CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 25

Real-life examples

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
- Our microbiome – another organ of our body

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.

Assembly challenges

- **Not all organisms at same level of coverage**. Statistical repeat detection prevents most represented organisms from assembling

EXISTING ASSEMBLERS DESIGNED FOR 1 SINGLE DNA MOLECULE AND UNIFORM COVERAGE

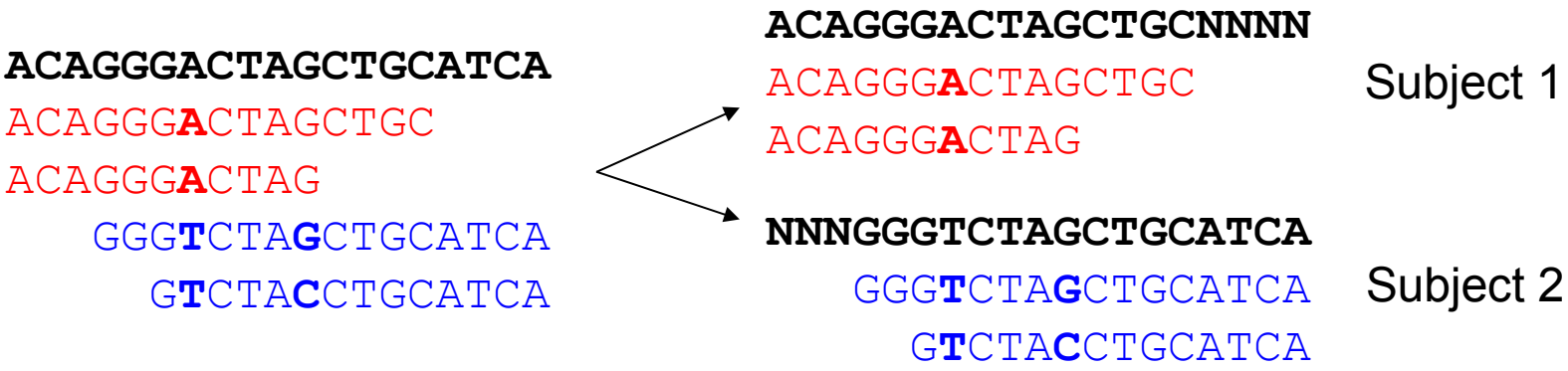
- **rDNAs are likely to co-assemble** leading to chimeric contigs/scaffolds.
- **Low depth of coverage** for the majority of organisms leads to poor assemblies.
- **Rearrangements and mobile element insertions** in closely related organisms further complicate assembly.

Assembly strategy

- Combined assembly of Subjects 1 and 2, followed by separation into specific assemblies.
- Turn down/off statistical repeat detection in Celera Assembler (A-stat cutoff set to -20)
- Comparative assembly (using AMOScmp) of organisms expected to be present within the samples.

	# contigs	# bases	# singletons	# bases	avg. coverage
Subject 1	9,237	15,938,119	28,611	23,680,659	1.75
Subject 2	11,144	20,494,902	25,080	21,011,299	1.91
Total	17,668	33,753,108	53,691	44,691,958	1.99

Co-assembly & SNP rates



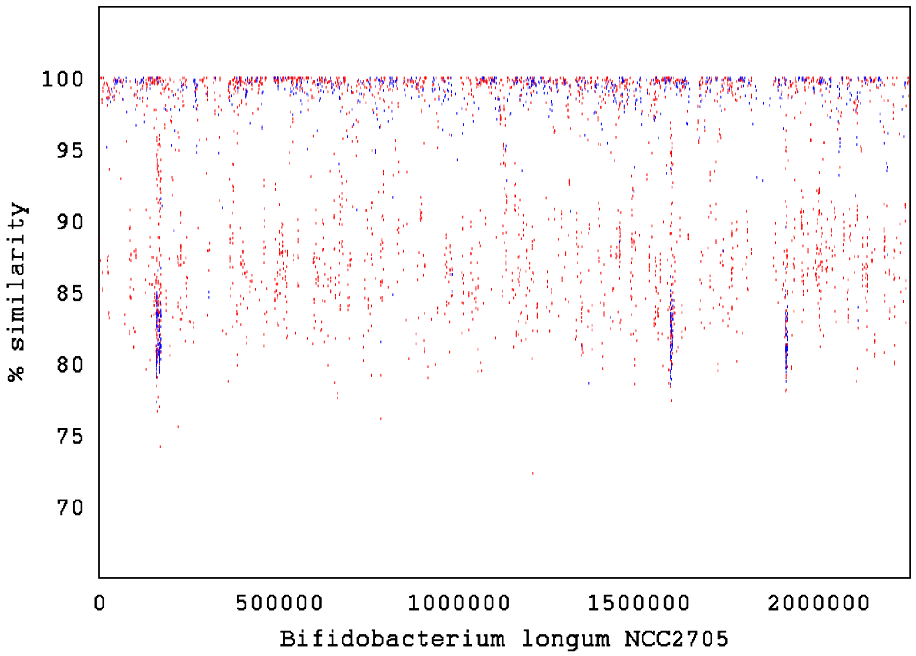
SNP – disagreement with phred q.v. of > 30 on both haplotypes

