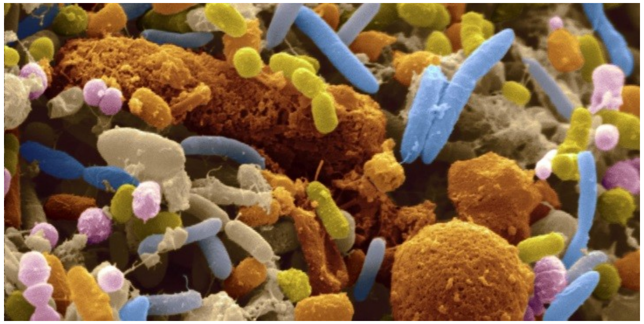
Metagenomics and RNA-seq

Tobias Österlund



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

NGS part of the course

Week 4	Friday 12/2	15.15-17.00	NGS lecture 1: Introduction to NGS, alignment, assembly
Week 6	Thursday 18/2	08.00-09.45	NGS lecture 2: RNA-seq, metagenomics
Week 6	Thursday 18/2	10.00-11.45	NGS computer lab: Resequencing analysis
Week 7	Thursday 3/3	10.00-11.45	Marcela: Exome sequencing
Week 8	Monday 7/3	23.59	Deadline: Essay on NGS and metagenomics
Week 8	Thursday	08.00-09.45	Fredrik: HMMer and Metagenomics

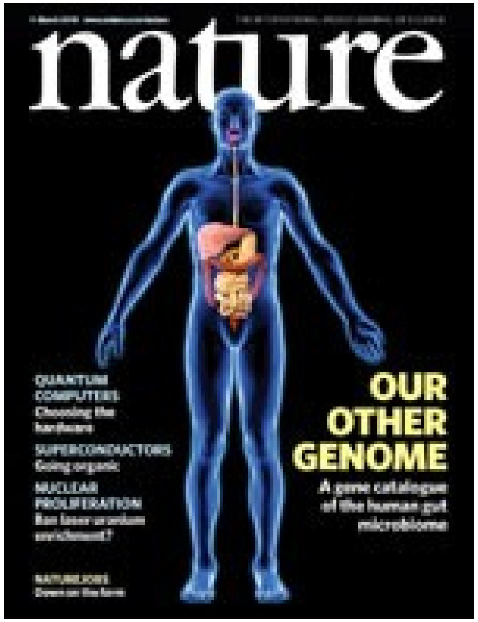
Today’s lecture

- 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

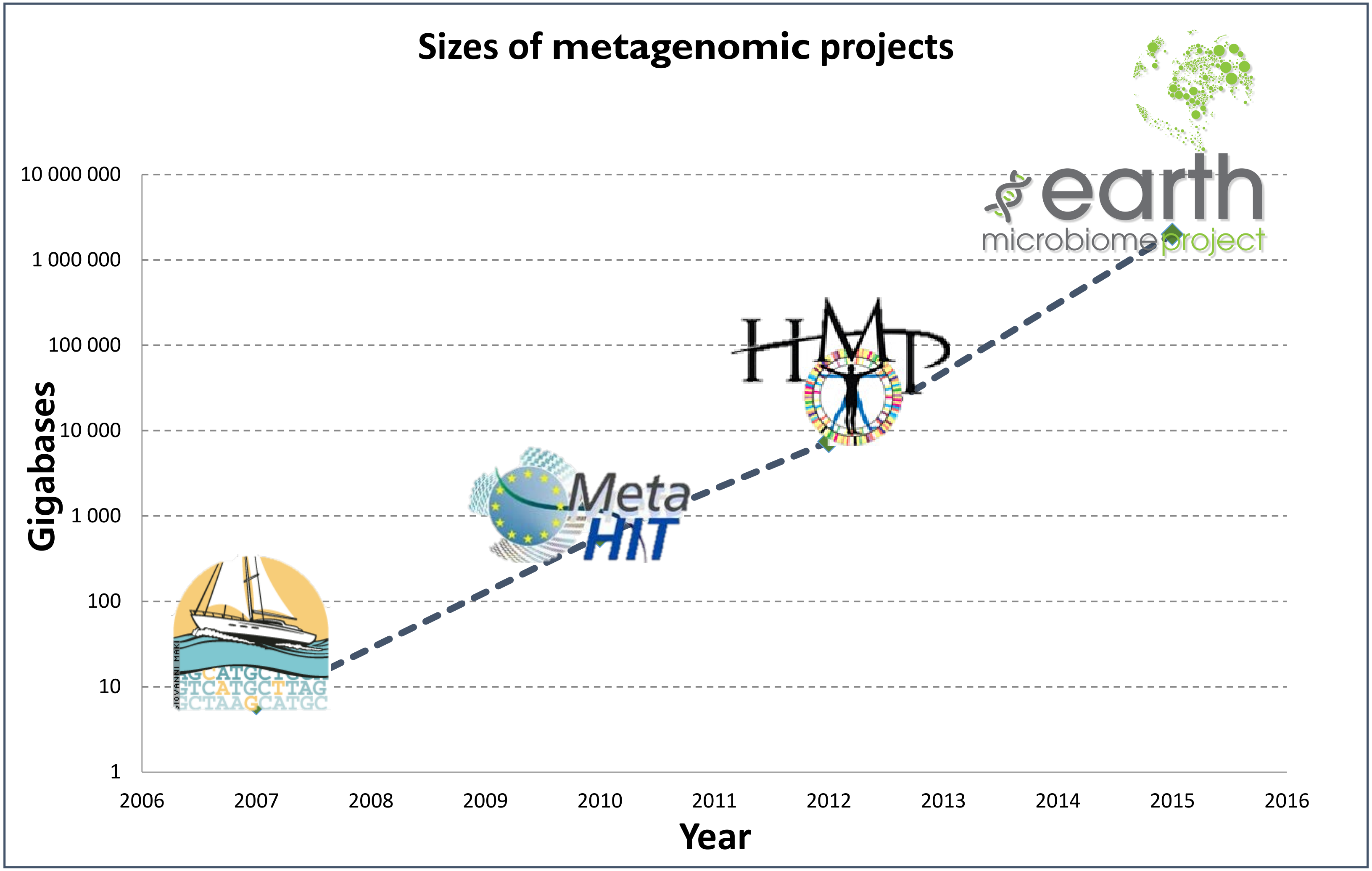
Metagenomics

- 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



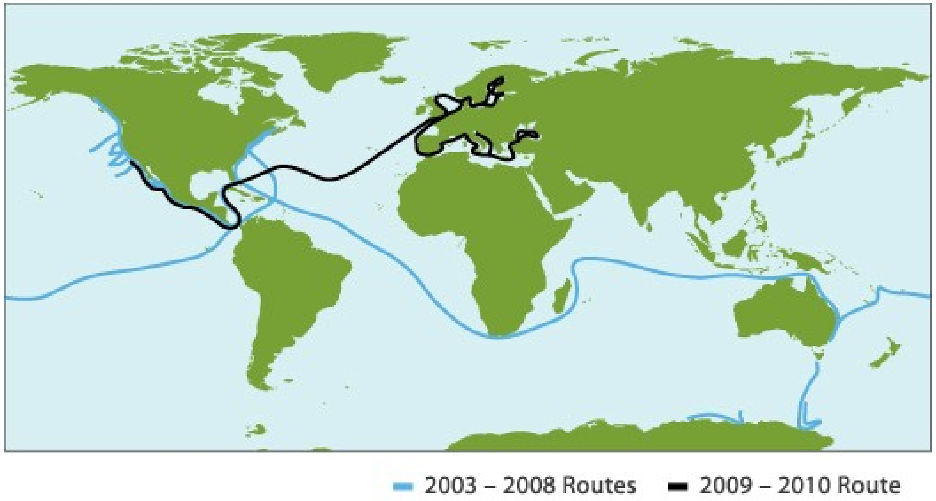
Metagenomic data revolution



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

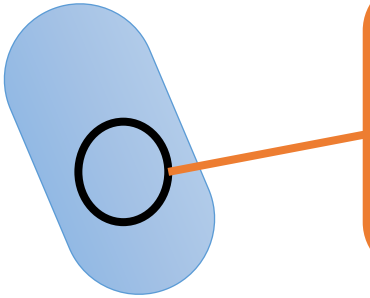
- Investigating microbial diversity in the ocean
- A sailing boat equipped with a sequencer



