Differential Analysis of Whole-genome Shotgun Sequences

Sohan Seth[1], Niko Välimäki[2,3], Samuel Kaski[1,3], Antti Honkela[3]

1 Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University, Espoo, Finland
2 Genome-Scale Biology Program and Department of Medical Genetics,
University of Helsinki, Helsinki, Finland
3 Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki, Helsinki, Finland

Background

- metagenomics is the study of genetic material collected from environment
- each whole-genome shotgun (WGS) sample is a large ($\sim 10^7$) collection of short ($\sim$100 bp) DNA fragments, or reads from multiple ($\sim$100) species, e.g.,

$$\underbrace{\text{ATGCGTGA...GAACCGTAC}}_{\text{100 bp}}$$

- given two groups of metagenomic samples, differential analysis explores what attributes (compositional, functional) is over or under expressed

  e.g., Bacteroides sp. 20-3 is enriched in T2D patients

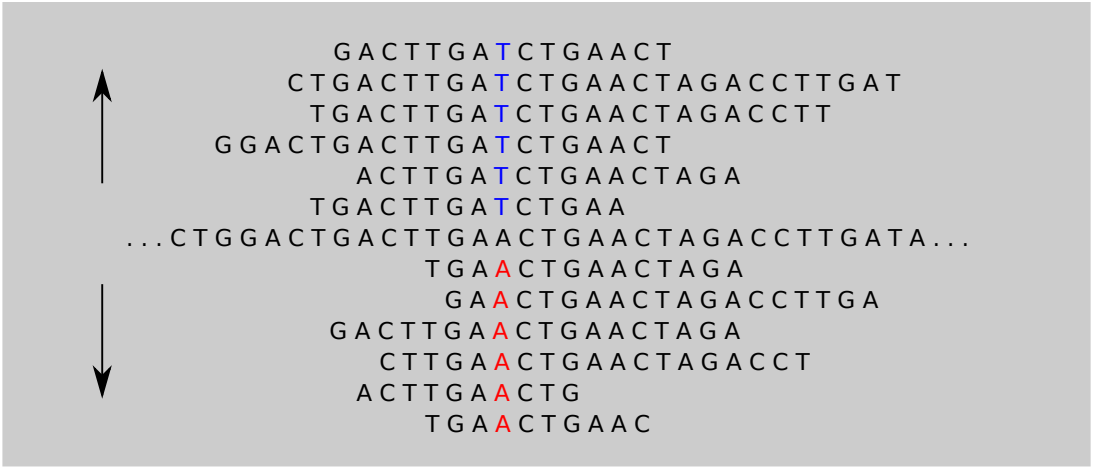
- differential analysis of metagenomic samples is usually done at the species or strains level,

  e.g., find relative abundances of taxa $\rightarrow$ do hypothesis testing

- our goal is to study local changes at the base pair level, e.g., mutations, insertion, deletion, duplication, etc., as well as global changes at the taxa level

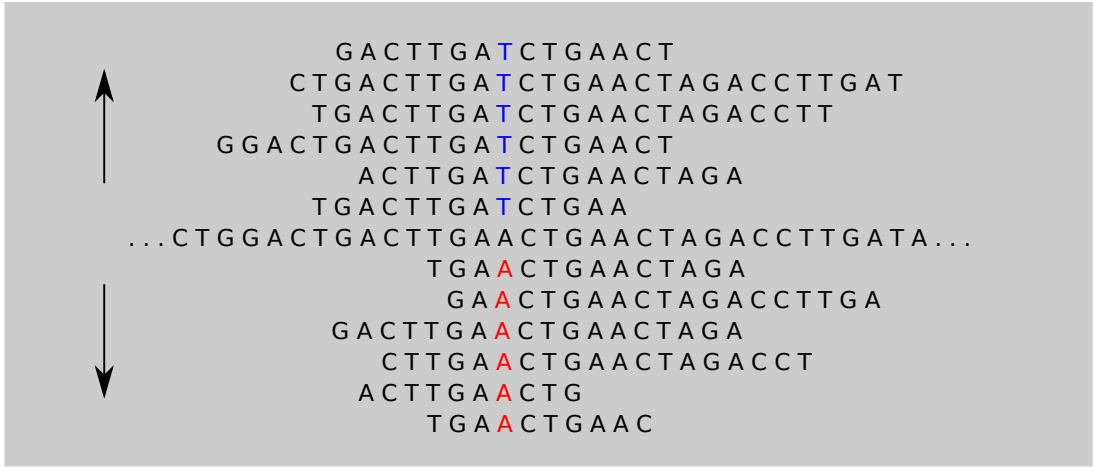
Motivation

- longer k-mers are usually unique at the species and even strain level
- mutations, deletions etc. change the k-mer abundances locally, e.g.,



