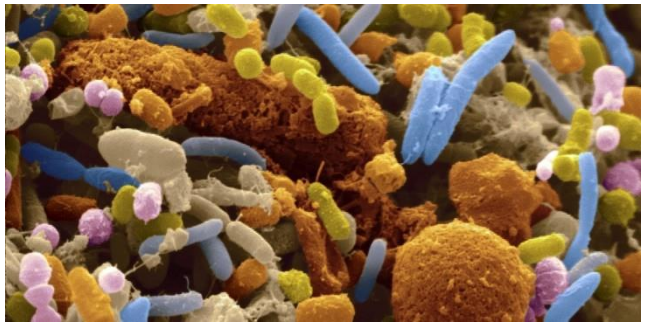
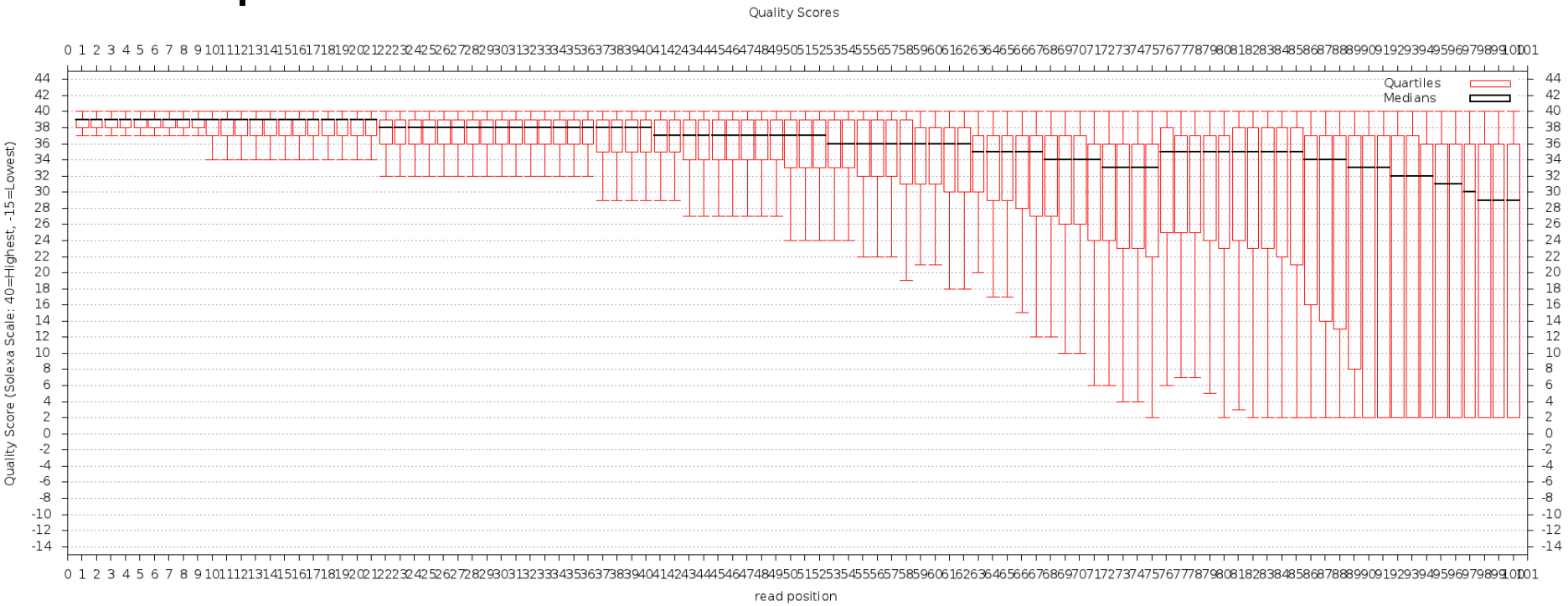
Metagenomics and RNA-seq

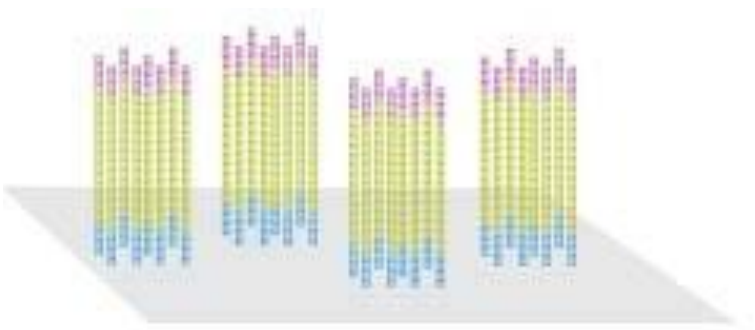
Tobias Österlund




```
GCAACAGTTTGGCGGTAATTCAATTGT
CAGTTTACGGATTCCTTGATTGGATAA
TCCAGTCTGCCCCCAGGCTGCAGTTGC
AAAAGAAAGAAACGACTATGAATAAAC
GACTTCGGATCATTGGACTGTTTGCTG
TGTTCTTTGGCCAGATGATCCACGCGC
AGACCACAGCGTTCACTTATCAGGGGC
GTCTCAATGACAACGGCGCGCTGGCCA
ACGGCATTTATGATTTGAAATTTTCAC
TATACACCGTGGCGACCAATGGCAGTG
CCTCATCGTCGCGGTCAAATGCCGCCA
CCGTCGTCAG
```

Remark on quality of Illumina sequences





Today's lecture

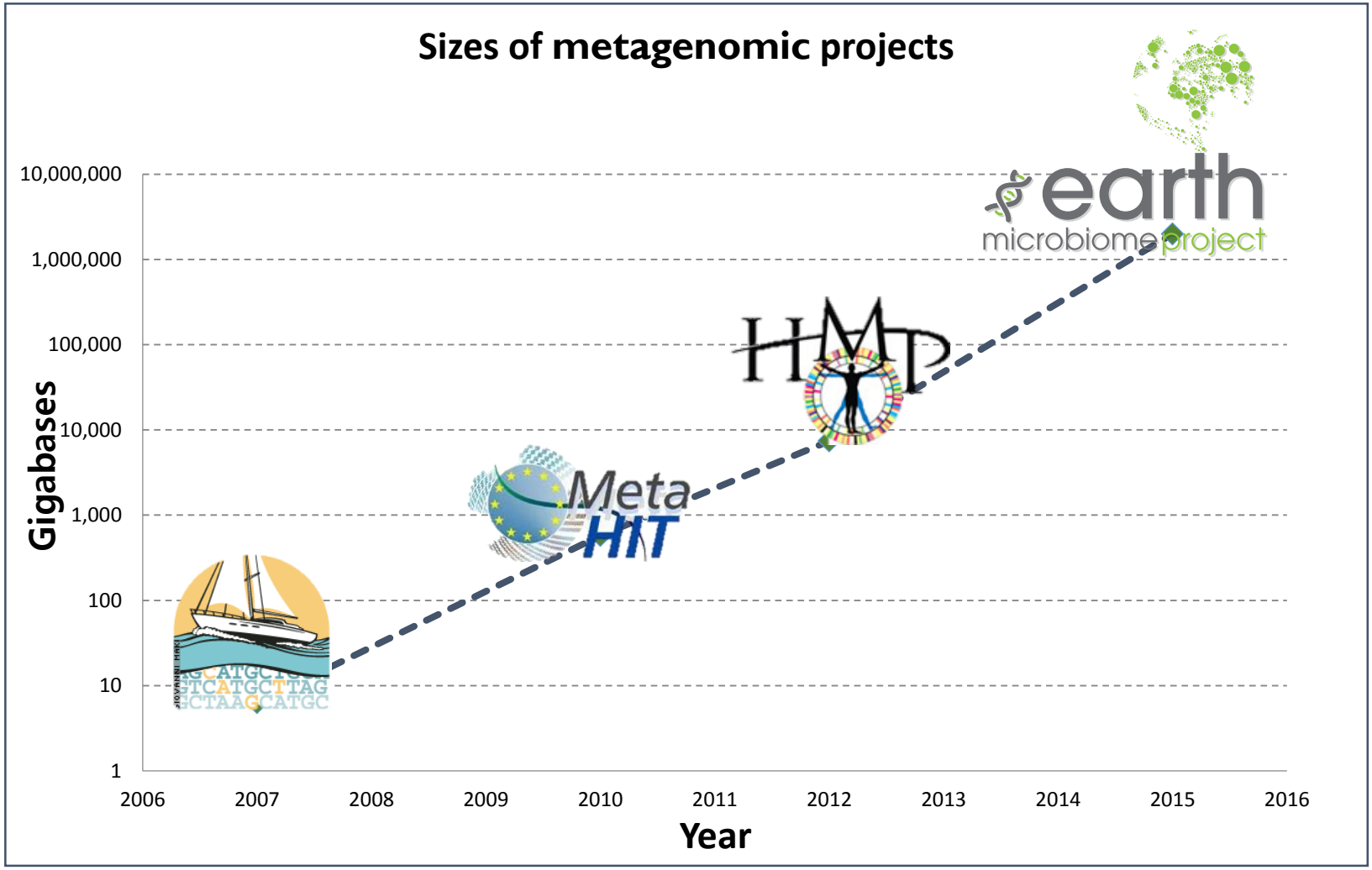
- Metagenomics analysis
  - On the species level: Who's there?
  - On the gene/functional level: What are they doing?
- RNA-seq analysis
  - Data normalization
  - Finding differentially expressed genes
- Computer exercise
  - Whole genome sequencing for variant detection

Metagenomics

- Some facts about microbes

Number of microbes on Earth	5×10^{30}
Number of microbes in all humans	6×10^{23}
Number of stars in the universe	7×10^{21}
Number of bacterial cells in one human gut	10^{14}
Number of human cells in one human	10^{13}
Number of bacterial genes in one human gut	3,000,000
Number of genes in the human genome	21,000

