UNIVERSITÉ DE NANTES
UFR SCIENCES PHARMACEUTIQUES ET BIOLOGIQUES

ANNÉE 2015 **N° 054**

MÉMOIRE
DU DIPLÔME D'ÉTUDES SPÉCIALISÉES DE BIOLOGIE MEDICALE

Soutenu devant le jury interrégional

Le *18 Septembre 2015*

Par *Benjamin Cogné*

Conformément aux dispositions du Décret n° 2012-172 du 3 février

THÈSE
POUR LE DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE

Caractérisation des virus adéno-associés recombinants par séquençage haut débit : limites et perspectives

---

Président : **Mme Virginie FERRE, Professeur de Virologie**

Membres du jury : **M Philippe MOULLIER, Directeur de recherche**
M Adrien LEGER, Ingénieur de recherche
M Stéphane BEZIEAU, Professeur de Génétique
Mme Marianne COSTE-BUREL, Praticien hospitalier

Remerciements

A Philippe, merci d'avoir accepté d'accueillir un pharmacien dans tes murs pendant ces 2 ans et demi, j'ai beaucoup appris de tes conseils. Merci de m'avoir toujours soutenu, même quand les résultats n'étaient pas forcément ceux attendus, ce qui a été finalement ma plus belle expérience de la recherche…

Je remercie Madame Virginie FERRE, pour avoir accepté de juger et de présider cette thèse.

Je remercie Monsieur Stéphane BEZIEAU, pour avoir accepté de juger ce travail et de faire partie de mon jury, et pour m'avoir accueilli dans son laboratoire après ces travaux.

Je remercie Madame Marianne COSTE-BUREL, pour avoir accepté de juger ce travail et d'être membre de mon jury.

Je remercie Adrien, pour m'avoir proposé un projet de recherche intéressant et qui m'a appris bien plus que ce que le sujet pouvait laisser penser au début. Pour ton effet « geekisant » et l'initiation à linux, la bioinfo, inkscape… J'espère que les figures de cette thèse, faites sous inkscape, sont à la hauteur… Merci aussi d'avoir corrigé cette thèse avec autant de pertinence. Et puis, le travail ensemble a été plutôt fructueux, au plaisir de remettre ça peut-être un jour…

Je remercie aussi toute l'équipe de pharmaco : JB, Benoit, Emilie et les stagiaires qui ont ponctué ici et là le journal club d'histoires parfois surprenantes. Merci pour tous ces bons moments.

Je remercie Aurélien, Baptiste, Kizito, JB et toutes celles et ceux qui ont fait de ces petits verres en terrasse après le travail de bons moments de détente. Mention spéciale à Marine, pour ta bonne humeur au quotidien.

Et plus généralement, merci à toutes les personnes de la 1089 et de la 703 pour tous les agréables moments passés.

Merci à tous mes amis de Carquefou : Gaël, Jérèm, Xav, Jess, Romain, Marco, Alex, Brice, Benoit… c'est si important de pouvoir compter sur vous depuis le collège.

Merci à ma maman, j'ai toujours admiré ton courage, c'est grâce à toi que j'en suis là aujourd'hui.

Merci à ma petite sœur Marion, on est si différent mais pourtant si proche.

A Hélène, pour tout ce que tu m'apportes au quotidien, je ne m'imagine pas la suite sans toi.

A mon papa, parce que tout ça, c'est aussi pour toi.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

AAP	*Assembly Acting Protein*
AAV	*Adeno-associated virus*
AAVr	*Adeno-associated virus* recombinants
AAVS1	AAV site d'intégration 1
ADN	Acide désoxyribonucléique
AMM	Autorisation de mise sur le marché
ARN	Acide ribonucléique
ATM	*Ataxia telangiectasia mutated*
CHC	Carcinome hépato-cellulaire
CRISPR/Cas9	*Clustered regularly interspaced short palindromic repeat-associated nuclease Cas9*
DMD	Dystrophie musculaire de Duchenne
EMA	*European Medicine Agency*
FDA	*Food and drug administration*
GFP	*Green Fluorescent Protein*
HSPG	Protéoglycanes à héparanes sulfates
HSV	Herpes Simplex Virus
hTK	hygromycine thymidine kinase
ITR	*Inverse Terminal Repeats*
LAM-PCR	*Linear amplification mediated PCR*
LM-PCR	*Linker-mediated PCR*
LTR	*Long Terminal Repeats*
MLV	*Murine Leukemia Virus*
NGS	*Next-generation sequencing*
NHEJ	Recombinaison non-homologue
ORF	*Open Reading Frame*
pb	paires de base
PCR	*Polymerase Chain Reaction*
qPCR	PCR quantitative
RBE	*Rep Binding Element*
RH	Recombinaison homologue
RSV	*Rous Sarcoma Virus*
SI	Site d'intégration
SPRI	*Solid Phase Reversible Immobilization*
SSV-Seq	*Single Stranded Virus - Sequencing*
TA	*tibialis anterior*
TALEN	*Transcription activator–like effector nucleases*
TLR	*Toll-like receptors*
trs	*terminal resolution site*
VEGF	*Vascular Endothelial Growth Factor*
vg/kg	Génomes de vecteur par kilogramme
VHC	Virus de l'hépatite C
VIH	Virus de l'immunodéficience humaine
VP	Protéines de capside
ZFN	*Zinc Finger Nucleases*

Avant-Propos

Depuis longtemps espérée, parfois décriée, la thérapie génique s'inscrit aujourd'hui de plus en plus comme une piste thérapeutique crédible pour de nombreuses maladies. Dans quelques années, les médicaments de thérapie génique seront peut-être une alternative aux traitements pharmacologiques, pour lesquels le peu de nouvelles classes thérapeutiques récemment mises sur le marché illustre une innovation décroissante. En attendant, bien que certaines compagnies pharmaceutiques commencent à investir dans ces nouveaux traitements, de nombreux obstacles subsistent et leur développement reste encore majoritairement du domaine de la recherche académique. En effet, une meilleure connaissance de leurs mécanismes d'action et une amélioration de leur caractérisation est nécessaire pour garantir leur efficacité et leur innocuité. Alors que de nombreux essais cliniques apportent régulièrement de nouveaux éléments sur ces deux aspects, de nouvelles méthodes de séquençage de l'ADN sont apparues et se placent aujourd'hui comme un puissant outil pour améliorer la caractérisation des médicaments de thérapie génique. Ce travail de thèse s'inscrit dans cette volonté d'améliorer notre connaissance de ces médicaments innovants en profitant des possibilités offertes par le séquençage de nouvelle génération.

I. Introduction

I.1 La thérapie génique

I.1.1 Solutions thérapeutiques pour les maladies génétiques

Une maladie génétique est la conséquence de l'altération de l'expression d'un ou plusieurs gènes causée par la présence de mutations génétiques, entrainant un défaut de fonctionnement de certaines cellules de l'organisme. Il faut différencier les maladies génétiques monogéniques, dont l'absence d'expression d'un seul gène est directement responsable de l'apparition des symptômes, des maladies multifactorielles à prédisposition génétique comme le cancer ou les maladies auto-immunes. Si pour ces dernières des formes héréditaires d'origine génétique ont été identifiées (ex : mutation de BRCA1/2 dans les cancers du sein et de l'ovaire), des facteurs environnementaux jouent souvent un rôle important dans le processus pathologique. Cette différence physiopathologique implique des différences fondamentales dans l'approche diagnostique et thérapeutique.

Parmi les 25000 gènes annotés du génome humain, des mutations dans plus de 3000 ont été liées à des phénotypes pathologiques (www.omim.org/statistics/geneMap). Concernant les maladies génétiques monogéniques, il en existerait plus de 6000 différentes. Elles sont rares individuellement mais combinées elles toucheraient une naissance sur 100. L'atteinte tissulaire d'une maladie génétique est guidée par l'importance fonctionnelle du gène considéré dans les différents types cellulaires. Ainsi, nous pouvons par exemple distinguer les pathologies neuromusculaires, regroupant des atteintes des motoneurones (ex : Amyotrophie spinale) et des muscles (ex : Myopathie de Duchenne), les pathologies rétiniennes (ex : Amaurose congénitale de Leber) ou encore les pathologies systémiques (métaboliques, …).

L'identification des protéines responsables des phénotypes pathologiques a permis le développement de traitements par enzymothérapie de substitutions pour certaines maladies monogéniques (hémophilies, maladie de Pompe, maladie de Gaucher, maladie de Fabry...). Cette approche thérapeutique est cependant lourde car elle nécessite des administrations répétées d'une enzyme recombinante par voie parentérale tout au long de la vie. De plus, les coûts de ces traitements à long terme sont très élevés et l'apport exogène de la protéine déficiente ne corrige pas l'anomalie génétique sous-jacente, ce qui aboutit à une efficacité thérapeutique souvent modeste (Lim et al., 2014). Pour les nombreuses maladies génétiques non éligibles à ce type d'approche, la prise en charge actuelle par traitement pharmacologique

consiste à améliorer la qualité de vie des patients et à retarder l'apparition des complications les plus graves.

Dans ce contexte, la correction/compensation de la déficience au niveau génétique se place comme une piste thérapeutique attirante qui permettrait un traitement à long terme après une unique intervention.

I.1.2 Historique de la thérapie génique

La thérapie génique est une stratégie thérapeutique qui consiste à traiter des maladies par transfert d'acides nucléiques (ARN ou ADN) dans les cellules ou les tissus d'un individu. Pour les maladies génétiques, elle consiste plus précisément à permettre la réexpression d'une protéine fonctionnelle. La première observation de transfert de gène a eu lieu en 1952 avec la transmission d'une résistance aux antibiotiques entre deux souches de Salmonelles, *via* un bactériophage (virus procaryote). Le terme de transduction était né. En 1961, une observation similaire a été réalisée sur des cellules eucaryotes suite à leur transduction par un RSV (*Rous Sarcoma Virus*) et l'intégration stable des gènes viraux. Enfin, c'est lors d'un symposium sur la médecine du futur en 1966, que pour la première fois, Edward Tatum évoqua la possibilité d'utiliser des virus pour modifier des cellules somatiques (Tatum, 1966). Cependant, les moyens technologiques n'étaient pas encore disponibles et il fallut attendre l'amélioration des techniques de biologie moléculaire permettant la manipulation et l'amplification de l'ADN (les enzymes de restriction, le clonage moléculaire et la PCR) pour voir émerger les premières expériences de transfert d'un transgène thérapeutique. Finalement, c'est en 1990 que le premier essai clinique de thérapie génique a été approuvé aux Etats-Unis pour le traitement du déficit en adénosine désaminase chez deux enfants.

Aujourd'hui, plus de 2000 essais cliniques de thérapie génique ont été réalisés (http://www.abedia.com/wiley/years.php). La majorité de ces essais concernent le traitement du cancer (**Figure 1**), largement devant les maladies monogéniques, infectieuses et cardio-vasculaires. Cela reflète le plus grand nombre de patients en cancérologie, permettant une constitution plus facile des cohortes nécessaires pour les essais cliniques. De plus, il est évident que l'impact d'une thérapeutique oncologique en termes de santé publique sera beaucoup plus important, ce qui a poussé l'investissement dans ce domaine. Ainsi, le premier traitement de thérapie génique, approuvé en 2003 par la Chine, était un Adénovirus portant le gène p53 (Gendicine) indiqué dans les cancers tête et cou. Cette autorisation, accordée sans résultat de phase III démontrant son efficacité a été vivement critiquée et encore aujourd'hui, aucune preuve de son efficacité n'existe.

En parallèle, les premiers « vrais » succès de la thérapie génique ont été obtenus dans le domaine des maladies génétiques monogéniques. En 2000, une équipe parisienne a traité des enfants atteints de déficit immunitaire combiné sévère lié à l'X (DICS-X) (OMIM #300400) avec un vecteur rétroviral codant pour la chaine gamma du récepteur aux chimiokines aboutissant à la restauration d'un système immunitaire fonctionnel (Cavazzana-Calvo et al., 2000), malgré de lourds effets indésirables qui seront discutés dans la partie I.1.5.2. Ce succès a eu un écho médiatique important et a depuis été suivi par de nombreux autres essais cliniques. Un palier important a été franchi en 2012 avec l'autorisation de mise sur le marché d'un vecteur viral indiqué dans le traitement du déficit en lipoprotéine lipase (Glybera, UniQure) par l'agence européenne des médicaments (Bryant et al., 2013). C'était une étape cruciale vers la considération de la thérapie génique comme réelle solution thérapeutique. Cependant, la longue procédure d'obtention de l'AMM et les critiques émises par les autorités réglementaires ont également mis en lumière les nombreuses étapes qu'il reste à franchir avec ces nouveaux médicaments notamment en termes d'efficacité et de biosécurité.

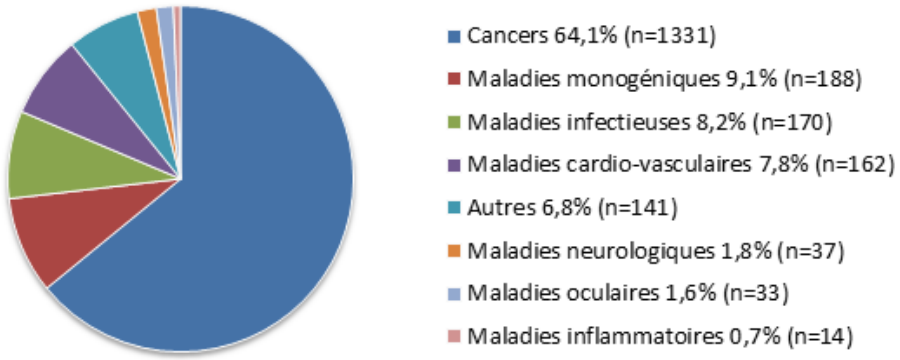


Figure 1 : Indications des essais cliniques de thérapie génique réalisés jusqu'à présent. Source : http://www.abedia.com/wiley/indications.php

I.1.3 Méthodes de vectorisation

De nombreuses barrières biologiques se dressent devant l'acheminement et l'expression intracellulaire d'un acide nucléique exogène (**Figure 2**). Plusieurs types de vecteurs de transfert ont donc été développés pour contourner certains de ces obstacles. Cependant, aucun ne les maitrise tous à la perfection et le choix du vecteur, qu'il soit dérivé de virus ou non, dépendra de l'objectif à atteindre (**Figure 3**).

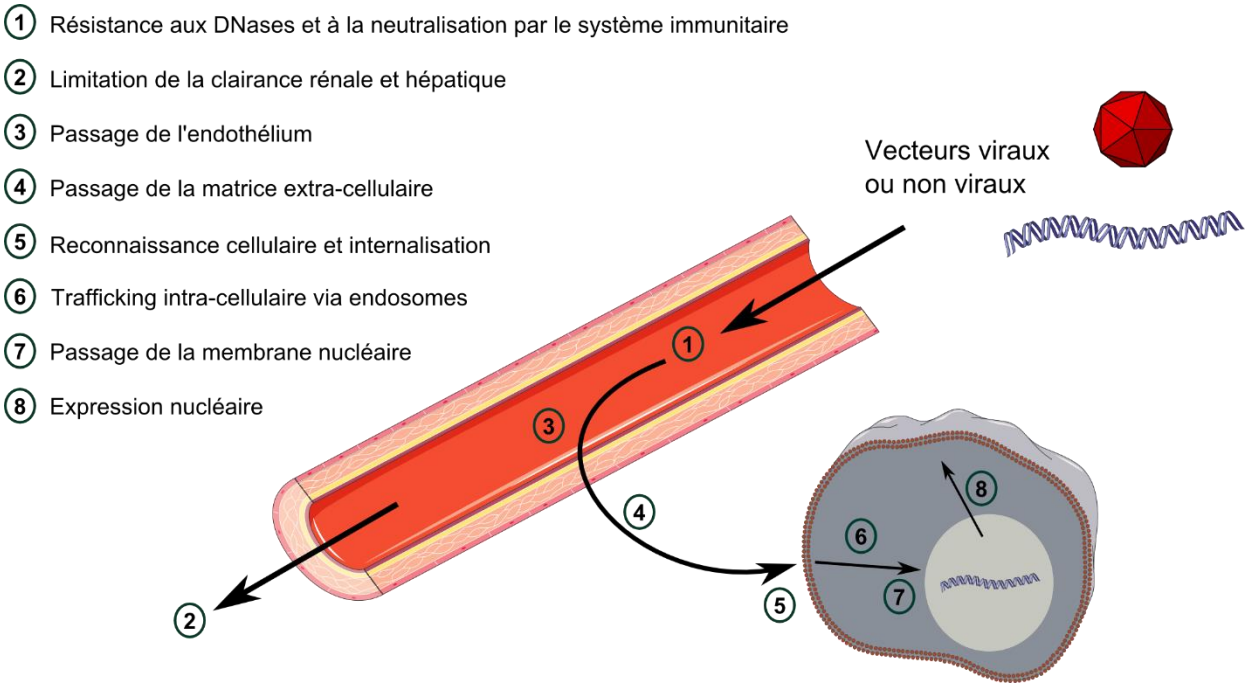


Figure 2 : Principaux obstacles rencontrés par un vecteur après injection systémique. Chaque étape est susceptible de réduire l'efficacité du transfert de gène. Le mode d'injection et le type de vecteur utilisé peut permettre de limiter l'impact de certaines de ces étapes.

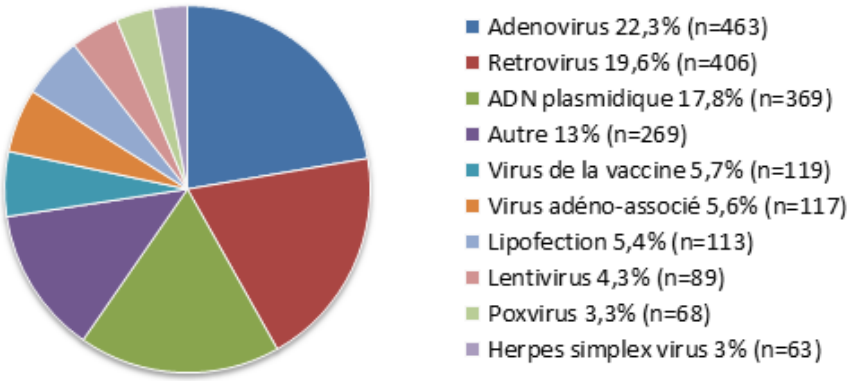


Figure 3 : Vecteurs utilisés dans les essais cliniques de thérapie génique. http://www.abedia.com/wiley/vectors.php.

- **Vectorisation physique et synthétique**

L'injection de plasmides nus permet d'obtenir l'expression d'un transgène, comme montré par l'expression de béta-glucosidase dans le muscle chez la souris après injection intramusculaire (Wolff et al., 1990). Mais le faible niveau d'expression obtenu n'est pas compatible avec une application thérapeutique. Des **méthodes physiques** ont donc été développées pour forcer l'introduction d'ADN dans le compartiment intracellulaire (Mehier-Humbert and Guy, 2005): l'injection sous pression de microparticules assemblées avec l'ADN (« gene gun »), l'injection hydrodynamique en augmentant le volume injecté, la perméabilisation membranaire à l'aide d'un courant électrique (électroporation) ou d'ultrasons

(sonoporation) ou encore l'association de l'ADN avec une nanoparticule magnétique et application d'un champs magnétique (magnétofection). L'électroporation permet l'expression la plus efficace (Heller and Heller, 2010), mais elle est peu applicable *in vivo* chez l'homme sur la totalité du corps et est relativement traumatique. Le manque de reproductibilité de ces méthodes physiques et l'expression localisée à faible niveau du transgène a nécessité de développer des vecteurs plus efficaces.

Les **vecteurs synthétiques** consistent à assembler l'acide nucléique avec des lipides (lipoplexes) ou des polymères (polyplexes). Cette structure protège l'acide nucléique de la dégradation par les nucléases circulantes et permet de neutraliser les charges négatives du squelette phosphate et ainsi de favoriser sa diffusion. La lipofection semble plus efficace, car la fusion des lipides cationiques avec l'endosome favoriserait l'accessibilité nucléaire (Yin et al., 2014). L'expression à court terme et à faible niveau obtenue avec ces vecteurs limite leur utilisation à la cancérologie, aux maladies cardio-vasculaires et à la vaccination génétique. Certaines de ces applications bénéficient de cette expression transitoire, comme dans l'infarctus du myocarde, où l'expression de VEGF (*Vascular endothelial growth factor*) par transfert d'ARNm permet d'améliorer la vascularisation de la zone infarcie en évitant les complications d'une expression à long terme (Zangi et al., 2013). Une autre approche, qui est aujourd'hui très répandue, y compris pour le traitement des maladies monogéniques, est le transfert de petits ADN ou ARN modifiés chimiquement appelés oligonucléotides antisens qui vont interagir avec l'ARN pré-messager ou messager (Meng et al., 2015). Leurs principaux avantages, par rapport aux vecteurs viraux, sont une faible immunogénicité, qui permet une réinjection régulière, une facilité de production et une absence de génotoxicité. Ils ne sont pas pour autant dénués de toxicité, puisque lorsque de fortes doses doivent être injectées, des atteintes rénales ont été observées (Voit et al., 2014).

- **Vectorisation virale**

Les virus sont naturellement incapables de se multiplier de façon autonome, ils sont dépendant de leur capacité à infecter une cellule hôte et se sont ainsi spécialisés dans le transfert de gène. La connaissance approfondie de certains de ces virus a permis d'en dériver des vecteurs de transfert d'acide nucléique qui conservent les capacités d'expression du matériel génétique propres au virus dont ils sont issus, tout en étant dépourvus des gènes impliqués dans la pathogénicité et la réplication. Les vecteurs viraux sont aujourd'hui les plus utilisés en thérapie génique, largement devant les approches physiques et synthétiques. Les principaux sont dérivés des adénovirus, des rétro/lentivirus, des virus adéno-associés (AAV) et des virus de l'herpès (HSV) (**Tableau 1**).

Vecteurs	Matériel génétique	Capacité d'encapsidation	Persistance	Limitations	Avantages	Applications
Gamma-rétrovirus	ARN	8kb	Intégré	Transduit uniquement cellules en division. Fort pouvoir oncogénique	Transfert de gène persistant	Maladies monogéniques ex vivo, cancérologie
Lentivirus	ARN	8kb	Intégré	Risque oncogénique	Transfert de gène persistant dans de nombreux tissus	Maladies monogéniques *ex vivo,* cancérologie
AAV	ADNsb	<5kb	Episomale > 99,9%	Faible capacité d'encapsidation	Large tropisme, peu pathogène, expression à long terme dans cellules quiescentes	Maladies monogéniques in *vivo*
Adénovirus	ADNdb	30kb	Episomale	Provoque une forte réaction immunitaire, expression transitoire	Transduction efficace de nombreux tissus	Cancérologie, maladies cardio-vasculaires
HSV-1	ADNdb	40kb	Episomale	Forte réaction inflammatoire, expression transitoire hors neurones	Grande capacité d'encapsidation, forte transduction neuronale	Cancérologie

Tableau 1: Caractéristiques des principaux vecteurs viraux utilisés en thérapie génique. ADNsb : ADN simple brin, ADNdb : ADN double brin, kb : kilobases. Adapté de Thomas et al. 2004

Une approche séduisante en **cancérologie**, et pourvoyeuse d'un grand nombre d'essais cliniques de thérapie génique, est le développement de virus modifiés pour se répliquer sélectivement dans les cellules cancéreuses et les lyser (virus oncolytiques). Les principaux utilisés sont les adénovirus (Heise et al., 2000), le HSV-1 et le virus de la vaccine (Poxvirus) (Zeh et al., 2015). Une monothérapie est cependant rarement efficace contre le cancer et la thérapie génique ne déroge pas à la règle. La stimulation concomitante du système immunitaire est actuellement étudiée pour améliorer l'efficacité anticancéreuse (Farzad et al., 2014).

En **infectiologie**, la capacité des adénovirus à transduire des cellules présentatrices d'antigène et à stimuler le système immunitaire leur permet d'être utilisés comme vaccins (Appaiahgari and Vrati, 2015). Par ailleurs, la vectorisation directe d'un gène codant pour un anticorps par administration intra-nasale d'un vecteur AAV recombinant (AAVr) a également permis d'obtenir une réponse contre le virus de la grippe (Limberis et al., 2013)**.**

Le traitement de certaines **maladies monogéniques** nécessite une expression à long terme d'un transgène dans des cellules en division. Dans cette optique, les vecteurs issus des gamma-rétrovirus et des lentivirus représentés respectivement par le virus de la leucémie murine (MLV) et le virus de l'immunodéficience 1 (VIH1) sont les plus appropriés. En effet, ces vecteurs intègrent leur matériel génétique directement dans le génome cellulaire, permettant ainsi de conserver le transgène malgré les divisions cellulaires. Ils sont utilisés en thérapie génique dite ***ex vivo*** qui consiste à successivement prélever des cellules de patients, les transduire *in vitro* avant de les réimplanter chez le même patient (autogreffe après modification génique). Cette approche, largement utilisée sur les cellules souches hématopoïétiques, a permis d'obtenir des résultats cliniques très probants dans le traitement du déficit immunitaire combiné sévère lié à

l'X (Cavazzana-Calvo et al., 2000), du déficit en adénosine désaminase (OMIM #102700) (Aiuti et al., 2002), du syndrome de Wiskott-Aldrich (OMIM #301000) (Boztug et al., 2010), de l'adrénoleucodystrophie lié à l'X (OMIM #300100) (Cartier et al., 2009) et de la leucodystrophie métachromatique (OMIM #250100) (Biffi et al., 2013). L'expérience de ces études cliniques montre que la persistance des cellules souches transduites dépend de l'avantage sélectif induit par le transfert de gène. Pour certaines pathologies, comme la granulomatose septique chronique (OMIM #306400) (Grez et al., 2011), en l'absence de conditionnement, l'avantage sélectif est insuffisant pour maintenir un niveau d'expression thérapeutique. De plus, derrière ces succès, le principal inconvénient de ces vecteurs est leur pouvoir oncogénique, qui sera discuté par la suite.

De nombreuses maladies monogéniques ne sont pas éligibles aux approches *ex vivo* et nécessitent une thérapie génique ***in vivo*** qui consiste à injecter le vecteur viral directement dans l'organisme par voie locale, locorégionale ou systémique. Les vecteurs AAVr sont alors les plus pertinents au vu des récents succès cliniques obtenus dans l'hémophilie B (OMIM #306900) (Nathwani et al., 2011), l'amaurose congénitale de Leber (OMIM #204100) (Maguire et al., 2008) et la choroidérémie (OMIM #303100) (MacLaren et al., 2014). Ils seront discutés plus en détail dans la partie qui leur est consacrée (I.2).

I.1.4 Stratégies thérapeutiques

L'évolution rapide des connaissances en biologie cellulaire et moléculaire a permis le développement de différentes stratégies pour corriger ou compenser l'anomalie génétique au niveau moléculaire (**Figure 4**).

- **Apport d'un transgène thérapeutique**

La première approche thérapeutique développée en thérapie génique, qui a permis ses premiers succès avec les vecteurs rétroviraux ou AAVr, est l'apport d'un transgène codant pour une protéine fonctionnelle. Le transgène est souvent une copie saine d'un gène déficient, notamment dans le cas des maladies monogéniques. De façon moins fréquente, il peut également coder pour une protéine activant des voies annexes (ex : follistatine dans la myopathie de Duchenne), ayant montrées une amélioration phénotypique. Ces approches sont néanmoins inapplicables aux maladies autosomiques dominantes et présentent certaines limites : (i) le gène déficient peut être d'une taille trop importante pour permettre sa vectorisation, (ii) la copie artificielle du gène n'est pas soumise à la même régulation transcriptionnelle que le gène d'origine, (iii) il est difficile d'obtenir un niveau d'expression thérapeutique du transgène, et (iv) une persistance d'expression à long terme reste difficile à

obtenir. Si cette approche reste majoritaire, des stratégies complémentaires ont été développées pour palier à ses limitations.

- **Stratégies d'édition de l'ARN**

La connaissance des différents mécanismes physiologiques permettant la synthèse, la maturation et la régulation des ARN a permis le développement de stratégies thérapeutiques par action sur l'ARN pré-messager et messager.

Au niveau de l'**ARN pré-messager**, il s'agit principalement d'interférer avec le processus d'épissage. Cette stratégie du « saut d'exon » est notamment à l'étude dans le traitement de la myopathie de Duchenne et consiste à restaurer un cadre ouvert de lecture (ORF) fonctionnel sur l'ARNm. Le masquage d'un site d'épissage permet la délétion d'un ou plusieurs exons responsables du décalage de l'ORF et de l'introduction d'un codon stop. Ce masquage s'effectue soit par le transfert de petits oligonucléotides antisens modifiés chimiquement, soit par la vectorisation d'un U7 snRNA (*small nuclear RNA*) modifié par un AAVr (Le Guiner et al., 2014).

La régulation post-transcriptionnelle par l'intermédiaire de micro-ARN (ou miRNA) a été décrite en 1993 (Lee et al., 1993). Les miRNA agissent en s'hybridant sur l'**ARN messager mature**, dont ils induisent le clivage ou l'inhibition de la traduction par l'intermédiaire d'un important complexe protéique. L'utilisation thérapeutique de ces ARN interférants est actuellement étudiée (Borel et al., 2014) par vectorisation virale de précurseurs nucléaires de miRNA ou par transfert de siRNA (*small interfering RNA*) qui correspondent à de petits ADN double brins directement actifs dans le cytoplasme. Ils sont notamment à l'étude en infectiologie (ex : clivage de l'ARN viral du VIH ou du VHC), en cancérologie et pour cliver spécifiquement l'allèle pathogène dans les maladies monogéniques dominantes (ex : Maladie de Huntington). Ils présentent néanmoins une certaine toxicité lorsqu'ils sont exprimés à forte dose, en saturant les voies de maturation des miRNA endogènes (van Gestel et al., 2014).

- **Stratégies d'édition du génome**

Lorsqu'une cassure double brin de l'ADN intervient dans le génome, sa réparation est effectuée soit par une voie permettant une réparation précise par utilisation de l'information présente sur la chromatine sœur (**recombinaison homologue**), soit par une voie moins fiable qui lie les deux extrémités libres en entrainant fréquemment des insertions/délétions (**recombinaison non-homologue**). L'induction de telles cassures double brin à des endroits précis du génome est aujourd'hui possible par la synthèse à façon de nucléases tels que les ZFN (*Zinc finger nucleases*), les TALEN (*Transcription activator–like effector nucleases*) et les CRISPR/Cas9 (*Clustered regularly interspaced short palindromic repeat-associated nuclease*

Cas9). D'un point de vue thérapeutique, ces nucléases peuvent être synthétisées pour cliver spécifiquement un gène muté, et lorsqu'une matrice ADN possédant des régions d'homologies avec les séquences de part et d'autre de la cassure est présente, la matrice peut être incorporée par recombinaison homologue et ainsi corriger ce gène de façon permanente. Une preuve de concept thérapeutique a été apportée en 2011 chez des souris modèles d'hémophilie B dans lesquelles la réparation de la mutation génétique a permis de restaurer partiellement la synthèse de facteur IX (Li et al., 2011a). En l'absence de matrice de recombinaison, la recombinaison non homologue aboutit fréquemment à l'inhibition du gène en induisant un décalage de l'ORF. Le premier essai clinique basé sur ce principe a consisté, en transduisant *ex vivo* des cellules souches hématopoïétiques avec un vecteur adénoviral codant pour des ZFN, à rendre non fonctionnel le gène codant pour le co-récepteur CCR5 du VIH (Tebas et al., 2014).

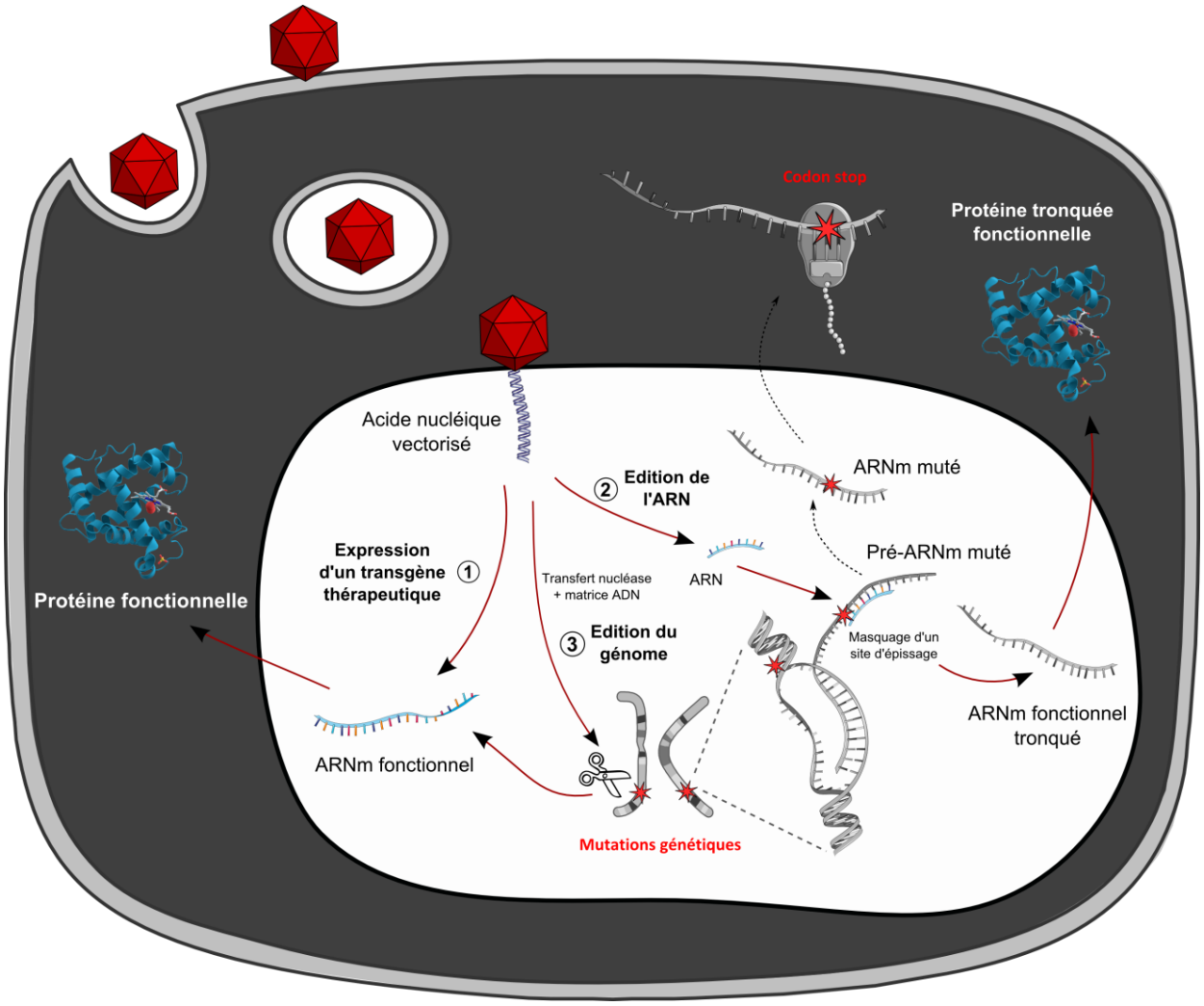


Figure 4 : Schéma des principales stratégies thérapeutiques utilisées en thérapie génique pour traiter une maladie monogénique. 1) La vectorisation d'un gène thérapeutique permet l'expression directe d'une protéine fonctionnelle. Une persistance à long terme du vecteur au niveau nucléaire est requise pour maintenir le bénéfice thérapeutique. **2)** Le masquage d'un site d'épissage par hybridation d'un ARN non codant permet la délétion d'un exon au niveau de l'ARNm et la restauration de l'expression d'une protéine fonctionnelle. **3)** La coupure du génome au niveau du gène muté par vectorisation de nucléases permet, en apportant une matrice ADN sans la mutation, de corriger définitivement l'expression de la protéine.

I.1.5 Effets indésirables des vecteurs viraux

L'introduction de matériel génétique exogène dans notre génome *via* des particules virales est responsable d'effets indésirables spécifiques, qui n'ont jamais été rencontrés par les traitements pharmacologiques classiques. Il est donc nécessaire de les caractériser et de les anticiper avant d'administrer ces nouvelles thérapies aux patients. Cependant, les modèles animaux utilisés pendant les études de toxicité ont échoué dans l'anticipation des deux principaux effets indésirables observés en thérapie génique: la génotoxicité des vecteurs rétroviraux et l'immunotoxicité des vecteurs adénoviraux. Ainsi, l'enjeu de la thérapie génique est aujourd'hui de réduire la toxicité tout en assurant un transfert de gène efficace : la marge thérapeutique apparait parfois étroite (**Figure 5**).

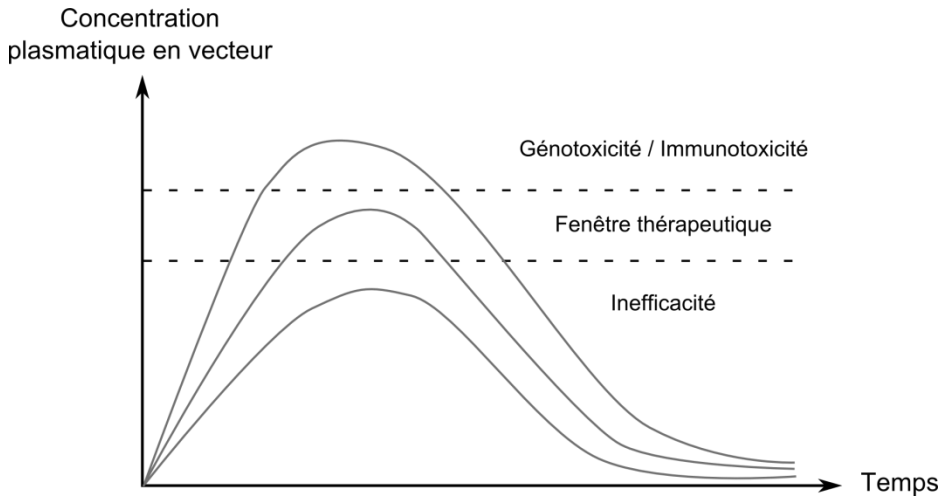


Figure 5 : Illustration de la marge thérapeutique étroite des vecteurs viraux. Ce schéma s'applique principalement pour la thérapie génique *in vivo.* Les limites de la fenêtre thérapeutique ne sont pas encore bien définies, elles peuvent varier selon les vecteurs et les patients. Il est possible que pour certaines pathologies nécessitant une forte dose de vecteur, les seuils inférieurs et supérieurs de la fenêtre thérapeutique se confondent.

I.1.5.1 Immunotoxicité

La réponse immunitaire est un obstacle majeur en thérapie génique *in vivo* utilisant des vecteurs viraux non enveloppés. En effet, les antigènes présents à la surface des capsides des vecteurs AAVr et adénoviraux sont très immunogènes et activent à la fois l'immunité innée et l'immunité adaptative (humorale et cellulaire).

Après injection systémique, les vecteurs adénoviraux sont reconnus par des récepteurs de l'**immunité innée** (TLR9) et transduisent efficacement les cellules dendritiques (Zhu et al., 2007)**.** Ainsi, lorsqu'ils sont injectés en grande quantité, la libération immédiate de cytokines pro-inflammatoires peut avoir des conséquences cliniques dramatiques. Cette toxicité a été révélée en 1999 lors d'un essai clinique visant à traiter le déficit en Ornithine transcarbamylase (Raper et al., 2003) dans lequel un jeune homme de 18 ans est décédé d'un syndrome de réponse

inflammatoire systémique 98 heures après l'injection systémique d'un vecteur adénoviral. Cet évènement a révélé l'extrême hétérogénéité du système immunitaire avec à la fois une barrière d'espèce, les études pré-cliniques menées chez la souris et le primate n'avaient pas prédit cette toxicité, et une variabilité inter-individuelle puisqu'un autre patient ayant reçu la même dose a bien toléré le traitement. Les vecteurs AAVr, bien que capables de transduire les cellules dendritiques à faible niveau (Gernoux et al., 2015), n'ont jamais déclenchés de telles réactions dans les essais clinique réalisés jusqu'à présent.

