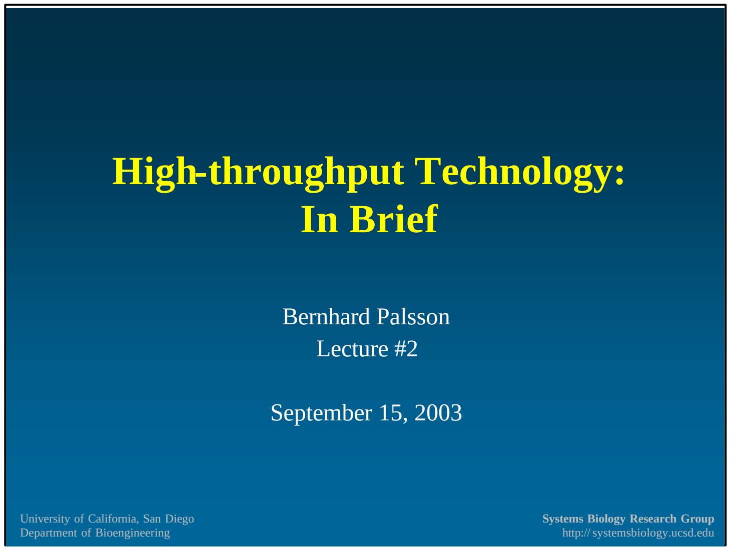
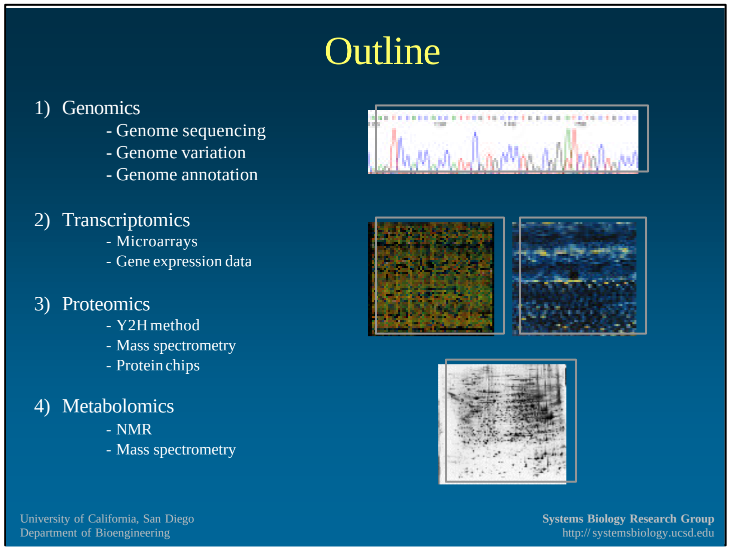


High-throughput Technology: In Brief

Bernhard Palsson
Lecture #2

September 15, 2003



Outline

1) Genomics
   - Genome sequencing
   - Genome variation
   - Genome annotation

2) Transcriptomics
   - Microarrays
   - Gene expression data

3) Proteomics
   - Y2H method
   - Mass spectrometry
   - Protein chips

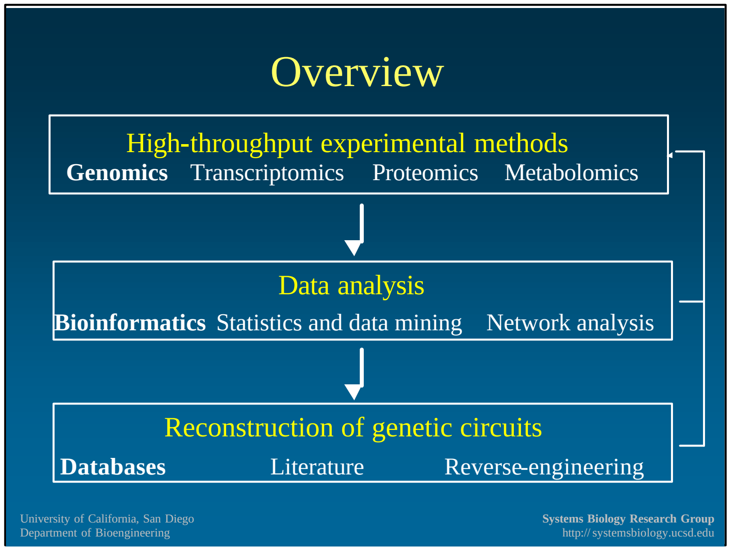
4) Metabolomics
   - NMR
   - Mass spectrometry

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

If a particular protein encoded in the genome is a metabolic enzyme, it will then catalyze a specific reaction(s). Heterogeneous data types associated with this dogma include genomic data (which defines the messages and the resulting protein sequences), transcriptomic data (which reveals the levels of messages present), proteomic data (which gives the levels of each protein present), and “fluxomic” data (which, if it existed, would provide measurements of intracellular fluxes on a complete scale).

The reaction associations shown here refer to: **ACALD + NAD -> AC + NADH** *(ALDD2x)*



Overview

High-throughput experimental methods

Genomics Transcriptomics Proteomics Metabolomics

Data analysis

Bioinformatics Statistics and data mining Network analysis

Reconstruction of genetic circuits

Databases Literature Reverse-engineering

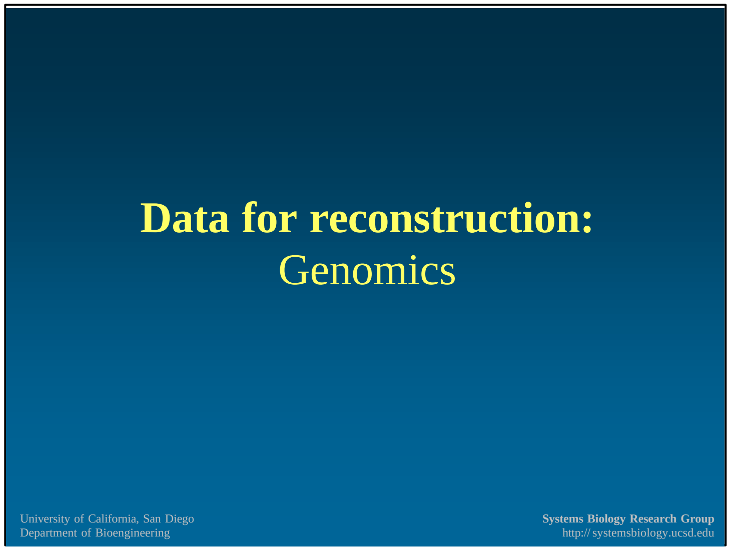
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

OVERVIEW

The utilization of “omics” data in genetic circuit reconstruction can be conceptually organized in three stages:

1) Generating the data using high-throughput experiments including genomics, transcriptomics, proteomics, and metabolomics methods
2) Analyzing the raw data using various bioinformatics, statistics, and data mining methods
3) Reconstruction of genetic circuits using results from the data analysis process that are typically either deposited in databases or published in research papers. In addition for some data types (transcriptomics) it is possible to use the data directly to reverse-engineer (or “back-calculate”) genetic circuits to a certain degree.

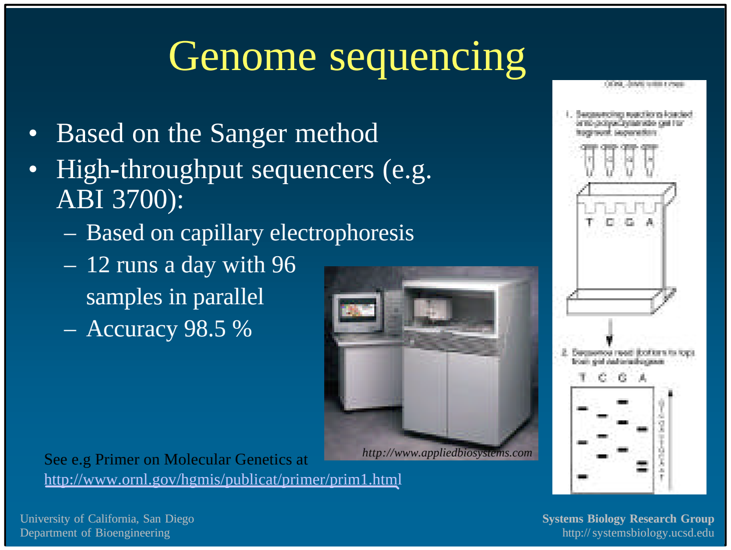


Data for reconstruction:
Genomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

The following five topics will describe in more detail high-throughput experimental methods, how data from these experiments is analyzed, and how this data can be utilized in reconstructing genetic circuits. The focus in this first topic will be on genomics - genome sequencing, annotation and analysis. The discussion will mainly be at a fairly high level and the emphasis will be in conceptual understanding of the experimental techniques and computational methods. Of particular interest will be the kinds of errors in databases that could occur due to problems in sequencing or annotation.



Genome sequencing

- Based on the Sanger method
- High-throughput sequencers (e.g. ABI 3700):
  - Based on capillary electrophoresis
  - 12 runs a day with 96 samples in parallel
  - Accuracy 98.5 %

http://www.appliedbiosystems.com

See e.g Primer on Molecular Genetics at http://www.ornl.gov/hgmis/publicat/primer/prim1.html

1. Sequencing reactions loaded onto polyacrylamide gel for fragment separation
T C G A
2. Sequence read (bottom to top) from gel autoradiogram
T C G A

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

GENOME SEQUENCING

Modern sequencer such as the ABI 3700 automate and multiplex the Sanger method so that it can be utilized to sequence whole genomes. An important advance is the use of capillary electrophoresis instead of gel electrophoresis, which removes the problem of gel lane migration. The ABI 3700 has the capacity to run 12 runs a day with 96 samples of ~500 nucleotides long amplified DNA fragments in parallel. This results in a nominal sequencing capacity of 576 kb a day (the whole human genome is of the order of 3 billion bases). The reported accuracy of the ABI 3700 is 98.5 % indicating that there are less than 2 errors per 100 bases sequenced.

