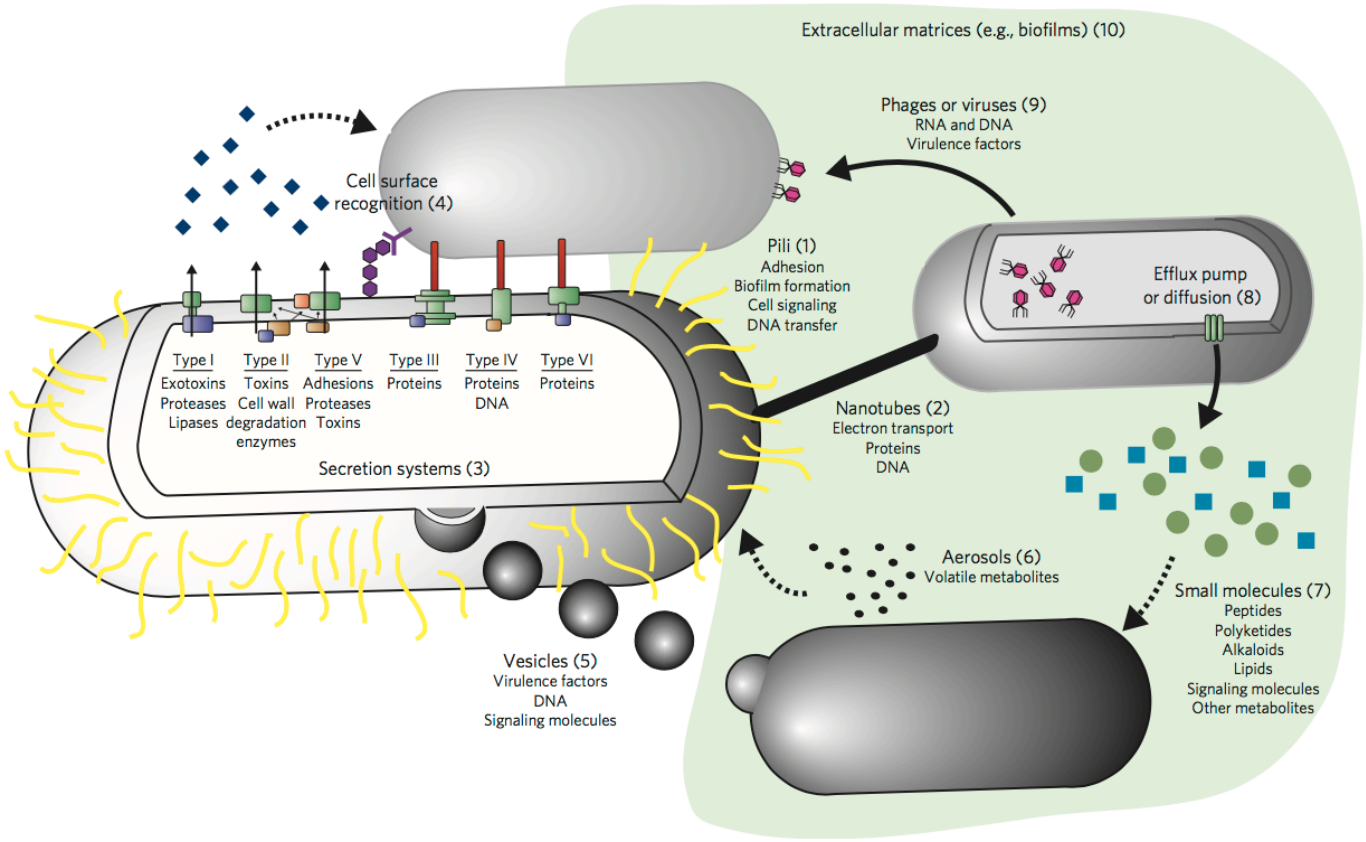
Microbiomes and metabolomes

Michael Inouye
Baker Heart and Diabetes Institute
Univ of Melbourne / Monash Univ

Summer Institute in Statistical Genetics 2017
Integrative Genomics Module
Seattle

@minouye271
www.inouyelab.org

Interactions between microbes and metabolites

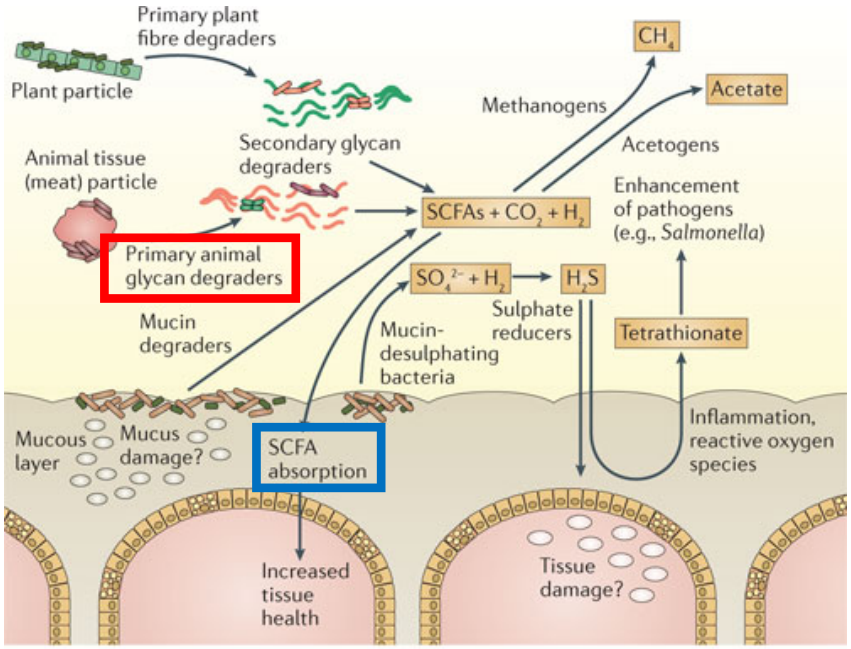


Metabolites are…

- Nutrients
- Signals between cells (microbe-microbe, microbe-host)
- Control of multicellular/community behaviour

