SYLICA 2013
Bowater lectures

Using ‘Omic Technologies to Investigate Gene Function

Bowater Lectures in Brno, Feb. 2013

4 lectures on linked topics will be delivered during the coming week:

- *Contemporary DNA Sequencing Technologies* – 26/2/2013 @ 10:00
- *Using 'Omic Technologies to Investigate Gene Function* – 26/2/2013 @ 14:00
- *Biophysical Methods to Study Molecular Interactions* – 27/2/2013 @ 10:00
- *Synthetic Biology & Nanotechnology: Tomorrow's Molecular Biology?* – 28/2/2013 @ 10:00

Genomics, 'Omics & Technology

- Molecular biology: major scientific discipline for past ~50 years
- *Genomics* = "analysis of genomes": became important science during 1990's
- Analyses of various other biological molecules have developed into their own scientific disciplines; e.g. *Metabolomics* = "analysis of metabolites", etc.
- *Transcriptomics/Proteomics*: developed during past 10-15 years
- Bioinformatics: has developed as major branch of science - enables efficient analysis of data from "omics" experiments

Genomics & Technology

- Significance of “omics” coincides with dramatic improvements in different technologies:
  - molecular biology: increased range of approaches for purification and manipulation of proteins and nucleic acids
  - computers: required for gathering and analysis of data
  - internet: allows data to be shared, quickly and easily
- All developments have increased speed and cost-effectiveness - available to much wider audience

Transcriptomes

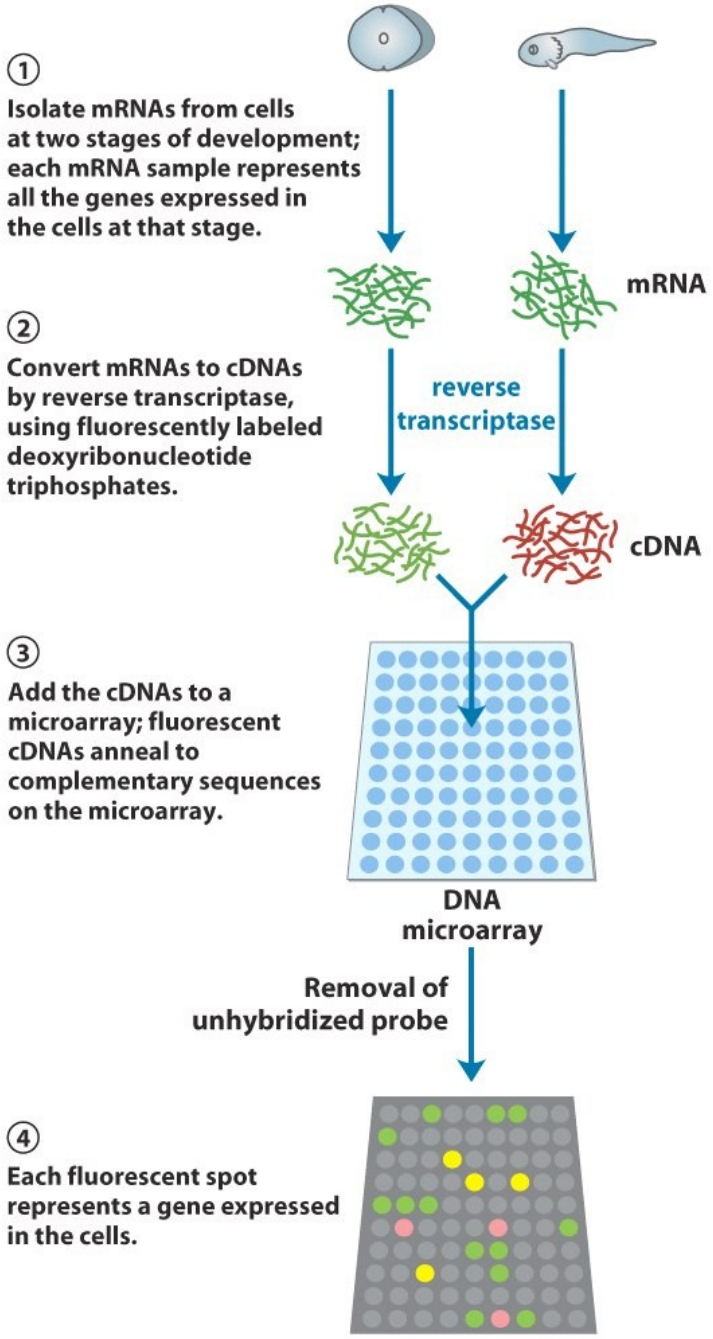
- *Genome*: all of hereditary information encoded in the DNA (or RNA)
- *Transcriptome*: set of all mRNAs ("transcripts") produced from a genome
- Term can be applied to:
  - complete set of transcripts for a given organism
  - specific subset of transcripts present in a particular cell type or under specific growth conditions
- Transcriptome varies because it reflects genes that are actively expressed at any given time

DNA Microarrays Show Differences in Gene Expression

- Microarray chips contain fragments from genes in the group to be analyzed
  - Full genome of bacteria or yeast, or protein families from larger genomes
- mRNA or cDNA from different samples are differentially tagged
- Analysis on the same chip shows differences

Transcriptomics

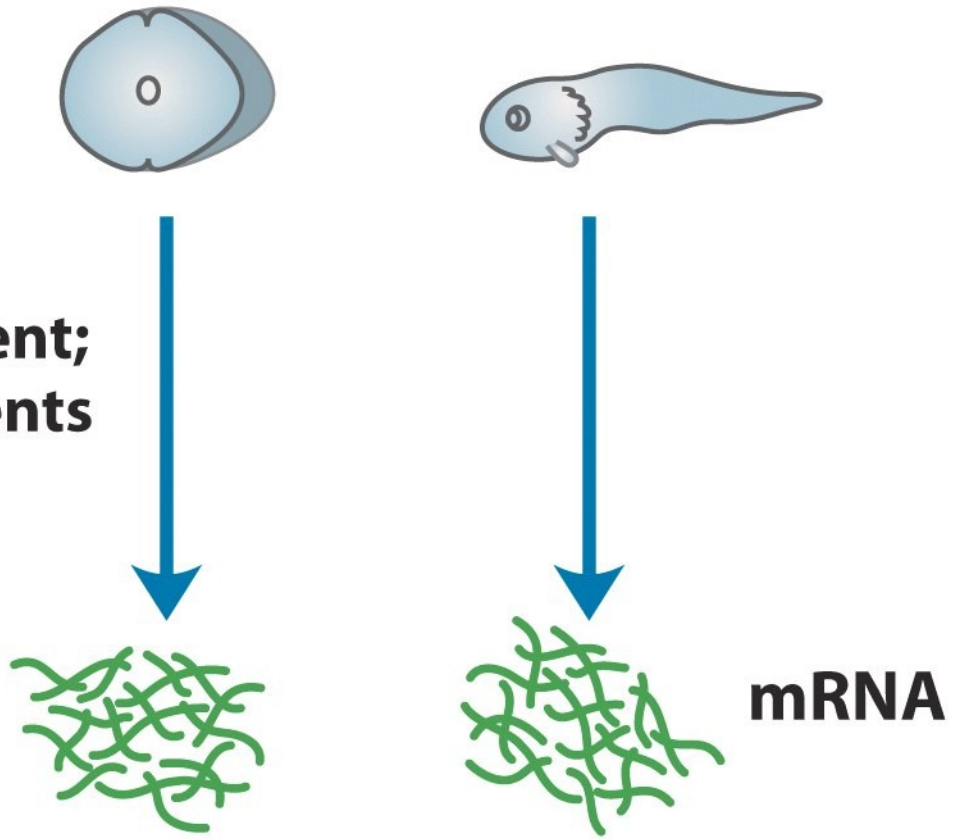
- *Transcriptomics* uses high-throughput techniques based on DNA microarrays
- For further details about microarrays see Lucchini *et al.*, *Microbiology*, **147**, 1403-1414 (2001)



Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

Transcriptomics

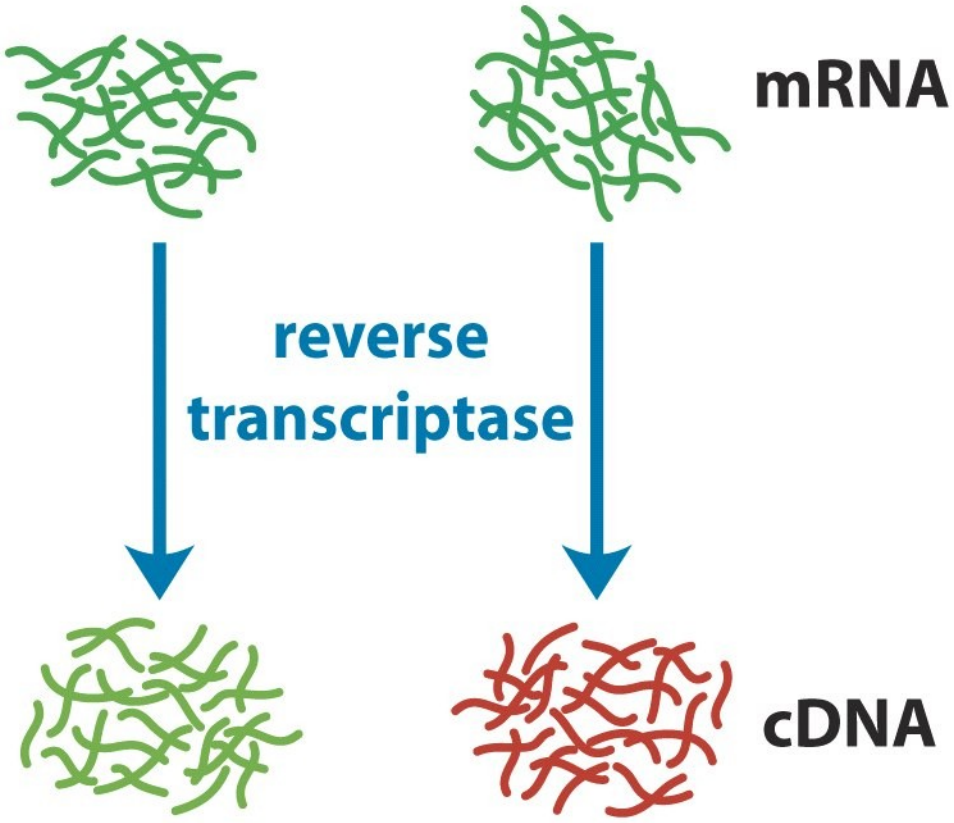
① **Isolate mRNAs from cells at two stages of development; each mRNA sample represents all the genes expressed in the cells at that stage.**



Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

Transcriptomics

② **Convert mRNAs to cDNAs by reverse transcriptase, using fluorescently labeled deoxyribonucleotide triphosphates.**