La production d'anticorps neutralisant est naturellement destinée à éliminer les particules virales lors d'une infection. Ainsi, lorsque les patients ont préalablement été en contact avec un virus sauvage similaire à celui utilisé pour la vectorisation ou lorsqu'ils ont déjà été traités une première fois par ce même vecteur, l'**immunité humorale** pré-existante peut neutraliser le vecteur et réduire drastiquement son efficacité thérapeutique. Ce phénomène est d'autant plus important que le virus est largement répandu dans la population, comme c'est le cas pour les adénovirus et les AAV (Calcedo et al., 2009). La découverte de nouveaux sérotypes et la génération de capsides modifiées permettent de limiter la neutralisation par les anticorps circulants et sont ainsi l'enjeu d'un développement actif en thérapie génique.

Un essai clinique utilisant des vecteurs AAVr et destiné à traiter l'hémophilie B a mis en évidence une perte d'expression du transgène suite à la lyse des hépatocytes transduits par des **lymphocytes T cytotoxiques** (Manno et al., 2006). Cette réponse était dirigée contre des antigènes de capside de l'AAVr et a, par la suite, pu être limitée par l'utilisation de glucocorticoïdes dès l'augmentation des transaminases (Nathwani et al., 2011). La réponse cellulaire anti-capside a depuis été montrée comme dose-dépendante (Nathwani et al., 2014). Une première étape serait donc de diminuer les doses d'AAVr injectées mais l'obtention d'un effet thérapeutique est souvent conditionnée par l'injection de fortes doses. La question est donc maintenant de savoir si la corticothérapie doit être instaurée de façon systématique chez tous les patients ou uniquement au moment de la détection d'une cytolyse hépatique, avec le risque de diminuer l'efficacité thérapeutique par la destruction des hépatocytes transduits. Par ailleurs, une réponse cellulaire dirigée contre des nouveaux épitopes apportés par le transgène peut également limiter l'efficacité thérapeutique (Mendell et al., 2010).

I.1.5.2 Génotoxicité

Comme introduit précédemment, l'un des premier succès en thérapie génique fut le traitement d'enfants atteints de DICS-X dont les résultats furent publiés en 2000 (Cavazzana-Calvo et al., 2000). Cependant, 3 ans après, 5 des 20 enfants traités conjointement en France et au Royaume-Uni ont développé une leucémie aiguë lymphoblastique T, entrainant 1 décès. Le

même type de cancer a depuis été détecté dans d'autres essais cliniques, visant à traiter différentes maladies monogéniques (**Tableau 2**) et utilisant les mêmes vecteurs dérivés du virus de la leucémie murine (MLV). Jusqu'à ces observations, l'utilisation de vecteurs viraux intégratifs non-réplicatifs n'était pas considérée comme cancérigène, ce risque ayant été mis en évidence tardivement chez les modèles animaux (Li et al., 2002). Depuis, la génotoxicité observée dans ces études a été reliée à la trans-activation d'oncogènes comme LMO2, un facteur de transcription hématopoïétique, par l'activité *enhancer* des extrémités du génome viral ou LTR (long terminal repeat). En effet, l'activation de l'oncogène mène à un état prolifératif pré-leucémique qui prédispose à l'accumulation d'autres anomalies génétiques comme une sur-expression d'autres oncogènes et une sous-expression d'anti-oncogènes, respectivement Notch1 et CDKN2A pour Howe et al. (Howe et al., 2008), entrainant secondairement une leucémisation. De façon surprenante, aucun des patients traités pour un déficit en adénosine désaminase n'a développé de leucémie. Bien qu'aucune explication définitive n'ait été apportée jusqu'à présent, cela souligne l'importance du transgène et de la pathologie sous-jacente.

Le risque oncogénique associé à un vecteur viral est directement lié à sa probabilité d'intégration près d'oncogènes et de séquences de régulation transcriptionnelle. Les gamma-rétrovirus ont une forte préférence d'intégration pour les marques épigénétiques actives associées aux unités transcriptionnelles (**Figure 6**) (LaFave et al., 2014). Ainsi, le risque génotoxique est maximal lorsqu'une séquence *enhancer* présente sur ces vecteurs est capable de stimuler la transcription d'un gène voisin du site d'intégration. Les lentivirus ont un tropisme préférentiel pour les gènes actifs, mais s'intègrent dans les régions intragéniques plutôt qu'au niveau des séquences régulatrices (Cattoglio et al., 2010). Ce profil d'intégration réduit leur potentiel génotoxique et jusqu'à présent, aucun évènement génotoxique n'a été rapporté avec les vecteurs lentiviraux. Parallèlement, les profils de biosécurité des vecteurs lentiviraux et gamma-rétroviraux ont été grandement améliorés en supprimant les séquences *enhancers* des LTR (ΔU3 LTR), responsables d'une importante trans-activation (Montini et al., 2006).

Pathologie	Essai clinique	Patients inclus	Cancers développés	Décès	Oncogènes impliqués
Déficit immunitaire combiné sévère lié à l'X	Hacein-Bey-Abina et al. 2008	10	4 LAL-T	1	3 LMO2, 1 CCND2, 1(LMO2+BMI1)
	Gaspar et al. 2011	10	1 LAL-T	0	LMO2
Wiskott Aldrich	Braun et al. 2014	10	6 LAL-T, 1 LAM	2	LMO2 (LAL-T), MDS1 et MN1 (LAM)
Granulomatose septique chronique	Stein et al. 2010	2	1 Myélodysplasie	1	EVI1
	Kang et al. 2011	2	0	0	MDS1-EVI1, PRDM16, CCND2 sans clonalité
Déficit en adénosine désaminase	Aiuti et al. 2009	10	0	0	/
	Gaspar et al. 2011	6	0	0	/

Tableau 2 : Evènements génotoxiques observés lors d'essais cliniques utilisant un vecteur gamma-rétroviral. LAL-T : leucémie aiguë lymphoblastique T, LAM : leucémie aiguë myéloblastique, LMO2 : LIM domain-only 2, CCND2 : cyclin D2, MDS1 : myelodysplasia syndrome 1, EVI1 : ecotropic viral integration site 1, MN1 : meningioma 1, PRDM16 : PR domain-containing protein 16

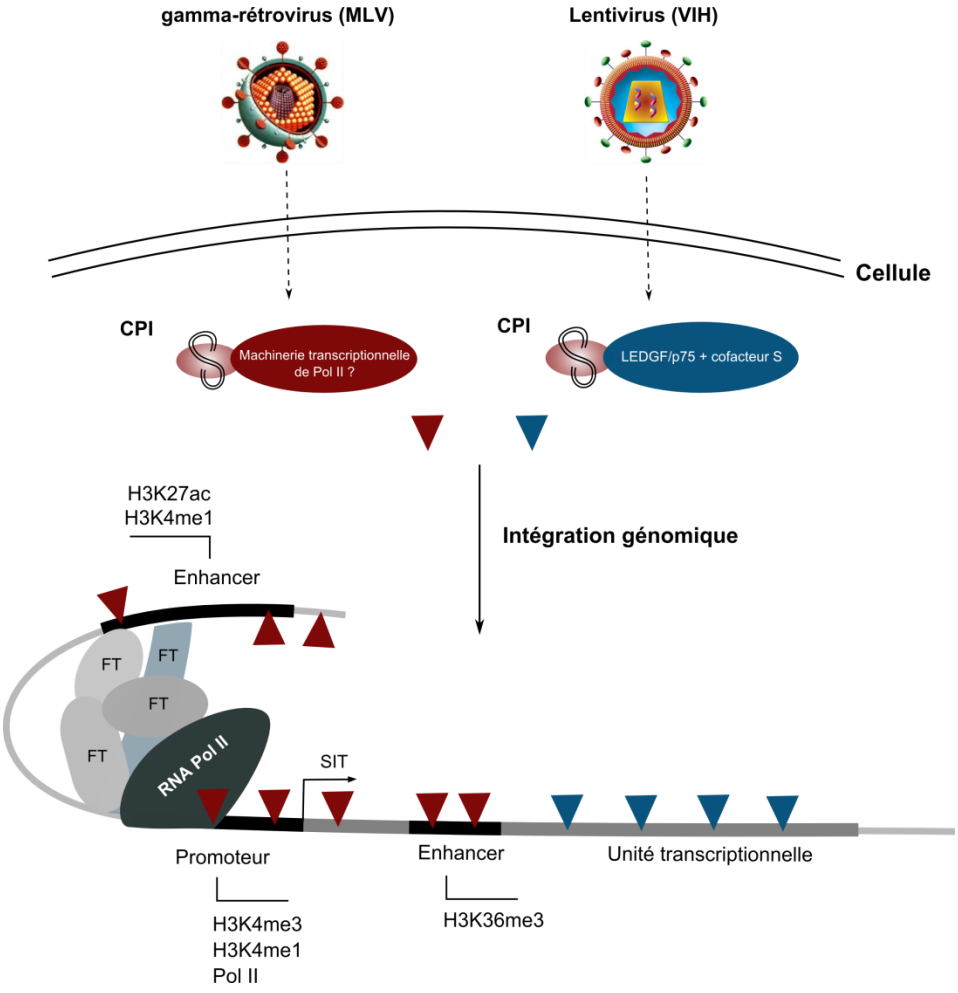


Figure 6 : Tropisme préférentiel des gamma-rétrovirus (MLV) et des lentivirus (HIV). CPI : complexe de pré-intégration, FT : facteur de transcription, SIT : site d'initiation de la transcription, LEDGF/P75 : lens epithelium–derived growth factor p75 splice variant. Adapté de Cavazza *et al.* 2013.

Pour les vecteurs AAVr, aucun évènement aussi grave n'a été observé. Cependant, une étude a impliqué les AAVr dans le développement de carcinomes hépato-cellulaires chez la souris (Donsante et al., 2007). La plupart des essais cliniques ont depuis été rassurants mais la génotoxicité demeure une préoccupation majeure en thérapie génique et le risque associé doit être évalué à plus grande échelle, dans différents contextes et pour tous les types de vecteurs utilisés, y compris pour ceux dits « non intégratifs » tels que les vecteurs AAVr ou Adenoviraux.

I.1.5.3 Transmission verticale

L'efficacité thérapeutique d'un traitement de thérapie génique apparait de plus en plus conditionnée par la prise en charge précoce des patients atteints de maladies génétiques, dès la jeune enfance dans certain cas. Pour des raisons éthiques et de biosécurité, il est indispensable que le transfert de gène soit restreint aux cellules somatiques chez ces jeunes patients qui sont en âge de procréer. En effet, la transduction de cellules germinales entrainerait la possibilité de générer des organismes humains génétiquement modifiés, ce qui est formellement interdit dans la majorité des pays. L'injection systémique d'un vecteur viral est l'approche la plus à risque et nécessite, pour chaque sérotype utilisé, l'étude du tropisme pour la lignée germinale.

Après injection systémique d'un AAV2, le génome du vecteur a été détecté par PCR jusqu'à 2 mois post-injection dans le sperme de patients hémophiles (Manno et al., 2006). Cette présence transitoire dans le sperme a également été observée chez le lapin avec un AAV2 (Schuettrumpf et al., 2006) et un AAV8 (Favaro et al., 2009) et semble plus refléter une voie d'excrétion du vecteur que le résultat d'une transduction cellulaire. Cependant, les vecteurs AAVr ont été utilisés pour transduire des cellules souches germinales *ex vivo* lors d'essais concluants de transgénèse chez la souris, la chèvre (Honaramooz et al., 2008) et le cochon (Zeng et al., 2013). Dans ces études, la transmission du transgène des cellules souches aux spermatozoïdes implique une intégration stable dans le génome cellulaire, évènement rare pour les AAVr. Cependant, étant donné que les doses utilisées *ex vivo* sont nettement supérieures à la fraction des AAVr injectés *in vivo* qui atteignent les cellules germinales, il semble que le risque de transmission verticale soit minime. Il convient néanmoins d'être prudent et une congélation du sperme ainsi qu'un moyen de contraception adapté sont recommandés pour les patients inclus dans des essais de thérapie génique *in vivo* (Favaro et al., 2009).

I.2 Les virus adéno-associés sauvages et recombinants

I.2.1 Les virus adéno-associés (AAV)

I.2.1.1 Caractéristiques

En 1965, Atchison et al. (Atchison et al., 1965) ont observé la présence de petites particules de tailles homogènes en microscopie électronique, lors de la culture d'un adénovirus simien (SV15) sur des cellules de rein de singe. Leur capside en capsomère et l'absence d'homologie antigénique avec les adénovirus ont permis à cette équipe d'identifier un nouveau virus : les AAV. Ils ont également observé que la réplication des AAV nécessitait la présence d'adénovirus, ce qui leur a valu plus tard d'être classés dans le genre Dependovirus de la famille des Parvoviridae.

Avec ses 22nm de diamètre, l'AAV est un des plus petit virus connus. Il est non-enveloppé et constitué d'une capside icosahédrique formée de l'assemblage de 60 sous-unités protéiques. L'AAV possède un génome à ADN simple brin de 4,7 kilobases comprenant 3 promoteurs et 2 gènes viraux (*rep* et *cap*) (**Figure 7**) entourés par deux régions inversées et répétées appelées ITR (*Inverse Terminal Repeats*). Les ITR contiennent 145 nucléotides dont 125 sont palindromiques et forment une structure secondaire en "T" qui sert notamment d'origine de réplication aux polymérases cellulaires (**Figure 8**).

Le gène *rep*, par l'intermédiaire des promoteurs p5 et p19 et d'un site donneur et accepteur d'épissage, code pour 4 protéines régulatrices nommées selon leur masse moléculaire : Rep78, Rep68, Rep52 et Rep40. Les Rep78/68 participent à la réplication du génome viral *via* (i) leur site de reconnaissance de l'ADN qui lie le RBE (*Rep binding element*) et le RBE' présents sur l'ITR, (ii) leur activité hélicase qui déroule la structure secondaire de l'ITR et (iii) leur activité endonucléasique ATP-dépendante (Snyder et al., 1993) qui clive l'ITR au niveau du trs (*terminal resolution site*). Les Rep52/40 permettent l'accumulation de génomes viraux simple brin et leur encapsidation à l'aide d'une activité hélicase (Gonçalves, 2005).

Les protéines de capside VP1, VP2 et VP3 sont produites à partir du promoteur p40 par épissage alternatif et par différents codons d'initiation, et s'assemblent avec un ratio de 1:1:10. Un second cadre ouvert de lecture a récemment été décrit (Sonntag et al., 2010) et permet la synthèse de la protéine chaperonne AAP (*assembly-activating protein*), qui a un rôle fondamental dans l'assemblage des VP pour former la capside.

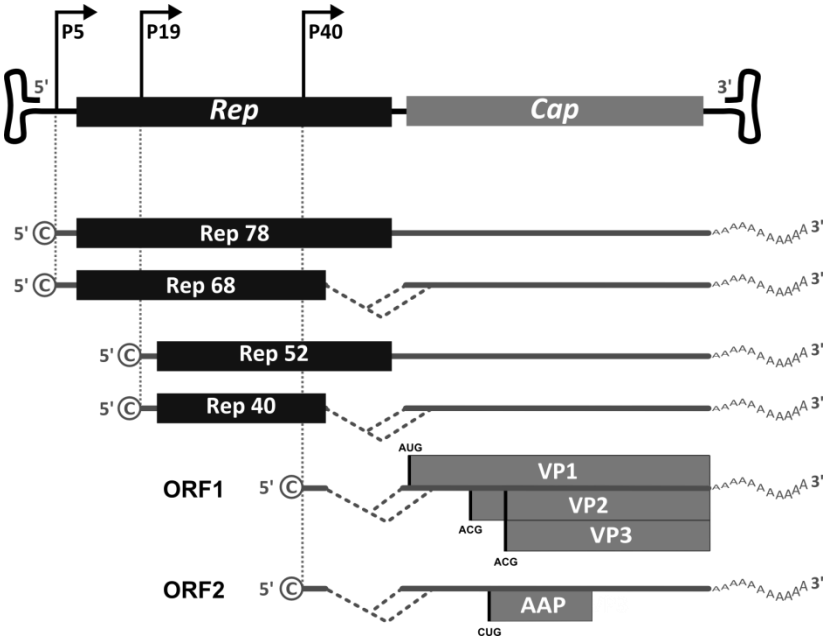


Figure 7 : Organisation du génome de l'AAV. Les promoteurs p5 et p19 codent pour les protéines Rep78 et Rep52 puis Rep68 et Rep40 par épissage alternatif. Le promoteur p40 code pour les 3 protéines de capsides *via* différents codons d'initiation de la traduction (VP1, VP2 et VP3) et pour la protéine chaperonne AAP par un second cadre ouvert de lecture (ORF). Adapté de la thèse A.Léger 2012.

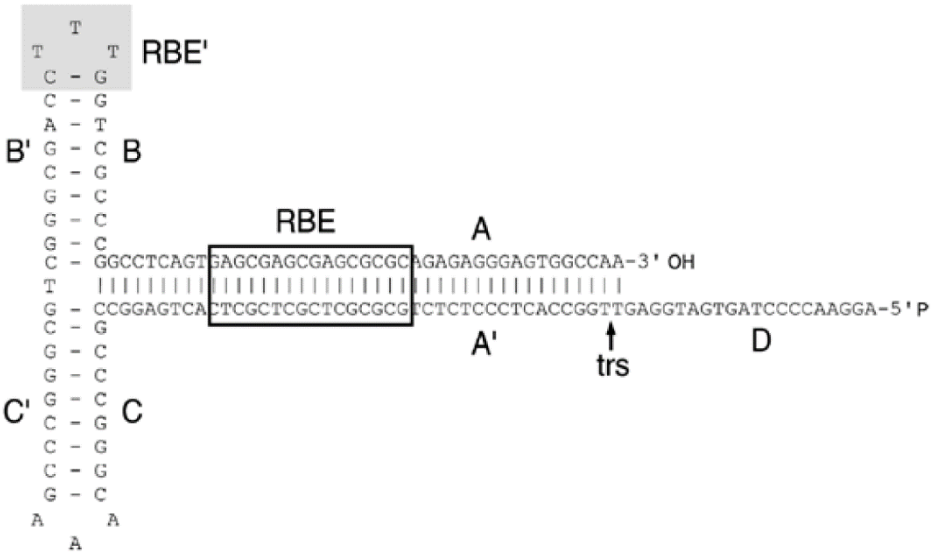


Figure 8 : Structure secondaire de l'extrémité palindromique de l'AAV (ITR). Les régions palindromiques B-B', C-C' et A-A' permettent le repliement caractéristique en "T". La région D correspondant aux 20 derniers nucléotides de l'ITR est simple brin. L'origine de réplication en 3' ainsi formée permet aux polymérases cellulaires de répliquer le génome de l'AAV. Les protéines Rep78/68 se fixent à la région RBE (*Rep Binding Element*) et RBE', clivent l'ADN par leur activité endonucléase au niveau du trs (*terminal resolution site*) et déroulent l'ITR par leur activité hélicase permettant la résolution du second brin. Tiré de Goncalves *et al.* 2005.

I.2.1.2 Sérotypes et pathogénicité

Les AAV ont été retrouvés chez de nombreuses espèces animales dont l'Homme, les bovins, les ovins ou les oiseaux (Bossis and Chiorini, 2003; Schmidt et al., 2004). Chez l'Homme et les primates, une dizaine de sérotypes différents et de nombreux variants ont été

identifiés (**Figure 9**). Les premiers sérotypes d'AAV (1, 2, 3, 4 et 6) ont été découverts de façon fortuite comme contaminants de cultures d'adénovirus (Atchison et al., 1965; Hoggan et al., 1966). L'AAV 5 a été isolé chez l'Homme d'un condylome pénien (Bantel-Schaal and zur Hausen, 1984) et les sérotypes 7, 8, 9 et 10 ont été identifiés par PCR dans du tissu cardiaque de primates non humains (Gao et al., 2002). En terme de séquence protéique, l'AAV6 est très proche de l'AAV1 (96% d'homologie), il a probablement été généré en laboratoire par recombinaison entre l'AAV1 et 2 (Rutledge et al., 1998). L'AAV3 est également très proche de l'AAV2. De manière générale, un degré important de cross-réactivité entre les sérotypes a été observé (Calcedo et al., 2009; Thwaite et al., 2015).

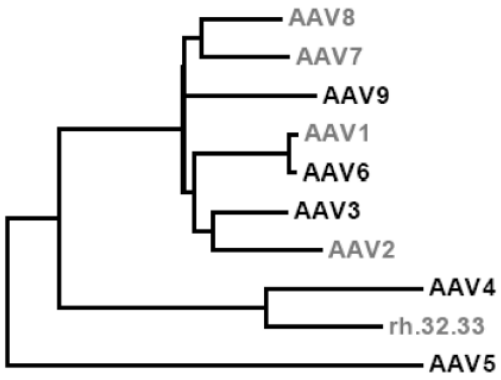


Figure 9 : Arbre phylogénétique des principaux sérotypes d'AAV humains et simiens. Données issues de la comparaison de la séquence en acides aminés de VP1. Tiré de Calcedo *et al*. 2009.

Chez l'Homme, le sérotype 2 semble le plus répandu avec une séroprévalence estimée de 30%, 35% et 60% respectivement aux Etats-Unis, en Europe et en Afrique (Calcedo et al., 2009). L'AAV2 a été retrouvé au niveau pharyngé, anal (Blacklow et al., 1967, 1968), génital (Bantel-Schaal and zur Hausen, 1984; Coker et al., 2001) et dans le liquide amniotique (Arechavaleta-Velasco et al., 2008). La transmission de l'AAV, intimement liée à celle de ses virus auxiliaires dont ils sont dépendants pour leur réplication, peut ainsi avoir lieu *in utero*, pendant l'enfance (voie aérienne et digestive) et chez le jeune adulte (voie sexuelle). Aucune pathologie humaine n'a été associée aux AAV jusqu'à présent. Cependant, une corrélation a été observée entre la présence d'IgM anti-AAV pendant le premier trimestre de la grossesse (primo-infection ou réactivation) et l'augmentation du risque de pré-éclampsie et d'accouchement prématuré (Arechavaleta-Velasco et al., 2008; Burguete et al., 1999)**.** Ces corrélations sont à prendre avec précaution puisque leur présence peut témoigner de la présence d'autres virus comme le HSV ou le papillomavirus (Erles et al., 2001). Un rôle protecteur contre le cancer du col de l'utérus a également été évoqué (Coker et al., 2001).

I.2.1.3 Cycle infectieux des AAV

L'AAV2 utilise comme récepteur primaire les protéoglycanes à héparanes sulfates (HSPG) (Summerford and Samulski, 1998), auquel il se lie grâce à cinq résidus basiques présents dans les VP (Opie et al., 2003). Les HSPG sont des polysaccharides sulfatés présents à la surface des cellules et dans la matrice extracellulaire des eucaryotes supérieurs, ils sont ubiquitaires et expliquent le large tropisme cellulaire de ce virus. L'internalisation est ensuite favorisée par des co-récepteurs : l'intégrine αvβ5 (Summerford et al., 1999), le récepteur au facteur de croissance des fibroblastes (Qing et al., 1999) et le récepteur au facteur de croissance hépatocytaire (Kashiwakura et al., 2005). Les sérotypes 1, 4, 5, 6 lient différents types de glycanes sialylés (Kaludov et al., 2001; Wu et al., 2006) tandis que le sérotype 9 se fixe sur des glycanes possédant des N-galactoses terminaux (Shen et al., 2011). L'AAV5 cible également le récepteur au facteur de croissance plaquettaire (Di Pasquale et al., 2003).

Suite à l'internalisation, les AAV peuvent établir une infection lytique ou latente en fonction de la présence ou non de facteurs auxiliaires. En effet, en l'absence de co-stimulation, l'AAV est incapable de se répliquer et établit une infection latente en s'intégrant dans le génome cellulaire. Dans les premières études réalisées, l'intégration du génome de l'AAV a été localisé spécifiquement dans le chromosome 19 (19q13.4), dans le premier exon de la protéine PPP1R12C (*protein phosphatase 1 regulatory subunit 12C*) au sein d'un locus qui a été dénommé AAVS1 (Kotin et al., 1990; Samulski et al., 1991). Ce mécanisme d'intégration ciblée, unique parmi les virus eucaryotes, a depuis été affiné grâce à de nouvelles méthodes d'analyse utilisant le séquençage haut débit (**Figure 10a**). L'intégration de l'AAV2 se fait ainsi préférentiellement au niveau de répétitions GAGC, séquences qui ressemblent au motif RBS de l'ITR (Janovitz et al., 2013). Les protéines Rep78/68 lient ces séquences, avec une affinité d'autant plus forte qu'il y a de répétitions (Hüser et al., 2014), ce qui favorise l'intégration du génome viral (**Figure 10b**). Il faut noter que l'AAV5, le plus divergent des sérotypes, code pour des protéines Rep ayant un domaine de reconnaissance de l'ADN particulier, il présente ainsi un profil d'intégration génomique différent (Janovitz et al., 2014). Cette intégration ciblée n'a pour l'instant pas été mise en évidence *in vivo*. Par ailleurs, une équipe a suggéré une possible persistance du génome viral sous forme épisomale (extra-chromosomique) dans les tissus humains (Schnepp et al., 2005).

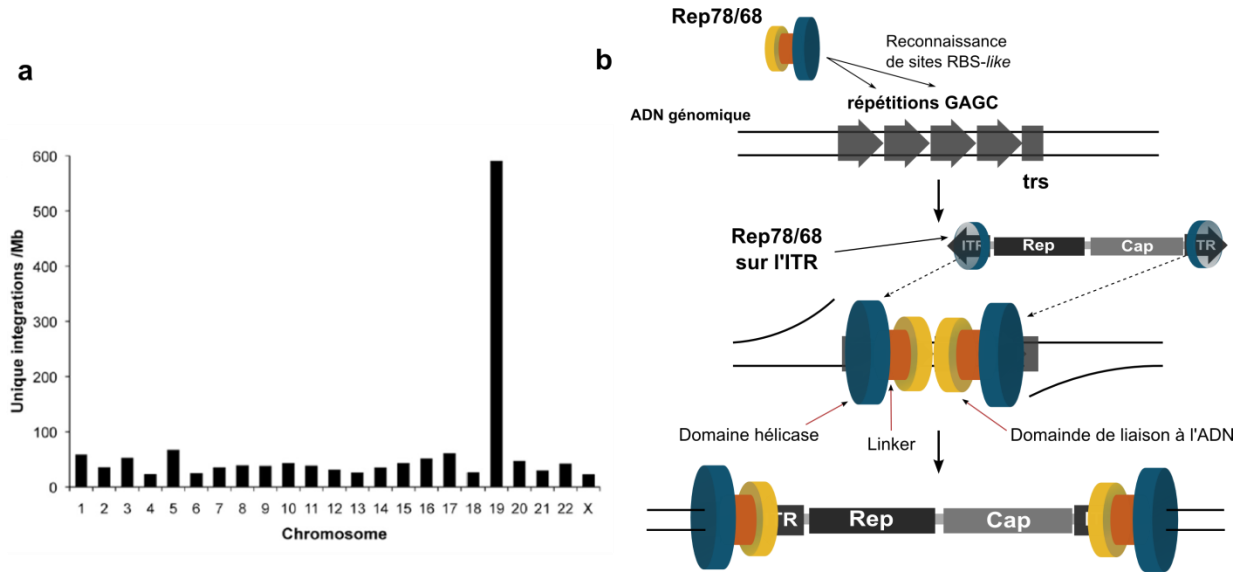


Figure 10 : Intégration ciblée de l'AAV2 dans le génome humain. (**a**) En culture de cellules humaines, environ 45% des évènements intégratifs sont retrouvés dans le chromosome 19 au niveau de l'AAVS1. Le reste des intégrants sont répartis dans le génome avec une préférence pour les régions présentant des répétitions GAGC. Tiré de Janovitz *et al.* 2013. (**b**) Modèle d'intégration génomique de l'AAV. Les protéines Rep78/68 lient des séquences analogues au RBS de l'ITR dans le génome (répétitions GAGC) puis leur domaine endonucléasique clive l'ADN au niveau de séquences *trs-like*. Les Rep 78/68 se dimérisent avec celles liant les ITR de l'AAV, permettant une intégration du génome de l'AAV après l'action des ligases. Adapté de Janovitz *et al.* 2013.

Plusieurs virus peuvent fournir les facteurs auxiliaires nécessaires à la réplication de l'AAV : les adénovirus (Atchison et al., 1965), les papillomavirus (Walz et al., 1997), les virus de l'herpès (Alazard-Dany et al., 2009) et le baculovirus. Lorsqu'une cellule infectée avec un AAV latent est co-infectée par un de ces virus auxiliaire, il se déclenche une infection lytique qui permet de produire de nouveaux virions AAV et de propager l'infection. Le génome de l'AAV est excisé et se réplique par l'intermédiaire des polymérases cellulaires et des protéines Rep (**Figure 11**). Les VP s'assemblent au niveau des nucléoles (Wistuba et al., 1997) et les génomes viraux simples brins sont encapsidés dans les capsides pré-formées par leur extrémité 3' dans le nucléoplasme. Lors de ce cycle, le rôle des facteurs auxiliaires a été peu documenté, cependant les gènes adénoviraux ont été les plus étudiés : E1a active la transcription des gènes Rep et Cap, E1b et E4 améliorent la vitesse de transport des ARNm dans le cytoplasme, E2a et l'ARN non-codant des adénovirus (VA-RNA) améliorent la stabilité de l'ARNm et l'efficacité de traduction (Xiao et al., 1998). L'identification des différents facteurs nécessaires à la réplication des AAV a permis à certaines équipes d'envisager l'utilisation de ces virus comme vecteurs géniques en remplaçant une partie du génome viral par un gène d'intérêt.

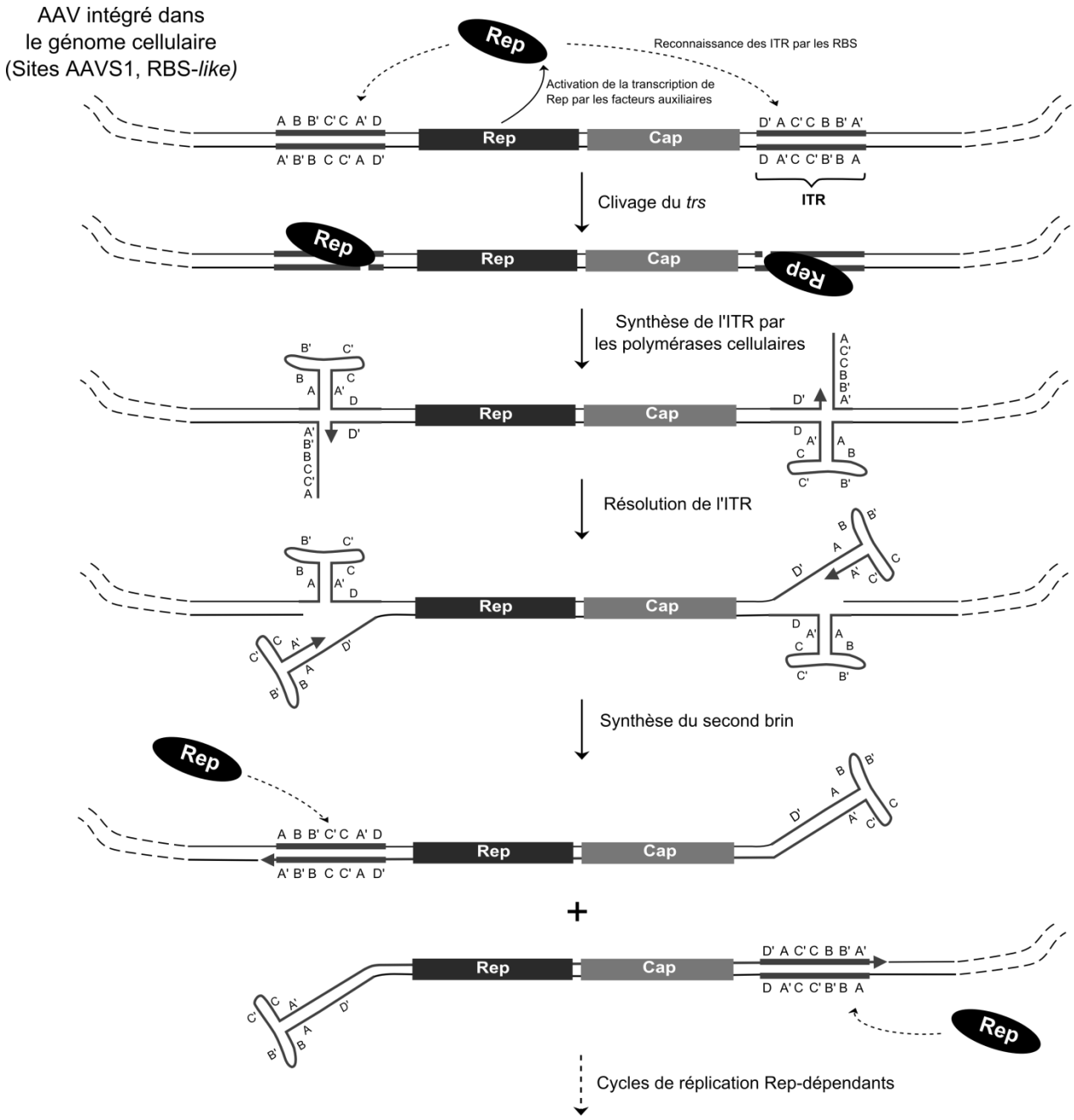


Figure 11 : Modèle d'excision par réplication d'un AAV intégré dans le génome cellulaire. Adapté de Ward *et al*. 2003

I.2.2 Les virus adéno-associés recombinants (AAVr)

I.2.2.1 Genèse des AAVr

En 1983, Samulski *et al.* ont décrit la production de particules virales d'AAV après transfection de cellules humaines par un plasmide contenant le génome de l'AAV et infection par un adénovirus (Samulski et al., 1983). Le génome de l'AAV pouvait ainsi être excisé et répliqué à partir d'un plasmide de la même façon qu'à partir d'un virus latent intégré. Suite à cette observation, les premières expériences de transfert de gène ont été réalisées en remplaçant le gène Cap, celui-ci étant apporté par un autre plasmide (le terme en *trans* sera utilisé par la suite), par un gène codant la résistance à la néomycine (Hermonat and Muzyczka, 1984) ou pour la chloramphénicole acétyltransférase (Tratschin et al., 1984). Les vecteurs AAVr produits avaient efficacement exprimé ces gènes dans des cellules humaines. Par la suite, seuls les ITR (145 bases terminales) se sont montrés indispensables en *cis* pour la réplication et l'encapsidation du génome (Samulski et al., 1983). Un transgène d'environ 4,5kb pouvait ainsi être inclus entre les ITR. L'inconvénient était la production concomitante d'adénovirus, qu'il fallait éliminer par purification. En 1998, une équipe a montré que les facteurs auxiliaires adénoviraux pouvaient être apportés par un troisième plasmide (Xiao et al., 1998), évitant ainsi la production d'adénovirus et ouvrant la voie à l'ère moderne de production des AAVr.

I.2.2.2 Systèmes de production actuels

La production de médicaments chimiques par l'industrie pharmaceutique a depuis des dizaines d'années été optimisée et développée à grande échelle. Avec l'arrivée des biothérapies et la production de protéines recombinantes, la production de médicaments s'est complexifiée par l'utilisation de cultures de bactéries ou de cellules eucaryotes. La production de vecteurs viraux recombinants amène un niveau supérieur de complexité en faisant intervenir un système viral d'encapsidation. La maitrise de ces systèmes de production est aujourd'hui un enjeu majeur en thérapie génique car, après avoir fait ses preuves de concept, le traitement d'un nombre important de patients nécessite aujourd'hui une grande quantité de vecteur. Par exemple, le Glybera, le premier AAVr ayant obtenu une AMM, doit être injecté chez les patients en intramusculaire à une dose de $3x10^{12}$ génomes de vecteur par kilogramme (Bryant et al., 2013), soit plus que le nombre total de cellules eucaryotes de l'organisme, estimé à $3,72x10^{13}$ (Bianconi et al., 2013). Ainsi, une optimisation constante des procédés de production des vecteurs AAVr est nécessaire. Aujourd'hui, les principaux utilisent la transfection de cellules humaines, le système baculovirus et les lignées stables (Satkunanathan et al., 2014), les deux premières sont les plus utilisées et sont décrites ci-dessous.

- **Transfection transitoire**

Les cellules humaines HEK293 (*Human embryonic kidney*, 293ème essai de transformation) ont été obtenues en 1973 par transformation de cellules de rein embryonnaire issues d'un avortement avec de l'ADN fragmenté d'un Adénovirus 5. C'est l'insertion du gène E1 (4,5kb) dans le chromosome 19 (19q13.2) qui a permis l'immortalisation de ces cellules (Louis et al., 1997). Elles ont depuis été largement utilisées en biotechnologie.

La production d'AAVr est actuellement effectuée par co-transfection des HEK293 avec un plasmide vecteur contenant le transgène entouré par les ITR et un plasmide auxiliaire contenant les gènes de l'AAV (Rep2/CapX) et les gènes adénoviraux (E2a, E4 et VA-RNA), E1 étant apporté par les HEK293 (**Figure 12a**). La transfection se fait lorsque les cellules sont en mitose (70 à 80% de confluence) pour permettre aux plasmides de traverser la membrane nucléaire, qui est alors rompue. L'expression des différents gènes dans ce système artificiel permet l'excision et la réplication du génome du vecteur puis son encapsidation. Si la plupart des vecteurs possèdent un ITR2 et sont ainsi produits avec l'expression du gène Rep2, différents sérotypes peuvent être produits selon le gène Cap exprimé (Cap1, Cap2,…). Selon le sérotype produit et la durée d'incubation (48h à 96h), les particules recombinantes sont extraites à partir du culot cellulaire ou du surnageant. Après une étape de précipitation avec du polyéthylène glycol, les AAVr sont purifiés selon différents procédés décrits plus bas.

Cette méthode de production est largement utilisée pour produire des vecteurs utilisés lors des essais pré-cliniques sur des modèles animaux. Mais l'utilisation de cellules adhérentes limite les capacités de production, qui s'avèrent insuffisantes lorsque des quantités très importantes de vecteurs sont nécessaires comme lors de la plupart des essais chez le gros animal et des essais cliniques. Dans ces situations, selon la dose utilisée, il est souvent nécessaire d'obtenir plus de 10^{15} particules de vecteur. Par transfection, cela nécessiterait plus de 10^{11} HEK293 sur une surface totale de flasques de 5000x175cm^2 (Urabe et al., 2002).

- **Système baculovirus**

Les baculovirus sont des virus d'insectes qui ont été utilisés depuis le début des années 1980 par l'industrie pharmaceutique pour produire de grandes quantités de protéines recombinantes. Le système a été adapté pour produire des AAVr et consiste à générer deux baculovirus : l'un comportant le transgène entre deux ITR et l'autre les gènes Rep et Cap (**Figure 12b**). Ces éléments sont insérés par transposition à la place d'un gène non essentiel pour la réplication du baculovirus dont le promoteur viral natif permet une forte expression du transgène incorporé. L'apport des gènes adénoviraux n'est pas nécessaire, les fonctions auxiliaires sont apportées par les baculovirus et les cellules Sf9, bien que les facteurs protéiques

ne soient pas identifiés. Lorsque les deux baculovirus recombinants infectent conjointement des cellules de *Spodoptera frugiperda* (Sf9) en suspension, le génome de l'AAVr est répliqué puis encapsidé.