Phelan et al, Nat Chem Biol 2011

Human microbiota and metabolism



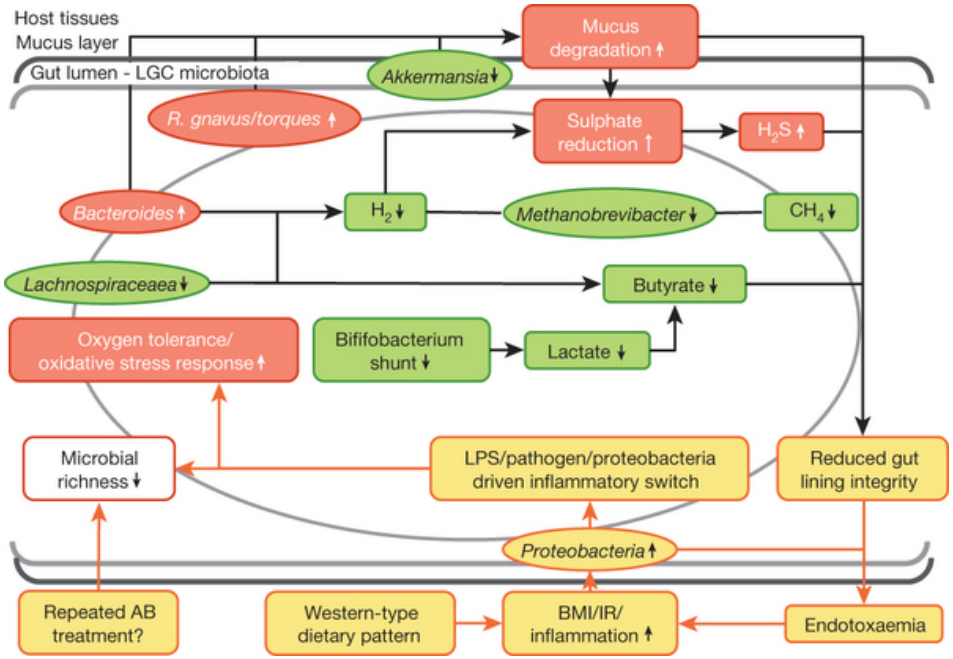
Nature Reviews | Microbiology

Roseburia, Eubacterium, Clostridium, Ruminococcus, Bifidobacterium

SCFA – Short chain fatty acids

Koropatkin et al, 2013

LGC – Low bacterial gene count ('richness')



Top, observed increase (red) or decrease (green) of functions and phylogenetic groups. Bottom, potential drivers (yellow) of inflammation related to decreased richness. Left, antibiotic-mediated perturbation of the richness; Right, proteobacterial lipopolysaccharide-mediated perturbation of the richness. AB, antibiotic; IR, insulin resistance.

Le Chatelier et al, Nature 2013

Background (microbiome)

- Culture
  - Looking only at a few candidate species
  - Need prior hypothesis
  - Akin to candidate gene approach

- Sequencing
  - >99% of microbes cannot be cultured.
  - Hypothesis free/unbiased approach
  - Akin to GWAS

Courtesy of Shu Mei Teo

Background

What bacteria?

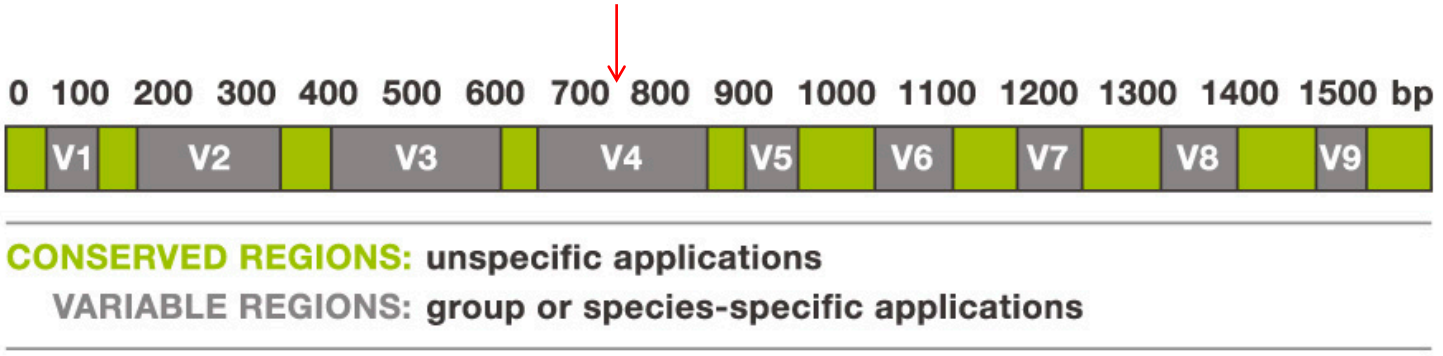
- Sequence all the genes

- Sequence a marker gene
  - Cost efficient
  - Depends on research question
  - 16S rRNA gene

Why 16S rRNA gene?

- Present in all bacteria
- Conserved regions → primer design
- Variable regions
- Large databases