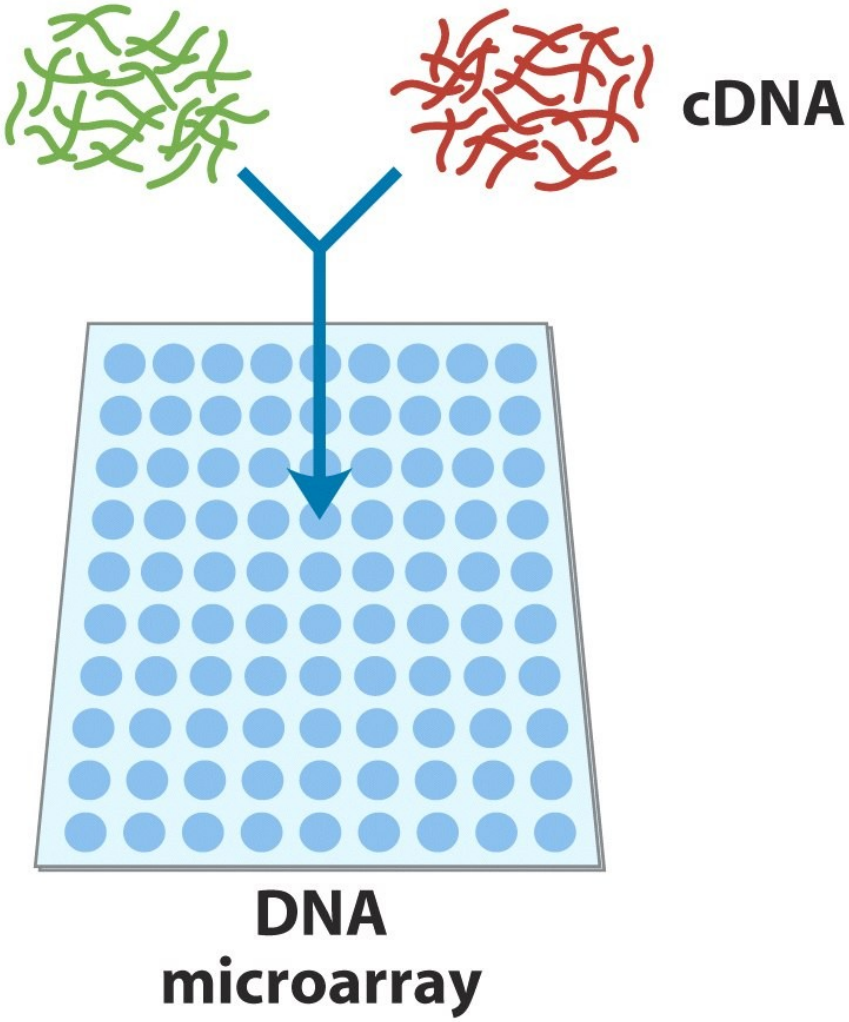


Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

Transcriptomics

③

Add the cDNAs to a microarray; fluorescent cDNAs anneal to complementary sequences on the microarray.

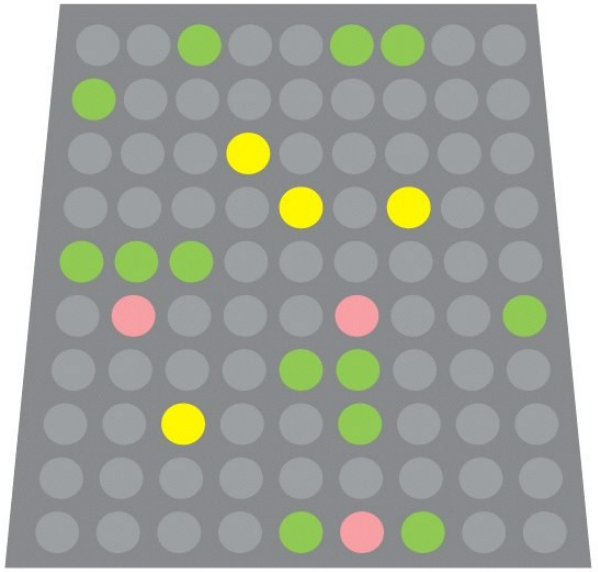


Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

Transcriptomics

DNA microarray

Removal of unhybridized probe

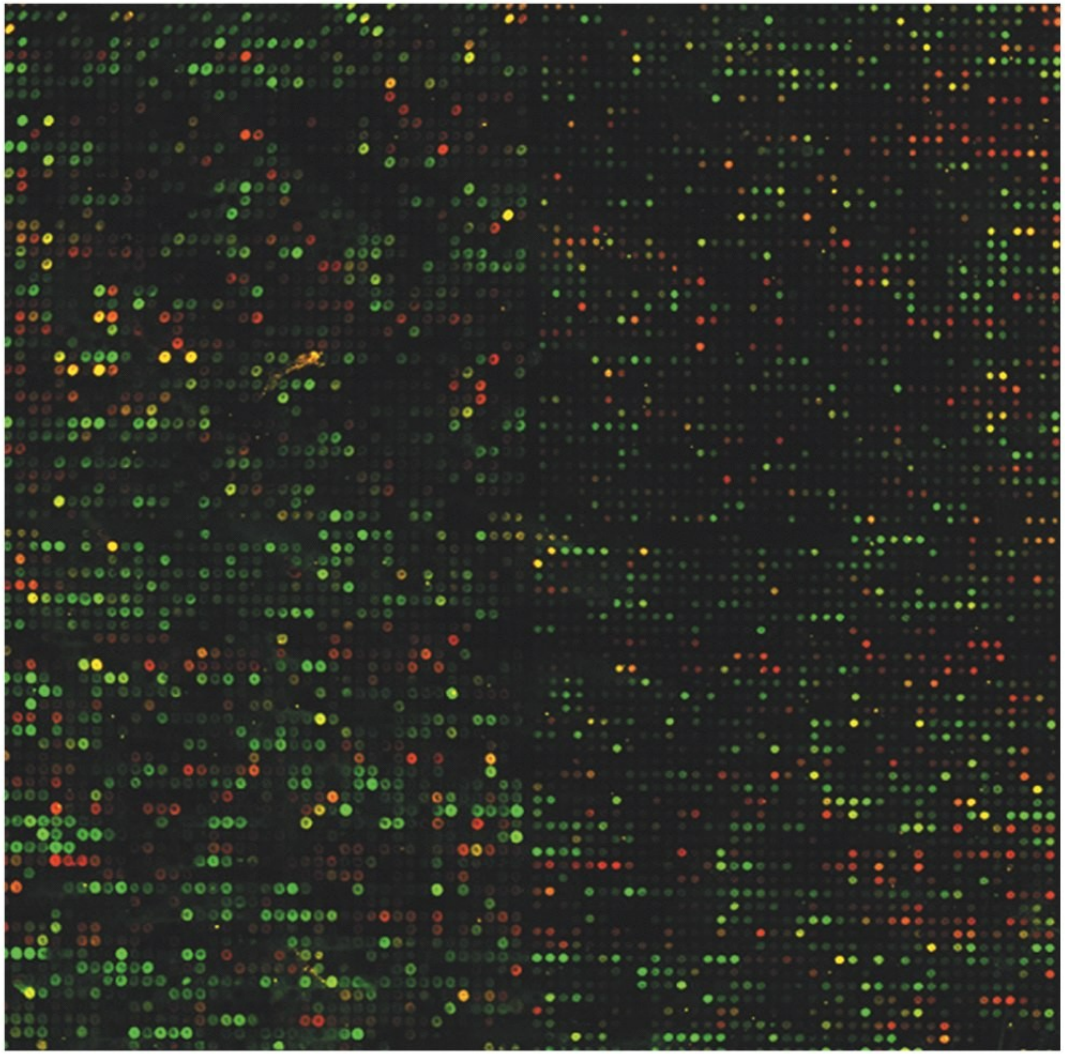


④

Each fluorescent spot represents a gene expressed in the cells.

Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

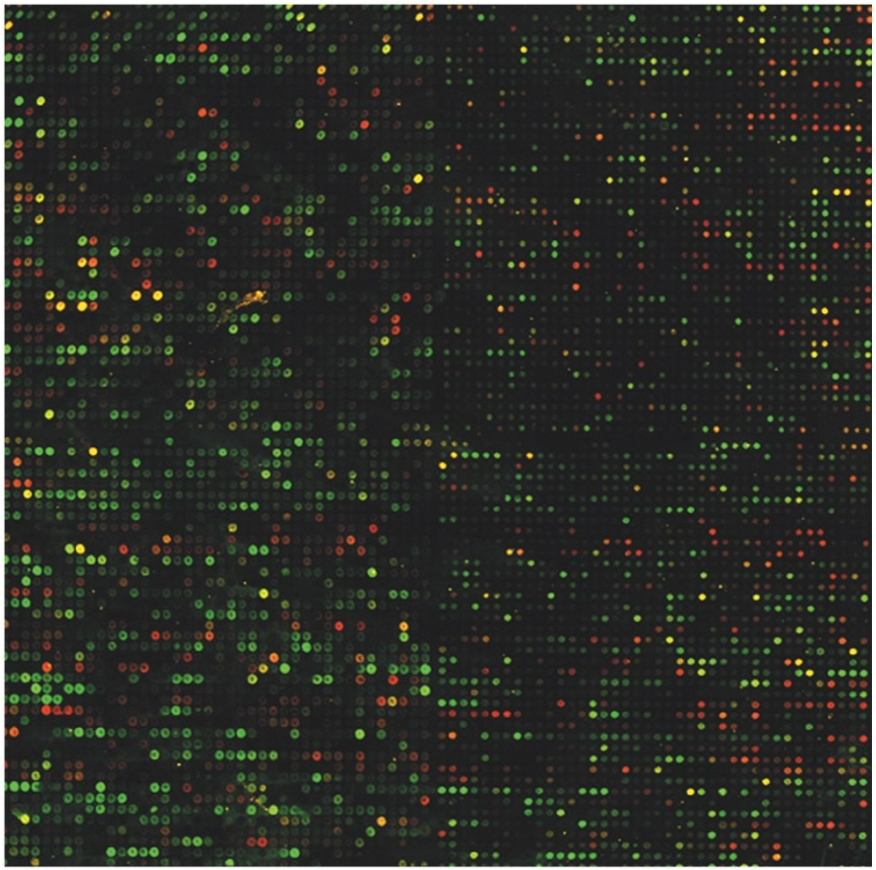
Transcriptomics



Nelson & Cox, “Lehninger, Principles of Biochemistry”, 4th edn, 2004, p. 328

Transcriptomics

- Experiments performed under different conditions
- Determines effect of conditions on expression
- Produces huge amount of data
- Lots of repeats required - expensive



Nelson & Cox, "Lehninger, Principles of Biochemistry", 4th edn, 2004, p. 328

Polymerase Chain Reaction (PCR)