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



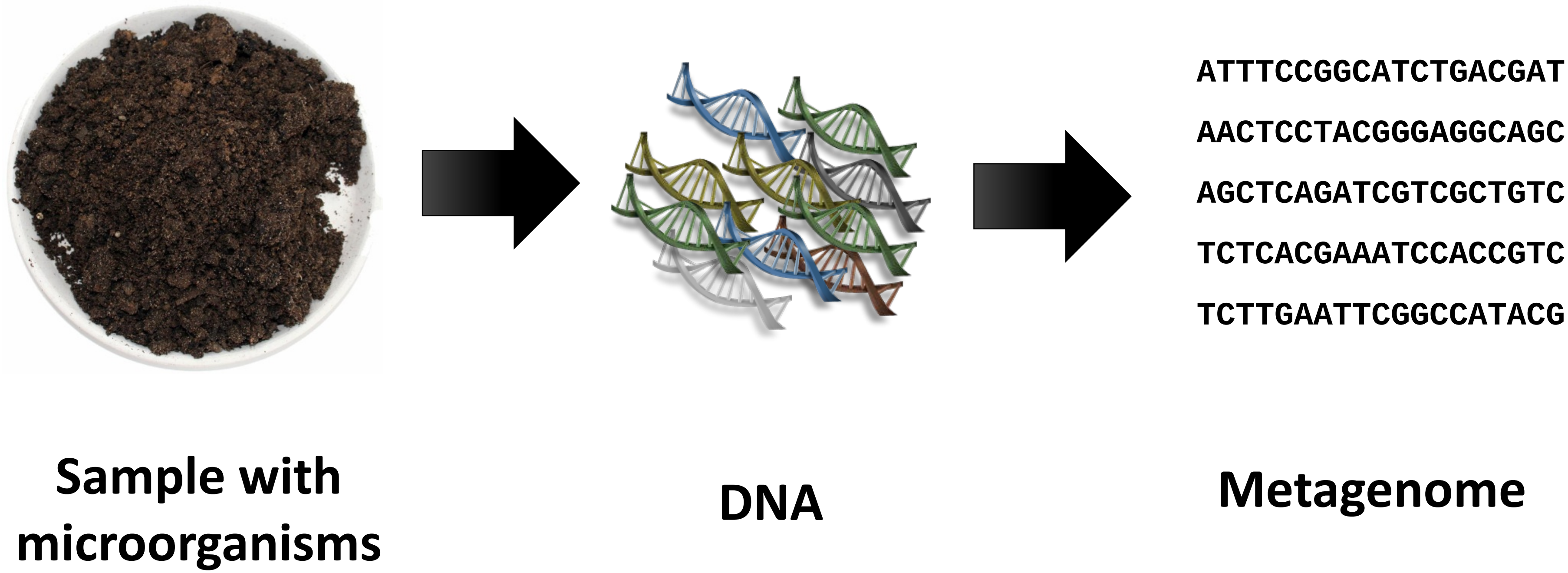
- 1-5 million bases
- 1000-5000 genes

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

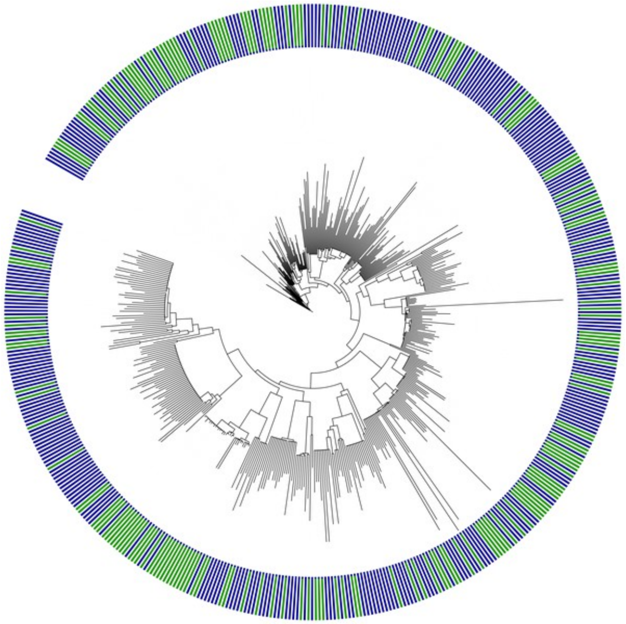
Two types of questions

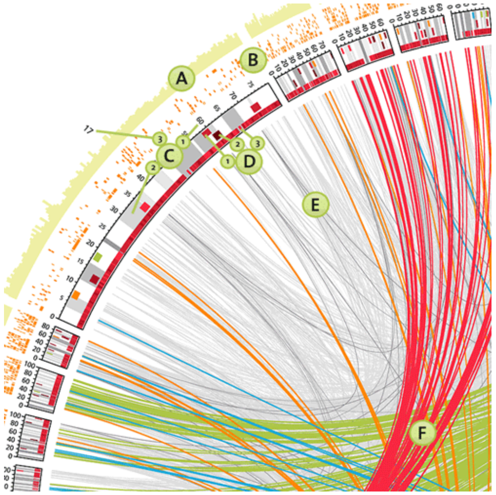
Who's there?

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

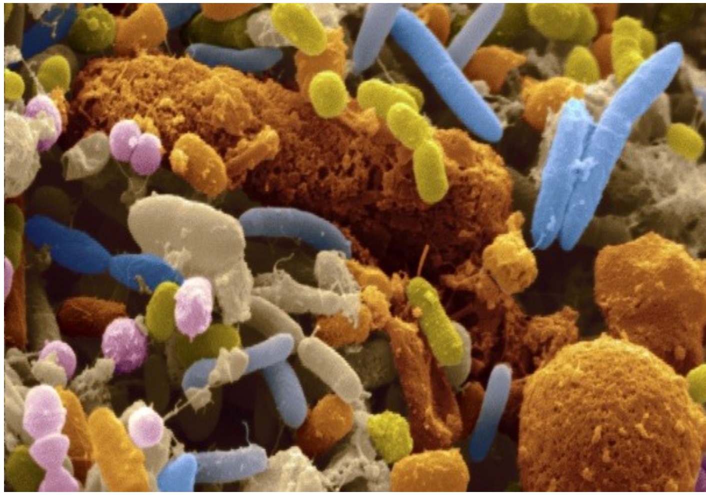
What are they doing?

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





Who is there?

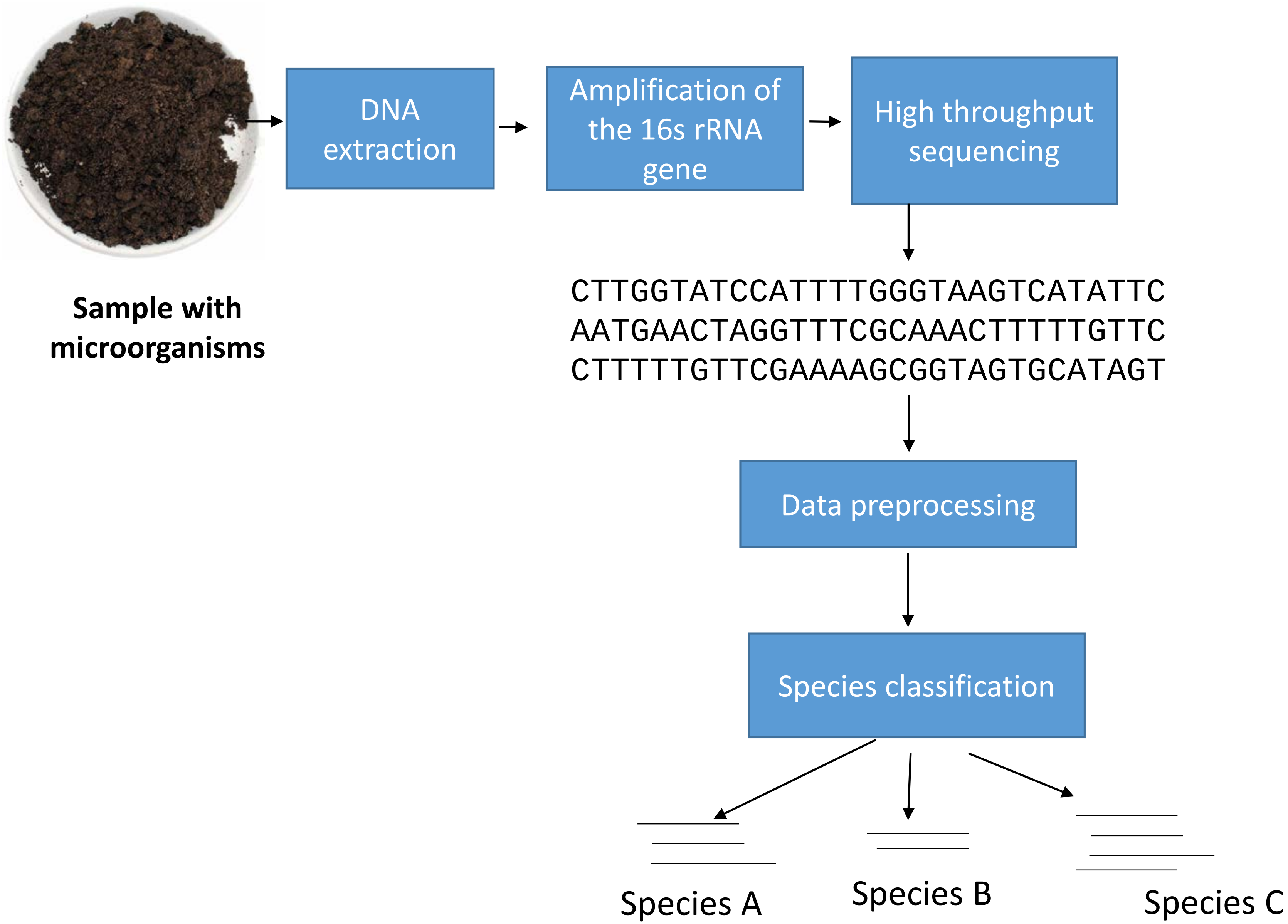


- How would you find that?
  - Amplicon sequencing of phylogenetic marker genes
  - Shotgun sequencing
    - Mapping reads to species with known genomes
    - Binning of reads