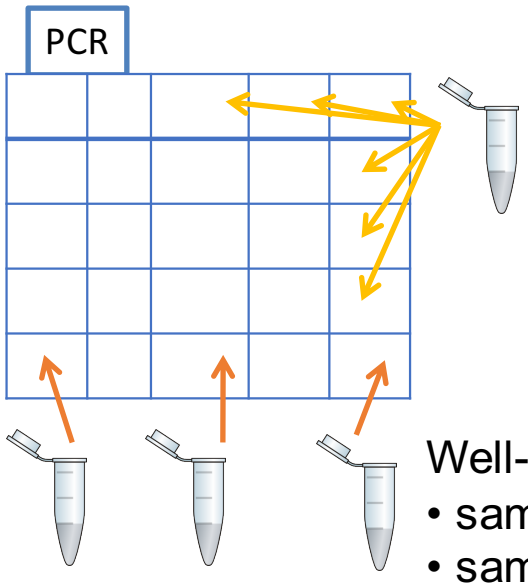
→ Taxonomic assignment



http://www.alimetrics.net/en/index.php/dna-sequence-analysis

Sample preparation for multiplex sequencing

- DNA extraction
  Bead beating step for lysis



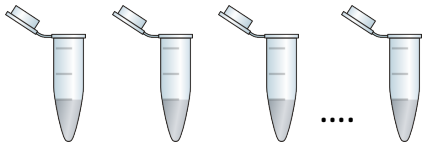
Reaction mix
- forward primer
- PCR mix

Well-specific mix
- sample
- sample specific tagged reverse primer

Sample preparation

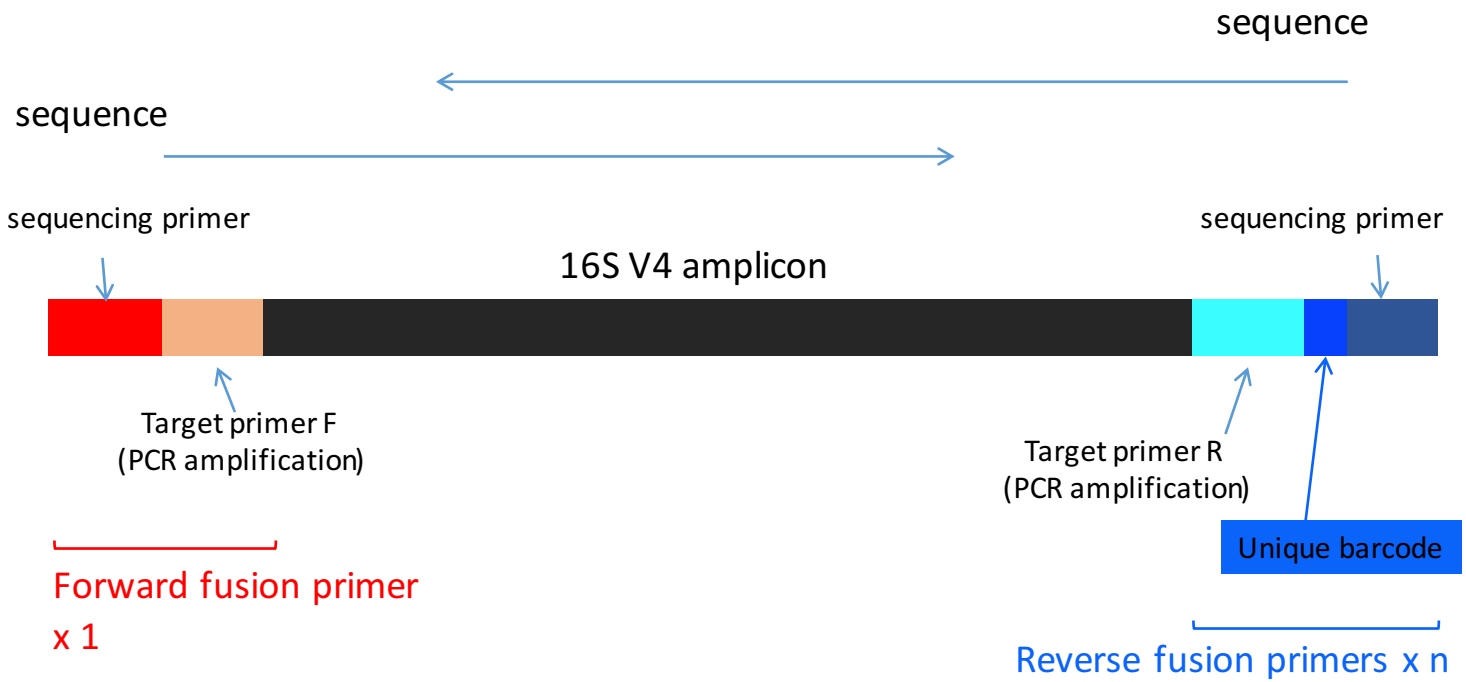
- DNA extraction
- PCR

Purify the sample
Quantify purified sample
Dilute samples so that each sample has same concentration



Pool equal amounts of DNA

Final pool for sequencing


sequence
sequence
sequencing primer
sequencing primer
16S V4 amplicon
Target primer F
(PCR amplification)
Target primer R
(PCR amplification)
Unique barcode
Forward fusion primer
x 1
Reverse fusion primers x n

Sample1.read1.fastq ~200,000 sequences × >1000 samples

@M00267:65:000000000-A5GLT:1:1101:16947:1503 1:N:0:27
TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTACAGCGCGCGCAGGCGGTTTTTTTTGTCTGATGTTACCGCCTCTGGCTTTACCTTTGTACTCCTTCTTATCCTTTTTTCCTTTCTTCCT
TTCGGGTGACTTGTACTTCCTGTT
+
>>>>>FFAFDFF1EFGEEGCF1GGG0EEH1AFF1/AEE///A1///A////////>>/>?>///B>>1121BF221//</!111??F01?11<11<111111?1111111<1111-00=00000000000.-.;-
.00:0;0:;0::00:0
@M00267:65:000000000-A5GLT:1:1101:13913:1503 1:N:0:27
TACGTAGGTCCCGAGCGTTGTCCGGCTTTATTGGGCGTACAGCGTGCTCTGTCGGTTTTTTTTTTCTTCTTTTTTTTGCTTTTTCTTTCCCTTTTTCTTCTTTTTTTTCTTTTTTTCTTTTCTTCCTTTCT
GTTTCCTTTTCTTCCCTTTT
+
>>>>AFFBFFFFGCFGGEGEGFHDGCAEH5D[redacted]2222??11//-011>=<--
/0=000=00000000000000;00:00;00/:0;0
@M00267:65:000000000-A5GLT:1:1101[redacted]
TACGTAGGTCCCGAGCGTTGTCCGGCTTTC[redacted]TTTTCTCTTTTTTCTCTTTTTTTCTTTTCTTCTTTTCT
TTTTTCTTGTTTTTCCTTTT
+
>3>>>C?AFFFFGCEEGEGGGBFAE2AAG5B[redacted]/01222?11<01222?<1/-01>111<111111111100-
/=00=0/<./000;0
@M00267:65:000000000-A5GLT:1:1101[redacted]
TACGTAGGTCCCGAGCGTTGTCCGGATTTA[redacted]TTTTTCTCTTTTTTTCTTTTTTTCTTTTCTTCCTTTC
TCTTTCCTTGTCTTTCCTTTT
+
>>>>ACFBFFFFGFGG?EFEGFHDE?AEH5D[redacted]01111>11<--<00;;0--
:0;0009000000000000000;00;00;09090;0
@M00267:65:000000000-A5GLT:1:1101:13603:1530 1:N:0:27
TACGTAGGTCCCGAGCGTTGTCCGGCTTTATTGTGCGTAAAGCGATCTCTTTCGTTTTTTTTTTTCTTCTTTTTTCTTCTTTTTCTTTACCCTTTTTTTTTTTTTTTTCTTTTTTTCTTTTGTTCTCTTCTT
TTTTCTTTTCTTTCCTTTT
+
>>>>>FFAFFBFEGCEAEEEFDGG22AEHDFG55552B21155111115555511BA11>>///>3B4444BF3/344444443B?B44?33?0111/////>--<--/00;<0--/0;900;/00000000090-
/900;00;00000;0
@M00267:65:000000000-A5GLT:1:1101:13615:1542 1:N:0:27
TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGTCTGTTTGTTTTTTCTTCTGTTTACTGCTTTTTCTTTACCCTTTTTTTCTTTTTTATCTTTTTTTCTTTTCTTCTTTT
GTCGTTAGTTTTCTTCCCTTTT
+

Merge reads (Software: flash)

