Introduction to Biological Annotation Resources

Shamith Samarajiwa

Integrative Systems Biomedicine Group

MRC Cancer Unit

University of Cambridge

Analysis of high-throughput sequencing data with R/BioConductor

June 2015

Email: ss861@mrc-cu.cam.ac.uk

MRC | Cancer Unit

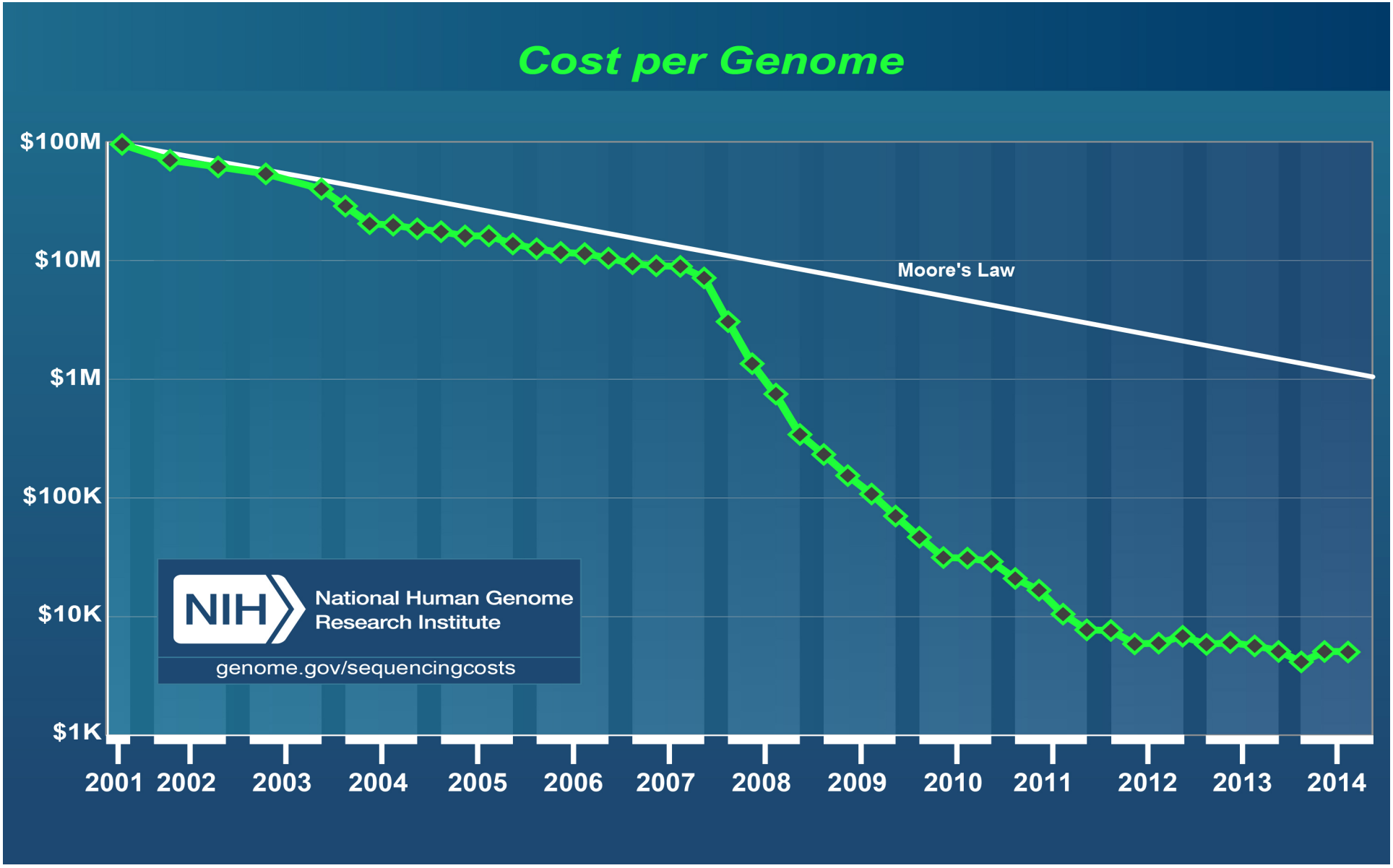
UNIVERSITY OF CAMBRIDGE

The $1000 Genome

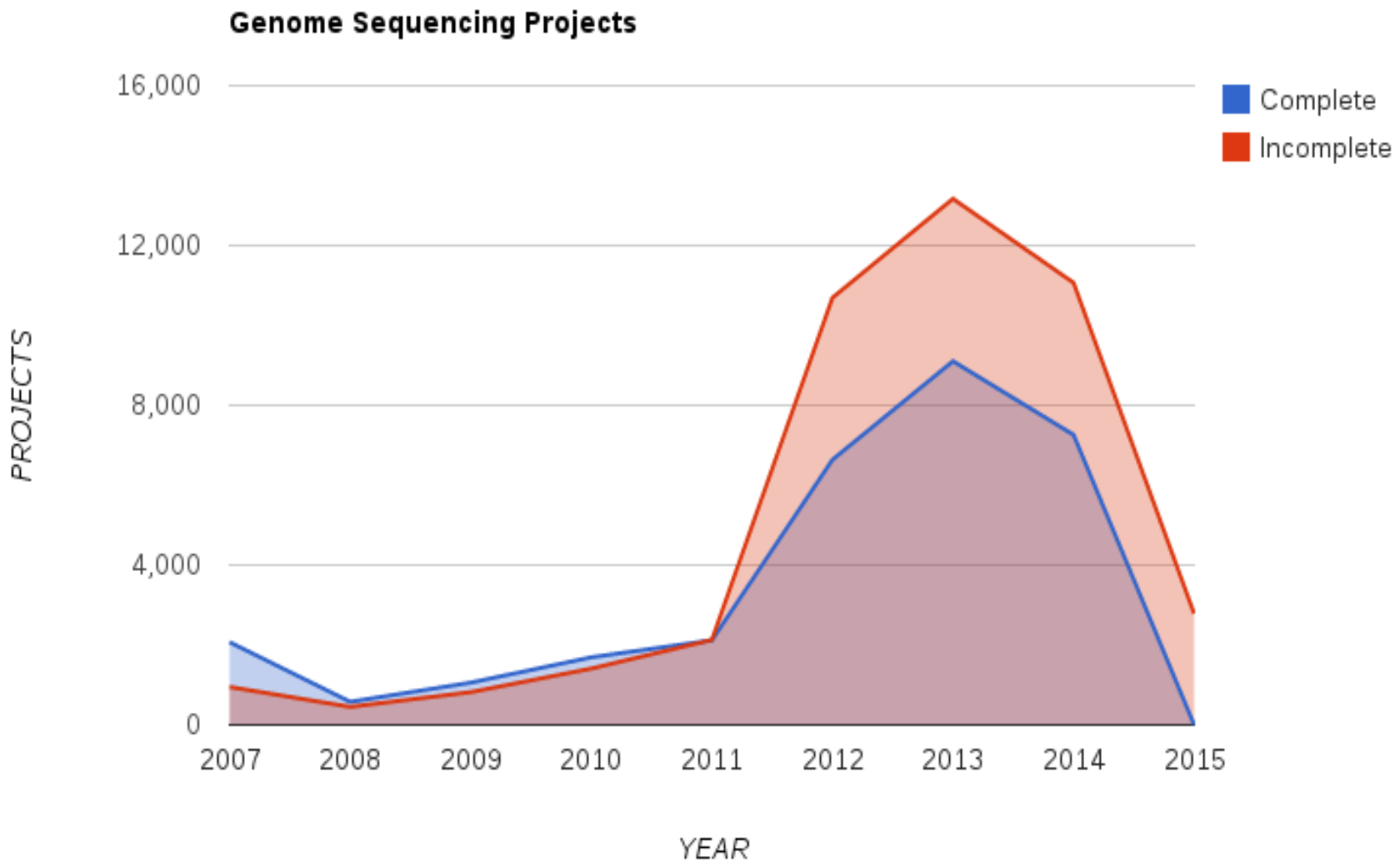
- Illumina HiSeq X Ten- (18,000 human genomes per year at the cost of $1000 per genome).
- Other Platforms:
- Illumina HiSeq 2500, MiSeq, NextSeq
- PacBio RSII, Oxford Nanopore MinION

Cost of sequencing

Moore's law : the simplified version of this law states that processor speeds, or overall processing power for computers will double every two years. http://genome.gov/sequencingcosts/

