NGS – part 2: applications

Tobias Österlund tobiaso@chalmers.se

NGS part of the course

Week 4	Friday 13/2	15.15-17.00	NGS lecture 1: Introduction to NGS, alignment, assembly
Week 6	Thursday 26/2	08.00-09.45	NGS lecture 2: RNA-seq, metagenomics
Week 6	Thursday 26/2	10.00-11.45	NGS computer lab: Resequencing analysis
Week 7	Thursday 5/3	10.00-11.45	Marcela: Exome sequencing
Week 8	Monday 9/3	17.00	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	nature
Number of microbes in all humans	6×10 ²³	magne
Number of stars in the universe	7×10 ²¹	
Number of bacterial cells in one human gut	10 ¹⁴	
Number of bacterial cells in one human gut	10 ¹³	
Number of bacterial genes in one human gut	3,000,000	Autoreal Aut
Number of genes in the human genome	21,000	BATURE COST

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 uncultivable and can not be grown in laboratories

Shotgun metagenomics

• Analysis of <u>all</u> DNA in a sample – the metagenome



Sample with high diversity

High throughput sequencing

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

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

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/

Species identification

- Prokaryots:
 - 16s rRNA gene
- Eukaryots:
 - 18s rRNA gene
- Can be amplified using PCR (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

Species abundance

0.0

-0.2

-0.4

-0.6

-0.8

-1.0

-1.2

-1.4

-1.6

PC.354

PC.481

(k Bacteria;p Firmicutes;c Clostridia) Qiime (Unassigned;Other;Other) (k Bacteria;p Bacteroidetes;c Bacteroidia) Bioinformatics (k_Bacteria;p_Firmicutes;c_Bacilli) program available (k_Bacteria;p_Firmicutes;c_Erysipelotrichi) at qiime.org (k_Bacteria;p_Proteobacteria;c_Deltaproteobacteria) Pick OTUs (k_Bacteria;p_Actinobacteria;c_Coriobacteriia) - Analysis of (k_Bacteria;p_Deferribacteres;c_Deferribacteres) species (k Bacteria;p Tenericutes;c Mollicutes) abundance (k Bacteria;p Proteobacteria;c Epsilonproteobacteria) Bioinformatics (k_Bacteria;p_Verrucomicrobia;c_Verrucomicrobiae) analysis (k Bacteria;p TM7;c TM7-3) (k Bacteria;Other;Other)

PC.634

PC.636

PC.635

PC.355

PC.593

PC.356

PC.607

Enterotypes of the human gut

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



Metagenome assembly

- Reference genomes are missing for majority of the bacteria we are studying
- Need to annotate the sequences
- Annotation (using e.g. blast) is easier for longer fragments
- De novo assembly of reads into longer contigs
- Can be hard due to large amount of data
 Need large amount of memory
- Can be hard due to undersampling of the metagenome

