Soutenance de stage

Création d'un pipeline d'analyse de variants génomiques dans le cadre d'une expérience de mutagenèse aléatoire.

Romuald Marin

Laboratoire d'accueil

Recherche et Développement des Plantes - RDP

Tutelle : ENS, CNRS, INRA ...

Domaine de recherche :

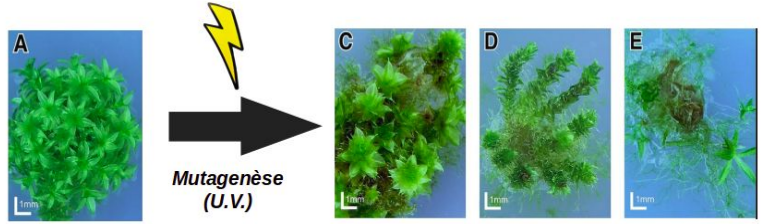
- Développement des plantes
- Evolution des structures reproductrices

Equipe **Signal** (Signalisation hormonale et développement)

Maitre de stage : **Fabrice Besnard**

Données Biologiques

Expérience de mutagenèse aléatoire



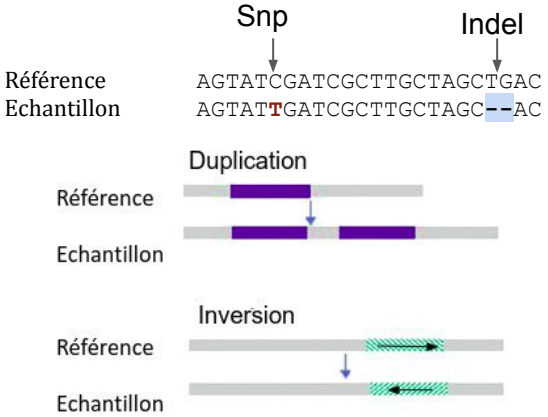
Plant normal

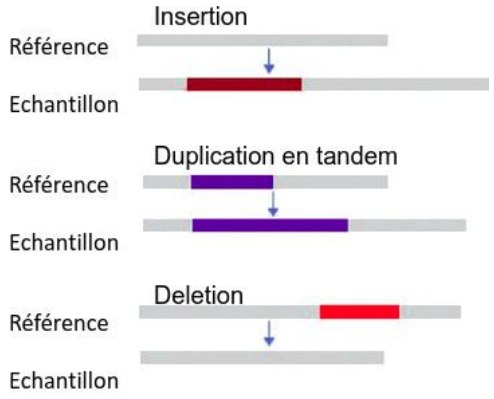
Plants mutés avec phénotype intéressant

Identifier les **mutations** génétiques causant ces phénotypes

↝ **Lien entre gène et phénotype**

Mutations génétiques





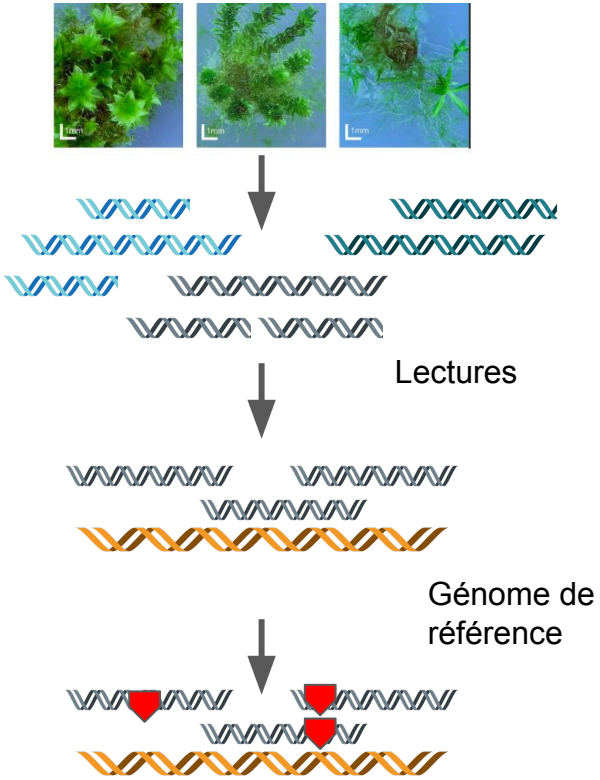
Analyse bio-informatique

Les grandes étapes

Séquencer l'ADN issu des plants mutés

Aligner les lectures des différents échantillons sur le génome de référence

Repérer les variations par rapport au génome de référence : "**Variant calling**"



