Introduction to Annotation Resources

Shamith Samarajiwa
Computational Biology and Statistics Group
University of Cambridge

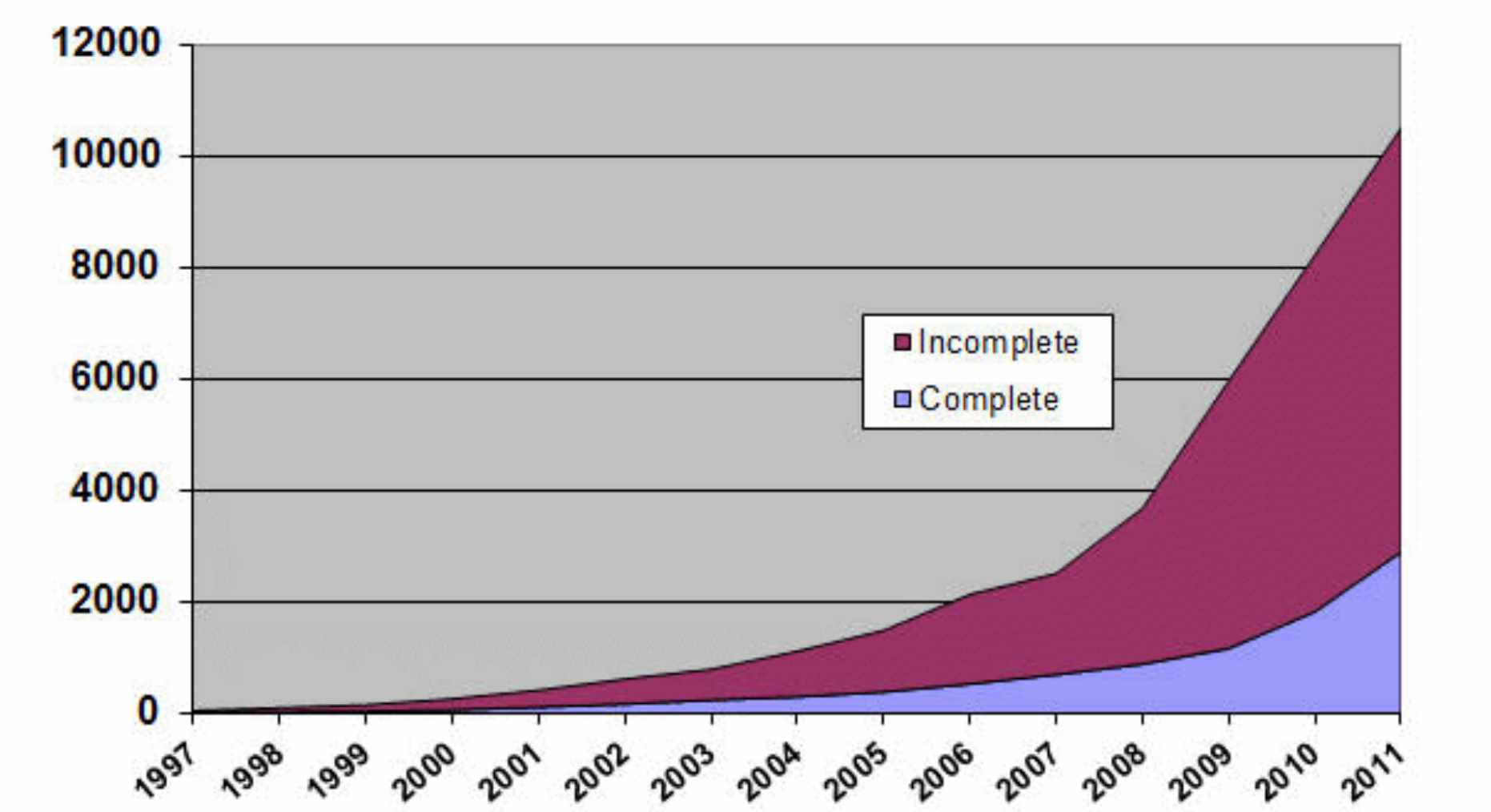
Analysisof High-throughput sequencing data withBioConductor

