Long-read metagenomics with MetaMaps

Alexander Dilthey (DPhil)

Heinrich Heine University Düsseldorf

UKD Universitätsklinikum Düsseldorf

hhu.

Long-read metagenomics

1. Potentially better classifiability
2. Long-range spatial information on genome structure
3. Methylation
4. Potential for real-time analysis

Read
Reference genomes

kMer transformation

kMer transformation

Minimizer selection & indexing

Within every window of *w* bases, keep kmer with lowest (hash) value

0	1	2	3	4	5	6	7	8
T	G	A	T	A	C	G	A	A
T	G	A	T	A				
	G	A	T	A	C			
		A	T	A	C	G		
			T	A	C	G	A	
				A	C	G	A	A

Reference genomes

kMer transformation

Minimizer selection & indexing

Car

Sco

$$E[J(A,B)] := \frac{E[|A \cap B|]}{E[|A \cup B|]} = \frac{n_{kmers} \times i^k}{2 \times n_{kmers} - (n_{kmers} \times i^k)}$$

Reference genomes

Ondov et al.

$x \in S(\mathrm{A} \cup \mathrm{B})$: $\mathrm{P}(x \in A \cap B) = Jaccard(\mathrm{Read}, \mathrm{Ref})$

$$x \in A \cap B \iff x \in S(A) \wedge x \in S(B).$$

Reference genomes

Ondov et al.

Estimated alignment identities

MashMap, Mash

Jain et al.

$x \in S(\mathrm{A} \cup \mathrm{B})$: $\mathrm{P}(x \in A \cap B) = Jaccard(\mathrm{Read}, \mathrm{Ref})$

$$x \in A \cap B \iff x \in S(A) \wedge x \in S(B).$$

Reference genomes

Ondov et al.

$x \in S(\mathrm{A} \cup \mathrm{B})$: $\mathrm{P}(x \in A \cap B) = Jaccard(\mathrm{Read}, \mathrm{Ref})$

$x \in A \cap B \iff x \in S(A) \wedge x \in S(B)$.

Mapping qualities

Probability distribution over mapping locations

- Ass
  Rea ce.
- Exp

  $E[J($ $\frac{s \times i^k}{n_{kmers} \times i^k})$
- P$(x$ d(Read, Ref)]
- Rea t identity
- Normalize over mapping locations.

Reference genomes

kMer transformation

Minimizer selection & indexing

Candidate mapping regions

Sco

Sar

Read mapping locations
(cond. on composition)

Genome frequencies

Reference genomes

1. Likelihood of a mapping: $F_{\mathbf{g}} \times \frac{1}{E_{r;g}} \times P_r(i)$

Probability that read emanates from genome g

Mapping quality of location i

Possible read start locations in genome g

2. $LC(r;\mathbf{F}) := \sum_{g \in G} \sum_{i \in map(r,g)} F_g \times \frac{1}{E_{r;g}} \times P_r(i)$

3. Optimization via EM.

kMer transformation

Minimizer selection & indexing

Candidate mapping regions

Scoring (MinHash)

- Alignment idt.
- Mapping quality

Sample composition

Read mapping locations (cond. on composition)

kMer transformation

Minimizer selection & indexing

Candidate mapping regions

Sco

Sar

Rec
(co

1. MetaMaps is an integrated metagenomic classification system.
2. MashMap [Jain et al. 2018], Mash [Ondov et al. 2016].
3. Auto-tuning: minimum read length & identity.
4. EM also used in Kallisto, Pathoscope, Centrifuge…

Standard DB & evaluation

Standard database – 25 gigabases

Bacteria	5774
Viruses/viroids	6059
Archaea	215
Fungi	7
Human	1
Other eukaryotes	2
Total	12058

- 2 simulation experiments
  - i100: 100 different species
  - p25: 25 pathogens with 3 x 5 closely related strains
- 2 real-data experiments
  - PacBio HMP data
  - Zymo Mock Community by Loman et al
- Contamination experiments and simulated CAMI mouse gut data

