# Analyse Intégrative

Dystrophie musculaire Emery-Dreifuss

Équipe 20

Bourgeois Jérémie
Warter Perrine
Hakobjanyan Tatevik

Nous allons nous intéresser à une mutation découverte sur une séquence partielle d'un transcrit Cette mutation est responsable de la maladie appelée dystrophie d'Emery-Dreifuss. Notre but sera de définir les caractéristiques de la séquence mutée impliquée dans cette maladie.

Pour ce faire, nous allons tout d'abord rechercher des similitudes avec d'autres séquences connues. On utilise alors l'outil Blast sur le serveur du NCBI.

Programme : Blastn car notre séquence est nucléotidique.
Base de données : Nucléotide collection Nr/Nt, on utilise cette base de données car elle est non-redondante et contient l'ensemble des séquences nucléotidiques des différentes bases de données.
Paramètres : par défaut

On cherche à trouver les séquences les plus proches, quels que soient les niveaux de similitudes, on utilise donc "somewhat similar sequences" comme optimisation de sélection de programme.

| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| Homo sapiens lamin A/C (LMNA), transcript variant 6, mRNA | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 2467 | NM_001282625.2 |
| Homo sapiens lamin A/C (LMNA), transcript variant 3, mRNA | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 3088 | NM_170708.4 |
| Homo sapiens lamin A/C (LMNA), transcript variant 2, mRNA | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 2029 | NM_005572.4 |
| Homo sapiens lamin A/C (LMNA), transcript variant 7, mRNA | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 3028 | NM_001282626.2 |
| Homo sapiens lamin A/C (LMNA), transcript variant 1, mRNA | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 3178 | NM_170707.4 |
| Homo sapiens lamin C mRNA, complete cds, alternatively spliced | Homo sapiens | 3169 | 3169 | 87% | 0.0 | 100.00% | 3521 | MZ054260.1 |

*Figure 1 : Résulats Blastn*

On obtient avec Blastn un % d'identité de 100% pour les 18 premiers résultats, ce pourcentage est supérieur au seuil de 30% requis pour pouvoir poser une hypothèse d'homologie. On observe ensuite une E value de 0 (arrondie à 0 car très faible) pour tous nos hits.

L'hypothèse d'homologie est donc fiable, car plus la E value est faible, plus un alignement est fiable et moins les chances d'avoir obtenu cet alignement au hasard sont élevées. Le taux de recouvrement est de 87% pour la majorité de nos hits.

On peut poser une forte hypothèse d'homologie qui concerne donc 87% de notre séquence. En effet, l'alignement des séquences est incomplet en 3'

Tous nos meilleurs hits correspondent au gène LMNA mais pour le reste de notre étude on utilisera le meilleur hit, correspondant au transcript variant 6 du gène LMNA de l'Homo Sapiens codant pour la synthèse des lamines A et C par épissage alternatif.

Pour déterminer la CDS de notre séquence on utilise tout d'abord l'outil PlotOrf du serveur EMBOSS, on apprend alors que notre ORF la plus probable se situe sur la phase F2, on utilise alors l'outil ShowOrf en se basant sur la phase F2. On obtient alors notre CDS qui commence en position 149 et finit au nucléotide 1862.

La séquence de notre variant fait 2467 nucléotides de long et son CDS se situe entre 647 et 2365 d'après la fiche GenBank.

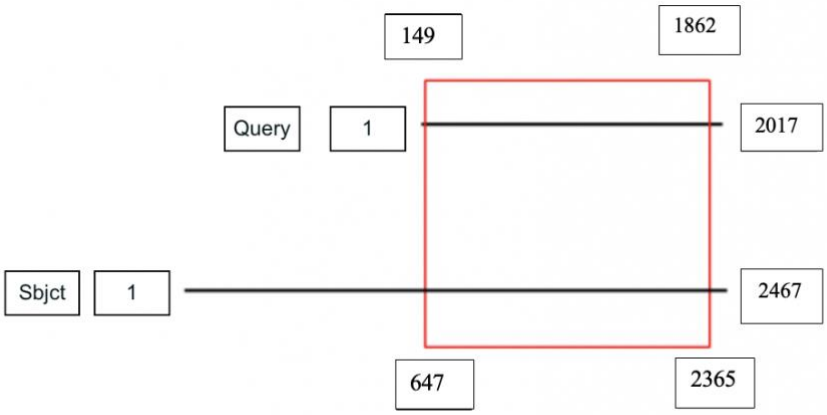



*Figure 2 : Diagramme de recouvrement*

Comme on retrouve le terme mRNA dans les descriptions de nos meilleurs hits, on peut affirmer que notre séquence est codante. On peut alors faire l'hypothèse de fonction que notre séquence complète code pour la lamine.