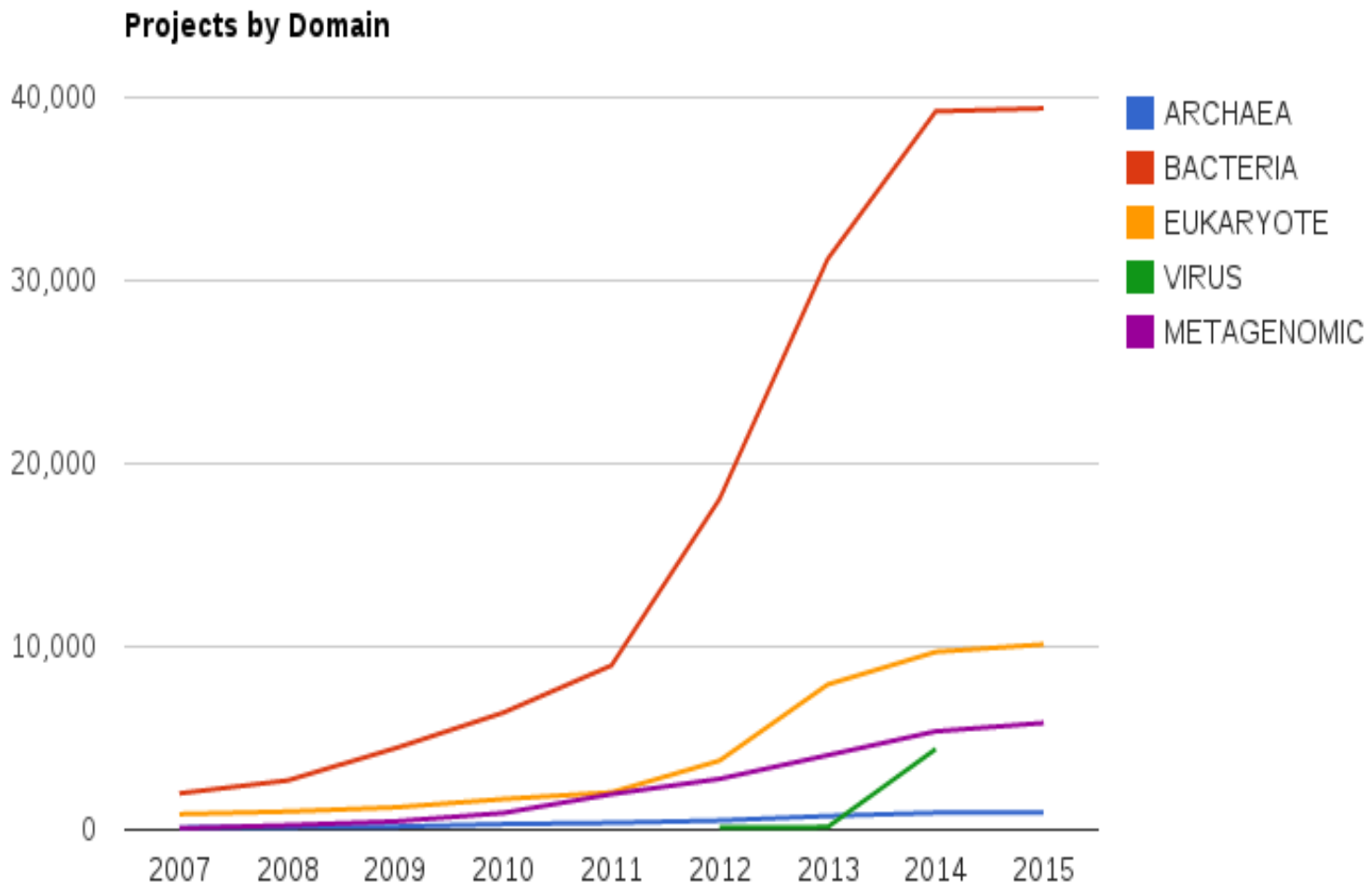


Genome sequencing projects



Genome sequencing projects

Genome sequencing projects



Why Annotate?

- Genome annotation is the process of attaching biological information to sequences or genomic regions. **Annotation provides biological context to data.**
- There are a multitude of annotation resources that are gene or genome centric.
- These are either web based or accessed via API's programmatically.
- Annotating datasets is a critical step in most bioinformatics workflows.
- **Bioconductor** provides and extensive number of resources for gene/genome annotation.

Types of Annotation

- Genome builds
- Organisms
- Genome and Epigenome centric
- Gene, Transcriptomic or Protein centric
- Regulation and Systems Biology
- **Bioconductor** contains more than a 1000 annotation related packages.

