National Czech Programme | eosc

The experiences and lessons learned from working in a fully open consortium analyzing genomics data

Monika Cechova

Faculty of Informatics, Masaryk University

Co-funded by the European Union

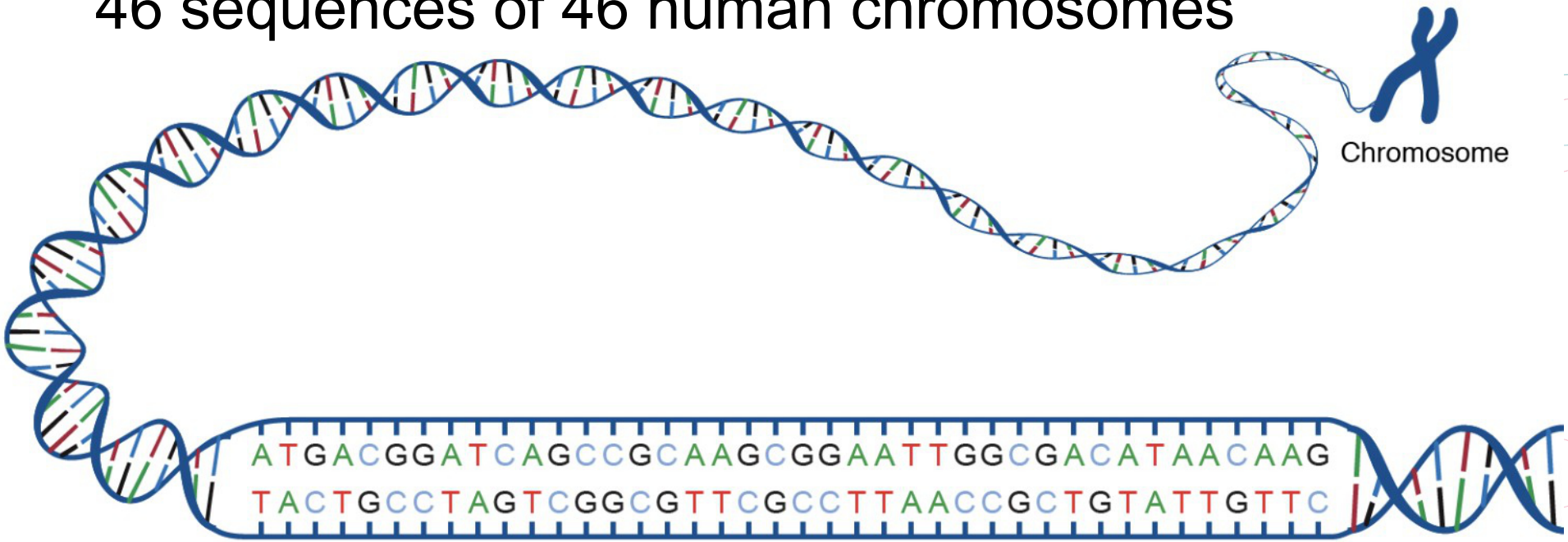
MINISTRY OF EDUCATION, YOUTH AND SPORTS

MUNI FAKULTA INFORMATIKY

UC SANTA CRUZ

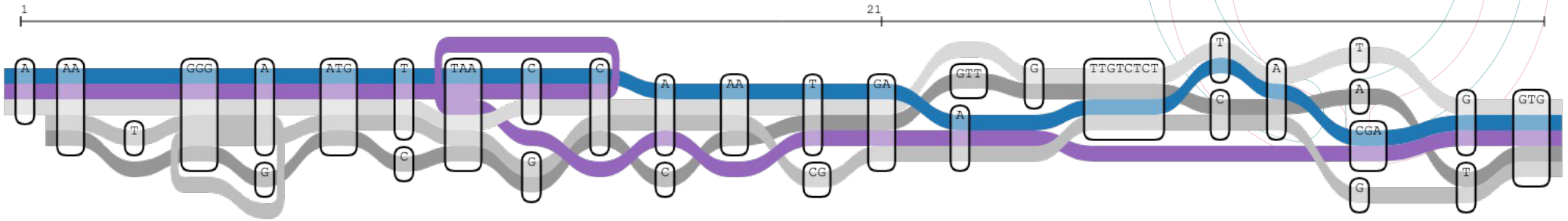
We need complete human genomes

- DNA is just a string of letters (ACTG) and can therefore be stored in text files
- The future is now: The genome assemblies can be represented by 46 sequences of 46 human chromosomes