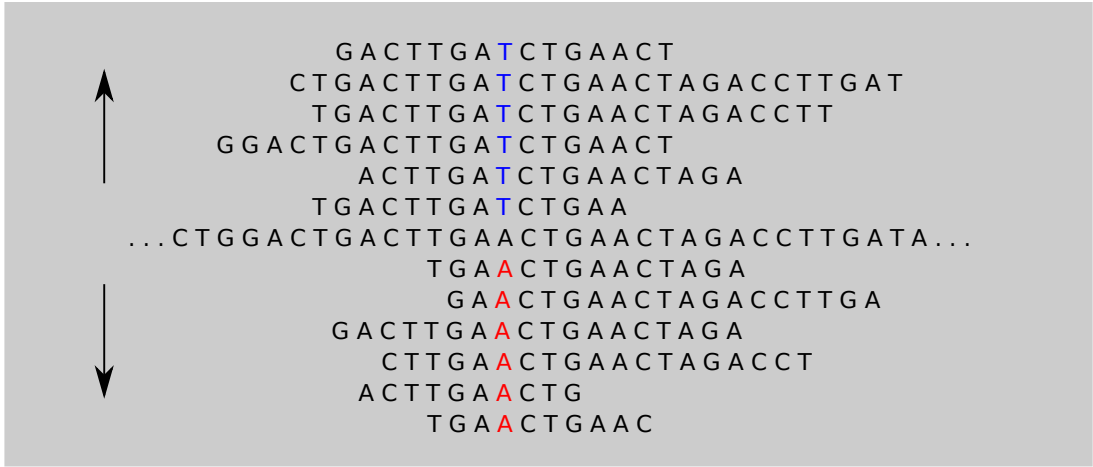
Motivation

- longer k-mers are usually unique at the species and even strain level
- mutations, deletions etc. change the k-mer abundances locally, e.g.,



