

Un pipeline reproductible pour la réanalyse de données d'exome en génétique

Alexis Praga (M.D, Ph.D)

Oncobiologie Génétique Bioinformatique, CHU Besançon

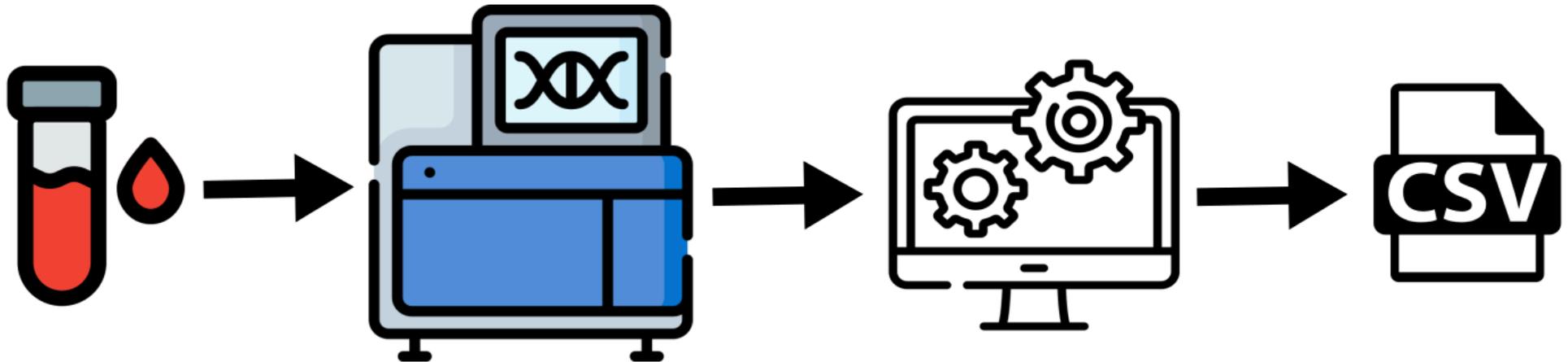
5 novembre 2024

Introduction

- Consultations de maladies rares (Centre de Génétique Humaine)

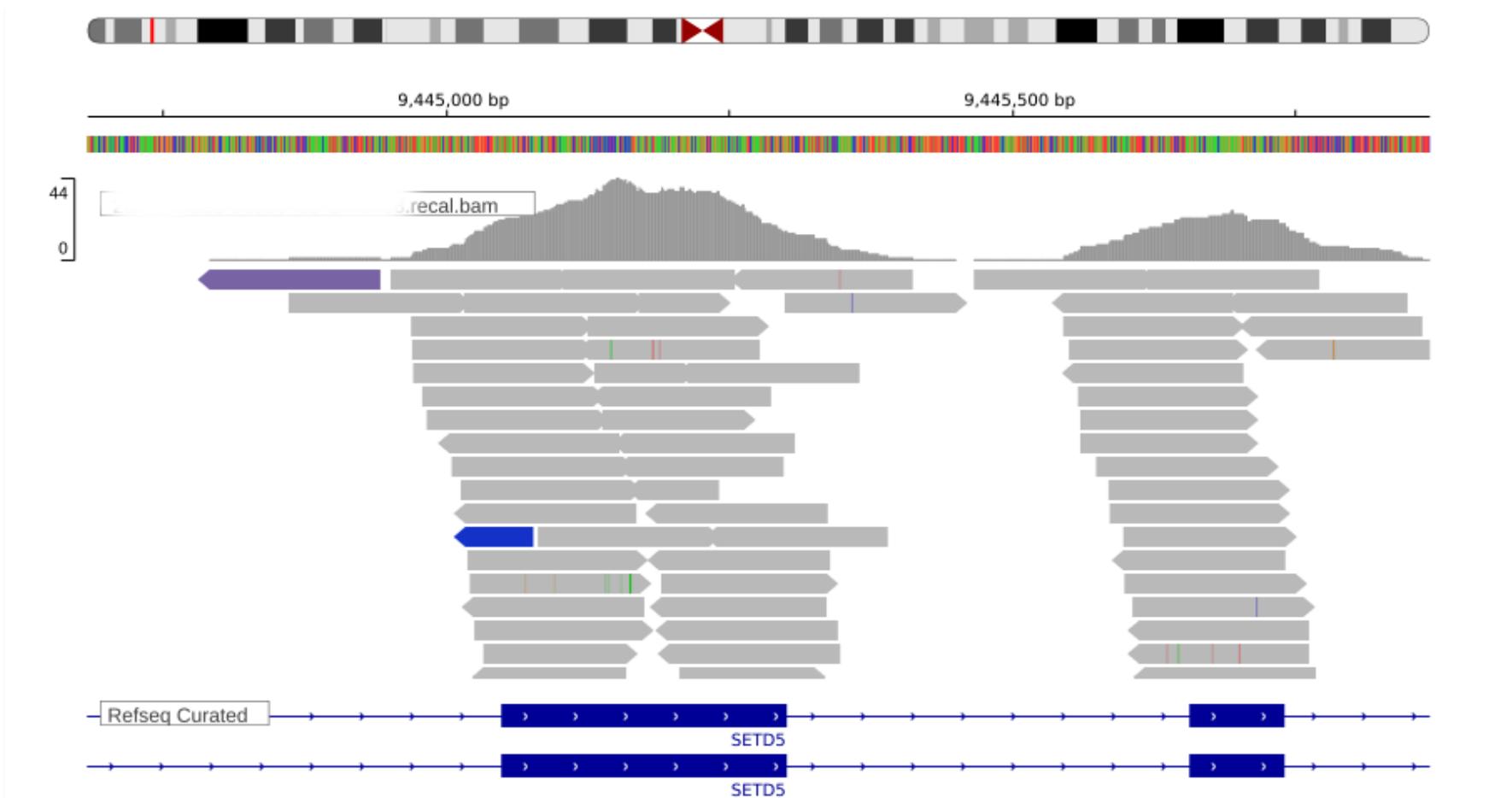
- Consultations de maladies rares (Centre de Génétique Humaine)
- Séquencage haut débit (NGS) en 2e ligne = *exome*
 - 1% de l'ADN
 - 32% diagnostics