- Used to amplify DNA in the test tube
  - Can amplify regions of interest (genes) within DNA
  - Can amplify complete circular plasmids
- Mix together
  - Target DNA
  - Primers (oligonucleotides complementary to target)
  - Nucleotides: dATP, dCTP, dGTP, dTTP
  - Thermostable DNA polymerase
- Place the mixture into thermocycler
  - Melt DNA at ~95°C
  - Cool to ~ 50–60°C, primers anneal to target
  - Polymerase extends primers in 5'→3' direction
  - After a round of elongation is done, repeat steps

General Steps of PCR

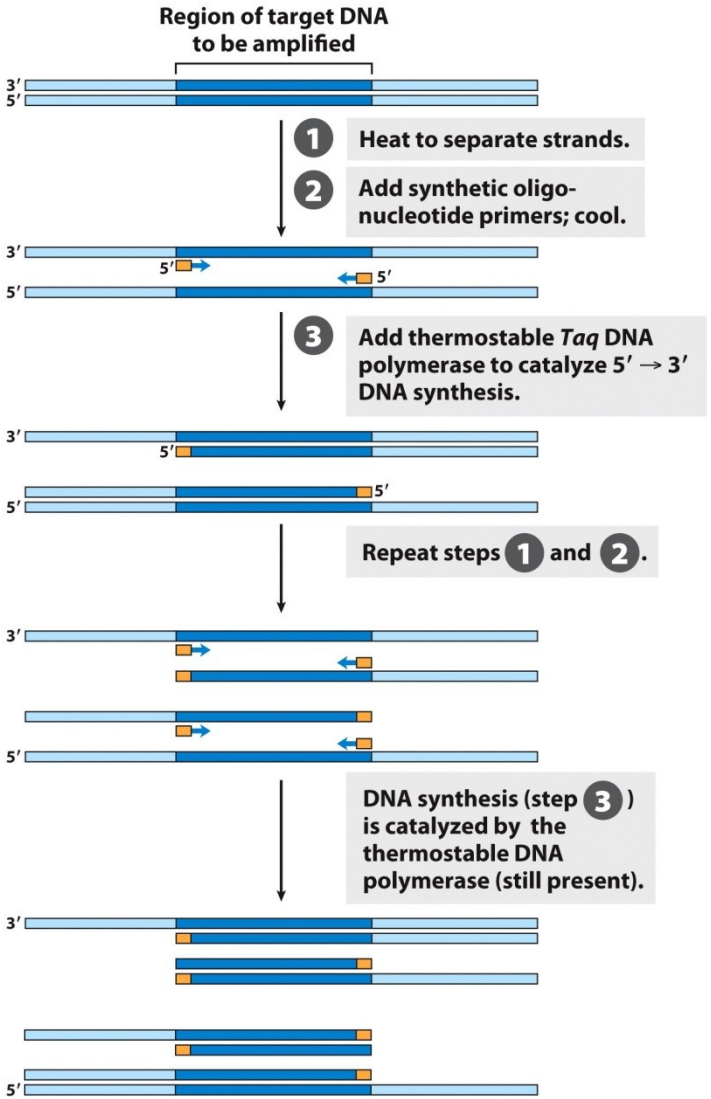


Figure 9-12a part 1
Lehninger Principles of Biochemistry, Sixth Edition

General Steps of PCR

- Repeat steps 1–3 many times:

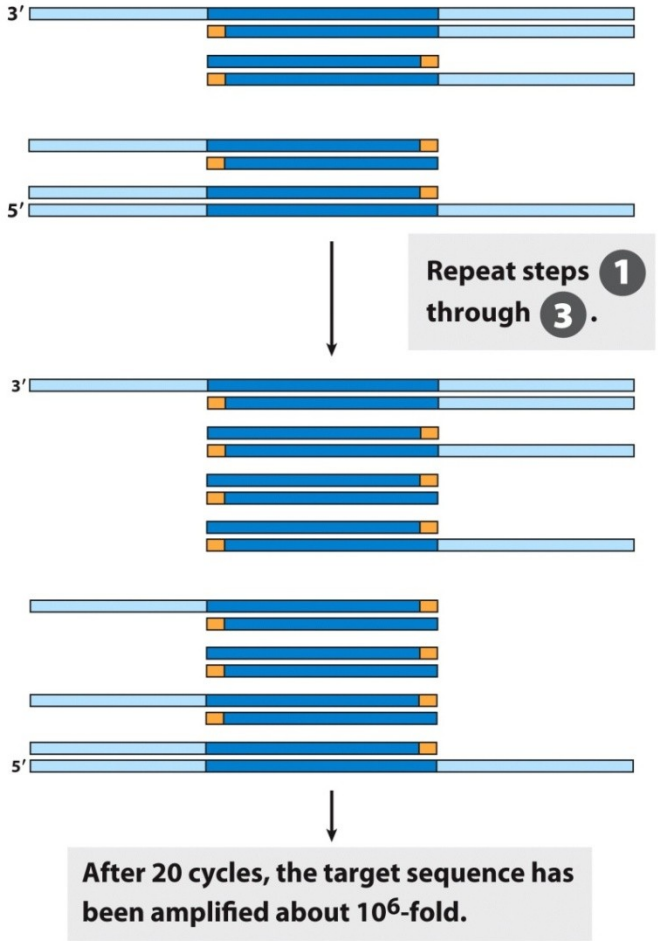


Figure 9-12a part 2
Lehninger Principles of Biochemistry, Sixth Edition

Photolitographic Synthesis of DNA

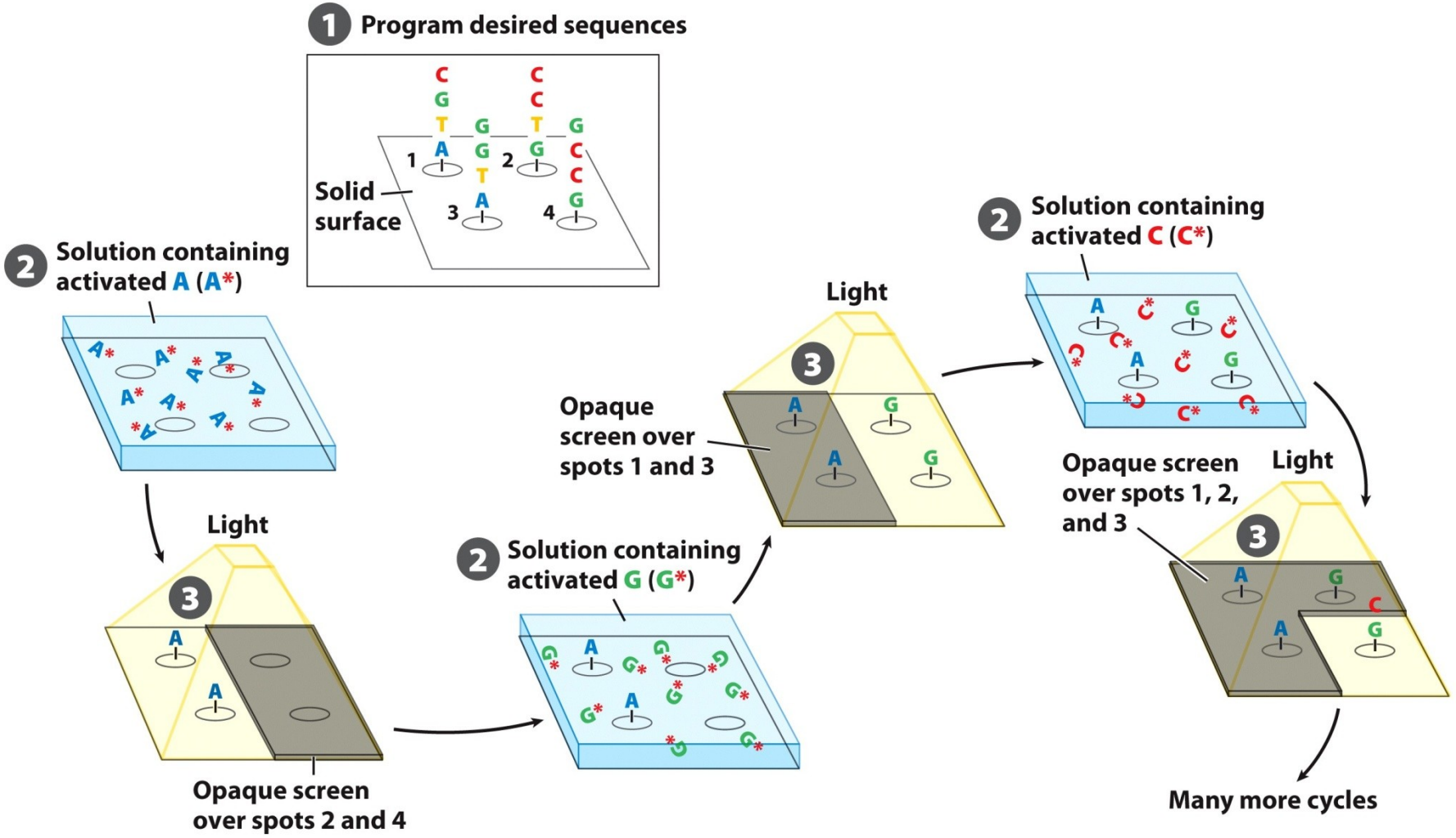


Figure 9-22
Lehninger Principles of Biochemistry, Sixth Edition

DNA Microarrays: Applications

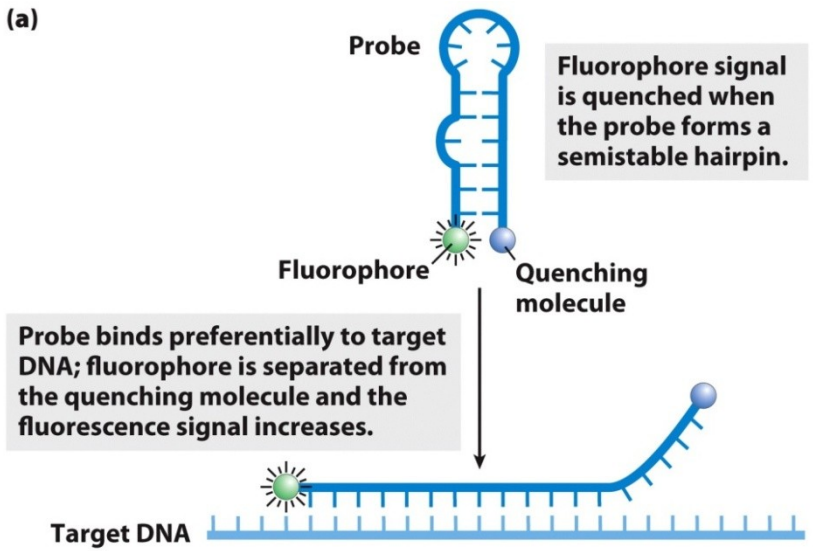
DNA microarrays allow simultaneous screening of many thousands of genes: high-throughput screening

- Genome-wide genotyping
  - Which genes are present in this individual?
- Tissue-specific gene expression
  - Which genes are used to make proteins?
- Mutational analysis
  - Which genes have been mutated?

