

Université de Rouen  
Master I BioInforMatique - Année 2014 / 2015

# DÉVELOPPEMENT D'UN PROTOCOLE D'ANALYSE DE DONNÉES DNA-SEQ POUR LA RECHERCHE DE VARIATIONS DU NOMBRE DE COPIES (CNV)

Stéphanie Levon

**Plateforme Biopuces et Séquençage**  
Institut de Génétique et de Biologie Moléculaire et Cellulaire  
Encadrante : Stéphanie Le Gras

# PLAN DE LA SOUTENANCE

## 1<sup>ère</sup> PARTIE

### Introduction et objectifs du stage

- Définition d'un CNV
- Détection par séquençage haut débit

## 2<sup>ème</sup> PARTIE

### Acquisition des données de tests

- Simulées
- Réelles

## 3<sup>ème</sup> PARTIE

### Recherche bibliographique sur les outils de détection de CNVs

- Méthodes de détection
- Outils retenus

## 4<sup>ème</sup> PARTIE

### Conclusion et perspectives

# Les variations du nombre de copies (CNV)

- *Définition :*

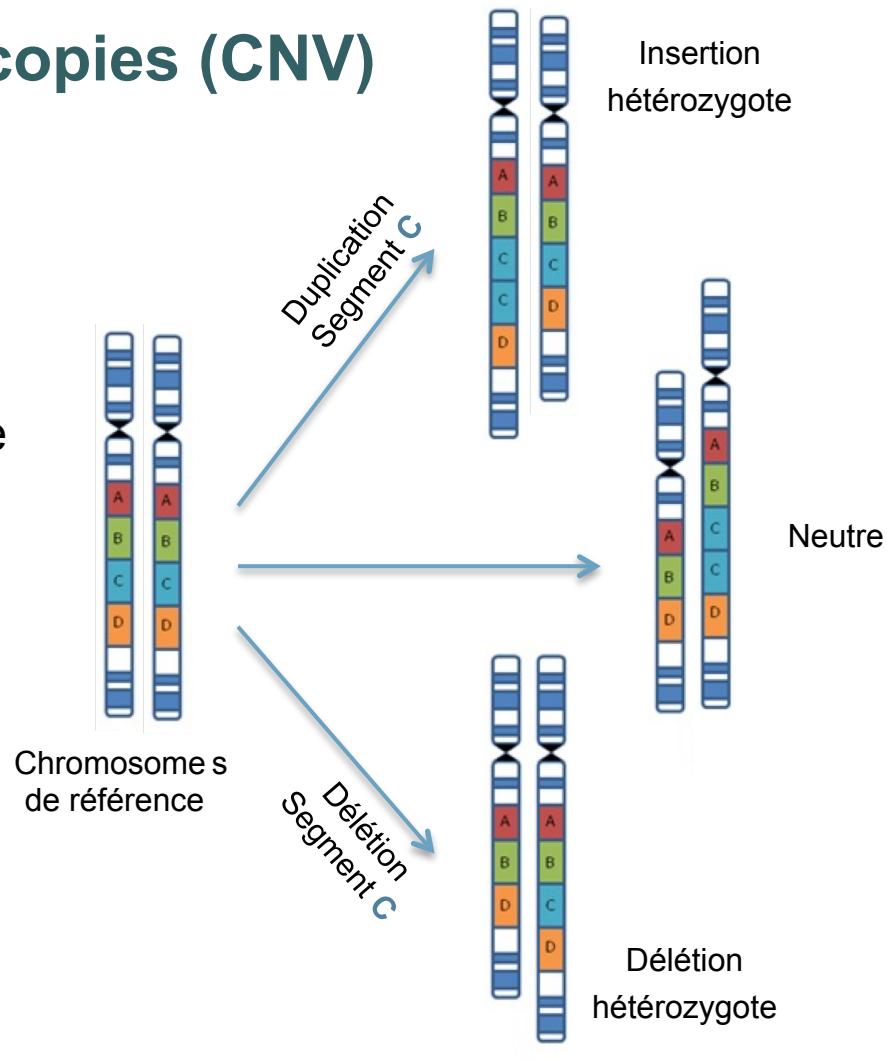
- Classe des variations structurales
- Variation du nombre de copies par rapport à un génome de référence
- À partir de 20 paires de base et jusqu'à plusieurs mégabases

- *Impact biologique :*

- Modifie l'expression d'un gène
- Polymorphisme

- *Détection :*

- FISH
- Puce CGH
- Séquençage haut débit



Source : modifié de <http://neurowiki2013.wikidot.com/individual:copy-number-variations>.