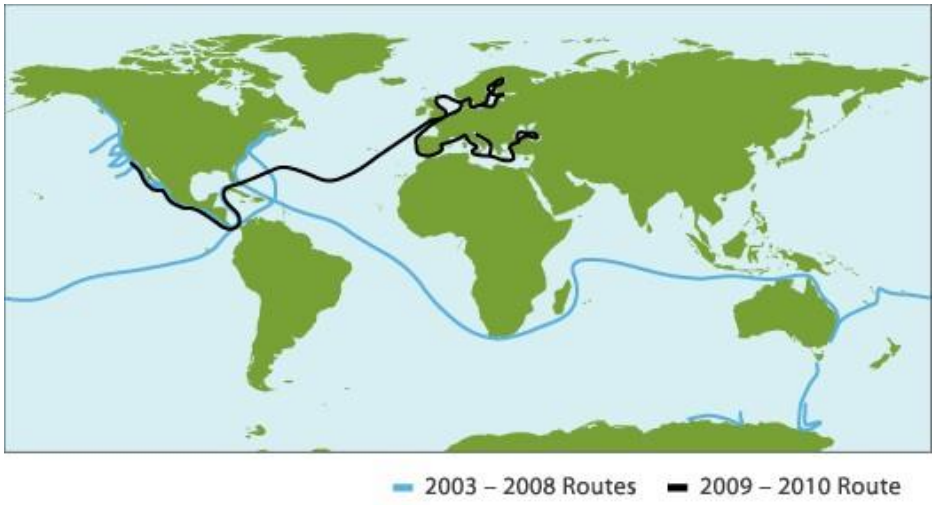
Metagenomic data revolution



GOS: http://www.jcvi.org/cms/research/projects/gos/overview/, Meta-HIT: http://www.metahit.e
HMP: http://www.hmpdacc.org/, EMP: http://www.earthmicrobiome.org/. Accessed 2013-10-

The global ocean sampling

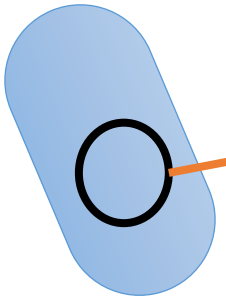
- Investigating microbial diversity in the ocean
- A sailing boat equipped with a sequencer



http://www.jcvi.org/cms/research/projects/gos/overview/

Microbial diversity

- Bacteria are present in every habitat on Earth
- There are up to 100 million bacterial species
  - only a small fraction of these are known
- More than 99% of all bacteria are not culturable under normal laboratory conditions



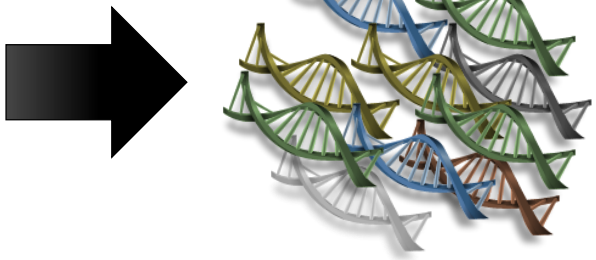
- 1-5 million bases
- 1000-5000 genes

1 gram of soil

- 10 000 species
- 100 million cells
- DNA: 100 terabases (10^{14})

Total sequencing to date: less than 1% of the DNA in 1 liter of ocean water.

Metagenomics



ATTTCCGGCATCTGACGAT

AACTCCTACGGGAGGCAGC

AGCTCAGATCGTCGCTGTC

TCTCACGAAATCCACCGTC

TCTTGAATTCGGCCATACG

Sample with microorganisms

DNA

Metagenome

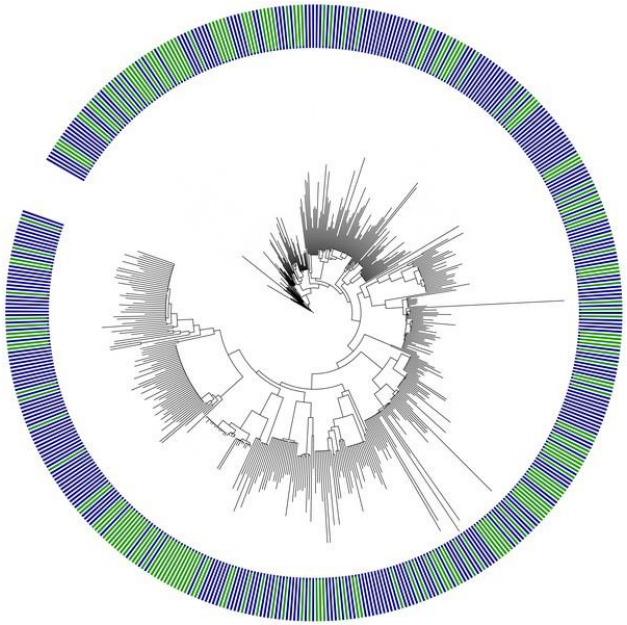
Metagenomics

- Metagenomics is used to study the unculturable organisms and viruses
  - ~50% of human gut bacteria are unculturable
  - <1% of environmental bacteria are unculturable
- Metagenomes are highly fragmented and undersampled
- The majority of DNA found in metagenomes is usually very hard to annotate

Two types of questions

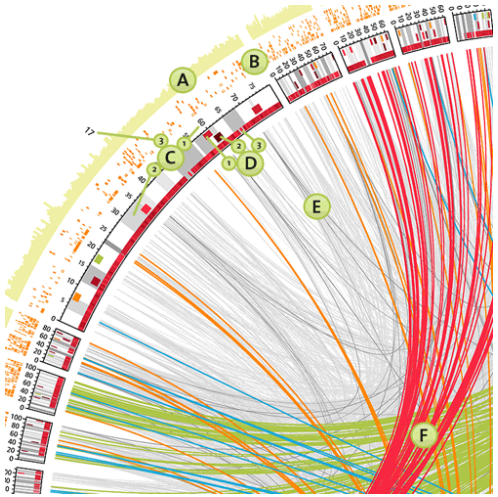
Who's there?

- Identification of species, phylum etc.
- Estimation of species abundance

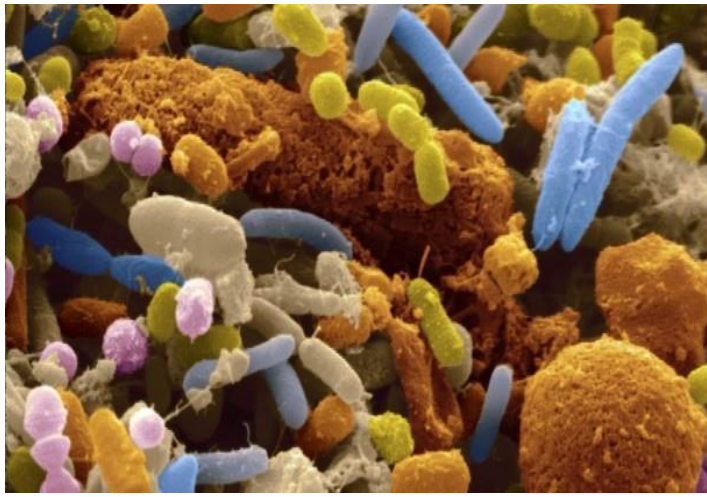


What are they doing?

- Functional annotation (gene families / pathways)
- Estimation of gene/ pathway abundance



Who is there?



- How would you find that?
  - Amplicon sequencing of phylogenetic marker genes
  - Shotgun sequencing
    - Mapping reads to species with known genomes
    - Binning of reads

Species identification using marker genes

- Prokaryots:
  - 16s rRNA gene
- Eukaryots:
  - 18s rRNA gene
- Can be amplified using amplicon sequencing

- Sequences mapped to known species using BLAST
- Operational taxonomic unit (OTU):
  - 97% sequence similarity for the 16s rRNA gene
  - Cluster based on sequence similarity using UCLUST

16s sequencing

Sample with microorganisms

DNA extraction

Amplification of the 16s rRNA gene

High throughput sequencing

```
CTTGGTATCCATTTTGGGTAAGTCATATTC
AATGAACTAGGTTTCGCAAACTTTTTGTTC
CTTTTTGTTCGAAAAGCGGTAGTGCATAGT
```