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

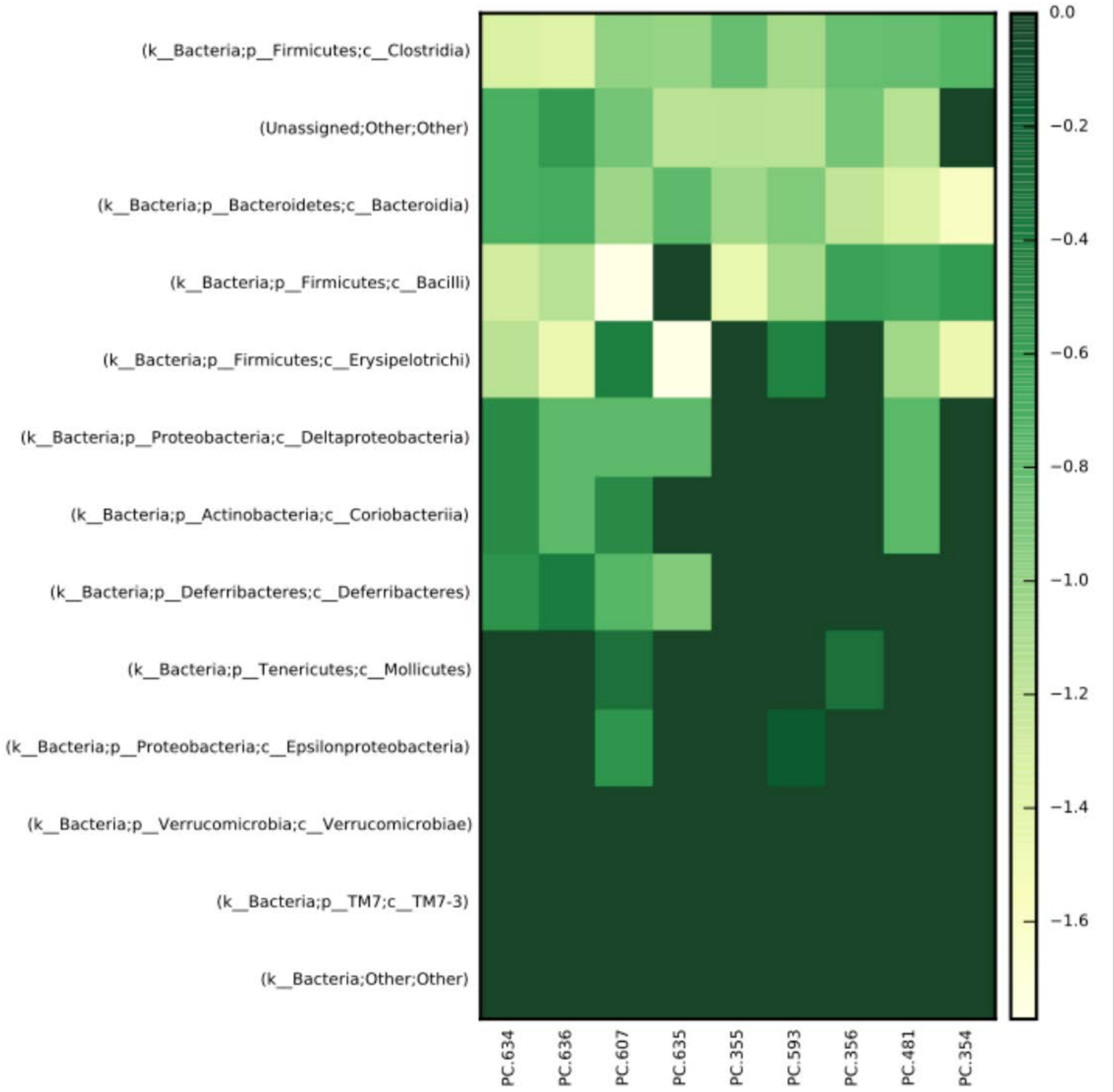
Species A

Species B

Species C

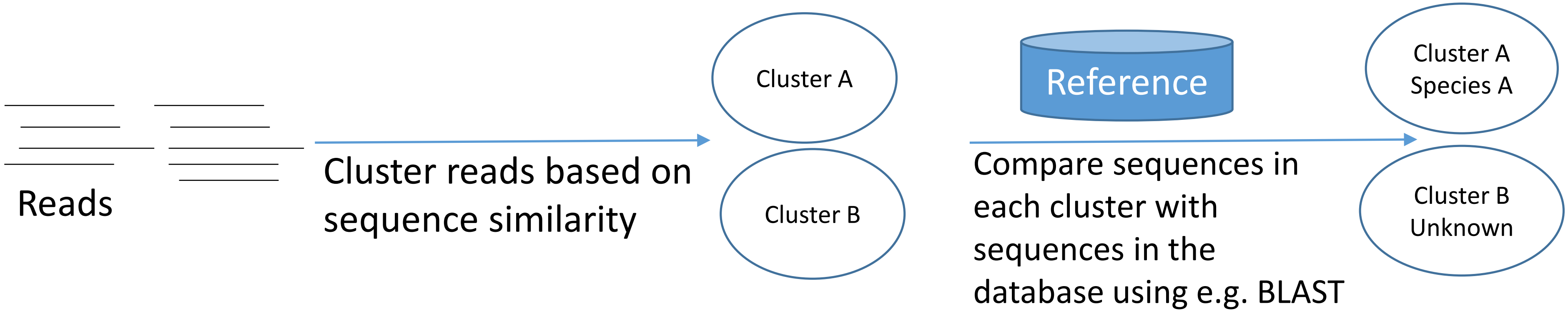
Species abundance

- 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:



Reads

Cluster reads based on sequence similarity

Cluster A

Cluster B

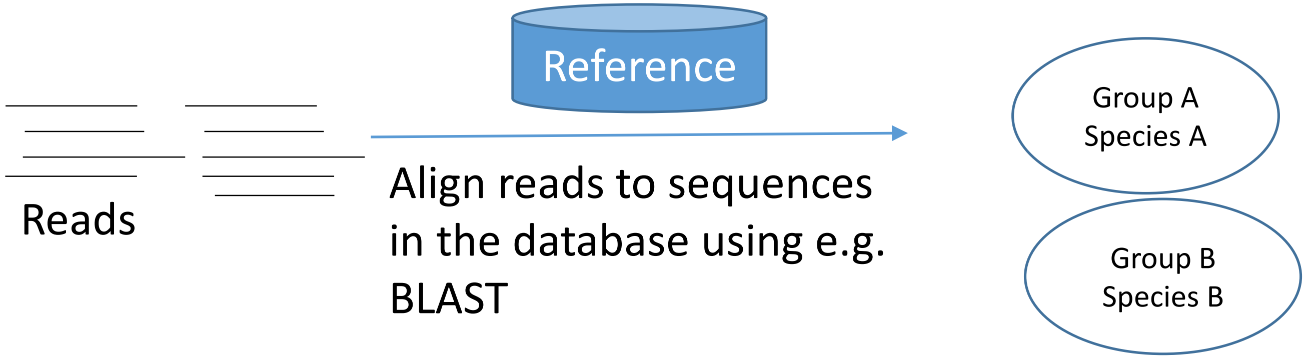
Reference

Compare sequences in each cluster with sequences in the database using e.g. BLAST

Cluster A
Species A

Cluster B
Unknown

- Closed reference OTU picking:



Reads

Reference

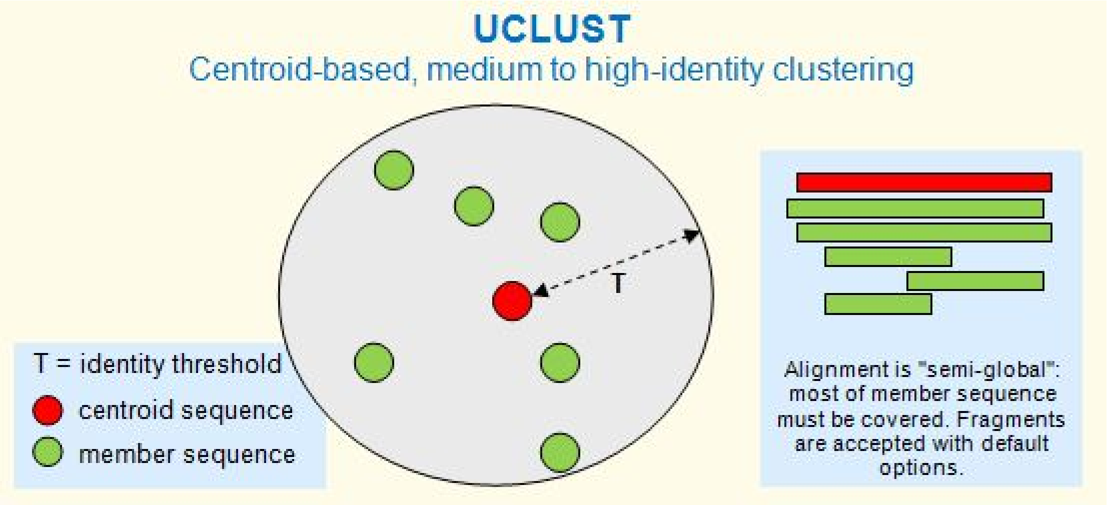
Align reads to sequences in the database using e.g. BLAST

Group A
Species A

Group B
Species B

Uclust

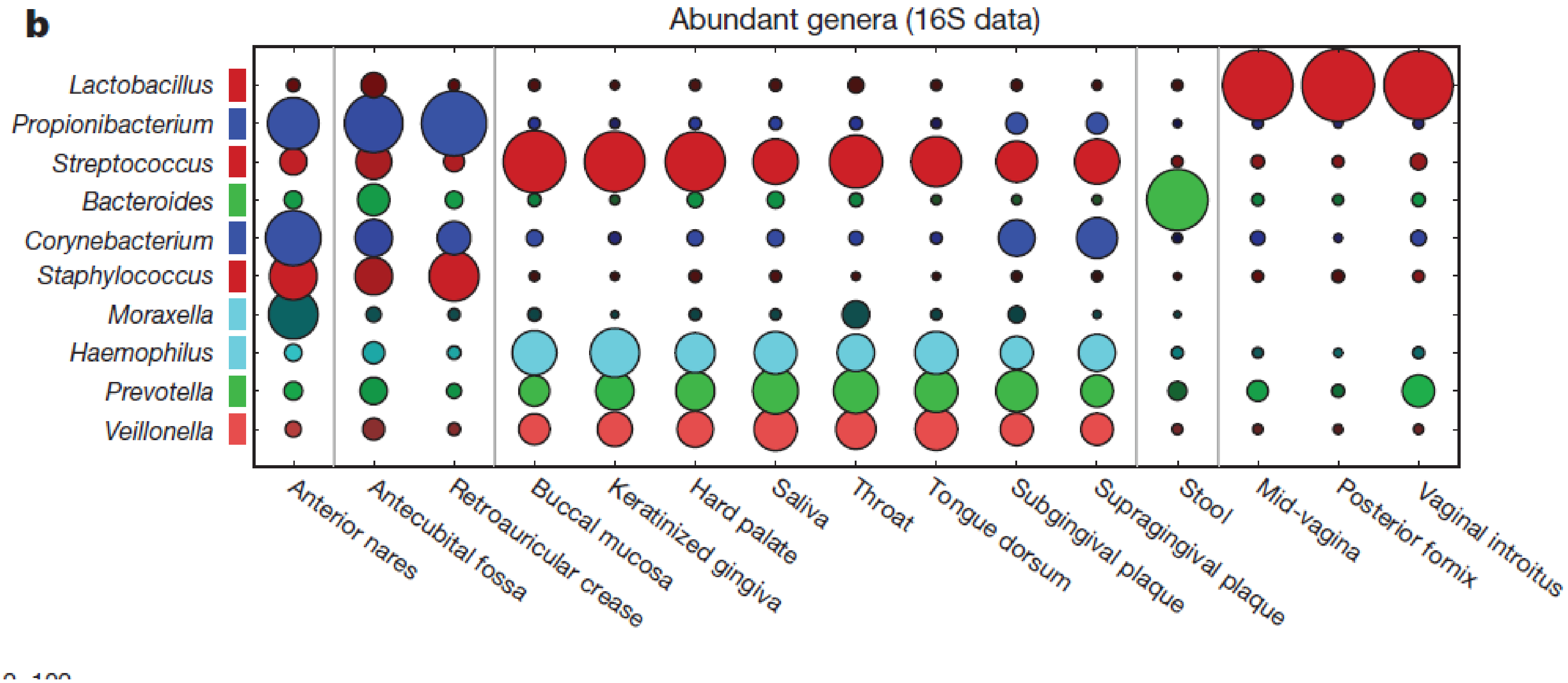
- Fast clustering of short sequences based on sequence identity



Alignment is "semi-global": most of member sequence must be covered. Fragments are accepted with default options.