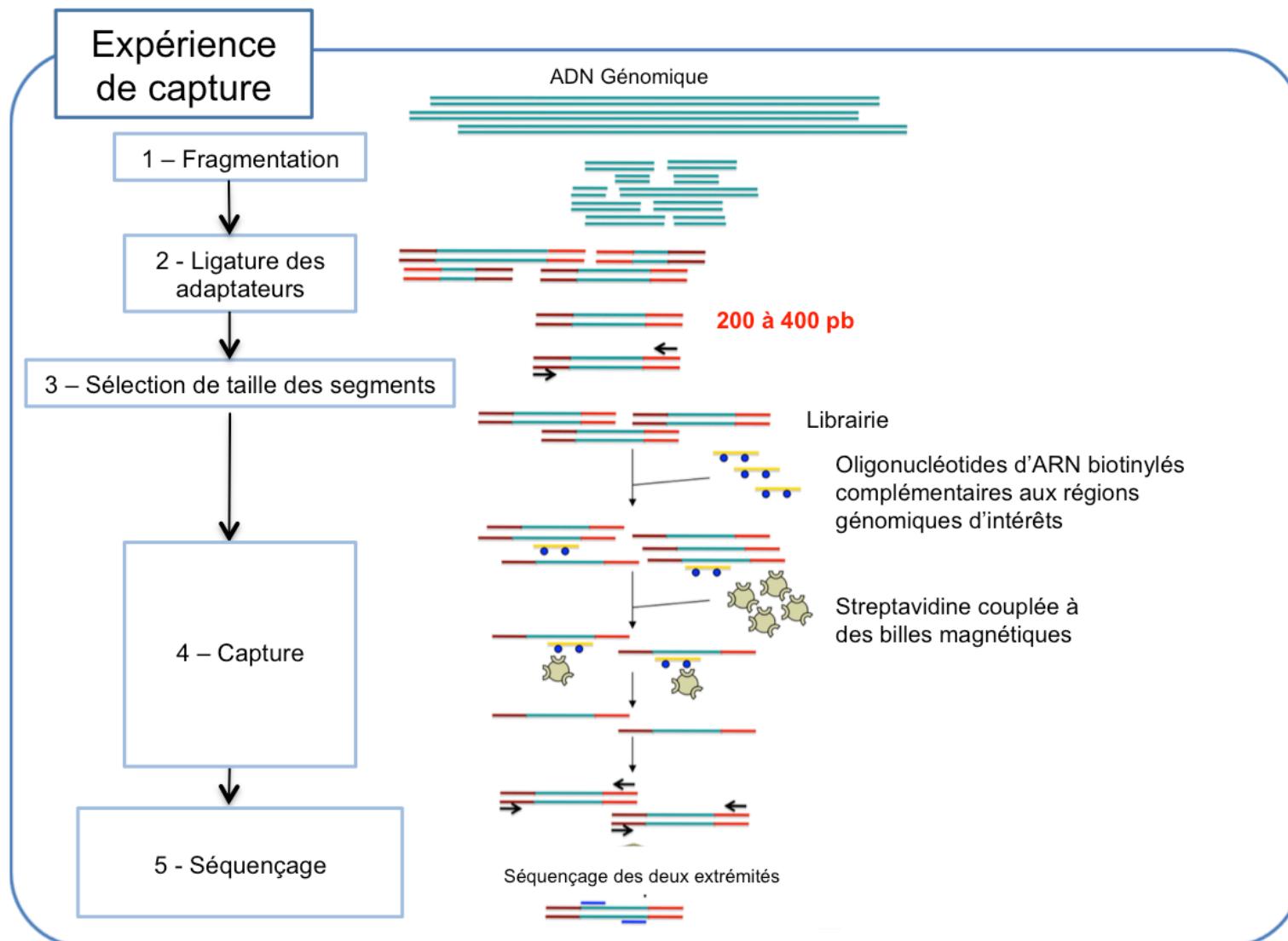
# Détection de CNVs grâce au séquençage haut débit