Cahier des charges

Problème :

- Procédure : 8 scripts différents
- Nombreux types de fichiers
- Quelques outils disponibles mais version non à jour
- Pas de variants structuraux
- Un seul design expérimental

Solution :

- Fichiers d'entrée et de sortie simples
- Amélioration et mise à jour des outils
- Pipeline automatisé

Avantage du pipeline Nextflow

Coté développeur :

- Créé pour la bioinformatique
- Différents environnements et langages de script
- Système de cache
- Gestion des containers docker

Coté utilisateur :

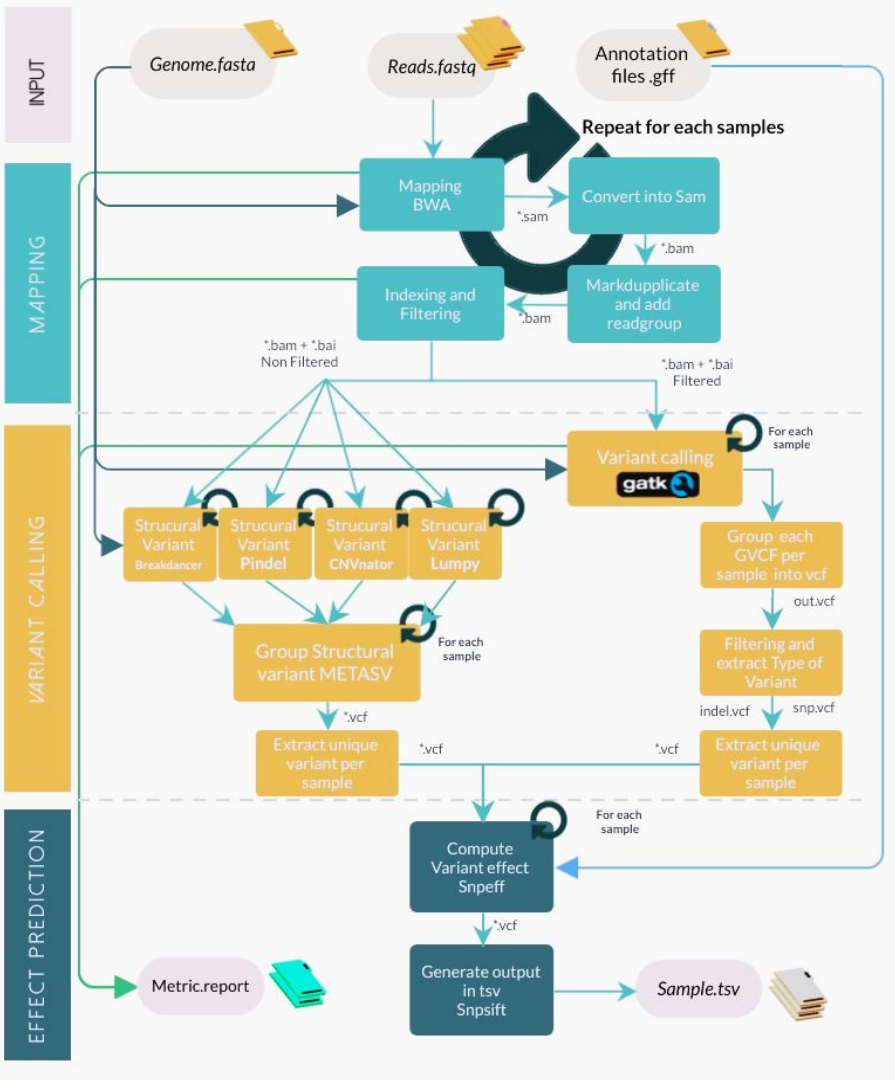
- Peu de connaissances en informatique nécessaires
- Une ligne de commande
- Entrée et sortie simples
- Suivi de l'analyse

