

SCBT401 BIOINFORMATICS

GENE FUNCTION & TRANSCRIPTION

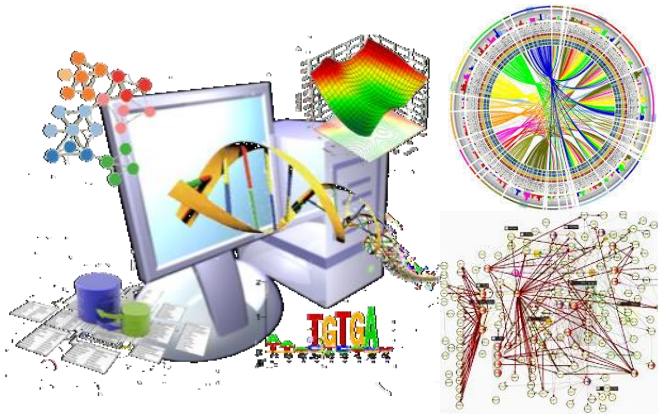
Asst.Prof. Adisak Romsang, Ph.D.

K610 EBI CENTER and SCBT

Faculty of Science, Mahidol University

Tel: 02-201-5962, Email: adisak.rom@mahidol.ac.th

Bioinformatics

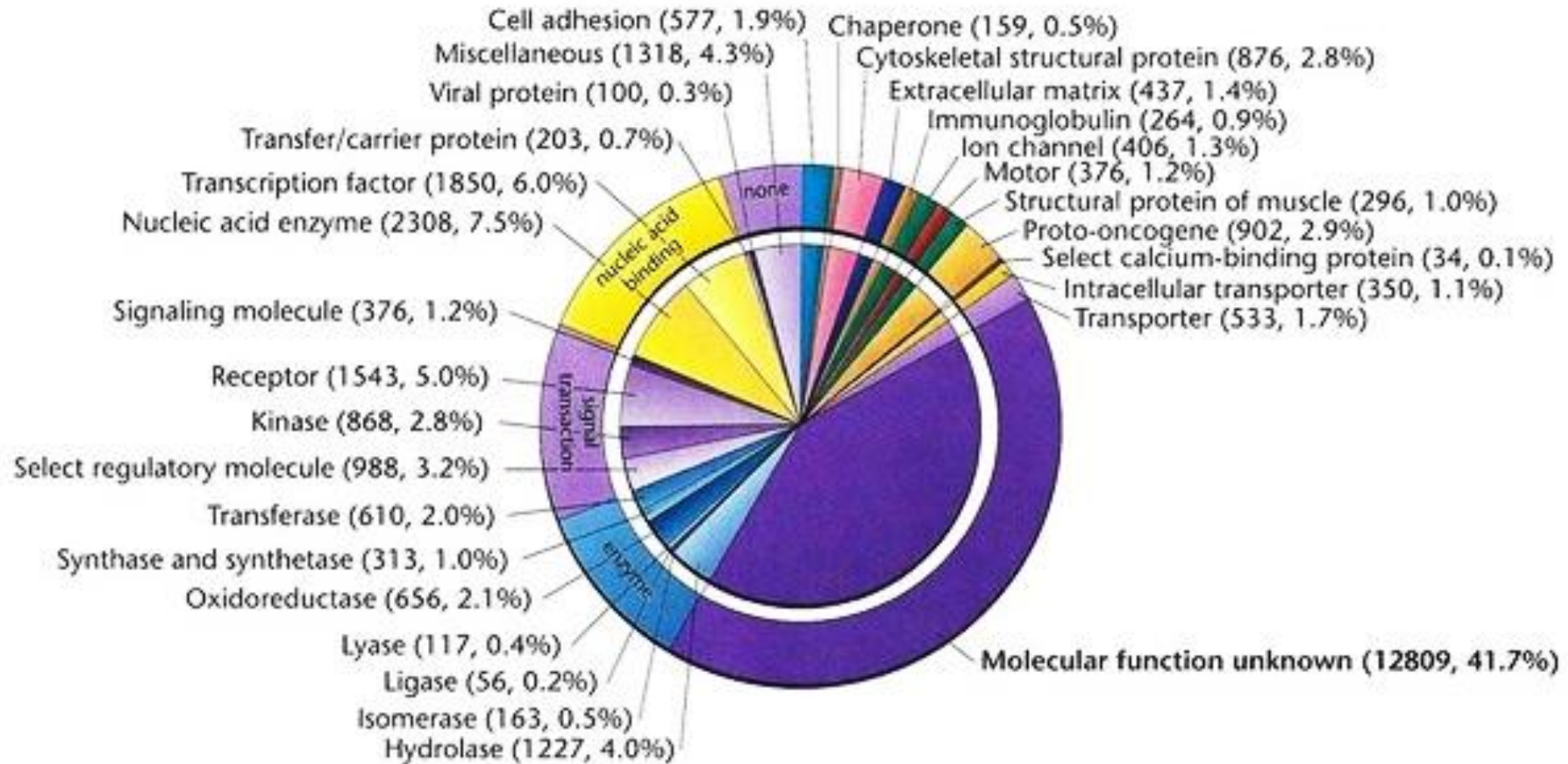


An interdisciplinary field that develops methods and software tools for studying biological data, which combine computer science, statistics, mathematics and engineering.



- ☐ Sequence analysis
- ☐ Gene and protein expression
- ☐ Structural bioinformatics
- ☐ Network and systems biology
- ☐ Databases
- ☐ Software and tools

Gene Function in Human



Its not difficult to introduce mutations, but its extremely difficult to bring desired powers through mutations.

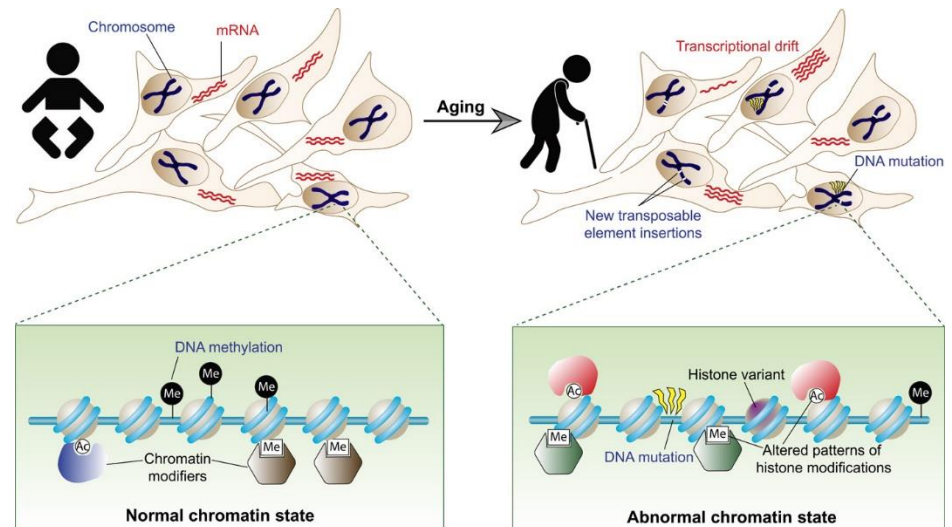
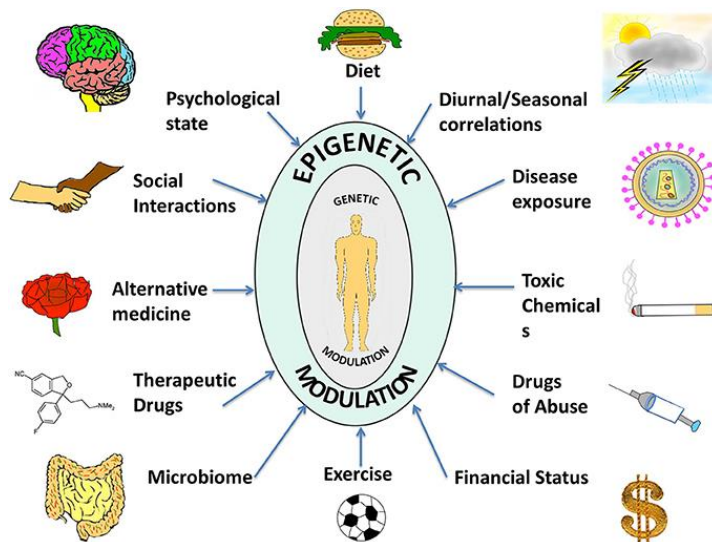
Transcription

Why is important?



There are around 4000,000,000 base pairs in the human genome, so you would expect between 10 and 100 new mutations per person that occur early enough in embryonic development to be present in most cells in the body.

Twins share the same genes but their environments become more different as they age.



Objectives

Students are able to...

- To explain currently molecular techniques for analyze gene expression and function
- To apply the most suitable techniques for analyze the gene expression and function in different research areas
- To synthesize the proper experimental procedures for working on research related to study the gene function and expression

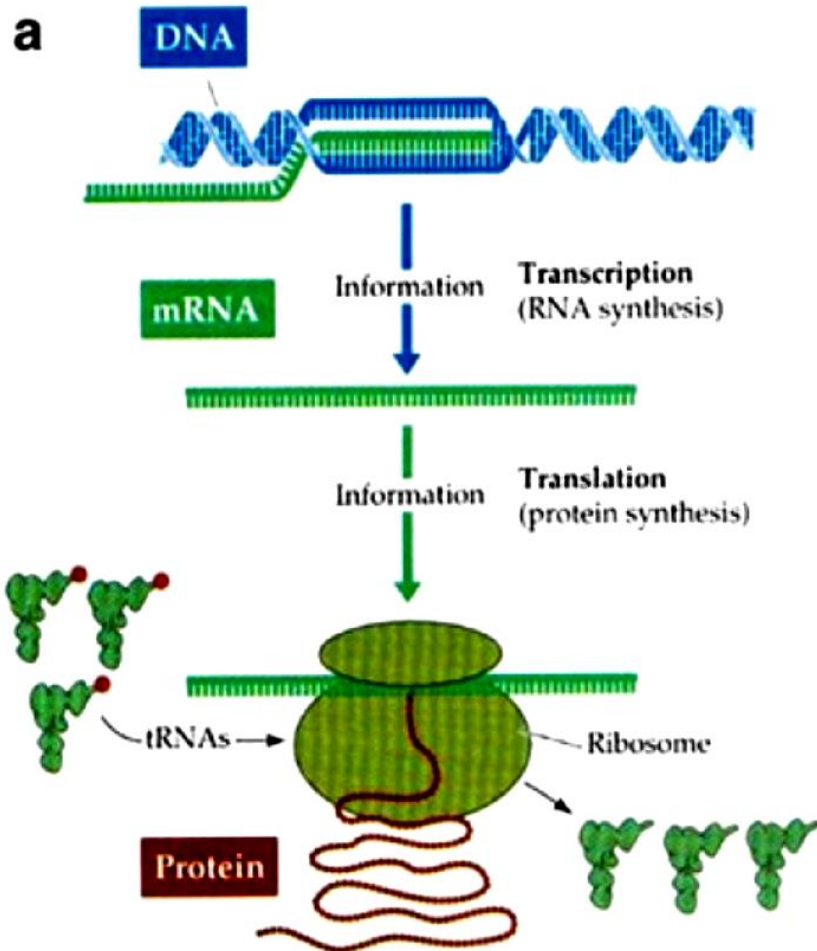
Outline

Students are able to...

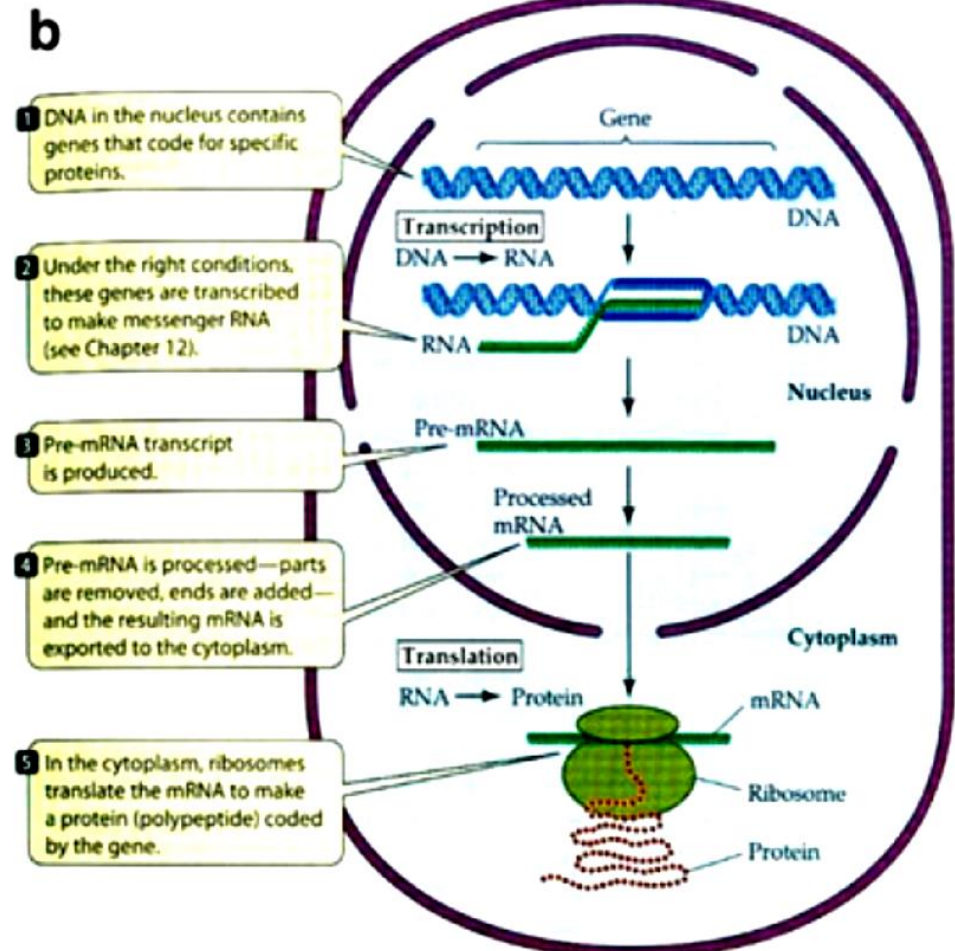
- To explain currently molecular techniques for analyze gene expression and function
- To apply the most suitable techniques for analyze the gene expression and function in different research areas
- To synthesize the proper experimental procedures for working on research related to study the gene function and expression

