# Proteomics, functional genomics, and systems biology

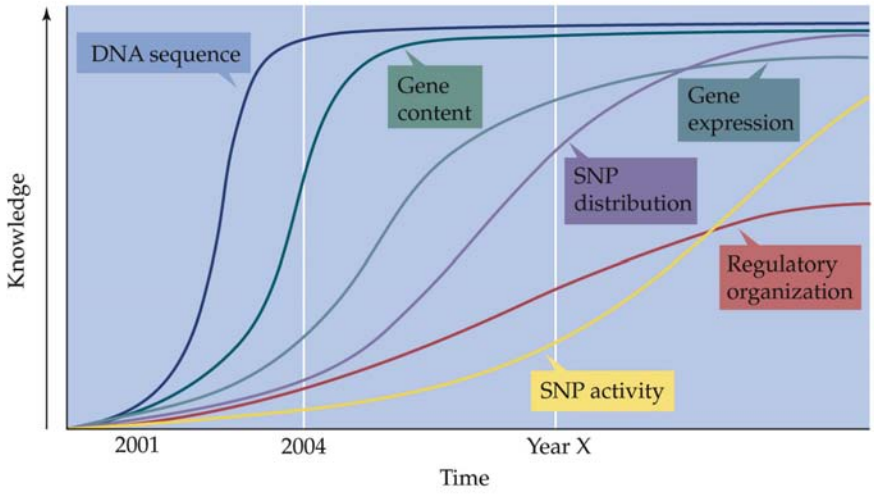
Biosciences 741: Genomics

Fall, 2013

Week 11

## Orthologs, paralogs, and synteny groups

- "Ortholog" refers to the relationship between two genes, in two species, such that those two genes last shared a common ancestor at the time of species divergence.
- "Paralog" refers to the relationship between two genes (in the same or different species) that last shared a common ancestor prior to a gene duplication (or other chromosomal rearrangement). Thus gene families are composed of paralogs.
- Orthologs generally retain similar functions. Paralogs usually (not always) evolve different or more specialized functions.
- Orthologs are often tentatively identified as the most closely related pair of genes between two genomes. However, genetic map order is usually a more sensitive and specific test, because orthologs (by definition) remain within the same synteny group.
- Gene duplication events create chromosomal rearrangements - tandem duplications, segmental duplications, transpositions, retrotranspositions, and so on. Thus paralogs generally have different genetic map locations.



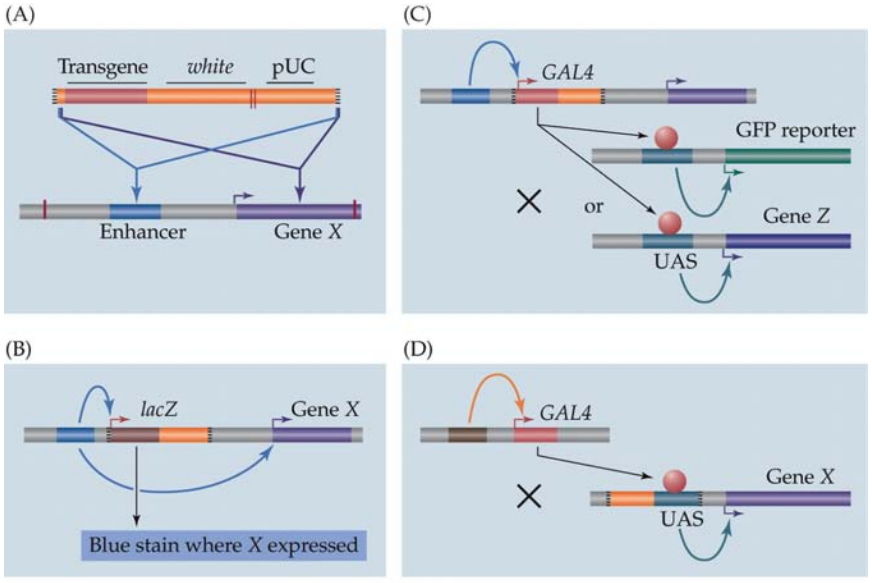

*A PRIMER OF GENOME SCIENCE 3e*, Figure 6.1



## Functional Genomics

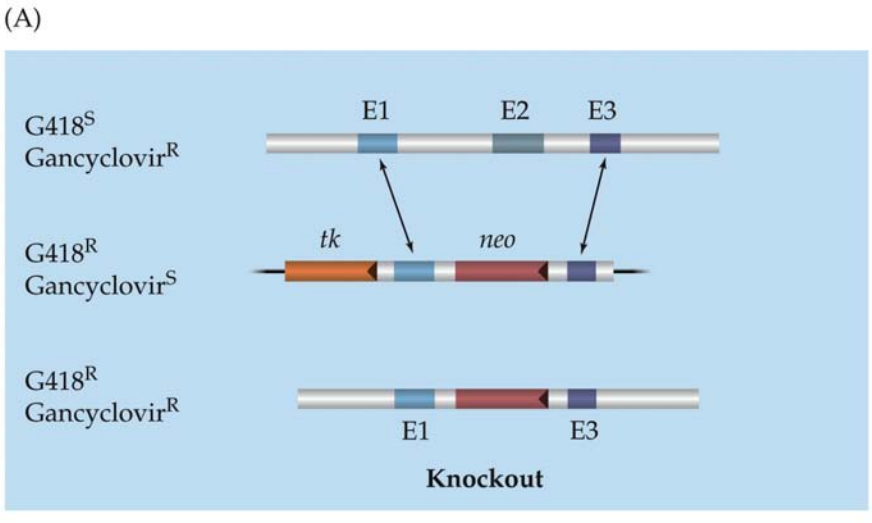
- The field of "functional genomics" represents the next stage in genomics, in which we will move beyond the (relatively) simple clerical exercise of "sequence annotation" to experimental determination of biological function(s) at the whole genome level.
- In most cases, functional genomics is based on the methods of molecular genetics, but adapted to high throughput, in one of the following ways:
- "forward genetics" uses random mutagenesis to identify the set of genes that affect any given trait.
- "reverse genetics" uses molecular sequence information to construct loss-of-function mutations without needing to know the phenotype.
- "fine structure genetics" is used to systematically manipulate the structure and function of genes, in order to obtain a more detailed picture of the functions of each part of the molecule, or the functions in each part of the body, etc.
- Chip-on-chip and DNA microarrays are also part of functional genomics.