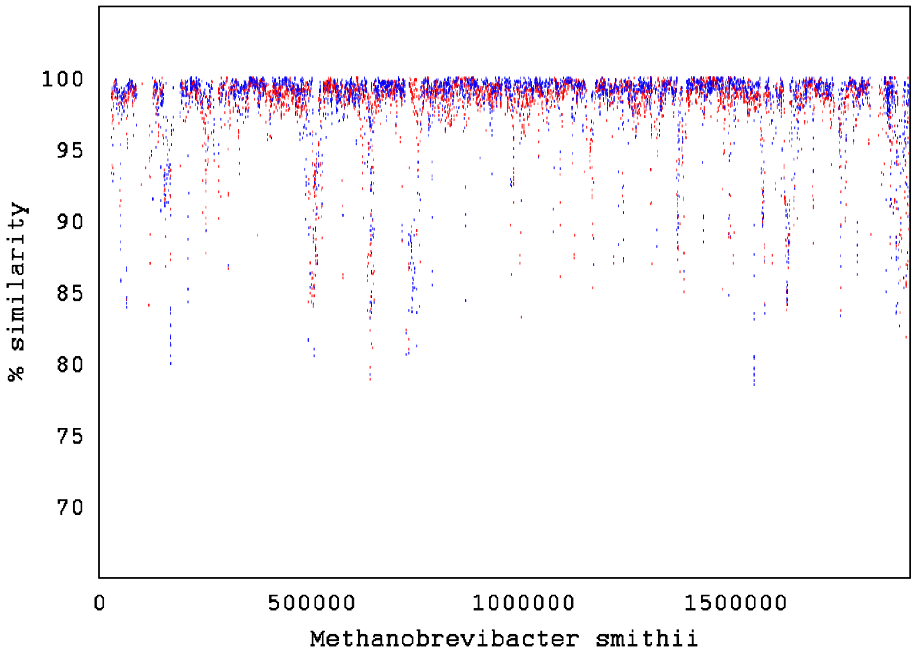
SNPs provide a measure of species diversity:

37,460 inter-subject vs. 34,545 intra-subject

SNPs/kb	Subject 1	Subject 2
Bacteria	1.7	0.6
Archaea	0.19	0.09

Comparative Assembly (AMOScmp)