The Central Dogma

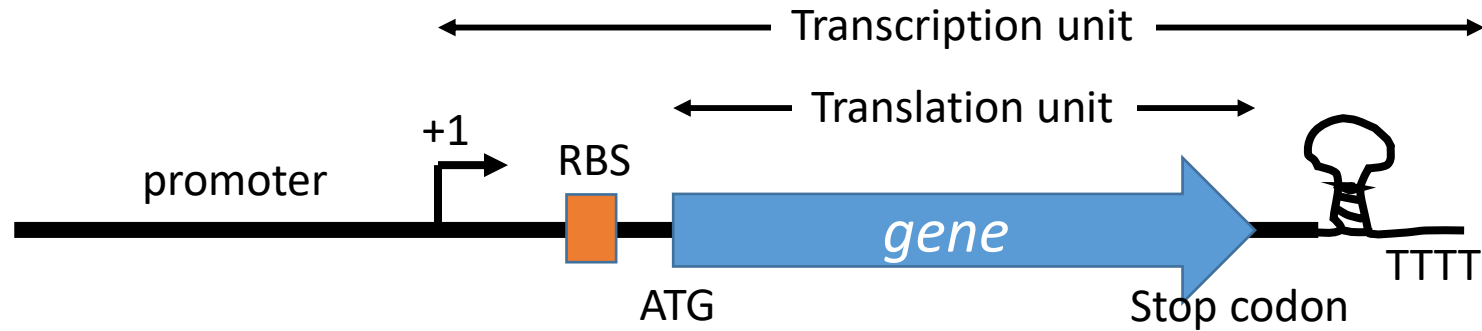
Prokaryote



Eukaryote

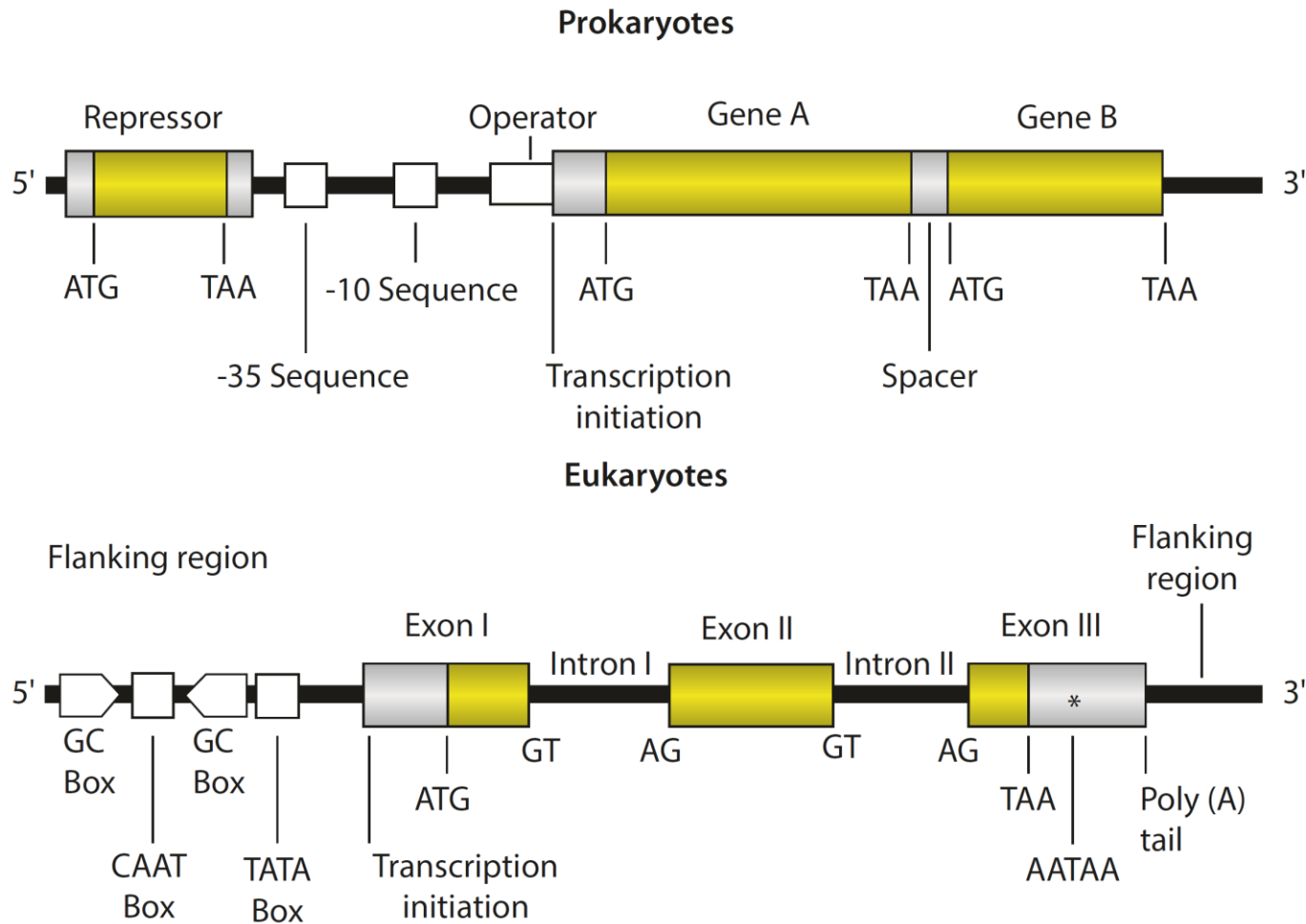


Gene prediction in Prokaryote



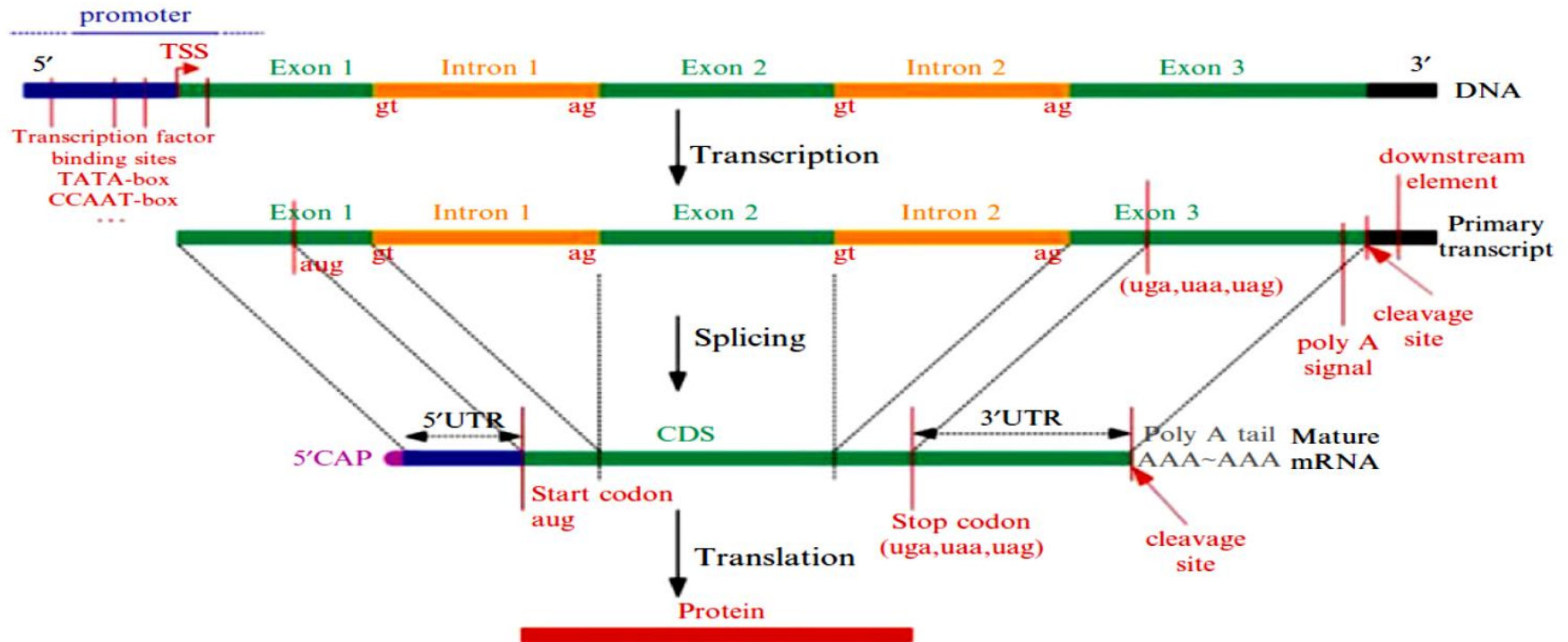
- Translational start codon:
 - ATG (alternatives as GTG or TTG)
- *Shine-Delgarno* sequence
 - purine-rich sequence complementary to 16S rRNA
 - Slightly upstream of translational start codon
 - Consensus sequence of AGGAGGT
- Stop codon
 - TGA, TAG, and TAA in DNA (UGA, UAG, and UAA in RNA)
- Transcriptional termination signal
 - *Rho-independent terminator*
 - stem loop structure followed by a string of Ts

Gene prediction in Eukaryote



■ Fig. 1.5 The structure of gene regions of prokaryotes and eukaryotes

Gene prediction in Eukaryote



Further Reading

Amino acids. https://en.wikipedia.org/wiki/Amino_acid

Biochemistry. <https://en.wikipedia.org/wiki/Biochemistry>

NCBI Books. <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books>

Protein structures. <http://www.rcsb.org/>

Gene prediction in Eukaryote

- Large nuclear genomes → Very low gene density
 - In humans, only 3% of the genome codes for genes, with about 1 gene per 100 kbp on average
 - Space between genes is very large and rich in repetitive sequences and transposable elements
- Mosaic organization
 - Gene is split into pieces (exons) by intervening noncoding sequences (introns)
- 5' capping for methylated initial residue of RNA
- Polyadenylation at 3' end with a consensus of CAATAAA(T/C)
- Splicing junctions of introns and exons: GT-AG rule
- Most vertebrate genes:
 - ATG as translation start codon with conserved flanking sequence (*Kozak sequence*, CCGCCATGG)
 - high density of CG dinucleotides near transcription start site (CpG island, ρ refers to the phosphodiester bond connecting two nucleotides)

Gene prediction program

- Homology-based programs
 - Exon structures and exon sequences of related species are highly conserved
 - Many homologous sequences to be compared with are derived from cDNA or expressed sequence tags (ESTs) of the same species
 - Novel genes in a new species cannot be discovered without matched in the database
- GenomeScan (<http://genes.mit.edu/genomescan.html>)
- EST2Genome (<http://emboss.bioinformatics.nl/cgi-bin/emboss/est2genome>)
- TwinScan (<http://mblab.wustl.edu/>)

Gene prediction program

- Consensus-based programs
 - Integrated approach: use several different programs to generate lists of predicted exons
 - Common predictions agreed by most programs and removing inconsistent predictions
 - GeneComber: combined HMMgene with GenScan
 - (<http://www.bioinformatics.ubc.ca/gencombver/index.php>)
 - DIGIT (<http://digit.gsc.riken.go.jp/cgi-bin/index.cgi>)
 - Combines FGENESH, GENSCAN and HMMgene

ORF Finder and BLAST

<https://www.ncbi.nlm.nih.gov/orffinder/>

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for [Linux x64](#).

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt



Enter Query Sequence

 Enter accession number, gi, or nucleotide sequence in FASTA format:

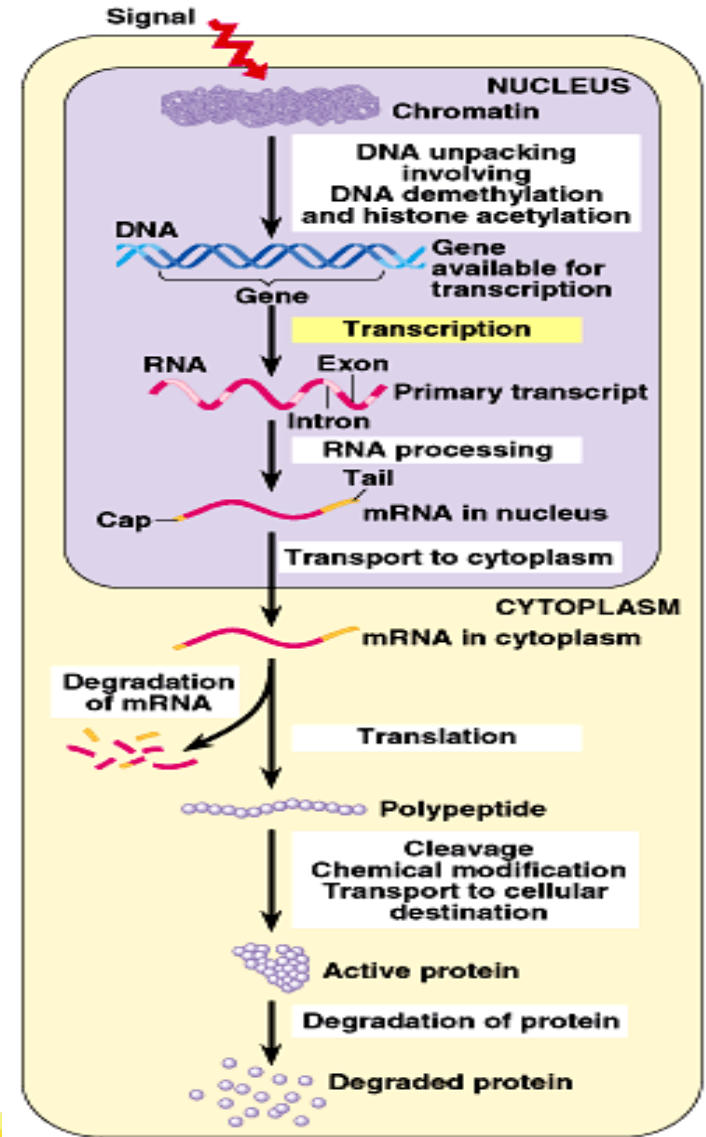
 From: To:

Confirmation of gene prediction ?

- Transcribe as mRNA
 - Gene expression analysis
 - Transcription start site
 - RNA detection
- Translate as protein
 - *In vitro* translation
 - Protein detection
- Protein activity (function)
 - Protein activity assay

Gene expression and regulation

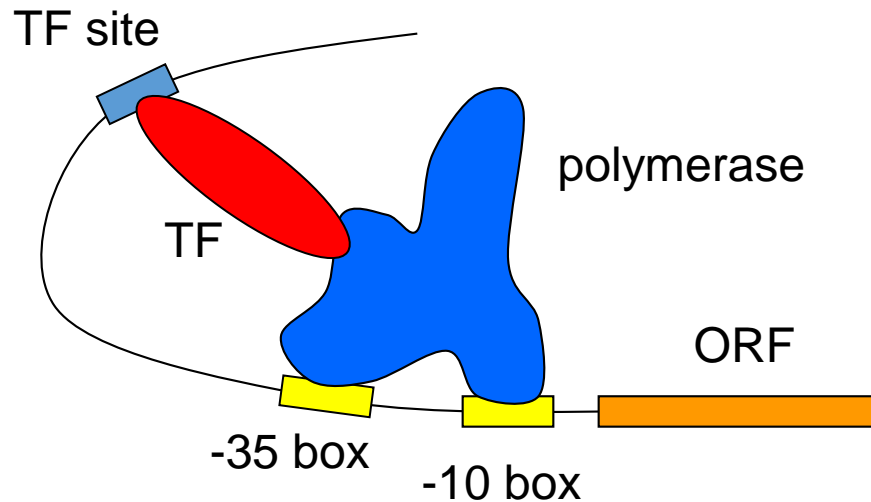
- Gene expression:
 - house-keeping (always express in the similar level)
 - inducible (regulated by regulatory elements)
- Gene regulation:
 - Prokaryote: Mainly transcriptional regulation and indirect regulation: small RNAs, tRNAs, and proteins
 - For eukaryote, many levels of control include chromatin packing, transcription, RNA processing, translation, and various alterations to the protein product.



Transcriptional regulation

- Direct transcriptional regulation
 - Protein-DNA binding affects the level of gene expression
 - Promoter analysis: sequence and motif
 - Regulatory elements
 - RNA polymerase and sigma factor
 - Regulatory proteins
 - Binding motif and structure of protein
 - Binding affinity and sequence homology
 - Sensor and response
- Indirect regulation
 - Change in gene expression via intermediates

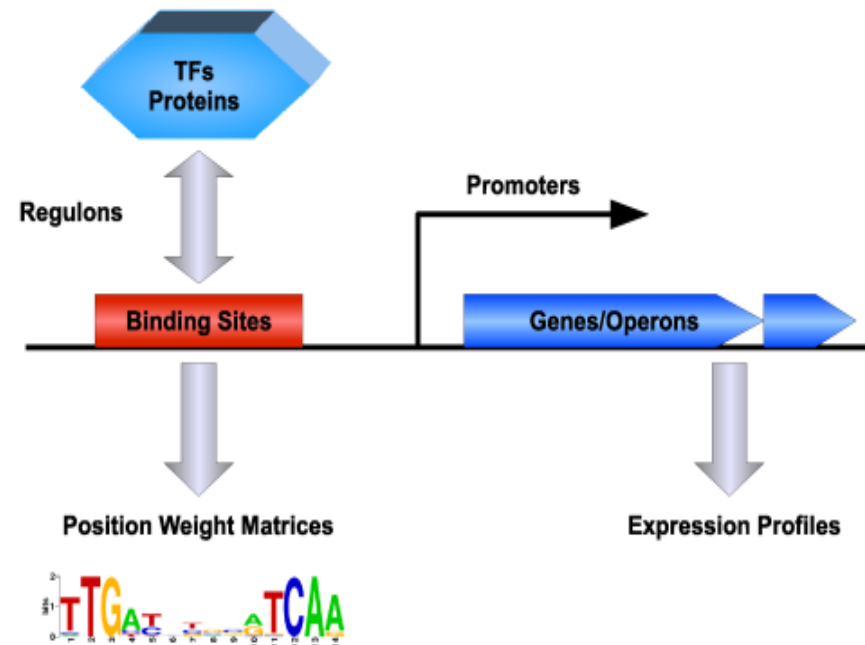
Promoter prediction: Prokaryotic gene



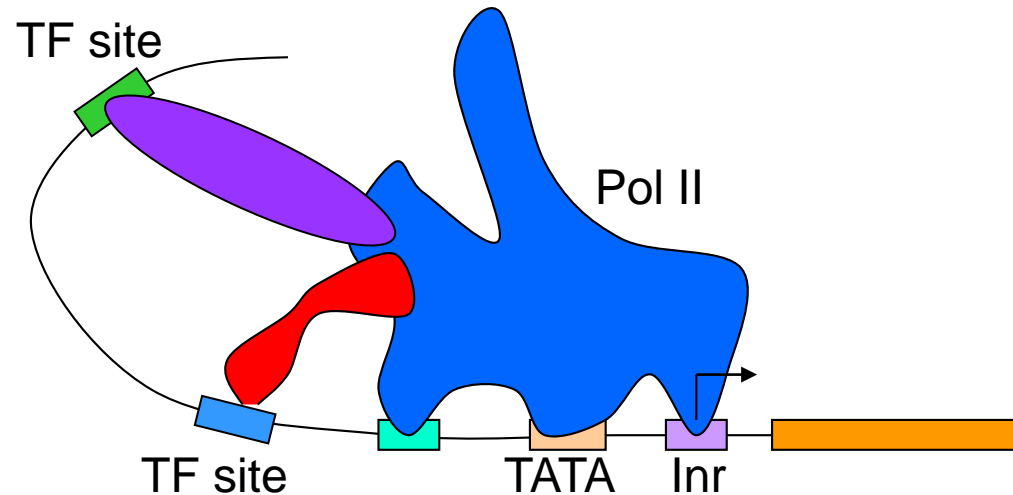
- σ^{70} factor binds to -35 and -10 boxes and recruit full polymerase enzyme
- - 35 box consensus sequence: TTGACA
- - 10 box consensus sequence: TATAAT
- Transcription factors that activate or repress transcription
- Bind to regulatory elements
- DNA loops to allow long-distance interactions

Promoter prediction: Prokaryotic gene

- Find operon and upstream of-first gene is promoter
- Wang rules (distance between genes, no ρ -independent termination, number of genomes that display linkage)
- BPRM (<http://www.softberry.com>)
- Based of arbitrary setting of operon open distances
- 200bop upstream of first gene
- many FPs
- FindTerm (<http://sun1.softberry.com>)
- Searches for ρ -independent termination signals
- Prodoric (<http://prodoric.tu-bs.de/>)
- A comprehensive database about gene regulation and gene expression in prokaryotes



Promoter prediction: Eukaryotic gene structure



Polymerase I, II and III

Basal transcription factors (TFIID, TFIIA, TFIIIB, etc.)