Transposon mutagenesis, enhancer trap screens, the GAL4-UAS expression system




*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.24






*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.20 (Part 1)

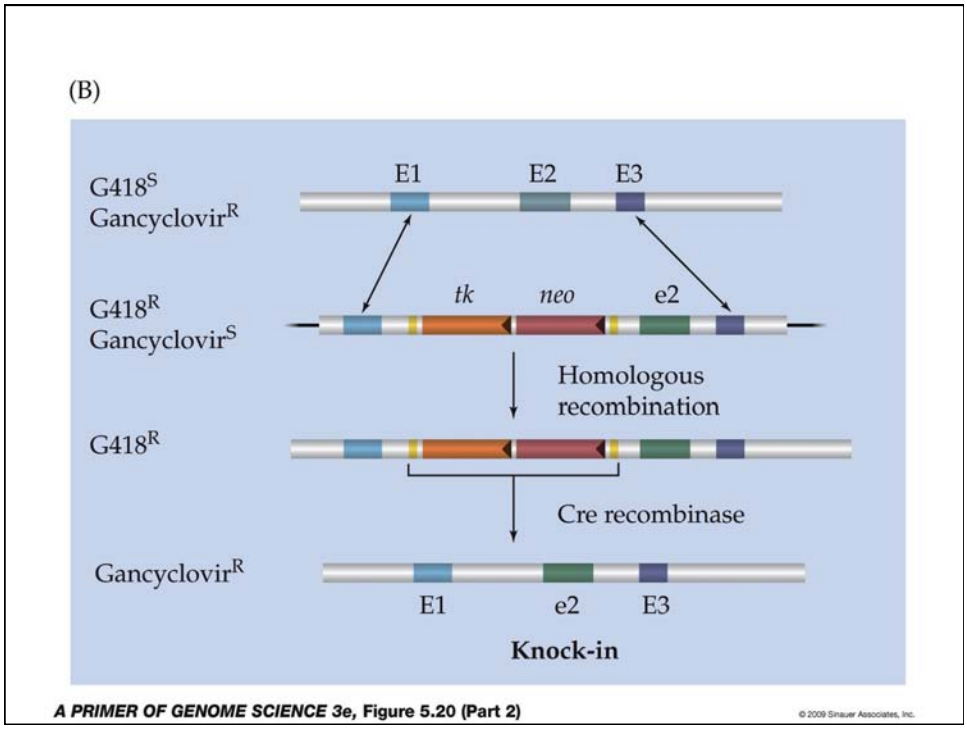



Figure 5.25 - Floxing mice to knock out gene function in a specific tissue or developmental stage

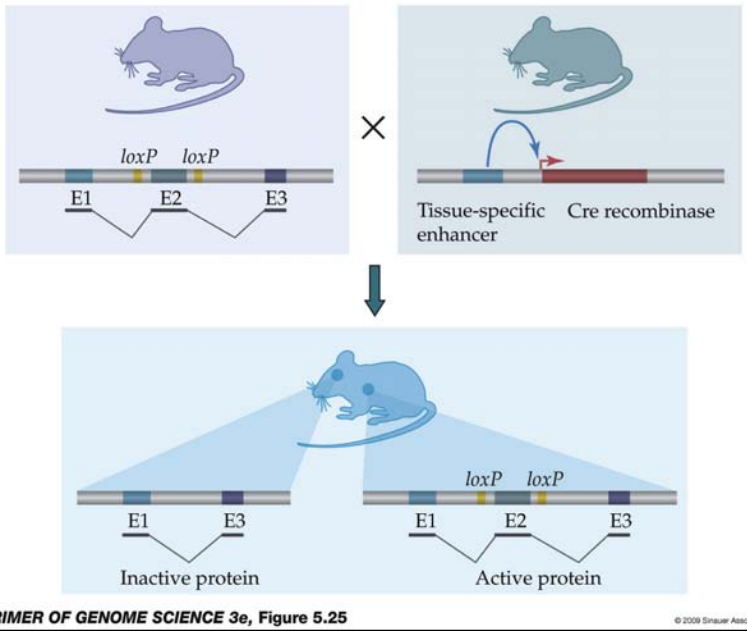
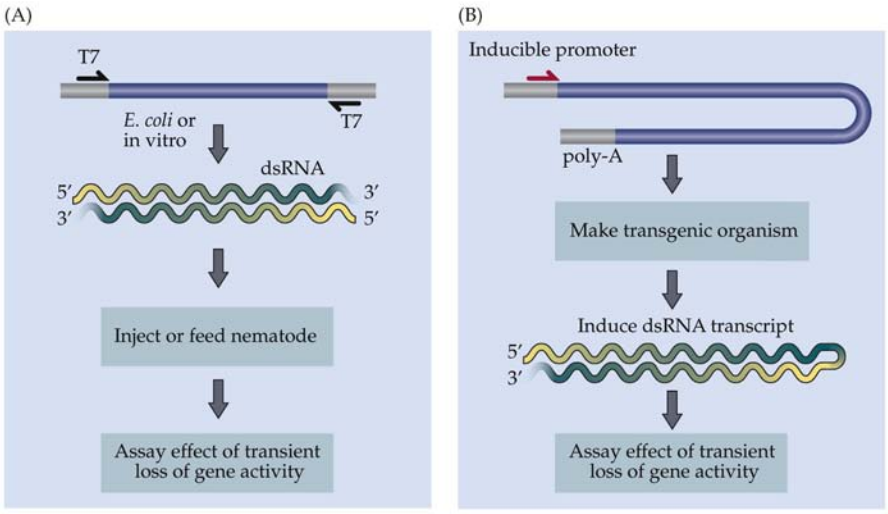


Figure 5.21 - RNA interference
(can also use opposite technique of "miRNA sponges")




*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.21



## Proteomics

- Proteomics as a complement to gene expression analysis
- Basic methods - 2d gels, affinity chromatography, immunohistochemistry
- Mass spectrometry
- Two-hybrid screens of protein binding
- Protein microarrays
- Computational prediction of protein structure & function
- Drug development as an example of an integrated genomics approach to problem solving.

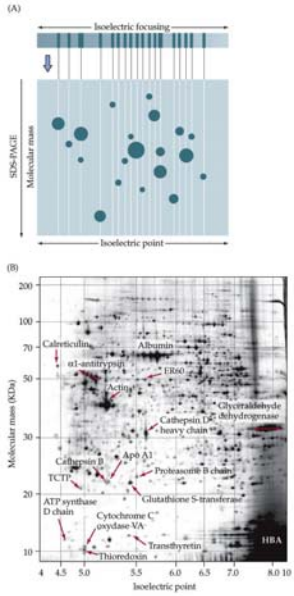
## Proteomics vs. DNA microarrays

- It has been estimated that translational + post-translational regulation accounts for about 1/3 of all gene regulation in yeast (L. Hood). These are often quantitative (1.5 fold) rather than qualitative (on or off).
- In principle, protein measurements could be more accurate because they are more directly related to the gene's biological function.
- In practice, practical issues can out-weigh the above advantages. For example, protein chips often have far fewer genes than DNA chips.
- Denaturation of proteins is irreversible.
- Post-translational modifications are complex, and may require specialized reagents adapted to specific gene products.
- The technology for working with integral membrane proteins is rudimentary at best.

## Two-dimensional gel electrophoresis

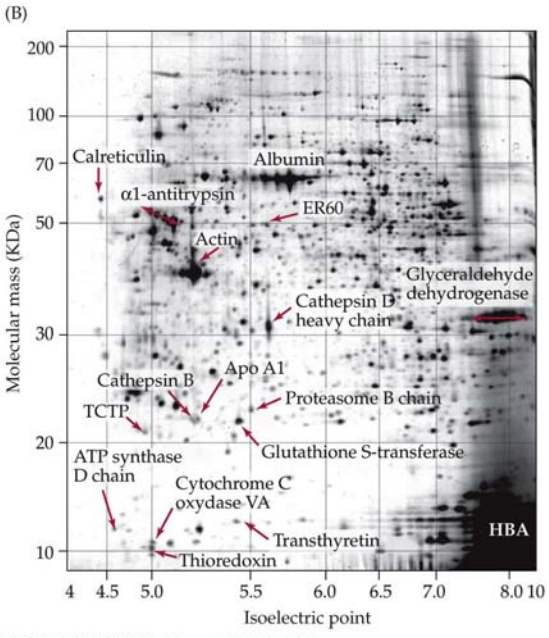
- First dimension is isoelectric focusing (separation by charge in a capillary).
- Second dimension is SDS-PAGE (separation by molecular weight in an acrylamide slab gel).
- The technique is capable of resolving several thousand proteins (this often requires multiple isoelectric focusing columns).
- Post-translational modifications are only partially resolved.
- Quantitation and sensitivity are limited.
- In principle, proteins can be identified by using specific antibodies, or specific mutants, or by obtaining N-terminal sequences from individual spots.
- In practice, amino acid sequencing from 2-d spots has given mixed results. Proteins are often blocked, impure, or not present in sufficient amounts for amino acid sequencing.

