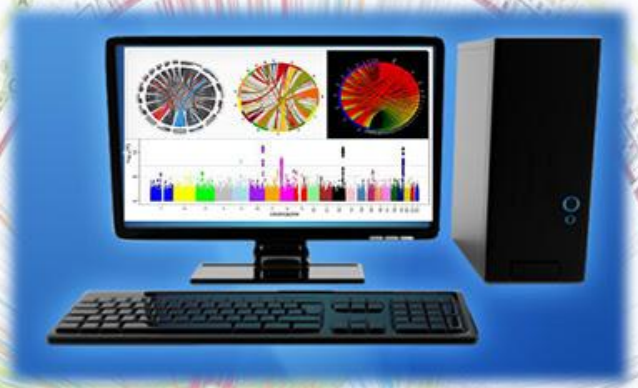
İTF Tıbbi Biyoloji AD

II.Uygulamalı Biyoinformatik Kursu



17-18 Nisan 2017

Mehmet Tevfik DORAK, MD PhD
Liverpool Hope University
U.K.

YOUR FUTURE
STARTS WITH HOPE

Liverpool Hope University EST. 1844

İTF Tıbbi Biyoloji AD

II.Uygulamalı Biyoinformatik Kursu

17-18 Nisan 2017

Prof. Dr. Mehmet Tevfik Dorak

Liverpool Hope University, UK

1. Gün (17/4/2017):	
09:30-10:25	2017 yılında uygulamalı biyoinformatik (son gelişmeler)
10:30-11:25	Epigenetik (metilasyon, histon, miRNA/lncRNA)
11:30-12:25	Sekans datasında rastlanan yeni mutasyonlarda patojenite değerlendirmesi
12:30-13:25	Öğlen yemeği
13:30-14:25	Galaxy'ye giriş
14:30-14:55	Araştırmalar için masif data kaynakları
15:00-16:30	Uygulama: Katılımcı projeleri
2. Gün (18/4/2017):	
09:30-12:25	İmmünolojik araştırmalarda biyoinformatik
12:30-13:25	Öğlen yemeği
13:30-14:55	Kanser araştırmalarında biyoinformatik
15:00-16:30	Uygulama: Katılımcı projeleri ve Soru/Cevap
Hazırlık:	
Genom biyolojisine genel giriş ve temel biyoinformatik uygulamaları için,	
bkz: I.Uygulamalı Biyoinformatik Kursu (2016) (http://www.dorak.info/genbiol)	

Düzenleme Kurulu
Prof.Dr. Fatma S. OĞUZ
Doç.Dr. Çiğdem KEKİK ÇINAR
PhD. Hayriye ÇİFTÇİ
Msc. Sebahat USTA AKGÜL
e-posta: tibbibiyolojiad@gmail.com
Tel: 212/6351168

Yer: İstanbul Tıp Fakültesi,
Hulusi Behçet Kitaplığı, Pembe Salon

Uygulamalı Kurs ücreti: 125TL

Day 1
Genome Biology in 2017
Bioinformatics Tools in Epigenetics
Pathogenicity Assessment of DNA Sequence Mutations
Introduction to Galaxy
Massive Data Sources
Practical

Day 2
Bioinformatic Tools for Immunologic Research
Bioinformatic Tools for Cancer Research
Practical

Aim

Immunology Taught by Viruses

Rolf M. Zinkernagel

Host genetics and viral infections: immunology taught by viruses, virology taught by the immune system

