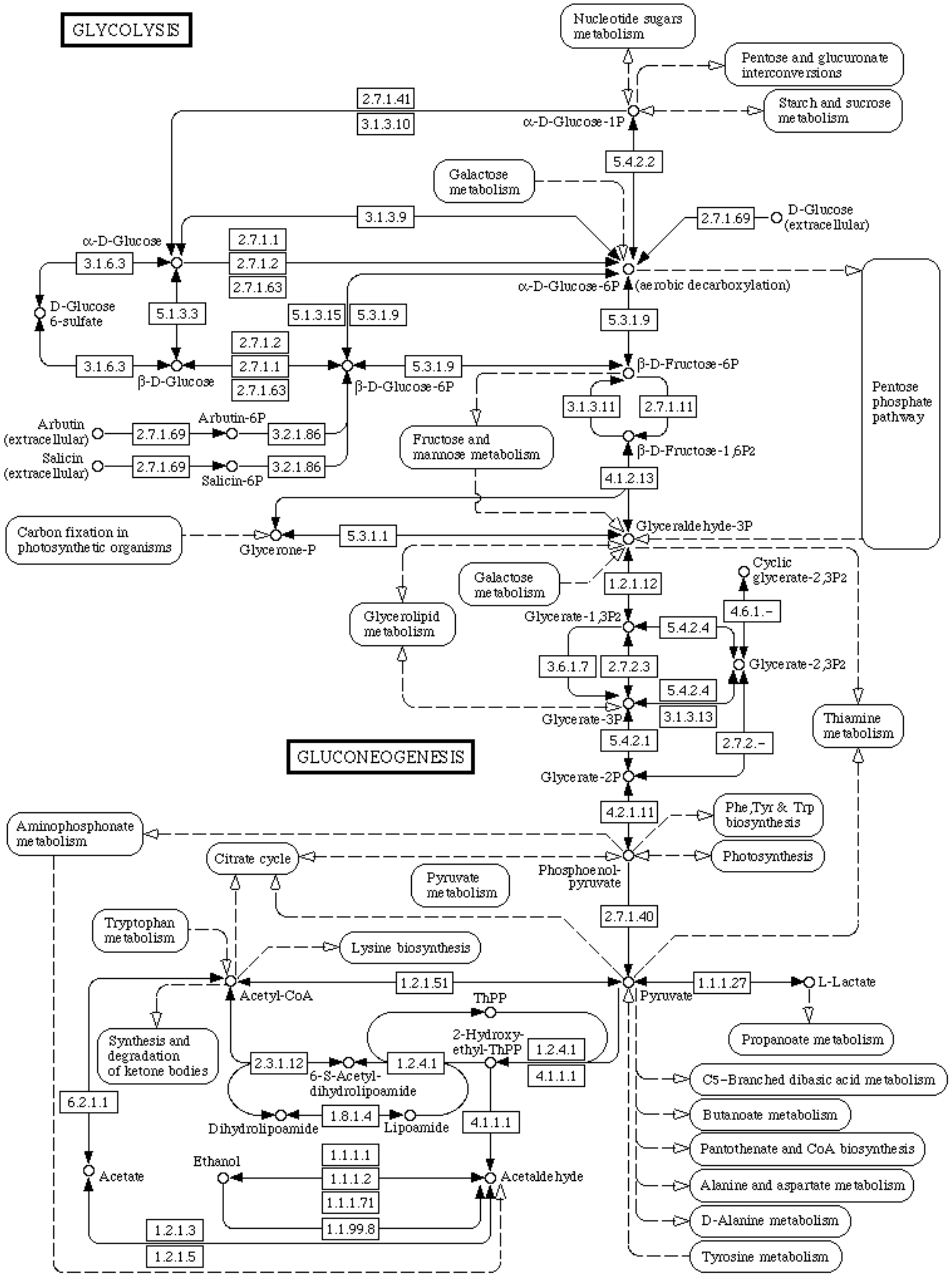
CMSC423: Bioinformatic Algorithms, Databases and Tools Lecture 22

