GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Deliverable D3.4
Operational analysis workflows in Embassy Cloud

Deliverable leader: EMBL
Authors: Peter Harrison (EMBL-EBI), Alexey Sokolov (EMBL-EBI), Cyril Kurylo (INRAE), Sylvain Foissac (INRAE), Sarah Djebali (INSERM), Cervin Guyomar (INRAE).

Version: 1.1

Due date of deliverable (as in DoA):	M24
Actual submission date:	24/06/2021, month M24

Dissemination level:

PU Public	X
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 817998

Table of contents

1 Summary

Within this document we report on the establishment of the Embassy Cloud GENE-SWitCH analysis platform and the implementation of production operational analysis workflows to process GENE-SWitCH data. The Embassy Cloud platform has been established and is currently processing the raw data generated in the GENE-SWitCH project. One of the operational analysis workflows that is being used to process the data is the TAGADA pipeline that was developed by researchers within the GENE-SWitCH project. The TAGADA pipeline is used in this report as an exemplar operational analysis workflow currently in operation on the GENE-SWitCH embassy analysis platform.

2 Introduction

The EMBL-EBI Embassy Cloud infrastructure (https://www.embassycloud.org/) is providing the GENE-SWitCH project with a powerful analysis environment that is co-localised with the GENE-SWitCH datasets stored for FAANG in the EMBL-EBI archives. The EMBL-EBI Data Co-ordination Centre established a GENE-SWitCH analysis platform within the Embassy Cloud infrastructure, for the consortium members to develop and then operate analysis workflows for the processing and analysis of GENE-SWitCH data.

3 Results

3.1 The Embassy cloud GENE-SWitCH analysis platform

The GENE-SWitCH analysis platform provides centralised compute for the entire consortium and ensures that data is uniformly processed using the same analysis workflows. The GENE-SWitCH analysis platform is deployed on version 4 of the Embassy Cloud infrastructure. Embassy v4 is a new implementation of Openstack based on Openstack Ussuri. For task orchestration we use Magnum which is an Openstack project for a Container Orchestration Engine to deploy and manage Kubernetes clusters (https://wiki.openstack.org/wiki/Magnum). This new version is architecturally distinct and provides a much smoother user experience with a number of key improvements. This includes improved options to self-manage Kubernetes nodes, with worker node auto-scaling so that the platform can adapt to the size of analysis as required. Container usage is also supported out of the box to ensure easy deployment of workflows developed by the wider community and EuroFAANG projects.

3.1.1 Technical specification of GENE-SWitCH analysis platform

The GENE-SWitCH analysis platform Kubernetes cluster has 3 master and 10 worker nodes. All nodes use Fedora CoreOS 32 operating system. Every master node has 4 CPU and 4GB of memory and every worker node has 8 CPU and 32GB of memory. For ReadWriteOnce persistent volume claims (PVC) cluster uses dynamic Persistent Volumes using Cinder (backed by Ceph). ReadWriteMany PVCs are achieved by installing an NFS server in a cluster backed by default-cinder PV (persistent volume) above. For backups and archiving we use an S3 compatible Object Storage backend that could also be used as a data lake to store unstructured data and large datasets. S3 object storage and internal cluster storage has 30TB of disk space available. To simplify data flow from EBI archives to the cluster FiRe Archive access was granted, so data could be transferred much faster without unnecessary DNS lookup overhead.

3.1.2 Operating workflows on the platform

To run pipelines on GENE-SWitCH analysis platform we use NextFlow (https://www.nextflow.io/). The built-in support for Kubernetes provided by Nextflow streamlines the execution of containerised workflows in Kubernetes clusters (Figure 1).