L'utilisation de cellules en suspension permet d'augmenter plus facilement la quantité de vecteurs produits. En effet, la capacité de production par cellule étant la même que par transfection (10^4 à 10^5 AAVr/cellule), l'augmentation du volume cellulaire total est le seul moyen d'augmenter le rendement. Les bioréacteurs actuels pouvant atteindre une capacité de 2000L, c'est la seule méthode de production pouvant fournir suffisamment de vecteurs pour les essais cliniques demandant de grandes quantités de vecteurs. Malgré cet avantage, le produit final est beaucoup plus hétérogène et moins caractérisé que par transfection, il est notamment nécessaire d'éliminer les baculovirus qui se sont multipliés pendant la production.

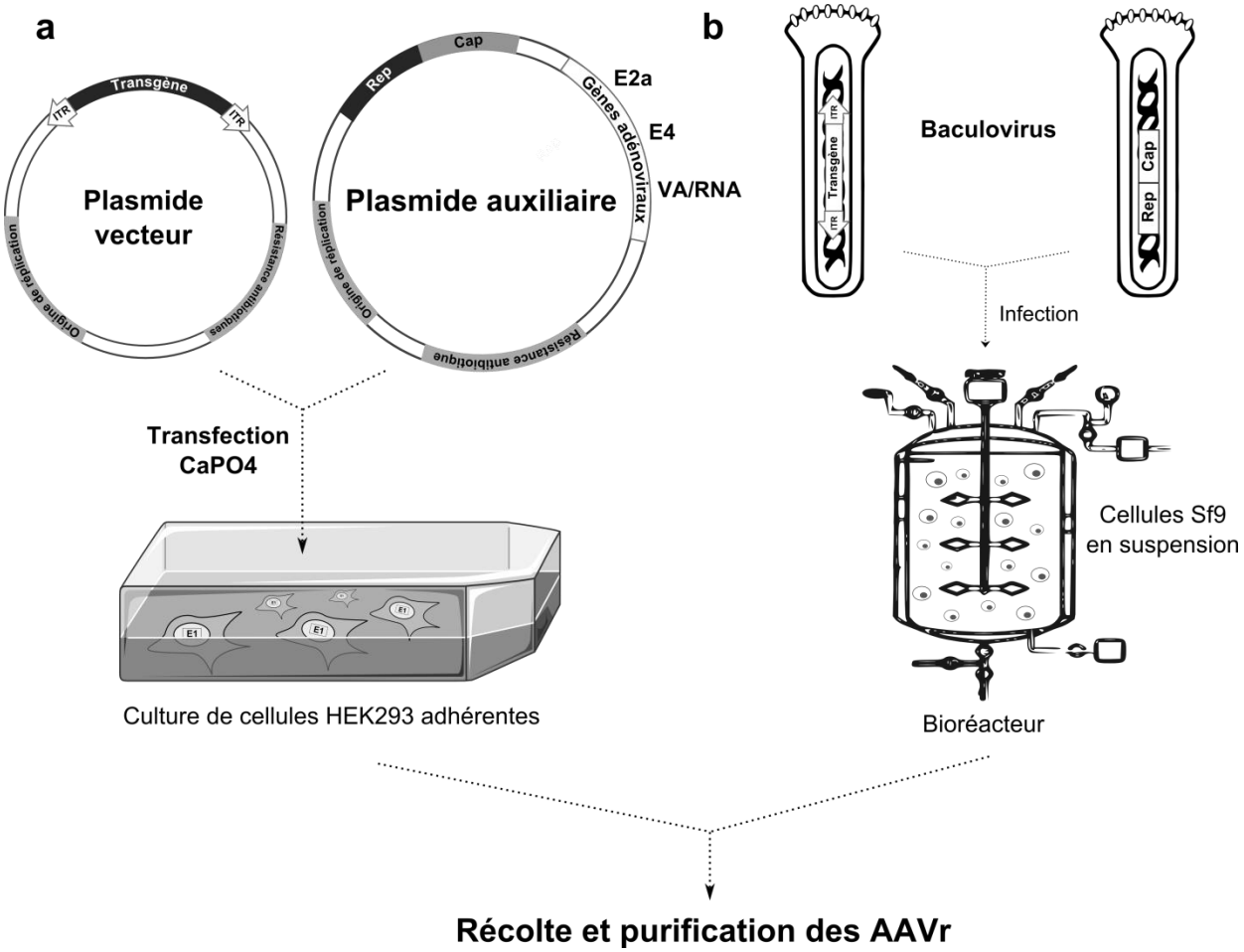


Figure 12 : Production des AAV recombinants par transfection transitoire et par système baculovirus. (a) Transfection transitoire : un plasmide contenant le transgène entouré par deux ITR et un plasmide comportant les gènes Rep/Cap et auxiliaires sont transfectés dans des cellules HEK293 adhérentes. **(b)** Système baculovirus : un baculovirus contenant le transgène entouré par deux ITR et un baculovirus contenant les gènes Rep/Cap infectent des cellules d'insecte en suspension. Différents sérotypes sont produits selon le gène Cap exprimé.

I.2.2.3 Purification et caractérisation des lots produits

Dans l'idéal, un lot de vecteur AAVr doit être une suspension pure de capsides ayant chacune internalisée une copie du génome du vecteur. Cependant, après production, un mélange complexe de capsides vides, de protéines cellulaires, de lipides et d'acides nucléiques d'origines plasmidique et cellulaire est présent en plus des capsides « pleines » fonctionnelles. Ainsi, des systèmes de purification et de caractérisation ont été développés respectivement pour éliminer ces impuretés et pour s'assurer qu'elles ne sont pas présentes dans le lot final au-dessus d'une certaine quantité. C'est à la fois un enjeu important de biosécurité pour les patients et un paramètre pouvant influer sur l'efficacité de la thérapie génique.

- **Systèmes de purification**

La purification par **ultracentrifugation sur double gradient de chlorure de Cesium** (Ayuso et al., 2010) est basée sur une séparation des constituants selon leur densité. Les capsides pleines peuvent ainsi être séparées des contaminants après prélèvement de la bande migrant à la densité attendue pour un sérotype donné. Cette méthode est efficace pour éliminer les capsides vides qui, en plus d'être inefficaces, peuvent contribuer à la stimulation de la réponse immune anti-capside et ainsi altérer la persistance du transgène (Nathwani et al., 2011). Néanmoins, si cette méthode est efficace lorsqu'elle est appliquée à une production par transfection transitoire, elle est rendue plus compliquée pour le système baculovirus dans lequel le profil de migration est beaucoup plus hétérogène. Elle est également difficile à transposer sur des grands volumes et surtout difficile à automatiser.

Les **méthodes chromatographiques** sont plus facilement applicables sur de larges productions, il existe la chromatographie d'échange ionique (Davidoff et al., 2004) et la chromatographie d'affinité (Smith et al., 2009). La première consiste à lier les capsides chargées sur des colonnes anioniques ou cationiques puis à les éluer et la seconde consiste à fixer les capsides sur une colonne par l'intermédiaire d'anticorps anti-capsides. Ces méthodes ont l'avantage d'être applicables sur de gros volumes, reproductibles et automatisables. Cependant, une mise au point importante est nécessaire pour chaque sérotype d'AAVr utilisé, voire pour chaque transgène produit.

- **Caractérisation des contaminants**

La **pureté protéique** est évaluée par migration sur gel d'acrylamide (SDS-PAGE) de la production d'AAVr après dénaturation thermique. Dans une production pure, seules les bandes correspondantes à VP1, VP2 et VP3 doivent être présentes. Si c'est souvent le cas après transfection transitoire, le système baculovirus produit de nombreuses autres bandes encore mal caractérisées. La génération de **virus réplicatifs** pendant la production, potentiellement

engendrés par recombinaison entre les différents plasmides de production, est étudiée par un test cellulaire (les systèmes actuels de production limitent de façon importante leur présence). Enfin, la proportion de **capsides vides** peut être estimée par microscopie électronique.

Une fraction de l'ADN nécessaire à la production des AAVr (plasmidique et cellulaire) se retrouve dans le lot final purifié. Ces contaminants, dont la concentration résiduelle est estimée par PCR quantitative, posent un important problème de biosécurité. En effet, la présence importante d'**ADN contaminant** dans la production d'AAVr était un des six points critiques relevés par l'agence européenne du médicament pendant la procédure d'attribution de l'AMM au Glybera (rapport EMA 2012). Une des craintes était le transfert de séquences possédant un cadre ouvert de lecture intact pouvant produire une protéine d'origine virale d'effet inconnu chez l'Homme. Par ailleurs, le transfert de gènes de résistance aux antibiotiques et d'oncogènes (E1) utilisés lors de la production d'AAVr pourrait également être préjudiciable pour le patient. Une étude menée au laboratoire a montré que les séquences d'origine procaryote représentent jusqu'à 10% de l'ADN total encapsidé et peuvent être transférées *in vivo* chez le chien et le primate, où elles sont détectables plusieurs mois après injection (Chadeuf et al., 2005). Etant donné les doses massives de vecteurs injectées chez les patients, cela représente une grande quantité d'ADN contaminant transférée. De façon moins importante, des séquences des gènes auxiliaires (Allay et al., 2011; Chadeuf et al., 2005; Ye et al., 2011) et de l'ADN génomique des cellules productrices (Allay et al., 2011) ont été retrouvés respectivement jusqu'à 0,5% de l'ADN encapsidé et 1ng/10^{12} génomes de vecteur. Cependant, les méthodes actuelles de détection des contaminants ADN, basées sur la PCR, ne fournissent pas une vision exhaustive de ce qui est réellement injecté chez les patients. Ainsi, un important travail de caractérisation des vecteurs AAVr persiste et celui-ci passera par le développement de nouvelles méthodes analytiques.

I.2.2.4 Etapes du transfert de gène

Les premières étapes de transduction des AAVr dépendent des propriétés du sérotype viral d'origine. En particulier le tropisme cellulaire, qui est modifié selon les récepteurs et co-récepteurs ciblés (voir I.2.1.3). L'**internalisation** des particules de vecteur semble varier selon le type cellulaire considéré, plusieurs voies sont possibles : la macropinocytose (Weinberg et al., 2014), l'endocytose dépendant de la clathrine et de la dynamine (Duan et al., 1999) ou encore par endocytose indépendante de la clathrine ou CLIC/GEEC (Nonnenmacher and Weber, 2011). Suite à leur internalisation, les AAVr traversent le cytoplasme par la voie endo-lysosomale, dont l'interaction avec le réseau de microtubules cellulaire entraine un transport rapide vers la zone péri-nucléaire (Xiao and Samulski, 2012). Puis, l'acidification des

endosomes permet un changement conformationnel de VP1 et l'exposition de son domaine N-terminal. Celui-ci possède une activité phospholipase A2 qui est responsable de la lyse de l'endosome et du relargage cytoplasmique de l'AAVr (Sonntag et al., 2006). Les signaux de localisation nucléaire présents sur VP1 semblent ensuite lier des protéines chaperonnes, les importin-beta, qui permettent l'internalisation des particules intactes dans le noyau par les pores nucléaires (Nicolson and Samulski, 2014) puis leur accumulation au niveau des nucléoles (Johnson and Samulski, 2009). Le relargage du génome du vecteur a sans doute lieu dans le nucléoplasme mais la cinétique de cette étape est encore mal décrite. Après toutes ces étapes, une équipe a montré *in vitro* que seuls 30% des AAVr atteignent le noyau (Xiao et al., 2012).

Lorsque les génomes recombinants sont relargués au niveau intra-cellulaire, en absence de gènes viraux, leur devenir moléculaire est presque totalement dépendant de l'interaction entre les ITR et les facteurs protéiques apportés par la cellule hôte. *In vitro,* le génome recombinant déclenche les mêmes évènements moléculaires que l'effondrement d'une fourche de réplication (Jurvansuu et al., 2005) : il s'assemble avec des protéines impliquées dans la réparation de l'ADN comme le complexe MRN (Mre11, Rad52, Nbs1) et MDC1, qui induisent la **recombinaison homologue (RH)**, et Ku86, qui active la **recombinaison non-homologue (NHEJ)** (Cervelli et al., 2008; Zentilin et al., 2001). Cette caractéristique est à la base d'une importante limite de transduction pour certaines cellules (**Figure 13**). En effet, ces deux voies sont activées différemment selon l'état prolifératif ou quiescent d'une cellule : les cellules en mitose activent la RH pour permettre une réparation fidèle d'une cassure double brin de l'ADN en utilisant l'information portée par la chromatine sœur tandis que les cellules en G0 ou post-mitotiques utilisent la NHEJ pour réparer rapidement la lésion par un processus non fidèle. Il a été montré que les AAVr transduisent efficacement les cellules quiescentes, au contraire des cellules en prolifération (Lovric et al., 2012). Cependant, la permissivité des cellules en mitose est restaurée après inhibition de Nbs1 et ATM par KO cellulaire (Cataldi and McCarty, 2013; Cervelli et al., 2008; Sanlioglu et al., 2000; Zentilin et al., 2001) et du complexe MRN (Lovric et al., 2012) et de MDC1 (Cervelli et al., 2008) par des ARN interférents. Lovric *et al.* ont confirmé cette observation *in vivo* chez la souris en augmentant l'efficacité du transfert de gène dans le foie en prolifération (néo-natal) en co-administrant des ARN interférents anti-complexe MRN, ouvrant une perspective thérapeutique. De façon intéressante, la protéine adénovirale E4orf6 induit une dégradation du complexe MRN pour promouvoir l'infection productive par un adénovirus et un AAV sauvage (Schwartz et al., 2007)**.** Pour résumer, la permissivité aux AAVr dans les cellules en mitose est restreinte par les protéines impliquées dans la RH, vraisemblablement par une interaction directe avec le génome simple brin. Elles peuvent notamment altérer la synthèse du second brin et la formation des formes moléculaires de

persistance de l'AAVr. Par ailleurs, la synthèse du second brin peut également être inhibée par l'action d'une protéine chaperonne (FKBP52) qui se lie à la région D de l'ITR sous sa forme phosphorylée (Qing et al., 1999).

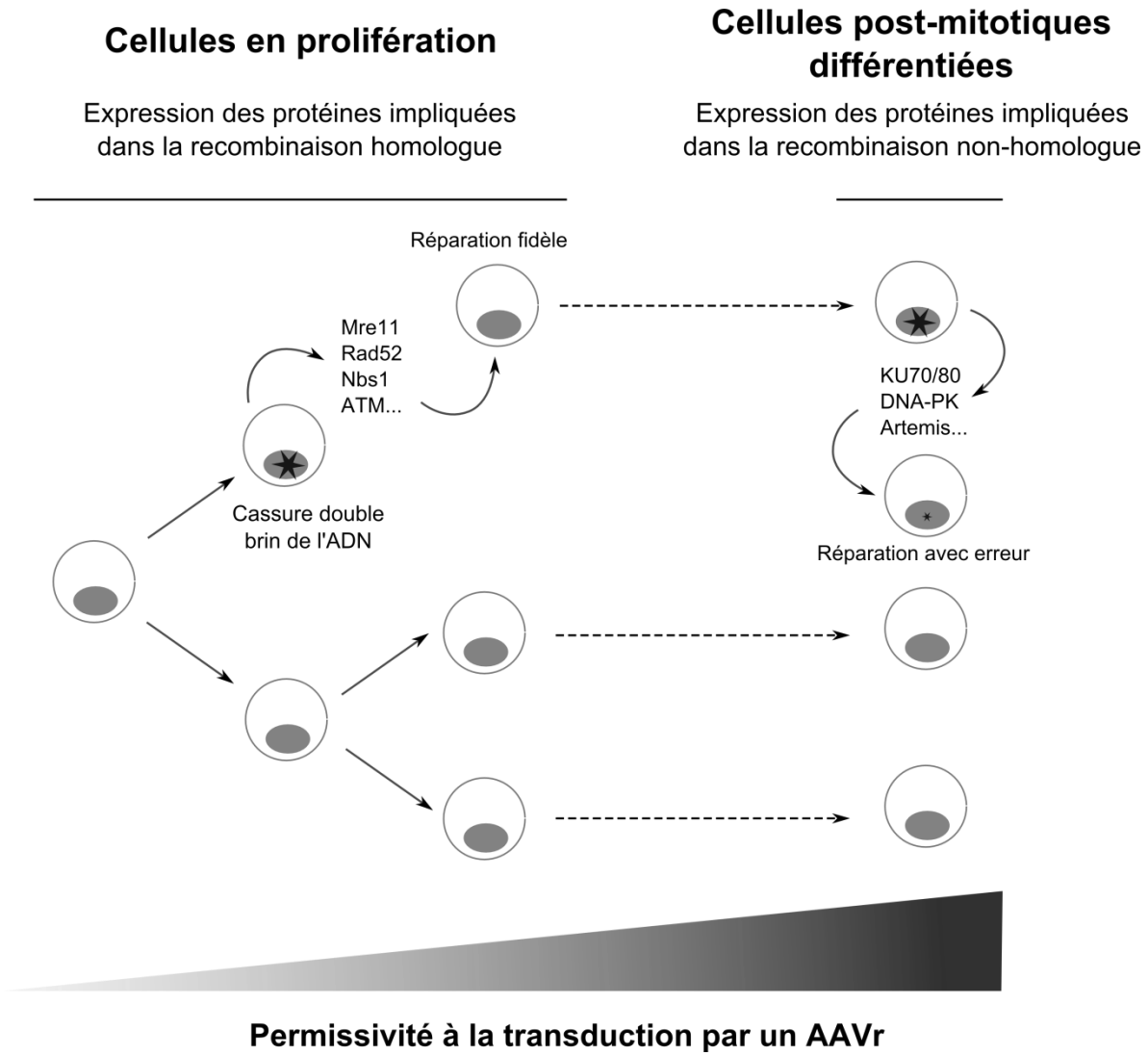


Figure 13 : Influence de la recombinaison homologue et non-homologue sur la transduction par un AAVr. * = dommages à l'ADN.

Suite à la synthèse du second brin, en l'absence des protéines Rep, les génomes recombinants vont pouvoir **persister** au niveau nucléaire majoritairement sous forme **épisomale** (ou extra-chromosomique). Ces épisomes peuvent être linéaires ou circulaires, monomériques ou concatémériques (assemblage de plusieurs génomes). Les concatémères circulaires, qui se forment progressivement, semblent être responsables de la persistance à long terme, comme montré dans le muscle de primate (Penaud-Budloo et al., 2008). La formation des concatémères dépendrait de la RH, qui assemblerait deux génomes suite à l'hybridation de leurs extrémités répétées (ITR). Par ailleurs, la persistance de formes linéaires chez des souris SCID (DNA-PK -/-) (Duan et al., 2003; Song et al., 2001) tend à montrer que la circularisation

des génomes serait quant à elle dépendante de la NHEJ. Cependant, comme évoqué précédemment, l'intervention de la RH et de la NHEJ est très dépendante du tissu considéré (Inagaki et al., 2007a). Cette persistance épisomale permet l'expression à long terme d'un transgène dans les tissus quiescents après une injection unique et est à l'origine des récents succès cliniques obtenus avec les AAVr dans la rétine (MacLaren et al., 2014; Maguire et al., 2008), le muscle (Childers et al., 2014; Le Guiner et al., 2014) et le foie (Nathwani et al., 2014).

I.2.2.5 Persistance de l'AAVr dans un phénotype pathologique

Nous avons vu que le devenir moléculaire du génome de l'AAVr est dépendant du niveau d'expression de différentes protéines cellulaires. Ainsi, en contexte pathologique, l'expression différentielle de ces protéines est susceptible d'altérer la persistance du vecteur donc de diminuer l'efficacité thérapeutique. Cependant, peu d'études ont étudié l'impact d'un phénotype pathologique sur l'efficacité thérapeutique d'une thérapie génique par AAVr. Dans une étude du laboratoire étudiant la persistance de l'AAVr dans des muscles de souris modèles de myopathie de Duchenne (DMD) (Dupont et al., 2015), il a été montré une perte des génomes AAVr dans les muscles déficients en dystrophine en comparaison avec des muscles sains. Cette observation a été corrélée avec la surexpression de protéines impliquées dans la RH et la NHEJ. Cette étude a également montré que l'efficacité de l'expression du vecteur était altérée par une diminution de la stabilité des ARNm causée par un important stress oxydatif.

Le stress oxydatif est une composante importante de la physiopathologie de nombreuses maladies, comme Alzheimer et Parkinson (Shan et al., 2003). Dans la DMD, le stress oxydatif participe à la dégénérescence progressive des fibres musculaires par oxydation des protéines, des lipides, de l'ADN et de l'ARN. Notamment, il a été montré que les espèces réactives de l'oxygène peuvent directement oxyder les bases azotées du génome des fibres musculaires dystrophiques avec un effet mutagène et une fréquente génération de cassures double brin de l'ADN (Schmidt et al., 2011). Dans ce contexte d'instabilité génomique, l'introduction d'ADN exogène sous la forme d'un génome AAVr pourrait avoir des conséquences néfastes sur l'intégrité du génome cellulaire. Ainsi, il parait important d'étudier l'interaction entre les AAVr et le génome cellulaire dans la DMD, notamment en termes de risque intégratif. C'est une question importante car les récents succès obtenus avec les AAVr sur des modèles canins de DMD (Le Guiner et al., 2014; Shin et al., 2013) laissent entrevoir des essais cliniques dans les prochaines années.

I.2.2.6 Génotoxicité

Malgré l'absence d'expression des protéines Rep, le génome du vecteur est capable de s'intégrer dans le génome cellulaire. Les AAVr présentent ainsi un risque génotoxique qui

convient d'évaluer, notamment en estimant la fréquence et le profil d'intégration génomique. La **fréquence d'intégration est faible**, elle a été estimée entre 10^{-4} et 10^{-5} par génome diploïde et entre 10^{-4} et 10^{-6} par génome de vecteur (Kaeppel et al., 2013; Li et al., 2011b; Nowrouzi et al., 2012). Certaines études ont montré une intégration préférentielle au niveau des **gènes actifs** (Nakai et al., 2003), des **îlots CpG**, des **régions riches en GC** (Li et al., 2011b) et des **palindromes** de plus de 20 bases (Inagaki et al., 2007b). Les études sur le primate (Nowrouzi et al., 2012) et l'Homme (Kaeppel et al., 2013) n'ont en revanche pas retrouvé de sites d'intégration (SI) préférentiels. Les résultats de ces études sont cependant limités par le faible nombre de jonctions retrouvées.

L'analyse par de nouvelles méthodes de séquençage à haut débit du profil d'intégration d'un AAVr 2 dans différents types cellulaires a révélé un enrichissement des SI au niveau des marques épigénétiques associées à une activité transcriptionnelle (H3K4me3, H3K27ac) (Hüser et al., 2014). Il est intéressant de noter que dans ces régions, les cassures double brin de l'ADN sont favorisées par l'activité des topoisomérases (Haffner et al., 2011). L'intégration génomique des AAVr semble ainsi s'effectuer au niveau des cassures double brin pré-existantes de l'ADN, vraisemblablement par interaction avec les protéines permettant la réparation de ces lésions. Cette hypothèse est corroborée par l'observation *in vitro* d'une augmentation de la fréquence d'intégration des AAVr lors de l'induction de cassures double brin dans l'ADN par des rayonnements gamma, des enzymes de restriction et de l'étoposide (Miller et al., 2004). Par ailleurs, nous avons vu précédemment que dans la DMD, l'important stress oxydatif était responsable d'une augmentation du nombre de cassures double brin de l'ADN. Il est donc possible que la fréquence d'intégration de l'AAVr soit supérieure dans les phénotypes pathologies où le stress oxydatif est exacerbé.

Jusqu'à présent, aucun évènement génotoxique n'a été rapporté lors des essais cliniques utilisant des AAVr. Leur fréquence d'intégration génomique est si faible que le risque est considéré comme minime par rapport aux vecteurs intégratifs. Cependant, une étude publiée en 2007 dans *Science* par Donsante *et al*. a impliqué les AAVr dans le développement de **carcinomes hépatocellulaires** (CHC) (Donsante et al., 2007). Une forte dose de vecteurs (10^{14} vg/kg) avait été injectée en intraveineuse chez des souris modèles de mucopolysaccharidose de type VII âgées de 2 jours (Daly et al., 2001). Plusieurs SI avaient été retrouvés dans le locus Dlk1-Dio3 du chromosome 12, au niveau du gène de miR 341. Cette région complexe code pour de nombreux ARN non codants ayant un rôle régulateur important notamment lors de la différenciation cellulaire (Benetatos et al., 2013). L'implication de ce locus dans certains cancers humains (Kawakami et al., 2006) a mis en évidence la nécessité d'étudier le potentiel génotoxique des AAVr (**Tableau 3).** L'intégration des AAVr dans le locus Dlk1-Dio3 a été

observée une nouvelle fois par Zhong *et al.* en 2013 (Zhong et al., 2013) et par une étude récente qui a montré, pour la première fois, une augmentation dose-dépendante et promoteur-dépendante du risque tumoral après injection intrahépatique de souris en période néonatale (Chandler et al., 2015), définissant probablement le cadre général de la génotoxicité des AAVr (**Figure 14**). Le plus intéressant dans cette étude est la relation qu'elle souligne entre le niveau d'expression d'un gène et le nombre de SI détecté. Ainsi, l'albumine et l'alpha-foetoprotéine, les plus exprimées en période néo-natale, présentaient le plus grand nombre de SI, et c'est l'intégration dans le troisième gène le plus exprimé, Rian (locus Dlk1-Dio3), qui a une nouvelle fois été à l'origine des tumeurs. Cela conforte l'hypothèse d'une intégration des AAVr au niveau des cassures double brin de l'ADN engendrées lors d'une forte activité transcriptionnelle.

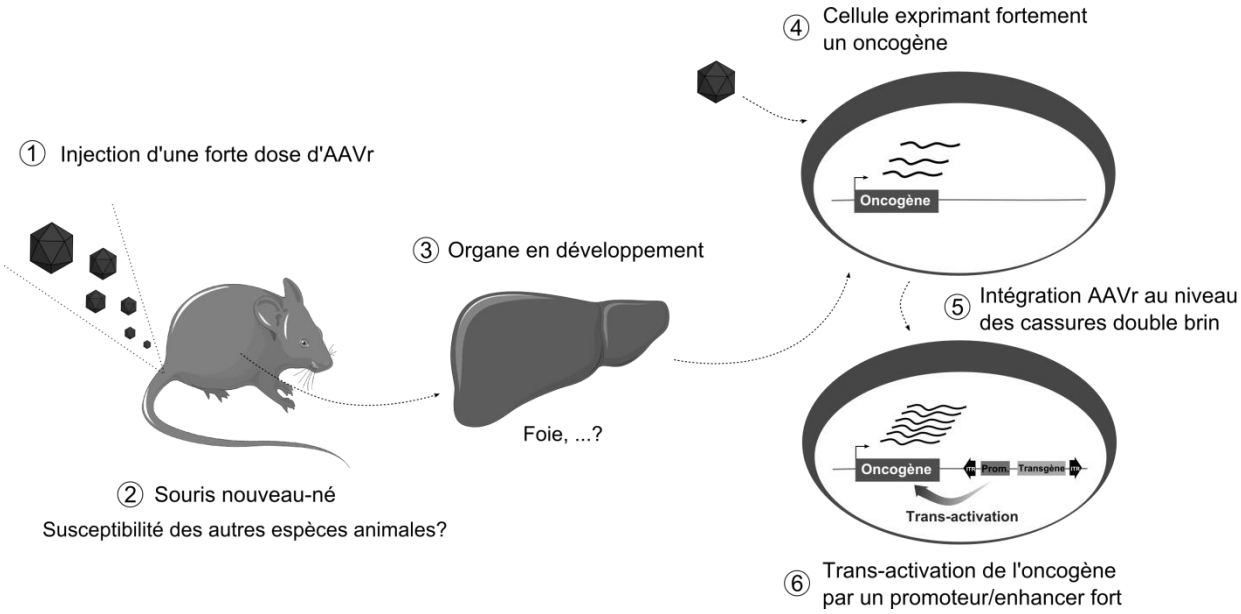


Figure 14 : Modèle de carcinogénèse après injection d'un AAVr. D'après les études pré-cliniques réalisées jusqu'à maintenant.

La majorité des essais pré-cliniques n'ont cependant pas mis en évidence de risque génotoxique (Bell et al., 2005; Gauttier et al., 2013; Li et al., 2011b; Nowrouzi et al., 2012) et il n'est pour l'instant pas déterminé si les observations faites chez la souris peuvent êtres extrapolées chez l'Homme. Cependant, la nécessité d'injecter de fortes doses d'AAVr chez de jeunes enfants, voire des nourrissons, pour traiter certaines maladies génétiques implique de continuer les études de génotoxicité dans différents modèles animaux et chez l'Homme.

Etude	Protocole	Techniques utilisées	Résultats
En faveur d'un risque génotoxique			
Donsante *et al.* 2001, Donsante *et al.* 2007	Injection **néonatale** IV, souris MPS VII, AAV βactin-CMV-GUSB - 1,5E11vg	Inverse-PCR, Séquençage Sanger	Souris MPSVII: 6/18 tumeurs hépatiques Vs 1/25 chez groupe contrôle. Excès de tumeur chez souris saines injectées (56% Vs 8%). 4 SI dans le locus Dlk1-Dio3 du chromosome 12: trans-activation
Rosas *et al.* 2012	Injection IV, souris SCID et C3H/HeJ de 8 à 12 semaines , 2E12vg/kg, scAAV CMV GFP ou scAAV-CBA-null	Inverse-PCR, LM-PCR, Séquençage Sanger	Pas d'augmentation des cancers chez SCID. Mâles C3H (souche prédisposée aux tumeurs hépatiques), groupe AAVr: CHC 14/25 Vs 4/21 groupe contrôle. scAAV-CMV-GFP: pas de SI retrouvés. scAAV CBA-null: SI près des oncogènes Hras1, Sos1, Fgf3, Fgf10.
Zhong *et al.* 2013, Souris issues de Bell *et al.* 2006	Injection IP, souris hybrides B6C3F1 Otc -/-, 1E12vg, AAVr 2/7/8 Otc ou lacz	LM-PCR, Pyroséquençage 454	Pas d'augmentation de l'incidence des CHC mais l'analyse des SI dans souris traitées par AAVr montre 4 SI dans locus Dlk1-Dio3 - 1 seul à env. 1 copie par cellule.
Chandler *et al.* 2014	Injection **néonatale** Intrahépatique, souris Mut-/- et +/-, AAVr 2/8/9, promoteurs: CBA/TBG/hAAT, transgènes: Mut et GFP, doses: 7E11 à 1E14 vg/kg	LM-PCR, Séquençage illumina	Augmentation CHC dans groupes traités par AAVr forte dose avec promoteur fort CBA. SI: Dlk1-Dio3 responsables trans-activation.
AAVr non génotoxique (liste non exhaustive)			
Bell *et al.* 2005	695 souris injectées à 6-8 sem, 7 sérotypes d'AAV et 8 transgènes. Souches: C57BL/6 et FIX KO, Rag-1, LDLR-déficient, NCr nude	Analyse histologique, Pas d'étude des sites d'intégration	Pas d'augmentation de l'incidence des tumeurs hépatiques par rapport au groupe contrôle de 226 souris C57BL/6.
Li *et al.* 2011	Injection IV, AAV2-AAT-hFIX, 80 mâles C57BL6, 2-3 mois, 60 souris à 5E12vg/kg et 20 à 1E14vg/kg	Histologie, LM-PCR, Pyroséquençage 454	Pas d'augmentation des tumeurs chez les groupes traités par AAVr (5,8% Vs 2%). Analyse SI: intégration préférentielle au niveau des gènes, des îlots CpG et des régions riches en GC. Fréquence d'intégration: 6E-4/dg et 7E-4/vg
Nowrouzi *et al.* 2012	Injection IM et IR, AAV 2/1 et AAV2/8 RSV GFP, 9 primates non-humains, 5E12 vg/kg	LAM-PCR, Pyroséquençage 454	Fréquence d'intégration de 10^{-4} à 10^{-6}/vg sans SI préférentiels.
Gauttier *et al.* 2013	Injection 10 rats Wistar nouveau né avec scAAV 2/8 RSV GFP, 5E12 vg/kg	LAM-PCR, Pyroséquençage 454	Pas d'augmentation des nodules pré-néoplasiques hépatiques par rapport au groupe PBS contrôle. Pas de SI préférentiels.
Kaeppel *et al.* 2013	Injection IM , 5 Hommes, AAV1-LPLS447X, 1E12 vg/kg	LAM-PCR, Séquençage illumina	Intégration aléatoire, enrichissement dans le génome mitochondrial

Tableau 3 : Résumé des principales études ayant étudiées le risque génotoxique des AAVr. IV : intraveineuse, IP : intrapéritonéale, IR : intraveineuse régionale, IM : intramusculaire.

I.2.2.7 Méthodes d'analyse des sites d'intégration génomique des AAVr

Les premières méthodes développées pour identifier les SI des vecteurs AAVr utilisaient le principe de "plasmid rescue". Les vecteurs contenaient une origine de réplication bactérienne qui permettait le sous-clonage des jonctions AAVr / génome cellulaire puis leur séquençage par technique Sanger (Inagaki et al., 2007b; Miller et al., 2005; Nakai et al., 2003). L'étape de sélection bactérienne entraînait cependant un biais lorsque la taille et la séquence des intégrants affectait négativement la croissance bactérienne. La structure secondaire de l'ITR est également

compliquée à séquencer par méthode Sanger (Penaud-Budloo et al., 2008). Cette méthode ne permettait donc pas une identification précise et exhaustive des SI des AAVr.

Par la suite, des méthodes plus élaborées basées sur la PCR couplée au séquençage à haut débit ont été développées afin de répondre à la nécessité d'identifier les SI des vecteurs rétroviraux responsables des leucémies dans plusieurs essais cliniques (**Tableau 2**). Les deux méthodes les plus employées sont actuellement la LM-PCR (*linker-mediated PCR*) (Li et al., 2011b) et la LAM-PCR (*linear amplification mediated PCR*) (Gauttier et al., 2013; Nowrouzi et al., 2012; Wang et al., 2012) (**Figure 15**). Elles ont été utilisées avec succès pour détecter les SI des rétrovirus permettant d'appréhender avec précision le risque génotoxique.

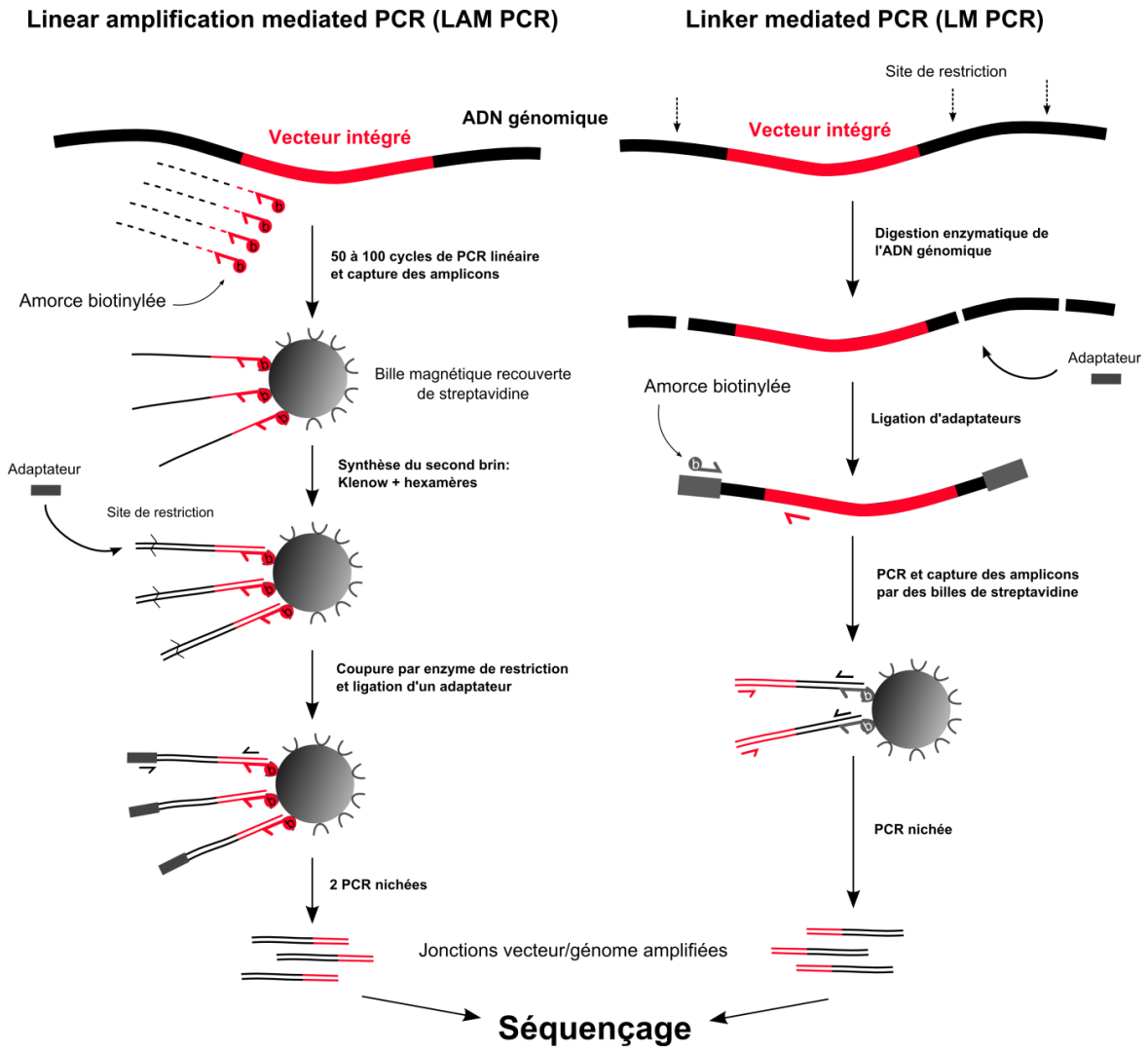


Figure 15 : Schéma de la LAM-PCR et de la LM-PCR.

Lorsque les premières études ont mis en évidence un possible effet génotoxique des AAVr, la LAM-PCR et la LM-PCR ont commencé à être utilisées pour ces vecteurs sans réelle optimisation. Cependant, le génome de l'AAVr présente des formes moléculaires de persistance variées et des particularités de séquence qui compliquent l'analyse de leur SI : (i) les formes

épisomales sont 10^4 à 10^5 fois plus nombreuses que les rares formes intégrées, (ii) les ITR présentent une forte structure secondaire difficile à traverser par les polymérases classiquement utilisées et (iii) le génome AAVr est souvent tronqué lors de l'intégration génomique. Ainsi, lors de l'utilisation de la LM ou de la LAM-PCR, les épisomes sont majoritairement amplifiés au détriment des formes intégrées. D'autre part, les génomes AAVr présentant des ITR complets sont difficilement amplifiables et les formes tronquées peuvent être ignorées si elles ne contiennent plus la zone d'hybridation des amorces de PCR. Enfin, l'utilisation d'enzymes de restriction ne permet pas une détection exhaustive des SI dans le génome. Dans ces conditions, un faible nombre de SI a pu être identifié *in vivo*. Par exemple, l'analyse par LAM-PCR de muscles et de foies de primates après injection d'un AAVr en intramusculaire ou en intraveineuse régionale n'a permis d'identifier que de 3 à 34 SI par tissu (Nowrouzi et al., 2012).

Les méthodes actuelles doivent donc être améliorées pour permettre d'évaluer avec précision le risque intégratif associé à une thérapie génique *in vivo* par AAVr. Les évolutions des techniques de biologie moléculaire et de séquençage d'ADN permettent, aujourd'hui, d'envisager des méthodes de détection des SI plus adaptées.

