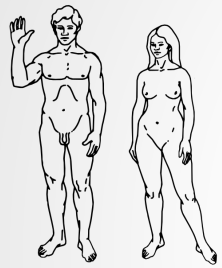
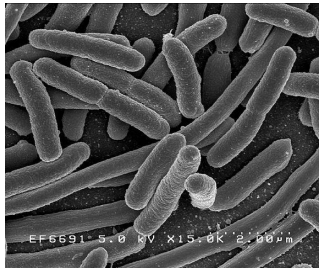
Infectome & Novel Bug Pipeline

Mark Kowarsky, Quake Lab

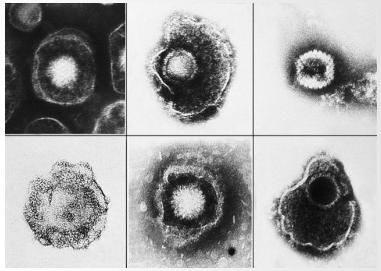
“Human” DNA



- 10^{13} cells
- Gb genome
- 10^{22} nucleotides



- 10^{14} cells
- Mb genome
- 10^{20} nucleotides



- 10^{14} virions
- kb genome
- 10^{17} nucleotides

- ~1% non-human, sequence cell-free DNA from blood

Pipeline overview

Samples

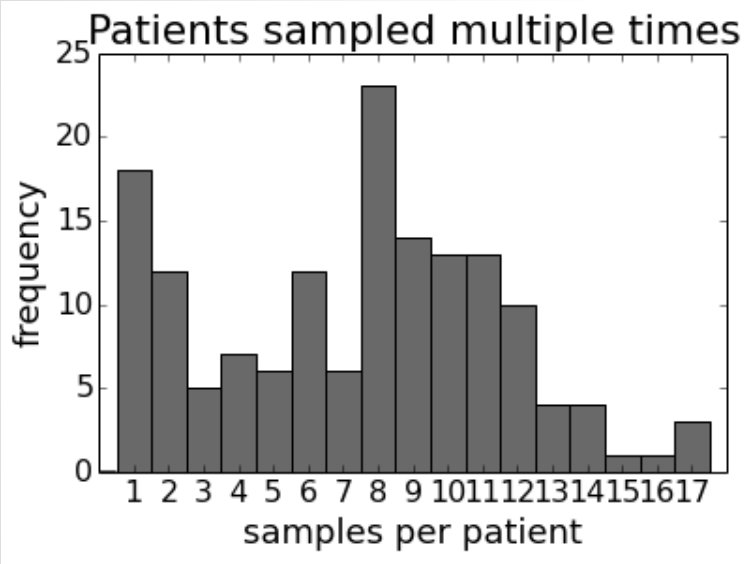
1. Preprocessing
2. Host removal
3. Infectome [known]
4. Finding the gold [unknown]

