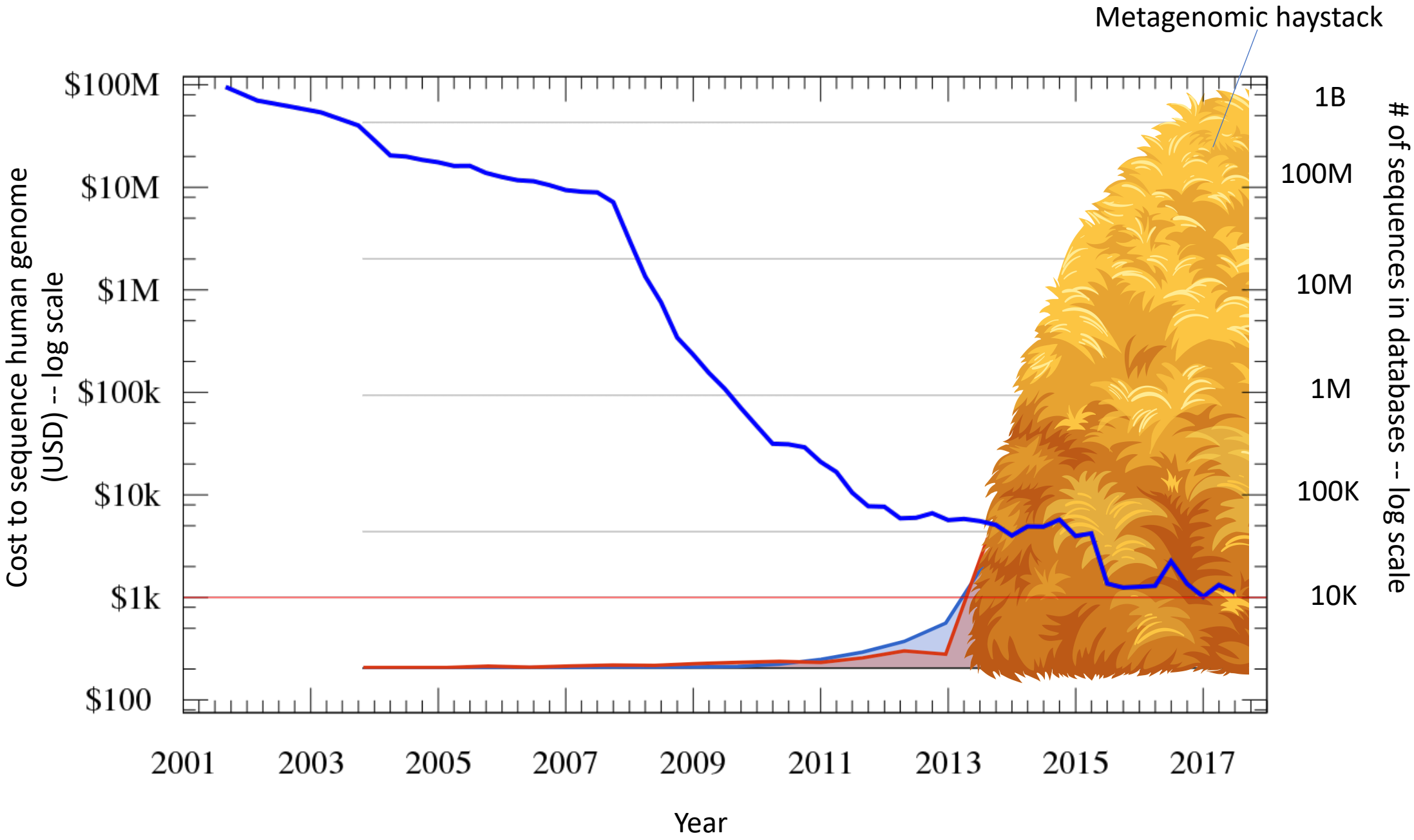
Part II

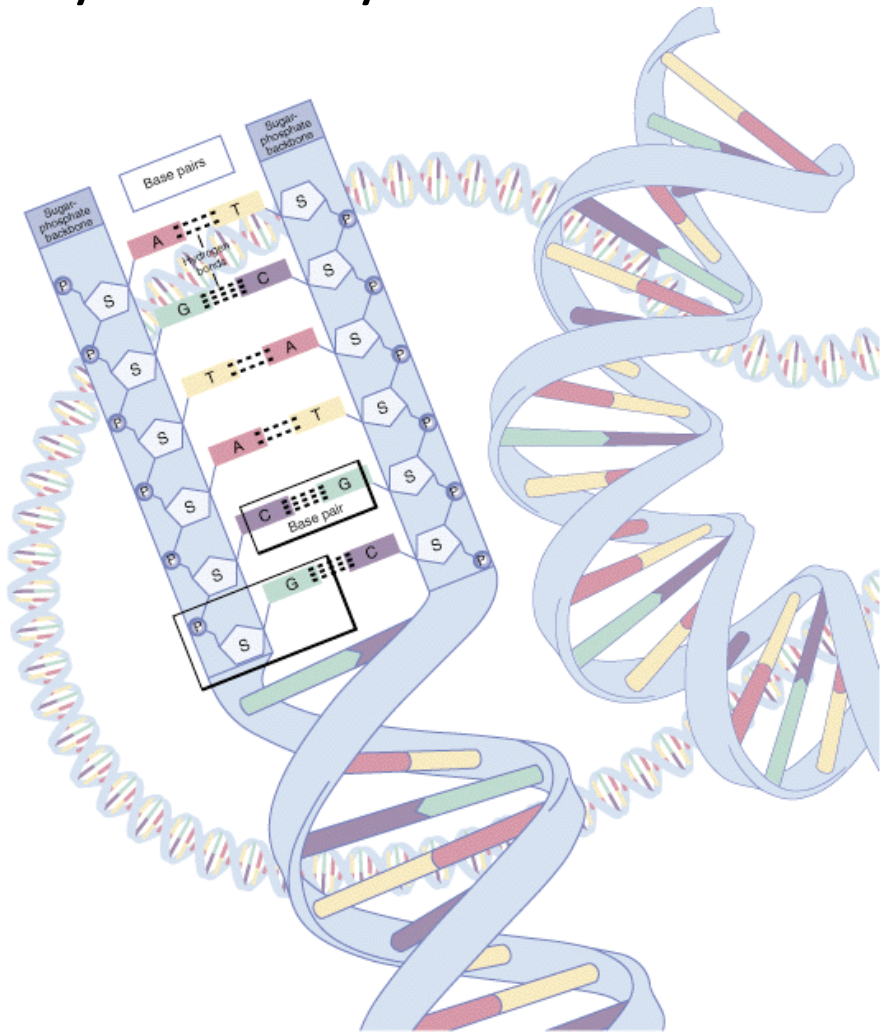
Strain-level analyses



https://en.wikipedia.org/wiki/DNA_sequencing

Documents of evolutionary history

- The **genome** of an organism contains its hereditary information and is encoded in its DNA.
- My research has focused on tracking evolution, one nucleotide at a time.
- How?



Zuckerandl and Pauling, “Documents of Evolutionary History” 1965.

How to detect strains from metagenomes ?

