Chapter 14: Functional Genomics
Learning objectives

Upon reading this chapter, you should be able to:

■ define functional genomics;
■ describe the key features of eight model organisms;
■ explain techniques of forward and reverse genetics;
■ discuss the relation between the central dogma and functional genomics; and
■ describe proteomics-based approaches to functional genomics.
Introduction

Relation between genotype and phenotype
Eight model organisms
  \textit{E. coli}; yeast; \textit{Arabidopsis}; \textit{C. elegans}; \textit{Drosophila};
  zebrafish; mouse; human

Functional genomics using reverse and forward genetics
  Reverse genetics: mouse knockouts; yeast; gene trapping; insertional mutagenesis; gene silencing
  Forward genetics: chemical mutagenesis

Functional genomics and the central dogma
  Approaches to function; Functional genomics and DNA; …and RNA; …and protein

Proteomic approaches to functional genomics
  CASP; protein-protein interactions; protein networks

Perspective
Albert Blakeslee (1874–1954) studied the effect of altered chromosome numbers on the phenotype of the jimson-weed *Datura stramonium*, a flowering plant.
Introduction: Functional genomics

Functional genomics is the genome-wide study of the function of DNA (including both genes and non-genic regions), as well as RNA and proteins encoded by DNA.

The term “functional genomics” may apply to
• the genome, transcriptome, or proteome
• the use of high-throughput screens
• the perturbation of gene function
• the complex relationship of genotype and phenotype
Functional genomics approaches to high throughput analyses

DNA → RNA → protein → phenotype

<table>
<thead>
<tr>
<th>Natural variation</th>
<th>DNA SNPs; epigenomics</th>
<th>RNA transcriptome profiling (RNA-seq)</th>
<th>protein localization; protein-protein interactions; pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>-- across development</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-- across body regions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-- across species, strains</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Functional disruptions       | DNA knockout collections transgenic animals Williams syndrome Down syndrome cancer chromosomal changes | RNAi; siRNA nonsense-mediated RNA decay | protein chemical modification myasthenia gravis |
| -- experimental             |                       |                                     |                                                             |
| -- in nature                |                       |                                     |                                                             |
Relationship between genotype and phenotype

The genotype of an individual consists of the DNA that comprises the organism. The phenotype is the outward manifestation in terms of properties such as size, shape, movement, and physiology. We can consider the phenotype of a cell (e.g., a precursor cell may develop into a brain cell or liver cell) or the phenotype of an organism (e.g., a person may have a disease phenotype such as sickle-cell anemia).

A great challenge of biology is to understand the relationship between genotype and phenotype. We can gather information about either one alone, but how they are connected very often remains obscure.
Outline: Functional genomics

Introduction
  Relation between genotype and phenotype

Eight model organisms
  \textit{E. coli}; yeast; \textit{Arabidopsis}; \textit{C. elegans}; \textit{Drosophila};
  zebrafish; mouse; human

Functional genomics using reverse and forward genetics
  Reverse genetics: mouse knockouts; yeast; gene trapping; insertional mutagenesis; gene silencing
  Forward genetics: chemical mutagenesis

Functional genomics and the central dogma
  Approaches to function; Functional genomics and DNA; …and RNA; …and protein

Proteomic approaches to functional genomics
  CASP; protein-protein interactions; protein networks

Perspective
We introduce 8 model organisms that have particularly important roles in functional genomics. The list is not comprehensive, but highlights important principles as well as advantages (and disadvantages) of studying various model systems.
Eight model organisms for functional genomics

Bacterium *Escherichia coli*

Yeast *Saccharomyces cerevisiae*

Plant *Arabidopsis thaliana*

Nematode *Caenorhabditis elegans*

Fruitfly *Drosophila melanogaster*

Zebrafish *Danio rerio*

Mouse *Mus musculus*

*Homo sapiens*: variation in humans
8 model organisms: (1) Bacterium *Escherichia coli*

The bacterium *Escherichia coli* serves as the best-characterized bacterial organism, if not the best-characterized living organism. For decades it served as a leading model organism for bacterial genetics and molecular biology studies.

- 4.6 megabase (million base pairs) genome was sequenced by Blattner *et al.* (1997) Principal website is **EcoCyc**, the Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism
- EcoCyc assigns a function to >75% of the 4501 annotated genes

[https://ecocyc.org/](https://ecocyc.org/)
The budding yeast *S. cerevisiae* is the best-characterized organism among the eukaryotes. It is a single-celled fungus and was the first eukaryote to have its genome sequenced. Its 13 megabase genome encodes 6000 proteins. The *Saccharomyces* Genome Database (SGD) is the principal database and community resource. There are approximately 6600 annotated open reading frames (ORFs, corresponding to genes), including approximately 5000 verified, 750 uncharacterized ones. Additionally, about 4200 gene products have been annotated to their root gene ontology terms, which includes molecular function, biological process, and cellular component.
**SGD (Saccharomyces Genome Database) entry for SEC1**

(www.yeastgenome.org)