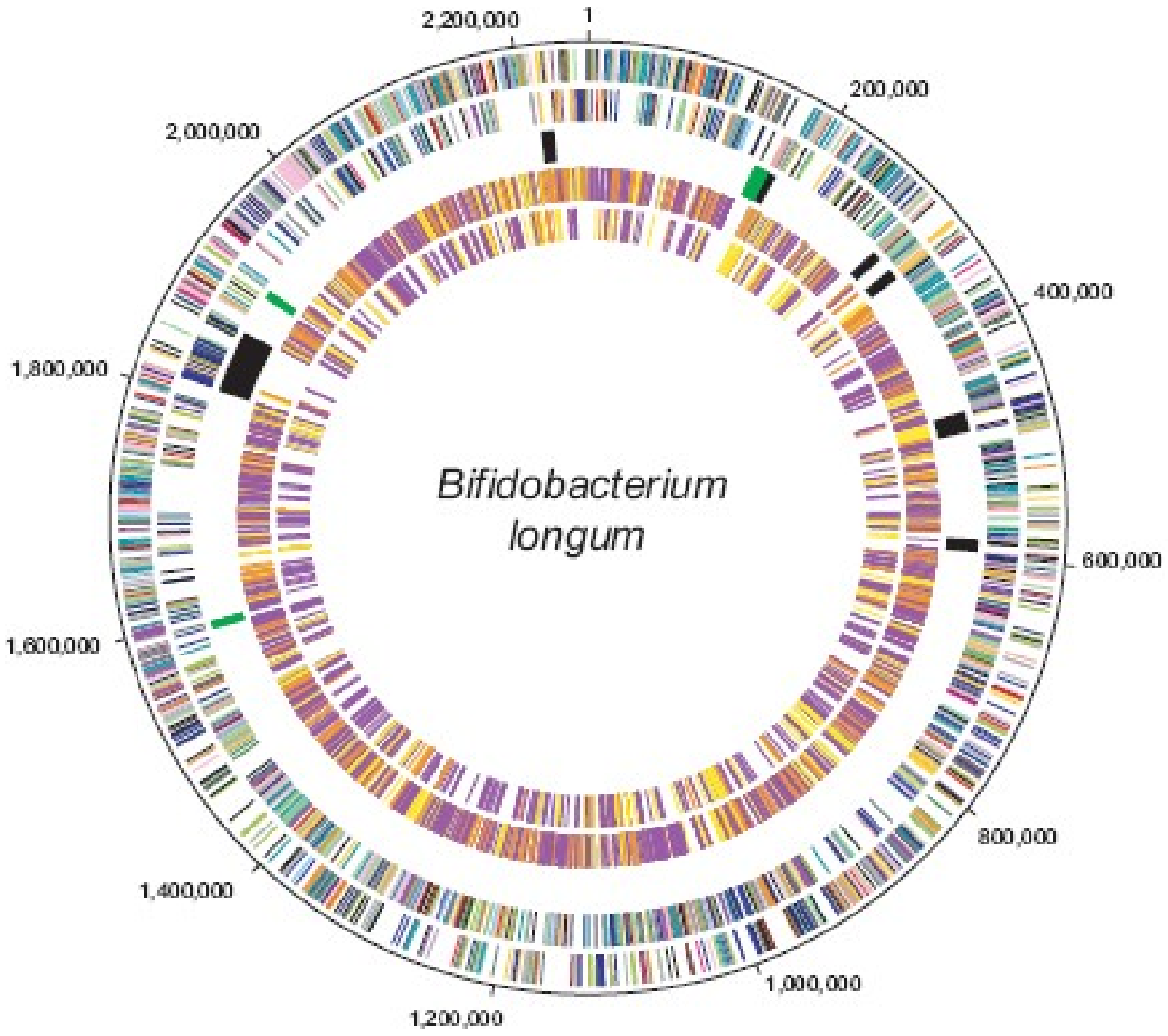




Genome size	2.26 MB	~1.9 MB
Coverage	0.7	3.5
# contigs	789	222
# bases	988,707	1,538,516

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

Mobile elements



Binning results

- Binned data
  - Contigs: 66% (S1), 55% (S2)
  - Singletons: 42% (S1) 40% (S2)
- 76 prokaryotic orders found in shotgun data
- 15 prokaryotic orders found in amplified 16S rRNA (3 sets of universal primers)
- 2 addl. orders found with specific 16S rRNA primers (*Bacteroidales*, *Desulfovibrionales*)

Taxonomy				# rDNA clones		# bases							
				universal primer rDNA		shotgun rDNA				blast best hit assembly data			
				1	2	1	%	2	%	1	%	2	%
BACTERIA	Firmicutes	Bacilli	Bacillales	4	0	1,201	0.81		0.00	2,469,250	11.63	2,867,781	13.34
			Lactobacillales	69	58	853	0.58	930	0.68	1,979,301	9.33	2,466,826	11.48
		Clostridia	Clostridiales	2,777	3,386	70,055	47.38	102,140	74.16	4,396,295	20.71	5,562,074	25.88
			Thermoanaerobacteriales	0	0		0.00		0.00	1,629,710	7.68	2,005,958	9.33
		Mollicutes		341	121		0.00		0.00	109,420	0.52	87,461	0.41
	Actinobacteria	Actinobacteria	Actinomycetales	0	0	332	0.22	793	0.58	820,619	3.87	548,724	2.55
			Bifidobacteriales	30	0	31,443	21.27	5,101	3.70	2,882,567	13.58	851,278	3.96
			Coriobacteriales	4	6	25,781	17.44	10,804	7.84	0	0.00	0	0.00
	Proteobacteria	Alpha proteobacteria		1	0		0.00		0.00	310,514	1.46	323,571	1.51
		Beta proteobacteria		0	0		0.00		0.00	549,867	2.59	497,139	2.31
		Delta proteobacteria		0	0		0.00		0.00	1,628,492	7.67	1,461,558	6.80
		Epsilon proteobacteria	Campylobacterales	0	0		0.00		0.00	126,688	0.60	147,525	0.69
		Gamma proteobacteria		0	0		0.00		0.00	858,159	4.04	873,932	4.07
	Fusobacteria	Fusobacteria	Fusobacterales	0	0		0.00		0.00	363,524	1.71	442,726	2.06
	Bacteroidetes			0	0		0.00		0.00	613,177	2.89	736,839	3.43
	Spirochaetes	Spirochaetes	Spirochaetales	0	0		0.00		0.00	490,790	2.31	467,356	2.17
	Cyanobacteria			0	3		0.00		0.00	209,737	0.99	231,994	1.08
	Chlamydiae /Verrucomicrobia	Chlamydiae	Chlamydiales	0	0		0.00		0.00	40,292	0.19	25,355	0.12
	Deinococci	Deinococci		0	0		0.00		0.00	90,923	0.43	86,268	0.40
	Thermotogae	Thermotogae	Thermotogales	0	0		0.00		0.00	114,817	0.54	155,759	0.72
	Planctomycetes	Planctomycetacia	Planctomycetales	0	0		0.00		0.00	54,210	0.26	61,673	0.29
	Chlorobi	Chlorobia	Chlorobiales	0	0		0.00		0.00	62,323	0.29	59,757	0.28
	Aquificae	Aquificae	Aquificales	0	0		0.00		0.00	40,770	0.19	40,707	0.19
ARCHAEA	Crenarchaeota	Thermoprotei		0	0		0.00		0.00	18,406	0.09	23,008	0.11
	Euryarchaeota	Methanobacteria	Methanobacteriales	0	0	18,188	12.30	17,970	13.05	943,256	4.44	946,329	4.40
		Other		0	0		0.00		0.00	420,096	1.98	523,221	2.43
			Totals:	3,226	3,574	147,853	100.00	137,738	100.00	21,223,203	100.00	21,494,819	100.00