*Type de séquençage :*

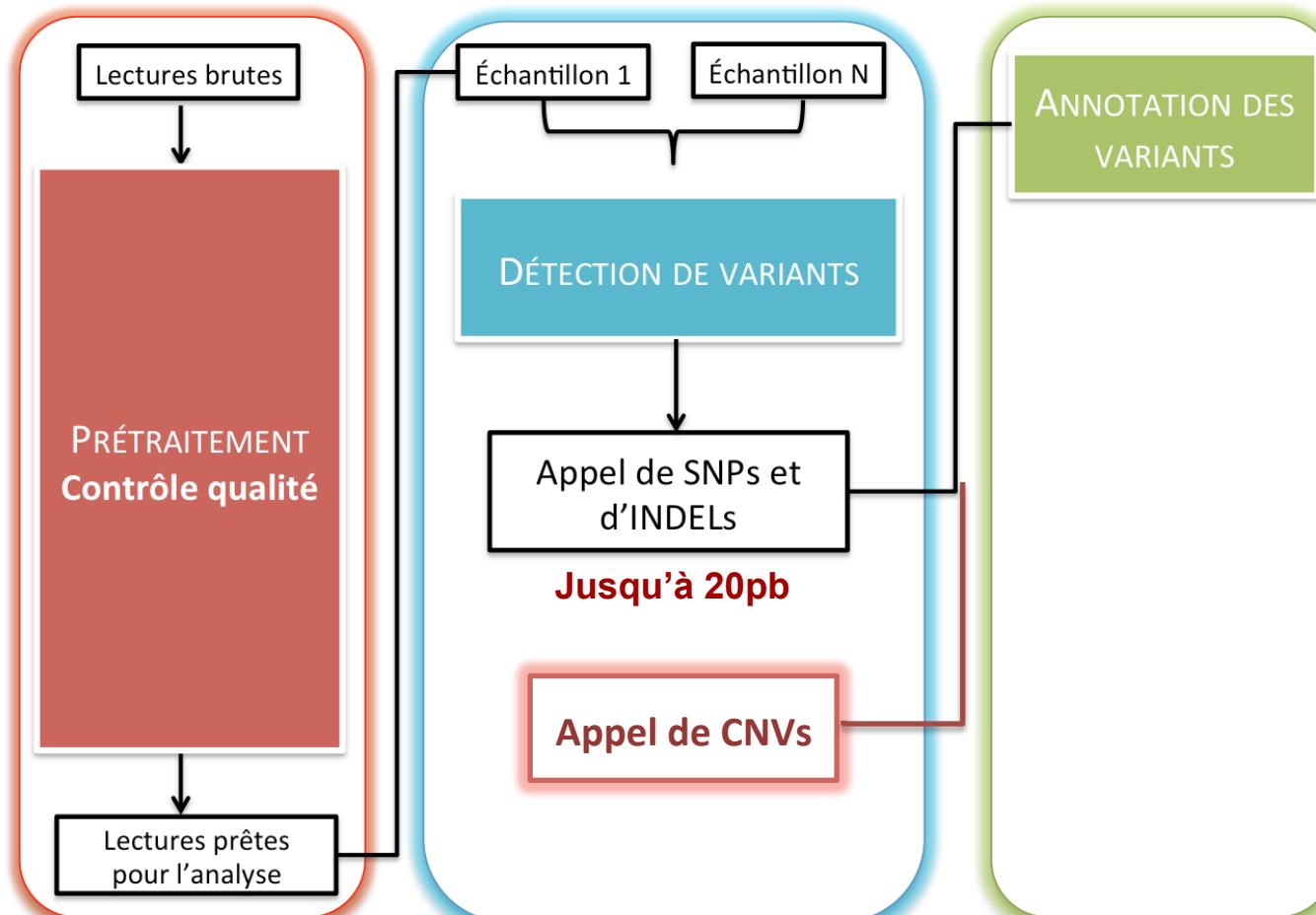
- DNA-seq et reséquençage de régions d'intérêt (**capture**)
- Exons – séquençage de l'exome
- Liste de gènes d'intérêts

➡ Préparation de librairies

# Détection de CNVs grâce au séquençage haut débit



# Protocole d'analyse DNA-seq



LATEX

## Objectifs du stage

**Implémenter un protocole de détection de CNVs  
au sein du protocole d'analyse DNA-seq**

- Familiarisation avec le protocole d'analyse de données DNA-seq
- Acquisition de données DNA-seq de test, simulées et réelles
- Recherche bibliographique des méthodes et des outils de détection de CNVs

# Obtention de données de test DNA-seq

→ Différents types de données pour sélectionner l'outil le plus performant

- Données simulées

Accès à tous les paramètres statistiques

		Vérité		Paramètres statistique	
		Présence de CNV			
Résultat de l'outil	CNV détecté	Vrais positifs (VP)	Faux positifs (FP)		
	CNV non détecté	Faux négatifs (FN)	Vrais Négatifs (VN)		
	Paramètres statistique				
		<ul style="list-style-type: none"> <li>■ Valeur prédictive positive (VPP)</li> </ul>		$VPP = \frac{VP}{VP + FP}$	
		<ul style="list-style-type: none"> <li>■ Valeur prédictive négative (VPN)</li> </ul>		$VPN = \frac{VN}{VN + FN}$	
		<ul style="list-style-type: none"> <li>■ Sensibilité</li> </ul>	<ul style="list-style-type: none"> <li>■ Spécificité</li> </ul>		
		$\frac{VP}{VP + FN}$	$\frac{VN}{VN + FP}$		