### SEC1/YDR164C Summary

#### Alternative single page format

**SEC1 BASIC INFORMATION**

<table>
<thead>
<tr>
<th>Standard Name</th>
<th>SEC1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic Name</td>
<td>YDR164C</td>
</tr>
<tr>
<td>Feature Type</td>
<td>ORF, Verified</td>
</tr>
<tr>
<td>Description</td>
<td>5-like protein involved in docking and fusion of exocytic vesicles through binding to assembled SNARE complexes at the membrane. Localization to sites of secretion (bud neck and bud tip) is dependent on SNARE function (2)</td>
</tr>
</tbody>
</table>

| Name Description | SECory 3 |

<table>
<thead>
<tr>
<th>GO Annotations</th>
<th>All SEC1 GO evidence and references</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th>Manually curated</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Manually curated</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Cellular Component</th>
<th>Manually curated</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Mutant Phenotype</th>
<th>SEC1 Phenotype details and references</th>
</tr>
</thead>
</table>

| Systematic deletion | Free text |

---

**SEC1 RESOURCES**

- **Literature**
  - Literature Guide
  - View

- **Retrieve Sequences**
  - Genomic DNA
  - View

- **Sequence Analysis Tools**
  - BLASTP
  - View

- **Protein Info & Structure**
  - Protein Info
  - View

- **Localization Resources**
  - GFP DB at UCSF
  - View

- **Interactions**
  - BioGRID (Toronto)
  - View

- **Phenotype Resources**
  - PROPHECY
  - View
SGD (*Saccharomyces Genome Database*) entry for *SEC1*

**Interactions**

**Physical Interactions**
- Affinity Capture-MS
- Affinity Capture-RNA
- Affinity Capture-Western
- Reconstituted Complex
- Two-hybrid

**Genetic Interactions**
- Dosage Lethality
- Dosage Rescue
- Phenotypic Suppression
- Synthetic Growth Defect
- Synthetic Lethality

**SEC1 All interactions details and references**

**SEC1 Physical Interactions details and references**

There are 7 total Affinity Capture-MS interactions
- There is 1 total Affinity Capture-RNA interactions
- There are 13 total Affinity Capture-Western interactions
- There are 7 total Reconstituted Complex interactions
- There are 3 total Two-hybrid interactions

**SEC1 Genetic interactions details and references**

There are 5 total Dosage Lethality interactions resulting in the following phenotype: inviable
- There are 13 total Dosage Rescue interactions resulting in the following phenotype: wildtype
- There are 6 total Phenotypic Suppression interactions resulting in the following phenotype: Not available
- There is 1 total Synthetic Growth Defect interactions resulting in the following phenotype: slow growth
- There are 38 total Synthetic Lethality interactions resulting in the following phenotype: inviable

**Sequence Information**

Chr IV: 784212 to 782038 | ORF Map | GBrowse

*Note: this feature is encoded on the Crick strand.*

Genetic position: 94.77 cM

Coordinates: 2006-04-13 | Sequence: 1996-07-31

**Relative Coordinates**

CDS: 1,2175

**Chromosomal Coordinates**

784212, 782038

**Most Recent Updates**

Coordinates: 2006-04-13 | Sequence: 1996-07-31

**External Links**

All Associated Seq | Entrez Gene | Entrez RefSeq
Protein | MIPS | UniProt/Swiss-Prot

**Primary SGID**

S000002571
**SSO1 BASIC INFORMATION**

**Standard Name**
SSO1

**Systematic Name**
YPL232W

**Feature Type**
ORF, Verified

**Description**
Plasma membrane t-SNARE involved in fusion of secretory vesicles at the plasma membrane and in vesicle fusion during sporulation; forms a complex with Sec9p that binds v-SNARE SnC2p; syntaxin homolog, functionally redundant with Sso2p (1, 2, 3, 4)

**GO Annotations**

All SSO1 GO evidence and references

View Computational GO annotations for SSO1

- SNAP receptor activity (IDA, IPI)
- Golgi to plasma membrane transport (TAS)
- membrane fusion (IDA, IMP)
- prospore formation (IMP)
- sporulation (sensu Fungi) (IMP)
- plasma membrane (IDA)
- prospore membrane (IDA)
- SNARE complex (IDA)

**Molecular Function**
Manually curated

**Biological Process**
Manually curated

**Cellular Component**
Manually curated

**Mutant Phenotype**

**SSO1 Phenotype details and references**
Order mutant strains used in systematic deletion project

- viable
- SSO1, SSO2 double null mutant is inviable; high copy number of either SSO1 or SSO2 suppresses mutations in late-acting sec genes (sec1,3,5,9,15)

**SSO1 RESOURCES**

- Literature
  - Literature Guide
  - View

- Retrieve Sequences
  - Genomic DNA
  - View

- Sequence Analysis Tools
  - BLASTP
  - View

- Protein Info & Structure
  - Protein Info
  - View

- Localization Resources
  - GFP DB at UCSF
  - View

- Interactions
  - BioGRID (Toronto)
  - View

- Phenotype Resources
  - PROPHECY
  - View
SSO1 All interactions details and references

SSO1 Physical Interactions details and references
There are 12 total Affinity Capture-Western interactions
There are 2 total Biochemical Activity interactions
There are 2 total Co-crystal Structure interactions
There are 7 total Reconstituted Complex interactions
There is 1 total Two-hybrid interactions

