# Quantifying genomic variation of gut microbiota across the human population

Stephen Nayfach

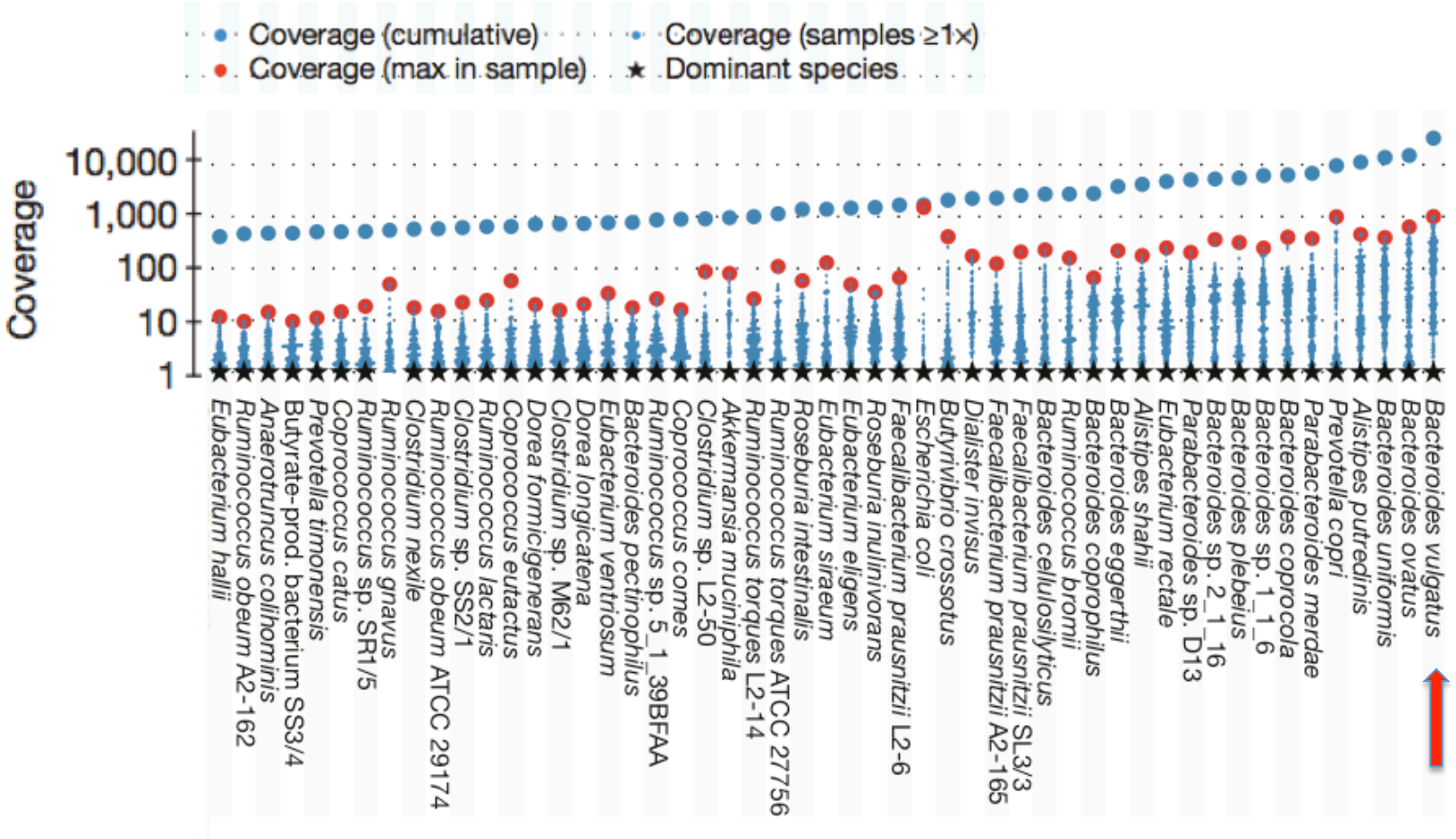
iSEEM2 Call

February 9, 2015

# Biological Motivation

- Evolutionarily similar organisms often differ in their gene content and ecology
- Unknown how genes from gut microbiota vary in copy number across the human population
- Is there strain-level biogeography?
- What genes are essential for gut microbiota? Do they vary across species?
- How stable are genomes from gut microbiota? Are some more stable than others?
- What can we learn by studying patterns of co-variation across genes? Are these genes acquired/lost as a unit via HGT?