- **Known microbe/virus/pathogen in compex sample**
  - Gene expression array cards (TLDA)
  - PCR-based, targeted
  - Computational approaches
    - Signature based
    - **Marker based (MetaCompass)**
    - **K-mer based (more about this tomorrow!)**
    - Phylogenetic placement

- **Interesting genome discovery/unknown (pathogen) in sample**
  - DB might be helpful, likely not
  - Approaches:
    - Profile HMM for distant homology detection
    - **Assembly -> Annotation -> Systems based approach**

Should strain detection methods operate on:

1. Bags of kmers?
2. Bags of genes?
3. Bags of replicons?
4. Signature sequences?

Choose your own adventure: strain detection methods should operate on:

1. Bags of kmers?
2. Bags of genes?
3. Bags of replicons?
4. Signature sequences?

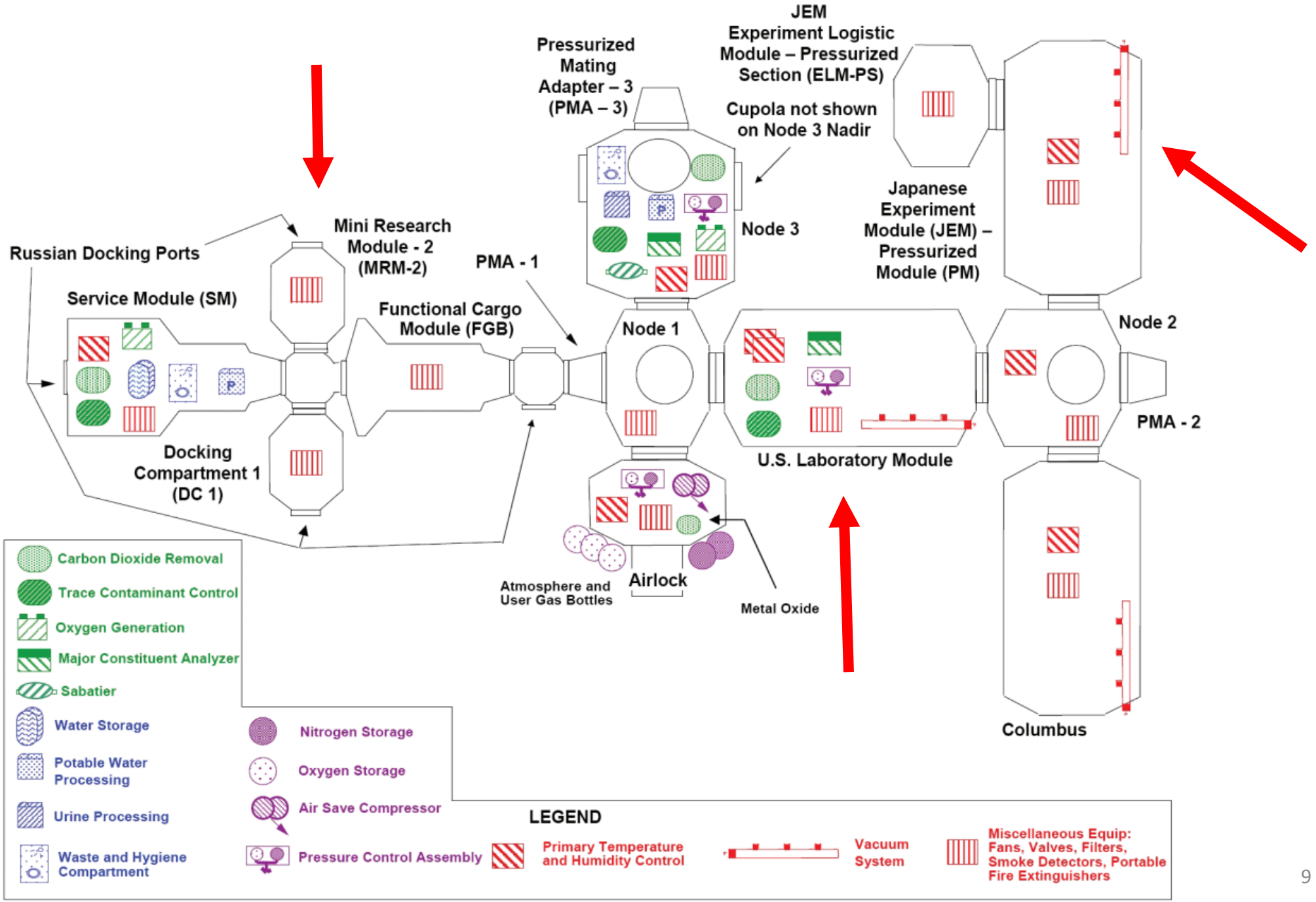
- http://tiny.cc/stamps19

Survey results:

- https://ql.tc/axwUgB

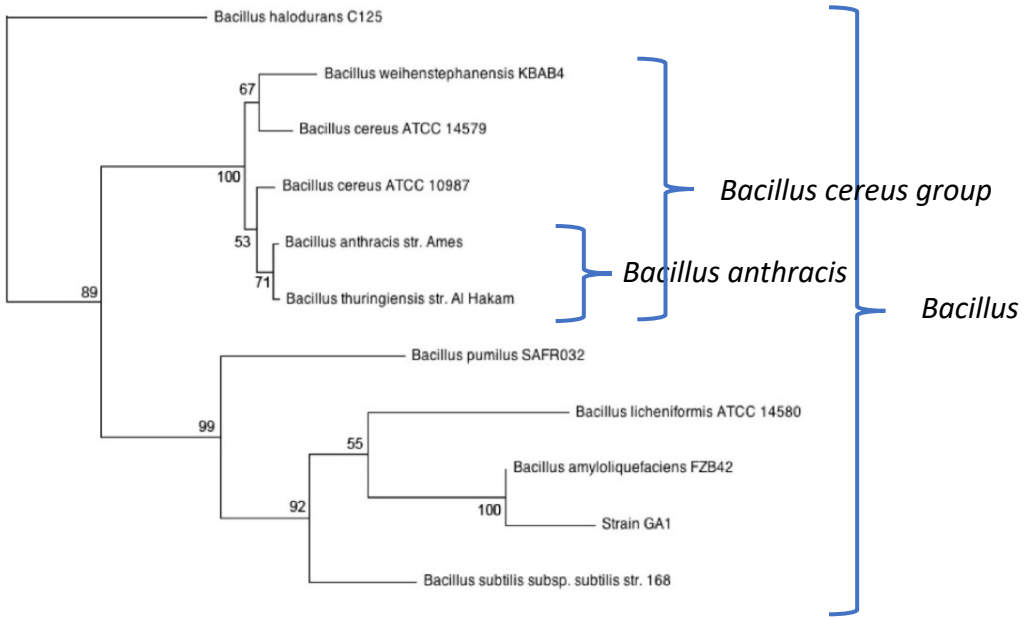
Investigation of *Anthrax* on the ISS

- Air filters from the International Space Station sent to NBACC for analysis
- Initial analysis suggested likely *Anthrax* presence
- Computing a multiple sequence alignment (MSA) allowed us to detect a single-nucleotide difference and determine Anthrax was not present on the International Space Station