Gene networks

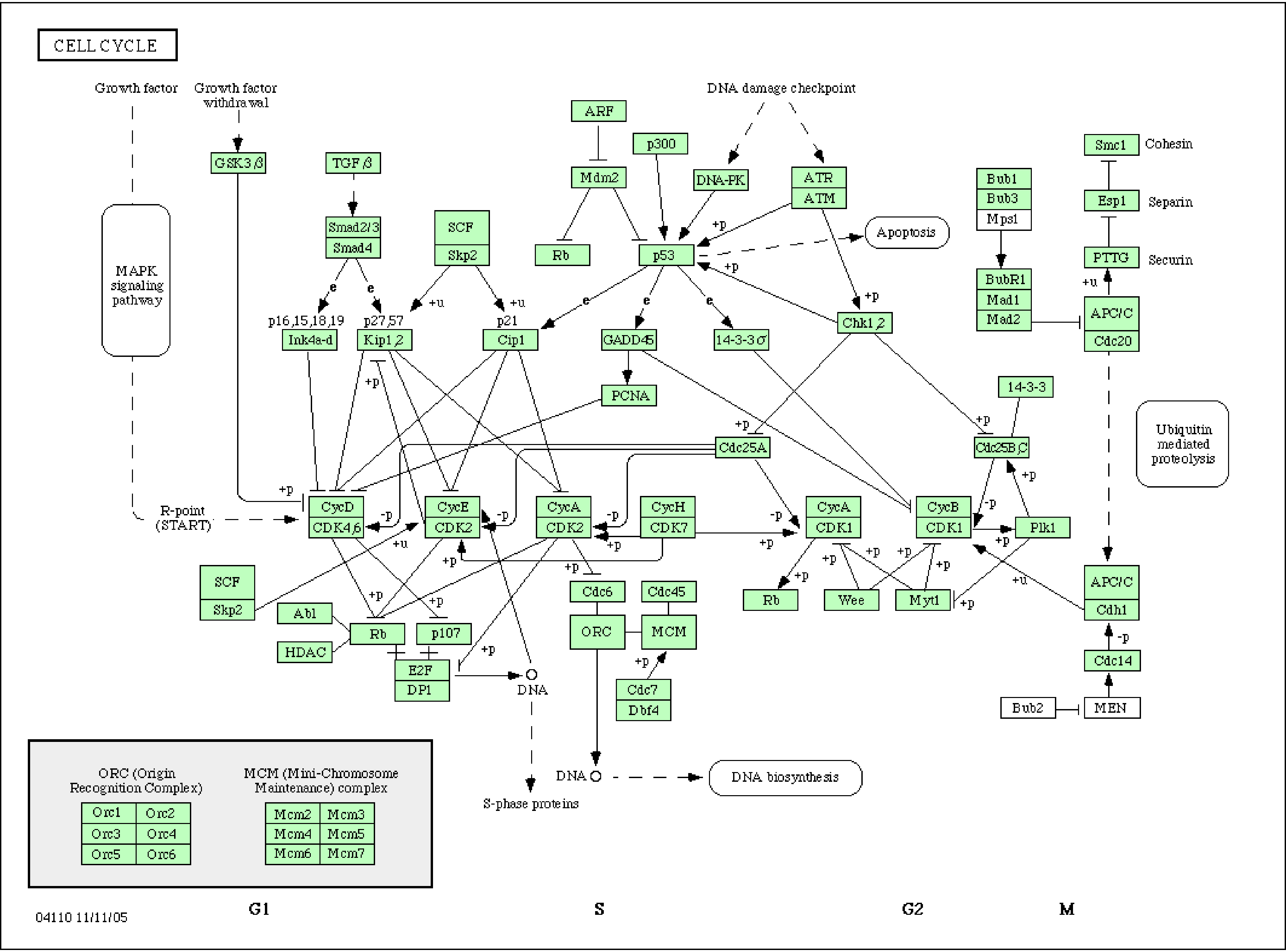
Real-life examples

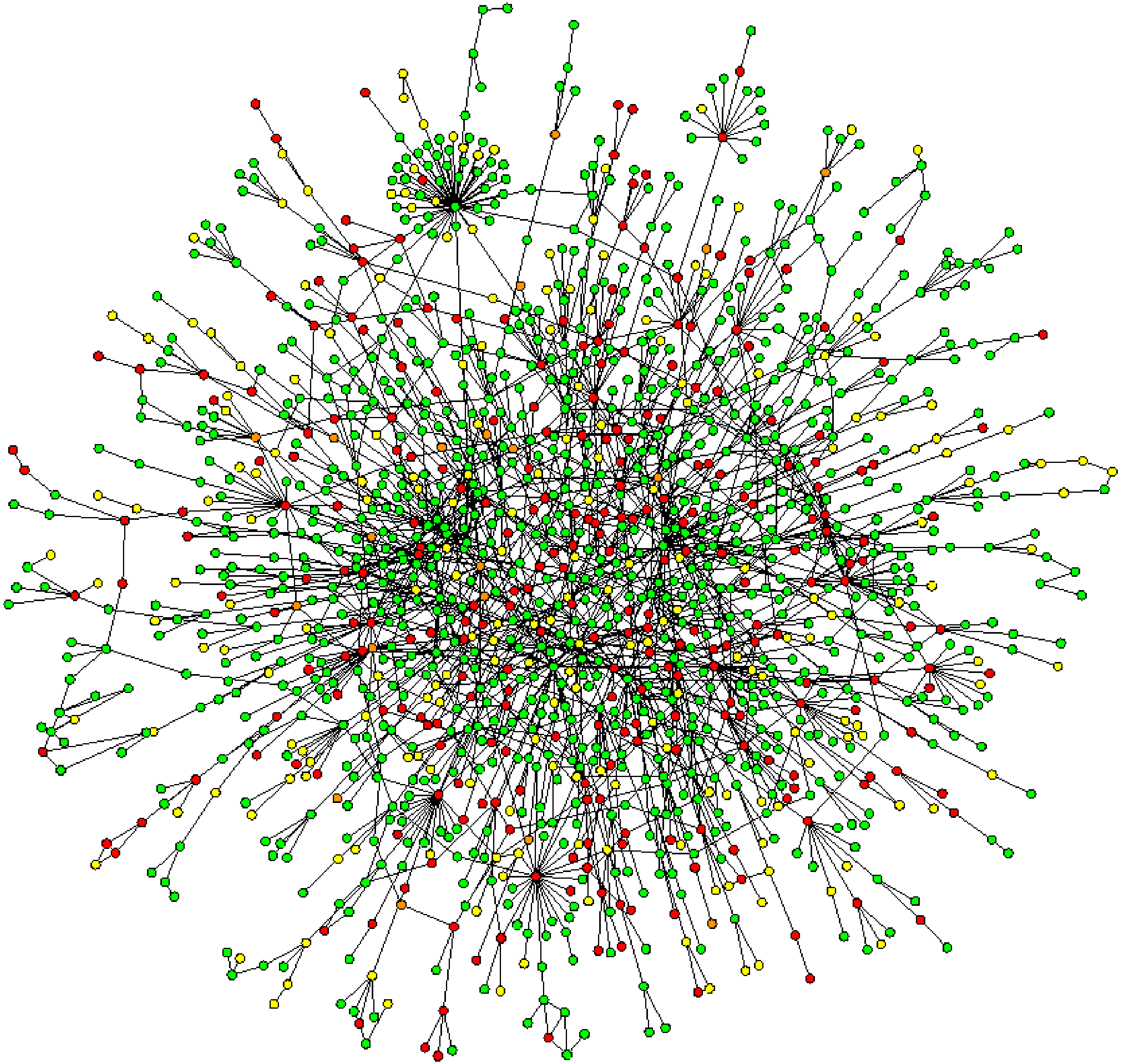
Biological networks

- Genes/proteins do not exist in isolation
- Interactions between genes or proteins can be represented as graphs
- Examples:
  - metabolic pathways
  - regulatory networks
  - protein-protein interactions (e.g. yeast 2-hybrid)
  - genetic interactions (synthetic lethality)