- Consultations de maladies rares (Centre de Génétique Humaine)
- Séquencage haut débit (NGS) en 2e ligne = *exome*
 - 1% de l'ADN
 - 32% diagnostics
- Séquencage, analyse **et** interprétation sous-traités



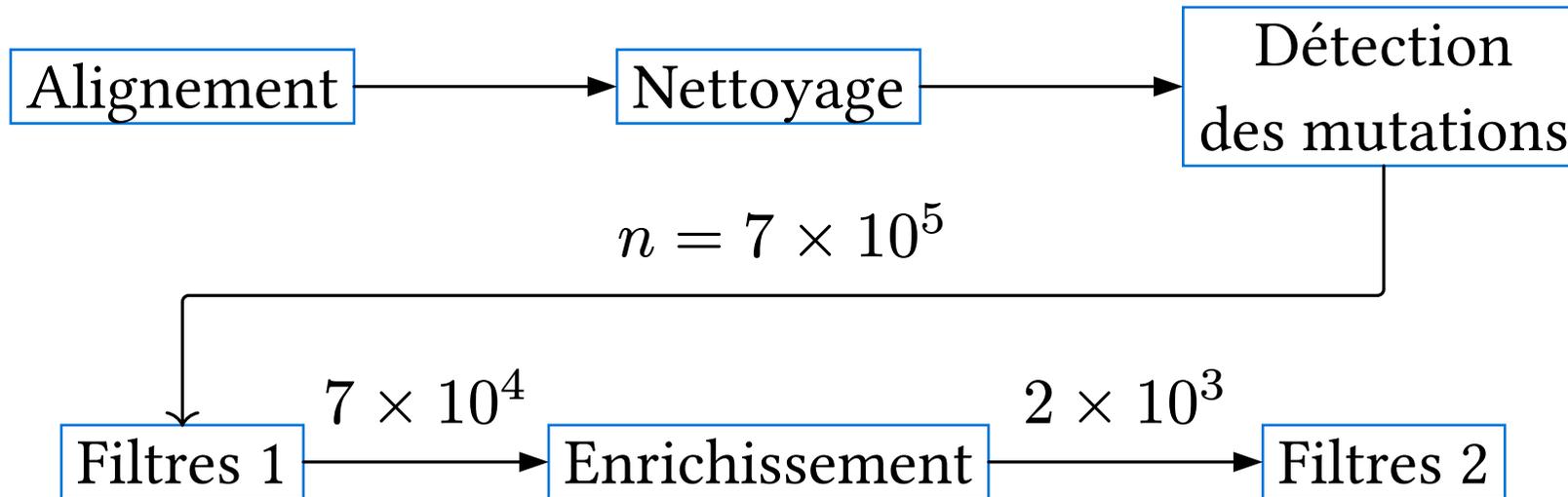
Et si l'on ne trouve pas de cause génétique ?