JEM
Experiment Logistic
Module – Pressurized
Section (ELM-PS)
Pressurized
Mating
Adapter – 3
(PMA – 3)
Cupola not shown
on Node 3 Nadir
Node 3
Japanese
Experiment
Module (JEM) –
Pressurized
Module (PM)
Russian Docking Ports
Mini Research
Module - 2
(MRM-2)
PMA - 1
Service Module (SM)
Functional Cargo
Module (FGB)
Node 1
Node 2
PMA - 2
U.S. Laboratory Module
Docking
Compartment 1
(DC 1)
Atmosphere and
User Gas Bottles
Airlock
Metal Oxide
Columbus
Carbon Dioxide Removal
Trace Contaminant Control
Oxygen Generation
Major Constituent Analyzer
Sabatier
Water Storage
Potable Water
Processing
Urine Processing
Waste and Hygiene
Compartment
Nitrogen Storage
Oxygen Storage
Air Save Compressor
Pressure Control Assembly
LEGEND
Primary Temperature
and Humidity Control
Vacuum
System
Miscellaneous Equip:
Fans, Valves, Filters,
Smoke Detectors, Portable
Fire Extinguishers

Environmental surveillance on the ISS

- Microbial environment on the ISS is monitored continually
  - Routine testing performed in space
  - Sample sets sent back to Earth for more detailed analyses
- Samples from ISS air filters in American, Japanese, and Russian modules produce bacterial growth consistent with *Bacillus spp.*



- **Initial testing**
  - **Morphology/phenotypic testing indicates *Bacillus* isolates are part of the *B. cereus* group**
  - **16S rRNA and *gyrB* sequencing consistent with *B. anthracis*, though PCR tests for virulence markers are negative**
  - **MLST and WGS/Avg Nucleotide Identity analyses also point to *B. anthracis***

B. anthracis on the ISS?

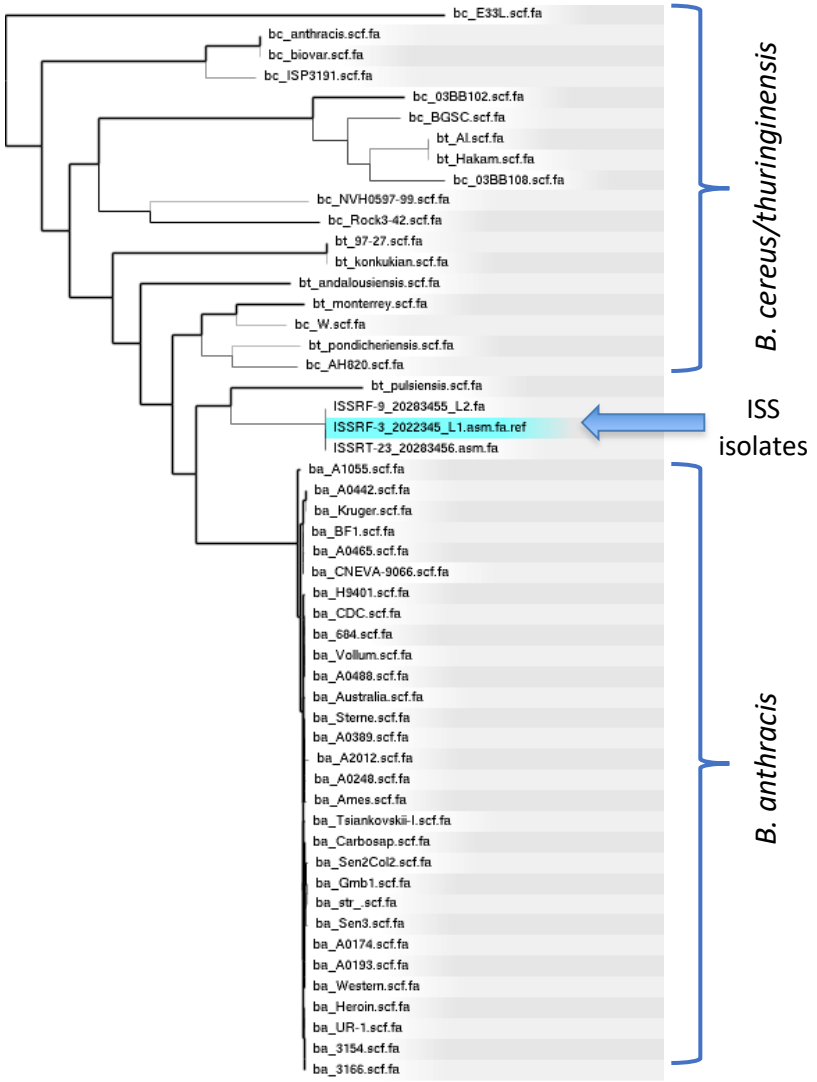
Bacillus anthracis + plcR biology

- Can a single SNP matter?
  - YES!

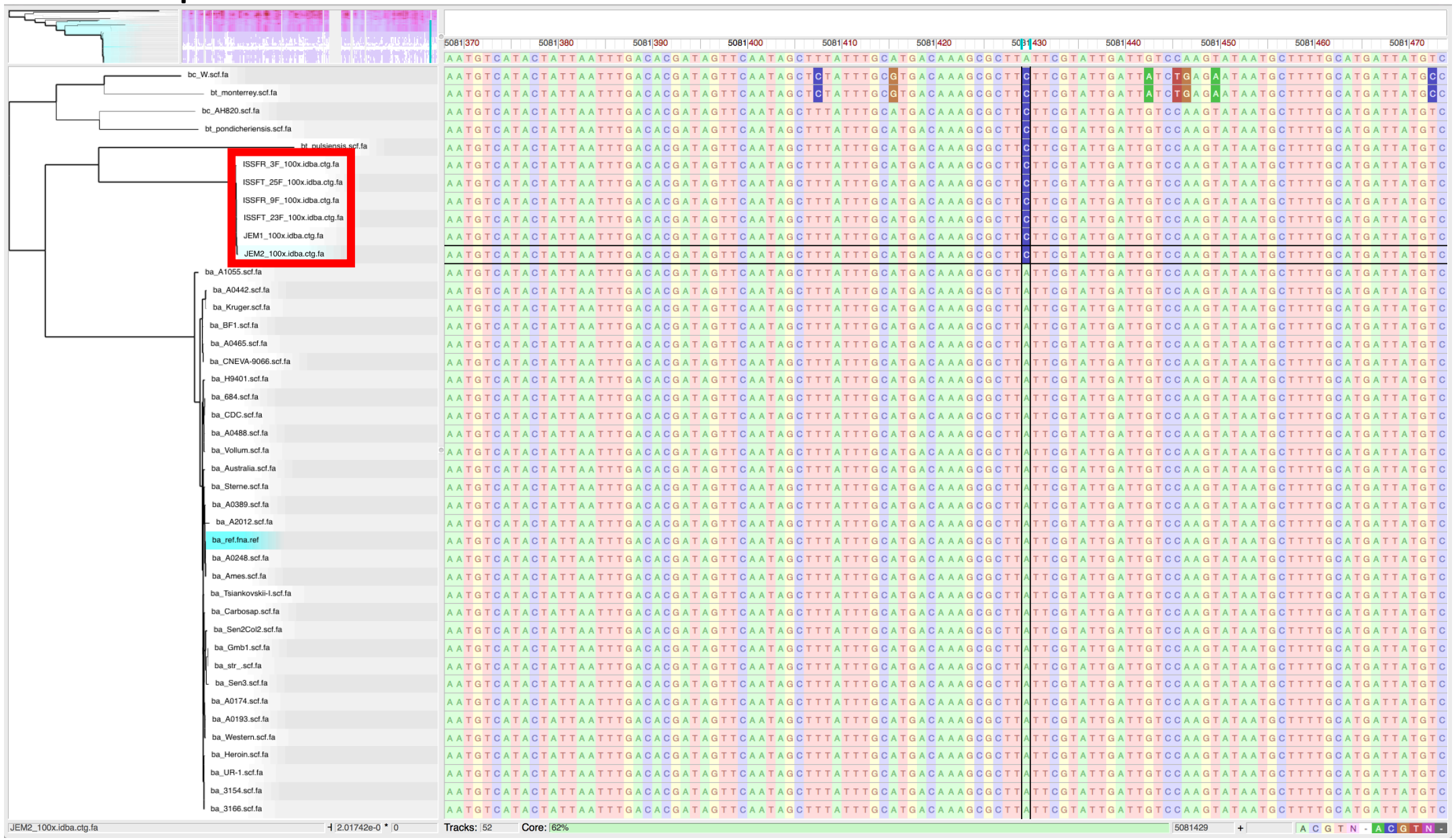
- plcR expression interferes with sporulation in B. anthracis.
  - Sporulation is a critical component of the ecology of B. anthracis
  - plcR nonsense mutation present in all known B. anthracis genomes