Fonctionnement du Pipeline

3 fichiers d'entrée :
Génome de référence
Lecture
Fichier d'annotation

2 fichiers de sortie :
Rapport .HTML
Variants en .tsv

3 étapes :
Alignement des lectures
Détection des variants
Prédiction des effets

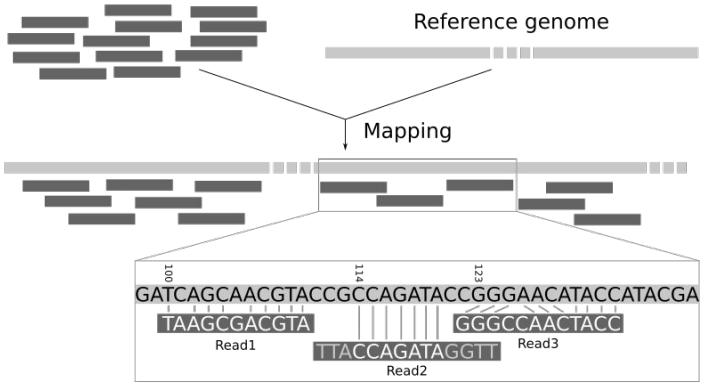


Etape 1 : Alignement des lectures

But de l'alignement

Aligner les lectures contre le génome de référence

Objectif : Avoir le plus de lectures alignées

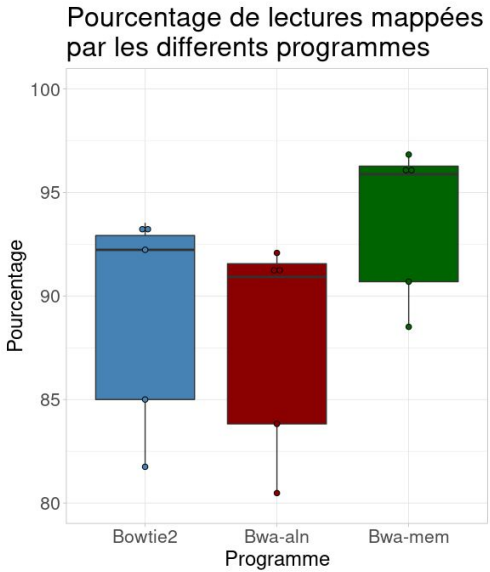


Test de différents programmes

Bowtie2

BWA aln

BWA mem



Traitement des lectures alignées

Picard Tools (filtration)
Samtools (filtration et conversion)