# Obtention de données de test DNA-seq

→ Différents types de données pour sélectionner l'outil le plus performant

- **Données simulées**

**Accès à tous les paramètres statistiques**

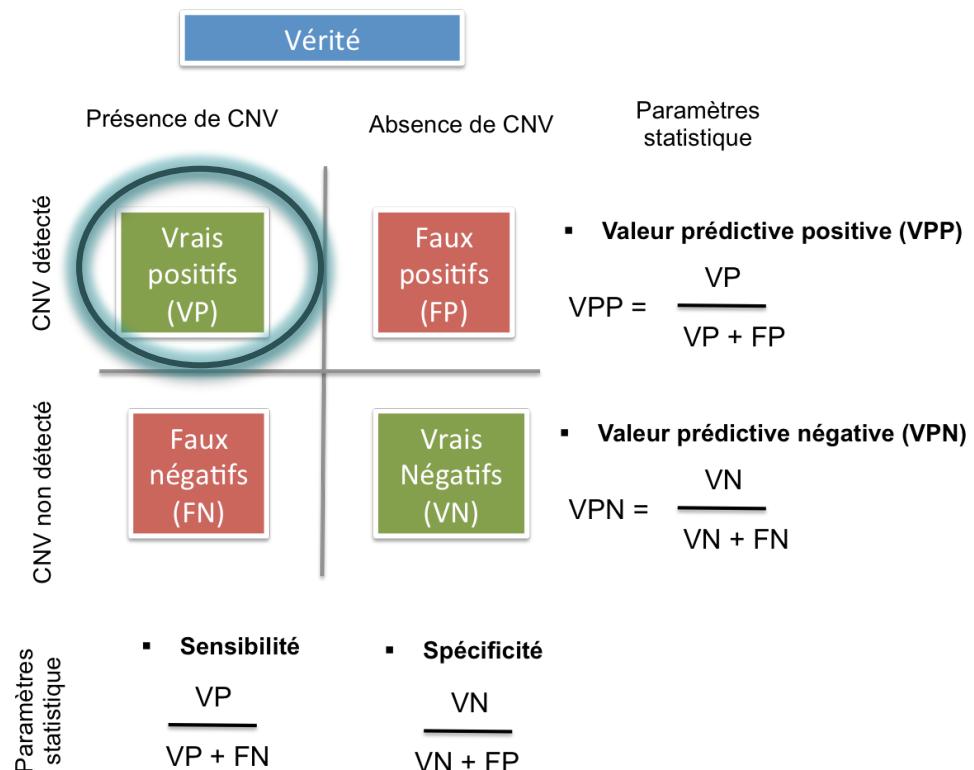
- **Données réelles**

1 - Séquencées au sein de la plateforme

Patients atteints de retards mentaux  
ou du syndrome de Bardet-Biedl  
(Redin et al., 2012)

**Accès aux VP**

Résultat de l'outil



# Obtention de données de test DNA-seq

→ Différents types de données pour sélectionner l'outil le plus performant

- **Données simulées**

**Accès à tous les paramètres statistiques**

- **Données réelles**

1 - Séquencées au sein de la plateforme

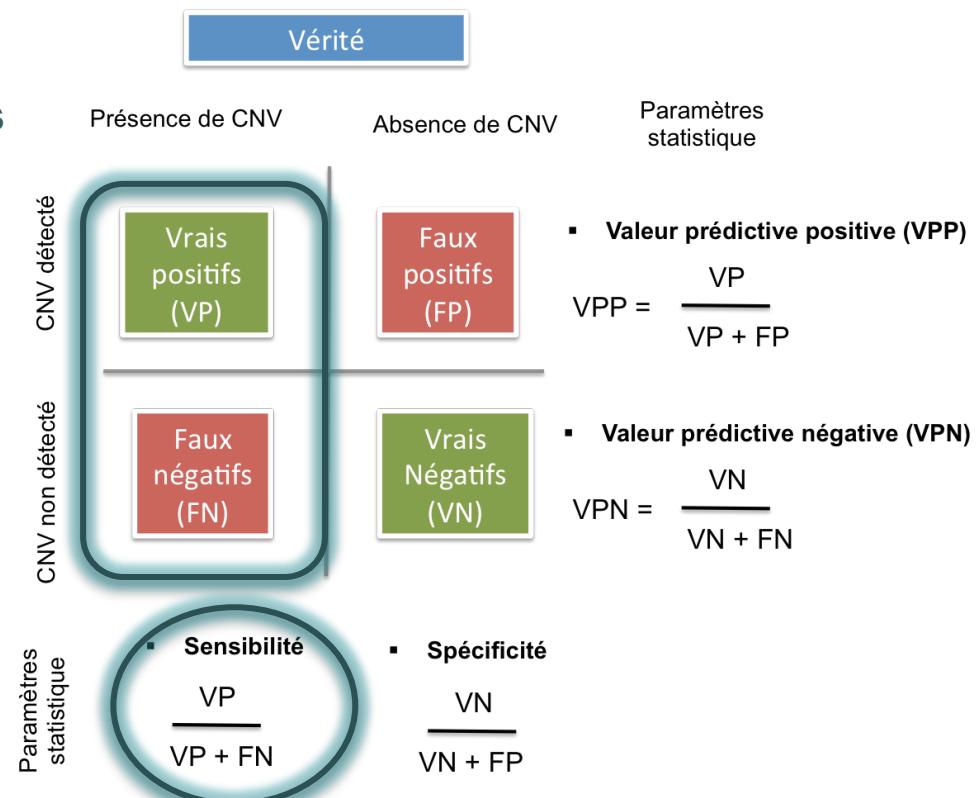
Patients atteints de retards mentaux  
ou du syndrome de Bardet-Biedl  
(Redin et al., 2012)

**Accès aux VP**

2 - Issues du projet 1000 Genomes

**Accès aux VP & aux FN**

Résultat de l'outil



# Recherche d'un simulateur de données

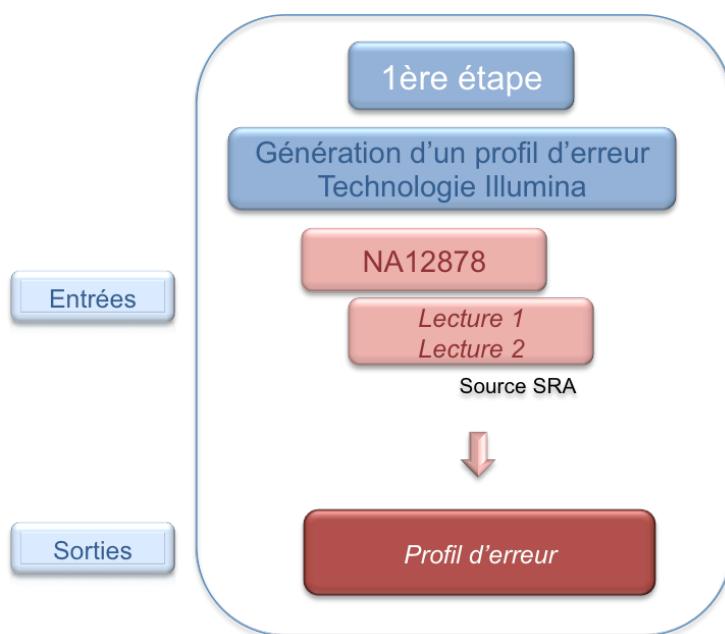
## Sources :