# Strain-level diversity of *Bacteroides vulgatus*: a pilot study




Schloissnig et al. (2013) Nature

# Strain-level diversity of *Bacteroides vulgatus*: a pilot study

Shotgun metagenomes from gut microbiome

Estimate abundance of bacterial species

Select samples with high relative abundance of *Bacteroides vulgatus*

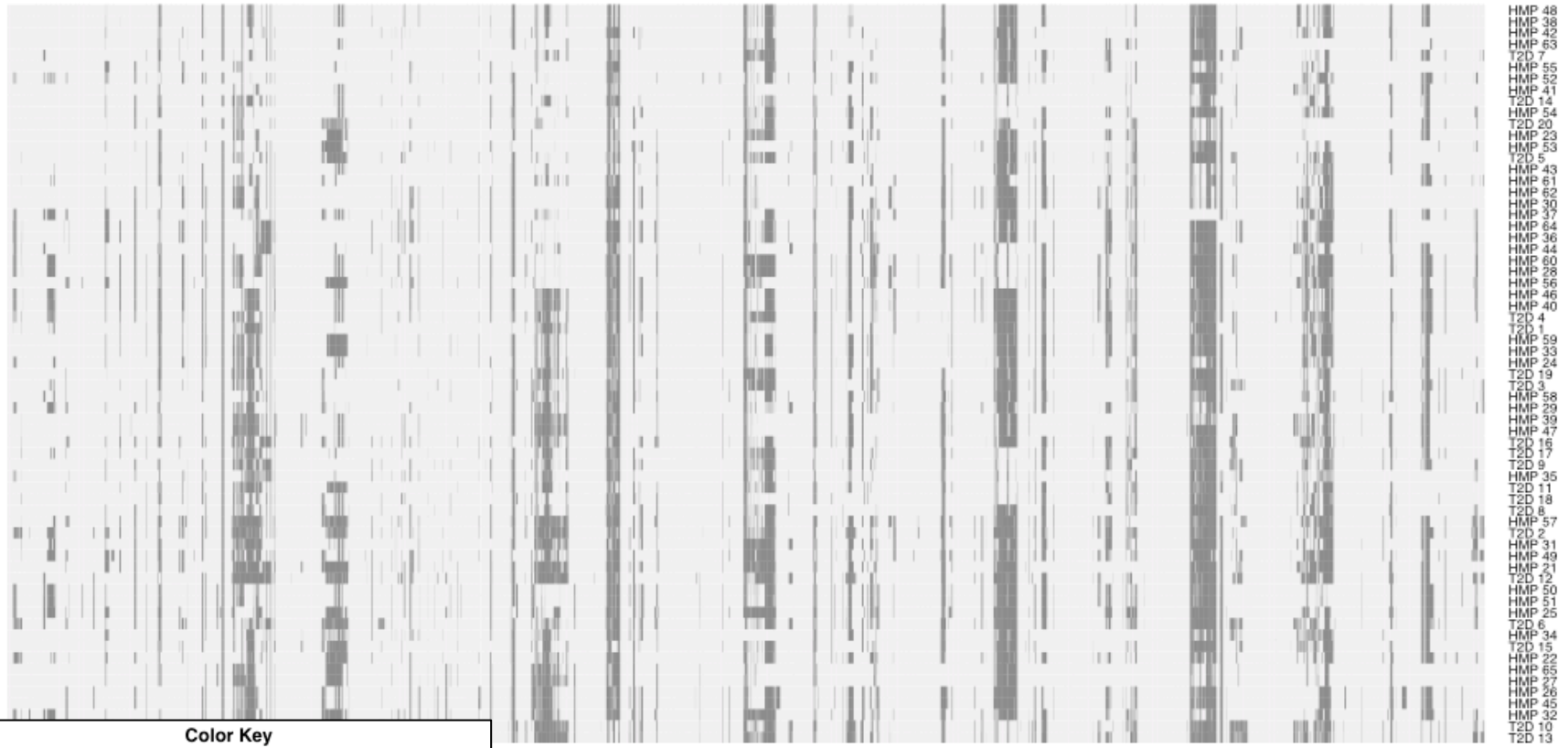
20,000 reference genomes

Select representative genome for *Bacteroides vulgatus*

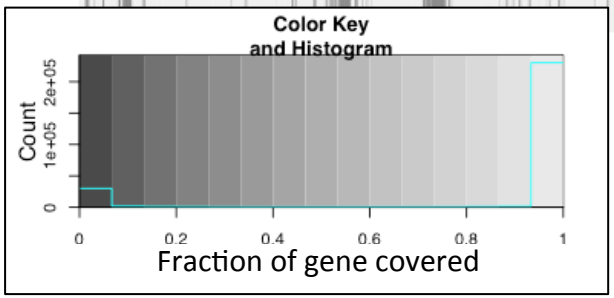
Map paired end reads with bowtie2

Estimate gene copy number variation and SNP frequencies

# Inter-sample gene content variability of *Bacteroides vulgatus*




Stool samples (N=65)