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

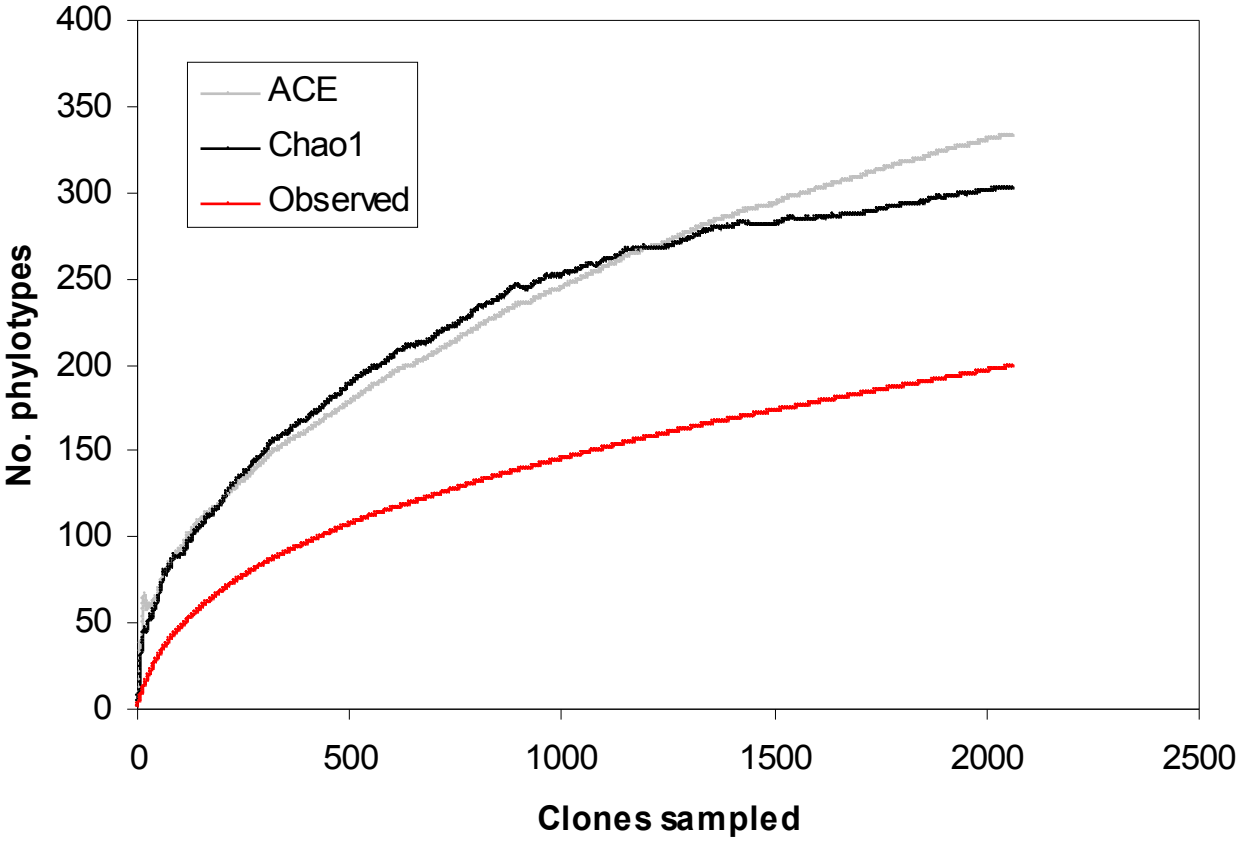
Abundance analysis

$$abundance \sim \frac{\#16S\_rDNA\_reads}{\#16S\_rDNA\_copies}$$

	Order (rDNA copy number)	Subject 1	Subject 2
Actinobacteria	Coriobacteriales (3.5)	12	5.15
	Bifidobacteriales (3.5)	12	0.85
Clostridia	Clostridiales (10)	10.4	16.1
Archaea	Methanobacteriales (2)	7	6.5