Alignement



Détection des mutations





Objectif

Patient

Ré-analyse données brutes de **patients sans diagnostic**

- *grep* pour nouveaux gènes
- *pipeline* maison : mutations manquées ?

Informatique

Améliorer la v0.1 en Bash (Dr. Alexis Overs)

1. reproductible, portable et parallélisé
2. testé et validé
3. ré-analyse de la cohorte bisontine



nextflow



Reproductibilité & Performance

Audit des laboratoires de biologie médicale



- ✓ Environnement logiciel reproductible
- ✓ Données externes reproductible
- ⌚ Environnement de production
- ⌚ Gestion des pannes
- ⌚ Intégration continue
- ⌚ Intégrité des données

Reproductibilité au *bit près*

- Compilation du code source
- Multiples versions en parallèle
- Sur mésocentre de Franche-Comté
- 100k paquets

Dans Nix :

✓ dépendances logicielles

✗ fichiers d'enrichissement (190Go)

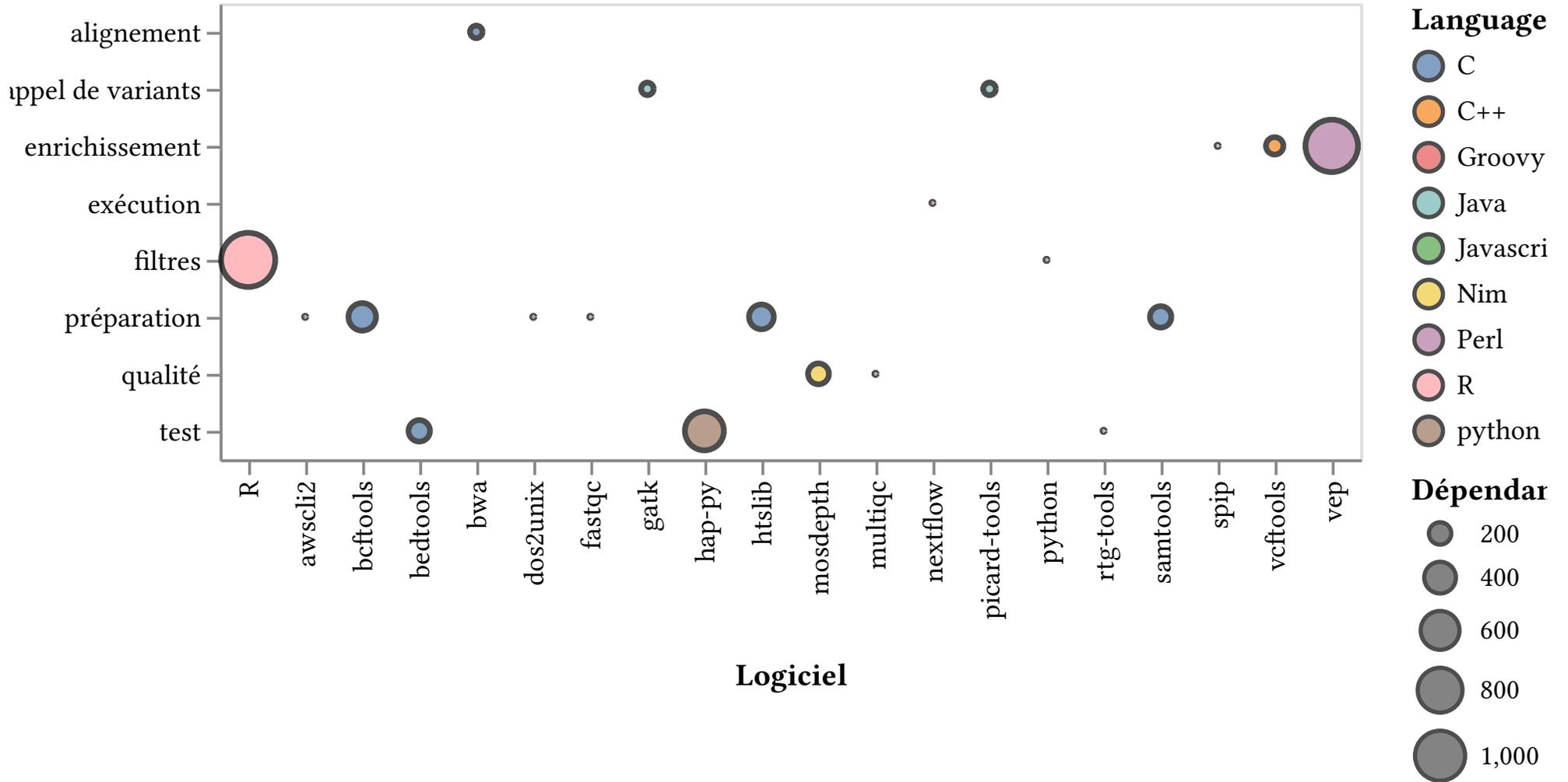
```
flake.nix
```

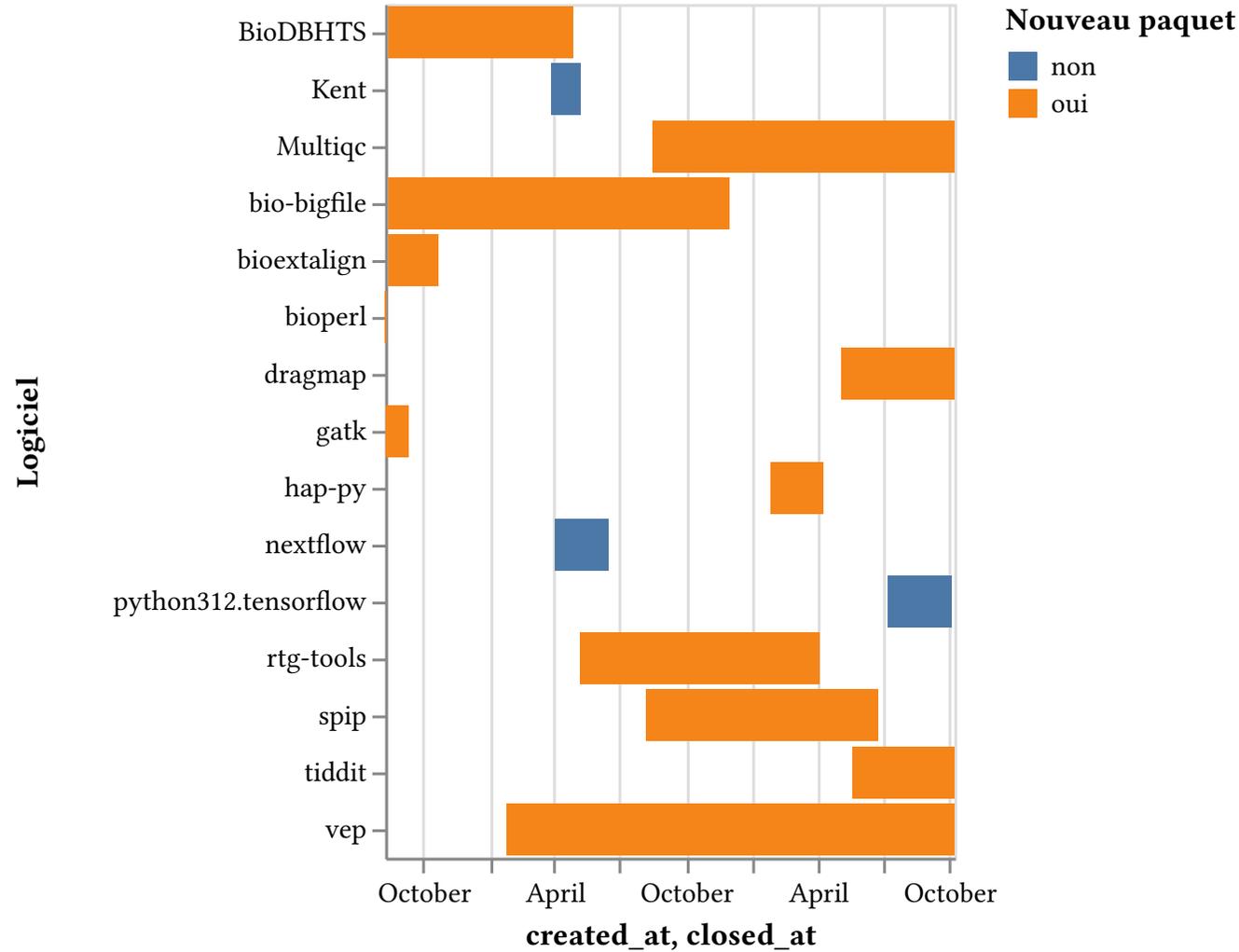
```
├─ biodbbigfile.nix  
├─ biodbhts.nix
```

```
...
```

```
└─ vep.nix
```

```
inputs = {  
  nixpkgs.url = "nixpkgs/  
nixos-23.05";  
};  
outputs = { self, nixpkgs }:  
  # ...  
  packages.${system} = {  
    gatk = pkgs.gatk;  
    hap-py = pkgs.callPackage pkgs/  
hap-py.nix {};  
  };
```