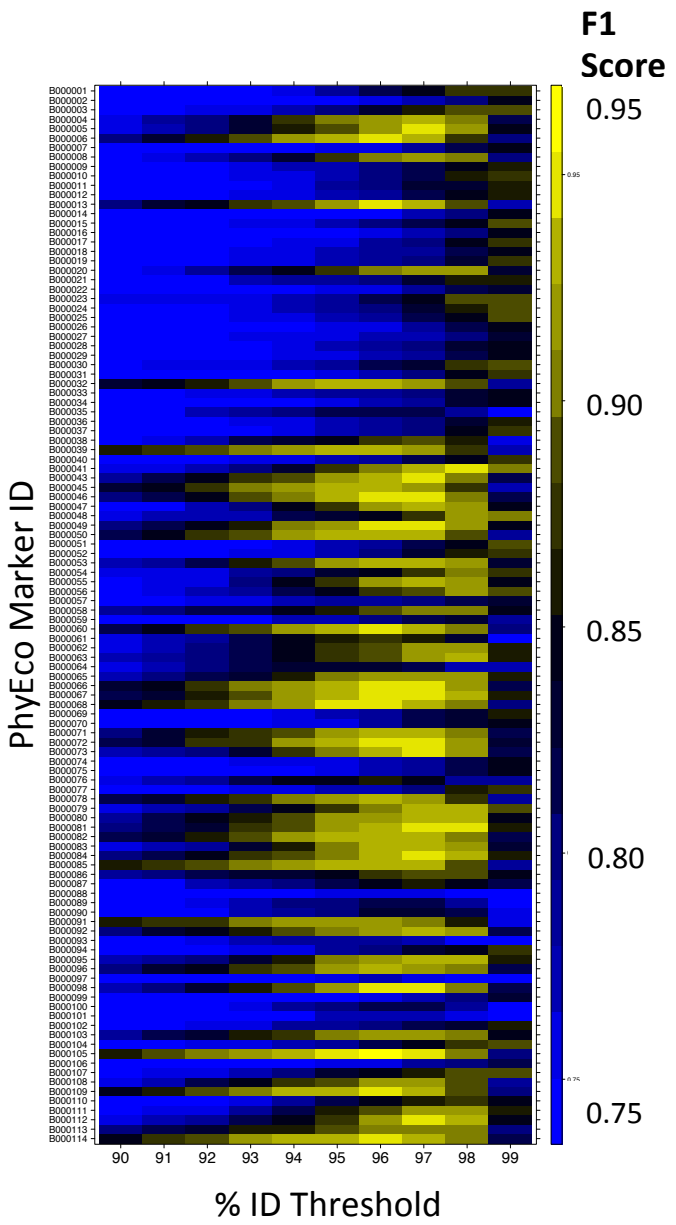

B. vulgatus ATCC 8482 genes
(N=4195; ordered by genomic position)

# Developing methods to scale up project

- Reference genome clustering (i.e. species level groups)
- Profiling abundance of genome clusters in metagenomes
- Mapping reads to abundant genome clusters
- Estimating copy number variation of pan-genomes

# PhyEco genome clusters

- Computed pairwise similarity of 112 Bacterial PhyEco marker genes across >18,000 high quality genomes
- Performed average linkage clustering for each marker genes at various %ID thresholds
- Evaluated performance against ANI (average nucleotide identity)
- Good performance for a subset of marker genes at ~96.5%ID threshold