TATA box (TATA(A/T)A(A/T))

“Housekeeping” genes often do not contain TATA boxes

Initiator site (Inr) (C/T) (C/T) CA(C/T) (C/T) coincides with transcription start

Many TF sites

Activation/repression

Promoter prediction: Eukaryotic gene structure

- Searching for consensus sequences in databases (TransFac)
- Increase specificity by searching for CpG islands
- High density of transcription factor binding sites
- CpGProD (<http://pbil.univ-lyon1.fr/software/cpgprod.html>)
- CG% in moving window
- Eponine (<http://servlet.sanger.ac.uk:8080/eponine/>)
- Matches TATA box, CCAAT box, CpG island to PSSM
- Cluster-Buster (<http://zlab.bu.edu/cluster-buster/cbust.html>)
- Detects high concentrations of TF sites
- FirstEF (<http://rulai.cshl.org/tools/FirstEF/>)
- QDA of first exon boundary
- McPromoter (<http://genes.mit.edu/McPromoter.html>)
- Neural net of DNA bendability, TAT box, initiator box
- Trained for *Drosophila* and human sequences

Transcriptional regulators

- [Tfsitescan](#) (*Institute for Transcriptional Informatics, Pittsburgh, U.S.A.*) - This tool is intended for promoter sequence analysis and works best with sequences of ~500 nt.
- [PLACE](#) (*National Institute of Agrobiological Sciences, Japan*) - Plant cis-acting regulatory DNA elements. [PlnTFDB](#) - Plant Transcriptional Factor Database - allows BLAST searching (Reference: P. Pérez-Rodríguez et al. 2009 Nucl. Acids Res. **38**: D822-D872) or [here](#) for related site.
- [DBD](#): Transcription factor prediction database (*Gesellschaft für Biotechnologische Forschung mbH (GBF), Braunschweig, Germany*) (Reference: D. Wilson et al. 2010. Nucl. Acids Res. **36**: D88-D92)
- [rVista 2.0](#) (*Comparative Genomics Center, Lawrence Livermore National Laboratory, U.S.A.*) - High-throughput discovery of functional regulatory elements in sequence alignments. Excluding up to 95% false positive transcription factor binding sites predictions while maintaining high sensitivity of the search.
- [TESS](#) (Transcription Element Search System) is a web tool for predicting transcription factor binding sites in DNA sequences. It can identify binding sites using site or consensus strings and positional weight matrices from the TRANSFAC, JASPAR, IMD, and our CBIL-GibbsMat database. This resource is best for scanning short DNA sequences.
- [PlantTFDB](#) Plant Transcription Factor Database (*Peking University, China*) - provides [search](#) and [Blast](#) search capability.
- Bmicc (<http://www.bmicc.org/web/english/home>) National Scientific Data Sharing Platform for Population and Health Biocline Information Center

Protein-DNA binding motif

| Gateway site | URL |
|--|---|
| NCBI Genomic Biology | http://www.ncbi.nlm.nih.gov/Genomes/index.html |
| GOLD (Genomes OnLine Database) | http://www.genomesonline.org/ |
| TIGR Microbial Database | http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi |
| Bacterial genomes | http://genolist.pasteur.fr/ |
| Yeast databases | http://genome-www.stanford.edu/Saccharomyces/yeast_info.html |
| Ensembl Genome Database Project | http://www.ensembl.org/ |
| MIPS (Munich Information Center for Protein Sequences) | http://mips.gsf.de |

- [TRANSFAC Matrix models](#)
- One important point to know is that not all the matrix models are experimentally curated or good. So not all models can be used from TRANSFAC.
- [JASPAR](#) is another one mentioned above with high quality matrix models.
- [UniPROBE](#)
- [Human Protein-DNA Interactome \(hPDI\)](#)
- [Factorbook](#)

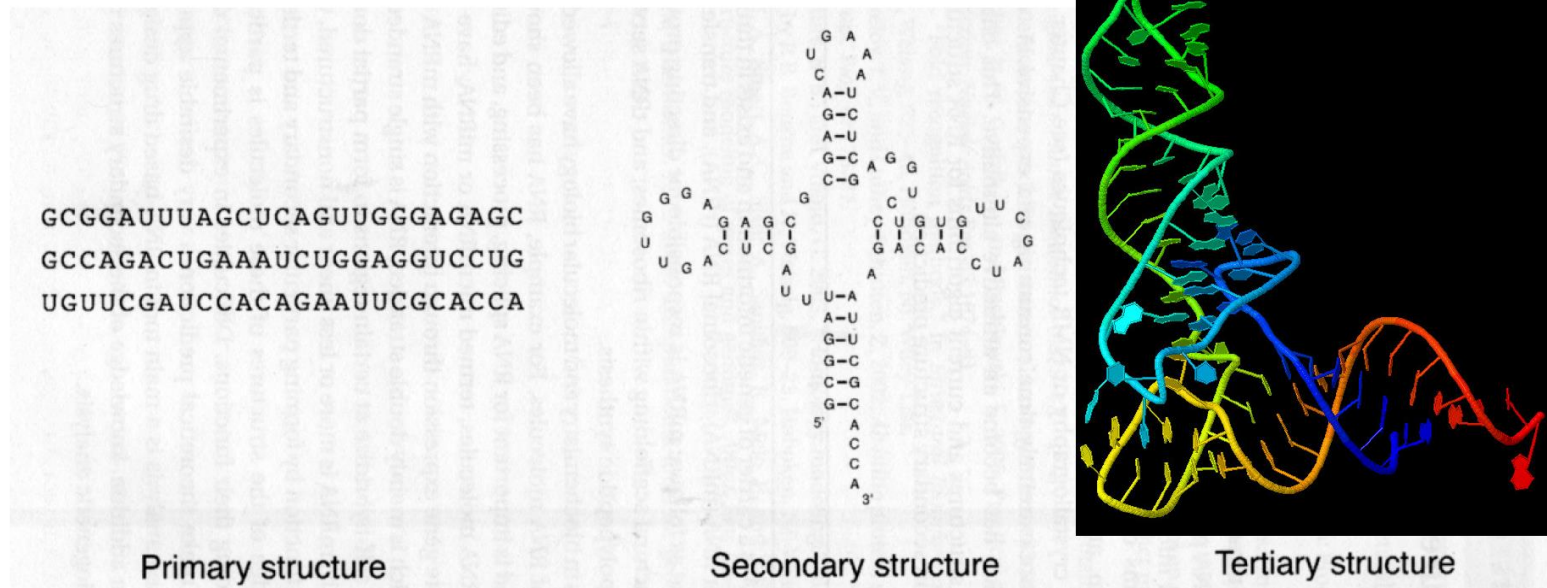
<http://www.gene-regulation.com/pub/databases.html>

MAPPER₂ - Multi-genome Analysis of Positions and Patterns of Elements of Regulation

MotifMap: genome-wide maps of regulatory elements

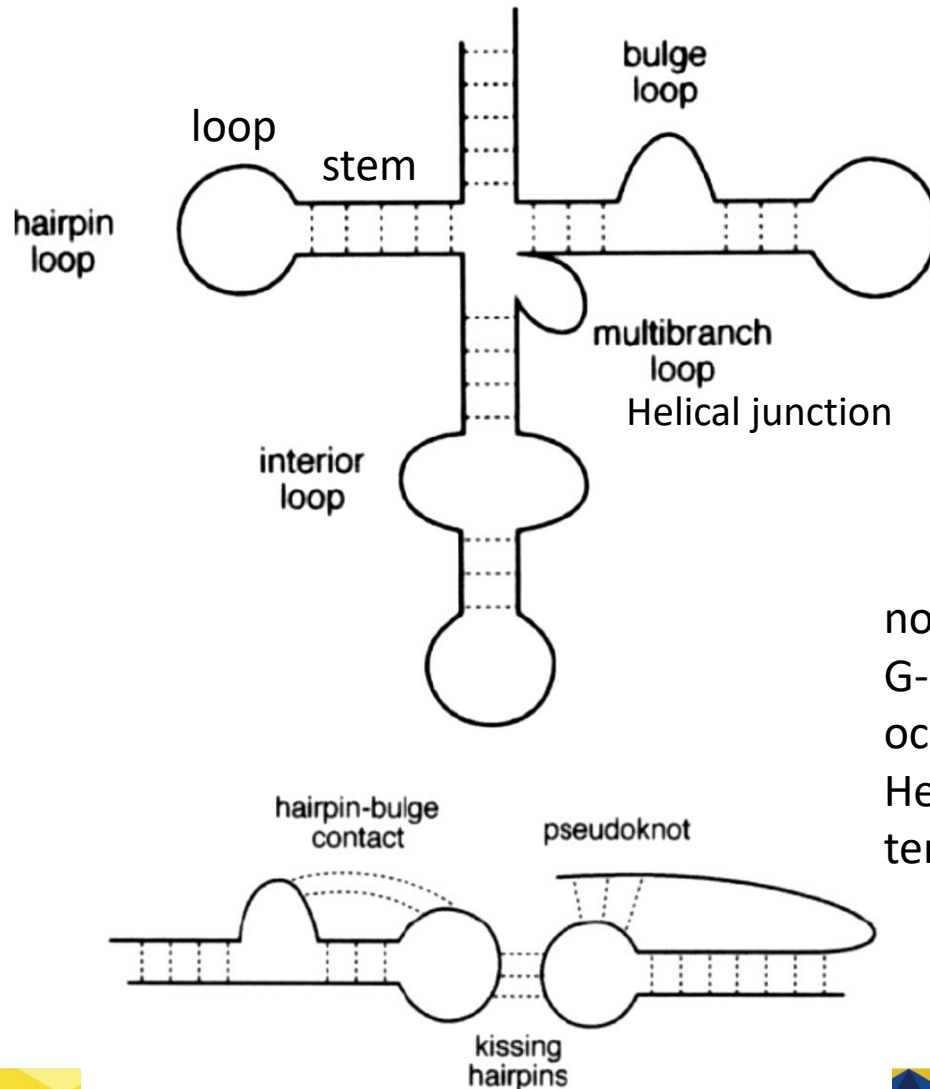
Gene expression: RNA structure

- RNA is single stranded, although some parts can self-hybridize to form partial double-stranded structures.
- mRNA is more or less linear and nonstructured, whereas rRNA and tRNA can only function by forming particular secondary and tertiary structures.



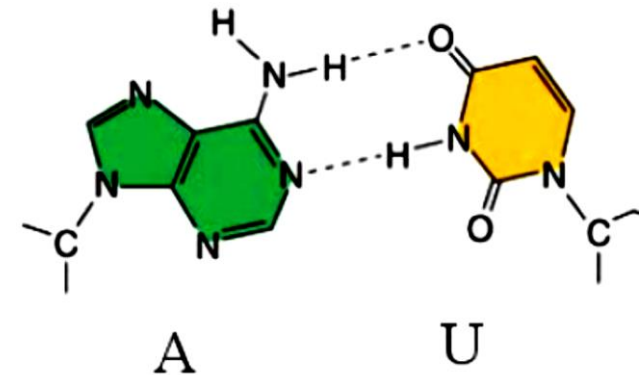
- RNA can act as enzymes (ribozymes) to speed chemical reactions. tRNA structure is responsible!
- The secondary structures of rRNA is key for RNA-based phylogenetic analysis.

Type of RNA structures

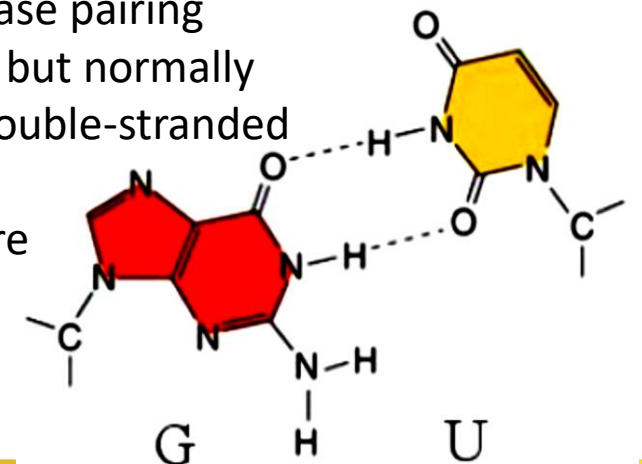


Watson-Crick base pairing
(canonical base pairing)

A-U, G-C



noncanonical base pairing
G-U, less stable but normally
occurs within double-stranded
Helix to form
tertiary structure



RNA structure prediction

Prediction based on a single RNA sequence

Search for RNA structure with lowest energy

Free energy calculated from $G-C < A-U < G-U < \text{unpaired pairs}$

Stacking between aromatic rings (*van der Waals* interactions) gives rise to cooperativity

Neighboring loops or bulges impose unfavorable entropic change

Find all possible base-pairing interaction

Calculate the energy of each and choose the lowest energy configuration

1. Dot Matrices

Plot all interactions in self alignment plot

Find diagonals after applying sliding window

2. Dynamic Programming

Find the single optimal match

Use Watson-Crick and wobble base pairing scores

Conformations with slightly higher energies may exist without optimal base pairing

RNA structure prediction program

3. Partition function

Use a probability distribution to generate sub-optimal structures within a given energy range

Mfold

<http://mfold.bioinfo.rpi.edu/applications/mfold/>

Dynamic programming and thermodynamic calculation

RNAfold

<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>

Extend alignment to more than one diagonal in dotplot to calculate thermodynamic stability of structures

RNA structure prediction program

4. Comparative Approach

- Assumption that homologous RNA sequences fold into same structure
- Covariant regions in homologous sequences are likely to be base-paired
- Predict consensus structure based on predictions for all aligned sequences

RNAalifold

<http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi>

Prealignment

Predictions based on covariance, minimum free energy, dynamic programming
finds optimal structure for entire alignment

Foldalign

<http://foldalign.ku.dk/>

No prealignment

Clustal alignment and dynamic programming

SOFT BERRY

<http://www.softberry.com/berry.phtml?topic=index&group=programs&subgroup=gfindb>

BP-ROM

<http://www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb>

FIND-TERM

<http://www.softberry.com/berry.phtml?topic=findterm&group=programs&subgroup=gfindb>



Virtual Footprint

<http://www.prodoric.de/vfp/>

Transcriptomics

Global expression analysis and DNA microarrays

RNA expression can be measured by hybridizing the RNA to other oligonucleotides. Analyzing signal intensities under different conditions can identify the levels of differential gene expression.

DNA microarrays

There are many types of microarray that are appropriate for different types of analysis. DNA microarray technology uses a grid of oligonucleotides on a chip that hybridize to the complementary target RNA. The level of transcript is determined by the hybridization signal of either Affymetrix- or Agilent-based microarrays. Affymetrix microarrays comprise 25-mer oligonucleotides and there are 11–20 probe-pairs per probe-set per gene. Agilent microarrays have longer oligonucleotide probes, which are more specific, but this method has only one probe per gene.

Tiling arrays

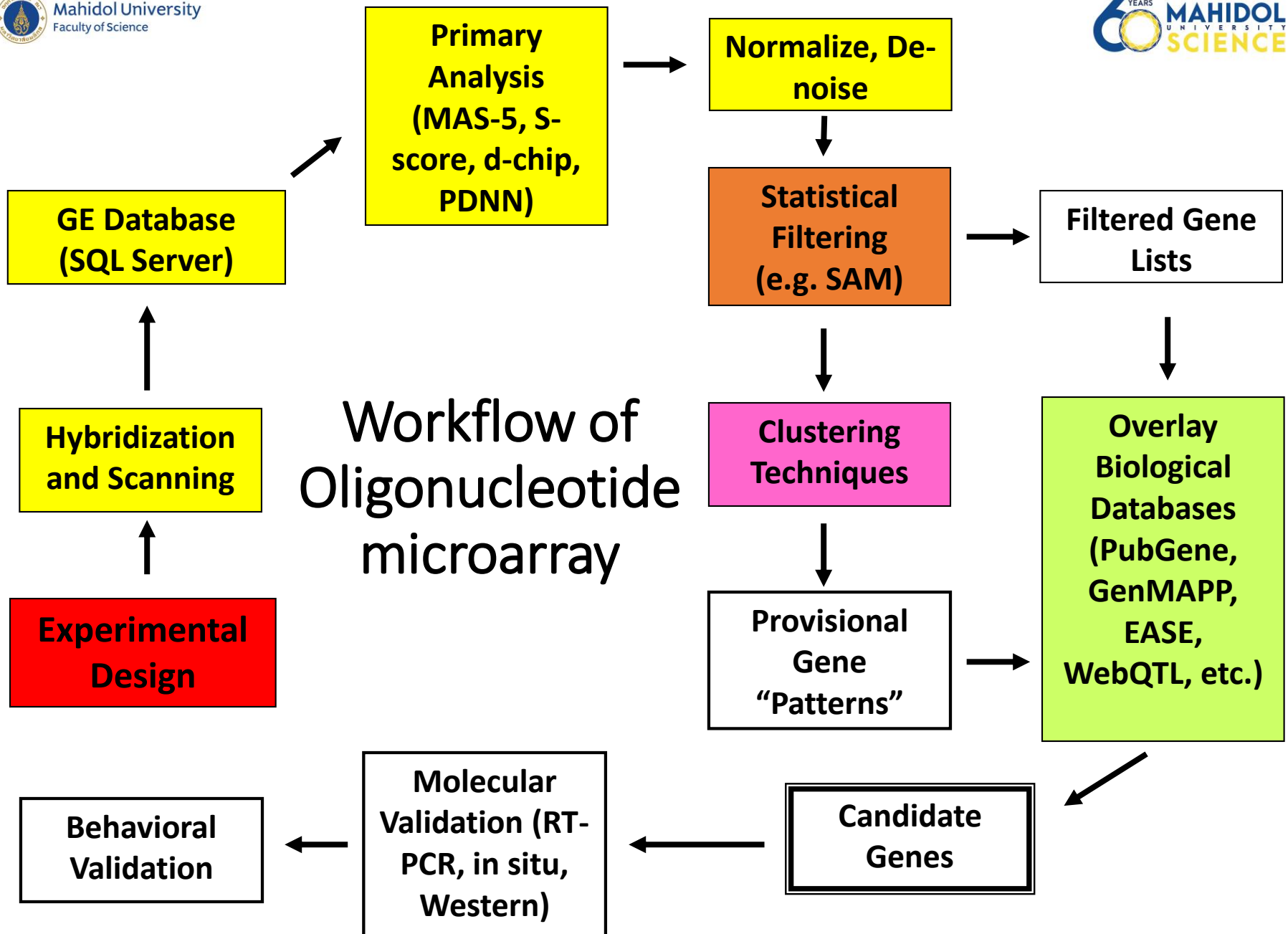
Tiling arrays have probes that target sections of the genome at a very high resolution. The tiling probes can be overlapping or have a short gap between them. Tiling arrays are used in ChIP-chip, MeDIP-chip, and DNase-chip studies.

SNP microarrays

SNP microarrays measure single nucleotide polymorphisms that occur in different diseases, or a comparison of tissues or treatments. They are the key technology in human disease genome-wide association studies and drug development studies.

Next generation sequencing

Next generation sequencing encompasses new technologies that provide fast and accurate sequence reads. The number of sequencing reads can be converted into levels of differential gene expression.