Cleaning the data

Roche 454/ Ion Torrent
~ 400 - 1000bp
single end
1 run > 160K reads

Illumina Miseq
2 X (150 – 250)bp
paired end
1 run > 16M reads

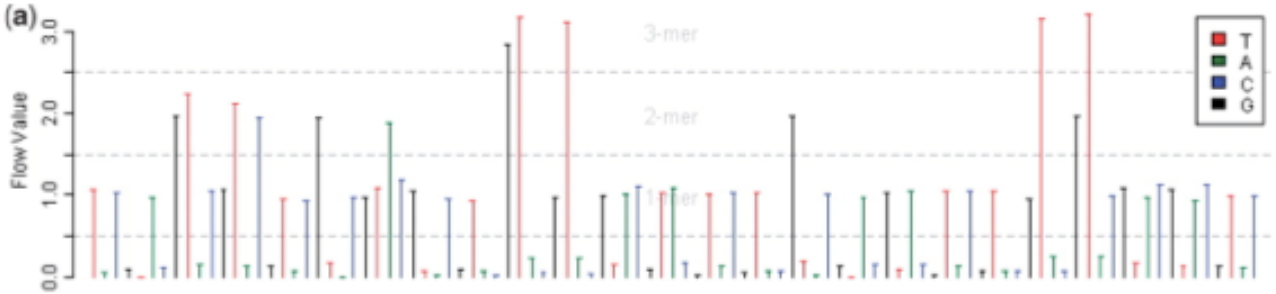
Cleaning the data

Roche 454/ Ion Torrent
~ 400 - 1000bp
single end
1 run > 160K reads

Illumina Miseq
2 X (150 – 250)bp
paired end
1 run > 16M reads

Acacia

Homopolymer error correction



Balzer S et al. Bioinformatics 2010;26:i420-i425

Cleaning the data

Roche 454/ Ion Torrent
~ 400 - 1000bp
single end
1 run > 160K reads

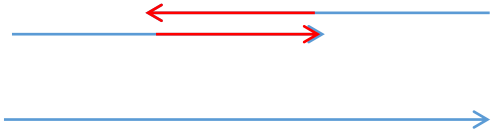
Homopolymer error correction

Illumina Miseq
2 X (150 – 250)bp
paired end
1 run > 16M reads

Merge reads

Flash

Stitch/merge the reads together



Cleaning the data

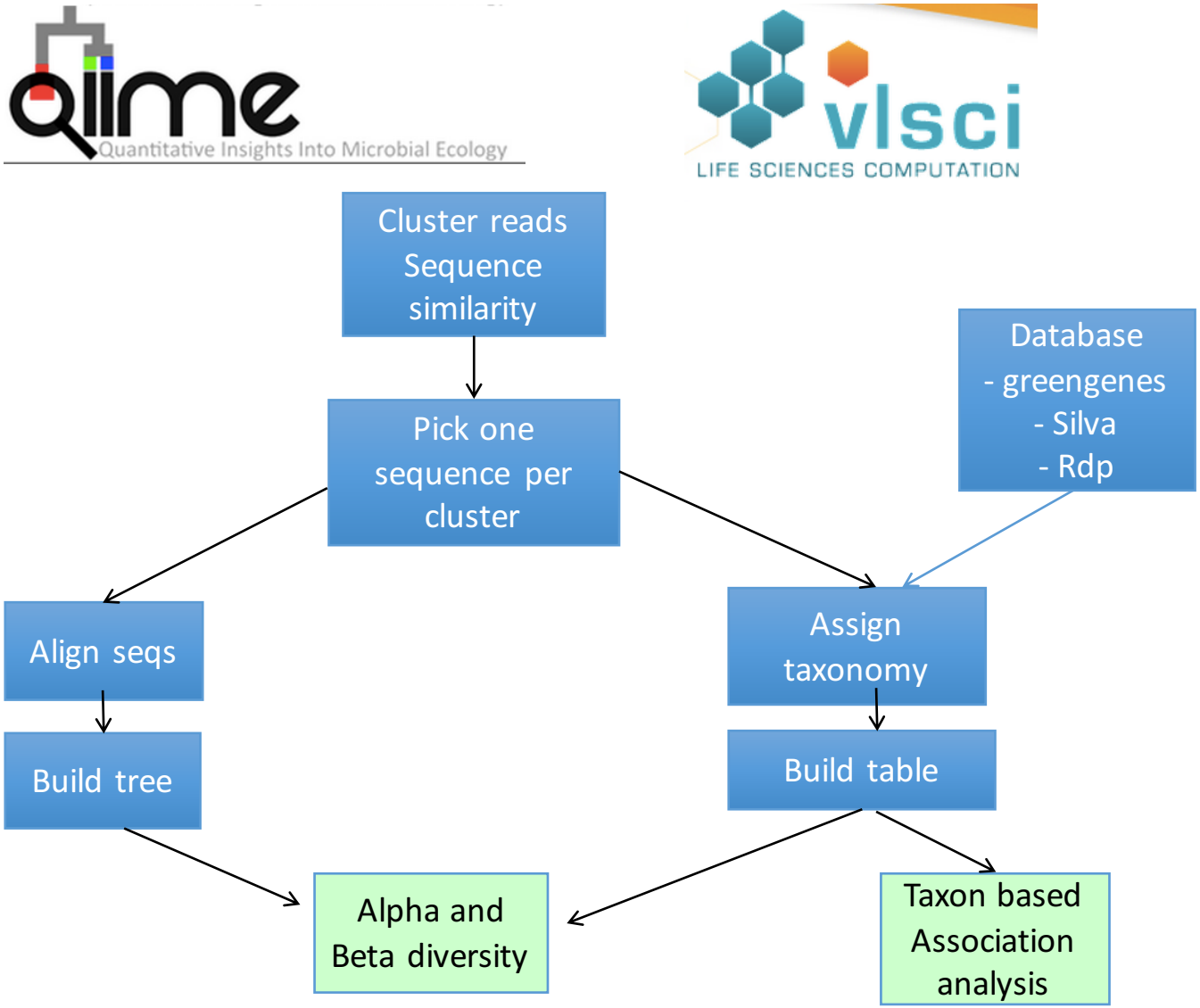
Roche 454/ Ion Torrent
~ 400 - 1000bp
single end
1 run > 160K reads

Illumina Miseq
2 X (150 – 250)bp
paired end
1 run > 16M reads

Homopolymer error correction

Merge reads

Remove Chimeras


qiime
Quantitative Insights Into Microbial Ecology
vlsci
LIFE SCIENCES COMPUTATION
Cluster reads Sequence similarity
Pick one sequence per cluster
Database - greengenes - Silva - Rdp
Align seqs
Build tree
Assign taxonomy
Build table
Alpha and Beta diversity
Taxon based Association analysis

Operational Taxonomic Units (OTUs)

- A cluster of highly similar sequences is termed an OTU.
- Typically.. Cluster all the sequences at a predefined similarity threshold (97%)
  - Within cluster(OTU), sequences are >=97% similar.
  - Between OTUs, the sequences are <97% similar.

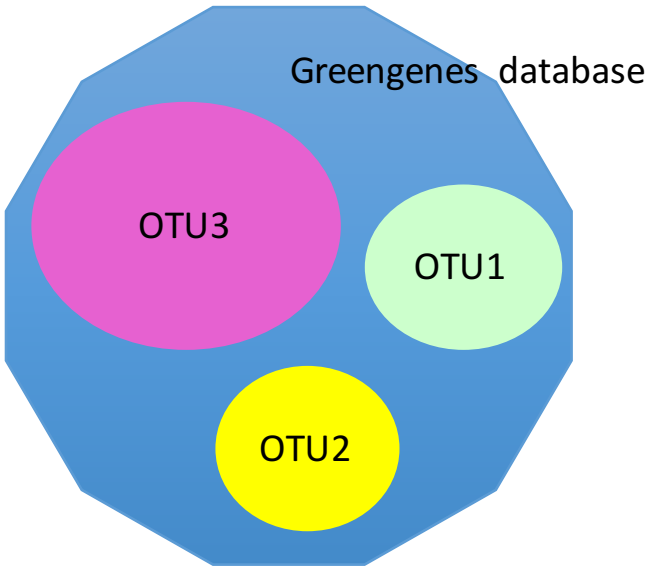
- Computationally demanding for large datasets (>200 million)