File Formats

- FASTA and FASTQ
- SAM (sequence alignment map) is a generic format for storing large nucleotide sequence alignments. (Li et al., 2009)
- BAM is a compressed binary version of SAM
- WIG (wiggle) is for display of continuous value information and is composed of declaration lines and data lines. There are two options for formatting wiggle data; **variableStep** and **fixedStep.**
- bigWig is for display of dense, continuous data that will be displayed in the Genome Browser as a graph.
- VCF (Variant Call Format) is a flexible and extendable format for variation data such as single nucleotide variants, insertions/deletions, copy number variants and structural variants.
- BED is used for chromosomal interval information and consists of **chromosome start, end** and other optional columns.
- bigBed and bedGraph
- GFF and GTF
- More information at https://genome.ucsc.edu/FAQ/FAQformat.html

External Annotation Resources

- Genome Reference Consortium (Genome builds)
- UCSC, Ensembl, NCBI, 1000 Genomes (Genome)
- GMOD, MGI, RGD, FlyBase (Genome)
- HGNC & EntrezGene (Gene specific)
- Uniprot & PFAM (Proteins)
- Biomart (Annotation from 44 databases)
- ENCODE & Ensembl (Gene Regulation)
- WashU, VizHub (Epigenomic)
- GO, NIH DAVID (SysBiol)
- KEGG, Reactome, PathwayCommons, BioModels, AnnotationHub (SysBiol)

Genome builds: Genome Reference Consortium

- http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
- The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome.
- Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis. GRC routinely releases patches and corrections.

Human:

GRCh38 = hg38 (2014)

GRCh37 = hg19 (2009)

Mouse:

GRCm38 = mm10

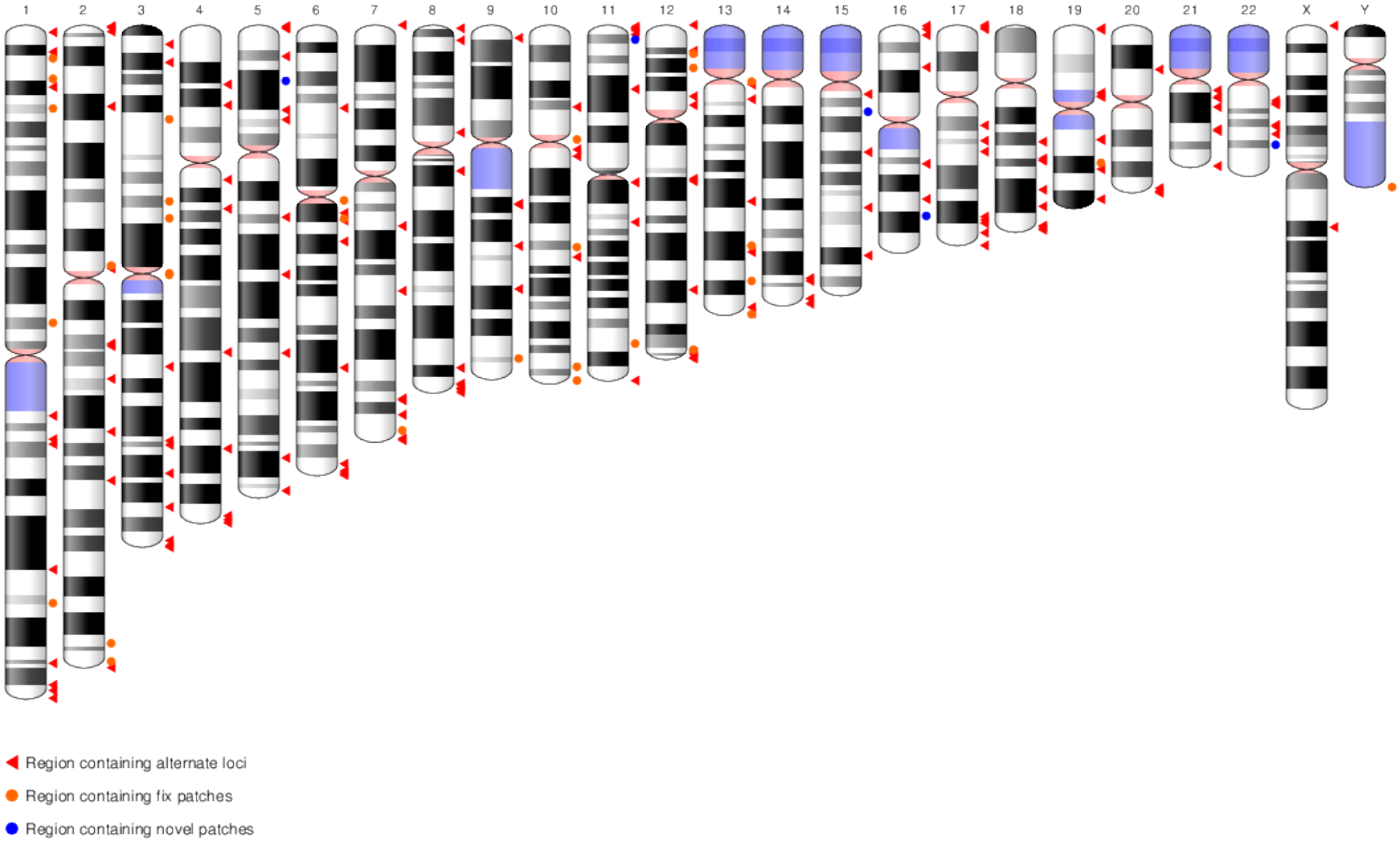
The Genome Reference Consortium consists of:

wellcome trust sanger institute — The Wellcome Trust Sanger Institute

The Genome Institute at Washington University — The Genome Institute at Washington University

EMBL-EBI — The European Bioinformatics Institute

NCBI — The National Center for Biotechnology Information


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Region containing alternate loci
Region containing fix patches
Region containing novel patches

Genome annotation: 1000 Genomes References

The use of **decoy** sequences:

- The reference human genome is incomplete, particularly around the centromeres and telomeres.
- Often reads which truly belong elsewhere are wrongly mapped to a specific place in the genome because the true match is missing from the reference.
- These cause false positive calls. The decoy sequences are a pragmatic solution to this and contain known true human genome sequence that is not in the reference genome, and will “collect” reads that would otherwise map with low quality in the reference.
- The **hs1k** contains: the 22 autosomes, X and Y chr, unlocalized and unplaced contig sequences and patches from GRCh37 or GRCh38, human herpesvirus 4 type 1 (NC_007605), concatenated decoy sequence derived from human, Human Bac and Fosmid clones NA12878 and the rCRS mitochondrial sequence (NC_012920).

Genome Annotation: Genomes

- **Bsgenome** provide Biostrings container based reference genome information.
- Bsgenome.Organism.Provider.Build Version.mask
- 76 pre-built genomes
- BSgenome.Hsapiens.UCSC.hg19 based on the *H. sapiens* UCSC hg19 build.
- Similar packages for other organisms.
- Used in analysis pipelines of other Bioconductor packages.
- Repeat masking, motif finding, SNPs etc.
- bsapply( ) - Apply function to each chromosome

