
- Transcriptome analysis tools:
  - edgeR and baySeq: identification of differentially expressed genes in cancer versus matching control tissues
  - WGCNA and GeneCAT: identification of co-expressed genes or genes with correlated expression patterns
  - CUFFLINK: inference of splicing variants from RNA-seq data

TABLE 3.4: Tools for Genomics Data Analysis and Pathway Prediction and Mapping

Database	Content	URL
edgeR	A tool for detection of differentially expressed genes	http://www.genomine.org/edge/
WGCNA	A tool for co-expression analysis of genes	http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork
CUFFLINK	A tool for transcript assembly and identification of splicing variants	http://cufflinks.cbcb.umd.edu/index.html
DAVID	A tool for pathways enriched with differentially expressed genes (or any specified set of genes)	http://david.abcc.ncifcrf.gov/
CHARM	An early and widely used package for DNA methylation analysis.	http://www.bioconductor.org/packages/release/bioc/html/charm.html
EpiExplorer	A web-based tool for identification of comparing epigenetic markers in a specific genome to reference human epigenomes	http://epiexplorer.mpi-inf.mpg.de/
Pathway tools	A website providing a wide ranges of pathway-related tools, including pathway construction, editing, prediction, and flux analysis.	http://bioinformatics.ai.sri.com/ptools/
BioCyc and pathway tools	A database providing a list of reconstruction and analysis tools of metabolic pathways	http://biocyc.org/publications.shtml
PathoLogic pathway prediction	A tool for prediction of metabolic pathways based on BioCyc database	http://g6g-softwaredirectory.com/bio/cross-omics/pathway-dbs-kbs/20235SRIPathoLogicPathwPredict.php
Metabolic pathways	A website providing a large collection of pathway-related tools	http://www.hsls.pitt.edu/obrc/index.php?page=metabolic_pathway

- Ref: Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.

Public Tools for Genomic Data Analysis

- Statistical analysis / Data analytic tools:
- In addition to the above data type-specific tools, there are large collections of other data analysis tools on the Internet for boarder uses of analyzing different omic data types.
- Bioconductor
  - It is a communitywide effort for developing and deploying open source bioinformatics software packages.
  - All the deployed tools are written in the statistical programming language R. Currently the website has about 750 software tools, covering a wide range of analysis and inference capabilities.
- Galaxy
  - Open, web-based platform that hosts a large collection of genomic data analysis tools
  - a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences
  - Implement workflows

References

- Chapter 6- Genomic Data Analysis for Personalized Medicine Juan Cui
  Reddy, Chandan K., and Charu C. Aggarwal, eds. *Healthcare data analytics*. Chapman and Hall/CRC, 2015.