Closed reference OTU picking

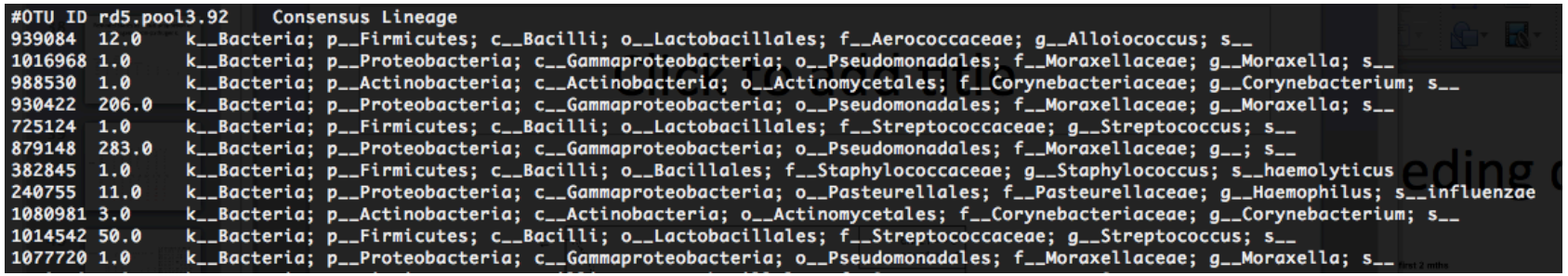
- Greengenes 99% OTUs

- Highly parallelizable
- Sequences that do not match reference are thrown

Compare each sequence →



Example output – 1 sample

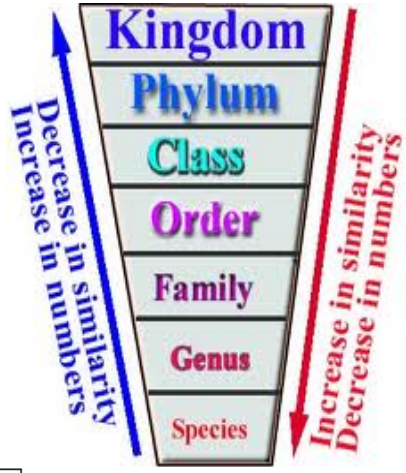


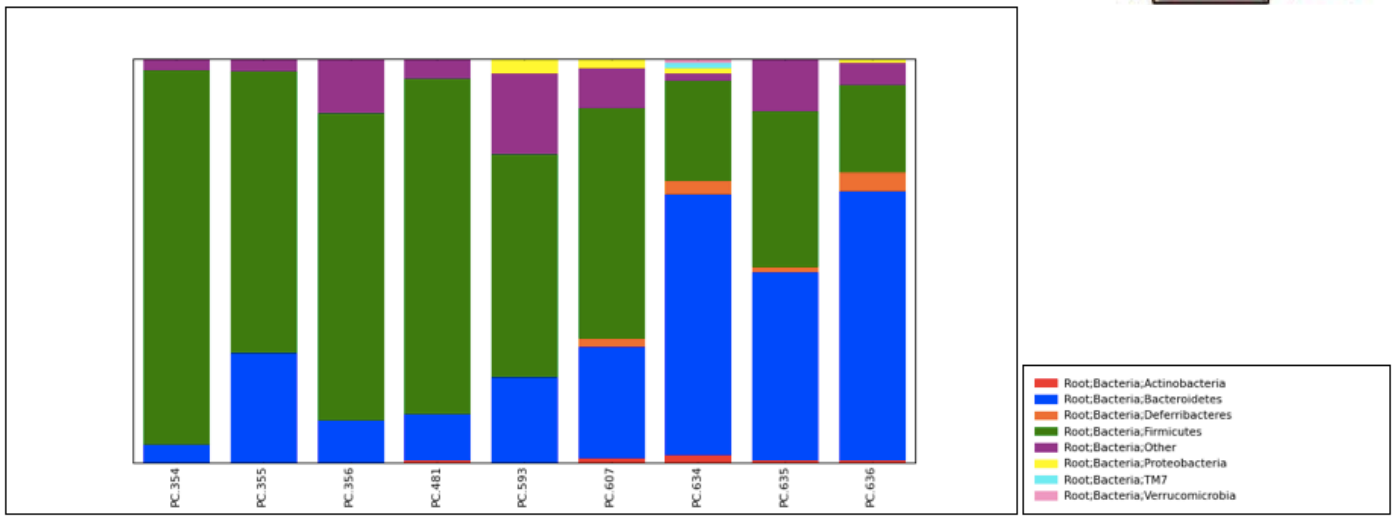
#OTU ID	rd5.pool3.92	Consensus Lineage
939084	12.0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Aerococcaceae; g__Alloiococcus; s__
1016968	1.0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Moraxella; s__
988530	1.0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Corynebacteriaceae; g__Corynebacterium; s__
930422	206.0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Moraxella; s__
725124	1.0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
879148	283.0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__; s__
382845	1.0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Staphylococcaceae; g__Staphylococcus; s__haemolyticus
240755	11.0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s__influenzae
1080981	3.0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Corynebacteriaceae; g__Corynebacterium; s__
1014542	50.0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
1077720	1.0	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Moraxella; s__

- Note different OTUs can have the same taxonomy
- Can also summarize in terms of taxonomy, example at the genus level

Summarized table

Phylum	Sample1	Sample2	Sample3
Acidobacteria	0	22	0
Firmicutes	291777	728	21

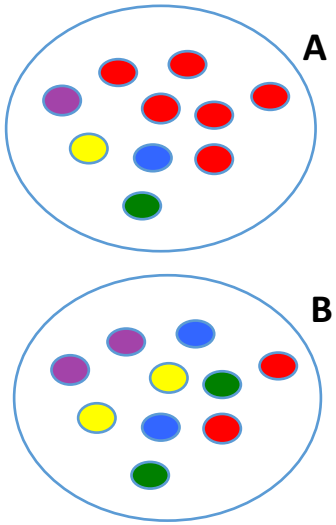




Alpha diversity

- Which sample has the most compositionally diverse microbiome?

- Rarefaction – subsample equal number of reads
- Qualitative measure – absence/presence
- Quantitative measure – considers relative abundance



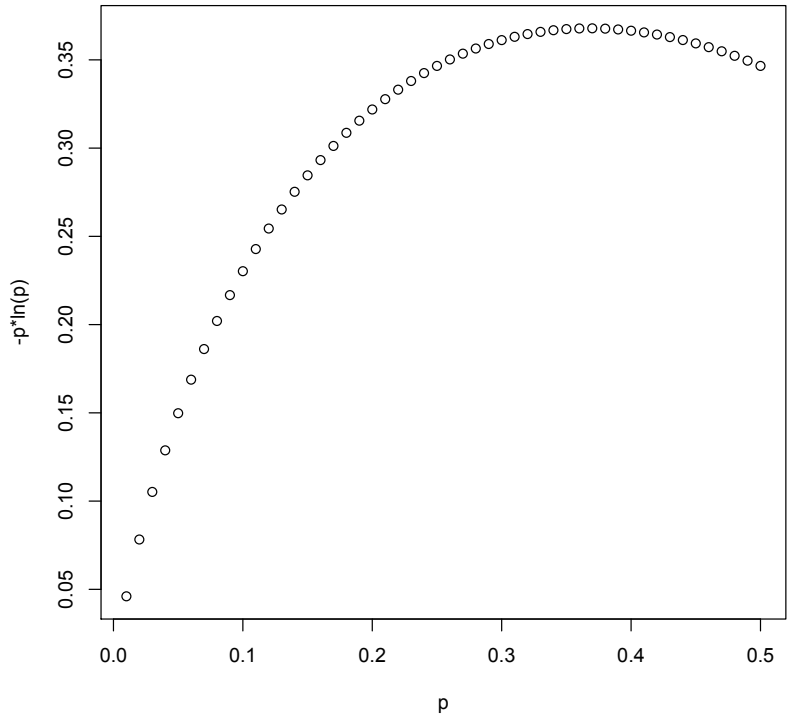
Qualitative – A and B equally diverse
Quantitative – B is more diverse