Data preprocessing

Species classification

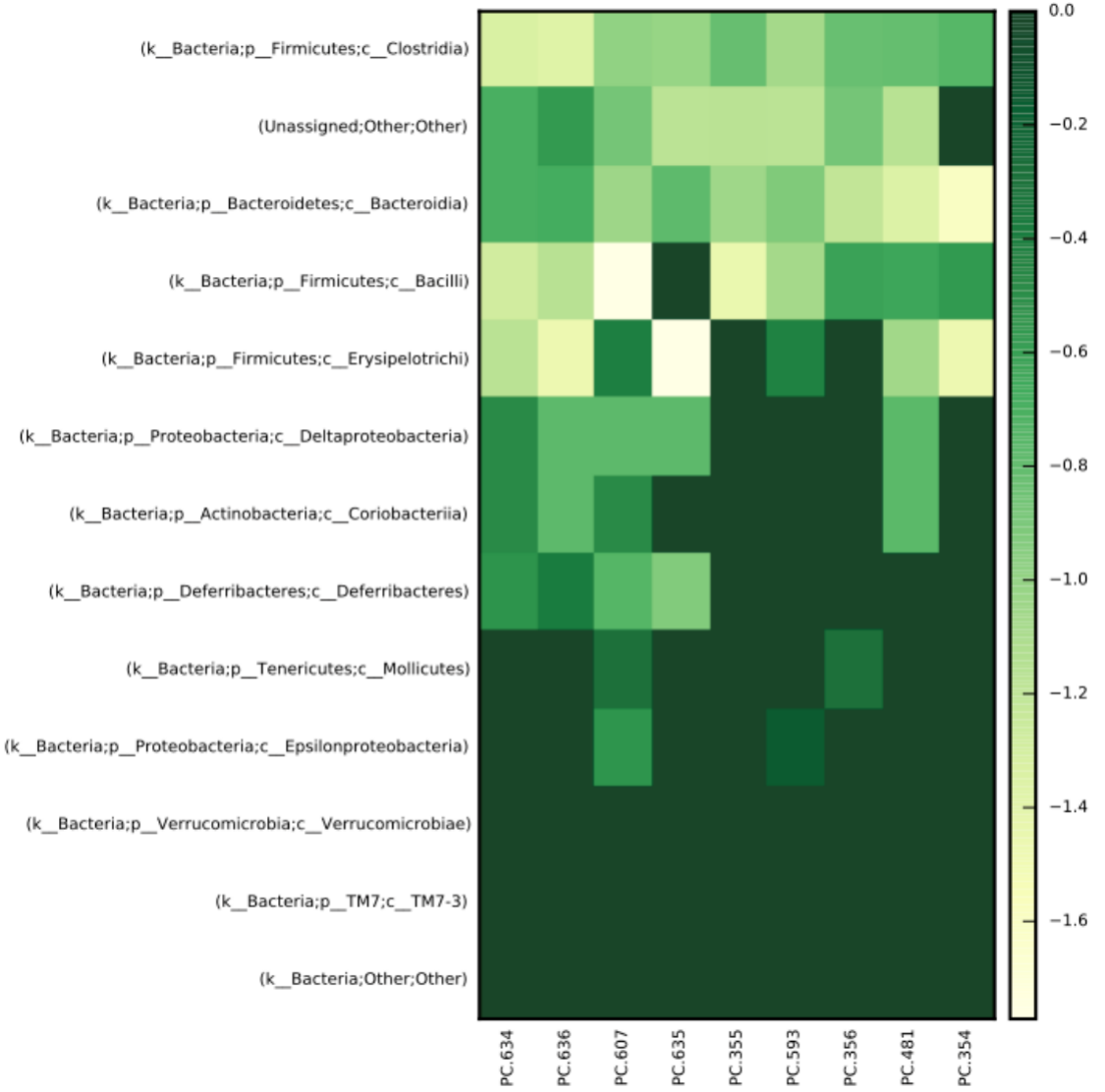
Species A

Species B

Species C

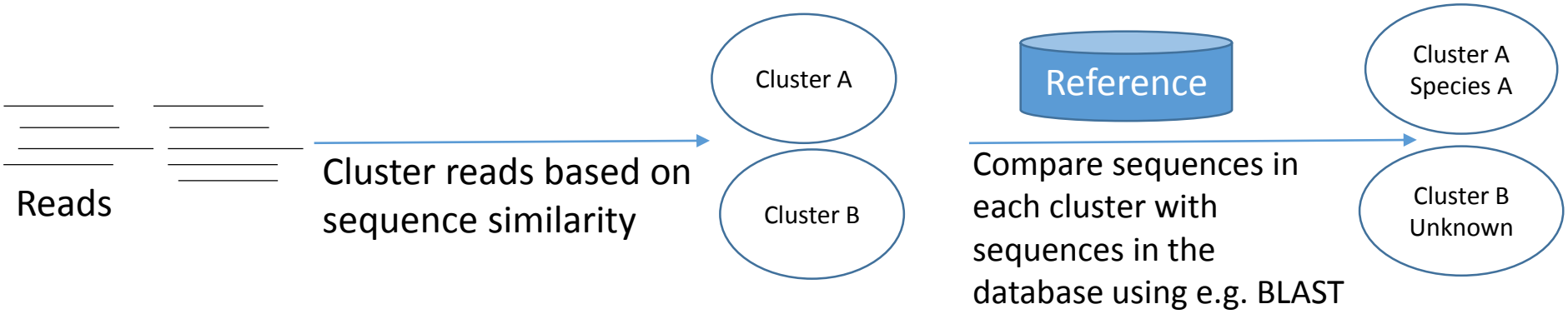
Species abundance

- Qiime
  - Bioinformatics program available at qiime.org
  - Pick OTUs
  - Analysis of species abundance
  - Bioinformatics analysis

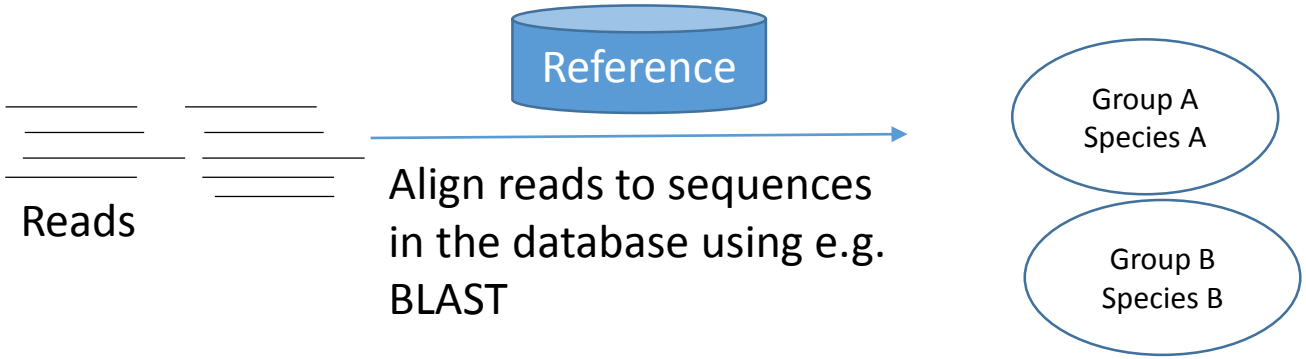


OTU picking

- Reference database with 16s sequences of known species
- Open OTU picking:

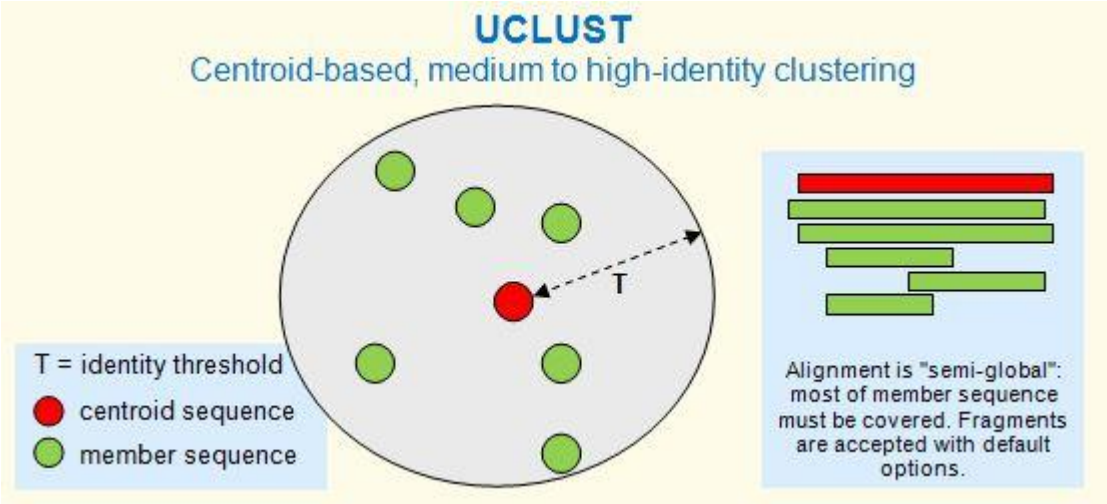


- Closed reference OTU picking:



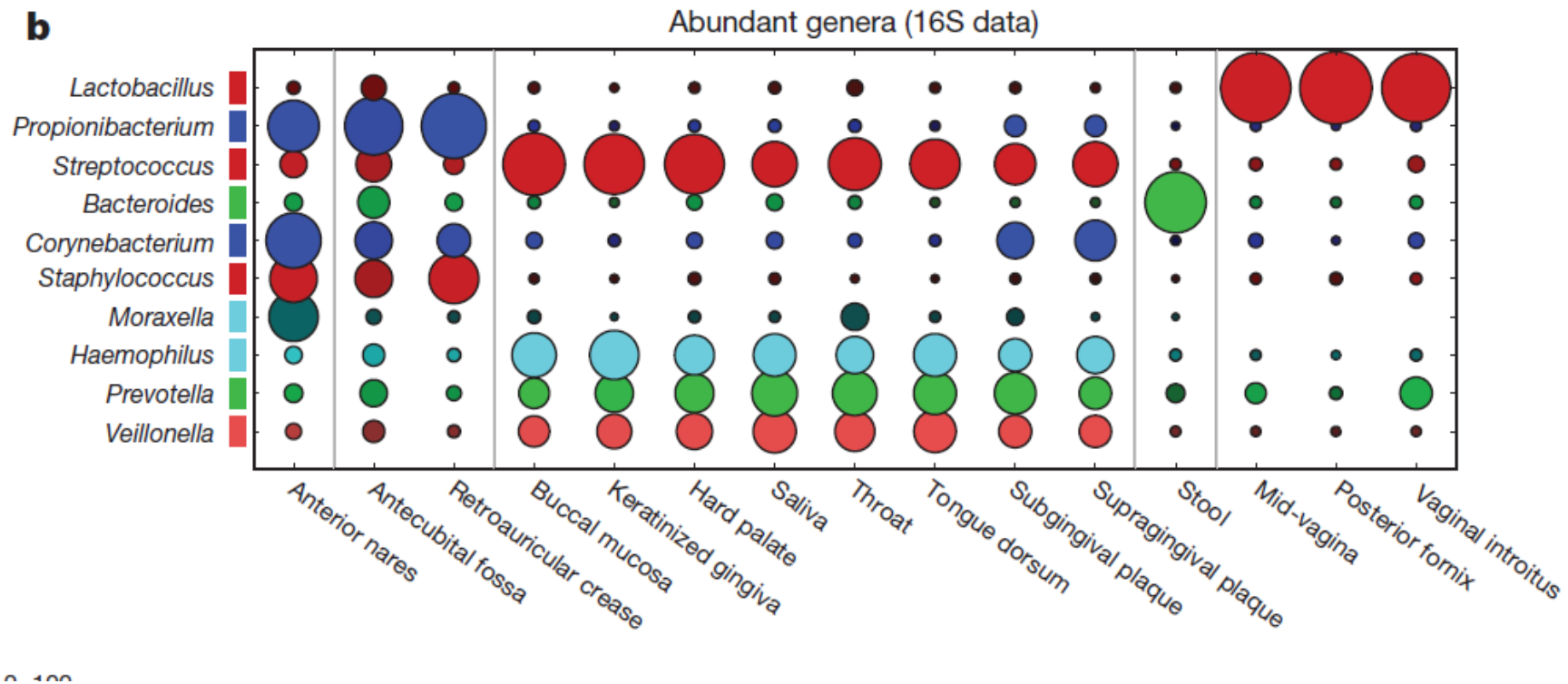
Uclust

- Fast clustering of short sequences based on sequence identity



Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics* 26(19), 2460-2461.