- graphe d'exécution
- nombreux *backends*

Schedulers



LSF

Platform Load Sharing
Facility



Cloud platforms



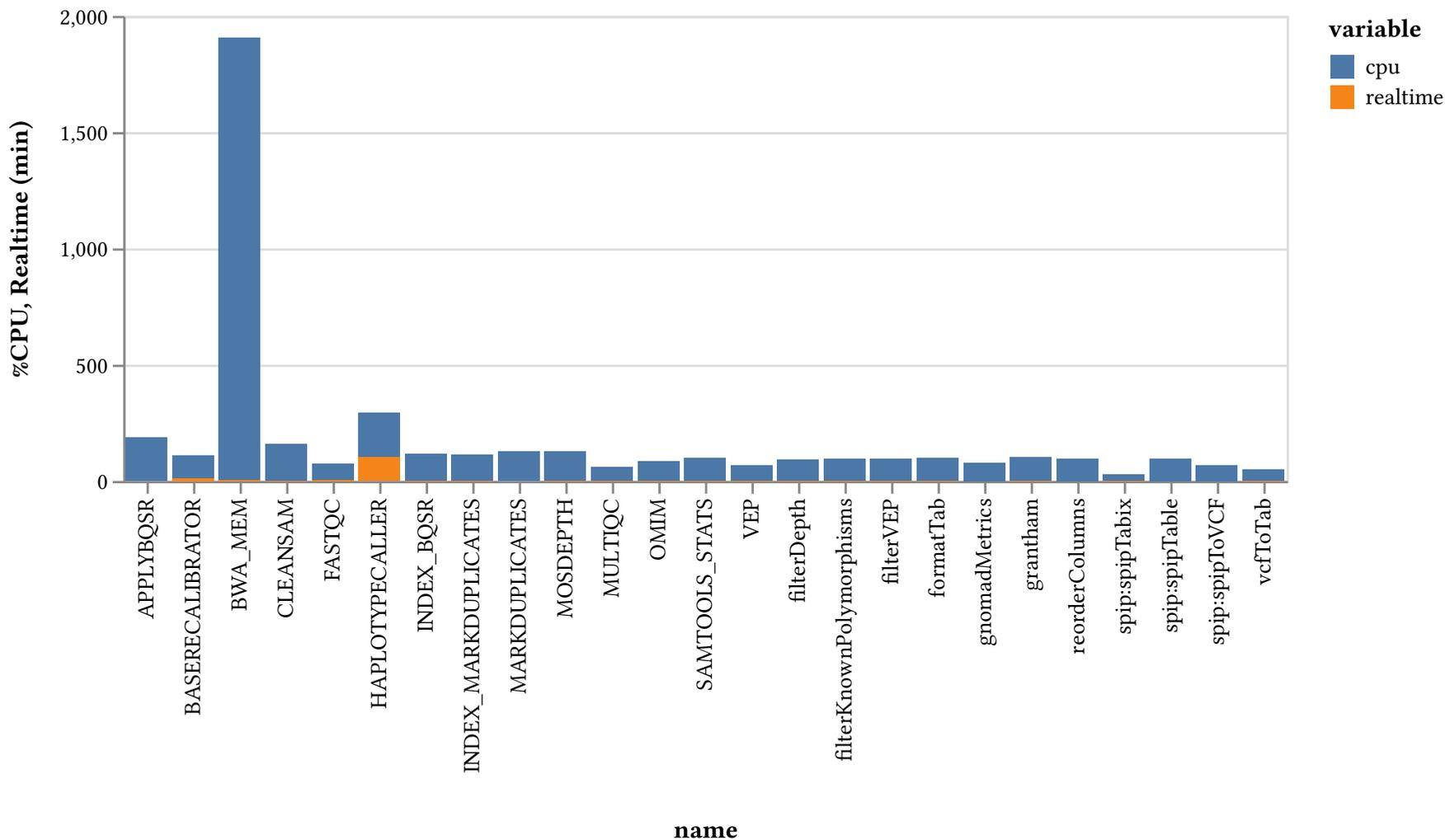
Un nouveau pipeline ?