# PhyEco genome clusters

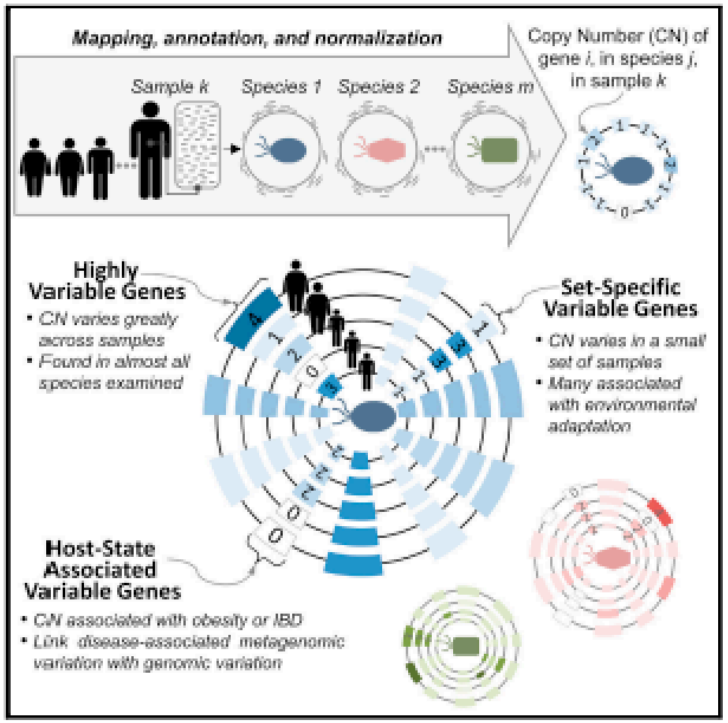
- Move on, or publish results and make resource available?

- Potential items:
  - Perpetually updated
  - Repeat for Archaea, and maybe Fungi
  - Webserver/ftp site

Cell

# Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species

## Graphical Abstract




## Highlights

- A metagenomic data analysis pipeline allows strain-level gene copy-number inference
- Copy-number variation (CNV) is widespread across many prevalent human gut species
- CNV involves mostly environment-related functions and is associated with disease
- Strain-level population structure reveals known and uncharacterized strains