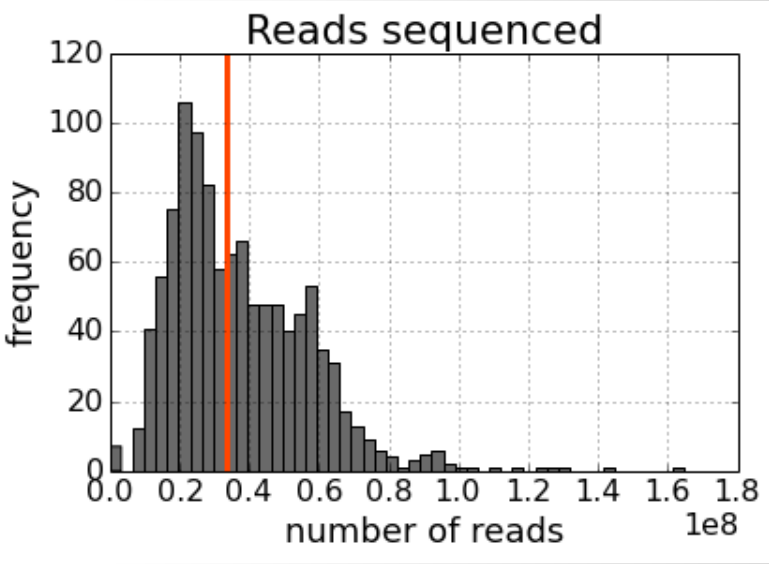
Sets of samples

1. Aggregated assembly
2. Realign for coverage
3. Identify contigs

Samples

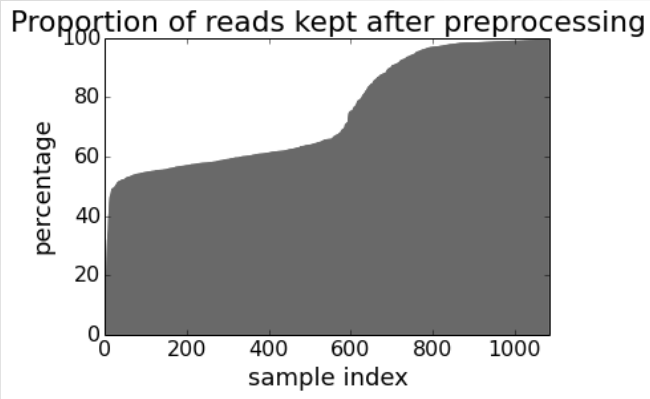
- 400+ people
- 1000+ samples
  - 642 heart transplant
  - 389 lung transplant
  - 73 bone marrow transplant
  - 18 chronic fatigue syndrome
- Illumina sequenced, 2x100bp or 1x50bp

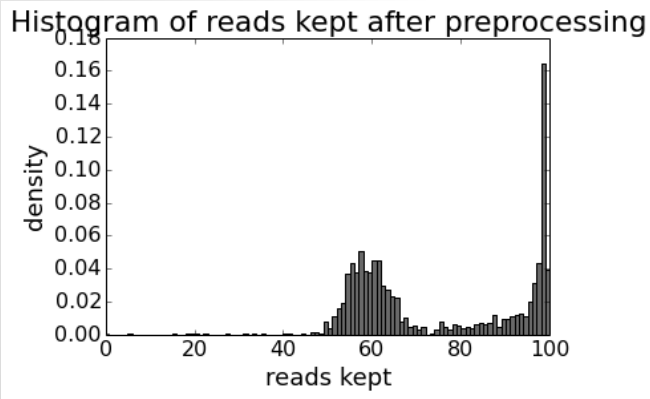

Patients sampled multiple times
frequency
0
5
10
15
20
25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
samples per patient

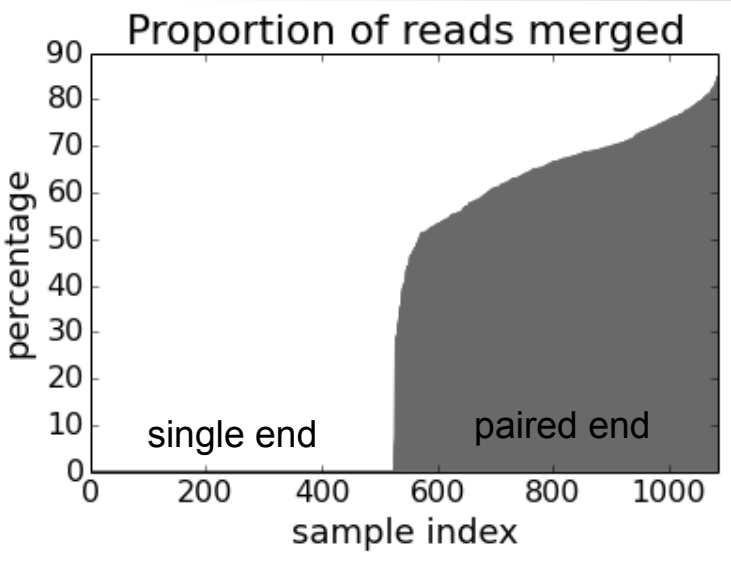

Reads sequenced
frequency
0
20
40
60
80
100
120
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8
number of reads
1e8

1. Preprocess

- Quality check (fastQC)
- Trim adapters
- Merge overlapping reads
- Set of ‘good’ reads to use