13-15 March 2014

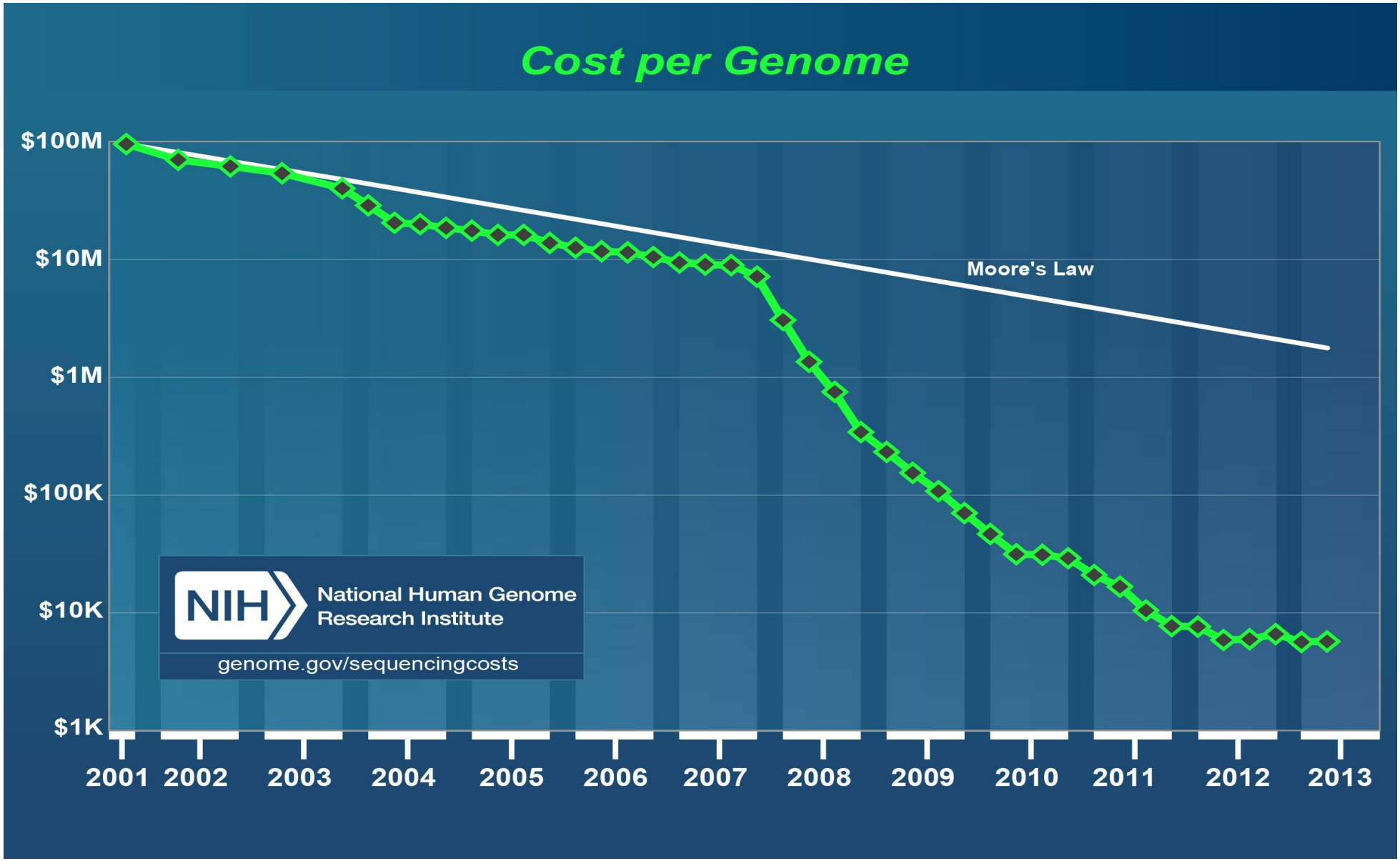
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

UNIVERSITY OF CAMBRIDGE

Genome sequencing projects



Cost of sequencing



http://genome.gov/sequencingcosts/

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 bioinformatic workflows.

- **Bioconductor** provides and extensive number of resources for gene or genome annotation.

Types of Annotation

- Genome builds
- Organism
- Platform specific
- Gene or Genome centric
- Transcriptomic or Protein centric
- Regulation and Systems Biology

External Annotation Resources

- GRC (Genome builds)
- UCSC/Ensembl (Genome)
- GMOD/MGI/RGD/FlyBase (Genome)
- HGNC /EntrezGene (Gene specific)
- Uniprot/PFAM (Proteins)
- Biomart (Annotation from 46 databases)
- ENCODE/modENCODE (Gene Regulation)
- GO/NIH DAVID (SysBiol)
- KEGG/Reactome (SysBiol)

File Formats

• SAM (sequence alignment map) is a generic format for storing large nucleotide sequence alignments. Li *et al.*, 2009 PMID:19505943

• BAM is a compressed binary version of SAM

• FASTA and FASTQ are text based sequence formats

• WIG 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.**

• The bigWig format is for display of dense, continuous data that will be displayed in the Genome Browser as a graph.