Global gene expression analysis

Gene expression matrices

The raw data from microarray experiments is converted into tables known as gene expression matrices. The rows represent genes and the columns represent experimental conditions. The values in the matrices are measurements of signal intensities, representing relative levels of gene expression.

Grouping expression data

Each gene in a gene-expression matrix has an expression profile, relative to the changes in expression measurement over a range of conditions. The analysis of microarray data involves grouping these data on the basis of similar expression profiles. If a pre-defined classification system is used to group the genes, the analysis is described as supervised. If there is no pre-defined classification, the analysis is described as unsupervised.

Tools for microarray data analysis

Many software applications are available for the analysis of microarray data and these can be downloaded and installed on local computers. Two examples of microarray analysis software platforms include GeneSpring and Bioconductor. There are also several resources available for the analysis of microarray data over the internet; Expression Profiler is the most widely used. Several gene expression databases have been constructed for the storage and dissemination of microarray data. These include the NCBI Gene Expression Omnibus and the EBI ArrayExpress database.

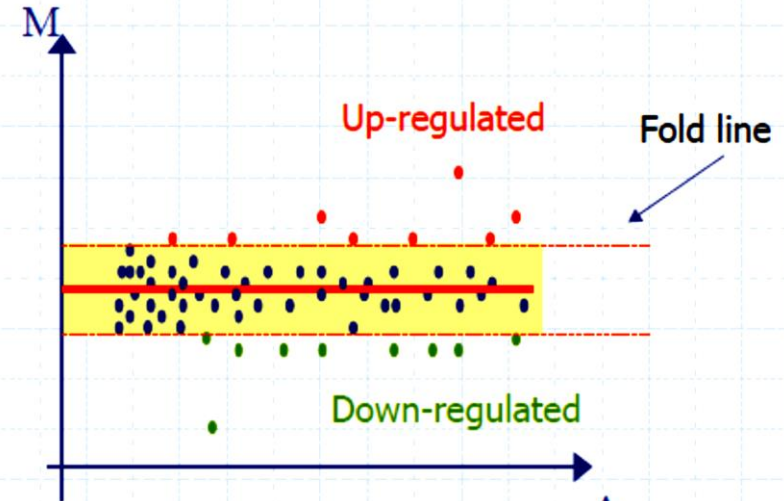
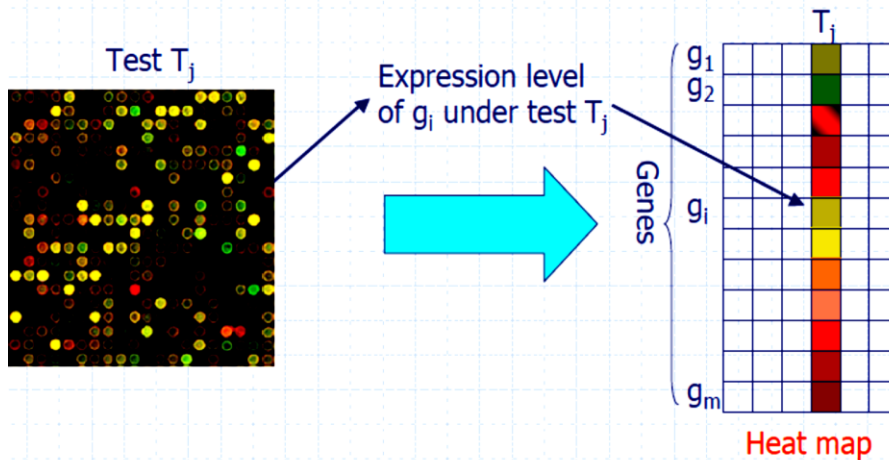
Differential expression

Differential expression refers to the up- and down-regulated genes a microarray experiment. The levels of expression are commonly determined by a fold-change at a set cutoff value. Volcano plots are a common representation of genes that are selected by fold-change and p value.

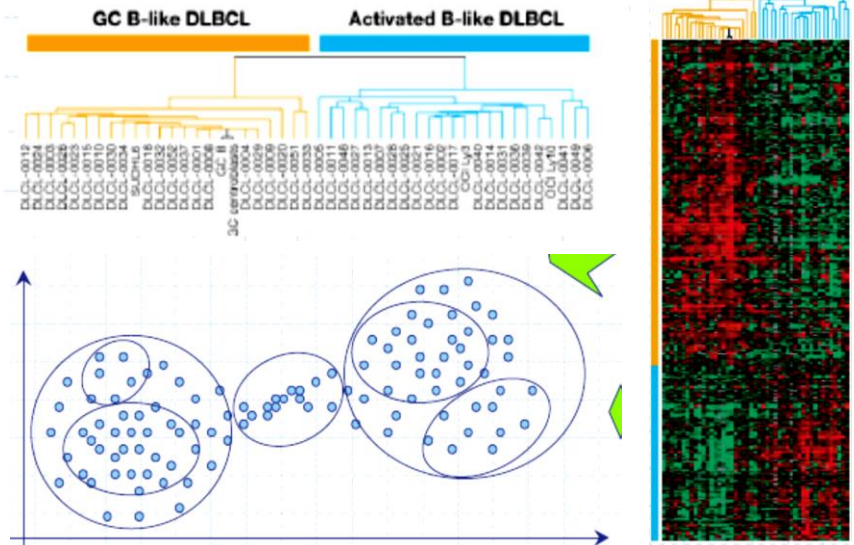
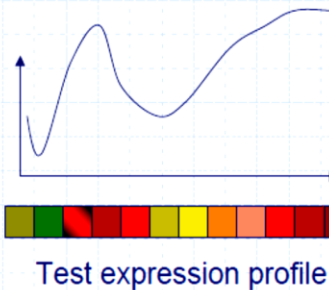
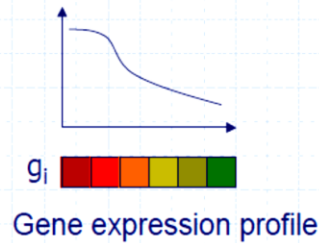
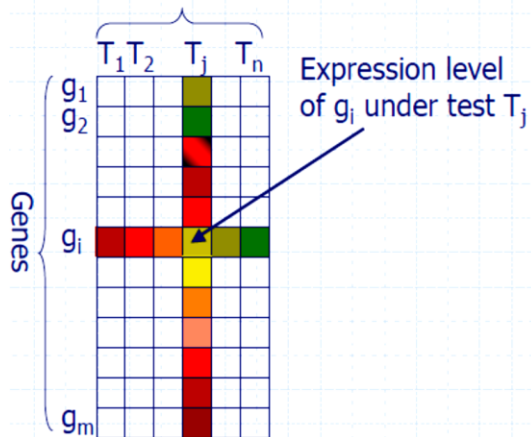
Mapping of expression data onto networks

Gene expression data can be mapped onto a network, which can be protein interaction, gene regulatory, or metabolic networks. Cytoscape is the main network visualization and analysis tool.

Microarray data interpretation



Tests/experiments/samples/conditions



- Heat map
- Gene profile \rightarrow co-expression
 - Test/sample profile \rightarrow sample similarity

Stepwise Analysis of Microarray Data

- Low-level analysis -- image analysis, expression quantitation
- Primary analysis -- is there a change in expression?
- Secondary analysis -- what genes show correlated patterns of expression? (supervised vs. unsupervised)
- Tertiary analysis -- is there a phenotypic “trace” for a given expression pattern?

Table 1. Microarray analysis software tools and internet resources for microarray expression analysis

| Product | Features | URL |
|---|--|---|
| Microarray analysis software tools | | |
| GeneSpring GX | Very popular and powerful tool for biologists. Full support for most microarray platforms, Affymetrix and Agilent are examples. Extensive analysis including clustering, PCA, and pathway analysis. License required | http://www.chem.agilent.com/en-US/products/software/lifesciencesinformatics/pages/gp35082.aspx |
| Bioconductor | R-based tool with many libraries for microarray analysis including extensive Affy and Agilent support. Free | http://www.bioconductor.org/ |
| DChip | Analysis and visualization of gene expression and SNP arrays | http://biosun1.harvard.edu/complab/dchip/ |

Examples of sites with extensive links to microarray analysis software and resources

| | | |
|--------------|--|---|
| Gene Pattern | Extensive list of software resources from Stanford University and other sources, both downloadable and WWW-based | http://www.broad.mit.edu/cancer/software/genepattern/ |
|--------------|--|---|

Examples of WWW-based microarray data analysis

| | | |
|---------------------|--|---|
| Expression profiler | Very powerful suite of programs from the EBI for analysis and clustering of expression data | http://www.ebi.ac.uk/expressionprofiler/index.html |
| EPClust | Generic data clustering, visualization, and analysis tool | http://www.bioinf.ebc.ee/EP/EP/EPCLUST/ |
| Genevestigator | Provides gene expression meta-profiles for animals and plants (e.g., human, mouse, rat, and <i>arabidopsis</i>) | https://www.genevestigator.ethz.ch/gv/index.jsp |

The major microarray databases

| | | |
|------------------------------------|--|---|
| NCBI GEO (Gene Expression Omnibus) | GEO is a gene expression and hybridization array database, which can be searched by accession number, through the contents page, or through the Entrez ProbeSet search interface | http://www.ncbi.nlm.nih.gov/geo/ |
| ArrayExpress | EBI microarray gene expression database. Developed by MGED and supports MIAME | http://www.ebi.ac.uk/microarray-as/ae/ |
| Stanford Microarray Database | Microarray database that provides a resource for the scientific community. Many tools to explore and analyze the data | http://genome-www5.stanford.edu/ |
| NASCArrays | Standard microarray database consisting of plant and other species data. Data mining tools and experiment search functions | http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl |

Microarray Analysis software tools

Resources of gene expression analysis

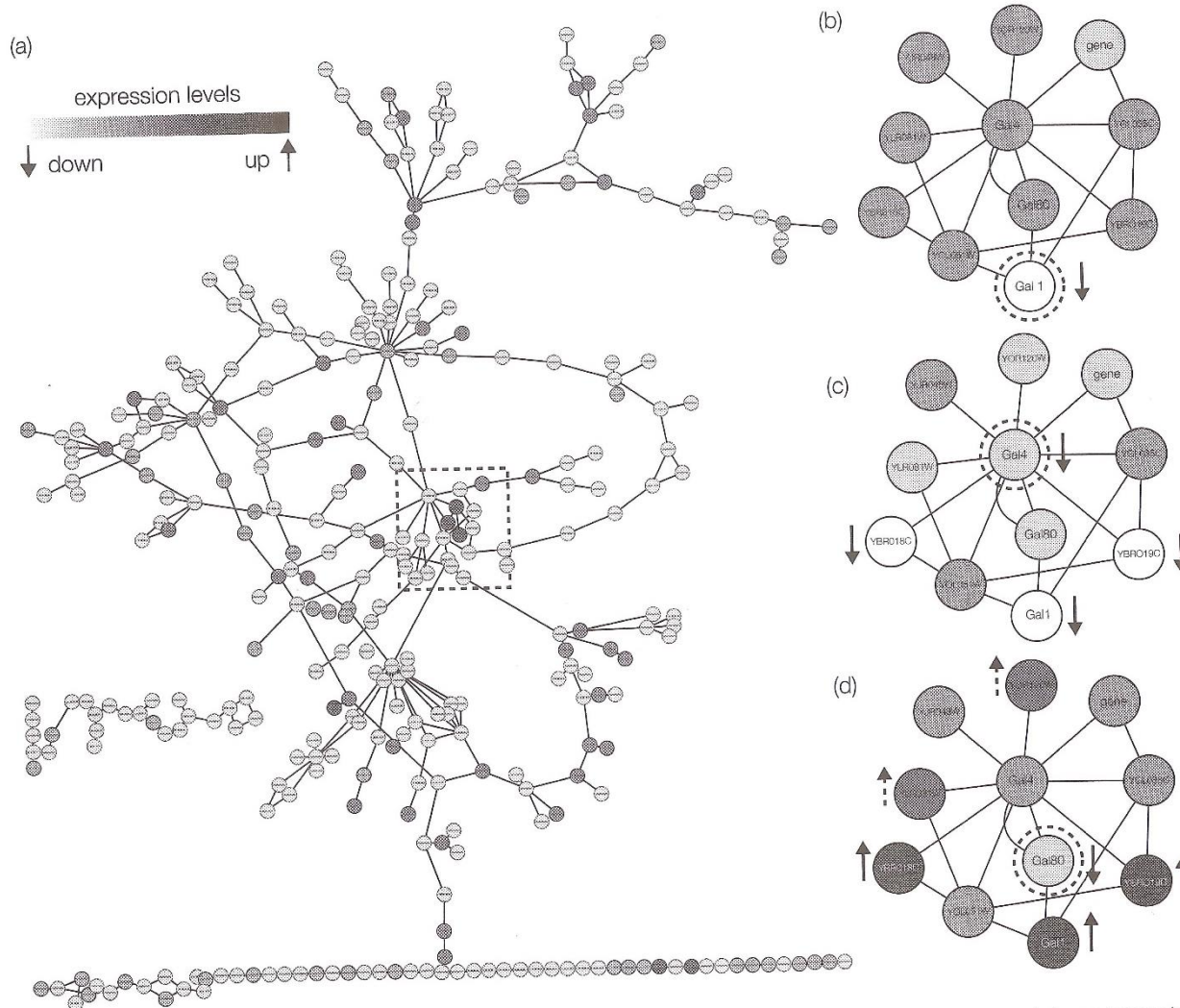


Fig. 4. This figure is based upon a yeast gene-regulatory network, where each node is a gene or protein, color-coded in the case of genes on the basis of their transcript levels, with darker shades representing higher expression levels. (a) The overall network of 331 nodes and 362 edges with key genes denoted in the black rectangle; (b), (c), and (d) show the same subnetwork when Gal1, Gal4, and Gal80 are knocked out, respectively. The dashed circle shows which gene has been knocked out, and the arrows indicate whether the result is a slight (dashed) or major (solid) increase (upward arrow) or decrease (downward arrow) in gene expression compared to the wild-type. (Permission to use the Cytoscape tutorial to produce this figure was kindly given by Trey Ideker, UCSD.)

Mapping of expression data on networks

“Cytoscape”
Online program
<http://cytoscape.org/>



GEO PROFILE, NCBI

<https://www.ncbi.nlm.nih.gov/geoprofiles/>

Genome technology

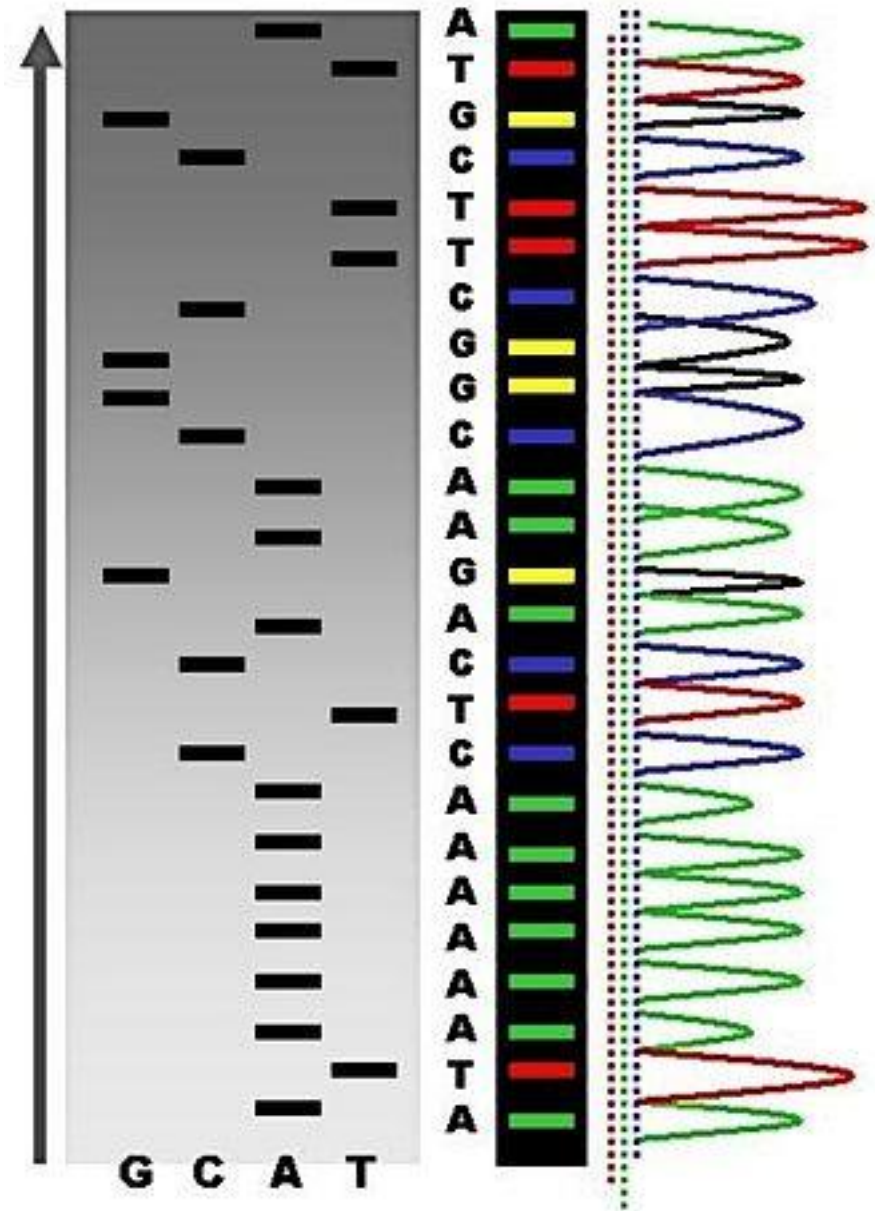
The **canonical structure of DNA** has four bases: thymine (T), adenine (A), cytosine (C), and guanine (G). Bacteriophage: hydroxy methyl or hydroxy methyl glucose cytosine. Mammalian DNA, variant bases with methyl groups or phosphosulfate may be found i.e. 5mC (5 methyl cytosine)

The chain-termination method developed by Frederick Sanger and coworkers in 1977 soon became the method of choice, owing to its relative ease and reliability:

Sanger Method

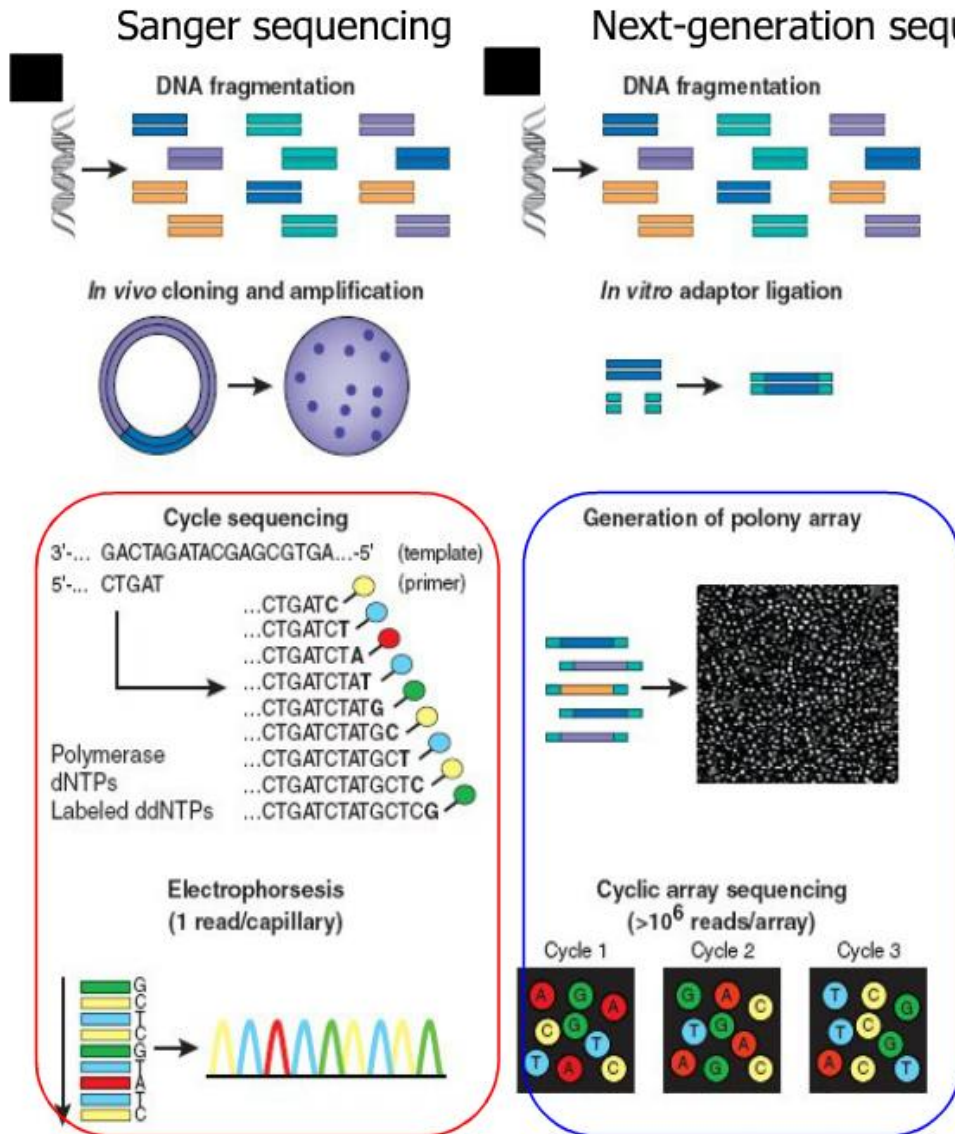
- Dominant for last ~30 years
- 1000bp longest read
- Based on primers so not good for repetitive or SNPs sites

The term "**de novo sequencing**" specifically refers to methods used to determine the sequence of DNA with no previously known sequence. Gaps in the assembled sequence may be filled by primer walking.



An example of the results of automated chain-termination DNA sequencing.

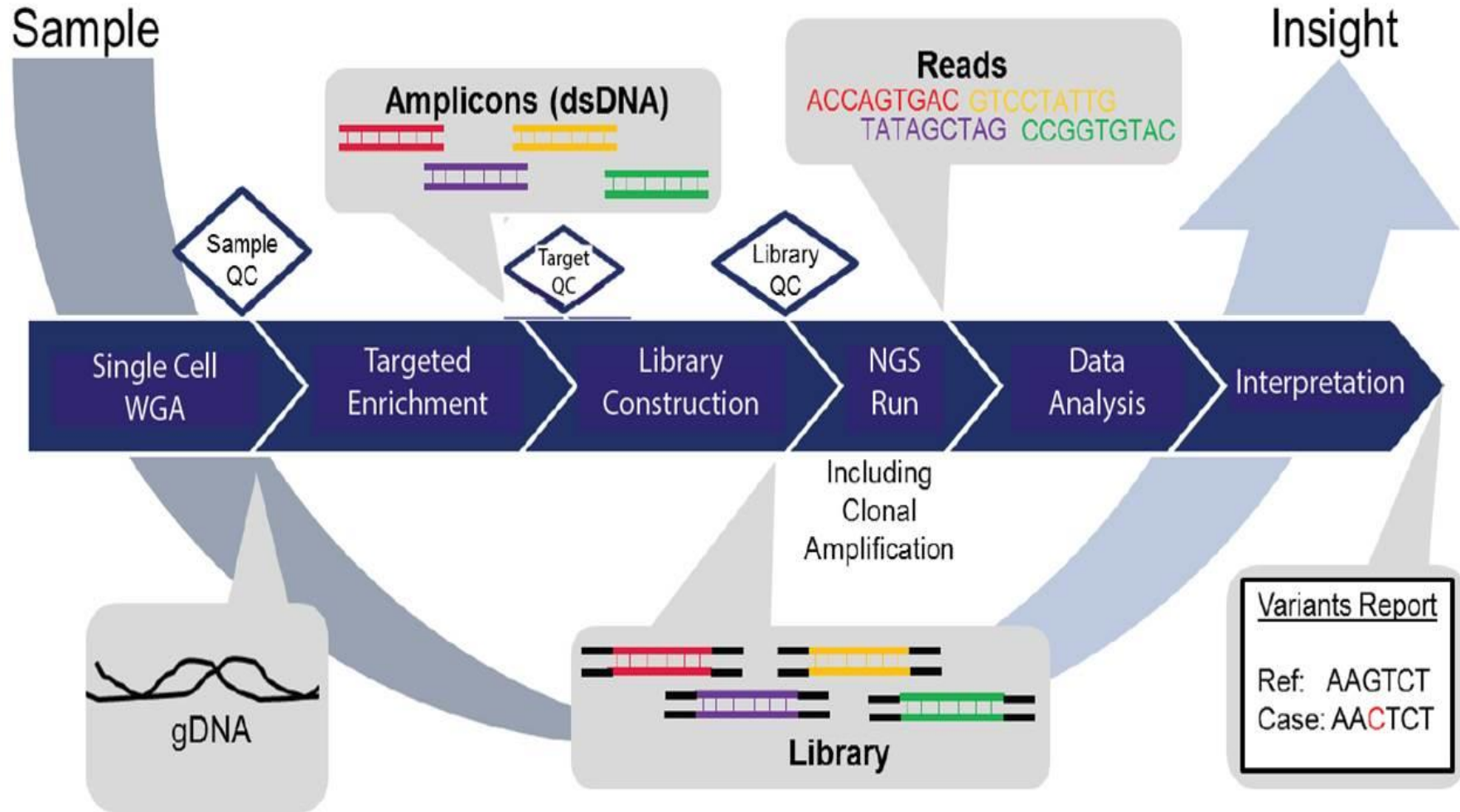
Next-generation DNA sequencing



Advantages:

- Construction of a sequencing library → clonal amplification to generate sequencing features
 - ✓ No in vivo cloning, transformation, colony picking...
- Array-based sequencing
 - ✓ Higher degree of parallelism than capillary-based sequencing

Example of NGS Work Flow

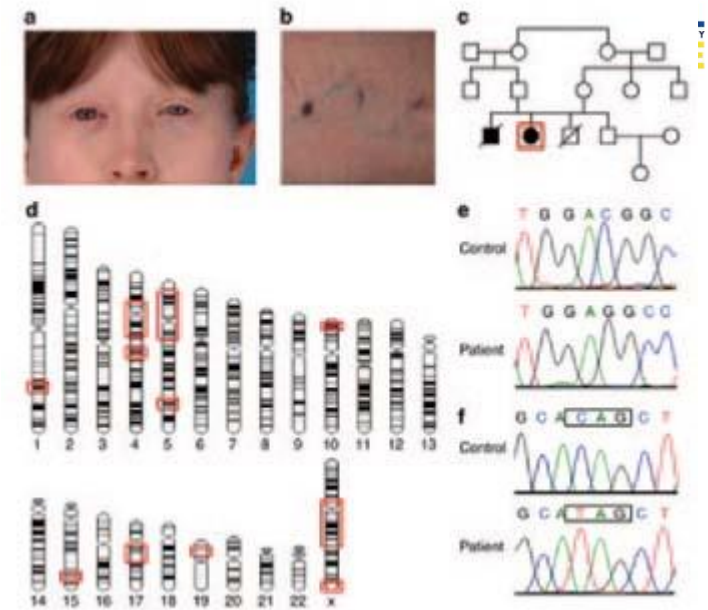


A.

Biological interpretation of human whole genome, exome, and targeted panel samples

Next Generation Sequencing

https://www.youtube.com/watch?v=MxkYa9XCvBQ&feature=player_embedded



re-sequence the genome of previously sequenced organisms (**re-sequencing**)
sequence the genomes of organisms with unknown sequences (**de novo sequencing**)
determine RNA abundance levels (**RNA-seq**)
determine protein–DNA binding regions (**ChIP-seq**)
determine protein–RNA binding sequences (**CLIP-seq**)
homozygosity mapping then **whole-exome seq**: disease-causing mutations in a patient
Targeted sequencing: more affordable, yields much higher coverage of genomic regions of interest, and reduces sequencing cost and time
in population genetic study/the status quo of integrative cancer genomic approaches
Having aligned the fragments of one or more individuals to a reference genome, '**SNP calling**' identifies variable sites, whereas '**genotype calling**' determines the genotype for each individual at each site.
and more...

Applications of Next-Generation

DNA Level

Whole genome resequencing (WGS)

- Discover the genetic variations in a genome-wide range.

Exome Seq

- Discover the causative, susceptibility loci
- Discover rare/novel variants
- More economical and efficient

Target Region Seq

- Find the novel variants or validate the candidate variants in the target regions

Genotyping

- SNP and CNV detection in a genome-wide range
- Customized array for personal usage which is more flexible
- Validation of candidate pathogenetic genes or loci in large amount of samples

Single Cell Seq

- Genetic variation research at single cell level
- Explore cancer cells evolution during tumor progression

RNA Level

Transcriptome Seq

- Comprehensive analysis of differential gene expression
- Discover novel genes
- RNA editing analysis(such as alternative splicing, cSNP, gene fusion, etc)

RNA-Seq (Quantification)

- Precise quantification of gene expression analysis that is suitable for large samples
- Discover disease-related functional genes

Small RNA Seq

- Gene expression analysis of miRNA
- Gene regulatory networks and targets study of mi RNA
- Discover disease-specific biomarkers

Non-coding RNA Seq

- Identify novel non-coding RNA
- Discover disease-specific biomarkers

Cell Line Seq

- Obtain a clear and comprehensive genetic patterns of the cell lines
- Obtain mutation information of high accuracy

Epigenetic Level

Whole Genome Bisulfite Seq (WGBS)

- DNA methylation research at whole genome-wide level
- High accuracy and high resolution(single-based)

MeDIP Seq

- Based on immunoprecipitation for methylated DNA enrichment
- Whole genome-wide DNA methylation research and cost-effective

RRBS Seq

- Methylation analysis of promoter regions with substantial genome coverage
- Based on enzyme digestion and bisulfite treatment
- Good repeatability

ChIP Seq

- Genome-wide protein-DNA interaction studies
- Higher resolution, more precise and abundant than ChIP-chip

Protein Level

Proteome Profiling

- Analyze the component of protein mixtures
- Obtain comprehensive information of protein category, metabolic pathways, etc

Quantitative Proteomics

- Fast and accurate protein differential analysis for multiple samples

Modification Proteomics

- Fast and comprehensive analysis of protein modification spectrum for multiple samples

Target Proteomics

- Based on the technology of Multiple Reaction Monitoring(MRM)
- Validate the discovered biomarkers
- Identify protein modification and low abundant proteins

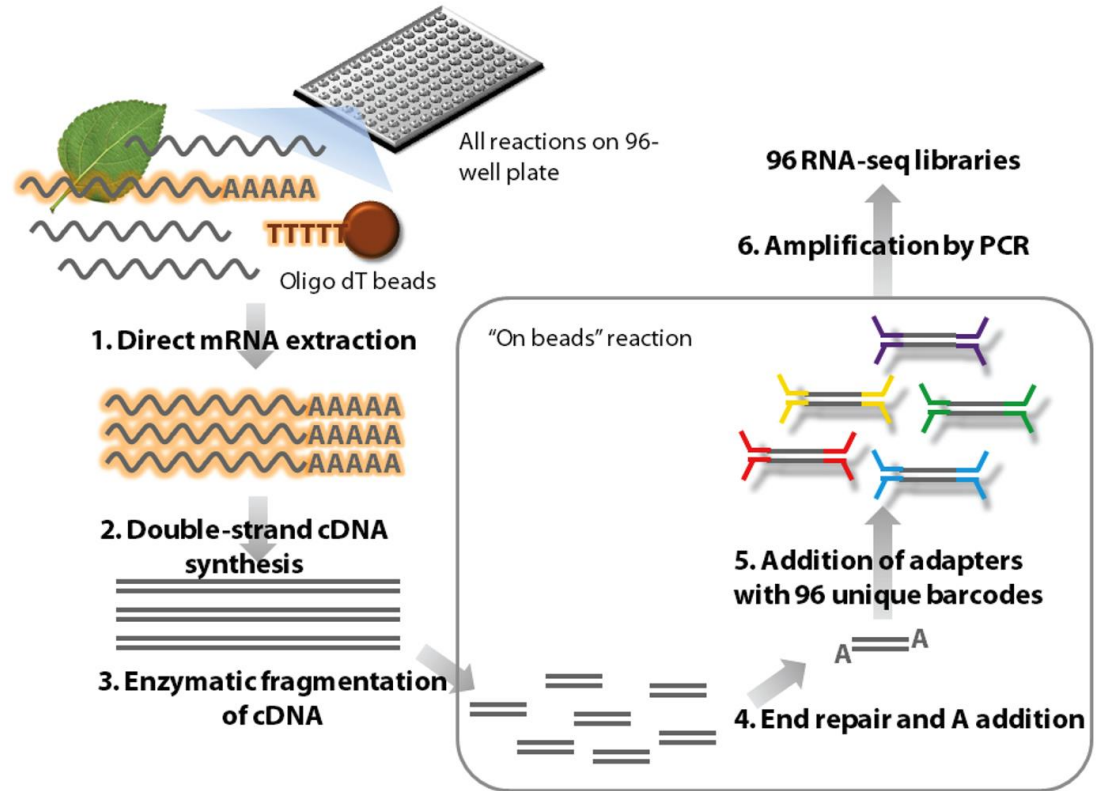
NGS by using RNA seq

Population of RNA (poly A+) converted to a library of cDNA fragments with adaptors attached to one or both ends

Solid Phase Amplification performed

Molecules sequenced from one end (Single End) or both ends (Pair End)

Reads are typically 30-400bp depending on sequence technology used



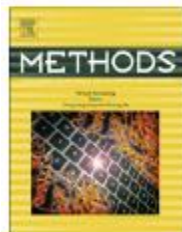
Article history:

Received 8 April 2015

Received in revised form 7 June 2015

Accepted 9 June 2015

Available online xxxx



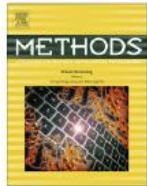
Methods. 2015 Jun 16. pii: S1046-2023(15)00254-6.

Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*.

Bischler T, Tan HS, Nieselt K, Sharma CM.

Research Center for Infectious Diseases (ZINF), University of Würzburg, Josef-Schneider-Str. 2 / Bau D15, 97080 Würzburg, Germany.