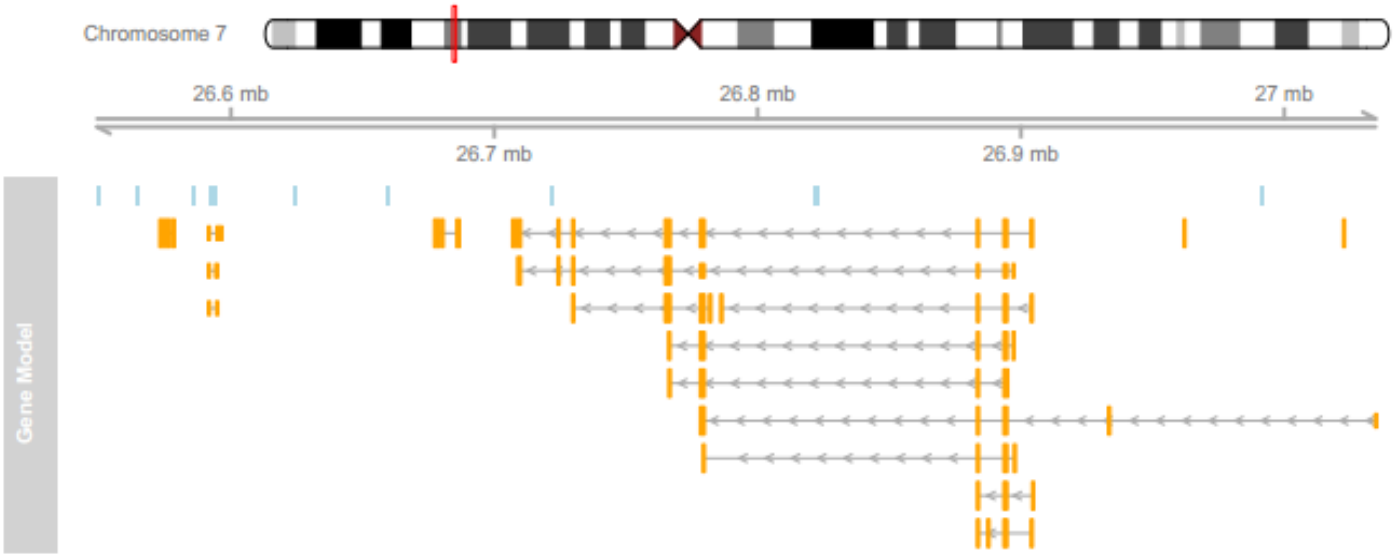
Genome annotation: UCSC genomes

- http://genome.ucsc.edu
- Provides reference sequences, draft assemblies and annotation for a large number of genomes.
- Multiple interfaces: Genome Browser, Table browser, MySQL server.
- The Table browser is useful for extracting annotation, and data can be exported to **GREAT**, **Galaxy** or **GenomeSpace** pipelines. More complex queries can be run via the MySQL server.
- **GenomicFeatures:** Retrieves transcript related features from UCSC and Biomart. Extracts genomic locations of the transcripts, exons and CDS of a given organism and stored as TxDb objects. Access to both UCSC and Biomart databases.
- **Rtracklayer:** Bioconductor package for visualizing browser tracks and an interface between R and genome browsers.

Genome annotation: Ensembl

- http://www.ensembl.org

- Web genome browser, Biomart interface and Perl API
- **Gviz**: Genomic data analyses requires integrated visualization of known genomic information and new experimental data. Gviz uses **biomaRt** package to perform live annotation queries to Ensembl and visualizes genes and transcripts.
- Supersedes GenomeGraphs package.