Figure caption: Sanger method for DNA sequencing

Dideoxy sequencing (also called chain- termination or Sanger method) uses an enzymatic procedure to synthesize DNA chains of varying lengths, stopping DNA replication at one of the four bases and then determining the resulting fragment lengths. Each sequencing reaction tube (T, C, G, and A) in the diagram contains

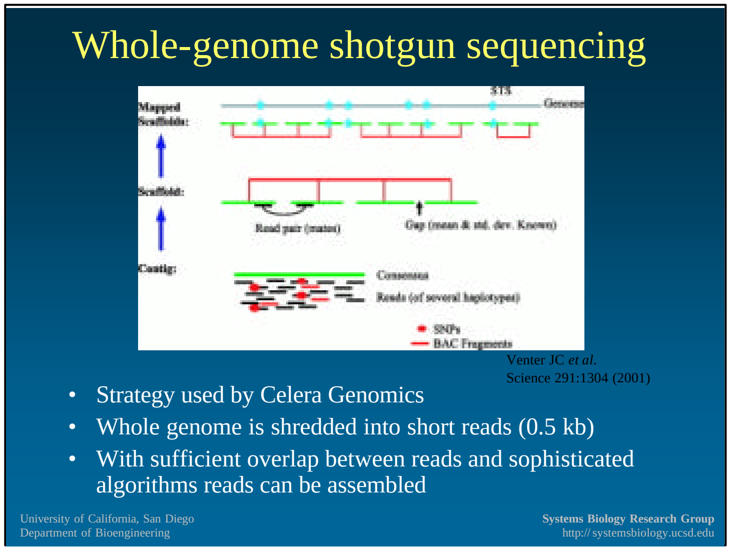
•a DNA template, a primer sequence, and a DNA polymerase to initiate synthesis of a new strand of DNA at the point where the primer is hybridized to the template;

•the four deoxynucleotide triphosphates (dATP, dTTP, dCTP, and dGTP) to extend the DNA strand;

•one labeled deoxynucleotide triphosphate (using a radioactive element or dye); and

•one dideoxynucleotide triphosphate, which terminates the growing chain wherever it is incorporated. Tube A has didATP, tube C has didCTP, etc.

For example, in the A reaction tube the ratio of the dATP to didATP is adjusted so that each tube will have a collection of DNA fragments with a didATP incorporated for each adenine position on the template DNA fragments. The fragments of varying length are then separated by electrophoresis (1) and the positions of the nucleotides analyzed to determine sequence. The fragments are separated on the basis of size, with the shorter fragments moving faster and appearing at the bottom of the gel. Sequence is read from bottom to top (2).



Whole-genome shotgun sequencing

Venter JC *et al.*
Science 291:1304 (2001)

- Strategy used by Celera Genomics
- Whole genome is shredded into short reads (0.5 kb)
- With sufficient overlap between reads and sophisticated algorithms reads can be assembled

University of California, San Diego
Department of Bioengineering

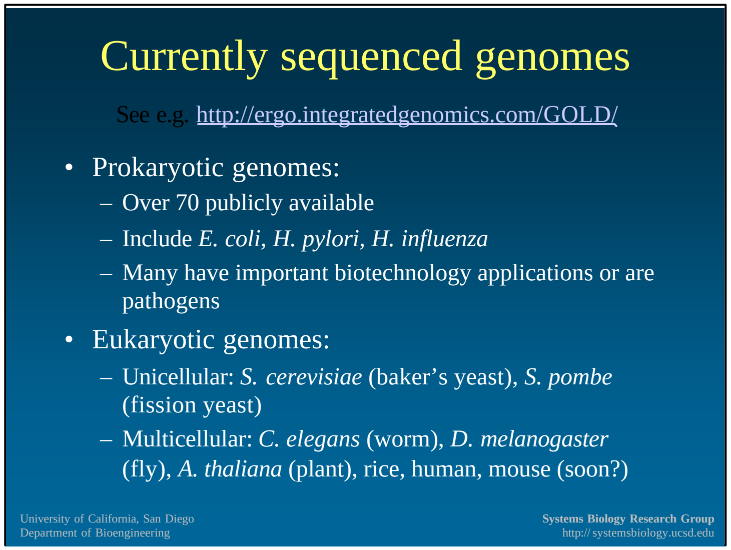
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

WHOLE-GENOME SHOTGUN SEQUENCING

Figure caption:

Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

--
Science. 2001 Feb 16;291(5507):1304-51.
The sequence of the human genome.
Venter JC et al. (Celera Genomics)



Currently sequenced genomes

See e.g. http://ergo.integratedgenomics.com/GOLD/

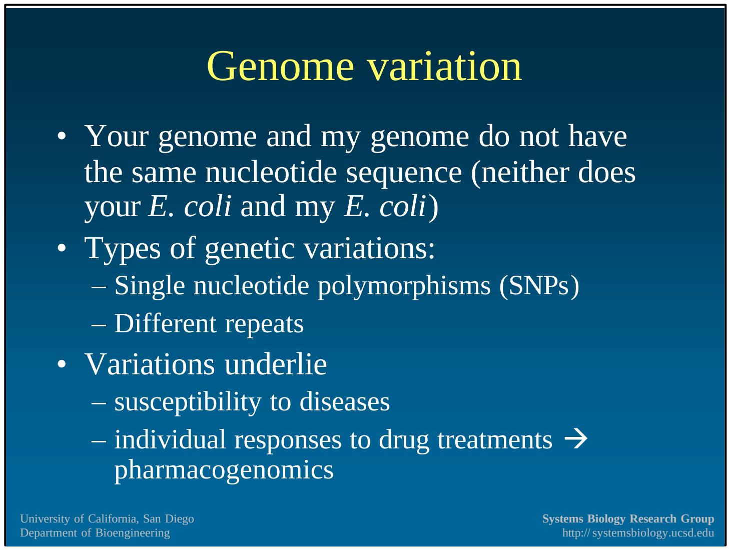
- Prokaryotic genomes:
  - Over 70 publicly available
  - Include *E. coli*, *H. pylori*, *H. influenza*
  - Many have important biotechnology applications or are pathogens
- Eukaryotic genomes:
  - Unicellular: *S. cerevisiae* (baker’s yeast), *S. pombe* (fission yeast)
  - Multicellular: *C. elegans* (worm), *D. melanogaster* (fly), *A. thaliana* (plant), rice, human, mouse (soon?)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

CURRENTLY SEQUENCED GENOMES

There are currently 531 known genome sequencing projects, 82 published complete genomes, 271 ongoing prokaryotic sequencing projects, and 178 eukaryotic sequencing projects (from GOLD on April 9th 2002). Published prokaryotic genomes include *E. coli*, *H. pylori*, and *H. influenza* whose metabolic networks have been reconstructed in the Genetic Circuits group at UCSD. Many of the sequenced prokaryotes are either pathogens (such as *H. pylori*) or have biotechnology applications (such as *E. coli*). Much fewer eukaryotic genomes have been sequenced – currently only genomes for prominent model organisms (the yeasts, worm, fly and *A. thaliana*) are available. And of course we have the two different sequences of the human genome available as well. The latest addition to the eukaryotic genomes is rice, whose genome sequence (for two different subspecies) was published in the April 5th issue of the journal Science (Yu J *et al.* A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*) Science 296:79 (2002) and Goff SA *et al.* A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*) Science 296:92 (2002)). The mouse genome sequencing effort is largely complete (both public and Celera), but the assembly and annotation are still under way.



Genome variation

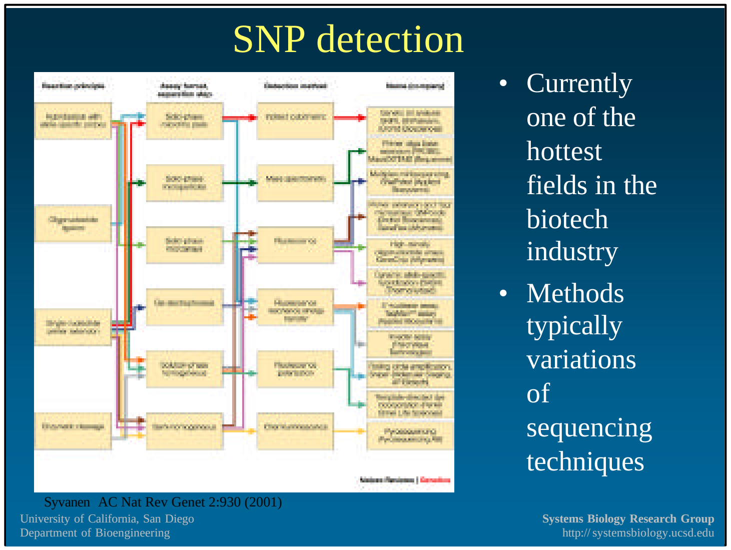
- Your genome and my genome do not have the same nucleotide sequence (neither does your *E. coli* and my *E. coli*)
- Types of genetic variations:
  - Single nucleotide polymorphisms (SNPs)
  - Different repeats
- Variations underlie
  - susceptibility to diseases
  - individual responses to drug treatments → pharmacogenomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

GENOME VARIATION

Once the genome of an organism is sequenced the next important task is the study of genetic or genome variation between individuals of the same species. This refers to both human genome variation and microbial genome variation. The types of variations that are commonly considered include single nucleotide polymorphisms (SNPs) and different types of repeats. Here we focus on human SNPs, which are defined as single base variations between individuals that occur at high enough frequency in a population to be considered to be non-random. The reason for interest in genomic variation is that these variations are a large part of what determines the difference between individuals especially when it comes to susceptibility to various diseases and responses to drug treatments. The latter aspect is considered to be important enough to be awarded its own name – pharmacogenomics.


Reaction principle
Assay format, separation step
Detection method
Name (company)
Hybridization with allele-specific probes
Oligonucleotide ligation
Single-nucleotide primer extension
Enzymatic cleavage
Solid-phase microtitre plate
Solid-phase microparticles
Solid-phase microarrays
Gel electrophoresis
Solution-phase homogeneous
Semi-homogeneous
Indirect colorimetric
Mass spectrometry
Fluorescence
Fluorescence resonance energy transfer
Fluorescence polarization
Chemiluminescence
Pyrosequencing (Pyrosequencing AB)
Nature Reviews | Genetics

SNP detection

- Currently one of the hottest fields in the biotech industry
- Methods typically variations of sequencing techniques

Syvanen AC Nat Rev Genet 2:930 (2001)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

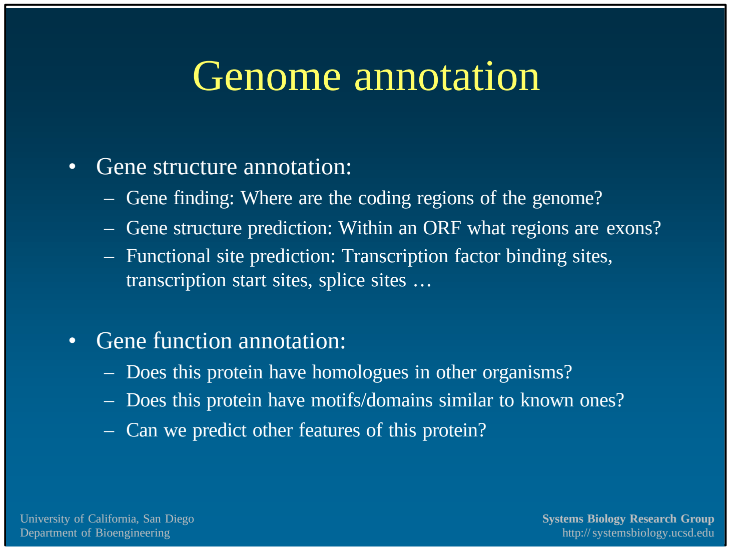
SNP DETECTION

Figure caption

Coloured arrows are used to show the reaction principles, assay format and detection methods that make up a particular genotyping method. For example, the TaqMan^{TM} assay involves hybridization with allele-specific oligonucleotides, a solution-phase assay and detection by fluorescence resonance energy transfer. The figure illustrates principles for assay design, and the list of assays is not intended to be comprehensive.