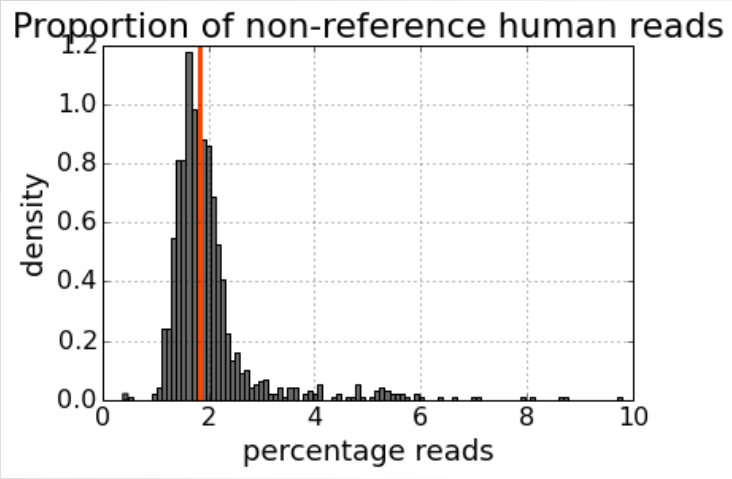

Proportion of reads kept after preprocessing
percentage
sample index
0
20
40
60
80
100
0
200
400
600
800
1000

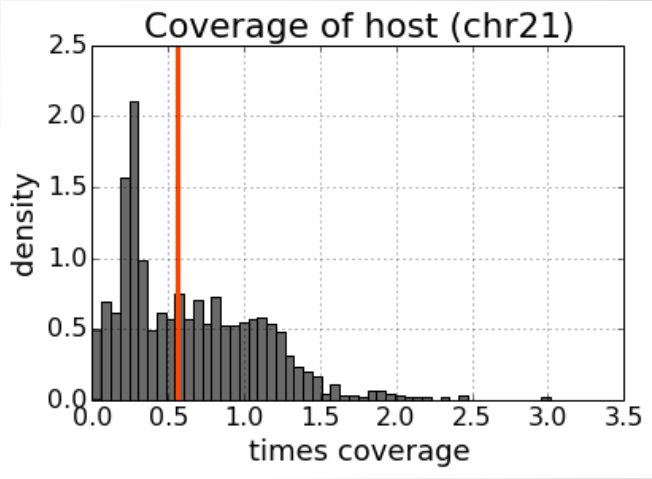

Histogram of reads kept after preprocessing
density
reads kept
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
0
20
40
60
80
100

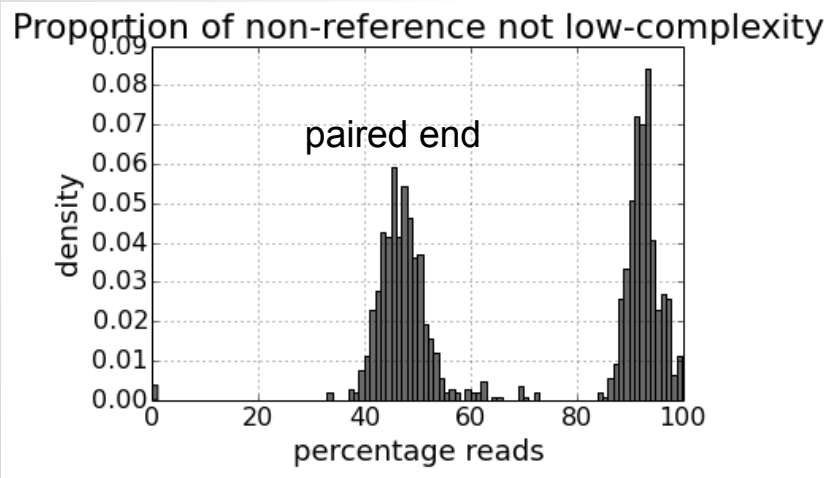

Proportion of reads merged
percentage
sample index
single end
paired end
0
10
20
30
40
50
60
70
80
90
0
200
400
600
800
1000

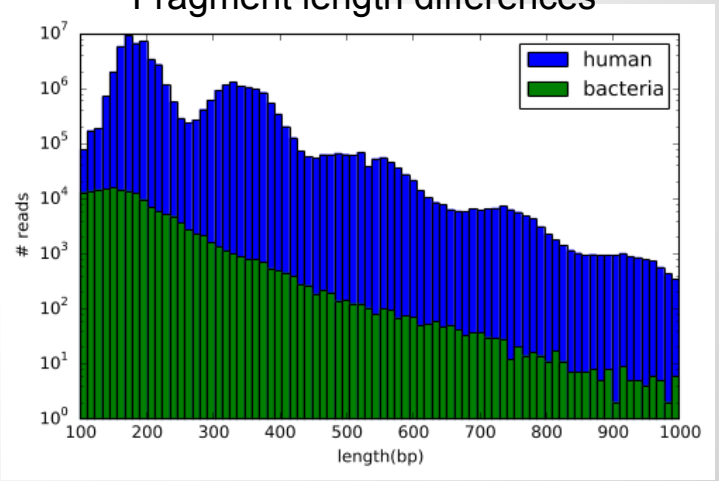
2. Host removal

- Align to hg19
- Align to phiX
- Remove low complexity reads
- ~2% of reads remain