EnsDb and ensembldb

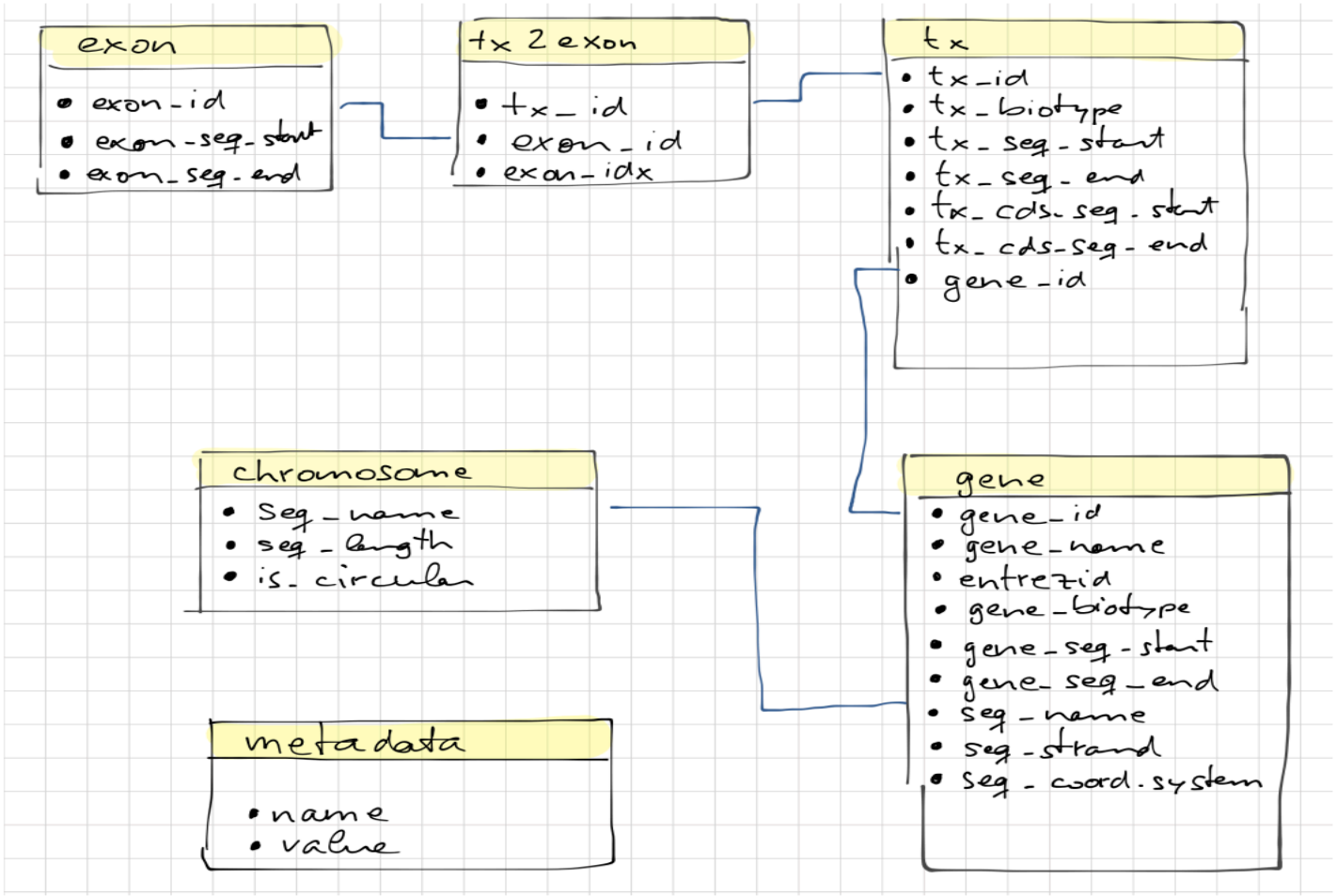


Figure 1: Database layout.

Gene annotation: HGNC & EntrezGene

- http://www.genenames.org
- The HUGO Gene Nomenclature Committee (HGNC) has assigned unique gene symbols and names to over 39,000 human loci, of which around 19,000 are protein coding genes.
- Useful for gene name disambiguation.

- http://www.ncbi.nlm.nih.gov/gene
- Supplies gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information.
- Entreze-Utils: programmatic access to EntrezGene
- **org.Hs.eg.db:** provides gene identifier mapping.

Orgdb example:

```
{
library(org.Hs.eg.db)
keys <- head(keys(org.Hs.eg.db), n=2)
cols <- c("PFAM","GO", "SYMBOL")
select(org.Hs.eg.db, keys, cols, keytype="ENTREZID")
}
```

Protein annotation: Uniprot & PFAM

- www.uniprot.org
- **Uniprot** provides comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
- Protein-family specific annotation at **PFAM** database:
- http://pfam.sanger.ac.uk/
- The **UniProt.ws** package provides a select interface to the UniProt web service. **PFAM.db** provides annotation from PFAM.

UniProtKB	Protein knowledgebase, consists of two sections: Swiss-Prot, which is manually annotated and reviewed. TrEMBL, which is automatically annotated and is **not** reviewed. Includes complete and reference proteome sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Biomart