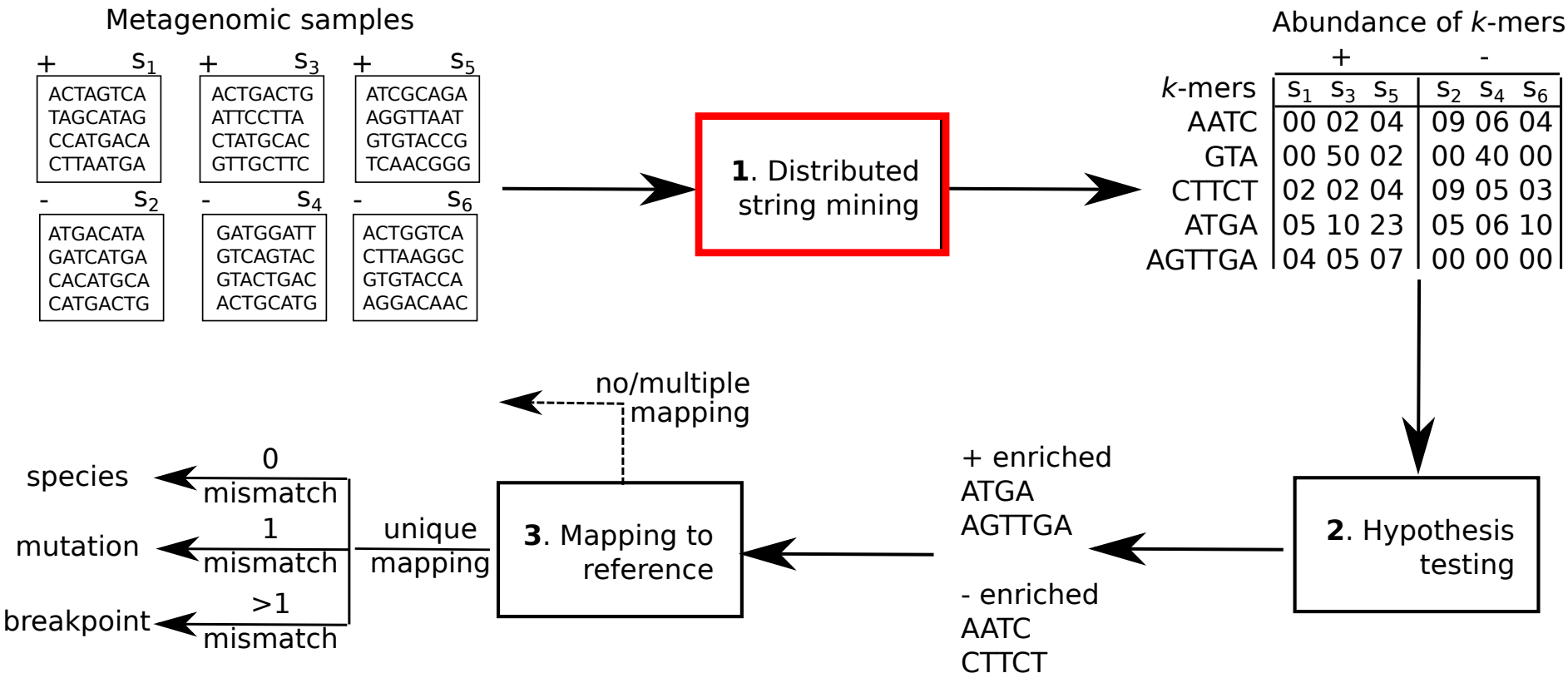
Motivation

- longer k-mers are usually unique at the species and even strain level
- mutations, deletions etc. change the k-mer abundances locally, e.g.,