David Nolan, Silvana Gaudieri and Simon Mallal

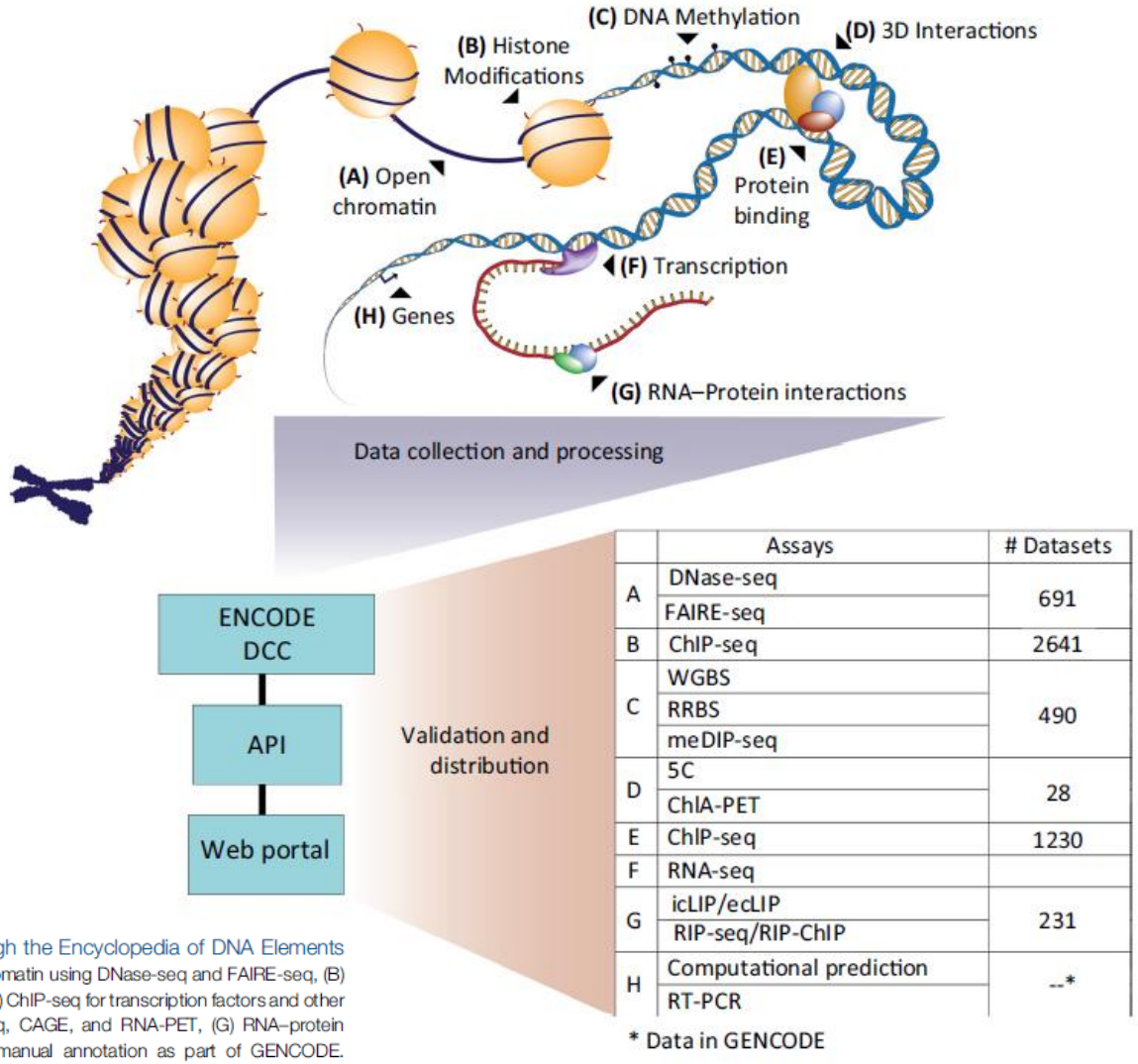
“Genome Biology Taught by Genetic Associations”

Genome Biology 101

Review

Deciphering ENCODE

Adam G. Diehl[1] and Alan P. Boyle[1,2,*]



	Assays	# Datasets
A	DNase-seq	691
	FAIRE-seq	
B	ChIP-seq	2641
C	WGBS	490
	RRBS	
	meDIP-seq	
D	5C	28
	ChIA-PET	
E	ChIP-seq	1230
F	RNA-seq	
G	icLIP/ecLIP	231
	RIP-seq/RIP-ChIP	
H	Computational prediction	--*
	RT-PCR	

* Data in GENCODE

Figure 1. An Illustrative Example of the Data Types Available through the Encyclopedia of DNA Elements (ENCODE) Project Portals. These include measurements of (A) Open chromatin using DNase-seq and FAIRE-seq, (B) ChIP-seq for histone modifications, (C) DNA methylation, (D) 3D interactions, (E) ChIP-seq for transcription factors and other chromatin-associated proteins, (F) transcriptional output including RNA-seq, CAGE, and RNA-PET, (G) RNA–protein interactions, and (H) gene body predictions through computational and manual annotation as part of GENCODE.

