Metagenomics and RNA-seq

Tobias Österlund

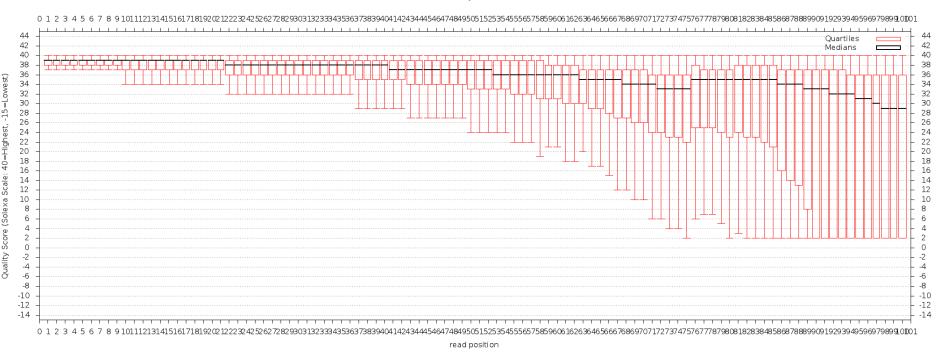


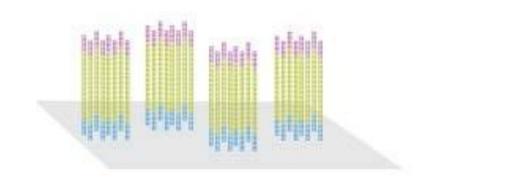


GCAACAGTTTGGCGGTAATTCAATTGT CAGTTTACGGATTCCTTGATTGGATAA TCCAGTCTGCCCCCAGGCTGCAGTTGC AAAAGAAAGAAACGACTATGAATAAAC GACTTCGGATCATTGGACTGTTTGCTG TGTTCTTTGGCCAGATGATCCACGCGC → AGACCACAGCGTTCACTTATCAGGGGC GTCTCAATGACAACGGCGCGCTGGCCA ACGGCATTTATGATTTGAAATTTTCAC TATACACCGTGGCGACCAATGGCAGTG CCTCATCGTCGCGGTCAAATGCCGCCA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

Remark on quality of Illumina sequences

Ouality Scores





Today's lecture

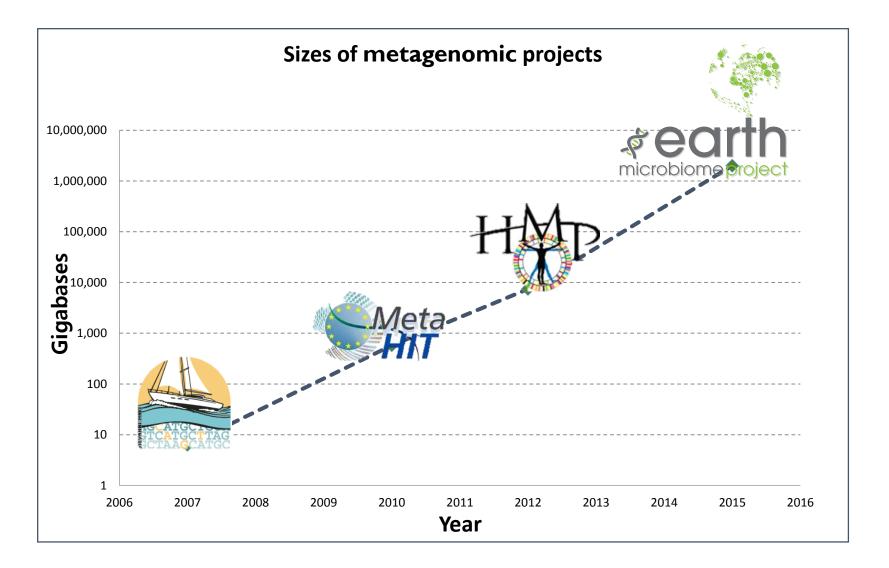
- Metagenomics analysis
 - On the species level: Who's there?
 - On the gene/functional level: What are they doing?
- RNA-seq analysis
 - Data normalization
 - Finding differentially expressed genes
- Computer exercise
 - Whole genome sequencing for variant detection

Metagenomics

• Some facts about microbes

Number of microbes on Earth	5×10 ³⁰	notire
Number of microbes in all humans	6×10 ²³	naune
Number of stars in the universe	7×10 ²¹	
Number of bacterial cells in one human gut	1014	
Number of human cells in one human	10 ¹³	
Number of bacterial genes in one human gut	3,000,000	A game catalogue MUCLIER PROLITER Res laser speaker and the human get microbiome
Number of genes in the human genome	21,000	Anti-Anti-Anti-Anti-Anti-Anti-Anti-Anti-

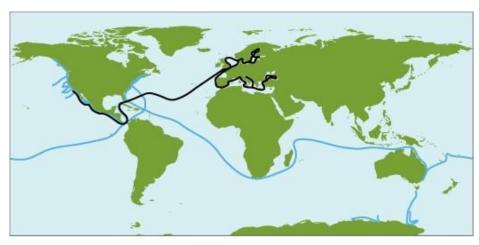
Metagenomic data revolution



The global ocean sampling

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





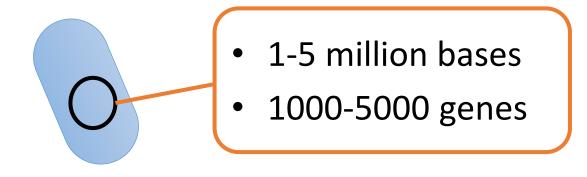
= 2003 – 2008 Routes = 2009 – 2010 Route