 Can't get a complete assembly

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	LE <mark>EIPSLVIDPGS</mark> CV	VT <mark>RFGYAGEE</mark> SI	<mark>Р</mark> М	TILPSYYG		<mark>g</mark> rn				K <mark>YVV</mark> DELQIH	AP	MEVKN <mark>g</mark> k.	. SN <mark>GII</mark> QDWE <mark>ST</mark>	LÝ <mark>TW</mark>
ARP4_SCHPO/8-433	GDEVSAIVIDPG <mark>S</mark> KV	V T <mark>R I G F S</mark> G E D I F	• <mark>K</mark>	. CVL <mark>PSY</mark> C <mark>g</mark>	E <mark>F</mark> S D <mark>G</mark> .					R R L F G E E Y I Y	κsΝ <mark>ρg</mark>	MEIKNAL.	. R N <mark>G W V</mark> E N W <mark>D</mark> V T	VDLW
ARP10_HUMAN/10-389	GGEKTAVVIDLGEAF	FT <mark>KCGFAGE</mark> TGI	• <mark>R</mark>	. CII <mark>P</mark> SVIK	(RA <mark>GMP</mark> K <mark>P</mark> V <u>R</u> V\	(Q <mark>Y N I</mark> N T <u>E</u>	<u>.</u>	<u>.</u> .	<u></u> EL <mark>Y</mark> S <mark>Y</mark> L	KEF I
Q4QF49_LEIMA/22-440	VLH <mark>TNAAVLD</mark> M <mark>gs</mark> h1	F T <mark>R L G F A G D T V I</mark>	• <mark>R</mark>	. MRQR <mark>T</mark> CVV	′К <mark>око</mark> тр	SDACDV				L D <mark>H V</mark> D D <mark>P</mark> A A A	. <mark>T</mark> T	<u></u> <u>. V</u> L	. E N <mark>G V I</mark> V D W E <mark>G Y</mark>	EELL
ARP6_ORYSJ/2-428	T <mark>gg<mark>s</mark>gvvvl<mark>d</mark>ngggl</mark>	L L <mark>kag f g g d</mark> m n <mark>f</mark>	<mark>РТ</mark>	. <u>av</u> v <mark>p</mark> ncma	∖K <mark>ppg</mark> s⊭	<				K <mark>wlv</mark> adqlqa	QDVDVT <mark>G</mark>	MTLRR <mark>P</mark> I.	DR <mark>gyl</mark> in <u>q</u> evq	REVW
ARP6_ASPFU/19-465	SLPEK <mark>T</mark> FII <mark>D</mark> NGAY1	F L <mark>kagyapg pi</mark>	<mark>PPE</mark> DL <mark>G</mark> QALSA	C <mark>STIP</mark> NAIA						R I <mark>Y I G</mark> AQ L N S	QVTDWNE	MVFRR <mark>P</mark> V.	. EK <mark>gyi</mark> vnweaq	KEIW
ARP6_NEUCR/16-439	A PPT T TLVL D NG AD 1	FI <mark>kagevsddk</mark>	3 <mark>0</mark> 6K	PRIIPNCLA		. <u>.</u> <u>.</u>	<u> </u>		<u>.</u>	K <mark>iyv</mark> gselek	(<mark>с</mark> к DF <u>s</u> e	L <mark>AF</mark> RR <mark>PV</mark> .	. EK <mark>gfi</mark> vnweaq	<u>k</u> e iw
Q4RMJ4_TETNG/4-389	D D E T T A L V C D N G S G L	. V <mark>k</mark> agfagddai	• <mark>R</mark>	. AVE <mark>PS</mark> IV <mark>G</mark>		.WPNSLVTG	→ VI <mark>G</mark> RH <mark>G</mark> WH <mark>G</mark> S		E <mark>G</mark> LL	R <mark>wg</mark> r <mark>gp</mark> eqkr	YSD <mark>P</mark> E	IPHRARH.	. H H Q L G T T W	EKIW
ACT25_DICDI/3-385	CE <mark>EVQAIVID</mark> N <mark>GS</mark> S\	/ C <mark>K</mark> AGF <u>G</u> GDDA <mark>I</mark>	P <mark>R</mark>	. <mark>T</mark> AF <mark>PS</mark> IV <mark>G</mark>	•	FIVDMDKK	KDSYFCKKNSC		FM <mark>G</mark> QK	DL <mark>YIG</mark> DEA	QSK.R <mark>G</mark> I	LNVKY <mark>PI</mark> .	. ER <mark>GIITNWN</mark> DM	EEIW.
Q94E75_ELAOL/4-378	AEDIQPLVCDNGTGN	MV <mark>KAGFAGDD</mark> AI	• <mark>R</mark>	. AVE <mark>PS</mark> IV <mark>G</mark>	• R <mark>P</mark> R H T <mark>G</mark>	VMV <mark>G</mark> MGQ k	<			D <mark>AYVG</mark> DEA	QSK.R <mark>G</mark> I	LTLKY <mark>P</mark> I.	. EH <mark>GIVNNWD</mark> DM	EKIW