Trends in Genetics

Genome Biology 101

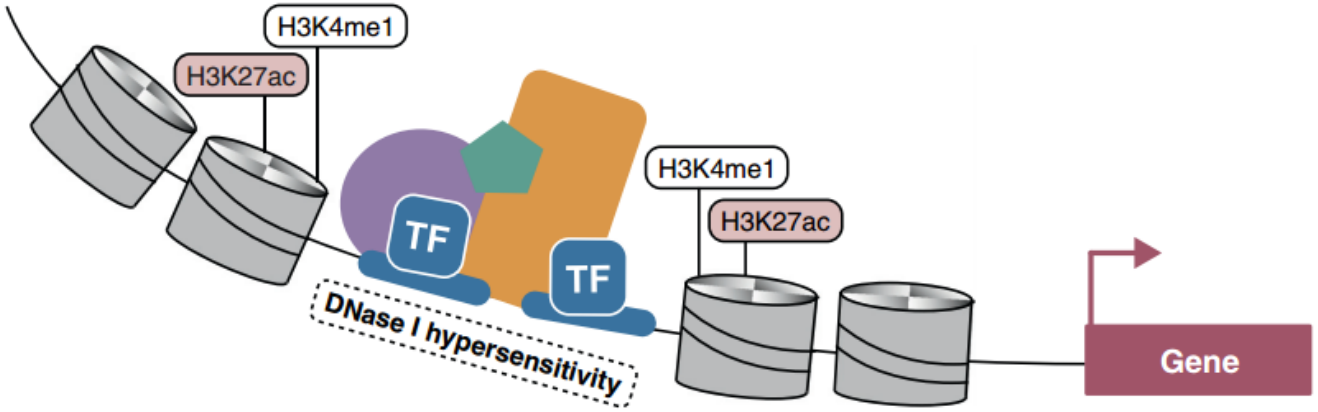


Figure 1 Model of enhancer function. Transcriptional enhancer elements are noncoding stretches of DNA that regulate gene expression levels, most often in *cis*. Active enhancer elements are located in open chromatin sensitive to DNase I digestion and flanked by histones marked with H3K4me1 and H3K27ac. Enhancers are often bound by a number of transcription factors (TF), such as p300 (blue). Mediator and cohesin are part of a complex (orange, green and purple) that mediates physical contacts between enhancers and their target promoters.

Ritchie and Flicek *Genome Medicine* 2014, **6**:87
http://genomemedicine.com/content/6/10/87

Genome Medicine

REVIEW

Computational approaches to interpreting genomic sequence variation

Graham RS Ritchie[1,2] and Paul Flicek[1,2*]

Human Genome (hg) Assembly Version

UCSC VERSION	RELEASE DATE	RELEASE NAME
hg38	Dec. 2013	Genome Reference Consortium GRCh38
hg19	Feb. 2009	Genome Reference Consortium GRCh37
hg18	Mar. 2006	NCBI Build 36.1
hg17	May 2004	NCBI Build 35
hg16	Jul. 2003	NCBI Build 34
hg15	Apr. 2003	NCBI Build 33
hg13	Nov. 2002	NCBI Build 31
hg12	Jun. 2002	NCBI Build 30
hg11	Apr. 2002	NCBI Build 29
hg10	Dec. 2001	NCBI Build 28
hg8	Aug. 2001	UCSC-assembled
hg7	Apr. 2001	UCSC-assembled
hg6	Dec. 2000	UCSC-assembled
hg5	Oct. 2000	UCSC-assembled
hg4	Sep. 2000	UCSC-assembled
hg3	Jul. 2000	UCSC-assembled
hg2	Jun. 2000	UCSC-assembled
hg1	May 2000	UCSC-assembled

Transitioning to GRCh38? Try the NCBI Remapping Service, which uses the same assembly-assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p11) will be a minor (patch) release in summer 2017.

File Formats in Bioinformatics

OEGE - Online Encyclopedia for Genetic Epidemiology studies

- Overview
- GWAS
- NGS
- Population Studies
- Case & Family Studies
- Genetic Databases
- Genotyping
- Phenotyping
- Software
- Disease Genes
- Genetic Diversity
- Journals

Next generation sequencing

NGS platforms

- Illumina® (Solexa) Genome Analyzer™ and HiSeq
- Roche 454 Sequencer™
- Applied Biosystems SOLiD™