--

Nat Rev Genet. 2001 Dec;2(12):930-42.
Accessing genetic variation: genotyping single nucleotide polymorphisms.
Syvanen AC



Genome annotation

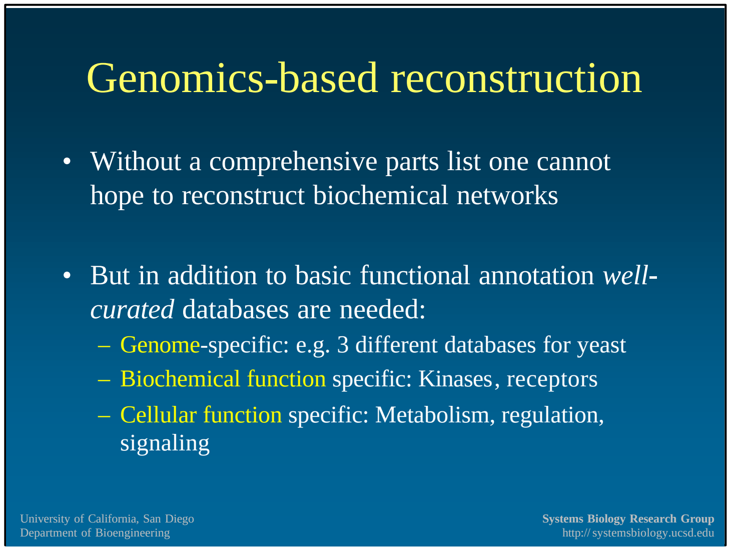
- Gene structure annotation:
  - Gene finding: Where are the coding regions of the genome?
  - Gene structure prediction: Within an ORF what regions are exons?
  - Functional site prediction: Transcription factor binding sites, transcription start sites, splice sites …
- Gene function annotation:
  - Does this protein have homologues in other organisms?
  - Does this protein have motifs/domains similar to known ones?
  - Can we predict other features of this protein?

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

GENOME ANNOTATION

Once a genome has been sequenced and assembled into reasonably large contigs the next task is to annotate the genome. This task can be roughly divided into two halves – annotation of gene structure (not the same thing as protein structure) and functional annotation. The most prominent task in gene structure annotation is finding all ORF's (open reading frame = potential gene coding region) in the genome. In addition to just knowing where a potential ORF begins and ends in eukaryotes one also needs to know where the exons (expressed regions) and introns within the ORF are. This can be a major difficulty since typically in higher eukaryotes there are tens of exons per gene and exons are very small compared to introns. Once a gene has been identified and its exon/intron structure determined the next step is to figure out what the function of the gene is (this of course assuming that this particular gene hasn't been already cloned and studied in detail in this particular organism). Functional annotation is usually done through sequence comparisons to protein sequence databases or protein domain/motif databases.



Genomics-based reconstruction

- Without a comprehensive parts list one cannot hope to reconstruct biochemical networks
- But in addition to basic functional annotation *well-curated* databases are needed:
  - Genome-specific: e.g. 3 different databases for yeast
  - Biochemical function specific: Kinases, receptors
  - Cellular function specific: Metabolism, regulation, signaling

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

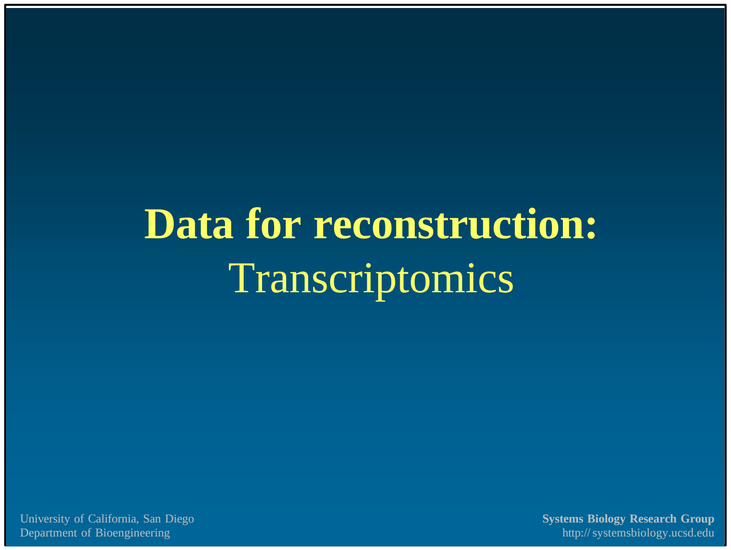
GENOMICS-BASED RECONSTRUCTION

All the annotation methods described in previous slides (both sequence homology-based and genome context-based) primarily aim to construct a comprehensive parts list of all the genes and proteins in an organism. This parts list forms the basis of genome-scale genetic circuit reconstruction, since obviously without knowing the parts its difficult to put together the whole circuitry. However, as already previously mentioned the largely automated genome annotation methodology described above has to be augmented with manual curation aimed at constructing databases specific to particular aspects of the biological system studied. Typical databases of this kind are

•Genome/organism-specific databases, which provide detailed information about the genes/proteins of a particular organism

•Biochemical function specific databases that focus on particular class of proteins and the specific aspects of this class (e.g. kinase substrate specificity)

•Cellular function-specific databases, which are the most useful ones for reconstruction purposes and will be described in detail later

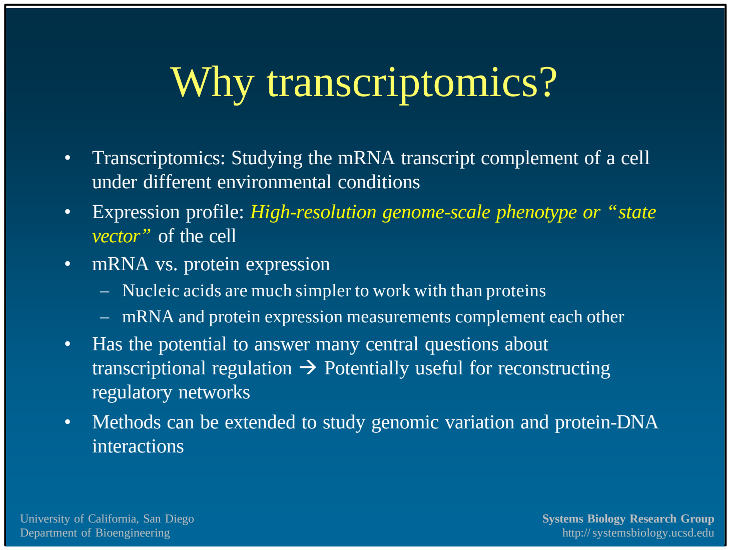


Data for reconstruction:
Transcriptomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

In this topic we continue the exploration of the ‘omics’ technologies and their relationship to reconstruction of genetic circuits. After genomics transcriptomics is probably the best developed of the different high-throughput technologies. Most of the high-throughput transcription profiling techniques were developed before the mid-nineties and these techniques are now used routinely by thousands of labs worldwide.



Why transcriptomics?

- Transcriptomics: Studying the mRNA transcript complement of a cell under different environmental conditions
- Expression profile: *High-resolution genome-scale phenotype or "state vector"* of the cell
- mRNA vs. protein expression
  - Nucleic acids are much simpler to work with than proteins
  - mRNA and protein expression measurements complement each other
- Has the potential to answer many central questions about transcriptional regulation → Potentially useful for reconstructing regulatory networks
- Methods can be extended to study genomic variation and protein-DNA interactions

University of California, San Diego
Department of Bioengineering

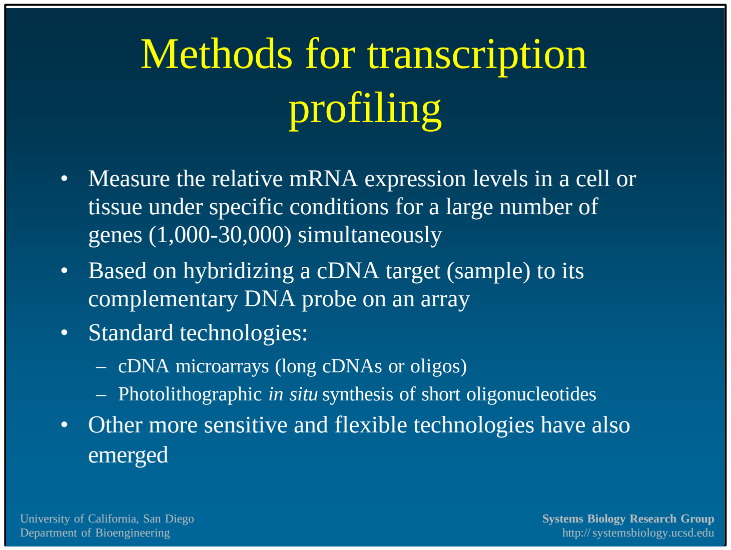
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

MOTIVATION FOR TRANSCRIPTOMICS

Transcriptomics could be defined as the study of the expressed mRNA transcript complement of a cell under different conditions. The central quantity in transcriptomics is the gene (or mRNA) expression profile of the cell. This profile can be considered to be a high-resolution (in terms of being at the single gene level) genome-scale phenotype (for biologist) or state vector (for engineers) of the cell.

While mRNAs do not play as important a role in cellular function as proteins, there are a number of reasons why one might prefer doing mRNA expression profiling as opposed to protein expression profiling. The principal reason is quite practical though – nucleic acids (such as mRNA) are much easier to separate, purify, detect and quantify than proteins (more on this in the proteomics topic). Also since protein concentrations can be considered to be integrals of mRNA concentrations, the variability at the mRNA level is usually larger than the variability at the protein level. A third reason is simply that mRNA and protein expression measurements complement each other.

The major attraction in transcriptomics is that the ability to measure mRNA concentrations of all genes under any condition allows studying regulation of gene expression at a genome-wide scale. This makes transcriptomics an essential component of the regulatory network reconstruction process. Even more importantly for this aim transcription profiling techniques have been extended to study protein-DNA interactions such as genome-wide detection of transcription factor binding sites.



Methods for transcription profiling

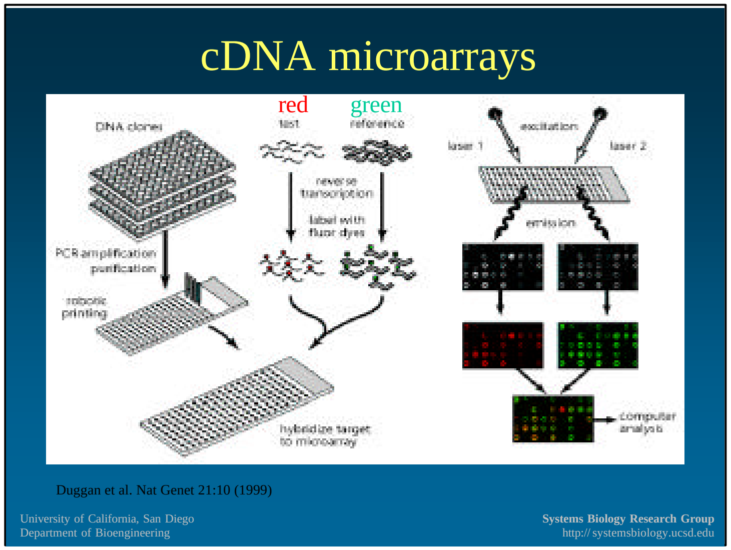
- Measure the relative mRNA expression levels in a cell or tissue under specific conditions for a large number of genes (1,000-30,000) simultaneously
- Based on hybridizing a cDNA target (sample) to its complementary DNA probe on an array
- Standard technologies:
  - cDNA microarrays (long cDNAs or oligos)
  - Photolithographic *in situ* synthesis of short oligonucleotides
- Other more sensitive and flexible technologies have also emerged

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

METHODS FOR TRANSCRIPTION PROFILING

The basic idea in transcription profiling is to measure (usually relative) mRNA expression levels of thousands of genes simultaneously in a cell or tissue sample under specific conditions. All transcription profiling techniques are based on the process of hybridization, in which a cDNA target from the sample to be studied is hybridized to its complementary single stranded DNA probe on an array. The target cDNA is created by extracting all mRNA from a sample, reverse transcribing the mRNAs to cDNAs, and simultaneously labeling the resulting cDNAs with a dye so that they can be detected and quantified. The two standard technologies for transcription profiling are cDNA microarrays where the DNA probe on the array is a long cDNA, and Affymetrix Gene Chips where the probe on the array is a short oligonucleotide. These two technologies will be discussed in more detail in the following two slides. In addition the these major techniques there are a number of more sensitive and flexible technologies that have been developed in recent years.



cDNA microarrays

Duggan et al. Nat Genet 21:10 (1999)

University of California, San Diego
Department of Bioengineering

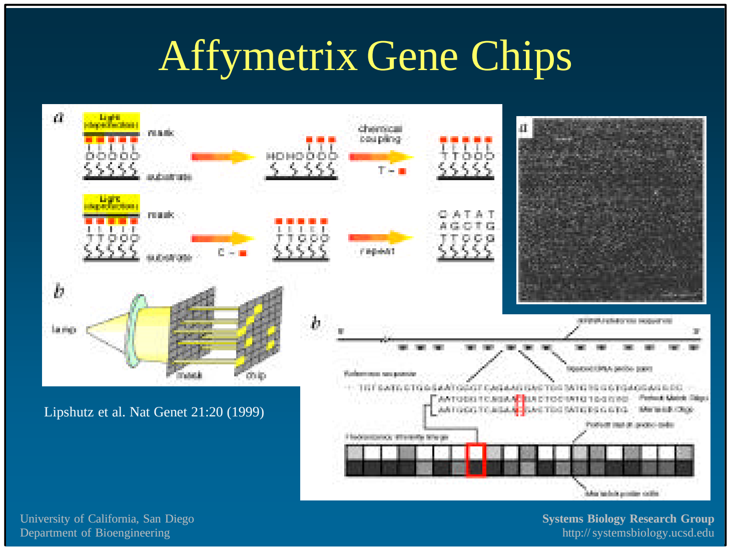
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

cDNA MICROARRAYS

Nat Genet 1999 Jan;21(1 Suppl):10-4
Expression profiling using cDNA microarrays.
Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM.

Abstract: cDNA microarrays are capable of profiling gene expression patterns of tens of thousands of genes in a single experiment. DNA targets, in the form of 3' expressed sequence tags (ESTs), are arrayed onto glass slides (or membranes) and probed with fluorescent- or radioactively-labelled cDNAs. Here, we review technical aspects of cDNA microarrays, including the general principles, fabrication of the arrays, target labelling, image analysis and data extraction, management and mining.

Figure caption: cDNA microarray schema
Templates for genes of interest are obtained and amplified by PCR. Following purification and quality control, aliquots (5 nl) are printed on coated glass microscope slides using a computer–controlled, high–speed robot. Total RNA from both the test and reference sample is fluorescently labelled with either Cye3– or Cye5–dUTP using a single round of reverse transcription. The fluorescent targets are pooled and allowed to hybridize under stringent conditions to the clones on the array. Laser excitation of the incorporated targets yields an emission with a characteristic spectra, which is measured using a scanning confocal laser microscope. Monochrome images from the scanner are imported into software in which the images are pseudo–coloured and merged. Information about the clones, including gene name, clone identifier, intensity values, intensity ratios, normalization constant and confidence intervals, is attached to each target. Data from a single hybridization experiment is viewed as a normalized ratio (that is, Cye3/Cye5) in which significant deviations from 1 (no change) are indicative of increased (>1) or decreased (<1) levels of gene expression relative to the reference sample. In addition, data from multiple experiments can be examined using any number of data mining tools.



Affymetrix Gene Chips

Lipshutz et al. Nat Genet 21:20 (1999)

University of California, San Diego
Department of Bioengineering

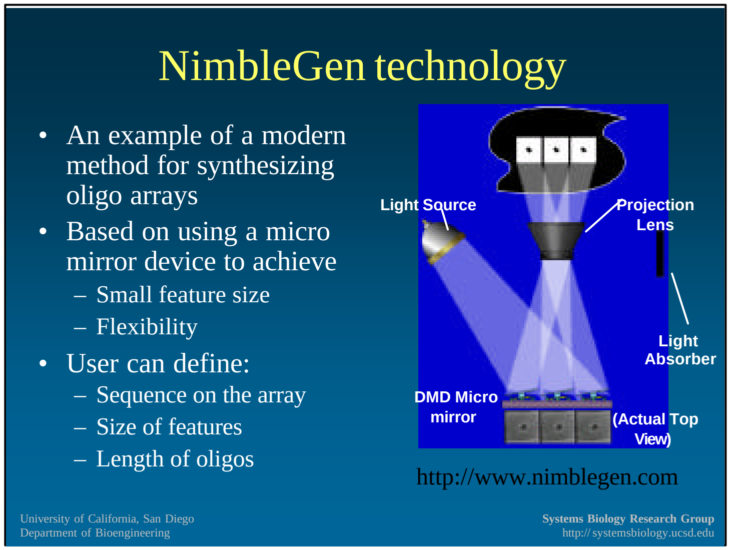
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

AFFYMETRIX GENE CHIPS

Nat Genet 1999 Jan;21(1 Suppl):20-4
High density synthetic oligonucleotide arrays.
Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ.

Abstract: Experimental genomics involves taking advantage of sequence information to investigate and understand the workings of genes, cells and organisms. We have developed an approach in which sequence information is used directly to design high-density, two-dimensional rays of synthetic oligonucleotides. The GeneChipe probe arrays are made using spatially patterned, light-directed combinatorial chemical synthesis and contain up to hundreds of thousands of different oligonucleotides on a small glass surface. The arrays have been designed and used for quantitative and highly parallel measurements of gene expression, to discover polymorphic loci and to detect the presence of thousands of alternative alleles. Here, we describe the fabrication of the arrays, their design and some specific applications to high-throughput genetic and cellular analysis.

Figure captions:
Light directed oligonucleotide synthesis.
a) A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group. Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites. The process is repeated, activating different sets of sites and coupling different bases allowing arbitrary DNA probes to be constructed at each site. **b)** Schematic representation of the lamp, mask and array.
Gene expression monitoring with oligonucleotide arrays.
a) A single 1.28 1.28 cm array containing probe sets for approximately 40,000 human genes and ESTs. This array contains features smaller than 22 22 m and only four probe pairs per gene or EST. **b)** Expression probe and array design. Oligonucleotide probes are chosen based on uniqueness criteria and composition design rules. For eukaryotic organisms, probes are chosen typically from the 3´ end of the gene or transcript (nearer to the poly(A) tail) to reduce problems that may arise from the use of partially degraded mRNA. The use of the PM minus MM differences averaged across a set of



NimbleGen technology

- An example of a modern method for synthesizing oligo arrays
- Based on using a micro mirror device to achieve
  - Small feature size
  - Flexibility
- User can define:
  - Sequence on the array
  - Size of features
  - Length of oligos

http://www.nimblegen.com

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

NIMBLEGEN MICROMIRROR TECHNOLOGY

Some specifications for the micromirror device:

•Mirrors spacing 17 um

•Mirror transit time <20 us

•Tilt angle ±10 degrees

•Five mirrors = diameter human hair

•Analog pictures from digital switches?

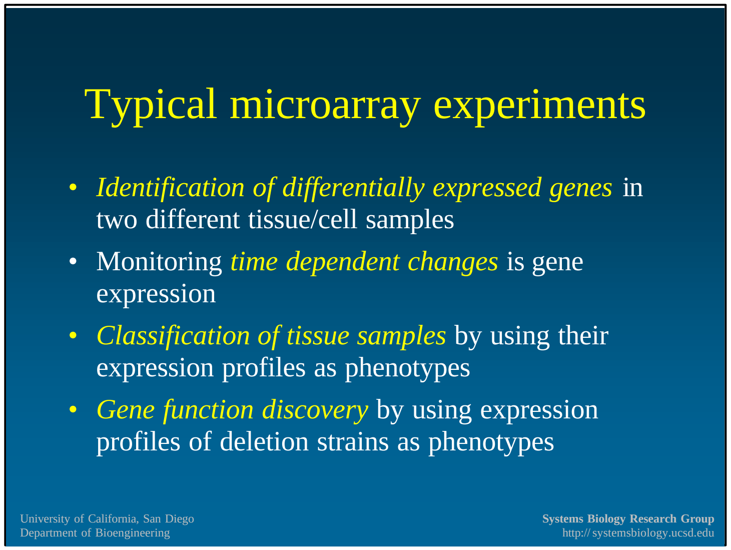
The user can define:

•Sequence on the arrays

•Controls on the array

•Number of replicates on the array

•Length of oligonucleotides on the array



Typical microarray experiments

- *Identification of differentially expressed genes* in two different tissue/cell samples
- Monitoring *time dependent changes* is gene expression
- *Classification of tissue samples* by using their expression profiles as phenotypes
- *Gene function discovery* by using expression profiles of deletion strains as phenotypes

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

TYPICAL MICROARRAY EXPERIMENTS

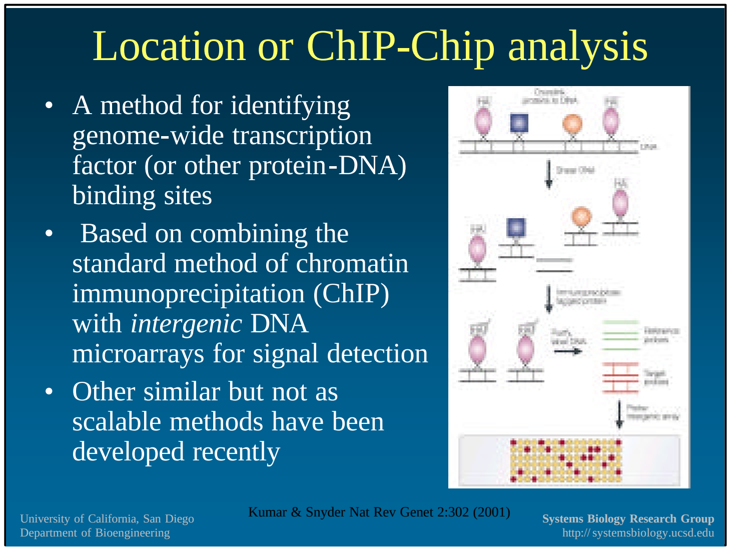
In order to be able to understand what types of computational methods are useful for analyzing microarray data, four typical microarray experiments will be described in the following slides:

•The simplest experiment is identification of differentially expressed genes in two different tissue/cell samples such as cells treated with a drug vs. untreated cells

•Many of the early experiments monitored time dependent changes in gene expression in a biological process such as the cell cycle or sporulation

•From the medical viewpoint an important application is classification of tissue samples such as different tumors using expression profiles

•Finally, for organisms where large-scale gene deletion studies are possible (e.g. yeast) expression profiles of deletion strains can be used as phenotypes to identify functions for unannotated genes



Location or ChIP-Chip analysis

- A method for identifying genome-wide transcription factor (or other protein-DNA) binding sites
- Based on combining the standard method of chromatin immunoprecipitation (ChIP) with *intergenic* DNA microarrays for signal detection
- Other similar but not as scalable methods have been developed recently

Kumar & Snyder Nat Rev Genet 2:302 (2001)

University of California, San Diego
Department of Bioengineering

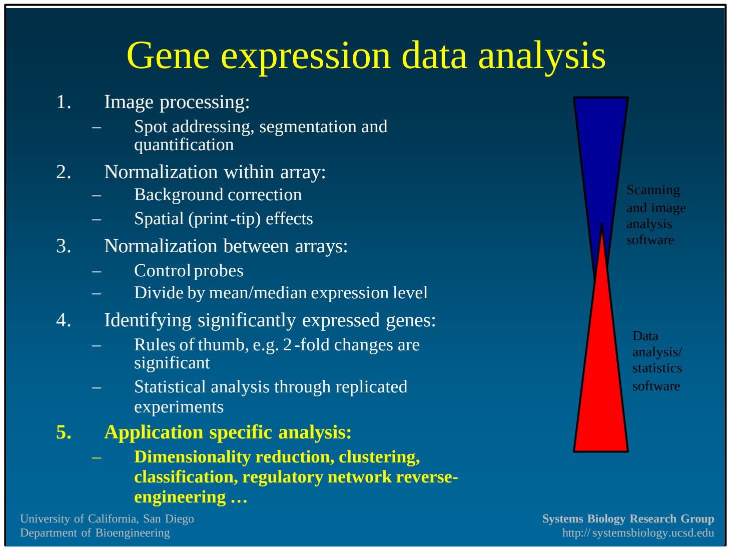
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

CHIP-CHIP METHOD

Figure caption: Genome-wide identification of protein–DNA interactions. Protein–DNA interactions are 'captured' *in vivo* by crosslinking proteins to their genomic binding sites. Crosslinked DNA is subsequently extracted, sheared and purified by immunoprecipitation with antibodies directed against an epitope-tagged protein of interest (such as the influenza virus haemagglutinin protein (HA epitope)). Purified DNA fragments are subsequently amplified and fluorescently labelled for use as target probes; labelled reference probes (shown in green) are often prepared from a strain deleted for the protein of interest. Probes are co-hybridized to an array of intergenic regions. The ratio of target probe to reference probe at each array 'spot' provides an indication of the frequency with which each corresponding genomic locus is bound by the tagged protein.

--

Nat Rev Genet 2001 Apr;2(4):302-12
Emerging technologies in yeast genomics.
Kumar A, Snyder M.



Gene expression data analysis

1. Image processing:
   - Spot addressing, segmentation and quantification
2. Normalization within array:
   - Background correction
   - Spatial (print-tip) effects
3. Normalization between arrays:
   - Control probes
   - Divide by mean/median expression level
4. Identifying significantly expressed genes:
   - Rules of thumb, e.g. 2-fold changes are significant
   - Statistical analysis through replicated experiments
5. **Application specific analysis:**
   - **Dimensionality reduction, clustering, classification, regulatory network reverse-engineering …**

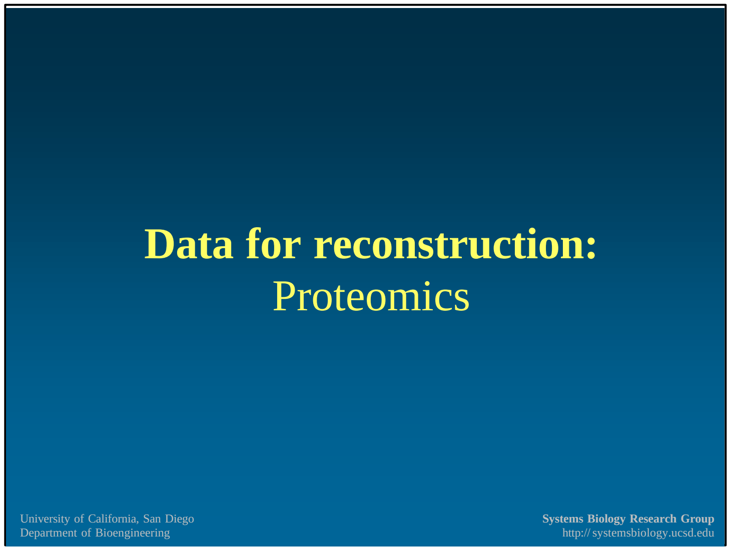
Scanning and image analysis software
Data analysis/ statistics software

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

OVERVIEW OF GENE EXPRESSION DATA ANALYSIS

The process of analyzing gene expression data starts with a scanned image of the array (or two images – red and green – for cDNA microarrays). The desired end result of the analysis is some kind of biologically meaningful conclusion such as which genes are upregulated or which genes are coexpressed. Typically many of the basic data analysis steps are done by the image analysis software that comes with the microarray scanner and require little user intervention unless a very careful analysis is required. While there are a lot of interesting image processing and statistics questions in the low level analysis, these questions are not addressed here due to lack of time. Where the user or experimentalist has to have more imput is in choosing the normalization method between arrays and in identifying what data points can actually be considered to be useful for further analysis (filtering out the background noise). For these tasks there are as many different approaches as there are people doing microarray experiments so that these methods are not reviewed in detail here. Instead, in this presentation we will focus in the last and probably most important and diverse step in microarray data analysis, which aims to convert a normalized and filtered set of numbers into useful biological knowledge or hypotheses. This is also the step where a large variety of different methods exists to answer different questions or to explore data in different ways.

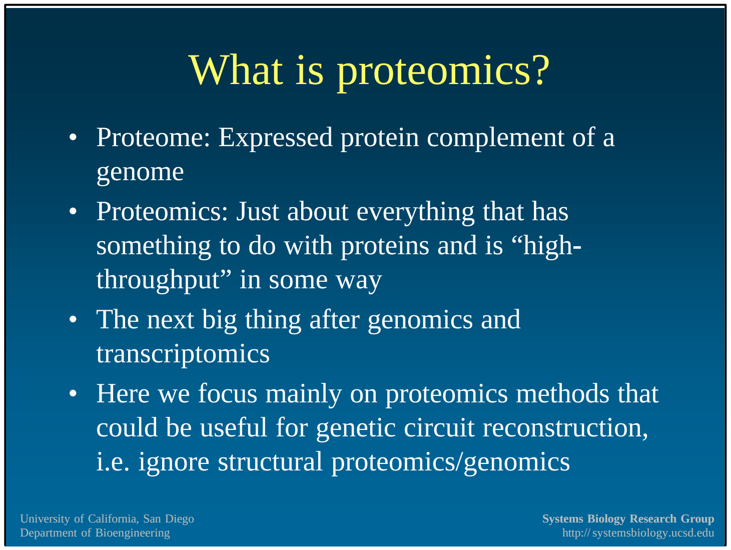


Data for reconstruction:
Proteomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

This topic will discuss mainly experimental methods for proteomics. Most of the methods have been developed in the last few years so that as a field proteomics is still in its infancy. However, there is great excitement about the prospects of proteomics research for understanding cellular function and reconstruction of genetic circuits. The hope is that just like genomics is today in a few years time we will be routinely be able to use data from proteomics technologies to reconstruct genetic circuits.



What is proteomics?

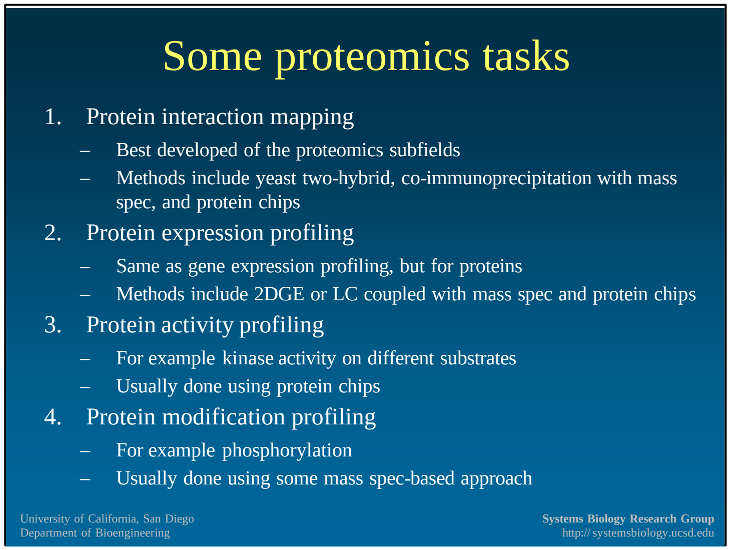
- Proteome: Expressed protein complement of a genome
- Proteomics: Just about everything that has something to do with proteins and is “high-throughput” in some way
- The next big thing after genomics and transcriptomics
- Here we focus mainly on proteomics methods that could be useful for genetic circuit reconstruction, i.e. ignore structural proteomics/genomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

WHAT IS PROTEOMICS?

Proteomics is not a very well defined topic – it could be described as a large-scale study of protein structure, expression, and function (including modifications and interactions). Proteomics is currently an extremely “hot” topic and people from many different fields – biology, bioengineering, bioinformatics, computer science, chemistry … - are actively developing ways to characterize the proteome. In this lecture we will ignore a major part of proteomics – large-scale structural characterization of proteins. Although knowing the structures of all proteins would certainly help in annotating their functions, this type of data is only indirectly connected to reconstruction of genetic circuits.



