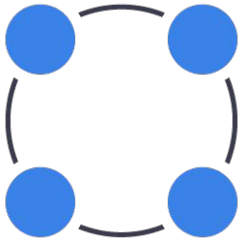
Разработка системы идентификации нуклеотидных последовательностей патогенных вирусов и бактерий в смесях на основе данных NGS

Александр Бебяков



Руководитель: Семенов Александр

Санкт-Петербургский НИИ эпидемиологии и микробиологии им. Пастера

Цель проекта

Построение системы обнаружения патогенных агентов на основе технологии секвенирования Illumina

Задачи проекта

Составить базу референсных последовательностей вирусов и бактерий, размеченных по группам патогенности (BSL) [СП 2013, HSE 2013]

Построить пайплайн идентификации патогенных агентов, аннотации доступных фрагментов из смешанных образцов

База последовательностей

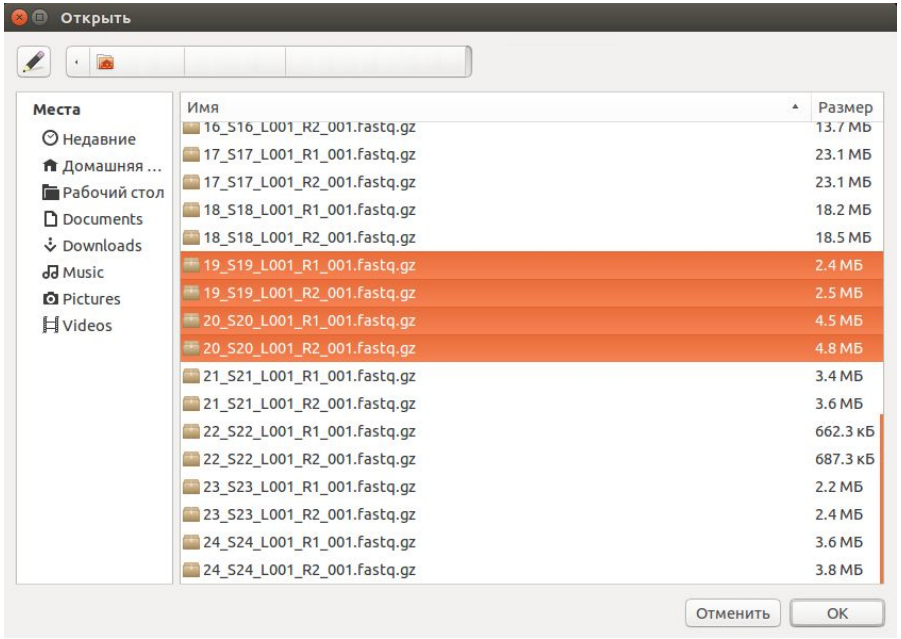
Entrez Direct

NCBI RefSeq, Nucleotide

	Вирусы	Бактерии
Последовательности	1322	3512
Таксоны	459	164

Kraken


Reads
Kraken
Report


Открыть
Места
Недавние
Домашняя ...
Рабочий стол
Documents
Downloads
Music
Pictures
Videos
Имя
Размер
16_S16_L001_R2_001.fastq.gz 13.7 МБ
17_S17_L001_R1_001.fastq.gz 23.1 МБ
17_S17_L001_R2_001.fastq.gz 23.1 МБ
18_S18_L001_R1_001.fastq.gz 18.2 МБ
18_S18_L001_R2_001.fastq.gz 18.5 МБ
19_S19_L001_R1_001.fastq.gz 2.4 МБ
19_S19_L001_R2_001.fastq.gz 2.5 МБ
20_S20_L001_R1_001.fastq.gz 4.5 МБ
20_S20_L001_R2_001.fastq.gz 4.8 МБ
21_S21_L001_R1_001.fastq.gz 3.4 МБ
21_S21_L001_R2_001.fastq.gz 3.6 МБ
22_S22_L001_R1_001.fastq.gz 662.3 кБ
22_S22_L001_R2_001.fastq.gz 687.3 кБ
23_S23_L001_R1_001.fastq.gz 2.2 МБ
23_S23_L001_R2_001.fastq.gz 2.4 МБ
24_S24_L001_R1_001.fastq.gz 3.6 МБ
24_S24_L001_R2_001.fastq.gz 3.8 МБ
Отменить
OK