Shannon's diversity index

N total number of species/OTUs in the community (richness)

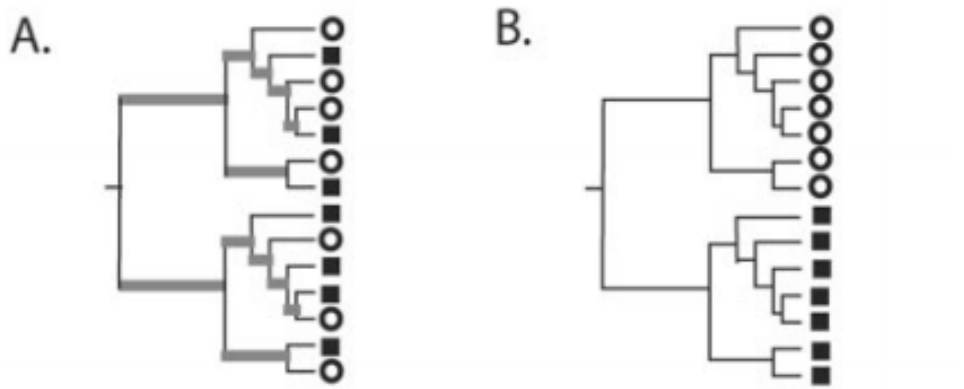
P_i proportion of species *i* relative to N

$$H = -\sum_{j=1}^{N} P_i ln P_i$$



Beta diversity

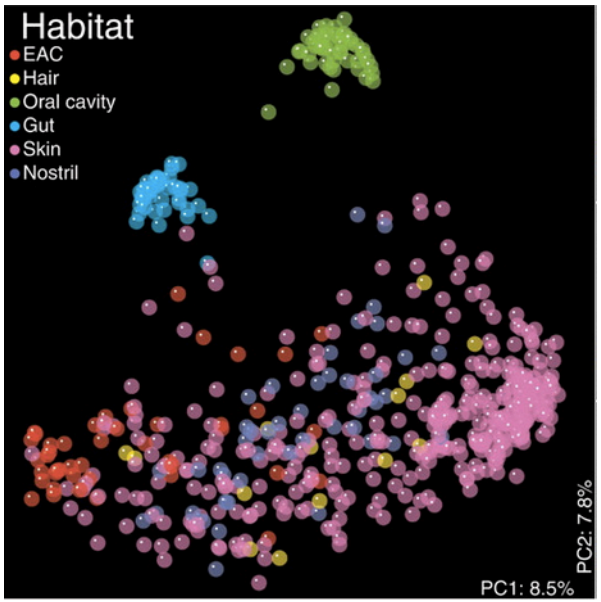
- "Between sample" diversity – how similar/dissimilar are two samples
- Unifrac distance = fraction of the total branch lengths that is unique to one community



Lozupone and Knight 2005

Beta diversity

- Principal coordinates analysis of Unifrac distance matrix
- Which factors correlate with differences in microbiota composition?