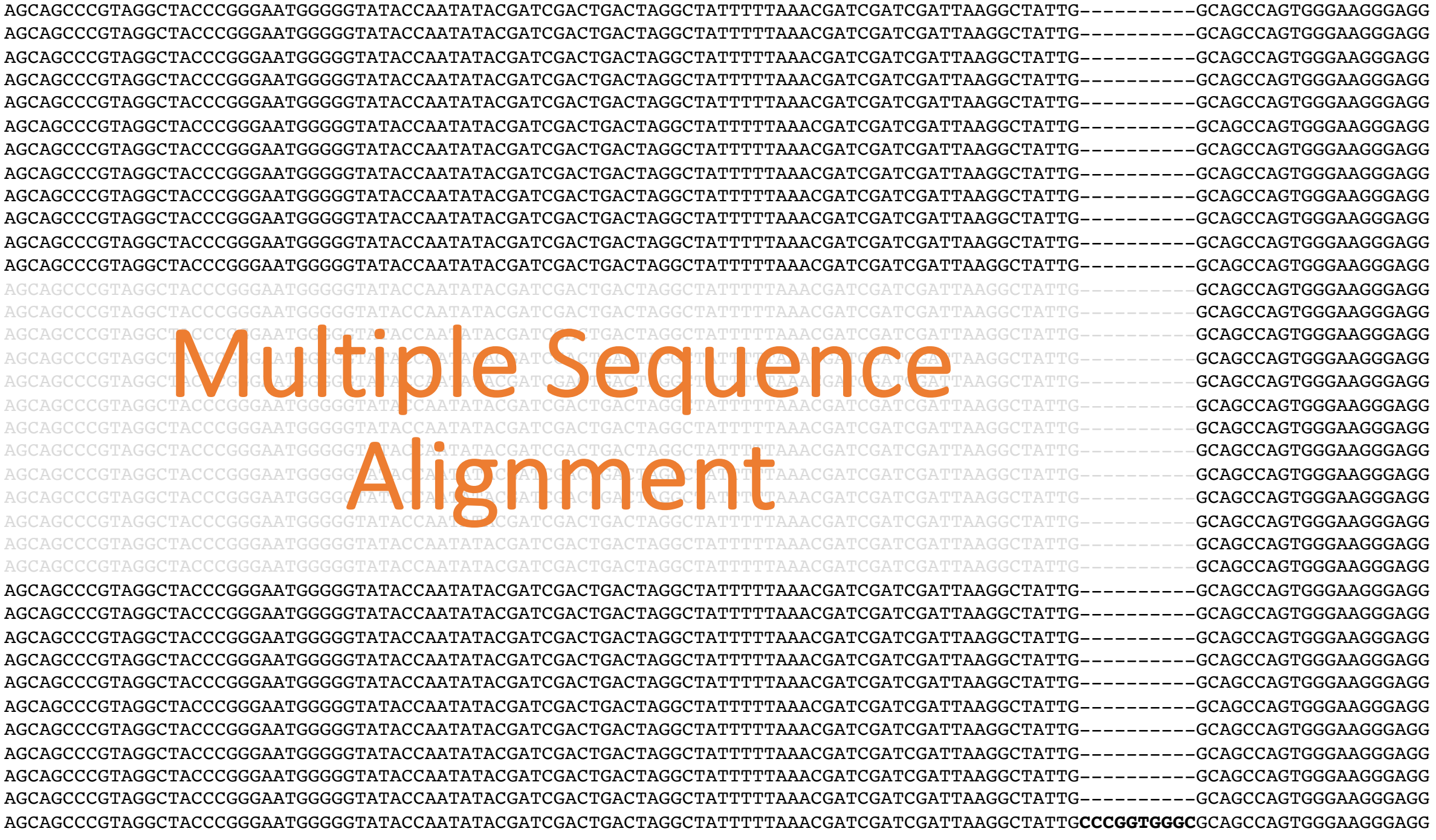
Genomic analysis

- Grow US, Japanese, and Russian isolates and extract genomic DNA
- WGS on multiple platforms
  - Illumina
  - Pac Bio (Oxford Nanopore)
- All isolates are very similar
  - <20 SNP's separate the most distant pair
- Closest to *B. anthracis* clade, but still ~50,000 SNP's away from nearest *B. anthracis* (not a member)
- No *B. anthracis* virulence genes
- New species
  - *Bacillus ???*



plcR nonsense mutation




Multiple Sequence
Alignment

Multiple Sequence Alignment (MSA): *another grand challenge*[1]

```
S1 = AGGCTATCACCTGACCTCCA          S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC              S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC                   S3 = TAG-CT-------GACCGC--
   …                               …
Sm = TCACGACCGACA          ⟶      Sm = -------TCAC--GACCGACA
```

Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

Given: Two strings

$$a = a_1a_2a_3a_4...a_m$$
$$b = b_1b_2b_3b_4...b_n$$

where a_i, b_i are letters from some alphabet like {A,C,G,T}.

Compute how similar the two strings are.

What do we mean by "similar"?

Edit distance between strings a and b = the smallest number of the following operations that are needed to transform a into b:

- Replace a character
- Delete a character
- Insert a character

Multiple Sequence Alignment

Input: Sequences S_1 , S_2 , … S_k;

Let $\boldsymbol{M}$ be a **MSA** between these sequences.

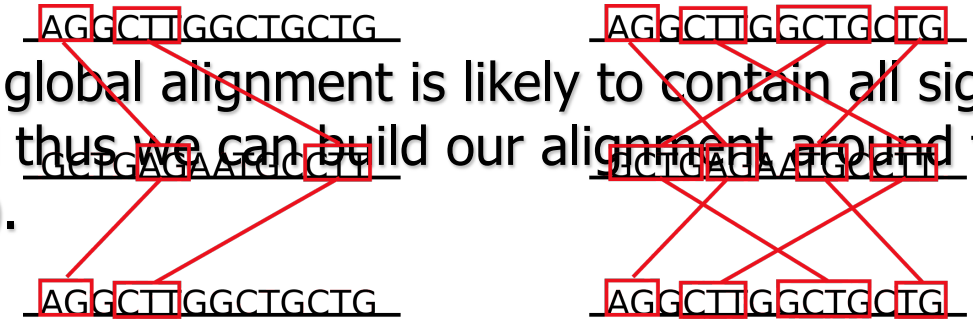
Let $\boldsymbol{d_M(Si\ \ Sj)}$ = be the edit distance of the alignment between $S_i^{'}$ and S_j.