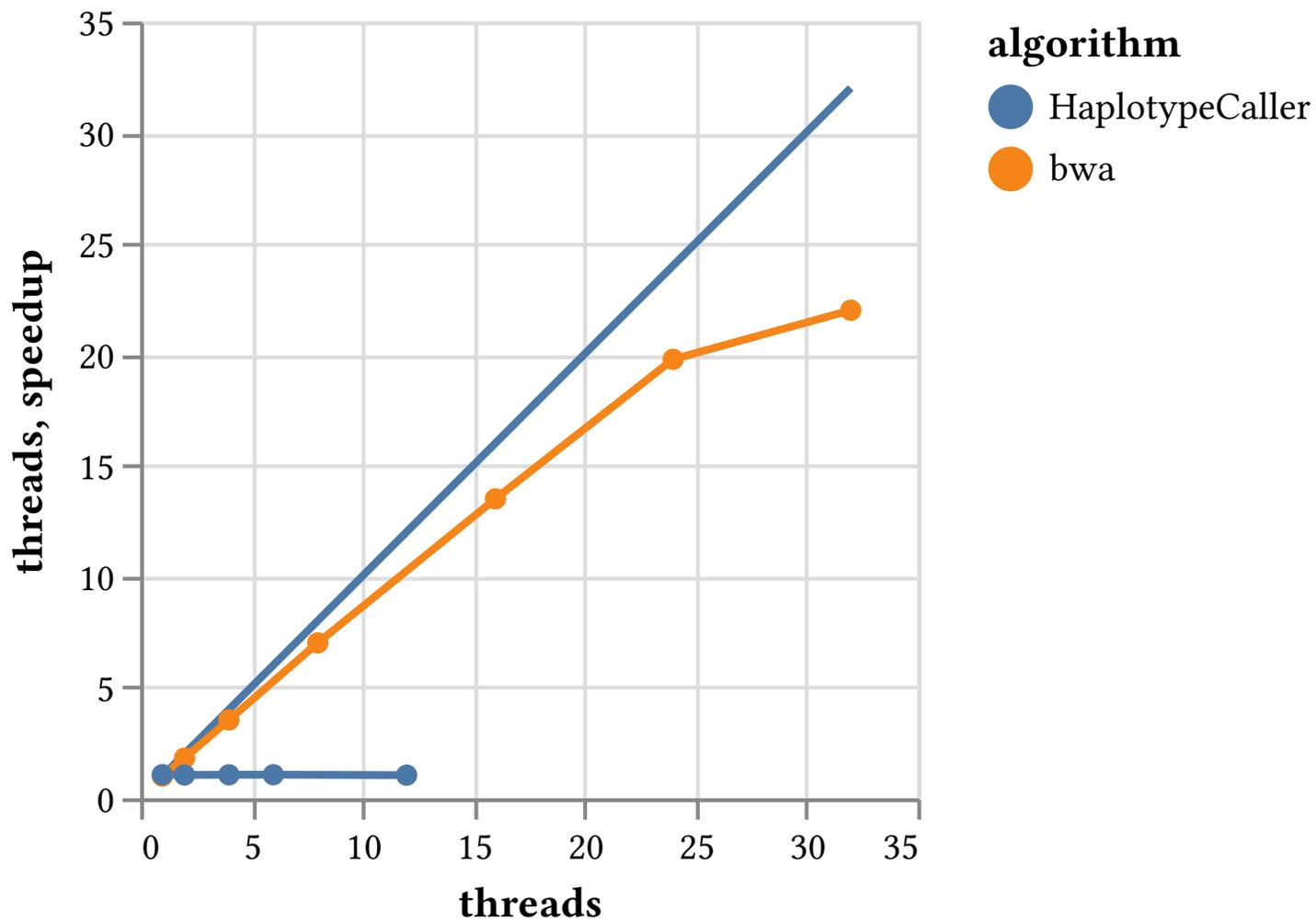
- **nf-core** 🍷 : sarek, rare-disease
- Bisonex :
 - téléchargement dépendances externes
 - adapté aux besoins du laboratoire

Un nouveau pipeline ?