Pour confirmer cette hypothèse de fonction, on effectue un Blastx, toujours sur le serveur du NCBI en utilisant cette fois la base de données Swiss-Prot.

Nos 4 meilleurs hits confirment bien que notre séquence code pour les lamines A et C (E value = 0, Query cover = 79%, similitude > 98%).

L'utilisation de cette base de données annotée et vérifiée par les chercheurs assure la fiabilité des résultats.

| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| RecName: Full=Prelamin-A/C; Contains: RecName: Full=Lamin-A/C; AltName: Full=70 kDa lamin; AltName: Fu... | Homo sapiens | 854 | 854 | 79% | 0.0 | 100.00% | 664 | P02545.1 |
| RecName: Full=Prelamin-A/C; Contains: RecName: Full=Lamin-A/C; Flags: Precursor [Sus scrofa] | Sus scrofa | 825 | 825 | 79% | 0.0 | 98.32% | 664 | Q3ZD69.1 |
| RecName: Full=Prelamin-A/C; Contains: RecName: Full=Lamin-A/C; Flags: Precursor [Mus musculus] | Mus musculus | 822 | 822 | 79% | 0.0 | 98.51% | 665 | P48678.2 |
| RecName: Full=Prelamin-A/C; Contains: RecName: Full=Lamin-A/C; Flags: Precursor [Rattus norvegicus] | Rattus norvegicus | 820 | 820 | 79% | 0.0 | 98.32% | 665 | P48679.1 |

*Figure 3 : Résultat Blastx*

Ces résultats confirment également que notre meilleur hit concerne les Homo Sapiens.

Les lamines A et C sont des protéines fibreuses impliquées dans la formation, la résistance et la stabilité de la membrane nucléaire. Elles ont également un rôle dans la régulation de l'expression génique et sont présentes dans le cytosquelette nucléaire.

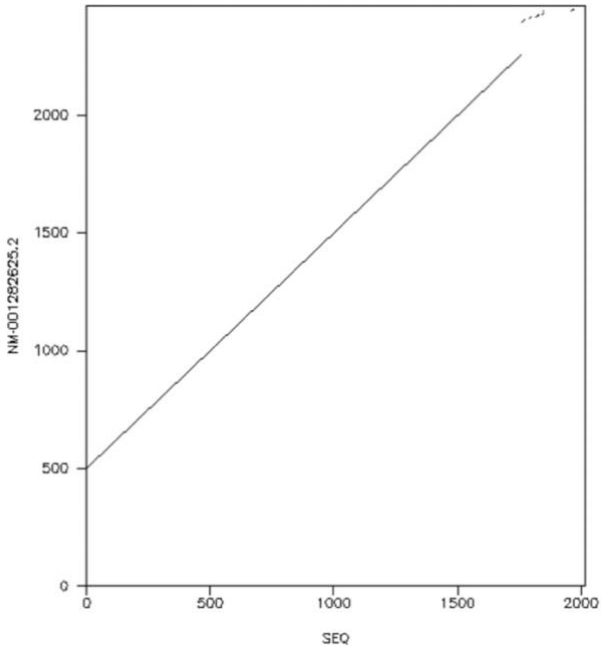
Nous allons maintenant localiser le gène sur le génome humain.

Pour étudier le génome de l'homme, on utilise le site UCSC genome browser. On recherche le gène LMNA et on apprend que ce dernier est localisé sur le locus 1q22 (bras long) du chromosome 1 et est situé sur le brin +. En amont, se trouve le gène MEX3A sur le brin-. En aval, SEMA4A sur le brin+. Les informations données initialement par le chercheur sont cohérentes.


*Figure 4 : Localisation du gène LMNA*

On apprend également que notre gène possède 42 transcrits dont 3 notés GOLD dans le cadre du projet HAVANA. Le variant 6 que nous utilisons possède 13 exons, donc 12 introns, et son TSS est situé en position 156.137.866.