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



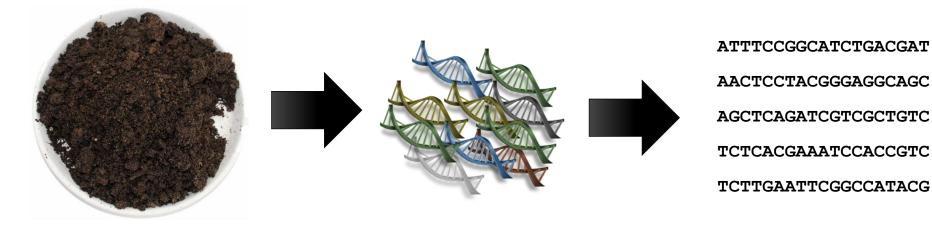


1 gram of soil

- 10 000 species
- 100 million cells
- DNA: 100 terabases (10¹⁴)

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

Metagenomics



Sample with microorganisms

DNA

Metagenome

Metagenomics

• Metagenomics is used to study the unculturable organisms and viruses

~50% of human gut bacteria are unculturable

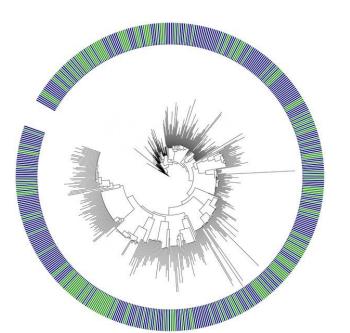
– <1% of environmental bacteria are unculturable</p>

- Metagenomes are highly fragmented and undersampled
- The majority of DNA found in metagenomes is usually very hard to annotate

Two types of questions

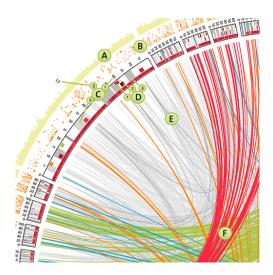
Who's there?

- 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?

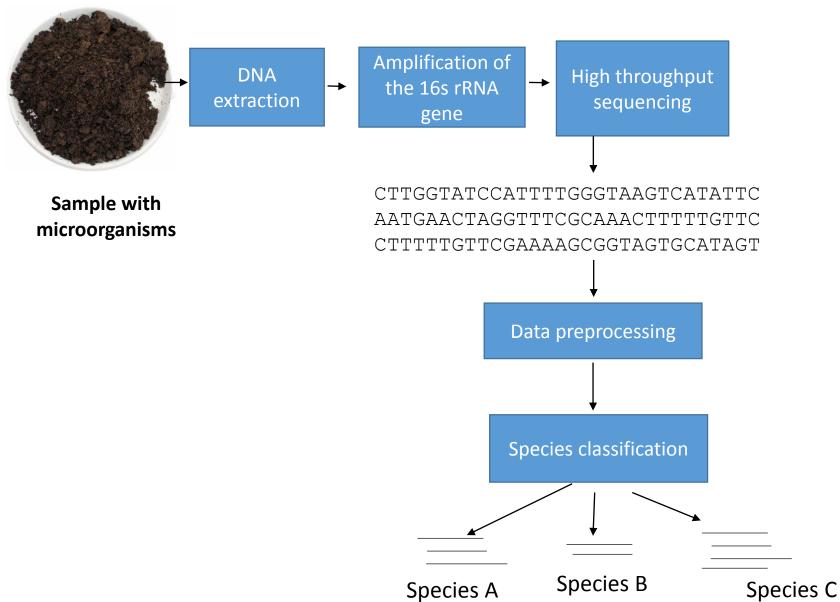


- Amplicon sequencing of phylogenetic marker genes
- Shotgun sequencing
 - Mapping reads to species with known genomes
 - Binning of reads

Species identification using marker genes

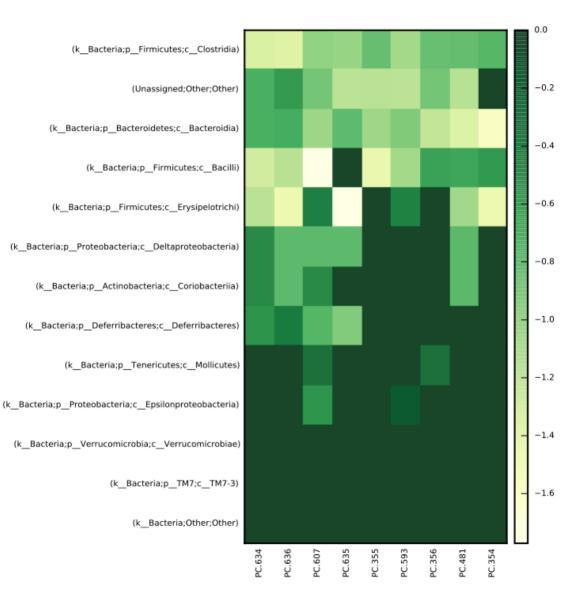
- Prokaryots:
 - 16s rRNA gene
- Eukaryots:
 - 18s rRNA gene
- Can be amplified using amplicon sequencing
- Sequences mapped to known species using BLAST
- Operational taxonomic unit (OTU):
 - 97% sequence similarity for the 16s rRNA gene
 - Cluster based on sequence similarity using UCLUST

