Figure 1: header image

Project - M1 GENIOMHE 2024/25:

Introduction on the species, structural genomics and goal of this project, then discuss potential issues that might arise due to complexity. Proceed by summarizing the desired workflow highlighting the criteria demanded by the instructor to be fulfilled

Exploration

Sequence properties

Checking GC content in this region to have an idea about potential gene desnitites. For that we run the script:

```
python src/GCcontent.py data/region8.fasta
```

The GC content of the DNA sequence is 48%.

Region localization

Want to localize this region by mapping agaisnt the reference sequence of *Triticum aestivum* (available on RefSeq at GCF_018294505.1)

Validation

BLAST

To validate the predicted genes, we will start of by blasting against the proteome of *Triticum aestivum* available on UniProt. We retrieved the list of proteins from the supplementary material of an International Wheat Genome Sequencing Consortium (IWGSC) published in *Science*[1] aiming to provide an annotated

[1]The International Wheat Genome Sequencing Consortium (IWGSC) et al. ,Shifting the limits in wheat research and breeding using a fully annotated reference genome.Science361,eaar7191(2018).DOI:10.1126/science.aar7191

reference sequence of the *Triticum aestivum* genome. The article is available here. We will access all the proteins sequences (including isoforms) using an api call to the UniProt database.

```
curl https://rest.uniprot.org/uniprotkb/stream?compressed=true&format=fasta&query=%28%28lit_
gunzip data/web_retrieved_sequences/proteins.fasta.gz
```

```
cat data/////////proteins.fasta | grep '>' | wc -l
130283
```

There is a total of 130283 proteins in the file. We will now perform a BLAST search against this database to see if our predicted genes are similar to any of the known annotated proteins of *Triticum aestivum*. It'll be a tblastn search, as we are looking for protein sequences that are similar to our DNA sequence (which increases the sensitivity of the search, as it takes into account the codon degenracy).

```
# --creating the local database
makeblastdb -in data/proteins.fasta -dbtype prot -out data/database/Triticum_aestivum_protei
# --running the blast search
tblastn -query data/region8.fasta -db data/database/Triticum_aestivum_proteins_uniprot -out
```

Transcriptome

The European Nucleotide Archive (ENA) comprises a large collection of sequencing data from raw sequences to assembly to functionally annotated ones. While looking for transcriptome studies for *Triticum aestivum* we find several projects (Total= 22, in this table[2])

[2]

Accession	Description
GAEF01000000	Triticum aestivum, TSA project GAEF01000000 data
GAJL01000000	Triticum aestivum, TSA project GAJL01000000 data
GBKH01000000	Triticum aestivum, TSA project GBKH01000000 data
GBKI01000000	Triticum aestivum, TSA project GBKI01000000 data
GBKJ01000000	Triticum aestivum, TSA project GBKJ01000000 data
GBKK01000000	Triticum aestivum, TSA project GBKK01000000 data
GBZP01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
GDTJ01000000	Triticum aestivum, TSA project GDTJ01000000 data
GEUX01000000	Triticum aestivum, TSA project GEUX01000000 data
GEWU01000000	Triticum aestivum, TSA project GEWU01000000 data
GFFI01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
GIJS01000000	Triticum aestivum, TSA project GIJS01000000 data
GILY01000000	Triticum aestivum, TSA project GILY01000000 data
GIXT01000000	TSA: Triticum aestivum cultivar TcLr19 isolate leaf, transcriptome shotgun assembly.
GJAR01000000	TSA: Triticum aestivum cultivar Avocet R, transcriptome shotgun assembly.
GJUY01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
HAAB01000000	Triticum aestivum, TSA project HAAB01000000 data
HCEC01000000	TSA: Triticum aestivum
HCED01000000	TSA: Triticum aestivum

TSA stands for Transcriptome Shotgun Assembly

One of them is published by Xiao et al. (2013) in BMC Genomics [REF1]. They have performed short read RNA-seq using Illumina Hi-Seq tech, and deposited the project's raw reads on the SRA database, project `SRX212270`. We will use this as trial to explore how we can validate using Whole Transcriptomes before optimizing our choice.

Trial 1: blasting against transcriptome As a first attempt, due to the high memory requirement (*e.g.,* one of them is 15GB of reads), we have tried performing BLAST on ncbi's server against this whole transcriptome in [3], with default parameters (can perform it here by just adding the region8 fasta file). The default search gave no significant results, we will try to relax the paramters (BLOSUM45 and lowering penalties, accepting lower thresholds...)

Trial 2: downloading the WTS data We will try downloading the reads of [4] to see how to manipulate such a large file. Since it surpasses the threshold to download a file from SRA webserver (which is 5GB), we will download it using `sra-toolkit`.
While running out of time and memory, we will try doing that using Galaxy[5][6].

Trial 3: Analysis Working on galaxy, first retrieve the SRA accession number from the project, tools > Get data > EBI SRA, copy the accession number and get the fastq in galaxy. After loading them (paired end so 2 fastq) > fastq groomer, to make sure the fastq format fits Galaxy's requirement and make it run. Meanwhile > FastQC to make sure the quality of the transcriptome is good or whether it's better to take another set of reads.

We will try now mapping: using Tophat2, we will map the reads to the reference genome of *Triticum aestivum* (available on ENSEMBL) to see how many reads are mapped and how many are not. We have taken the reference genome using

IAAK01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
IAAL01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
IAAM01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.

[3]Xiao, J., Jin, X., Jia, X., Wang, H., Cao, A., Zhao, W., ... & Wang, X. (2013). Transcriptome-based discovery of pathways and genes related to resistance against Fusarium head blight in wheat landrace Wangshuibai. BMC genomics, 14, 1-19.

[4]Xiao, J., Jin, X., Jia, X., Wang, H., Cao, A., Zhao, W., ... & Wang, X. (2013). Transcriptome-based discovery of pathways and genes related to resistance against Fusarium head blight in wheat landrace Wangshuibai. BMC genomics, 14, 1-19.

[5]The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update Nucleic Acids Research, gkae410 doi:10.1093/nar/gkae410

[6]The Galaxy server used for some calculations is partly funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

Trial 4: visualization *trying to perform RNA-seq aln and viz using IGV*

Supplementary

Resources

- Whole Genome (all 6n chr) of *triticum aestivum* on ENSEMBL : https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/gff3/triticum_aestivum/

- ENSEMBL in general : https://plants.ensembl.org/Triticum_aestivum/Info/Index

- ENA: https://www.ebi.ac.uk/ena/browser/view/Taxon:4565
- SRA: Sequence Read Archive, repository for seq data

- RNAseq reads fetch and viz: youtube video

- RefSeq: reference sequence v2.1 here, link to acces the dataset is https://api.ncbi.nlm.nih.gov/datasets/v2/genome/accession/GCF_018294505.1/download?include_annotation_type=GENOME_FASTA&include_annotation_type=GENOME_GFF&include_annotation_type=RNA_FASTA&include_annotation_type=CDS_FASTA&include_annotation_type=PROT_FASTA&include_annotation_type=SEQUENCE_REPORT&hydrated=FULLY_HYDRATED