SSO1 Genetic Interactions details and references
There are 3 total Dosage Lethality interactions resulting in the following phenotype: inviable
There are 6 total Dosage Rescue interactions resulting in the following phenotype: wildtype
There is 1 total Phenotypic Enhancement interactions resulting in the following phenotype: Not available
There are 2 total Phenotypic Suppression interactions resulting in the following phenotype: Not available
There is 1 total Synthetic Lethality interactions resulting in the following phenotype: inviable
There is 1 total Synthetic Rescue interactions resulting in the following phenotype: wildtype

Chromosome: 107275 to 108147 | ORF Map | GBrowse


Relative Coordinates Chromosomal Coordinates Most Recent Updates Coordinates Sequence
CDS 1.873 107275.108147 1996-07-31 1996-07-31

ORF Genomic DNA Get Sequence

SSO00008153

Additional Information for SSO1

Community wiki Domains/Motifs Expression Connection Function Junction
Gene/Sequence Resources Global Gene Hunter Locus History PDB Homologs
Protein Info Researchers
A constitutive trafficking pathway exists in yeast (left) with a set of proteins having orthologs in a regulated trafficking pathway in mammals (right).
Diagram of *S. cerevisiae* and mammalian proteins involved in secretion to illustrate functional genomics principles and approaches

(c) pathway

input (vesicle arrives)

\[ \text{Sec1p} \]
\[ \text{Sso1p} \]

\[ \text{Snc1p} \]
\[ \text{Sso1p} \]
\[ \text{Snc2p} \]
\[ \text{Sso2p} \]
\[ \text{Sec9p} \] output (cargo secreted)

Pathway diagram: parallel pathways

(d) protein interactions

\[ \text{Sec1p} \]
\[ \text{Sso1p} \]
\[ \text{Snc1p} \]
\[ \text{Sso1p} \]
\[ \text{Snc2p} \]
\[ \text{Sso2p} \]
\[ \text{Sec9p} \]

(e) genetic interactions

Biochemical protein interactions (left) and genetic interactions (above) are complementary methods to elucidate pathways
8 model organisms: (3) Plant *Arabidopsis thaliana*

- The thale cress *Arabidopsis thaliana* was the first plant to have its genome sequenced (and the third finished eukaryotic genome sequence).
- Model for eukaryotic functional genomics projects
- Principal web site is The *Arabidopsis* Information Resource (TAIR)
- Appealing features as a model plant: short generation time, prolific seed production, compact genome size, and opportunities for genetic manipulation.
The Arabidopsis Information Resource (TAIR): principal genome database for Arabidopsis

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community. Gene product function data is updated every two weeks from the latest published research literature and community data submissions. Gene structures are updated 1-2 times per year using computational and manual methods as well as community submissions of new and updated genes. TAIR also provides extensive linkouts from our data pages to other Arabidopsis resources.

The Arabidopsis Biological Resource Center at The Ohio State University collects, reproduces, preserves and distributes seed and DNA resources of Arabidopsis thaliana and related species. Stock information and ordering for the ABRC are fully integrated into TAIR.

Breaking News

Change to seed ordering process for European users. NASC can no longer accept orders placed through TAIR, because of changes to their database and ordering system.

AraCyc 4.1 release
23 pathways were significantly updated in the last release in October. More details.

New GO bar charts
Try our new bar charts to visualize GO annotation categories for your gene set or the whole genome. (see details)

GBrowse now at TAIR
View TAIR genome map data using the GMOD generic genome browser, or upload your own genome data track (see details)

Perlegen SNPs new available
243,052 high-quality SNPs from Perlegen resequencing arrays now available from TAIR polymorphism search and SeqViewer. Over 1 million
<table>
<thead>
<tr>
<th>Search</th>
<th>Tools</th>
<th>Stocks</th>
<th>Download</th>
<th>Portals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search Overview</td>
<td>Tools Overview</td>
<td>Stocks Overview</td>
<td>Download Overview</td>
<td>Portals Overview</td>
</tr>
<tr>
<td>DNA/Clones</td>
<td>GBrowse</td>
<td>ABRC Home</td>
<td>Genes</td>
<td>Clones/DNA Resources</td>
</tr>
<tr>
<td>Ecotypes</td>
<td>Seqviewer</td>
<td>Browse ABRC Catalog</td>
<td>GO and PO Annotations</td>
<td>Education and Outreach</td>
</tr>
<tr>
<td>Genes</td>
<td>Mapviewer</td>
<td>Supplement to ABRC Catalog</td>
<td>Maps</td>
<td>Gene Expression Resources</td>
</tr>
<tr>
<td>GO Annotations</td>
<td>AraCyc Metabolic Pathways</td>
<td>Search ABRC DNA/Clone Stocks</td>
<td>Pathways</td>
<td>Genome Annotation</td>
</tr>
<tr>
<td>Keywords</td>
<td>BLAST</td>
<td>Search ABRC Seed/Germplasm Stocks</td>
<td>Proteins</td>
<td>MASC/Functional Genomics</td>
</tr>
<tr>
<td>Locus History</td>
<td>WU-BLAST</td>
<td>ABRC Stock Order History</td>
<td>Proteins</td>
<td>Mutant and Mapping Resources</td>
</tr>
<tr>
<td>Markers</td>
<td>FASTA</td>
<td>ABRC Fee Structure</td>
<td>Proteins</td>
<td>Nomenclature</td>
</tr>
<tr>
<td>Microarray Element</td>
<td>Patmatch</td>
<td>Place ABRC Order</td>
<td>Software</td>
<td>Proteome Resources</td>
</tr>
<tr>
<td>Microarray Experiment</td>
<td>Motif Analysis</td>
<td>Search My ABRC Orders</td>
<td>User Requests</td>
<td></td>
</tr>
<tr>
<td>Microarray Expression</td>
<td>VxInsight</td>
<td>Search ABRC Invoices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>People/Labs</td>
<td>Java Tree View</td>
<td>How to Make Payments to ABRC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polymorphisms/Aleles</td>
<td>Bulk Data Retrieval</td>
<td>ABRC Stock Donation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteins</td>
<td>Chromosome Map Tool</td>
<td>Download Overview</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td>Gene Hunter</td>
<td>ABRC Home</td>
<td>Download Overview</td>
<td></td>
</tr>
<tr>
<td>Publication</td>
<td>Restriction Analysis</td>
<td>Browse ABRC Catalog</td>
<td>Genes</td>
<td></td>
</tr>
<tr>
<td>Seed/Germplasm</td>
<td>Gene Symbol Registry</td>
<td>Supplement to ABRC Catalog</td>
<td>GO and PO Annotations</td>
<td></td>
</tr>
<tr>
<td>Sequences</td>
<td></td>
<td>Search ABRC DNA/Clone Stocks</td>
<td>Maps</td>
<td></td>
</tr>
</tbody>
</table>