PubMed – OMIC Tools - SEQanswers

Nom	Simulation de lectures avec profil d'erreur caractéristique de la technologie Illumina	Simulation de variants de type CNVs	Particularités
<b>ART</b> Huang et al., 2006	✓		
<b>RSVSim</b> Bartenhagen et Dugas, 2013	✓	✓	Variations structurales globales
<b>SCNVSim</b> Quin et al., 2015	✓	✓	Echantillons tumoraux
<b>SInC</b> Pattnaik et al., 2014	✓	✓	
<b>Wessim</b> Kim et al., 2013	✓		

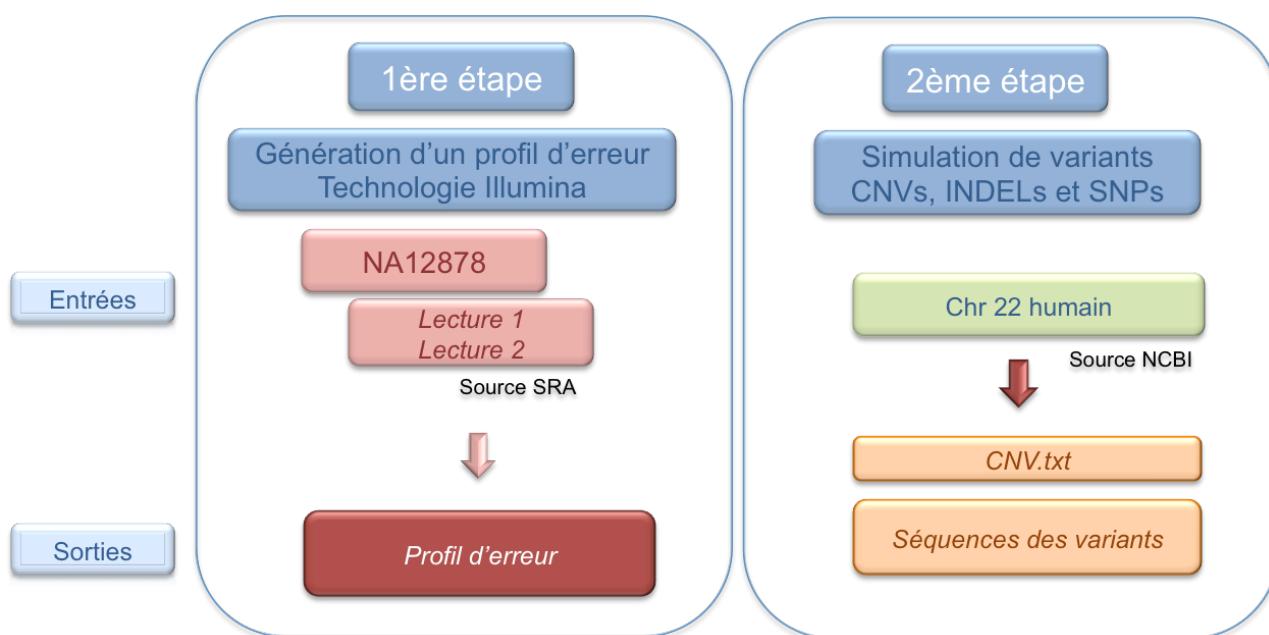
**SInC : SNP, INDEL et CNV**

# Fonctionnement de l'outil de simulation SInC



**SInC : SNP, INDEL et CNV**

# Fonctionnement de l'outil de simulation SInC

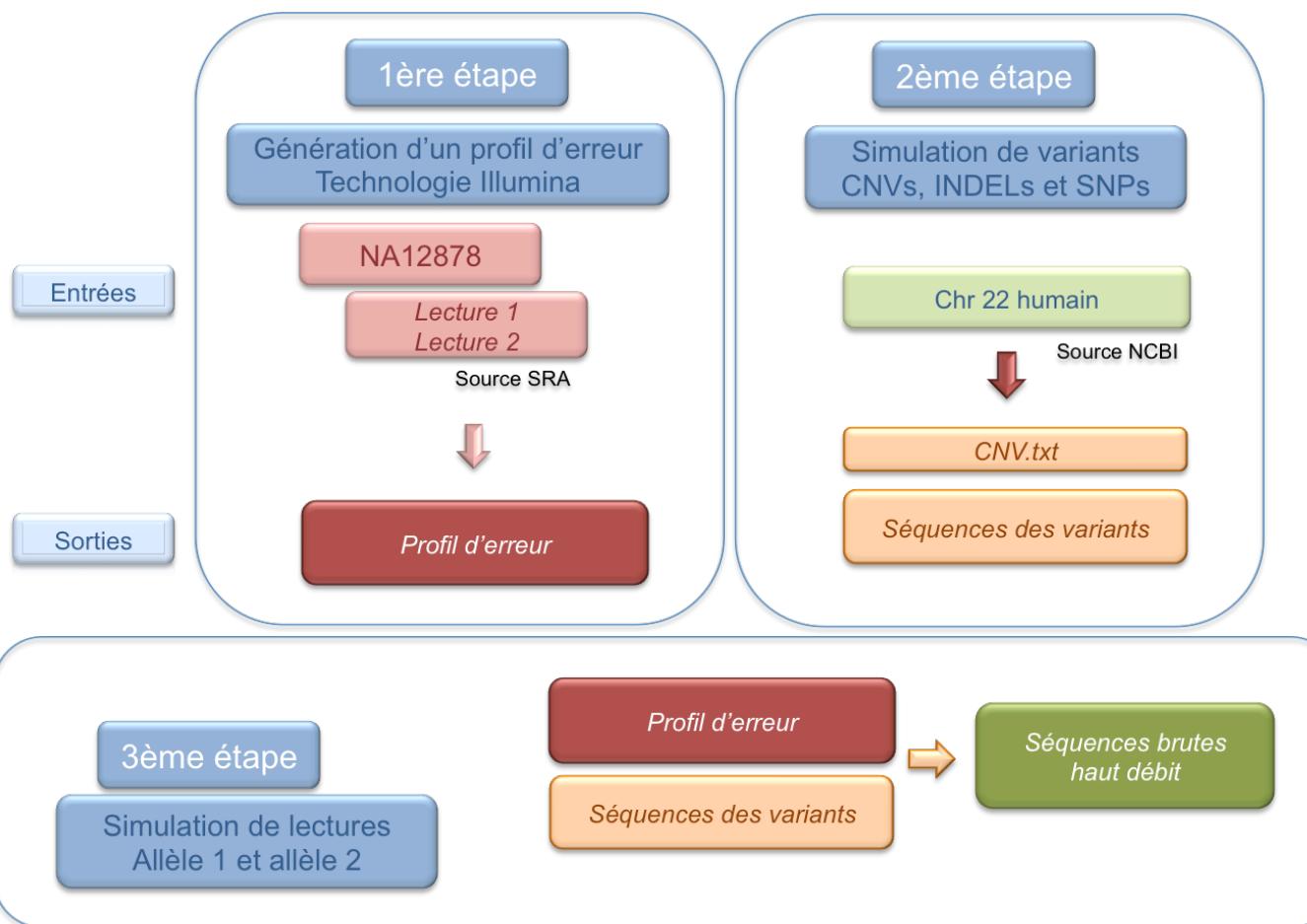


## Paramètres modifiés

- **Pourcentage du génome altéré par les CNVs**  
0% et 2%