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

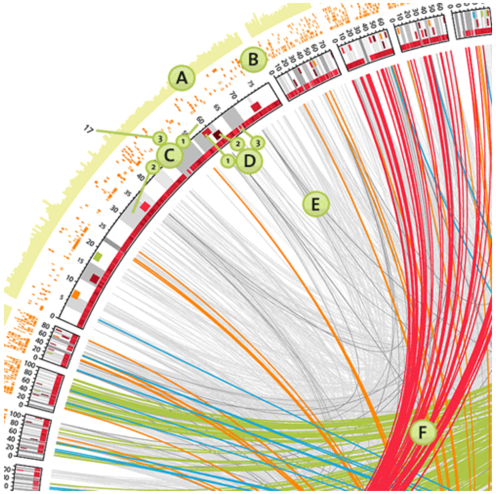
Example from the human gut microbiome



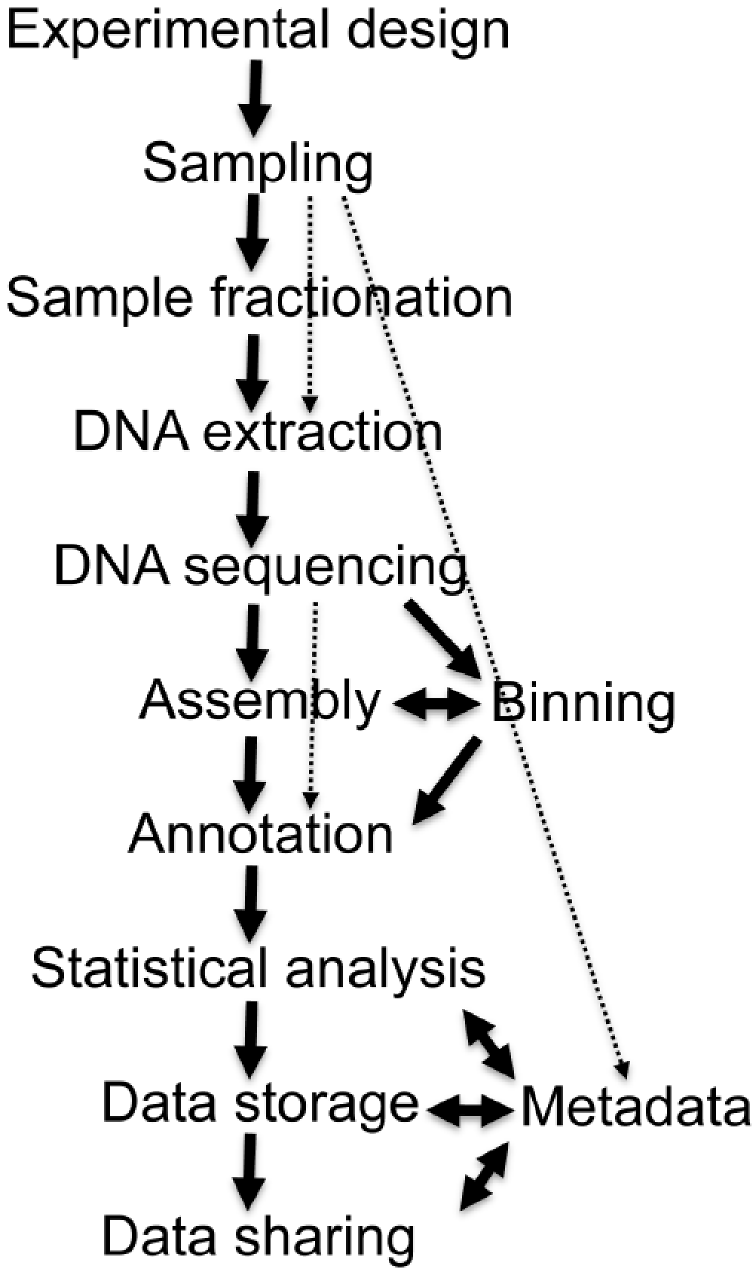
The Human Microbiome Project Consortium, Nature 486, 207–214 (14 June 2012)

What are they doing

- Shotgun metagenomics

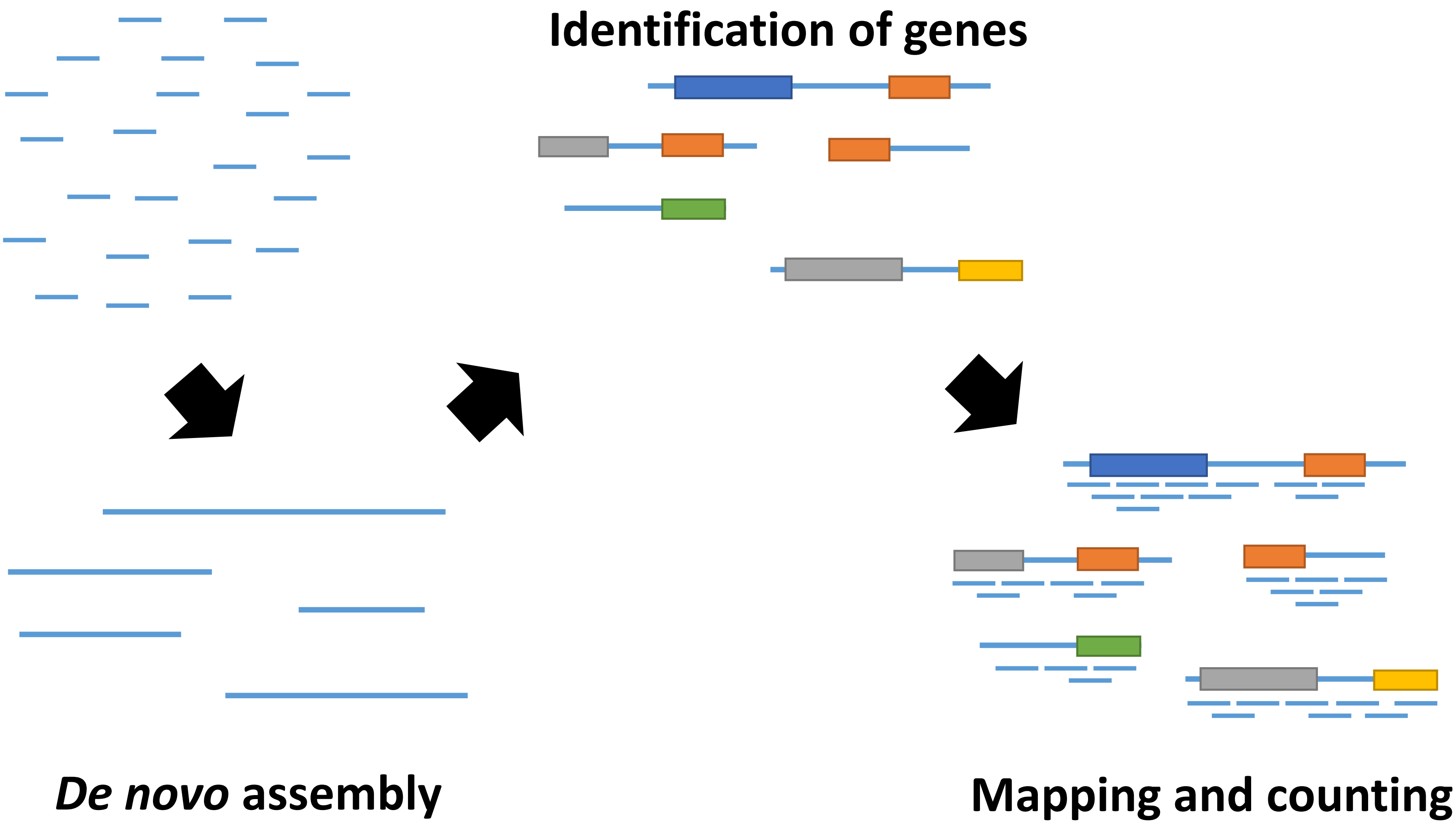


Analysis of a typical shotgun metagenomics dataset



Thomas et al. Microbial Informatics and Experimentation 2012

Binning (functional analysis)