NGS by using ChIP seq



Review Article

Defining bacterial regulons using ChIP-seq

Kevin S. Myers^{a,b}, Dan M. Park^c, Nicole A. Beauchene^d, Patricia J. Kiley^{d,b,*}

Table 1

File formats used in ChIP-seq analysis.

| File type | Brief description | Use in analysis |
|--------------|--|-------------------------------------|
| FASTQ | Illumina sequencing file from experimental run | Raw ChIP-seq data |
| FASTQC | Illumina quality control file for each Illumina sequencing run | Evaluating ChIP-seq sequencing data |
| SAM | Alignment file from Bowtie2 or BWA | Aligned ChIP-seq file |
| BAM | Binary SAM file | Aligned ChIP-seq file |
| Wiggle (WIG) | File containing results of enumerating read hits at each base location | Visualization file |
| ELAND | Another alignment file format, used as input in MOSAiCS | For peak calling |

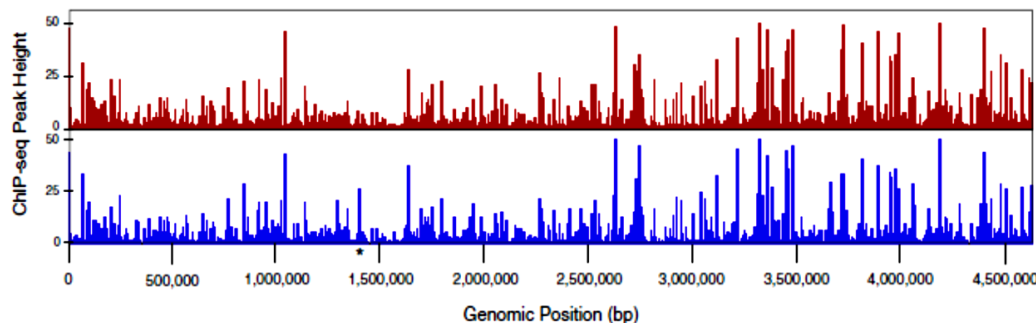
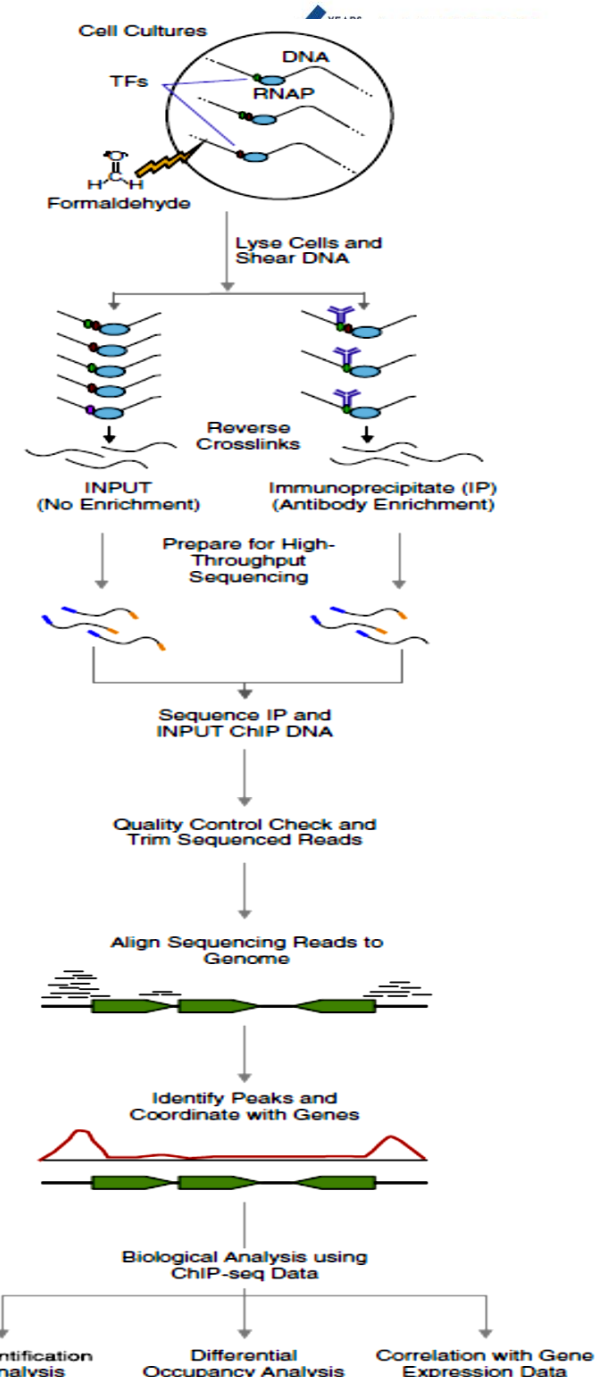
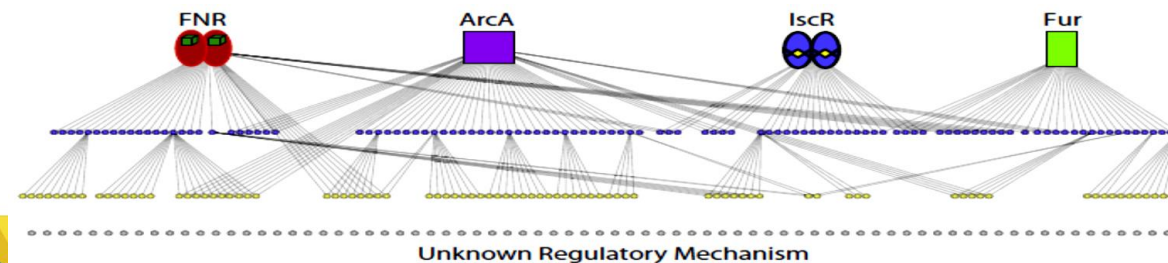


Fig. 2. Comparing ChIP-seq data between growth conditions. Shown are RNAP (σ^{70}) ChIP-seq data traces collected from cultures grown under aerobic (red) or anaerobic (blue) conditions [2]. ChIP-seq IP/INPUT ratio is shown on the y-axis and genomic position is shown on the x-axis. The asterisk indicates an example of differential binding between growth conditions. This figure was generated in the MochiView browser [53].



Confirmation of Transcriptional Control?

- Expression analysis
- Indirect control
 - Intermediate regulator
 - Sensor and response, enhancer or repressor
- Direct binding
 - Protein (Transcriptional regulator)
 - Protein-DNA binding
 - Foot-printing
 - Site-directed mutagenesis
 - Complementation
 - DNA (gene promoter)
 - Promoter analysis
 - Reporter enzyme activity assay
 - Site-directed mutagenesis

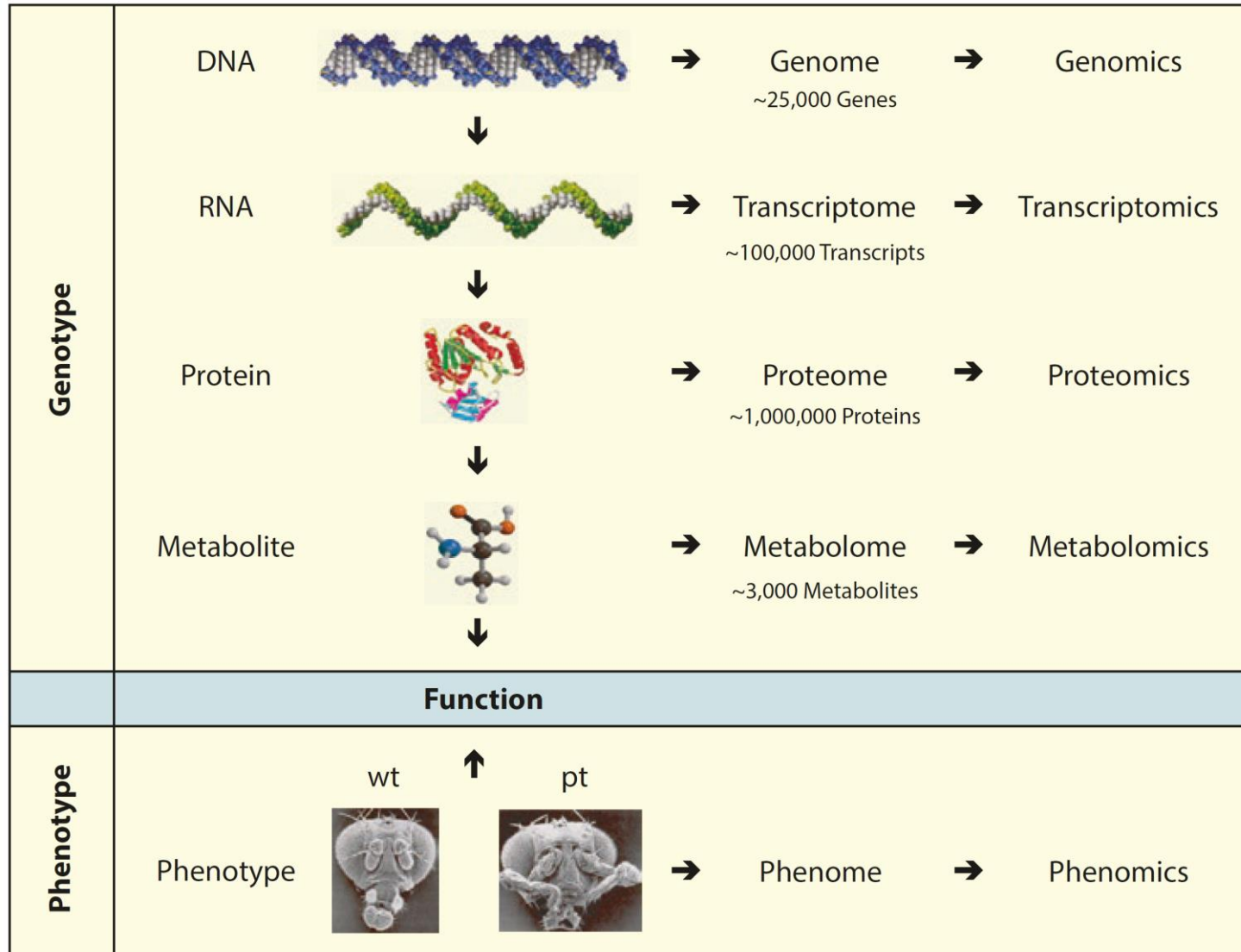
Article Data Analysis

Transcriptomics analysis

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137762>

<https://www.ncbi.nlm.nih.gov/pubmed/21749987>

The Functional Analysis of Genomes



Gene function

1. **Functional genomics** is the study of the function of genes and their products: Gene annotation, Gene ontology, and prediction of gene function
2. **Functional genomics** – analysis of genome wide gene expression and gene functions

DNA microarrays (“gene chips”) enable the analysis of gene expression at the whole-genome level;

- DNA fragments are deposited on a slide
- Probed with labeled mRNA from different sources
- Active/inactive genes are identified

Next generation Sequencing:

- “RNAseq” for Transcriptomic study
- “ChIPseq” for identifying the regulator’s targets

Genome Annotation

- Ab initio, i.e. based on sequence alone
 - INFERNAL/rFAM (RNA genes), miRBase (miRNAs), RepeatMasker (repeat families), many gene prediction algorithms (e.g. AUGUSTUS, Glimmer, GeneMark, ...)
 - Evidence-based
 - Require transcriptome data for the target organism (the more the better)
 - Align cDNA sequences to assembled genome and generate gene models: TopHat/Cufflinks, Scripture
1. Sequence
 2. Gene structures (GenScan, FgenesH)
 3. Predictions verified by BLAST against sequence database, cDNA and EST (GeneWise, Spidey, SIM4, EST2Genome)
 4. Manually verified by human curators
 5. Functional assignment of proteins by BLAST searches of protein database
 6. Further functional description from Pfam and InterPro and literature
-
- BLAST of gene models against protein databases
 - Sequence similarity to known proteins
 - InterProScan of predicted proteins against databases of protein domains (Pfam, Prosite, HAMAP, PANTHER, ...)
 - Mapping against Gene Ontology terms (BLAST2GO)

DATABASES and DATA Sources

- DNA sequencing
 - Sanger method, chain termination sequencing (dideoxy)
 - Shotgun sequencing and clone contig approach
 - Pyrosequencing (immobilized on beads)
 - Next generation sequencing
- RNA sequencing
 - Not well due to minor nucleotides and RNA editing
 - Next generation sequencing
- Protein sequencing
 - Edman degradation with labelled terminal residues
 - Mass spectrometry (MS, m/z ratio): Soft ionization methods (without degradation)
 - ESI (electrospray ionization)
 - MALDI (Matrix assisted laser desorption/ionization)

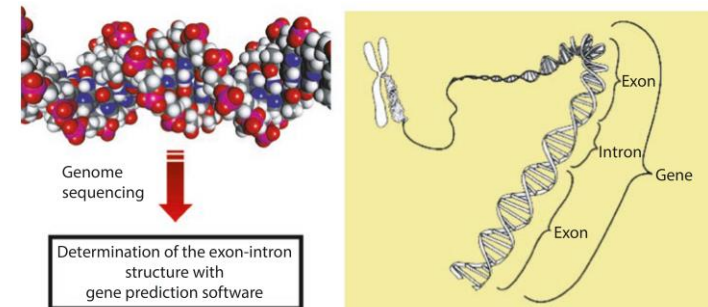
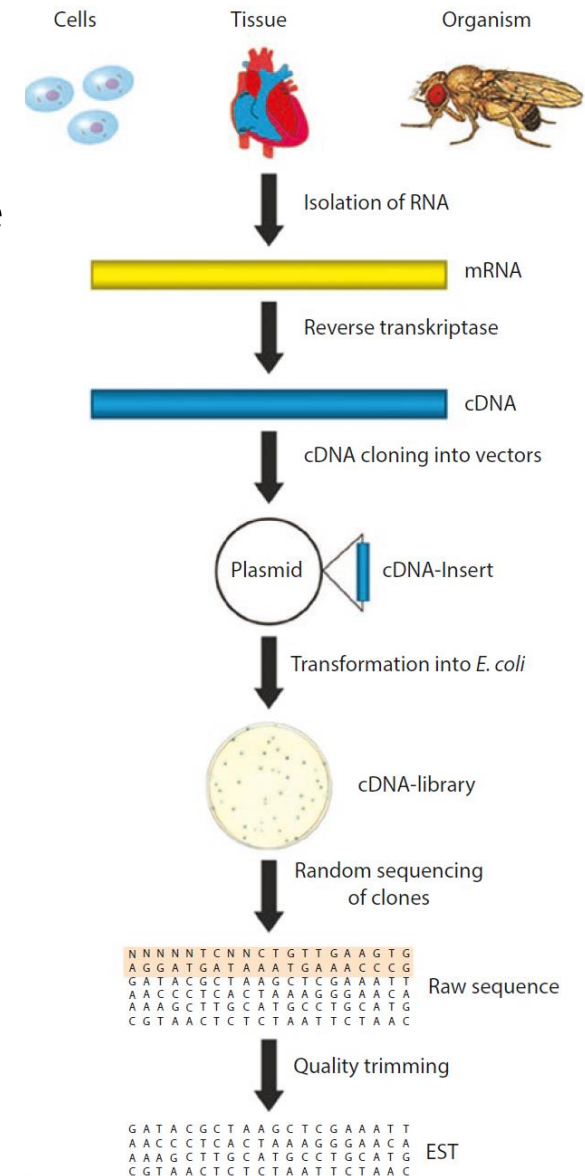
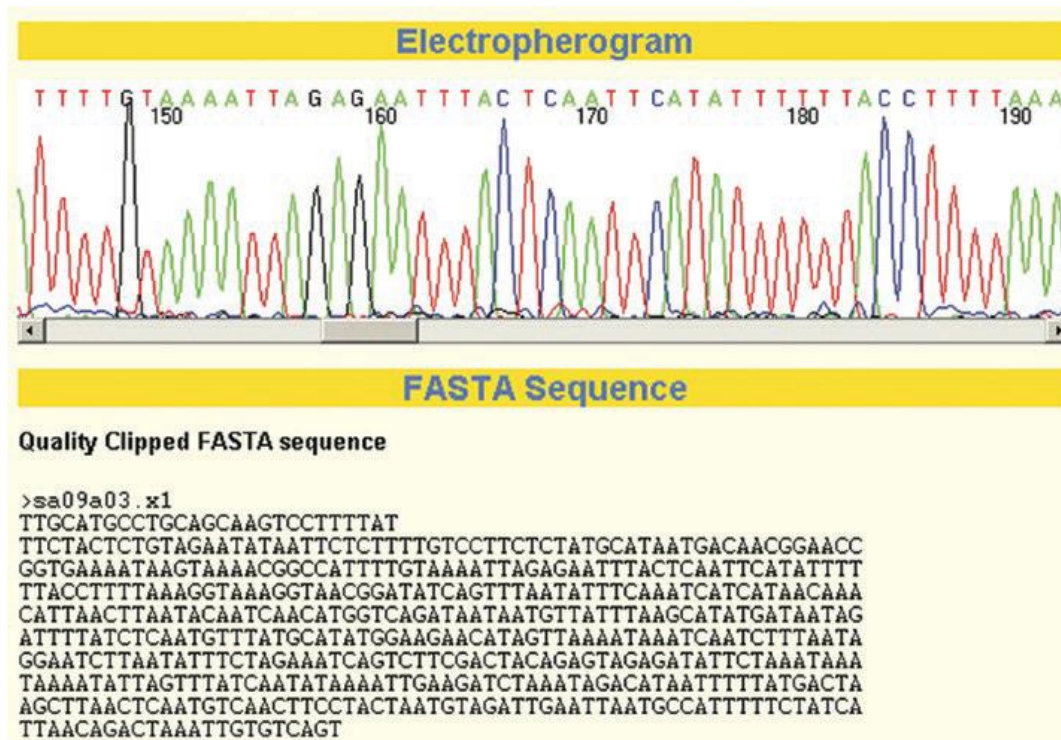


Fig. 3.9 Identification of new genes and proteins by genome sequencing

Expressed Sequence Tags

- partial sequences of cDNA clones could also be used in the discovery of new genes (Adams et al. 1991). Because cDNA clones are derived from expressed genes, the sequences were called expressed sequence tags (ESTs). ESTs are generated by the end-sequencing of cDNAs



Pharmacogenetics (or pharmacogenomics)