**The Arabidopsis Information Resource (TAIR)**
8 model organisms: (4) Nematode *C. elegans*

- First multicellular animal to have its genome sequenced
- Capable of complex behaviors
- Body is simple and all the 959 somatic cells in its body have been mapped including their lineages throughout development

- **Wormbase** is the main database/resource
- Genome encodes ~20,400 protein-coding genes (same number as in humans).
8 model organisms: (5) Fruitfly *Drosophila*

- Metazoan (animal) invertebrate
- Early studies of *Drosophila* resulted in the descriptions of the nature of the gene as well as linkage and recombination, producing gene maps a century ago
- Sequencing of many *Drosophila* genomes (and inbred lines) providing unprecedented insight into mechanisms of genome evolution
- Genomic changes can be induced with extreme precision, from single-nucleotide changes to introducing large-scale chromosomal deletions, duplications, inversions, or other modifications
8 model organisms: (6) zebrafish *Danio rerio*

- Lineages leading to modern fish and humans diverged approximately 450 million years ago
- Freshwater fish having a genome size of 1.8 billion base pairs (Gb) organized into 25 chromosomes
- >26,000 protein-coding genes
- Mutations in large numbers of human disease gene orthologs have been generated and characterized, using both forward and reverse genetic screens
- Short generation time
- Large numbers of progeny
- Developing embryo is transparent (transgenes can be visualized)
8 model organisms: (7) Mouse *Mus musculus*

- Shared common ancestor with humans ~90 million years ago
- Close structural and functional relationship between mouse and human genomes
- Relatively short generational span
- Powerful tools developed to manipulate its genome
- Main mouse genome website is the Mouse Genome Informatics (MGI)
- ~10,000 genes knocked out
- Collaborative Cross: 1000 recombinant inbred strains of mouse are being bred, producing large numbers of genetically related mice that have nonlethal phenotypic diversity
MGI database is the principal website for mouse genomics information. The home page provides a portal to a vast number of resources.
Mouse genome informatics (MGI) database

MGI offers customized Biomarts for mouse functional genomics projects.

International Knockout Mouse Consortium (IKMC)
8 model organisms: (8) humans

We do not think of humans as model organisms per se. But nature performs functional genomics experiments on us constantly.

Motivation for studying humans: to understand the causes of disease in order to search for more effective diagnoses, preventions, treatments, and ultimately cures.
Outline: Functional genomics

Introduction
  Relation between genotype and phenotype
Eight model organisms
  *E. coli*; *yeast*; *Arabidopsis*; *C. elegans*; *Drosophila*;
zebrafish; mouse; human

Functional genomics using reverse and forward genetics
  Reverse genetics: mouse knockouts; yeast; gene
  trapping; insertional mutagengesis; gene silencing
  Forward genetics: chemical mutagenesis

Functional genomics and the central dogma
  Approaches to function; Functional genomics and
  DNA; …and RNA; …and protein

Proteomic approaches to functional genomics
  CASP; protein-protein interactions; protein networks

Perspective
Functional genomics using reverse and forward genetics

**Reverse genetic screens:** a large number of genes (or gene products) is systematically inhibited one by one. This can be accomplished in many ways, for example by deleting genes using homologous recombination, gene trapping, or by selectively reducing messenger RNA abundance. One or more phenotypes of interest are then measured.

Main challenges of this approach:

- for some organisms it difficult to disrupt large numbers of genes (such as tens of thousands) in a systematic fashion.
- It can also be challenging to discern the phenotypic consequences for a gene that is disrupted.
Functional genomics using reverse and forward genetics

**Forward genetic screens:**

- the starting point is a defined phenotype of interest, such as the ability of plants to grow in the presence of a drug, neurons to extend axons to appropriate targets in the mammalian nervous system, or an eukaryotic cell to transport cargo.