## Authors

Sharon Greenblum, Rogan Carr, Elhanan Borenstein

## Correspondence

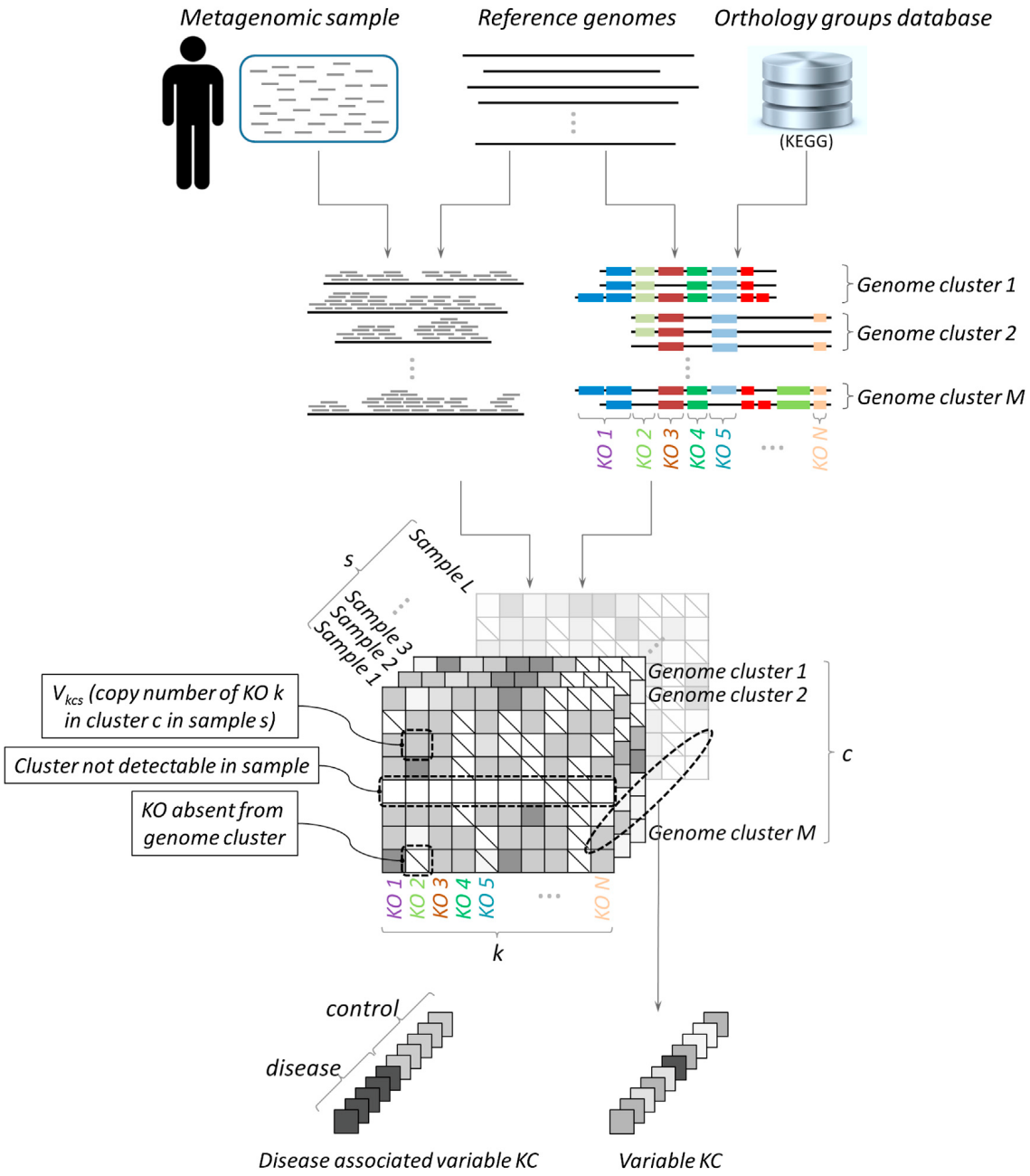
elbo@uw.edu

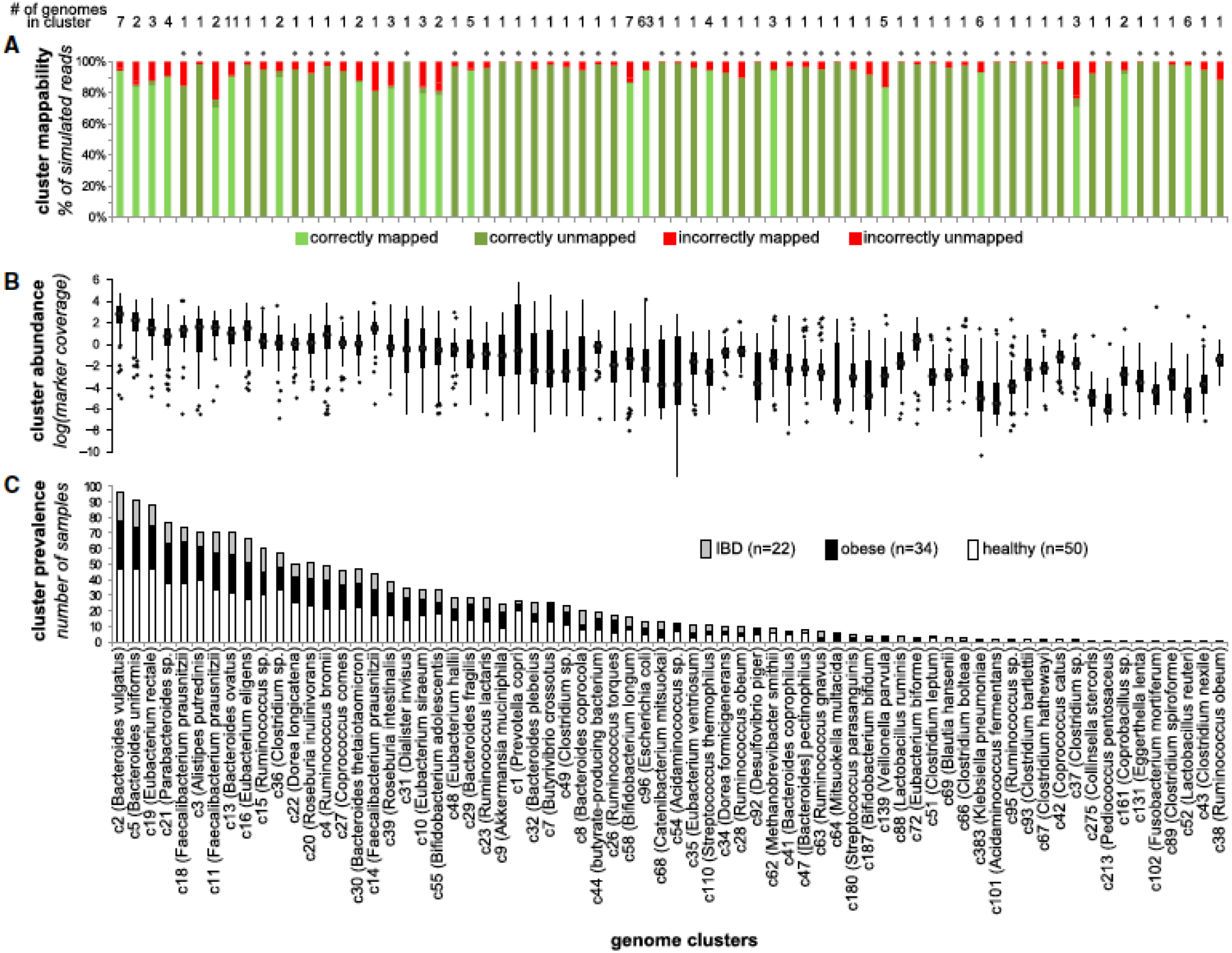
## In Brief

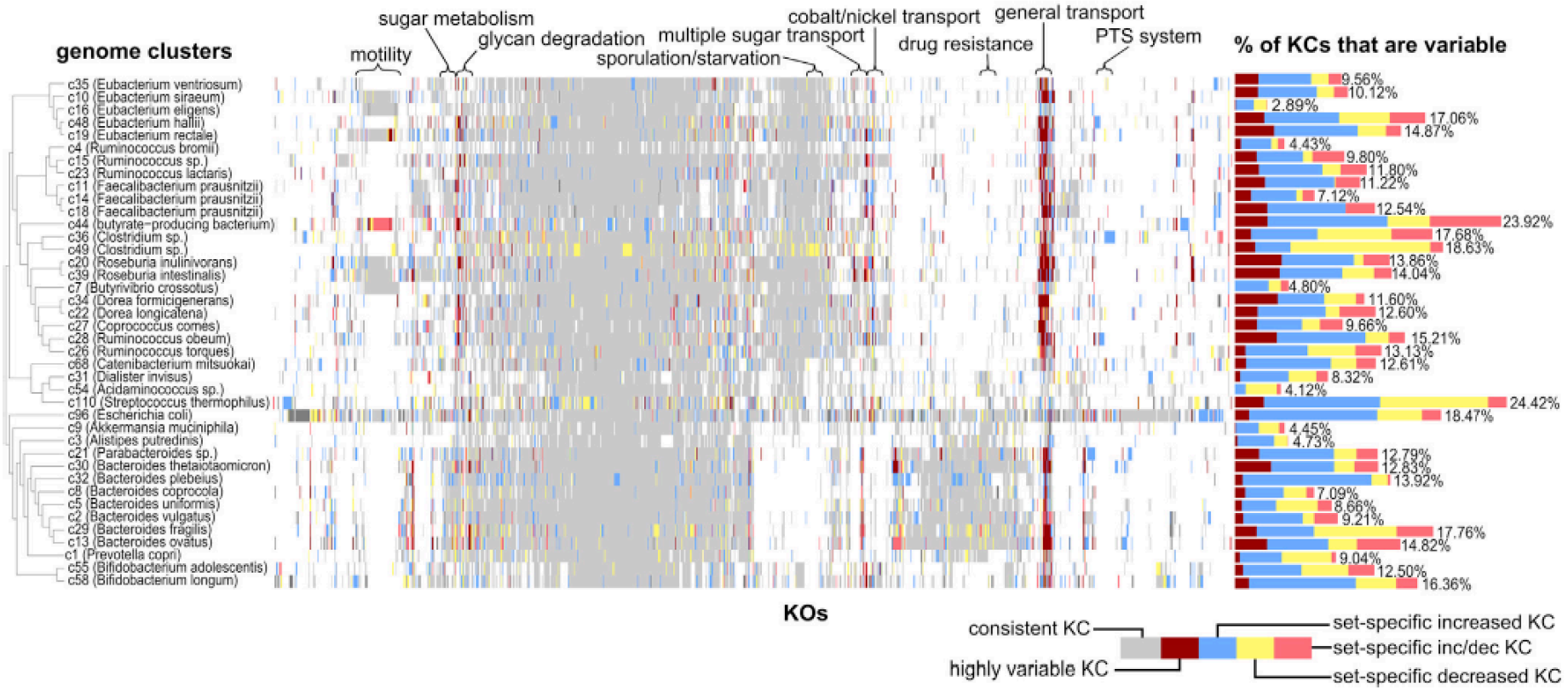
Extensive strain-level variation is detected in the human gut microbiome, with differences in gene copy-number impacting specific adaptive functions and linked to obesity and inflammatory bowel disease.