Figure 5.4 Two-dimensional polyacrylamide gel electrophoresis



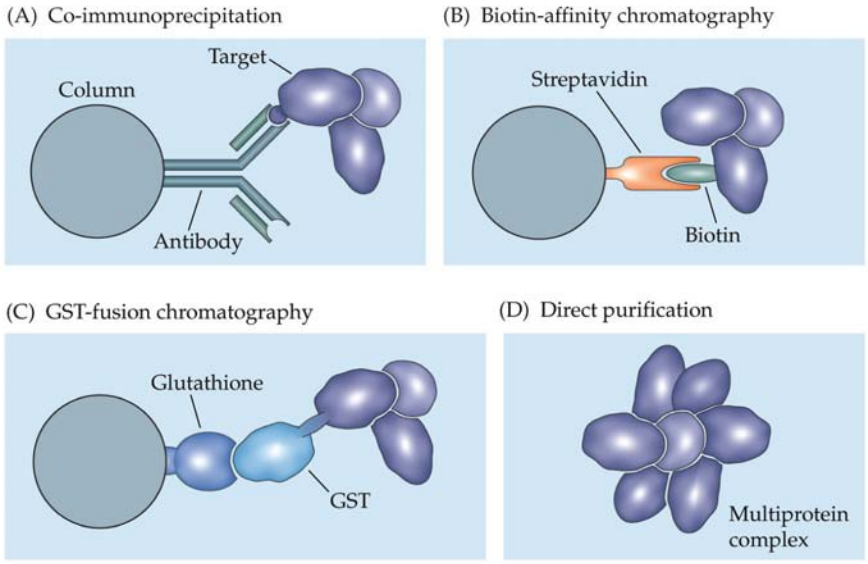

*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.4





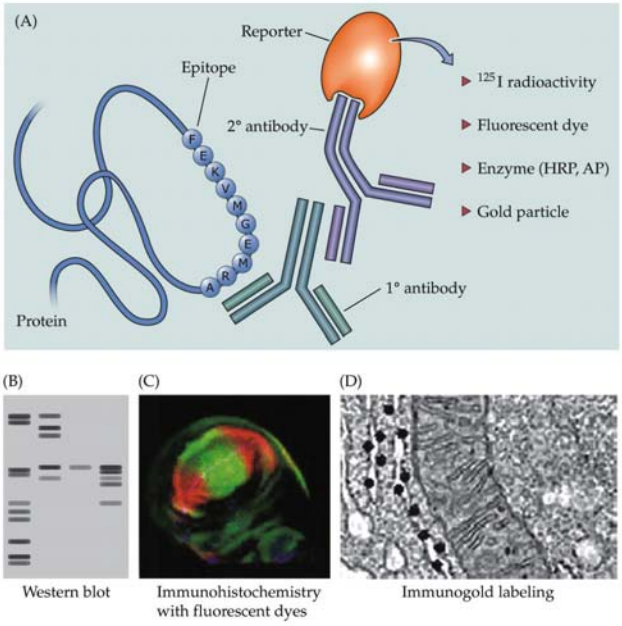

*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.4 (Part 2)

## Affinity purification of protein complexes




*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.5

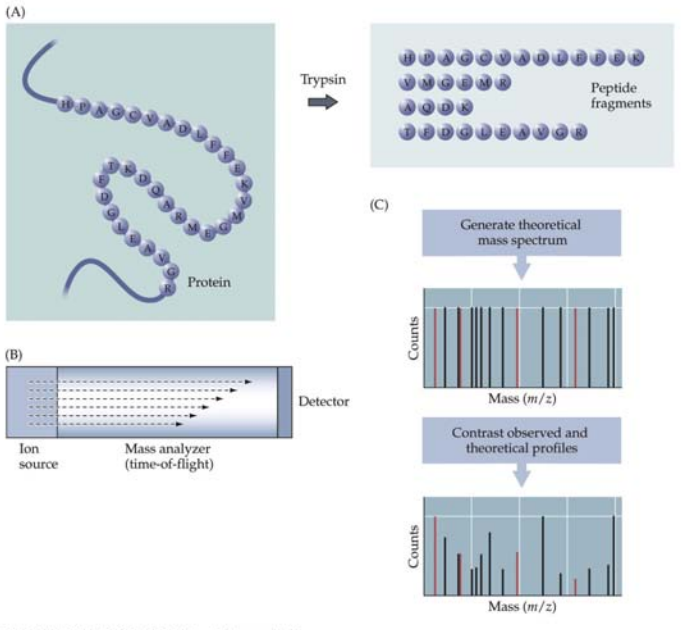





*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.9

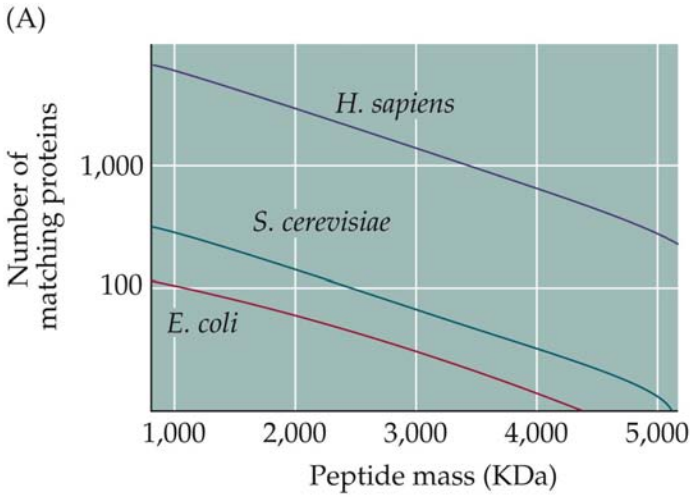
## Mass spectrometry

- In principle, ionization time-of-flight mass spectrometry provides an extremely accurate measurement of molecular weight (time of flight to detector) and abundance (charge transferred to the detector) of macromolecules, including small to medium-sized proteins.
- In practice, resolution is roughly comparable to that of 2-d gels.
- Quantitation is good but is limited by vagaries of surface chemistry and the requirements for extensive sample preparation.
- Speed and sensitivity are excellent.
- Identification of proteins are possible from the molecular weight, but post-translational modifications complicate this issue.
- Mass spec is not a preparative technique; other methods must be used to isolate specific proteins for sequencing.
- Membrane proteins are somewhat difficult to analyze.




