Shotgun Metagenomic

SLU intro to genomics course 2020

 @metamoritz

 @moritzbuck

 @metamoritz

Before the computer

- Get sample
- Extract DNA
- Prep library
- Send to Sequencing facility
- ????
- PROFIT

What you get from your shotgun?

- **Load and load of reads**

moritz@malas-castle: ~/repos/moritz/0038_GarbageSeq 211x53

```
@A00605:34:HHVWNDSXX:1:1101:7265:1016 2:N:0:ATTCAGAA+CAGGACGT
GAGACCGTGCTGGCGATCGTCGAAGGATGCACGGACGGCGTACCTGACGCGCTGGGAAGAAAGGCGGACTGGAAAGTACGCAAGCTGCGCTACCTCGACCACCTTGCGACGGCTTCTGACGCAATCATCTTGGTATCGGGGTCGGATAAGC
+
FFFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFFFFFFFFFFFFFFF:FFF:FFFFF:F:FFFFFFFFFFFFF,FF,FFFFF,FFFFFFFFFFFFFFFFFF:,:FFFFFFFFFFFF:FFFFFFFF,:,,FFFFFFFF,:FFFFF:,,FF
@A00605:34:HHVWNDSXX:1:1101:8766:1016 2:N:0:ATTCAGAA+CAGGACGT
CTGCTGTGATGAATAAATTTCTGGTATTTCTGGAGCAAAGAGAAGGTATTGTCAAACCATCCTCTATTTCTGTCTGGAACAGGGTTCAGGGGCTTGCGGCTTTAGATGGAAATTCGGTTGTTTCAGGAATTCTTGCCGGGCCGGTTGAAAT
+
FFFF:FFFFF:FFFFF:FFFF:FFFFFFFFFFFFFFFFFFFF:FFF:FFFFFFFFFFFFFFFFF:,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,:FFFF:FFFFFFFFFFFFFFFF:F:FFFFFFF,FFFFFFFFFF:FFFF
@A00605:34:HHVWNDSXX:1:1101:15130:1016 2:N:0:ATTCAGAA+CAGGACGT
AGTTTGAGTTCAAGTTCGCCCGATCTACCGATATCGTCGGCCACATGGCCTTGCATGCCTATGTCGAGGCTCAAGACACGAACGATATGGACCTCTTCGTGGGAATTCAAAAGCGCAATGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGT
+
FFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFF:FFFFFFFFFF::FF:FFF,F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFF:FF
@A00605:34:HHVWNDSXX:1:1101:16251:1016 2:N:0:ATTCAGAA+CAGGACGT
AGCGAGCTGTTGCAGTCCGCCGTCAAGAGCAAGGTAAAGAGATCGGAAGAGCGTCGTGTTGGGAAAGAGTGTACGTCCTGGTGTAGTTATCGGGGGTGGGCGGGGCAGTGAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F:FFFF,FF,,FFFFFFFFF:FFFFFFF:FF::::,,F,:,FF:FF,,F,,,:,,,:,:,,,,,,,::FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00605:34:HHVWNDSXX:1:1101:29234:1016 2:N:0:ATTCAGAA+CAGGACGT
ACGGCCGGTTCCACCCCGGCGACTCCGGAACTCTGGTACCCCTGGACTTCGACATCCCCCCGGGCACCACCGCCCTCGCGGTTCGGTTCCGGTATTCGCCGCTGGGCAGTCGCGACCGCGAACGCAACCGTATCGCCGTGCAAGCGGCCGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F:FF:FFFFFFFFFF:FFF,FFFFFFFF
@A00605:34:HHVWNDSXX:1:1101:29740:1016 2:N:0:ATTCAGAA+CAGGACGT
GGAGATACCAGCGGCAATGCCGATGGGGCGCCCGCTGATTCAGCCAGAAAAGAACGTGCCACGTTCTTTTCTGGTCGGCAAGCCAGACACTGCCGACAAGGTTGTTGCAGATCGGAAGAGCGTCGTGGAGGGAAAGAGTGTACGTCCTGGG
+
FFFFFFFFFF:FFFFF:FFFFFF:FFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF:FFFFFFFFFFFFFFF:F:FF:FF:F:F:FF:FFFF,F:FFFFFFFF,FFFFF:FFFFFFFFFF:FF,FF,
@A00605:34:HHVWNDSXX:1:1101:3079:1031 2:N:0:ATTCAGAA+CAGGACGT
ATTTTGGCTGTGGCTTTGGGGTCGAGCGCTGAGGTGGGCTCATCGAGCAGCAGCACTTCGGGCTCAACGGCAAGCGTGCGGGCGACACATAAGCGCTGCTGCTGGCCGCCGCTAAGCCCGAGAGCGTTTTTGTCGAGGCGGTCGCACACTT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00605:34:HHVWNDSXX:1:1101:3513:1031 2:N:0:ATTCAGAA+CAGGACGT
TCATCCTGGTGGCCATACTCAACACCGTTTCACACTCCCGACACCACCACAATACGGTGTAGCCTCTCCAGAGCTTACTCTTTGAAAACGTTGCTCGTTTGAGCGGCGTTTTCAAGTTAGTAAAAGTATATATTTAGGCGACTGCAGCCTT
```