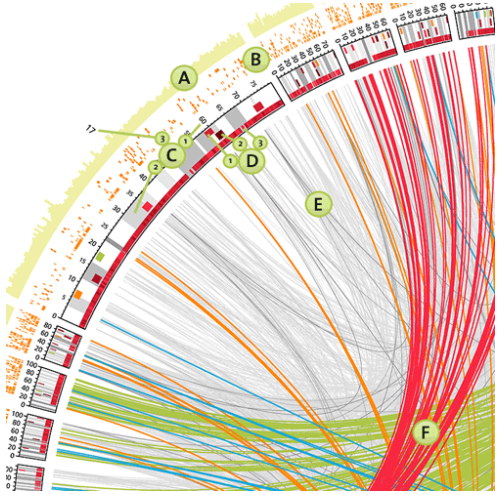
Example from the human gut microbiome



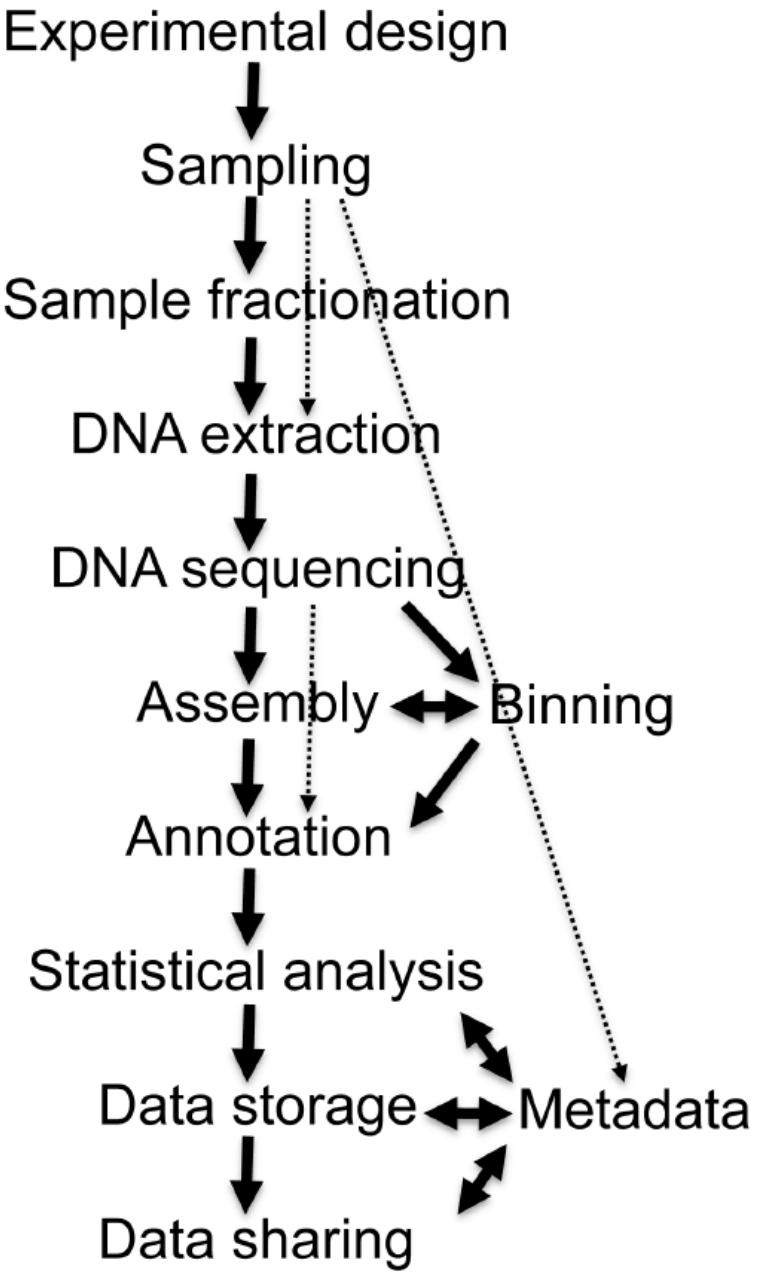
The Human Microbiome Project Consortium, Nature 486, 207–214 (14 June 2012)