D$(\boldsymbol{M}) = \sum_{ij} \boldsymbol{d_M(Si, Sj)}$ = Sum of all pairwise edit distances be the edit distance of the alignment between S_i and S_j.

- **Two important parameters: the number k of sequences, the length n of the longest sequence**.

Maximal Unique Matches (MUMs)

- Using unique matches as anchors, we can reduce the computational cost by partitioning large DNA sequences into tractable parts.
- The idea is that any optimal global alignment is likely to contain all significant common unique regions and thus we can build our alignment around these anchors (Delcher et al 1999).
- Anchor-based sequence alignment has been widely recognized as a reasonable heuristic for efficient multiple global alignment, and is featured in tools such as MUMmer and MAUVE.



How to detect unique matches in multiple DNA sequences?

- Several approaches exist, but the most efficient have been based around using suffix trees & suffix arrays.

- We use a compressed suffix structure, a specialization of a suffix tree, with increased efficiency and reduced space requirements with respect to the number of genomes being compared.

Suffix tree

Given string ababaabs#:

Suffixes:

1: ababaabs#

2: babaabs#

3: abaabs#

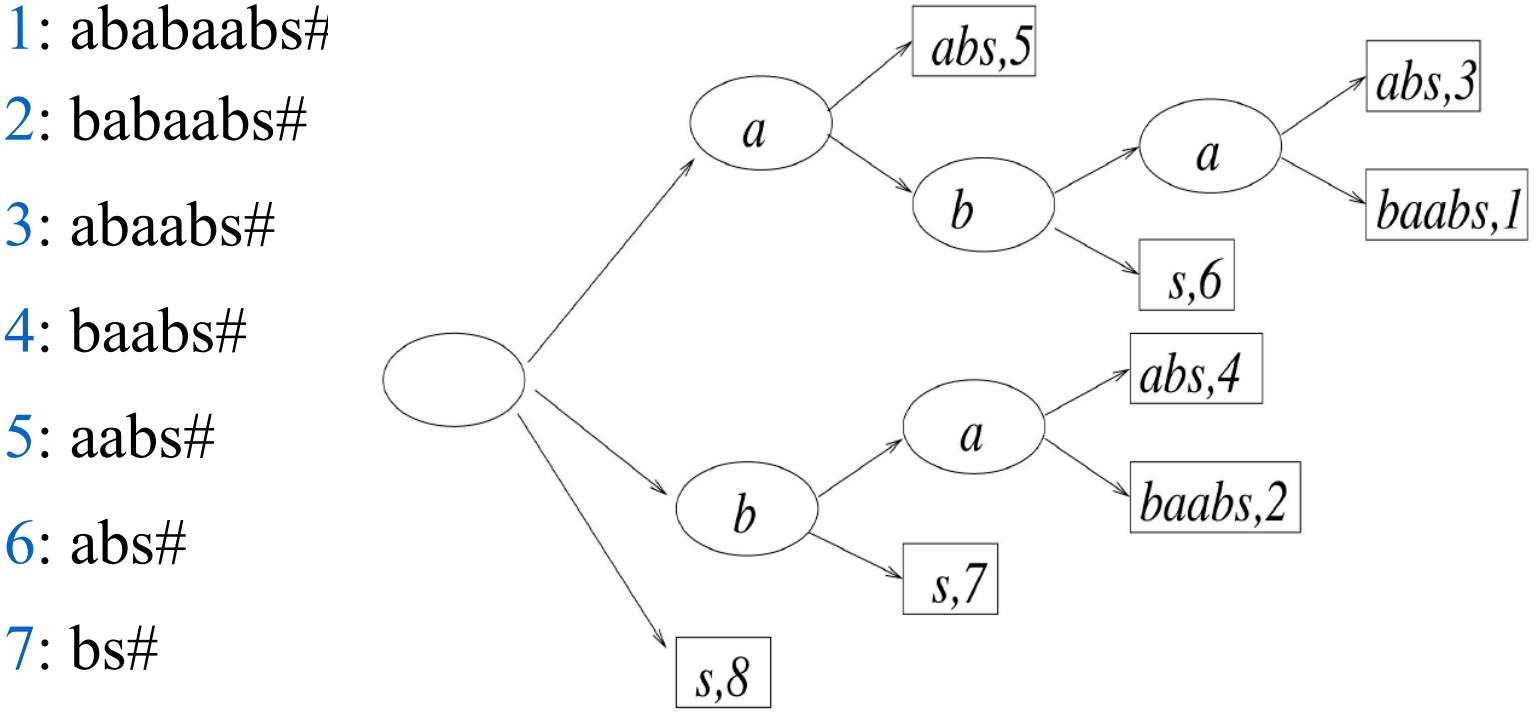
4: baabs#

5: aabs#

6: abs#

7: bs#

8: s#

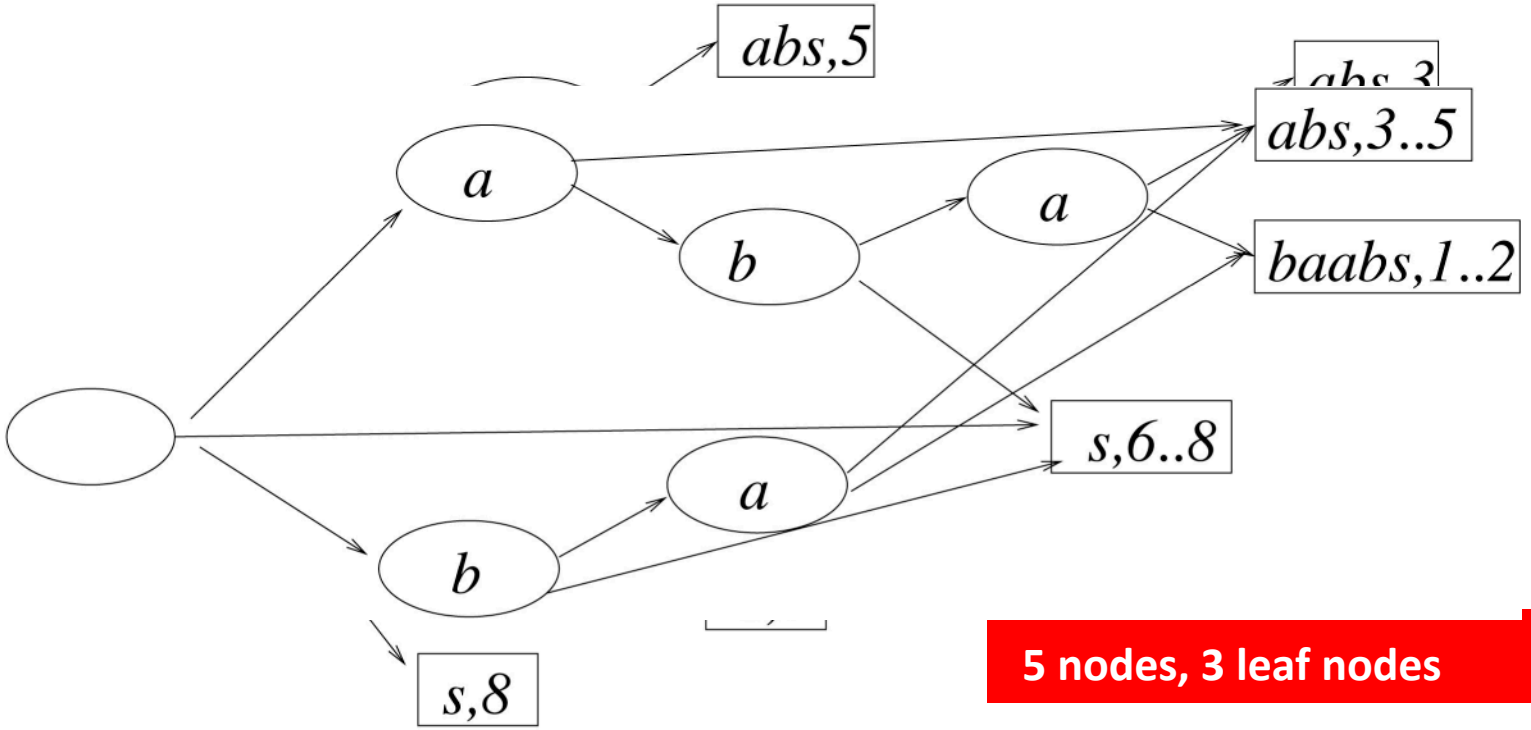


How can we generate a genome comparison framework efficiently?

1) Compress the suffix tree data structure.

2) Stream the remaining $k - 1$ sequences through the suffix structure for linear time complexity O($|S_1|+ \ldots +| S_k|$)

Compressing the suffix tree



Suffix tree of the string *abababaabs*.

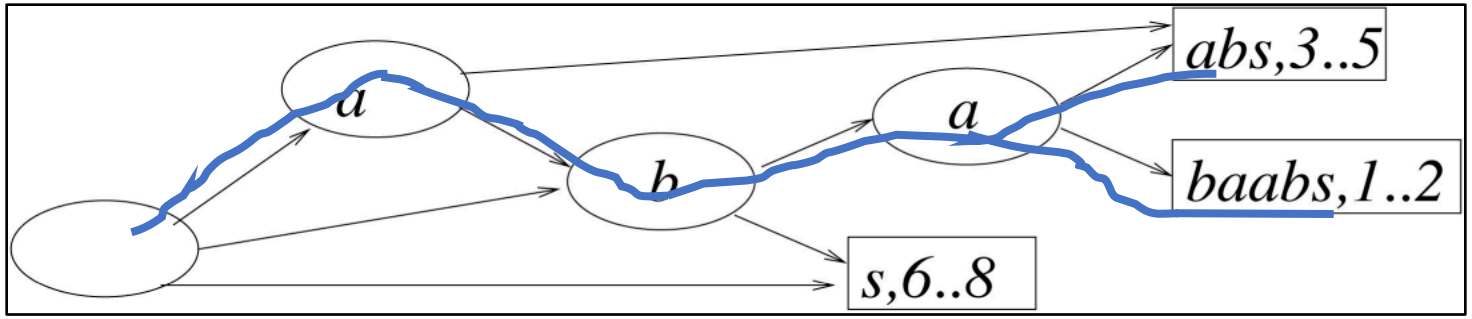
Treangen, Todd J., and Xavier Messeguer. "M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species." BMC bioinformatics 7.1 (2006): 433.APA

How can we generate a genome comparison framework efficiently?

1) Compress the suffix tree data structure (McCreight 1976, Ukonnen 1995).

2) Stream the remaining $k - 1$ sequences through the suffix structure for linear time complexity O($|S_1|+ \ldots +| S_k|$)

MUM search algorithm



Unique Matches

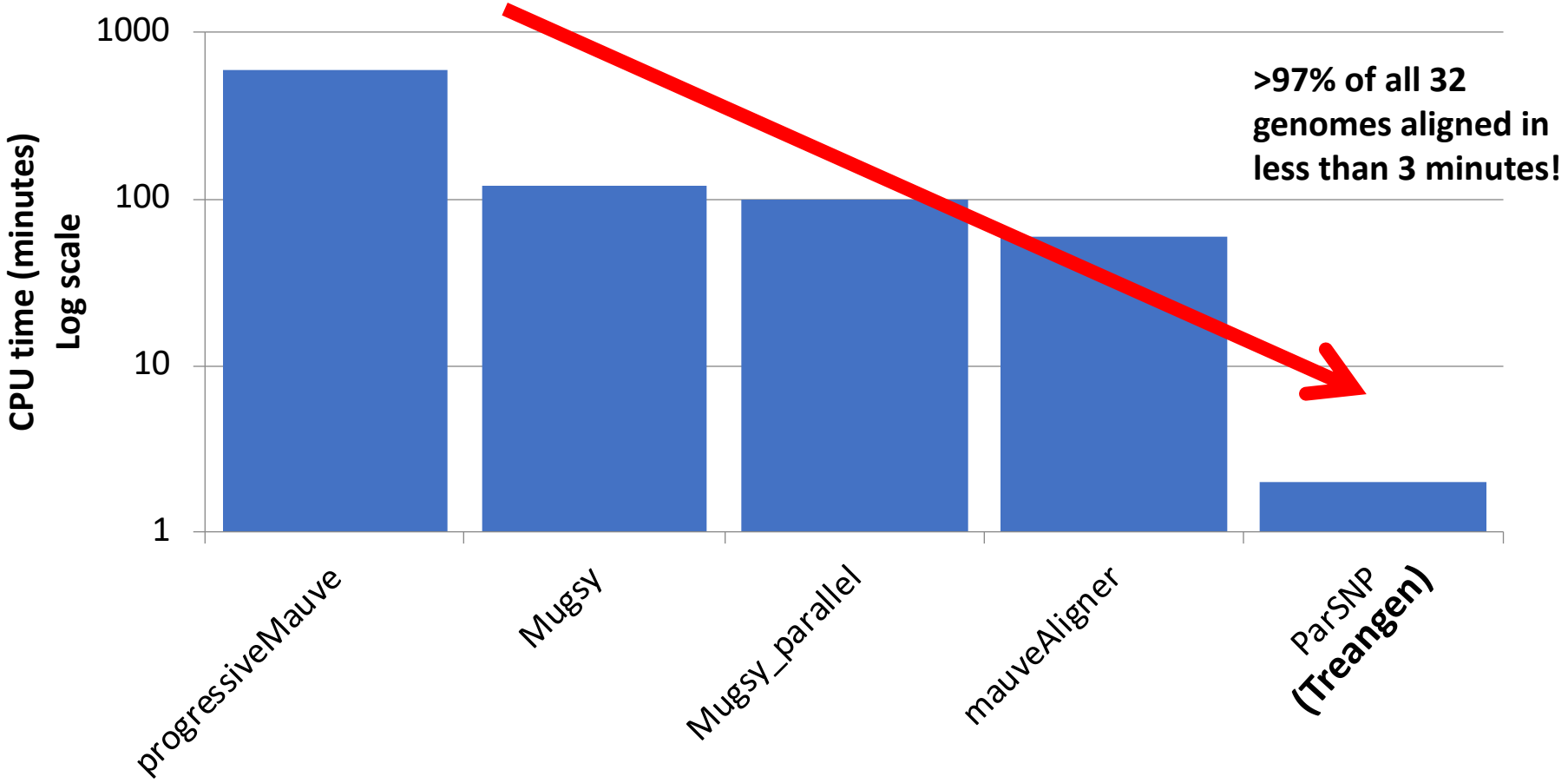
abaa
ababaabs

abaasbbababaabs

Treangen, Todd J., and Xavier Messeguer. "M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species." BMC bioinformatics 7.1 (2006): 433.APA