GLYCOLYSIS
GLUCONEOGENESIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
α-D-Glucose-1P
2.7.1.41
3.1.3.10
5.4.2.2
Galactose metabolism
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
α-D-Glucose
3.1.6.3
2.7.1.1
2.7.1.2
2.7.1.63
α-D-Glucose-6P
(aerobic decarboxylation)
D-Glucose 6-sulfate
5.1.3.3
5.1.3.15
5.3.1.9
Pentose phosphate pathway
β-D-Glucose
β-D-Glucose-6P
β-D-Fructose-6P
3.1.3.11
2.7.1.11
β-D-Fructose-1,6P2
Arbutin (extracellular)
Arbutin-6P
3.2.1.86
Salicin (extracellular)
Salicin-6P
Fructose and mannose metabolism
4.1.2.13
Carbon fixation in photosynthetic organisms
Glycerone-P
5.3.1.1
Glyceraldehyde-3P
Galactose metabolism
1.2.1.12
Cyclic glycerate-2,3P2
4.6.1.-
Glycerolipid metabolism
Glycerate-1,3P2
5.4.2.4
Glycerate-2,3P2
3.6.1.7
2.7.2.3
Glycerate-3P
3.1.3.13
5.4.2.1
2.7.2.-
Thiamine metabolism
Glycerate-2P
4.2.1.11
Phe,Tyr & Trp biosynthesis
Photosynthesis
Aminophosphonate metabolism
Citrate cycle
Pyruvate metabolism
Phosphoenol-pyruvate
2.7.1.40
Tryptophan metabolism
Lysine biosynthesis
Acetyl-CoA
1.2.1.51
Pyruvate
1.1.1.27
L-Lactate
Propanoate metabolism
ThPP
2-Hydroxy-ethyl-ThPP
1.2.4.1
4.1.1.1
Synthesis and degradation of ketone bodies
2.3.1.12
6-S-Acetyl-dihydrolipoamide
1.2.4.1
C5-Branched dibasic acid metabolism
Butanoate metabolism
6.2.1.1
Dihydrolipoamide
1.8.1.4
Lipoamide
4.1.1.1
Pantothenate and CoA biosynthesis
Acetate
Ethanol
1.1.1.1
1.1.1.2
1.1.1.71
1.1.99.8
Acetaldehyde
Alanine and aspartate metabolism
D-Alanine metabolism
1.2.1.3
1.2.1.5
Tyrosine metabolism

00010 3/23/06


CELL CYCLE
Growth factor
Growth factor withdrawal
MAPK signaling pathway
GSK3β
TGFβ
Smad2/3
Smad4
SCF
Skp2
ARF
Mdm2
p300
DNA damage checkpoint
DNA-PK
ATR
ATM
Rb
p53
Apoptosis
Bub1
Bub3
Mps1
BubR1
Mad1
Mad2
Smc1
Cohesin
Esp1
Separin
PTTG
Securin
APC/C
Cdc20
p16,15,18,19
Ink4a-d
p27,57
Kip1,2
p21
Cip1
GADD45
14-3-3σ
Chk1,2
PCNA
14-3-3
Cdc25A
Cdc25B,C
Ubiquitin mediated proteolysis
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
Abl
HDAC
Rb
p107
E2F
DP1
DNA
Cdc6
Cdc45
ORC
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
APC/C
Cdh1
Cdc14
Bub2
MEN
S-phase proteins
DNA
DNA biosynthesis
e
+p
-p
+u
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
04110 11/11/05
G1
S
G2
M



Gene networks research at UMD

- Active area of research in Carl Kingsford's lab
- Data will be generated in Najib El Sayed's lab
- My own research on microbial communities will translate into such data.

Metagenomics

Human microbiome

- Gill, S.R., et al., *Metagenomic analysis of the human distal gut microbiome.* Science, 2006. **312**(5778): p. 1355-9.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citatio
- Examine all bacteria in an environment (human gut) at the same time using high-throughput techniques

Why the gut biome?
We are what we eat

- Majority of human commensal bacteria live in the gut
  (more bacterial cells than human cells by an order of magnitude – 100 trillion bacterial cells)
- We rely on gut bacteria for nutrition
- Gut bacteria important for our development
- Imbalances in bacterial populations correlate with disease

Environment “exploration”

- Culture-based
  - heavily biased (1-5% bacteria easily cultured)
  - amenable to many types of analyses
- Directed rRNA sequencing
  - less biased
  - limited analyses possible
- Random shotgun sequencing
  - “differently” biased
  - amenable to many types of analyses
  - $$$

Project overview

- Collaboration between TIGR, Stanford, and Washington University (St. Louis)
- Sequenced fecal samples from two healthy individuals