Proportion of non-reference human reads
density
percentage reads
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10


Coverage of host (chr21)
density
times coverage
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Proportion of non-reference not low-complexity
paired end
density
percentage reads
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0
20
40
60
80
100


Fragment length differences
human
bacteria
# reads
length(bp)
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
100
200
300
400
500
600
700
800
900
1000

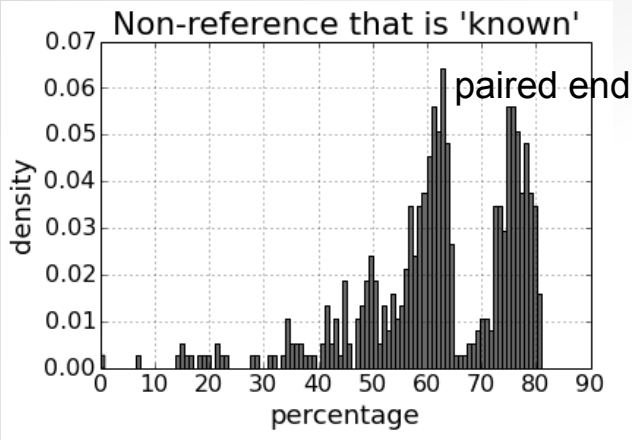
3. Known infectome

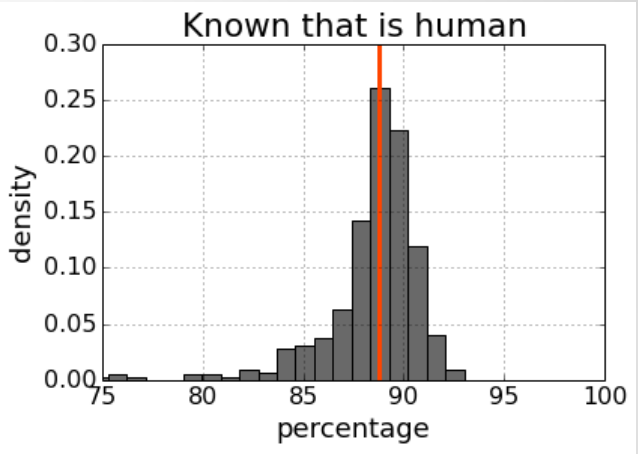
- BLAST filtered non-reference reads against bug database
- Use GRAMMy* to estimate abundance
- See Lance’s talk for details

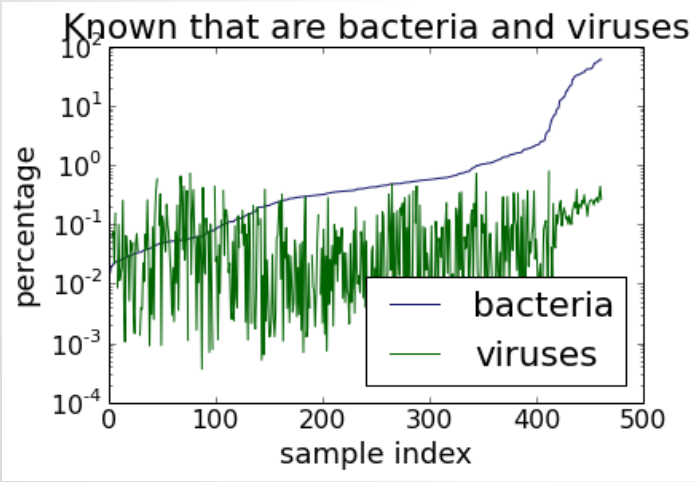
* Genome Relative Abundance using Mixture Models - https://bitbucket.org/charade/grammy/wiki/Home