File formats

- **FASTQ:** derived from FASTA format with the addition of quality scores. Each read from a sequencer comprises an identifier line, a sequence line, a second identifier line (or with a + character) and final a quality line. This typically forms the input to a mapping program (along with a FASTA reference genome). A typical human exome FASTQ file might be around 10-15GB, which can be compressed to 5-6GB using gzip.
- **SAM format:** mapped/aligned sequence containing detail about alignment, mapping quality etc. This usually contains a subset of the raw reads (as some will have been discarded at the mapping stage). The SAM (or BAM) file is typically used as the substrate for variant calling algorithms and other analyses.
- **BAM format:** binary version of SAM. A typical human exome BAM file might be around 2GB in size.
- **BED format:** annotation format that describes genome regions, with the optional addition of annotation data for display of genome browser tracks.
- **Other annotation formats:** UCSC describe a number of other formats suitable for generating tracks in genome browsers.
- **VCF:**variant call format - this contains details about the number of reads at variant sites in the genome, plus a range of quality information. A typical human exome VCF file might contain about 20,000 lines.

ENCODE at UCSC (2003-2012) - **Downloads**

Specifications of Common File Formats Used by the ENCODE Consortium

September 2013

The ENCODE consortium uses several file formats to store, display, and disseminate data:

- FASTQ
- BAM
- bigWig
- bigBed

FASTQ[1] is a text-based format for storing nucleotide sequences (reads) and their quality scores. The Sequence Alignment/Mapping (SAM)[2] format is a text-based format for storing read alignments against reference sequences and it is interconvertible with the binary BAM format. The bigWig format is an indexed binary format for rapid display of continuous and dense data in the UCSC Genome Browser. And the bigBed format is also an indexed binary format for rapid display of annotation items such as a linked collection of exons or the binding peaks of a transcription factor.

These file formats were originally designed to be generic and flexible. As the ENCODE consortium is a collaborative effort, the consortium has made several specifications on the file formats to facilitate data archival, presentation, and distribution, as well as integrative analysis on the data. The consortium considers FASTQ as the basic file format for archival purpose and thus the FASTQ format's specifications aim to preserve the raw sequence data. In comparison, the other file formats are geared towards data visualization and dissemination, thus their specifications aim to facilitate user-friendliness.

UCSC Genome Browser ENCODE-specific File Formats
References

Updated 4 Dec 2013

Liverpool Hope University EST.1844

File Formats in Bioinformatics

Galaxy Data Formats

Dataset missing?

If you have a dataset in your history that is not appearing in the drop-down selector for a tool, the most common reason is that it has the wrong format. Each Galaxy dataset has an associated file format recorded in its metadata, and tools will only list datasets from your history that have a format compatible with that particular tool. Of course some of these datasets might not actually contain relevant data, or even the correct columns needed by the tool, but filtering by format at least makes the list to select from a bit shorter.

Some of the formats are defined hierarchically, going from very general ones like Tabular (which includes any text file with tab-separated columns), to more restrictive sub-formats like Interval (where three of the columns must be the chromosome, start position, and end position), and on to even more specific ones such as BED that have additional requirements. So for example if a tool's required input format is Tabular, then all of your history items whose format is recorded as Tabular will be listed, along with those in all sub-formats that also qualify as Tabular (Interval, BED, GFF, etc.).

There are two usual methods for changing a dataset's format in Galaxy: if the file contents are already in the required format but the metadata is wrong (perhaps because the Auto-detect feature of the Upload File tool guessed it incorrectly), you can fix the metadata manually by clicking on the pencil icon beside that dataset in your history. Or, if the file contents really are in a different format, Galaxy provides a number of format conversion tools (e.g. in the Text Manipulation and Convert Formats categories). For instance, if the tool you want to run requires Tabular but your columns are delimited by spaces or commas, you can use the "Convert delimiters to TAB" tool under Text Manipulation to reformat your data. However if your files are in a completely unsupported format, then you need to convert them yourself before uploading.

Format Descriptions

- AB1
- AXT
- BAM
- BED
- BedGraph
- Binseq.zip
- FASTA
- FastqSolexa
- FPED
- gd_indivs
- gd_ped
- gd_sap
- gd_snp
- GFF
- GFF3
- GTF
- HTML
- Interval
- LAV
- LPED
- MAF
- MasterVar
- PBED
- pgSnp
- PSL
- SCF
- SFF
- Table
- Tabular
- Txtseq.zip
- VCF

All of these are flat text files!
Yo can open them on NotePad or other text editors.
They can be exported to MS Excel.

