



Plateforme de Recherche de Mutations



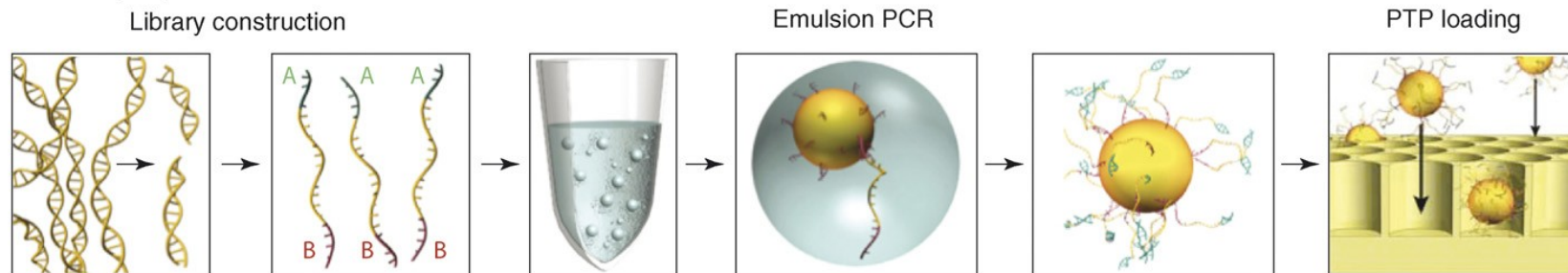
- ✓ Présentation des données produites par le GSFLX : type, qualité, ...
- ✓ Méthodes de détection de mutations
- ✓ Le projet 'cancer de la vessie'



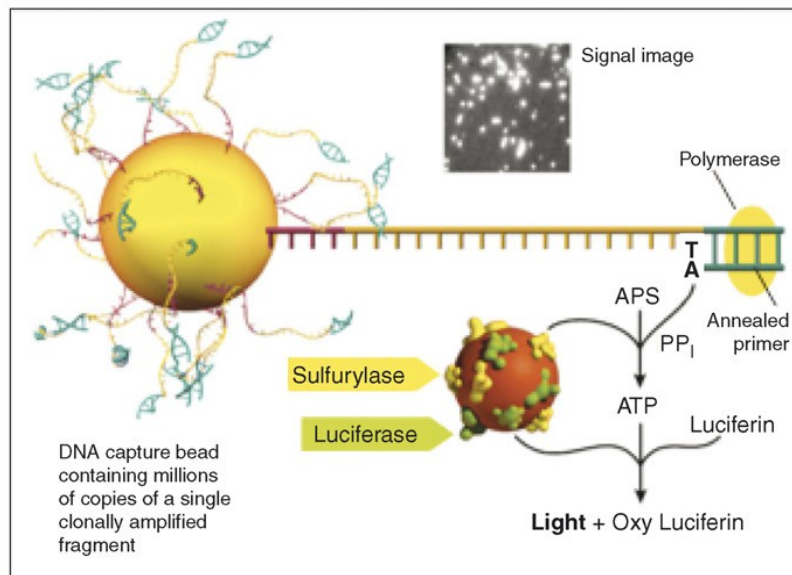


454 / Roche – Genome Sequence FLX

Roche (454) GSFLX Workflow:



1 fragment -> 1 bille
 1 bille -> 1 lecture

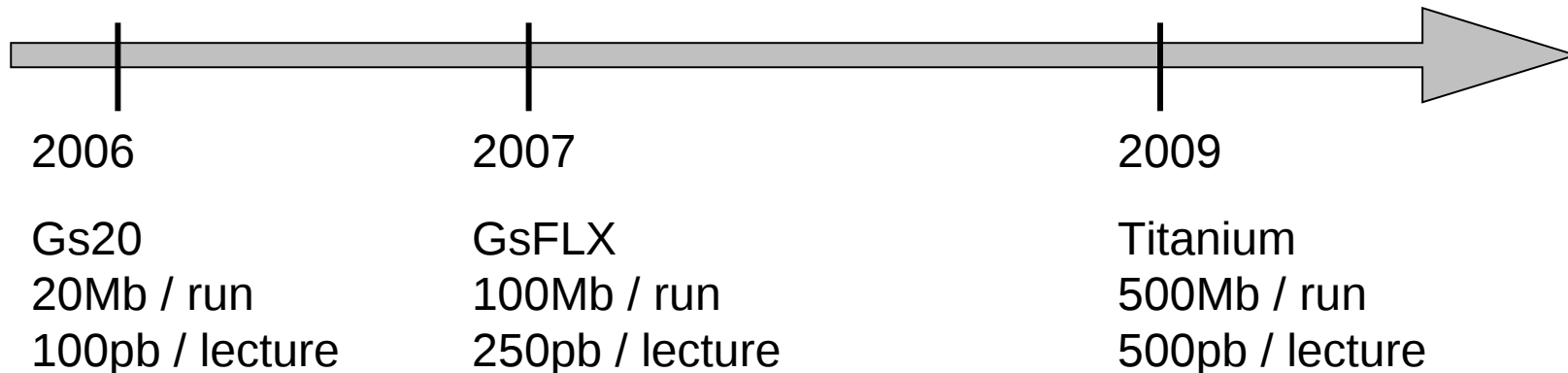


Pyrosequencing reaction

TRENDS in Genetics



454 / Roche – Genome Sequence FLX



✓ Version actuelle (GS FLX) :