16s sequencing



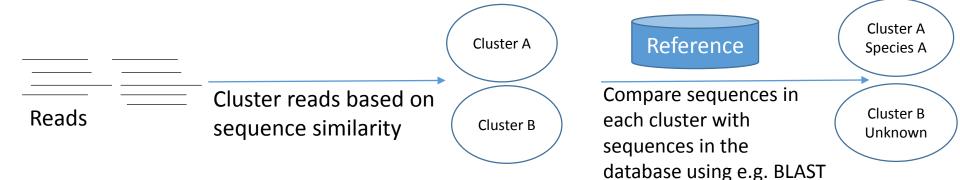
Species abundance

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

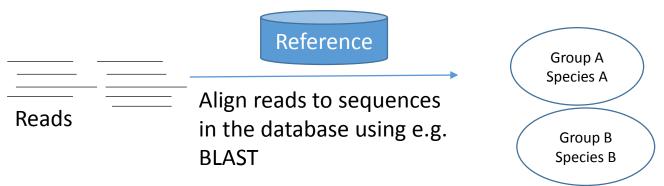


OTU picking

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

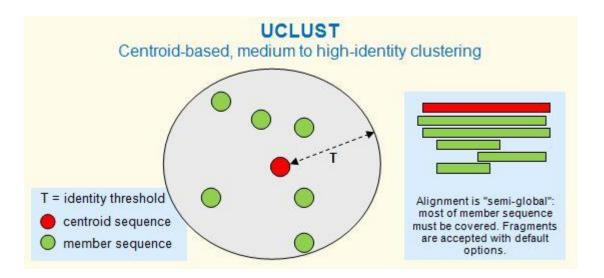


Closed reference OTU picking:



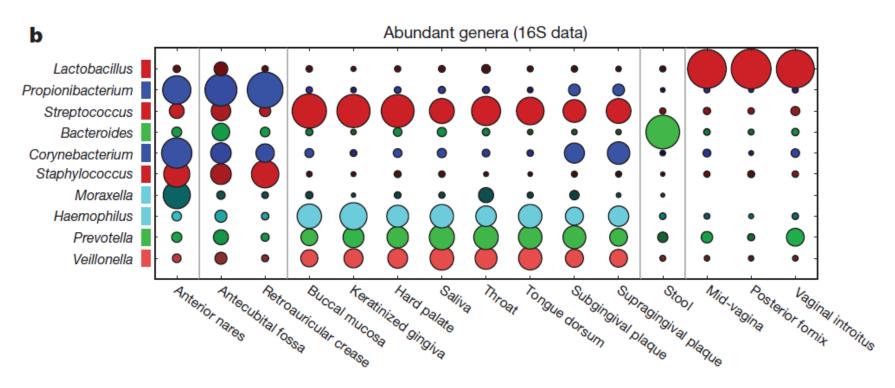
Uclust

 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

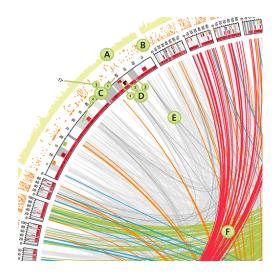


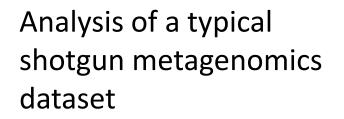
0 100

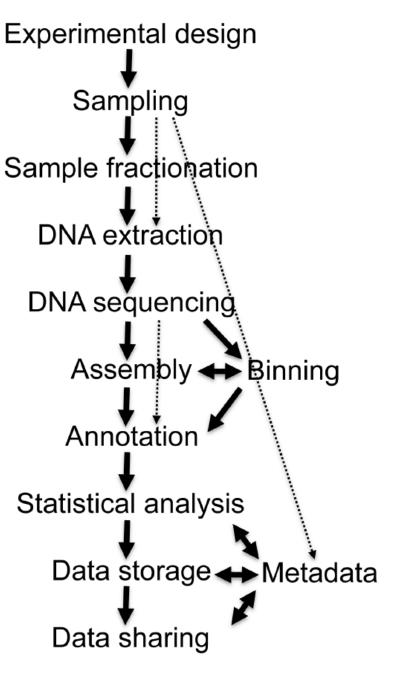
The Human Microbiome Project Consortium, Nature 486, 207–214 (14 June 2012)

What are they doing

• Shotgun metagenomics

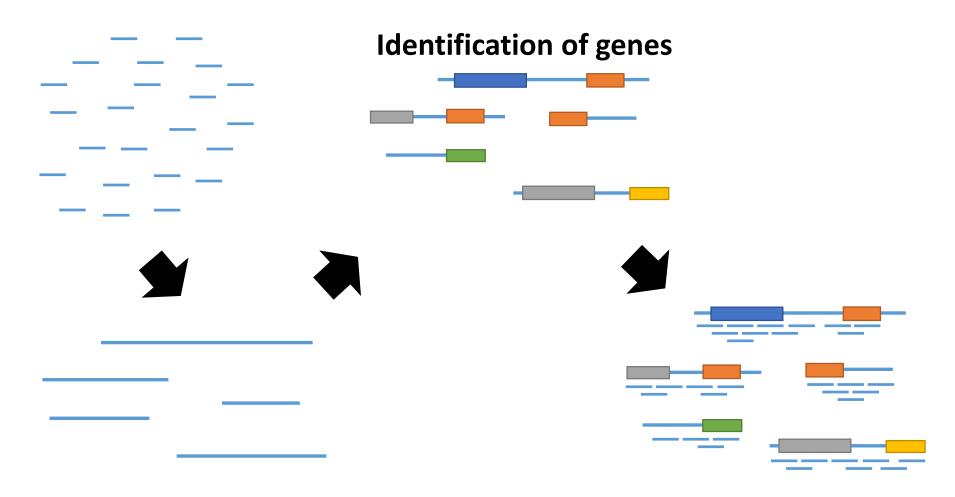






Thomas et al. Microbial Informatics and Experimentation 2012

Binning (functional analysis)