study all possible differentially expressed k-mers

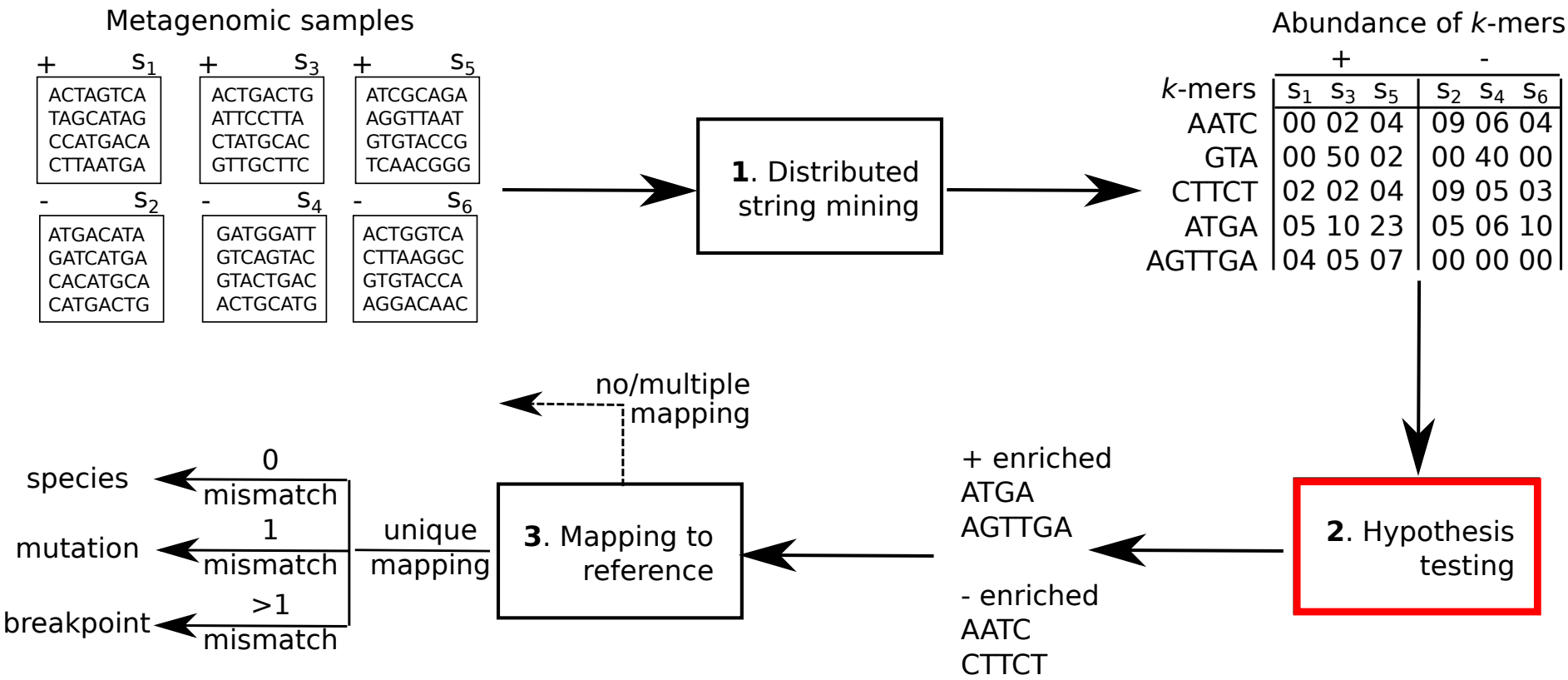
Pipeline



1 memory and time efficient computation of k-mer abundances,

- (minimal) sub-k-mers with same abundances as super-k-mers are ignored
- (robust) k-mers appearing only once in a sample are ignored
- (sequential) only enriched k-mers are stored for further analysis

Pipeline

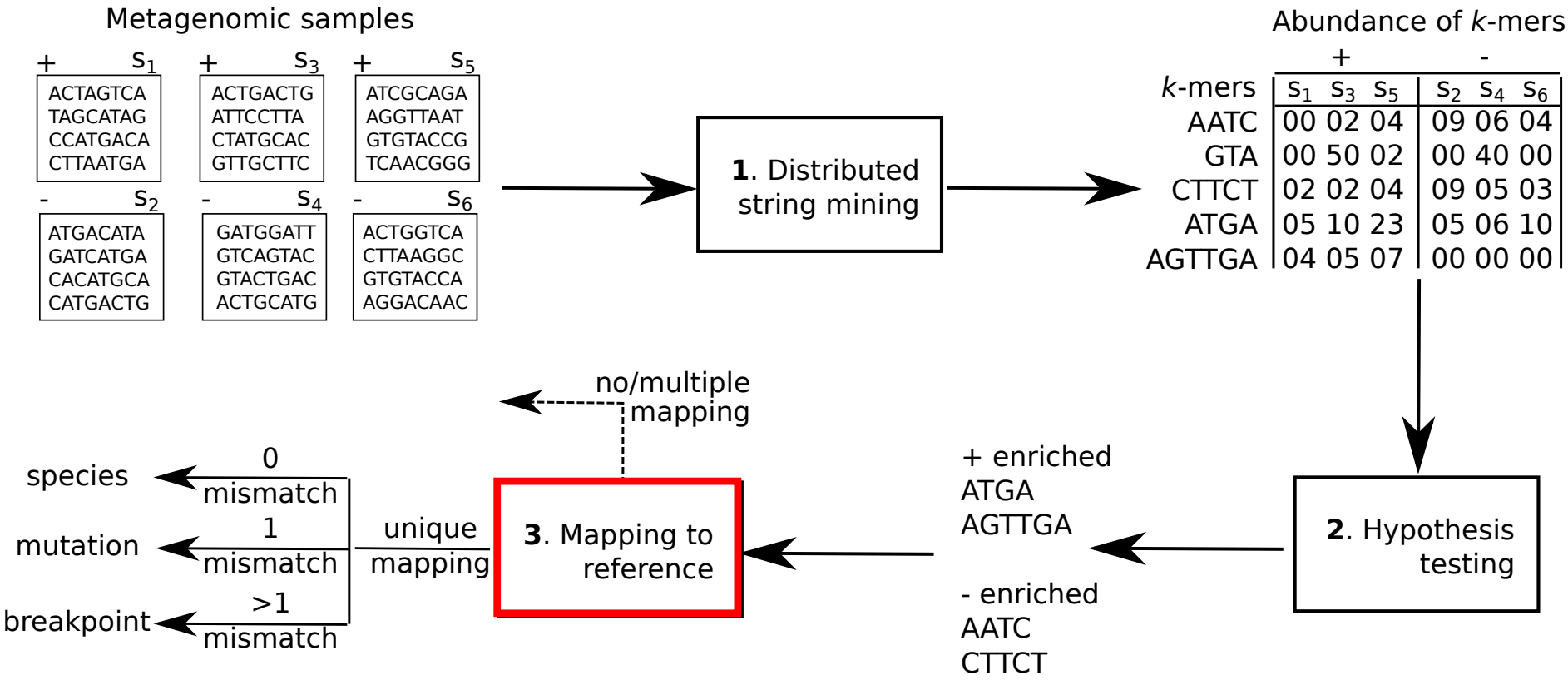


2 likelihood ratio test under model $n_j^c \sim \mathrm{NB}(s_j \mu^c, \alpha^c)$