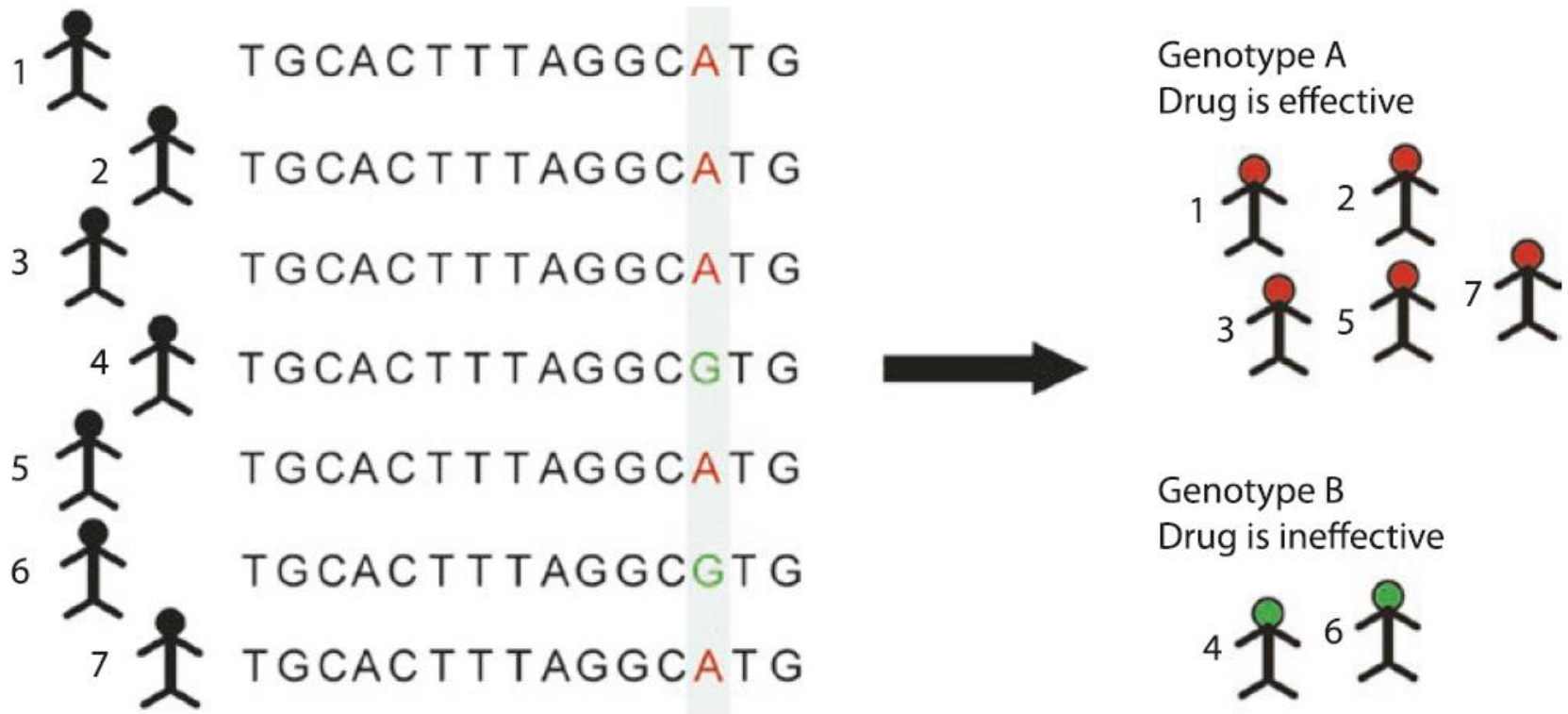


Fig. 4.9 Genotyping of patients by detecting SNPs

Personalized Medicine and Biomarkers

Next-Generation Sequencing (NGS)

Proteogenomics

Further Reading

cap. <http://doua.prabi.fr/software/cap3>
dbest. <https://www.ncbi.nlm.nih.gov/dbEST/>
dbgss. <https://www.ncbi.nlm.nih.gov/dbGSS/>
db SNP. <https://www.ncbi.nlm.nih.gov/SNP/>
dbSTS. <https://www.ncbi.nlm.nih.gov/dbSTS/>
ebi-gwas. <http://www.ebi.ac.uk/gwas/>
grailexp. <http://compbio.ornl.gov/grailexp/>
gwas. <http://www.gwascentral.org/>
helix-nebula. <http://www.helix-nebula.eu/usecases/embl-use-case>
homologene. <http://www.ncbi.nlm.nih.gov/homologene/>
hts-mapper. http://www.ebi.ac.uk/~nf/hts_mappers/
humatrix. <https://www.humatrix.de/>
image. <http://imageconsortium.org/>
nematode. <http://www.nematode.net/>
ngs-knowledge-base. <https://goo.gl/HlaY1W>
ngs-movie. <https://www.youtube.com/watch?v=jFCD8Q6qSTM>

DATABASES and DATA Sources

- Primary sequence databases (raw sequence data)
 - GenBank by National Center for Biotechnology Information (NCBI)
 - Nucleotide sequence database by European Molecular Biology Laboratory (EMBL)
 - Nucleotide sequence database by DNA Databank of Japan (DDBJ)
- Subsidiary sequencing databases
 - dbEST is a part of GenBank (EST = Expressed sequence tag)
 - dbGSS: GSS = genome survey sequences, single-pass genomic sequences
 - dbSTS : STS = sequence tagged sites (unique sequence for physical markers)
 - HTG = high-throughput genomic, unfinished genomic sequence data
- Protein sequence databases
 - SWISS-PROT (features table & sequence) + TrEMBL (Translated EMBL) = UniPort
 - UniPort (Universal protein resources): function, classification, and cross references
 1. UniRef combined similar sequences together in a single records
 2. UniParc keep a record of the history of the sequences
 3. UniMes a record of metagenomics and environmental data

TABLE 5.1 Three-Letter Abbreviations of GenBank Divisions

| | | |
|----|-----|---|
| 1 | PRI | Primate sequences |
| 2 | ROD | Rodent sequences |
| 3 | MAM | Other mammalian sequences |
| 4 | VRT | Other vertebrate sequences |
| 5 | INV | Invertebrate sequences |
| 6 | PLN | Plant, fungal, and algal sequences |
| 7 | BCT | Bacterial sequences |
| 8 | VRL | Viral sequences |
| 9 | PHG | Bacteriophage sequences |
| 10 | SYN | Synthetic sequences |
| 11 | UNA | Unannotated sequences |
| 12 | EST | Expressed sequence tag sequences |
| 13 | PAT | Patent sequences |
| 14 | STS | Sequence tagged sites sequences |
| 15 | GSS | Genome survey sequences |
| 16 | HTG | High-throughput genomic sequences |
| 17 | HTC | Unfinished high-throughput cDNA sequences |
| 18 | ENV | Environmental sampling sequences |

5.4.5 Sequence Accession Numbers and Redundancy in Primary Databases

Nucleotide: 1 letter + 5 numerals (e.g. J00750)
or 2 letters + 6 numerals (e.g. AF208545)

Protein: 3 letters + 5 numerals (e.g. AAG60350, CAB92299).

SWISS-PROT features table is started with 2 letters as means as:

| | |
|----|------------------------|
| ID | identity |
| AC | accession number |
| DT | date |
| DE | description |
| GN | gene name |
| CC | comment |
| | ! as continuation line |
| DR | reference |
| KY | key words |
| FT | features |

DATABASES and DATA Sources

- Organism-specific resources

Table 1. A small selection of organism-specific genomic databases available on the WWW. These databases are curated actively by members of the research community working on the particular organism of interest, and generally include links to organism-specific resources such as clone sets and mutant strains

| Organism | Database/Resource | URL |
|---------------------------------|--|---|
| <i>Escherichia coli</i> | EcoGene | http://ecogene.org/ |
| | EcoCyc (Encyclopedia of <i>E. coli</i> genes and metabolism) | http://www.ecocyc.org/ |
| | Colibri | http://genolist.pasteur.fr/Colibri/ |
| <i>Bacillus subtilis</i> | SubtiList | http://genolist.pasteur.fr/SubtiList/ |
| <i>Saccharomyces cerevisiae</i> | <i>Saccharomyces</i> Genome Database (SGD) | http://genome-www.stanford.edu/Saccharomyces/ |
| <i>Plasmodium falciparum</i> | PlasmoDB | http://PlasmoDB.org |
| <i>Arabidopsis thaliana</i> | MIPS <i>Arabidopsis thaliana</i> Database (MAtdB) | http://mips.gsf.de/proj/thal/db |
| | The <i>Arabidopsis</i> Information Resource (TAIR) | http://www.arabidopsis.org/ |
| <i>Drosophila melanogaster</i> | FlyBase | http://flybase.bio.indiana.edu/ |
| <i>Caenorhabditis elegans</i> | <i>A. C. elegans</i> DataBase (ACeDB) | http://www.acedb.org/ |
| Mouse | Mouse Genome Database (MGD) | http://www.informatics.jax.org/ |
| Human | OnLine Mendelian Inheritance in Man (OMIM) | http://www.ncbi.nlm.nih.gov/omim |

DATABASES and DATA Sources

Table 2. Useful gateway sites providing information and links to multiple, organism-specific, and genomic resources

| Gateway site | URL |
|--|---|
| NCBI Genomic Biology | http://www.ncbi.nlm.nih.gov/Genomes/index.html |
| GOLD (Genomes OnLine Database) | http://www.genomesonline.org/ |
| TIGR Microbial Database | http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi |
| Bacterial genomes | http://genolist.pasteur.fr/ |
| Yeast databases | http://genome-www.stanford.edu/Saccharomyces/yeast_info.html |
| Ensembl Genome Database Project | http://www.ensembl.org/ |
| MIPS (Munich Information Center for Protein Sequences) | http://mips.gsf.de |

Table 3. Database tools for displaying and annotating genomic sequence data

| Viewer format | URL for further information and tutorials |
|-----------------|---|
| Artemis | http://www.sanger.ac.uk/Software/Artemis |
| ACeDB | http://www.acedb.org/Tutorial/brief-tutorial.shtml |
| Apollo | http://apollo.berkeleybop.org/current/install.html |
| Ensembl | http://www.ensembl.org |
| NCBI map viewer | http://www.ncbi.nlm.nih.gov/mapview/ |
| GoldenPath | http://genome.ucsc.edu/ |

NCBI database: “Entrez”

Entrez is the common front-end to all the databases maintained by the NCBI and is an extremely easy system to use. The Entrez main page, as with all NCBI pages, is undemanding in its browser requirements and downloads quickly. Part of the front page is illustrated in *Fig. 8*. The databases available for searching can be accessed by hyperlinks, or by using the search box as shown. The search term

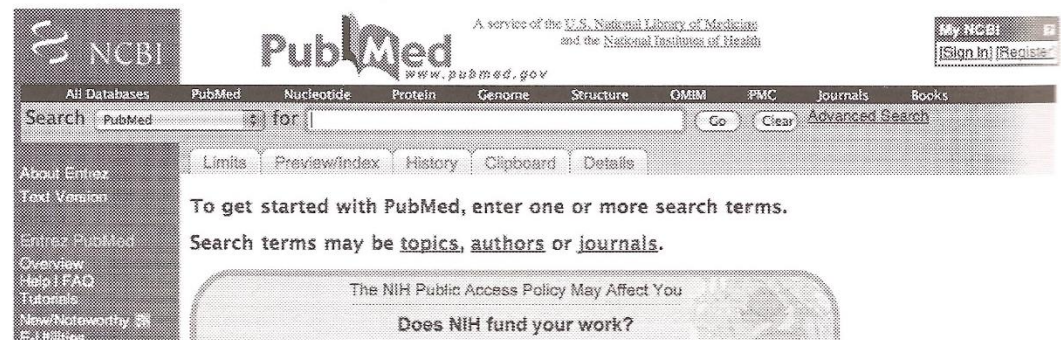


Table 4. The databases covered by Entrez, listed by category

| Category | Database |
|------------------------|--|
| Nucleic acid sequences | Entrez nucleotides: sequences obtained from Genbank, RefSeq, and PDB. Also UniGene, PopSet, Probe, Trace Archive, PA, UniST, dbEST, dbGSS, dbSNP, dbST, HomoloGene, and MGC |
| Protein sequences | Entrez protein: sequences obtained from SWISS-PROT, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq. Also 3D domains, Protein Clusters, and PROW |
| 3D structures | Entrez Molecular Modeling Database (MMDB). Also 3D domains |
| Genomes | Complete genome assemblies from many sources |
| OMIM | OnLine Mendelian Inheritance in Man |
| Taxonomy | NCBI Taxonomy Database |
| Books | Bookshelf |
| Expression databases | Gene Expression Omnibus (GEO), SAGE |
| Literature | PubMed |

DATABASES and DATA Sources

DBGET/LinkDB

DBGET is an integrated data retrieval system developed and jointly maintained by the Institute for Chemical Research (Kyoto University) and the Human Genome Center (University of Tokyo). It is integrated with more than 30 databases (*Table 5*), which can be searched one at a time or in combination. Hits are presented as a list of results together with any available associated information. **LinkDB** is an associated database of links (**binary relationships**) between entries in the different databases available to DBGET and also further organism-specific databases, such as AceDB, Flybase, and SGD. DBGET is associated closely with KEGG, the Kyoto Encyclopedia of Genes and Genomes, which is maintained by the same group.

Table 5. The databases covered by DBGET/LinkDB, listed by category

| Category | Database |
|---------------------------------|---|
| Nucleic acid sequences | GenBank, EMBL, RefSeq |
| Protein sequences | SWISS-PROT, PIR, PRF, PDBSTR, UniProt |
| 3D structures | PDB |
| Sequence motifs | PROSITE, EPD, TRANSFAC, BLOCKS, PRODOM, PRINTS, PFAM |
| Enzyme reactions | LIGAND |
| Metabolic pathways | KEGG |
| Amino acid mutations | PMD |
| Amino acid indices | AAindex |
| Genetic diseases | OMIM |
| Literature | LITDB Medline |
| Organism-specific gene catalogs | <i>E. coli</i> , <i>H.influenzae</i> , <i>M.genitalium</i> , <i>M.pneumoniae</i> , <i>M.jannaschii</i> , <i>Synechocystis</i> , <i>S.cerevisiae</i> |

DATABASES and DATA Sources

Table 6. The databases covered by the SRS at <http://srs6.ebi.ac.uk>, listed by category

| SRS description | Examples |
|---|---|
| Literature | MEDLINE TAXONOMY, OMIM |
| Nucleotide sequence databases | EMBL, RefSeq, |
| Nucleotide sequence related | TFSITE, TFFACTOR, REBASE |
| Protein sequence databases | UNIPROT, REFSEQ, IPI |
| Protein function, structure and interaction databases | INTERPRO, PRODOM, PRINTS, BLOCKS, PFAMHMMFS, PROSITE, PDB, FSSP, EXPERIMENT, INTERACTION, INTERACTOR |
| TransFac | TFSITE, TFFACTOR, TFCELL, TFCLASS, TFMATRIX ,TFGENE |
| Enzymes reactions and metabolic pathways | ENZYME, UPATHWAYM UENZYME, UREACTION, UPATHWAY |
| Mutation and SNP databases | HGVBASE |
| User-owned databanks | USERDNA, USERPROTEIN |
| Application results | FASTA, FASTX, FASTY, NFASTA, BLASTP, BLASTN, CLUSTALW, NCLUSTALW, PPSEARCH, RESTRICTIONMAP, PSIBLAST, HMMPFAM |
| EMBOSS result databases | Including: ANTIGENIC, BACKTRANSEQ, BIOSEDN, RESTRICT, MERGER, ETC. |



Primary Databases

Nucleotide Sequence Databases

.1 GenBank

GENBANK

```

LOCUS      SCU49845                      5028 bp    DNA      linear    PLN 14-JUL-2016
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.

ACCESSION  U49845
VERSION    U49845.1  GI:1293613
KEYWORDS   .
SOURCE     Saccharomyces cerevisiae (baker's yeast)
            ORGANISM  Saccharomyces cerevisiae
                        Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
                        Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
                        Saccharomyces.

REFERENCE  1 (bases 1 to 5028)
AUTHORS    Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE      Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein

[.]

FEATURES             Location/Qualifiers
     source            1..5028
                       /organism="Saccharomyces cerevisiae"
                       /mol_type="genomic DNA"
                       /db_xref="taxon:4932"
                       /chromosome="IX"

     mRNA              <1..>206
                       /product="TCP1-beta"

     CDS               <1..206
                       /codon_start=3
                       /product="TCP1-beta"
                       /protein_id="AAA98665.1"
                       /db_xref="GI:1293614"
                       /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
AEVLLRVDNIIRARPRTANRQHM"

[.]

ORIGIN
      1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
     61 ccgacatgag acagtttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
  
```

■ Fig. 2.1 Database record of GenBank database. The entry was shortened at some points, as indicated by [...]

ENTREZ

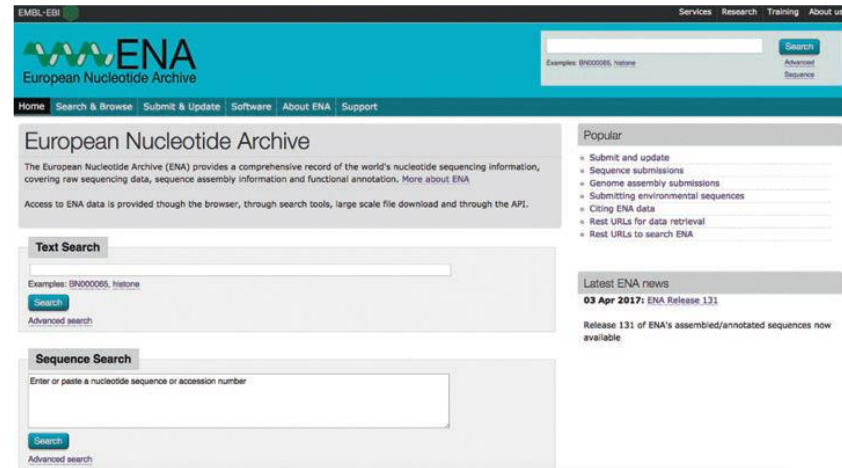
- Query of the GenBank database is carried out via the NCBI Entrez system [entrez], which is used to query all NCBI-associated databases (NCBI Resource Coordinators 2016).