- An experimental intervention is made, such as administering a chemical mutagen or radiation to cells (or to an organism). This results in the creation of mutants.

- The phenotype of interest is observed in rare representatives among a large collection of mutants.
Reverse genetics (mutate genes then examine phenotypes)

Strategy:
- Systematically inhibit the function of every gene in a genome
- Approach 1: gene targeting by homologous recombination
- Approach 2: gene trap mutagenesis
- Approach 3: inhibit gene expression using RNA interference
- Measure the effect of gene disruption on a phenotype

Forward genetics

Strategy:
- Identify a phenotype (e.g. growth in the presence of a drug)
- Mutate genomic DNA (e.g. by chemical mutagenesis)
- Identify individuals having an altered phenotype
- Identify the gene(s) that were mutated
- Confirm those genes have causal roles in influencing the genotype
Reverse genetics: mouse knockouts and the β-globin gene

- Knocking out a gene: create an animal model in which a homozygous deletion is created, that is, there are zero copies (denoted (−/−) and referred to as a null allele) instead of the wildtype situation of two copies in a diploid organism (+/+).
- In a hemizygous deletion, one copy is deleted and one copy remains (+/−).
- Use a targeting vector that includes the β-globin gene having a portion modified by insertion of the neo gene into exon 2.
- This targeting vector is introduced into embryonic stem cells by electroporation. When the cells are cultured in the presence of the drug G418, wildtype cells die whereas cells having the neo cassette (gene cassette) survive. Confirm by PCR.
The β globin locus

Human (85 kilobases on chr11:5,235,001–5,320,000, GRCh37/hg19 assembly)
The β globin locus

Mouse (65 kilobases on chr7:103,806,001–103,871,000, GRCm38/mm10 assembly)
The successfully targeted locus includes a β globin gene that is interrupted by the neo gene.
Mouse Genome Informatics (MGI) website entry for the major beta globin gene (**Hbb-b1**)

The entry summarizes molecular data on that gene and includes a phenotype category, indicating that seven mutant alleles are indexed (five targeted and two chemically induced).
The entry includes phenotypic data such as type of mutation, human disease relevance, genetic background.
Reverse genetics: knocking out genes in yeast using molecular barcodes

Knockout studies in the yeast *S. cerevisiae* are far more straightforward and also much more sophisticated than in the mouse:

- The yeast genome is extremely compact, having very short noncoding regions and introns in fewer than 7% of its ~6000 genes.
- Homologous recombination can be performed with high efficiency.
Targeted deletion of virtually all *S. cerevisiae* genes

Strategy: use gene replacement by homologous recombination. Each gene (e.g., *SSO1*) is deleted and replaced by a *KanR* gene.
Reverse genetics: random insertional mutagenesis (gene trapping)

- Insertional mutations are introduced across the genome in embryonic stem cells.
- Vectors insert into genomic DNA leaving sequence tags that often include a reporter gene.
- In this way, mutagenesis of a gene can be accomplished and the gene expression pattern of the mutated gene can be visualized.
## Reverse genetics techniques

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homologous recombination (e.g., gene knockouts)</td>
<td>A targeted gene can be replaced, deleted, or modified precisely; stable mutations are produced; specific (no off-target effects)</td>
<td>Low throughput; low efficiency</td>
</tr>
<tr>
<td>Gene silencing (e.g., RNAi)</td>
<td>Can be high-throughput; can be used to generate an allelic series; can restrict application to specific tissues or developmental stages</td>
<td>Unpredictable degree of gene silencing; phenotypes not stable; off-target effects are possible</td>
</tr>
<tr>
<td>Insertional mutagenesis</td>
<td>High-throughput; used for loss-of-function and gain-of-function studies; results in stable mutations</td>
<td>Random or transposon-mediated insertions target only a subset of the genome; limited effectiveness on tandemly repeated genes; limited usefulness for essential genes</td>
</tr>
<tr>
<td>Ectopic expression</td>
<td>Similar to gene silencing</td>
<td>Similar to gene silencing</td>
</tr>
</tbody>
</table>
Strategies for gene trap mutagenesis

An enhancer trap consists of a vector containing a promoter, a neo gene that confers antibiotic resistance (and therefore allows for selection of successfully integrated sequences), and a polyadenylation signal (polyA). This construct is activated by an endogenous enhancer, and disrupts the function of the endogenous gene.
Strategies for gene trap mutagenesis

A promoter trap lacks an exogenous promoter and instead depends on an endogenous enhancer and promoter. It includes a splice acceptor (SA), neo cassette, and polyadenylation site. Integration of this vector disrupts the expression of an endogenous gene.
Strategies for gene trap mutagenesis

A poly(A) trap vector includes its own promoter and neo cassette but depends on an endogenous polyadenylation signal for successful expression.
Reverse genetics: insertional mutagenesis in yeast

Two powerful approaches to gene disruption in yeast (in addition to homologous recombination) are:

(1) genetic footprinting using transposons; and

(2) harnessing exogenous transposons.
A population of yeast is selected (e.g., by changing the medium or adding a drug); some genes will be unaffected by the selection process.
Genetic footprinting

The yeast transposable element **Ty1** is present in about 35 copies per genome;

Random insertion of a transposon allows gene-specific PCR to be performed.