Some proteomics tasks

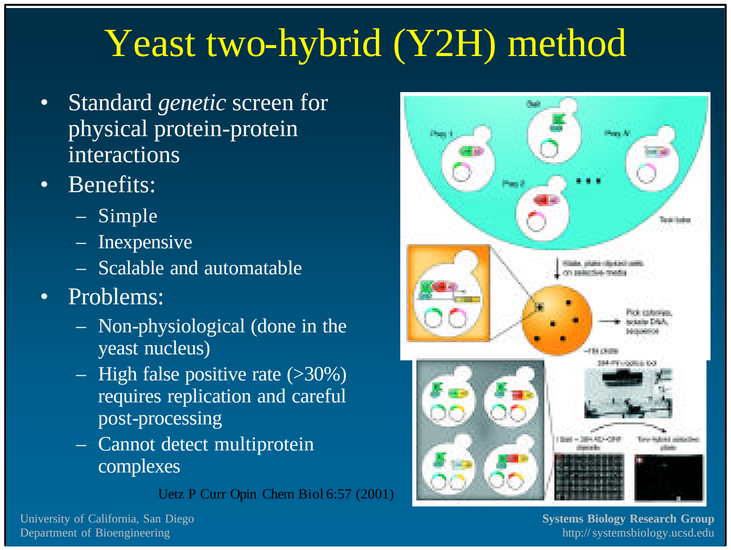
1. Protein interaction mapping
   - Best developed of the proteomics subfields
   - Methods include yeast two-hybrid, co-immunoprecipitation with mass spec, and protein chips
2. Protein expression profiling
   - Same as gene expression profiling, but for proteins
   - Methods include 2DGE or LC coupled with mass spec and protein chips
3. Protein activity profiling
   - For example kinase activity on different substrates
   - Usually done using protein chips
4. Protein modification profiling
   - For example phosphorylation
   - Usually done using some mass spec-based approach

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

TYPICAL PROTEOMICS TASKS

There are many possible tasks in proteomics out of which only four fairly well-developed ones are listed here. For most of these tasks there are more than one method and all the methods have their own advantages and disadvantages.



Yeast two-hybrid (Y2H) method

- Standard *genetic* screen for physical protein-protein interactions
- Benefits:
  - Simple
  - Inexpensive
  - Scalable and automatable
- Problems:
  - Non-physiological (done in the yeast nucleus)
  - High false positive rate (>30%) requires replication and careful post-processing
  - Cannot detect multiprotein complexes

Uetz P Curr Opin Chem Biol 6:57 (2001)

Bait
Prey 1
Prey N
Prey 2
Test tube
Mate, plate diploid cells on selective media
Pick colonies, isolate DNA, sequence
–His plate
Two-hybrid selection plate

University of California, San Diego
Department of Bioengineering

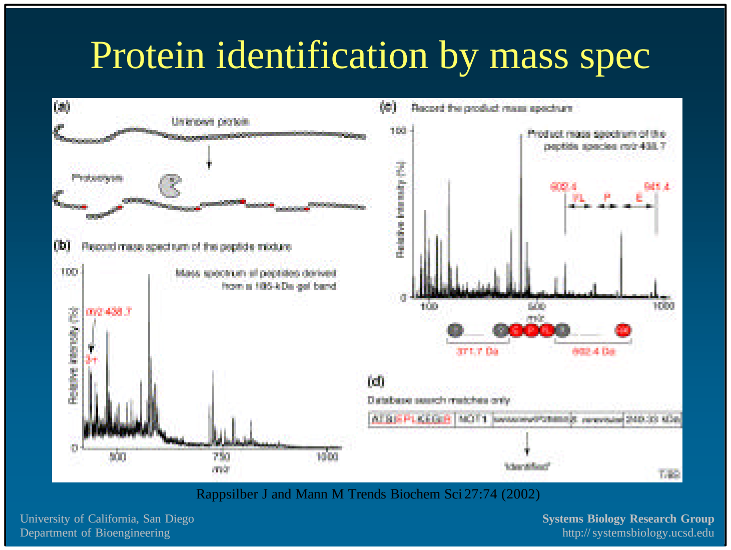
Systems Biology Research Group
http:// systemsbiology.ucsd.edu

YEAST TWO-HYBRID METHOD

Figure caption: Principle of two-hybrid library and array screens. **(a)** Typical two-hybrid screens use a library of random DNA or cDNA fused to a transcriptional activation domain (AD), expressed in yeast ('preys'; circles denote plasmids). The library clones are mated to a strain of opposite mating type that expresses a protein of interest ('bait', B) as a fusion to a DNA-binding domain (DBD). If bait and prey interact in the resulting diploid cells, they reconstitute a transcription factor, which activates a reporter gene whose expression allows the diploid cell to grow on selective media (here, without histidine). As an alternative to mating, prey libraries can also be transformed into the bait strain in order to express bait and prey in the same cell. In any case, positive clones have to be picked, their DNA isolated and the encoded plasmids sequenced in order to identify interacting proteins. **(b)** Array screens use defined sets of cloned prey ORFs or fragments thereof that are mated systematically to a certain bait strain. Matings and two-hybrid tests can be automated when large sets of preys have to be assayed, as in the case of whole genomes.

--

Curr Opin Chem Biol 2002 Feb;6(1):57-62
Two-hybrid arrays.



Protein identification by mass spec

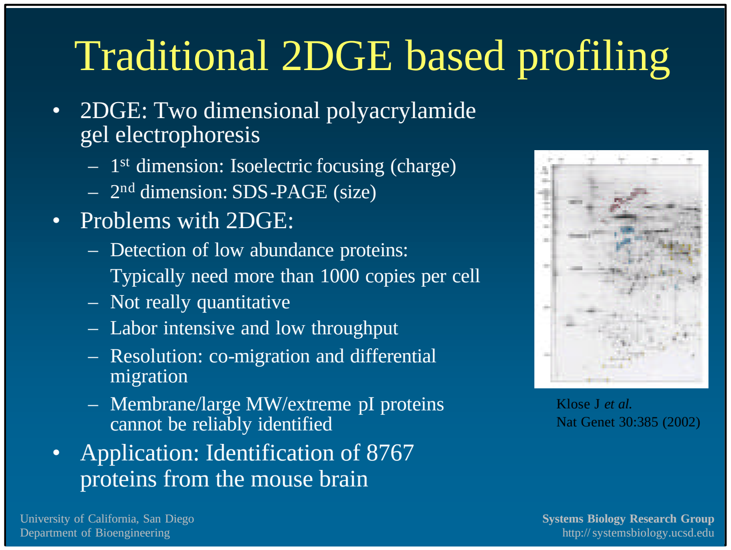
Rappsilber J and Mann M Trends Biochem Sci 27:74 (2002)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

PROTEIN IDENTIFICATION USING MASS SPEC

Figure caption: Mass spectrometric identification of a protein. A protein is digested using a highly specific protease (typically trypsin) (a) and the derived peptides are analysed in a mass spectrometer (b). One peptide species is selected for collision with an inert gas, such as argon, in the mass spectrometer. The derived fragments of this peptide are measured to give the product mass spectrum (c). Some of these fragments differ in their mass by individual amino acids – leucine and isoleucine having identical masses. Part of the sequence can therefore be read out from a series of peaks in the spectrum. This sequence information is placed in the peptide by the mass of the fragments, and is used in the `peptide sequence tag' in conjunction with the mass of the peptide and the specificity of the protease (tryptic digest results in K or R at the C terminus of the peptide) to search for a match in the database. A single peptide sequence tag is usually sufficient to unambiguously link a database entry with the investigated protein (d). Alternatively, fragment masses are measured and automatically compared with the predicted fragment masses of all peptides derived from a database to find the best match. In any case, the confidence increases with the number of fragmented peptides matching the same entry. Typically 2–70% of the sequence of the entry is covered by the experimental data, depending on the sample amount. Abbreviations: *m/z*, mass: charge ratio; *S. cerevisiae*, *Saccharomyces cerevisiae*.



Traditional 2DGE based profiling

- 2DGE: Two dimensional polyacrylamide gel electrophoresis
  - 1st dimension: Isoelectric focusing (charge)
  - 2nd dimension: SDS-PAGE (size)
- Problems with 2DGE:
  - Detection of low abundance proteins: Typically need more than 1000 copies per cell
  - Not really quantitative
  - Labor intensive and low throughput
  - Resolution: co-migration and differential migration
  - Membrane/large MW/extreme pI proteins cannot be reliably identified
- Application: Identification of 8767 proteins from the mouse brain

Klose J *et al.*
Nat Genet 30:385 (2002)

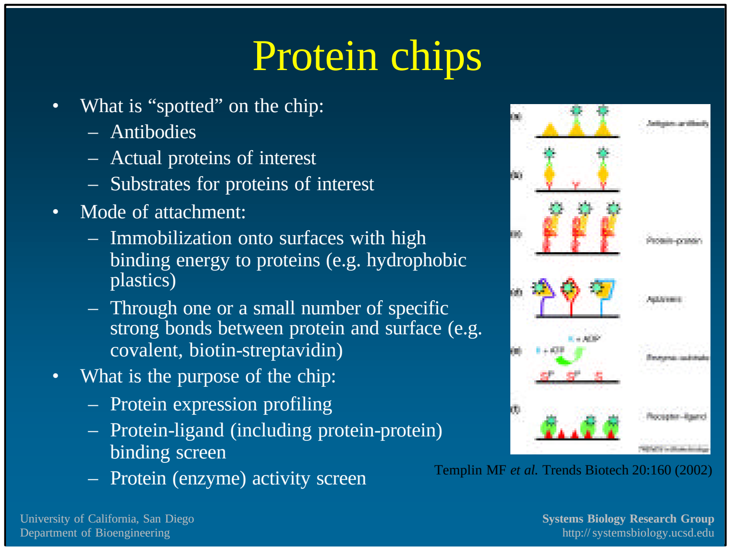
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

PROTEIN EXPRESSION PROFILING USING 2DGE

For a review on 2DGE in general see Lilley KS *et al*. Curr Opin Chem Biol 6:46 (2001)

Genetic analysis of the mouse brain proteome.
Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, Bluggel M, Krause E, Schalkwyk LC, Rastan S, Brown SD, Bussow K, Himmelbauer H, Lehrach H.
Abstract: Proteome analysis is a fundamental step in systematic functional genomics. Here we have resolved 8,767 proteins from the mouse brain proteome by large-gel two-dimensional electrophoresis. We detected 1,324 polymorphic proteins from the European collaborative interspecific backcross. Of these, we mapped 665 proteins genetically and identified 466 proteins by mass spectrometry. Qualitatively polymorphic proteins, to 96%, reflect changes in conformation and/or mass. Quantitatively polymorphic proteins show a high frequency (73%) of allele-specific transmission in codominant heterozygotes. Variations in protein isoforms and protein quantity often mapped to chromosomal positions different from that of the structural gene, indicating that single proteins may act as polygenic traits. Genetic analysis of proteomes may detect the types of polymorphism that are most relevant in disease-association studies.
Figure caption: 2-DE brain protein pattern from a B6-SPR hybrid.
From the whole 2-DE pattern, consisting of an acid half and a basic half (see Methods), only the acid half is shown. The three protein spot families show that spot families can be recognized in 2-DE patterns on the basis of genetic variation. The spot family of heat-shock 70-kD protein 4 (HSP70, red) consists of 52 'isospots', which occur in 52 double spots (hybrid spots). The two allelic forms vary in the vertical direction in all 52 double spots, with a spacing of 0.5 mm and the B6 spot on top. The -enolase 2 family (blue) shows 24 double spots, which vary again in the vertical direction with the B6 spot on top, but with a spacing of 1 mm. The lactate dehydrogenase 2 B chain spot family (LDH2, yellow) includes 17 double spots that form in each case horizontal spot pairs spaced at 25 mm, with the B6 spots always on the left side. This family also includes 10 paV spots, which are interpreted as a difference in degradation rate between the two allelic forms[11]. Four spots were present on the basic half of the 2-DE pattern (data not shown).
Nat Genet. 2002 Apr;30(4):385-93