VCF

Variant Call Format (VCF) is a tab delimited text file with specified fields. It was developed by the 1000 Genomes Project. Field specifications.

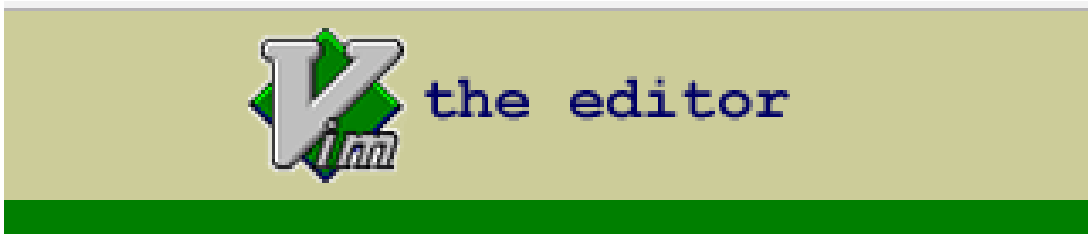
Can be converted to:

- pgSnp
  Convert Formats → VCF to pgSnp

Liverpool Hope University EST. 1844

Handling Large Files

If you do not use Unix, Python or R, you will find that MS Excel can still handle a lot, but no more than 1,000,000 rows in each worksheet.
You can ask for help from people who can handle large datasets and extract data for you even without opening them.
If you want to do it yourself, try the following freeware.



Bioinformatics Links

HLA MHC Genetics Evolution Biostatistics Population Genetics Genetic Epidemiology Homepage

BIOINFORMATICS TOOLS

(for Genetic Epidemiologists)

Mehmet Tevfik DORAK, M.D., Ph.D.

Open Access Reviews

ENCODE Resources (Pazin MJ. CSHP 2015); Database Tools (Bianco, Genomics 2013; Zou, 2015); UCSC Browser (Mangan, 2014)

Journals: Nucleic Acids Research: Database Issue (2017) & Web Server Issue (2016); Database: Biomart Issue & Ensembl Issue

Nucleic Acids Research Databases Catalog

Societies

ISCB ISMB AMIA

Online Suites

Galaxy (101 - Support) ExPASy (ref) Taverna 2.5.0/Archive (ref) Apache-Taverna BioMart (ref) Artemis (manual) Babelomics (ref / tutorial)

Bioconductor (manual) BioPython (tutorial & cookbook / PDF)

EDGE: Empowering the Development of Genomics Expertise (ref)

Biomarts

Ensembl BioMart (YouTube tutorial) UCSC Table Browser GWAS Central GWAS Mart (tutorial) SPS Mart (allele freqs) Fantom5

Genome Browsers

UCSC Human Genome BioInformatics

Ensembl Genome Browser (Human) Vega Genome Browser (Human)

NIH RoadMap Epigenomic Browser (tutorial) WashU EpiGenome Human Epigenome Atlas NCBI Epigenomics (help) NGSmethDB (ref)

1KG (Browser - Quick Start Guide)-(NCBI) UK10K (ref) GENCODE Browser SNiPA

CNV Browser DECIPHER eQTL Browser seeQTL GTEx eQTL Browser

StarBase (tutorial / ref) IncRNome IncRBase NonCode Integrative Genomics

VISTA Enhancer Browser Swiss Regulon Reactome Pathway Browser ImmunoBase Browser (HaemAtlas) T1Dbase

General Resources

NCBI Global Search GoPubMed HuGE LabMeeting NCBI Tutorials / Handbook / Help Manual NIH Helix NCBI Tools for Data Mining

Sanger: Software / Databases EMBL-EBI Swiss Institute of Bioinformatics: ExPASy

NCBI Gene .. GeneCards .. GeneMap .. GenAtlas .. GenCanVas .. iHOP .. NetAffX .. BioGPS Gene Annotation

NCBI (tutorials on YouTube) .. BLAST .. MimicMe (ref) .. e-PCR .. In Silico PCR .. GeneNote .. GeneLoc .. CGAP-GAI .. Aceview

ENCODE (ref) GENCODE RegulomeDB FANTOM/Gateway (papers / software)

… Looking forward

Day 1
Genome Biology in 2017
Bioinformatics Tools in Epigenetics
Pathogenicity Assessment of DNA Sequence Mutations
Introduction to Galaxy
Massive Data Sources
Practical