Visualization of DNA products electrophoresed on a gel. Some genes will be unaffected by the selection process (panel at left). Other genes, tagged by the transposition, will be associated with a reduction in fitness. Less PCR product will be observed, therefore identifying this gene as necessary for survival of yeast in that selection condition.
Another approach to identifying gene function is to **disrupt the messenger RNA** rather than the genomic DNA. RNA interference (RNAi) is a powerful, versatile technique that allows **genes to be silenced** by double-stranded RNA.
Forward genetics: chemical mutagenesis

• Forward genetics approaches are sometimes called phenotype-driven screens.
• *N*-ethyl-*N*-nitrosurea (ENU) is a powerful chemical mutagen used to alter the male germline to induce point mutations (applied to mouse, *Arabidopsis*, other organisms).
• After ENU is given a phenotype of interest is observed. Recombinant animals are created by inbreeding and the phenotype can then be demonstrated to be heritable.
• The mutagenized gene is mapped by positional cloning and identified by sequencing the genes in the mapped interval.
Comparison of reverse and forward genetics

• Reverse genetics asks “What is the phenotype of this mutant?” Forward genetics asks “What mutants have this particular phenotype?”
• Reverse genetics approaches attempt to generate null alleles as a primary strategy (and conditional alleles in many cases).
• Forward genetics strategies such as chemical mutagenesis are “blind” in that multiple mutant alleles are generated that affect a phenotype.
Outline: Functional genomics

Introduction
  Relation between genotype and phenotype

Eight model organisms
  *E. coli*; yeast; *Arabidopsis*; *C. elegans*; *Drosophila*; zebrafish; mouse; human

Functional genomics using reverse and forward genetics
  Reverse genetics: mouse knockouts; yeast; gene trapping; insertional mutagenesis; gene silencing
  Forward genetics: chemical mutagenesis

Functional genomics and the central dogma
  Approaches to function; Functional genomics and DNA; ...and RNA; ...and protein

Proteomic approaches to functional genomics
  CASP; protein-protein interactions; protein networks

Perspective
The ENCODE project claimed that >80% of genomic DNA is functional.

We now consider three different definitions of function:
• evolutionary selected effect
• causal role
• inferred selected effect

And consider three approaches to studying function:
• genetic
• evolutionary
• biochemical
Distinguishing different approaches to function (columns) from definitions of function (rows)

<table>
<thead>
<tr>
<th>Definition of function</th>
<th>Genetic</th>
<th>Evolutionary</th>
<th>Biochemical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Establish consequence of sequence alterations</td>
<td>Comparative genomics: align DNA, proteins</td>
<td>Measure an activity in a given cell type</td>
</tr>
<tr>
<td>Evolutionary selected effect</td>
<td>Naturally occurring or targeted mutations can be a “gold standard”</td>
<td>&lt;15% of genome under constraint</td>
<td>There are increasing numbers of examples of mutations in enhancer regions that cause disease</td>
</tr>
<tr>
<td></td>
<td>Possible to infer function based on selection</td>
<td>Noncoding regions often hard to align</td>
<td></td>
</tr>
<tr>
<td>Causal role</td>
<td>Example: knockout generates a phenotype</td>
<td>Many conserved loci functionally important</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caveat: some phenotypes depend on a particular condition to be identified</td>
<td>Caveat: some ultra-conserved loci dispensable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caveat: some poorly conserved loci are functionally equivalent</td>
<td></td>
</tr>
<tr>
<td>Inferred selected effect</td>
<td>Question inspired by ENCODE biochemical map: do most biochemical signatures correspond to functional sites that impact fitness?</td>
<td>Creation of ENCODE biochemical map may inspire new discoveries of sequence conservation in biochemically functional noncoding regions</td>
<td>Majority of genome functional</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>An uncertain % drift, noise</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ENCODE biochemical map will facilitate hypothesis testing</td>
</tr>
</tbody>
</table>
Distinguishing different approaches to function from definitions of function

Three circles corresponding to the magnitude of functional findings in ENCODE
Functional genomics and DNA: integrating information

A goal of functional genomics is to provide integrated views of DNA, RNA, protein, and pathways. Many resources (such as those at Ensembl, EBI, and NCBI) offer this integrated view.

An example is the Frequency weighted links (FLink) tool at NCBI. Input a list of genes (or proteins or small molecules) and obtain a ranked list of biosystems.
NCBI FLINK: identify connections between an input list of proteins, genes, or other molecules and associated database entries

FLINK: select database

FLINK: input identifiers or search terms
NCBI FLink: identify connections between an input list of proteins, genes, or other molecules and associated database entries

FLink: table of globin results
Surveys of RNA transcript levels across different regions (for multicellular organisms) and times of development provide fundamental information about an organism’s program of gene expression.