# Overview of methods

- Used genome clusters from Schloissnig et al 2012
- Annotated genes with KEGG Orthology database
- Mapped reads from Qin 2010 (~100 samples) to 70 “detectable” genome clusters
- Estimated copy number coverage normalizing by 13 single copy genes


Metagenomic sample
Reference genomes
Orthology groups database
(KEGG)
Genome cluster 1
Genome cluster 2
Genome cluster M
KO 1
KO 2
KO 3
KO 4
KO 5
KO N
s
Sample L
Sample 3
Sample 2
Sample 1
Genome cluster 1
Genome cluster 2
Genome cluster M
c
k
$V_{kcs}$ (copy number of KO k in cluster c in sample s)
Cluster not detectable in sample
KO absent from genome cluster
control
disease
Disease associated variable KC
Variable KC


# of genomes in cluster
7 2 3 4 1 1 2 11 1 1 2 1 1 1 1 2 1 3 1 3 2 1 5 1 1 1 1 1 1 1 1 1 7 63 1 1 1 4 1 1 1 3 1 1 1 1 1 1 5 1 1 1 1 1 1 6 1 1 1 1 1 3 1 1 2 1 1 1 6 1 1
A
cluster mappability
% of simulated reads
100%
80%
60%
40%
20%
0%
correctly mapped
correctly unmapped
incorrectly mapped
incorrectly unmapped
B
cluster abundance
log(marker coverage)
6
4
2
0
−2
−4
−6
−8
−10
C
cluster prevalence
number of samples
IBD (n=22)
obese (n=34)
healthy (n=50)
c2 (Bacteroides vulgatus)
c5 (Bacteroides uniformis)
c19 (Eubacterium rectale)
c21 (Parabacteroides sp.)
c18 (Faecalibacterium prausnitzii)
c3 (Alistipes putredinis)
c11 (Faecalibacterium prausnitzii)
c13 (Bacteroides ovatus)
c16 (Eubacterium eligens)
c15 (Ruminococcus sp.)
c36 (Clostridium sp.)
c22 (Dorea longicatena)
c20 (Roseburia inulinivorans)
c4 (Ruminococcus bromii)
c27 (Coprococcus comes)
c30 (Bacteroides thetaiotaomicron)
c14 (Faecalibacterium prausnitzii)
c39 (Roseburia intestinalis)
c31 (Dialister invisus)
c10 (Eubacterium siraeum)
c55 (Bifidobacterium adolescentis)
c48 (Eubacterium hallii)
c29 (Bacteroides fragilis)
c23 (Ruminococcus lactaris)
c9 (Akkermansia muciniphila)
c1 (Prevotella copri)
c32 (Bacteroides plebeius)
c7 (Butyrivibrio crossotus)
c49 (Clostridium sp.)
c8 (Bacteroides coprocola)
c44 (butyrate-producing bacterium)
c26 (Ruminococcus torques)
c58 (Bifidobacterium longum)
c96 (Escherichia coli)
c68 (Catenibacterium mitsuokai)
c54 (Acidaminococcus sp.)
c35 (Eubacterium ventriosum)
c110 (Streptococcus thermophilus)
c34 (Dorea formicigenerans)
c28 (Ruminococcus obeum)
c92 (Desulfovibrio piger)
c62 (Methanobrevibacter smithii)
c41 (Bacteroides coprophilus)
c47 ([Bacteroides] pectinophilus)
c63 (Ruminococcus gnavus)
c64 (Mitsuokella multacida)
c180 (Streptococcus parasanguinis)
c187 (Bifidobacterium bifidum)
c139 (Veillonella parvula)
c88 (Lactobacillus ruminis)
c72 (Eubacterium biforme)
c51 (Clostridium leptum)
c69 (Blautia hansenii)
c66 (Clostridium bolteae)
c383 (Klebsiella pneumoniae)
c101 (Acidaminococcus fermentans)
c95 (Ruminococcus sp.)
c93 (Clostridium bartlettii)
c67 (Clostridium hathewayi)
c42 (Coprococcus catus)
c37 (Clostridium sp.)
c275 (Collinsella stercoris)
c213 (Pediococcus pentosaceus)
c161 (Coprobacillus sp.)
c131 (Eggerthella lenta)
c102 (Fusobacterium mortiferum)
c89 (Clostridium spiroforme)
c52 (Lactobacillus reuteri)
c43 (Clostridium nexile)
c38 (Ruminococcus obeum)
genome clusters