- **Taille minimale et maximale**  
20pb à 6 000pb

**SInC : SNP, INDEL et CNV**

# Fonctionnement de l'outil de simulation SInC

Paramètres modifiés

- Couverture**  
60X – capture d'exomes
- Taille des lectures**  
100pb

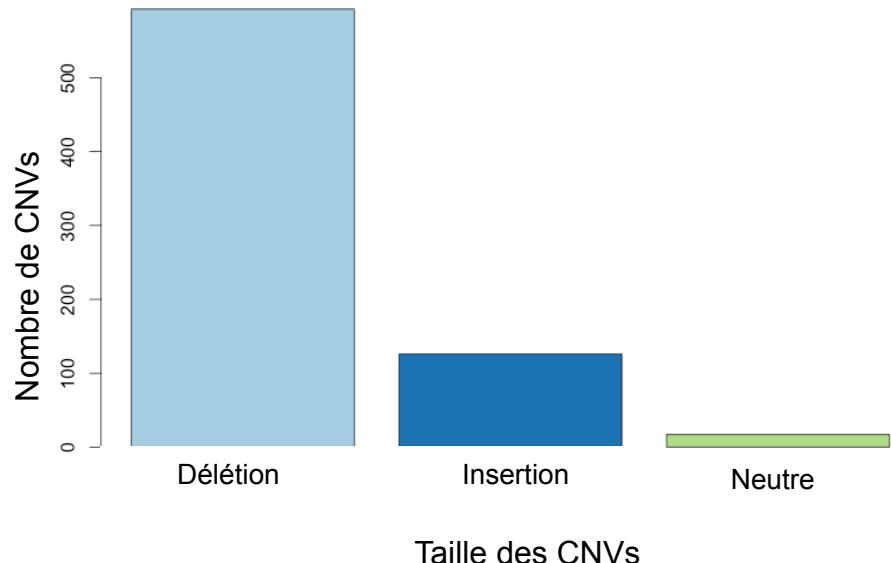
**Deux jeux de données simulées : 0% et 2% de CNVs**

## Génération des données simulées

- ✓ Nombre de lectures générées
  - 0% de CNVs
  - 2% de CNVs } ~40 millions de lectures / jeu
- ✓ Qualité des données (FASTQC, Andrews, 2010)
  - qualité des données simulées similaire à celle des données sources utilisées pour l'établissement du profil d'erreur
- ✓ Données d'alignement – BWA (Li et al., 2010)
  - 0% de CNVs
  - 2% de CNVs } > 99,99 % de lectures alignées
- ✓ Statistiques descriptives des CNVs