• Variant Call Format (VCF) 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.

• More information at :http://genome.ucsc.edu/FAQ/FAQformat.html#format5.1

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.

GRCh37 = hg19
GRCh38
GRCm38 = mm10

The Genome Reference Consortium consists of:

The Wellcome Trust Sanger Institute

The Genome Institute at Washington University

The European Bioinformatics Institute

The National Center for Biotechnology Information

Genome annotation: Genomes

- BSgenome BioConductor packages provide Biostrings based reference genome information.

- Pre-built genomes, e.g., BSgenome.Hsapiens.UCSC.hg19 based on the *H. sapiens* UCSC hg19 build.

- Similar packages are available for other organisms.

- Used in analysis pipelines of other Bioconductor packages.

- Repeat masking, motif finding, SNPs etc.

Genome annotation : UCSC genomes

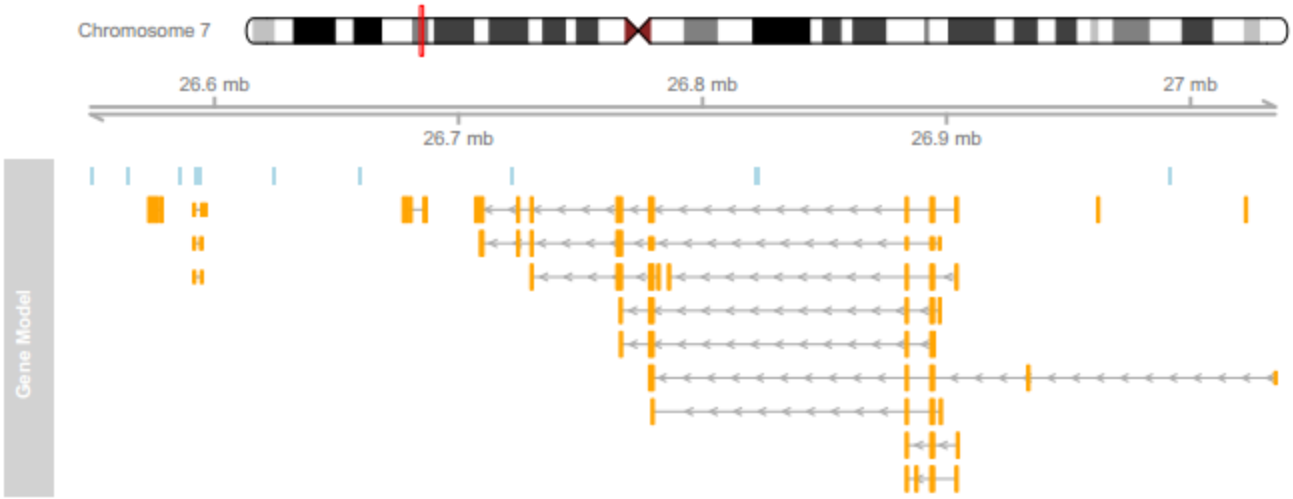
- http://genome.ucsc.edu

- 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** tools. More complex queries can be run via the MySQL server.
- GenomicFeatures: BioConductor package retrieves transcript related features from UCSC genomes and Biomart.
- Rtracklayer: BioConductor package for visualizing browser tracks. The Rtracklayer package is 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 the biomaRt package to run live annotation queries to Ensembl databases and visualizes genes and transcripts.
- Supersedes GenomeGraphs BioCpackage.



Gene annotation: HGNC & Entrez Gene

- HGNC: http://www.genenames.org/
- The HUGO Gene Nomenclature Committee (HGNC) has assigned unique gene symbols and names to over 37,000 human loci, of which around 19,000 are protein coding.
- Gene name disambiguation.

- Entrez Gene: 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.
- Entrez e-Utils: programmatic access to Entrez Gene resource.

Protein annotation: UniProt & PFAM

- www.uniprot.org
- **Uniprot** provides comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

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.

- Protein-family specific annotation at **PFAM** database:
- http://pfam.sanger.ac.uk
- The UniProt.ws Bioconductor package provides a select interface to the UniProt web service.
- PFAM.db provides annotation from PFAM.