Table 2.1 Field IDs to restrict search terms to certain database fields in the Entrez system

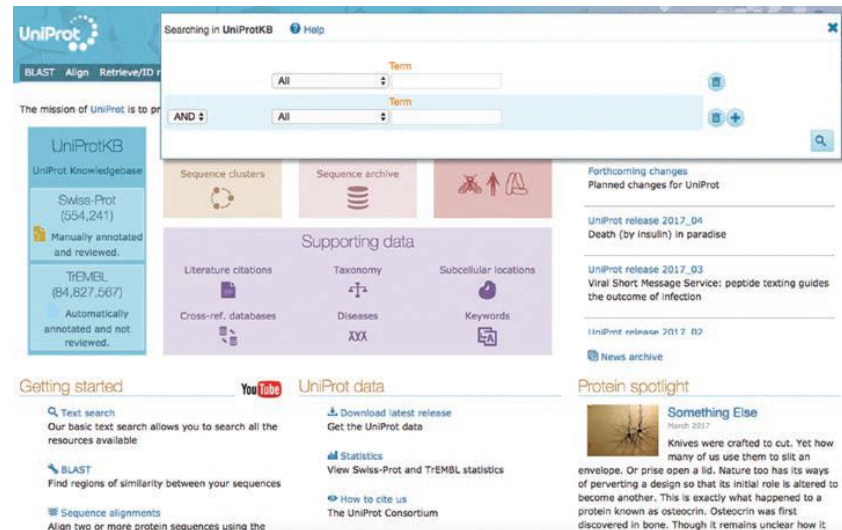
| Field ID | Database field |
|----------|---|
| ACC | Accession number |
| AU | Author name |
| DP | Publication date |
| GENES | Gene name |
| ORGN | Scientific and common name of the organism |
| PT | Publication type, e.g., review, letter, technical publication |
| TA | Journal name, official abbreviation, or ISSN number |

Other Primary Databases

- EMBL and DDBJ
- ENA Online Retrieval



- Protein sequence databases
 - UniProt
 - NCBI Protein Database



Secondary Databases

- Prosite
- PRINTS
- Pfam
- Interpro
- Genotype-Phenotype Databases
 - PhenomicDB
 - Molecular Structure Databases
 - Protein Data Bank
 - CATH
 - PubChem

 proSite

Entry: PS01159

General information about the entry

| | |
|---------------------|--|
| Entry name [info] | WW_DOMAIN_1 |
| Accession [info] | PS01159 |
| Entry type [info] | PATTERN |
| Date [info] | 01-NOV-1995 CREATED; 01-DEC-2004 DATA UPDATE; 12-APR-2017 INFO UPDATE. |
| PROSITE Doc. [info] | PD0050020 |

Name and characterization of the entry

| | |
|--------------------|---|
| Description [info] | WW/rsp5/WWP domain signature. |
| Pattern [info] | M-k(9,11)-[VFY]-[FYV]-k(6,7)-[GSTNE]-[GSTQCR]-[FYV]-[K]-[SA]-P. |

Numerical results [info]


Numerical results for UniProtKB/Swiss-Prot release 2017_04 which contains 554241 sequence entries.

| | |
|---|--------------------------------|
| Total number of hits | 327 in 227 different sequences |
| Number of true positive hits | 275 in 175 different sequences |
| Number of 'unknown' hits | 0 |
| Number of false positive hits | 52 in 52 different sequences |
| Number of false negative sequences | 56 |
| Number of 'partial' sequences | 0 |
| Precision (true positives / (true positives + false positives)) | 84.10 % |
| Recall (true positives / (true positives + false negatives)) | 83.08 % |

Comments [info]

| | |
|--------------------------------------|------------|
| Taxonomic range [info] | Eukaryotes |
| Maximum number of repetitions [info] | 4 |

RCDB PDB Deposit · Search · Visualize · Analyse · Download · Learn · More · [MyPDB Login](#)

 **PDB**
PROTEIN DATA BANK

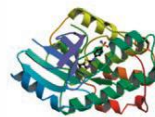
An Information Portal to
129387 Biological
Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands

Advanced Search | Browse by Annotations

Structure Summary 3D View Annotations Sequence Sequence Similarity Structure Similarity Experiment Literature

Biological Assembly 1



2BTS
STRUCTURE OF CDK2 COMPLEXED WITH PNU-230032
DOI: 10.2210/pdb2bts/pdb

Classification: TRANSFERASE
Deposited: 2005-05-06 Released: 2005-11-09
Deposition author(s): Valpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Peverello, R.
Organism: Homo sapiens
Expression System: TRICHOPUSIA NI
Structural Biology Knowledgebase: 2BTS (v24 annotations) [Annotations](#)

Experimental Data Snapshot
Method: X-RAY DIFFRACTION
Resolution: 1.99 Å
R-Value Free: 0.261
R-Value Work: 0.214

wwPDB Validation

| Metric | Percentile Ranks | Value |
|-----------------------|------------------|-------|
| R-factor | 99.9 | 0.268 |
| Clashscore | 99.9 | 0.0 |
| Ramachandran outliers | 99.9 | 0.0% |
| Solvent accessibility | 99.9 | 4.9% |
| RSCC outliers | 99.9 | 0.0% |

[Literature](#)

Structure-Based Drug Design to the Discovery of New 2-Aminothiazole Cdk2 Inhibitors.
Valpetti, A., Casale, E., Roletto, F., Amici, R., Villa, M., Peverello, R.
(2008) J Mol Graph Model 24: 341
PubMed (12360160) [Open in PubMed](#)
DOI: 10.1016/j.jmg.2008.09.012
Primary Citation of Related Structures: 2BTS 2BTS

PubMed Abstract:
N-(5-Bromo-1,3-thiazol-2-yl)butanamide (compound 1) was found active (IC50=608 nM) in a high throughput screening (HTS) for CDK2 inhibitors. By exploiting crystal structures of several complexes between CDK2 and inhibitors and applying structure-based drug design (SBDD), we rapidly discovered a very potent

Further Reading

bankit.

<http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank>

cath. <http://www.cathdb.info/>

dbgap. <http://www.ncbi.nlm.nih.gov/gap>

ddbj. <http://www.ddbj.nig.ac.jp/>

ebi. <http://www.ebi.ac.uk/>

ebi-manual.

[http://www.ebi.ac.uk/embl/Documentation/User_manual/
usrman.html](http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html)

ena. <http://www.ebi.ac.uk/ena/>

entrez. <http://www.ncbi.nlm.nih.gov/nucleotide>

entrez-help.

[http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/
helpdoc.html](http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html)

expasy. <http://www.expasy.org/>

flybase. <http://www.flybase.org/>

gb-sample.

<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

genbank. <http://www.ncbi.nlm.nih.gov/Genbank/>

homologene. <http://www.ncbi.nlm.nih.gov/homologene>

interpro. <http://www.ebi.ac.uk/interpro/>

mgd. <http://www.informatics.jax.org/>

nar. <http://nar.oxfordjournals.org/>

ncbi. <http://www.ncbi.nlm.nih.gov/>

nig. <https://www.nig.ac.jp/nig/>

omia. <http://omia.angis.org.au/home/>

omim.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

pdb. <http://www.rcsb.org/pdb/home/home.do>

pdb-models.

<http://www.rcsb.org/pdb/search/searchModels.do>

pfam. <http://pfam.xfam.org/>

phenomicdb. <http://www.phenomicdb.de/>

pir.

http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml

prints. <http://bioinf.man.ac.uk/dbbrowser/PRINTS/>

prosite. <http://prosite.expasy.org/>

prosite-manual. <http://prosite.expasy.org/prosuser.html>

pubchem. <http://pubchem.ncbi.nlm.nih.gov/>

scop. <http://scop.mrc-lmb.cam.ac.uk/scop/>

scop2. <http://scop2.mrc-lmb.cam.ac.uk/>

sequin. <http://www.ncbi.nlm.nih.gov/Sequin/>

swissprot. <http://www.expasy.org/sprot/>

tigr. <http://maize.jcvi.org/>

uniprot. <http://www.uniprot.org/>

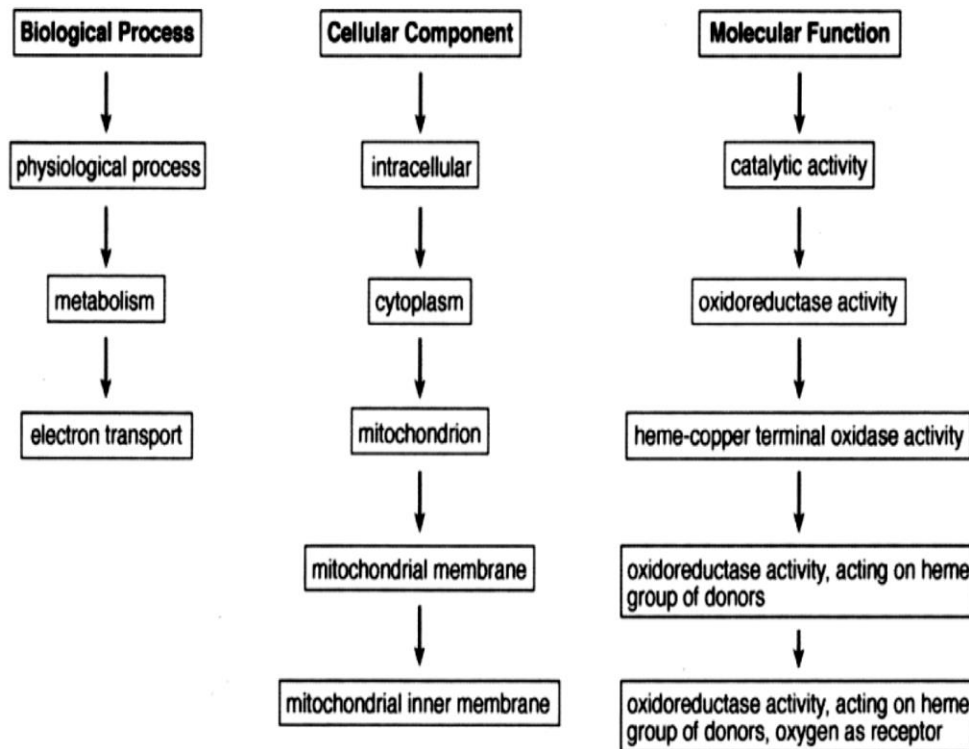
wormbase. <http://www.wormbase.org/>

wwpdb. <http://www.wwpdb.org/>

Gene Ontology

- Uses limited vocabulary to describe: cellular components, biological processes, and molecular functions
- Vocabulary arranged in a hierarchical manner from widest to most specific description

Cytochrome c oxidase



GO: "cytochrome c oxidase gene" in Ensembl

biological_process [GO:0008150] [607 gene\(s\)](#)
 cellular process [GO:0009987] [14 gene\(s\)](#)
 cellular metabolic process [GO:0044237] [34 gene\(s\)](#)
 generation of precursor metabolites and energy [GO:0006091] [61 gene\(s\)](#)
 electron transport chain [GO:0022900] [3 gene\(s\)](#)
 energy derivation by oxidation of organic compounds [GO:0015980] [3 gene\(s\)](#)
 cellular respiration [GO:0045333] [9 gene\(s\)](#)
 energy reserve metabolic process [GO:0006112] [7 gene\(s\)](#)
 glycogen metabolic process [GO:0005977] [23 gene\(s\)](#)
 fermentation [GO:0006113] [9 gene\(s\)](#)
 acetate fermentation [GO:0019654] [3 gene\(s\)](#)
 anaerobic amino acid catabolic process [GO:0019665] [3 gene\(s\)](#)
 glucose catabolic process to butyrate [GO:0030645] [3 gene\(s\)](#)
 glucose catabolic process to lactate [GO:0019659] [3 gene\(s\)](#)
 glycerol biosynthetic process [GO:0006114] [3 gene\(s\)](#)
 glycolytic fermentation [GO:0019660] [3 gene\(s\)](#)
 malolactic fermentation [GO:0043464] [3 gene\(s\)](#)
 nitrogenous compound catabolic process [GO:0019666] [3 gene\(s\)](#)
 non-glycolytic fermentation [GO:0019662] [3 gene\(s\)](#)
 regulation of fermentation [GO:0043465] [3 gene\(s\)](#)
 energy derivation by oxidation of reduced inorganic compounds [GO:0015975] [3 gene\(s\)](#)
 aerobic respiration, using ammonium as electron donor [GO:0019409] [3 gene\(s\)](#)
 aerobic respiration, using arsenite as electron donor [GO:0043554] [3 gene\(s\)](#)
 aerobic respiration, using carbon monoxide as electron donor [GO:0019410] [3 gene\(s\)](#)
 aerobic respiration, using ferrous ions as electron donor [GO:0019411] [3 gene\(s\)](#)
 aerobic respiration, using hydrogen as electron donor [GO:0019412] [3 gene\(s\)](#)
 aerobic respiration, using nitrite as electron donor [GO:0019332] [3 gene\(s\)](#)
 aerobic respiration, using sulfur or sulfate as electron donor [GO:0019414] [3 gene\(s\)](#)
 anaerobic respiration, using ammonium as electron donor [GO:0019331] [3 gene\(s\)](#)
 •
 •

DATABASES and DATA Sources

Table

Algorithm

blastp

blastn

blastx

tblastn

tblastx

Further Reading

bioedit. <http://www.mbio.ncsu.edu/bioedit/bioedit.html>

blast. <https://blast.ncbi.nlm.nih.gov>

clustalomega. <http://www.ebi.ac.uk/Tools/msa/clustalo/>

ddbj-blast. <http://ddbj.nig.ac.jp/blast/blastn?lang=en>

embnet. <http://www.embnet.org/>

embl-blast. <https://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html>

emboss. <http://emboss.sourceforge.net/>

expasy. <https://www.expasy.org/>

genscan. <http://genes.mit.edu/GENSCAN.html>

glimmer. <http://ccb.jhu.edu/software/glimmer/index.shtml>

ncbi. <http://www.ncbi.nlm.nih.gov/>

ncbi-blast. <http://www.ncbi.nlm.nih.gov/blast/>

six

ding

ncbi-blast. <http://www.ncbi.nlm.nih.gov/blast/>

Annotation of hypothetical proteins

In newly sequences, genome as much as 40% of protein are “hypothetical”

To assign function:

- Homology searches in databases
- Search for similar motifs, domains and secondary structures
- Identify conserved functional sites by HMM
- Predict structure with fold recognition or threading
- Assign broad function to protein
- Test assigned function experimentally

YEL008W BASIC INFORMATION

| | |
|-----------------|---|
| Systematic Name | YEL008W |
| Feature Type | ORF, Dubious |
| Description | Hypothetical protein predicted to be involved in metabolism (1) |
| GO Annotations | All YEL008W GO evidence and references View Computational GO annotations for YEL008W |

Genome Economy

One gene → one protein is not true

EST suggests >100,000 proteins in humans (from 25,000 genes)

Alternative splicing

- Joining different exons from a single transcript to form different proteins

Exon shuffling

- Joining exons from different genes
- *Drosophila* Dscam gene contains 115 exons, 20 of which are constitutively spliced and 95 of which are alternatively spliced
- Expresses 38,016 different mRNAs by virtue of alternative splicing

Trans-splicing










- *Drosophila* mdg4 gene
- Joins 4 exons on sense strand and 2 exons on anti-sense strand
- Single transcript of encodes dentin phosphoprotein and sialoprotein. Protein is cleaved to form two different proteins

Gene order comparisons

Rhodobacter capsulatus

bch   //    //  







Heliobacillus mobilis

bch   //   //     

Chlorobium tepidum

bch   //   //   

Chloroflexus aurantiacus

bch   //   //  

- Where gene order is conserved between genomes, it is called **synteny**
- Synteny may indicate functional relationships and physical interaction of proteins
- Genes encoding proteins catalyzing consecutive steps of metabolic pathway sometimes are ordered – co-regulation of “operon”?
- MAL cluster in yeast: multigene complex that encodes the MAL23 trans-acting MAL-activator, MAL21 maltose permease, and MAL22 maltase in order on chromosomes 2, 3, 7, 9 and 10

Confirmation of Gene Function?

- Gene mutation
 - Gene knock-out
 - Gene deletion
 - Transposable element
- Alteration of gene expression
 - Gene overexpression
 - Expression plasmid/vector
 - Integrated chromosomal expression
 - Gene silencing
 - RNAi: microRNA, siRNA
- Protein activity assay (enzyme/cofactor/complex)
- Physiological function analysis (direct and indirect)

...further reading...

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

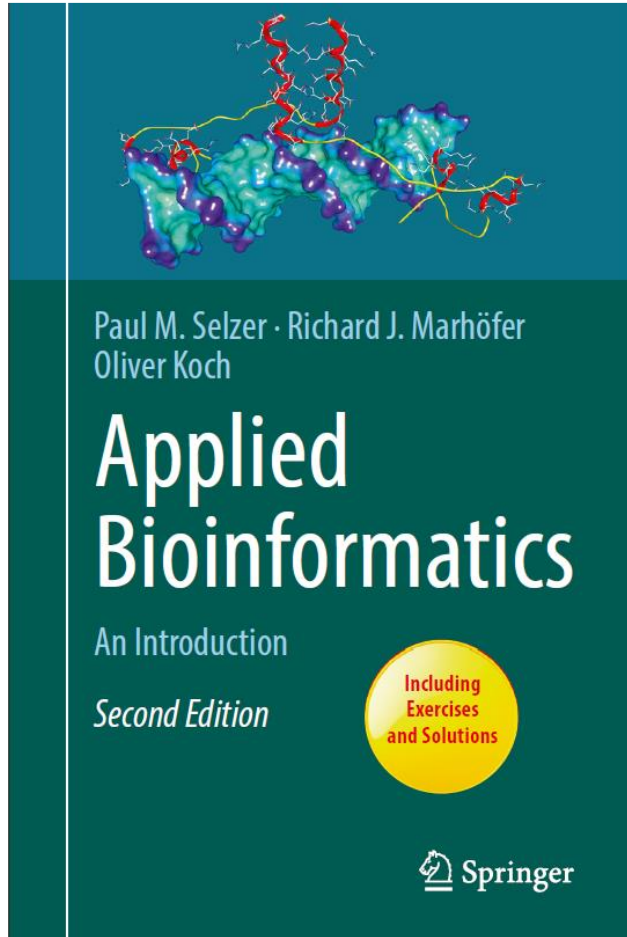
Training & Tutorials

Variation

1. Choudhuri S. Bioinformatics for beginners: Gene, Genome, Molecular Evolution, Databases, and Analytical tools. Elsevier Inc, 2014, 225 pp.
2. Jin Xiong. Essential Bioinformatics. Cambridge University Press, 2006, 339 pp.
3. Hodgman TC, French A, and Westhead DR. Bioinformatics. Taylor & Francis Group, 2010, 340 pp.

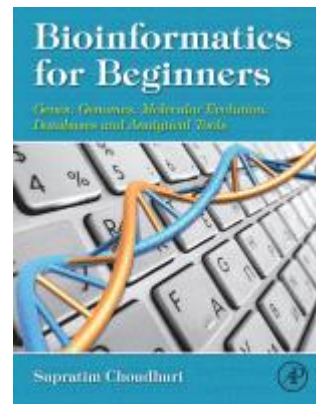


Recommended Reference Books



[Applied Bioinformatics - An Introduction | Paul M. Selzer | Springer](#)

- Use for educational purpose only in SCBT401, MUSC-BT
- <https://drive.google.com/drive/folders/1OIhG8HWxuPhPUtjFAT6NNVXVuzV7dz2l?usp=sharing>



<https://doi.org/10.1016/C2012-0-07153-0>

<https://doi.org/10.1016/C2020-0-03935-6>