I.3 Le séquençage d'acides nucléiques

I.3.1 Séquençage de première génération

En 1977, l'équipe de Frederick Sanger a décrit une nouvelle technique de séquençage de l'ADN utilisant des nucléotides terminateurs de chaines (Sanger et al., 1977a). Ces 2'3' dideoxynucléotides, qui ont la particularité de ne pas posséder de groupement OH en 3', bloquent la progression de la polymérase. La séquence d'ADN était déduite par migration des fragments générés sur gel d'agarose. Cette équipe a séquencé le premier génome en 1977 (le bactériophage ΦX174, 5386pb) (Sanger et al., 1977b) puis le bactériophage lambda et le génome mitochondrial humain en 1980. Pour ces derniers, ils ont séquencé pour la première fois des librairies générées aléatoirement (« shotgun library »), principe largement utilisé par la suite pour séquencer des organismes plus complexes comme Haemophilus influenzae (1,8Mb) en 1995 (Fleischmann et al., 1995). Les améliorations de la technologie initiale de séquençage avec l'utilisation de ddNTP marqués avec 4 fluorochromes (Prober et al., 1987), de l'électrophorèse capillaire (Luckey et al., 1990) et de l'automatisation (Meldrum et al., 2000) ont permis le séquençage du **génome humain en 2001** après une dizaine d'années d'étude par une collaboration internationale (Human Genome Project) (Lander et al., 2001) et une société privée (Celera Genomics) (Venter et al., 2001). Le coût du séquençage d'un génome humain était alors d'environ 100 millions de dollars.

I.3.2 Séquençage à haut débit ou séquençage nouvelle génération (NGS)

I.3.2.1 Les différentes plateformes NGS

Les développements de la microfluidique, de la luminométrie et de la bio-informatique ont révolutionné les techniques de séquençage d'ADN il y a une dizaine d'années. En 2005, Margulies *et al.* ont décrit une nouvelle méthode de séquençage basée sur une évolution du pyroséquençage initialement décrit en 1996, nommé **séquenceur 454** (Margulies et al., 2005). L'ADN était d'abord fragmenté aléatoirement et des séquences universelles ou « adaptateurs » étaient liguées. Après une PCR en émulsion sur des billes (qui remplace alors le clonage bactérien), celles-ci étaient individualisées dans des puits d'un volume de l'ordre du picolitre sur une lame de fibre optique, ce qui permettait d'appliquer le pyroséquençage en parallèle sur des milliers de fragments. Cette technologie est à l'origine du séquençage haut débit ou nouvelle génération (NGS pour *next generation sequencing*). Le marché du NGS a dès lors explosé et de nombreuses technologies concurrentes se sont développées, au dépend du 454 qui tend à disparaitre. Le séquençage **Illumina** (Bentley et al., 2008) est aujourd'hui largement dominant

devant les plus récents **Ion Torrent** et **Pacific Biosciences** (PacBio). Les différentes étapes de ces trois technologies sont décrites sur la **figure 16**.

Toutes les technologies NGS débutent par une étape de préparation de librairie qui aboutit à l'ajout de séquences connues ou « adaptateurs » sur l'ADN double brin fragmenté. Ensuite, pour les séquençages dits de seconde génération, une étape d'amplification clonale sur surface solide (Illumina) ou sur bille (Ion Torrent) est nécessaire. Au contraire, le séquençage de troisième génération (PacBio) possède un système de séquençage suffisamment sensible pour utiliser un seul fragment comme matrice. Lors de la réaction de séquençage, le signal émis par l'incorporation des nucléotides est de nature physique par émission de fluorescence (Illumina, PacBio) ou chimique par détection d'une variation de pH (Ion Torrent). Le **tableau 4** compare les principales propriétés de chacune de ces plateformes. Les séquenceurs Illumina, principalement grâce à leur débit très élevé qui leur permet d'être utilisés dans de nombreuses applications, sont aujourd'hui les plus utilisés. Les termes techniques spécifiques du NGS sont expliqués sur la **figure 17.**

	Illumina		Ion Torrent		PacBio RSII
	MiSeq	HiSeq 2500	PGM318	Ion Proton	
Coût de l'instrument	$125 000	$740 000	$50 000	$149 000	$700 000
Coût d'un *run*	$1400	$8000 (R) $29 000 (HD)	$750	$1000	$400
Durée *run*	48h-72h	40h (R) - 6j (HD)	4-7h	2-4h	30min-4h
Taille insert	150-1000pb		200-400pb		200pb à 20-30kb
Taille *reads*	2x300pb	2x150pb (R) 2x125pb (HD)	400pb	200pb	10-15kb (moyenne)
Nombre *reads* max	25 millions	600 millions (R) 4 milliard (HD)	5,5 millions	82 millions	50 000
Taux d'erreur	0,8% (substitutions)		1,7% (Insertions/délétions)		13% (Insertions/délétions)
Applications de séquençage	Petits génomes Exome Ciblé Chip-Seq	Grands génomes Exome Transcriptome Métagénomique	Petits génomes Ciblé	Exome Petits génomes Transcriptome Chip-seq	Petits génomes Ciblé
Avantages	Facilité d'accès à une technologie très répandue Gros débit de séquençage *paired-end* Faible taux d'erreur		Prix Rapidité		Longs reads qui facilitent l'assemblage des séquences Bonne couverture des régions riche en GC et AT
Inconvénients	Longueur des *runs* Difficulté de séquençage des homopolymères >20 (répétitions de la même base)		Encore peu d'appareils sur le marché Difficulté de séquençage des homopolymères >8 et régions riches en AT		Prix machine, taux d'erreur, faible débit Technologie peu dispoinible

Tableau 4 : Comparatif des trois principales technologies de NGS. (R) : mode rapide, (HD) : mode haut débit. Illumina propose d'autres appareils (NextSeq500, HiSeq X Ten…) dont le débit de séquençage varie pour s'adapter aux différentes besoins des laboratoires. Données issues de http://allseq.com/knowledgebank/sequencing-platforms et de Quail *et al.* 2012.

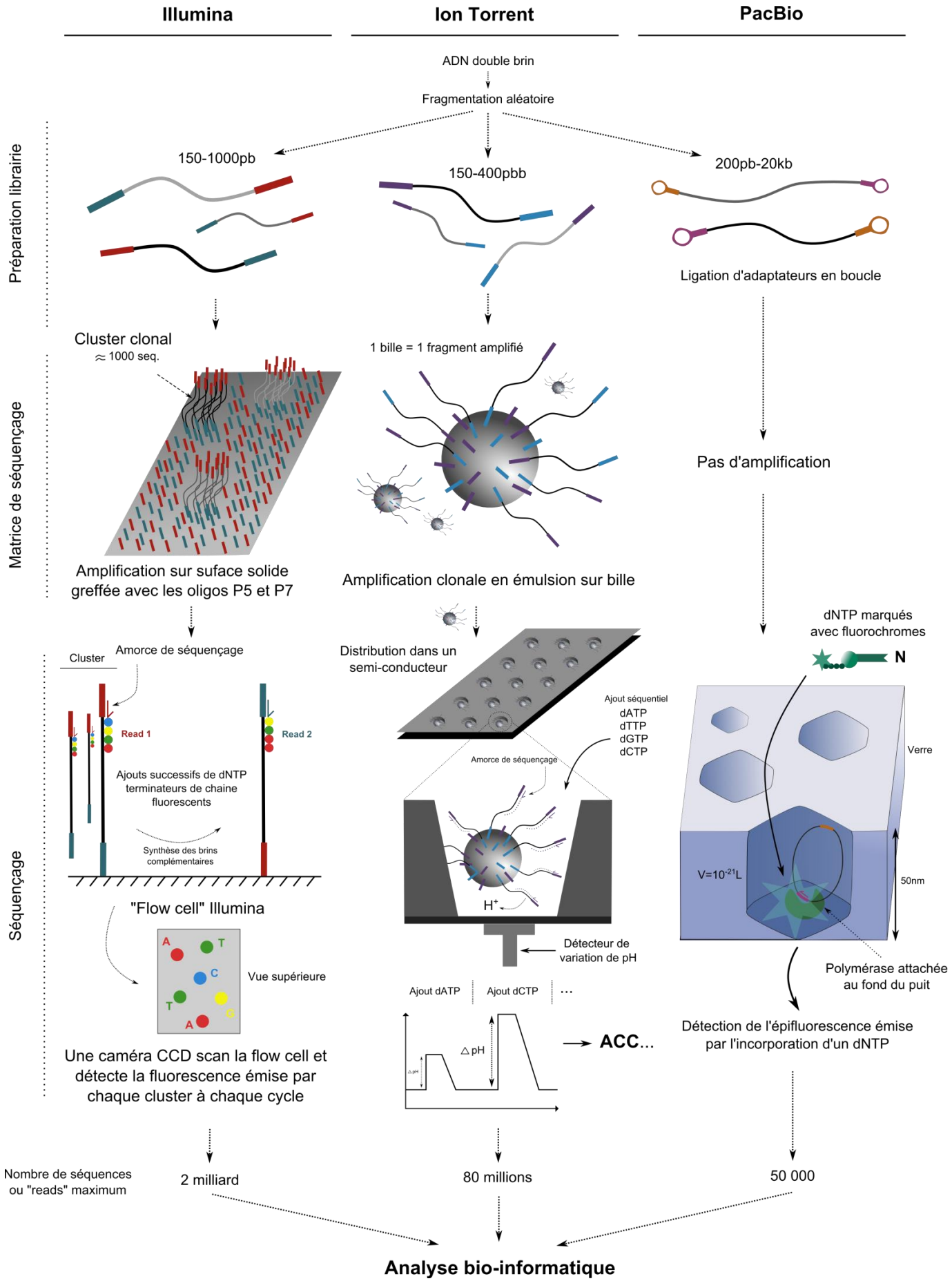


Figure 16 : Schéma des trois principales méthodes de séquençage haut débit.

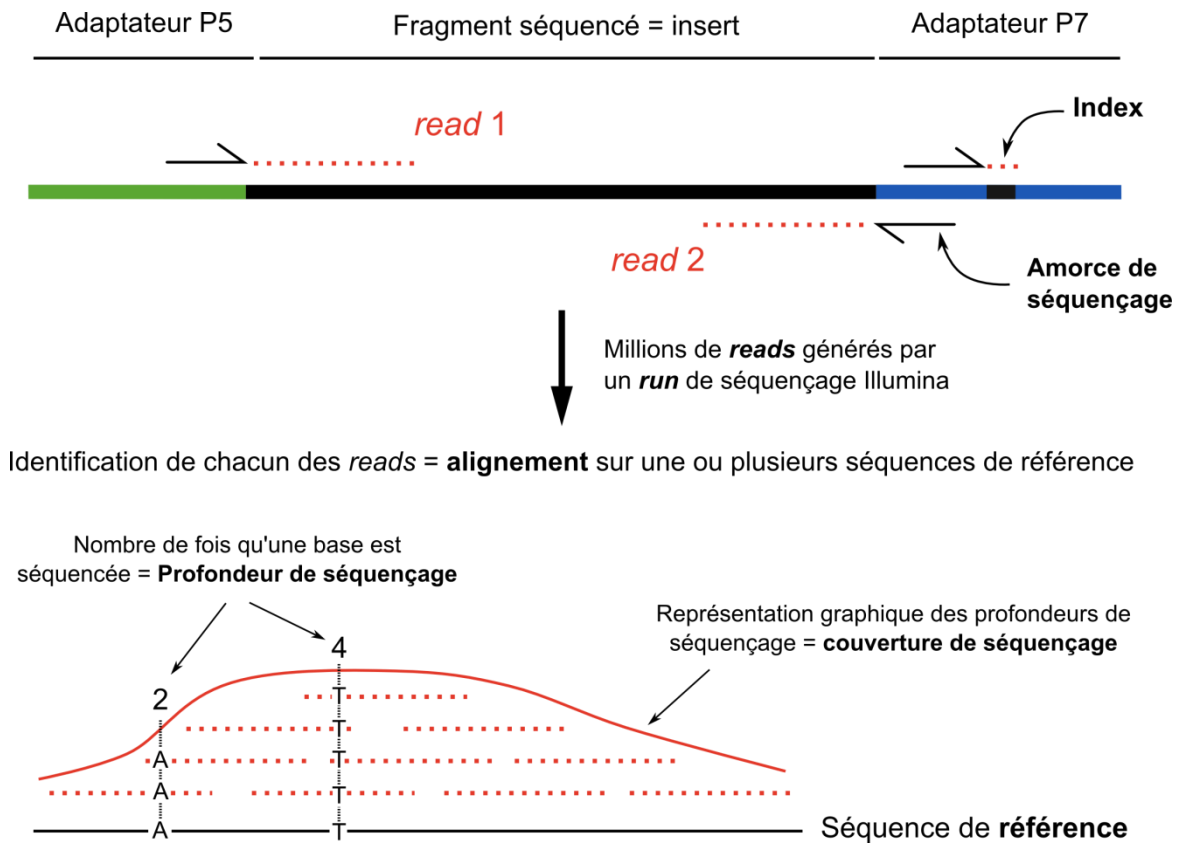


Figure 17 : Terminologie du NGS : exemple du séquençage Illumina. Pour qu'un fragment d'ADN puisse être séquencé par NGS, il doit posséder des séquences à ses extrémitées (**adaptateurs**), spécifiques de la technologie de séquençage utilisée (ici P5 et P7 sont les noms donnés aux adaptateurs Illumina)**.** Le fragment est ensuite nommé **insert**. Les séquences issues du séquençage sont appelées ***reads***, ils peuvent être d'une taille choisie entre 50 et 350 bases (Illumina)**.** Un séquençage ***paired-end*** correspond au séquençage des deux extrémités de l'insert, ce qui génère un ***read* 1** associé à un ***read* 2.** Un séquençage ***single-end*** correspond au séquençage d'une seule extrémité (*read* 1 uniquement). L'**index** dans l'adaptateur correspond à une séquence connue de 6 à 8 bases, il permet d'identifier l'origine des *reads*. Lorsque différents index sont attribués à plusieurs échantillons, ces échantillons peuvent être mélangés et séquencés en même temps (**multiplexage**). Le séquençage en parallèle de millions de fragments sur la *flow-cell* illumina lors d'un ***run*** de séquençage génère des millions de *reads*. Ensuite, par **bio-informatique**, les *reads* sont **alignés** sur une séquence connue de **référence**. Les caractéristiques de cet alignement (ex : coordonnées de début et fin d'alignement) permettent de calculer une **profondeur de séquençage** pour chaque base de la séquence de référence et de représenter la **couverture de séquençage** de la référence.

I.3.2.2 Le séquençage ciblé

La réduction des coûts de séquençage a dans un premier temps été permis par le « multiplexage » de plusieurs échantillons au sein d'un même *run*. Il consiste à ajouter dans les adaptateurs une séquence spécifique « code barre » qui permet de relier la séquence à un échantillon. De plus, la fraction du génome humain codant pour des protéines, qui est la partie la plus pertinente pour le diagnostic des maladies génétiques, est de l'ordre de 1%. Ainsi, afin de séquencer uniquement ces régions, Gnirke *et al.* ont mis au point en 2009 une technique de capture d'ADN permettant de séquencer seulement ce qu'ils nomment l'exome (Gnirke et al., 2009). Pour ce faire, ils ont synthétisé des milliers d'ARN biotinylés (ou sondes) complémentaires des séquences codantes du génome permettant leur extraction avec des billes magnétiques recouvertes de streptavidine puis leur séquençage (**Figure 18**). Le multiplexage

de plusieurs exomes permet, en plus de réduire les coûts, de simplifier l'analyse bio-informatique. De nombreuses maladies génétiques sont actuellement diagnostiquées grâce à cette technologie (Bolze et al., 2010; Isidor et al., 2011; Mercier et al., 2013). Cette capture est aujourd'hui possible avec des sondes ADN et ARN, en solution (Agilent - Sureselect, Mycroarray - Mybaits) et sur puce (Roche - Nimblegen).

Les séquences des sondes pouvant être modifiées pour cibler n'importe quelle séquence d'ADN, d'autres applications ont été développées. Duncavage *et al.* ont identifié les sites d'intégration génomique du polyomavirus de Merkel dans des tissus tumoraux cutanés après capture du génome viral par des sondes d'ADN biotinylées générées par PCR (Duncavage et al., 2011). Certains fragments capturés, donc présentant du génome viral, comportaient une jonction génomique (partie qualifiée de « off-target ») dont les coordonnées génomique ont été identifiées par bio-informatique. La même méthode a été appliquée en 2012 pour identifier les sites d'intégration de vecteurs lentiviraux en culture cellulaire (Ustek et al., 2012). Après hybridation avec des sondes d'ADN fixes (Nimblegen) complémentaires du génome du vecteur et séquençage par pyroséquençage 454, ils ont obtenu 9,5% des séquences totales comportant le génome du vecteur. Il est donc possible d'utiliser cette technologie pour étudier les sites d'intégration de n'importe quel transgène dans un génome (Dubose et al., 2013).

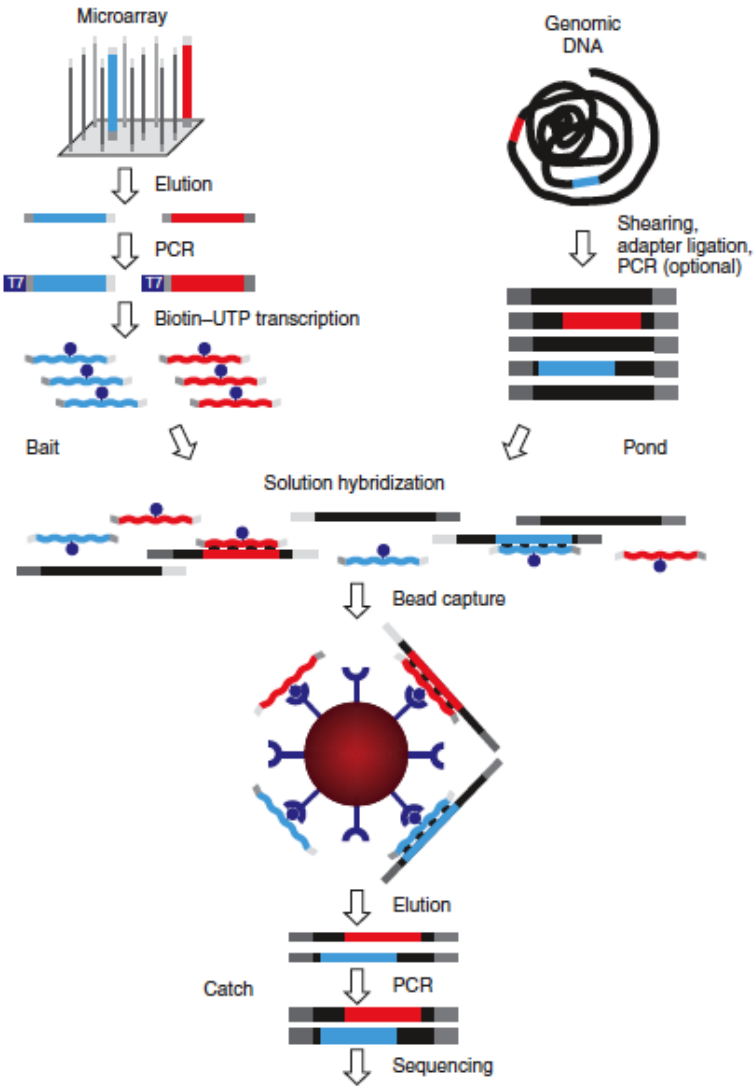


Figure 18 : Séquençage ciblé par hybridation de sondes d'ARN biotinylées. Des oligonucléotides d'ADN (100 – 200pb) complémentaires de séquences d'intérêt sont synthétisés en parallèle sur une puce. Une transcription *in vitro* est ensuite réalisée avec une ARN polymérase T7 en incorporant de l'uracile biotinylé. Les sondes générées sont ensuite incubées avec une librairie NGS (ADN fragmenté avec des adaptateurs ligués). Les hybrides ADN/ARN biotinylés sont extraits avec des billes magnétiques recouvertes de streptavidine puis, après une étape de PCR, la librairie enrichie en séquences d'intérêt est séquencée par NGS. Tiré de Gnirke *et al.* 2009.

I.3.2.3 Applications du NGS

L'augmentation des débits de séquençage a dans un premier temps permis de séquencer les génomes de nombreux organismes (près de 3000 espèces eucaryotes en 2014) (Reddy et al., 2014) et d'améliorer le diagnostic des maladies génétiques. Cependant, la plupart des génomes séquencés, notamment humain, sont en réalité un consensus de plusieurs individus qui ne reflète pas les nombreux polymorphismes présents dans les populations. Une avancée majeure a été réalisée en 2008 avec le séquençage en deux mois du génome personnel de James D Watson par pyroséquençage 454 (Wheeler et al., 2008), qui a révélé plus de 3 millions de polymorphismes nucléotidiques et de nombreux réarrangements par rapport à la séquence de référence. Depuis, le cap du séquençage de génomes individuels humains pour 1000 dollars a

été franchi en 2014 par un nouveau modèle de séquenceur à très haut débit (HiSeq X Ten, Illumina). Un centre qui possède ce type de séquenceurs est désormais capable de séquencer près de 18000 génomes humains par an. Avec de tels débits, les applications du NGS s'en trouvent élargies notamment avec la médecine personnalisée, dont le but peut être de prédire l'activité d'un médicament sur une personne pour mieux anticiper son efficacité et ses éventuels effets indésirables (pharmacogénomique) ou encore de développer des traitements anticancéreux ciblant spécifiquement le type de tumeur présente chez le patient. Le but ultime étant de séquencer les génomes directement « au lit du patient ». Le développement actuel de techniques de séquençage par nanopores (Ashton et al., 2015), dont l'appareillage est de la taille d'une clé USB, fait figure de pionnier dans le domaine. Parallèlement à toutes ces avancées, l'extraction d'une information fiable et pertinente des importantes données générées par analyse bio-informatique est un enjeu crucial qui nécessite toujours plus d'espace de stockage et de puissance de calcul.

La diminution des coûts de séquençage a également permis à d'autres domaines que la génétique et la cancérologie de bénéficier de ces avancées. En bactériologie, le séquençage de micro-organismes directement dans leur environnement complexe naturel (métagénomique) révèle leur incroyable biodiversité. La virologie bénéficie également du génotypage rapide de populations virales. Par exemple, l'analyse d'une centaine de génomes de virus Ebola isolés à partir de différents patients a permis d'établir l'origine de l'épidémie actuelle et les mutations virales associées à son évolution (Gire et al., 2014).

D'un point de vue thérapeutique, le NGS est un outil important en thérapie génique. Il a permis d'analyser et de comprendre les mécanismes génotoxiques à l'origine des leucémies provoquées par les premières générations de vecteurs rétroviraux (Schmidt et al., 2007). Cependant, le NGS n'est pas encore utilisé en routine dans les laboratoires de thérapie génique. La technologie est récente et commence tout juste à être accessible à un coût raisonnable. De plus, il n'est pas toujours facile de rassembler les compétences biologiques et bio-informatiques au sein d'un même laboratoire, nécessaires au développement de nouveaux protocoles techniques et informatiques adaptés au type de données générées. Ainsi, un travail important est encore à réaliser dans le domaine pour tirer le maximum des capacités du NGS.

II. Objectifs de l'étude et méthodologie

Les récents succès cliniques obtenus en thérapie génique *in vivo* avec les AAVr (MacLaren et al., 2014; Maguire et al., 2008; Nathwani et al., 2014) concrétisent des décennies de recherche fondamentale pour trouver des solutions thérapeutiques aux maladies génétiques. Demain, il est possible qu'un grand nombre de patients soient traités avec ces nouveaux agents, y compris pour des maladies à plus forte prévalence (maladies infectieuses, cancéreuses…). Dans cette optique, l'efficacité du transfert de gène doit encore être amélioré, notamment en développant de nouveaux sérotypes d'AAV qui possèderaient des capacités de transduction plus efficaces et plus spécifiques d'un tissu cible (Lisowski et al., 2014) et en trouvant des solutions aux limites imposées par la réaction immunitaire anti-vecteur (Manno et al., 2006). Parallèlement à ces avancées thérapeutiques, la **biosécurité** de ces nouveaux médicaments doit être minutieusement étudiée. Leur origine virale implique un mode de production complexe et des effets biologiques potentiellement délétères (immunotoxicité et génotoxicité) qui n'ont pas d'équivalent dans l'arsenal thérapeutique actuel. Ces différentes problématiques sont exacerbées par la nécessité d'injecter des doses massives d'AAVr pour obtenir l'effet thérapeutique escompté.

Les récentes avancées du NGS offrent l'opportunité d'étudier avec plus de précision les risques biotechnologiques associés à l'utilisation des AAVr. **Mon projet de thèse a donc consisté à développer de nouvelles méthodes NGS permettant (1) de détecter efficacement les sites d'intégration génomiques des AAVr et (2) de caractériser l'ADN encapsidé dans les particules AAVr pendant leur production.**

Pour ce travail, nous avons choisi d'utiliser la technologie Illumina car c'est la plateforme NGS qui permet de générer le plus de *reads*, paramètre important dans nos deux objectifs pour identifier les évènements les plus rares. Nous avons donc collaboré avec l'Infrastructure Nantaise de Génomique et Bio-informatique qui possède les séquenceurs MiSeq et HiSeq d'Illumina. La première étape de ce travail a été de mettre en place la préparation des échantillons pour permettre leur séquençage sur plateforme Illumina. Cette **préparation de librairie NGS** a été commune aux deux protocoles plus spécifiques que nous avons mis en place par la suite et qui sont décrits sur la **figure 19**.

- **<u>Projet 1</u> : Détection des sites d'intégration (SI) génomiques des AAVr**

Les limites des techniques de PCR actuellement utilisées pour détecter les SI des vecteurs viraux nous ont amené à développer une nouvelle méthode basée sur l'utilisation du séquençage

ciblé. Cette approche permet de séquencer uniquement des séquences d'intérêt et a déjà été utilisée pour étudier l'intégration des virus sauvages (Duncavage et al., 2011) et des vecteurs lentiviraux (Ustek et al., 2012) mais jamais sur des vecteurs AAVr. Cette technologie présente plusieurs avantages théoriques : (i) la couverture complète du génome en évitant l'utilisation d'enzymes de restriction, (ii) la réduction du nombre d'étapes de PCR qui peuvent biaiser l'amplification à travers la forte structure secondaire de l'ITR et (iii) la détection des génomes intégrés partiellement délétés par capture de la totalité du génome. Nous avons donc adapté le **séquençage ciblé** sur les AAVr *in vivo* pour permettre une analyse non biaisée des SI génomiques.

Nous avons choisi d'évaluer ce protocole dans un contexte pertinent pour le laboratoire, à savoir l'évaluation du risque génotoxique dans la DMD. En effet, suite aux succès pré-cliniques obtenus chez le chien GRMD (modèle de DMD) avec les stratégies du saut d'exon (Le Guiner et al., 2014) et de microdystrophine (Shin et al., 2013), des essais cliniques sont programmés dans les années à venir. Jusqu'à présent, aucune donnée n'a été générée sur le risque intégratif des AAVr dans un contexte déficient en dystrophine. En effet, les muscles dystrophiques présentent d'importants désordres métaboliques avec notamment un fort impact du stress oxydatif sur l'intégrité du génome cellulaire (Schmidt et al., 2011). L'augmentation du nombre de cassures double brin de l'ADN pourrait augmenter la fréquence d'intégration des AAVr donc également leur potentiel génotoxique, dans une pathologie où les patients sont déjà susceptibles aux tumeurs (Schmidt et al., 2011).

Pour évaluer les caractéristiques de l'intégration génomique des AAVr dans un contexte dystrophique, nous avons travaillé avec des **souris mdx^{4CV}**, un modèle de DMD couramment utilisé qui reproduit certaines caractéristiques importantes de la maladie, notamment au niveau du stress oxydatif, en comparaison à un groupe contrôle composé de souris saines C57BL/6J qui présentent le même fond génétique que les mdx^{4CV}. Un AAVr sinple brin de serotype 8 codant pour la GFP sous contrôle du promoteur ubiquitaire RSV (**ssAAVr 2/8 RSVp GFP)** a été choisi pour cette étude. Le sérotype 8 possède un bon tropisme pour le muscle squelettique et est largement utilisé lors des études pré-cliniques et cliniques. Le choix d'un transgène rapporteur a été fait pour évaluer le risque intégratif dans un contexte défavorable sans amélioration phénotypique éventuelle apportée par un transgène thérapeutique. Les mises au point de la méthode ont été réalisées sur un nombre limité de souris, avant une éventuelle application à plus large échelle sur différents modèles animaux

- **<u>Projet 2</u> : Caractérisation de l'ADN encapsidé dans les particules AAVr**

La complexité de production et de purification des vecteurs AAVr fait qu'aujourd'hui, les patients reçoivent un mélange complexe de particules thérapeutiques et de molécules ADN indésirables. Ces séquences contaminantes peuvent contenir des gènes de résistance aux antibiotiques, des gènes d'origine virale et potentiellement, des oncogènes ou des gènes codant pour des ARN possédant une activité de régulation transcriptionnelle. La persistance à long terme de ces séquences chez les patients est potentiellement un important problème de biosécurité et les agences réglementaires sont très clairement demandeuses de tests de contrôle qualité plus évolués pour les des futures demandes de mise sur le marché de traitement de thérapie génique par des vecteurs AAVr (Bryant et al., 2013).

Aujourd'hui, les contaminants ADN sont quantifiés par qPCR. C'est une technique simple, peu coûteuse, rapide et très répandue dans les laboratoires de biologie moléculaire. Cependant, la qPCR nécessite le « design » d'amorces spécifiques pour chaque cible préalablement déterminée. Cette méthode ne permet donc de mettre en évidence que les contaminants spécifiquement recherchés. De plus, une étude internationale impliquant 16 laboratoires qui ont dosé par qPCR un même lot de vecteur AAVr a révélé l'importante variabilité inter-laboratoire des titres obtenus (Ayuso et al., 2014). Cette variabilité peut être expliquée par le système de quantification de la qPCR qui est basé sur une comparaison avec une gamme plasmidique elle-même souvent dosée par un système photométrique imprécis.

Dans ce contexte, le NGS apparait comme un complément idéal de la qPCR pour caractériser les contaminants ADN dans les productions d'AAVr. Il permettrait de mettre en évidence des contaminants non ciblés par la qPCR (exemple : couverture de la totalité du génome des cellules de production) et de quantifier de façon relative les différentes populations ADN au sein d'un même *run*, sans comparaison indirecte avec une gamme. Ce projet a donc consisté à adapter la technologie NGS afin d'améliorer notre connaissance de l'ADN encapsidé pendant la production des AAVr.

Ce projet a consisté à développer un protocole technique prenant en compte les caractéristiques des productions d'AAVr (présence d'ADN à l'extérieur des capsides, génome de l'AAVr simple brin…) pour permettre leur séquençage par NGS. Après séquençage Illumina, une analyse bio-informatique a été développée pour quantifier les différents constituants (génome AAVr et contaminants), analyser les séquences préférentiellement encapsidées et comparer les résultats obtenus avec la qPCR. La preuve de concept du protocole développé a été réalisée sur un **AAV 2/8 CMV GFP hTK** produit en transfection transitoire.

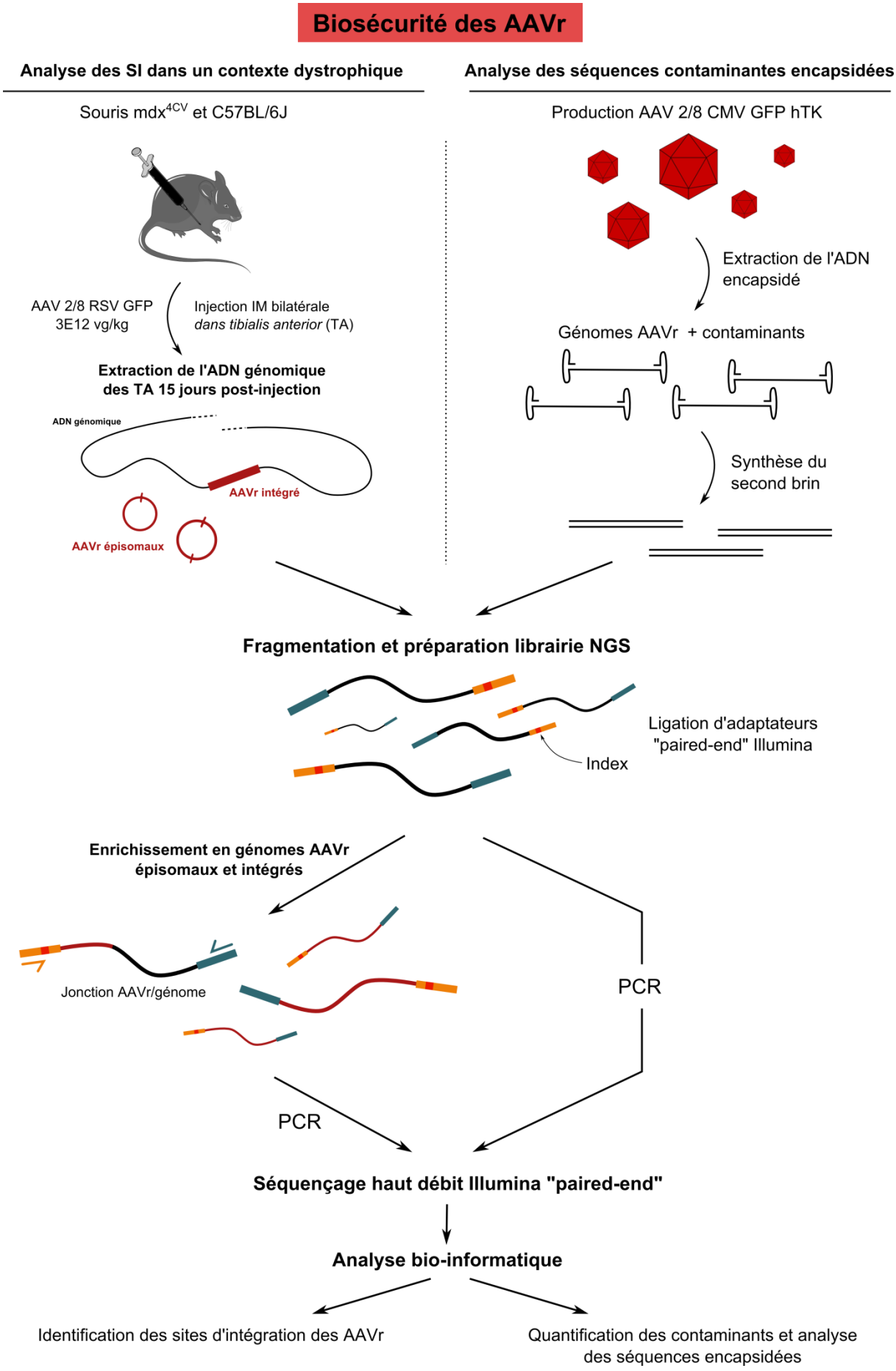


Figure 19 : Protocoles NGS développés pendant ma thèse.

III. Résultats

III.1 Mise au point de la préparation de librairie NGS

Le séquençage d'ADN par les plateformes utilisant la technologie Illumina requiert une préparation spécifique des échantillons, qui permet de générer des librairies NGS. Elles sont composées d'ADN double brin fragmenté (taille maximale de 1000pb) comportant de chaque côté des séquences spécifiques (ou adaptateurs) permettant l'hybridation, l'amplification et le séquençage sur la « flow cell » illumina. Les adaptateurs peuvent être ajoutés par PCR (séquence des adaptateurs ajoutée en 5' des amorces), par transposition (ex : Nextera, Agilent) ou par ligation enzymatique (ex : TruSeq, Illumina). Cette dernière option a été choisie car elle est applicable sur différentes quantités d'ADN de départ et adaptée aux deux projets NGS que nous avons développés.

La première étape a consisté à fragmenter aléatoirement l'ADN. Nous avons choisi une méthode de sonication par ultrasons en bain (Bioruptor) qui permet de fragmenter 12 échantillons en parallèle. Le type de tube utilisé pendant la fragmentation s'est avéré très important puisque les premiers tests effectués ont montré des tailles de fragmentation très variables. L'utilisation des tubes recommandés par le fournisseur ayant une épaisseur de plastique calibrée a permis d'obtenir des résultats satisfaisant. La **figure 20** représente la reproductibilité de fragmentation obtenue avec les conditions optimales déterminées pour une taille centrée sur 500pb (tests effectués sur de l'ADN génomique extrait de muscles de souris). Cette taille permettant de récupérer un nombre optimal de jonctions AAVr/génome cellulaire, ces paramètres ont été utilisés pour fragmenter l'ADN génomique dans l'étude des SI des AAVr. Les conditions ont été adaptées pour fragmenter l'ADN extrait des particules AAVr autour de 300pb.

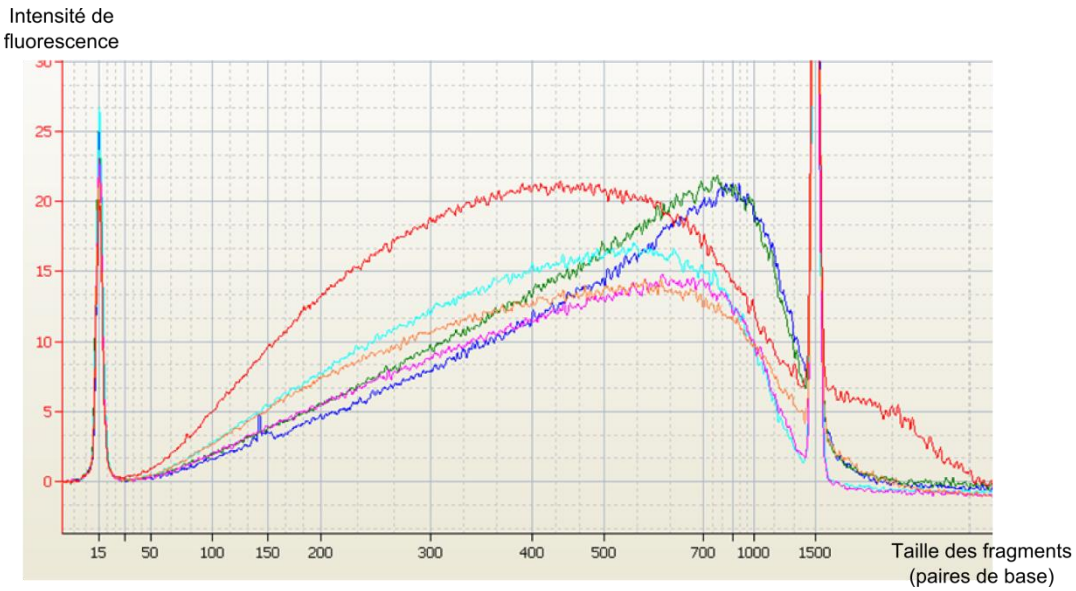


Figure 20 : Tests de fragmentation d'ADN génomique par le Bioruptor. Six échantillons d'ADN génomique extrait de muscles de souris ont été fragmentés avec les conditions suivantes : 4 cycles 30s ON/ 90s OFF, faible intensité, tubes diagénode 0,5mL. Les tracés de migration sur Bioanalyzer 2100 (puce D1K) sont représentés. Les pics à 15pb et 1500pb correspondent aux marqueurs de taille ajoutés avec les échantillons pour comparer les différents profils de migration.