As an example, the *Saccharomyces* Genome Database (SGD) offers many resources to describe gene expression in yeast. For each gene, an expression summary plots the log2 ratio of gene expression (x axis) versus the number of experiments. That plot is clickable, so experiments in which SEC1 RNA is dramatically up- or down- regulated can be quickly identified.
Outline: Functional genomics

Introduction
  Relation between genotype and phenotype
Eight model organisms
  *E. coli*; yeast; *Arabidopsis*; *C. elegans*; *Drosophila*; zebrafish; mouse; human
Functional genomics using reverse and forward genetics
  Reverse genetics: mouse knockouts; yeast; gene trapping; insertional mutagenesis; gene silencing
  Forward genetics: chemical mutagenesis
Functional genomics and the central dogma
  Approaches to function; Functional genomics and DNA; …and RNA; …and protein
Proteomic approaches to functional genomics
  CASP; protein-protein interactions; protein networks
Perspective
Proteomic approaches to functional genomics

Basic features of proteins include their sequence, structure, homology relationships, post-translational modifications, localization, and function. In addition to the study of individual proteins, high throughput analyses of thousands of proteins are possible. We describe three approaches:

• identifying pairwise interactions between protein using the yeast two-hybrid system;
• identifying protein complexes involving two or more proteins using affinity chromatography with mass spectrometry; and
• analyzing protein pathways.

While protein studies have been studied in depth in a variety of model organisms, studies in *S. cerevisiae* are particularly advanced.
Proteomic approaches to functional genomics

We usually think of forward and reverse approaches in terms of genetics, but these terms can apply to proteomics.

**Forward proteomics:**
- Select experimental system (e.g. normal versus diseased tissue).
- Proteins are extracted and may be labeled with fluorescent dyes or other tags.
- Proteins are separated and analyzed by techniques such as mass spectrometry.
- Spectra are analyzed and differentially regulated proteins are identified.
- These regulated proteins may reflect functional differences in the comparison of the original samples.
Proteomic approaches to functional genomics

We usually think of forward and reverse approaches in terms of genetics, but these terms can apply to proteomics.

Reverse proteomics:
• A genome sequence of interest is analyzed and genes, transcripts, and proteins are predicted.
• Complementary DNAs (cDNAs) are cloned based on information about open reading frames.
• cDNAs are validated by sequence analysis and expressed in systems such as E. coli (for the production of recombinant proteins), mammalian cells, or other model organism systems.
• Functional assays are performed; assays include the yeast two-hybrid system or other protein interaction assays.
Forward and reverse proteomics

Forward proteomics

1. Biological system
2. Protein isolation
3. Sample preparation
4. Mass spectrometry
5. Identify proteins
6. Data analysis; bioinformatics
7. Infer protein function

Reverse proteomics

1. Biological system
2. Data analysis; bioinformatics
3. Cloning
4. Expression
5. Data analysis; bioinformatics
6. Functional assays
7. Infer protein function
Critical assessment of protein function annotation

CAGI involved many challenges inherent in the nature of protein function:

• Protein function is defined at multiple levels, involving the role of a protein on its own and in pathways, cells, tissues, and organisms.
• Protein function is context dependent (e.g., many proteins change function in the presence of a signal such as calcium or a binding partner).
• Proteins are often multifunctional.
• Functional annotations are often incomplete and may be incorrect.
• Curation efforts map protein function to gene names, but multiple isoforms of a gene may have different functions.
Protein-protein interactions

Most proteins perform their functions in networks associated with other proteins and other biomolecules. As a basic approach to discerning protein function, pairwise interactions between proteins can be characterized.

Proteins often interact with partners with high affinity. (The two main parameters of any binding interaction are the affinity, measured by the dissociation constant $K_D$, and the maximal number of binding sites $B_{max}$.)
Protein-protein interactions

The interactions of two purified proteins can be measured with dozens of techniques such as the following:

- **Co-immunoprecipitation**: specific antibodies directed against a protein are used to precipitate the protein along with any associated binding partners.

- **Affinity chromatography**: a cDNA construct encodes a protein of interest in frame with glutathione S-transferase (GST) or some other tag. A resin to which glutathione is covalently attached is incubated with a GST fusion protein, and it binds to the resin along with any binding partners. Irrelevant proteins are eluted and then the specific binding complex is eluted and its protein content is identified.
Protein-protein interactions

• **Cross-linking with chemicals or ultraviolet radiation:** a protein is allowed to bind to its partners and then cross-linking is applied and the interactors are identified.

• **Surface plasmon resonance** (with the BIAcore technology of GE Healthcare): a protein is immobilized to a surface and kinetic binding properties of interacting proteins are measured.

• **Equilibrium dialysis and filter binding assays**, in which bound & free ligands are separated and quantitated.

• **Fluorescent resonance energy transfer** (FRET): two labeled proteins yield a characteristic change in resonance energy upon sharing a close physical interaction.
Yeast two-hybrid system

(a) DNA binding without activation

(b) Prey bound to activation domain

(c) Transcription activation upon prey binding to bait
## Protein–protein interaction databases

<table>
<thead>
<tr>
<th>Database</th>
<th>Comment</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive Yeast Genome Database (CYGD)</td>
<td>From the Munich Information Center for Protein Sequences (MIPS)</td>
<td><a href="http://mips.helmholtz-muenchen.de/genre/proj/yeast/">http://mips.helmholtz-muenchen.de/genre/proj/yeast/</a></td>
</tr>
<tr>
<td>IntAct</td>
<td>At the European Bioinformatics Institute</td>
<td><a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a></td>
</tr>
<tr>
<td>Molecular Interactions (MINT) Database</td>
<td>Rome</td>
<td><a href="http://mint.bio.uniroma2.it/mint/">http://mint.bio.uniroma2.it/mint/</a></td>
</tr>
<tr>
<td>PDZBase</td>
<td>Database of PDZ domains</td>
<td><a href="http://abc.med.cornell.edu/pdzbase">http://abc.med.cornell.edu/pdzbase</a></td>
</tr>
</tbody>
</table>

There are many prominent protein–protein interaction databases.
Example of a protein-protein interaction database entry: BioGrid network map for syntaxin and its binding partners
A typical mammalian genome has \(~20,000\) to \(25,000\) protein-coding genes, a subset of which (perhaps \(10,000\) to \(15,000\)) are expressed in any given cell type. These proteins are localized to particular compartments (or are secreted) where many of them interact as part of their function.