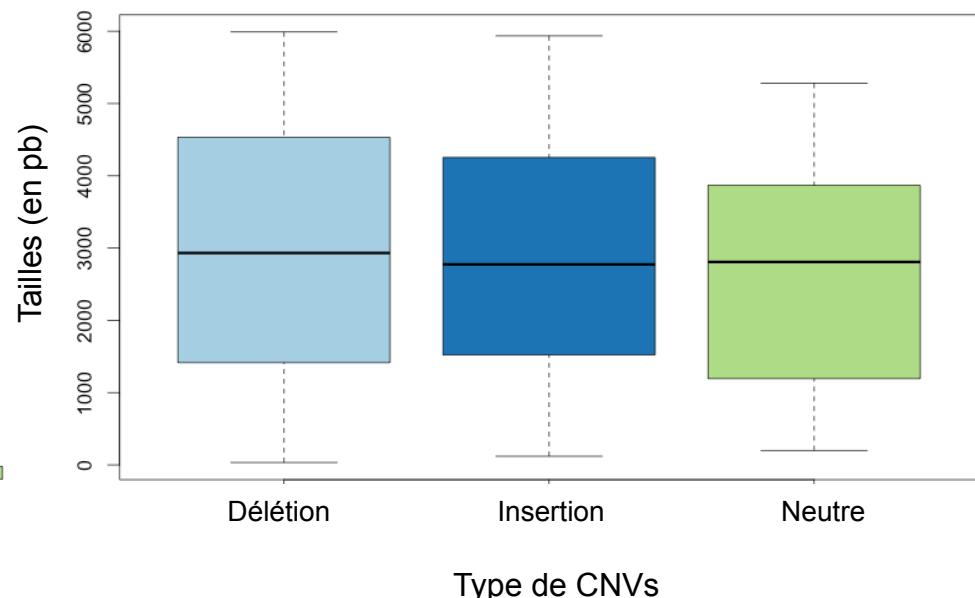
# Description des données simulées

Répartition des types de CNVs



Déséquilibre délétion / insertion  
Distribution similaire au génome humain

Distribution des tailles en fonction  
des types de CNVs



Couverture des tailles fidèle aux  
paramètres choisis

→ Données tests pour les outils de détection de CNVs

# Quelles sont les méthodes de détection de CNVs ?

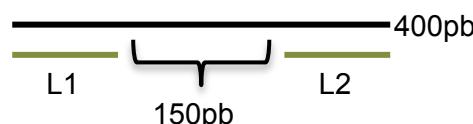
4 méthodes de détection  
de CNVs :

- Lectures chimériques
  - Assemblage
  - Distance entre lectures pairees
  - Profondeur de séquençage
- ]} Méthodes les plus utilisées

# Quelles sont les méthodes de détection de CNVs ?

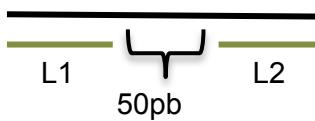
## Distance entre lectures pairees

Fragment d'ADN de référence



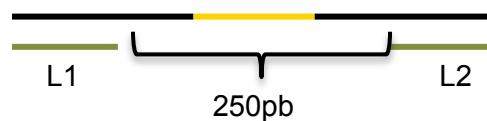
Délétion

Individu 1



Insertion

Individu 2



## Profondeur de séquençage



Duplication

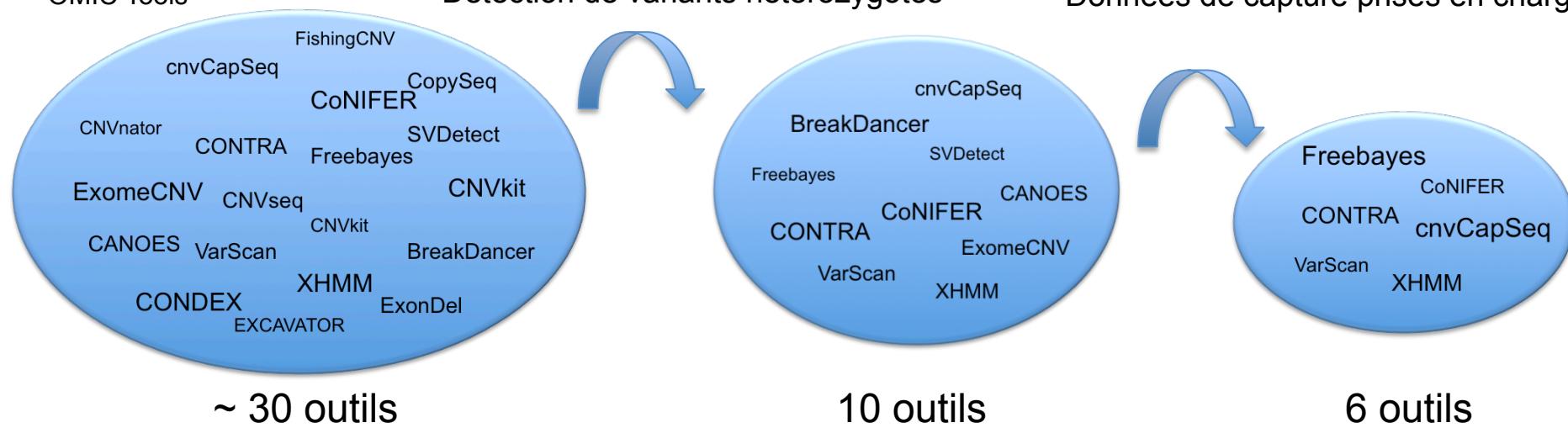
Délétion

# Critères de sélection des outils à tester

**Sources :**  
 PubMed  
 SEQanswer  
 OMIC Tools

Disponibilité  
 Lectures paireées Illumina  
 Détection de variants hétérozygotes