- **nf-core** 🍷 : sarek, rare-disease
- Bisonex :
 - téléchargement dépendances externes
 - adapté aux besoins du laboratoire

```
workflow bisonex {  
  p = prepare()  
  preprocess(reads, ...)  
  callVariant(preprocess.out.bam, ...)  
  annotate(callVariant.out, ...)  
}
```





Validation

1. “Gold standard” : échantillons de référence (mutations connues)
2. Second temps : données simulées

Séquencage et pipeline

Patient NA12878

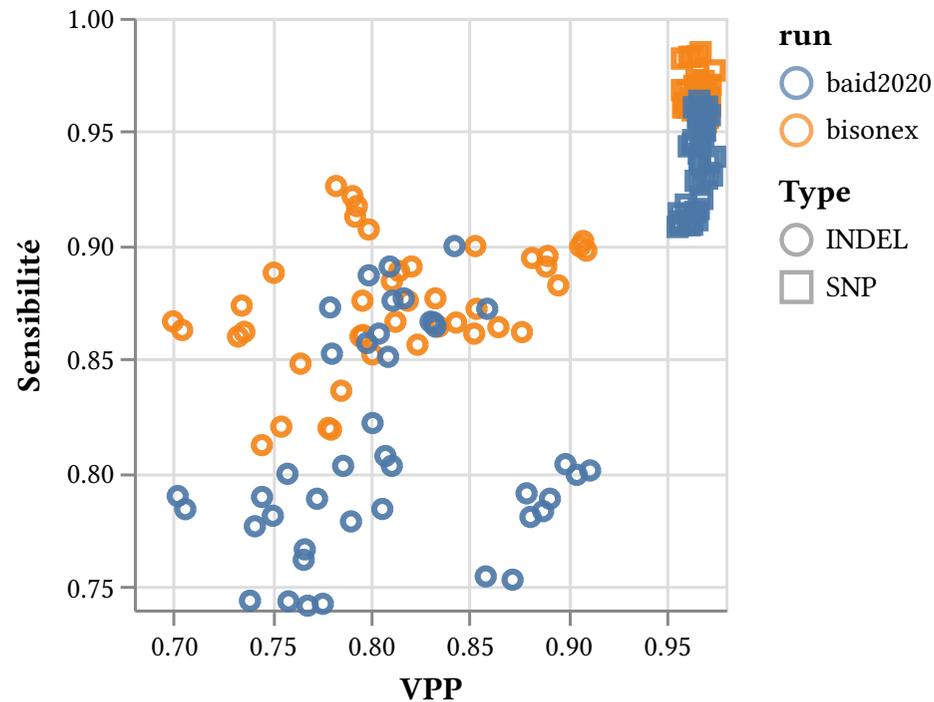
- ADN commandé à Coriell
- séquencé au laboratoire sous-traitant
- analysé par notre pipeline

Type	Sensibilité	Valeur Prédictive Positive
Indel	0.954	0.775
SNV	0.983	0.965

1. Patient de référence

Pipeline seul

Données brutes de Google sur 7 patients de référence



2. Données simulées

Base de test : 126 mutations confirmées

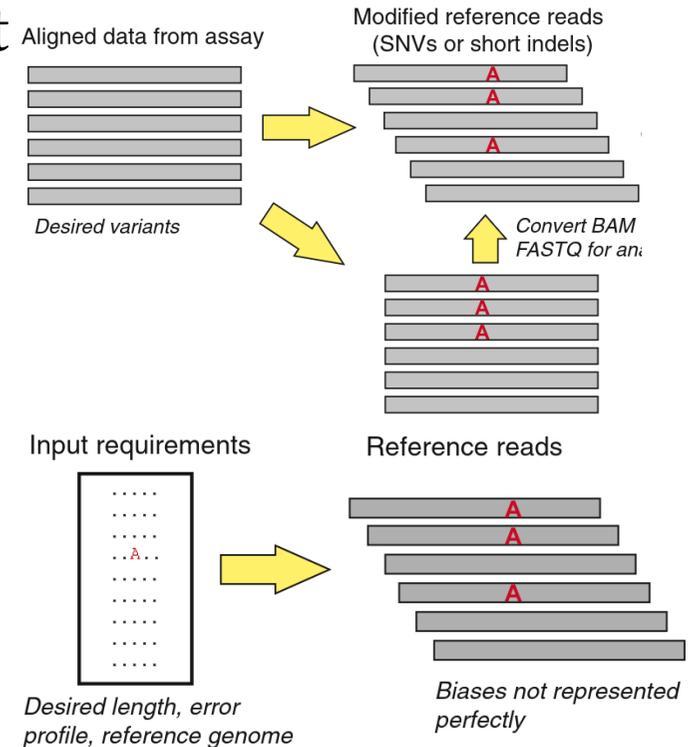
1. insérée dans données brutes d'un patient