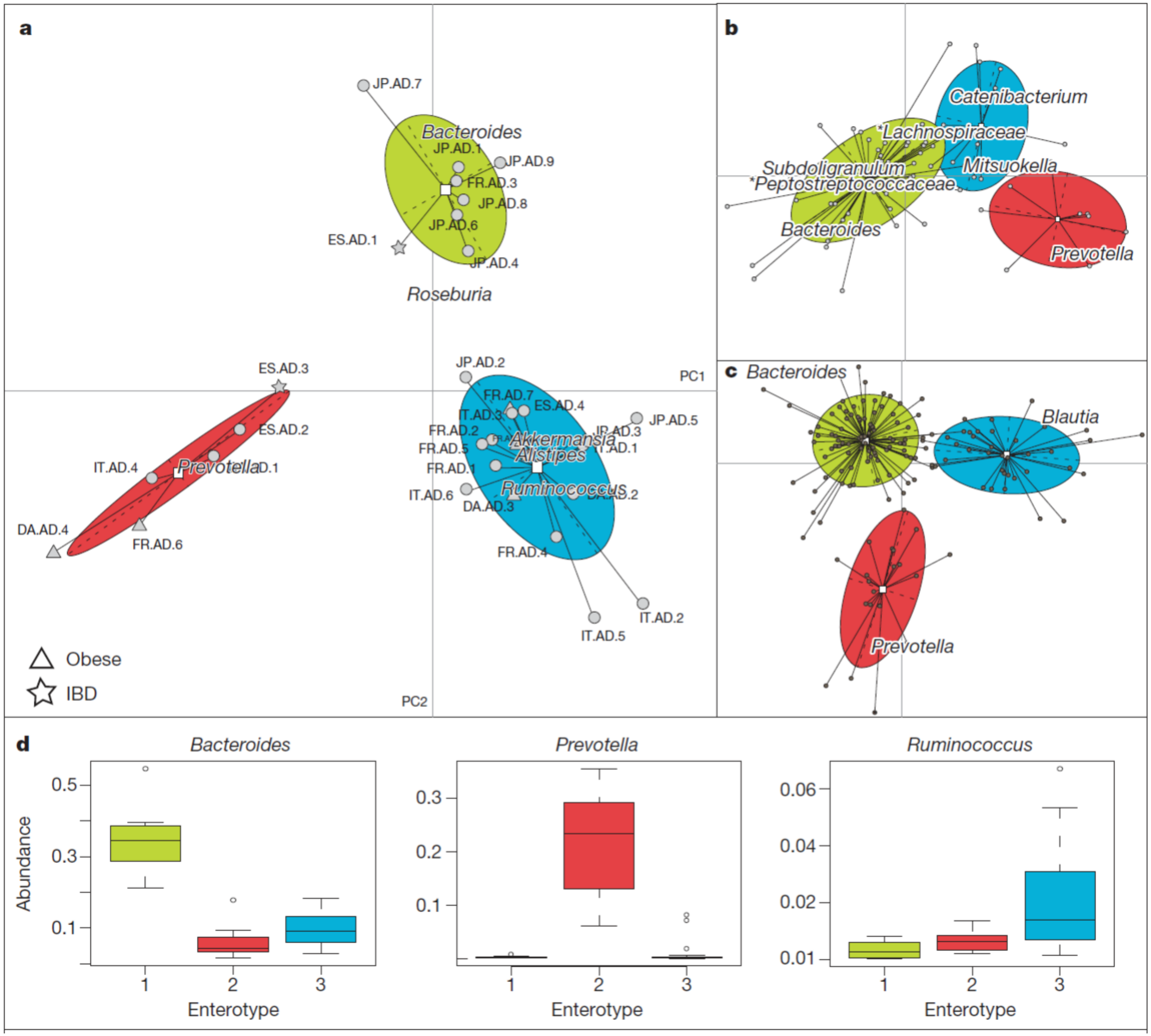
Identification of genes

***De novo* assembly**

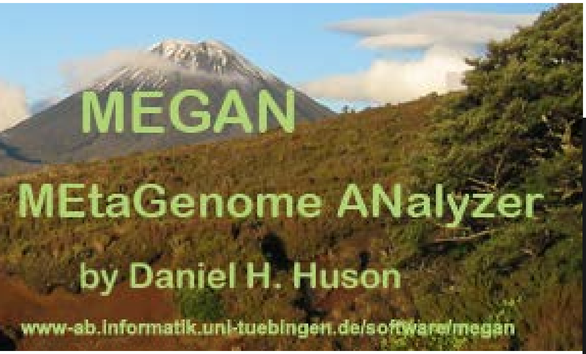
Mapping and counting

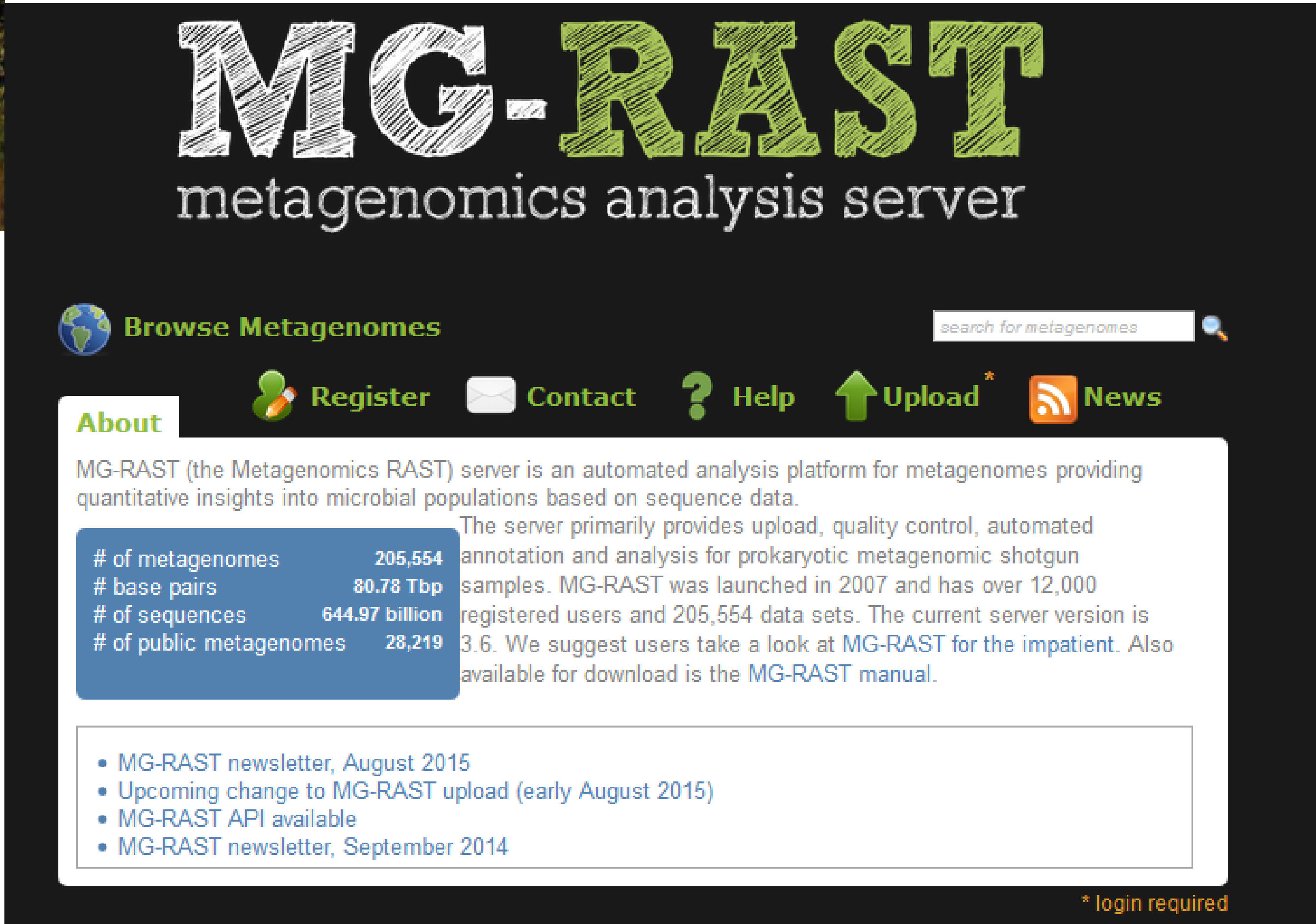
Enterotypes of the human gut

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



Metagenomics analysis software/server




MG-RAST

metagenomics analysis server

Browse Metagenomes

search for metagenomes

Register Contact Help Upload* News

About

MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

# of metagenomes	205,554
# base pairs	80.78 Tbp
# of sequences	644.97 billion
# of public metagenomes	28,219

The server primarily provides upload, quality control, automated annotation and analysis for prokaryotic metagenomic shotgun samples. MG-RAST was launched in 2007 and has over 12,000 registered users and 205,554 data sets. The current server version is 3.6. We suggest users take a look at MG-RAST for the impatient. Also available for download is the MG-RAST manual.

- MG-RAST newsletter, August 2015
- Upcoming change to MG-RAST upload (early August 2015)
- MG-RAST API available
- MG-RAST newsletter, September 2014

* login required

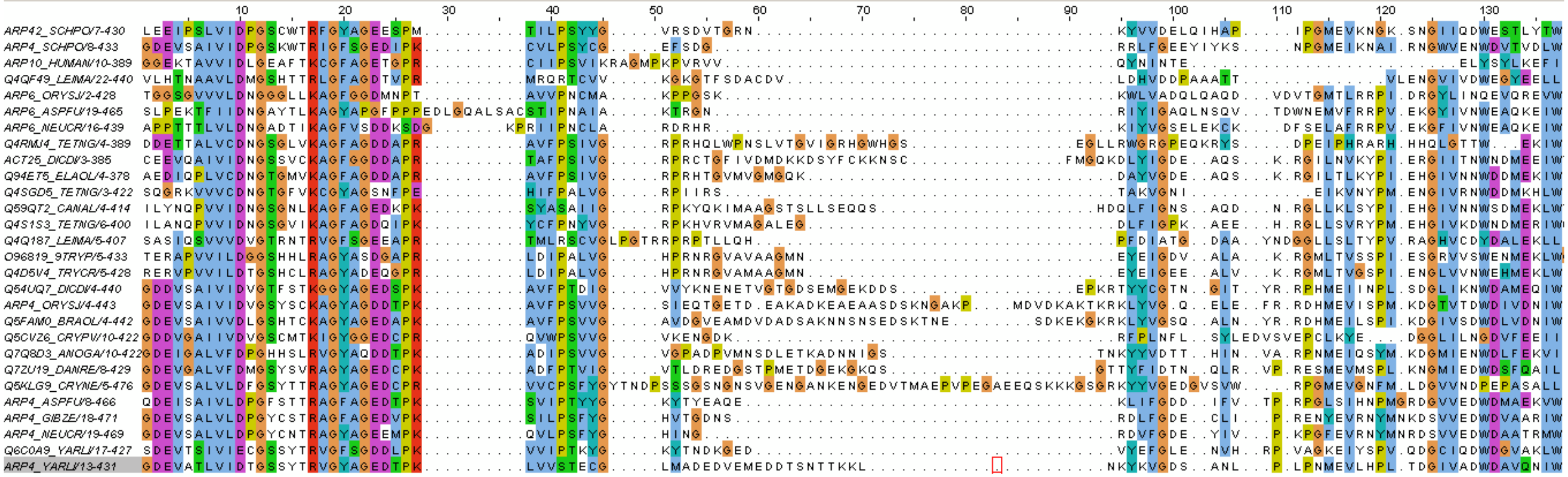
Metagenome assembly software

- 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)



```
                            10        20        30        40        50        60        70        80        90       100       110       120       130
ARP42_SCHPO/7-430    LEEIPSLVIDPGSCWTRFGYAGEESPM..........TILPSYYG.....VRSDVTGRN...................................KYVVDELQIHAP.....IPGMEVKNGK.SNGIIQDWESTLYTW
ARP4_SCHPO/8-433     GDEVSAIVIDPGSKWTRIGFSGEDIPK..........CVLPSYCG.....EFSDG.......................................RRLFGEEYIYKS.....NPGMEIKNAI.RNGWVENWDVTVDLW
ARP10_HUMAN/10-389   GGEKTAVVIDLGEAFTKCGFAGETGPR..........CIIPSVIKRAGMPKPVRVV......................................QYNINTE..........................ELYSYLKEFI
Q4QF49_LEIMA/22-440  VLHTNAAVLDMGSHTTRLGFAGDTVPR..........MRQRTCVV.....KGKGTFSDACDV................................LDHVDDPAAATT..............VLENGVIVDWEGYEELL
ARP6_ORYSJ/2-428     TGGSGVVVLDNGGGLLKAGFGGDMNPT..........AVVPNCMA.....KPPGSK......................................KWLVADQLQAQD...VDVTGMTLRRPI.DRGYLINQEVQREVW
ARP6_ASPFU/19-465    SLPEKTFIIDNGAYTLKAGYAPGFPPPEDLGQALSACSTIPNAIA.....KTRGN.......................................RIYIGAQLNSQV...TDWNEMVFRRPV.EKGYIVNWEAQKEIW
ARP6_NEUCR/16-439    APPTTTLVLDNGADTIKAGFVSDDKSDG.......KPRIIPNCLA.....RDRHR.......................................KIYVGSELEKCK....DFSELAFRRPV.EKGFIVNWEAQKEIW
Q4RMJ4_TETNG/4-389   DDETTALVCDNGSGLVKAGFAGDDAPR..........AVFPSIVG.....RPRHQLWPNSLVTGVIGRHGWHGS................EGLLRWGRGPEQKRYS.....DPEIPHRARH.HHQLGTTW...EKIW
ACT25_DICDI/3-385    CEEVQAIVIDNGSSVCKAGFGGDDAPR..........TAFPSIVG.....RPRCTGFIVDMDKKDSYFCKKNSC...............FMGQKDLYIGDE..AQS...K.RGILNVKYPI.ERGIITNWNDMEEIW
Q94ET5_ELAOL/4-378   AEDIQPLVCDNGTGMVKAGFAGDDAPR..........AVFPSIVG.....RPRHTGVMVGMGQK..............................DAYVGDE..AQS...K.RGILTLKYPI.EHGIVNNWDDMEKIW
Q4SGD5_TETNG/3-422   SQGRKVVVCDNGTGFVKCGYAGSNFPE..........HIFPALVG.....RPIIRS......................................TAKVGNI............EIKVNYPM.ENGIVRNWDDMKHLW
Q59QT2_CANAL/4-414   ILYNQPVVIDNGSGNLKAGFAGEDKPK..........SYASAIIG.....RPKYQKIMAAGSTSLLSEQQS.....................HDQLFIGNS..AQD...N.RGLLKLSYPI.EHGIVNNWSDMEKLW
Q4S1S3_TETNG/6-400   ILANQPVVIDNGSGVIKAGFAGDQIPK..........YCFPNYVG.....RPKHVRVMAGALEG..............................DLFIGPK..AEE...H.RGLLSVRYPM.EHGIVKDWNDMERIW
Q4Q187_LEIMA/5-407   SASIQSVVVDVGTRNTRVGFSGEEAPR..........TMLRSCVGLPGTRRPRPTLLQH...................................PFDIATG..DAA..YNDGGLLSLTYPV.RAGHVCDYDALEKLL
O96819_9TRYP/5-433   TERAPVVILDGGSHHLRAGYASDGAPR..........LDIPALVG.....HPRNRGVAVAAGMN..............................EYEIGDV..ALA...K.RGMLTVSSPI.ESGRVVSWENMEKLW
Q4D5V4_TRYCR/5-428   RERVPVVILDTGSHCLRAGYADEQGPR..........LDIPALVG.....HPRNRGVAMAAGMN..............................EYEIGEE..ALV...K.RGMLTVGSPI.ENGLVVNWEHMEKLW
Q54UQ7_DICDI/4-440   GDDVSAIVIDVGTFSTKGGYAGEDSPK..........AVFPTDIG.....VVYKNENETVGTGDSEMGEKDDS.................EPKRTYYCGTN..GIT..YR.RPHMEIINPL.SDGLIKNWDAMEQIW
ARP4_ORYSJ/4-443     GDEVSAIVIDVGSYSCKAGYAGDDTPK..........AVFPSVVG.....SIEQTGETD.EAKADKEAEAASDSKNGAKP....MDVDKAKTKRKLYVG.Q..ELE..FR.RDHMEVISPM.KDGTVTDWDIVDNIW
Q5FAM0_BRAOL/4-442   GDEVSAIVVDLGSHTCKAGYAGEDAPK..........AVFPSVVG.....AVDGVEAMDVDADSAKNNSNSEDSKTNE........SDKEKGKRKLYVGSQ..ALN..YR.RDHMEILSPI.KDGIVSDWDLVDNIW
Q5CVZ6_CRYPV/10-422  GDDVGAIIVDVGSCMTKIGYGGEDCPR..........QVWPSVVG.....VKENGDK.....................................RFPLNFL..SYLEDVSVEPCLKYE....DGGLILNGDVFEEII
Q7Q8D3_ANOGA/10-422  GDEIGALVFDPGHHSLRVGYAQDDTPK..........ADIPSVVG.....VGPADPVMNSDLETKADNNIGS....................TNKYYVDTT..HIN..VA.RPNMEIQSYM.KDGMIENWDLFEKVI
Q7ZU19_DANRE/8-429   GDEVGALVFDMGSYSVRAGYAGEDCPK..........ADFPTVIG.....VTLDREDGSTPMETDGEKGKQS....................GTTYFIDTN..QLR..VP.RESMEVMSPL.KNGMIEDWDSFQAIL
Q5KLG9_CRYNE/5-476   GDEVSALVLDFGSYTTRAGYAGEDCPR..........VVCPSFYGYTNDPSSSGSNGNSVGENGANKENGEDVTMAEPVPEGAEEQSKKKGSGRKYYVGEDGVSVW.....RPGMEVGNFM.LDGVVNDPEPASALL
ARP4_ASPFU/8-466     QDEISAIVLDPGFSTTRAGFAGEDTPK..........SVIPTYYG.....KYTYEAQE....................................KLIFGDD..IFV..TP.RPGLSIHNPMGRDGVVEDWDMAEKVW
ARP4_GIBZE/18-471    GDEVSALVLDPGYCSTRAGFAGEDVPK..........SILPSFYG.....HVTGDNS.....................................RDLFGDE..CLI...P.RENYEVRNYMNKDSVVEDWDVAARIW
ARP4_NEUCR/19-469    GDEVSALVLDPGYCNTRAGYAGEEMPK..........QVLPSFYG.....HING........................................RDVFGDE..YIV...P.KPGFEVRNYMNRDSVVEDWDAATRMW
Q6C0A9_YARLI/17-427  SDEVTSIVIECGSSYTRVGFSGDDLPK..........VVIPTKYG.....KYTNDKGED...................................VYEFGLE..NVH..RP.VAGKEIYSPV.QDGCIQDWDGVAKLW
ARP4_YARLI/13-431    GDEVATLVIDTGSSYTRVGYAGEDTPK..........LVVSTECG.....LMADEDVEMEDDTSNTTKKL.......................NKYKVGDS..ANL...P.LPNMEVLHPL.TDGIVADWDAVQNIW
```