- http://www.biomart.org

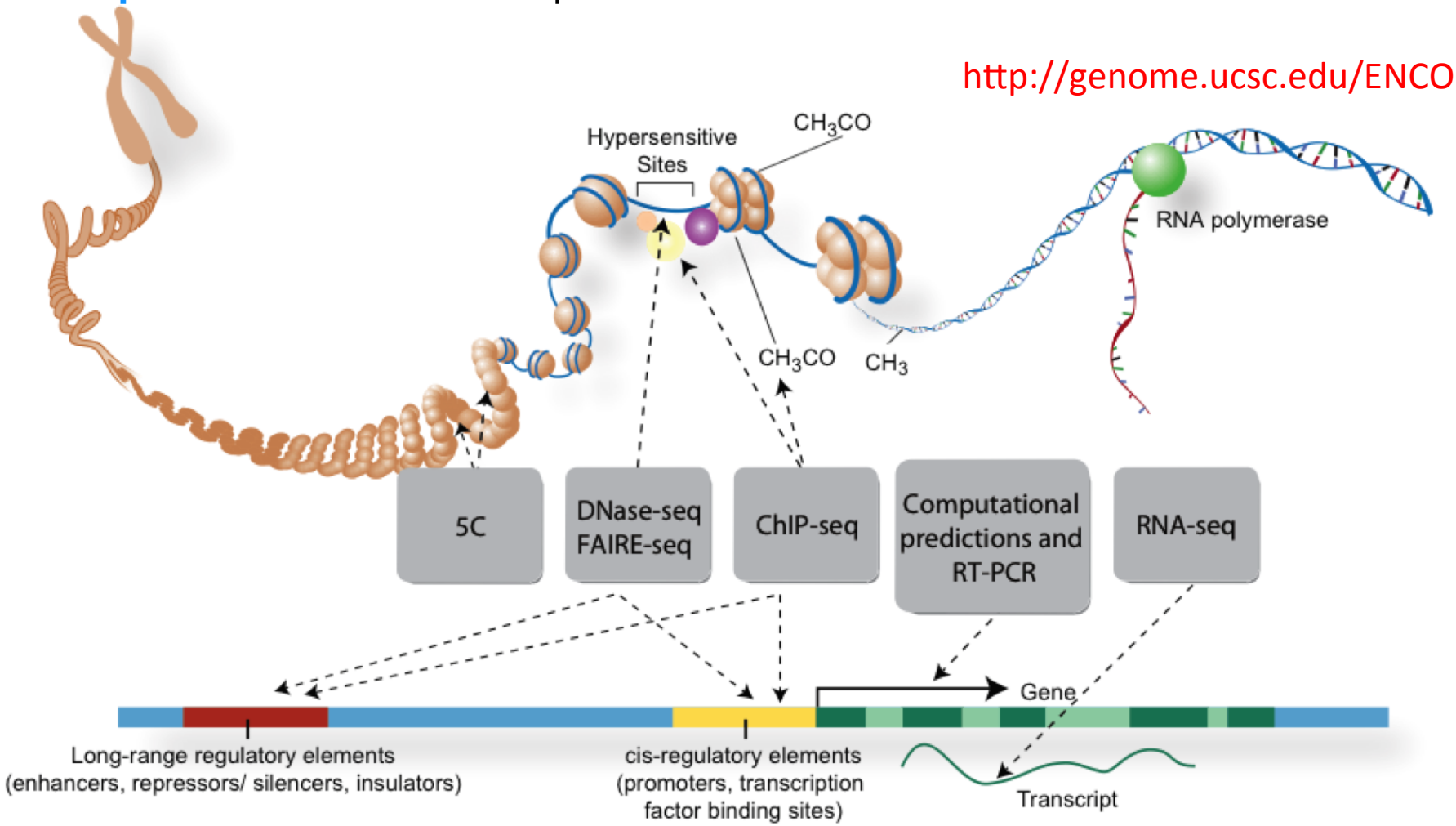
- Web, REST API, local database
- **biomaRt**: Bioconductor package provides an interface to a 68 databases implementing the BioMart software suite.
- These include **ENSEMBL, Uniprot, REACTOME, HGNC, COSMIC** and **HapMap** marts.
- Enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write complex SQL queries.

Gene Regulation: ENCODE

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute. The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODExplorer: build metadata sql database from Encode data.

http://genome.ucsc.edu/ENCODE/



SysBiol: Gene Ontology

http://www.geneontology.org

- The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.
- GO consists of 3 structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.
- Provides ontologies and annotation (updated daily).
- **GO.db:** Bioconductor resource, based on GO data. Limited functionality.

SysBiol: DAVID: The Database for Annotation, Visualization and Integrated Discovery

http://david.abcc.ncifcrf.gov

- A comprehensive set of functional annotation tools to understand biological meaning behind large list of genes.
- Gene-annotation enrichment analysis, functional classification, ID conversion etc.
- **DAVIDQuery:** Bioconductor interface with limited functionality.

SysBiol: KEGG: Kyoto Encyclopaedia of Genes and Genomes

- http://www.genome.jp/kegg

- KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
- Provides well annotated metabolic pathways.
- KEGG.db: out of date
- **KEGGGraph:** depends on KEGG XML
- **KEGGREST:** Useful Bioconductor interface to KEGG. Allows access to the REST API

KEGGREST example:

```
{
library(KEGGREST)
listDatabases()
query <-keggGet(c("hsa:10458", "ece:Z5100"))
png<-keggGet("hsa05130", "image")
}
```

SysBiol: Reactome

- http://www.reactome.org

- Reactome is an open-source, open access, manually curated and peer-reviewed pathway database.
- **Reactome.db**: search Reactome annotation
- **ReactomePA**: pathway enrichment, gene set enrichment analysis and visualization methods.

SysBiol: Annotation Hubs

- This package provides a client for the Bioconductor **AnnotationHub** web resource.
- The AnnotationHub web resource provides a central location where genomic files (e.g., VCF, bed, wig) and other resources from standard locations (e.g., UCSC, Ensembl) can be discovered.
- The resource includes metadata about each resource, e.g., a textual description, tags, and date of modification.
- The client creates and manages a local cache of files retrieved by the user, helping with quick and reproducible access.