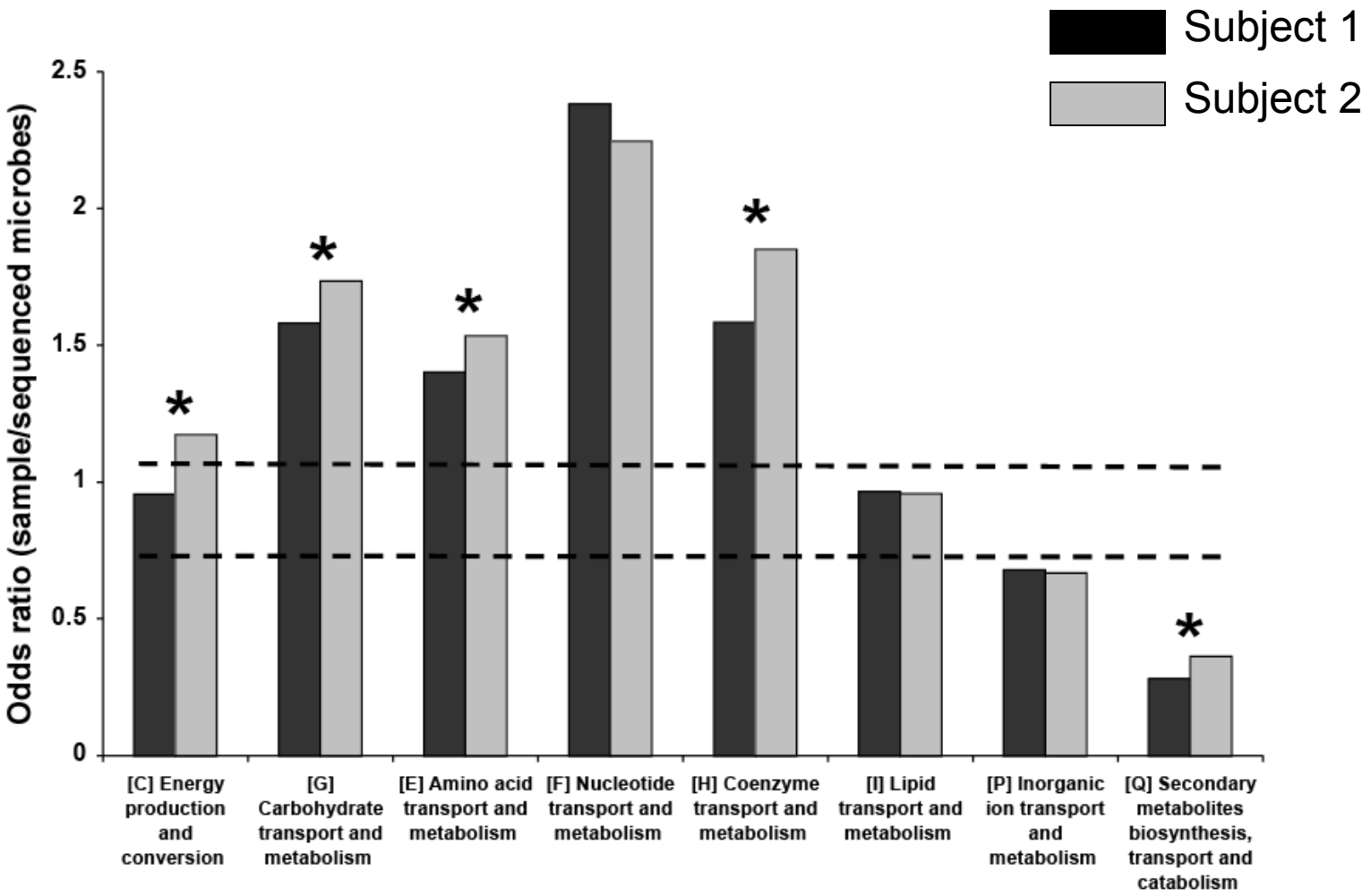
Assuming random libraries. Some assembly required. Past results are not indicative of future performance.

Richness of environment



ACE – 334 phylotypes
Chao1 – 303 phylotypes

Metabolome



Enrichment of cellular processes in GI tract bacteria w.r.t. other sequenced bacteria

What's new?

What have we learned that we didn't know from rDNA studies?

- Unexpected abundance of archaea (cell lysis artifact?)
- Antibiotic resistance genes identified in GIT
  - Tetracycline 19
  - Vancomycin 8
- Mobile elements that influence bacterial diversity
- Strong enrichment of MEP (2-methyl-D-erythritol 4-phosphate) pathway (vs. other bacteria)