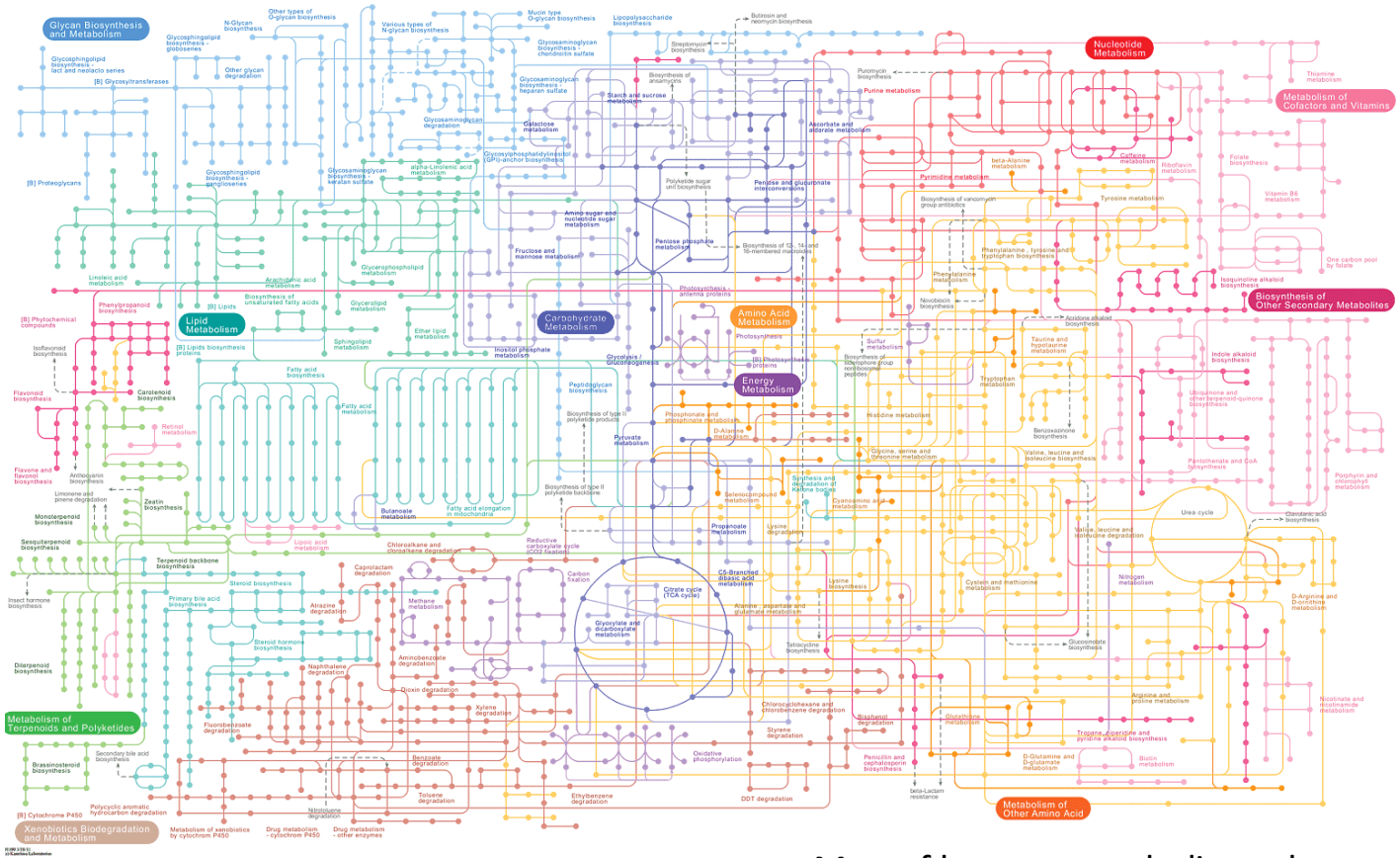


Costello et al. (2009), Science 326:1694

Concepts recap

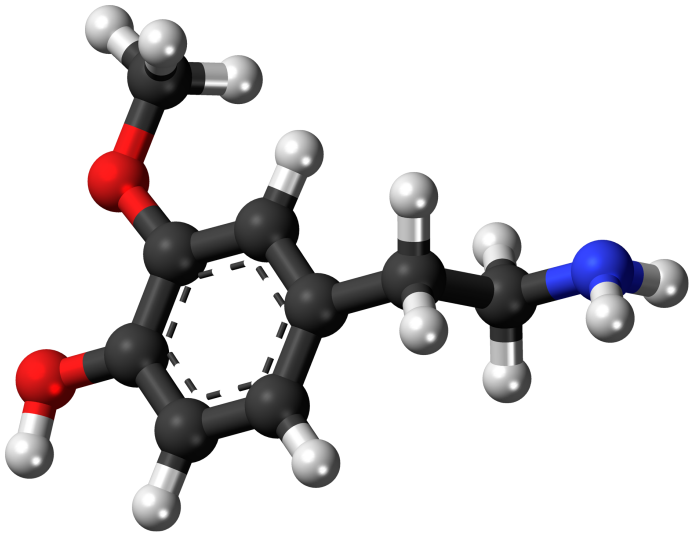
- What kind of animals are there? (Taxonomy)
- How many kinds of animals can I find where I hunt? (Alpha diversity)
- How different is one place from another? (Beta diversity)

Metabolism

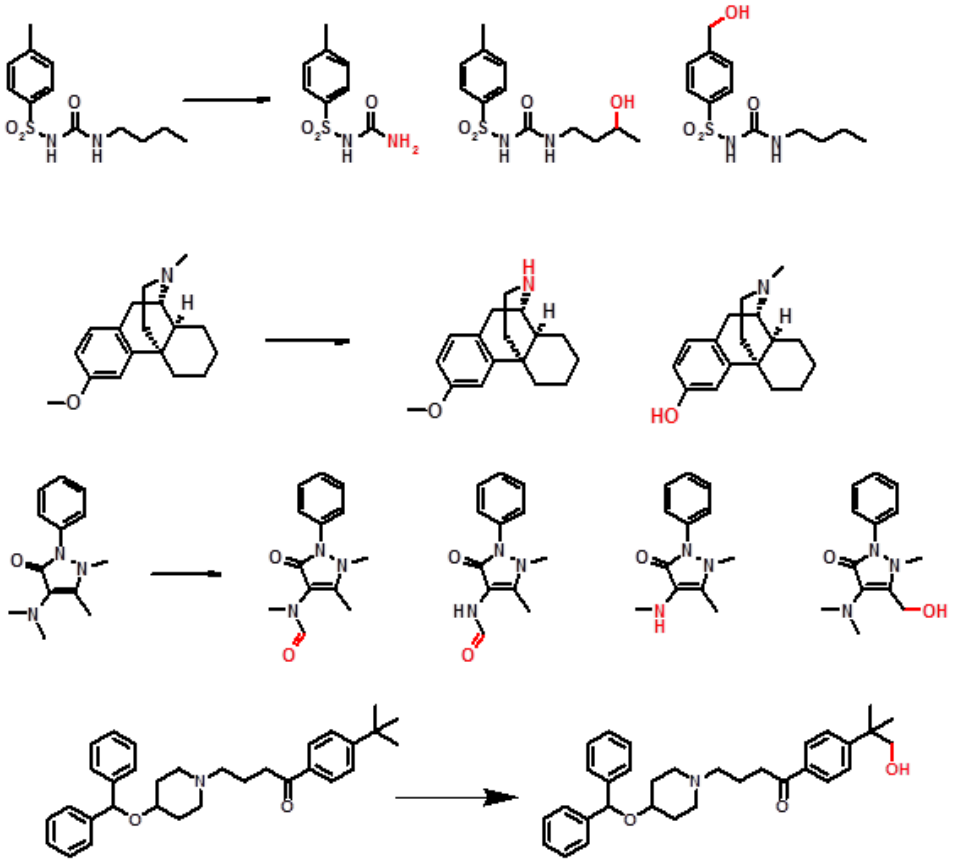


Map of human metabolic pathways (it's complicated)