(a)
Antigen–antibody
(b)
(c)
Protein–protein
(d)
Aptamers
(e)
K + ATP
K + ADP
SP
SP
S
Enzyme–substrate
(f)
Receptor–ligand

Protein chips

- What is "spotted" on the chip:
  - Antibodies
  - Actual proteins of interest
  - Substrates for proteins of interest
- Mode of attachment:
  - Immobilization onto surfaces with high binding energy to proteins (e.g. hydrophobic plastics)
  - Through one or a small number of specific strong bonds between protein and surface (e.g. covalent, biotin-streptavidin)
- What is the purpose of the chip:
  - Protein expression profiling
  - Protein-ligand (including protein-protein) binding screen
  - Protein (enzyme) activity screen

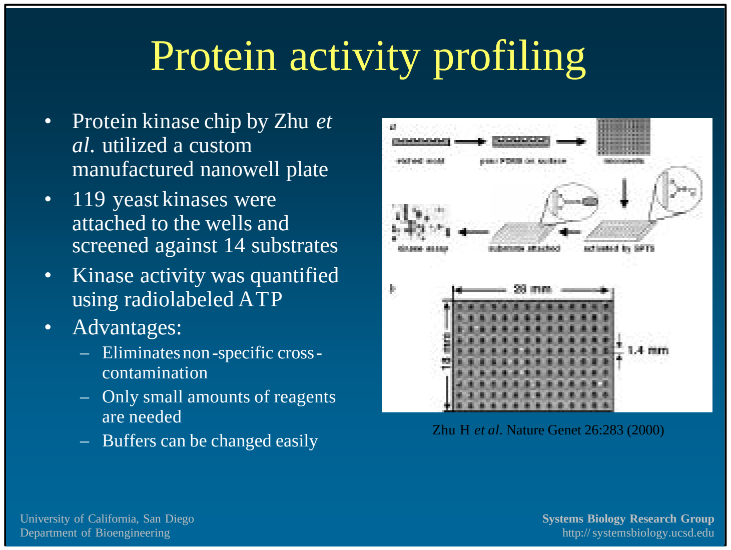
Templin MF *et al.* Trends Biotech 20:160 (2002)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

PROTEIN CHIPS

Figure caption: Classes of capture molecules for protein microarrays. For specific interaction analysis, different classes of molecules can be immobilized on a planar surface to act as capture molecules in a microarray assay. (a) Illustrates antigen–antibody interaction and (b) shows a scheme of a Sandwich immunoassay. In (c), a specific protein–protein interaction is shown. A different class of binders is shown in (d), where synthetic molecules referred to as aptamers act as capture molecules. They can be composed of nucleotides, ribonucleotides or peptides. Interactions of enzymes with their specific substrates are shown in (e), where a substrate (s) for kinases is immobilized and phosphorylated (P) by the respective kinase. A typical example for a receptor–ligand interaction is given in (f), where synthetic low molecular mass compounds are immobilised as capture molecules.


a
etched mold
pour PDMS on surface
microwells
kinase assay
substrate attached
activated by GPTS
b
28 mm
18 mm
1.4 mm

Protein activity profiling

- Protein kinase chip by Zhu *et al*. utilized a custom manufactured nanowell plate
- 119 yeast kinases were attached to the wells and screened against 14 substrates
- Kinase activity was quantified using radiolabeled ATP
- Advantages:
  - Eliminates non-specific cross-contamination
  - Only small amounts of reagents are needed
  - Buffers can be changed easily

Zhu H *et al*. Nature Genet 26:283 (2000)

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

APPLICATIONS OF PROTEIN CHIPS: PROTEIN ACTIVITY PROFILING

Nat Genet 2000 Nov;26(3):283-9

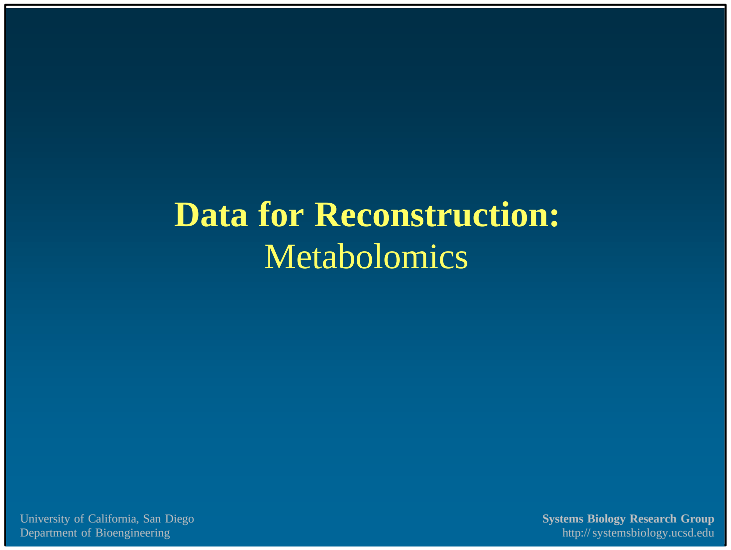
Analysis of yeast protein kinases using protein chips.

Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M

Abstract: We have developed a novel protein chip technology that allows the high-throughput analysis of biochemical activities, and used this approach to analyse nearly all of the protein kinases from Saccharomyces cerevisiae. Protein chips are disposable arrays of microwells in silicone elastomer sheets placed on top of microscope slides. The high density and small size of the wells allows for high-throughput batch processing and simultaneous analysis of many individual samples. Only small amounts of protein are required. Of 122 known and predicted yeast protein kinases, 119 were overexpressed and analysed using 17 different substrates and protein chips. We found many novel activities and that a large number of protein kinases are capable of phosphorylating tyrosine. The tyrosine phosphorylating enzymes often share common amino acid residues that lie near the catalytic region. Thus, our study identified a number of novel features of protein kinases and demonstrates that protein chip technology is useful for high-throughput screening of protein biochemical activity.

Figure caption: Protein chip fabrication and kinase assays.

a, Kinase activities were detected using protein chips. PDMS was poured over the acrylic mold. After curing, the chip containing the wells was peeled away and mounted on a glass slide. The next step included modification of the surface and then attachment of proteins to the wells. Wells were blocked with 1% BSA before kinase, ^{33}P -ATP and buffer were added. After incubation for 30 min at 30 °C, the chips were washed extensively and exposed to both X-ray film and a phosphoimager, which has a resolution of 50 m and is quantitative. For 12 substrates each kinase assay was repeated at least twice; for the remaining 5 the assays were performed once. ***b*** , An enlarged picture of the protein chip.



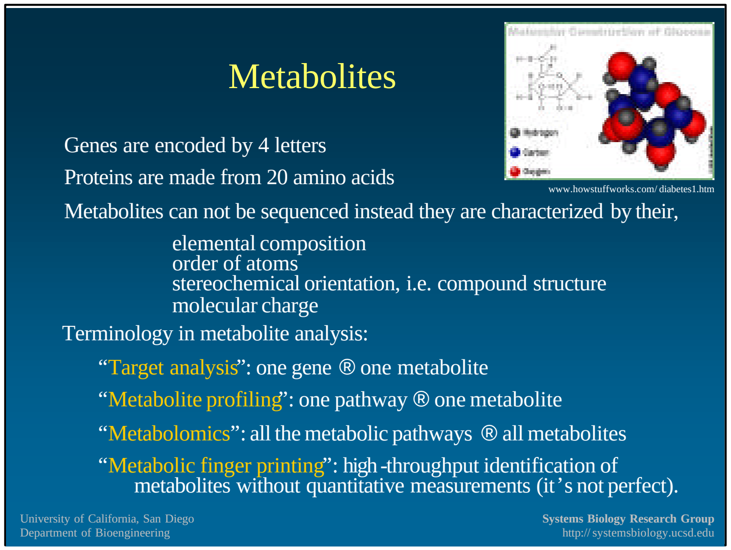
Data for Reconstruction:
Metabolomics

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

Data for Reconstruction: Metabolomics

In addition to genomics, trancriptomics, and proteomics data, the changes in metabolite concentration levels in the cell can be used for network reconstruction of biological systems and analysis of phenotypic behavior in the cell. “Metabolomics”, or the whole-cell analysis of metabolite concentrations, is thus the topic of the following lecture.



Molecular Construction of Glucose
Hydrogen
Carbon
Oxygen

Metabolites

Genes are encoded by 4 letters

Proteins are made from 20 amino acids

www.howstuffworks.com/ diabetes1.htm

Metabolites can not be sequenced instead they are characterized by their,

elemental composition
order of atoms
stereochemical orientation, i.e. compound structure
molecular charge

Terminology in metabolite analysis:

"Target analysis": one gene ® one metabolite

"Metabolite profiling": one pathway ® one metabolite

"Metabolomics": all the metabolic pathways ® all metabolites

"Metabolic finger printing": high-throughput identification of metabolites without quantitative measurements (it's not perfect).

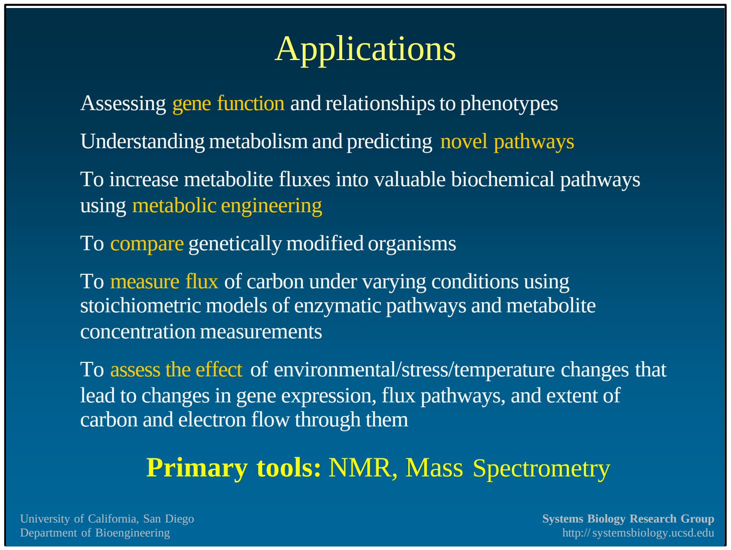
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

Metabolites

Unlike genes that are encoded by 4 letters, or proteins that are made from 20 amino acids, metabolites don't have a set of codons and thus cannot be sequenced. Instead, they are characterized by their elemental composition, order of atoms, stereochemical orientation, and molecular charge.

