Premières Analyses des Données de Séquençage

Samuel BARRETO

19 novembre 2015

I Qualité des données

I.A *qualité du séquençage*

Globalement la plupart des séquences était de bonne qualité. Sur les 192 envoyées à séquencer, 179 ont été retenues pour l'analyse, soit 93%.

Étant donnée la faible qualité des bases en début et en fin de séquence, elles ont été tronquées. Le score 28 semblait le seuil naturel de qualité. De plus, toutes les séquences qui avaient une longueur inférieure à 620 étaient généralement mal alignées. Elles ont été éliminées de l'analyse.

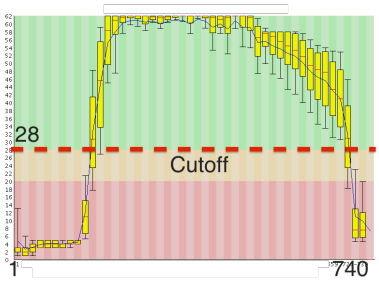

FIGURE 1: Qualité des séquences *avant* d'être trimmées et filtrées sur la qualité

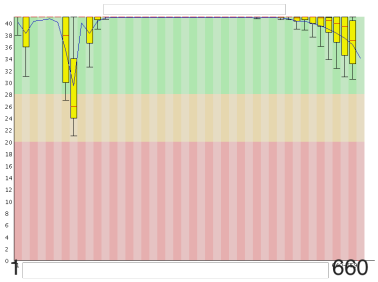

FIGURE 2: Qualité des séquences *après* avoir été trimmées et filtrées sur la qualité

I.B *Présence de contaminations ?*

Pas évident à déterminer : voir après.

I.C *Observations générales*

nombre de SNP par gene synthétique	**moyen**	**sd**	**median**
global	14.4	6.4	15.0
strong	15.5	6.2	15.5
weak	13.3	6.5	13.0

I.D *Nombre de SNPs*

	strong	**weak**
nombre de SNP par gène synthétique	1337	1162
nombre de substitutions	**1791**	708

Pour un nombre de SNPs par mutant sensiblement équivalent, il y a 2.53 fois plus de substitutions *strong* que *weak* !

II Distribution des SNPs

II.A *Distribution globale*

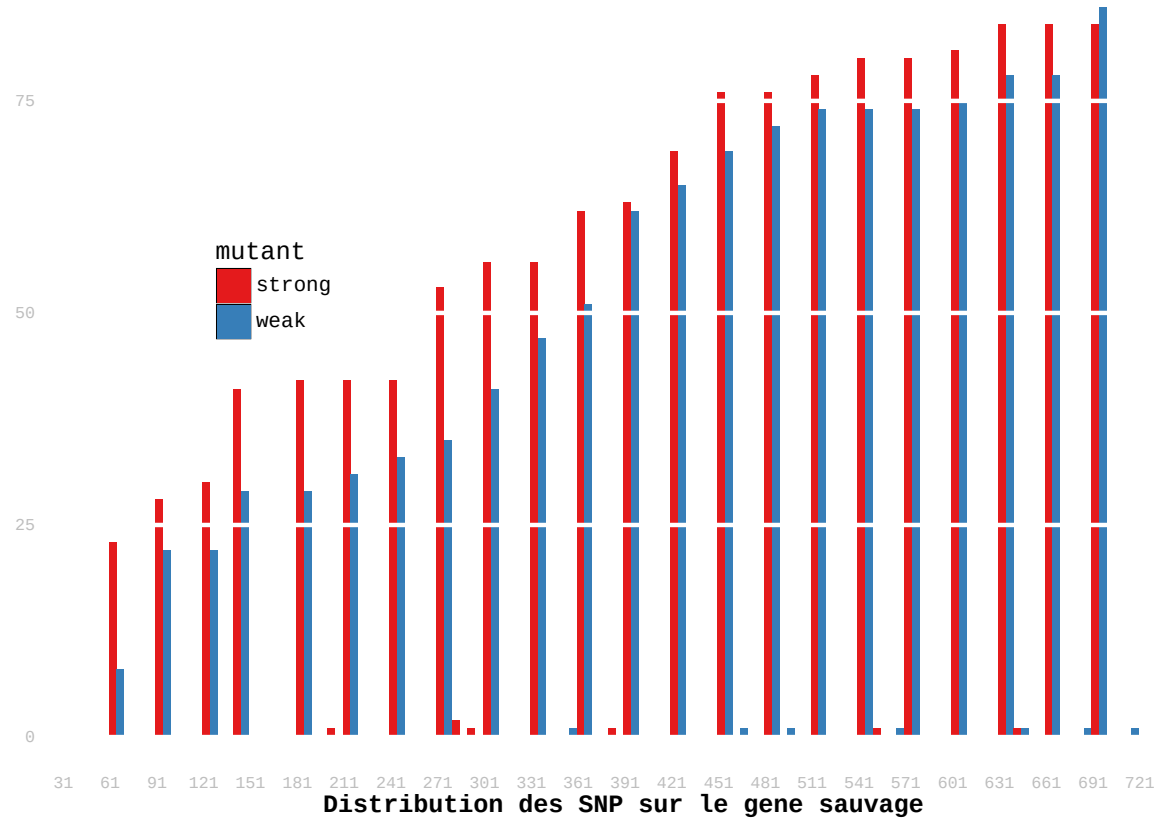

FIGURE 3: La distibution des SNPs, sans tenir compte de la qualité de la mutation. La couleur représente le mutant d'origine, qu'il soit sensé être Weak ou Strong.

Sans tenir compte de la qualité des mutations, on obtient le résultat en figure 3. Il semble qu'il n'y ait pas de différence significative dans la position des mutations entre les deux types de mutants.

Malgré les filtres, il reste du bruit de fond, avec quelques SNP qui ne sont pas à leur place normale.

Principale conclusion : il y a plus de substitutions dans les régions 3' que 5', sur la fin de la conversion tract. Où se fait le switch ?

À noter qu'on n'a pas de SNP avant la position 61. C'est dû au trimming *des séquences. On perd l'information des premiers SNP.*

II.B *Distribution de la qualité des mutation*

Le graphe suivant montre un résultat surprenant. Lorsque le gène synthétique est de type Strong, les substitutions occasionnées sont — quasiment — exclusivement de type *strong*.

Mais lorsque le mutant est de type Weak, les substitutions occasionnées sont à la fois de type *weak* et de type *strong*, en fonction de la localisation du SNP. Les SNP autour de la position 241 et 601

sont tous *strong*. Comment interpréter ça ?

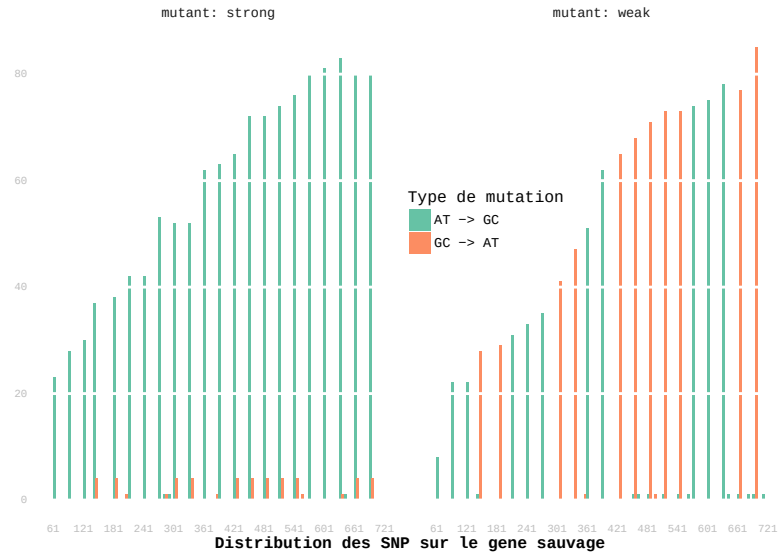

FIGURE 4: **Distribution des SNP par position sur la séquence de référence.**

On retrouve bien les positions des polymorphismes "artificiels", toutes les 30 paires de bases. En vert les mutations *strong* et en rouge les mutations *weak*. Les mutants Strong montrent exclusivement des substitutions *strong*. Les mutants Weak montrent cependant des choses différentes. Il y a beaucoup de mutations *strong*, contrairement à l'attendu.

MONTRÉ AUTREMENT, on voit le problème plus clairement.

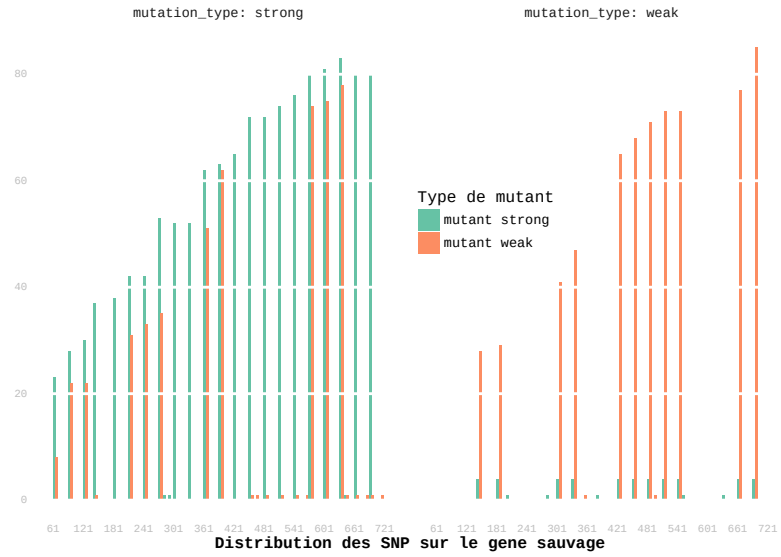

FIGURE 5: **Distribution de la qualité des substitutions**.

À gauche la distribution des substitutions vers GC, à droite celle des substitutions vers A ou T. On voit bien que les mutations *weak* sont quasiment exclusivement dans les mutants de type Weak, alors qu'on retrouve des mutations *strong* dans les deux types de mutants.

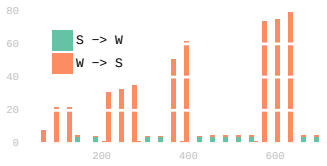

FIGURE 6: Avec ici un focus sur les *outliers qui n'en sont pas*

III Distribution de la position de basculement

III.A *Basculement global*

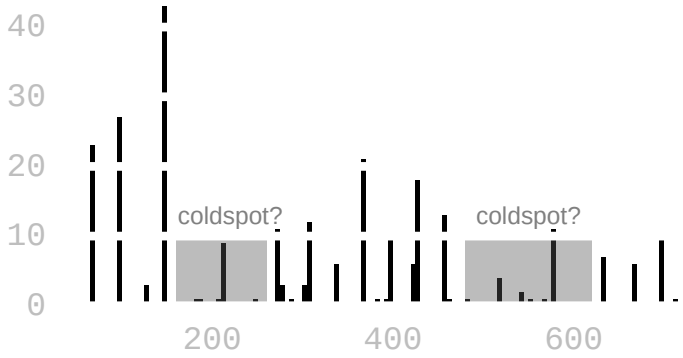

FIGURE 7: **Position des switch, indifféremment de la qualité de la substition ou du mutant**.

Il y a des disparités dans la distribution des positions de basculement. Il y a beaucoup de basculement dès le début, moins vers la fin. Il semble y avoir une sorte de *coldspot* local, autour de 500bp et 200bp sur la séquence de référence.

Il y a des disparités dans la distribution des positions de basculement. Il y a beaucoup de basculement dès le début, moins vers la fin. Il semble y avoir une sorte de coldspot local, autour de 500bp sur la séquence de référence.

III.B *Position initiale de basculement par type de mutation*

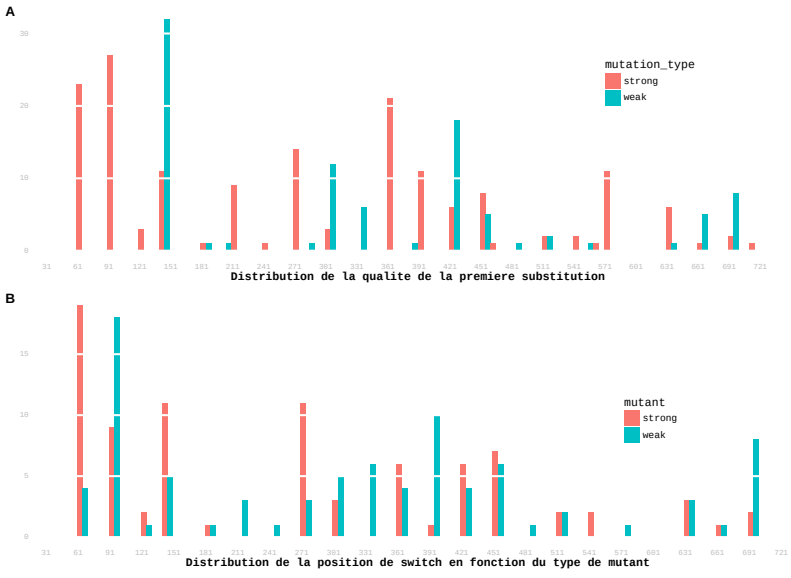

FIGURE 8: Position des switch en fonction du type de mutant.

Le graphe A représente la distribution et la qualité du premier SNP, AT $\mapsto$ GC est *strong* et GC $\mapsto$ AT est *weak*. Le graphe B représente la distribution du premier SNP par clone, en fonction de la qualité du clone, Strong ou Weak.

On ne semble pas voir de différence significative. Dans les deux cas, les distributions sont assez similaires pour le *weak* et le *strong*. Cependant, des différences existent entre les graphes A et B : toutes les premières substitutions sont de type *strong*.

Il y a toujours le même patron de coldspot autour de 541bp.

Le graphe A a été obtenu en filtrant le jeu de donnée de la façon suivante :

— groupe par clone et par type de mutation.

— demande la première position de SNP “groupwise”.

Le graphe B a été obtenu en filtrant le jeu de donnée de la façon suivante :

— groupe par clone

— demande la première position de SNP “groupwise”.

À vue d’œil, il n’y a pas de variation significative sur la distribution des SNPs, quelle que soit la qualité du gène synthétique ou de la substitution.

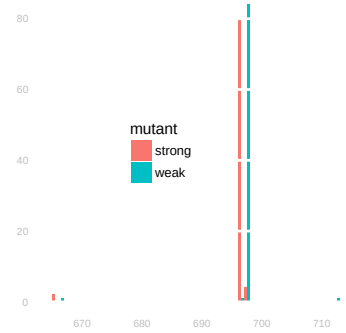

FIGURE 9: Position du dernier SNP.
Pas de variation là dessus. À priori les deux mutants terminent au même endroit, c’est à dire au dernier site avant le cutoff de trimming.