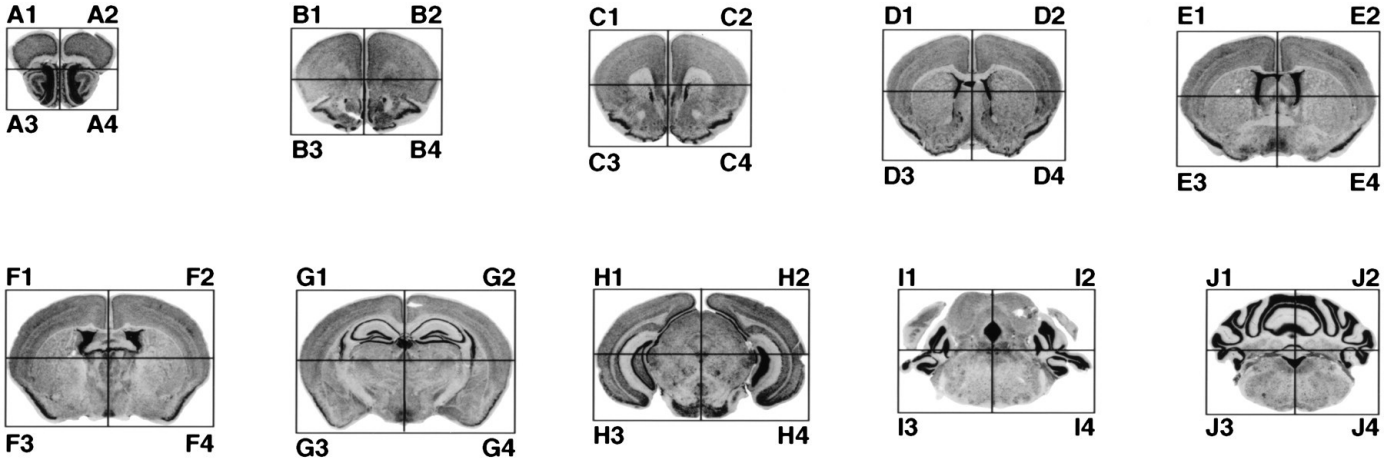
Voxelation

- 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

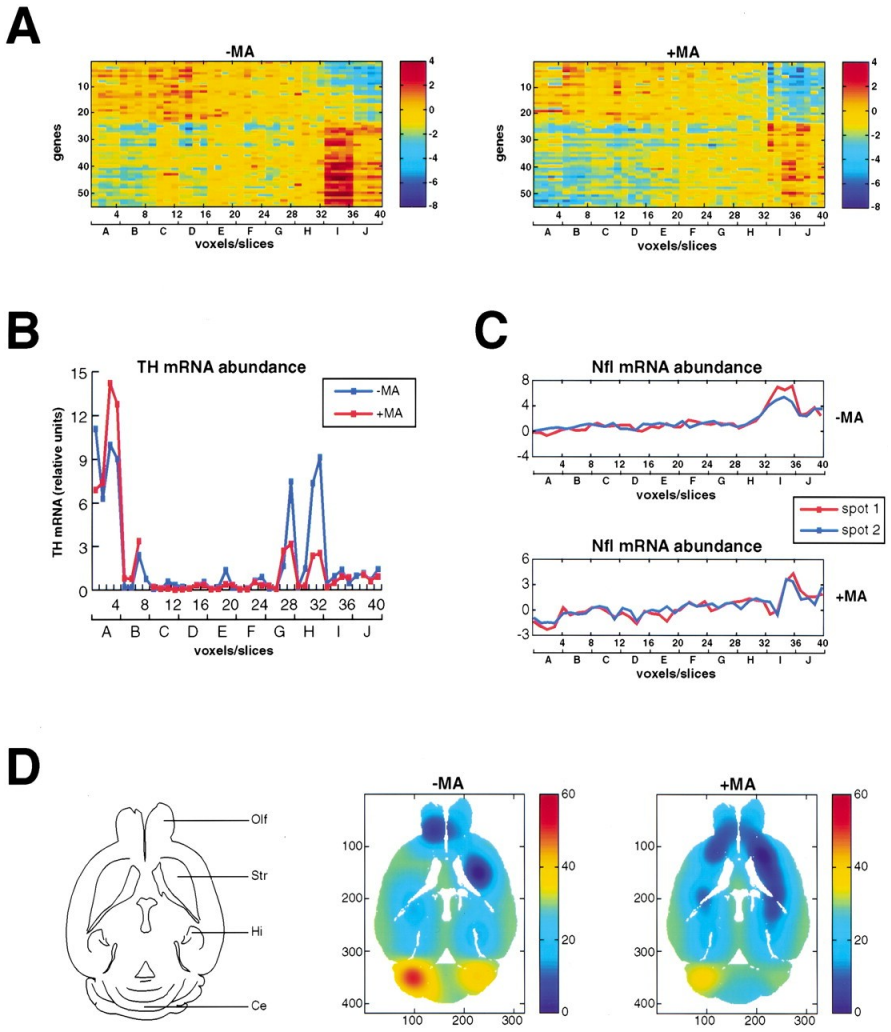
Figure 2 Voxelation scheme



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

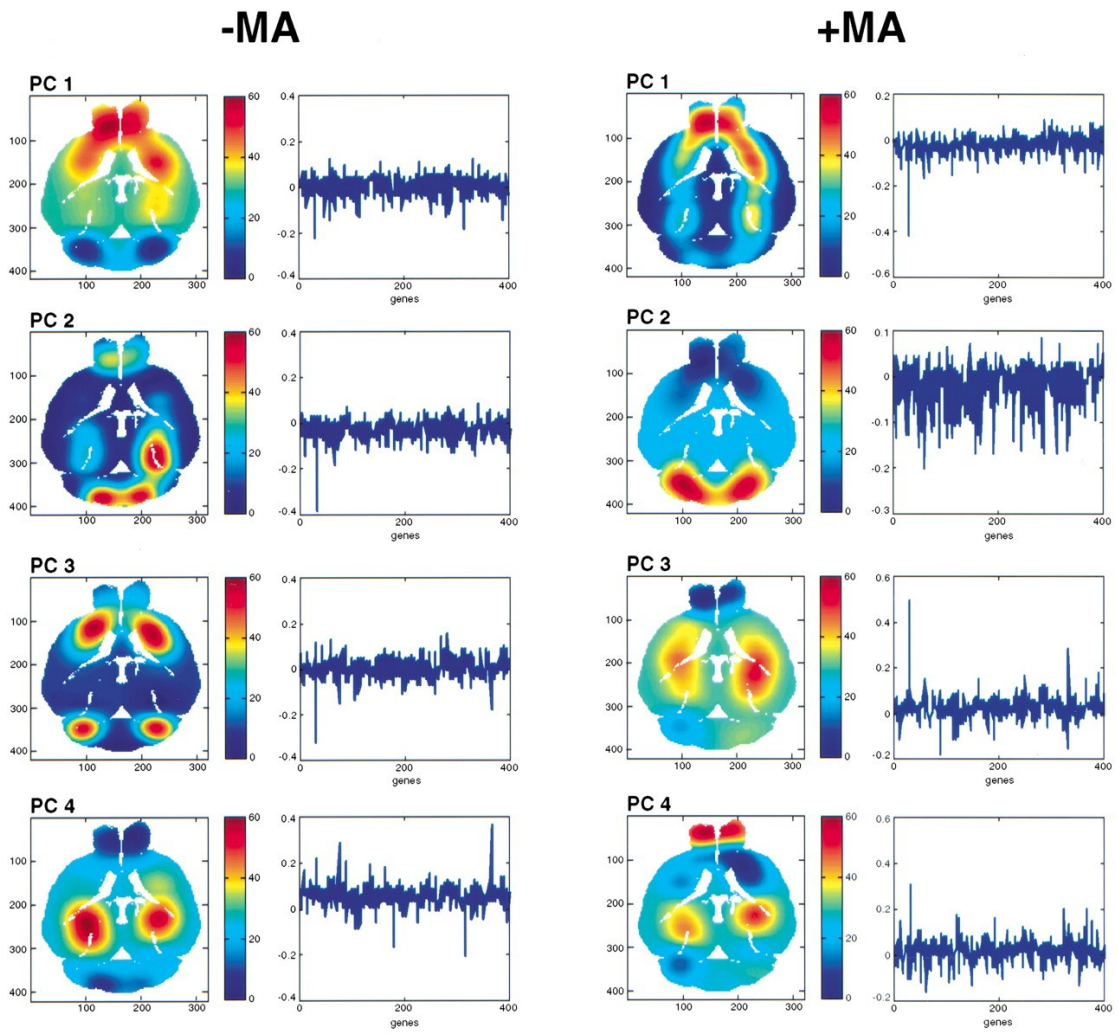