Nous allons maintenant définir les particularités de notre séquence en la comparant aux séquences homologues trouvées dans les bases de données, puis nous allons déterminer sur quels exons de la séquence se situent les éventuelles mutations.




Afin d'analyser les différentes mutations on utilise l'outil Dotplot sur le serveur EMBOSS. Les décalages entre les diagonales correspondent à des insertions ou à des délétions.
On peut voir que la séquence commence vers le 500ème nucléotide de notre séquence query et on observe des cassures sur la courbe localisée vers la fin de notre séquence, représentatives de mutations en 3' de notre séquence ce qui confirme notre Blast

*Figure 5 : DotPlot*

Afin de définir ces mutations, on utilise Needle (outil d'alignement 2 à 2, serveur emboss), on peut voir que les mutations apparaissent au 1758ème nucléotide de notre séquence requête. (586ème acide aminé).

La séquence query s'aligne à partir du 499ème nucléotide de la séquence subject.

Nous avons 720 gaps et un score d'alignement de 8952 avec needle. Ce score est obtenu avec la matrice BLOSUM62 (gap_open=10, gap_extend=0.5, gap penalty=false)

En faisant un alignement de nos deux séquences protéiques on peut voir qu'il y a une délétion de 43 acides aminés par rapport à la séquence subject en position Cterm.

```
SEQ          1703 ACCTGGGGCTGCGGGAACAGCCTGCGTACGGCTCTCATCAACTCCACTGG  1752
                  ||||||||||||||||||||||||||||||||||||||||||||||||||
NM_001282625.  2201 ACCTGGGGCTGCGGGAACAGCCTGCGTACGGCTCTCATCAACTCCACTGG  2250

SEQ          1753 GGAAGTAACTAGGCCTGGGCCTGGCTGCTTGCTGGACGAGGCTCCCCCTG  1802
                  |||||.|..       ||||..|| ||..|||||.    ||.|.|..||
NM_001282625.  2251 GGAAGAAGT--------GGCCATGC-GCAAGCTGGT----GCGCTCAGTG  2287

SEQ          1803 A---------TGGCCAAC---------ATCGGAG----CCAGCTGCCCCC  1830
                  |         .||.|.||         ...||||    ||.||| ||..|
NM_001282625.  2288 ACTGTGGTTGAGGACGACGAGGATGAGGATGGAGATGACCTGCT-CCATC  2336

SEQ          1831 AAC-CCAAGT------TTGCCAATTCAGGGCCCCTTTCTAGAGCT--CTC  1871
                  |.| |||.||      |.|||....|.|.|.||       ||||.  |.|
NM_001282625.  2337 ACCACCACGTGAGTGGTAGCCGCCGCTGAGGCC--------GAGCCTGCAC  2379

SEQ          1872 TGTTGCAGGCTCCAGACTTCTCCACCCAGTAGGCAAACCAAAAGATGCTT  1921
                  ||..||   |.||...|    |...||.|..||| |.||..       |
NM_001282625.  2380 TGGGGC----CACCCAGC-----CAGGCCTGGGGGC-AGCCTC--------T  2413

SEQ          1922 CCTCAACAGCACAAGGGGTGGAAGTTAGACAGTGAGGATTGTTAAAGGCA  1971
                  ||.||.|..|.|...|......||..|...||.|.|.||.||||..   .
NM_001282625.  2414 CCCCAGCCTCCCCGTGCCAAAAATCTTTTCATTAAAGAATGTTTT----G  2459

SEQ          1972 GAGCCATACTCCTACCCGGAGAGCTTGACAGTGTCCCTCTGGGGTG   2017
                  ||.|..||
NM_001282625.  2460 GAACTTTA----------------------------------------  2467
```