sample j, group c, mean and dispersion (μ, α), and sequencing depth s_j

- Newton's method is used to find the maximum likelihood solution
- test is sensitive to change in both mean and dispersion $\Rightarrow$ simpler

Pipeline

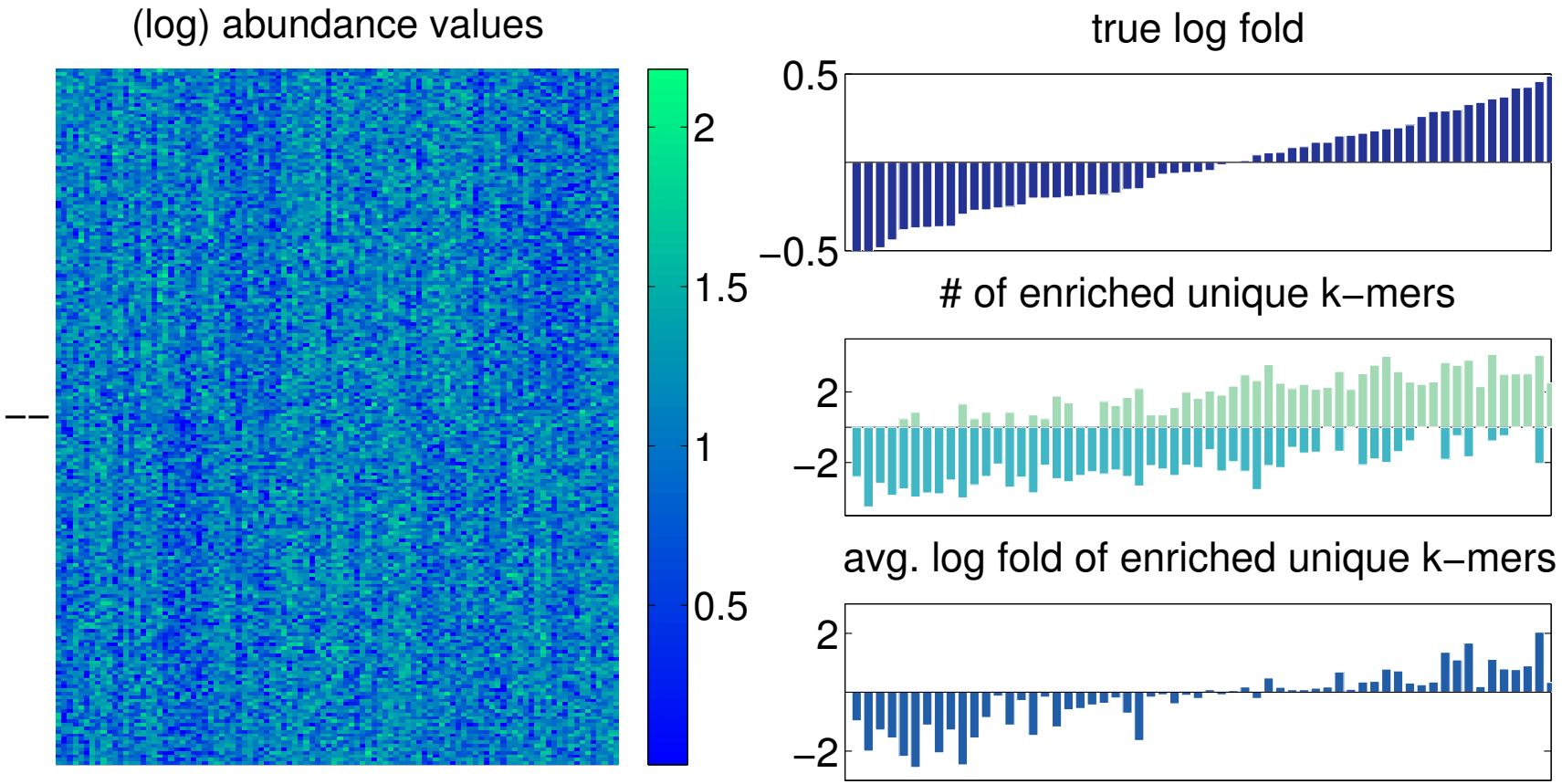


3 mapping to suitable reference database with standard tools, e.g., BWA

- none hits and multiple best hits are (so far) ignored
- unique mappings (single best hit) are stored and investigated