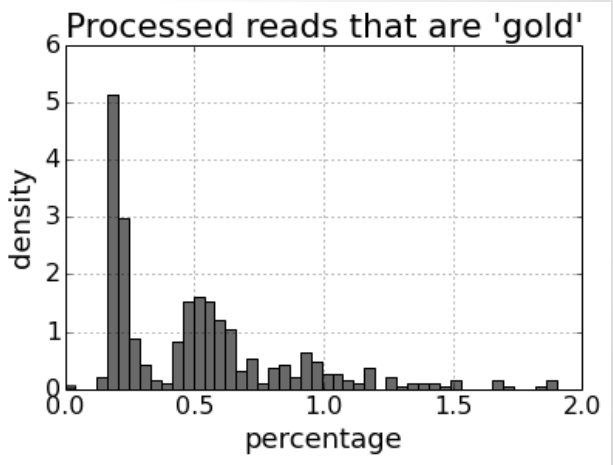
4. Novel bugs - “gold”

- BLAST against NCBI’s NT database
- Filter reads that have excellent alignments
- Remnants are ‘gold’


Non-reference that is 'known'
paired end
density
percentage
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0 10 20 30 40 50 60 70 80 90

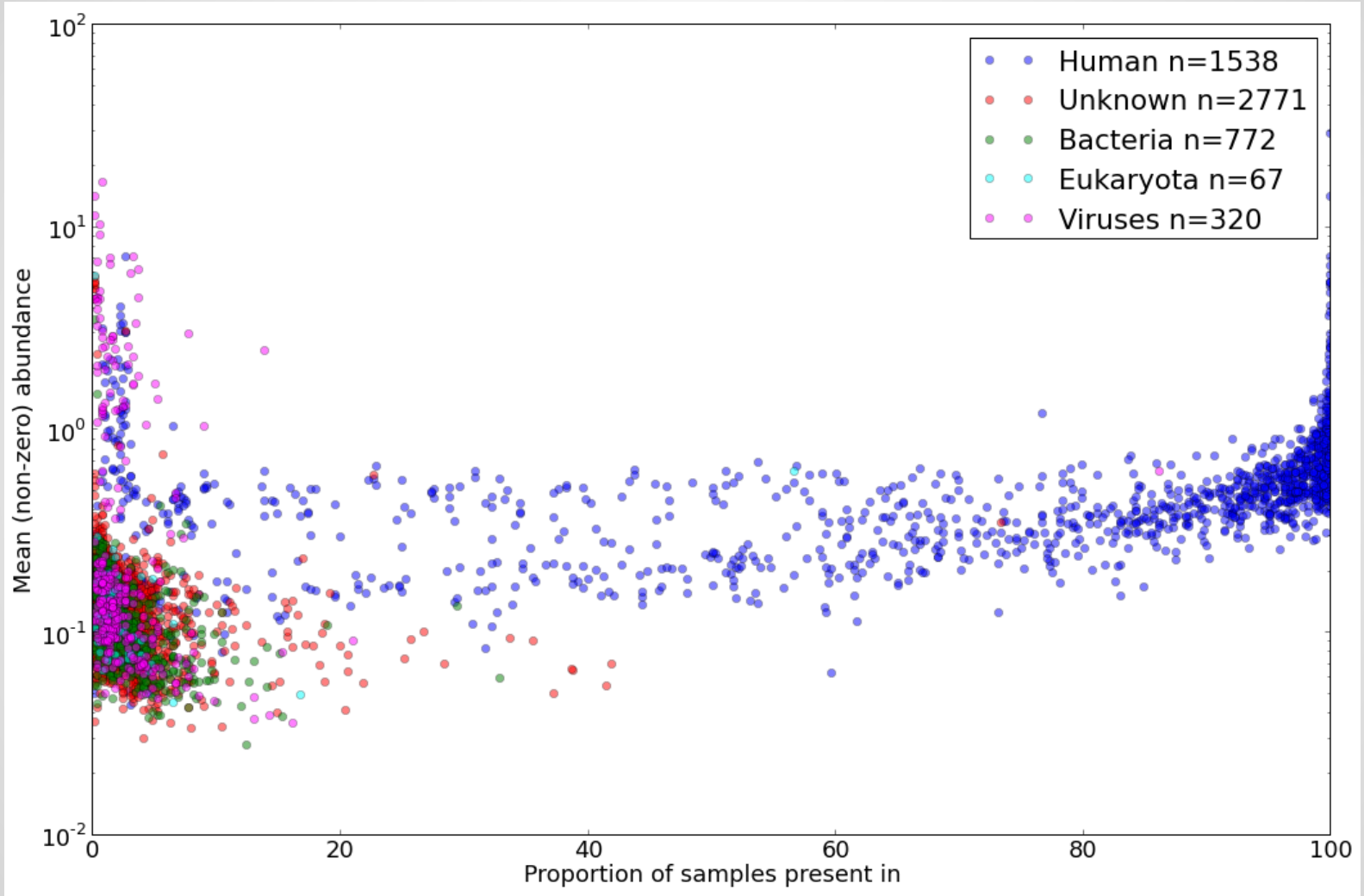

Known that is human
density
percentage
0.00
0.05
0.10
0.15
0.20
0.25
0.30
75 80 85 90 95 100


Known that are bacteria and viruses
percentage
sample index
bacteria
viruses
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0 100 200 300 400 500


Processed reads that are 'gold'
density
percentage
0 1 2 3 4 5 6
0.0 0.5 1.0 1.5 2.0