(XX, XY) (veg+,veg-) correlation lost due to IRB

- Also performed “traditional” amplified 16S rDNA sequencing

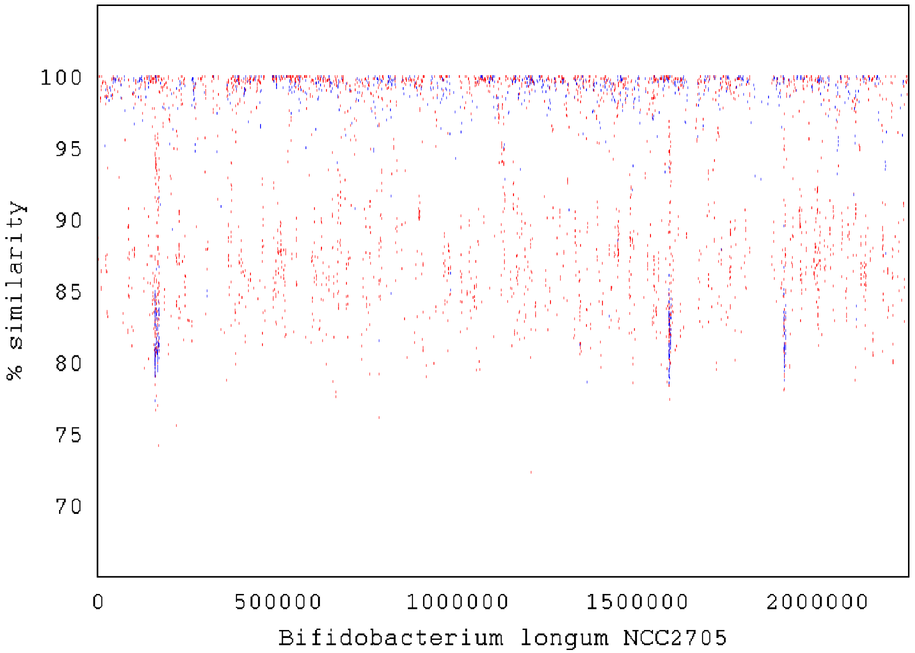
	Subject 1	Subject 2	Total
Shotgun reads	65,059	74,462	139,521
amplified 16S rDNA clones	3,514	3,601	7,115

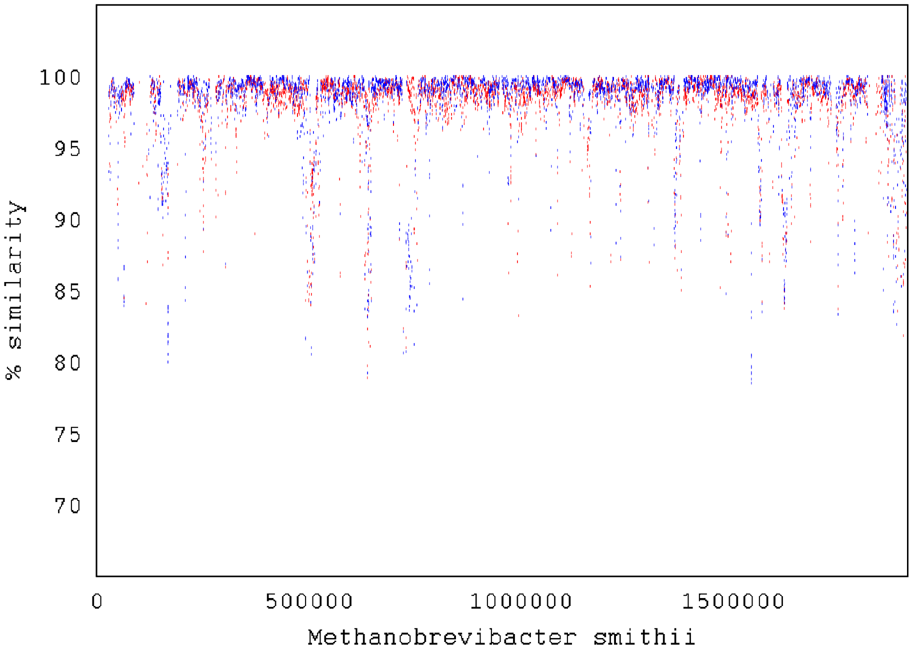
All shotgun reads from ~ 2 kbp library

Metagenomic pipeline

- Assembly (graph theory, string matching)
  - puzzle-together shotgun reads into contigs and scaffolds
- Gene finding (machine learning)
- Binning (clustering, statistics)
  - assign each contig to a taxonomic unit
- Annotation (natural language processing)
  - gene roles, pathways, orthologous groups, etc
- Analysis (statistics, graph theory, data visualization)
  - diversity
  - comparison between environments
  - metabolic potential
  - etc.

Comparative Assembly (AMOScmp)





me size	2.26 MB	~1.9 MB
rage	0.7	3.5
tigs	789	222
es	988,707	1,538,516

> 50% of archaeal contigs are likely *M. smithii*

Binning results

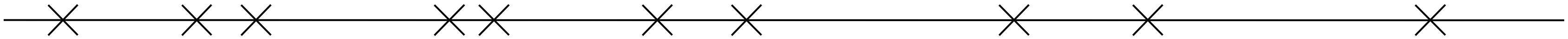
Order	amplified rRNA clones		shotgun rRNA (bases)		shotgun blastx(bases)	
Subject	1	2	1	2	1	2
Clostridiales	2,777 <	3,386	70,055 <	102,140	4,396,295 <	5,562,074
Bifidobacteriales	30 >	0	31,443 >	5,101	2,882,267 >	851,278
Coriobacteriales	4	6	25,781	10,804	**0**	**0**
Methanobacteriales	**0**	**0**	18,188	17,970	943,256	946,329

Metagenomics...