Bien que de nombreux kits commerciaux soient disponibles pour préparer des librairies NGS Illumina, nous avons fait le choix d'adapter un protocole récemment publié (Kozarewa and Turner, 2011), afin de disposer d'une plus grande flexibilité de mise au point. La préparation des librairies NGS a ainsi été réalisée par plusieurs réactions enzymatiques permettant successivement de réparer et de phosphoryler les extrémités des fragments générés par la sonication, d'ajouter une Adénine en 3', de liguer des adaptateurs illumina et enfin d'amplifier les librairies pour obtenir assez de matériel pour le NGS (**Figure 21**). Entre chaque étape, une purification par des billes magnétiques (SPRI) permet le changement de milieu réactionnel et laisse la possibilité d'effectuer une sélection de taille (élimination des petits et/ou des grands fragments) de façon simple, avec un meilleur rendement que la traditionnelle extraction sur gel d'agarose.

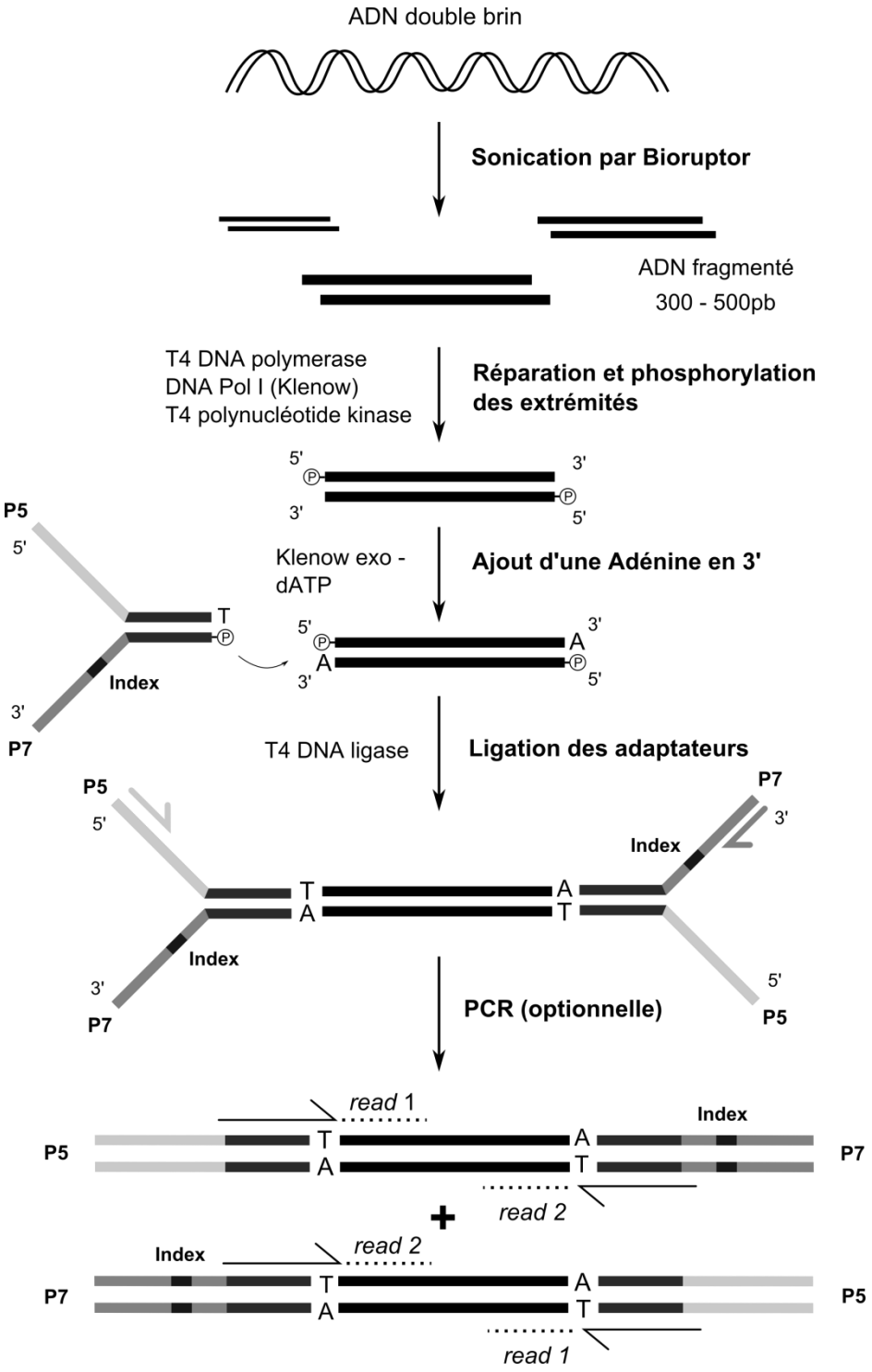


Figure 21 : Etapes de préparation des librairies pour le séquençage Illumina. L'ajout d'une adénine permet d'augmenter l'efficacité de la ligation avec l'adaptateur qui possède un T débordant.

L'adaptateur utilisé (**Figure 22**) contient les séquences permettant l'amplification des fragments sur la puce Illumina (génération des « clusters »), le séquençage de leurs deux extrémités (séquençage « paired-end ») et la lecture d'un index de 6 bases permettant de mélanger plusieurs échantillons dans un même *run* de séquençage (multiplexage). Ce type d'adaptateur rend optionnel l'amplification finale par PCR si la quantité d'ADN de départ est suffisante (>1µg) car leur structure en « Y » apporte toutes les séquences nécessaires au

séquençage Illumina (P5 et P7). Le protocole sans PCR est particulièrement intéressant pour éviter les biais induits par une étape d'amplification : induction de mutations par la polymérase, difficulté d'amplification de certaines structures ADN et génération de duplicats de PCR (fragments identiques séquencés).

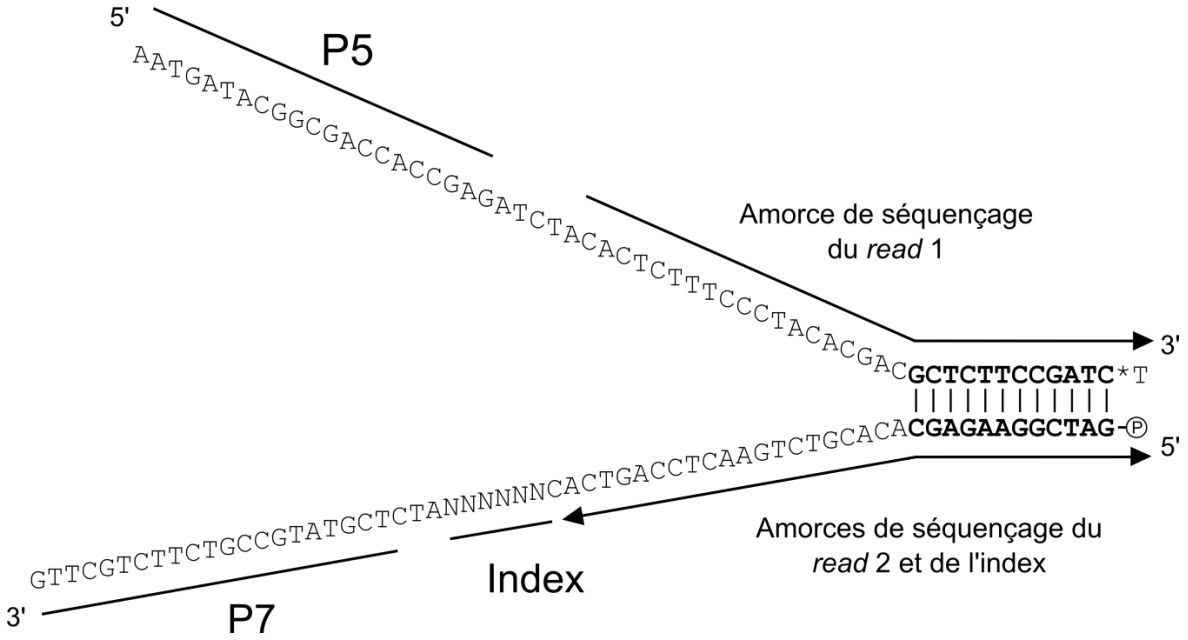


Figure 22 : Séquence de l'adaptateur Illumina « paired-end » utilisé. P5 et P7 correspondent aux séquences permettant l'hybridation et la formation des *clusters* sur la « flow cell » d'Illumina. Les zones d'hybridation des amorces de séquençage sont indiquées. * : liaison phosphorothioate qui permet d'éviter une dégradation prématurée du T débordant par des nucléases. P : Phosphate permettant la ligation sur les fragments d'ADN double brin.

L'efficacité des différentes réactions enzymatiques a été contrôlée par analyse du profil de migration après chaque étape (**Figure 23**). Le décalage de migration après ligation des adaptateurs correspond à la ligation des adaptateurs de part et d'autre des fragments (ajout de 122 bases). La quantification des aires sous courbe a permis d'évaluer le rendement de chaque cycle de réaction enzymatique/purification à 80%. Les fragments sont ainsi efficacement ligués et séquençables par ce protocole de préparation de librairie NGS.

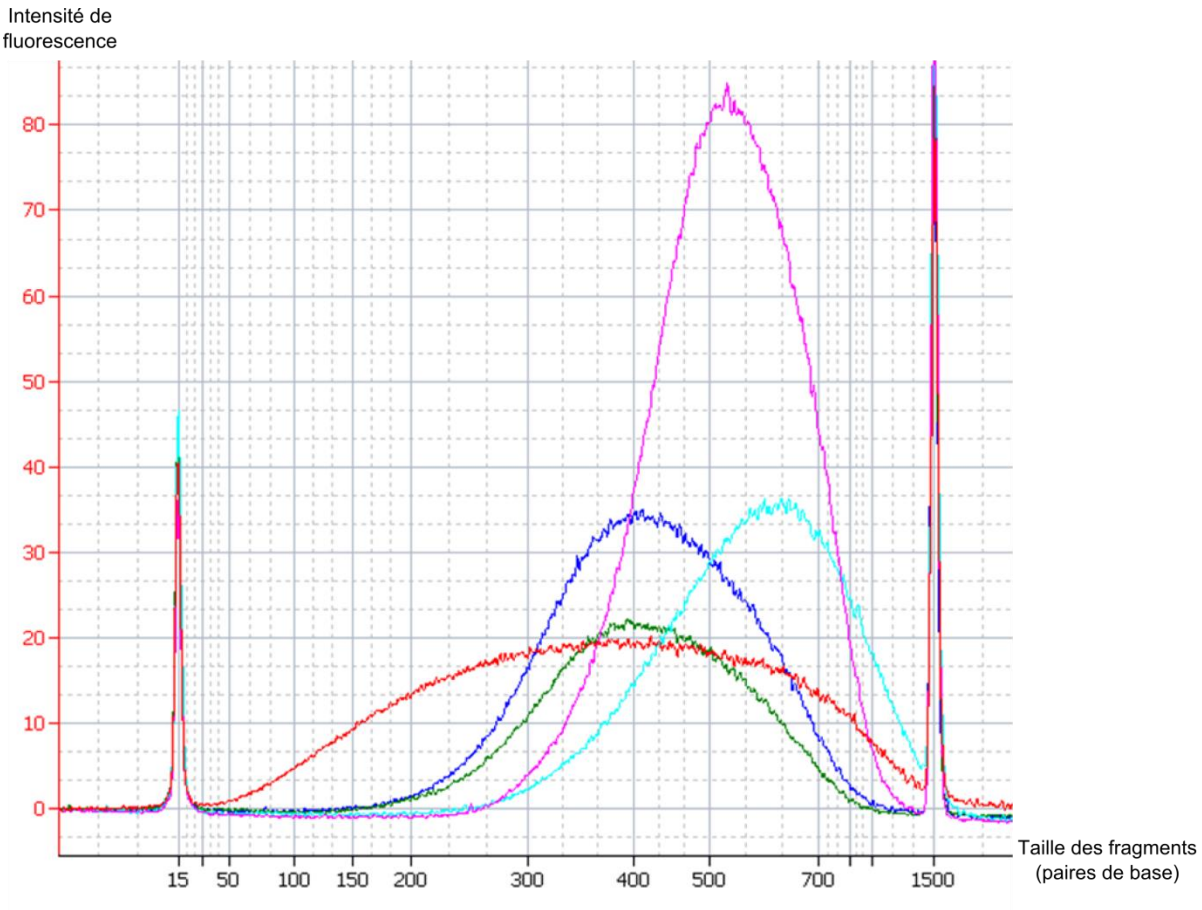


Figure 23 : Contrôle de la distribution des tailles des fragments et des rendements pendant la préparation d'une librairie NGS. Les profils de migration sur Bioanalyzer 2100 (puce D1K) après fragmentation (rouge), sélection de taille par SPRI pour éliminer les fragments <200pb et >1000pb (bleu foncé), ajout d'adénine (vert), ligation des adaptateurs (bleu ciel) et PCR (rose) sont représentés.

III.2 Analyse des sites d'intégration (SI) génomiques des vecteurs AAVr

Le risque génotoxique des vecteurs AAVr n'a été mis en évidence que récemment par des études pré-cliniques chez la souris (Chandler et al., 2015; Donsante et al., 2007). Ce risque peut être apprécié en évaluant deux paramètres : la fréquence et la localisation préférentielle des SI. Ils n'ont pour l'instant pas pu être évalués avec précision *in vivo* par les méthodes actuelles basées sur la PCR car elles présentent un certain nombre de biais (voir I.2.2.6). Ainsi, nous avons choisi de développer une nouvelle méthode profitant des avancées du NGS pour permettre une analyse plus précise de ce risque. Pour évaluer cette méthode sur un matériel biologique pertinent, nous avons dans un premier temps injecté une forte dose ($2{,}4x10^{13}$ vg/kg) d'un AAVr 2/8 RSV GFP en intramusculaire chez des souris mdx^{4CV}, modèles de myopathie de Duchenne, et des souris saines C57BL/6J.

III.2.1 Quantification des vecteurs AAV dans les muscles injectés

Les souris ont été euthanasiées 15 jours post-injection et l'ADN des muscles injectés a été extrait. Nous avons ensuite quantifié le nombre de génome de vecteur AAVr par génome diploïde (vg/dg) à l'aide de PCR quantitatives (qPCR). Le nombre de copies de vecteur a été évalué par une qPCR au niveau de la GFP, cette séquence étant unique au vecteur, et une seconde qPCR au niveau de l'albumine a permis de quantifier le nombre de génomes haploïdes (**Tableau 5**). Les groupes de souris mdx^{4CV} et C57Bl6 présentent respectivement un vg/dg moyen de 3,55 et 8,11. Ces valeurs, bien qu'obtenues sur un petit nombre de souris, montrent la même tendance que de précédents résultats au laboratoire. En effet, le génome de l'AAVr semble être moins stable dans un muscle dystrophique, aboutissant à une diminution du nombre de génomes d'AAVr persistants.

Souris	Nombre de copies d'albumine	Nombre de copies de GFP	Copies de vecteurs par génome diploïde
mdx^{4CV}	$1{,}71x10^5$	$3{,}64x10^5$	4,26
	$2{,}06x10^5$	$3{,}65x10^5$	3,54
	$2{,}85x10^5$	$4{,}06x10^5$	2,85
C57BL/6J	$1{,}77x10^5$	$1{,}07x10^6$	12,12
	$1{,}68x10^5$	$5{,}54x10^5$	6,6
	$1{,}15x10^5$	$3{,}21x10^5$	5,61

Tableau 5 : Quantification des vecteurs AAVr dans les muscles des souris mds^{4CV} et C57BL/6J 15 jours post-injection d'un AAV2/8 RSV GFP.

Par ailleurs, il a été montré qu'au sein des muscles déficients en dystrophine, l'important stress oxydatif aboutit à une augmentation des cassures doubles brin de l'ADN (Schmidt et al., 2011). Les vecteurs AAVr s'intégrant préférentiellement au niveau de ces brèches génomiques, il est possible que, malgré un nombre absolu de vecteurs inférieur, le nombre de de vecteurs AAVr intégré soit supérieur. Afin d'évaluer le pouvoir oncogénique des vecteurs AAVr dans ce contexte, nous avons donc cherché à comparer les sites d'intégration (SI) présents chez les souris saines et dystrophiques.

III.2.2 Séquençage ciblé sur les vecteurs AAVr

La taille du génome de l'AAV est très inférieure à celle du génome cellulaire. D'après les quantifications réalisées sur les muscles injectés, après fragmentation, les échantillons contiennent environ un fragment d'AAVr pour 10^6 fragments d'ADN génomique. Malgré l'augmentation importante des capacités du NGS, le séquençage direct de ces échantillons ne permettrait de séquencer, tout au plus, qu'une centaine de vecteurs. Etant donnée la faible fréquence d'intégration des vecteurs AAVr (10^{-4} et 10^{-6} par génome de vecteur) (Nowrouzi et al., 2012), la probabilité de séquencer un SI est quasi-nulle dans ces conditions. Il faut donc une méthode d'enrichissement en vecteur avant le NGS.

En alternative aux techniques de PCR, nous avons choisi de réaliser un séquençage ciblé sur les AAVr (et des SI associés) grâce à un système d'enrichissement par hybridation avec des sondes d'ARN. L'avantage d'un système d'hybridation par rapport aux techniques telles que la LAM-PCR et la LM-PCR (cf I.2.2.7) est la capture des génomes d'AAVr tronqués qui ne possèdent plus la zone d'hybridation des amorces de PCR. Nous avons donc fait synthétiser des ARN biotinylés de 80 bases, se chevauchant sur 40 bases, complémentaires du brin *sens* de notre vecteur AAV RSV GFP (**Figure 24**). Pour une meilleure efficacité, les ARN ont été synthétisés sans possibilité de cross-hybridation, les extrémités répétées du vecteur (ITR) ont donc été partiellement couvertes.

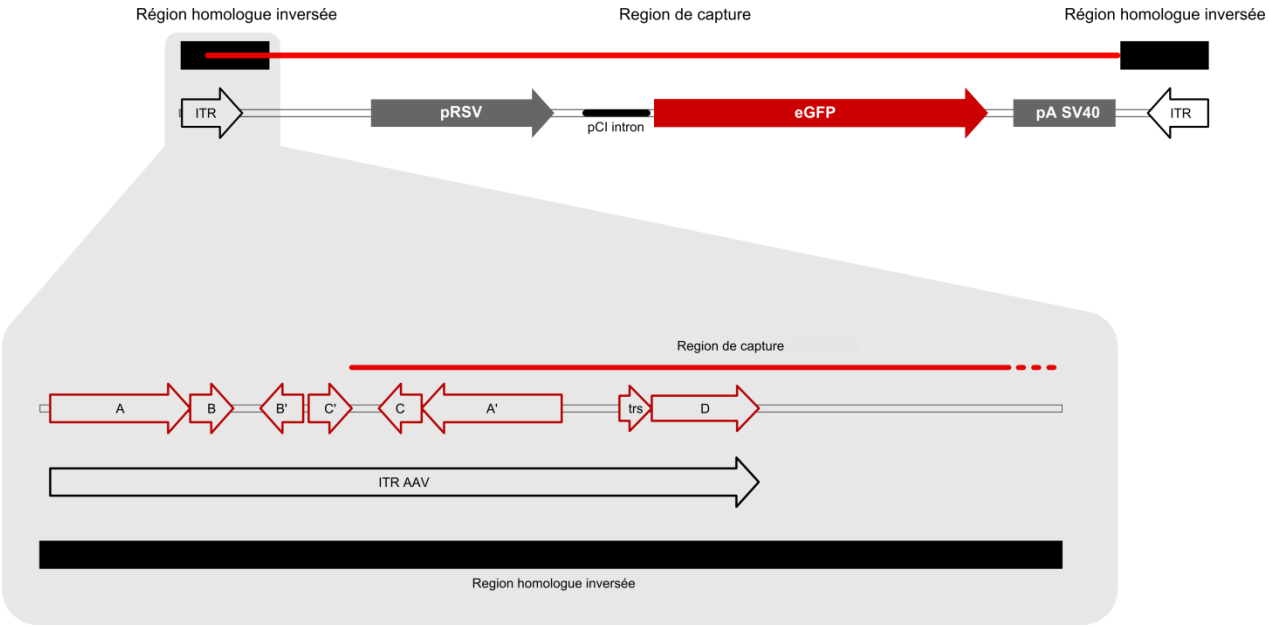


Figure 24 : Région du vecteur capturée par les ARN biotinylés. 49 ARN biotinylés chevauchant de 80 bases ont été produits sur une puce. Leur séquence a été déterminée pour qu'ils s'hybrident sur la région de capture représentée en rouge. Les ARN sont complémentaires du brin *sens*, l'ITR gauche étant couvert jusqu'au milieu du premier palindrome et l'ITR droit étant exclu de la capture pour éviter que les ARN puissent s'hybrider entre eux.

Jusqu'à présent, le séquençage ciblé n'a jamais été appliqué sur des vecteurs AAVr. Afin d'évaluer les biais potentiels pouvant intervenir lors des nombreuses étapes du protocole, des échantillons contrôles destinés à être analysés parallèlement aux échantillons de souris ont été synthétisés.

- **Echantillon « artificiel » contrôle**

Ce contrôle a été développé pour mimer les formes moléculaires de persistance des vecteurs AAVr (monomères et concatémères circulaires et formes intégrées). Il est ainsi composé d'ADN génomique extrait des muscles de souris non injectés (sans vecteurs) auquel a été ajouté des formes épisomales circulaires d'AAVr synthétisées *in vitro* et des plasmides mimant les formes intégrées.

Les concatémères ont été réalisés par digestion du plasmide pAAV RSV GFP par l'enzyme de restriction Xma1 qui coupe dans les ITR au niveau du palindrome C-C' en générant des extrémités cohésives. Les génomes des AAVr d'une taille de 2,2kb ont été extraits sur gel d'agarose puis mis en présence de ligase. Nous avons alors observé un assemblage des génomes formant des structures moléculaires linéaires et circulaires allant des monomères aux longs concatémères (**Figure 25**). Afin d'obtenir uniquement des formes circulaires, qui représentent les formes majoritaires de persistance des vecteurs AAVr, un traitement par la Plasmid-Safe DNAse (PS DNAse) a été réalisé. En effet, cette exonucléase digère uniquement les fragments

linéaires. Les différentes formes d'assemblage circulaires des génomes AAVr obtenues ont ensuite été dosées par qPCR au niveau de la GFP.

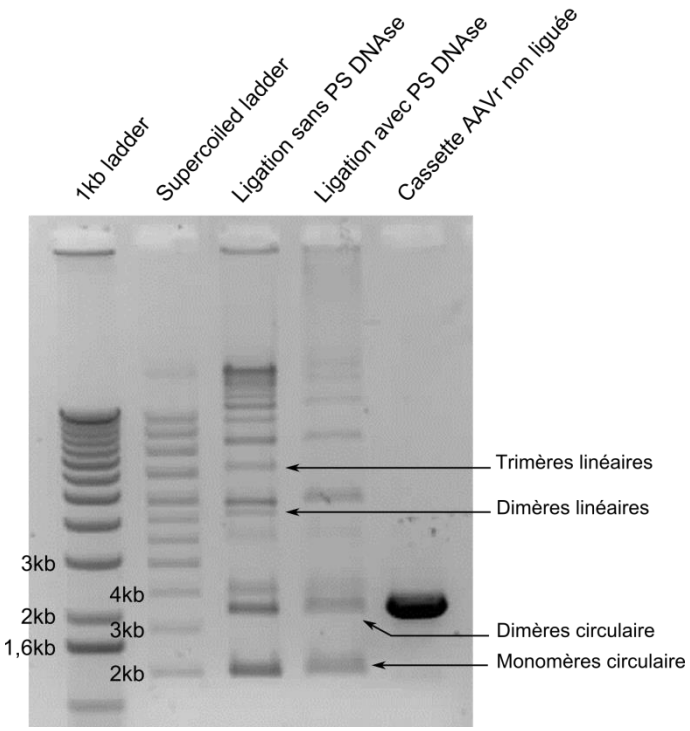


Figure 25 : AAVr épisomaux simulés *in vitro* par ligation des cassettes AAVr RSV GFP. Sur le gel d'agarose, nous avons fait migrer un marqueur de taille pour ADN linéaire (1kb ladder) et circulaire (supercoiled ladder) avec les échantillons correspondant aux produits de ligation avec et sans Plasmid-Safe DNase. L'échantillon contrôle sans ligation présente une bande attendue à 2,2kb correspondant au génome AAVr linéaire.

Pour mimer les vecteurs AAVr intégrés dans le génome cellulaire, nous avons choisi de cloner la cassette AAV RSV GFP dans un plasmide utilisé pour la production d'adénovirus, d'une longueur totale de 36kb. Cette taille de plasmide nous permet de reproduire, au moment de la fragmentation, le comportement d'un site d'intégration génomique. Ces formes ont également été dosées par une qPCR dans le plasmide.

Les différents composants ont ensuite été mélangés de façon à reproduire nos échantillons biologiques, qui contiennent environ 5 vecteurs AAVr par génome diploïde (**Figure 26**). Les AAVr intégrés ont été ajoutés à une fréquence de $2x10^{-4}$/vg, pour se placer dans la limite haute de ce qui a été décrit comme fréquence d'intégration. Cet échantillon contrôle a permis d'évaluer la sensibilité et la spécificité de la méthode de séquençage ciblé.

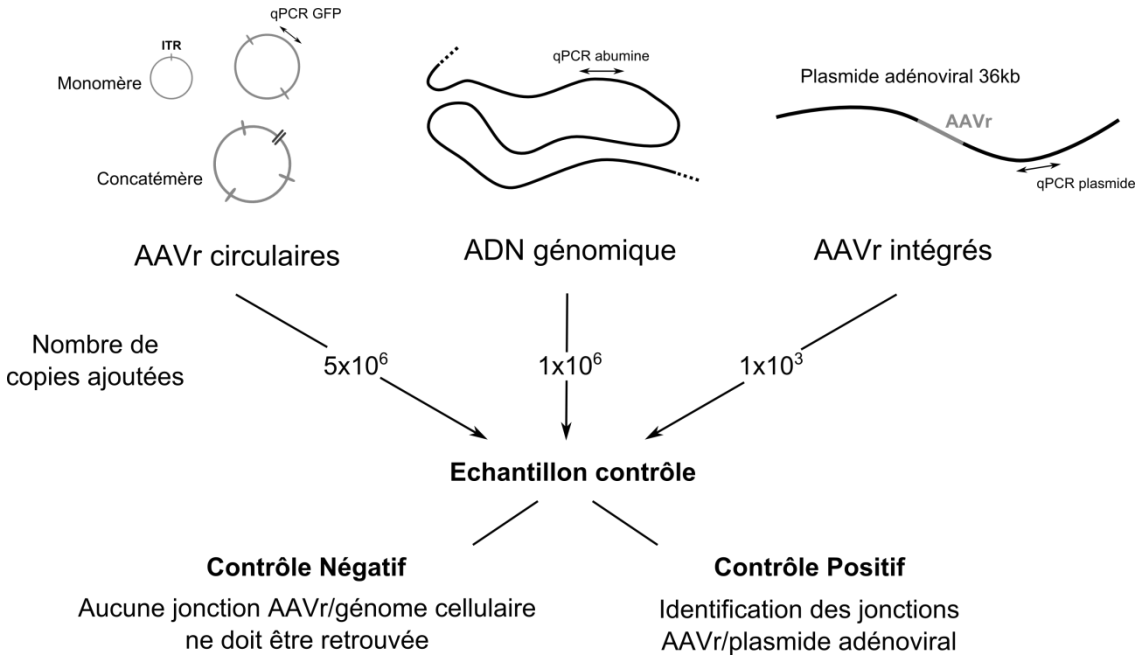


Figure 26 : Composition de l'échantillon contrôle.

- **Séquençage ciblé sur le génome AAVr**

Cinq microgrammes d'ADN extraits de deux souris mdx^{4CV}, de deux souris saines et de deux contrôles ont été utilisés pour tester le protocole d'enrichissement. Les échantillons ont été fragmentés à une taille centrée sur 500pb et les librairies NGS générées sans PCR finale. Puis, 500ng de ces librairies dénaturées ont été incubées pendant 36h à 65°C avec les ARN biotinylés complémentaires du génome AAVr. Les duplex ADN/ARN ont ensuite été extraits par des billes magnétiques de streptavidine et les ADN simple brin récupérés après lyse alcaline des ARN. Après 18 cycles de PCR « P5-P7 », les échantillons ont été mélangés de façon équimolaire (« multiplexage ») et séquencés sur plateforme MiSeq en mode *paired-end* (2x150pb).

Nous avons obtenu un total de 6 700 000 *paired-reads*, soit 13 400 000 *reads* totaux. Après « démultiplexage » bio-informatique par lecture des index, de 800 000 à 1 500 000 *paired-reads* ont été attribués aux échantillons. Le nombre de séquence variable obtenu entre les échantillons est vraisemblablement le résultat d'imprécisions lors du pipetage ou lors de la quantification des librairies par qPCR. Les *reads* ont ensuite été alignés par bio-informatique sur les séquences de référence du génome murin et de l'AAVr (**Figure 27**). Pour les échantillons murins, 36% à 57% des *reads* ont été alignés sur l'AAVr (entre 500 000 et 600 000 *reads* par échantillon). Avant capture, les fragments d'AAVr avaient été estimés être 10^6 fois moins nombreux que les fragments d'ADN génomique. Après enrichissement, d'après les données de séquençage, le ratio AAVr/génome cellulaire est situé entre 0,56 et 1,36. Le séquençage ciblé a donc permis d'enrichir un million de fois les échantillons en AAVr. Aucune publication

scientifique n'a pour le moment mentionné le séquençage d'un aussi grand nombre de vecteurs AAVr *in vivo* sans amplification par une méthode de PCR ciblée sur le vecteur. La majorité des génomes séquencés proviennent des épisomes mais l'importante profondeur de séquençage obtenue est sensée permettre d'étudier les rares évènements intégratifs.

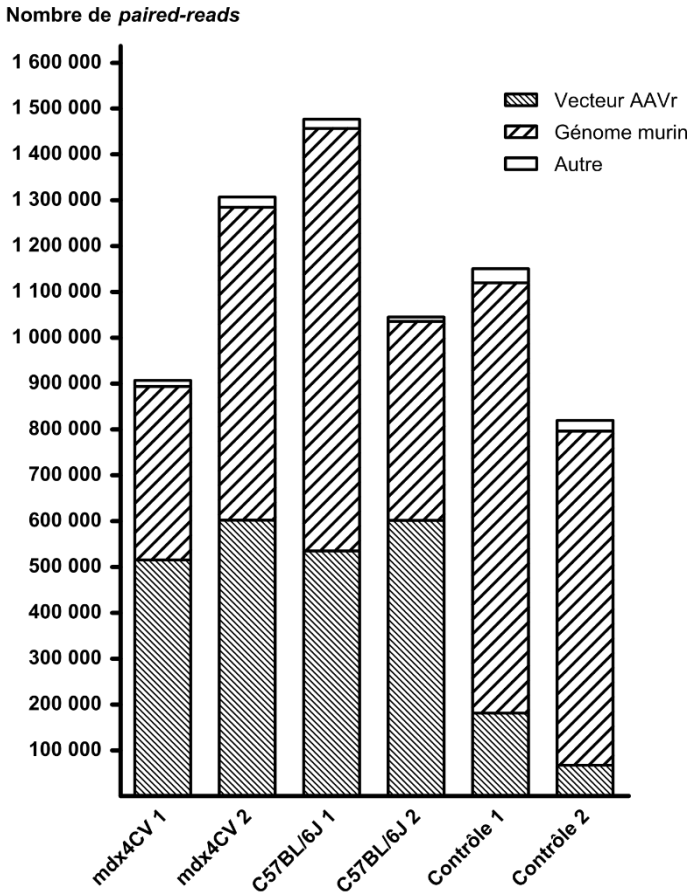


Figure 27 : Alignement des *reads* générés par MiSeq sur le génome AAVr et le génome murin. L'enrichissement en AAVr obtenu pour les contrôles est inférieur à celui obtenu pour les échantillons murins. C'est vraisemblablement la conséquence d'une sous-estimation par qPCR des épisomes générés *in vitro*.

La majorité des séquences « Autre » (ni génome, ni vecteur), a été identifiée comme correspondant au plasmide pAAV-RSV-GFP utilisé pour produire les vecteurs. Selon les échantillons, 3000 à 6000 de ces séquences ont été détectées. Le fait qu'elles aient été retrouvées après capture du génome AAVr implique leur co-encapsidation avec le génome du vecteur en tant que contaminants pendant la production des vecteurs (cf I.2.2.3). Leur présence justifie l'étude des contaminants ADN encapsidés dans les particules AAVr réalisée pendant ma thèse et décrite dans le chapitre suivant.

L'alignement des *reads* sur le génome du vecteur AAVr nous a permis d'obtenir une représentation graphique de sa couverture de séquençage (**Figure 28B**). L'analyse des polymorphismes nucléotidiques n'a pas révélé de variations majeures par rapport à la séquence de référence hormis une substitution nucléotidique présente à la fois dans les échantillons

contrôles et expérimentaux. Sa détection dans l'échantillon contrôle indique qu'il s'agit vraisemblablement d'une erreur dans la séquence de référence du génome du vecteur.

Les extrémités apparaissent sous-représentées dans les échantillons et les contrôles. Cette faible couverture observée peut être expliquée par l'étape de fragmentation (perte des courtes extrémités lors des purifications) et la capture partielle des ITR par les ARN. Au contraire, la région de l'AAVr codante pour la GFP est sur-représentée, notamment dans les échantillons contrôles. Il pourrait s'agir d'une contamination lors de la manipulation des échantillons dans le laboratoire par des plasmides ou des amplicons contenant la séquence de la GFP. En effet, une contamination environnementale par la GFP avait été mise en évidence lors de qPCR pendant la même période au laboratoire.

L'analyse plus précise de la couverture au niveau des ITR (**Figure 28C**) a révélé une chute de la profondeur de séquençage à la base 67 à partir de la fin de l'ITR, dans la région B'. Elle a été mise en évidence dans toutes les souris de façon très similaire, avec des baisses de couverture allant de 42% à 45%. Le fait que ce phénomène n'ait pas été observé dans les contrôles laisse penser qu'il s'agit d'un remaniement moléculaire des extrémités de l'AAVr *in vivo*. D'autres équipes ont déjà mis en évidence des recombinaisons au niveau des ITR *in vivo*, notamment dans la région du palindrome B/B' (Nowrouzi et al., 2012). Elles permettraient notamment la formation des concatémères par recombinaison homologue inter-vecteurs. Nous avons donc observé des réarrangements des ITR 15 jours après l'injection intramusculaire d'un vecteur AAVr, indifféremment du phénotype des souris. Sur ce point, le muscle déficient en dystrophine n'a pas provoqué de remaniements alternatifs des vecteurs détectables par le séquençage ciblé.

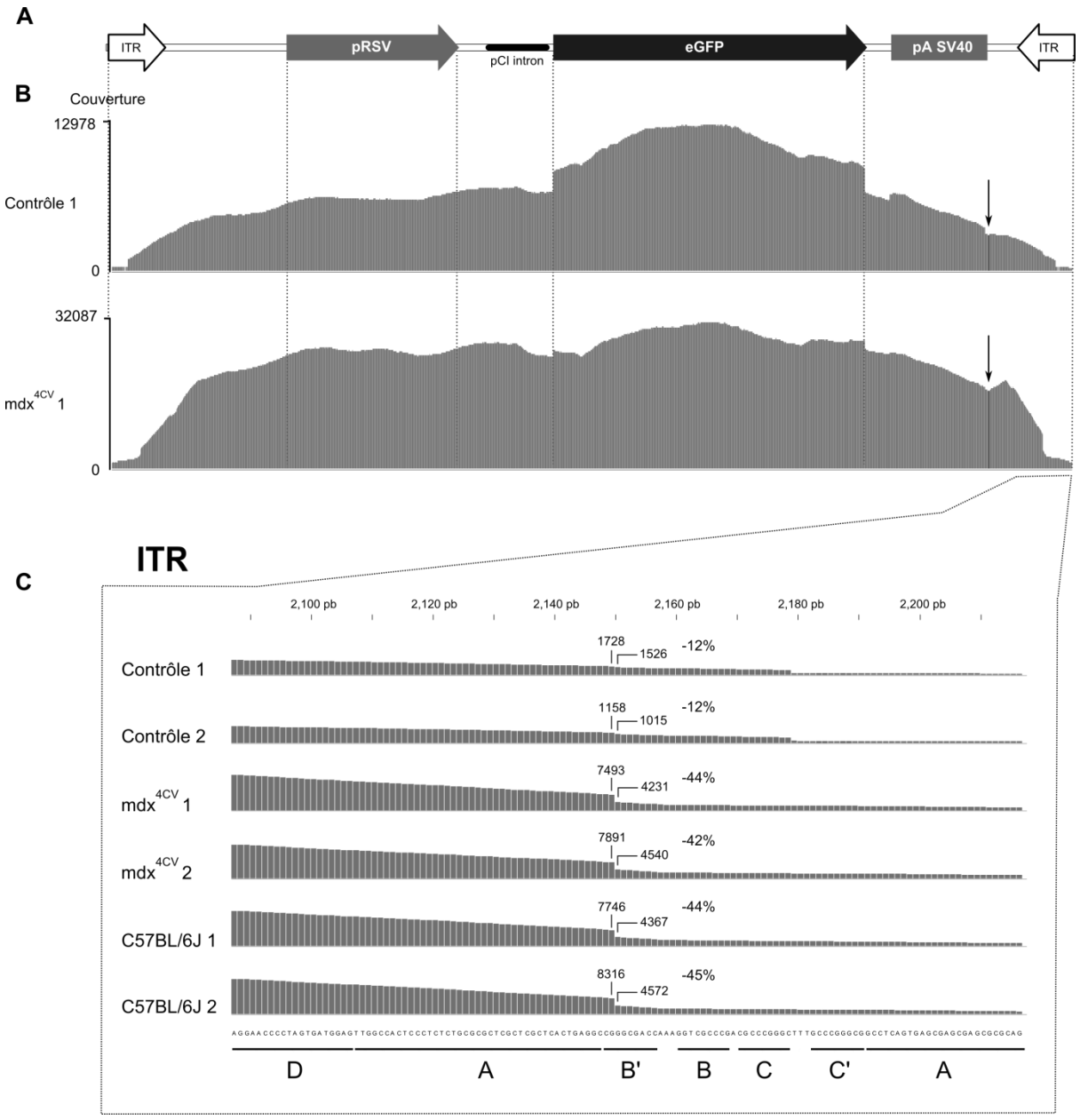


Figure 28 : Couverture du génome de l'AAVr après séquençage ciblé *in vivo*. Le schéma à l'échelle du génome de l'AAVr est représenté (**A**) ainsi que les couvertures de séquençage sur le génome AAVr pour le contrôle 1 et la mdx^{4CV} 1 (**B**). Ces deux profils sont représentatifs des autres échantillons. Les valeurs des couvertures maximales obtenues sont représentées sur l'axe des ordonnées. Les flèches représentent une substitution nucléotidique par rapport à la séquence de référence. Un zoom sur la région de l'ITR droit est également représenté (**C**), les deux valeurs indiquées correspondent aux valeurs de couverture obtenues au niveau du saut de couverture détecté dans les échantillons (entre les bases 67 et 68 à partir de la région A). Le pourcentage de diminution de la couverture est indiqué ainsi que la localisation des différentes régions de l'ITR.

- **Analyse bio-informatique des sites d'intégration**

L'analyse des sites d'intégration (SI) des AAVr dans le génome cellulaire a été effectuée par bio-informatique. Après séquençage *paired-end*, deux types de situations peuvent être rencontrées pour identifier un SI : soit la jonction est présente au sein d'un *read*, soit la jonction est déduite de l'alignement d'un *read* sur le génome du vecteur et du deuxième *read* associé sur l'ADN génomique (**Figure 29**).