Dream-quest of unknown metagenome

From an environment with the power of genomics we want to know:

- Who is in the environment? (taxonomy)
- How many of each? (abundance)
- What can they do they? (genetic potential)
- What are they doing? (expression analysis)
- What do they eat? (metabolism)
- Where do they come from? (evolution and ecology)

Two main roads

Assembly based :

- Assembling
- Binning
- Annotating
- Slow, tricky, more specific

Assembly-free :

- map/classify reads
- Direct annotation
- Heavily database reliant
- Faster, easier, more noisy

The tutorials

- **Let’s assemble some metagenome!**
- **And get some genomes out of it!**
- **Check how good they are!**
- **And then let’s compare some metagenomes directly assembly free**

Assembly based approaches

Turning the reads into genes and more! (aka The Contig):

- **Full length coding sequences**
- **Pieces of genomes**
- **Or anything else made of DNA**

Increases specificity of previous approaches, but might lose some sensitivity (e.g. not everything assembles)

Linking reads to contigs

The dark art of mapping.

Normally, quantitative information gets lost during assembly, we need to “map back”. Read based approaches “map” to databases

Mappers :

- **bowtie/bwa/bwa2**
- **Bbmap**
- **Squeeze…**

RNA/DNA euk or prok, all different!

=> pile-ups!

The gene atlas approach

If you only care about genes:

- **Predict genes and annotate genes:**
  - Functionally
  - Taxonomically
- **Gene clustering:**
  - By similarity
  - Into COGs
  - Into families
- **Quantify and analyse diversity**