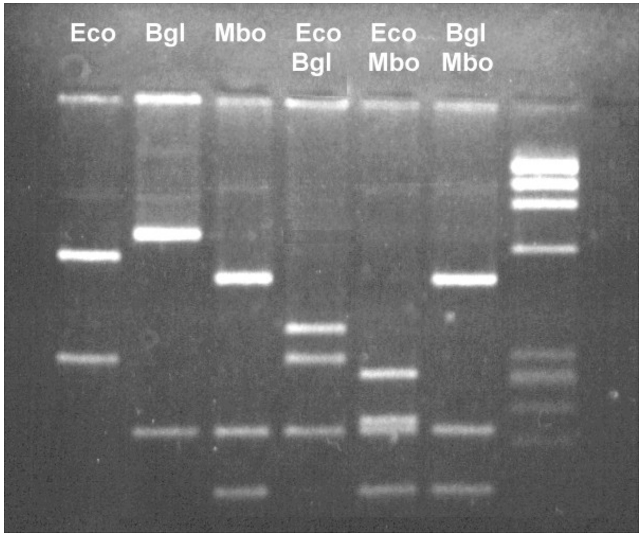
- This work is ongoing at UMD with support from NSF and NIH
- Paid summer internships available – contact me if you are interested.

Assembly with optical maps

Optical mapping data



- Restriction mapping (set/bag of fragment sizes)
  - restriction digest
  - spectrum of sizes defines “fingerprint”

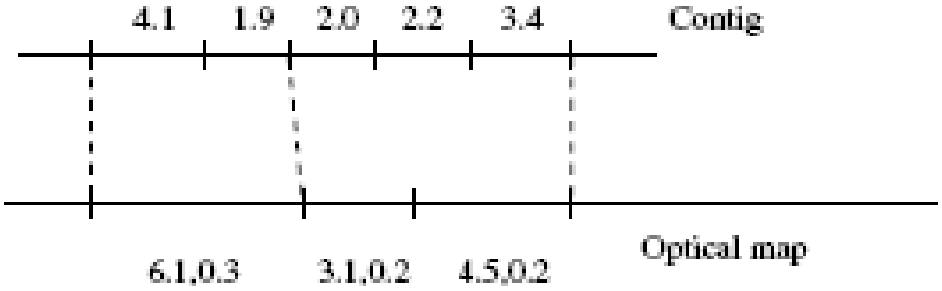


#.	size	(stdev)
1.	1.2	(0.3)
2.	4.1	(0.8)
3.	2.2	(0.5)
...		

- Optical mapping (list/array of fragment

Contig matching problem

- Find “best” placement of a contig on the map



$$\chi^2 \text{score} = \sum_{k=1}^{j} \left(\frac{c_k - o_k}{\sigma_k} \right)$$

- by best we mean:
  - most matched sites
  - best correspondence between fragment sizes

$$\left| \sum_{i=s}^{t} c_i - \sum_{j=u}^{v} o_j \right| \le C_\sigma \sqrt{\sum_{j=u}^{v} \sigma_j^2}$$

Solution to the matching problem

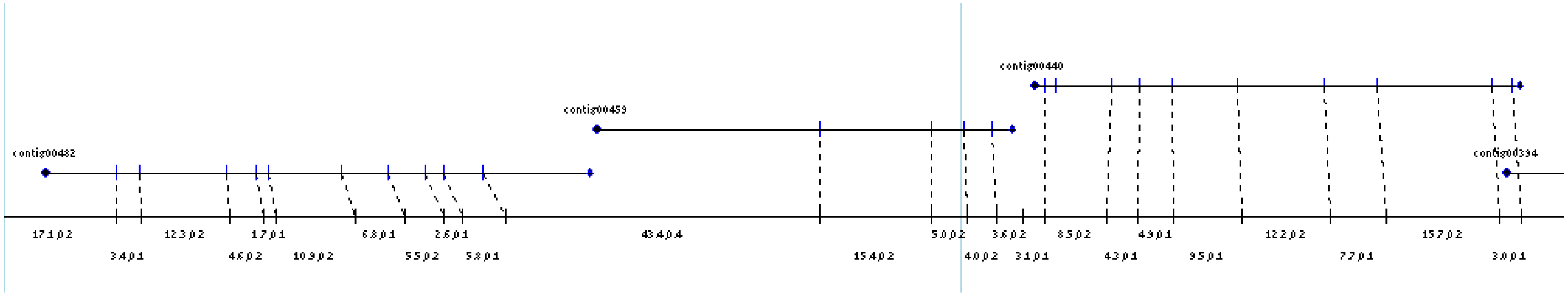
- Simple dynamic programming ($O(m^2n^2)$)

$$S[i,j]=max_{0\leq k\leq i,0\leq l\leq j}-C_r\times(i-k+l-j)-\frac{(\sum_{s=k}^{i}c_s-\sum_{t=l}^{j}o_t)^2}{\sum_{t=l}^{j}\sigma_t^2}+S[k-1,l-1]$$

- Main challenge: this procedure always returns a “best” match

- Solution:

Results – real data



Yersinia kristensenii

Optical map: 350 sites (AFLII)

Assembly: 86 contigs, 404 sites

48 contigs have > 1 site

45 contigs can be placed

30 unique matches
15 placed by greedy

4.4Mb (93%) in scaffold

Yersinia aldovae

Optical map: 360 sites (AFLII)

Assembly: 104 contigs, 411 sites

58 contigs have > 1 site

52 contigs can be placed

31 have unique matches
21 placed by greedy

3.7Mb (88%) in scaffold

Un-placed contigs appear to be mis-assemblies

Nagarajan, Read, Pop. Bioinformatics 2008.