Adaptations to PCR

- Reverse Transcriptase PCR (RT-PCR)
  - Used to amplify RNA sequences
  - First step uses reverse transcriptase to convert RNA to DNA
- Quantitative PCR (Q-PCR)
  - Used to show quantitative differences in gene levels

qPCR


(a)
Probe
Fluorophore signal is quenched when the probe forms a semistable hairpin.
Fluorophore
Quenching molecule
Probe binds preferentially to target DNA; fluorophore is separated from the quenching molecule and the fluorescence signal increases.
Target DNA

Proteomes

- *Proteome*: set of all proteins produced under a given set of conditions
- Term can be applied to:
  - complete set of proteins for a given organism
  - specific subset of proteins present in a particular cell type or under specific growth conditions
- Proteome varies because it reflects genes that are actively expressed at any given time
- *Proteomics* analyses many samples using 2D-electrophoresis and mass spectrometry
- High-throughput, but less than transcriptomics

Gel Electrophoresis

- Electrophoresis separates molecules by size
- Resolution is limited

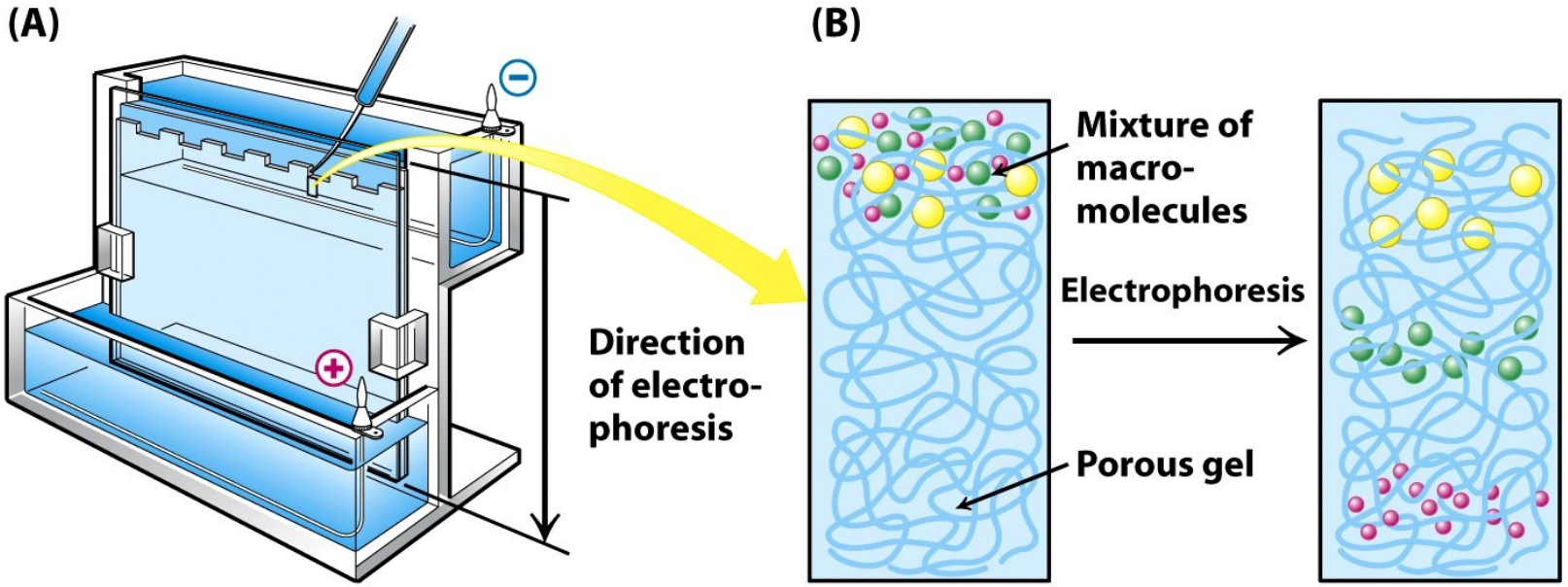


Figure 3-7
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 71

Isoelectric Focusing

- Electrophoresis across a pH gradient
- Proteins migrate to their isoelectric pH

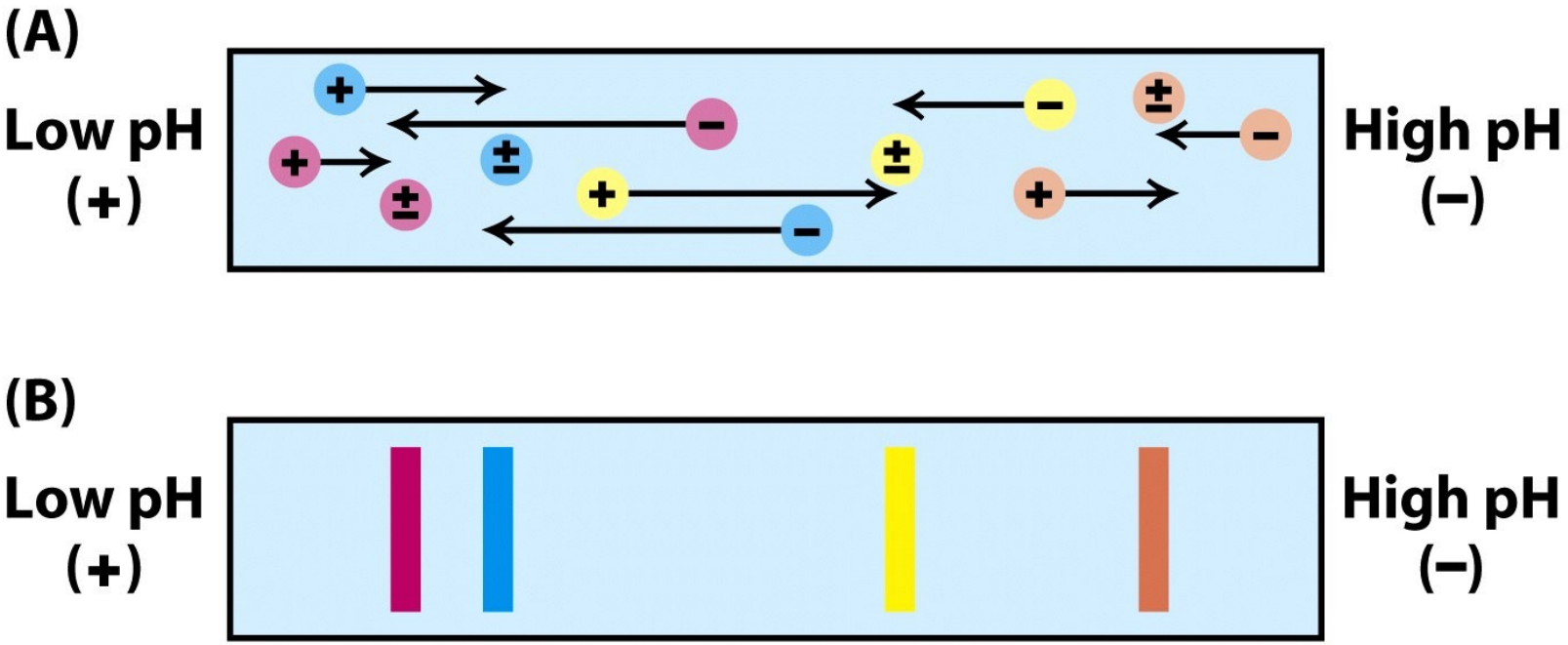


Figure 3-11
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 73

Two-dimensional Gel Electrophoresis

- Protein sample initially fractionated in one dimension by isoelectric focusing
- SDS-PAGE performed perpendicular to original direction
- Separates proteins according to pI and mass

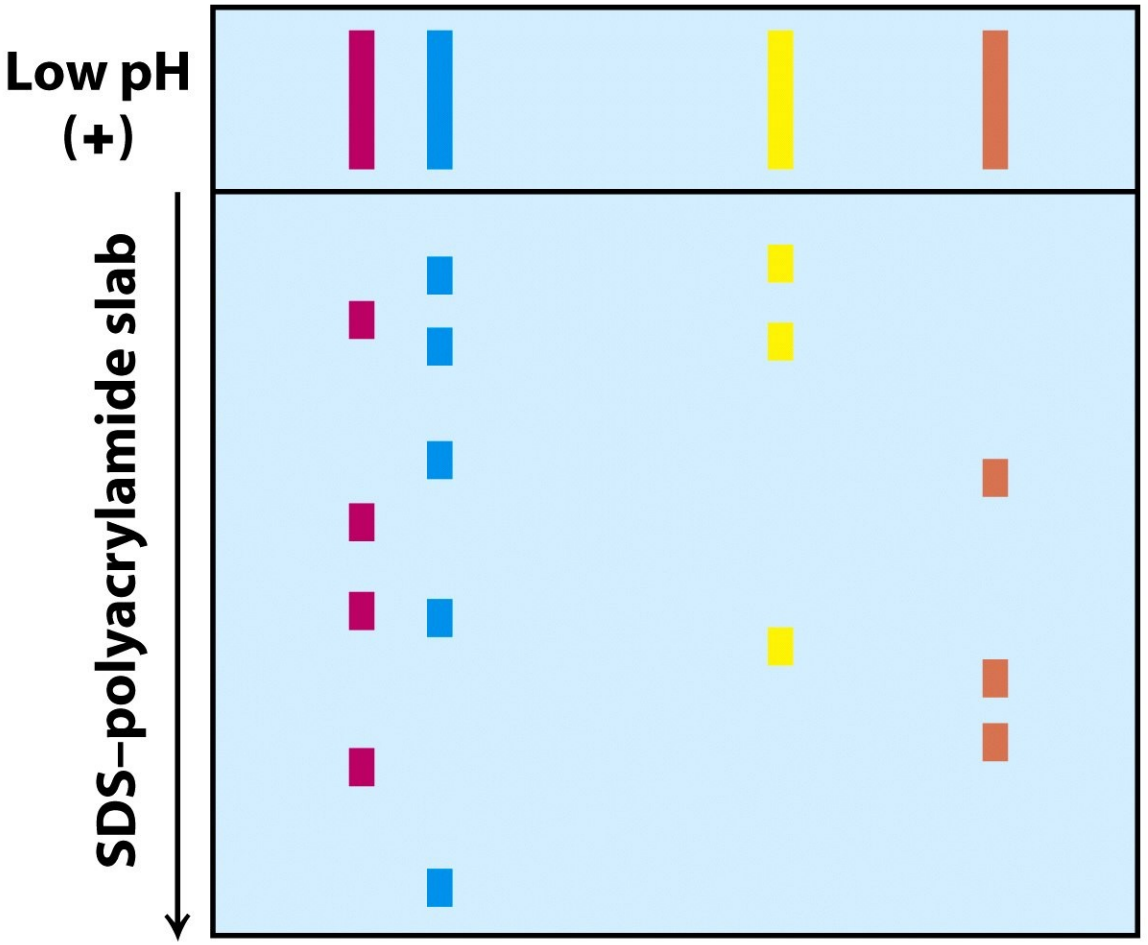


Figure 3-12a
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 74

Two-dimensional Gel Electrophoresis

Figure 3-12b
Biochemistry, Sixth Edition

- Proteins from *E. coli* separated by 2D-electrophoresis
- >1,000 proteins can be resolved