The king discipline

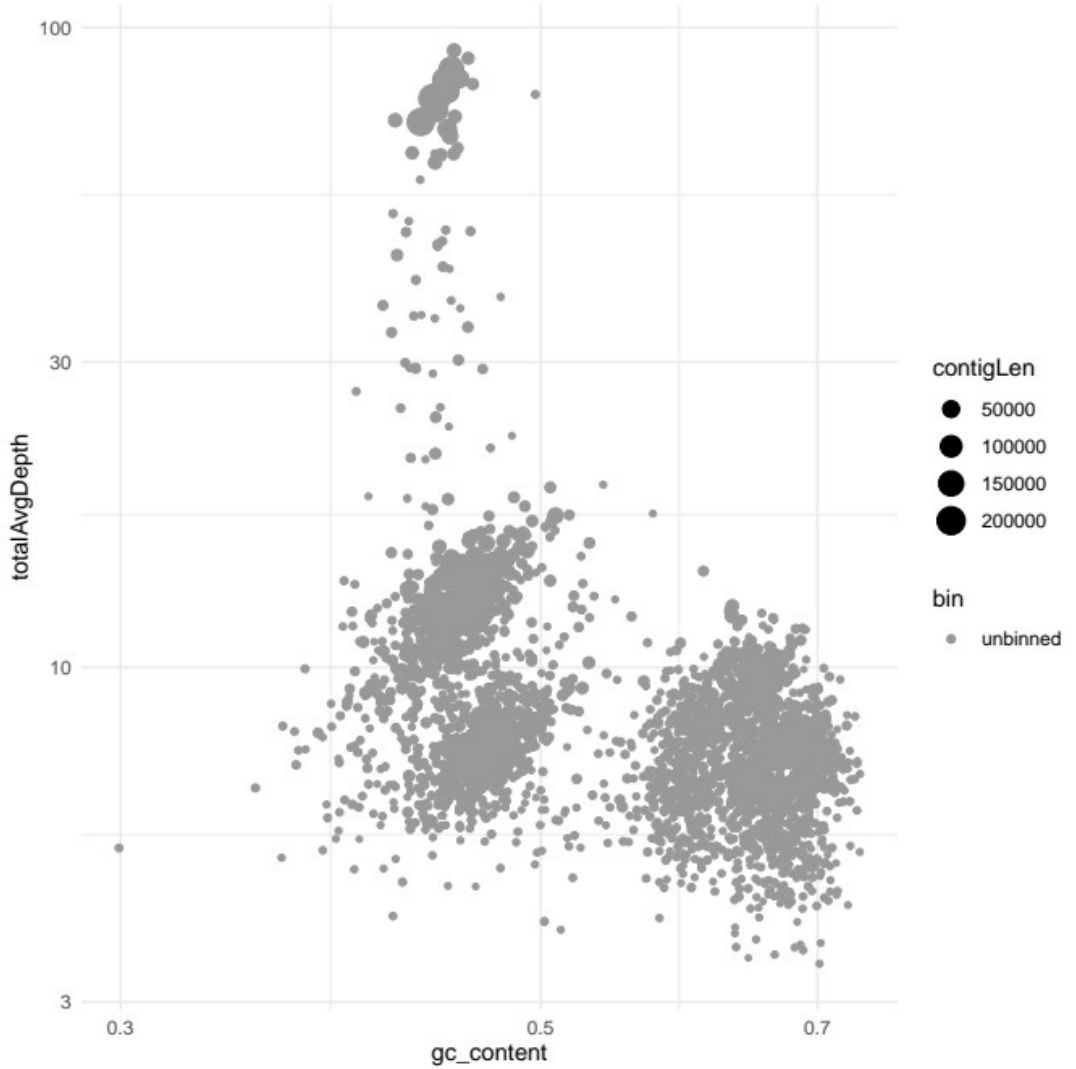
Genome-resolved metagenomics:

- **Cluster contigs into bins**
- **Identify possible genomes**
- **Quantify completeness/redundancy**
- **Assign taxonomy**
- **Have fun?**

So vamonos!