Run time performance

(32 simulated E. coli W3110 genomes)



Treangen, Ondov, Koren, & Phillippy, A. M. (2014). Rapid and interactive microbial genome alignment of thousands of genomes. *Genome Biology.*

Sustainable and organic strain-level harvesting with ParSNP

Motivation

- Hundreds to thousands of closely-related strains now available from public genome DBs and Metagenomes
  - Only to increase with 100K Foodborne Pathogen Genome Project (http://100kgenome.vetmed.ucdavis.edu), etc

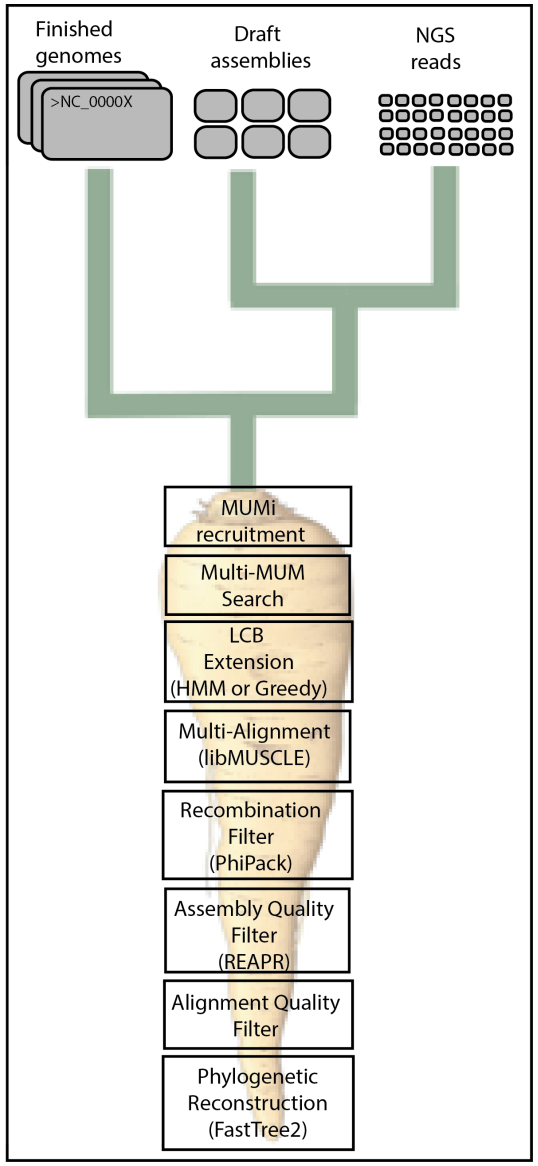
- Large-scale whole genome SNP typing and phylogenetic reconstruction is becoming a common task

ParSNP

1. NGS reads, draft assemblies, and/or finished genomes as input

1. Near-neighbor genomes are recruited

2. **Efficient Multi-MUM search to identify locally collinear blocks**



4. Multi-Alignment of LCBs with Muscle

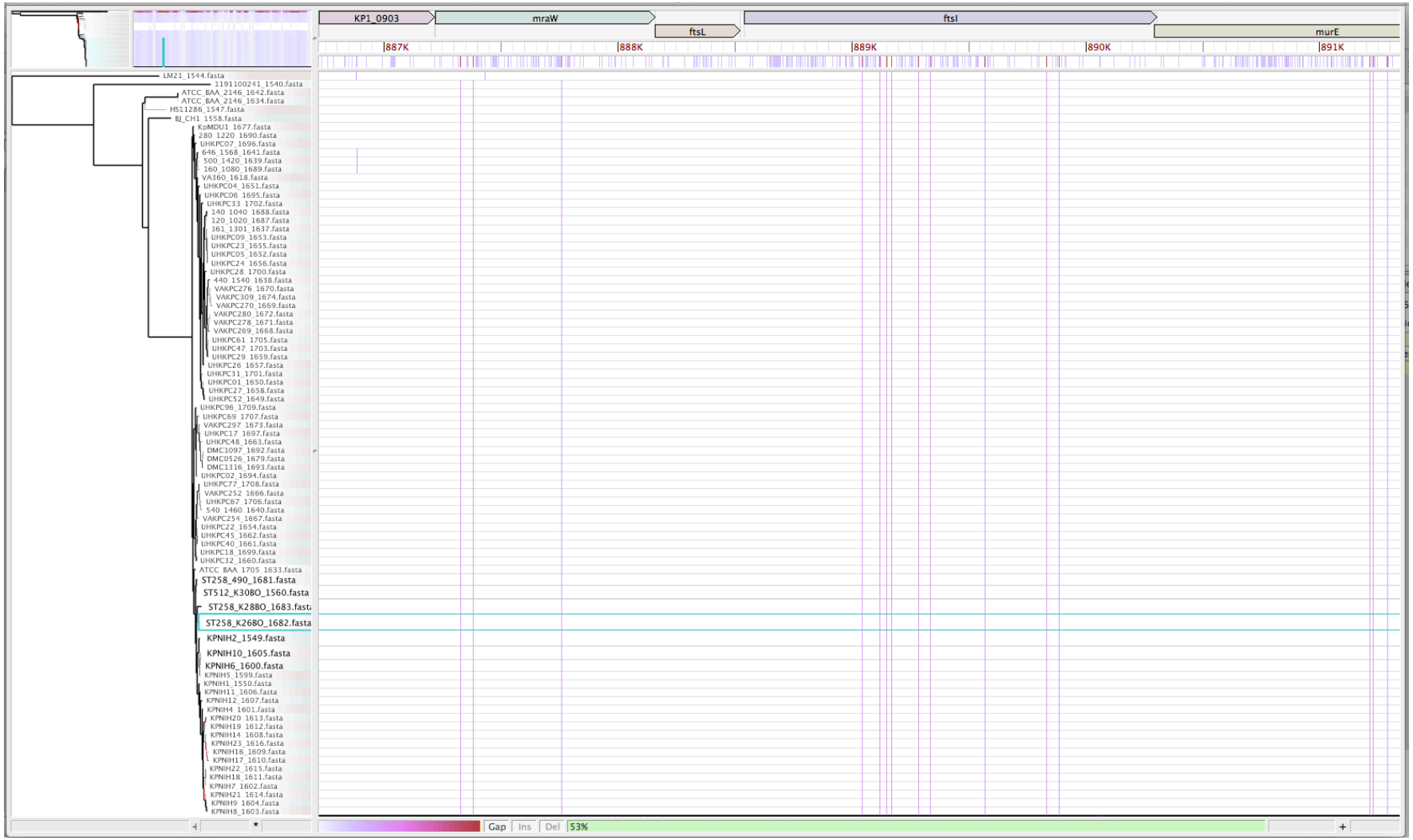
5. HMM-based LCB extension

6. Apply quality filters

7. Phylogenetic reconstruction with FastTree2

...with a touch of Gingr

100+ *Klebsiella pneumoniae genomes* (Primary developer: Brian Ondov)