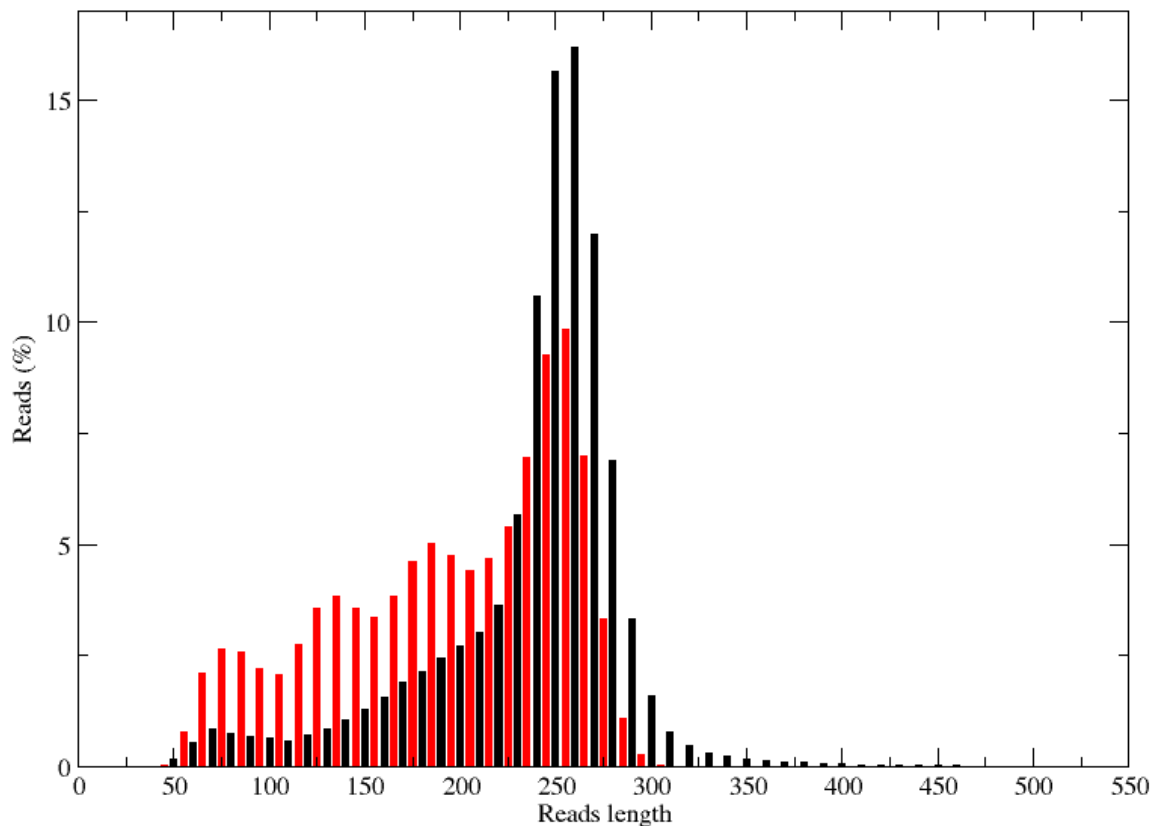
- ✓ Majorité des lectures à 250bp
- ✓ Environ 500.000 lectures / run et 100Mbp / run
- ✓ Durée du run : 8h

- ✓ Taux d'erreurs non négligeable dans les homopolymères
- ✓ Assemblage de qualité à environ 20X
- ✓ Pas de biais de clonage



454 / Roche – Genome Sequence FLX

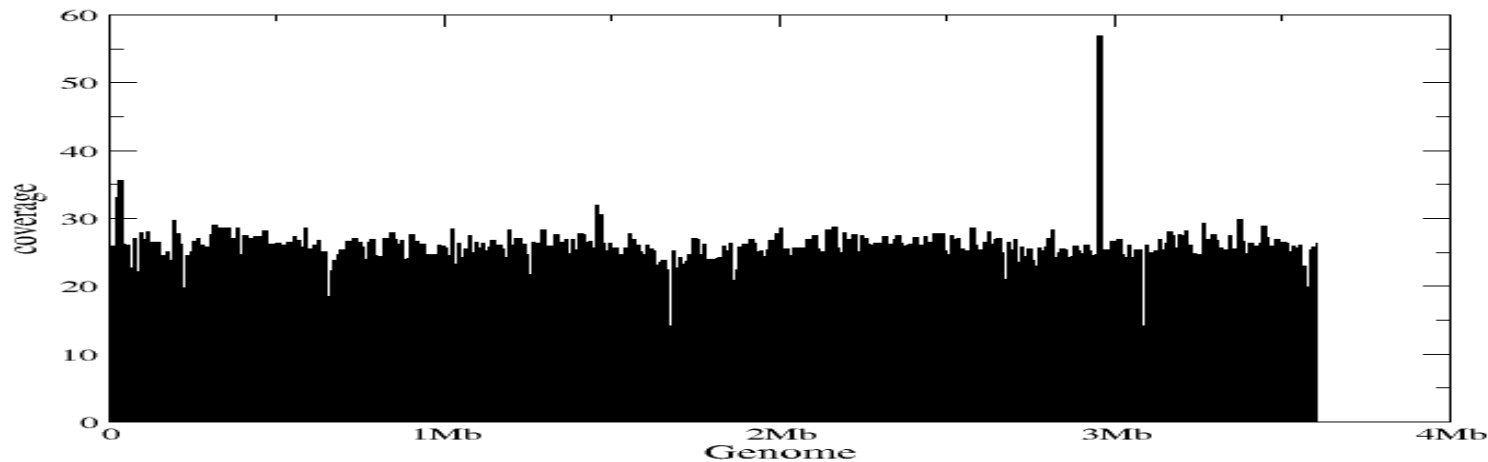
- ✓ Un run sur *Acinetobacter* (3,5Mbp) :
 - 522.876 lectures
 - taille cumulée de 96Mb, soit 26,7 équivalents génome (26,7X)





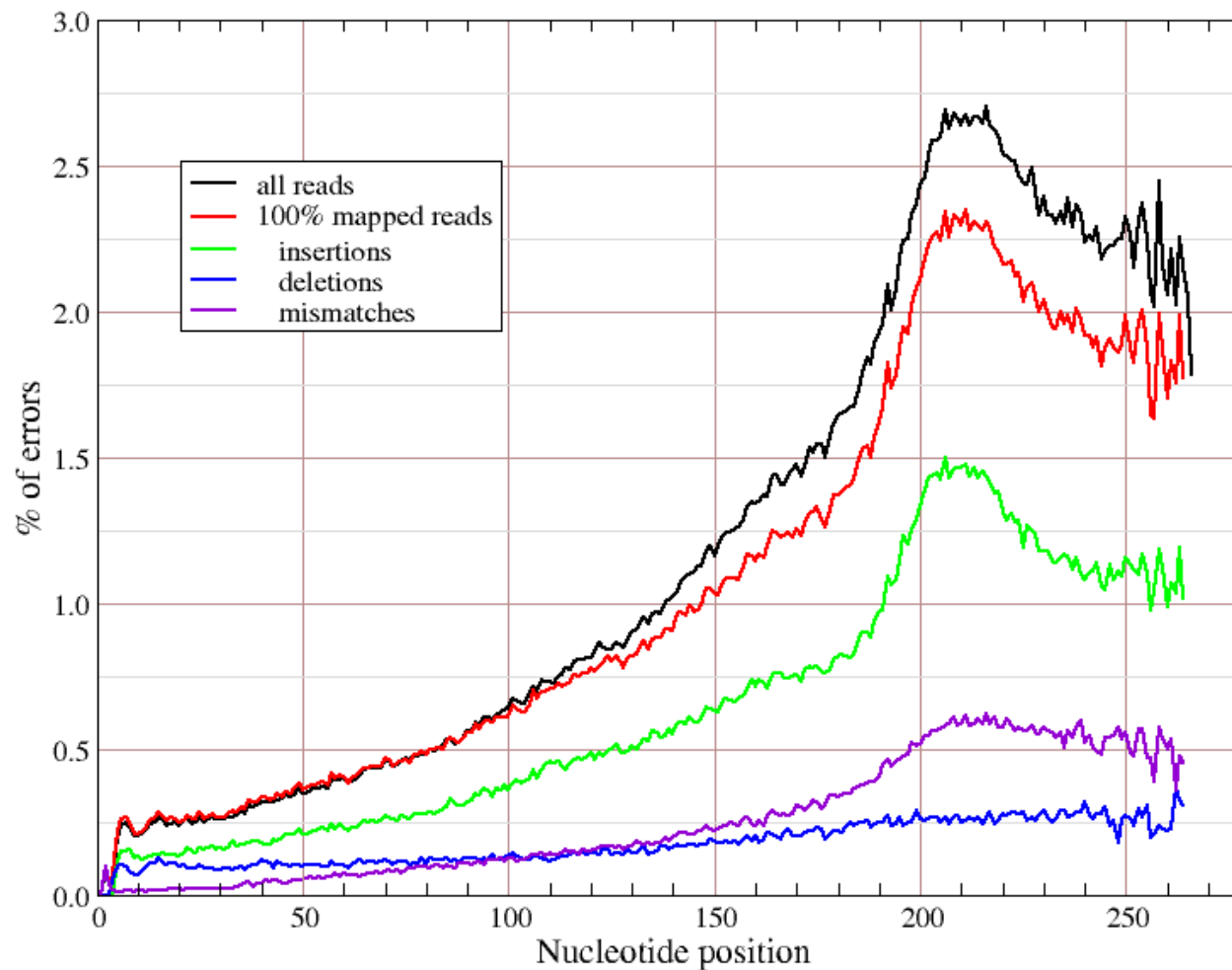
454 / Roche – Genome Sequence FLX

- ✓ Alignement des lectures au niveau nucléotidique
- ✓ 521.193 lectures mappées (soit 99,68%)
- ✓ 93.553.967 nt alignés contenant 800.295 erreurs (soit $8,6 \cdot 10^{-3}$ erreurs)
- ✓ Sur les 800.295 erreurs : 17% délétions, 62% insertions, 21% mismatches (12% de Ns).
- ✓ Sur les 348.796 lectures mappées à 100%, 109.637 sont sans erreurs (31%)





454 / Roche – Genome Sequence FLX



Neighbourhood Quality Standard (NQS)

Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513-516 (2000)

Méthode basée sur les scores qualités attribués à chaque base des lectures.

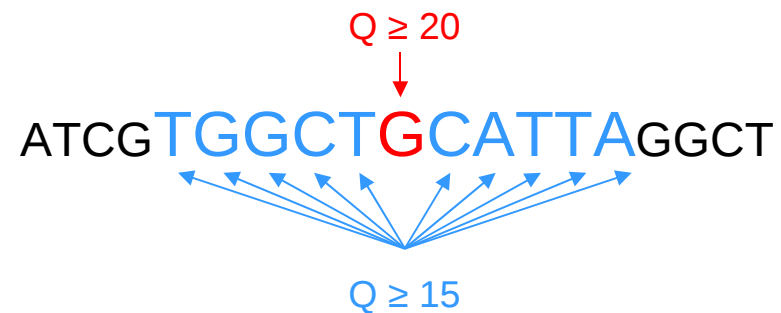
Le NQS a pour objectif de diminuer l'impact du taux d'erreurs du 'base calling' sur l'identification des sites polymorphes

$Q = 10 \Rightarrow p(\text{base incorrecte}) = 0,1$

$Q = 20 \Rightarrow p(\text{base incorrecte}) = 0,01 \dots$

Une base est marquée NQS si :

- ✓ son score qualité $Q \geq 20$
- ✓ les 5 bases de chaque coté ont chacune un score $Q \geq 15$



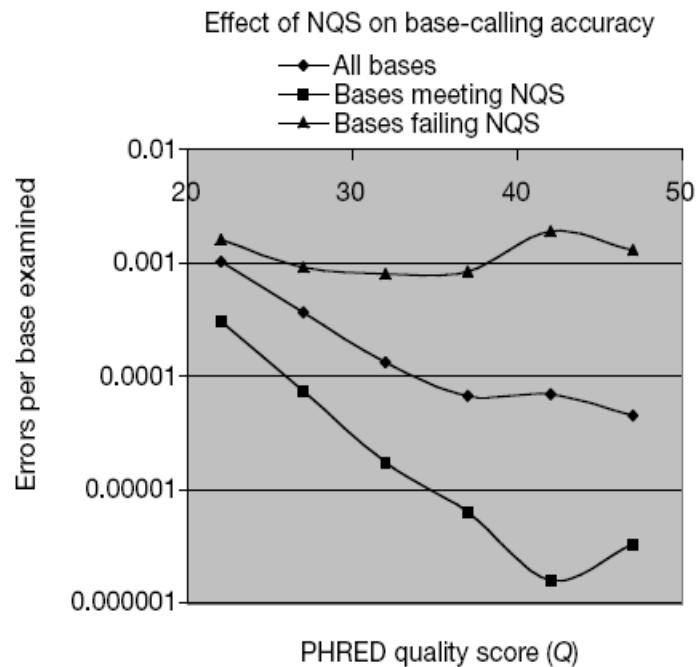


Figure 1 Impact of quality criteria on error rates. Data are plotted according to the PHRED Q score of each base^{8,9} and reported in bins of five PHRED Q units; only base substitution errors were counted. As previously reported⁹, overall PHRED scores accurately predict the observed rates of base-calling errors; however, bases meeting the NQS display substantially lower error rates than are predicted by their PHRED scores. Although the effect is proportionally greatest for high PHRED scores, the bulk of errors avoided are found in bases with lower PHRED scores (that is, those with the highest error rates).

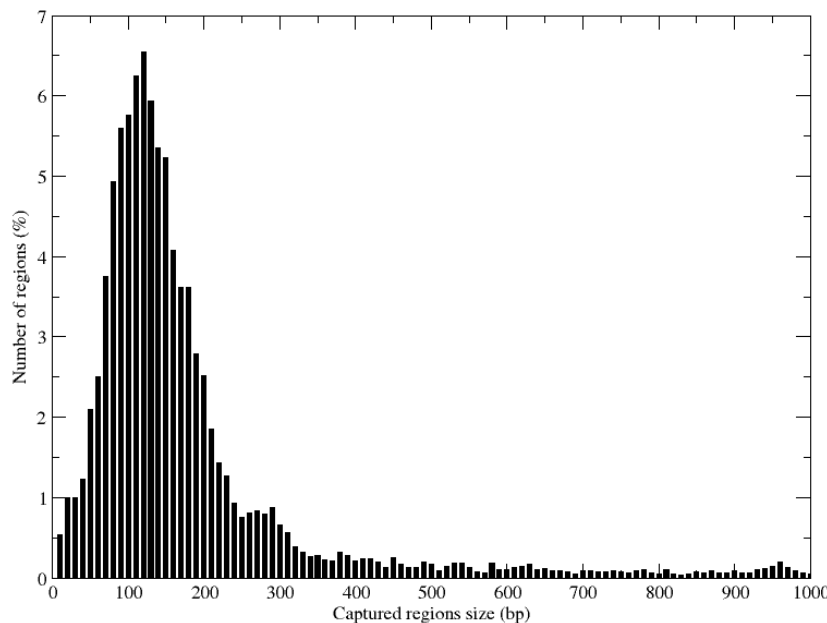
Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516 (2000)

- ✓ Avec les nouvelles technologies de séquençage (NTS) les lectures sont plus courtes
- ✓ Le principal problème devient l'alignement des lectures (mauvais placement sur le génome de référence)
- ✓ D'où la nécessité d'utiliser des algorithmes qui tiennent compte des spécificité des données issues des NTS.
- ✓ Algorithmes existants :
 - ✓ gsMapper (Roche/454)
 - ✓ SSAHA_pileup (Sanger), version adaptée de SSAHA_snp
 - ✓ GigaBayes (Boston College), version adaptée de PolyBayes

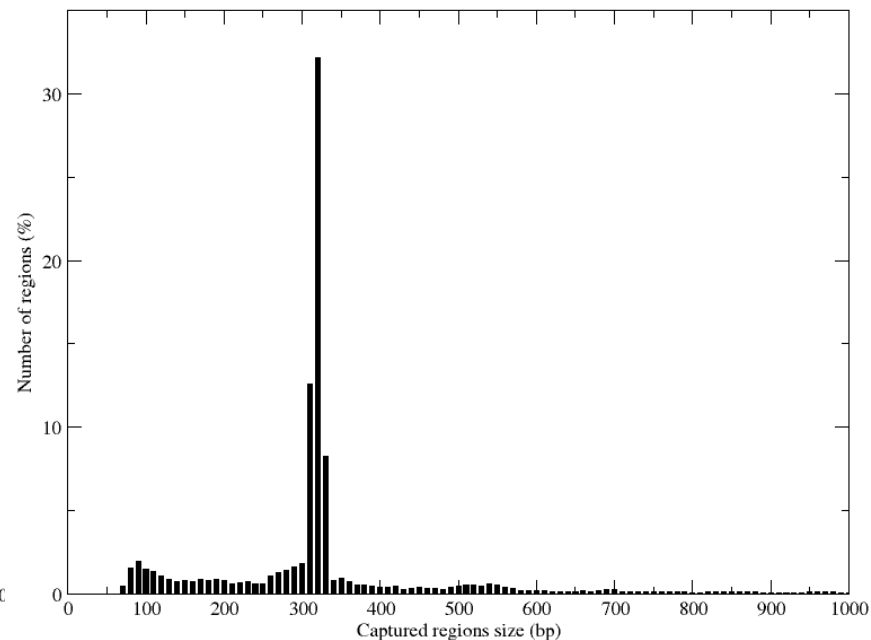
- ✓ Objectif du projet : recherche de nouveaux gènes suppresseurs de tumeurs impliqués dans les cancers de la vessie par l'identification de mutations.
- ✓ Collaboration avec F. Radvanyi (Institut Curie)
- ✓ Sélection des gènes :
 - ✓ gènes présents dans les régions récurrentes de perte : régions génomiques fréquemment perdues dans ces les cancers de la vessie.
 - ✓ gènes connus par la base de données 'Cancer Gene Census' comme ayant des mutations impliqués dans la progression tumorale
 - ✓ sélection de 1.251 gènes , 13.315 exons , taille cumulée d'environ 4 Mb
- ✓ 8 échantillons : 4 échantillons tumoraux et 4 échantillons normaux appariés

- ✓ prévision : 8 échantillons avec 1 run GSFLX par individu (soit ~ 100Mb)
- ✓ 13.315 régions ciblées : 3,97Mb (moyenne de 300pb)
- ✓ Après passage chez NimbleGen : 13.944 régions ; 5,6Mb (moyenne de 400pb)

Régions sélectionnées



Régions capturées

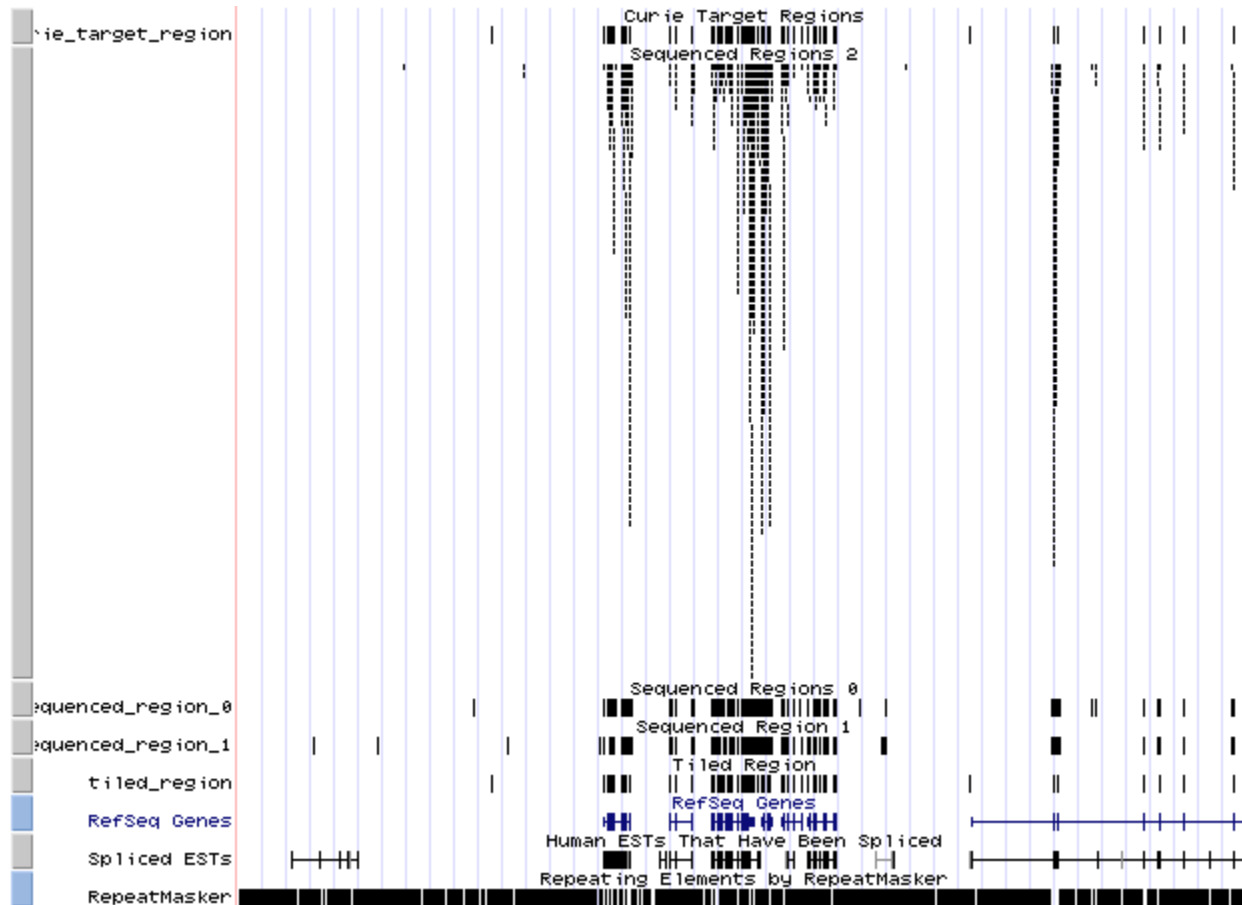


- ✓ Alignement des lectures provenant des 8 échantillons sur le génome humain



- ✓ Calcul de la sensibilité et de la spécificité de la capture

- ✓ Alignement des lectures provenant des 8 échantillons sur le génome humain



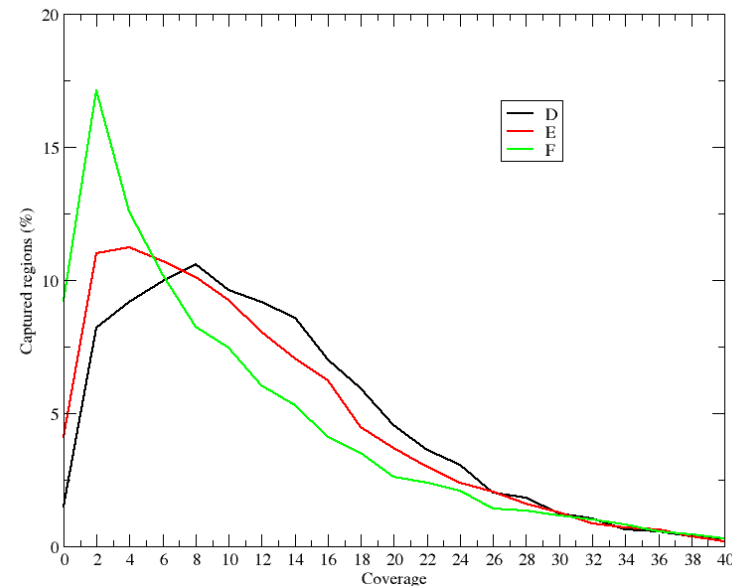
✓ Alignement des lectures provenant des 8 échantillons sur le génome humain

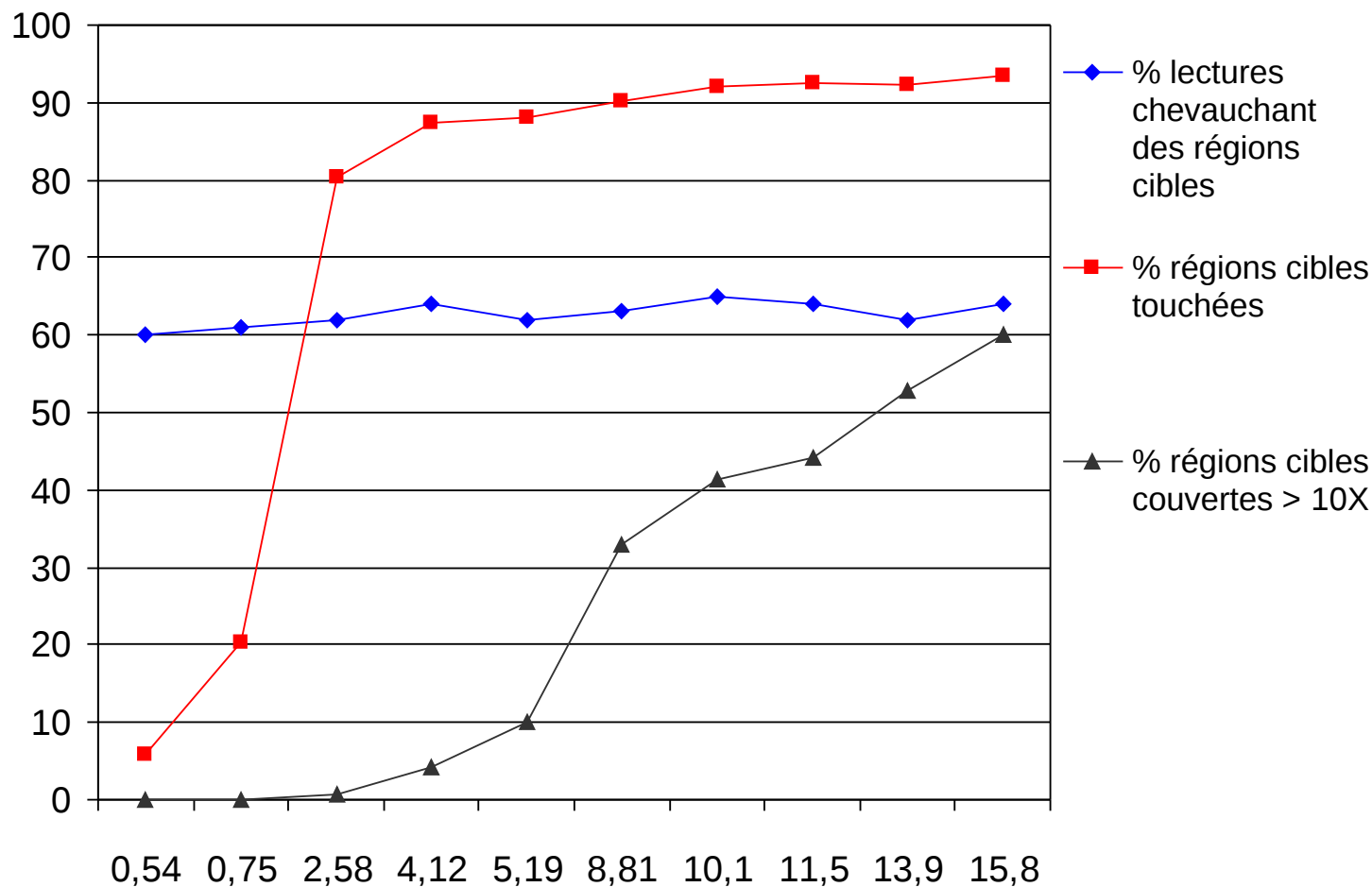
	B	C	D	E	F	G	H	I
# lectures	740.642	964.866	602.719	601.841	683.096	42.947	480.811	59.167
# lectures alignées	649.017 (88%)	822.999 (85%)	564.580 (94%)	531.657 (88%)	607.093 (89%)	32.755 (76%)	431.060 (90%)	53.022 (90%)
# lectures chevauchant des régions cibles	450.267 (69%)	525.778 (64%)	353.295 (63%)	348.492 (66%)	269.594 (44%)	977 (3%)	297.016 (69%)	4.422 (8%)
# lectures incluses dans des régions cibles	220.646 (49%)	260.185 (49%)	175.029 (50%)	160.027 (46%)	119.974 (45%)	424 (43%)	131.609 (44%)	1.729 (39%)
# régions cibles touchées	12.275 (92%)	12.434 (93%)	12.796 (96%)	12.325 (93%)	10.574 (79%)	783 (6%)	12.261 (92%)	2.699 (20%)
# régions cibles entièrement couvertes	10.932 (82%)	11.405 (86%)	11.091 (83%)	10.856 (82%)	8610 (65%)	142 (1%)	10.862 (82%)	622 (5%)

Projet cancer de la vessie

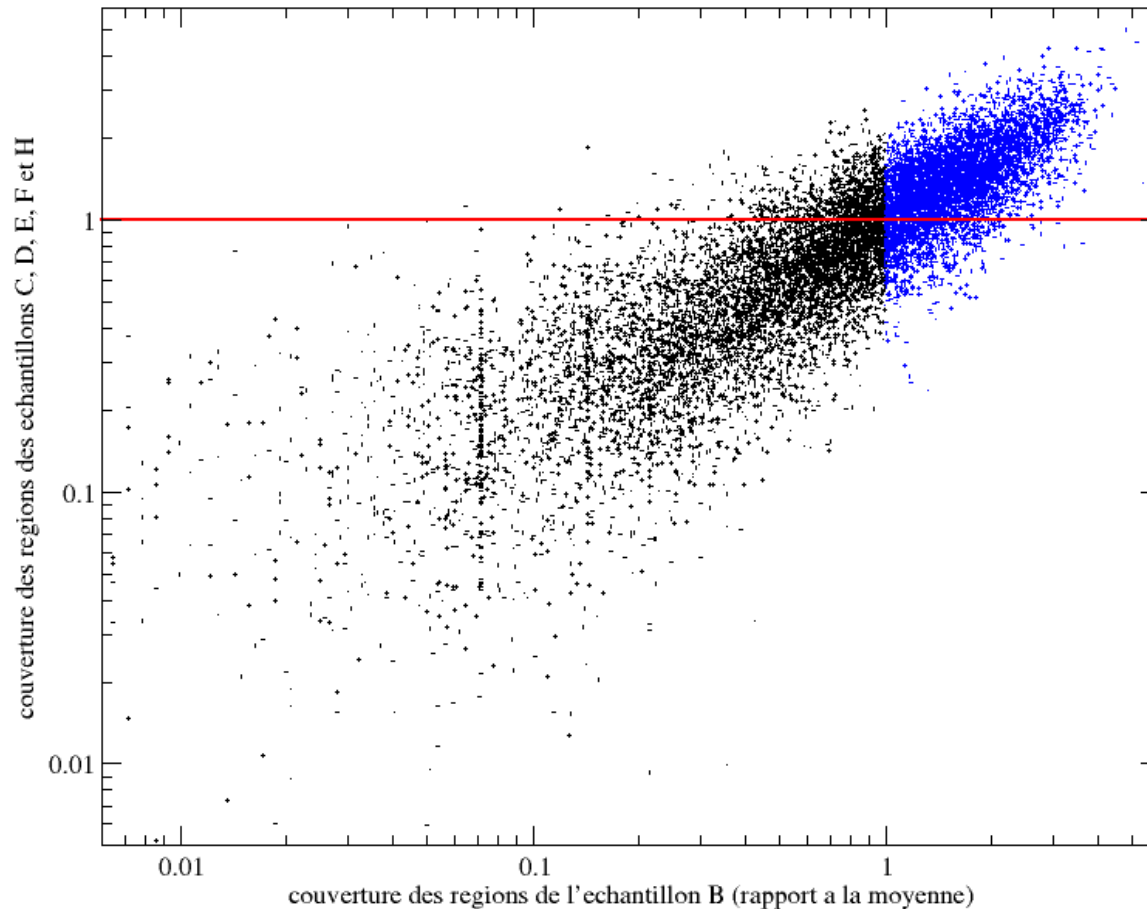
	B	C	D	E	F	H
Couverture initiale	42,9	53,3	35,1	35,1	41,1	29,6
Couverture moyenne	13,9	15,8	12,7	11,5	10,5	10,1
Couverture minimale	0	0	0	0	0	0
Couverture maximale	80,7	102,2	102,1	113,0	111,0	86,0
# régions couvertes à 10X	7.026 (53%)	7.985 (60%)	7.123 (54%)	5.886 (44%)	4.097 (31%)	5.502 (41%)

Avec >30X initialement, on ne couvre qu'environ 50% des régions avec une couverture supérieure à 10X





Les régions faiblement couvertes sont souvent communes à différents échantillons => biais de capture



✓ Variations détectées

✓ All_diff : différences reportée par au moins deux lectures indépendantes

```
chr1          1673394+  ATAAATGTAAACATTGAATGGCAGACGACTCCCTTCCCCTTGAAATCTTAA 1673452
                                     * * *
FJIN7VJ01DFLX9      42+  ATAAATGTAAACATTGAATGGCAGACA ACTCCCTTCCCCTTGAAATCTTAA 100
FJIN7VJ01ESS91     30+  ATAAATGTAAACATTGAATGGCAGACA ACTCCCTTCCCCTTGAAATCTTAA 88
                                     * * *
```

✓ HC_diff : différence de haute qualité : reportée par au moins 3 lectures, avec au moins une lecture alignée en forward et une en reverse

```
chr1          38111352+  CACCCGAGTTGAGCTTTGTAGCTGTCATCTTATTTACGAAGAGCTCCAA 38111411
                                     * * *
FJIN7VJ01AIOPN      224+  CACCCGAGTTGAGCTTTGTAGCTGTCGTCTTATTTACG          265
FJIN7VJ01C8XXY     166+  CACCCGAGTTGAGCTTTGTAGCTGTCGTCTTATTTACGAAGAGCTCCAA 225
FGXZNY01B97AX     104-  CACCCGAGTTGAGCTTTGTAGCTGTCGTCTTATTTACGAAGAGCTCCAA 45
FIZ3LQF01DFXB9    112-  CACCCGAGTTGAGCTTTGTAGCTGTCGTCTTATTTACGAAGAGCTCCAA 53
```

Echantillon	All_diff	HC_Diff
B	19.567	4.523
C	23.012	5.325
D	15.915	4.490
E	15.793	3.826
F	13.710	3.460
H	13.731	3.706
moyenne	16.955	4.222

- ✓ localisation des variations (exon, intron, intergénique) et comparaison entre échantillon
- ✓ Sélection :
 - ✓ variations de haute qualité présentes dans les échantillons tumoraux
 - ✓ absentes des échantillons normaux (apparié et autres)
 - ✓ absentes des base de données de variations connues
 - ✓ localisées dans les exons ou à proximité (50pb)

Exemple sur un échantillon tumoral	HC diff dans exons (1)	(1) + Absence des autres échantillons (2)	(2) + Inconnus	
	# variations	# variations	# variations	# gènes
Couverture				
2	3.092	406	360	244
3	3.092	406	360	244
4	2.721	218	183	132
5	2.439	141	120	96
6	2.202	110	93	77
10	1.457	53	45	40

- ✓ Initialement environ 20.000 variations de haute qualité

- ✓ Une 50aine de variations à valider par re-séquençage après classification et sélection

- ✓ Les critères de sélection importants :
 - ✓ qualité de la variation (profondeur de séquence)
 - ✓ localisation de la variation
 - ✓ comparaison entre échantillon et avec les variations connues

- ✓ Couverture de 10X environ nécessaire pour une détection optimale des mutations
- ✓ Le nombre de lectures à séquencer dépend des régions ciblées (nombre, taille) et de la qualité du run (taille moyenne des lectures, qualités)
- ✓ Détection des mutations basée sur la qualité des nucléotides et la longueur des lectures (unicité du placement)
- ✓ Insertions/délétions plus difficiles à détecter avec GSFLX à cause du taux d'erreurs dans les homopolymères
- ✓ Nécessité de classifier les variations pour réduire l'espace de recherche et de valider chaque variation

contact: pfm@genoscope.cns.fr