Q4SGD5_TETNG/3-422	SQ <mark>g</mark> rk <mark>vvvcdngtg</mark> f	FV <mark>KCGYAGSNFI</mark>	• <mark>E</mark>	. HIF <mark>P</mark> ALVG	R <mark>P</mark> IIRS	3 <mark>.</mark>				T <mark>akvg</mark> ni	<u>.</u> E	I KVNY <mark>P</mark> M.	. EN <mark>GIVRNWD</mark> DM	KHLW
Q59QT2_CANAL/4-414	ILYNQ <mark>PVVID</mark> N <mark>GS</mark> GN	N L <mark>kag faged</mark> k <mark>i</mark>	• <mark>K</mark>	. <mark>SYAS</mark> ATIG	R <mark>P</mark> KYQK	(IMAA <mark>G</mark> STS	SLLSEQQS		H D	Q L F I <mark>G N</mark> S A	QDN.R <mark>G</mark> L	LKLSY <mark>P</mark> I.	. EH <mark>GIV</mark> NNWSDM	EKLW.
Q4S1S3_TETNG/6-400	ILANQ PVVIDNGSGV	/ I <mark>K</mark> AGFAGDQ I I	• <mark>K</mark>	. YCF <mark>PN</mark> YVG		RVMA <mark>G</mark> ALE <mark>G</mark>	<mark>></mark>			DLFIGPKA	EEH.R <mark>G</mark> L	LSVRY <mark>P</mark> M.	. EH <mark>giv</mark> kdwndm	E R I WI
Q4Q187_LEIMA/5-407	SASIQ <mark>S</mark> VVV <mark>DVGT</mark> RM	NT <mark>RVGFSGE</mark> EAI	° <mark>R</mark>	. TMLRSCVG	L PG TRRPRPTL	. L Q H				PFDIATG.D	AAYND <mark>GG</mark> L	LSLTY <mark>PV</mark> .	. RA <mark>gh</mark> vcd <mark>yd</mark> al	EKLL
O96819_97RYP/5-433	TERAPVVILDGGS	IL RAGYASDGAI	° R	. LDIPALVG	· H P R N R G	VAVAA <mark>G</mark> MN	4			E <mark>YEIG</mark> DVA	LAK.R <mark>G</mark> M	LTVSS <mark>PI</mark> .	. ES <mark>GRV</mark> VSWENM	EKLW
Q4D5V4_TRYCR/5-428	RERVPVVILDTGSHO	CL <mark>RAGYADEQG</mark>	° R	. LDIPALVG	• H <mark>P</mark> R N R <mark>G</mark>	VAMAAGMN	• <u>.</u>		<mark>.</mark>	E <mark>MEIG</mark> EEA	LVK.R <mark>G</mark> M	LTV <mark>G</mark> S <mark>P</mark> I.	. EN <mark>GLVVNWEH</mark> M	EKLW
Q54UQ7_DICDI/4-440	GDDVSAIVIDVGTFS	6 T <mark>KG G Y</mark> A G E D <mark>S</mark> I	° <mark>K</mark>	. AVF PTDIG	V <mark>V</mark> YKNE	ENETV <mark>G</mark> T <mark>G</mark> D	SEM <mark>g</mark> ekdds.	· · · · · · · · · · · ·	E <mark>P</mark> KR	TYYCGTNG	ITYR.R <mark>P</mark> H	METIN <mark>P</mark> L.	. SD <mark>GLIKNWD</mark> AM	<mark>e</mark> q iw
ARP4_ORYSJ/4-443	GDEVSAIVIDV6 <mark>S</mark> YS	S C <mark>kagyag d d T</mark> f	° <mark>K</mark>	. AVF <mark>PS</mark> VV <mark>G</mark>	SIEQTG	ETD.EAKA	ADKEAEAASDS	SKN <mark>g</mark> ak <mark>p</mark>	. MDVDKAKTKR	KLYVG.QE	L E F R . R D H	MEVIS <mark>PM</mark> .	. KD <mark>G T</mark> VTDW <mark>D</mark> I V	DNTW
Q5FAM0_BRAOL/4-442	GDEVSAIVVDLOSH1	r c <mark>kagyagedai</mark>	° K	. AVF <mark>PS</mark> VVG	• A <mark>V</mark> D <mark>G</mark> VE	AMDVDADS	SAKNNSNSEDS	SKTNE	SDKEK <mark>g</mark> kr	KLYV <mark>G</mark> SQA	LN YR . RDH	MEILS <mark>PI</mark> .	. KD <mark>g i v</mark> sdw <mark>d l</mark> v	DNTW
Q5CVZ6_CRYPV/10-42	2 <mark>g d d vg</mark> a i i v <mark>d vg</mark> scn	AT KI GYGGEDCI	° R	. QVWPSVVG	VKENG	°К	· · · · · · · · <mark>·</mark> · · ·			R <mark>FP</mark> LNFLS	MLEDVSVE <mark>P</mark> C	L KYE	. D <mark>GGLILNGDVF</mark>	EEII