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

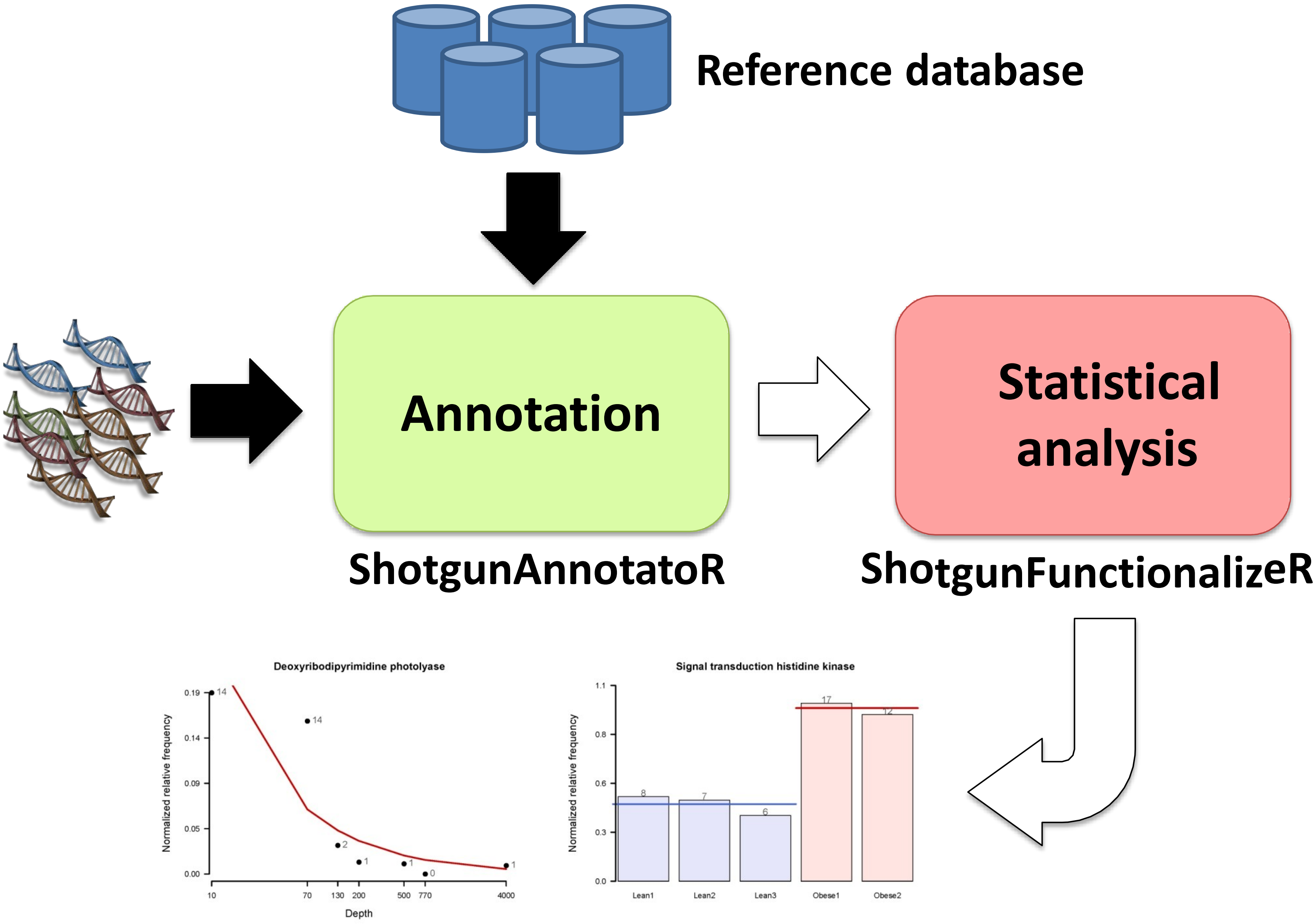
  ~4200 conserved protein families, mainly bacterial

  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

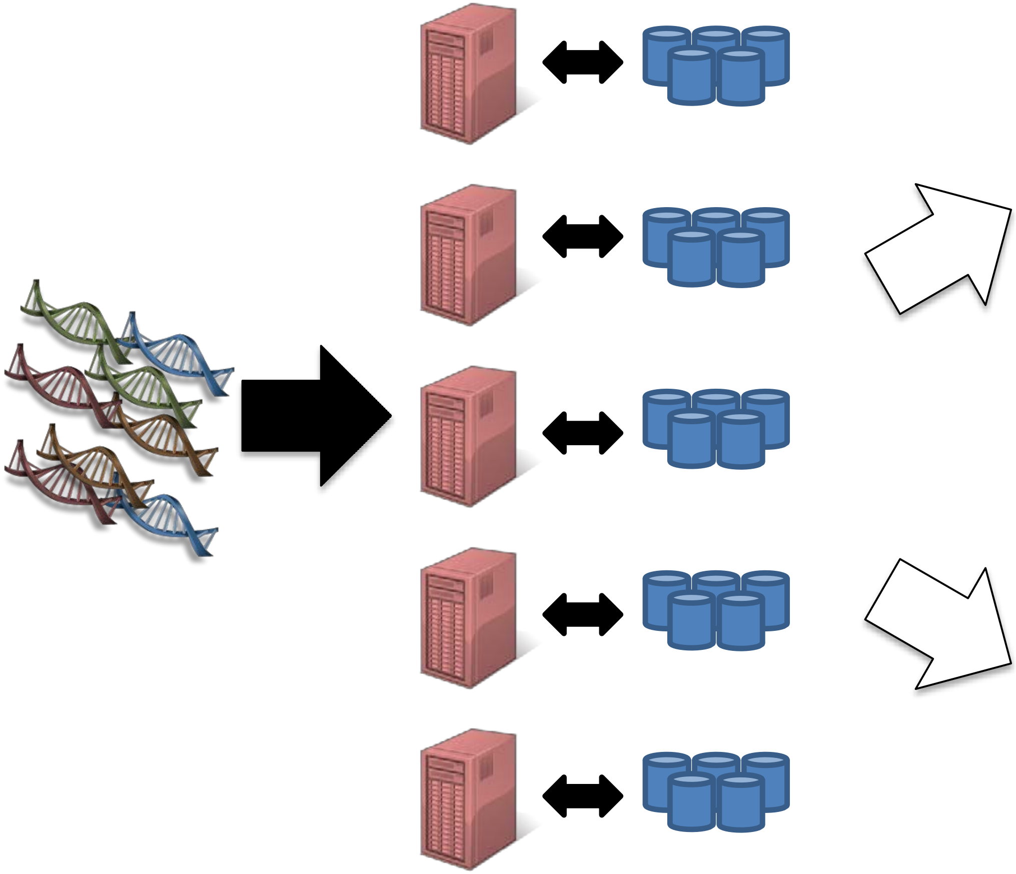
Statistical analysis

ShotgunAnnotatoR

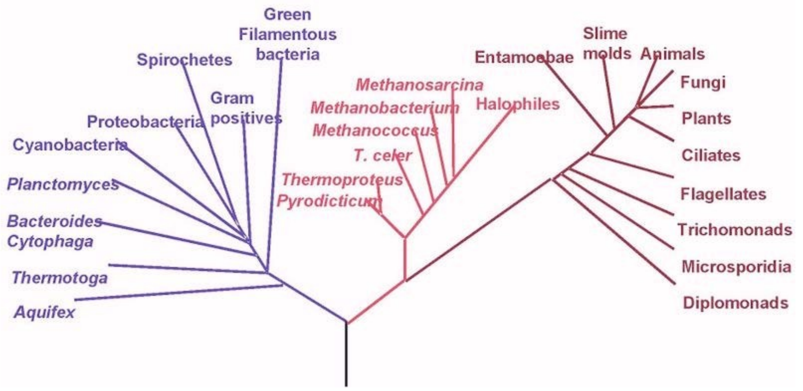
ShotgunFunctionalizeR

Deoxyribodipyrimidine photolyase
Normalized relative frequency
0.19
0.14
0.09
0.05
0.00
14
14
2
1
1
0
1
10
70
130
200
500
770
4000
Depth
Signal transduction histidine kinase
Normalized relative frequency
1.1
0.8
0.6
0.3
0.0
8
7
6
17
12
Lean1
Lean2
Lean3
Obese1
Obese2