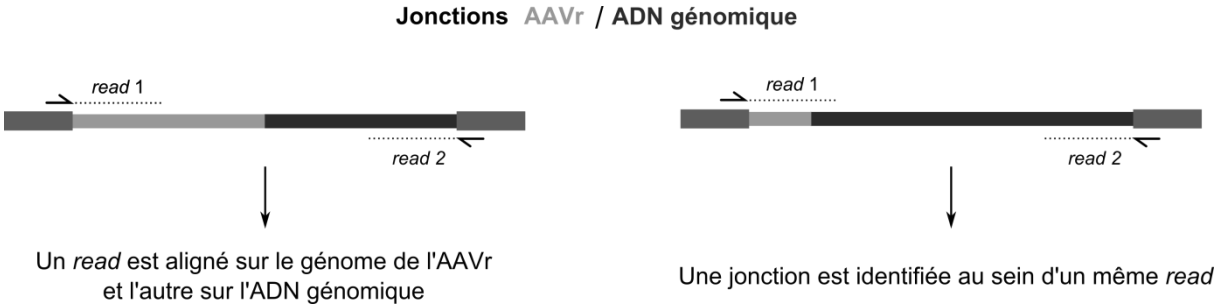


Figure 29 : Deux situations permettant l'identification des SI des AAVr.

L'analyse des SI dans nos échantillons de souris a identifié de 1583 à 1781 SI par souris, sans différence entre les phénotypes sauvages et dystrophiques (**Tableau 6**). La fréquence d'intégration a été évaluée à 3.10^{-3} par génome AAVr, ce qui est 10 à 100 fois supérieur à ce qui a été décrit jusqu'à présent. Parallèlement, nous avons analysé de la même façon nos échantillons contrôles, dans lesquels aucune jonction n'est censée être présente (**spécificité**). En effet, ils étaient composés d'ADN génomique sans AAVr intégrés (contrôles négatifs). De façon surprenante, nous avons obtenu 722 et 373 jonctions AAVr/génome murin respectivement pour les contrôles 1 et 2 (duplicats techniques). Nous avons identifié moins de jonctions dans les contrôles que dans les échantillons, mais étant donné le nombre inférieur de *reads* alignés sur le génome du vecteur dans les contrôles, les fréquences calculées des jonctions par AAVr ont été retrouvées équivalentes (moyenne contrôles : $4{,}66.10^{-3}$ Vs moyenne échantillons : $2{,}86.10^{-3}$). Cette fréquence similaire de jonctions AAVr/génome murin obtenue dans les échantillons murins et contrôles nous a indiqué que ces jonctions ont probablement été générées artéfactuellement durant le protocole de séquençage ciblé.

Nous avons ensuite cherché à identifier d'éventuels « vrais » SI présents dans nos échantillons en même temps que les jonctions artéfactuelles. En effet, les artéfacts étant *a priori* représentés par un seul évènement, l'identification de plusieurs jonctions présentant les mêmes coordonnées génomiques pouvait témoigner d'un SI présent à l'origine dans plusieurs cellules clonales. Néanmoins, l'analyse des jonctions dans les contrôles et les échantillons a révélé qu'elles étaient toutes uniques. De plus, la distribution des jonctions dans le génome murin est

apparue similaire entre les contrôles et les échantillons (**Figure 30**). Cette observation va dans le sens de la génération d'artéfacts techniques (fragments chimériques) pendant notre protocole. Ainsi, aucune caractéristique n'a permis l'identification de SI des AAVr dans le génome murin.

Echantillon	*Reads* totaux	% des *reads*			Jonctions AAVr/ADN génomique	
		AAVr	Génome murin	Autres	Nombre total	Jonction par AAVr
Contrôle 1	2 318 465	16,22%	81,17%	2,54%	722	3,88E-03
Contrôle 2	1 653 084	8,48%	88,73%	2,75%	373	5,44E-03
mdx^{4CV} 1	1 859 653	57,98%	39,16%	2,69%	1583	2,96E-03
mdx^{4CV} 2	2 666 955	47,26%	50,06%	2,57%	1525	2,44E-03
C57BL/6J 1	3 003 764	37,08%	60,60%	2,20%	1 781	3,22E-03
C57BL/6J 2	2 139 007	58,48%	38,98%	2,37%	1763	2,83E-03

Tableau 6 : Jonctions AAVr / ADN génomique identifiées dans les muscles de souris et les contrôles. Le nombre total de jonctions inclus les *reads* chimériques et les jonctions supportées par deux *reads* différents.

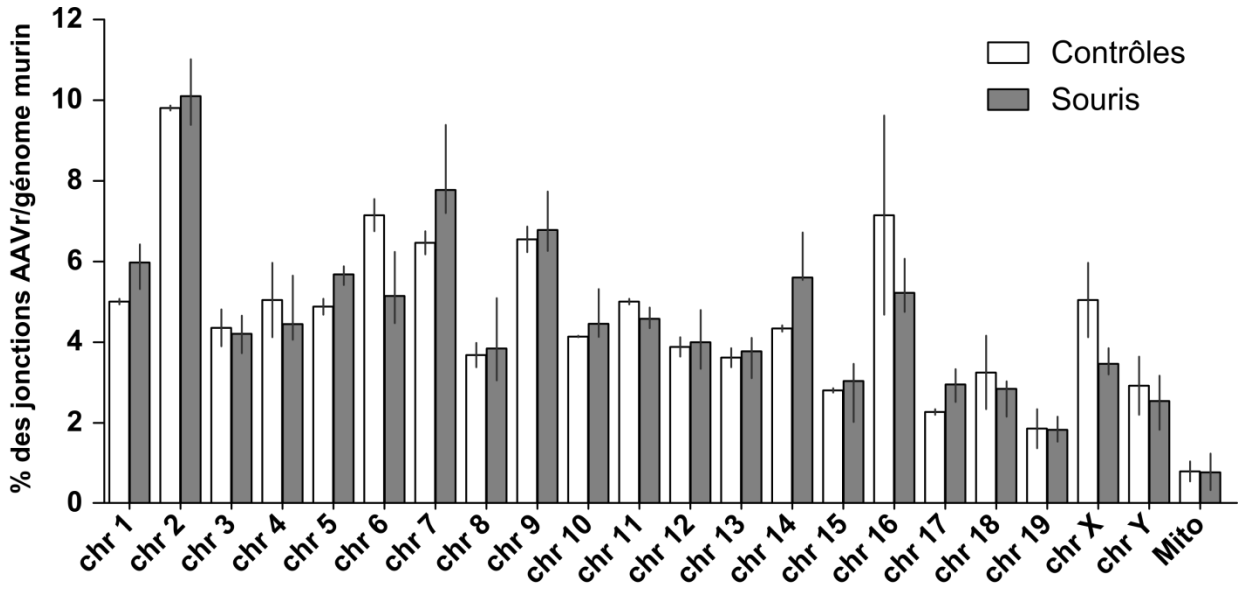


Figure 30 : Distribution génomique des jonctions AAVr/ADN dans le génome murin. Les barres représentent la médiane des valeurs obtenue dans les contrôles (n=2) et les échantillons (n=4). Les barres d'erreur représentent les valeurs inférieures et supérieures. chr = chromosome ; Mito = ADN mitonchondrial. Génome murin de référence : mm10.

Les contrôles étaient également des témoins positifs puisque les plasmides adénoviraux comportaient des génomes AAVr, créant ainsi des jonctions AAVr/plasmide qui permettaient de tester la capacité de notre méthode à détecter des SI (**sensibilité**). Nous avons détecté pour les contrôles 1 et 2, respectivement 8 et 4 jonctions AAVr/plasmide dans les données NGS. Le séquençage ciblé est donc capable d'identifier les SI des AAVr. Cependant, 150 jonctions

AAVr/plasmide étaient théoriquement présentes dans chacun des contrôles. Ainsi, notre protocole de séquençage ciblé est capable d'identifier 1 SI parmi environ 30 SI présents dans les échantillons. Cette sensibilité apparait insuffisante pour identifier les rares AAVr intégrés *in vivo* en l'absence d'évènement tumoral. En effet, dans le cas d'un tissu tumoral, de nombreux clones cellulaires possèdent le même SI, permettant l'identification du SI par des méthodes peu sensibles. Dans notre cas, le tissu musculaire murin 15 jours post-injection ne possède vraisemblablement que des évènements uniques d'intégration, qui nécessitent une méthode très sensible pour les identifier.

Au final, nous avons appliqué avec succès le séquençage ciblé sur les AAVr *in vivo*. Cette méthode nous a permis de séquencer un nombre très important de vecteurs AAVr présents dans les muscles des souris 15 jours post-injection. C'est la première fois qu'un tel enrichissement a été obtenu sans utiliser une méthode de PCR (LM-PCR, LAM-PCR). Cependant, le séquençage ciblé s'est avéré inadapté pour détecter les SI génomiques des AAVr. L'analyse des contrôles spécifiquement développés pour tester sa spécificité a révélé un nombre important d'artéfacts générés par le protocole, indistinguables des éventuels vrais SI par bio-informatique. Ainsi, aucune donnée concernant la fréquence et le tropisme d'intégration des AAVr dans les muscles sains et dystrophiques n'ont pu être générées lors de cette étude. Les raisons de l'échec du séquençage ciblé seront discutées par la suite.

III.3 Analyse de l'ADN encapsidé dans les particules AAVr

La caractérisation de l'ADN encapsidé dans les vecteurs AAVr est un autre aspect de la biosécurité qui a été étudié pendant cette thèse. La technique actuellement utilisée pour détecter les contaminants ADN dans les productions d'AAVr, la qPCR, ne permet pas d'avoir une vision exhaustive des différentes populations de contaminants présentes ni de comprendre les mécanismes menant à leur encapsidation. Le NGS promet aujourd'hui d'apporter des réponses à ces problématiques. Cependant, aucun protocole permettant d'appliquer le NGS à des vecteurs AAVr n'a été publié jusqu'à présent. Nous avons donc développé une méthode, nommée SSV-Seq (*Single Stranded Virus - Sequencing*), prenant en compte les spécificités biologiques des vecteurs AAVr pour permettre leur analyse par NGS.

III.3.1 Mise au point du protocole de SSV-Seq

Un paramètre important dans l'évaluation du potentiel délétère d'un contaminant ADN dans une production AAVr est de savoir s'il est présent à l'intérieur ou à l'extérieur des capsides AAVr. Dans le premiers cas, il pourrait être transféré et s'exprimer dans les cellules des patients alors que dans le cas contraire, il serait probablement rapidement dégradé par les nucléases présentes dans la circulation sanguine et dans la matrice extra-cellulaire. La première étape du développement de SSV-Seq a donc consisté à digérer efficacement l'ADN non éliminé par l'étape de purification afin d'analyser uniquement la fraction la plus pertinente des contaminants. Plusieurs DNases ont été testées, tout d'abord sur du plasmide seul en solution. Dans un second temps, du plasmide a été mélangé avec des AAVr pour s'assurer que la forte interaction de l'ADN avec les acides aminés chargés présents à la surface de la capside des AAVr n'altère pas l'efficacité de la digestion. L'association d'une endonucléase (Baseline-ZERO) et d'une exonucléase (Plasmid-Safe DNase) a montré la meilleure efficacité (**Figure 31**) et a été choisie pour digérer l'ADN non encapsidé lors de la première étape de SSV-Seq.

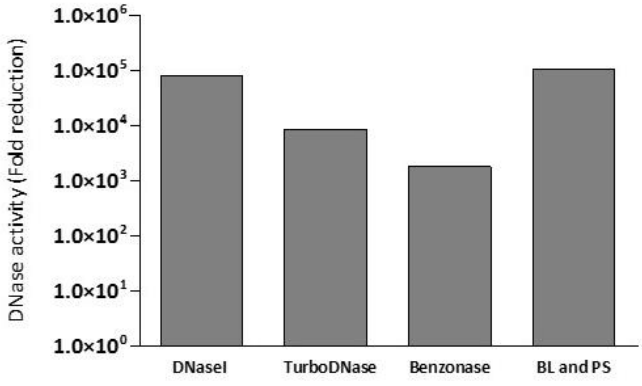


Figure 31 : Test de l'efficacité de digestion de l'ADN non encapsidé par plusieurs nucléases. Un plasmide a été mélangé à une production d'AAVr, l'efficacité de sa digestion par les nucléases a été évaluée par une qPCR ciblant le plasmide. Les rapports entre les titres qPCR obtenus avant et après digestion enzymatique sont représentés pout chaque DNase et couple de DNases testées. BL : Baseline-ZERO, PS : Plasmid Safe DNase.

Par la suite, les capsides virales ont été digérées par de la protéinase K puis, l'ADN encapsidé a été extrait par un système de précipitation. Etant donnée la nature simple brin du génome de l'AAVr, il a fallu mettre au point une étape de synthèse du second brin afin de pouvoir préparer une librairie NGS. Celle-ci a été réalisée par hybridation d'hexamères dégénérés (5'-NNNNNN-3') sur la matrice simple brin puis par synthèse du brin complémentaire par une DNA Polymerase I qui présente une activité 5'-3' exonucléase (digestion des brins rencontrés pendant la synthèse qui évite une amplification des fragments par déplacement de brin) et une activité 3'-5' exonucléase (faible taux d'erreur). L'efficacité de cette étape a été systématiquement contrôlée par incorporation de dUTP couplé à la fluorescéine dans le brin néo-synthétisé, qui a été visualisé par Southern Immunoblot et quantifié par fluorimétrie.

Après fragmentation de l'ADN double brin à une taille centrée sur 300pb, une librairie NGS a été réalisée comme décrit précédemment (cf IV.1) avec une amplification finale par 15 cycles de PCR. Les échantillons ont ensuite été séquencés en mode *paired-end* sur une plateforme de séquençage NGS Illumina HiSeq, dont le débit élevé permet de caractériser les contaminants les plus rares.

A l'issue du *run* de séquençage et du démultiplexage informatique un couple de fichier au format Fastq (**Figure 32**) contenant chacun plusieurs millions de *reads* est obtenu pour chaque échantillon. Un programme d'analyse a été développé au laboratoire (ContaVect) pour traiter par bio-informatique les fichiers Fastq obtenus. Dans un premier temps, l'utilisateur fourni toutes les séquences informatiques (nommées par la suite « **références** ») correspondant aux séquences réelles pouvant se retrouver dans la production d'AAVr (ex: les séquences du génome AAVr, du plasmide vecteur, du plasmide auxiliaire et du génome des cellules

productrices). Les *reads* des fichiers Fastq sont ensuite alignés contre les références et les caractéristiques de cet alignement sont utilisées pour générer les résultats finaux : comptage du nombre de *reads* provenant de chaque référence et analyse de la couverture de séquençage.

Afin de s'assurer de l'efficacité de l'analyse bio-informatique, un contrôle *in silico* a été développé. Il a consisté à simuler par informatique le résultat d'un *run* de séquençage en générant aléatoirement des fichiers Fastq contenant des *reads* provenant de chaque référence dans des proportions attendues. Le test d'alignement de ce contrôle avec ContaVect (**Tableau 7**), a montré une sensibilité et une spécificité de l'analyse bio-informatique proche de 100%. Seules les sensibilités observées pour le plasmide auxiliaire (91%) et le génome humain (98%) sont apparues légèrement plus faibles, vraisemblablement à cause de la présence de séquences répétées au sein de leur séquence de référence, sans que cela n'ait d'impact significatif sur les résultats.

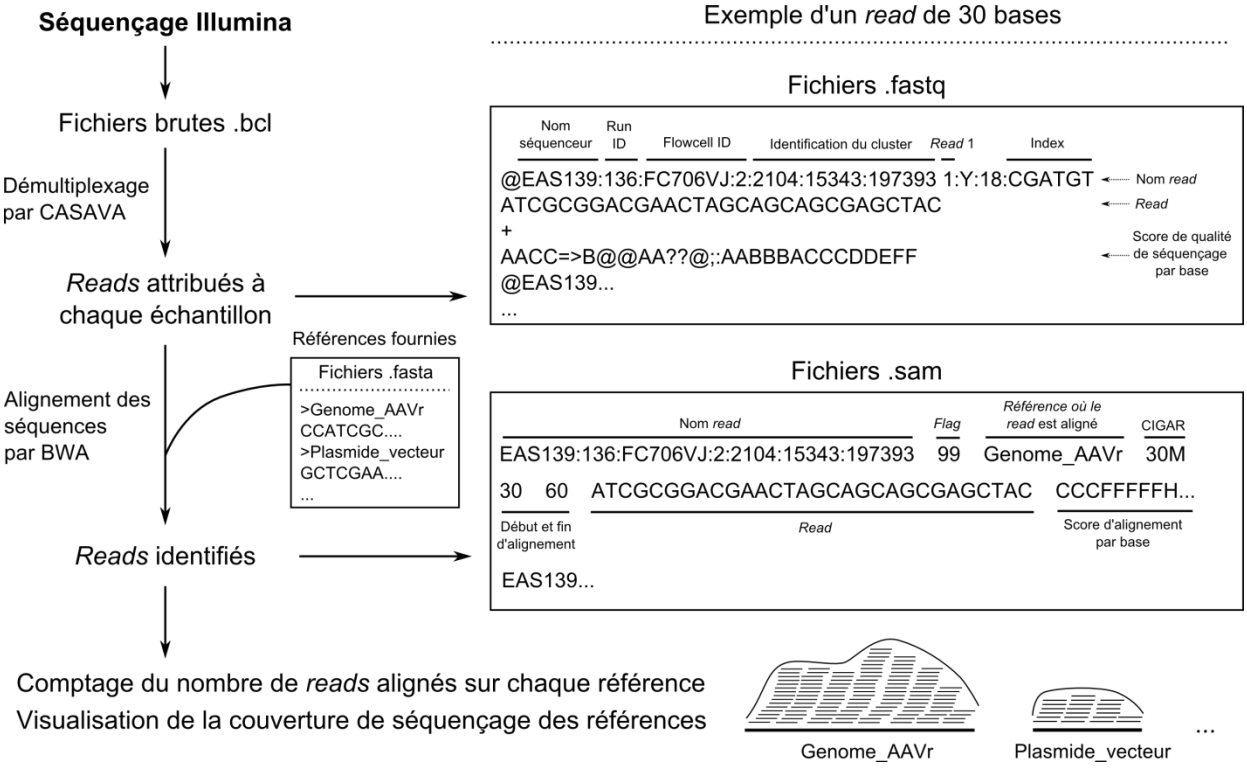


Figure 32 : Schéma simplifié du mode de fonctionnement de ContaVect. Les fichiers brutes sont traités par un logiciel d'Illumina (Casava) qui permet de générer des fichiers Fastq comprenant les *reads* associés à un score de séquençage par base. Lorsque plusieurs échantillons ont été multiplexés, la lecture des 6 bases spécifiques de l'adaptateur (index) permet de réattribuer chaque *read* à son échantillon. Les fichiers Fastq sont ensuite alignés par BWA (*Burrow Wheeler Aligner*) sur les références fournies par l'utilisateur (sous forme de fichier .fasta). Les fichiers .sam qui en résultent contiennent, pour chaque *read*, leurs caractéristiques d'alignement : un *flag* indiquant des informations générales sur l'alignement, la référence attribuée (le cas échéant), le CIGAR indique les propriétés de l'alignement (dans l'exemple, les 30 bases du *read* sont parfaitement alignées – 30 *Match*), les coordonnées de début et fin de l'alignement sur la référence, la séquence du *read* et enfin un score d'alignement attribué pour chaque base. Tous ces paramètres permettent ensuite à ContaVect d'extraire les résultats finaux.

		ATTENDU					
		AAVr	Plasmide vecteur	Plasmide auxiliaire	Génome humain	Non attribué	**Total observé**
OBSERVE	AAVr	14000000	0	0	0	0	**14000000**
	Plasmide vecteur	0	420000	455	0	0	**420455**
	Plasmide auxiliaire	0	0	4543	0	0	**4543**
	Génome humain	0	0	0	28952	0	**28952**
	Non attribué	0	0	2	572	200000	**200574**
	Total attendu	**14000000**	**420000**	**5000**	**29524**	**200000**	

Tableau 7 : Test de l'efficacité de ContaVect avec un contrôle *in silico*. Des *reads* ont été générés aléatoirement sur les 4 références**,** les résultats attendus et observés de leur alignement sont indiqués en nombre de *reads*.

III.3.2 Application de SSV-Seq sur une production d'AAVr

Afin de tester l'efficacité de SSV-Seq, nous avons appliqué le protocole développé sur des préparations de vecteurs AAVr 2/8 CMV GFP hTK produites par transfection transitoire dans des HEK293 et purifiées séparément par les trois méthodes de référence actuelles : l'ultracentrifugation sur gradient de chlorure de Césium (ClCs), la chromatographie d'échange ionique (IEX) et la chromatographie d'affinité (AVB). Les AAVr purifiés ayant été entièrement caractérisés par les contrôles de qualité classiquement utilisés, le but était de comparer les résultats obtenus par le NGS avec ces méthodes et ainsi d'évaluer à la fois son efficacité et la pertinence des informations complémentaires potentiellement générées.

L'étude précédente sur l'intégration des vecteurs AAVr a montré l'importance de développer des contrôles adaptés pour chaque projet de NGS. Nous avons donc décidé d'inclure dans SSV-Seq le traitement, parallèlement aux échantillons, d'un **contrôle positif** et d'un **contrôle négatif**. Le premier a consisté à mélanger les différents composants ADN attendus dans la production d'AAV2/8 étudiée dans des proportions déduites des résultats de qPCR, à savoir : le génome AAVr (2^E11 copies), le plasmide vecteur (1^E10 copies), le plasmide auxiliaire (4^E9 copies) et de l'ADN génomique fragmenté des HEK293 (1^E2 copies). Le second a été mis en place pour mettre en évidence une éventuelle contamination environnementale des échantillons pendant le protocole et a consisté à ajouter de l'ADN génomique de phage lambda (masse correspondante à 2^E11 copie de vecteur), car celui-ci ne présentait pas d'homologie avec les séquences étudiées.

Le protocole appliqué aux échantillons et aux contrôles est décrit sur la **figure 33**. Afin de séquencer la totalité de l'ADN présent dans la production d'AAVr ou uniquement la fraction encapsidée, deux volumes correspondants à 2^E11 copies de vecteur ont été prélevés dans chaque lot purifié d'AAVr et un des deux a été traité par les DNases. Pour contrôler l'efficacité de cette étape dans les données de NGS, de l'ADN de phage lambda a été ajouté aux échantillons avant le traitement par DNase.

Les différentes étapes du protocole ont été réalisées en aveugle sur les échantillons et les contrôles jusqu'aux résultats finaux générés par ContaVect. Afin d'étudier la reproductibilité du SSV-Seq, le protocole a été répliqué de façon indépendante, *runs* de séquençage inclus. Pour le premier et le second *run,* respectivement 5% et 1% d'ADN de Phage PhiX174 (sous forme de librairie NGS, commercialisé par Illumina) ont été mélangés avec les échantillons avant le séquençage. C'est une procédure classiquement utilisée lors du séquençage Illumina pour générer de la diversité de séquence. En effet, nos échantillons contenaient une forte proportion de fragments provenant du génome AAVr, donc avec une homogénéité de séquence potentiellement préjudiciable pour la technologie Illumina (Krueger et al., 2011).

Les deux *runs* de séquençage ont généré entre 12 et 17 millions de *reads* par échantillon (**Tableau 8**), avec un score de qualité de séquençage moyen supérieur à Q30 (moins d'une erreur de séquençage toute les 1000 bases). Avant d'analyser nos échantillons, nous voulions nous assurer de la qualité de nos données NGS en analysant la présence de séquences non désirées (artéfacts) dans les échantillons et les contrôles positifs et négatifs et d'en évaluer l'impact éventuel sur l'interprétation de nos résultats.

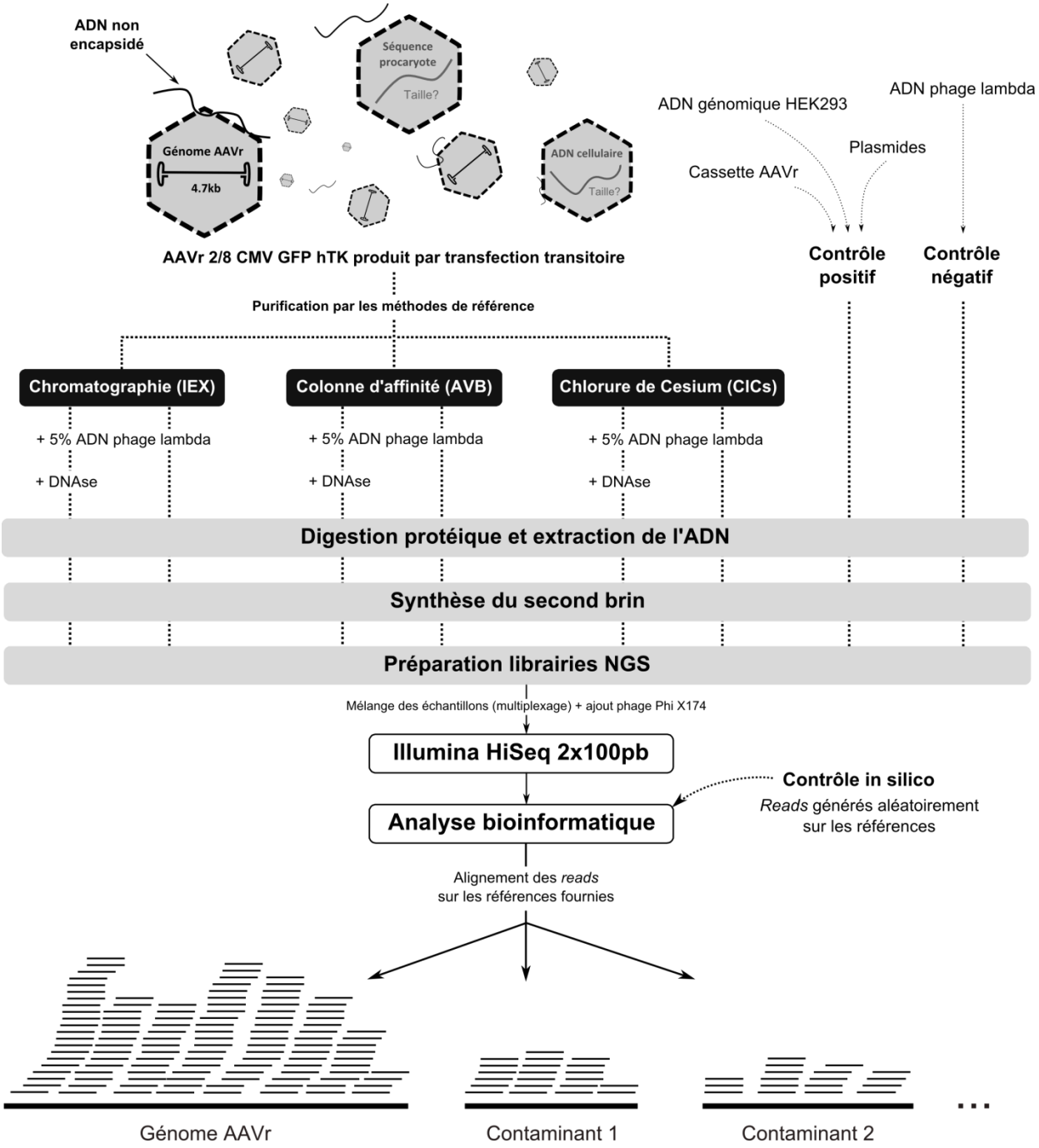


Figure 33 : Test du SSV-Seq sur une production d'AAVr 2/8 CMV GFP hTK. La production a été purifiée par les trois méthodes de référence puis, après ajout d'ADN de phage lambda, chaque produit de purification a été traité ou non par les DNases. L'ADN des échantillons et des contrôles a été extrait puis, après dénaturation thermique, une étape de synthèse du sécond brin a été réalisée. L'ADN double brin a ensuité été fragmenté à une taille d'environ 300pb et des libraires NGS compatibles avec le séquençage Illumina ont été générées (cf III.1.). Un index unique de 6 bases a été attribué à chaque échantillon, ce qui a permis de les multiplexer et de les séquencer lors d'un *run* HiSeq en mode *paired-end*. L'analyse bio-informatique a consisté à attribuer chaque *read* à une séquence de référence et à analyser de façon quantitative et qualitative les différentes populations d'ADN présentes dans les productions d'AAVr. Le protocole a été réalisé en aveugle et en duplicat technique.

Après analyse par ContaVect, nous avons trouvé de façon inattendue un nombre important de *reads* alignés sur le Phi X174 dans chaque échantillon et contrôle, avec une moyenne de 160 000 et 25 000 *reads* respectivement pour les *runs* 1 et 2. Les fragments de Phi X174 ne possédant pas d'index, ils ne devraient pas être attribués à des échantillons. Cette observation

témoigne d'un artéfact lors de la réaction de séquençage. Une partie des *reads* sont attribués à un mauvais échantillon. Ce phénomène a déjà été mis en évidence par une autre équipe (Kircher et al., 2012), bien que la plupart des publications utilisant la technologie Illumina n'aient pas étudié cet artéfact. La présence de Phi X174 n'était pas délétère en soi, mais elle témoignait de la possible cross-contamination entre échantillons multiplexés lors d'un même *run*. Des analyses ultérieures effectuées en excluant les *reads* possédant un index séquencé avec une mauvaise qualité (<Q30), comme suggéré par Kirsher *et al.*, ont montré que ce phénomène n'a pas eu d'impact sur nos résultats (données non montrées). Néanmoins, il nous a semblé important d'en évaluer l'impact pour chaque *run* et pour chaque application du NGS.

L'inclusion d'un contrôle négatif dans un protocole NGS n'est pas souvent réalisé mais il nous a semblé important d'évaluer l'importance de la cross-contamination entre les échantillons car celle-ci pouvait impacter de façon importante la quantification des différents contaminants dans nos productions d'AAVr. L'analyse de ces contrôles a révélé un certain niveau de contamination pour chaque référence, qui a pu se produire pendant les étapes techniques du protocole ou lors de la réaction de séquençage. Nous avons défini ces valeurs comme nos seuils de positivité pour nos échantillons. Dans les contrôles négatifs, quelques *reads* ont été alignés sur le génome humain (2292 et 3140), il est probable que ce soit en partie dû à une contamination par l'ADN génomique du manipulateur lors de la préparation des échantillons (Jun et al., 2012). Par ailleurs, l'analyse du nombre important de *reads* non identifiés dans le contrôle positif du second *run* a révélé une contamination avec de l'ADN bactérien, phénomène déjà décrit par plusieurs équipes (Laurence et al., 2014; Strong et al., 2014). Ces contaminations environnementales, d'origine humaine et bactérienne, doivent être évaluées à chaque *run* car elles peuvent biaiser la quantification des contaminants présents dans les productions d'AAVr.

Pour chaque système de purification, le nombre de *reads* alignés sur le phage lambda a été retrouvé au-dessous du seuil de positivité dans les échantillons traités par les DNases. Cette digestion totale de l'ADN du phage lambda nous a permis de confirmer que nous avons séquencé uniquement l'ADN encapsidé (ou très fortement associé à la capside) dans les échantillons traités par les DNases (+DNase).

	Nom référence	Contrôle négatif	Contrôle positif	ClCs		AVB		IEX	
				- DNase	+ DNase	- DNase	+ DNase	- DNase	+ DNase
***Run* 1**	Phi X174	155 022	131 849	157 572	214 214	137 587	127 001	174 238	201 407
	Phage lambda	15 820 832	**1 993**	1 317 316	509	1 567 573	270	1 863 260	631
	Génome AAVr	**17 217**	8 014 026	12 513 125	16 768 399	11 820 182	11 207 153	13 302 866	14 536 121
	Plasmide vecteur	**913**	1 138 694	148 331	142 147	533 912	348 283	737 458	815 318
	Plasmide auxiliaire	**54**	3 937 978	1 472	1 763	8 119	5 267	7 167	7 807
	Génome humain	**2 292**	196 882	13 210	6 361	36 122	24 708	28 946	26 333
	Non aligné	68 642	374 962	99 746	101 967	213 741	143 356	182 021	171 241
	Total	16 064 972	13 796 386	14 250 772	17 235 360	14 317 236	11 856 038	16 295 956	15 758 858
***Run* 2**	Phi X174	18 440	21 624	25 431	26 479	22 447	31 976	30 181	20 004
	Phage lambda	12 524 254	**838**	1 118 877	357	1 074 966	515	1 284 740	401
	Génome AAVr	**1 205**	7 996 633	12 848 053	13 281 337	9 729 696	13 446 419	11 698 973	11 248 707
	Plasmide vecteur	**366**	1 031 197	167 598	116 191	460 698	485 013	745 067	491 676
	Plasmide auxiliaire	**22**	628 280	1 622	1 348	6 671	8 028	6 601	9 321
	Génome humain	**3140**	263 076	17 749	5 804	30 671	39 382	28 913	19 637
	Non aligné	203 077	2 414 472	260 792	210 724	287 211	537 789	319 873	292 122
	Total	12 750 506	12 356 124	14 440 122	13 642 240	11 612 362	14 549 124	14 114 348	12 081 868

Tableau 8 : Nombre de *reads* attribués à chaque référence pour les deux *runs* de séquençage. En plus des séquences des différents composants ADN retrouvés dans la production d'AAVr (le génome humain étant une approximation du génome des cellules HEK293), les séquences de référence du génome du phage PhiX174 et du phage lambda ont été ajoutées lors de l'analyse par ContaVect. Les valeurs en gras représentent les seuils de positivité (niveau de contamination).

- **Comparaison de la quantification des contaminants par SSV-Seq et qPCR**

Les proportions des différents contaminants ADN et du génome AAVr ont été déterminées dans les productions d'AAVr purifiées (**Tableau 9**). La comparaison des valeurs obtenues dans les deux réplicats techniques a montré une très bonne reproductibilité du NGS, malgré les nombreuses étapes enzymatiques du protocole. Indépendamment du système de purification, nous avons obtenu une large majorité de génome AAVr (93,75% à 99,11%) suivi par le plasmide vecteur (0,84% à 5,97%), le génome humain (0,04% à 0,30%) et le plasmide auxiliaire (0,01% à 0,08%). Néanmoins, les vecteurs purifiés par ultracentrifugation sur gradient de chlorure de Césium se sont avérés moins contaminés que ceux purifiés par les méthodes chromatographiques. Le principe de purification de ces dernières entrainant une séparation moins efficace entre les capsides pleines et les vides, nous pouvons supposer qu'elles éliminent également moins efficacement les particules contenant des ADN contaminants, celles-ci possédant des propriétés physico-chimiques qui permettraient leur élimination lors d'une séparation basée sur la densité (ClCs). Par ailleurs, les proportions des contaminants ont été faiblement diminuées par le traitement par les DNases, entre 15% et 30% pour le plasmide vecteur, ce qui suggère que la majorité de ces contaminants sont encapsidés, donc transférables *in vivo*.

Methode	Référence	CICs		AVB		IEX	
		- DNase	+ DNase	- DNase	+ DNase	- DNase	+ DNase
SSV-Seq	Génome AAVr	98,71% 98,57%	99,11% 99,08%	95,34% 95,13%	96,74% 96,19%	94,50% 93,75%	94,48% 95,58%
	Plasmide vecteur	1,17% 1,29%	0,84% 0,87%	4,31% 4,50%	3,01% 3,47%	5,24% 5,97%	5,30% 4,18%
	Plasmide auxiliaire	0,01% 0,01%	0,01% 0,01%	0,07% 0,07%	0,05% 0,06%	0,05% 0,05%	0,05% 0,08%
	Génome HEK293	0,10% 0,14%	0,04% 0,04%	0,29% 0,30%	0,21% 0,28%	0,21% 0,23%	0,17% 0,17%
qPCR	Génome AAVr	99,48%	99,56%	98,16%	98,32%	97,65%	98,00%
	Plasmide vecteur	0,52%	0,44%	1,83%	1,67%	2,34%	1,99%
	Plasmide auxiliaire	0,004%	0,004%	0,012%	0,010%	0,010%	0,009%
	Génome HEK293	< Seuil	< Seuil	< Seuil	< Seuil	< Seuil	< Seuil

Tableau 9 : Quantification relative des contaminants ADN présents dans les productions d'AAVr par SSV-Seq et par qPCR. Pour SSV-Seq, les pourcentages ont été calculés en divisant le nombre de *reads* alignés sur chaque référence par le nombre total de *reads* alignés sur les 4 références. Les valeurs obtenues pour les deux *runs* sont indiquées. Pour permettre de comparer avec les valeurs de NGS, les titres de qPCR ont été normalisés par la taille des références : 4794pb pour le génome AAVr, 2209pb pour le plasmide vecteur, 22757pb pour le plasmide auxiliaire et $3^{E}9$pb pour le génome humain.

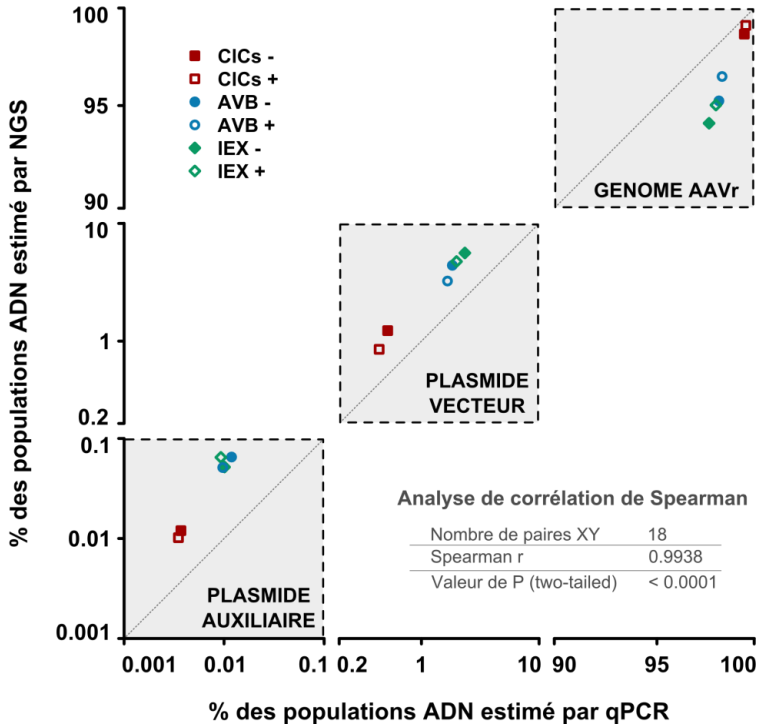


Figure 34 : Corrélation entre les quantifications obtenues par NGS et qPCR. Les axes des abscisses et des ordonnées ont été séparés pour pouvoir visualiser sur un même graphique les pourcentages des différents composants ADN dans la production d'AAVr.

Nous avons ensuite cherché à comparer les résultats du NGS avec ceux obtenus par la méthode de référence, la qPCR (**Tableau 9**, **Figure 34**). Le principe des deux techniques étant complètement différent, afin de comparer les quantifications relatives des contaminants, nous avons normalisé les titres de qPCR par la taille des références. Nous avons mis en évidence une très bonne corrélation entre le NGS et la qPCR (coefficient de corrélation de Spearman = 0.9938, $p < 0.0001$). Ainsi, malgré les étapes d'amplification du protocole, les résultats du NGS peuvent être considérés comme quantitatifs. Les valeurs obtenues par le NGS sont cependant deux à trois fois supérieures pour les contaminants plasmidiques. Dans ces données préliminaires, il n'est pour l'instant pas possible d'évaluer quelle méthode donne la quantification la plus fiable. Cependant, la comparaison directe des proportions au sein d'un même *run* suggère un avantage du NGS par rapport à la comparaison de deux qPCR indépendantes, chacune possédant un degré de variabilité (Ayuso et al., 2014). En revanche, l'avantage du NGS a été clairement établi pour la détection des contaminants provenant de l'ADN génomique des cellules productrices. La qPCR ne ciblant qu'une infime partie du génome (gène de l'albumine), la faible contamination en ADN génomique n'a pas pu être mise en évidence. Au contraire, le NGS détecte la totalité des fragments d'origine génomique, ce qui permet d'estimer la contamination en ADN cellulaire de façon exhaustive.