Q7Q8D3_ANOGA/10-42	22 G D E I G A L V F D P G H H S	S L R V G Y A Q D D T I	° K	. AD I <mark>P S</mark> V V G	· V <mark>GP</mark> AD <mark>F</mark>	VMNSDLET	FKADNNI <mark>g</mark> s		TN	KYYVDTTH	IN VA. R <mark>P</mark> N	MEIQS <mark>Y</mark> M.	. KD <mark>G</mark> MIENWDLF	EKVI
Q7ZU19_DANRE/8-429	GDEVGALVFDMGSYS	6 V RAGYAGED CI	° <mark>K</mark>	. ADF PTVIG	VTLDRE	DGSTPMET	rd <mark>g</mark> ek <mark>g</mark> kqs		GT	TYFIDTN.Q	LR. VP. RES	MEVMSPL.	. KNGMIEDWDSFI	
Q5KLG9_CRYNE/5-476	GDEVSALVLDFGSY1	T T R AG Y AG E D C I	° R	. VVCPSFYG	YTND PSSSSSN	I <mark>G</mark> NSV <mark>G</mark> ENG	ANKENGEDVT	MAE <mark>P</mark> VPEGA	EEQSKKK <mark>g</mark> s <mark>g</mark> r	K YYVG ED <mark>G</mark> VS	VW	MEVGNEM.	LDGVVNDPEPA	SALL
ARP4_ASPFU/8-466	QDE ISAIVLDPG FS1	ſ T <mark>R A G F A G E D T</mark> I	° K	. SVIPTYYG	K Y TYEA	QE				K <mark>LIFG</mark> DDI	FVTP.RPG	LSIHN <mark>P</mark> MG	RDGVVEDWDMA	EKVW
ARP4_G/BZE/18-471	GDEVSALVLDPGYCS	6 TRAGE AGED VI	K	. SILPSFYG	HVTGDN	18				RDLFGDEC	LIP.REN	YEVRNYMN	IKDSVVEDW <mark>D</mark> VA.	ARIW
ARP4_NEUCR/19-469	GDEVSALVLDPGYCN	I I RAGYAGEEMI	K	. UVLPSFYG						RDVFGDEY	TVP.KPG	FEVENYMN	IRDSVVEDWDAA	TRMW
Q6C0A9_YARL#17-427	SDEVISIVIECOSSI	TRVGFSGDDL	K	. VVIPTKYG		(<mark>G</mark> ED	· · · · · · · · · · · ·	· · · · · · · · · ·		VYEFGLE N	MH RP . VAG	KETYS <mark>P</mark> V.	. QDGCIQDWDGV	AKLW
ARP4_YARL¥13-431	GDEVATLVIDTGSS	TTRVGYAGEDT	[•] K	LVVSTECG	L <mark>M</mark> ADED	VEMEDDTS	SNITKKL		N	KYKVGDSA	(N.L	MEVLH <mark>PL</mark> .	. TD <mark>GIV</mark> ADW <mark>D</mark> AVI	

PFAM domain for actin.



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

A-T A-T



Gene occurrences

Identification of significant genes

Gi	roup 1		Group 2				
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

S.	ample 1	Sample 2	Sample 3	mple 4	Samples		
 Gene1	591	536	1260	284	19		
Gene2	28	21	19	36	10		
Gene3	53	51	97	118	36		
Gene4	106 _r	149	266	47	11		
••••	4	$X_{i,j}$					
Gene1312	243	362	163	258	423		
Gene1313	13	43	23	67	34		
Total	132 567	80 456	197 723	73 491	134 513		
	Ĩ	l_{i}					
$X_{i,j}$ -number of reads matching gene i in sample j $oldsymbol{p}$ $oldsymbol{D}$ $oldsymbol{D}$ $oldsymbol{L}$							
n_j -normalization factor per sample n_i							

Normalization

	ample 1	am _{ple 2}	Sample Sal	n _{ple 4}	mples
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
••••					
Total	1 32 567	80 456	1 97 723	73 491	1 34 513

Statistical analysis

- Data from metagenomics is descrete (counts per gene/species)
- Not normally distributed

•
$$X_{i,j} \sim \text{Poisson}(\lambda_i)$$

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

Statistical analysis

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

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

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	4	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_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)