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

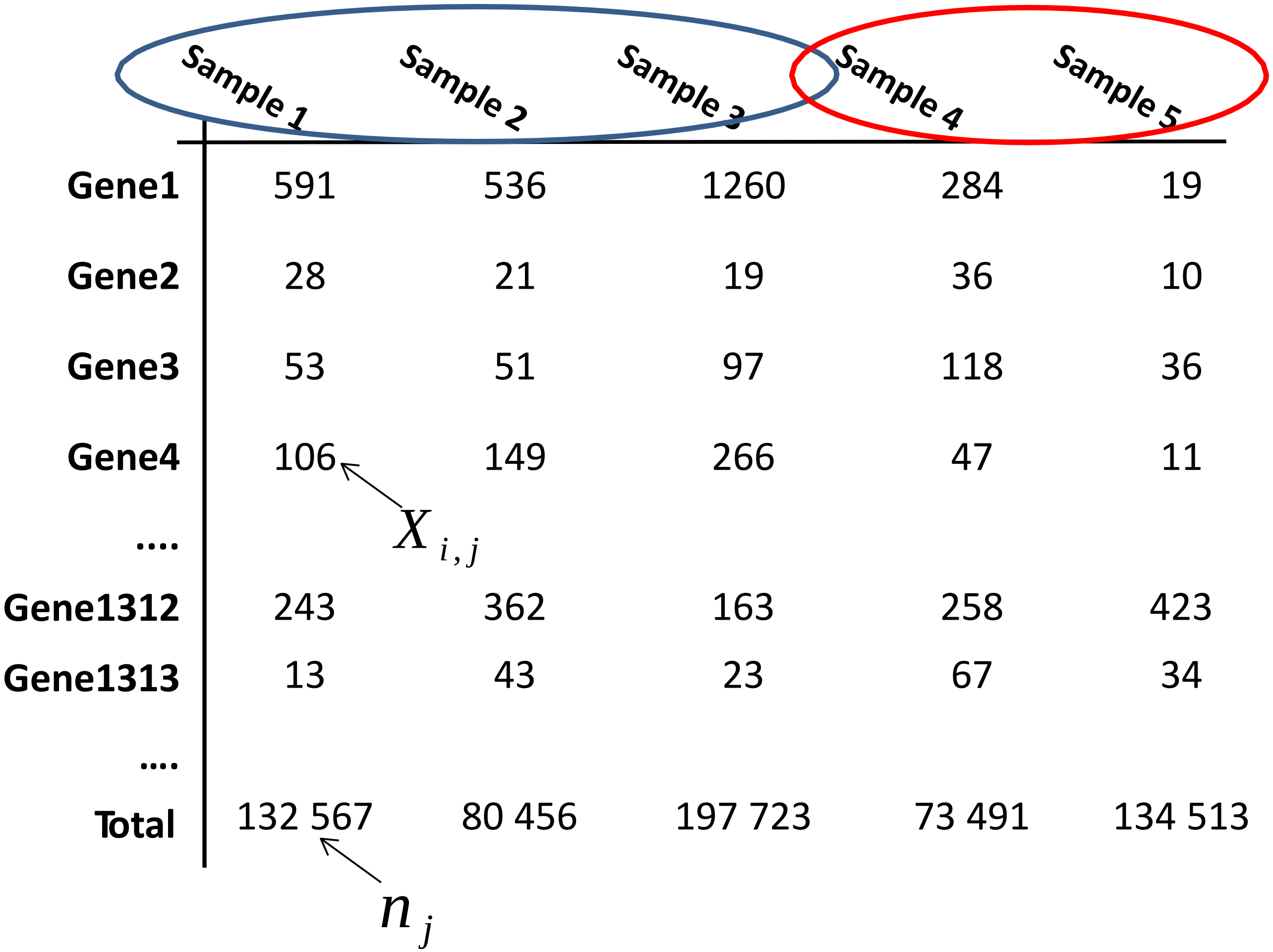
Identification of significant genes

Group 1

Group 2

	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

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

$X_{i,j}$ -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

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
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

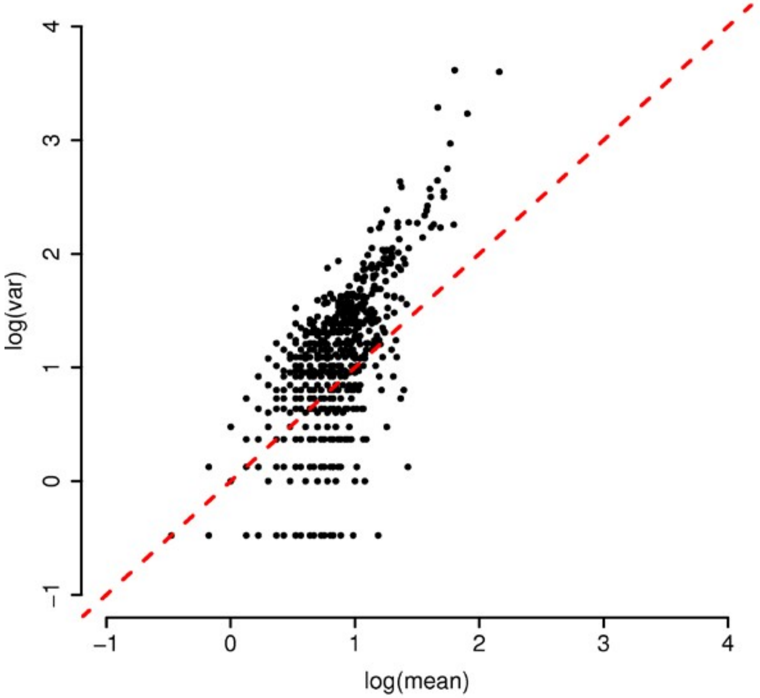
$$\log\left(\frac{\mathrm{E}\left[X_{i,j}\right]}{\mathrm{n}_{\mathrm{j}}}\right)=\underbrace{\alpha_0}_{\text{Baseline}}+\underbrace{\sum\alpha_k y_k}_{\text{Covariates (groups)}}$$

Statistical analysis

- Data from metagenomics is descrete (counts per gene/species)
- Not normally distributed
- $$X_{i,j} \sim \mathrm{Poisson}(\lambda_i)$$
  $$\mathrm{E}[X_{i,j}] = \lambda_i$$
  $$\mathrm{Var}[X_{i,j}] = \lambda_i$$

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

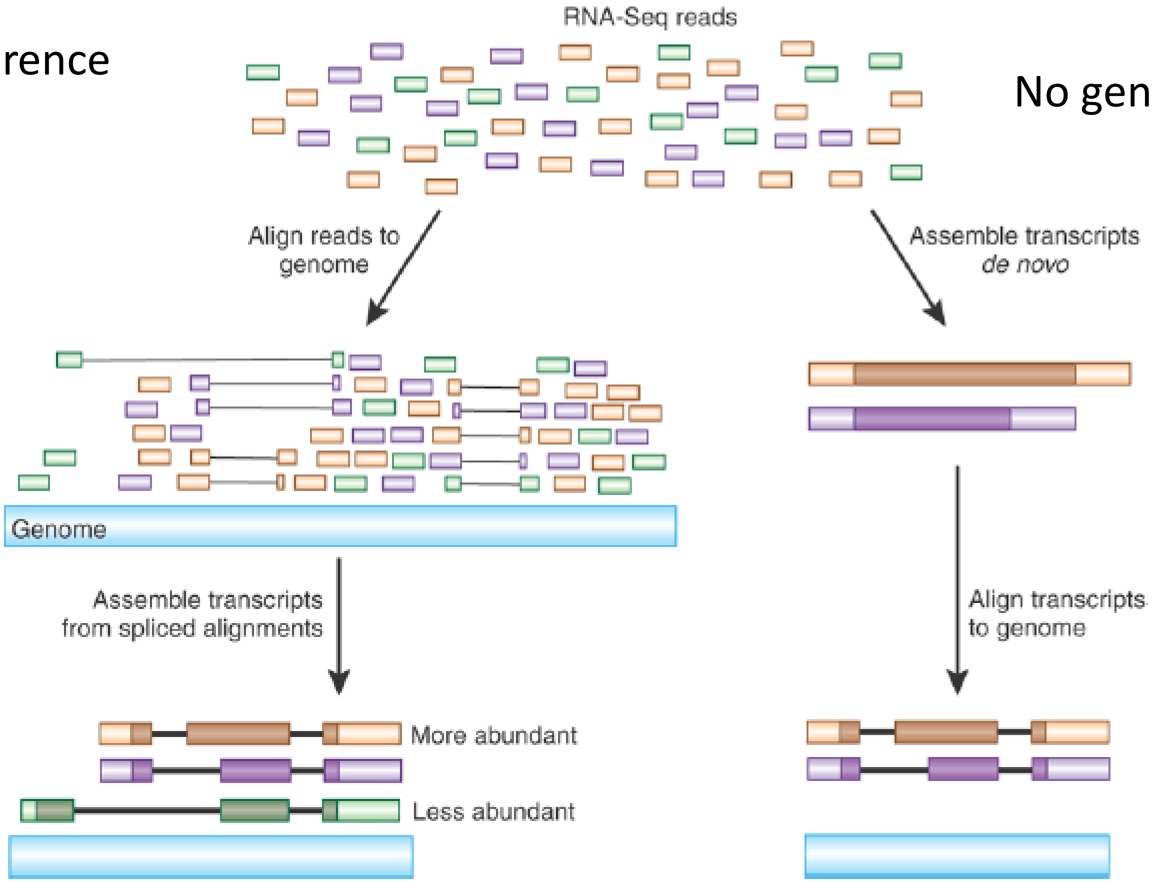
- Large-scale mRNA quantification
  - Identification of differentially expressed genes
    - Sequence all mRNA and map to reference sequence