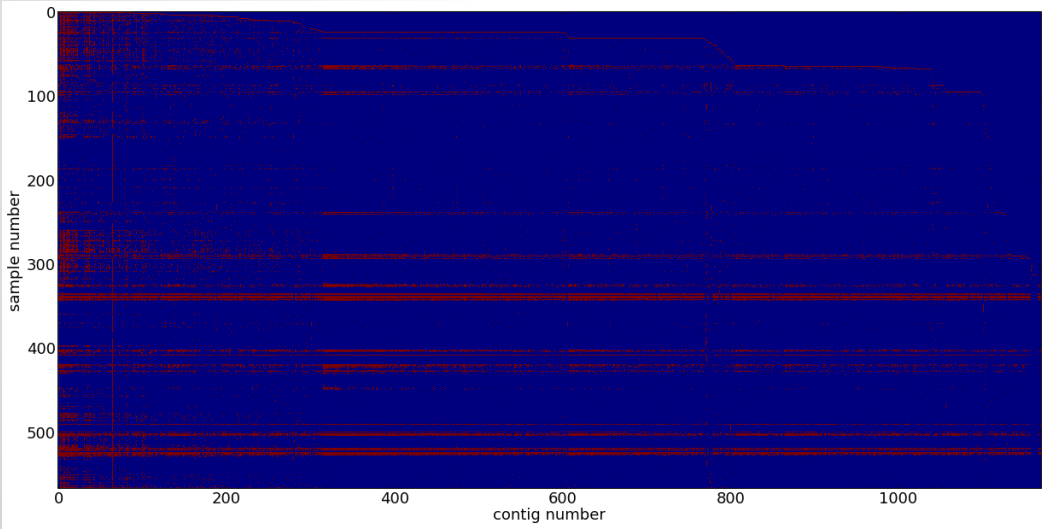
Sets

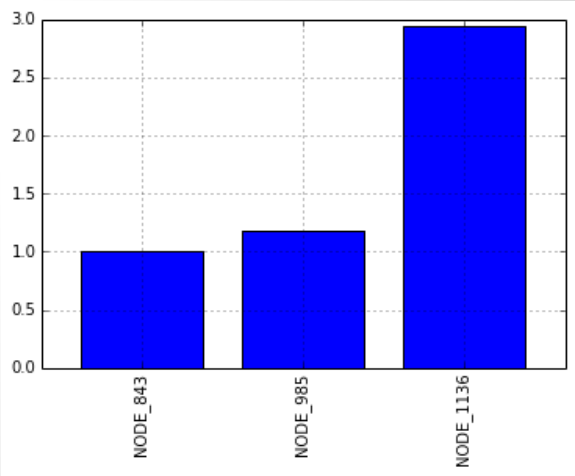
- Aggregate reads from groups of samples, e.g.
  - same transplant type
  - same patient
  - hospital
- Assemble
- Realign
- Identify


Human n=1538
Unknown n=2771
Bacteria n=772
Eukaryota n=67
Viruses n=320
Mean (non-zero) abundance
Proportion of samples present in
10^2
10^1
10^0
10^{-1}
10^{-2}
0
20
40
60
80
100

Clustering

- Correlate novel contigs
- Find clusters highly associated





Further work

- Finish running pipeline on data (~1-2 weeks)
- Find good bug candidates
  - high coverage
  - clinical correlation
- PCR to validate assembly

Acknowledgements

- Mickey Kertesz
- Lance Martin
- Iwjin De Vlaminck