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 74

Mass Spectrometry

- MALDI-TOF mass spectrometry
- Protein sample is ionized and exposed to electrical field
- Ions travel according to size

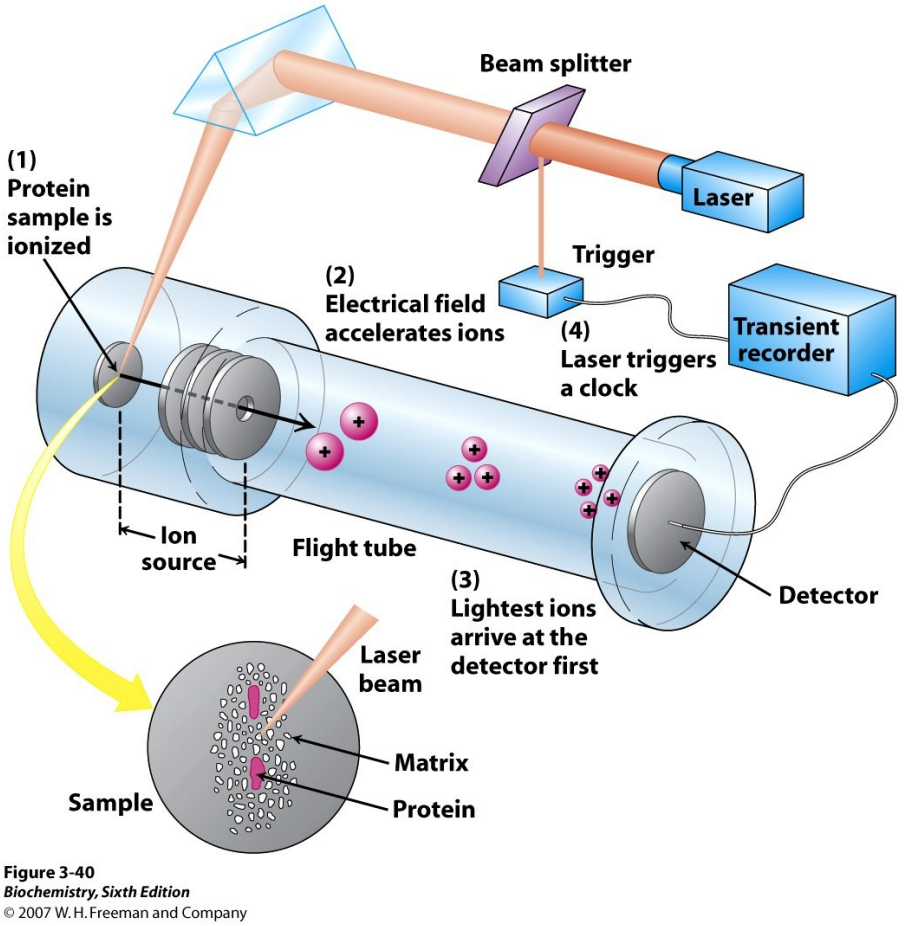


Figure 3-40
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 94

MALDI-TOF Mass Spectrum

- MALDI-TOF gives good estimates of molecular weights
- Can be used to identify a few proteins within a mixture

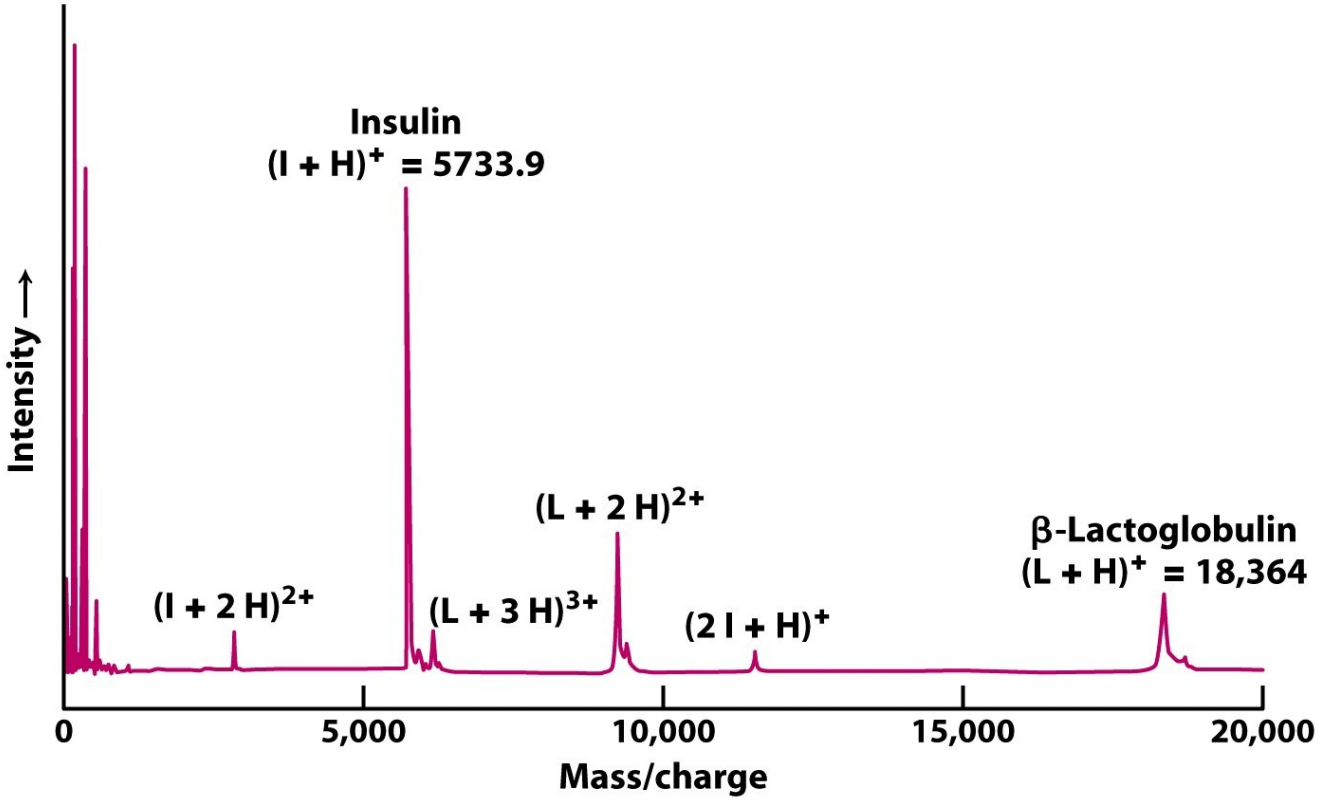


Figure 3-41
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 94

Proteomic Analysis by Mass Spectrometry

- Proteins separated by 2D electrophoresis
- Single proteins eluted
- Digestion with trypsin will give fragments with unique set of sizes
- Sizes identified by mass spectrometry and matched to database
- Allows identification of unknown proteins

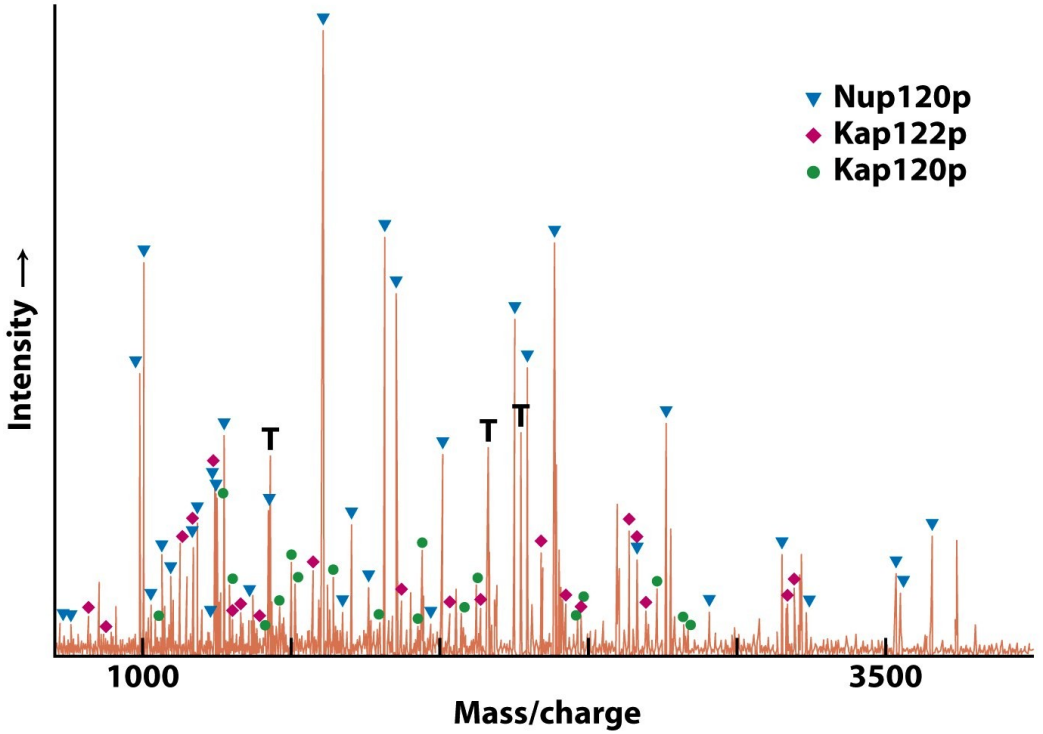


Figure 3-42
Biochemistry, Sixth Edition

Berg, Tymoczko & Stryer, "Biochemistry", 6th edn, 2006, p. 95

Transcriptomics v Proteomics

- Transcriptomics and proteomics are both very powerful
- Differences in their practical application:
  - Transcriptomics is robust, relatively cost-effective and user-friendly
  - Proteomics still relatively limited – problems can remain with purification and stability of proteins
- Increasing potential to combine and compare data sets - for discussion see Hegde *et al*., *Curr. Opin. Biotech.*, **14**, 647-651 (2003)