Etape 2 : Petits variants avec GATK

Description des étapes

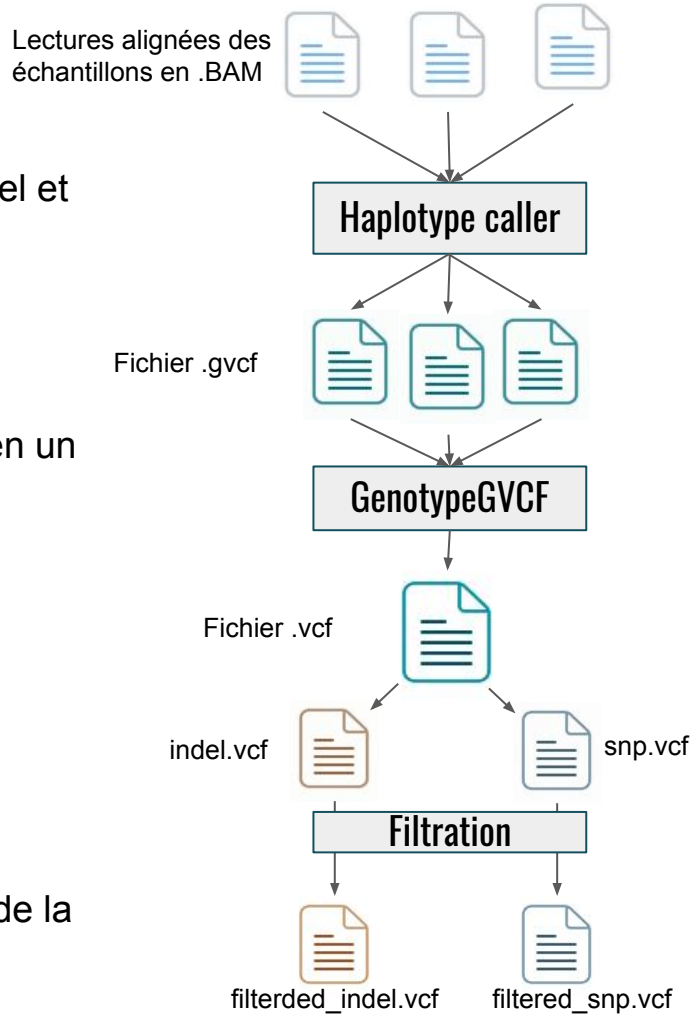
HaplotypeCaller : détecte snp et indel et fournit des fichiers gvcf

GenotypeGVCF : combine les gvcf en un seul vcf

Séparation des **snps et indels**

Filtration selon snps et indels

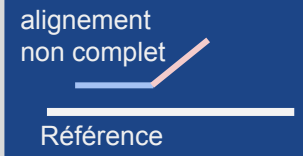
Sélection des **variants spécifiques** de la mutagenèse grâce à un script dédié



Etape 2 : Variants structuraux avec MetaSV

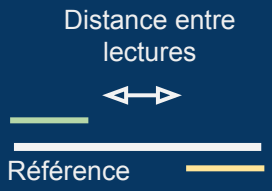
Pindel

Méthode :
"splits reads"
Datamining



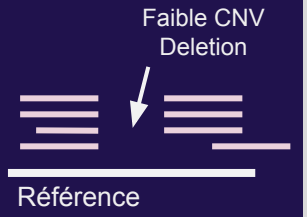
Breakdancer

Méthode :
Distances anormales entre lectures appariées



CNVnator

Méthode :
Copy Number Variation



Lumpy

Méthode :
Différents signaux de variations structurelles

Méthode :
Combiner les résultats des différents programmes

Avantage :
Variants **imprécis** ou de **faibles qualités**
Supprimer possible **faux variants**

MetaSV

\+ Sélection des **variants spécifiques**

Etape 3 : Prédiction de l'impact fonctionnel avec snpeff

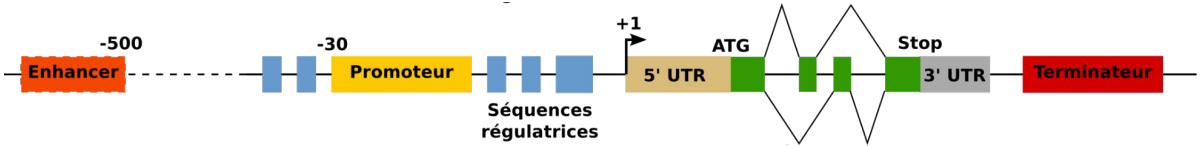
But : Prédire l'impact fonctionnel pour avoir une liste restreinte de gènes candidats

Fort (HIGH) : Cadre de lecture, site d'epissage, perte/gain codon stop

Modéré (MODERATE) : Modification d'un ou plusieurs codons

Faible (LOW) : Modification synonyme

Nul (MODIFIER) : Région inter-génique



Exemple de fichier de sortie

GENE	CHROM	POS	REF	ALT	DP	EFFECT	IMPACT	.BIOTYPE
ATOZI1	4	36920	CAA	C	69	upstream_gene_variant	MODIFIER	protein_coding
WRKY42	4	2221	TCCCCCC	T	35	splice_region_variant	LOW	protein_coding
BHLH14	6	363920	CAA	C	69	upstream_gene_variant	MODIFIER	protein_coding
AT4G1025	6	857995	G	GT	65	frameshift_variant	HIGH	protein_coding
AT3G1100	7	3448	C	A	56	stop_gained	HIGH	protein_coding