Simulation: species level change

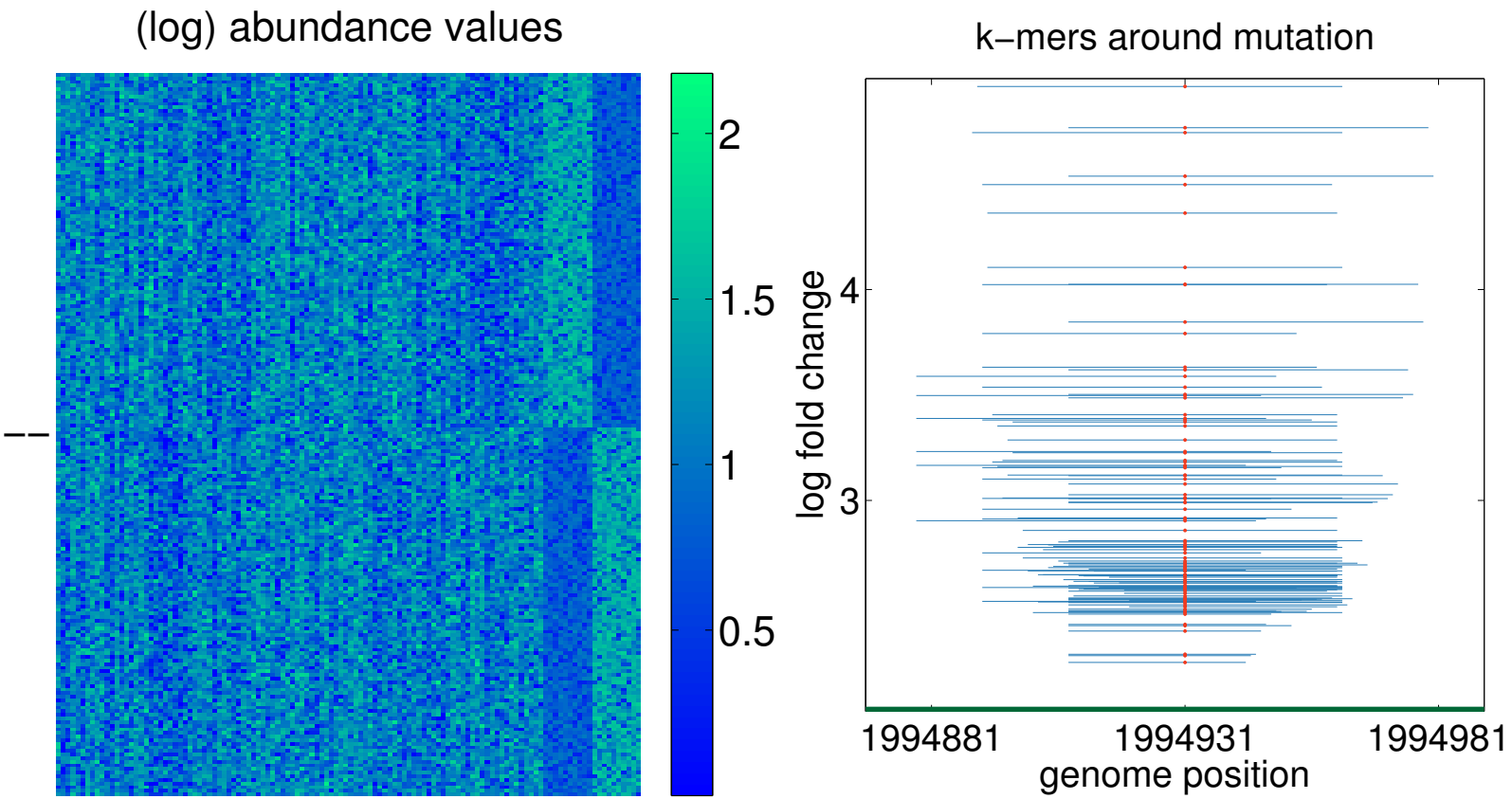
- proportion of exactly mapped enriched *k*-mers in two groups provides evidence toward enrichment at taxa level
- mock metagenomic samples: 100 species/strains, 100 samples in each group



proportion of over/under expressed *k*-mers are related to true fold change

Simulation: *k*-mer level change

- multiple enriched unique *k*-mers with single mismatch at a common genome position provides evidence toward mutation
- additional mock species with mutations (5), deletion (5), etc.