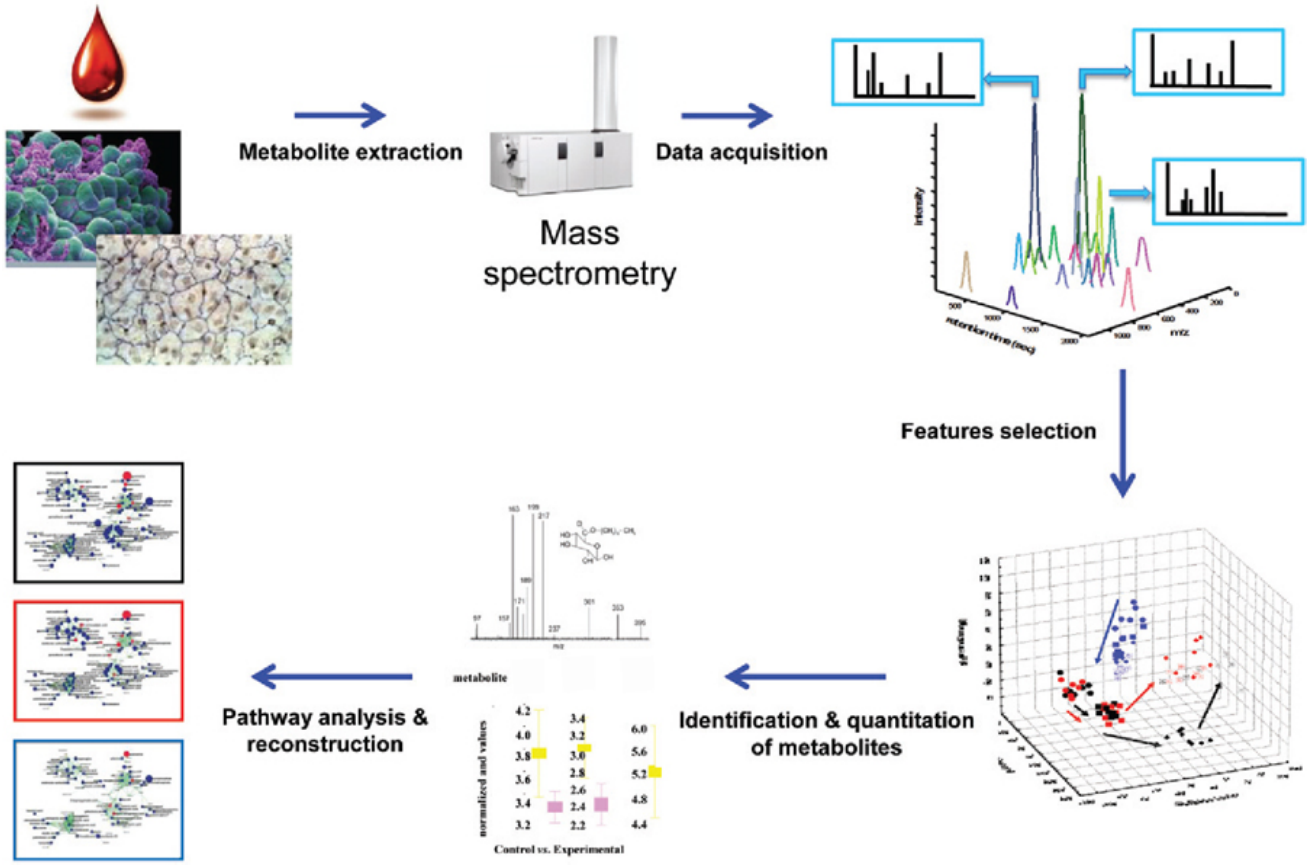
Metabolites



3-methoxytyramine



Metabolomics



Metabolomics data

- Right skewed
- All the usual technical effects

Metabolomics normalisation

- No internal standard(s)
  - Divide by the total sum of metabolite abundances in each sample
  - Divide by the median
- For each metabolite, subtract (log) abundance of internal standard
- Multiple internal standards
  - Selection of internal standard (across all metabolites or per metabolite)
  - Model variation in internal standards and remove it from each metabolite

metabolite_j standards error

mean_j coeffs

$$y_{ij} = \mu_j + r_i\delta_j + e_{ij}$$

$$\tilde{y}_{ij} = y_{ij} - r_i\hat{\delta}_j$$

Subtract 'unwanted' variation from original metabolite levels

De Livera et al, Ana Chem 2012

Metabolomics normalisation

- Usual approaches result in one normalised value for each metabolite, for each sample ('gobal' normalisation)

- These values can then be used in downstream analysis
  - PCA / clustering
  - Association analysis (the standard approaches apply)
  - Classification

- What if we combined normalisation and association analysis?
  - Thus minimising chance that (biological) variation of interest is not removed

De Livera et al, Ana Chem 2012

Modeling and removing 'unwanted' variation

- The 2-step Remove Unwanted Variation approach (RUV2)

- Use of non-changing metabolites as internal standards
  - These metabolite should be present in the sample, exposed to unwanted variation, and not associated with the (biological) factors of interest

(nonchanging) metabolite_j

(unwanted) variance component

coeffs

error

$$y_{ij} = w_i \tilde{\alpha}_j + \tilde{\varepsilon}_{ij} \qquad j = 1, \ldots, n_c$$

Number of non-changing metabolites.
Which to choose?

De Livera et al, Ana Chem 2012

Modeling and removing 'unwanted' variation

- Selecting the number of unwanted factors (eg non-changing metabolites)
  - Metabolites known *a priori* not to change
  - Metabolites derived from QC samples (e.g. replicates)
  - Spike-in metabolites
  - PCA

De Livera et al, Ana Chem 2012

Modeling and removing 'unwanted' variation

(unchanging) metabolite_j

(unwanted) variance component

coeffs

error

$$y_{ij} = w_i \tilde{\alpha}_j + \tilde{\varepsilon}_{ij} \qquad j = 1, \ldots, n_c$$

Factors of interest!

coeffs

(unwanted) variance component

coeffs

error

(changing) metabolite_j

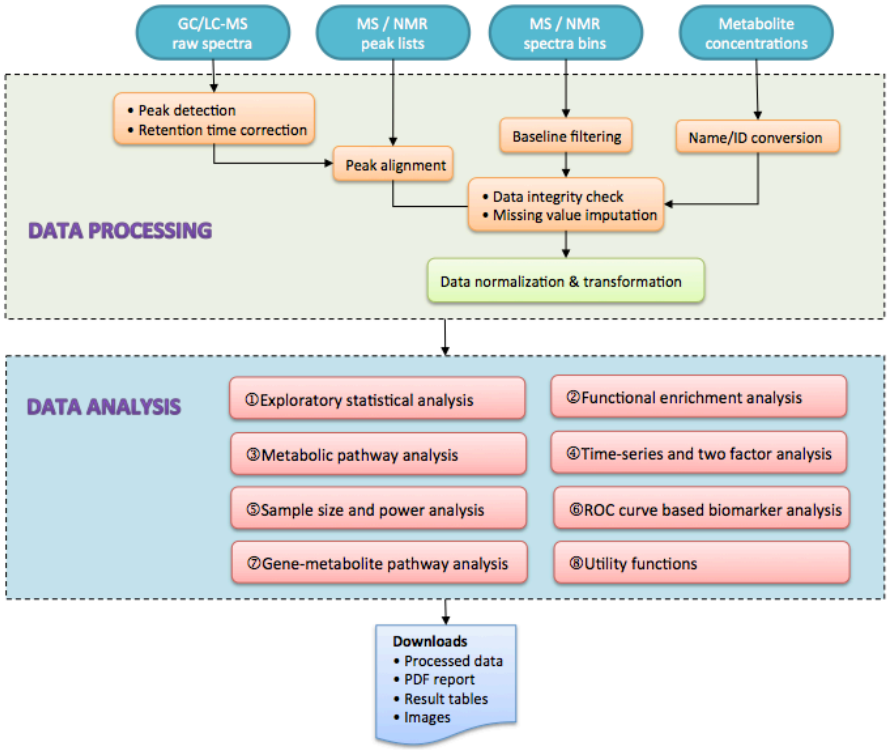
$$y_{ij} = x_i \beta_j + \hat{w}_i \alpha_j + \varepsilon_{ij}$$

Can also adjust for additional covariates

De Livera et al, Ana Chem 2012

Metabolomics pathway analysis

- Similar to gene set-based enrichment approaches
- MetaPA (now integrated into MetaboAnalyst)



BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 18 2010, pages 2342–2344
doi:10.1093/bioinformatics/btq418

Systems biology Advance Access publication July 13, 2010

MetPA: a web-based metabolomics tool for pathway analysis and visualization

Jianguo Xia[1] and David S. Wishart[1,2,3,*]

- KEGG database
  - Fisher exact test
  - Hypergeometric
  - GSEA
  - Network centrality approaches

Published online 20 April 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W251–W257***
doi: 10.1093/nar/gkv380

MetaboAnalyst 3.0—making metabolomics more meaningful

Jianguo Xia[1,2,*], Igor V. Sinelnikov[3], Beomsoo Han[3] and David S. Wishart[3,4,5]

Metabolic 'potential' of microbial communities

- Infer what metabolites (and levels thereof) are associated with sequences from any microbial community

- HUMAnN (Abubucker et al, PLOS Comp Bio 2012)
  - Input: Metagenomic sequences
  - Output: Estimates of gene and pathway abundances

HUMAnN

- Input QC'ed (non-human) metagenomic sequences
- Blast against a protein sequence database (e.g. KEGG)
- Estimate gene family abundances, normalise by gene family sequence length
- Assign genes to pathways (e.g. using MinPath)
- Use inferred microbial taxa to normalise for gene copy number and remove unlikely pathways
- 'Fill in' abundant pathways which may be missing a few genes
- Assign each pathway scores for presence/absence and for abundance