With Niranjan Nagarajan

Voxelation

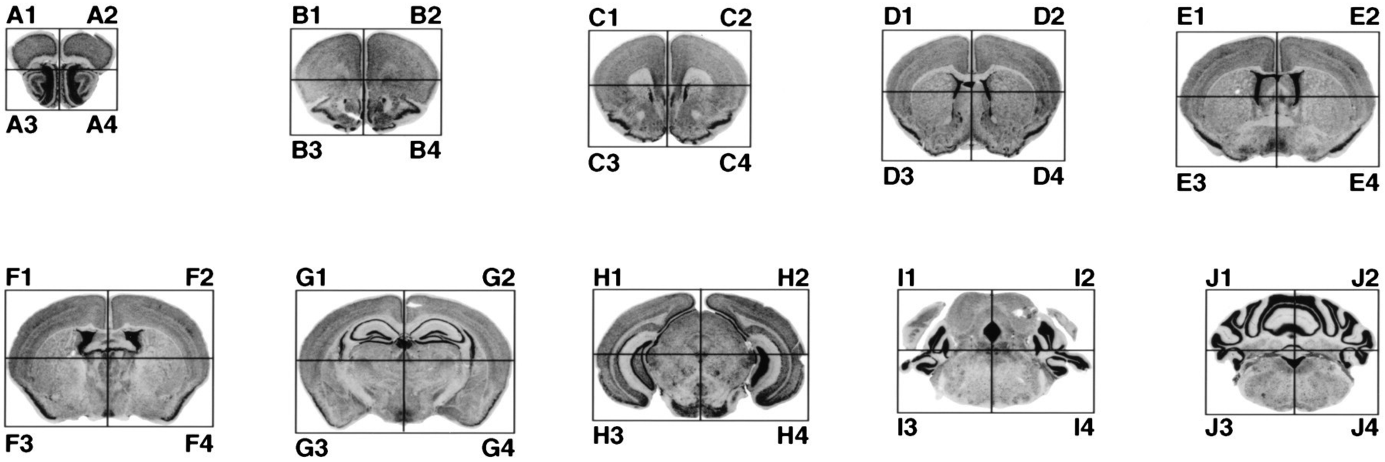
Voxelation

- Brown, V.M., et al., *High-throughput imaging of brain gene expression.* Genome Res, 2002. **12**(2): p. 244-54.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citatio

- Brown, V.M., et al., *Multiplex three-dimensional brain gene expression mapping in a mouse model of Parkinson's disease.* Genome Res, 2002. **12**(6): p. 868-84.
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citatio

- Gene expression information in a spatial context
- Combines microarray analysis with computer graphics

Figure 2 Voxelation scheme

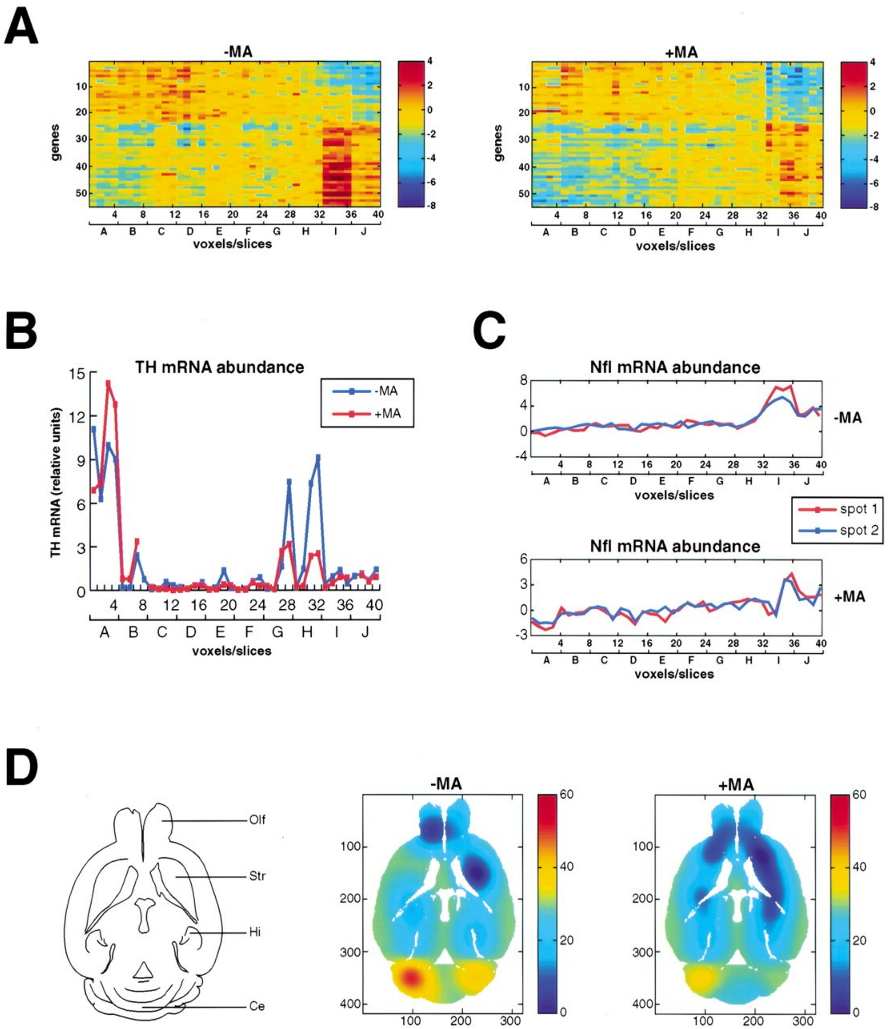


Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

- Mouse brain cut up into voxels
- Run a separate microarray experiment on each voxel