Bioinformatics: Mining the Data

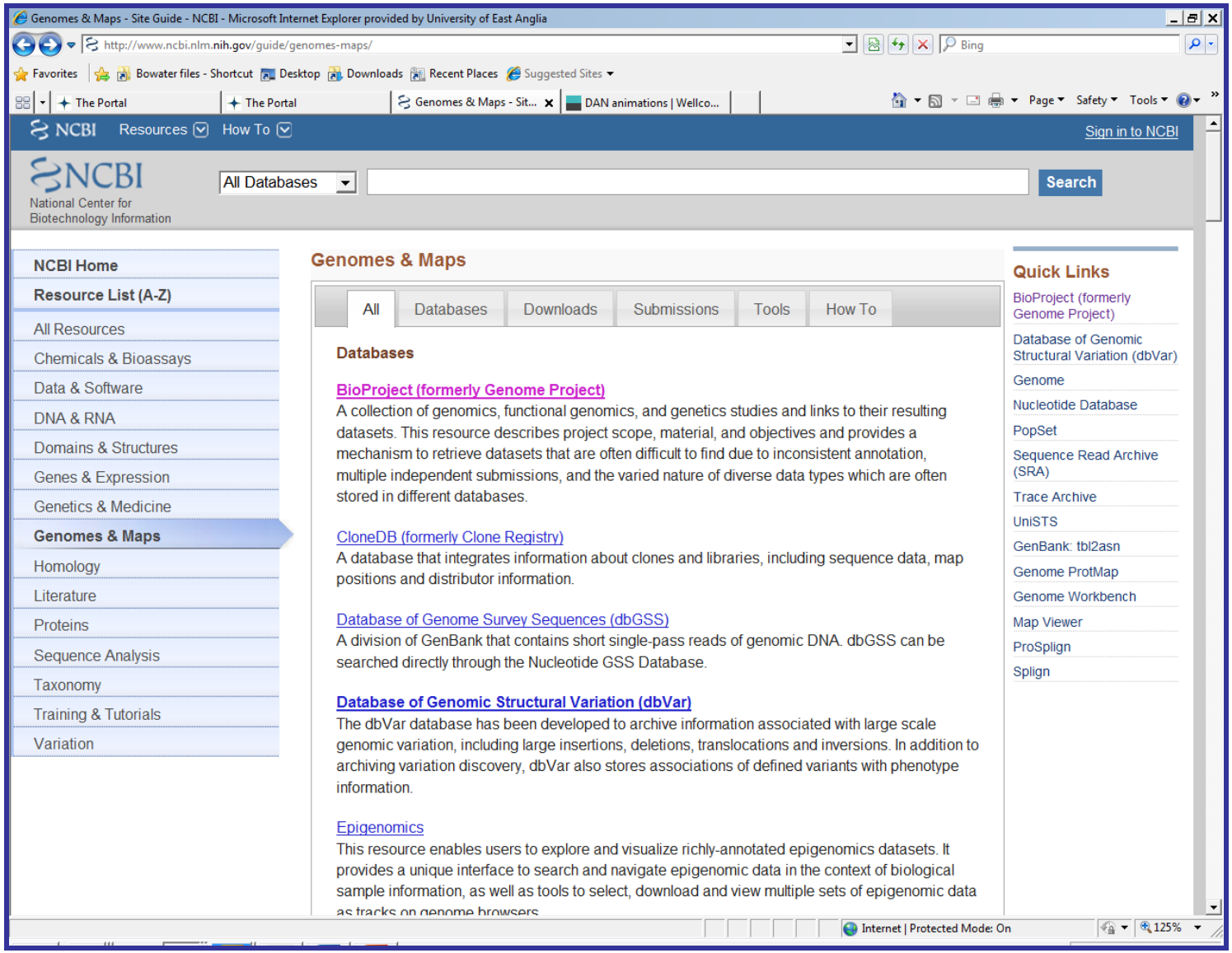
Bioinformatics & Databases

- Latest biological data is gathered, organised and disseminated through large databases
- Databases include:
  - EBI, NCBI, Pfam, SMART, SWISS-PROT, TAIR
- Information in bioinformatic databases:
  - sequences, structures, homology searches
- Fast search engines allow searches by all with internet access – databases are as useful as the results they help generate!
- Improved tools for analysis of sequences

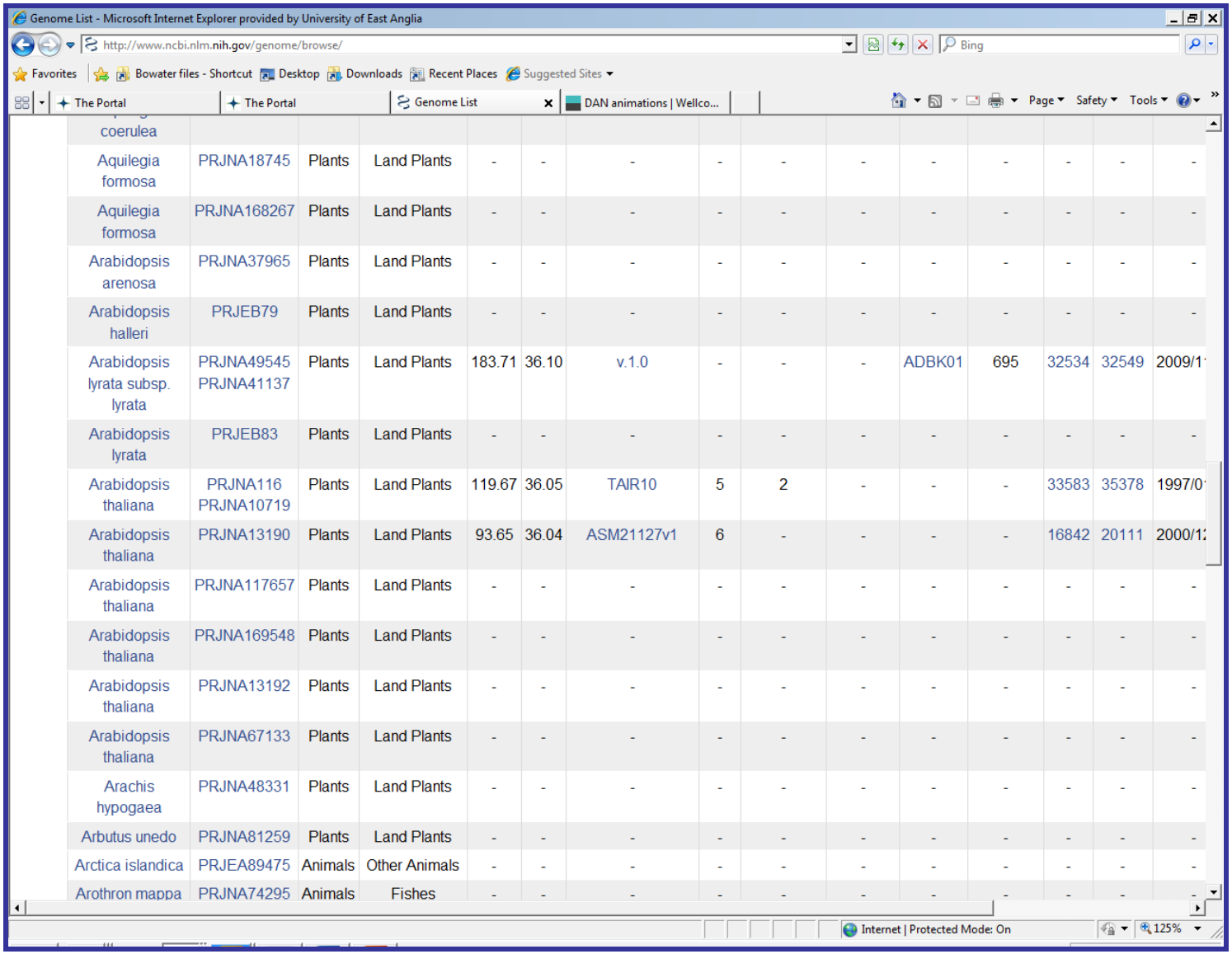
Databases – Some URLs

Resource	URL
European Bioinformatics Institute	www.ebi.ac.uk/
GenBank	www.ncbi.nlm.nih.gov/Genbank/
NCBI	www.ncbi.nlm.nih.gov/
Protein DataBank	http://www.rcsb.org/pdb/home/home.do
Sanger Centre	www.sanger.ac.uk/
SMART	smart.embl-heidelberg.de
The Arabidopsis Information Resource (TAIR)	www.arabidopsis.org/