genome clusters
motility
sugar metabolism
glycan degradation
sporulation/starvation
multiple sugar transport
cobalt/nickel transport
drug resistance
general transport
PTS system
% of KCs that are variable
c35 (Eubacterium ventriosum)
c10 (Eubacterium siraeum)
c16 (Eubacterium eligens)
c48 (Eubacterium hallii)
c19 (Eubacterium rectale)
c4 (Ruminococcus bromii)
c15 (Ruminococcus sp.)
c23 (Ruminococcus lactaris)
c11 (Faecalibacterium prausnitzii)
c14 (Faecalibacterium prausnitzii)
c18 (Faecalibacterium prausnitzii)
c44 (butyrate-producing bacterium)
c36 (Clostridium sp.)
c49 (Clostridium sp.)
c20 (Roseburia inulinivorans)
c39 (Roseburia intestinalis)
c7 (Butyrivibrio crossotus)
c34 (Dorea formicigenerans)
c22 (Dorea longicatena)
c27 (Coprococcus comes)
c28 (Ruminococcus obeum)
c26 (Ruminococcus torques)
c68 (Catenibacterium mitsuokai)
c31 (Dialister invisus)
c54 (Acidaminococcus sp.)
c110 (Streptococcus thermophilus)
c96 (Escherichia coli)
c9 (Akkermansia muciniphila)
c3 (Alistipes putredinis)
c21 (Parabacteroides sp.)
c30 (Bacteroides thetaiotaomicron)
c32 (Bacteroides plebeius)
c8 (Bacteroides coprocola)
c5 (Bacteroides uniformis)
c2 (Bacteroides vulgatus)
c29 (Bacteroides fragilis)
c13 (Bacteroides ovatus)
c1 (Prevotella copri)
c55 (Bifidobacterium adolescentis)
c58 (Bifidobacterium longum)
9.56%
10.12%
2.89%
17.06%
14.87%
4.43%
9.80%
11.80%
11.22%
7.12%
12.54%
23.92%
17.68%
18.63%
13.86%
14.04%
4.80%
11.60%
12.60%
9.66%
15.21%
13.13%
12.61%
8.32%
4.12%
24.42%
18.47%
4.45%
4.73%
12.79%
12.83%
13.92%
7.09%
8.66%
9.21%
17.76%
14.82%
9.04%
12.50%
16.36%
KOs
consistent KC
highly variable KC
set-specific increased KC
set-specific inc/dec KC
set-specific decreased KC

# What to focus on moving forward?

- Generate re-usable method for profiling strain level variation
  - Phyeco genome clusters
  - Genome-cluster profile
  - Mapping reads to present genome-clusters
  - Predict CNVs and SNPs
  - Build SNP and CNV trees

- Generate data resources
  - Up-to-date database of Phyeco genome clusters
  - Identify “core” genome for many gut species
  - Co-variation of genes to identify genomic islands

- Improve previous analysis
  - Way more reference genomes and metagenomes
  - Look at all genes, not just annotated ones (KEGG Orthologs)
  - Collect more information (SNPs, phylogenies)
  - More careful analysis

- Address unanswered biological questions
  - Biogeography of strains
  - Comparative analysis of core across species
  - Compare gene flux across species
  - Co-variation of genes to identify genomic islands

- Perform experimental validation
  - KO core genes?