NEXT Platform

A National Genomic Data Sharing Platform
Specifications v0.3 – Confidential

Author: Charles Vesteghem

Use cases

1 - Storing and management of genomic data

A centre has produced a .vcf file of either somatic mutations, germline variants or trio. One of the bioinformaticians of centre logs in to the NEXT Platform, accessible online within the hospital network, to upload the file. In order to proceed with the upload, metadata must be provided, some being mandatory (cpr, project_id, date, lab_info_id), the rest being optional (sex, diagnostic, topographies, etc).
Once the file has been uploaded together with its metadata, it appears in the list of entries of the centre. This interface is only available to users of the centre. Only the admin user of the centre can edit an entry.

2 - Finding participants for clinical trials

A pharmaceutical company would like to investigate the feasibility of organising a clinical trial in Denmark for a targeted drug. The company gets in contact with one centre and presents its needs. A user of the centre logs in to the NEXT Platform and search for potential participants among participating centres based on criteria such as age range, diagnosis and its recency, genes mutations, etc. A result page is displayed showing how many participants in each centre fits with the requirements and for each centre an admin user to be contacted. Once contacted, the admin users of the other centres can give permission to the cases of interest for a limited duration.

Tables description

User

Health care professional who has access to the platform through a login mechanism. The users are identified by their email address. A user belongs to a centre.

Centre

Research/Clinical unit. Initially, a user of a centre will have access to all the data related to this centre and limited access to other centres data. Roles of users within a centre should be defined better in a later phase.

Project

The data is structured in projects. A project belongs to a centre and includes cases.

Patient

The person providing the main tissues from which the genomic files are generated. A patient can have multiple cases associated. Typically, in case of cancer, a patient's relapse is seen as a new case.

Case

The occurrence during the patient's disease trajectory from which the tissues originate. The case can contain basic clinical information about the occurrence, i.e. the diagnostic and its date. In case of cancer some additional information can be provided such as topography, and relapse number.

In order to enforce interoperability, the diagnostic will be based on the 2016 version of the ICD-10 classification (http://apps.who.int/classifications/icd10/browse/2016/en/). In case of cancer, the topography will be based on the ICD-O-3.1 classification (http://codes.iarc.fr/topography). A more agnostic solution for diagnostic could be implemented in a later phase to facilitate the collection of these information. In case of genetic or chronic diseases there will be a unique case per patient.

Permission

The admin user of a centre can give access to a case, and its associated data, from their centre to a user from another centre. This permission can be time limited.

File

Genomic file associated to the case. The only type of data supported initially will be .vcf files (https://github.com/samtools/hts-specs). This file can contain somatic SNVs and small indels (in case of cancer), germline mutations or de novo mutations (for trios).

Variant

Single entry in the .vcf file showing a variant of enough confidence in the genetic material. This variant can be linked to dbSNP (https://www.ncbi.nlm.nih.gov/snp/) and, for cancer, COSMIC (http://cancer.sanger.ac.uk/cosmic) databases to ease annotations and findability.

Lab info

High level Information about the whole process applied to the tissue, i.e. the capture kit and the NEXT pipeline used.

Pipeline

Common set of bioinformatics workflows applied to the raw genomic data in the context of the NEXT project. Summarized as a version number and a URL to the definition/specification.

Ref Genome

Reference genome used in the pipeline for the alignment of the reads from the sequencing of the tissue. The platform could handle multiple reference genomes following the development of the pipeline.

Gene Location

Location of a gene in a reference genome.

Gene

Gene defined by a HGNC symbols (http://www.genenames.org/). Its location might change in different reference genomes.

Overview of the database structure