NCBI: Complete Genomes

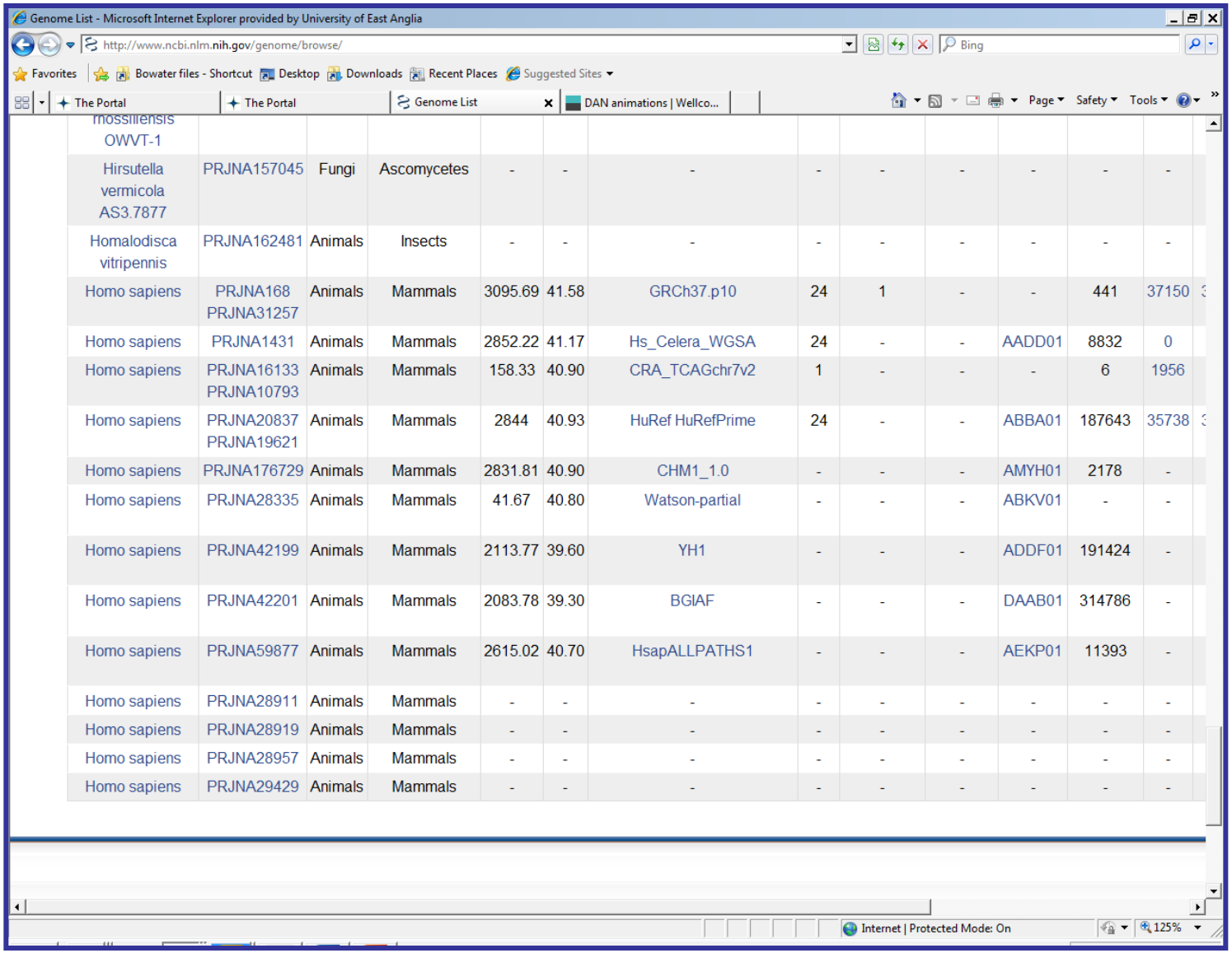


NCBI: Eukaryotic Genomes



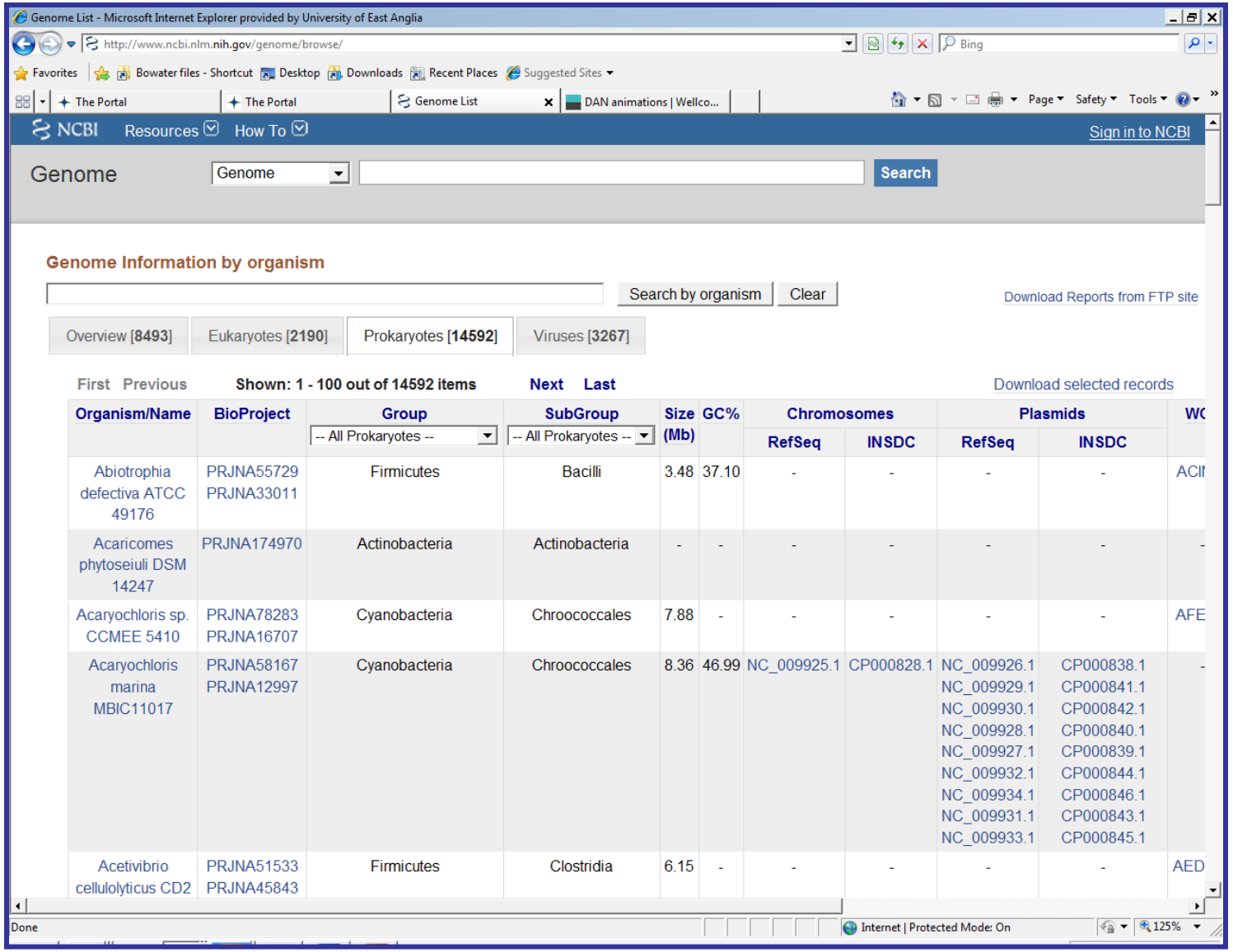
coerulea														
Aquilegia formosa	PRJNA18745	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Aquilegia formosa	PRJNA168267	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis arenosa	PRJNA37965	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis halleri	PRJEB79	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis lyrata subsp. lyrata	PRJNA49545 PRJNA41137	Plants	Land Plants	183.71	36.10	v.1.0	-	-	-	ADBK01	695	32534	32549	2009/1
Arabidopsis lyrata	PRJEB83	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis thaliana	PRJNA116 PRJNA10719	Plants	Land Plants	119.67	36.05	TAIR10	5	2	-	-	-	33583	35378	1997/0
Arabidopsis thaliana	PRJNA13190	Plants	Land Plants	93.65	36.04	ASM21127v1	6	-	-	-	-	16842	20111	2000/1
Arabidopsis thaliana	PRJNA117657	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis thaliana	PRJNA169548	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis thaliana	PRJNA13192	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arabidopsis thaliana	PRJNA67133	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arachis hypogaea	PRJNA48331	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arbutus unedo	PRJNA81259	Plants	Land Plants	-	-	-	-	-	-	-	-	-	-	-
Arctica islandica	PRJEA89475	Animals	Other Animals	-	-	-	-	-	-	-	-	-	-	-
Arothron mappa	PRJNA74295	Animals	Fishes	-	-	-	-	-	-	-	-	-	-	-