*Figure 6 : Alignement séquence nucléotidiques*

```
SEQ          485 TSGRVAVEEVDEEGKFVRLRNKSNEDQSMGNWQIKRQNGDDPLLTYRFPP  534
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001   651 TSGRVAVEEVDEEGKFVRLRNKSNEDQSMGNWQIKRQNGDDPLLTYRFPP  700

SEQ          535 KFTLKAGQVVTIWAAGAGATHSPPTDLVWKAQNTWGCGNSLRTALINSTG  584
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001   701 KFTLKAGQVVTIWAAGAGATHSPPTDLVWKAQNTWGCGNSLRTALINSTG  750

SEQ          585 E-----------------------------------------------VTRPGP  591
                 |                                              .|:||.
EMBOSS_001   751 EEVAMRKLVRSVTVVEDDEDEDGDDLLHHHHVSGSRR*GRACTGATQPGL  800

SEQ          592 GCLLDEAPPDGQHRSQLPPTQVCQFRAPF*SSLLQAPDFSTQ*ANQKMLP  641
                 |.    |.|.      ||      ||   .:|.....|.|.
EMBOSS_001   801 GA----ASPQ-------PP-------RA---KNLFIKECFGTL--------  822

SEQ          642 QQHKGWKLDSEDC*RQSHTPTRRA*QCPSGV    672
                 ||||||||||||| ||||||||||||||||
EMBOSS_001   823 ----------------------------------  822
```

*Figure 7 : Alignement séquence protéique*

Nous cherchons maintenant à placer cette mutation précisément sur notre chromosome.

Pour cela nous utilisons le serveur UCSC browser qui permet d'étudier en détails la constitution du génome et en reprenant les résultats de nos alignements obtenus grâce à Needle, on observe que notre mutation commence à partir de l'acide glutamique, soit à la fin de notre 12ème exon.

Cela nous permet de déterminer que la mutation est présente sur le 12ème intron du variant qui correspond à l'intron 9 du gène LMNA

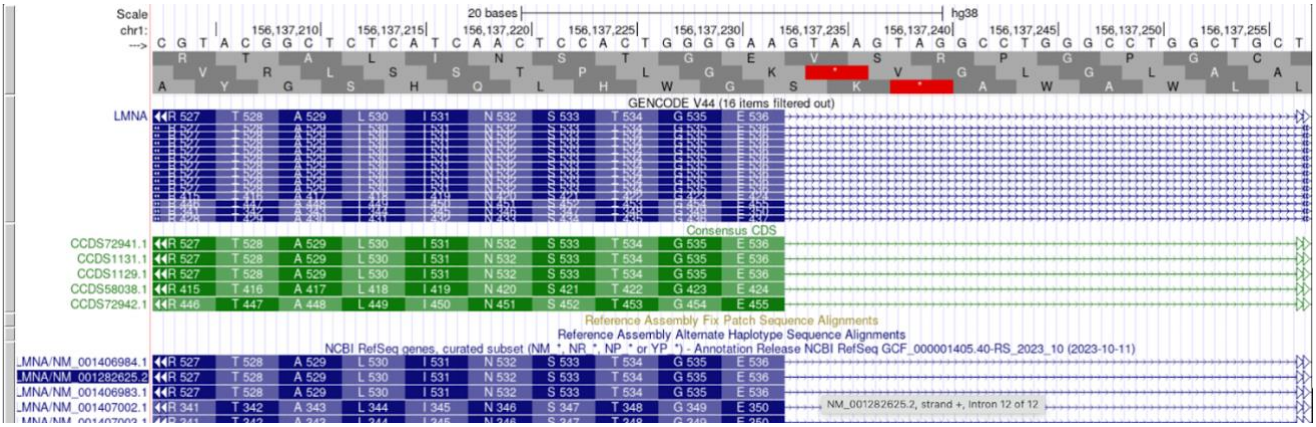



*Figure 8 : Localisation de notre mutation*

Maintenant que nous avons déterminé les régions mutées et qu'on les a placées sur notre séquence, on va chercher à savoir si notre gène pourrait être impliqué dans la maladie identifiée.

Puis nous chercherons à déterminer précisément l'origine et la nature de la mutation responsable de la maladie.

Tout d'abord, on observe que l'îlot CPG de notre gène n'est pas à proximité de notre mutation. Notre mutation n'a donc à priori pas de lien avec la région promotrice du gène.

On déplace le track OMIM pour la placer au-dessus du track NCBI RefSeq et en se plaçant au niveau de notre mutation on trouve le code OMIM correspondant précisément à notre mutation. Le track OMIM gêne nous confirme que le gène est bien impliqué dans la maladie d'Emery-Dreifuss comme annoncé par le chercheur.

Cette pathologie a une transmission autosomique dominante, ce qui implique que les membres de la famille d'un patient atteint sont souvent touchés également.

