Decontamination of bacterial isolates

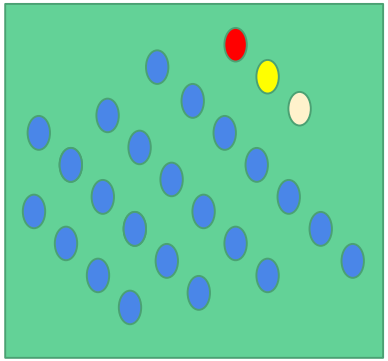
What types of contamination exists?

- Cross contamination
  - Other bacteria
  - Human, virus etc.
- Intra species contamination

Case scenarios

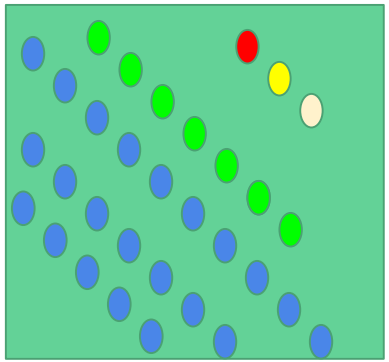
- Primary organism is when >75% of our DNA can be linked to a reference
- Secondary organism is when > 1% of our DNA can be linked to a reference
- Noise is the rest where we can't determine the origin

Good



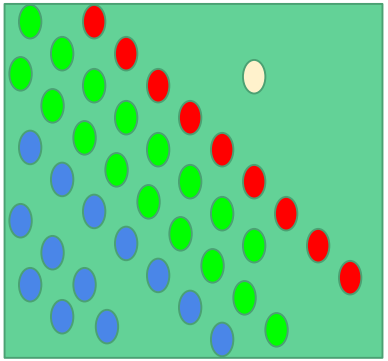
Primary isolate + noise

Fixable



Primary isolate + secondary + noise

Likely back to the lab :(



multiple secondary isolates + noise

What would we like our decontamination tool to do?

- Input: Raw DNA reads
- Some magic
  - Figure out the scope of the problem
    - Multiple organisms? One + noise? Several of the same organism?
  - Remove unwanted reads
- Output: Decontamination report, clean reads, contaminated reads

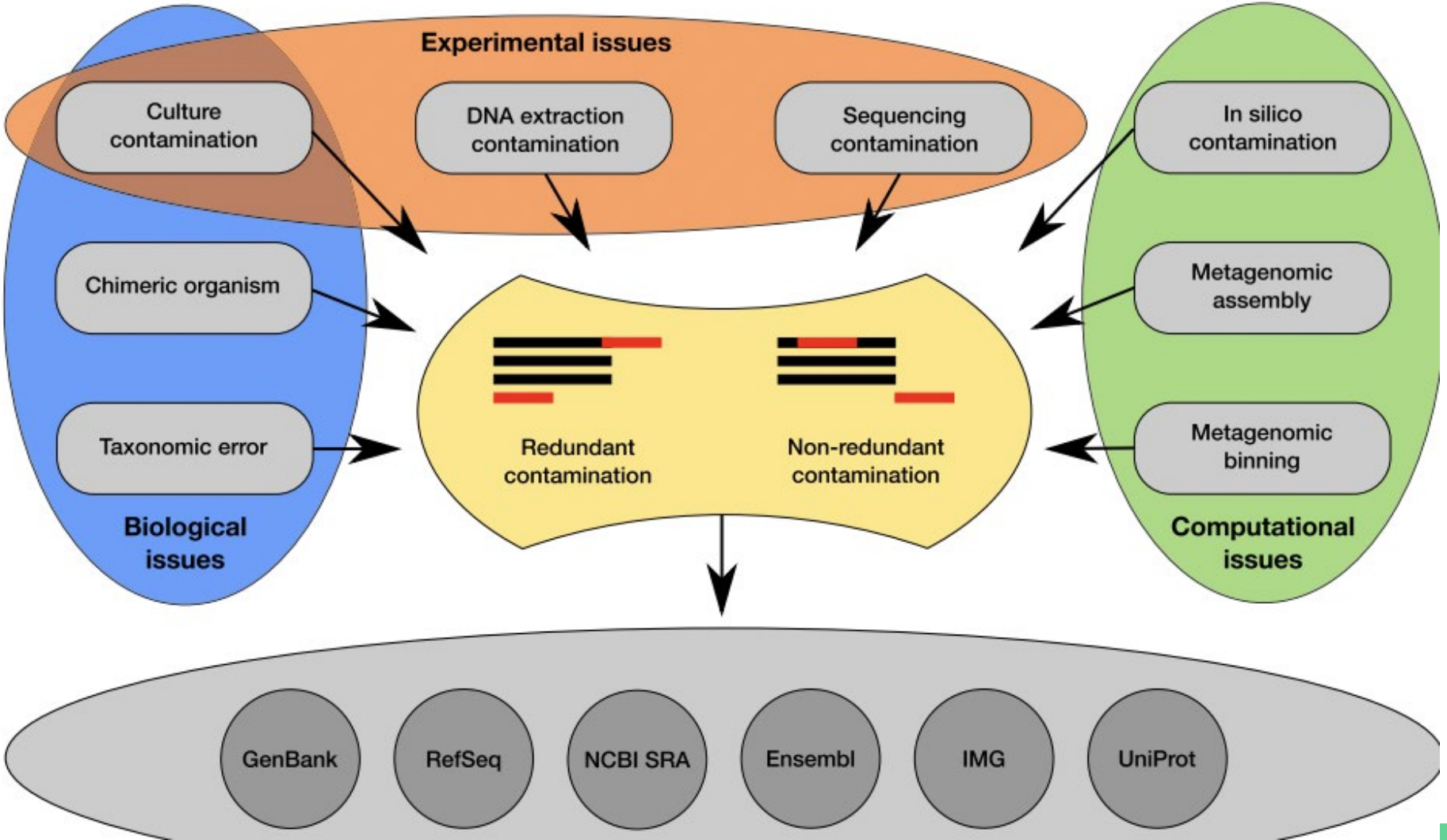
What existing tools are there?