Conclusions

① ParSNP can align hundreds closely-related microbial strains in *minutes*

- 200 *K. pneumoniae* genomes in 20 minutes on 8 cores

② Gingr provides an interactive interface for the simultaneous visualization of:

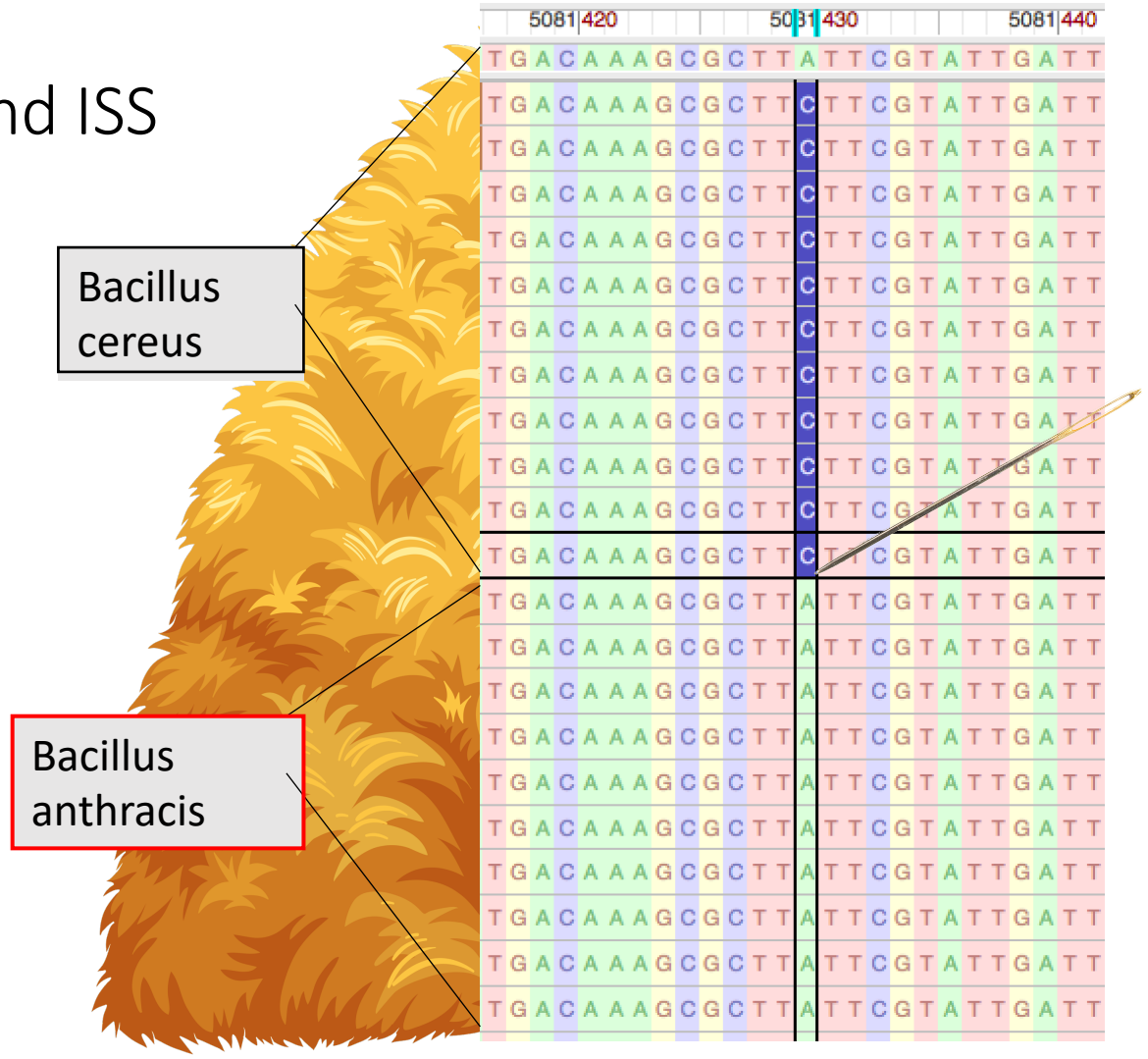
- *core genomes*
- *SNPs*
- *clade phylogeny*

③ Assembly + multiple alignment as a sustainable path to core genome SNP typing:

- 825 genomes 1500 GB (reads) 3.3 GB (asms) 0.13 GB (aln)

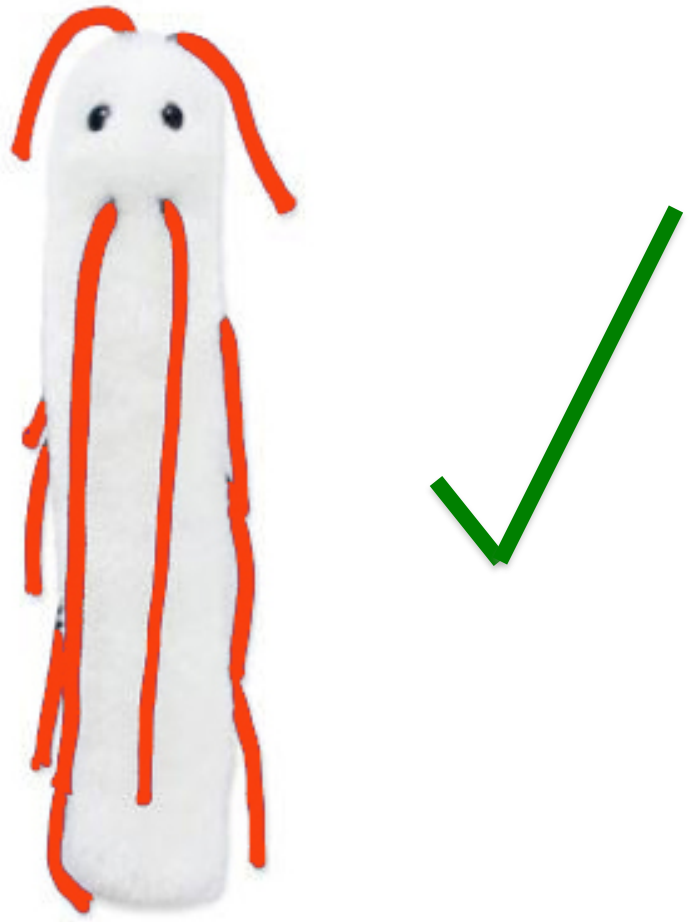
MSA and ISS

- Confirmation of single nucleotide difference via MSA
- 1 single nucleotide difference (needle) out of 5,000,000 nucleotides (haystack)



Conclusions

B. anthracis



***Novel B. cereus group* strain**

Your turn!

- https://gitlab.com/treangen/stamps2019/README.md#part3

Final exercise: AliView

- https://gitlab.com/treangen/stamps2019/README.md#part4