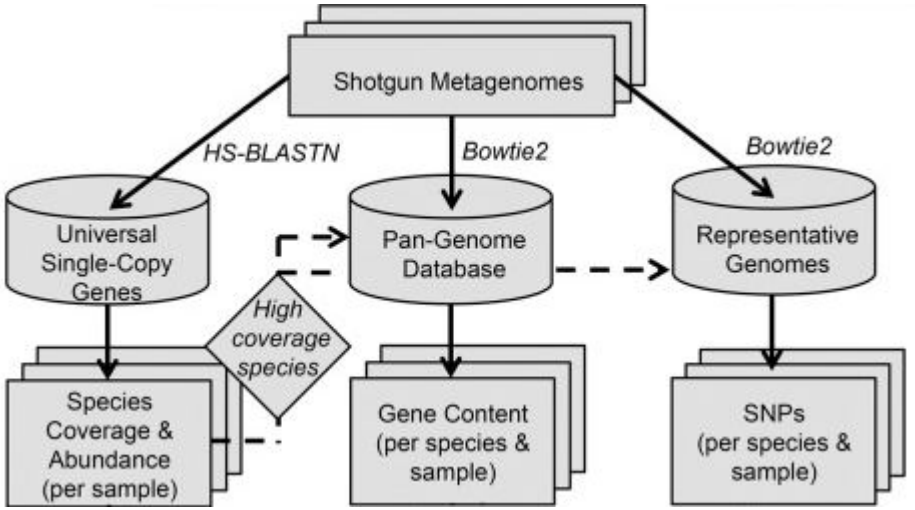
Kraken

	A	B	C	D	E	F	G	H	I	J
1	91.17	55518	55518	U	0	unclassified				
2	8.83	5375	595	-	1	root				
3	7.85	4780	3	D	10239	Viruses				
4	5.51	3355	0	-	439488	ssRNA viruses				
5	5.37	3271	0	-	35301	ssRNA negative-strand viruses				
6	5.37	3271	0	O	11157	Mononegavirales				
7	5.37	3271	0	F	11266	Filoviridae				
8	5.37	3270	0	G	186537	Marburgvirus				
9	5.37	3270	3006	S	11269	(I)Marburg marburgvirus				
10	0.37	227	227	-	448086	(I)Lake Victoria marburgvirus - Ci67				
11	0.06	37	37	-	1126254	(I)Lake Victoria marburgvirus - Leiden				
12	0.00	1	0	G	186536	(I)Ebolavirus				
13	0.00	1	1	S	186538	(I)Zaire ebolavirus				
14	0.14	84	1	-	35278	ssRNA positive-strand viruses	no DNA stage			
15	0.12	71	0	F	11050	Flaviviridae				
16	0.11	69	0	G	11051	Flavivirus				
17	0.11	69	69	S	64320	(II)Zika virus				
18	0.00	1	0	O	76804	Nidovirales				
19	0.00	1	0	F	11118	Coronaviridae				
20	0.00	1	0	-	693995	Coronavirinae				
21	0.00	1	0	G	693996	Alphacoronavirus				
22	0.00	1	1	S	11137	(IV)Human coronavirus 229E				
23	2.34	1422	1	-	35237	dsDNA viruses	no RNA stage			
24	2.33	1417	0	O	548681	Herpesvirales				
25	2.33	1417	0	F	10292	Herpesviridae				