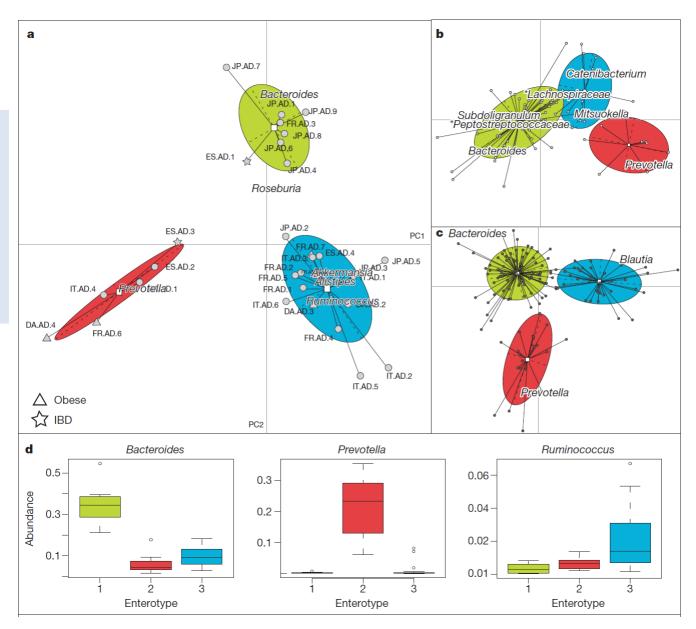


De novo assembly

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





metagenomics analysis server



# of metagenomes 205,554	annotat
# base pairs 80.78 Tbp	sample
# of sequences 644.97 billion	register
# of public metagenomes 28,219	3.6. We
	availabl