What are they doing

- Shotgun metagenomics

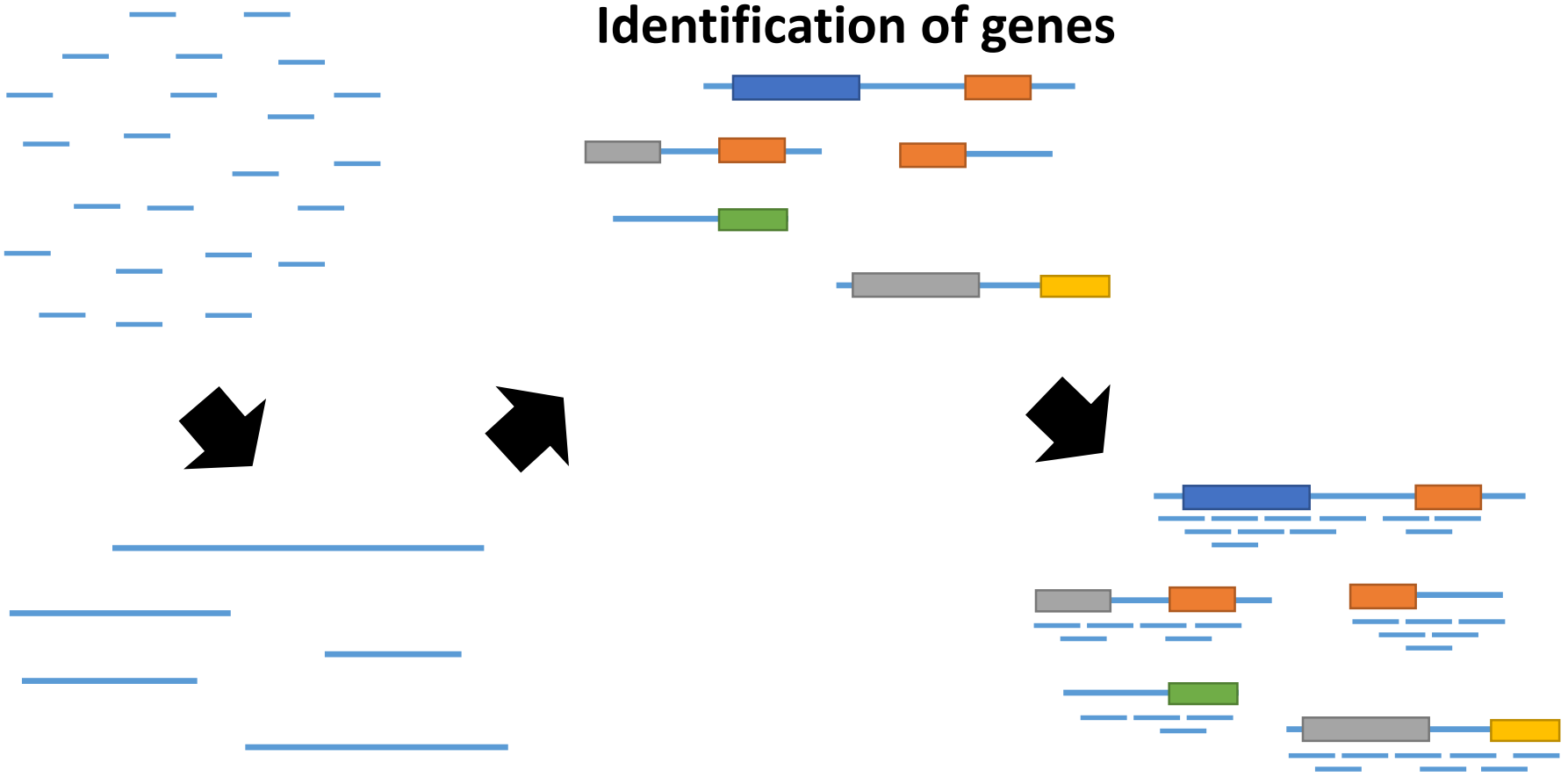


Analysis of a typical shotgun metagenomics dataset



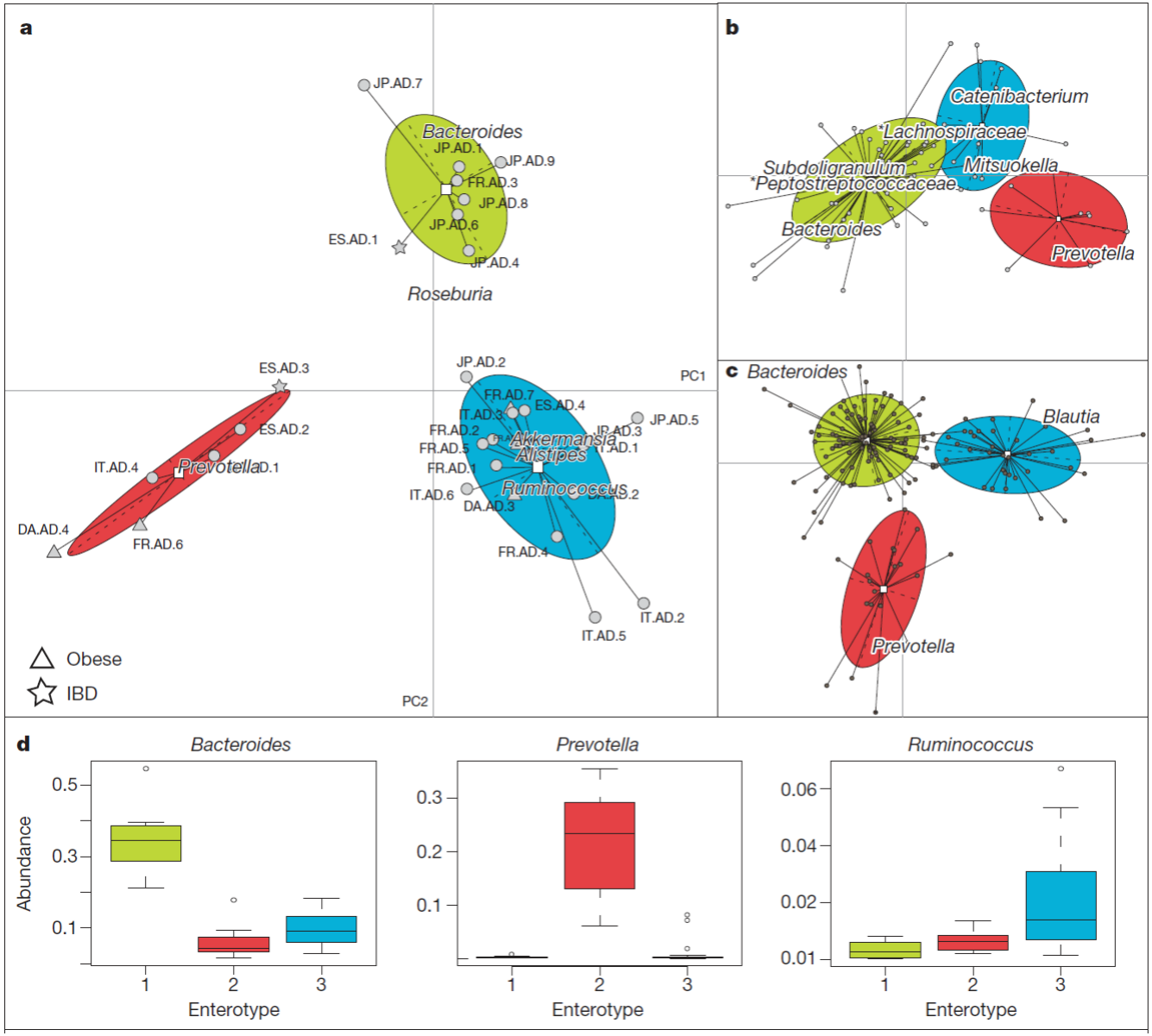
Thomas et al. Microbial Informatics and Experimentation 2012

Binning (functional analysis)

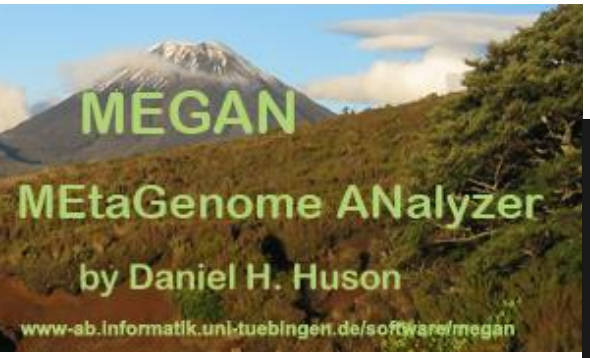


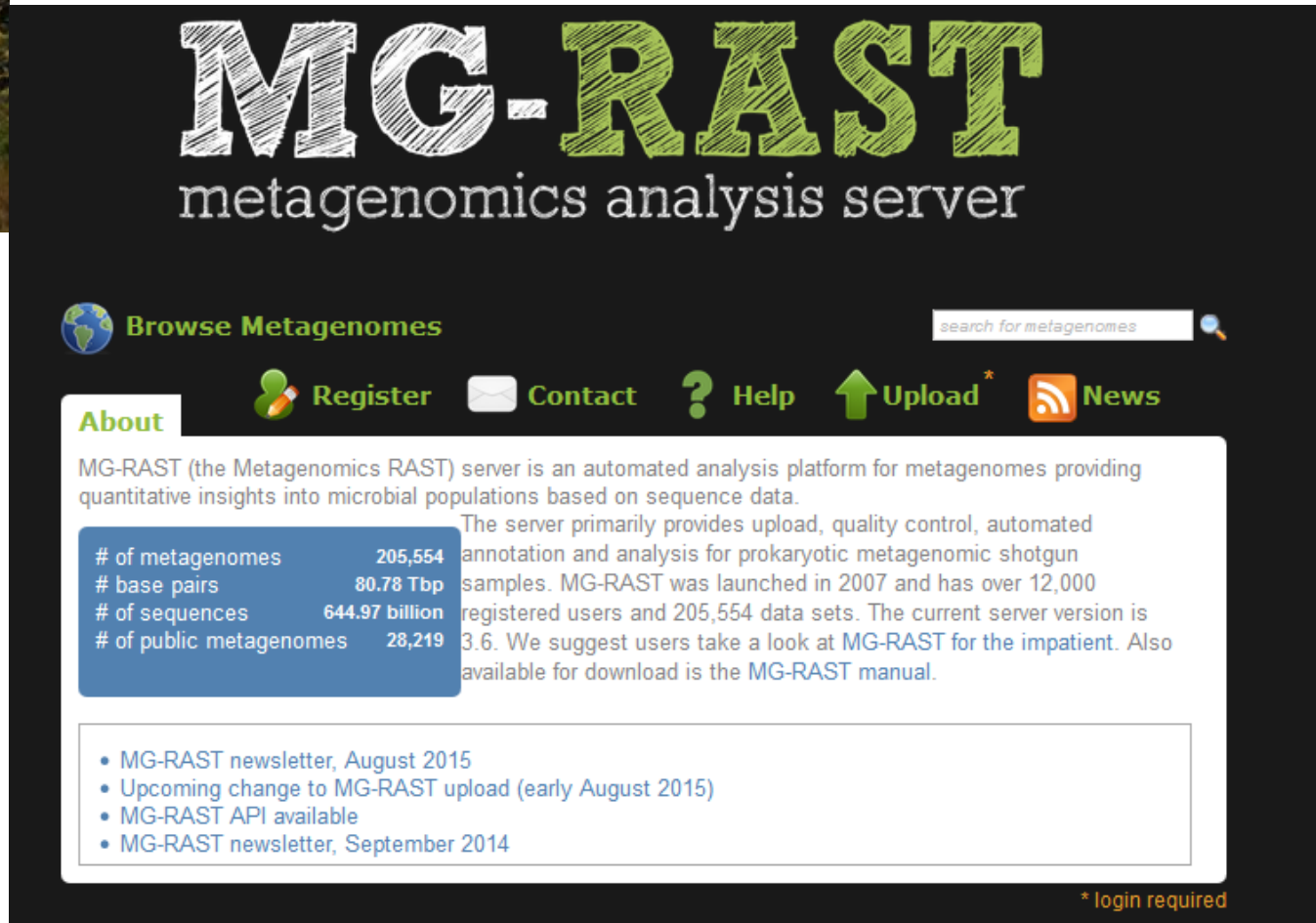
Enterotypes of the human gut

- Map reads to a gene catalog with 1500 known species
- Cluster based on species abundance



Metagenomics analysis software/server





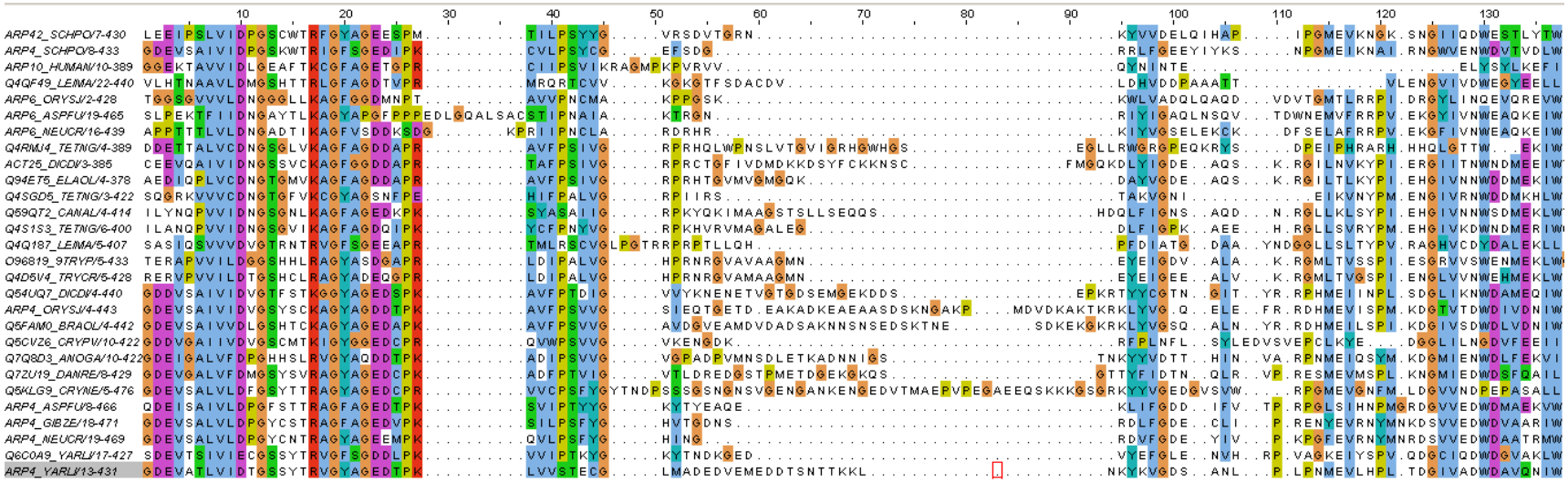
Metagenome assembly software

- Velvet
- Metavelvet
- MAQ
- SOAP *de novo*
- Etc.