- 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

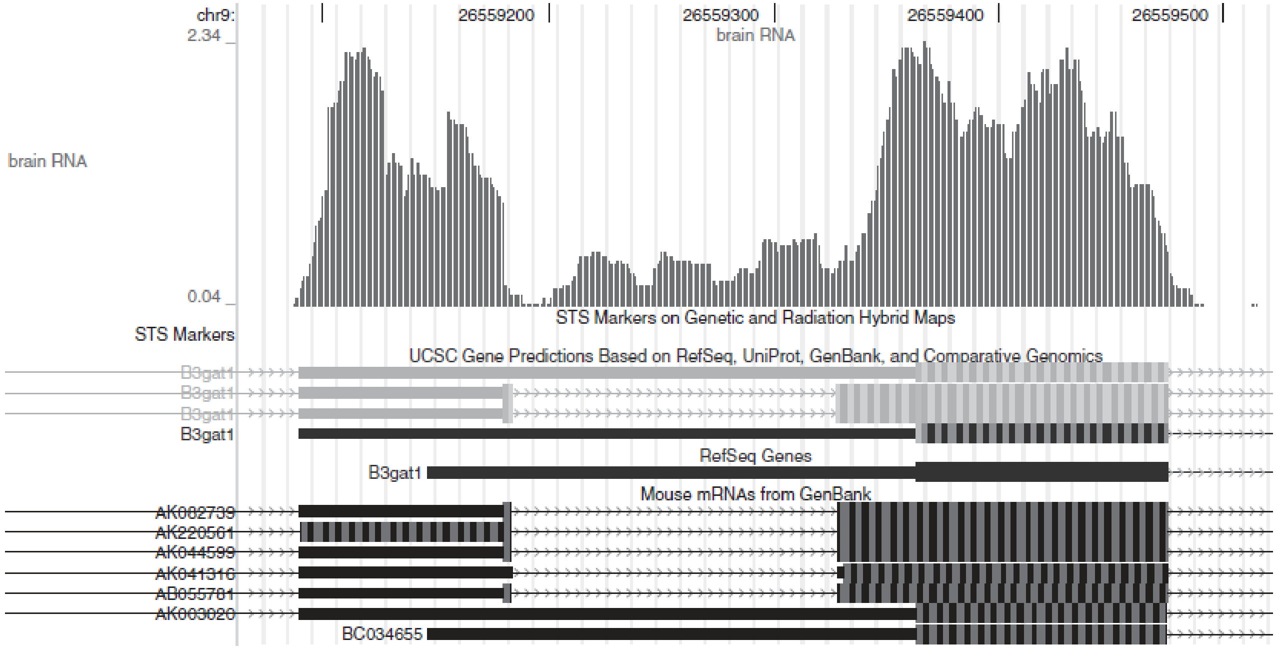
No genome


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

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

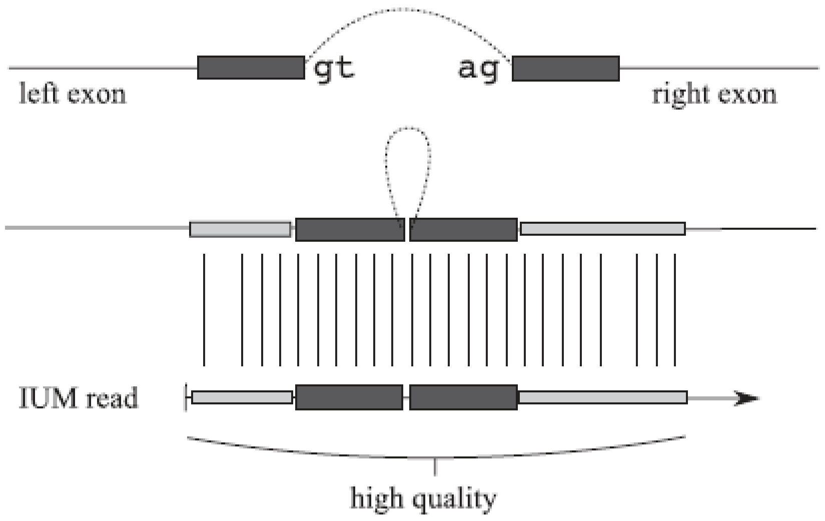
Alignment

- Using a splice-aware aligner



TopHat aligner (Trapnell et al. Bioinformatics 2009)

Alignment



TopHat aligner (Trapnell et al. Bioinformatics 2009)

De novo transcriptome assembly

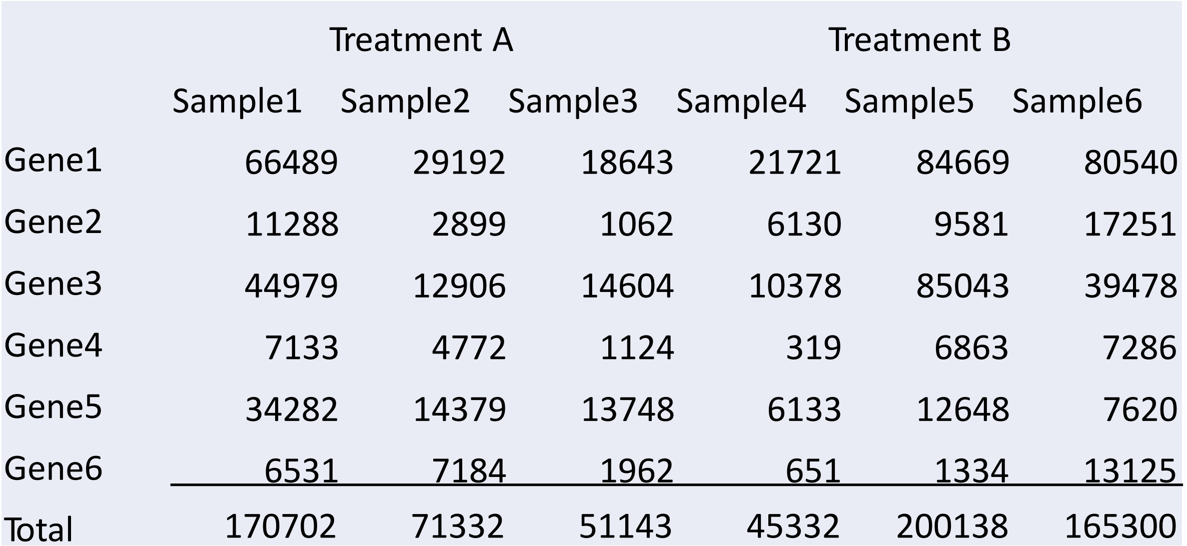
Trinity command line example:

```
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

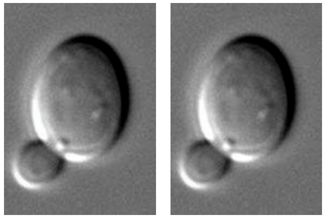
- How much should I sequence?
  - Depends on your question

  - 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

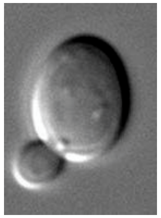
- Biological Replicates!!!

Sequencing lab

Genome sequencing of amylase producing yeast strains



UV1052 and UV715 mutation strains



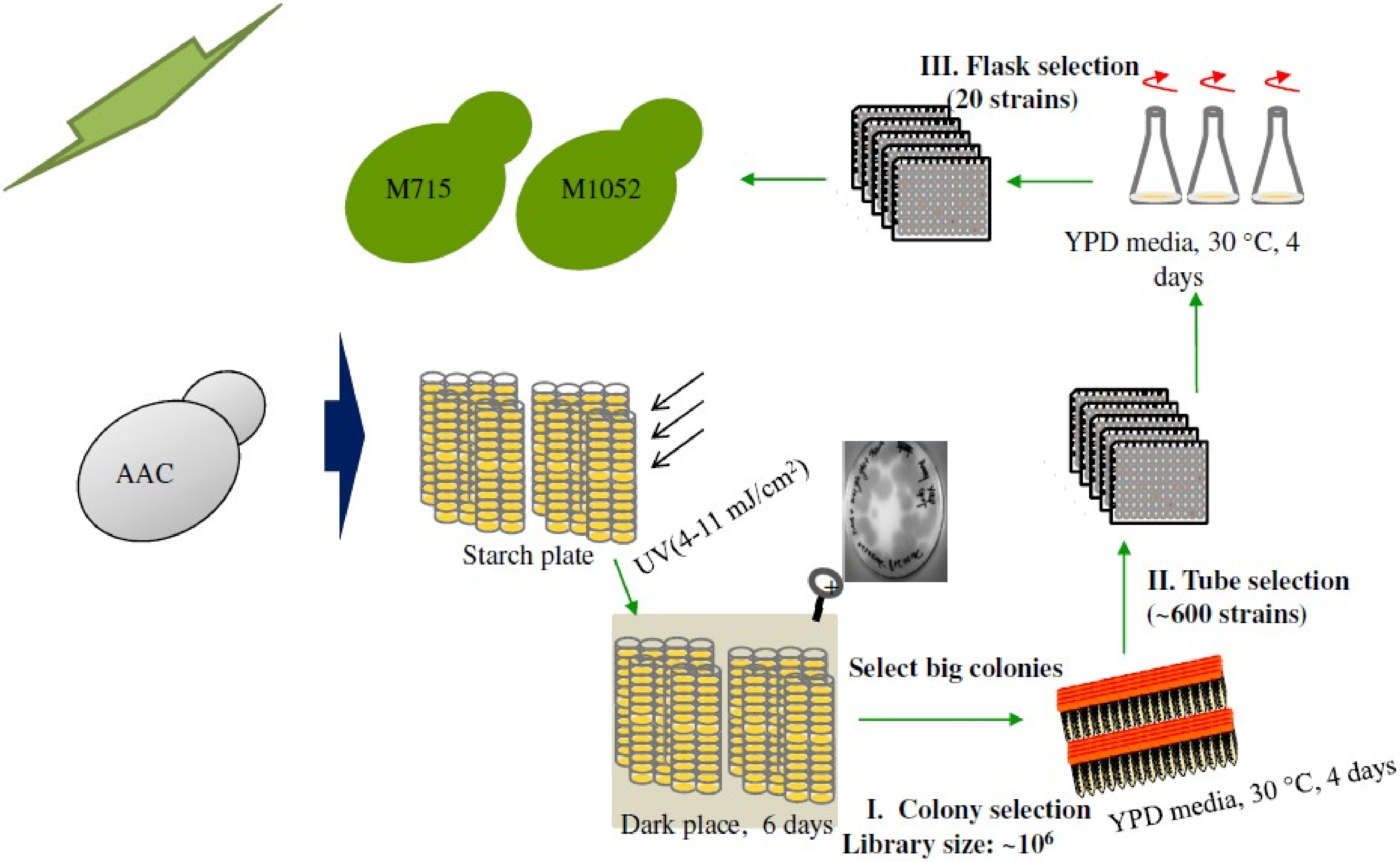
WT CEN.PK

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)

YPD media, 30 °C, 4 days

Starch plate

UV(4-11 mJ/cm^2)

Dark place, 6 days

Select big colonies

I. Colony selection
Library size: ~10^6

YPD media, 30 °C, 4 days

II. Tube selection (~600 strains)

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)