many *k*-mers of arbitrary lengths and locations overlap a common mutated base

T2D P2: Species level change

3.2 million k-mers with p-value < 0.000005

Organism	# of unique k-mers total	T2D	CON	k-mer len mean ± std	log-fold mean ± std
Bacteroides coprocola	27870	71	27799	47.35±17.90	-4.18±2.51
Clostridium bartlettii	31104	44	31060	44.64±17.09	-3.75±0.87
* Roseburia intestinalis	2259855	471	2259384	42.62±16.86	-3.34±1.19
* Faecalibacterium prausnitzii	61406	90	61316	31.91±12.85	-2.22±0.78
* Roseburia inulinivorans	18056	259	17797	33.58±16.00	-1.95±1.38
Bacteroides pectinophilus	19427	91	19336	40.36±18.36	-1.91±0.82
* Faecalibacterium prausnitzii	18882	383	18499	29.02±10.67	-1.80±1.09
* Eubacterium rectale	181315	108	181207	38.04±13.96	-1.74±0.87
* Eubacterium eligens	81902	72	81830	46.14±19.74	-1.53±0.46
Ruminococcus obeum	10667	843	9824	28.12±10.29	-1.33±1.34
Ruminococcus sp. 5-1-39	16475	952	15523	28.39±11.53	-1.25±1.05
Clostridium methylpentosum	10828	10485	343	43.07±14.08	2.04±1.14
Eubacterium hallii	17282	11784	5498	25.40±8.93	2.63±3.44
* Bacteroides sp. 20-3	23884	23710	174	46.87±18.12	3.05±1.53
* Clostridium bolteae	15877	14631	1246	29.51±13.27	3.80±2.82
* Bacteroides caccae	46884	46389	495	65.27±12.89	3.82±1.73
* Bacteroides cellulosilyticus	27089	26705	384	50.01±17.95	4.19±2.02

enrichment directions tally with previously reported findings $*$

T2D P2: k-mer level change

mutation locations covered by at least two k-mers

Organism	mutations
Faecalibacterium prausnitzii M21/2	1107
Faecalibacterium prausnitzii A2	892
Roseburia intestinalis L1 82	857
Eubacterium rectale ATCC 33656	825
Bacteroides coprocola DSM 17136	574
Roseburia inulinivorans DSM 16841	291
Eubacterium eligens ATCC 27750	266

T2D P2: k-mer level change

mutation locations covered by at least two k-mers

Organism	mutations
Faecalibacterium prausnitzii M21/2	1107
Faecalibacterium prausnitzii A2	892
Roseburia intestinalis L1 82	857
Eubacterium rectale ATCC 33656	825
Bacteroides coprocola DSM 17136	574
Roseburia inulinivorans DSM 16841	291
Eubacterium eligens ATCC 27750	266

- metagenomics is gaining substantial popularity in recent years
- understanding differential abundance is a crucial problem, e.g., understanding the difference between patients and healthy microbiome
- efficiently exploring local changes opens up new avenues
- our future goal is to draw viable conclusion without the reference database

T2D P2: k-mer level change

mutation locations covered by at least two k-mers

Organism	mutations
Faecalibacterium prausnitzii M21/2	1107
Faecalibacterium prausnitzii A2	892
Roseburia intestinalis L1 82	857
Eubacterium rectale ATCC 33656	825
Bacteroides coprocola DSM 17136	574
Roseburia inulinivorans DSM 16841	291
Eubacterium eligens ATCC 27750	266

- metagenomics is gaining substantial popularity in recent years
- understanding differential abundance is a crucial problem, e.g., understanding the difference between patients and healthy microbiome
- efficiently exploring local changes opens up new avenues
- our future goal is to draw viable conclusion without the reference database

Thank You! Question?