- Most assemblers uses deBruijn graphs
  - Kmers
  - Need to specify k

Functional analysis

- ”Gene centric analysis” (What are they doing?)
- Only a small fraction of the bacterial genomes have been sequenced.
- Annotation done using protein profiles catching the variability (PFAM, TIGRFAM, COG, etc)



PFAM domain for actin.

Databases for functional domains / orthologous groups

- PFAM

  ~ 10,000 conserved functional domains, eukaryots and prokaryots

  Identification using hidden Markov models (HMM) based tools.

- TIGRFAM

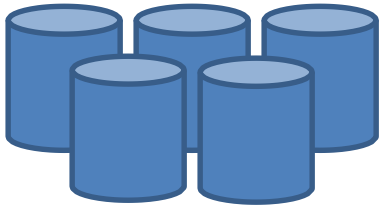
  ~4200 conserved protein families, mainly bacterial

  Identification using HMM

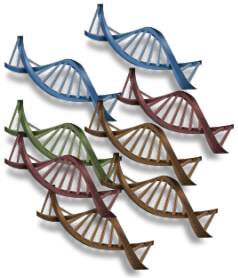
- COG
  - Clusters of orthologous groups, mainly bacterial
  - Identification using position specific weight matrices (PSWM)

Other functional annotation

- KEGG pathways
- GO-terms
- SEED classification



Reference database

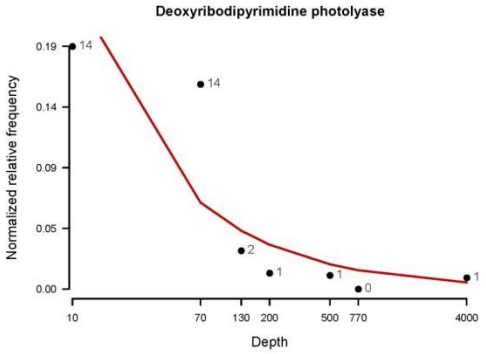


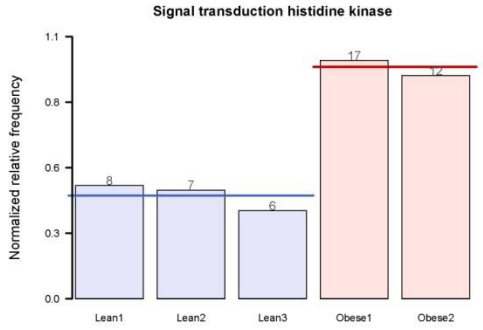
Annotation

Statistical analysis

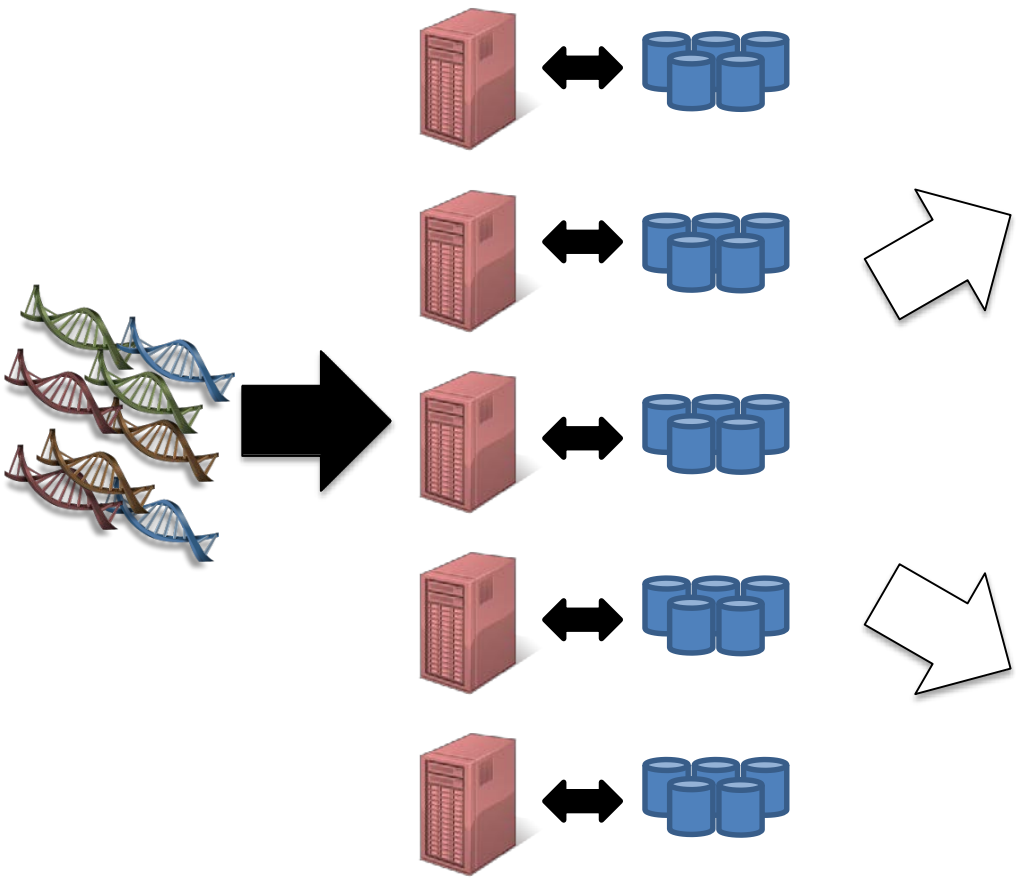
ShotgunAnnotatoR

ShotgunFunctionalizeR

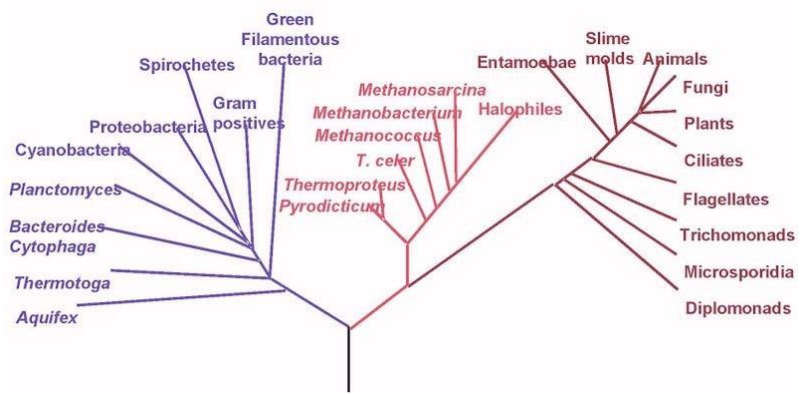




Kristiansson, E., Hugenholtz, P., Dalevi, D. (2009). ShotgunFunctionalizeR – an R-package for functional analysis of metagenomes. Bioinformatics 25(20). http://shotgun.zool.gu.se



Annotation

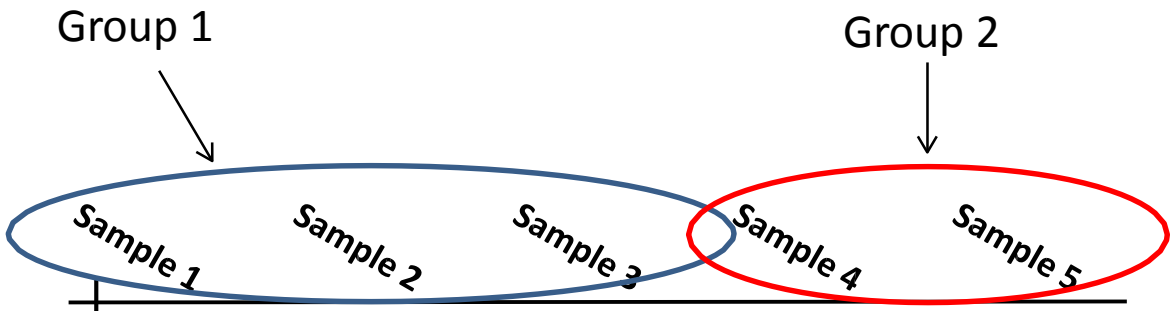


Taxonomic affiliation

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene1	591	536	1260	284	19
Gene2	28	21	19	36	10
Gene3	53	51	97	118	36
Gene4	106	149	266	47	11
....					
....					