There are common terminologies used in analysis of metabolite concentrations. "Target analysis" is referred to the process of perturbing one gene and measuring the effect of this perturbation on the concentration of a target metabolite (i.e. the metabolite of interest). If more than one gene is perturbed, for instance, in a given pathway and the changes of a target metabolite is measured following such perturbations, the analysis is referred to as "metabolite profiling". "Metabolomics" is a whole-cell measurement of all the metabolites and it is considered to be equivalent to transcriptomics in mRNA expression analysis. Metabolite concentration levels can also be measure in a high-throughput and qualitative fashion. This is referred to as "metabolic finger printing" and it is used routinely in disease detection and in comparing transgenic organisms to wild-type. For example, 50 different human diseases can be easily detected using "metabolic finger printing" and cancer cells are compared to the normal cells for metabolite comparison.



Applications

Assessing gene function and relationships to phenotypes

Understanding metabolism and predicting novel pathways

To increase metabolite fluxes into valuable biochemical pathways using metabolic engineering

To compare genetically modified organisms

To measure flux of carbon under varying conditions using stoichiometric models of enzymatic pathways and metabolite concentration measurements

To assess the effect of environmental/stress/temperature changes that lead to changes in gene expression, flux pathways, and extent of carbon and electron flow through them

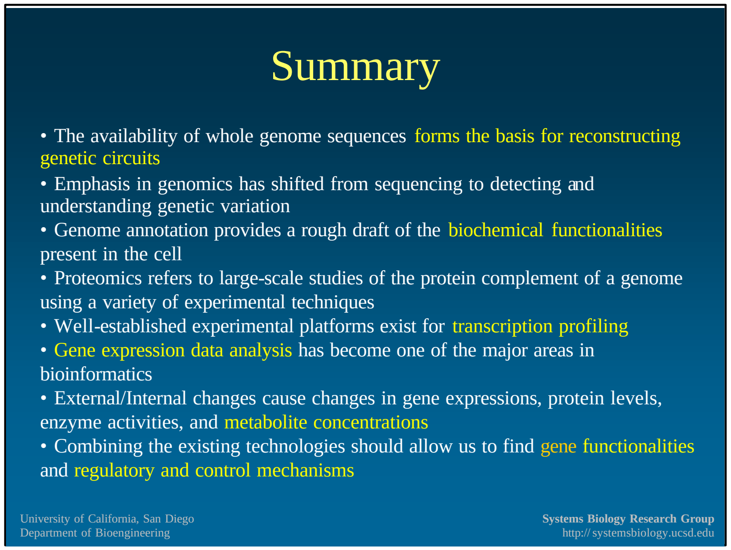
Primary tools: NMR, Mass Spectrometry

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http:// systemsbiology.ucsd.edu

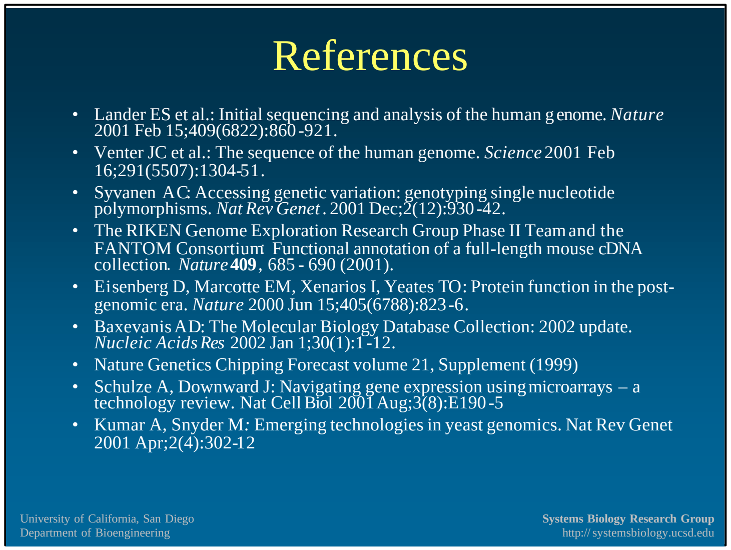
Applications

Applications of metabolite analysis in biology and metabolic engineering are numerous. For example, one can assess the function of a gene by genetically perturbing it and examining the effect of this perturbation on metabolite concentrations. This has been used in annotating “silent mutations” in *Saccharomyces cerevisiae* (Raamsdonk 2001). Metabolite measurements can also be use to better understand metabolism and to discover novel pathways in organisms such as *B. subtilis* (Szyperski, 1996). Since it is believed that the genetic changes affect the level and activity of proteins, which in turn affect the level of metabolites in the cell, metabolomics can be used in metabolic engineering of valuable and desirable byproducts. Genetically modified organisms can be compared to their wild-type strains using metabolite screening for toxic intermediates. In modeling metabolism, metabolite analysis and stoichiometric models can be combined to measure fluxes of carbon in the cell and the effect of internal and/or external inputs for a given cell can also be assessed using metabolite concentration measurements.



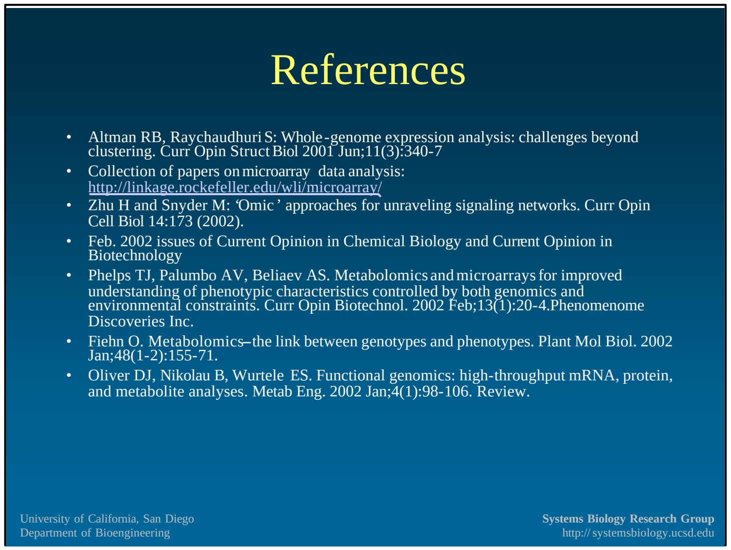
Summary

- The availability of whole genome sequences forms the basis for reconstructing genetic circuits
- Emphasis in genomics has shifted from sequencing to detecting and understanding genetic variation
- Genome annotation provides a rough draft of the biochemical functionalities present in the cell
- Proteomics refers to large-scale studies of the protein complement of a genome using a variety of experimental techniques
- Well-established experimental platforms exist for transcription profiling
- Gene expression data analysis has become one of the major areas in bioinformatics
- External/Internal changes cause changes in gene expressions, protein levels, enzyme activities, and metabolite concentrations
- Combining the existing technologies should allow us to find gene functionalities and regulatory and control mechanisms



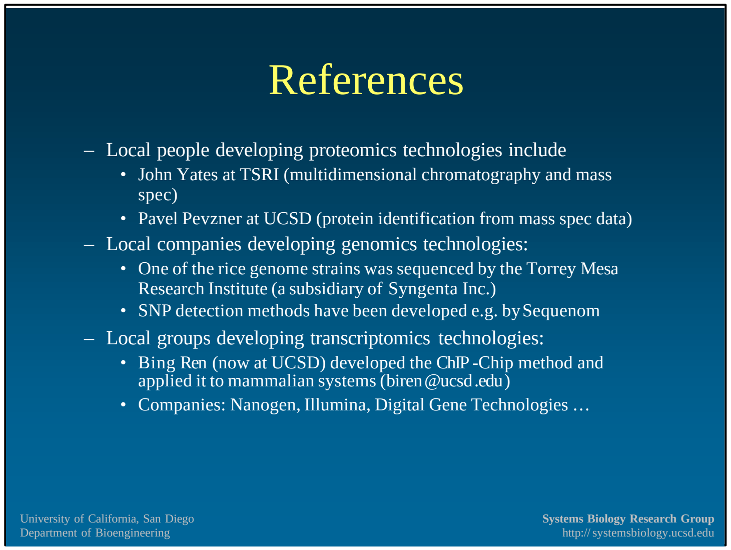
References

- Lander ES et al.: Initial sequencing and analysis of the human genome. *Nature* 2001 Feb 15;409(6822):860-921.
- Venter JC et al.: The sequence of the human genome. *Science* 2001 Feb 16;291(5507):1304-51.
- Syvanen AC: Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*. 2001 Dec;2(12):930-42.
- The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium: Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685 - 690 (2001).
- Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: Protein function in the post-genomic era. *Nature* 2000 Jun 15;405(6788):823-6.
- Baxevanis AD: The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res* 2002 Jan 1;30(1):1-12.
- Nature Genetics Chipping Forecast volume 21, Supplement (1999)
- Schulze A, Downward J: Navigating gene expression using microarrays – a technology review. Nat Cell Biol 2001 Aug;3(8):E190-5
- Kumar A, Snyder M: Emerging technologies in yeast genomics. Nat Rev Genet 2001 Apr;2(4):302-12



References

- Altman RB, Raychaudhuri S: Whole-genome expression analysis: challenges beyond clustering. Curr Opin Struct Biol 2001 Jun;11(3):340-7
- Collection of papers on microarray data analysis: http://linkage.rockefeller.edu/wli/microarray/
- Zhu H and Snyder M: 'Omic' approaches for unraveling signaling networks. Curr Opin Cell Biol 14:173 (2002).
- Feb. 2002 issues of Current Opinion in Chemical Biology and Current Opinion in Biotechnology
- Phelps TJ, Palumbo AV, Beliaev AS. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. Curr Opin Biotechnol. 2002 Feb;13(1):20-4.Phenomenome Discoveries Inc.
- Fiehn O. Metabolomics–the link between genotypes and phenotypes. Plant Mol Biol. 2002 Jan;48(1-2):155-71.
- Oliver DJ, Nikolau B, Wurtele ES. Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metab Eng. 2002 Jan;4(1):98-106. Review.



References

- Local people developing proteomics technologies include
  - John Yates at TSRI (multidimensional chromatography and mass spec)
  - Pavel Pevzner at UCSD (protein identification from mass spec data)
- Local companies developing genomics technologies:
  - One of the rice genome strains was sequenced by the Torrey Mesa Research Institute (a subsidiary of Syngenta Inc.)
  - SNP detection methods have been developed e.g. by Sequenom
- Local groups developing transcriptomics technologies:
  - Bing Ren (now at UCSD) developed the ChIP -Chip method and applied it to mammalian systems (biren@ucsd.edu)
  - Companies: Nanogen, Illumina, Digital Gene Technologies …