124/126 retrouvées

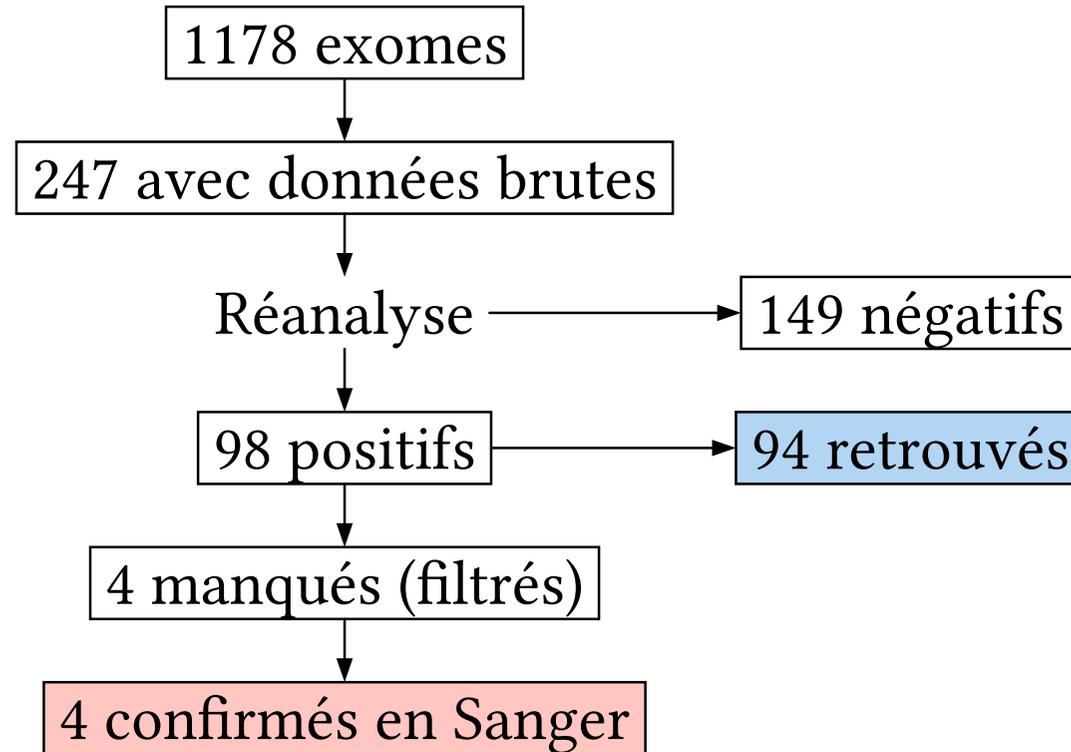
2. données simulées 100%

122/126 retrouvées

Manqués : erreur de méthode



Réinterprétation



- 4 manqués car filtrés : 3 sur la profondeur, 1 sur le nombre de reads porteurs

3 candidats très intéressants :

- trouble du neurodéveloppement : *ADNP*, *MED13*
- encéphalopathie épileptique : *KCNQ2*

Conclusion

Reproductibilité et validation : les 2 mamelles de la bioinformatique

- Nix : pour empaquetter autant de dépendances
- Nextflow : pour des chaînes de traitement bioinformatiques
- Résultat de validations encourageants
- Impact patient++

Futur :

- intégration nixpkgs en cours
- validation automatique (“CI”)
- utiliser les bases de tests récentes (2023)
- porter à d’autres analyses (génomome...) ?

Remerciements

- L'équipe du mésocentre (notamment Kamel Mazouzi)
- Laboratoire de génétique, notamment
 - Dr Alexis Overs (directeur de thèse)
 - Pr Paul Kuentz (encadrant)
 - Dr Éric Dahlen

Contact: alexis@praga.dev