Gene occurrences

Identification of significant genes



Gene1	591	536	1260	284	19
Gene2	28	21	19	36	10
Gene3	53	51	97	118	36
Gene4	106	149	266	47	11
....					
Gene1312	243	362	163	258	423
Gene1313	13	43	23	67	34
....					
Total	132 567	80 456	197 723	73 491	134 513

Normalization

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene1	591	536	1260	284	19
Gene2	28	21	19	36	10
Gene3	53	51	97	118	36
Gene4	106	149	266	47	11
….					
Gene1312	243	362	163	258	423
Gene1313	13	43	23	67	34
….					
Total	132 567	80 456	197 723	73 491	134 513

$X_{i,j}$

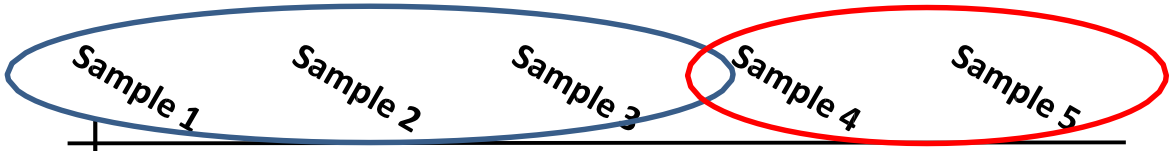
n_j

$X_{i,j}$ -number of reads matching gene *i* in sample *j*

n_j -normalization factor per sample

$$R_{i,j} = \frac{X_{i,j}}{n_j}$$

Normalization



Gene1	0.004458	0.006662	0.006373	0.003864	0.000141
Gene2	0.000211	0.000261	9.61E-05	0.00049	7.43E-05
Gene3	0.0004	0.000634	0.000491	0.001606	0.000268
Gene4	0.0008	0.001852	0.001345	0.00064	8.18E-05
....					
Gene1312	0.001833	0.004499	0.000824	0.003511	0.003145
Gene1313	9.81E-05	0.000534	0.000116	0.000912	0.000253
....					
Total	1	1	1	1	1

How to normalize metagenomic data?

$$R_{i,j} = \frac{X_{i,j}}{n_j}$$

- n_j – normalization factor per sample
- Divide with total number of reads mapped in each sample?
- Divide with the total number of reads in each sample
- Divide with the total number of reads mapping to the 16s rRNA gene in each sample?
- More advanced method?

Identification of significant genes

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene1	591	536	1260	284	19
Gene2	28	21	19	36	10
Gene3	53	51	97	118	36
Gene4	106	149	266	47	11
….					
Gene1312	243	362	163	258	423
Gene1313	13	43	23	67	34
….					
Total	1 32 567	80 456	1 97 723	73 491	1 34 513

$$\log\left(\frac{\mathrm{E}\left[X_{i,j}\right]}{\mathrm{n}_{\mathrm{j}}}\right)=\alpha_0+\sum\alpha_k y_k$$

Baseline Covariates (groups)

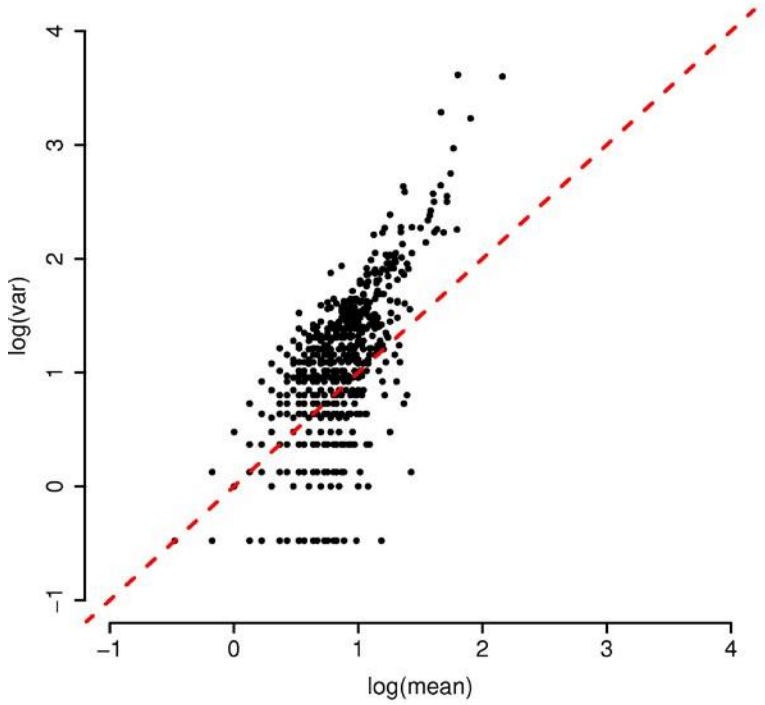
Statistical analysis

- Data from metagenomics is descrete (counts per gene/species)
- Not normally distributed
- $X_{i,j} \sim \text{Poisson}(\lambda_i)$

  $\text{E}[X_{i,j}] = \lambda_i$

  $\text{Var}[X_{i,j}] = \lambda_i$

Statistical analysis



- $\mathrm{Var}\left[X_{i,j}\right] > \mathrm{E}\left[X_{i,j}\right]$
- Overdispersed data!

$$\mathrm{Var}\left[X_{i,j}\right] = \phi\lambda_i$$

Estimated from the total residual sum

- The proportion of false positives are estimated using Benjamini-Hochberg's false discovery rate.

Summary metagenomics

- Metagenomics provides a powerful way to do culture-independent analysis of bacterial communities
- The low cost of next generation sequencing have increased the power of metagenomics substantially
- Examples of metagenomics studies of microbial communities in the human gut and from environmental samples

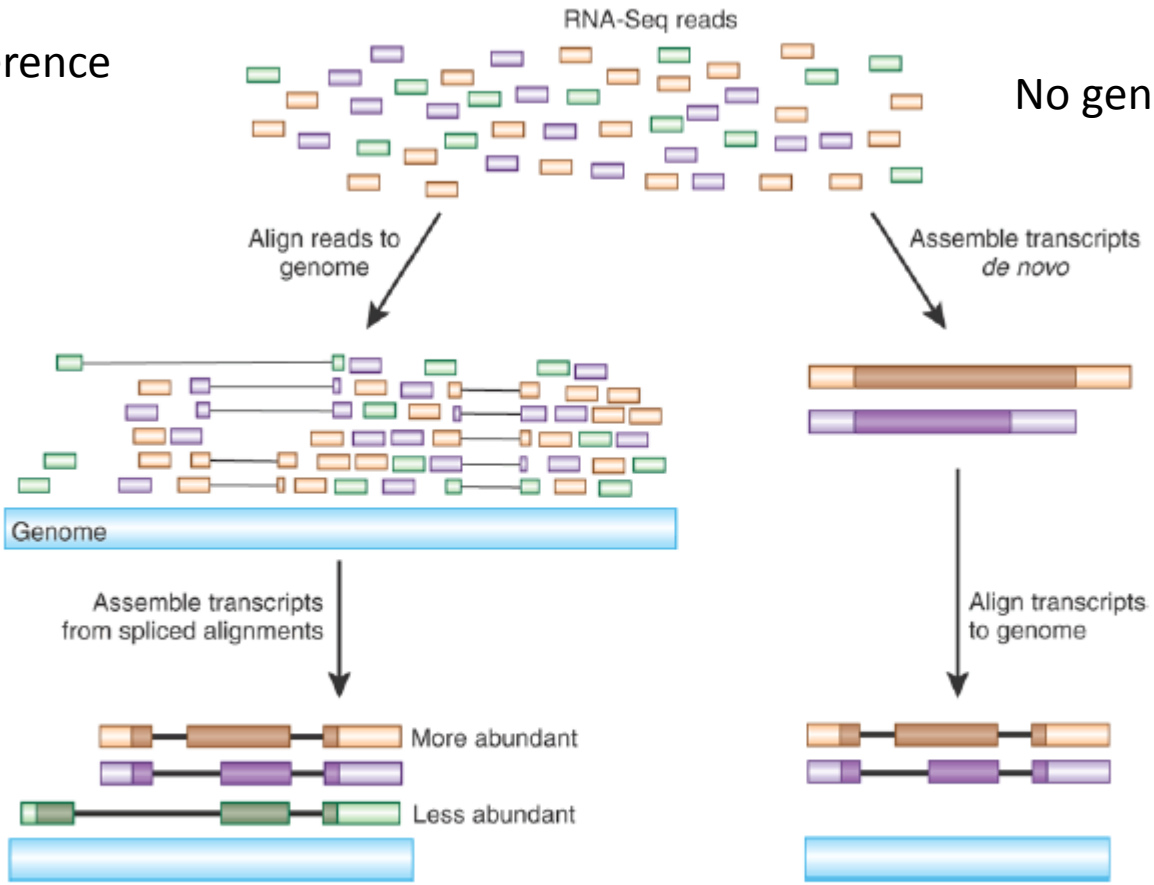
RNA-seq

- Large-scale mRNA quantification
  - Identification of differentially expressed genes
    - Sequence all mRNA and map to reference sequence
- De novo transcriptome assembly
  - Find new transcripts
  - Alternative splicing
  - When no reference sequence is available
    - Map the reads back to the newly assembled contigs
  - Can help in genome annotation