- Kraken2
  - Catches a lot but also missed a lot. Good at identifying potential organisms in sample, but not for correctly determining read origin
- checkm
  - Not trained for isolate sample and constantly overestimates contamination rate. Build for assembled data too.
- ConFindr
  - rMLST tool build only for illumina data.
- EukCC
  - Marker gene based tool for Eukaryotes
- CONSULT
  - k-mer based read matching against references.

Why not just assemble the reads???

Why not assemble and then figure contamination out?

Chimeric Assembly


Experimental issues
Culture contamination
DNA extraction contamination
Sequencing contamination
In silico contamination
Chimeric organism
Metagenomic assembly
Taxonomic error
Metagenomic binning
Redundant contamination
Non-redundant contamination
Biological issues
Computational issues
GenBank
RefSeq
NCBI SRA
Ensembl
IMG
UniProt

Combining it all:

- KMA alignment against selected databases:
  - Human, virus etc for obvious contaminations
  - Bacteria whole genome database
- rMLST, GTDB mapping
  - This provides a list of possible bacteria that could be in the sample
  - Has enough sensitivity to detect intra-species contamination
- Based off alignment hits and marker genes we can now more confidently determine the origin of our reads
  - Remove reads from highly confident sources of contamination, but keep reads that could potentially originate from the primary isolate

rmlst/MLST

#Template	Score	Expected	Template_length	Template_Identity	Template_Coverage	Query_Identity	Query_Coverage	Depth	q_value	p_value
adk_14	18378	21	477	100.00	100.00	100.00	38.30	18314.87		1.0e-26
adk_55	159662	20	477	100.00	100.00	100.00	332.52	159601.21		1.0e-26
atpG_7	4326	19	447	95.08	95.08	100.00	105.18	9.59	4266.89	1.0e-26
atpG_11	157467	19	447	100.00	100.00	100.00	100.00	350.53	157410.00	1.0e-26
atpG_14	3855	19	447	99.55	100.00	99.55	100.00	8.70	3795.92	1.0e-26
atpG_142	3086	19	447	95.30	95.53	99.77	104.68	6.65	3027.01	1.0e-
frdB_16	173367	20	489	100.00	100.00	100.00	351.00	173304.93		1.0e-26
frdB_59	9070	21	489	99.80	99.80	100.00	18.41	9005.23		1.0e-26
fucK_14	110568	14	345	100.00	100.00	100.00	316.06	110523.43		1.0e-26
fucK_15	2777	15	345	100.00	100.00	100.00	7.85	2731.41		1.0e-26
fucK_59	2404	15	345	97.68	98.26	99.41	101.77	6.94	2358.46	1.0e-26
mdh_81	4420	17	405	99.75	100.00	99.75	100.00	10.63	4366.41	1.0e-26
mdh_89	134579	17	405	100.00	100.00	100.00	100.00	328.15	134527.02	1.0e-26
pgi_113	171552	19	468	100.00	100.00	100.00	100.00	363.54	171492.56	1.0e-26
recA_1	3452	18	426	100.00	100.00	100.00	100.00	8.11	3395.72	1.0e-26
recA_3	153512	18	426	100.00	100.00	100.00	100.00	356.01	153457.62	1.0e-26

Perhaps we don’t need perfect rmlst/mlst to conclude that there is likely a intraspecies contamination

Total work flow

1. Trim input reads and filter for q-score
2. Map trimmed reads against EuDB, ViralDB and bacteriaDB (Maybe also others)
3. Determine marker genes for bacterial organism
    a. If based on mapping &marker genes that intra species contamination is taking place; sort them and access depth.
4. Remove minimal amount of only highly confident contamination DNA
    a. If it maps to something else and not our primary source, remove it.
5. Finally, divide trimmed reads into new output files; One for the cleaned reads, one for contaminated ones.

Concluding remarks

- DNA contamination is super important and currently no effort anywhere is done to decontaminate long read sequencing data
- We must decontaminate directly on raw reads to avoid chimeric assemblies
- rMLST enables us to screen for intra species contamination
- Remove only DNA we are 100% sure is not from our primary source, as too much removal can also result in fragmented genomes
- Now it is time to validate the tool on some data and validate that it works :)
- Finally, maybe write a docker image/conda package so everyone can use it