- Mouse brain cut up into voxels
- Run a separate microarray experiment on each voxel

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



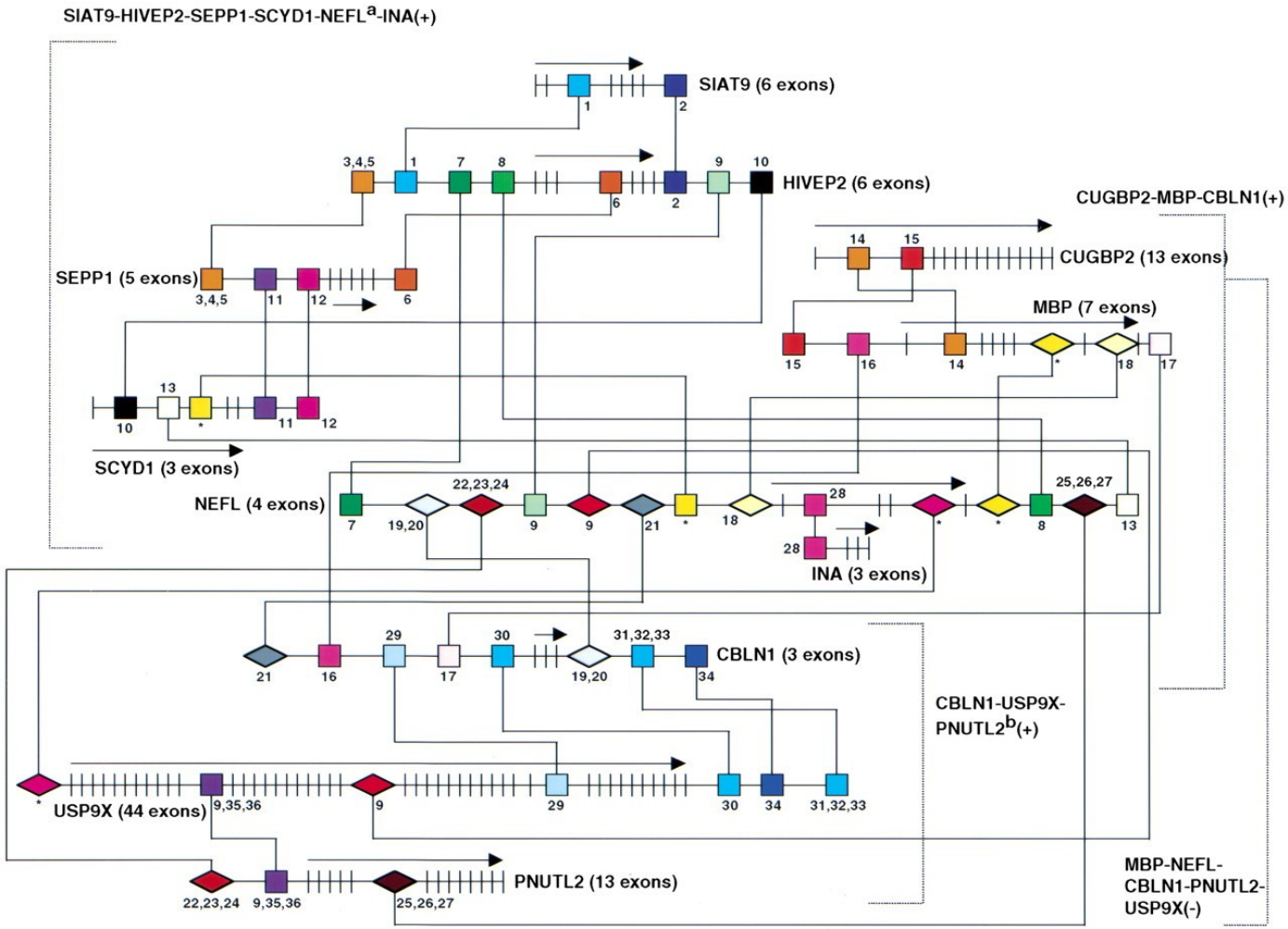
Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