Cold Spring Harbor Laboratory Press

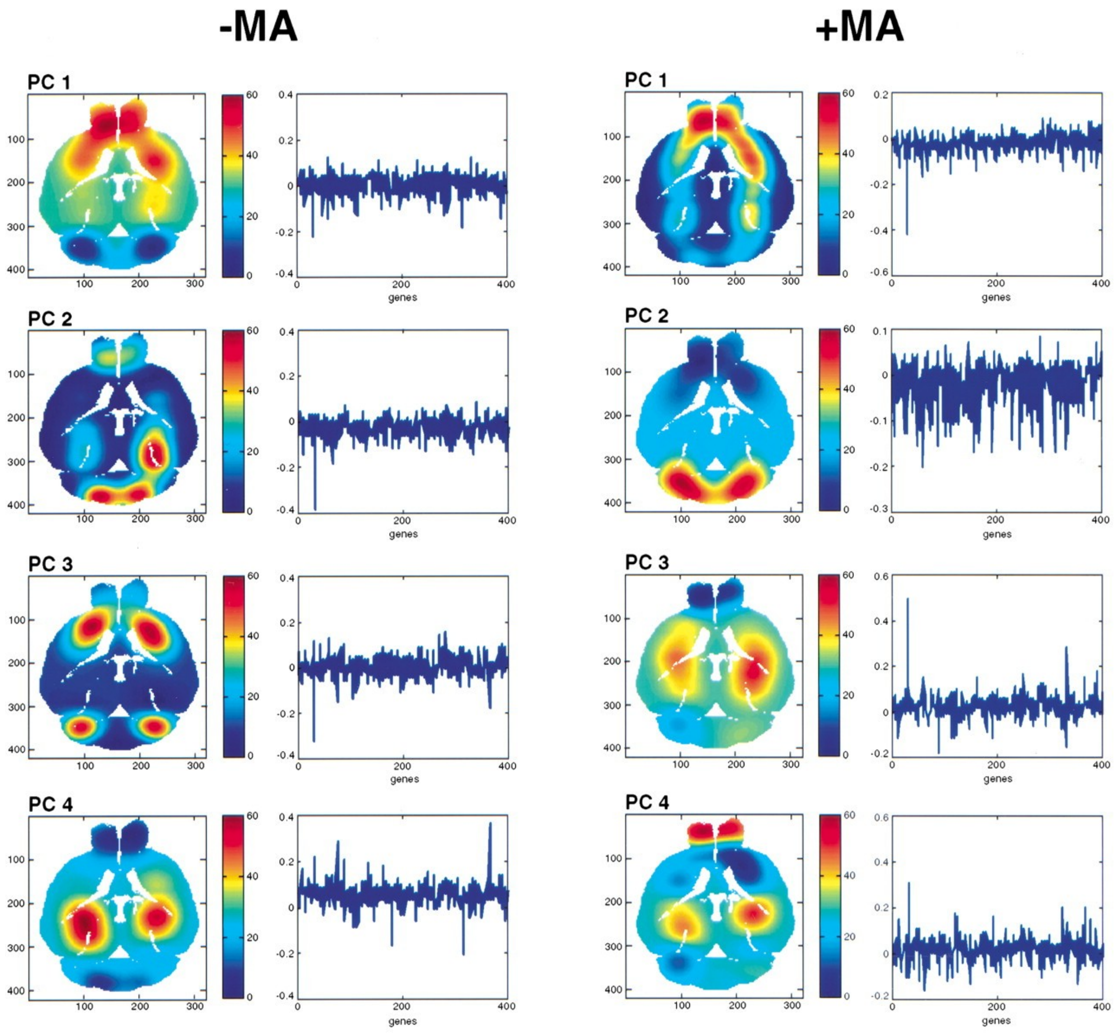
Figure 4 Spatial gene expression patterns for the subset of correlated genes



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Cold Spring Harbor Laboratory Press

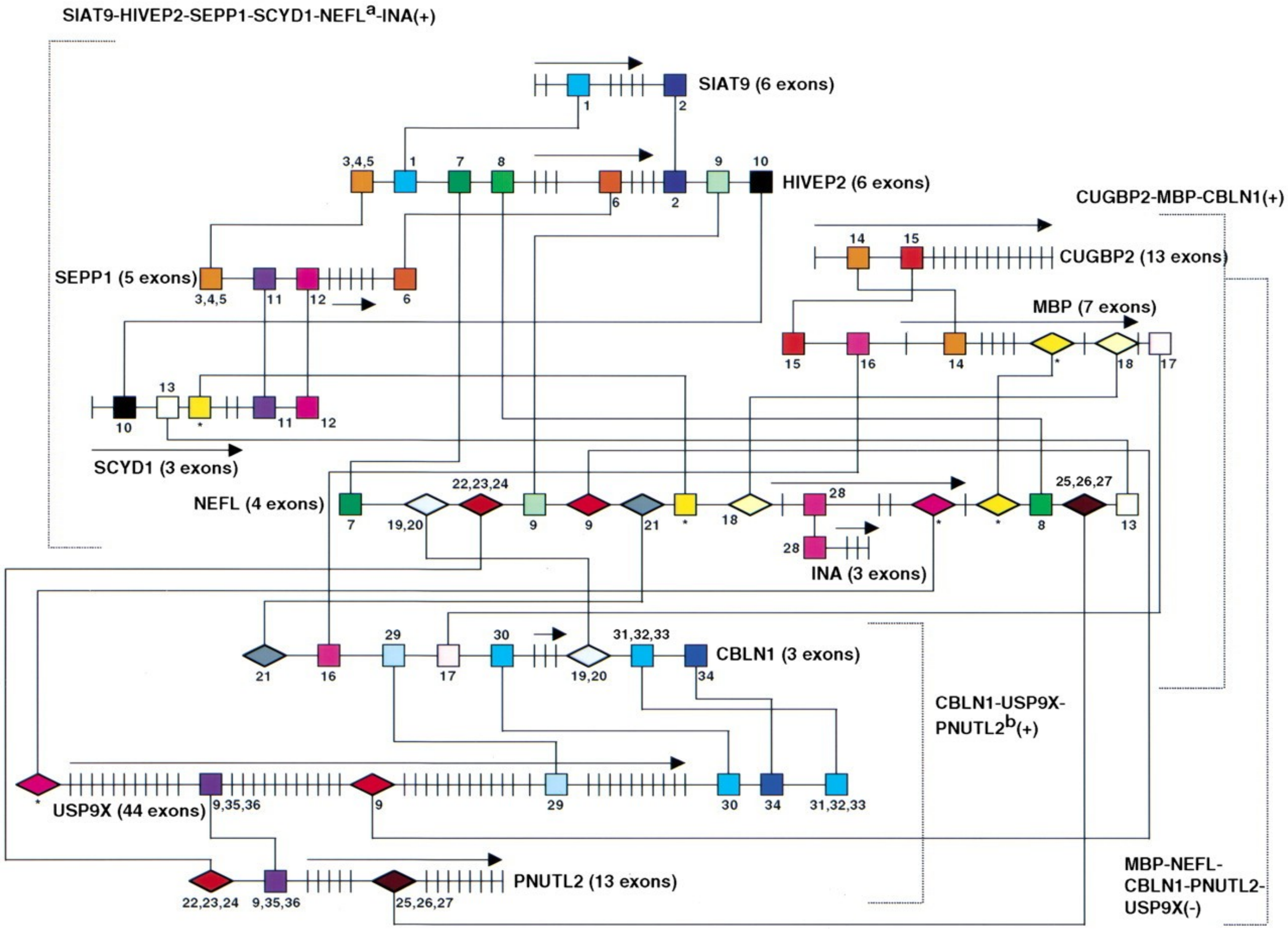
Figure 7 SVD delineates anatomical regions of the brain