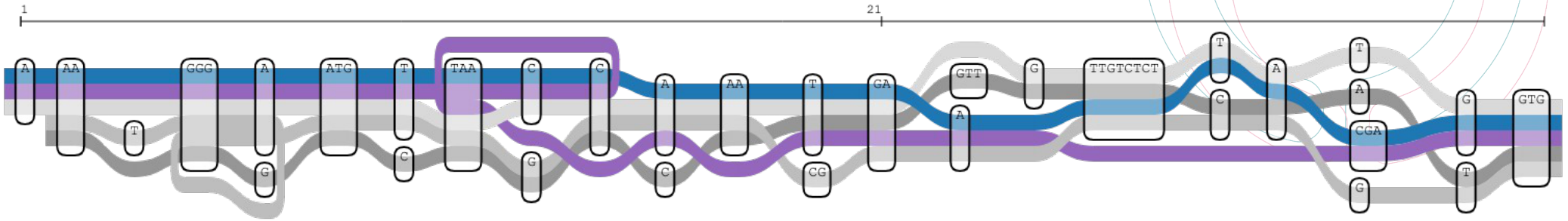
https://www.genome.gov/genetics-glossary/acgt

- Let’s understand the biology of complete human genomes => HPRC consortium



Re-using the existing resources

- Let’s understand the biology of complete human genomes => HPRC consortium
- The 1000 Genomes Project (1KGP) already samples **diverse samples from all around the world**
  - this landmark initiative serves as an extraordinarily rich source of knowledge, and a benchmark dataset
  - sequencing the already available cell lines with the modern state-of-the-art technologies and algorithms allows to build on this existing resource



Human Pangenome Reference Consortium (HPRC)

Heather Lawson

Karen Miga

Ann Mc Cartney

Erich Jarvis

Sadye Paez

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

We would like to acknowledge the National Genome Research Institute (NHGRI) for funding the following grants which are in support of creating the human pangenome reference: 1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963, and the Human Pangenome Reference Consortium (https://humanpangenome.org/)

- Slide from Karen Miga

The huge amound of data – what now?

- All the data is stored in the AWS bucket
- All wrangled data with the links is stored as a **Release Data Table on GitHub**
- This versioned table is shared with all the collaborators

genbank_accession	assembly_md5	assembly_fai
GCA_041900255.1	s3://human-pangenomics/working/HPRC/HG00408/assemblies/release2/HG00408_pat_hprc_r2_v1.0.1.fa.gz.md5	s3://human-pangenomics/workir
GCA_041900365.1	s3://human-pangenomics/working/HPRC/HG00597/assemblies/release2/HG00597_pat_hprc_r2_v1.0.1.fa.gz.md5	s3://human-pangenomics/workir
GCA_041900145.1	s3://human-pangenomics/working/HPRC/HG01192/assemblies/release2/HG01192_pat_hprc_r2_v1.0.1.fa.gz.md5	s3://human-pangenomics/workir
GCA_041900235.1	s3://human-pangenomics/working/HPRC/HG01261/assemblies/release2/HG01261_pat_hprc_r2_v1.0.1.fa.gz.md5	s3://human-pangenomics/workir
GCA_041900165.1	s3://human-pangenomics/working/HPRC/HG02015/assemblies/release2/HG02015_pat_hprc_r2_v1.0.1.fa.gz.md5	s3://human-pangenomics/workir

National Czech Programme | eosc

Release 2 Table

Last Modified	Size	Key
		../
	0	cat/
	0	censat/
	0	chains/
	0	chrom_assignment/
	0	liftoff/
	0	methylation/
	0	repeat_masker/
	0	segdups/

human-pangenomics / hprc_intermediate_assembly

Code | Issues 5 | Pull requests 3 | Actions | Projects | Security | Insights

Files

main

Go to file t

- assembly
- assembly_qc
- data_tables
  - annotation
  - assembly_qc
  - pangenomes
  - sample
  - sequencing_data
  - README.md
  - assemblies_release2_v1.0.ind...