ARTICLE
https://doi.org/10.1038/s41467-019-10934-2 OPEN

Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps

Alexander T. Dilthey[1,2], Chirag Jain[2,3], Sergey Koren[2] & Adam M. Phillippy[2]

High accuracy on most simulated datasets

only 28% of species and 63% of genera present in the sample are represented in the database

First, the median estimated MetaMaps alignment identity in the CAMI experiment is around 82%; this is

Very high degrees of divergence are problematic (CAMI simulated data)

- 28% of species and 63% of genera present in the MetaMaps database
- Median estimated alignment identity ~82%.

Real data: High precision

Read assignment accuracy HMP

Read assignment accuracy Zymo

Recall reduced due to minimum read length

1. Recall of MetaMaps reduced due to minimizer-based mapping strategy with minimum read length requirement (here: 1000bp).
2. HMW DNA extraction recommended, methods improving.
3. When measured at the base level *over all reads*, recall is competitive.

High accuracy for „medium-length“ reads

- In general: longer reads are easier to classify.
- MetaMaps achieves high accuracy from the minimum read length onwards.
- Likely cause: EM.

Alignment-based method is more robust against large-genome contaminants

High accuracy of compositional estimation

Database-sample mismatches become apparent

Database contains an Actinomyces genome, but not the right species

Database-sample mismatches become apparent

Reads	Reads %		COG group
390	0.14%	A	RNA processing and modification
399	0.15%	B	Chromatin structure and dynamics
47263	17.31%	C	Energy production and conversion
12889	4.72%	D	Cell cycle control, cell division, chromosome partitioning
68289	25.02%	E	Amino acid transport and metabolism
29636	10.86%	F	Nucleotide transport and metabolism
58230	21.33%	G	Carbohydrate transport and metabolism
31347	11.48%	H	Coenzyme transport and metabolism
26591	9.74%	I	Lipid transport and metabolism
49020	17.96%	J	Translation, ribosomal structure and biogenesis
70430	25.80%	K	Transcription
61699	22.60%	L	Replication, recombination and repair
59690	21.87%	M	Cell wall/membrane/envelope biogenesis
12514	4.58%	N	Cell motility
37550	13.76%	O	Post-translational modification, protein turnover, and chaperones
55657	20.39%	P	Inorganic ion transport and metabolism
13808	5.06%	Q	Secondary metabolites biosynthesis, transport, and catabolism
154932	56.76%	S	Function unknown
33782	12.38%	T	Signal transduction mechanisms
18251	6.69%	U	Intracellular trafficking, secretion, and vesicular transport
24103	8.83%	V	Defense mechanisms
408	0.15%	W	Extracellular structures
116	0.04%	Y	Nuclear structure
131	0.05%	Z	Cytoskeleton

Long-read benchmarking

- Difference between „reads correct“ and „bases correct“

- Even with no DB/sample mismatches, simulated accuracy > empirical accuracy
  [sensitivity and recall, >5%].

- Rapid development: Good simulators for current Nanopore and PacBio CCS?

- More complex empirical datasets with known truth urgently needed (ultra-deep PromethION + *de novo* assembly?)

Conclusion

- Integrated solution: Assignment, composition, mapping locations.
- Fast minimzer-based mapping plus probabilistic mapping qualities.
  Enables statistical modeling of sample composition via EM.
  Unique / high-confidence alignments serve as „attractors“ for ambiguous reads.
- Auto-tuning: Will improve with longer, higher-quality reads (e.g. PacBio CCS).
- Future work: speed; prediction of novel species.

Acknowledgements and funding

https://github.com/DiltheyLab/MetaMaps

NHGRI-NIH:
- **Chirag Jain**
- **Sergey Koren**
- **Adam Phillippy**
- **Brian Ondov**
- Arang Rhie
- Brian Walenz

HHU Düsseldorf:
- Sebastian Scharf
- Daniel Strelow
- Alona Tyshaieva
- Birgit Henrich

Georgia Institute of Technology:
- **Chirag Jain**
- **Srinivas Aluru**

NIH National Human Genome Research Institute

DFG

JÜRGEN MANCHOT STIFTUNG

Ondov et al.

$x \in S(\mathrm{A} \cup \mathrm{B})$: $\mathrm{P}(x \in A \cap B) = Jaccard(\mathrm{Read}, \mathrm{Ref})$

$x \in A \cap B \iff x \in S(A) \wedge x \in S(B).$

Mapping qualities

Probability distribution over mapping locations

- P(
  Re
- Ja
  (N
- Re
- $E[J$
- Normalize over mapping locations.

rd(Read,

ate

dentity

Reference genomes

Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data

Kristen Curry, Rice University PhD student
Treangen Lab; Dept. Computer Science
Dilthey Lab; Institute of Medical Microbiology, University Hospital of Düsseldorf Germany

Shotgun [WGS] vs. 16S Sequencing

	Shotgun	16S
kingdoms	all	bacteria & archea
composition	definite	relative
sequencing depth	high	shallow
computation	complex	direct
cost	$$$	$
accessbility	outsource	in-house

Sequencing technologies

r_1: ATGCT..
r_2: ATTGG..
⋮
r_n: AGTGC..

Reference-based classification

reads

reference database

classification

r_1: ATGCT..
r_2: ATTGG..
⋮
r_n: AGTGC..

Reference-based classification

r_1: ATGCTG..
r_2: ATGCC..
⋮
r_n: CTCCG..

Reference-based classification with errors

*Oxford Nanopore Technologies MinION**
"anywhere, anyone, anytime"
- portable
- real-time
- cost efficient
- **high error**

"An unknown read is **MORE** likely to be a bacteria that is already in the sample"

*Andrea D. Tyler et al., "Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications," *Scientific Reports* 8, no. 1 (July 19, 2018): 10931

EM in Emu

Alignment scores

best hit

truth

iteration 1

classification likelihood

iteration 2

iteration 3

"A read is **MORE** likely to be a bacteria that is in the sample in **HIGHER** abundance"

trim noise

Emu output

Test data results

Error throughout EM iterations in Emu

Classification software comparison

Critical Assessment of Metagenome Interpretation (CAMI) 2 Mouse Gut

Emu in gut microbiome studies

Alzheimer's Disease

Soriano, Sirena et al. "Fecal Microbiota Transplantation Derived from Alzheimer's Disease Mice Worsens Brain Trauma Outcomes in Young C57BL/6 Mice," *medRxiv*, November 23, 2021.

Concussion

Soriano, Sirena, et al. "Alterations to the Gut Microbiome after Sport-Related Concussion and Subconcussive Impacts in a Collegiate Football Players Cohort," *medRxiv*, July 14, 2021.

COVID-19 severity

Villapol Lab, Houston Methodist Research Institute

Acknowledgements

Emu – profile microbial communities with error-prone 16S reads

Treangen Lab - Rice University
Qi Wang
Michael G. Nute
Elizabeth Reeves
Todd J. Treangen

Heinrich Heine University Düsseldorf
Alexander Dilthey
Alona Tyshaieva
Enid Graeber
Patrick Finzer

Helios University Clinic Wuppertal
Werner Mendling

Villapol Lab - Houston Methodist Research Institute
Sirena Soriano
Sonia Villapol

Savidge Lab - Baylor College of Medicine
Qinglong Wu
Tor Savidge

Article | Published: 30 June 2022

Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data

Kristen D. Curry, Qi Wang, Michael G. Nute, Alona Tyshaieva, Elizabeth Reeves, Sirena Soriano, Qinglong Wu, Enid Graeber, Patrick Finzer, Werner Mendling, Tor Savidge, Sonia Villapol, Alexander Dilthey & Todd J. Treangen

Nature Methods **19**, 845–853 (2022) | Cite this article

6941 Accesses | **21** Citations | **166** Altmetric | Metrics

https://gitlab.com/treangenlab/emu