**A PRIMER OF GENOME SCIENCE 3e, Figure 5.6**

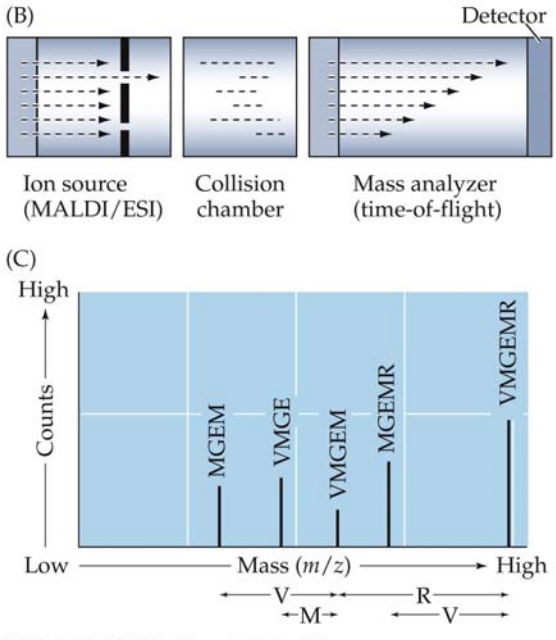
Predicted number of fragments of each size -
shows why proteomic identification is easier on small genomes




*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.7 (Part 1)



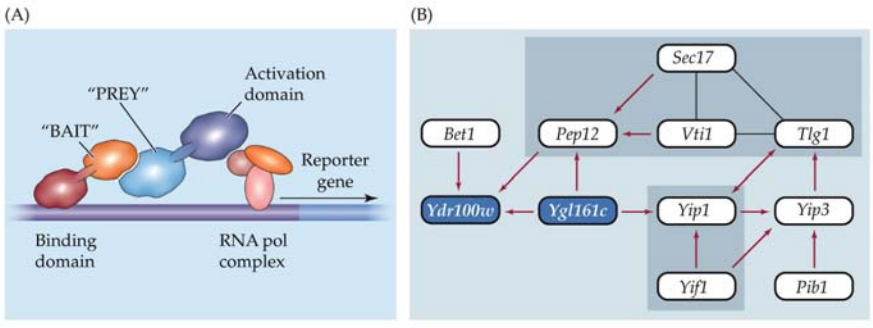
Figure 5.7 Peptide sequencing by tandem mass spectrometry




*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.7 (Part 2)

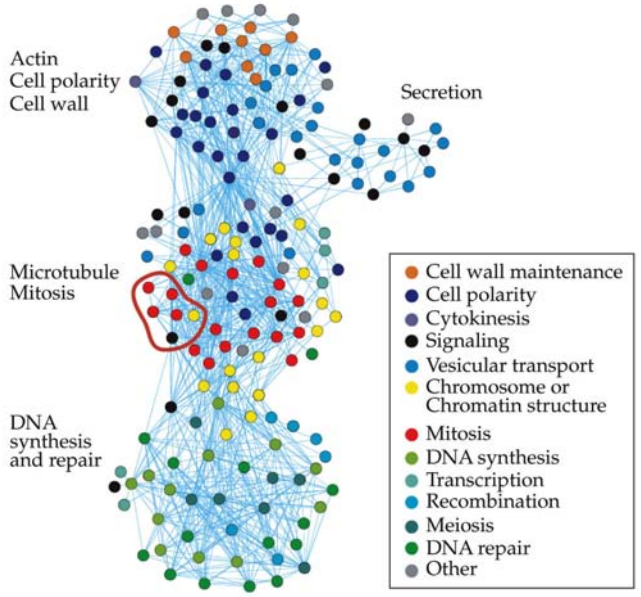
## Two-hybrid screens of protein binding

- Two-hybrid screens are based on cDNA expression libraries, in which the insert is expressed (usually in yeast) as a gene fusion to a DNA binding domain or a transcriptional activator domain.
- A second such library can be transformed into the same yeast cells (usually one clone at a time) by selecting for two antibiotics (one on each plasmid).
- Positive interactions are detected by expression of the reporter gene (such as lacZ).
- Positive colonies are selected visually and sequenced to determine the identity of the two interacting genes.
- Two-hybrid screens have been completed on a genomic scale in E. coli, yeast, and Drosophila. Similar results will also be forthcoming for vertebrate proteomes.
- Two-hybrid screens do not work well for membrane proteins, or proteins regulated by post-translational modifications.




**A PRIMER OF GENOME SCIENCE 3e, Figure 5.12**

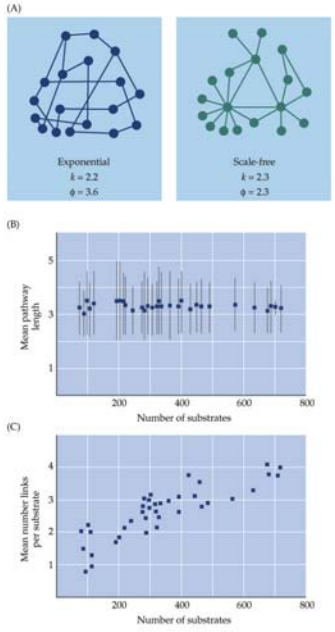
Inferring gene networks from two hybrid data
plus gene ontogeny annotations




*A PRIMER OF GENOME SCIENCE 3e*, Box 5.2, Figure C



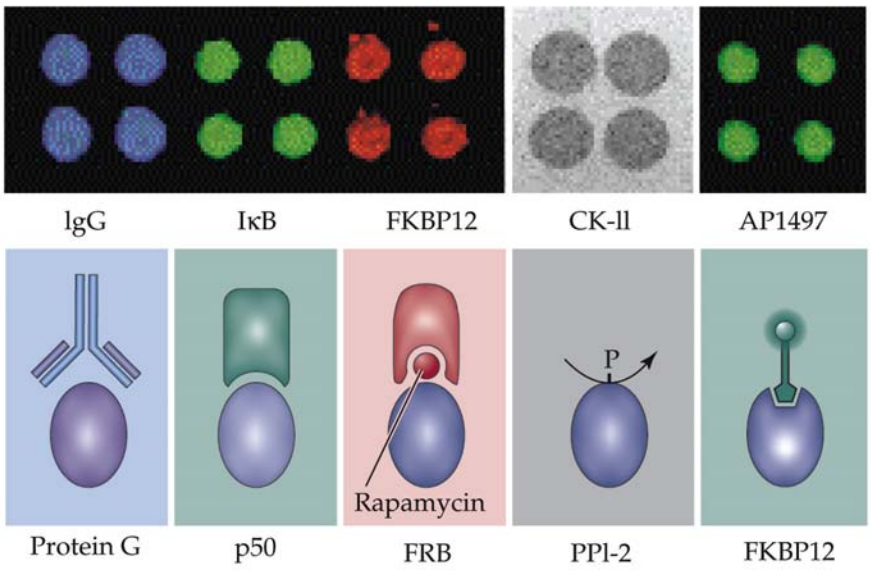
Figure 6.8 Large-scale organization of metabolic networks




*A PRIMER OF GENOME SCIENCE 3e*, Figure 6.8

## Protein microarrays

- Proteins can be printed on microarrays, and cross-linked to the substrate using techniques similar to those used to print DNA microarrays.
- The problem of protein denaturation can be minimized by selecting antibodies that are capable of binding to the denatured protein.
- Antibodies can be used to detect critical post-translational modifications (such as phosphorylation).
- The development and testing of specific monoclonal antibodies to each protein is expensive and time-consuming, but once developed they can be mass-produced.
- So far, attempts to do this on a large scale have been limited to about 2,000 proteins per slide in yeast (L. Hood), but further attempts can be expected.