RNA-seq analysis strategy

Good reference

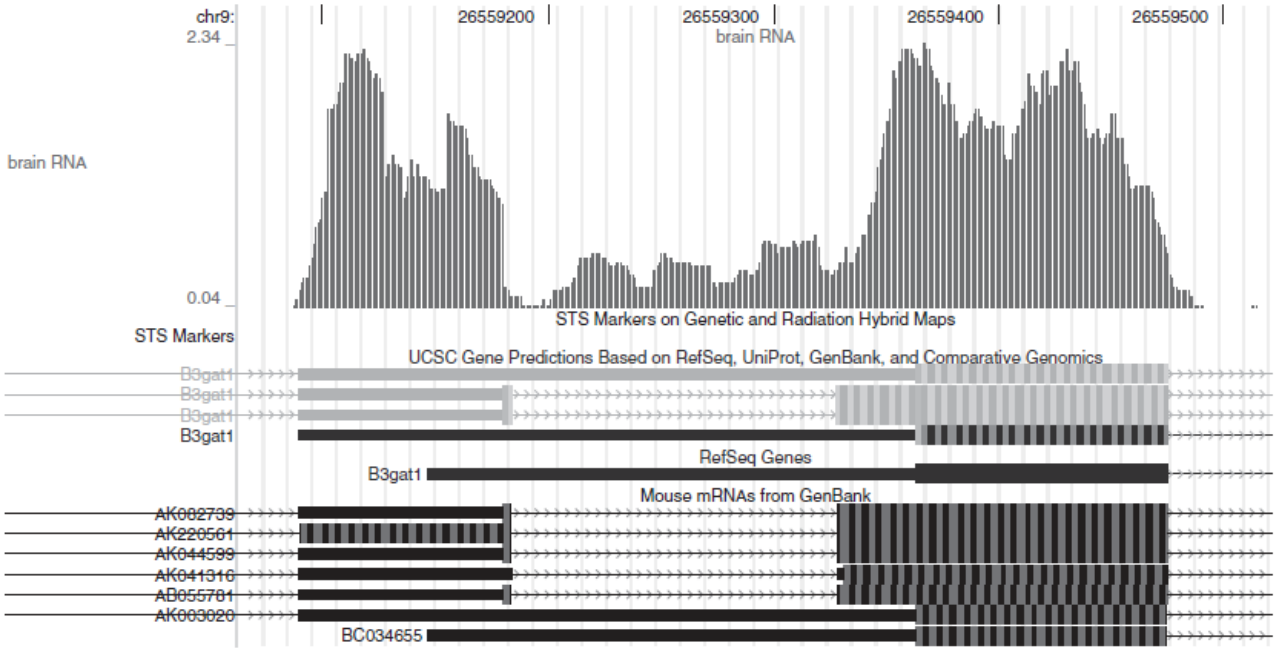
No genome



Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

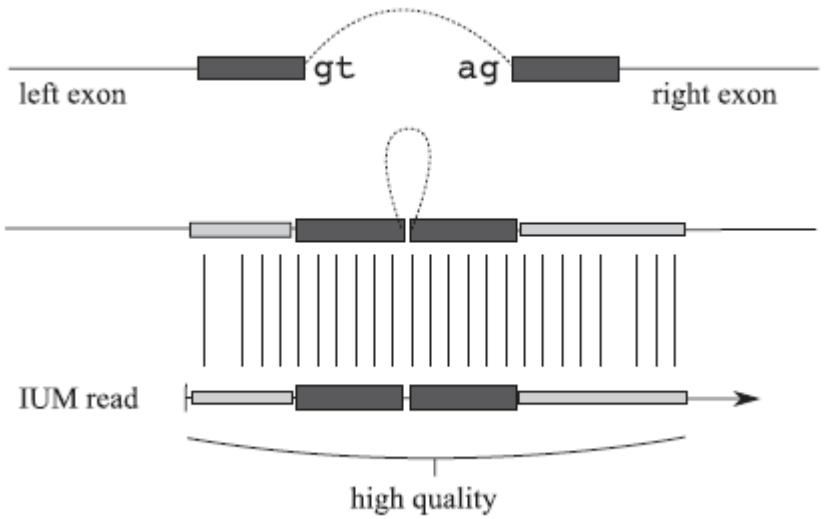
Alignment

- Using a splice-aware aligner



TopHat aligner (Trapnell et al. Bioinformatics 2009)

Alignment



TopHat aligner (Trapnell et al. Bioinformatics 2009)

De novo transcriptome assembly

Trinity command line example:

```
Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory  20G
```

- Inchworm assembles the transcripts
- Chrysalis and Butterfly estimates possible splice variants from the data

Statistical analysis

- Data from RNA-seq comes as reads/fragments per gene
  - $X_{i,j}$ = number of reads matching gene i in sample j

	Treatment A			Treatment B		
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Gene1	66489	29192	18643	21721	84669	80540
Gene2	11288	2899	1062	6130	9581	17251
Gene3	44979	12906	14604	10378	85043	39478
Gene4	7133	4772	1124	319	6863	7286
Gene5	34282	14379	13748	6133	12648	7620
Gene6	6531	7184	1962	651	1334	13125
Total	170702	71332	51143	45332	200138	165300

Data normalization

$$R_{i,j} = \frac{X_{i,j}}{n_j}$$

- n_j – normalization factor per sample
- Divide with total number of reads mapped in each sample?

- House keeping genes have a large influence on the normalization
- Robust scaling (Anders and Huber 2010)

$$n_j = median_i \frac{X_{i,j}}{\left(\prod_{j=1}^{m} X_{i,j} \right)^{1/m}}$$

RNA-seq is semi-quantitative

- Compare the same gene over different conditions

    - calculate fold-change and p-value

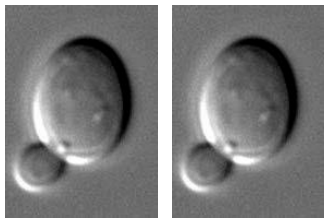
- Difficult to compare two genes from the same samples
  - Genes have different lengths
  - Genes have different GC-content (PCR-bias)

Study design

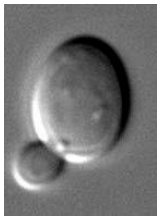
- How much should I sequence?
  - Depends on your question
  - Metagenomics: Sequence as much as possible
    - Your metagenome will still be undersampled
    - Need a lot of sequence to do assembly
  - RNA-seq: Sequence deep enough (enough coverage) to be able to detect both highly expressed transcript and rare transcripts
- Biological Replicates!!!

Sequencing lab

Genome sequencing of amylase producing yeast strains



UV1052 and UV715 mutation strains



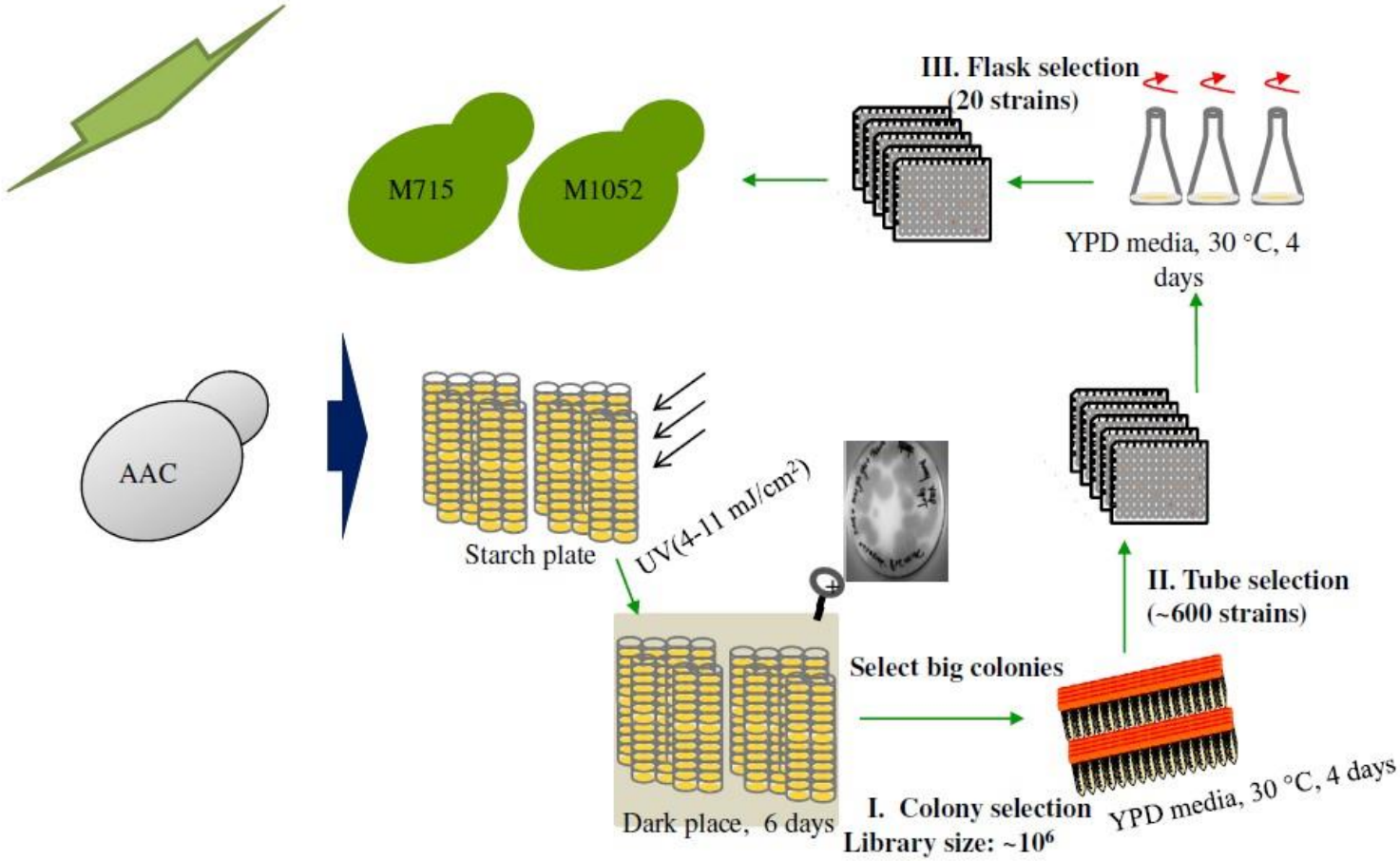
WT CEN.PK

Illumina deep sequencing (100x – 200x coverage)

- Alignment of reads to reference
- mutation calling / variant detection
- Identification of chromosomal rearrangements

Comparison of changes on genomic and transcriptomic level

Transcriptome analysis


III. Flask selection
(20 strains)
M715
M1052
YPD media, 30 °C, 4
days
AAC
Starch plate
UV(4-11 mJ/cm²)
II. Tube selection
(~600 strains)
Select big colonies
Dark place, 6 days
I. Colony selection
Library size: ~10⁶
YPD media, 30 °C, 4 days

Software used in lab

- Fastx toolkit – programs for preprocessing and quality control of Fastq and fasta files
- BWA – short read aligner
- Samtools – handling SAM and BAM files
- Integrative Genomics Viewer (IGV) – A genome browser viewing alignments (BAM-files)