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

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
RP42_SCHP0/7-430	LE <mark>EIPS</mark> LVI <mark>DPGS</mark> CWT	R F G Y A G E E <mark>S</mark> P M	<u>t</u>	. TIL <mark>PSYYG</mark>	V <u>R</u>	SDVT <mark>G</mark> RN				K <mark>YVVD</mark> ELQIH <mark>AP</mark>	I <mark>P</mark>	GMEVKNGK.	. SN <mark>gii</mark> qd <mark>west</mark> l'
RP4_SCHP0/8-433	<mark>gdev</mark> salvi <mark>dpgs</mark> kwt	RIGF <mark>S</mark> GEDIPK		. CVL <mark>PS</mark> YC <mark>g</mark>	<u>.</u> . <u>.</u> E F	S D <mark>G</mark>				R R L F <mark>G</mark> E E Y I Y K S	N <mark>P</mark>	GMEIKNAI.	. R N <mark>G W V</mark> E N W <mark>D V T</mark> V
RP10_HUMAN/10-389	<mark>gge</mark> kt <mark>avvid</mark> lgeaft	K C G F A G E T G P R		. CIT <mark>ps</mark> vik	RA <mark>GMP</mark> K <mark>P</mark>	VRVV				2 <mark>Y N I </mark> N T E			EL <mark>y</mark> s <mark>yl</mark> k
4QF49_LEIMA/22-440	VLH <mark>T</mark> NAAVL <mark>D</mark> M <mark>gs</mark> htt	R L G F A G D T V P R		. MRQR <mark>T</mark> CVV	K <mark>g</mark>	K <mark>g</mark> tfsdacdv.				L D <mark>H V</mark> D D <mark>P</mark> A A A <mark>T</mark> T		<mark>. v</mark> t	. EN <mark>gvi</mark> vdwe <mark>gye</mark>
RP6_ORYSJ/2-428	T <mark>gg<mark>s</mark>gvvvl<mark>d</mark>ngggll</mark>	K <mark>ag f g g d</mark> m n <mark>p</mark> t		. AVV <mark>P</mark> NCMA						K <mark>WLV</mark> ADQLQAQD	VDVT	GMTLRR <mark>P</mark> I.	. DR <mark>gyl</mark> inqe <mark>v</mark> qr
RP6_ASPFU/19-465	SLPEK <mark>TFIID</mark> NGAYTL	KAGYAPGFPPP	EDL <mark>G</mark> QALSA	C <mark>STIP</mark> NAIA	K <mark>T</mark>	R <mark>G</mark> N				R I <mark>Y I G</mark> AQLNSQV	T D W N	EMVFRR <mark>P</mark> V.	. EK <mark>gyi</mark> vnweaqk
RP6_NEUCR/16-439	A <mark>ppt</mark> t <mark>tlvld</mark> ngadti	K <mark>ag f vsddk</mark> sd	G K	PRIIPNCLA						KIYV <mark>g</mark> selek <mark>c</mark> k	DFS	ELAFRR <mark>P</mark> V.	. EK <mark>gfivnw</mark> eaqk
4RMJ4_TETNG/4-389	DDETTALVCDNGSGLV	KAGFAGDDAPR		. AVE <mark>PS</mark> IV <mark>G</mark>	B <mark>P</mark>	RHQLW <mark>P</mark> NSLV1	GVIGRHGWHGS.		E <mark>g</mark> ll	R <mark>WG</mark> R <mark>GP</mark> EQKR <mark>Y</mark> S	D <mark>P</mark>	EIPHRARH.	. H H Q L G T T W E
CT25_D/CD//3-385	CEEVQAIVIDNGSSVC	K A G F G G D D A P R		. TAF <mark>PS</mark> IV <mark>g</mark>	R <mark>P</mark>	RCTGFIVDMDP	KDSYFCKKNSC.		F М <mark>с</mark> Q К	DLYIGDE AQS	K . R <mark>g</mark>	ILNVKY <mark>P</mark> I.	. ER <mark>GII</mark> TNWND <mark>ME</mark>
94ET5_ELAOL/4-378	AEDIQPLVCDNGTGMV	K A G F A G D D A P R		. AVE <mark>PS</mark> IV <mark>G</mark>	R <mark>P</mark>	RHT <mark>G</mark> VMV <mark>G</mark> MGC	ΩК			D <mark>ayvg</mark> deaqs	K . R <mark>g</mark>	ILTLKY <mark>P</mark> I.	. EH <mark>GIVNNWD</mark> D <mark>ME</mark>
4SGD5_TETNG/3-422	SQGRK <mark>VVVCD</mark> NGTGFV	KCGYAGSNFPE		. HIFPALVG	R <mark>P</mark>	I I R S				T <mark>akvg</mark> ni		EIKVNY <mark>P</mark> M.	. EN <mark>givrnwd</mark> dmk
259QT2_CANAL/4-414	ILYNQ <mark>PVVID</mark> NG SGNL	K <mark>agfagedkpk</mark>		. SYASAIIG	R <mark>P</mark>	KYQKIMAA <mark>g</mark> st	SLLSEQQS		нр	LFIGNS AQD	N . R <mark>G</mark>	LLKLSY <mark>P</mark> I.	. EH <mark>giv</mark> nnwsdm <mark>e</mark>
45153_TETNG/6-400	ILANQ <mark>PVVID</mark> N <mark>gSg</mark> VI	KAGFAGDQIPK		. YCFPNYVG	R <mark>P</mark>	KHVRVMA <mark>g</mark> ale	G			D <mark>lfigp</mark> k Aee	H . R <mark>G</mark>	LLSVRY <mark>P</mark> M.	. EH <mark>giv</mark> kd <mark>w</mark> nd <mark>me</mark>