hprc_intermediate_assembly / data_tables / assemblies_release2_v1.0.index.csv

juklucas point to release locations; rename index files 2a7e3a5 · 2 months ago History

Preview | Code | Blame 467 lines (467 loc) · 223 KB Raw

Search this file

1	sample_id	haplotype	phasing	assembly_method	assembly_method_version	assembly_date	assembly_name	source	genbank_acce
2	HG00408	1	trio	hifiasm	0.19.7	2024-08	HG00408_pat_hprc_r2_v1.0.1	hprc	GCA_0419002
3	HG00597	1	trio	hifiasm	0.19.7	2024-08	HG00597_pat_hprc_r2_v1.0.1	hprc	GCA_0419003
4	HG01192	1	trio	hifiasm	0.19.7	2024-08	HG01192_pat_hprc_r2_v1.0.1	hprc	GCA_04190014
5	HG01261	1	trio	hifiasm	0.19.7	2024-08	HG01261_pat_hprc_r2_v1.0.1	hprc	GCA_0419002
6	HG02015	1	trio	hifiasm	0.19.7	2024-08	HG02015_pat_hprc_r2_v1.0.1	hprc	GCA_0419001
7	HG02056	1	trio	hifiasm	0.19.7	2024-08	HG02056_pat_hprc_r2_v1.0.1	hprc	GCA_0419000

National Czech Programme | eosc

HPRC Data Explorer

Human Pangenome **Data Explorer**

Sequencing Data | Assemblies | Annotations | Alignments | HPRC Website | Help & Documentation

Filters | Clear All

Search all filters...

SRA

- Accession (2220)
- Biosample Accession (232)

SAMPLE

- Sample ID (232)

SEQUENCING

- Platform (3)
- Instrument Model (6)
- Library Strategy (3)
- Filetype (3)
- Basecaller (5)
- Basecaller Version (11)

Sequencing Data

Table | Graph | Results 1 - 6048 of 6048 | Download TSV | Group By | Edit Columns

Filename	Accession	Sample ID	Family ID	Population Abbreviation	Population Descriptor	Biosample Accession
BN001_R1.trimmed.fastq.gz	Unspecified	HG00733	Unspecified	PUR	Puerto Rican in Puerto Rico	SAMN00006581
BN001_R2.trimmed.fastq.gz	Unspecified	HG00733	Unspecified	PUR	Puerto Rican in Puerto Rico	SAMN00006581
GM20503.deepconsensus.fastq.gz	Unspecified	NA20503	Unspecified	TSI	Toscani in Italia	SAMN41021630
GM20762.deepconsensus.fastq.gz	Unspecified	NA20762	Unspecified	TSI	Toscani in Italia	SAMN41021652
GM20806.deepconsensus.fastq.gz	Unspecified	NA20806	Unspecified	TSI	Toscani in Italia	SAMN41021648
GM20827.deepconsensus.fastq.gz	Unspecified	NA20827	Unspecified	TSI	Toscani in Italia	SAMN41021650
HG002.HiC_1_NovaSeq_1_S1_L001_R1_001.fastq.gz	Unspecified	HG002	Unspecified	Unspecified	Unspecified	SAMN03283347

Why can the genetic data be freely shared?

- The broad research consent allows data download and re-analysis
- International partners are welcome to participate in the data sequencing and analysis (low barrier to entry)
- Partnership with AWS allows (sometimes) petabytes of the raw data to be directly available to users
- Centralized pre-processing and analysis provides annotations and outputs useful for biologists and bioinformaticians

Implemented open science strategies

- All the raw data is publicly available (open-access)
- All genome assemblies (sequences) are versioned and uploaded to the Genbank repository and available for immediate download
- All the code is open (the production workflows are available on Dockstore)
- All publications are pre-printed on BioRxiv
- Call for companion papers in a joint submission

The lessons learned

- Yearly release with versioned data tracking seems to work well
- The data managment can be a full time job
  (two ~full time data wranglers for the HPRC consortium)
- Heterogenous data sources add lots of overhead
- Open consent for re-use and re-analysis helps (a lot)
- Centralized data (pre)processing can save lots of time and resources
- To coordinate: quick communication (Slack), zoom, and bi-yearly F2F

National Czech Programme | eosc

Thank you for your attention

Big thank you to everyone on the T2T and HPRC team

https://humanpangenome.org/

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS