LA STRUCTURATION ET L'INTÉGRATION DE DONNÉE OMIQUES: LE PROJET ODIN

FORMATION

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Plan

- **Gustave Roussy**
- **Définitions (prélèvement, donnée)**
- **Cycle de vie de la donnée : traçabilité**
- **Cas d'utilisation: le projet ODIN**
  - ODIN : catalogue de données
  - ODIN : architecture d'un système informatique
  - ODIN : Modules

- **Atelier : Initiation base de donnée**

- **Bonus1: Travail en équipe**
- **Bonus2: Développement informatique**

Gustave Roussy : Centre de lutte contre le cancer

- **Soin**
- **Recherche translationelle**
  - Médecine personalisée (de précision)
- **Recherche fondamentale**

- **Métiers**
  - Médecin
  - Biologiste
  - Chercheur
  - Technicien
  - Attaché de recherche clinique (ARC)
  - Biostatisticien
  - Bioinformaticien
  - Informaticien

Définition - échantillon

- **Identifiants**
  - Dossier patient (NIP = numéro d'identité personnelle)
  - Prélèvement
    - Chirurgie (biopsie, résection) : numéro histopathologie
    - Sang : date
- **Logiciel de laboratoire**
  - SGL: Système de gestion de laboratoire
  - LIMS: Laboratory Information Management System

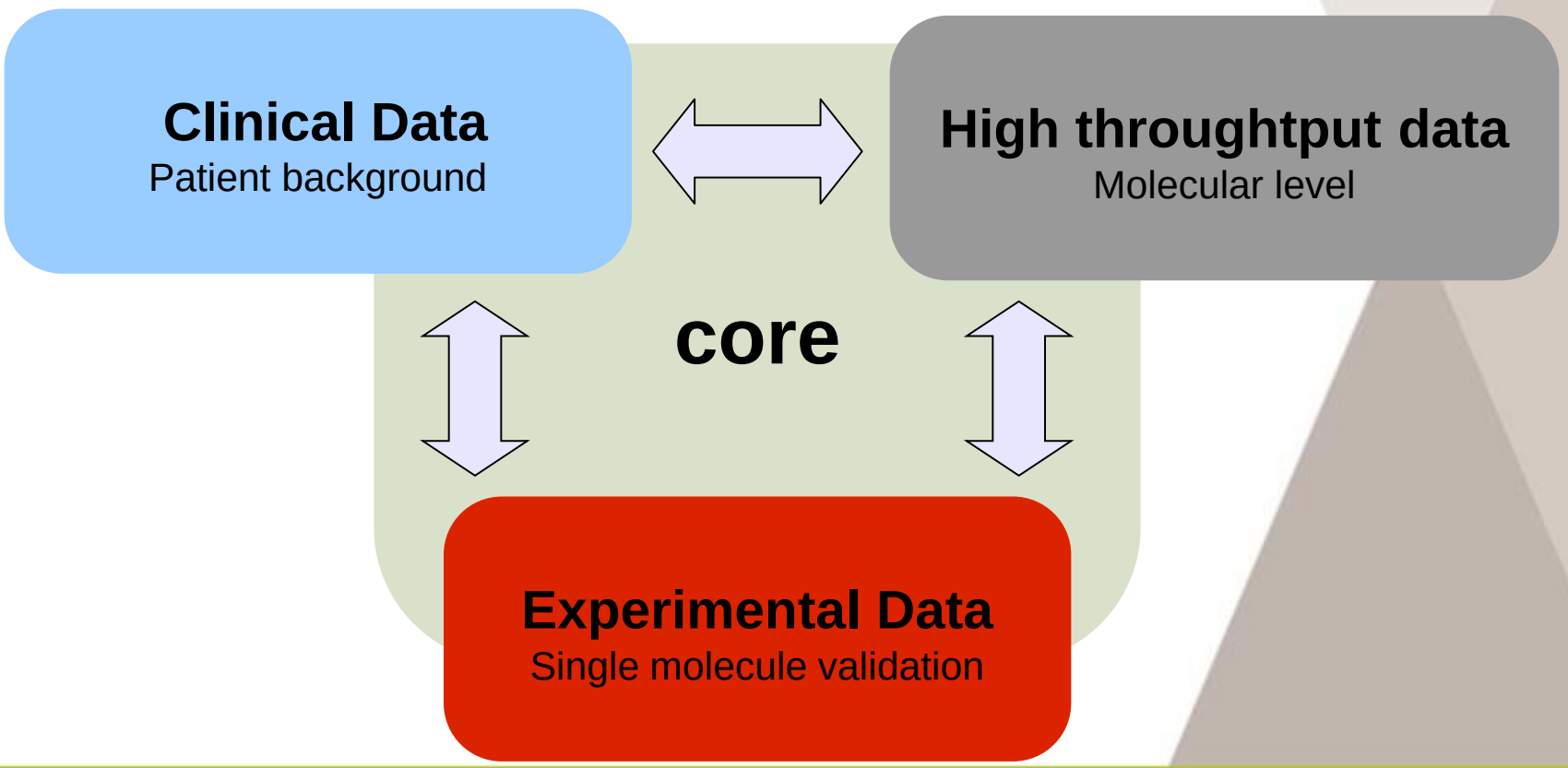
Définition - Structuration de données

- **Donnée**
  - Rapport
  - Tableau excel != Base de données
  - Analyses moléculaires
- **Gestion de données (Data Management)**
  - Métadonnées (catalogue)
  - Plan de gestion de la donnée
    - Quelle données, Où (source, destination, transformations)
    - Qui (profils: consultation / modification)
  - FAIR data : Findable, Accessible, Interoperable, Re-usable
- **Logiciels**

Définition - Intégration de données

Query: Pathology **AND** (Secreted molecules **OR** Modulated Genes)

Breast Cancer TNF secretion TNF expression



Cycle de vie de la donnée (Flux)

- *Prélèvement (biopsie)*
- *Produits dérivé (ADN)*
- Mesure (DNA-seq)
- Répertorier (fichiers)
- Structurer, contrôle d'intégrité (bioinfo pipeline)

- Base de résultat, portail d'interprétation

- Rapports

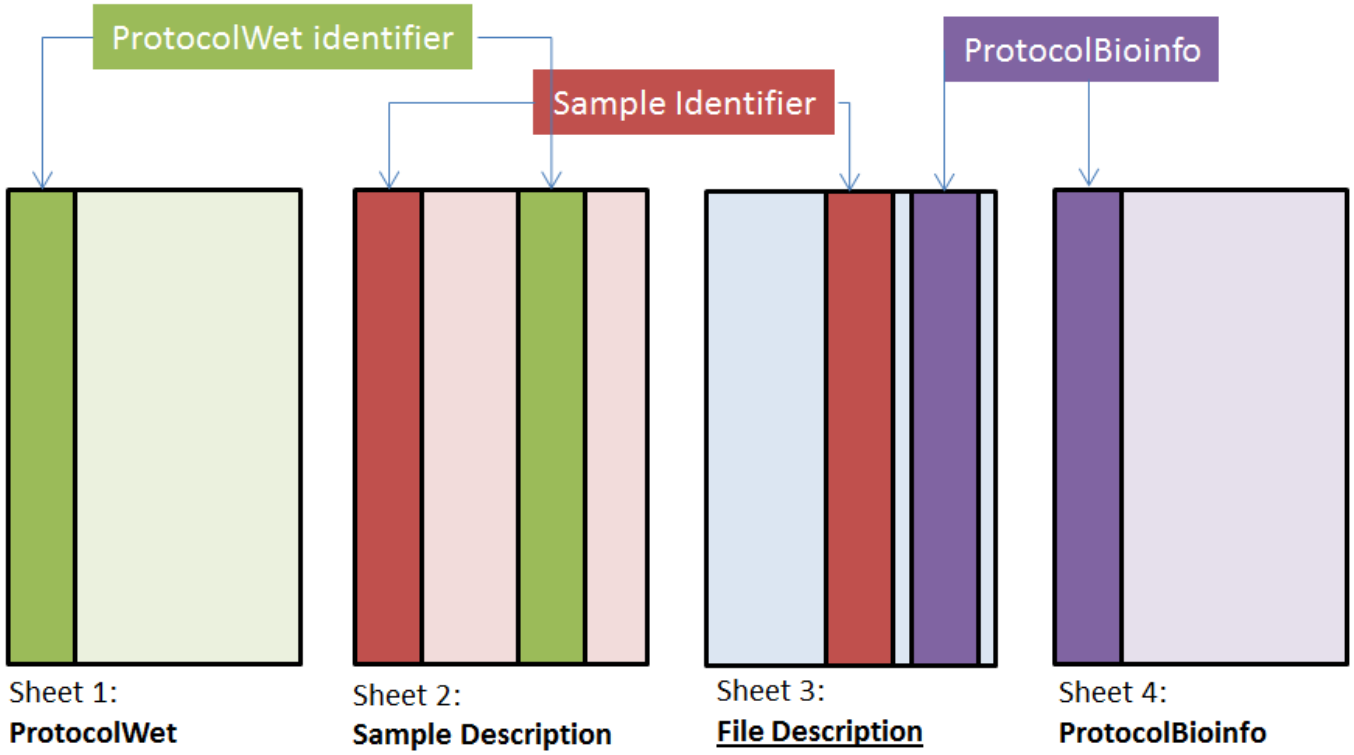
Enjeux de la gestion de données

- **Notion de droit**
  - Type de droit : lecture / écriture / suppression / mise à jour (CRUD)
  - Rôle – selon les métiers (designer, team member, administrateur)
  - Utilisateur
- **Intégrité des données**
  - Vocabulaire contrôlé
  - Exemple : md5sum – s'assurer qu'un fichier n'est pas modifié.
- **Etre capable de retouver la données en fonction de concept métier (Patient, échantillon, étude, projet)**
  - Cataloguer la donnée

Cas d’utilisation

- **Projet de médecine personnalisé: détection de mutation dans le but de:**
  - orienter le traitement.
  - Inclusion dans un essai thérapeutique
- **Analyses ancillaires**
  - Cohortes virtuelles : à partir de plusieurs projets ressources
  - Recherche translationelle : i.e. modèles xénogreffes
  - Recherche fondamentale

Cas d'utilisation : Donné produite par des collaborateurs, recueil de métadonnées



Describe how the samples have been prepared and processed. There is one line by type of measurement.

Describe the sample, one line correspond to one sample.

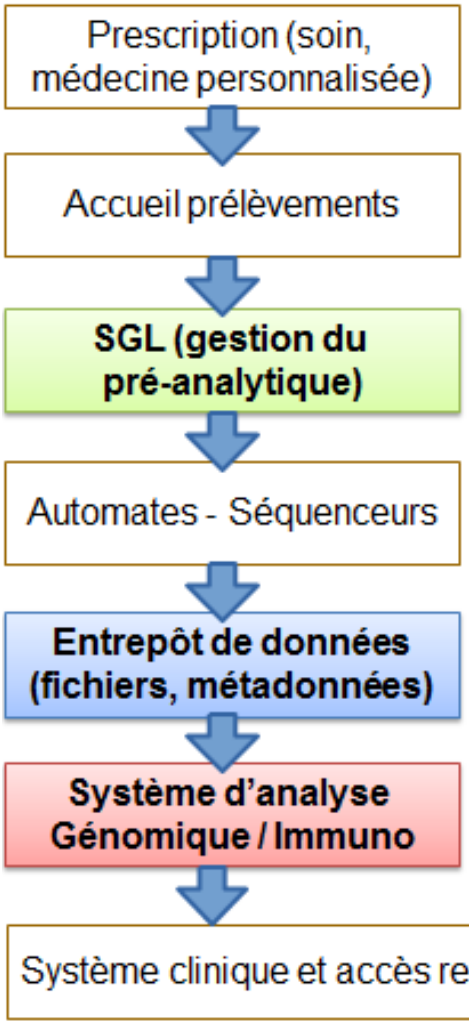
one line of this table correspond to each file shared. Other column refers to the previously described sheets.

if data file is not raw (file are described in "File Description" sheet), it describes the bioinformatic analysis that leads to processed data.

ODIN - Organisation des Données génomiques et d'immunologie

- **Objectifs**
  - Structurer la donnée à différents niveaux au sein d'un système unique
  - Donner accès aux données à des fins de recherche
  - Assurer l'interopérabilité des différents outils utilisés, depuis la réception de la prescription et de l'échantillon jusqu'au rapport d'analyse et aux bases de données
- **Périmètre**
  - Génomiques
    - Soin et med perso (rapport clinique)
    - Tous types d'altérations et technologie utilisées
    - Génétique constit et génomique somatique (biopsies, plasma)
  - Immunomonitoring : suivi longitudinal de patient (comptage cellulaire, mesure de concentration)
- **Un projet transverse**
  - Plateforme bioinfo
  - DTNSI (Direction de la transformation numérique et des services informatiques)
  - Biopath et BMO (Laboratoire de biopathologie)
  - LIO (Laboratoire d'Immunomonitornig)

ODIN 3 sous-systèmes principaux : implémenter le cycle de vie de la donnée



Les principes :

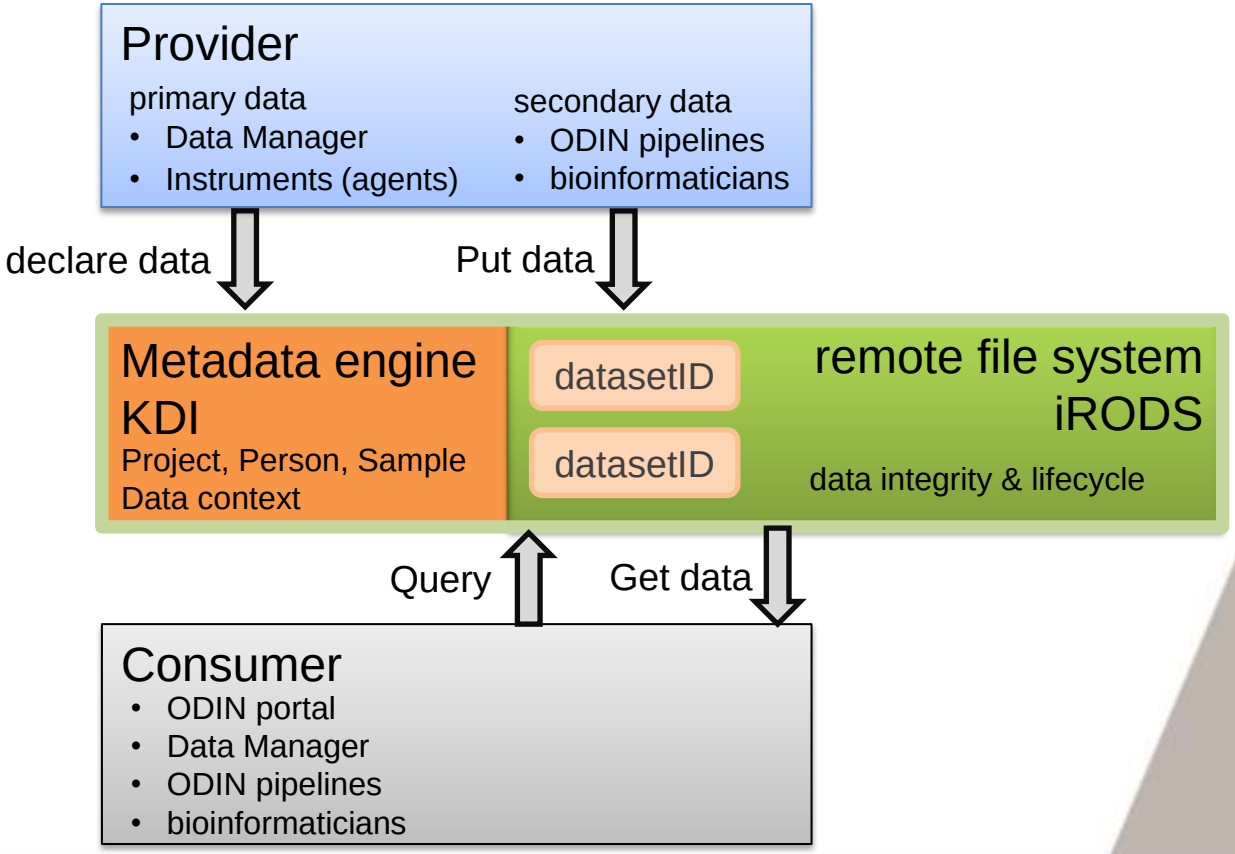
- Faire reposer le pré-analytique sur le SGL d'un éditeur
- S'inscrire d'emblée dans un management de la qualité et de certification
- Regrouper les applications au sein d'un nombre limité de systèmes

- Optimiser l'infrastructure
- Constituer un catalogue de donnés

- Garder la maitrise du système d'analyse génomique par un développement interne
- Enrichir une base de connaissance / de résultats

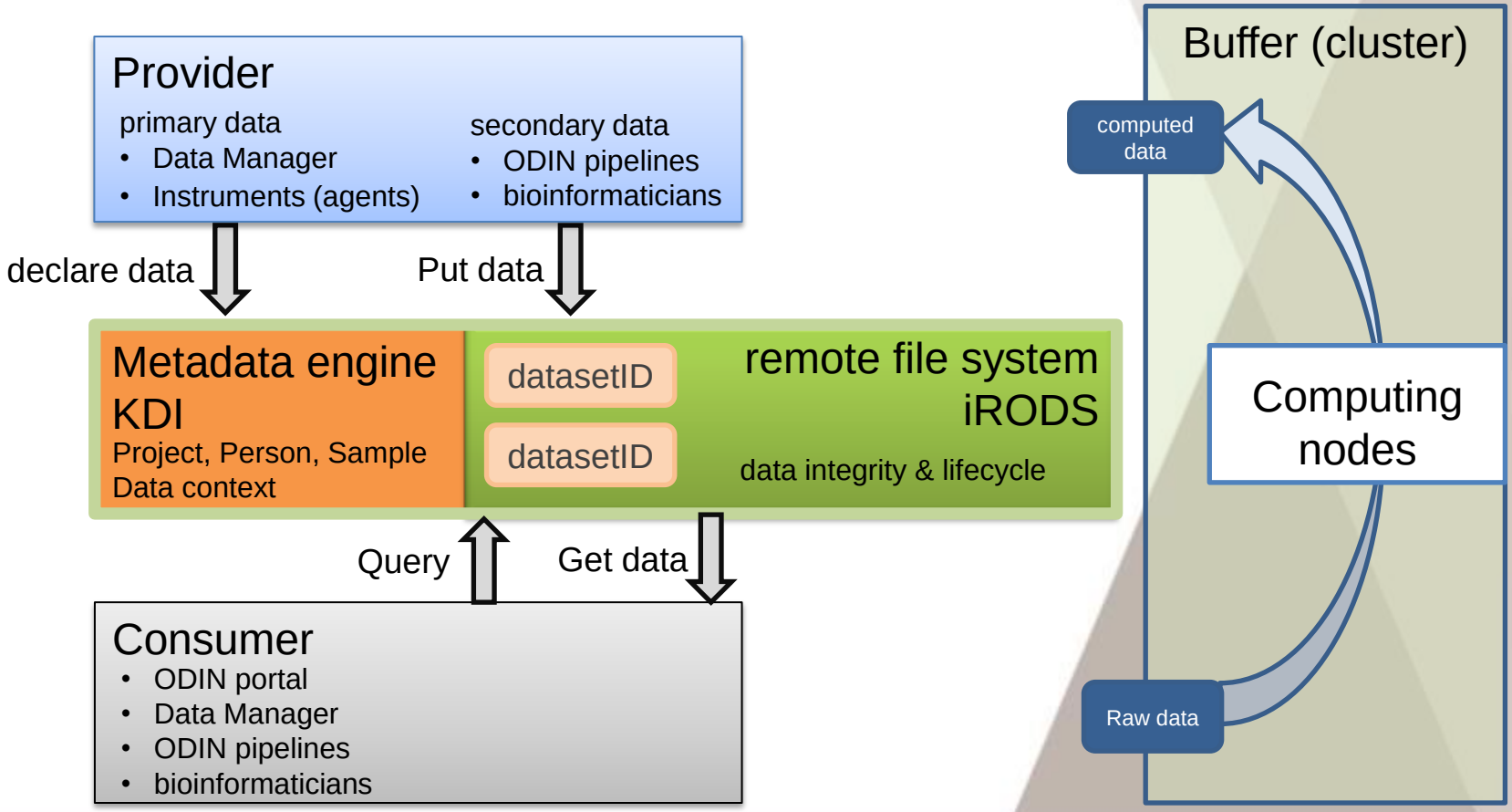
ODIN file warehouse

- **data traceability & availability**
- **storage unit and attached metadata**
  - implement study design
  - re-use data
- **fine grain user right**



ODIN file warehouse

- **data traceability & availability**
- **storage unit and attached metadata**
  - implement study design
  - re-use data
- **fine grain user right**



Portail d'interprétation de variants 1/2

ODIN
QUALIFICATION

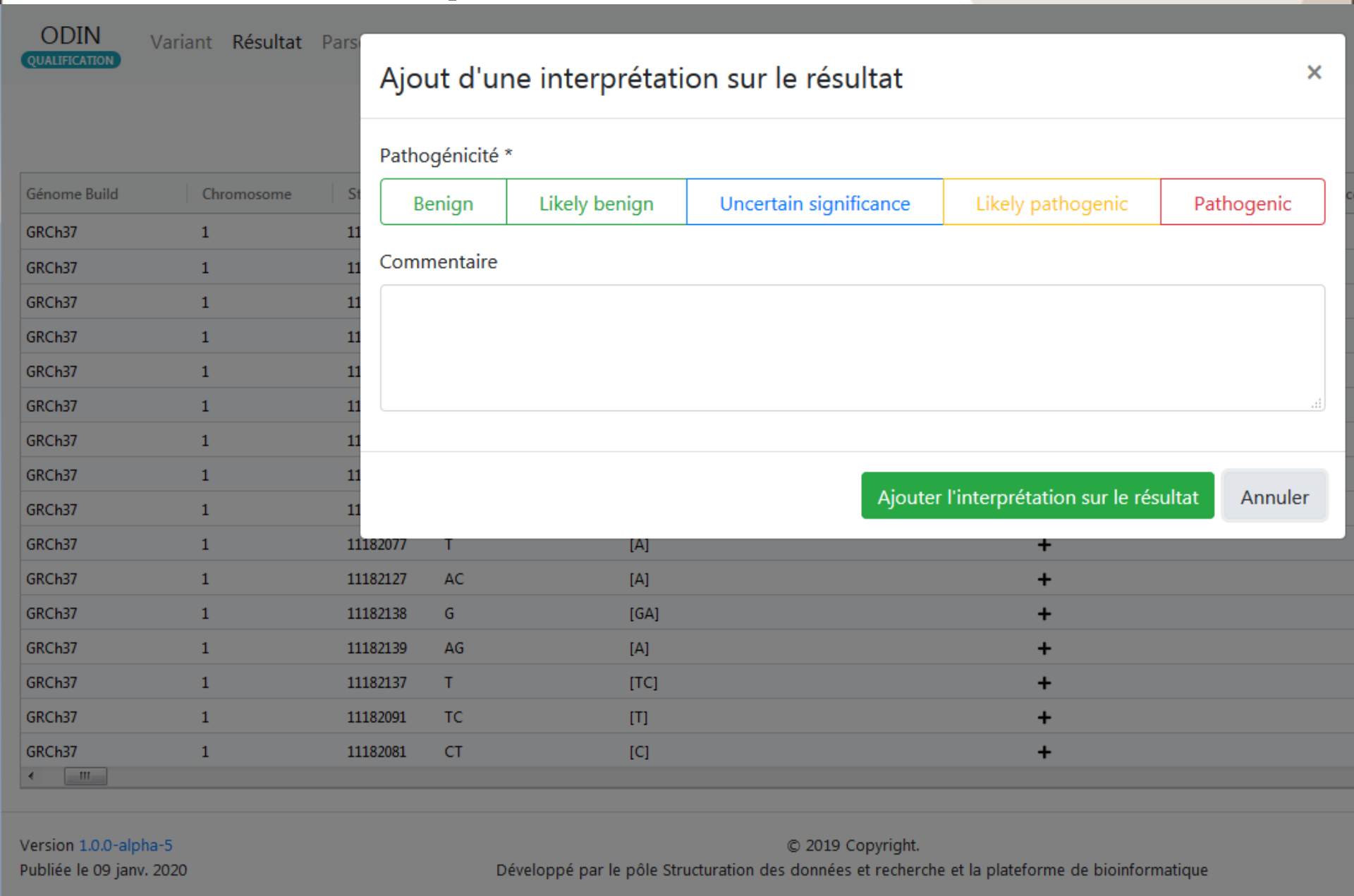
Variant Résultat Parseur

Génome Build	Chromosome	Start	Allèle de référence	Allèle alternatif	Pathogénicité	Pathogénicité (connaissance)
GRCh37	1	11177054	T	[TGA]	+	
GRCh37	1	11177075	TA	[T]	+	
GRCh37	1	11177075	T	[TA]	+	
GRCh37	1	11177083	A	[AC]	+	
GRCh37	1	11182075	CT	[C]	+	
GRCh37	1	11182070	A	[G]	+	
GRCh37	1	11182072	C	[A]	+	
GRCh37	1	11182077	T	[G]	+	
GRCh37	1	11182078	C	[A]	+	
GRCh37	1	11182077	T	[A]	+	
GRCh37	1	11182127	AC	[A]	+	
GRCh37	1	11182138	G	[GA]	+	
GRCh37	1	11182139	AG	[A]	+	
GRCh37	1	11182137	T	[TC]	+	
GRCh37	1	11182091	TC	[T]	+	
GRCh37	1	11182081	CT	[C]	+	

Version 1.0.0-alpha-5
Publié le 09 janv. 2020

Développé par le pôle Structuration des données et recherche et la plateforme de bioinformatique

Portail d'interprétation de variants 2/2



1 million de Génomes Humain d'ici 2022

1. ACGT, Czechia (2018)
2. FarGen, Denmark (2011)
3. France Médecine Génomique 2025 (2016)
4. Danish National Genome Center (2019)
5. GoNL, BBMRI-NL BIOS (2010)
6. Genomics England (2013)
7. FinnGen (2017) and SISu (2015)
8. Estonian Biobank (2000)
9. Scottish Genomes Partnership (2015)
10. UK Biobank (2006)
11. NCER-PD, Luxembourg (2015)
12. NCMG, Czechia (2015)
13. Genomic Medicine Sweden (2018)
14. Slovenian National contact point for rare diseases (2016)
15. Swiss Personalised Health Network (2017)
16. Welsh Genomics for Precision Medicine Strategy (2016)
17. Northern Ireland Genomic Medicine Centre (2017)
18. National Bionics programme, Hungary (2018)
19. National Oncology programme, Hungary (2015)

Fig. 1 | **Examples of current health care-focused and genomics-based national initiative projects across ELIXIR**

https://www.ncbi.nlm.nih.gov/pubmed/31455890

Partage de données

- **Volumes pour un échantillon:**
  - WES paired-end: 10Go
  - RNAseq paired-end: 10Go
  - Whole Genome: 100 Go
  - ChipSeq paired-end : 5Go
- **Programmes de Médecine personnalisé = 100aines To**
  - Limitation Technique: utilisation des réseaux
  - Limitation Réglementaire: données à protéger

- **Solution: Partage du catalogue uniquement**
  - Dénombrements
  - Permet de s'adapter aux legislations selon les partenaires
  - RGPD

TRAVAIL EN EQUIPE

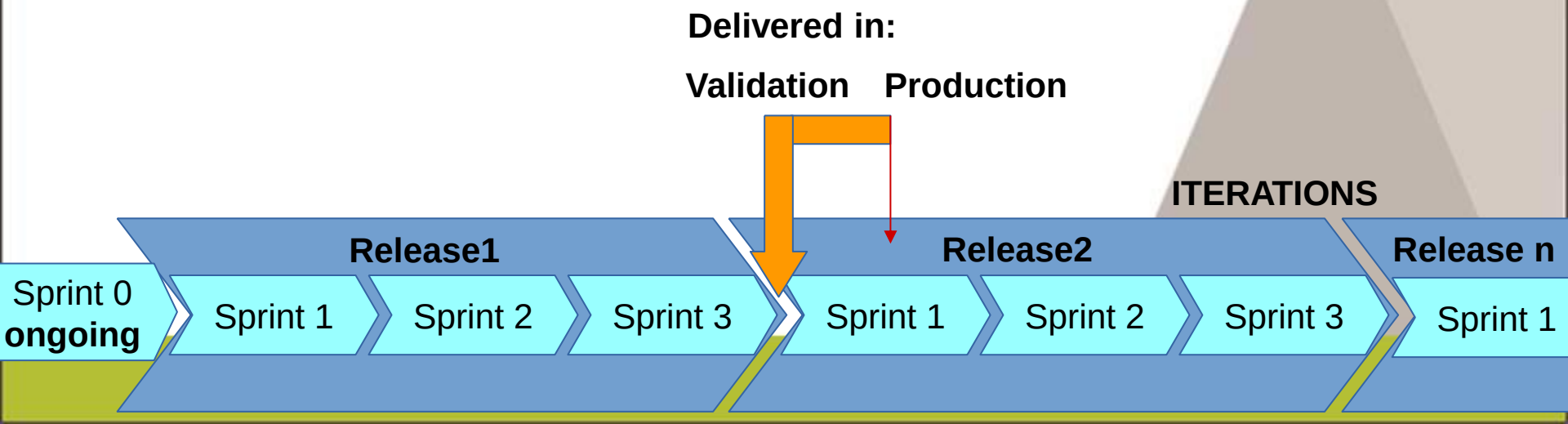
Méthodologie Agile SCRUM

SCRUM : main concepts

Gestion de projet :
- Définition du produit en continu
- Livraison régulière : identifier des versions intermédiaires fonctionelles (stables)
- Amélioration continue

Finir les sujets définis

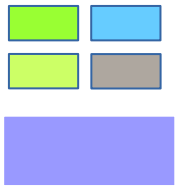
Outil : icescrum, Trello



SCRUM : main concepts

Story (working unit)

- Topic defined and estimated by the team (refinement meeting), with functional and technical area
- Sorted by value
- Have a definition of done

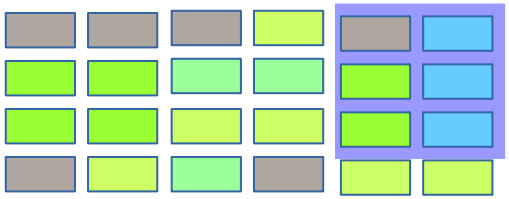


As a <user profile>, I can <do something>
<detail use case, screens>

Develop a secured REST API for <application>
<API endpoint description>

- User stories
- Technical stories
- Epics / Features (category of stories)

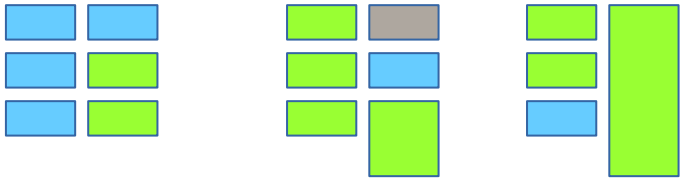
Product Backlog (list of stories)