	SASIQ <mark>SVVVD</mark> V <mark>gt</mark> rnt			. TML R S C V G	L <mark>PG</mark> TRR <mark>P</mark>	R <mark>P</mark> TLLQH				PFDIAT <mark>GDA</mark> A	YND <mark>gg</mark>	LLSLTY <mark>P</mark> V.	. RA <mark>ghvcd<mark>yd</mark>ale</mark>
096819_97RYP/5-433	TER <mark>APVVILDGGS</mark> HHL	RAGYASDGAPR		. LDIPALVG	H <mark>P</mark>	R N R <mark>g</mark> v a v a a <mark>g</mark> n	4N			EYEIGDV ALA	K. R <mark>g</mark>	MLTVSS <mark>P</mark> I.	. ES <mark>grv</mark> vswenme
4D5V4 TRYCR/5-428	RER <mark>VPVVILD</mark> TGSHCL	RAGYADEQGPR		. LDIPALVG	H <mark>P</mark>	R N R <mark>g</mark> v a ma a <mark>g</mark> n	4N			E <mark>YEIG</mark> EEALV	K . R <mark>g</mark>	MLTVGS <mark>P</mark> I.	. EN <mark>glvvnwehme</mark>
254UQ7 DICDV4-440	<mark>gddvsaivid</mark> v <mark>gt</mark> fst	KGGYAGEDSPK		. AVE <mark>PT</mark> DIG	v <mark>v</mark>	YKNENETV <mark>g</mark> t <mark>g</mark>	DSEM <mark>g</mark> ekdds		E <mark>p</mark> kr	Τ <mark>ΥΥ</mark> С <mark></mark> ТΝ	YR . R <mark>P</mark>	HMETINPL.	. SD <mark>GLIKNWDAME</mark>
RP4_ORYSJ/4-443	GDEVSAIVIDVGS YSC	KAGYAGDDTPK		. AVF <mark>PS</mark> VV <mark>G</mark>	s <mark>I</mark>	EQT <mark>G</mark> ETD.EAK	ADKEAEAASDSK	N <mark>gak</mark> pMi					. KD <mark>gtv</mark> tdw <mark>divd</mark>
SFAMO_BRAOL/4-442	GDEVSAIVVDLG <mark>S</mark> HTC	KAGYAGEDAPK		. AVF <mark>PS</mark> VV <mark>G</mark>	A <mark>v</mark>	DGVEAMDVDAD	SAKNNSNSEDSK	TNE	. SDKEK <mark>g</mark> kr	K <mark>lyvg</mark> sqaln	YR . RD	HMEILS <mark>P</mark> I.	. KD <mark>givsdwdlvd</mark>
25CVZ6 CRYPV/10-422	GDDVGALIVDVGSCMT	KIG <mark>Y</mark> GGEDCPR		. Q VW <mark>P S</mark> V V G	vk	EN <mark>G</mark> DK				R <mark>FPL</mark> NFLS <mark>Y</mark> L	E D V S V E <mark>P</mark>	CLKYE	. D <mark>gglilngdvfe</mark>
7Q8D3_ANOGA/10-42	2 <mark>g de ig</mark> alvf <mark>dpg</mark> hhsl	RVGYAQDDTPK		. AD I PSVVG	v <mark>g</mark>	PADPVMNSDLE	TKADNNI <mark>g</mark> s		. T N	K <mark>YY</mark> VDTTHIN	VA . R <mark>P</mark>	NMEIQS <mark>Y</mark> M.	. KD <mark>g</mark> mienw <mark>dlfe</mark>
772019_DANRE/8-429	<mark>g d e vg</mark> al vf <mark>d mg s</mark> ysv	RAGYAGEDCPK		. ADF <mark>PT</mark> VIG	v <mark>t</mark>	L D R E D <mark>G</mark> S T <mark>P</mark> M E	TD <mark>gekg</mark> kus		<mark>6</mark> т	T <mark>YFI</mark> DTNQLR	V <mark>P</mark> . RE	SMEVMS <mark>P</mark> L.	. KN <mark>g</mark> miedw <mark>dsfq</mark>
5KLG9 CRYNE/5-476	GDEVSALVLDFG <mark>S</mark> YTT	RAGYAGEDCPR		. VVC <mark>PS</mark> FYG	YTND <mark>P</mark> S <mark>S</mark>	S <mark>g</mark> s ng n s v g e n	I <mark>g</mark> anken <mark>g</mark> edvtn	IAE <mark>PVPEG</mark> AEEC	2 S K K K <mark>g</mark> s <mark>g</mark> r	K <mark>YYVG</mark> ED <mark>G</mark> VSVW	R <mark>P</mark>	GMEVGNEM.	LD <mark>GVV</mark> ND <mark>PEPA</mark> S.
RP4 ASPFU/8-466	QDEISAIVLDPGFSTT	RAGFAGEDTPK		. SVIPTYYG	<mark>.</mark> κ <mark>γ</mark>	TYEAQE	<u> </u>		.	K <mark>lifg</mark> dd F v	T <mark>P</mark> . R P	GLSIHNPMG	RDGVVEDWDMAE
RP4_GIBZE/18-471	GDEVSALVLDPGYCST	RAGFAGEDVPK		. SIL <mark>P</mark> SFYG	н 🗸	T <mark>G</mark> DNS				RDLF <mark>G</mark> DECLI			NKDSVVEDW <mark>D</mark> VAA
	GDEVSALVLDPGYCNT	RAGYAGEEMPK		QVLPSFYG	н I	N <mark>G</mark>				RD <mark>VFG</mark> DEYIV	<mark>P</mark> . K <mark>P</mark>	GFEVRNYMN	NRDSVVEDW <mark>D</mark> AAT
	SDEVTSIVIECOSSYT	RVGFSGDDLPK		. VVIPTKYG		TNDK <mark>g</mark> ed				VYEFGLE NVH			QD <mark>GCIQDWDGV</mark> A
_	GDEVATLVIDTGS SYT	RVGYAGEDTPK		LVVSTECG	L M	ADEDVEMEDDI	SNTTKKL		N	KYKVGDS. ANL			TDGIVADWDAVQ