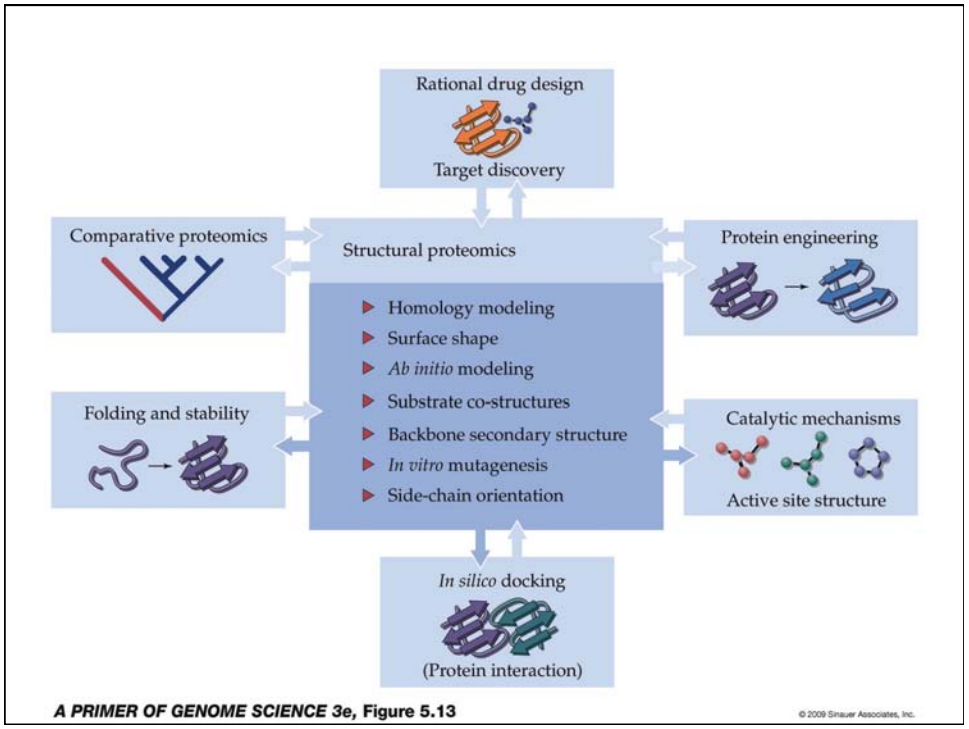

*A PRIMER OF GENOME SCIENCE 3e*, Figure 5.10

## Structure & Function Prediction

- Functions can be predicted from amino acid sequences of proteins, based on the known functions of gene family members.
- Functions can also be predicted, in a more general way, based on the presence of particular peptide motifs that occur in multiple gene families.
- Functions can be tested in transgenic mutants (mouse or fly) or by RNA interference (usually in cell culture).
- Stereotyped secondary structures (alpha helix, transmembrane segments, beta-pleated sheet) depend on local amino acid composition in a fairly well-understood way and can be predicted with about 80% confidence from amino acid sequences.
- Tertiary structures can be experimentally determined by X-ray crystallography.
- Predicting tertiary structures is still difficult, but a combination of energy minimization and phylogenetic comparison to proteins of known structure works better than either method alone.

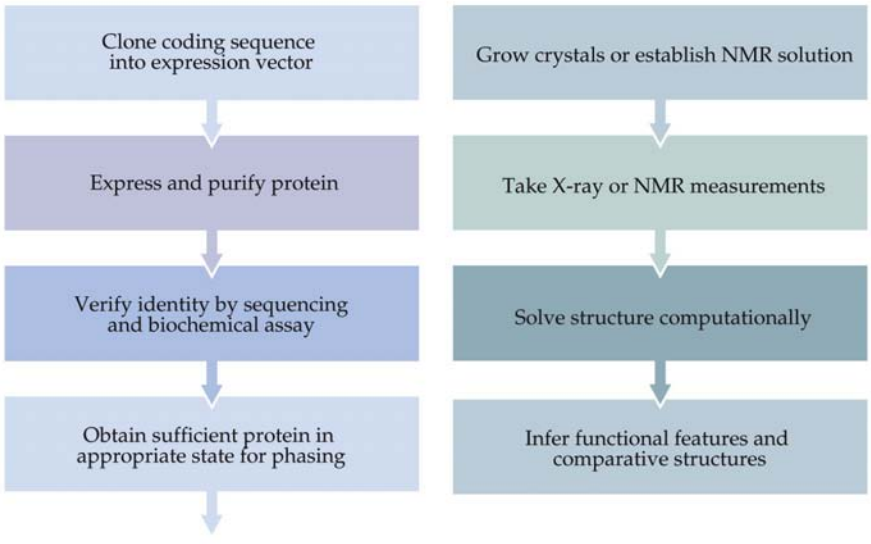
## Drug development

- The vast majority of all pharmacological drugs work by binding to a specific protein (often a G-protein coupled receptor).
- Similarly, most side effects of drugs are attributable to binding to other proteins, in addition to the desired target.
- Thus, a considerable amount of drug design is really a proteomics problem, and can (should) be assayed by proteome-level binding studies with trial drugs.
- Many of the recent generation of drugs (such as Ventolin and Viagra) were designed with the benefit of sequence data that allowed investigators to test previously unknown gene family members that would be the most likely to cause side effects.
- One recent refinement of this technique is the use of X-ray crystallography to determine the molecular structure of the candidate drug while it is bound to its correct (or incorrect) ligand (Don Abraham).
- Another refinement is the design of ligands for “orphan receptors” of unknown function (but promising parentage, such as leptin).