Mise à l'épreuve

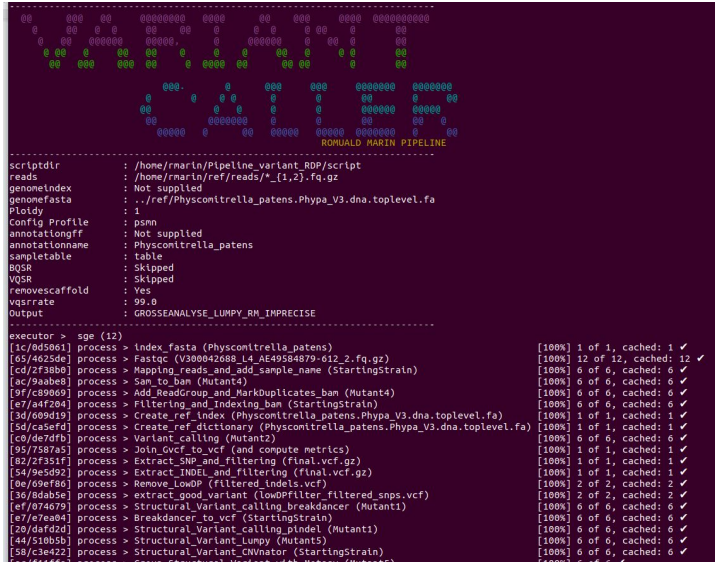
Simplicité d'exécution

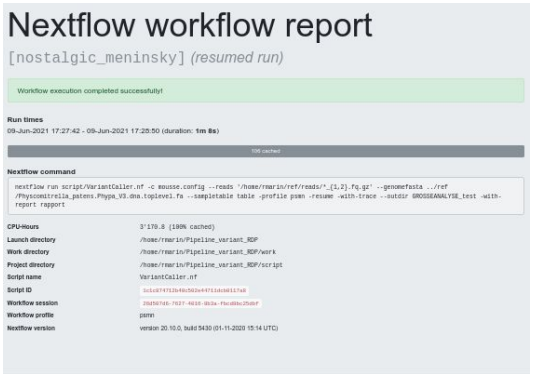
Une seule ligne de commande

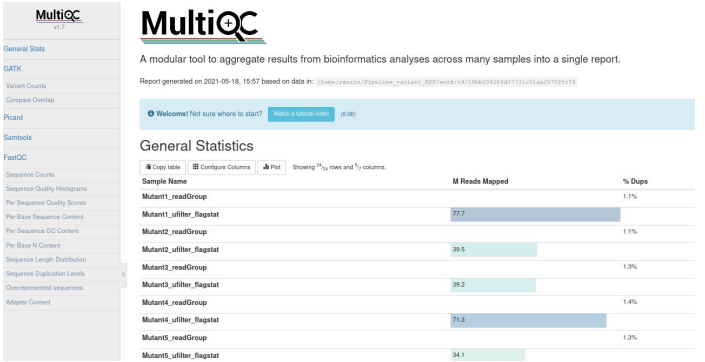
Informations sur l'avancement du pipeline

Temps d'exécution 30-35h

Rapport d'exécution concernant la partie technique du pipeline et concernant l'analyse



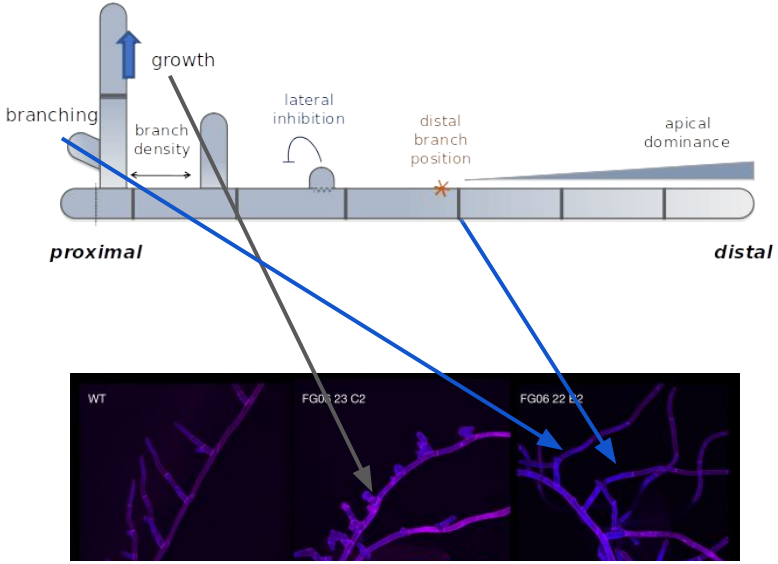




Mise à l'épreuve

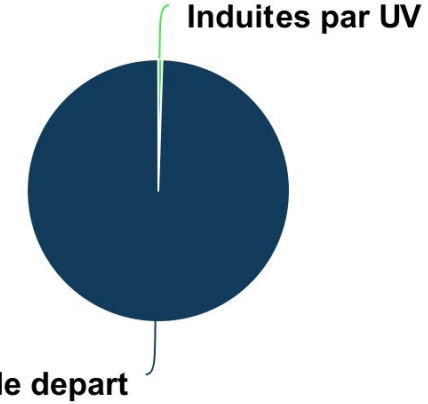
Physcomitrium patens

- Données issues de **Mousse**
- Mutations induites par **UV**
- **5 mutants** générés
- Développement anormale
- **Phyllotaxie**



Résultats

	Coverage	indel	snp	SV	Total
Mutant 1	**17,4**	234	1079	633	**1946**
Mutant 2	**8,9**	709	1470	1060	**3239**
Mutant 3	**8,8**	653	1774	1064	**3491**
Mutant 4	**15,9**	219	961	846	**2026**
Mutant 5	**7,7**	898	1916	986	**3800**



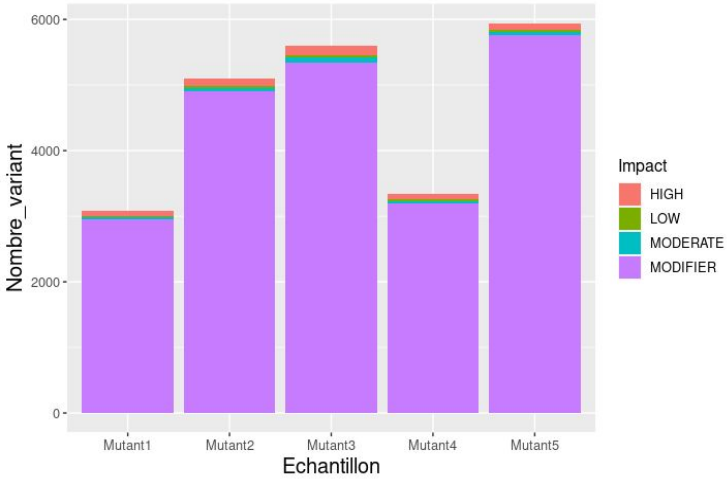
Mise à l'épreuve

Analyse impact fonctionnel des variants

Majorité de variants nul “**Modifier**”
→ Intergénique

Peu de variant à fort impact

Permet au généticien d’avoir une **liste restreinte** de variants candidats



Présence de variants intéressants :

Processus métabolique de phosphorylation

Le réseau métabolique de l’auxine

La fabrication de la membrane cellulaire

La fabrication de la pectine

Qualité de l’annotation :

Nombreux gènes prédits

Aucune information sur la fonction

Conclusion

Objectifs du cahier des charges validés :

✅ Programme avec **version à jour** (docker)

✅ Pipeline **facile d'utilisation**

✅ **Large spectre de variants** détectés

✅ Classification de l'**impact fonctionnel** des variants

Limite du pipeline :

Qualité initiale des données
Annotation et information fiable sur les gènes

Perspectives :

Données d'*Arabidopsis thaliana* (Meilleure annotation)
Ajout des informations dans rapport HTML

Merci de votre attention

Partie Discussion

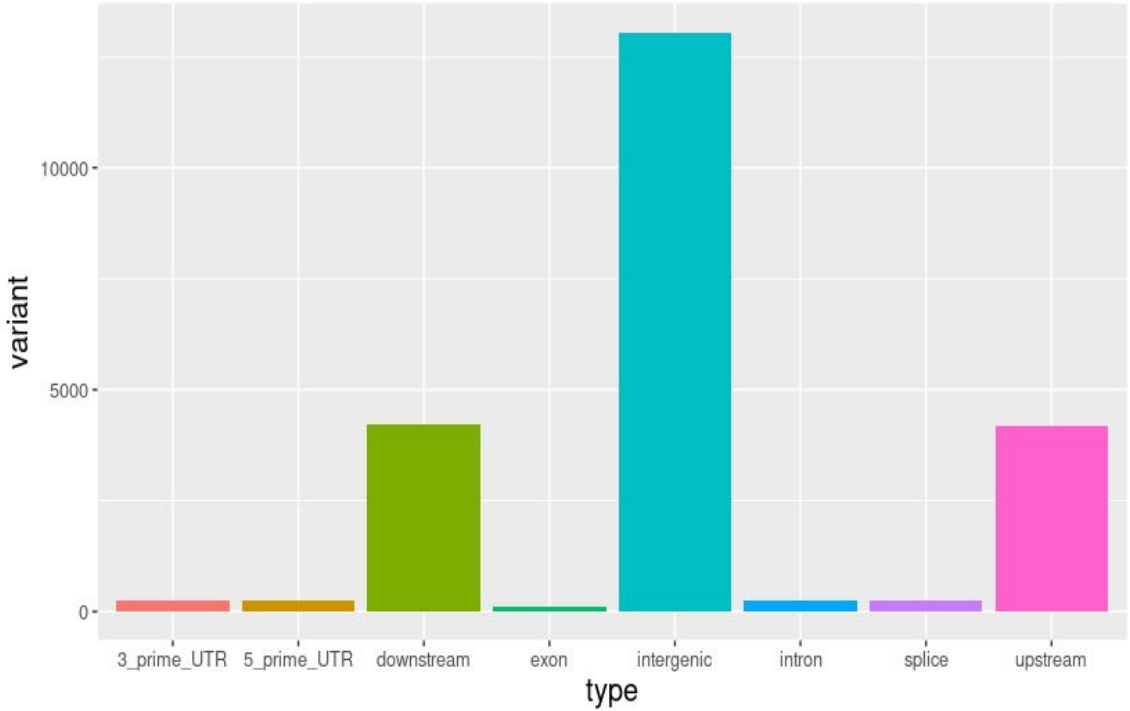
Mutagenèse par UV

Attendu :	Trouvé :
G -> A	20%
C -> T	16%

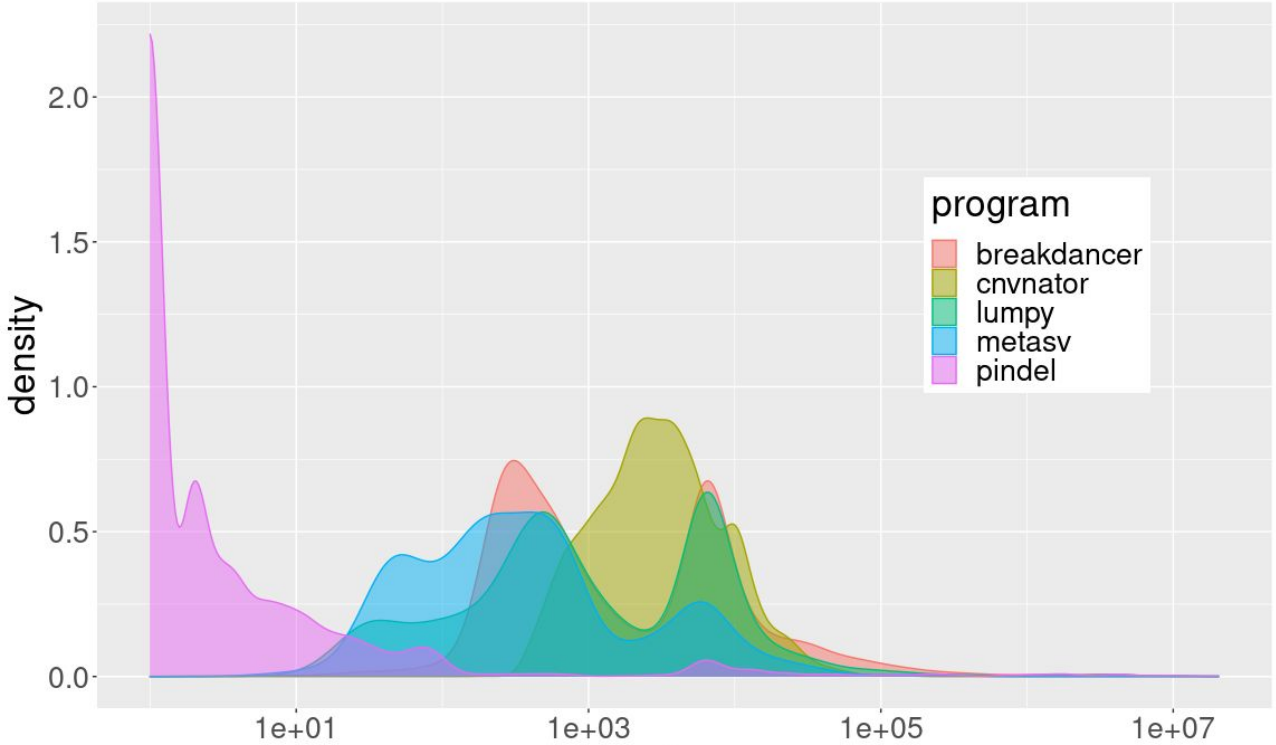
Taux “normal” de mutation d’une base vers une autre est d’environ 8% (1 / 12 nucléotides possibles)

Nakamura, M., Nunoshiba, T. & Hiratsu, K. Detection and analysis of UV-induced mutations in the chromosomal DNA of Arabidopsis. Biochem. Biophys. Res. Commun. 554, 89–93 (2021)

Catégorie des variants détectés (snpeff)



Spectre de détection



2 11033989 . G <DEL> . PASS CIEND=-9,6;END=11034921;SVLEN=-932;SVTYPE=DEL;CIPOS=-5,7;SVTOOL=MetaSV;SOURCES=2-11033989-2-11034921-932-Pindel,2-11033993-2-11034921-928-Lumpy,2-11034001-2-11035000-999-CNVnator,2-11034037-2-11034922-899-BreakDancer;NUM_SVMETHODS=3;NUM_SVTOOLS=4;VT=SV;SVMETHOD=RD,RP,SR;BD_CHR1=2;BD_POS1=11034036;BD_ORI1=10+0-;BD_CHR2=2;BD_POS2=11034922;BD_ORI2=0+10-;BD_SCORE=99.0;BD_SUPPORTING_READ_PAIRS=9 GT 1/1

3 25474944 . T <DEL> . PASS END=25481413;SVLEN=-6057;SVTYPE=DEL;SVTOOL=MetaSV;SOURCES=3-25474907-3-25481390-6565-BreakDancer,3-25475044-3-25481090-6046-Pindel,3-25475256-3-25481313-6057-Pindel;NUM_SVMETHODS=2;NUM_STOOLS=2;VT=SV;SVMETHOD=RP,SR;BD_CHR1=3;BD_POS1=25474906;BD_ORI1=31+0-;BD_CHR2=3;BD_POS2=25481390;BD_ORI2=1+28-;BD_SCORE=99.0;BD_SUPPORTING_READ_PAIRS=28 GT 1/1