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.

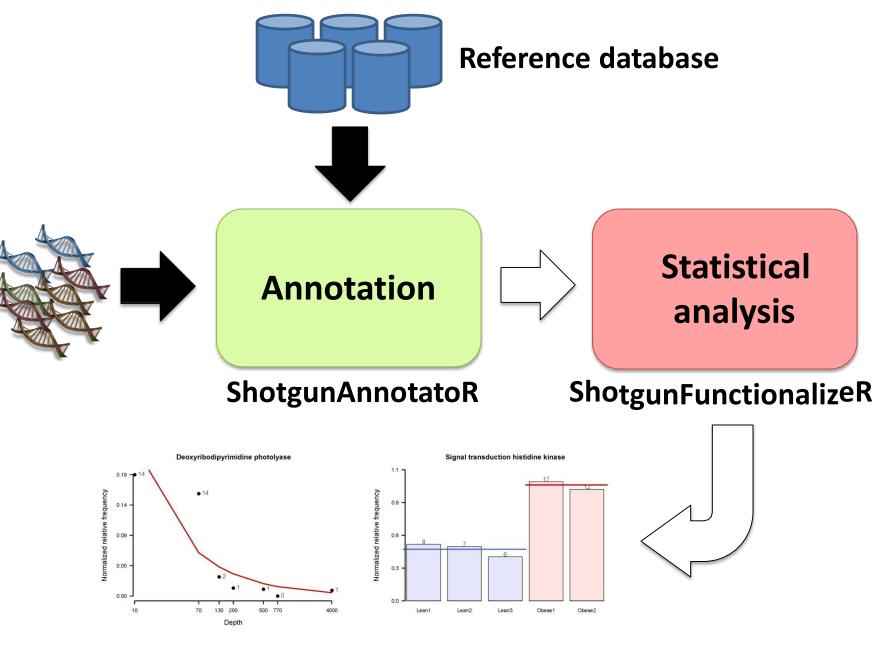
• TIGRFAM

~4200 conserved protein families, mainly bacterial Identification using HMM

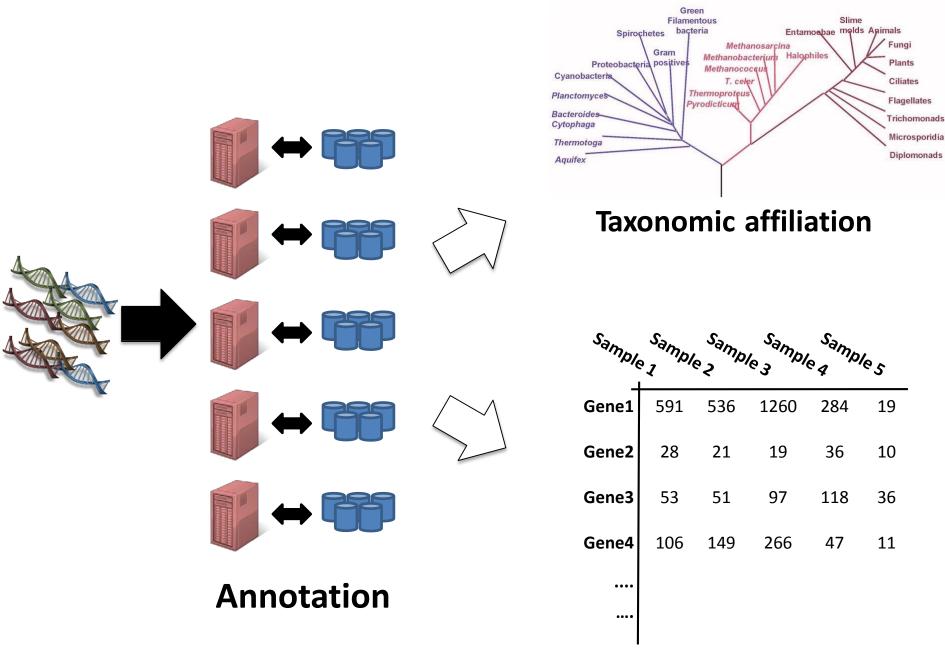
- COG
 - Clusters of orthologous groups, mainly bacterial
 - Identification using position specific weight matrices (PSWM)

Other functional annotation

- KEGG pathways
- GO-terms
- SEED classification







Gene occurrences

Identification of significant genes

Gi	roup 1		Group 2				
	7			\checkmark			
S	ample 1	Sample 2	Sample 3	mple q	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

Gene1	^m øl _{e z} 591	Sample 2	Sample 3	mple 4	Sample 5			
		536	1260	284	19			
Gene2	28	21	19	36	10			
Gene3	53	51	97	118	36			
Gene4	106 _r	149	266	47	11			
		$X_{i,j}$						
Gene1312	243	362	163	258	423			
Gene1313	13	43	23	67	34			
 Total	132 567	80 456	197 723	73 491	134 513			
n_{j} $X_{i,j}$ -number of reads matching gene <i>i</i> in sample <i>j</i> n_{j} -normalization factor per sample $R_{i,j} = \frac{X}{n}$								

Normalization

s	ample 1	n _{ple 2}	nple 3 Sample	Sam, le q	bles
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

	ample 1	Sample 2	Sample 3	ample 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
••••		00 45 6		70.404	
Total	1 32 567	80 456	1 97 723	73 491	1 34 513

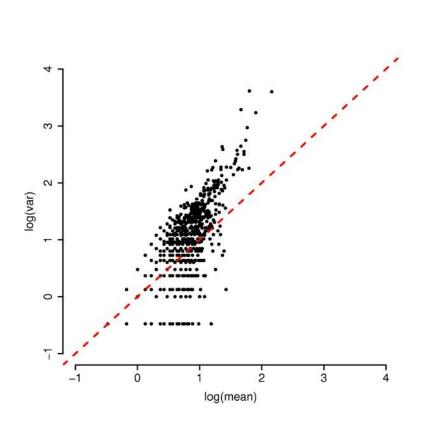
$$\log\left(\frac{\mathbf{E}[X_{i,j}]}{\mathbf{n}_{j}}\right) = \alpha_{0} + \sum \alpha_{k} y_{k}$$

Baseline Covariates (groups)

Statistical analysis

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

Statistical analysis



- Var $\begin{bmatrix} X_{i,j} \end{bmatrix} > E \begin{bmatrix} X_{i,j} \end{bmatrix}$
- Overdispersed data!