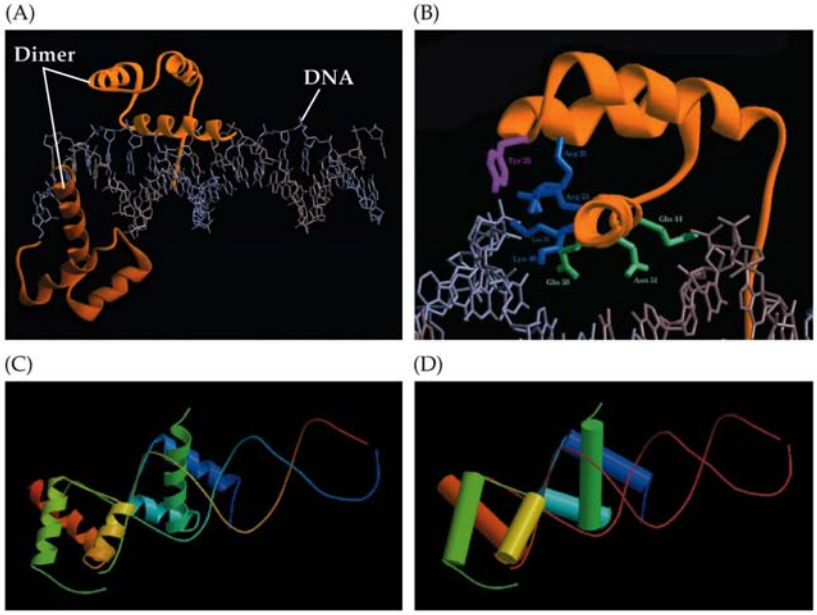

**A PRIMER OF GENOME SCIENCE 3e, Figure 5.13**





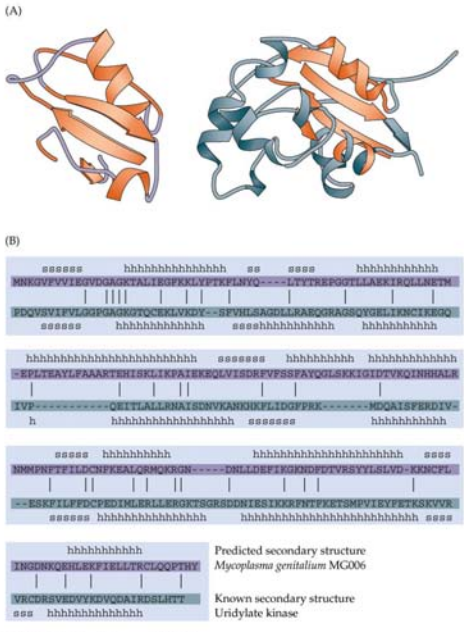

**A PRIMER OF GENOME SCIENCE 3e, Figure 5.14**

Figure 5.15 The *Drosophila* Engrailed homeodomain dimer bound to DNA

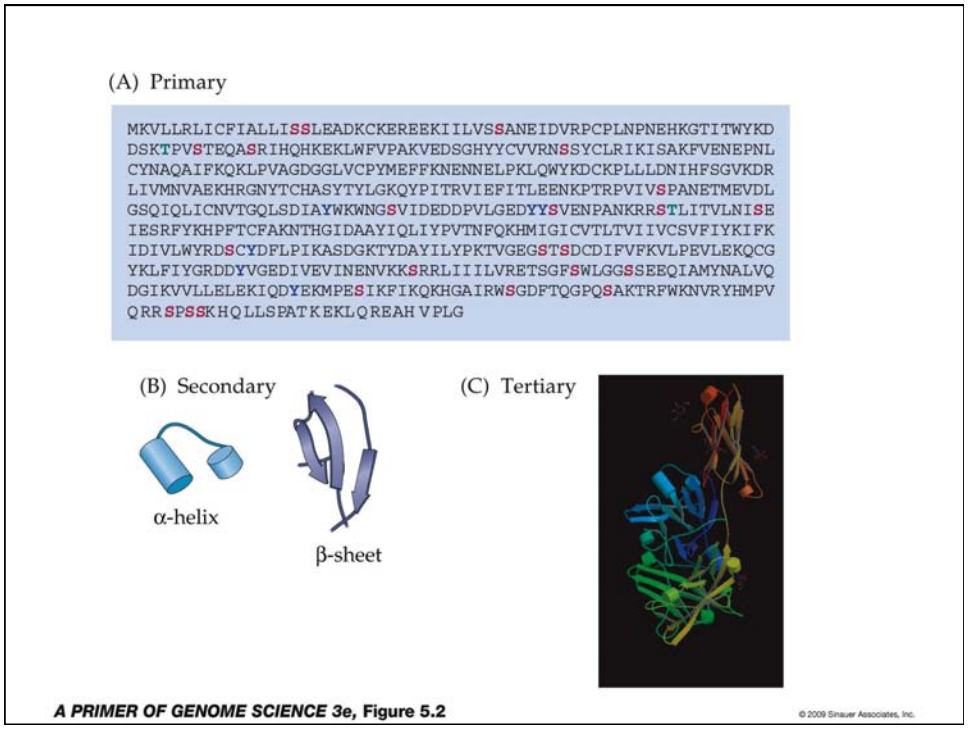



**A PRIMER OF GENOME SCIENCE 3e, Figure 5.15**



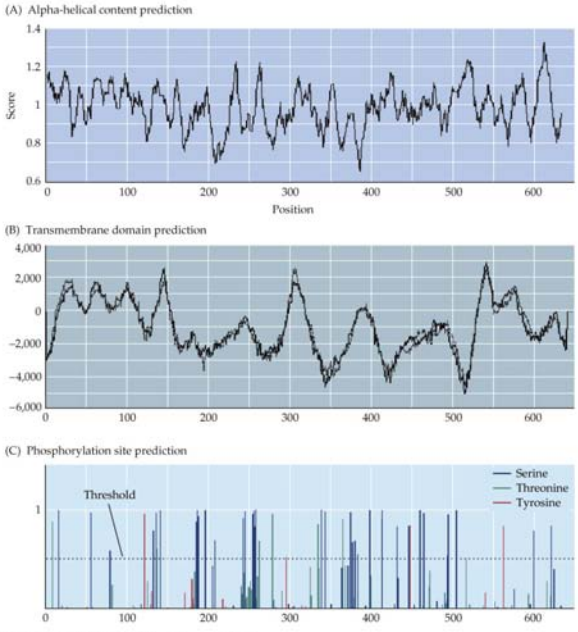



**A PRIMER OF GENOME SCIENCE 3e, Figure 5.16**

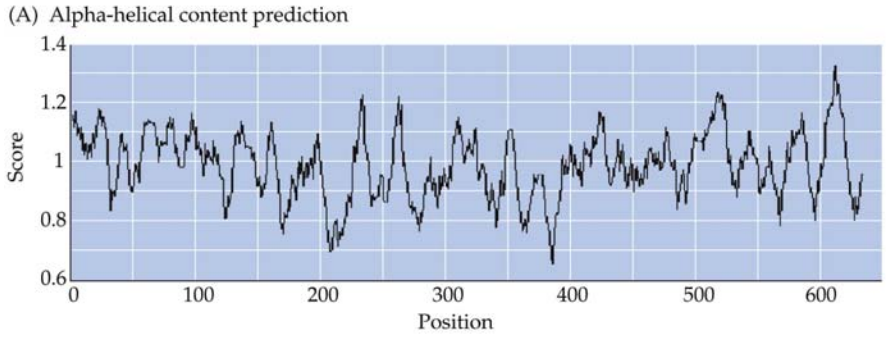



**A PRIMER OF GENOME SCIENCE 3e, Figure 5.2**



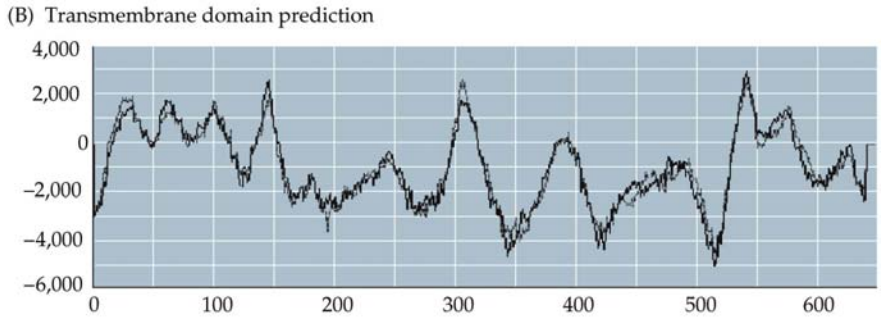



**A PRIMER OF GENOME SCIENCE 3e, Figure 5.3**

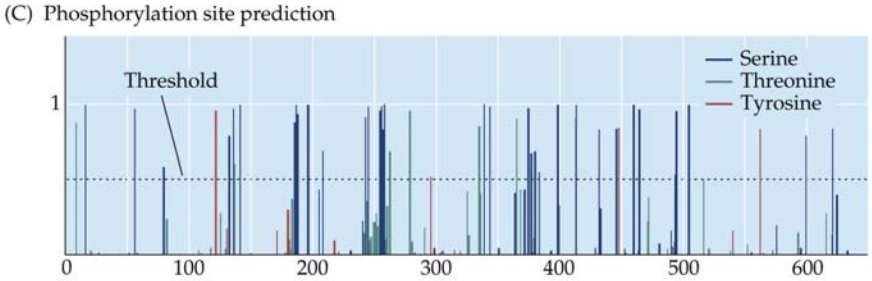



***A PRIMER OF GENOME SCIENCE 3e*, Figure 5.3 (Part 1)**






***A PRIMER OF GENOME SCIENCE 3e*, Figure 5.3 (Part 2)**




***A PRIMER OF GENOME SCIENCE 3e*, Figure 5.3 (Part 3)**



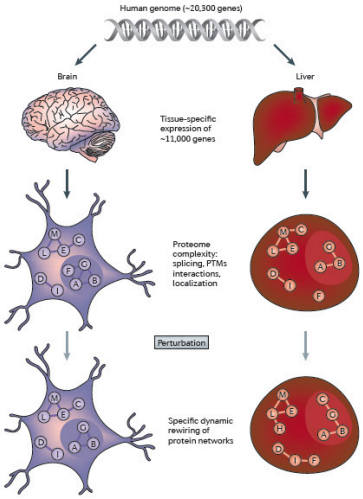
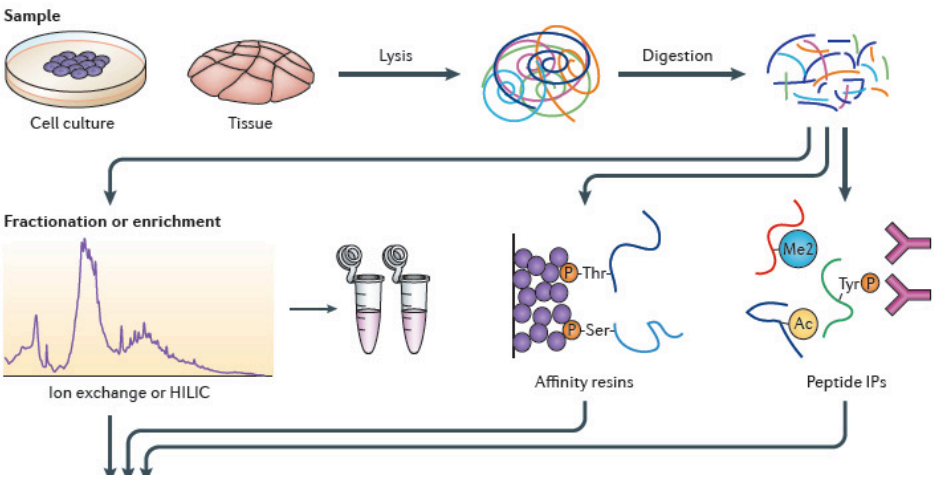



Figure 1 | **The diverse and dynamic mechanisms of proteome regulation provide a higher order of complexity to the human genome.** The human genome contains approximately 20,300 genes. The specific expression of a subset of the genome (~11,000 genes) determines the molecular backbone of the cellular phenotype (that is, the tissue cell types). A much higher order of complexity is achieved by the intricate mechanisms of protein regulation, including