Biomart

- http://www.biomart.org

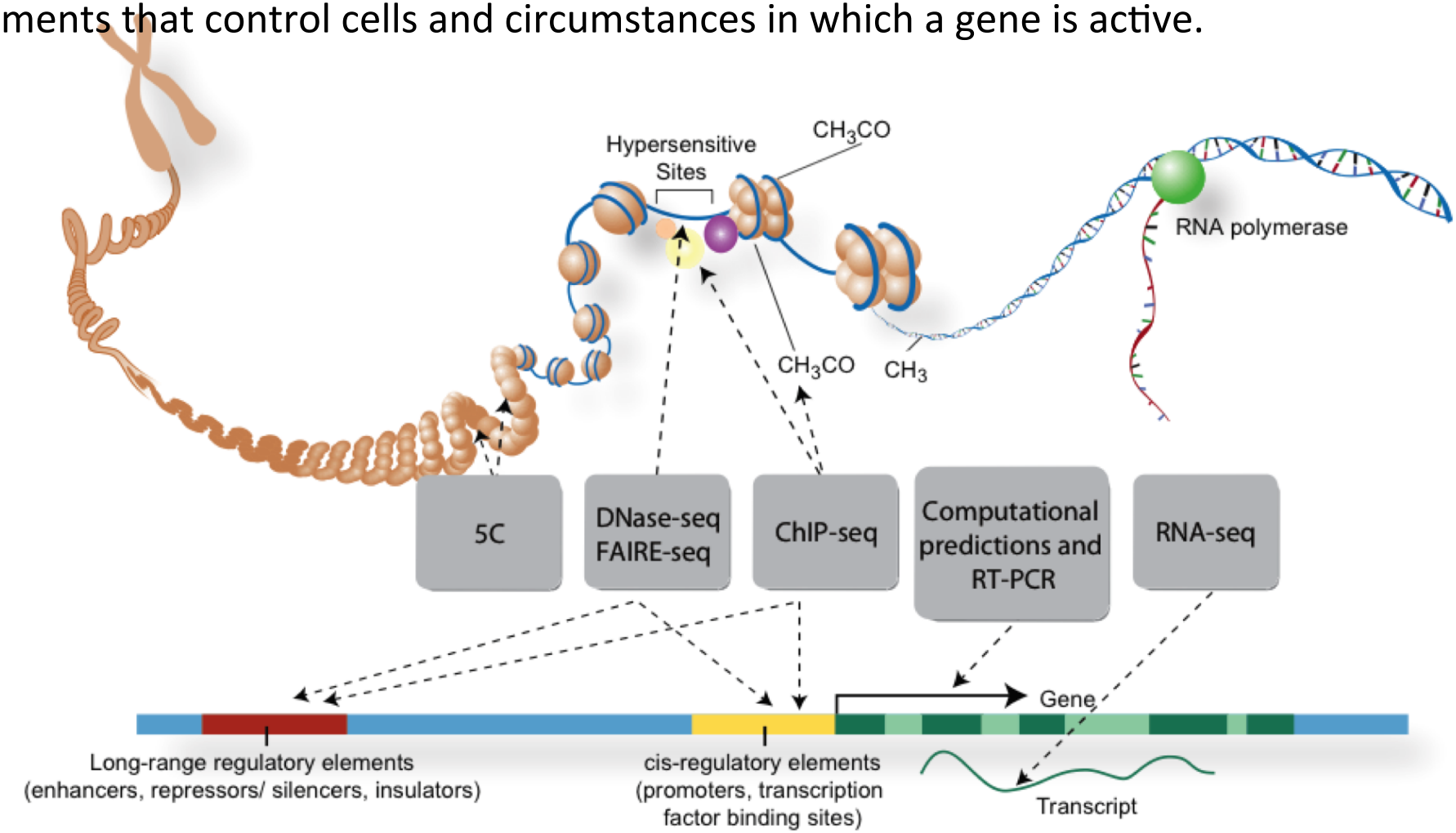
- Web, REST API, local database
- biomaRt: BioConductorpackage provides an interface to a 68 databases implementing the BioMart software suite. These iniclude **ENSEMBL, Uniprot, REACTOME, HGNC, COSMIC** and **HapMap** databases.

- The package enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write any complex SQL queries.

Gene Regulation: ENCODE

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

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (US). 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.



extract ENCODE meta-data

```
library (RMySQL)

connect = dbConnect (“MySQL”, host=“genome-mysql.cse.ucsc.edu”,
user=“genome”, Password=“”, dbname=“hg19”)

query = “select * from metaDb”
# query = “select * from metaDb limit 1000”

res= dbGetQuery (connect, query)
head(res)
```

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 in use.

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

- http://david.abcc.ncifcrf.gov

- Provides a comprehensive set of functional annotation tools to understand biological meaning behind large list of genes.
- Comprehensive gene-annotation enrichment analysis, functional classification, ID conversion etc.

- DAVIDQuery: Bioconductor interface. Limited in function.

SysBiol: KEGG: Kyoto Encyclopedia 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 and signaling 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("has:05130", "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.