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Cold Spring Harbor Laboratory Press

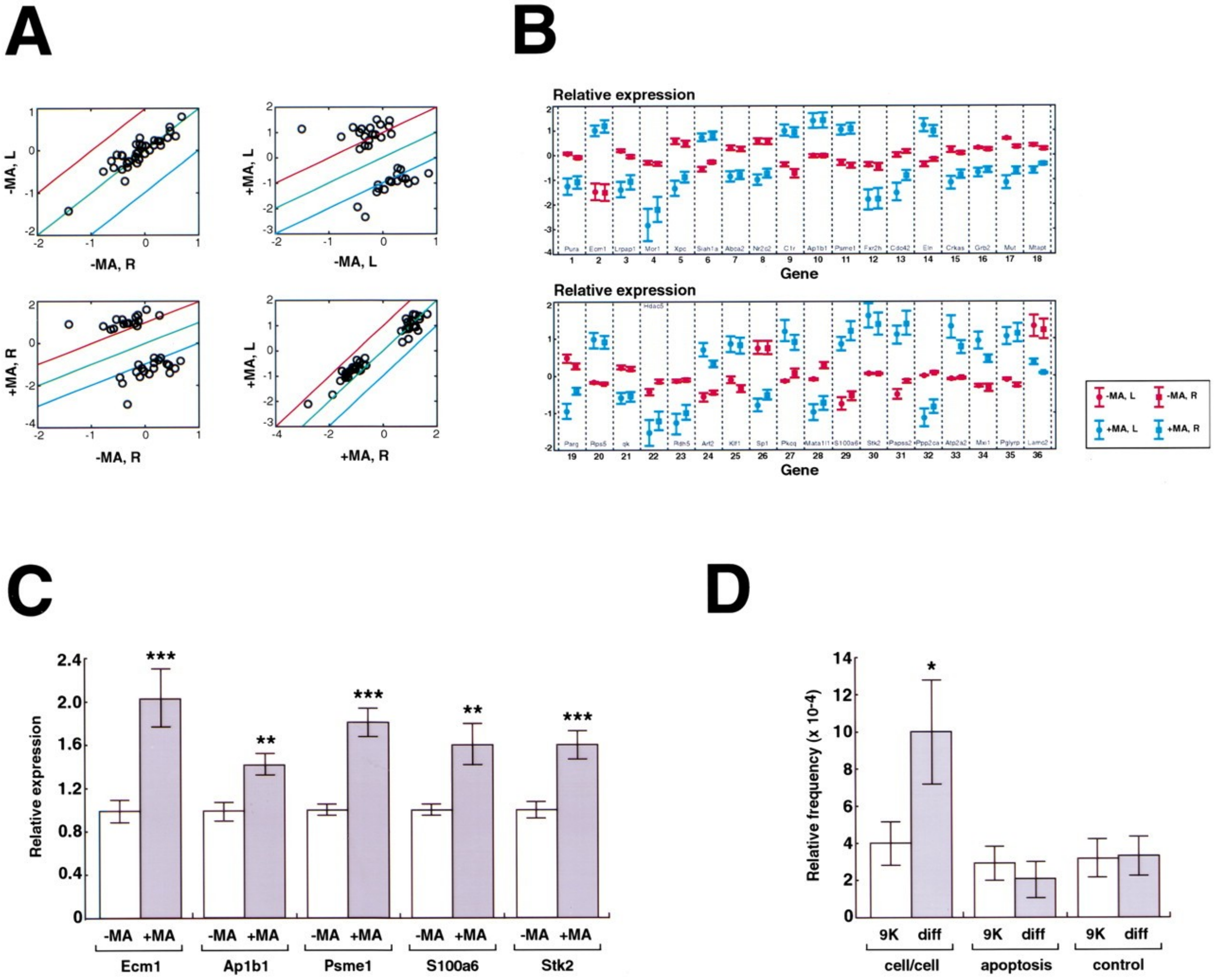
Figure 5 Putative regulatory elements shared between groups of correlated and anticorrelated genes



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Cold Spring Harbor Laboratory Press

Figure 6 Differentially expressed genes



Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Cold Spring Harbor Laboratory Press

Research at UMD

- Possible future work with Amitabh Varshney (CS) and Cristian Castillo-Davis (Biology)