Figure 7 SVD delineates anatomical regions of the brain



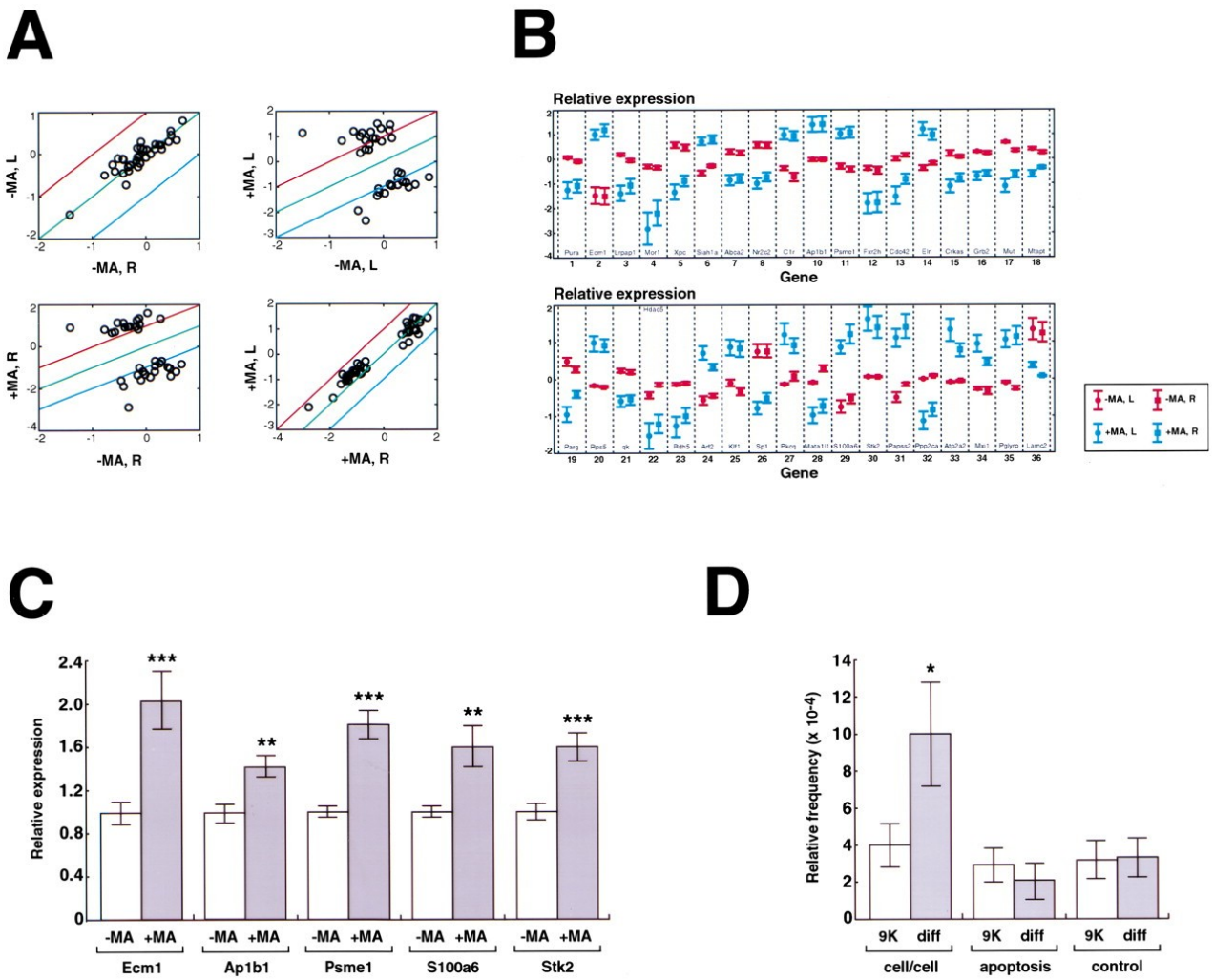
Vanessa M. Brown et al. Genome Res. 2002; 12: 868-884

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



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

Figure 6 Differentially expressed genes



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