- **Identification de contaminants par NGS ignorés par qPCR**

Le NGS permet, contrairement à la qPCR, d'étudier la séquence des contaminants encapsidés. Parmi les *reads* qui ont été alignés sur le génome humain, nous avons analysé leur provenance précise et étudié la représentation de chaque chromosome (**Figure 35a**). Pour étudier si l'ADN génomique était encapsidé de façon aléatoire, nous avons normalisé les distributions obtenues dans les échantillons par les valeurs obtenues dans notre contrôle positif, celui-ci étant composé d'ADN cellulaire fragmenté aléatoirement. Nous avons observé une représentation aléatoire dans tous les chromosomes, mis à part une sur-représentation du chromosome 15 dans la production purifiée par AVB et une sur-représentation du génome mitochondrial dans la production purifiée par ClCs.

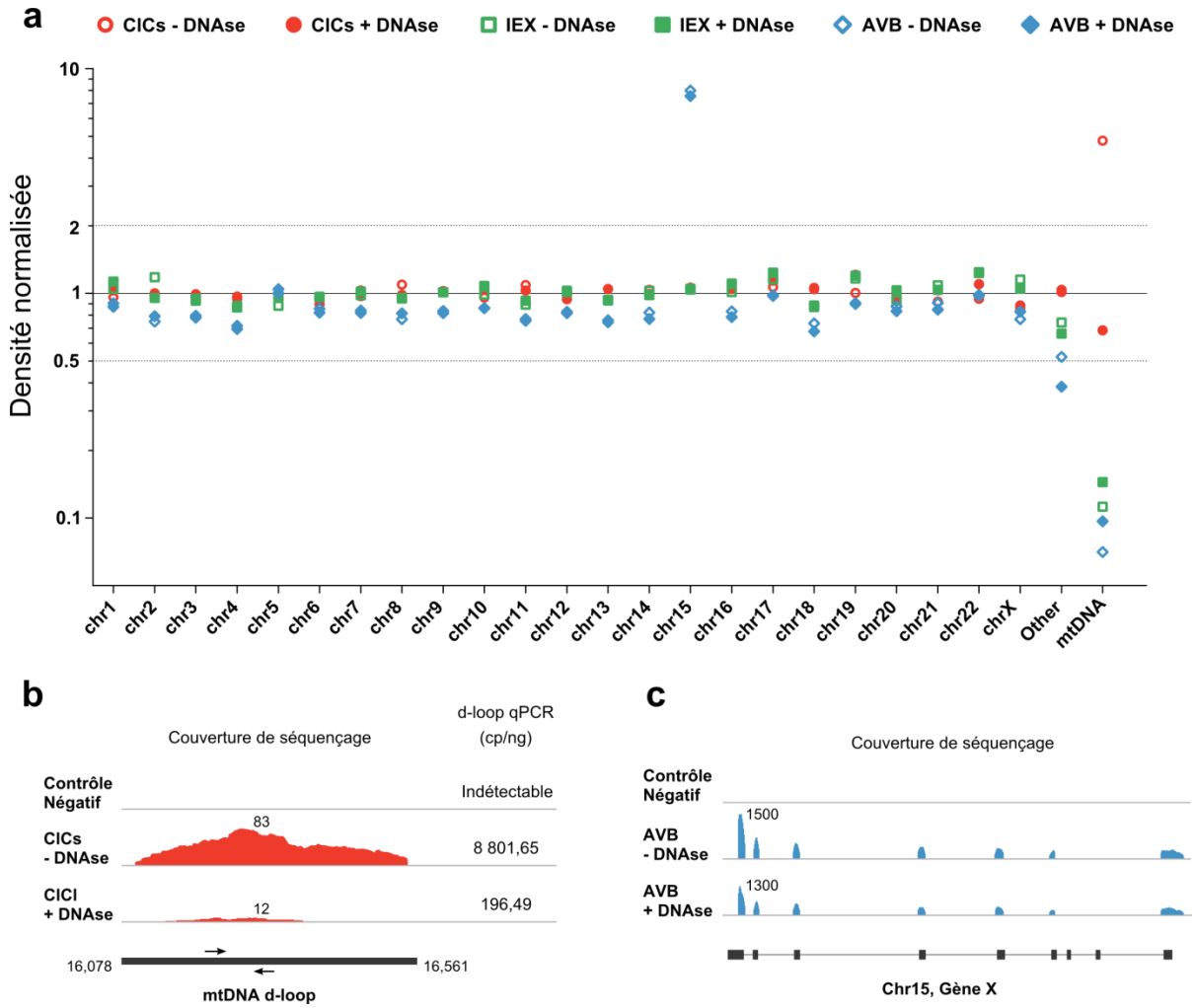


Figure 35 : Contaminations inattendues détectées dans la production d'AAVr. (**a**) Le nombre de *reads* alignés dans chaque chromosome a été divisé par le nombre total de *reads* alignés sur le génome humain afin d'obtenir une fréquence de représentation de chaque chromosome. Ces valeurs ont été normalisées par le même calcul effectué pour le contrôle positif, dont la représentation du génome est homogène (fragmentation aléatoire du génome des HEK293). Ces densités normalisées sont représentées pour chaque chromosome et le génome mitochondrial (mtDNA). Une valeur proche de 1 indique une représentation aléatoire du chromosome et des valeurs supérieures ou inférieures représentent respectivement une sur ou une sous-représentation. Détails des couvertures des deux séquences sur-représentées : la d-loop du génome mitochondrial pour ClCs (**b**) et un gène du chromosome 15 pour AVB (**c**). Les valeurs de couverture maximales sont indiquées. La contamination par la d-loop mitochondriale a été confirmée par une qPCR dont les amorces et les titres obtenus sont indiqué à droite du graphique.

L'analyse de la couverture du génome mitochondrial a montré qu'une région spécifique était responsable de son enrichissement : la d-loop (**Figure 35b**). Cette région possédant une origine de réplication et de transcription, son transfert *in vivo* pourrait être délétère pour les patients. Dans l'échantillon traité par les DNases, une nette diminution de sa couverture a été observée, ce qui a suggéré que cette contamination n'était pas encapsidée. Cette contamination a été retrouvée spécifiquement dans les échantillons purifiés par ClCs, ceux purifiés par les méthodes chromatographiques ont au contraire montré une déplétion d'ADN mitochondrial par rapport au contrôle. En effet, l'encapsidation de l'AAVr s'effectuant dans le noyau, une présence négligeable d'ADN mitochondrial était attendue dans les particules AAVr. Il est donc

probable que la purification par ClCs, basée sur une séparation par la densité, ait entrainé une co-purification du génome mitochondrial. Celui-ci a ensuite dû être partiellement digéré par la DNase inclue pendant cette purification, sauf au niveau de la d-loop qui présente une structure secondaire particulière (triple hélice ADN). Une qPCR a été mise au point au niveau de la d-loop pour valider les résultats de NGS, les titres obtenus ont été cohérents avec les couvertures observées. De plus, en considérant la couverture maximale obtenue dans l'échantillon sans DNase (12) comme la limite basse de détection du NGS, les valeurs de qPCR nous ont permis d'estimer la sensibilité du NGS à 10^4 copies d'un contaminant pour 10^{11} copies d'AAVr.

Un gène spécifique (identité non dévoilée pour cause de confidentialité) a été identifié comme responsable de l'enrichissement du chromosome 15 après purification par AVB. La **figure 35c** montre que seuls les exons du gène sont couverts et que l'échantillon traité par les DNases présente une couverture similaire. Cette contamination était donc vraisemblablement présente à l'intérieur des capsides. La purification par AVB, qui consiste à capturer les capsides de vecteur sur une colonne recouverte d'anticorps spécifiques anti-capside, a été réalisée par un collaborateur qui rince les colonnes entre deux purifications de vecteurs. Ainsi, nous avons pu identifier que cette contamination était due à un lot de vecteur AAVr précédemment purifié sur la même colonne. Ce vecteur comportant un transgène humain, nous avons pu l'identifier lors de l'alignement sur le génome humain.

Ces deux exemples démontrent le potentiel du NGS pour mettre en évidence des contaminations inattendues non détectées par qPCR et qui pourraient présenter un problème de biosécurité. Cependant, ces contaminants ont été identifiés car leur séquence était similaire au génome humain que nous avons utilisé comme référence. Une contamination par un vecteur comportant un gène d'une espèce éloignée de l'homme n'aurait pas été détectée aussi efficacement. Une solution plus exhaustive de détection des contaminants pourrait consister à analyser tous les *reads* non alignés sur les séquences attendues en les comparer à une base de donnée comportant toutes les séquences de contaminants possibles.

- **Analyse du génome AAVr encapsidé**

La fonctionnalité des vecteurs AAVr peut être altérée lorsque les génomes recombinants sont partiellement encapsidés (Kapranov et al., 2012) ou lorsque des mutations dans le transgène apparaissent. L'importante profondeur de séquençage du génome AAVr obtenu par NGS permet d'analyser avec précision ces deux paramètres (**Figure 36**).

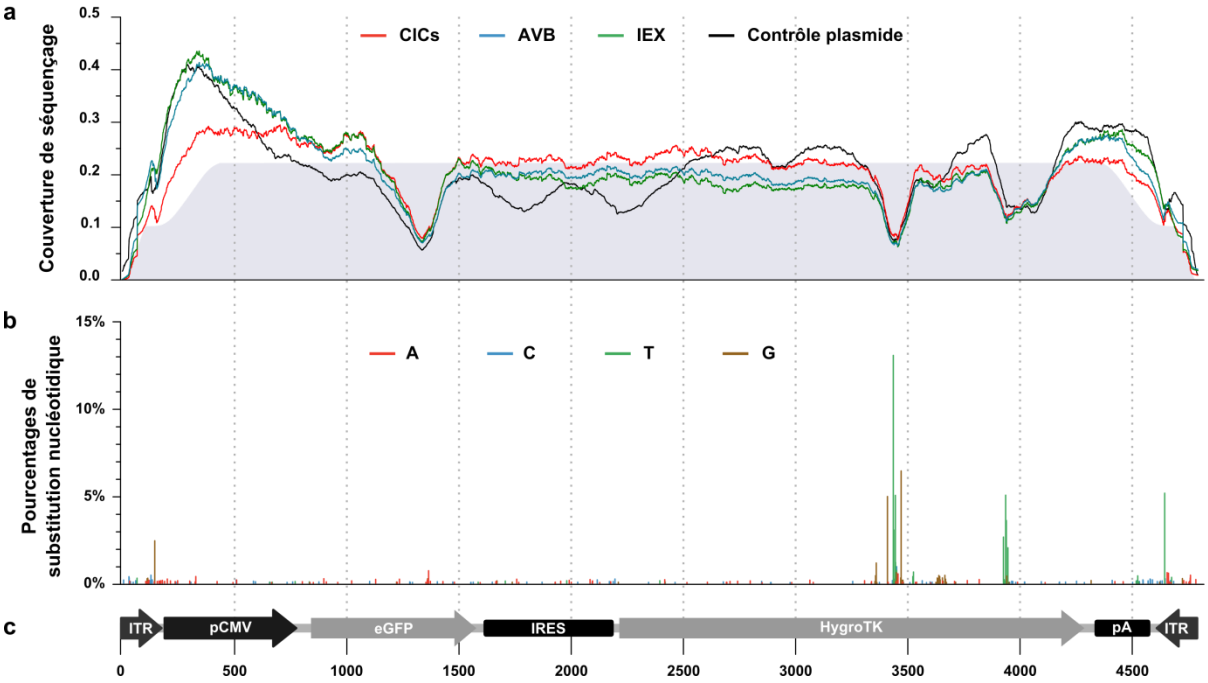


Figure 36 : Identité génomique du vecteur AAVr. (**a**) La couverture de séquençage a été calculée pour chaque base en comptant le nombre de *reads* comportant chacune d'elles, ces valeurs ont été normalisées en les divisant par le nombre total de bases alignées (x1000). Les couvertures sont représentées pour les échantillons non traités par DNase et le contrôle plasmidique. La zone grisée en arrière-plan représente la couverture du contrôle *in silico*. (**b**) Les pourcentages des substitutions nucléotidiques retrouvés pour chaque base par rapport à la séquence de référence sont représentés pour les échantillons. La même analyse a été réalisée sur le contrôle plasmidique et a montré le même profil. (**c**) Représentation à l'échelle du génome AAVr permettant de localiser la couverture de séquençage et l'analyse des polymorphismes.

La couverture de séquençage des échantillons a montré des zones sur-représentées, comme au niveau du promoteur CMV, et des parties moins représentées, notamment au niveau des régions codantes pour la GFP et l'hygroTK. Ces variations ont été retrouvées de façon similaire dans tous les échantillons. Cependant, l'analyse du contrôle plasmidique, pour lequel nous attendions une couverture homogène, a montré les mêmes variations. Le contrôle *in silico* ayant confirmé l'efficacité de l'alignement bio-informatique, ces variations de couverture résultent vraisemblablement de biais d'amplification lors des étapes de PCR du protocole de préparation des librairies NGS ou lors du séquençage Illumina. Afin d'identifier si des séquences particulières pouvaient être à l'origine de ce biais, nous avons analysé la distribution des pourcentages en GC le long du génome en AAVr. Aucune corrélation n'a été mise en évidence entre les régions riches en GC et les régions faiblement couvertes. L'analyse des structures secondaires (serveur web mfold ; Zuker, 2003) n'a également pas révélé de motifs permettant d'expliquer un décrochement des polymérases dans ces régions. En conclusion, nous pouvons suggérer en comparant les couvertures des productions de vecteur et des plasmides que le génome du vecteur est encapsidé de façon homogène dans les particules AAVr. Cette observation réalisée avec un vecteur présentant un transgène d'une taille optimale pour les AAVr (4,5kb) nécessitera d'être comparé aux résultats obtenus avec d'autre transgènes.

Les polymorphismes nucléotidiques par rapport à la séquence de référence du génome AAVr ont été analysés dans les échantillons et le contrôle positif. Pour éliminer le bruit de fond correspondant aux erreurs de PCR et de séquençage (env 1/1000 bases), seuls les polymorphismes présents dans plusieurs échantillons et contrôles ont été pris en compte. Dans les échantillons, nous avons retrouvé 162, 13 et 1 substitutions nucléotidiques respectivement à une fréquence supérieure à 0,1%, 1% et 10% des bases séquencées. Des substitutions dans les mêmes bases et aux mêmes fréquences ont été retrouvées dans les contrôles plasmidiques. Ces variants n'ont donc pas été générés pendant la production des AAVr mais étaient présents dans les plasmides de production. Si leur fréquence individuelle est faible, leur effet cumulé pourrait avoir un impact fonctionnel sur le vecteur, ce qui pourrait entrainer la nécessité d'augmenter les doses de vecteurs injectés pour obtenir une efficacité thérapeutique. Le séquençage des plasmides de production des AAVr par NGS pourrait permettre d'aller plus loin dans la caractérisation des différentes populations de plasmides que l'actuel séquençage Sanger. Par ailleurs, il est intéressant de noter que la région présentant le plus de substitutions nucléotidiques correspond à la zone de faible couverture de l'hygroTK. Il est donc possible que les polymorphismes aient altéré l'alignement des *reads* par bio-informatique, contribuant à la baisse de couverture observée.

Au final, nous avons développé et validé le SSV-Seq, une méthode permettant de caractériser l'ADN encapsidé dans les particules AAVr par NGS (article accepté pour publication dans *Molecular Therapy - Nucleic Acids*). Son application permet d'aller plus loin que l'actuelle qPCR en analysant la couverture et les polymorphismes du transgène encapsidé et en permettant une analyse qualitative et quantitative des contaminants encapsidés rares et parfois inattendus.

IV. Discussion

Les techniques de séquençage de l'ADN permettent aujourd'hui de séquencer des milliards de bases en un temps réduit et pour un coût qui ne cesse de diminuer. Leur accessibilité est facilitée pour les laboratoires de recherche et leur utilisation permet maintenant de répondre à des questions biologiques variées dans de nombreux domaines, en dehors même de la génétique. Les deux projets menés pendant ma thèse ont eu pour but de développer le séquençage haut débit dans le domaine de la thérapie génique, en l'appliquant plus spécifiquement aux vecteurs AAVr. Le NGS a été utilisé depuis ses débuts en thérapie génique pour identifier les sites d'intégration génomique des vecteurs intégratifs responsables du développement de leucémies lors des premiers essais cliniques. Cependant, les techniques de biologie moléculaire ont été transposées aux technologies de NGS sans réelle adaptation, alors que le matériel biologique analysé est très différent de celui pour lequel le NGS a été initialement développé. Aujourd'hui, l'apparition de nouvelles méthodes de NGS, comme le séquençage ciblé, permet d'envisager des stratégies plus adaptées pour détecter les SI des vecteurs viraux. Les résultats de l'étude des SI et de la caractérisation des vecteurs AAVr par NGS vont être discutés, en développant notamment un point clé que ces études ont en commun : **la nécessité de développer des contrôles de nouvelle génération pour valider les découvertes faites en NGS.**

- **Origine des SI artéfactuels générés par le séquençage ciblé**

Longtemps controversé, le potentiel génotoxique des vecteurs AAVr est aujourd'hui avéré. Des études pré-cliniques chez la souris ont mis en évidence des carcinomes hépato-cellulaires après injection néo-natale d'une forte dose d'AAVr (Chandler et al., 2015; Donsante et al., 2007; Rosas et al., 2012). Ces observations entrainent la nécessité d'étudier la fréquence et le tropisme préférentiel d'intégration dans différentes espèces animales et dans différents cadres pathologiques. Or, les vecteurs AAVr présentent des caractéristiques qui limitent l'efficacité des méthodes actuelles de détection des SI : (1) une persistance très majoritaire sous forme épisomale, (2) de fortes structures secondaires au niveau des extrémités ITR du génome recombinant et (3) des délétions terminales fréquentes lors de la concaténérisation et de l'intégration. Nous avons donc appliqué le séquençage ciblé sur les AAVr et testé la méthode sur du tissu musculaire de souris injecté avec un AAVr.

Nous avons obtenu un important enrichissement en génome de vecteur, ce qui a permis de générer une importante profondeur de séquençage du vecteur par NGS. L'analyse des SI dans les données de NGS par bio-informatique a cependant révélé une fréquence des SI similaire

dans nos échantillons murins et dans les contrôles négatifs. Ainsi, aucun SI n'a pu être identifié avec certitude dans nos échantillons. Ces artéfacts ont vraisemblablement été générés lors de la préparation des librairies NGS. C'est l'étape de ligation des adaptateurs qui apparait comme critique car des fragments provenant du génome AAVr (formes épisomales) et du génome murin peuvent se liguer entre eux et former des fragments chimériques (**Figure 37**). Bien que largement ignoré de la plupart des études, cet artéfact a été décrit dès 2008 par une équipe du *Sanger Institute* qui a montré que ce phénomène impactait jusqu'à 5% des *reads* (Quail et al., 2008). Ils attribuaient la formation des fragments chimériques à l'adénylation incomplète des fragments qui aboutirait à l'inter-ligation de fragments à bouts francs. Ils ont proposé un système de double sélection de taille sur gel d'agarose avant et après l'étape de ligation pour éliminer les fragments chimériques qui sont deux fois plus longs que les fragments « normaux ». Avec ce système, ils réduisaient à 0,02% la proportion de fragments chimériques. Cependant, dans notre cas, ce système serait certainement insuffisant car les SI sont attendus en si faible quantité que même un faible bruit de fond empêcherait leur identification. Nous ne pouvons pas exclure la possibilité que la PCR ait généré des fragments chimériques par hybridation des fragments au niveau de micro-homologies. Ce phénomène semble toutefois négligeable par rapport aux artéfacts de ligation au vu du faible nombre de jonctions retrouvées au sein d'un même *read* par rapport à celles supportées par deux *reads paired-end* (données non montrées).

L'impossibilité de distinguer les fragments chimériques des SI vient de notre matériel biologique initial. En effet, les muscles de souris injectés avec un AAVr ont été extraits peu de temps post-injection et aucun phénomène cancéreux n'a pu se développer dans ce cours laps de temps. Chaque SI est donc présent de façon unique dans les cellules musculaires. Il en résulte l'impossibilité technique de les distinguer des fragments chimériques après séquençage ciblé. Dans la plupart des applications de NGS, ces fragments chimériques ne sont pas un problème puisque les évènements uniques sont filtrés par les logiciels de bio-informatique comme lors de l'identification des SI génomiques des virus (Li et al., 2013; Wang et al., 2013) ou dans l'analyse des translocations chromosomiques (Hayes and Li, 2013). Dans notre étude, la filtration des évènements uniques entraine également l'élimination des SI, comme en témoigne l'absence de SI retrouvés lors de l'analyse de nos données avec les logiciels développés pour détecter les SI viraux.

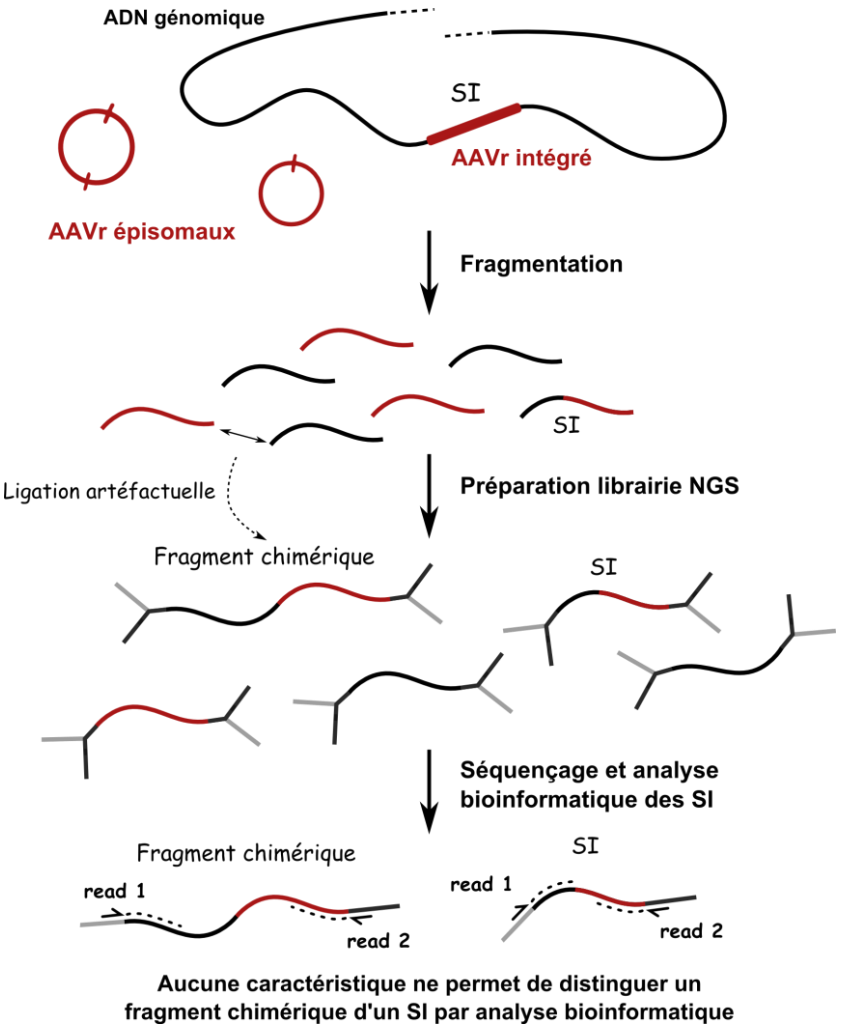


Figure 37 : Génération des fragments chimériques lors du protocole NGS. Lors de l'ajout de ligase, un fragment d'AAVr peut se lier à un fragment d'ADN génomique et créer un fragment chimérique qui a les mêmes caractéristiques qu'un vrai SI.

En 2013, l'équipe de Manfred Schmidt à Heidelberg a publié dans *Nature medicine* une étude des SI musculaires et hépatiques du vecteur AAV1-LPL^{S447X} (Glybera) chez les premiers patients injectés et chez des souris saines (Kaeppel et al., 2013). A l'aide de la LAM-PCR, ils ont conclu à une intégration aléatoire des vecteurs AAVr dans le génome humain et murin. Ils ont également mis en évidence des SI mitochondriaux, ce qui était surprenant au vu des connaissances actuelles du *trafficking* intra-cellulaire des AAVr. Afin de confirmer cette dernière observation ils ont par la suite procédé à un séquençage direct (Illumina) de l'ADN musculaire provenant d'un extrait enrichi en mitochondries d'un des animaux ayant préalablement reçu le vecteur AAVr par voie intramusculaire. 120 SI ont été identifié, dont 20 dans la mitochondrie. A partir de ces résultats, ils ont émis 3 conclusions fortes: (1) des cassures génomiques peuvent se produire tout le long du génome des AAVr lors de leur intégration, (2) la fréquence d'intégration est 10 fois supérieure à ce qui est actuellement décrit et (3) les pathologies mitochondriales pourraient bénéficier de l'intégration mitochondriale des AAVr. En analysant leurs données et au regard des résultats que nous avions générés en utilisant une

méthode d'analyse assez semblable, nous avons pu conclure que les SI identifiés par cette équipe après séquençage direct étaient très certainement de même nature artéfactuelle que dans notre étude. En effet, nous obtenons une répartition similaire des SI tout le long du génome de l'AAVr et une fréquence d'intégration similaire à environ 10^{-3} par génome AAVr. En réponse à ces conclusions vraisemblablement erronées nous avons publié un commentaire dans *Nature Medicine* montrant nos résultats et l'origine des artéfacts observés (Cogné et al., 2014) (Article en annexe).

L'importante différence entre l'étude de l'équipe de Manfred Schmidt et la nôtre est la présence d'un contrôle approprié. Dans notre étude, il n'était pas seulement composé d'ADN génomique de souris non injectées ou de plasmides seuls, lesquels n'auraient jamais révélés les artéfacts, mais il était une reconstitution *in vitro* des formes moléculaire de persistance des AAVr. C'est seulement dans cette situation que les fragments provenant des nombreux concatémères artificiels (mimant les formes épisomales du vecteur, **Figure 25**) peuvent se lier à des fragments d'ADN génomique pour former des jonctions chimériques AAV/génome cellulaire. Sans ce contrôle, nous aurions pu émettre de nombreuses conclusions erronées qui, au vu de l'inexpérience légitime des *reviewers* de thérapie génique en NGS, auraient pu être publiées.

Il est également important de filtrer par bio-informatique les évènements uniques potentiellement artéfactuels. Une équipe a utilisé une approche similaire à celle de l'équipe de Manfred Schmidt pour mettre en évidence des AAVr intégrés dans le génome mitochondrial, ils n'ont pas retrouvé de SI mais aucun filtre n'ayant été mentionné pour éliminer le bruit de fond, leur méthode n'était pas adaptée (Yu et al., 2013). Par ailleurs, une étude récemment publiée dans *Cell* s'est intéressée au profil d'intégration du VIH chez des patients en utilisant une LAM-PCR modifiée (Cohn et al., 2015). Ils ont mis en évidence des SI clonaux, supportés par plusieurs jonctions provenant de plusieurs lymphocytes CD4, et des SI uniques, supportés par une seule jonction. Pour détecter ces derniers, ils n'ont évidemment pas filtré les évènements uniques, sinon ils ne les auraient pas identifiés. Leur méthode avait déjà été utilisée par Janovitz *et al*. en 2013, mais leur analyse éliminait les jonctions supportées par un seul évènement (Janovitz et al., 2013). Je ne peux pas affirmer que les SI uniques observés par Cohn *et al.* sont de nature artéfactuelle, mais les contrôles utilisés (ADN de patients non infectés) ne préjugent pas de l'absence d'artéfacts générés par les étapes de ligation et d'amplification en présence de virus et d'ADN génomique. Un contrôle composé d'ADN de patients non infectés auquel aurait été ajouté des plasmides contenant le génome du VIH aurait été, de mon point de vue, plus approprié pour affirmer l'existence de SI non clonaux du VIH.

- **Avenir du séquençage ciblé en thérapie génique**

Le séquençage ciblé a déjà été utilisé avec succès pour caractériser les SI de virus (Depledge et al., 2011; Duncavage et al., 2011; Li et al., 2014), de vecteurs viraux (Ustek et al., 2012) et de transgènes (Dubose et al., 2013). Néanmoins, ces études ont en commun la clonalité de leur matériel biologique de départ (prélèvements tumoraux, cultures cellulaires ou animaux transgéniques). Cela signifie qu'un même évènement intégratif est présent dans de nombreux noyaux cellulaires. Dans cette situation, la présence des SI est confirmée par la détection de plusieurs fragments représentant le même SI. Ainsi il est aisé de filtrer les artéfacts dans les données de séquençage en éliminant les évènements uniques par bio-informatique. Dans notre étude, nous avons choisi de prédire le risque intégratif des AAVr en étudiant les caractéristiques de leur intégration polyclonale au sein de tissus non tumoraux. Dans ce contexte, nous avons vu que le séquençage ciblé n'est pas adapté *in vivo* pour des vecteurs peu intégratifs tels que les AAVr.

En revanche, le séquençage ciblé peut remplacer les techniques actuelles à base de PCR pour détecter les SI *in vitro* (culture cellulaire) et en situation clonale *in vivo*. L'équipe de Manfred Schmidt a indiqué dans une communication orale lors du congrès 2014 de l'*European society of gene and cell therapy* avoir récemment adopté le séquençage ciblé et identifié des SI de vecteurs lentiviraux. La limite de détection mentionnée était cependant de l'ordre de 100 à 1000 SI par microgramme, ce qui est cohérent avec ce que nous avons observé avec notre contrôle plasmidique. La sensibilité semble limitée par l'efficacité de l'étape d'hybridation et par les étapes de lavage du protocole de séquençage ciblé. L'optimisation de ces étapes pourrait permettre d'améliorer la sensibilité mais elle restera certainement en dessous des techniques de LAM-PCR et de LM-PCR. Au final, les récentes améliorations apportées à la LM-PCR permettent à cette technique, lorsqu'elle est applicable, d'être plus efficace et moins chère que le séquençage ciblé. Par exemple, la fragmentation aléatoire par sonication permet de remplacer l'utilisation d'enzymes de restriction et donc de quantifier les différents clones au sein d'une population cellulaire (Gillet et al., 2011; Hacein-Bey-Abina et al., 2014). Cependant, elle reste surtout applicable lorsque la forme moléculaire intégrée du virus/vecteur est stable, comme c'est le cas pour le VIH. Lorsqu'un phénomène clonal est étudié avec des provirus réarrangés et partiellement tronqués lors de l'intégration, le séquençage ciblé peut s'avérer utile en capturant la totalité du génome.

Finalement, en ce qui concerne les vecteurs AAVr, aucune méthode actuelle n'est réellement adaptée en l'absence d'évènement tumoral, lorsque les épisomes prédominent. La LM-PCR est probablement le meilleur candidat, mais il est encore nécessaire d'apporter des

améliorations spécifiques comme l'utilisation de polymérases permettant de traverser efficacement la forte structure secondaire des ITR.

- **Fiabilité du NGS pour caractériser les productions d'AAVr ?**

A l'heure où les vecteurs AAVr commencent à être utilisés chez les patients pour traiter un large éventail de pathologies héréditaires et acquises, une meilleure connaissance du produit thérapeutique injecté devient indispensable. Les agences réglementaires exigent aujourd'hui des analyses plus poussées que la qPCR pour caractériser l'ADN contaminant résiduel dans le produit fini. En effet, l'EMA avait contraint Uniqure, la société qui commercialise le Glybera (premier AAVr approuvé en Europe), à mieux caractériser les contaminants dans leurs AAVr produits par système baculovirus (*Glybera: EPAR - Public assessment report. 2012*). Ils ont ainsi analysé les ADN contaminants par NGS, sans pour autant publier la méthode ni les résultats. Dans ce contexte, il nous a semblé important de développer et d'optimiser un protocole NGS qui soit proposé à la communauté comme contrôle efficace des productions AAVr.

Les premiers tests du protocole de SSV-Seq que nous avons développé ont montré une très bonne **reproductibilité** de la quantification des différents composants ADN présents dans les préparations d'AAVr. Il était important de vérifier l'aspect quantitatif du NGS car les étapes de synthèse du second brin, de PCR, de génération des *clusters* et de séquençage sont toutes susceptibles de biaiser la représentation relative des différents composants. La très bonne corrélation obtenue entre les quantifications par qPCR et par SSV-Seq témoigne de la **fiabilité** des résultats obtenus. Néanmoins, les proportions des contaminants obtenus par NGS sont 2 à 3 fois supérieures aux résultats générés par qPCR. S'il n'est pour l'instant pas déterminé quelle est la méthode de quantification la plus fiable, la qPCR a montré, dans une étude récente, une variabilité de 2log dans les titres obtenus par différents laboratoire avec un même protocole (Ayuso et al., 2014). La comparaison par rapport à une gamme quantifiée de façon approximative apparait comme le point critique. En NGS, l'avantage est la comparaison directe des proportions au sein d'un même *run*. De plus, une étude a montré une bonne reproductibilité inter-laboratoire du NGS avec un protocole de séquençage d'ARN présentant des étapes enzymatiques similaires à notre protocole ('t Hoen et al., 2013). L'aspect quantitatif du NGS est primordial dans le domaine du séquençage ARN, pour lequel un contrôle externe composé d'ARN de tailles différentes et présents en concentrations déterminés a été développé (Jiang et al., 2011). Un travail similaire sur l'ADN serait profitable pour tester notre protocole dans des conditions contrôlées. C'était en partie l'objectif de notre contrôle positif, mais des imprécisions dans les quantifications initiales ont altéré sa pertinence. L'apport du NGS est néanmoins indéniable dans la quantification des larges contaminants (plasmide auxiliaire,

génome des cellules productrices) pour lesquels la qPCR ne cible qu'une infime partie. Le NGS pourra permettre de comparer plus efficacement l'élimination de ces contaminants par les différents procédés de production et de purification.

Le contrôle négatif utilisé dans notre étude, composé d'ADN de phage lambda, a été essentiel. Il était destiné à évaluer le niveau de contamination environnementale lors de la préparation des échantillons (contamination exogène, contamination inter-échantillon…). Son analyse a montré un bruit de fond pour toutes les références, ce qui nous a permis de fixer une limite de détection spécifique de chaque référence par SSV-Seq. La **contamination en génome humain** a été la plus importante (2000 à 3000 *reads)* au regard du faible nombre de *reads* alignés sur le génome humain dans nos échantillons (5000 à 40000 *reads*). Son origine est vraisemblablement une contamination par de l'ADN génomique des manipulateurs lors de la préparation des échantillons. En effet, la présence d'ADN humain contaminant a déjà été mise en évidence dans les données générées par NGS (Cibulskis et al., 2011; Gansauge and Meyer, 2014; Jun et al., 2012; Longo et al., 2011). Par ailleurs, nous avons retrouvé un nombre important de *reads* non alignés dans un échantillon de cette étude et plus récemment dans tous les échantillons d'un nouveau *run*. Après analyse, ces *reads* non identifiés se sont avérés provenir de génomes d'entérobactéries (*Klebsiella sp., E.coli*). Cette famille bactérienne provient de la flore commensale endogène, la contamination est donc vraisemblablement d'origine manuportée lors de la préparation des échantillons. En plus de réduire le nombre de *reads* utiles, nous avons remarqué que les nombreux fragments bactériens pouvaient s'aligner partiellement sur le génome humain par des zones d'homologies d'environ 20pb, aboutissant à une surestimation du génome humain dans les échantillons. La filtration des *reads* s'alignant sur de courtes séquences a permis de réduire l'effet de ce phénomène, sans pour autant résoudre le problème. La contamination bactérienne a déjà été décrite en NGS (Strong et al., 2014), des séquences bactériennes ayant même été retrouvées dans les génomes de référence d'espèces eucaryotes (Laurence et al., 2014). Cependant, les contaminations décrites sont nettement moins importantes que celles que nous avons observées, de l'ordre du millier de *reads* contre environ 1 million dans certains de nos échantillons. Ainsi, afin de réduire les contaminations environnementales dans nos échantillons (ADN humain et bactérien), nous avons mis en place de nouvelles mesures de manipulations : le travail sous hotte à flux laminaire, une importante décontamination de surface avant manipulation et un traitement par ultraviolets des réactifs et matériels utilisés. Dans le récent *run* réalisé avec ces mesures, une nette diminution du bruit de fond dans le contrôle négatif a été observée (une dizaine de *reads* dans l'ADN génomique).

La technologie Illumina permet le séquençage simultané de plusieurs échantillons au cours d'un même *run* en identifiant chaque échantillon par une courte séquence ADN de 6 bases dans

l'adaptateur NGS, appelée index. Cependant, le contrôle négatif a mis en évidence qu'une mauvaise attribution d'un *read* à son échantillon est possible, ce qui aboutit à une **contamination croisée entre échantillons**. Malgré une fréquence faible, cela peut avoir un impact non négligeable sur des références faiblement représentées dans notre étude (ADN génomique humain), lors de l'étude de rares mutations somatiques en cancérologie (Campbell et al., 2010) ou encore lors de l'étude de l'hétéroplasmie mitochondriale (He et al., 2010). L'origine de cet artéfact est sans doute la présence de *clusters* mixes sur la *flowcell* Illumina, composés d'un mélange de deux fragments, pour lesquels un index est attribué au mauvais *read* (Kircher et al., 2012). Kircher *et al*. ont estimé l'incidence du phénomène à 0,3% des séquences. Pour y remédier ils proposent (1) de filtrer les *reads* ayant un index avec une mauvaise qualité de séquençage par bio-informatique et (2) de préparer les librairies NGS en identifiant les échantillons par deux index au lieu d'un seul (double indexage). Nous avons récemment expérimenté les 2 suggestions de Kircher *et al* lors d'un nouveau *run* de SSV-Seq non présenté dans cette thèse. L'utilisation conjointe du filtrage des index et du double indexage a permis une très nette diminution des contaminations croisées entre échantillons. Cela a contribué, avec les mesures prises pour la manipulation des échantillons, à réduire considérablement le bruit de fond dans le contrôle négatif (diminution de 3 logs de la contamination en PhiX). Nous allons donc par la suite continuer à utiliser la technologie du double index.

La possibilité d'analyser les séquences d'ADN encapsidées dans les particules AAVr est un avantage certain du NGS qui pourrait permettre de mieux comprendre le processus d'encapsidation du génome recombinant, qui reste pour le moment peu connu. La couverture de séquençage du génome AAVr et des contaminants doit cependant être étudiée avec précaution et toujours en comparaison avec un contrôle plasmidique. En effet, certaines séquences ADN ont une tendance à être systématiquement moins bien séquencées par NGS (Ross et al., 2013). Bien que l'étape de PCR de la préparation des librairies NGS soit la première cause suspectée (Aird et al., 2011), le pourcentage de GC et les structures secondaires prédites n'expliquent pas ces variations dans nos échantillons. Ainsi, il est fort probable que d'autres étapes du protocole participent à ce biais. Par exemple, la fragmentation par sonication n'est pas totalement aléatoire et semble privilégier des ruptures au niveau de séquences particulières (Poptsova et al., 2014). De plus, comme montré en séquençage ARN, la synthèse du brin complémentaire par des hexamères aléatoires aboutit à une représentation biaisée des références (Hansen et al., 2010). Enfin, le séquençage Illumina lui-même pourrait induire un biais, lors de la génération des clusters ou lors de la réaction de séquençage (Nakamura et al., 2011). Ainsi, il est encore nécessaire de continuer les efforts de normalisation et de standardisation des résultats de NGS.