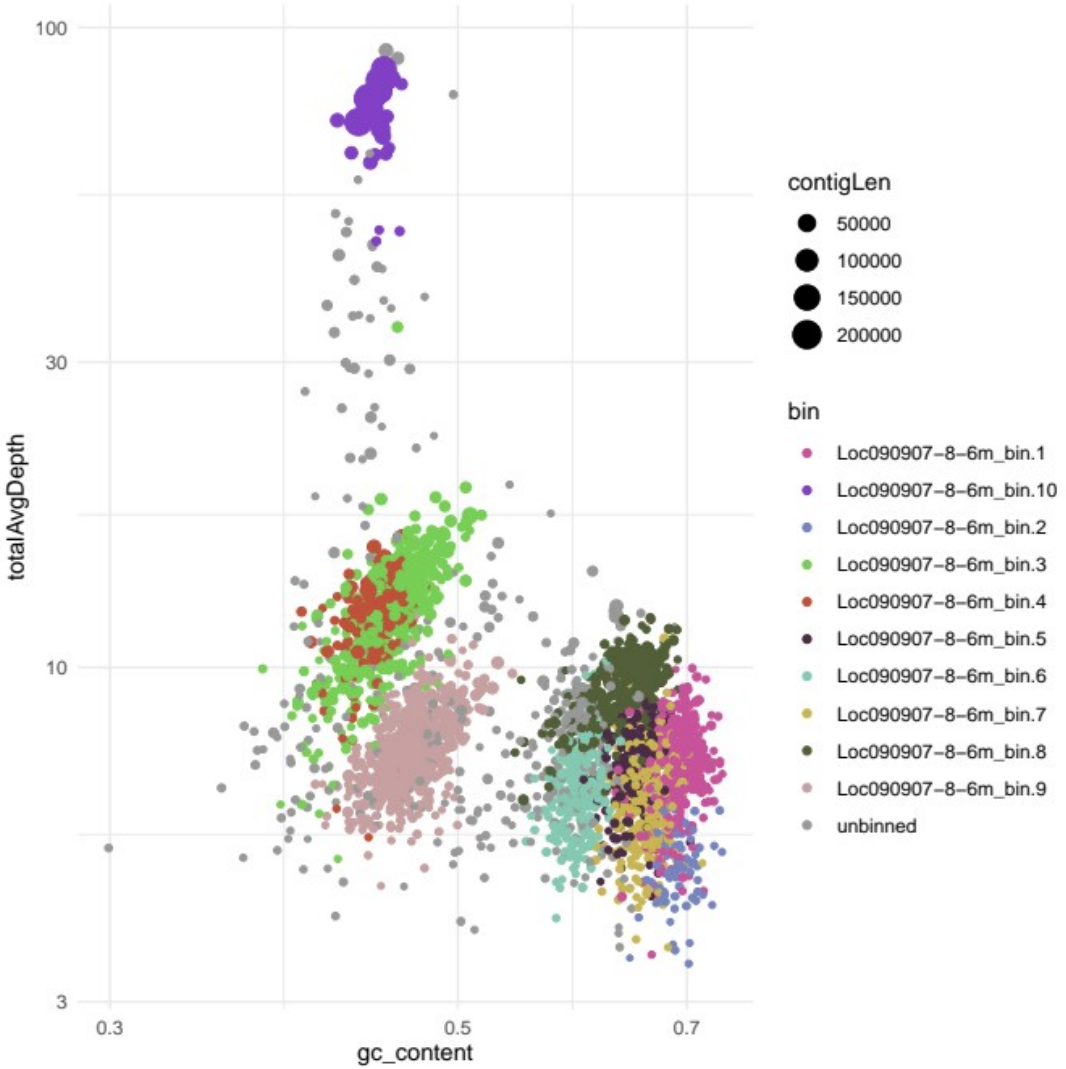
Clustering contigs

Binning:



Clustering contigs

Binning:



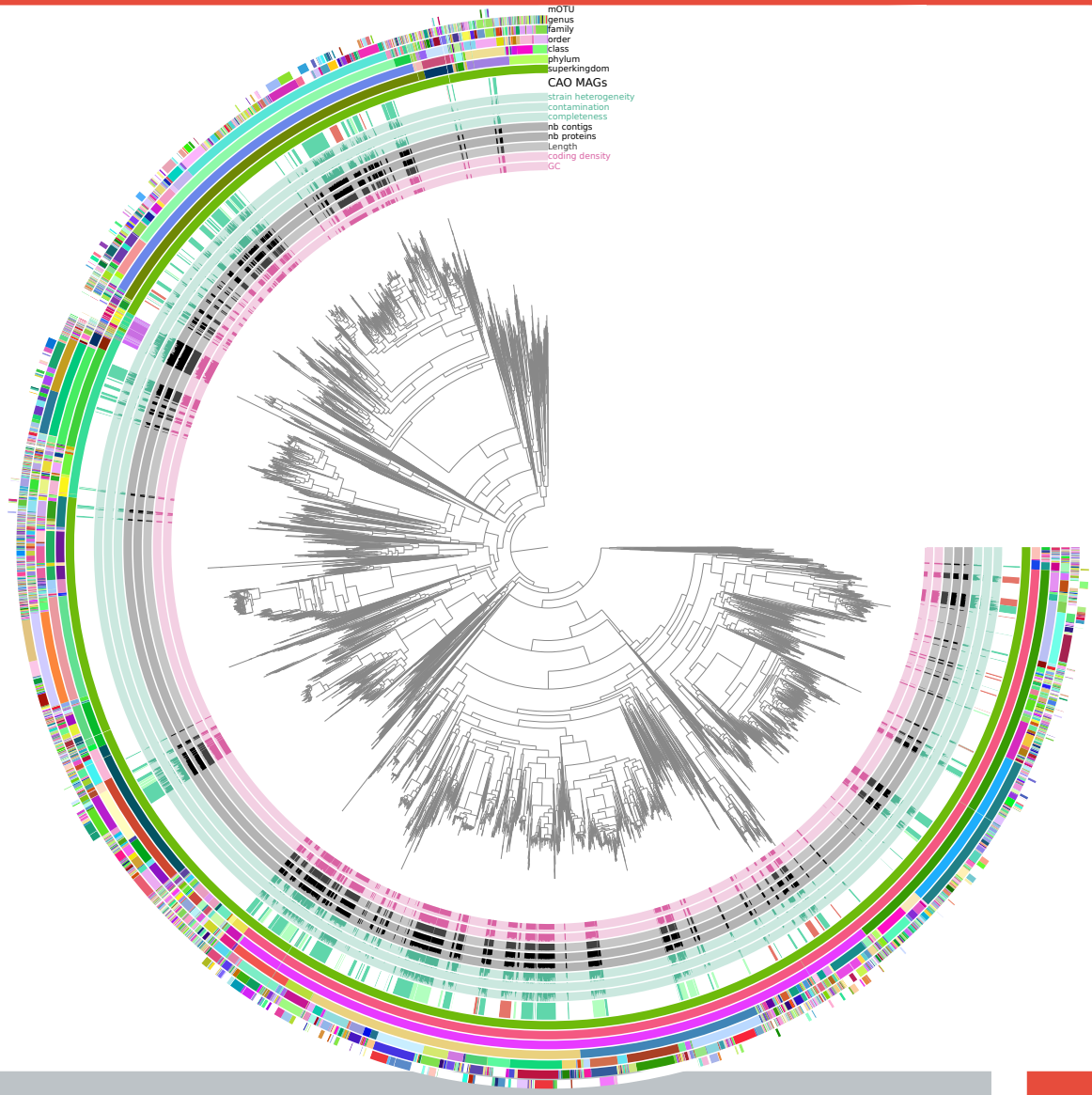
What is a genome?

How to recognize a genome:

- Size and coding-density
- Similarity to known genome
- Universal single copy genes

```
>k141_3
GCCATCCGAGGTGGGCCGGCCGCTCCGCGGGCGGCGGATTGGAGAGAGAGGCGTGGCCGG
CGAAGCCGCGCGCGGAGCGCACGGCCCACCCGGGGACACCGGACCCTGCACGAAACTGAG
ATGCGCCACCCTCACGCAGCGGCCGACCCGTGGCGAAATCAGCGATCGAACGGCGGGGCG
TTTTTTCACGCGCCCATCTCAGGTCGGCGGTTGGGAAACCGCCGGTCCTCCTCCGCCGCT
CACAGCGGCCGGTTGGAGAGGACGACGTCGTCCACGAACAGCGTCTGCGCCTGATTGTGG
CCGTTGTTGGCGGTGAGTCCGACCTGCATCCGGTCGTAGACCGTGCGGGAGCGCGGCAGC
GTTTTGCCGGCTTCGTCGAGCACCTTGACGCCGTCCTGCCAGACTTGCATCCGGCCGTCC
GCTCCTTCGGAGAGGCGCAGGTGGATTTTCACCTCGACCCACTTGTCCTTGGGGAACTTC
GGCGCACCGACGACCGCGCGGAATTTTTTCGCCGTCCACCACTTGCCCAGATCGGAGGCG
ACCATTTCGCCGTTCTGCAGAAAGAGGCGGCGGCCGGGCGTGTTGCGCAGCGACGTCGAC
TCCAGATCCCAGAGAAAAATCGAGGACGATTCGGTGCCGCCTTCCAGCCAGAACCAGGCG
CTGAACCACACATGGTCGCCTTTCACGAACCGCAATCCCTCGCGCTCGATATCTGCCTCG
GAGGCCCGCCGGCCGTCGAACGGGGCCGCGACGCATTTCAGAGCGTTGGTTCCCGAGTGA
ACCCGTTCGGTGGTCAGCGTAACCGAATTGGACGCGAGACCCGTCGTGGACTCGATTTCC
AGGCCCTGCCACCGGCTCGAATCCGCGGGGAACAAATCTTCGGTCTTCGCCGAACCTTCG
AACCCATCGCGAAAGACGAATGGCGCTGCCGGGGCCGGCAGCGGCCGCGCCACCCAAAAC
CACCACGCCGCCCCCGCACTCACGGCCAAGGCGGAAAATGCTCCCCCGGCGACCAGCCAA
CTCACCCAGCGGGGCGTTGCTCGCGGACTCGGGGAGGGCGGAGTGGTCGGCGTCATCAGG
GTATGGGGGTCTTAGGAACAAAAGCGCTGGGCGGAAAATCGCAACAAGTGTTCTGCGGGC
CATCGCAGCTCTTTCTGTTTTCCCGGCTTGCGCCGCCCGCGGCTCTTTGCGGCCATGTCT
CACGCATGAGAATCACCGGACTCGCCCTCGTTCCGCCGCCGACCGCCGCTGACCTCCCGC
AGGTCACTCCCGAGCTCCTCGCGTCCGTTCTCGCCCGTTACTCCCGCAGCAATCAGGGCA
TCGCCGCCATCCTGTCCAAGGTCGATCTCGCCAACCCCGACGCTTCGATCGACCGCATCC
TCAGTTTCGTCGACTACGGCCATGCGTCCATCGGCGGCCTCACCGGCGGTCTCGCGGTGG
CCCTCGACGATGTGTCGATGTGGCTGGCCTACAAAATTTTTGAAATCGCCACCATGGCCG
ACGGCCAGGAGTCGAGCACCCGCTACATCACGATGGACGCGGCGAACCTCCCGACCGCCG
CCGAGTTGGGGATCCCCGACGATCTGGCGCCGCGCTGGTCCGCTGTGATGGCCAAGTCCT
TCGCCGCCTACCACGCGGAGTATACCCGGCTCGATGCCCTCGCGACCGCCCACCCCGGCC
TCGTCCGCCTCCCGCCGGACGCCAAACCCGCCGTCGTCACGCGCCTCCGCAAGAACTACG
CGCTCGACCGCGCCCGCTACTTCATCCCGCTCGCCACGCGCACCAATGTCGGCCTCGTGC
AGTCGTCGCGCATGTGGGCGATGACGGTGAAGCACCTCGATTCGCTGCCGCACCCCGAGG
CCCGGGCGGCCGCGCGGCTGATCCGCGACGAACTGCTGAAGCTTTCCCCGCGGCTGATGC
GCCACAGTTCGGCGGAACAGTCTTACACGGCCCAGGCCGCCCAAGAGCTGGCGACCTCGT
```

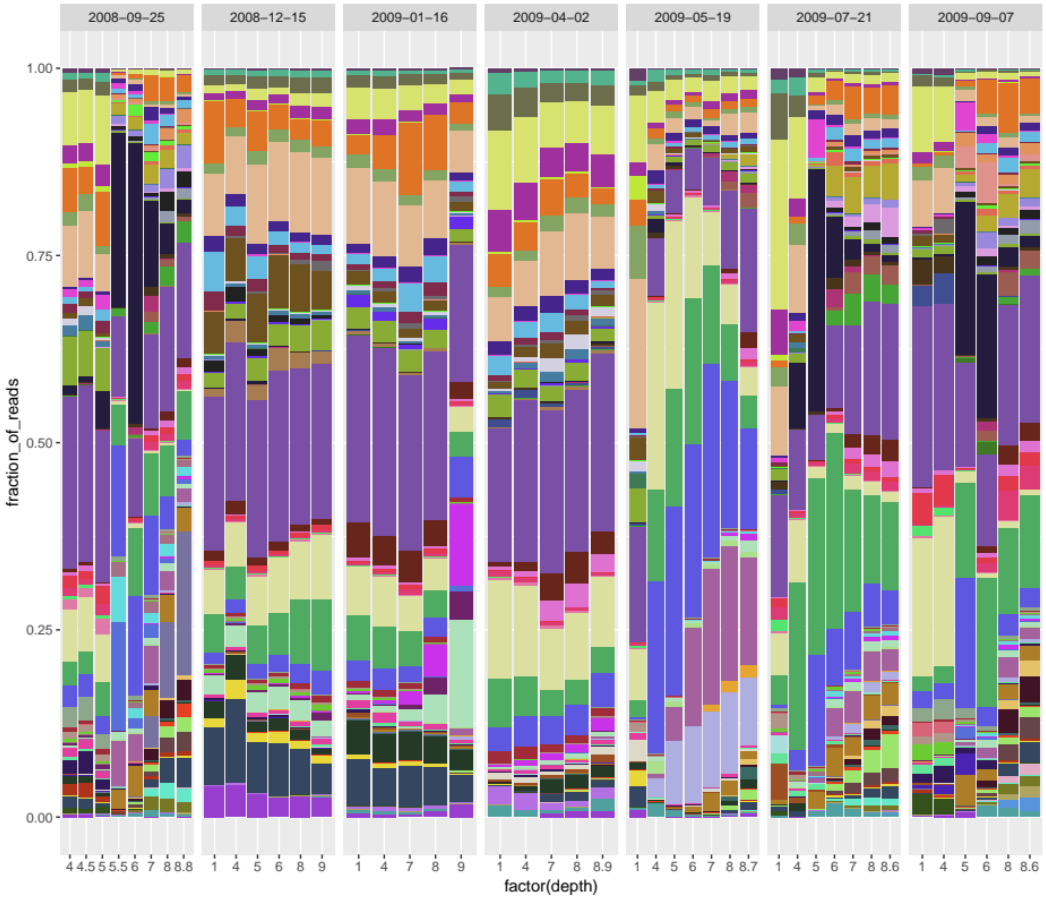
Putting it into phylogenomics context!



Quantifying it

Mapping read to the bins:

- Statistics
- Diversity metrics
- Network analysis



What are those reads?

The inventory approach:

- **Annotating the reads directly**
- **Heavily relying on databases**
- **Good for well known microbiomes (western European poop, agricultural soils)**
- **Dangerous for others as methods tend to overclassify**

What are those reads?

- **Taxonomically :**
  - Kraken
  - Kaiju
- **Functionally :**
  - Blast …
  - Megan
  - Humann