MIDAS

Metagenomic Intra-species Diversity Analysis System

- Species abundance
- Gene content
- SNPs

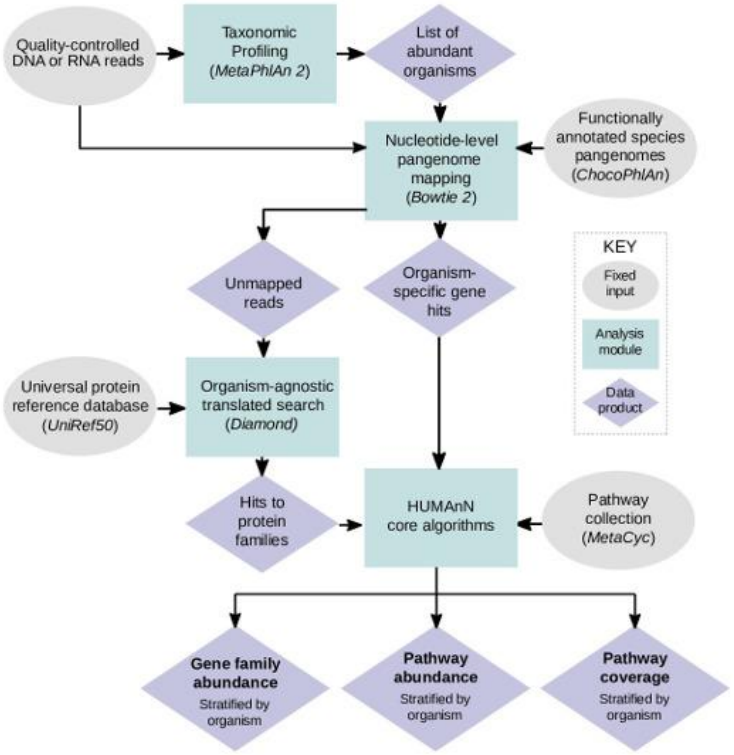


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088602/figure/NAYFACHGR201863F2/

HUMAnN2

HMP Unified Metabolic Analysis Network

- Species abundance (Archaea, Bacteria, Eukaryotes, Viruses)
- Gene families
- Pathway abundance



http://huttenhower.sph.harvard.edu/sites/default/files/humann2_flow_resume_option_no_text.png

METACYC™

HUMAnN2

k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Yersinia	38.16895
k__Bacteria|p__Proteobacteria|c__Betaproteobacteria|o__Burkholderiales|f__Alcaligenaceae|g__Bordetella	27.98239
k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Leptospiraceae|g__Leptospira	20.7263
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Salmonella	8.49172
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Escherichia	3.58191
k__Bacteria|p__Firmicutes|c__Bacilli|o__Bacillales|f__Listeriaceae|g__Listeria	1.04872
k__Bacteria|p__Proteobacteria|c__Betaproteobacteria|o__Burkholderiales|f__Alcaligenaceae|g__Bordetella|s__Bordetella_pertus…	27.98239
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Yersinia|s__Yersinia_en…	26.3766
k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Leptospiraceae|g__Leptospira|s__Leptospira_interrogans	20.7263
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Salmonella|s__Salmon…	8.49172
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Yersinia|s__Yersinia_p…	3.72465
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Escherichia|s__Esche…	3.58191
k__Bacteria|p__Firmicutes|c__Bacilli|o__Bacillales|f__Listeriaceae|g__Listeria|s__Listeria_monocytogenes	1.04872
k__Bacteria|p__Proteobacteria|c__Betaproteobacteria|o__Burkholderiales|f__Alcaligenaceae|g__Bordetella|s__Bordetella_pertus…	27.98239
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Yersinia|s__Yersinia_en…	26.3766
k__Bacteria|p__Spirochaetes|c__Spirochaetia|o__Spirochaetales|f__Leptospiraceae|g__Leptospira|s__Leptospira_interrogans|t__…	20.7263
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Salmonella|s__Salmon…	8.49172
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Yersinia|s__Yersinia_p…	3.72465
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacteriales|f__Enterobacteriaceae|g__Escherichia|s__Esche…	3.58191

# Gene Family	20_Abundance-RPKs
UNMAPPED	61784.0000000000
UniRef90_unknown	26324.9620077738
UniRef90_unknown|g__Yersinia.s__Yersinia_enterocolitica	7247.1636567391
UniRef90_unknown|g__Leptospira.s__Leptospira_interrogans	5551.0335537455
UniRef90_unknown|g__Escherichia.s__Escherichia_coli	5244.0035253125
UniRef90_unknown|g__Bordetella.s__Bordetella_pertussis	3755.6351172122
UniRef90_unknown|g__Salmonella.s__Salmonella_enterica	2927.8706226176
UniRef90_unknown|g__Yersinia.s__Yersinia_pseudotuberculosis	770.2204738465
UniRef90_unknown|g__Listeria.s__Listeria_monocytogenes	164.0607016243
UniRef90_A8ACR3	2195.9595959596
UniRef90_A8ACR3|g__Escherichia.s__Escherichia_coli	2195.9595959596
UniRef90_V8UZI1	1052.6315789474
UniRef90_V8UZI1|g__Bordetella.s__Bordetella_pertussis	1052.6315789474
UniRef90_A0A058XNH5	1000.0000000000
UniRef90_A0A058XNH5|g__Bordetella.s__Bordetella_pertussis	1000.0000000000
UniRef90_A8AKR9	1000.0000000000
UniRef90_A8AKR9|g__Escherichia.s__Escherichia_coli	1000.0000000000
UniRef90_B1JRI3	1000.0000000000
UniRef90_B1JRI3|g__Yersinia.s__Yersinia_pseudotuberculosis	1000.0000000000
UniRef90_K6DVP6	1000.0000000000
UniRef90_K6DVP6|g__Leptospira.s__Leptospira_interrogans	1000.0000000000

PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)	17.2370681554
PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)|g__Salmonella.s__Salmonella_enterica	7.3154221014
PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)|g__Escherichia.s__Escherichia_coli	5.6518298740
PWY-6630: superpathway of L-tyrosine biosynthesis	17.2273918514
PWY-6630: superpathway of L-tyrosine biosynthesis|g__Escherichia.s__Escherichia_coli	2.1444447806
PWY0-1297: superpathway of purine deoxyribonucleosides degradation	16.4483280244
PWY0-1297: superpathway of purine deoxyribonucleosides degradation|g__Yersinia.s__Yersinia_enterocolitica	5.6321045325
PWY-7664: oleate biosynthesis IV (anaerobic)	16.2320972221
PWY-7664: oleate biosynthesis IV (anaerobic)|g__Escherichia.s__Escherichia_coli	2.2809663700
HISDEG-PWY: L-histidine degradation I	16.1141206756
HISDEG-PWY: L-histidine degradation I|g__Yersinia.s__Yersinia_enterocolitica	9.5827247495
HISDEG-PWY: L-histidine degradation I|g__Salmonella.s__Salmonella_enterica	1.7370779111
AST-PWY: L-arginine degradation II (AST pathway)	16.0838375875
AST-PWY: L-arginine degradation II (AST pathway)|g__Yersinia.s__Yersinia_enterocolitica	8.8854138906
AST-PWY: L-arginine degradation II (AST pathway)|g__Yersinia.s__Yersinia_pseudotuberculosis	1.6366710655
PWY-6628: superpathway of L-phenylalanine biosynthesis	16.0037339007
PWY-6628: superpathway of L-phenylalanine biosynthesis|g__Escherichia.s__Escherichia_coli	2.1727290772
PYRIDNUCSYN-PWY: NAD biosynthesis I (from aspartate)	15.6439477544
FASYN-ELONG-PWY: fatty acid elongation -- saturated	14.6594981434
FASYN-ELONG-PWY: fatty acid elongation -- saturated|g__Escherichia.s__Escherichia_coli	2.2880025438
PWY-5188: tetrapyrrole biosynthesis I (from glutamate)	14.1477732424
PWY-5188: tetrapyrrole biosynthesis I (from glutamate)|g__Escherichia.s__Escherichia_coli	1.5098406742

UniRef to

- MetaCyc Reactions
- KEGG Orthogroups (KOs)
- Pfam domains
- Level-4 enzyme commission (EC) categories
- EggNOG (including COGs)
- Gene Ontology (GO)
- Informative GO

Тестовые данные

Бактерии	**Вирусы**
Bordetella pertussis	Zaire ebolavirus
Yersinia enterocolitica	Marburg marburgvirus
Yersinia pseudotuberculosis	Zika virus
Listeria monocytogenes	Hepatitis B virus
Leptospira interrogans	
Escherichia coli	
Salmonella enterica	

Результаты

- Размечено 623 tax_id
- Kraken pipeline
- HUMAnN2

- Проверка на данных секвенирования

https://github.com/ABapmath/2018_spring_SEM

Литература

Приложение 3 к СП 1.3.3118-13

The Approved List of biological agents: http://www.hse.gov.uk/pubns/misc208.htm

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research. 2016;26(11):1612-1625. doi:10.1101/gr.201863.115.

Abubucker S et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012 Jun 13; 8(6):e1002358.

Спасибо за внимание