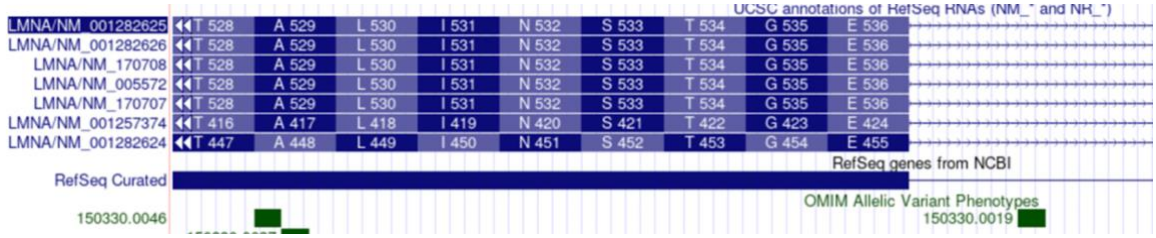



*Figure 9 : Code OMIM associé à notre mutation*

OMIM Allelic Variant 150330.0019 EMERY-DREIFUSS MUSCULAR DISTROPHY 2, AUTOSOMAL DOMINANT
OMIM 150.300: LAMIN A/C
AMINO ACID REPLACEMENT/ IVS9DS, G-C, +5
dbSNP/ClinVar: rs267607539

En additionnant ces informations à celles recueillies sur le serveur du NCBI ainsi que OMIM on apprend que notre mutation est un variant de transcription intronique court de type SNV (single nucléotide variant). Cette mutation correspond à la substitution d'une guanine par une cystéine en position 156.137.237 sur l'intron 9 du gène LMNA.

Une substitution G>A est également possible et peut potentiellement mener à une pathologie mais dans notre cas il s'agit d'une substitution G>C (forcément pathogène).

Cette mutation peut avoir différentes conséquences mais la plus importante est une variation de séquence entraînant un changement à la 5ème paire de base à partir du début de l'intron, ce qui va entraîner un défaut d'épissage.

Nous avons vu en cours de SEG que l'épissage alternatif est un procédé selon lequel, à partir d'un ARNm précurseur, il y a formation de plusieurs combinaisons d'exons différentes. L'épissage alternatif est un phénomène naturel permettant d'avoir une grande diversité du protéome. Ces combinaisons peuvent mener à des mutations de différents types, dans notre cas nous avons à faire à un phénomène de rétention d'intron qui entraîne l'apparition d'un codon stop prématuré (PTC) dans la suite de la séquence. Le processus d'épissage est régulé par différents facteurs. Dans le cas de la rétention d'intron, les éléments régulateurs sont appelés Intronique Splicing Silencers (ISS), ils interfèrent avec la reconnaissance et l'accessibilté des sites d'épissages en recrutant des protéines inhibitrices de l'épissage nommées hnRNP. Ce sont des protéines associées à l'ARN immature qui se fixent sur les silencers. Cela concorde avec l'exemple donné par MME PRESSE sur le gène LMNA en cours.

Nous allons maintenant définir précisément l'impact fonctionnel de cette mutation sur notre protéine.

Au niveau de notre séquence, un nucléotide G est substitué en C, ce qui change notre codon AGT en ACT. Ce codon sera ensuite traduit en UGA, un codon stop. La rétention d'intron liée à l'épissage alternatif implique la présence de cet intron 9 dans l'ARNm durant la traduction. La présence du codon stop prématuré induit un arrêt de traduction, le dernier exon n'est donc pas traduit ce qui entraine la formation d'une protéine tronquée dysfonctionnelle ou non-fonctionnelle.

Notre pathologie fait partie du groupe des laminopathies, ces maladies sont dues à des anomalies du gène LMNA, codant pour les lamines A et C. Les lamines sont des protéines d'union ayant un rôle dans l'ancrage des protéines membranaires au cytosquelette d'actine. Les mutations du gène LMNA induiraient des défauts structurels des lamines A/C. Or, les lamines sont les principales composantes de la lamina nucléaire, leur mutation engendre une perte de résistance mécanique des cellules musculaires donc une déstructuration. Cette dénaturation perturbe la communication entre le noyau et le cytoplasme. Tous ces éléments compromettent la fonction cellulaire normale.

L'étude de notre séquence nous a appris que la maladie d'Emery-Dreifuss est liée à une mutation du gène LMNA ce qui entraine la dégradation des cellules des muscles striés provoquant une dystophie musculaire.

https://www.omim.org/entry/181350