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

Estimated from the total residual sum

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

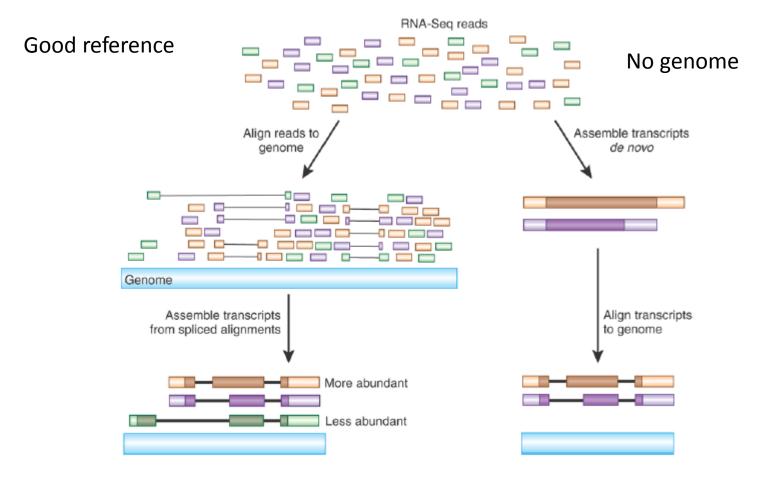
Summary metagenomics

- 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

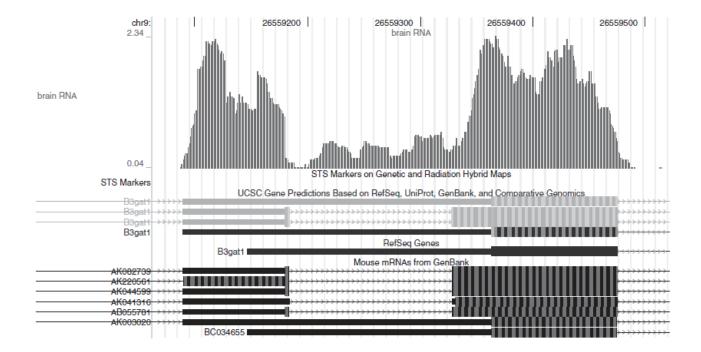
RNA-seq analysis strategy



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

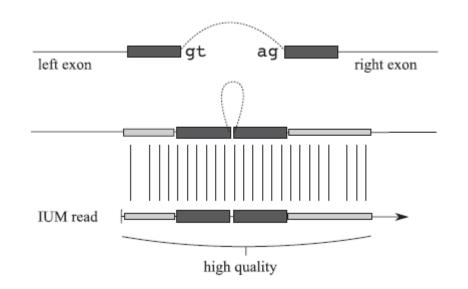
Alignment

• 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	2 18643	3 21721	84669	80540
Gene2	11288	3 2899	9 1062	2 6130	9581	. 17251
Gene3	44979	12906	5 14604	4 10378	85043	39478
Gene4	7133	3 4772	2 112	4 319	6863	7286
Gene5	34282	2 14379	9 1374	8 6133	12648	7620
Gene6	6531	7184	4 1962	2 651	1334	13125
Total	170702	71332	2 51143	3 45332	200138	165300

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

- n_i 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?

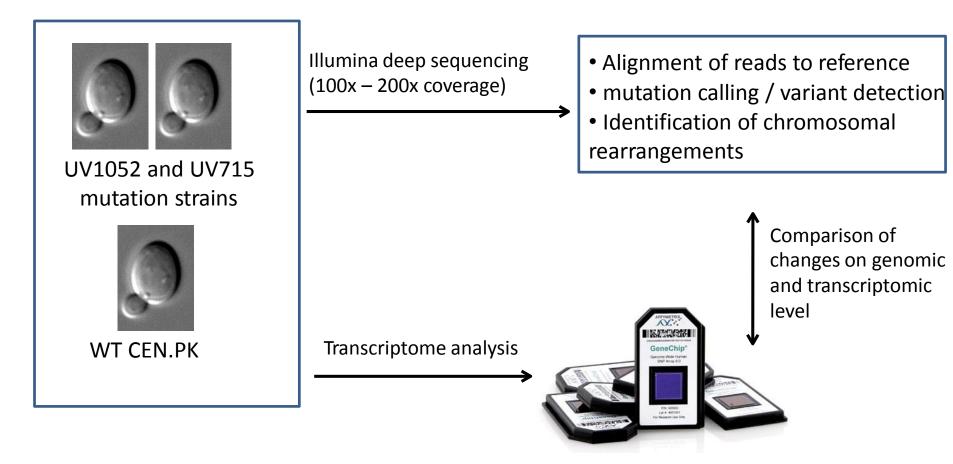
– Depends on your question

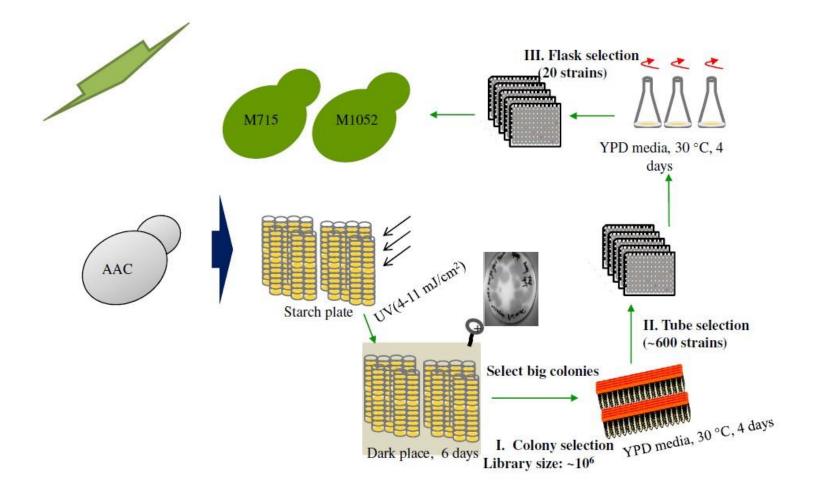
– Metagenomics: Sequence as much as possible

- 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





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 (BAMfiles)