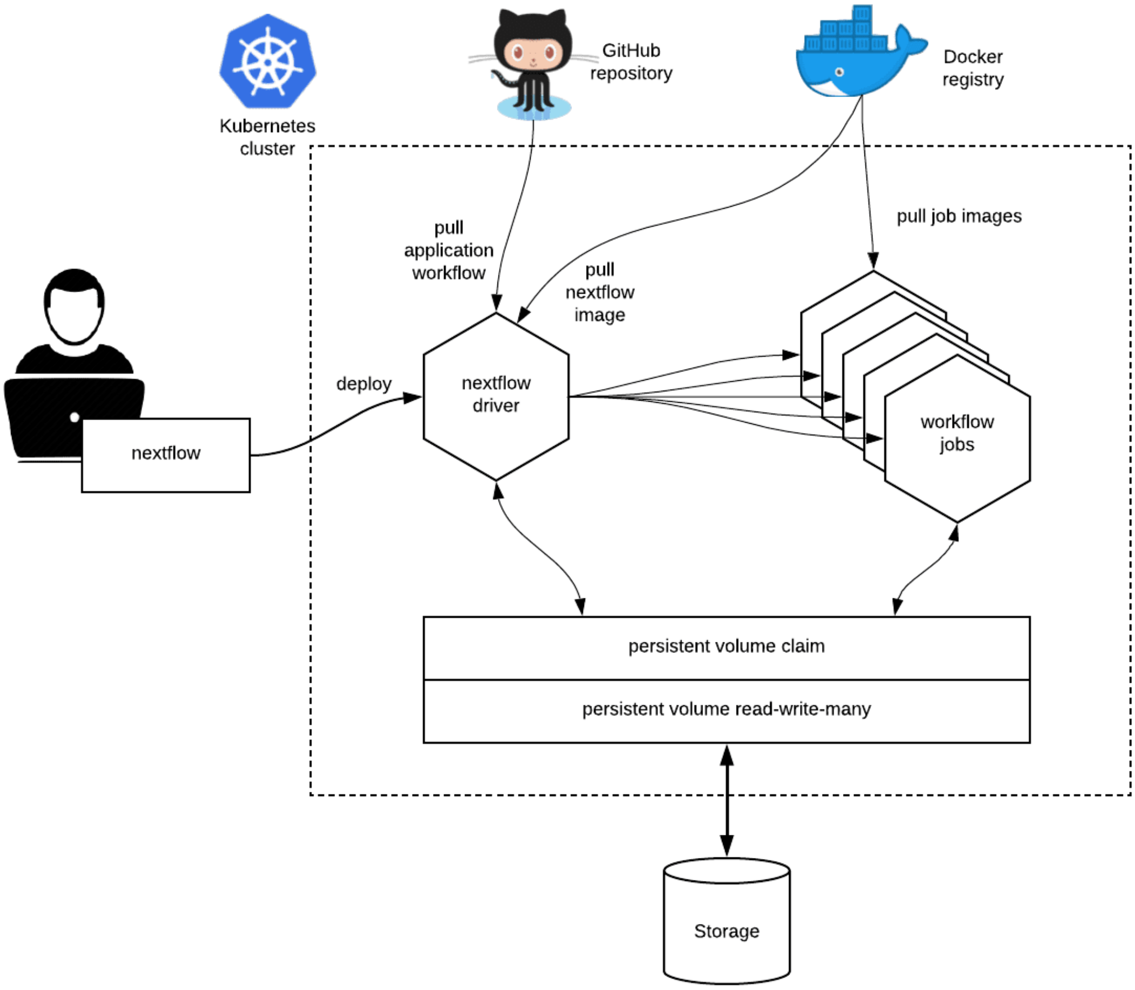


Figure 1. NextFlow built-it support for Kubernetes (https://www.nextflow.io/docs/latest/kubernetes.html)

Upon deployment the system creates a single primary pod that keeps reference of a whole workflow and for every process of the NextFlow script creates a separate (secondary) pod inside k8s (Kubernetes) cluster. Upon execution the secondary pod could report “Success” or “Failure”. By default the system will try to re-run failing pods 5 times and only then report final “Failure”. Kubernetes automatically checks for any failing nodes and could re-assign pods from failing nodes to the healthy nodes.

In the next reporting period we will implement further improvements for systematic analysis of raw data. This will implement a parallel work processing queue approach (https://kubernetes.io/docs/tasks/job/fine-parallel-processing-work-queue/) to run analyses automatically. Using this approach we will have multiple worker pods that will constantly pull messages from the Redis queue and start NextFlow processes. All logs will be stored in the central MongoDB database that will have Angular UI on top of it.

3.2 Operational workflow exemplar: The TAGADA RNA-Seq analysis pipeline

The TAGADA RNA-seq pipeline, for Transcripts And Genes Assembly, Deconvolution, Analysis, has been specifically developed for the GENE-SWitCH project. While many RNA-seq pipelines are currently available to build *de novo* gene models or quantify gene expression levels using a provided gene annotation, few allow both transcript reconstruction and expression assessment from RNA-seq data in a reproducible way. This is why a dedicated tool was needed.

In brief, TAGADA combines several reference RNA-seq bioinformatics tools into a containerized pipeline. It maps reads with STAR, reconstructs and quantifies genes and transcripts with StringTie and detects and characterizes long non-coding RNAs (lncRNAs) with FEELnc. TAGADA uses the Nextflow framework in line with the nf-core specifications. It provides a containerized environment that makes it compatible with a variety of high-performance computing platforms and workload orchestrators. It is designed to be easy to use and flexible. As such it requires a minimal set of inputs: a set of RNA-seq read files, a reference genome and its gene annotation. Optionally, a simple tabulated metadata file can also be provided to describe the experimental design and seamlessly merge samples according to specified factors.

The pipeline automatically generates a large variety of quality controls in the form of interactive charts and tables with statistics and metrics for various steps of the workflow. Expression tables are also provided with read counts for annotated and predicted genes and transcripts allowing further comparative expression analyses. We believe that the TAGADA pipeline offers a useful, powerful and easy-to-use way to process RNA-seq data, nicely complementing the existing nf-core catalogue of bioinformatics tools and contributing to the FAANG global action.

3.2.1 Analysis workflow

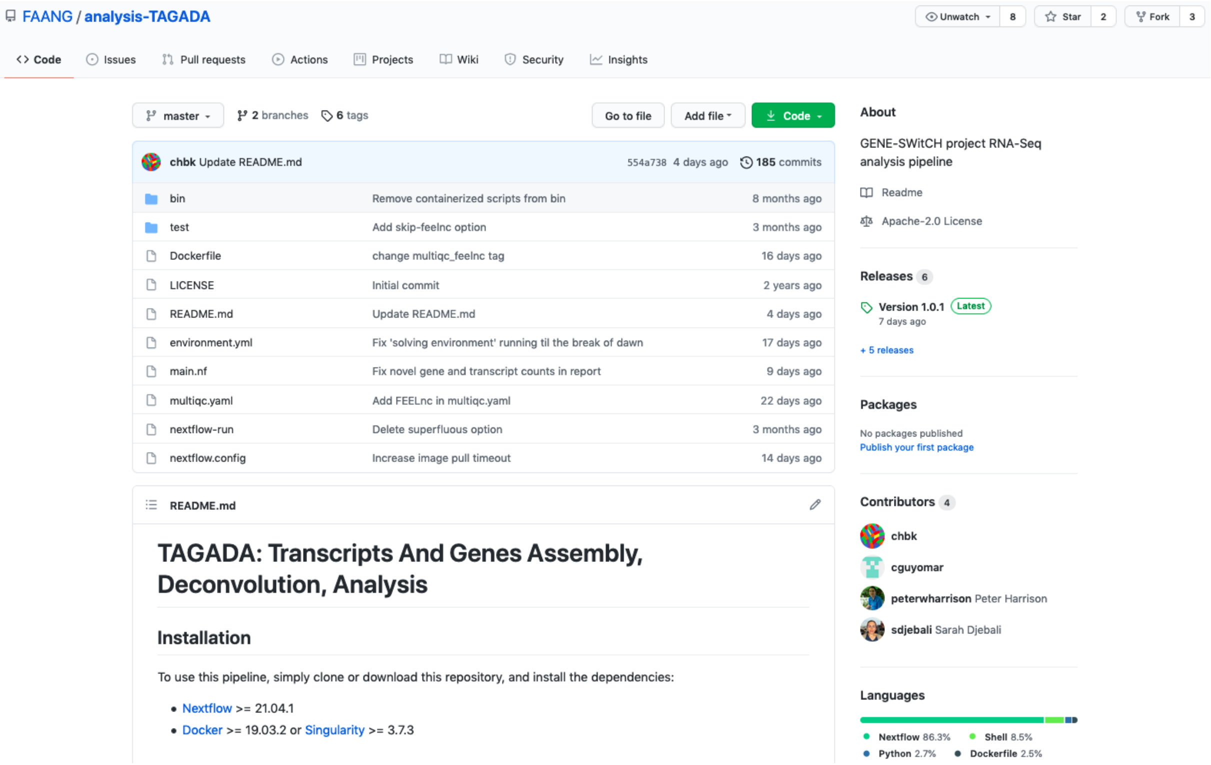
The TAGADA pipeline executes the following main processes:

1. Control reads **quality** with FastQC.
2. **Trim** adaptators from reads with Trim Galore.
3. Estimate **overhang** length of splice junctions, and **index** the genome sequence with STAR. The indexed genome is saved to output/index.
4. **Map** reads to the indexed genome with STAR. The mapped reads are saved to output/maps.
5. Estimate **direction** and **length** of mapped reads, and compute genome **coverage** with Bedtools. The coverage files are saved to output/coverage.
6. **Merge** mapped reads by factors with Samtools.
7. **Assemble** transcripts and **combine** them into a novel annotation with StringTie. The novel annotation is saved to output/annotation.
8. **Detect** long non-coding RNAs with FEELnc.
   The annotations of long non-coding RNAs are saved to output/annotation.
9. **Quantify** genes and transcripts with StringTie, and **format** them into tabulated files. The TPM values and read counts for each annotation are saved to output/quantification.

10. 10. Aggregate quality controls into a **report** with MultiQC. The report is saved to output/control.

3.2.2 Code repository

The TAGADA analysis pipeline is developed in the form of an open-source software. Its code is publicly available on the Github code repository at https://github.com/FAANG/analysis-TAGADA (Figure 2). It is developed under the Apache-2-0 License. It is mainly coded using the NextFlow language (https://www.nextflow.io/) and offers a containerized execution, as per the GENE-SWitCH WP2 Code and Procedure's agreement.



chbk Update README.md 554a738 4 days ago 185 commits

bin	Remove containerized scripts from bin	8 months ago
test	Add skip-feelnc option	3 months ago
Dockerfile	change multiqc_feelnc tag	16 days ago
LICENSE	Initial commit	2 years ago
README.md	Update README.md	4 days ago
environment.yml	Fix 'solving environment' running til the break of dawn	17 days ago
main.nf	Fix novel gene and transcript counts in report	9 days ago
multiqc.yaml	Add FEELnc in multiqc.yaml	22 days ago
nextflow-run	Delete superfluous option	3 months ago
nextflow.config	Increase image pull timeout	14 days ago

README.md

TAGADA: Transcripts And Genes Assembly, Deconvolution, Analysis

Installation

To use this pipeline, simply clone or download this repository, and install the dependencies:

- Nextflow >= 21.04.1
- Docker >= 19.03.2 or Singularity >= 3.7.3

About

GENE-SWitCH project RNA-Seq analysis pipeline

Readme

Apache-2.0 License

Releases 6

Version 1.0.1 Latest
7 days ago

+ 5 releases

Packages

No packages published
Publish your first package

Contributors 4

chbk
cguyomar
peterwharrison Peter Harrison
sdjebali Sarah Djebali

Languages

● Nextflow 86.3% ● Shell 8.5%
● Python 2.7% ● Dockerfile 2.5%

Figure 2. Main page of the code repository on Github

3.2.3 Usage

To run the pipeline a command-line execution is required under the form ./nextflow-run FAANG/analysis-TAGADA --revision 1.0.1 --output directory --profile test docker The following arguments can be provided to the pipeline:

Option	**Parameter(s)**	**Description**	**Requirement**
--profile	profile1 profile2 ...	Profile(s) to use when running the pipeline. Specify the profiles that fit your infrastructure among singularity, docker, kubernetes, slurm.	Required
--output	directory	Output directory where all temporary files, logs, and results are written.	Required

--reads	reads.fq *.bam ...	Input fastq file(s) and/or bam file(s).	Required
--annotation	annotation.gtf[.gz]	Input reference annotation file or url.	Required
--genome	genome.fa[.gz]	Input genome sequence file or url.	Required
--index	directory[.tar.gz]	Input genome index directory or url.	Optional to skip genome indexing.
--metadata	metadata.tsv	Input tabulated metadata file or url.	Required if --merge is provided.
--merge	factor1 factor2 ...	Factor(s) to merge mapped reads.	Optional
--max-cpus	16	Maximum number of CPU cores that can be used for each process. This is a limit, not the actual number of requested CPU cores.	Optional
--max-memory	64GB	Maximum memory that can be used for each process. This is a limit, not the actual amount of alloted memory.	Optional
--max-time	12h	Maximum time that can be spent on each process. This is a limit and has no effect on the duration of each process.	Optional
--resume		Preserve temporary files and resume the pipeline from the last completed process. If this option is absent, temporary files will be deleted upon completion, and the pipeline will not be resumable.	Optional
--feelnc-args	'--mode shuffle ...'	Custom arguments to pass to FEELnc's coding potential script when detecting long non-coding RNAs.	Optional
--skip-feelnc		Skip the detection of long non-coding RNAs with FEELnc.	Optional

The --merge and --metadata options can be used together to merge mapped reads. This results in genes and transcripts being quantified by **factors** rather than by **inputs**.

The metadata file consists of tab-separated values describing your inputs. The first column must contain file names without extensions. There is no restriction on column names or number of columns.

For example, given the following tabulated metadata file: input diet tissue

1. A corn
2. B corn

3. C wheat
4. D wheat

With the following arguments:
--reads A_R1.fq A_R2.fq B.fq C.fq D.bam --metadata metadata.tsv --merge diet
A and **B** mapped reads will be merged, resulting in gene and transcript counts for the **corn** diet. **C** and **D** mapped reads will be merged, resulting in gene and transcript counts for the **wheat** diet. With the following arguments:

--reads A_R1.fq A_R2.fq B.fq C.fq D.bam --metadata metadata.tsv --merge diet tissue

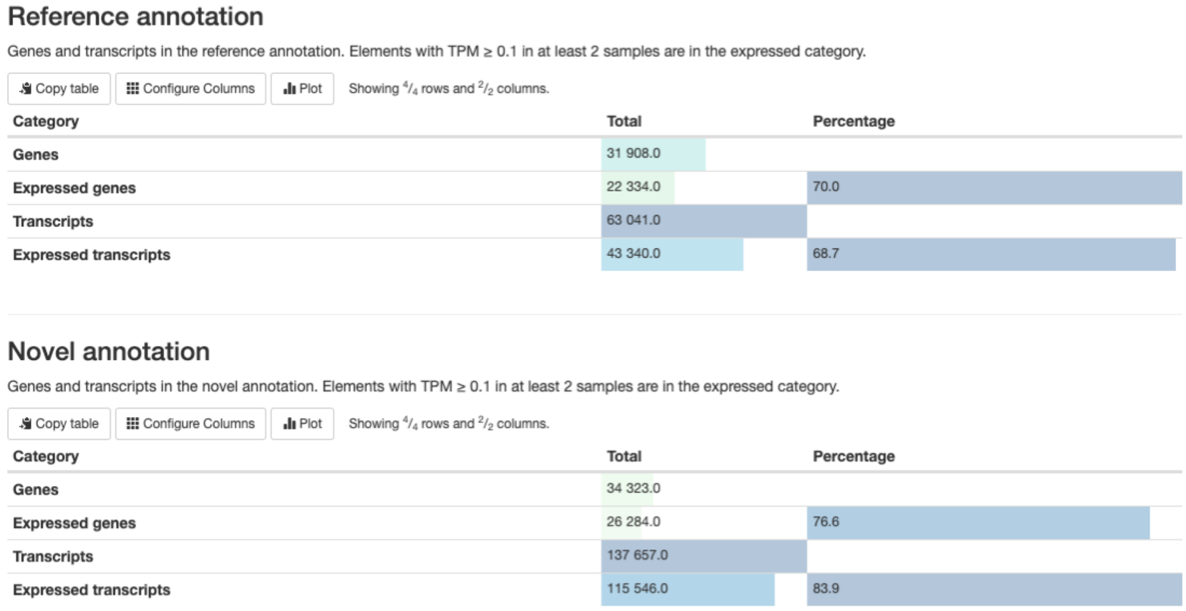
A and **B** mapped reads will be merged, resulting in counts for the **corn** diet and **liver** tissue pair. **C** mapped reads will be left alone, resulting in counts for the **wheat** diet and **liver** tissue pair.
D mapped reads will be left alone, resulting in counts for the **wheat** diet and **muscle** tissue pair.

3.2.4 Results

In addition to the results and log files, an interactive html-based report is provided for each execution to summarize the main Quality Controls (QC) and analysis results. This report is a stand-alone file that can easily be shared with the project's partners, sent to external collaborators, or published online.

The report presents the following QCs and results, corresponding to successive analysis steps:

- A comparison between genes and transcript counts in the reference annotation provided to the pipeline and the novel annotation generated by the pipeline:



Reference annotation

Genes and transcripts in the reference annotation. Elements with TPM ≥ 0.1 in at least 2 samples are in the expressed category.

Showing 4/4 rows and 2/2 columns.

Category	Total	Percentage
Genes	31 908.0	
Expressed genes	22 334.0	70.0
Transcripts	63 041.0	
Expressed transcripts	43 340.0	68.7

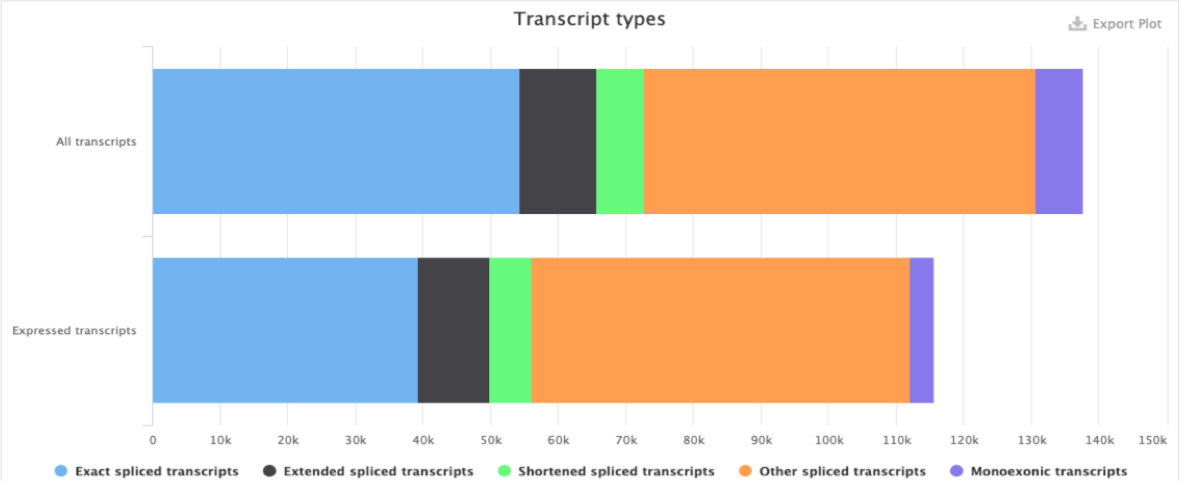
Novel annotation

Genes and transcripts in the novel annotation. Elements with TPM ≥ 0.1 in at least 2 samples are in the expressed category.

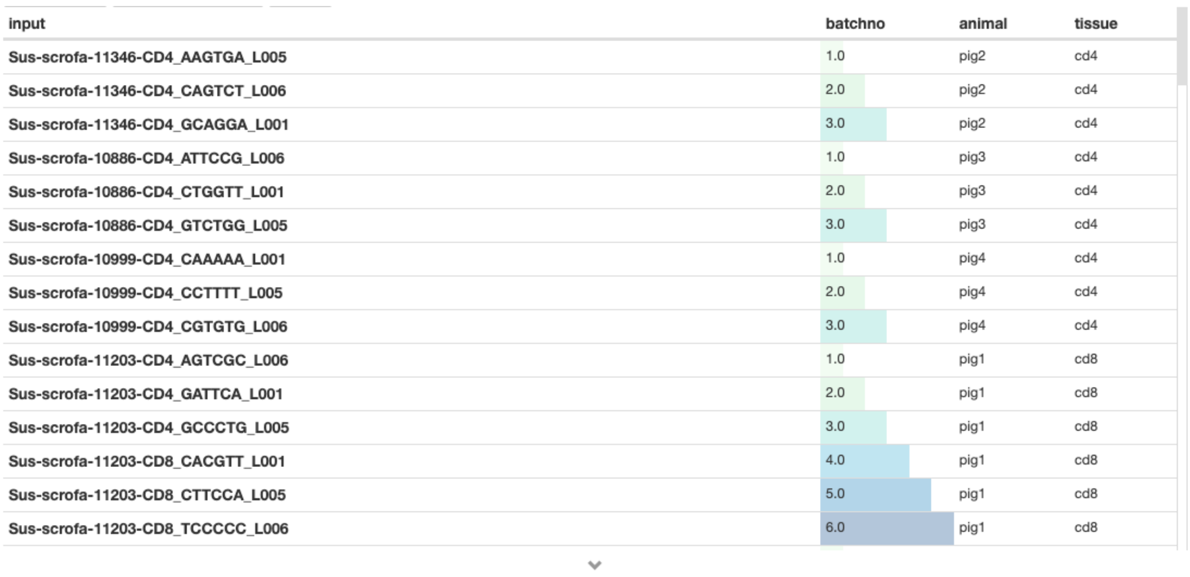
Showing 4/4 rows and 2/2 columns.

Category	Total	Percentage
Genes	34 323.0	
Expressed genes	26 284.0	76.6
Transcripts	137 657.0	
Expressed transcripts	115 546.0	83.9

- How transcripts from the novel annotation differ from the transcripts of the reference annotation:

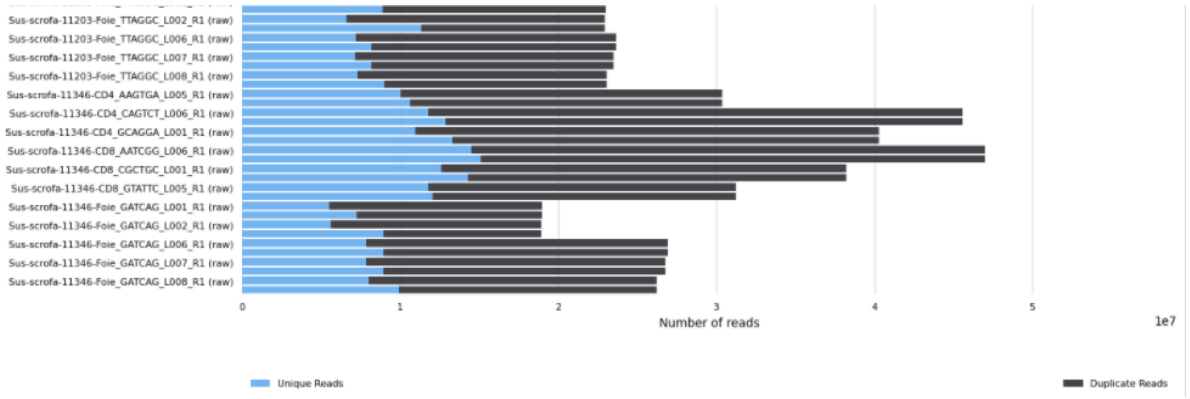


- A summary of the metadata describing the samples:

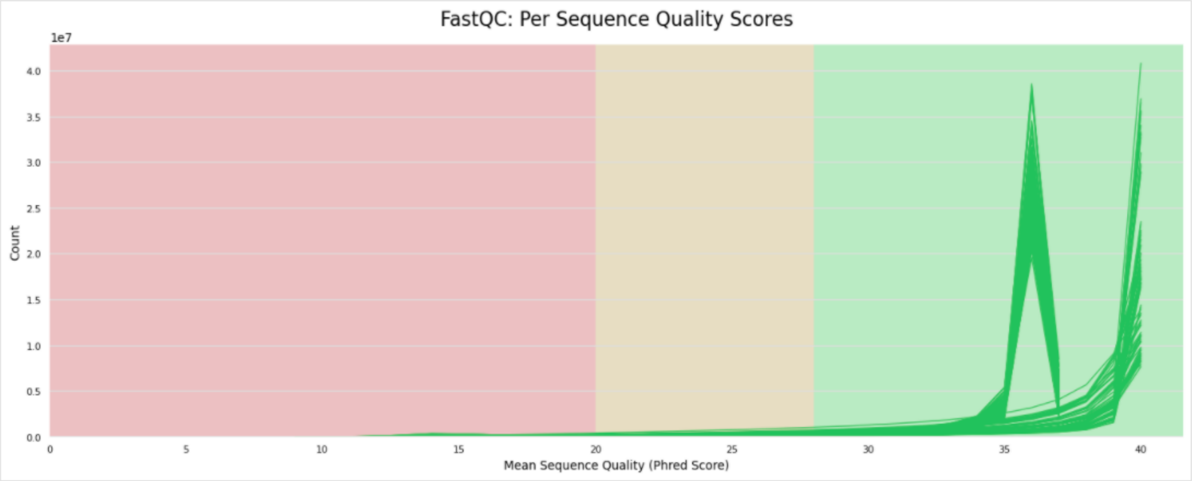


input	batchno	animal	tissue
Sus-scrofa-11346-CD4_AAGTGA_L005	1.0	pig2	cd4
Sus-scrofa-11346-CD4_CAGTCT_L006	2.0	pig2	cd4
Sus-scrofa-11346-CD4_GCAGGA_L001	3.0	pig2	cd4
Sus-scrofa-10886-CD4_ATTCCG_L006	1.0	pig3	cd4
Sus-scrofa-10886-CD4_CTGGTT_L001	2.0	pig3	cd4
Sus-scrofa-10886-CD4_GTCTGG_L005	3.0	pig3	cd4
Sus-scrofa-10999-CD4_CAAAAA_L001	1.0	pig4	cd4
Sus-scrofa-10999-CD4_CCTTTT_L005	2.0	pig4	cd4
Sus-scrofa-10999-CD4_CGTGTG_L006	3.0	pig4	cd4
Sus-scrofa-11203-CD4_AGTCGC_L006	1.0	pig1	cd8
Sus-scrofa-11203-CD4_GATTCA_L001	2.0	pig1	cd8
Sus-scrofa-11203-CD4_GCCCTG_L005	3.0	pig1	cd8
Sus-scrofa-11203-CD8_CACGTT_L001	4.0	pig1	cd8
Sus-scrofa-11203-CD8_CTTCCA_L005	5.0	pig1	cd8
Sus-scrofa-11203-CD8_TCCCCC_L006	6.0	pig1	cd8

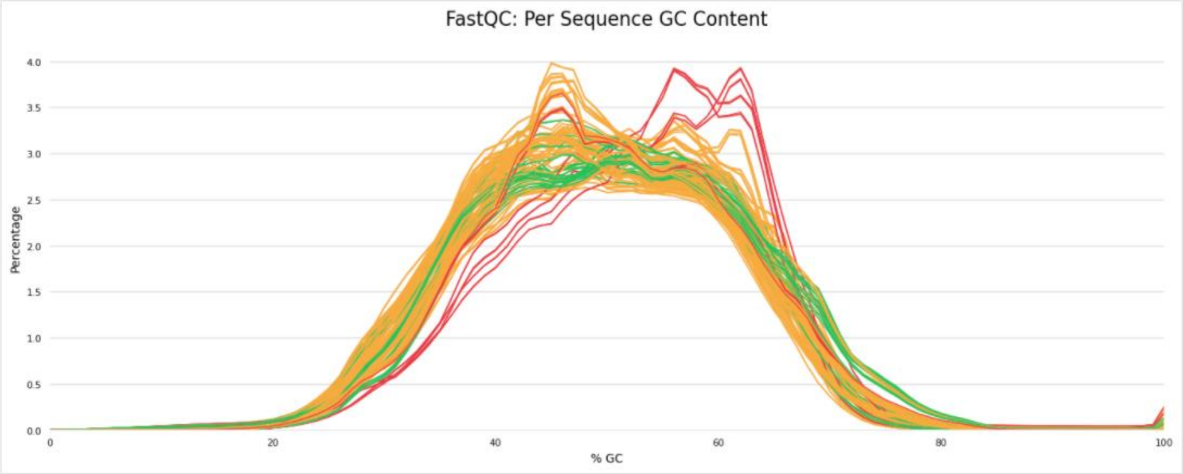
- The counts of unique (blue) and duplicated (black) reads:



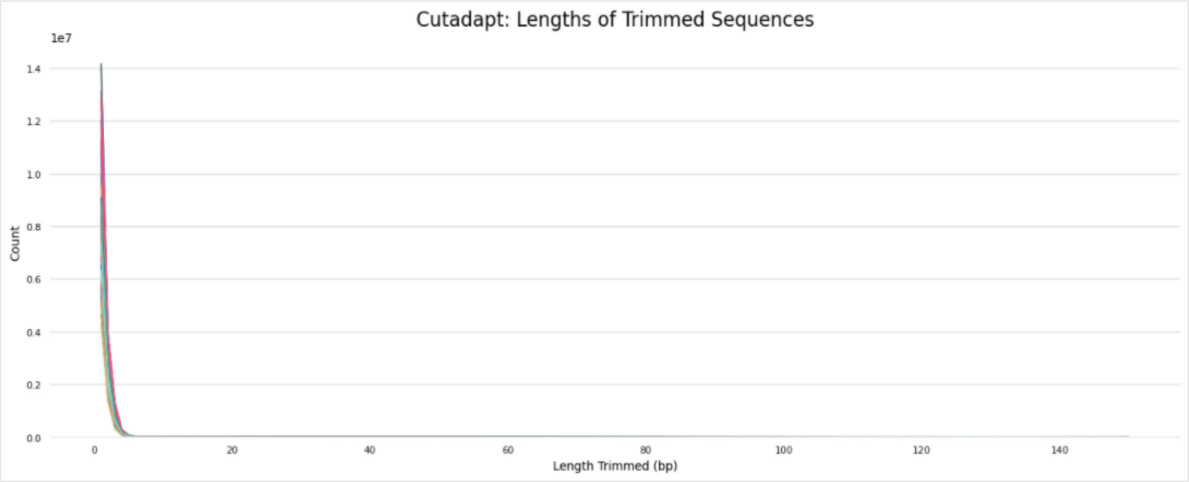
- The distribution of average sequence quality scores:



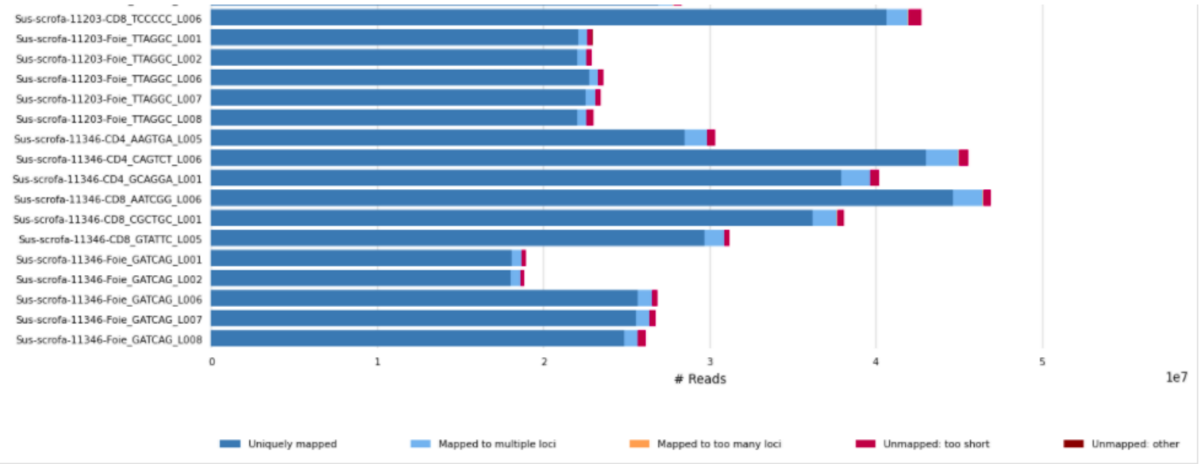
- The percentage of GC bases per sequence:



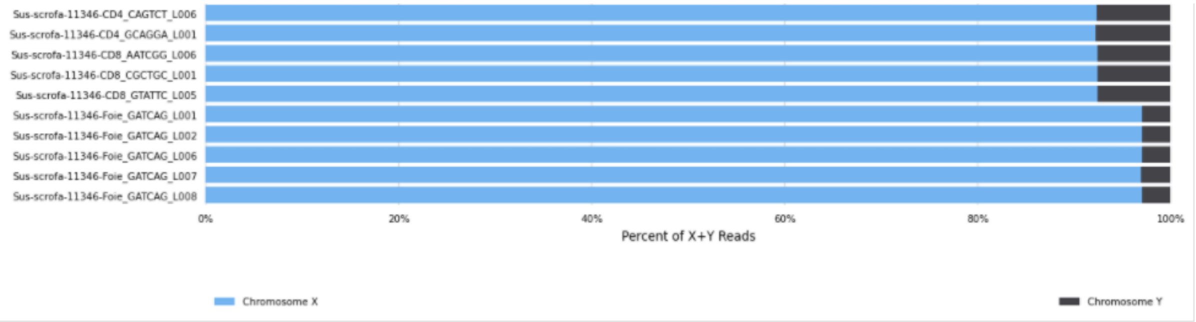
- The length of trimmed sequences (adapters):



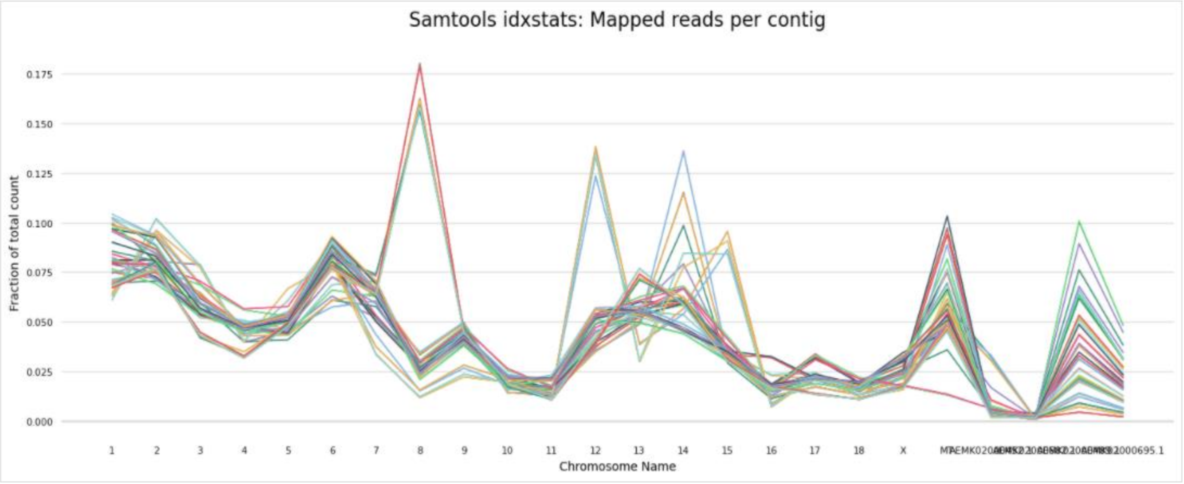
- The proportions of mapped and unmapped reads:



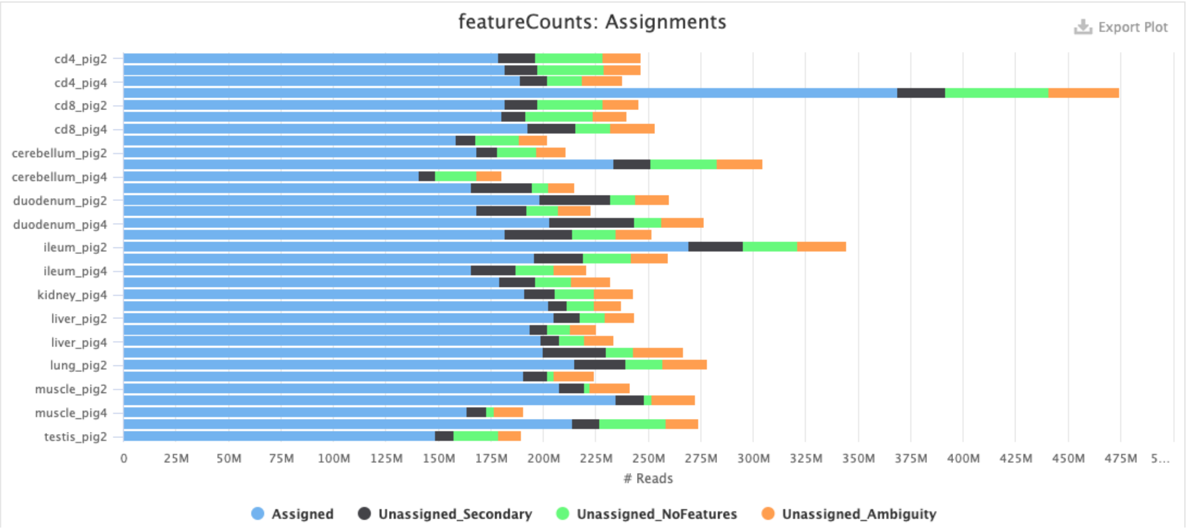
- The proportions of X chromosome and Y chromosome reads:



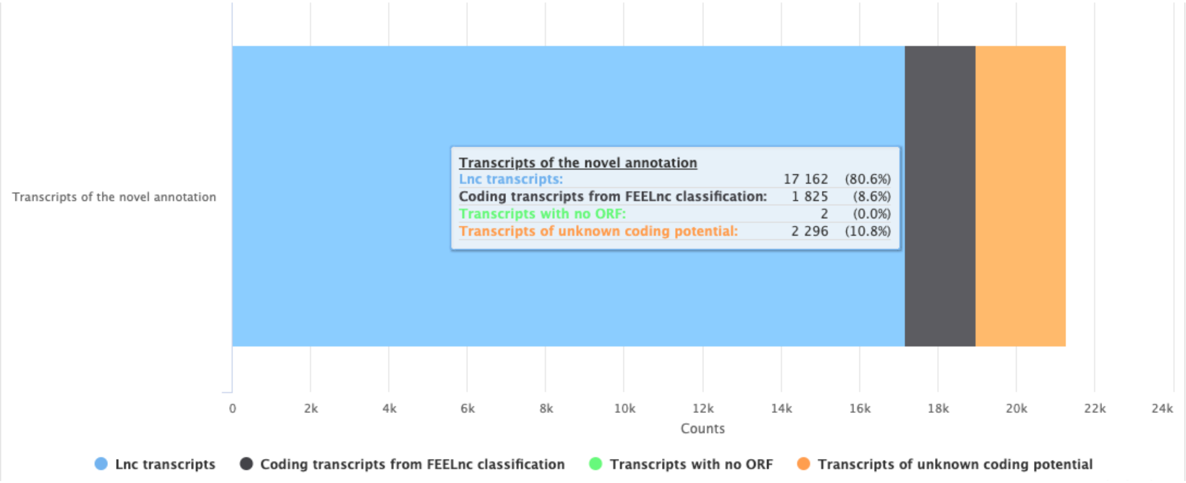
- The fraction of mapped reads per contig:



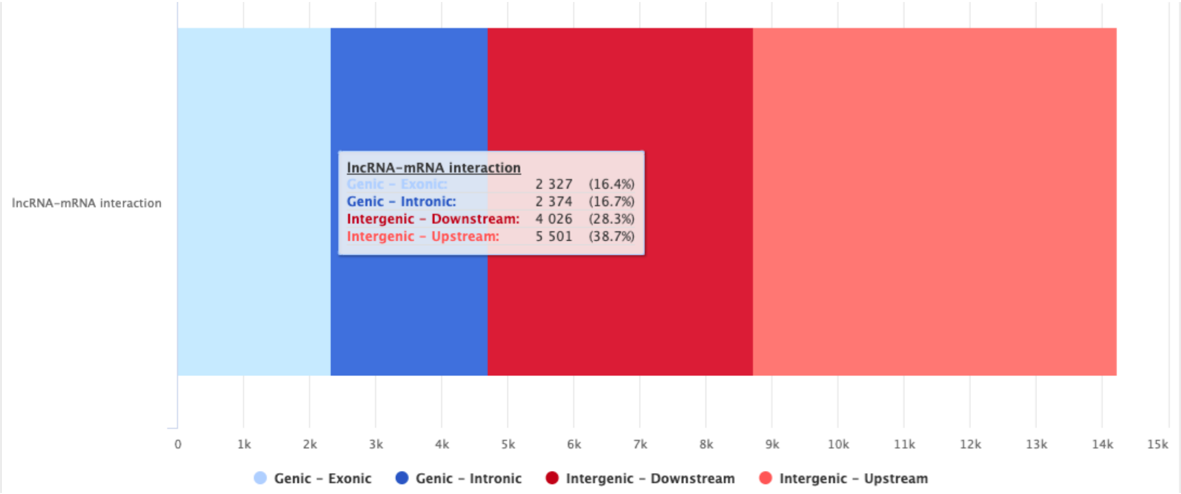
- The number of exonic reads (blue) per experiment group:



- The number of long non-coding transcripts in the novel annotation:



- The positions of long non-coding transcripts relative to the closest coding transcript:



The pipeline was presented during the BovReg Nextflow workshop organized by the CRG on 17-20 November 2020: https://www.bovreg.eu/nextflow-and-nf-core-workshop-by-crg/

4 Conclusion

The TAGADA pipeline is currently being used on EMBL-EBI's Embassy Cloud infrastructure to process RNA-Seq chicken data and pig data delivered by WP1. The processes launched by the pipeline are parallelised to make use of all resources simultaneously and accelerate the delivery of results. Following extensive testing during the early part of the project, the platform is now in production use processing the primary analysis data from work package 1 to make a standardised product for the consortia's downstream analyses.

In parallel to the RNA-Seq TAGADA pipeline that we have exemplified in this report, work package 2 are also preparing for the production analysis of the other GENE-SWitCH data that is being generated. This includes:

- Iso-Seq analysis using the TAMA workflow https://github.com/GenomeRIK/tama that will liekleu become the nf-core (https://nf-co.re/) nf-isoseq pipeline.
- Bisulphite analysis using the Methyl-Seq workflow https://github.com/FAANG/proj-gs-meth
- Remaining pipelines will be adopted from nf-core (https://nf-co.re/) that are being collaborated on with EuroFAANG colleagues.

5 Deviations or delays

This deliverable was delayed by six months due to the secondment of EMBL DCC staff from March to September 2020 to support the emergency European COVID-19 scientific response to develop the European COVID-19 data portal (https://www.covid19dataportal.org/).

6 Glossary

DCC	Data Coordination Centre
EURO-FAANG	The European node of the Functional Annotation of Animal Genomes Data Coordination Centre
FAANG	Functional Annotation of Animal Genomes
FAIR	Findable, Acessible, Interoperbale and Reusable