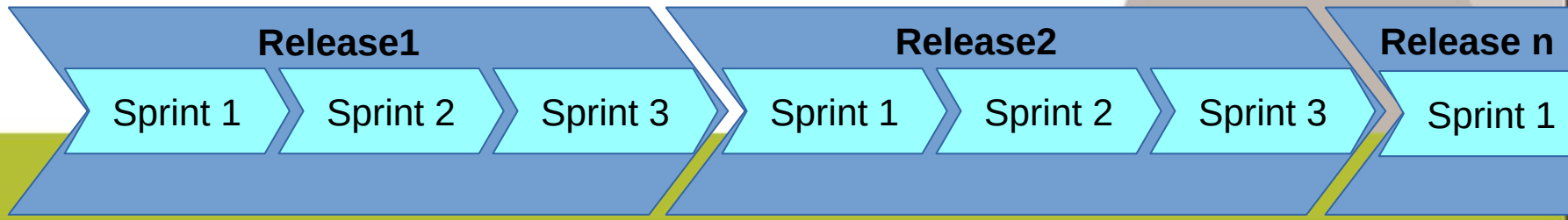
Color code:

- KDI
- Infrastructure
- Application CNV
- Pipeline

Sprint Backlog (selected for development)



ITERATIONS



SCRUM roles

Business Owner:

Give the main directions, prioritize the features for the next release

Product Owner:

Describe the product (application) = write the features and the user stories, know the business logic

Team member:

Write the user stories and technical stories, implement the stories

Scrum Master:

Keeper of the method, plan the meetings and topics

SCRUM meetings & timeline

Team member located in bioinformatics plateform and biopatho:

-Interaction within team: slack (live chat)

-Interaction with Product Owners:

daily stand up meeting (15 minutes) :

what has been done, what will be done

weekly refinement meeting (1 hours):

Discuss & estimate the stories

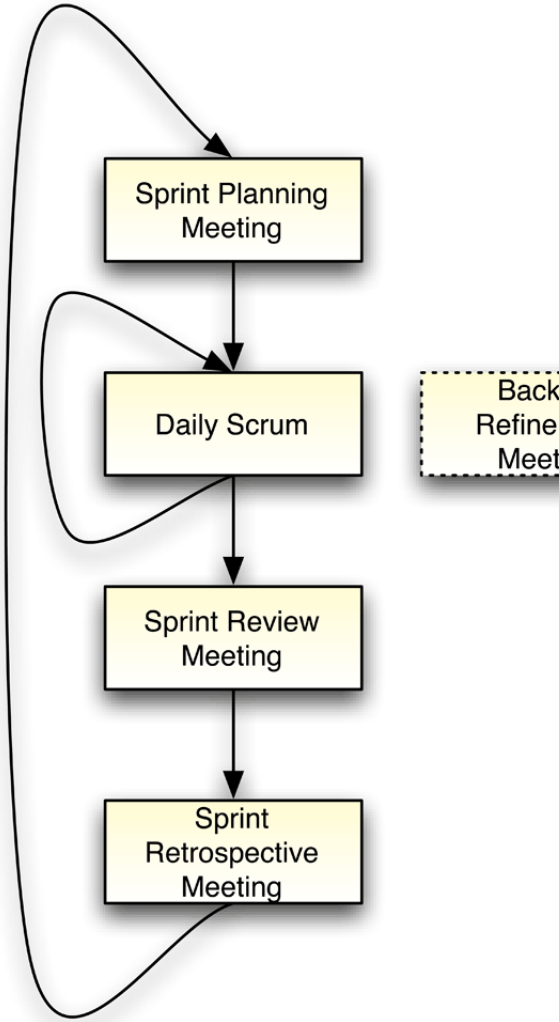
monthly sprint demo:

Demo of the stories delivered during the sprint

-Interaction with Business Owners:

release demo **Every 3 month** :

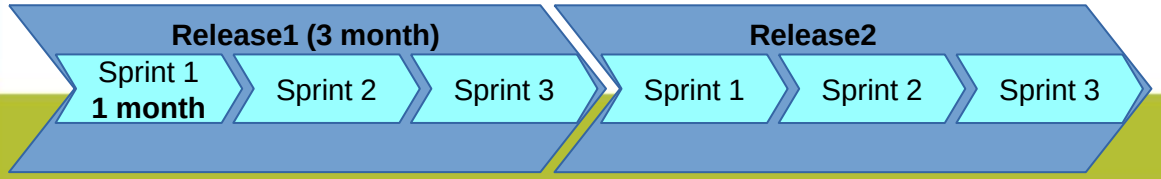
Demo the features that correspond to the stories delivered.



At the beginning of each sprint a **planning meeting** allows to select the stories ready in the backlog.

End of each sprint a **retrospective meeting** allows **continuous improvement** : identify and solve the stress in the team, adapt the method if needed.

Timeline:



GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Data management

gerome.jules-clement@gustaveroussy.fr

114, rue Édouard-Vaillant
94805 Villejuif Cedex - France
www.gustaveroussy.fr