splicing variants, post-translational modifications (PTMs), protein–protein interactions (PPIs) and subcellular localization. This creates time-dependent tissue- and organelle-specific protein networks that respond differently to perturbations (for example, ageing or drug treatment).

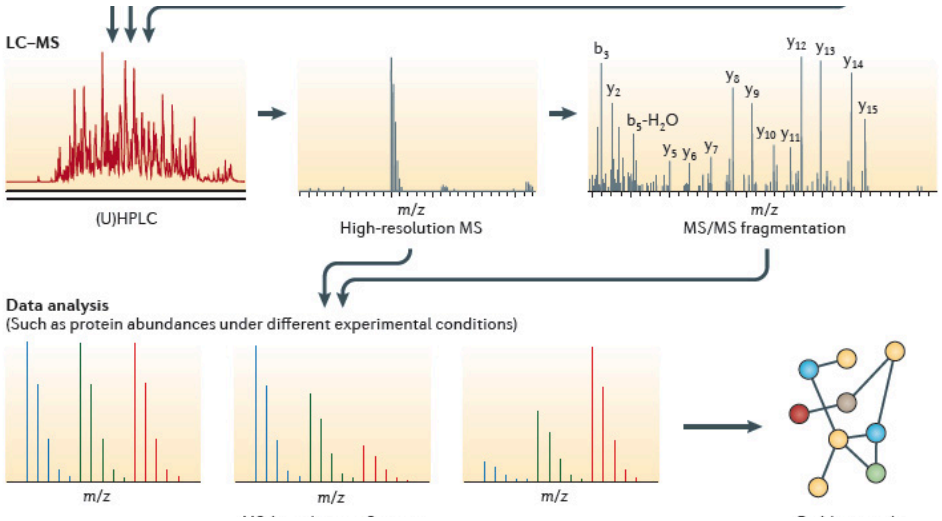
Maarten Altelaar et al. (2013) *Nat. Rev. Genet.* 14, 35-48.

## Sample preparation for mass spec

- Specific cell populations can be targeted with fluorescence-activated cell sorting (FACS) or tissue micro-dissection. A similar but more powerful method at the mRNA level is provided by "polysome tagging", based on ribosomal protein gene fusions driven by cell-type-specific promoters.
- Reduction in complexity is achieved by conventional HPLC (hydrophilic stationary phase, proteins eluted by a salt gradient), ion-exchange chromatography (charged stationary phase, eluted by a pH gradient), or reversed-phase chromatography (hydrophobic stationary phase, eluted by decreasing the polarity of the solvent, i.e. with a gradient of organic solvent concentration).
- Identification efficiency can be improved with higher mass resolution and accuracy, and/or electron transfer dissociation, but most still use defined proteolysis steps.



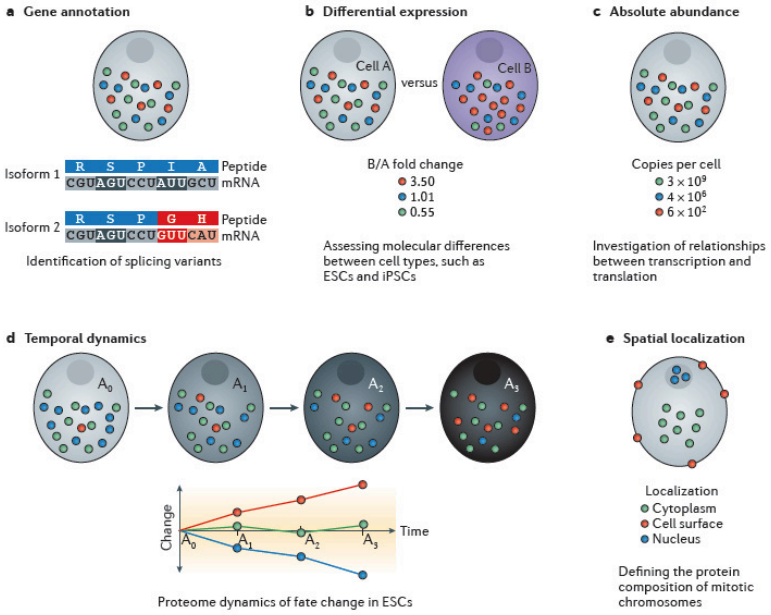

Maarten Altelaar et al. (2013)
*Nat. Rev. Genet.* 14, 35-48.



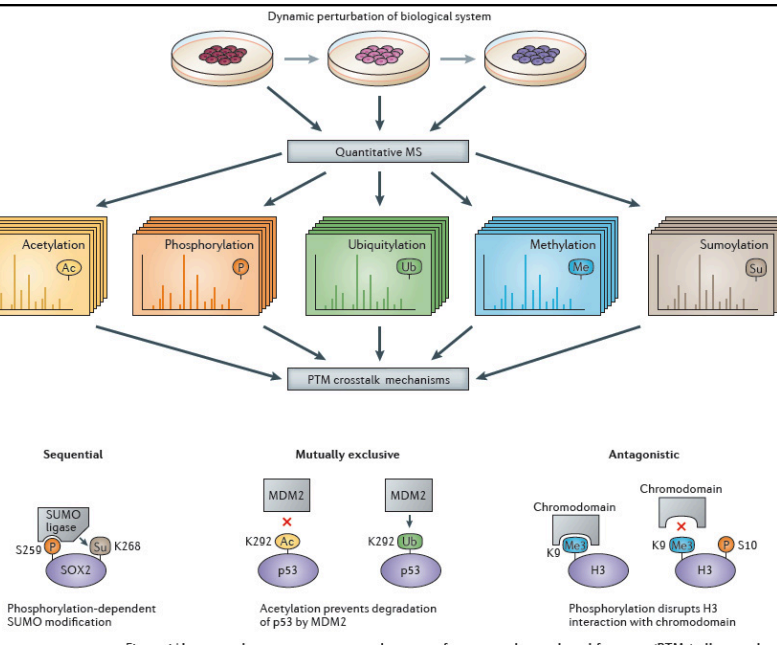

Maarten Altelaar et al. (2013)
*Nat. Rev. Genet.* 14, 35-48.

## Mass spec - data reporting and sharing

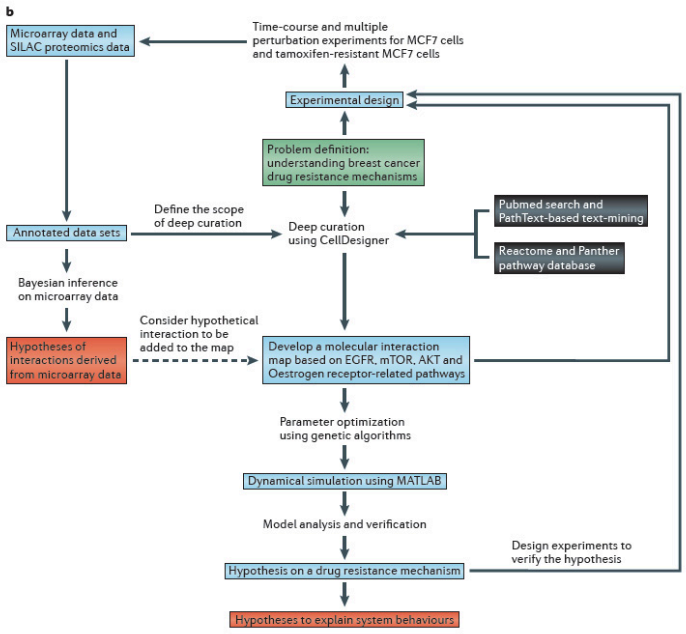
- Most raw mass spectrometry data files are presently in proprietary formats, and can not be readily shared.
- The European Bioinformatics Institute and the NCBI are working on standard (XML) formats; however the experimenter would have to translate their data before submission.
- In the meantime, conclusions (as opposed to raw data) can be presented in published data tables or supplementary files with Genbank (or Refseq) ID numbers.
- Many public databases (PhosphoSitePlus, etc) specialize in data related to specific post-translational modifications. Likewise, databases such as String, Intact and BioGRID are devoted to protein-protein interactions based on affinity purification followed by MS studies. UniProt attempts to unify proteomics databases.




Maarten Altelaar et al. (2013)
*Nat. Rev. Genet.* 14, 35-48.




Maarten Altelaar et al. (2013)
*Nat. Rev. Genet.* 14, 35-48.




Ghosh et al. (2011)
Nat. Rev. Genet. 12, 821-832.

Table 1 | **A resource matrix of software tools and data resources**

| | Tools | | Standards | | | Projects |
|---|---|---|---|---|---|---|
| | **Software** | **Resources** | **Ontologies** | **File format** | **Minimum information** | |
| **Data and knowledge management** | MAGE-TAB, ISA-TAB, KNIME, caGrid, Taverna, Bio-STEER | BioCatalogue | SBO, OBO, NCBO | MGED (MAGE), PSI, MSI | MIAME, MIAPE, MIBBI, ISO MDR, DCMI | |
| **Data-driven network inference** | R, MATLAB, BANJO | | | | | DREAM Initiative, Sage Bionetworks |
| **Deep curation** | CellDesigner, EPE, Jdesigner, PathVISIO | KEGG, Reactome, Panther pathway database, BioModels.net, WikiPathways | | SBML, SBGN, CellML, BioPAX, PSI-MI | MIRIAM | |
| ***In silico* simulation** | COPASI, SBW, JSim, Neuron, GENESIS, MATLAB, ANSYS, FreeFEM, ePNK, ina, WoPeD, Petri nets, OpenCell, CellDesigner + COPASI, CellDesigner + SOSlib, PhysioDesigner (formerly *insilico*IDE) | | | SED-ML, SBRML, PNML, SBML | MIASE | |
| **Model analysis** | MATLAB, Auto, XPPAut, BUNKI, ManLab, ByoDyn, SenSB, COBRA, MetNetMaker, DBSolve Optimum, Kintecus, NetBuilder, BooleanNet, SimBoolNet | | | | | |
| **Physiological modelling** | JSim, PhysioDesigner (formerly *insilico*IDE), CellDesigner (cellular modelling), FLAME, OpenCell, Virtual Physiology (produced by cLabs), GENESIS, Neuron, Heart Simulator, AnyBody | | | CellML, SBML, NeuroML, MML | | IUPS Physiome Project, Virtual Physiological Human, High-Definition Physiology |
| **Molecular interaction modelling** | AutoDock Vina, GOLD, eHiTS | RCSB PDB, ZINC, PubChem, PDBbind | | | | |

This table summarizes the tools and resources that correspond to each step in a systems biology workflow; please refer to FIG. 1 for an overview of the workflow and to Supplementary information S1 (table) for additional information and Weblinks to these resources.

Ghosh et al. (2011) Nat. Rev. Genet. 12, 821-832.

## Discussion questions

- Discuss the advantages and disadvantages of (1) Chip-on-chip, (2) two-hybrid screens of protein binding, and (3) MS together with HPLC and/or affinity purification, for the purpose of building models of gene interaction networks.
- Why aren't transcript and protein expression profiles always in agreement? Which is more closely related to biological function? How does this depend on which function you have in mind?
- Discuss some of the advantages and disadvantages of RNA interference (and micro RNA "sponges") as a way of manipulating gene expression.
- Discuss the use of comparative genomics, structure prediction, X-ray crystallography, and expression cloning in drug discovery and design.