Nombre de citations  
 IF du journal de publication  
 Date de dernière mise à jour  
 Données de capture prises en charge



## Caractéristiques des outils choisis

Nom	Nombre de citation	IF du journal de publication	Méthodologie <b>Profondeur de séquençage</b>	Particularité de l'outil
<b>CONTRA</b> Li et al., 2012	35	5,3	✓	
<b>cnvCapSeq</b> Bellos et al., 2014	0	8,8	✓	Combinaison avec distance entre les séquences pairées
<b>CoNIFER</b> Krumm et al., 2012	61	13,8	✓	Combinaison avec méthode de décomposition en valeur singulière
<b>XHMM</b> Fromer et al., 2012	38	11,2	✓	Combinaison avec modèle de Markov caché
<b>Freebayes</b> Garrison et Marth, 2012	80	/		Méthode statistique Bayésienne
<b>VarScan</b> Koboldt et al., 2009	226	4,9		Méthode de réalignement



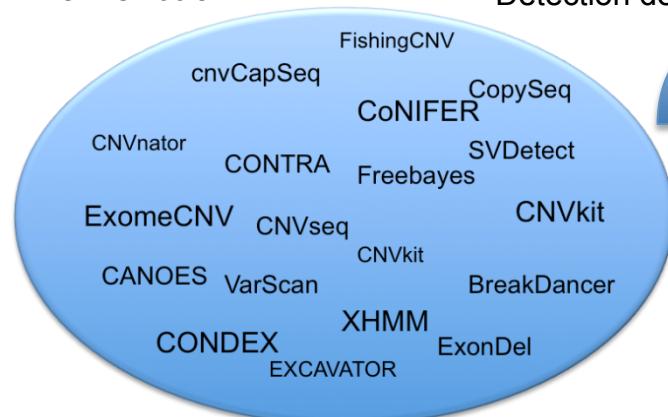
Installés sur le serveur et prêts à être testés

# Critères de sélection des outils à tester

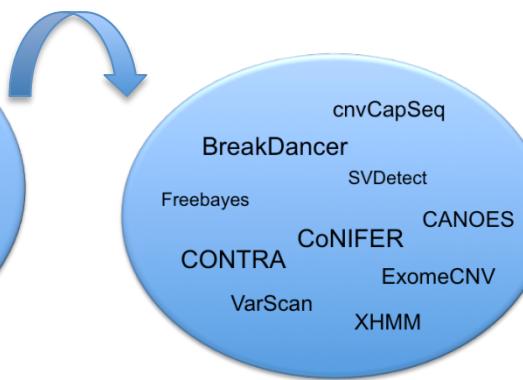
**Sources :**  
 PubMed  
 SEQanswer  
 OMIC Tools

Disponibilité  
 Lectures paireées Illumina  
 Détection de variants hétérozygotes

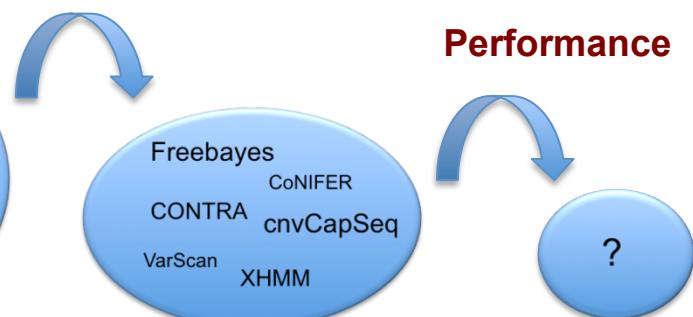
Nombre de citations  
 IF du journal de publication  
 Date de dernière mise à jour  
 Données de capture prises en charge



~ 30 outils



10 outils



6 outils



1 outil



Tests de performance des outils

# Conclusion

## *Jeux de données tests*

- ✓ Jeux de données réelles à disposition pour le test des outils
- ✓ Simulation de données grâce à l'outil SInC

## *Outils de détection de CNVs*

- ✓ Six outils sélectionnés : cnvCapSeq, CoNIFER, CONTRA, Freebayes, VarScan, XHMM.

## *Familiarisation avec le protocole de développement*

- ✓ Réalisation d'un script d'entraînement au développement de protocole d'analyse (bash, Perl, Git)

# Perspectives

✧ Test de comparaison des outils

Évaluation des critères de performance : sensibilité, spécificité

✧ Implémentation dans le protocole d'analyse DNA-seq

# Merci de votre attention

Merci à

Céline Keime

Christelle Thibault-Carpentier

Stéphanie Le Gras

Serge Uge

Matthieu Jung

Arjun Rao

Constance Vagne

Tao Ye

Amandine Velt

...ainsi qu'à tous les membres de la plateforme Biopuces et Séquençage