NCBI: Eukaryotic Genomes

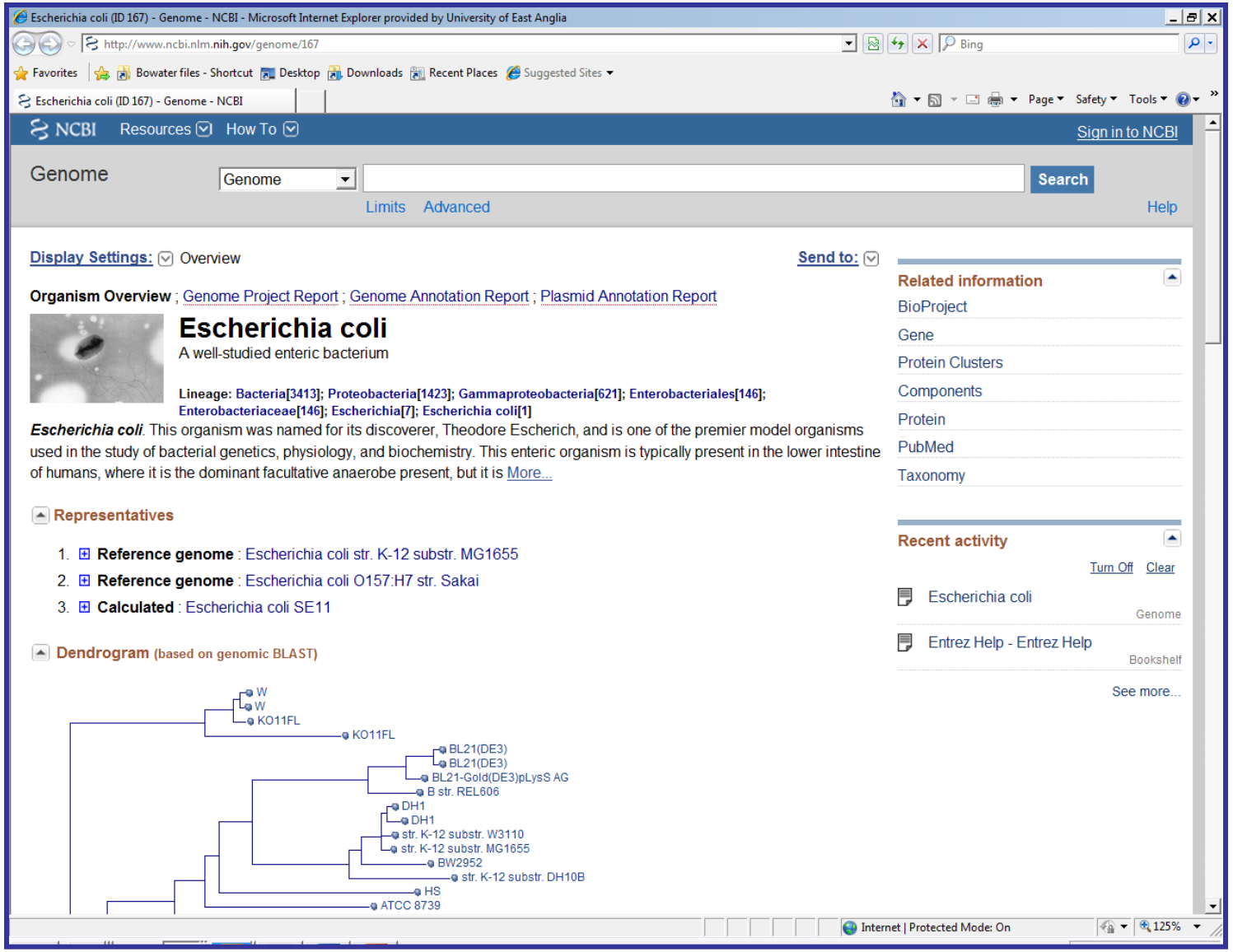


mossiliensis OWVT-1												
Hirsutella vermicola AS3.7877	PRJNA157045	Fungi	Ascomycetes	-	-	-	-	-	-	-	-	-
Homalodisca vitripennis	PRJNA162481	Animals	Insects	-	-	-	-	-	-	-	-	-
Homo sapiens	PRJNA168 PRJNA31257	Animals	Mammals	3095.69	41.58	GRCh37.p10	24	1	-	-	441	37150
Homo sapiens	PRJNA1431	Animals	Mammals	2852.22	41.17	Hs_Celera_WGSA	24	-	-	AADD01	8832	0
Homo sapiens	PRJNA16133 PRJNA10793	Animals	Mammals	158.33	40.90	CRA_TCAGchr7v2	1	-	-	-	6	1956
Homo sapiens	PRJNA20837 PRJNA19621	Animals	Mammals	2844	40.93	HuRef HuRefPrime	24	-	-	ABBA01	187643	35738
Homo sapiens	PRJNA176729	Animals	Mammals	2831.81	40.90	CHM1_1.0	-	-	-	AMYH01	2178	-
Homo sapiens	PRJNA28335	Animals	Mammals	41.67	40.80	Watson-partial	-	-	-	ABKV01	-	-
Homo sapiens	PRJNA42199	Animals	Mammals	2113.77	39.60	YH1	-	-	-	ADDF01	191424	-
Homo sapiens	PRJNA42201	Animals	Mammals	2083.78	39.30	BGIAF	-	-	-	DAAB01	314786	-
Homo sapiens	PRJNA59877	Animals	Mammals	2615.02	40.70	HsapALLPATHS1	-	-	-	AEKP01	11393	-
Homo sapiens	PRJNA28911	Animals	Mammals	-	-	-	-	-	-	-	-	-
Homo sapiens	PRJNA28919	Animals	Mammals	-	-	-	-	-	-	-	-	-
Homo sapiens	PRJNA28957	Animals	Mammals	-	-	-	-	-	-	-	-	-
Homo sapiens	PRJNA29429	Animals	Mammals	-	-	-	-	-	-	-	-	-

NCBI: Microbial Genomes



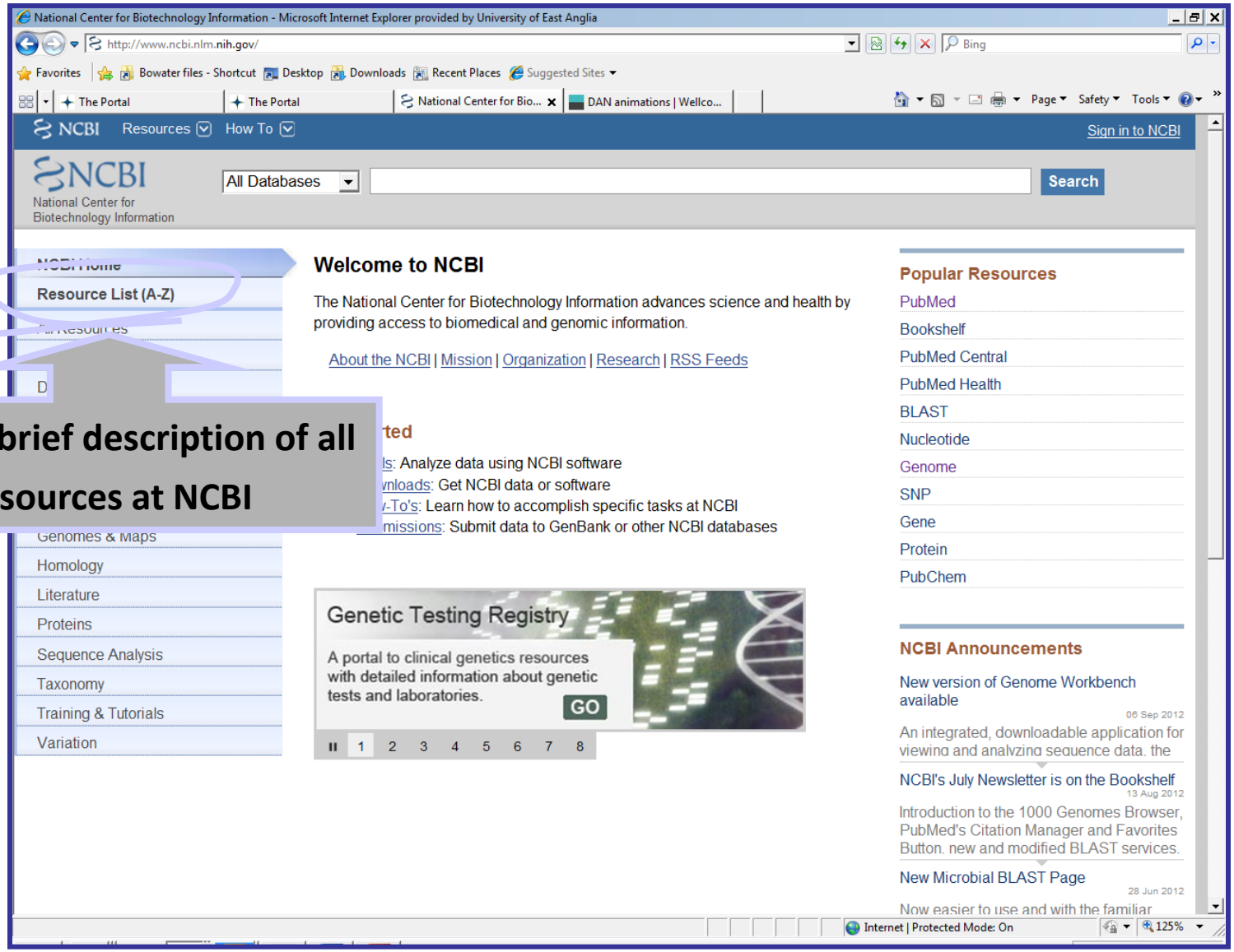
NCBI: Microbial Genomes



Databases - The Caveats

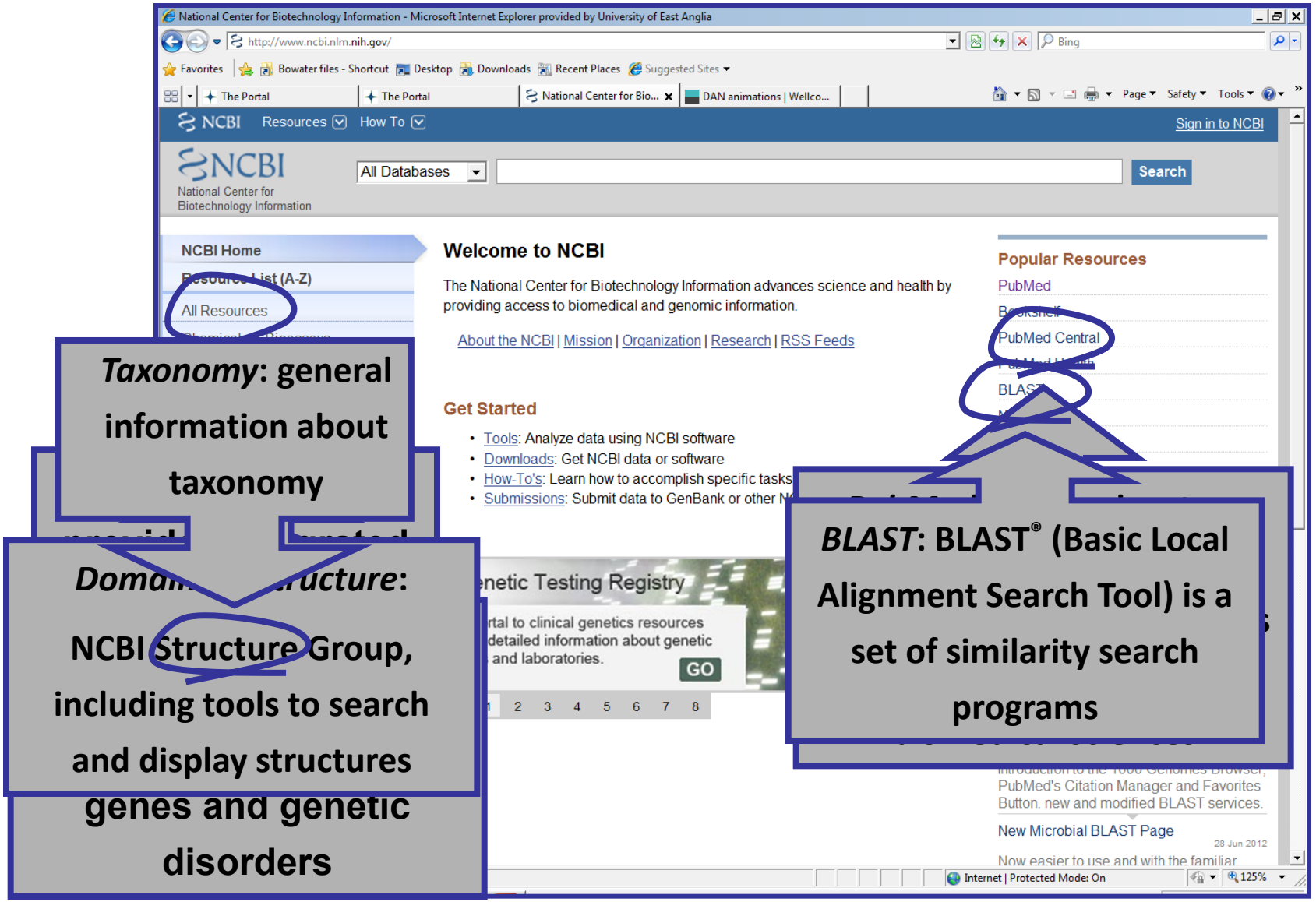
- Databases contain mistakes (low as a proportion of total data)
  - primary data errors
  - data analysis errors
  - annotation errors
- Errors are difficult to correct
- Make the interpretation of data your own responsibility!!

NCBI – Useful Links



Links to brief description of all resources at NCBI

NCBI – Useful Links



Taxonomy: general information about taxonomy

Domain Structure: NCBI Structure Group, including tools to search and display structures

genes and genetic disorders

BLAST: BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs

Databases Summary

- Many databases are available
  - some have lot of general information (NCBI, EBI)
  - some have specific data (Pfam, SWISS-PROT)
  - some relate to specific research interests (TAIR)
- Become well acquainted with specific databases
- Wide range of databases, web sites and other resources are available for *in silico* analysis of biological data
- Great advantages, but beware caveats and potential pitfalls – understand capabilities and limitations!
- Use information intelligently:
  - always ask if the conclusions make biological sense
  - may require further analyses or experimentation

“Omics” Overview

- Analyses of various biological molecules have developed into their own scientific disciplines; e.g. *Metabolomics* = “analysis of metabolites”, etc.
- *Transcriptome*: set of all mRNAs ("transcripts”) produced from a genome
- *Proteome*: set of all proteins produced under a given set of conditions
- Both can vary because they reflect genes that are actively expressed at any given time
- Transcriptomics and proteomics are both powerful, but are used differently: transcriptomics is cheaper and more user friendly than proteomics