Many databases show protein network data. We next show PSICQUIC and Cytoscape as examples.
Protein interaction networks

<table>
<thead>
<tr>
<th>Database</th>
<th>Database</th>
<th>Database</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>APIID</td>
<td>BAR</td>
<td>bhf-ucl</td>
<td>BIND</td>
</tr>
<tr>
<td>BindingDB</td>
<td>BioGrid</td>
<td>ChEMBL</td>
<td>DIP</td>
</tr>
<tr>
<td>DrugBank</td>
<td>GeneMANIA</td>
<td>HPIDb</td>
<td>I2D</td>
</tr>
<tr>
<td>I2D-IMEx</td>
<td>InnateDB</td>
<td>InnateDB-IMEx</td>
<td>IntAct</td>
</tr>
<tr>
<td>Interoporc</td>
<td>iRefindex</td>
<td>MatrixDB</td>
<td>IntAct</td>
</tr>
<tr>
<td>mentha</td>
<td>MINT</td>
<td>MolCon</td>
<td>MBInfo</td>
</tr>
<tr>
<td>Reactome</td>
<td>Reactome-FIs</td>
<td>Spike</td>
<td>MPIDB</td>
</tr>
<tr>
<td>TopFind</td>
<td>UniProt</td>
<td>VirHostNet</td>
<td>STRING</td>
</tr>
</tbody>
</table>

PSICQUIC databases of protein interactions.
Protein interaction networks

PSICQUIC display of Cytoscape network for syntaxin

Cytoscape data import
Protein interaction networks

Zoom of Cytoscape diagram showing syntaxin binding partners
From pairwise interactions to protein networks

There are many issues regarding protein interaction networks.

- **Assessment of accuracy.** How likely is it that a false positive or false negative error has occurred? Benchmark ("gold standard") datasets are required that consist of trustworthy pathways.

- **Choice of data.** Many researchers integrate data from genomic sequences, expression of RNA transcripts, and protein measurements. But RNA and protein levels may be poorly correlated.

- **Experimental organism.** Function may be better conserved between paralogs than between orthologs!
From pairwise interactions to protein networks

There are many issues regarding protein interaction networks.

• **Variation in Pathways.** Some pathways (e.g. Krebs cycle) are characterized in great detail; many not. Some are transient, others stable.

• **Categories of maps.** Maps may be of metabolic pathways, physical and/or genetic interaction data, summaries of the scientific literature, or signalling pathways. **Maps may be based on experimental data or inferred relationships.**
There are many database resources.

- **PathGuide** lists >500 biological pathway resources.
- **BioGRID** database provides manual curation of ~32,000 publications describing physical and genetic interactions.
- **MetaCyc** is a database of metabolic pathways. [https://metacyc.org/](https://metacyc.org/)
- **Kyoto Encyclopedia of Genes and Genomes (KEGG)** contains a detailed map of metabolism based on 120 metabolic pathways, with links to various organisms.
- **KEGG pathways** are a collection of manually drawn maps in six areas: metabolism; genetic information processing; environmental information processing; cellular processes; human diseases; and drug development.
KEGG database

KEGG includes pathway maps, data for a broad range of organisms, and a variety of analysis tools.
KEGG database

KEGG includes maps and data for a broad range of organisms. This pathway shows SNARE function including syntaxin.
A pathway for amyotrophic lateral sclerosis (ALS; Lou Gehrig’s disease) is shown.
Introduction
   Relation between genotype and phenotype
Eight model organisms
   \textit{E. coli}; yeast; \textit{Arabidopsis}; \textit{C. elegans}; \textit{Drosophila};
   zebrafish; mouse; human
Functional genomics using reverse and forward genetics
   Reverse genetics: mouse knockouts; yeast; gene trapping;
   insertional mutagenesis; gene silencing
   Forward genetics: chemical mutagenesis
Functional genomics and the central dogma
   Approaches to function; Functional genomics and DNA; …and RNA; …and protein
Proteomic approaches to functional genomics
   CASP; protein-protein interactions; protein networks
Perspective
The field of functional genomics is broad, and can be considered using many different categories.

• What type of organism do we wish to study? We highlighted eight model organisms, although many other models are commonly used.

• What type of questions do we want to address: natural variation or experimental manipulations used to elucidate gene function?

• What type of experimental approach do we wish to apply (e.g., forward versus reverse genetics)?

• What type of molecules do we wish to study (i.e., from genomic DNA to RNA to protein or metabolites)?

• What types of biological questions are we trying to address?
We are beginning to confront a problem that is perhaps even harder than identifying genes: identifying their function. Function has many definitions.