Pour conclure, nous avons donc développé un protocole NGS adapté à la caractérisation de l'ADN encapsidé dans les AAVr. Sous réserve d'utiliser des contrôles appropriés, son application sur diverses productions d'AAVr pourra permettre la mise au point de méthodes optimisées de production et de purification des vecteurs. De plus, en réponse aux exigences des autorités réglementaires, sa capacité à quantifier les larges contaminants et à détecter des contaminants non attendus en fait un test important pour caractériser les lots d'AAVr destinés à être injectés chez les patients lors des essais cliniques et lors de la future mise sur le marché de nouveaux médicaments de thérapie génique.

V. Conclusion

Les travaux réalisés pendant cette thèse ont mis en évidence à la fois la puissance et les limites actuelles du séquençage haut débit. Le virage technologique incarné par ces nouveaux séquenceurs ne doit pas faire oublier que les nombreuses réactions enzymatiques encore nécessaires en amont pour séquencer les échantillons sont toutes susceptibles d'induire des biais lors de l'analyse des résultats NGS. Ainsi, pour chaque application, une attention particulière doit être apportée au développement de contrôles appropriés. Il faut également garder un œil critique sur les technologies de séquençage utilisées, car chacune d'elles présentent des limites qu'il faut connaitre pour en évaluer l'impact éventuel sur la qualité des résultats générés. Enfin, l'analyse bio-informatique est un point qui est loin d'être anecdotique, l'utilisation rigoureuse des programmes existants et le développement de nouveaux programmes doit se faire en ayant préalablement évalué les risques de faux négatifs et de faux positifs afin d'appliquer des filtres adaptés aux données de NGS.

Malheureusement, les artéfacts du NGS combinés à la nécessité de détecter de très rares évènements dans nos échantillons biologiques ne nous ont pas permis d'analyser le risque intégratif des AAVr dans un contexte déficient en dystrophine. De nouvelles méthodes plus appropriées devront être développées pour estimer ce risque dans différents contextes pathologiques. Cependant, même si l'inquiétude est légitime après la publication d'études montrant le développement de tumeurs hépatiques chez la souris, il n'est pour le moment pas déterminé si ces résultats sont transposables chez l'Homme. En effet, aucun évènement tumoral n'a pour l'instant été observé chez l'Homme malgré un grand nombre d'essais cliniques réalisés. Il faudra cependant être vigilent dans des contextes particuliers, notamment lorsque de fortes doses d'AAVr seront injectées chez des nouveaux-nés. En ce qui concerne la myopathie de Duchenne, aucune tumeur n'a été détectée chez les chiens dystrophiques du laboratoire injectés avec une forte dose d'AAVr à plus d'un an post-injection. Le risque tumoral semble ainsi minime et cet aspect ne devrait pas empêcher le déroulement de futurs essais cliniques et la possible mise sur le marché de nouveaux AAVr thérapeutiques.

Lorsque les biais sont maitrisés, le NGS a en revanche le potentiel de révolutionner de nombreux domaines, comme c'est le cas actuellement en génomique. Le protocole que nous avons développé pour caractériser l'ADN encapsidé dans les particules AAVr est un premier pas important dans l'optique de son adoption en contrôle qualité pour la thérapie génique. En effet, l'importante augmentation du nombre d'essais cliniques utilisant les AAVr impose de développer des technologies permettant de garantir la sécurité des patients. Notre méthode

pourrait accompagner les laboratoires de recherche et développement afin d'améliorer les processus de production et de purification des vecteurs recombinants AAV. Du point de vue des autorités règlementaires le SSV-Seq permettrait de s'assurer de la qualité des préparations de vecteurs en termes de contaminants ADN résiduels. Cependant, il est important de souligner que notre méthode est seulement descriptive et ne peut en aucun cas fournir d'information fonctionnelle, il faudra développer de nouveaux tests pour déterminer leur éventuel impact biologique (exemple : projet en cours au laboratoire pour déterminer l'impact fonctionnel des microARN encapsidés dans les AAVr).

Il est également important de noter que le NGS est encore onéreux et difficile à analyser en routine pour beaucoup de laboratoires. Pour un temps encore, les techniques standards de biologie moléculaire vont continuer à être largement utilisées. Néanmoins, au vu du développement extrêmement rapide du NGS et de l'enjeu économique considérable, il est fort à parier que de nouvelles technologies de séquençage à haut débit trouveront leur place au côté de la PCR dans l'arsenal analytique de nombreux laboratoires de biologie moléculaire.

D'un point de vue personnel, cette thèse m'aura permis de développer un regard critique sur les nouvelles technologies. S'il est naturel d'en envisager le meilleur, il est indispensable d'en attendre le pire. Mais au-delà de la complexité d'analyse, les nouvelles techniques de séquençage de l'ADN ont pourtant bien le pouvoir de révolutionner la recherche biomédicale et la médecine. Elles s'inviteront bientôt dans le quotidien des chercheurs et des médecins et la question ne sera plus de savoir si elles sont fiables, mais quel impact sur la prise en charge des patients et sur les projets de recherche elles pourront avoir.

VI. Matériel et méthodes

VI.1 Production des vecteurs AAVr

L'**AAV2/8-pRSV-GFP-SV40** contient le promoteur du Rous Sarcoma Virus (pRSV) suivi par un court intron synthétique (plasmide pCI, Promega), la séquence codante de l'eGFP (*Green fluorescent protein*) et le signal de polyadénylation SV40. L'**AAV2/8-pCMV-GFP-hTK-BGH** contient le promoteur du cytomégalovirus (pCMV), les séquences codantes pour la GFP et l'hygromycine thymidine kinase (hTK) et le signal de polyadénylation de l'hormone de croissance bovine (BGH). Les cassettes d'expression ont été clonées entre les ITR de l'AAV2.

Les vecteurs de grade recherche simple brin AAV2/8 ont été produits par la plate-forme de production de vecteurs viraux de l'UMR1089 à Nantes. Brièvement, des cellules HEK293 ont été transfectées avec le plasmide vecteur et le plasmide pDP8 contenant les gènes Rep de l'AAV2, Cap de l'AAV8 et les gènes auxiliaires adénoviraux. Les cellules ont été récoltées 48h à 72h post-transfection, centrifugées à faible vitesse et les AAVr2/8 contenus dans le surnageant ont été extraits. L'AAV2/8-pRSV-GFP-SV40 a été purifié par une double ultracentrifugation sur gradient de chlorure de césium (ClCs) suivi par une dialyse avec du dPBS. L'AAV2/8-pCMV-GFP-hTK-BGH a été purifié par (1) chromatographie d'échange ionique, (2) chromatographie d'affinité (AVB Sepharose High Performance, GE Healthcare) et (3) ClCs. Puis les vecteurs ont été concentrés par filtration tangentielle et formulés dans du dPBS avec 0,001% de pluronic.

VI.2 Préparation des librairies NGS

Toutes les étapes de la préparation de la librairie ont été effectuées dans des tubes non-adhérents de 1,5mL (Non-stick RNase-free Microfuge, Life technologies). Toutes les enzymes et les tampons utilisés ont été obtenus chez New Englands BioLabs. Ce protocole a été adapté de Kozarewa and Turner, 2011**.**

Synthèse des adaptateurs : Six adaptateurs contenant des index différents (séquences ADN de 6 paires de bases) ont été générés (**Tableau 10**), ils permettent le multiplexage de 6 échantillons différents dans un même *run* NGS. Pour synthétiser les adaptateurs, 20µL d'oligonucléotide « P5 » 100µM, 20µL d'oligonucléotide « P7 » indexé 100µM, 5µL de tampon de T4 DNA ligase (New England Biolabs) et 5µL d'eau ont été mélangés dans un tube pour PCR de 0,2mL (Axygen). Puis dans un thermocycleur Step One Plus (Applied Biosystems) le programme suivant a été lancé : +0,5°C/s jusqu'à 97,5°C, maintien de 97,5°C pendant 150s et décente en température de 0,1°C toute les 5s jusqu'à 4°C. L'hybridation a été vérifiée sur puce

High Sensitivity (Bioanalyzer 2100, Agilent). Les adaptateurs à 40µM ont été conservés à -20°C.

Oligonucléotide « P5 »		
5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC**GCTCTTCCGATC***T 3'		
Oligonucléotide « P7 »	**Séquence de l'index**	**Oligonucléotide « P7 »**
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	ACAGTG	ATCTCGTATGCCGTCTTCTGCTTG 3'
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	GCCAAT	ATCTCGTATGCCGTCTTCTGCTTG 3'
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	CTTGTA	ATCTCGTATGCCGTCTTCTGCTTG 3'
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	GTGAAA	ATCTCGTATGCCGTCTTCTGCTTG 3'
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	CGATGT	ATCTCGTATGCCGTCTTCTGCTTG 3'
5' P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC	CAGATC	ATCTCGTATGCCGTCTTCTGCTTG 3'

Tableau 10 : Séquences des oligonucléotides utilisés pour la synthèse des adaptateurs. Les bases en gras des oligonucléotides P5 et P7 s'hybrident pour former un adaptateur indexé en « Y ». La liaison * est une liaison phosphorothioate. P : phosphate.

Sonication : L'ADN génomique ou extrait des particules AAVr (200ng à 5µg) ont été placés dans des microtubes 0,5mL Bioruptor (Diagénode) et le volume ajusté à 100µl avec du TE (10 :1). La sonication s'est effectuée en bain avec le Bioruptor UCD200 (Diagenode). Les tubes ont été placés dans un portoir contenant 12 positions, pour assurer une meilleure reproductibilité les places vides ont été remplies avec des tubes contenant 100µl de TE (10 :1). La température du bain a été maintenue à +4°C par le Water Cooler (Diagenode). Les conditions de sonication ont été adaptées en fonction de la taille des fragments désiré, pour 500pb : 4 cycles 30sec ON / 90sec OFF et pour 300pb : 12 cycles 30sec ON / 90sec OFF, à sur le réglage « low intensity ». Le profil de fragmentation a été systématiquement vérifié par migration de 1µL d'échantillon sur Bioanalyzer 2100 (Agilent) avec une puce D1K.

La **réparation des extrémités et leur phosphorylation** en 5' ont été effectuées en ajoutant aux échantillons: 50U de T4 Polynucléotide Kinase, 5U de fragment Klenow de DNA polymérase I, 15U de T4 DNA Polymérase, 4µL d'un mix dNTP 10mM, 10uL de tampon de T4 DNA ligase 10x et 30µL d'eau. L'incubation était de 30min à température ambiante (RT).

L'ajout d'un A en 3' a été réalisé en ajoutant aux échantillons purifiés: 15U de fragment de Klenow (3'-5' exo-), 10µL d'ATP 1mM et 5µL de NEBuffer 2 10x. L'incubation était de 30min à 37°C, réalisée dans un bain à sec chauffant.

Entre ces étapes enzymatiques, une purification par billes magnétiques (SPRIselect - Beckman Coulter) a été réalisée selon les recommandations du fournisseur.

Pour la **ligation des adaptateurs**, 2µL d'adaptateurs 40µM (un index différent pour chaque échantillon) ont été mélangés aux 15µL d'ADN purifié de l'étape précédente suivi d'une incubation de 15min à RT après l'ajout de 25µL de tampon de Quick ligase, 10000U de T4 DNA ligase et 3µL d'eau. Afin d'éliminer les dimères d'adaptateurs potentiellement formés, deux purifications successives avec 50µL de SPRI (1x) ont été réalisées.

Une **amplification** des fragments ligués a ensuite été réalisée avec une PfuUltra II Fusion Hotstart DNA polymerase (Agilent technologies) et les amorces P5-F (5'-AATGATACGGCGACCACCG-3') et P7-R (5'-CAAGCAGAAGACGGCATAC-3'). Le programme de PCR était de 2 min à 95°C, suivi de 15 à 20 cycles d'amplification (20s à 95°C, 20s à 60°C et 15s à 72°C) et d'une élongation finale de 3min à 72°C.

L'efficacité de la préparation des librairies a été évaluée sur puce D1K (Bioanalyzer 2100). Leur quantification a été effectuée sur ce même tracé par intégration de l'aire sous courbe.

VI.3 Protocole de séquençage ciblé sur les AAVr

VI.3.1 Expérimentation animale

Les expériences ont été réalisées sur des souris mâles C57BL/6J et mdx^{4CV} âgées de 5 à 7 semaines. Elles ont été anesthésiées avec une injection péritonéale de Kétamine/Xylazine puis injectées en intra-musculaire (IM) bilatérale dans les *Tibialis anterior* (TA). Les souris ont reçu deux fois 30µL de dPBS (1 souris) ou de vecteur (3 C57BL/6J et 3 mdx^{4CV}) soit une dose de 7,2x10^{11} vg ou 2,4x10^{13} vg/kg. Les animaux ont été euthanasiés 15 jours post-injection par exsanguination sous anesthésie par inhalation d'isoflurane. Les TA injectés ont été récupérés, congelés instantanément dans de l'azote liquide avant conservation jusqu'à extraction à -80°C.

VI.3.2 Extraction de l'ADN et quantification du vecteur

Extraction de l'ADN génomique : L'ADN a été extrait à partir de 50 à 100 mg de muscle avec le kit Gentra Puregene Blood Kit (Qiagen). Les tissus congelés ont été mélangés avec une solution de lyse cellulaire puis broyés mécaniquement par le Tissue Lyser II (Qiagen) 30s à 30Hz. L'ADN a été récupéré suivant les instructions du fournisseur et conservé à -20°C. La concentration et la pureté (ratio A260/A280) de l'ADN ont été mesurées par Nanophotomètre (Implen).

Quantification du vecteur par PCR quantitative (qPCR). L'albumine a été quantifiée par qPCR Sybr Green sur 5µL d'ADN à environ 10ng/µL dans un volume total de 20µL. Le cycle suivant a été réalisé dans un Step One Plus (Applied Biosystems) : 20s à 95°C puis 45 cycles de 3s à 95°C et 30s à 60°C. Une courbe de fusion a été réalisée pour vérifier la spécificité

de l'amplification. Une qPCR Taqman pour la GFP a été réalisée dans les mêmes conditions avec un cycle de 20s à 95°C et 40 cycles de 1s à 95°C et 20s à 60°C. Amorces utilisées pour l'albumine : *sens* 5' acatagcttgcttcagaacggt 3' et *antisens* 5' agtgtcttcatcctgccctaaa 3'. Amorces utilisées pour la GFP : *sens* 5' actacaacagccacaacgtctatatca 3' et *antisens* 5' ggcggatcttgaagttcacc 3'.

VI.3.3 Librairie artificielle

Les **formes intégrées des AAVr** ont été simulées par clonage moléculaire de la cassette AAV RSV GFP dans un grand plasmide (pAd easy, 36kb), qui a ensuite été linéarisé.

Les **concatémères** ont été réalisés par digestion de 10µg du plasmide AAV-RSV-GFP avec XmaI (New England Biolabs). La bande correspondante au genome AAVr (ITR à ITR) a été extraite gel d'agarose 1% puis purifiée par Nucleotrap (Macherey Nagel). Pour l'étape de ligation, 17µL de cassette AAVr purifiée (34ng/µL), 2µL de tampon de T4 DNA ligase 10x et 1µL de T4 DNA ligase 400U/µL ont été incubés 16h à 16°C. La ligase a ensuite été inactivée 10min à 60°C. Puis un traitement avec de la Plasmid-Safe ATP-dependent DNase (Epicentre) a permis de digérer les formes linéaires : 20µL du produit de ligation ont été mélangés avec 3µL de tampon de PS DNase 10x, 1,2µL d'ATP 25mM, 4,8µL d'eau et 1µL de PS DNase. L'incubation était de 2h à 37°C suivi par 30min à 70°C pour inactiver la PS DNase. La présence des concatémères a ensuite été visualisée par électophorèse sur gel d'agarose 0,7%.

VI.3.4 Capture des génomes AAVr

Après extraction de l'ADN génomique des TA injectés et préparation de la librairie NGS (sans PCR finale), 500ng d'ADN a été déshydraté par SpeedVac Plus (Savant) pendant environ 2h à faible température. Le culot a ensuite été reconstitué dans 3,4µL d'eau, volume exigé par le protocole de capture. La capture a été effectuée avec le kit MYbaits (MYcroarray). Des sondes d'ARN d'une longueur de 80 bases se chevauchant sur 40 bases ont été synthétisées de façon à recouvrir la quasi-totalité du génome du vecteur. Le protocole a été réalisé selon les instructions du fournisseur (manuel 1.3.8). L'incubation permettant l'hybridation entre les ARN et le vecteur a été effectuée dans un thermocycleur Veriti (Applied Biosystems) pendant 36h. Les billes magnétiques recouvertes de streptavidine (Dynabeads MyOne Streptavidin C1, Invitrogen) ont été utilisées pour capturer le duplex ADN/ARN biotinylé. Les ARN ont ensuite été lysés par du NaOH 100mM et l'ADN purifié par SPRI.

Les librairies enrichies en génomes d'AAVr ont ensuite été amplifiées par PCR selon le protocole décrit dans la préparation de librairie (III.2).

VI.4 Protocole de caractérisation de l'ADN encapsidé

A partir de la production d'AAV2/8-pCMV-GFP-hTK-BGH purifié par les trois méthodes de référence, le volume correspondant à $2x10^{11}$ copies de vecteur a été prélevé et mélangé avec 5% de phage lambda (en masse d'ADN totale). Afin de digérer l'ADN présent à l'extérieur des capsides, les échantillons ont ensuite été traités avec deux nucléases (10 U Baseline Zero – Epicentre et 40U Plasmid-Safe DNAse – Epicentre) pendant 2h à 37°C. La réaction a été arrêtée en ajoutant 3mM d'EDTA et en incubant 30min à 75°C. Pour évaluer l'efficacité du traitement par DNAse, pour chaque condition, un échantillon a subit cette digestion et l'autre non. Ensuite, tous les échantillons ont été traités avec 0,5mg de Proteinase K (Macherey Nagel) pendant 3h à 55°C puis avec 10U de RNAseA (Qiagen) pendant 15min à 37°C. Enfin, l'ADN total a été extrait avec le kit Gentra Puregene kit (Qiagen) selon les recommandations du fournisseur.

Pour la **synthèse du second brin,** l'ADN a été dénaturé en chauffant à 95°C pendant 5min puis, 58µM d'hexamères dégénérés (NEB), 2mM de dNTPs et 10U de DNA polymerase I (NEB) ont été ajoutés. Les échantillons ont été incubés 1h à 37°C puis la réaction a été arrêtée par ajout de 0,1mM d'EDTA. Une préparation de librairie NGS a ensuite été réalisée sur ces ADN double brin comme décrit dans la section III.2.

VI.5 Séquençage NGS et analyses bio-informatiques

Les différents échantillons présentant chacun un index ont été mélangés à une concentration finale de 2nM (multiplexage). Du phiX (Illumina) à hauteur de 1 à 5% a été ajouté, permettant de calculer le nombre de clusters formés et d'apporter de la diversité de séquence. Après une dénaturation selon le protocole d'Illumina, 6 à 10 pM de la librairie multiplexe ont été injectés sur la flow cell d'un MiSeq ou d'un HiSeq (Illumina). Un séquençage « paired-end » de 2x100pb ou 2x150pb a été réalisé.

Les séquences générées ont été alignées sur les séquences de référence avec le logiciel BWA (Li and Durbin, 2009). Les séquences de référence pour l'analyse des SI des AAVr étaient le génome murin (*mm10*) et le génome de l'AAV RSV GFP et pour la caractérisation de l'ADN encapsidé : le génome humain (*hg18*, approximation du génome des cellules HEK293), le génome de l'AAV CMV GFP hTK, les séquences des plasmides de production et les génomes du PhiX et du Phage lambda.

L'analyse bio-informatique des sites d'intégration a été réalisée par deux stratégies, la première permettant la détection des jonctions AAV/génome par analyse de l'information portée par les séquences « paires » (les deux extrémités des fragments) par le programme **FindMyVirus** (https://github.com/lindenb/jvarkit/wiki/FindMyVirus) et la seconde en

analysant individuellement chaque séquence pour trouver des fragments chimériques par le programme **Chimera_Finder** (http://a-slide.github.io/Chimera_Finder/).

Pour l'analyse des séquences issues des productions d'AAVr, les fichiers brutes BCL ont été démultiplexés avec le logiciel CASAVA (Illumina, San Diego, CA, USA) grâce aux index fournis pour chaque échantillon. Les fichiers fastq ont été analysés par le programme **ContaVect** (v0.2, https://github.com/a-slide/ContaVect) en utilisant le fichier de configuration. Le programme génère des fichiers standards de génomique (bam, bed, bedgraph). Les variants nucléotidiques du génome AAVr ont été calculés à partir des fichiers BAM par le programme **MiniCaller (**https://github.com/lindenb/jvarkit/wiki/MiniCaller) qui génère des fichiers VCF.

VII. Bibliographie

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Aiuti, A., Slavin, S., Aker, M., Ficara, F., Deola, S., Mortellaro, A., Morecki, S., Andolfi, G., Tabucchi, A., Carlucci, F., et al. (2002). Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science *296*, 2410–2413.

Alazard-Dany, N., Nicolas, A., Ploquin, A., Strasser, R., Greco, A., Epstein, A.L., Fraefel, C., and Salvetti, A. (2009). Definition of herpes simplex virus type 1 helper activities for adeno-associated virus early replication events. PLoS Pathog. *5*, e1000340.

Allay, J.A., Sleep, S., Long, S., Tillman, D.M., Clark, R., Carney, G., Fagone, P., McIntosh, J.H., Nienhuis, A.W., Davidoff, A.M., et al. (2011). Good manufacturing practice production of self-complementary serotype 8 adeno-associated viral vector for a hemophilia B clinical trial. Hum. Gene Ther. *22*, 595–604.

Appaiahgari, M.B., and Vrati, S. (2015). Adenoviruses as gene/vaccine delivery vectors: promises and pitfalls. Expert Opin. Biol. Ther. *15*, 337–351.

Arechavaleta-Velasco, F., Gomez, L., Ma, Y., Zhao, J., McGrath, C.M., Sammel, M.D., Nelson, D.B., and Parry, S. (2008). Adverse reproductive outcomes in urban women with adeno-associated virus-2 infections in early pregnancy. Hum. Reprod. Oxf. Engl. *23*, 29–36.

Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., and O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat. Biotechnol. *33*, 296–300.

Atchison, R.W., Casto, B.C., and Hammon, W.M. (1965). ADENOVIRUS-ASSOCIATED DEFECTIVE VIRUS PARTICLES. Science *149*, 754–756.

Ayuso, E., Mingozzi, F., Montane, J., Leon, X., Anguela, X.M., Haurigot, V., Edmonson, S.A., Africa, L., Zhou, S., High, K.A., et al. (2010). High AAV vector purity results in serotype- and tissue-independent enhancement of transduction efficiency. Gene Ther. *17*, 503–510.

Ayuso, E., Blouin, V., Lock, M., McGorray, S., Leon, X., Alvira, M.R., Auricchio, A., Bucher, S., Chtarto, A., Clark, K.R., et al. (2014). Manufacturing and characterization of a recombinant adeno-associated virus type 8 reference standard material. Hum. Gene Ther. *25*, 977–987.

Bantel-Schaal, U., and zur Hausen, H. (1984). Characterization of the DNA of a defective human parvovirus isolated from a genital site. Virology *134*, 52–63.

Bell, P., Wang, L., Lebherz, C., Flieder, D.B., Bove, M.S., Wu, D., Gao, G.P., Wilson, J.M., and Wivel, N.A. (2005). No evidence for tumorigenesis of AAV vectors in a large-scale study in mice. Mol. Ther. J. Am. Soc. Gene Ther. *12*, 299–306.

Benetatos, L., Hatzimichael, E., Londin, E., Vartholomatos, G., Loher, P., Rigoutsos, I., and Briasoulis, E. (2013). The microRNAs within the DLK1-DIO3 genomic region: involvement in disease pathogenesis. Cell. Mol. Life Sci. CMLS *70*, 795–814.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., et al. (2013). An estimation of the number of cells in the human body. Ann. Hum. Biol. *40*, 463–471.

Biffi, A., Montini, E., Lorioli, L., Cesani, M., Fumagalli, F., Plati, T., Baldoli, C., Martino, S., Calabria, A., Canale, S., et al. (2013). Lentiviral Hematopoietic Stem Cell Gene Therapy Benefits Metachromatic Leukodystrophy. Science *341*, 1233158.

Blacklow, N.R., Hoggan, M.D., and Rowe, W.P. (1967). Isolation of adenovirus-associated viruses from man. Proc. Natl. Acad. Sci. U. S. A. *58*, 1410–1415.

Blacklow, N.R., Hoggan, M.D., and Rowe, W.P. (1968). Serologic evidence for human infection with adenovirus-associated viruses. J. Natl. Cancer Inst. *40*, 319–327.

Bolze, A., Byun, M., McDonald, D., Morgan, N.V., Abhyankar, A., Premkumar, L., Puel, A., Bacon, C.M., Rieux-Laucat, F., Pang, K., et al. (2010). Whole-exome-sequencing-based discovery of human FADD deficiency. Am. J. Hum. Genet. *87*, 873–881.

Borel, F., Kay, M.A., and Mueller, C. (2014). Recombinant AAV as a platform for translating the therapeutic potential of RNA interference. Mol. Ther. J. Am. Soc. Gene Ther. *22*, 692–701.

Bossis, I., and Chiorini, J.A. (2003). Cloning of an avian adeno-associated virus (AAAV) and generation of recombinant AAAV particles. J. Virol. *77*, 6799–6810.

Boztug, K., Schmidt, M., Schwarzer, A., Banerjee, P.P., Díez, I.A., Dewey, R.A., Böhm, M., Nowrouzi, A., Ball, C.R., Glimm, H., et al. (2010). Stem-cell gene therapy for the Wiskott-Aldrich syndrome. N. Engl. J. Med. *363*, 1918–1927.

Bryant, L.M., Christopher, D.M., Giles, A.R., Hinderer, C., Rodriguez, J.L., Smith, J.B., Traxler, E.A., Tycko, J., Wojno, A.P., and Wilson, J.M. (2013). Lessons learned from the clinical development and market authorization of Glybera. Hum. Gene Ther. Clin. Dev. *24*, 55–64.

Burguete, T., Rabreau, M., Fontanges-Darriet, M., Roset, E., Hager, H.D., Köppel, A., Bischof, P., and Schlehofer, J.R. (1999). Evidence for infection of the human embryo with adeno-associated virus in pregnancy. Hum. Reprod. Oxf. Engl. *14*, 2396–2401.

Calcedo, R., Vandenberghe, L.H., Gao, G., Lin, J., and Wilson, J.M. (2009). Worldwide epidemiology of neutralizing antibodies to adeno-associated viruses. J. Infect. Dis. *199*, 381–390.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Cartier, N., Hacein-Bey-Abina, S., Bartholomae, C.C., Veres, G., Schmidt, M., Kutschera, I., Vidaud, M., Abel, U., Dal-Cortivo, L., Caccavelli, L., et al. (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science *326*, 818–823.

Cataldi, M.P., and McCarty, D.M. (2013). Hairpin-end conformation of adeno-associated virus genome determines interactions with DNA-repair pathways. Gene Ther. *20*, 686–693.

Cattoglio, C., Pellin, D., Rizzi, E., Maruggi, G., Corti, G., Miselli, F., Sartori, D., Guffanti, A., Di Serio, C., Ambrosi, A., et al. (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. Blood *116*, 5507–5517.

Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., Yvon, E., Nusbaum, P., Selz, F., Hue, C., Certain, S., Casanova, J.L., et al. (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. Science *288*, 669–672.

Cervelli, T., Palacios, J.A., Zentilin, L., Mano, M., Schwartz, R.A., Weitzman, M.D., and Giacca, M. (2008). Processing of recombinant AAV genomes occurs in specific nuclear structures that overlap with foci of DNA-damage-response proteins. J. Cell Sci. *121*, 349–357.

Chadeuf, G., Ciron, C., Moullier, P., and Salvetti, A. (2005). Evidence for encapsidation of prokaryotic sequences during recombinant adeno-associated virus production and their in vivo persistence after vector delivery. Mol. Ther. J. Am. Soc. Gene Ther. *12*, 744–753.

Chandler, R.J., LaFave, M.C., Varshney, G.K., Trivedi, N.S., Carrillo-Carrasco, N., Senac, J.S., Wu, W., Hoffmann, V., Elkahloun, A.G., Burgess, S.M., et al. (2015). Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. J. Clin. Invest. *125*, 870–880.

Childers, M.K., Joubert, R., Poulard, K., Moal, C., Grange, R.W., Doering, J.A., Lawlor, M.W., Rider, B.E., Jamet, T., Danièle, N., et al. (2014). Gene therapy prolongs survival and restores function in murine and canine models of myotubular myopathy. Sci. Transl. Med. *6*, 220ra10.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601–2602.

Cogné, B., Snyder, R., Lindenbaum, P., Dupont, J.-B., Redon, R., Moullier, P., and Leger, A. (2014). NGS library preparation may generate artifactual integration sites of AAV vectors. Nat. Med. *20*, 577–578.

Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., Hahn, B.H., Czartoski, J.L., McElrath, M.J., Lehmann, C., et al. (2015). HIV-1 integration landscape during latent and active infection. Cell *160*, 420–432.

Coker, A.L., Russell, R.B., Bond, S.M., Pirisi, L., Liu, Y., Mane, M., Kokorina, N., Gerasimova, T., and Hermonat, P.L. (2001). Adeno-associated virus is associated with a lower risk of high-grade cervical neoplasia. Exp. Mol. Pathol. *70*, 83–89.

Daly, T.M., Ohlemiller, K.K., Roberts, M.S., Vogler, C.A., and Sands, M.S. (2001). Prevention of systemic clinical disease in MPS VII mice following AAV-mediated neonatal gene transfer. Gene Ther. *8*, 1291–1298.

Davidoff, A.M., Ng, C.Y.C., Sleep, S., Gray, J., Azam, S., Zhao, Y., McIntosh, J.H., Karimipoor, M., and Nathwani, A.C. (2004). Purification of recombinant adeno-associated virus type 8 vectors by ion exchange chromatography generates clinical grade vector stock. J. Virol. Methods *121*, 209–215.

Depledge, D.P., Palser, A.L., Watson, S.J., Lai, I.Y.-C., Gray, E.R., Grant, P., Kanda, R.K., Leproust, E., Kellam, P., and Breuer, J. (2011). Specific capture and whole-genome sequencing of viruses from clinical samples. PloS One *6*, e27805.

Donsante, A., Miller, D.G., Li, Y., Vogler, C., Brunt, E.M., Russell, D.W., and Sands, M.S. (2007). AAV vector integration sites in mouse hepatocellular carcinoma. Science *317*, 477.

Duan, D., Li, Q., Kao, A.W., Yue, Y., Pessin, J.E., and Engelhardt, J.F. (1999). Dynamin is required for recombinant adeno-associated virus type 2 infection. J. Virol. *73*, 10371–10376.

Duan, D., Yue, Y., and Engelhardt, J.F. (2003). Consequences of DNA-dependent protein kinase catalytic subunit deficiency on recombinant adeno-associated virus genome circularization and heterodimerization in muscle tissue. J. Virol. *77*, 4751–4759.

Dubose, A.J., Lichtenstein, S.T., Narisu, N., Bonnycastle, L.L., Swift, A.J., Chines, P.S., and Collins, F.S. (2013). Use of microarray hybrid capture and next-generation sequencing to identify the anatomy of a transgene. Nucleic Acids Res. *41*, e70.

Duncavage, E.J., Magrini, V., Becker, N., Armstrong, J.R., Demeter, R.T., Wylie, T., Abel, H.J., and Pfeifer, J.D. (2011). Hybrid capture and next-generation sequencing identify viral integration sites from formalin-fixed, paraffin-embedded tissue. J. Mol. Diagn. JMD *13*, 325–333.

Dupont, J.-B., Tournaire, B., Georger, C., Marolleau, B., Jeanson-Leh, L., Ledevin, M., Lindenbaum, P., Lecomte, E., Cogné, B., Dubreil, L., et al. (2015). Short-lived recombinant adeno-associated virus transgene expression in dystrophic muscle is associated with oxidative damage to transgene mRNA. Mol. Ther. — Methods Clin. Dev. *2*, 15010.

Erles, K., Rohde, V., Thaele, M., Roth, S., Edler, L., and Schlehofer, J.R. (2001). DNA of adeno-associated virus (AAV) in testicular tissue and in abnormal semen samples. Hum. Reprod. Oxf. Engl. *16*, 2333–2337.

Farzad, L., Cerullo, V., Yagyu, S., Bertin, T., Hemminki, A., Rooney, C., Lee, B., and Suzuki, M. (2014). Combinatorial treatment with oncolytic adenovirus and helper-dependent adenovirus augments adenoviral cancer gene therapy. Mol. Ther. — Oncolytics *1*, 14008.

Favaro, P., Downey, H.D., Zhou, J.S., Wright, J.F., Hauck, B., Mingozzi, F., High, K.A., and Arruda, V.R. (2009). Host and vector-dependent effects on the risk of germline transmission of AAV vectors. Mol. Ther. J. Am. Soc. Gene Ther. *17*, 1022–1030.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496–512.

Gansauge, M.-T., and Meyer, M. (2014). Selective enrichment of damaged DNA molecules for ancient genome sequencing. Genome Res. *24*, 1543–1549.

Gao, G.-P., Alvira, M.R., Wang, L., Calcedo, R., Johnston, J., and Wilson, J.M. (2002). Novel adeno-associated viruses from rhesus monkeys as vectors for human gene therapy. Proc. Natl. Acad. Sci. U. S. A. *99*, 11854–11859.

Gauttier, V., Pichard, V., Aubert, D., Kaeppel, C., Schmidt, M., Ferry, N., and Conchon, S. (2013). No tumour-initiating risk associated with scAAV transduction in newborn rat liver. Gene Ther. *20*, 779–784.

Gernoux, G., Guilbaud, M., Dubreil, L., Larcher, T., Babarit, C., Ledevin, M., Jaulin, N., Planel, P., Moullier, P., and Adjali, O. (2015). Early interaction of adeno-associated virus serotype 8 vector with the host immune system following intramuscular delivery results in weak but detectable lymphocyte and dendritic cell transduction. Hum. Gene Ther. *26*, 1–13.

Van Gestel, M.A., van Erp, S., Sanders, L.E., Brans, M. a. D., Luijendijk, M.C.M., Merkestein, M., Pasterkamp, R.J., and Adan, R. a. H. (2014). shRNA-induced saturation of the microRNA pathway in the rat brain. Gene Ther. *21*, 205–211.

Gillet, N.A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., Berry, C., Bushman, F.D., Taylor, G.P., and Bangham, C.R.M. (2011). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood *117*, 3113–3122.

Gire, S.K., Goba, A., Andersen, K.G., Sealfon, R.S.G., Park, D.J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science *345*, 1369–1372.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182–189.

Gonçalves, M.A.F.V. (2005). Adeno-associated virus: from defective virus to effective vector. Virol. J. *2*, 43.

Grez, M., Reichenbach, J., Schwäble, J., Seger, R., Dinauer, M.C., and Thrasher, A.J. (2011). Gene therapy of chronic granulomatous disease: the engraftment dilemma. Mol. Ther. J. Am. Soc. Gene Ther. *19*, 28–35.

Le Guiner, C., Montus, M., Servais, L., Cherel, Y., Francois, V., Thibaud, J.-L., Wary, C., Matot, B., Larcher, T., Guigand, L., et al. (2014). Forelimb treatment in a large cohort of dystrophic dogs supports delivery of a recombinant AAV for exon skipping in Duchenne patients. Mol. Ther. J. Am. Soc. Gene Ther. *22*, 1923–1935.

Hacein-Bey-Abina, S., Pai, S.-Y., Gaspar, H.B., Armant, M., Berry, C.C., Blanche, S., Bleesing, J., Blondeau, J., de Boer, H., Buckland, K.F., et al. (2014). A Modified γ-Retrovirus Vector for X-Linked Severe Combined Immunodeficiency. N. Engl. J. Med. *371*, 1407–1417.

Haffner, M.C., De Marzo, A.M., Meeker, A.K., Nelson, W.G., and Yegnasubramanian, S. (2011). Transcription-induced DNA double strand breaks: both oncogenic force and potential therapeutic target? Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *17*, 3858–3864.

Hansen, K.D., Brenner, S.E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. *38*, e131.

Hayes, M., and Li, J. (2013). Bellerophon: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data. BMC Bioinformatics *14 Suppl 5*, S6.

He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz Jr, L.A., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature *464*, 610–614.

Heise, C., Hermiston, T., Johnson, L., Brooks, G., Sampson-Johannes, A., Williams, A., Hawkins, L., and Kirn, D. (2000). An adenovirus E1A mutant that demonstrates potent and selective systemic anti-tumoral efficacy. Nat. Med. *6*, 1134–1139.

Heller, L.C., and Heller, R. (2010). Electroporation gene therapy preclinical and clinical trials for melanoma. Curr. Gene Ther. *10*, 312–317.

Hermonat, P.L., and Muzyczka, N. (1984). Use of adeno-associated virus as a mammalian DNA cloning vector: transduction of neomycin resistance into mammalian tissue culture cells. Proc. Natl. Acad. Sci. U. S. A. *81*, 6466–6470.

't Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. *31*, 1015–1022.

Hoggan, M.D., Blacklow, N.R., and Rowe, W.P. (1966). Studies of small DNA viruses found in various adenovirus preparations: physical, biological, and immunological characteristics. Proc. Natl. Acad. Sci. U. S. A. *55*, 1467–1474.

Honaramooz, A., Megee, S., Zeng, W., Destrempes, M.M., Overton, S.A., Luo, J., Galantino-Homer, H., Modelski, M., Chen, F., Blash, S., et al. (2008). Adeno-associated virus (AAV)-mediated transduction of male germ line stem cells results in transgene transmission after germ cell transplantation. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *22*, 374–382.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.

Hüser, D., Gogol-Döring, A., Chen, W., and Heilbronn, R. (2014). Adeno-associated virus type 2 wild-type and vector-mediated genomic integration profiles of human diploid fibroblasts analyzed by third-generation PacBio DNA sequencing. J. Virol. *88*, 11253–11263.

Inagaki, K., Ma, C., Storm, T.A., Kay, M.A., and Nakai, H. (2007a). The role of DNA-PKcs and artemis in opening viral DNA hairpin termini in various tissues in mice. J. Virol. *81*, 11304–11321.

Inagaki, K., Lewis, S.M., Wu, X., Ma, C., Munroe, D.J., Fuess, S., Storm, T.A., Kay, M.A., and Nakai, H. (2007b). DNA palindromes with a modest arm length of greater, similar 20 base pairs are a significant target for recombinant adeno-associated virus vector integration in the liver, muscles, and heart in mice. J. Virol. *81*, 11290–11303.

Isidor, B., Lindenbaum, P., Pichon, O., Bézieau, S., Dina, C., Jacquemont, S., Martin-Coignard, D., Thauvin-Robinet, C., Le Merrer, M., Mandel, J.-L., et al. (2011). Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nat. Genet. *43*, 306–308.

Janovitz, T., Klein, I.A., Oliveira, T., Mukherjee, P., Nussenzweig, M.C., Sadelain, M., and Falck-Pedersen, E. (2013). High-throughput sequencing reveals principles of adeno-associated virus serotype 2 integration. J. Virol. *87*, 8559–8568.

Janovitz, T., Oliveira, T., Sadelain, M., and Falck-Pedersen, E. (2014). Highly divergent integration profile of adeno-associated virus serotype 5 revealed by high-throughput sequencing. J. Virol. *88*, 2481–2488.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Johnson, J.S., and Samulski, R.J. (2009). Enhancement of adeno-associated virus infection by mobilizing capsids into and out of the nucleolus. J. Virol. *83*, 2632–2644.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am. J. Hum. Genet. *91*, 839–848.

Jurvansuu, J., Raj, K., Stasiak, A., and Beard, P. (2005). Viral transport of DNA damage that mimics a stalled replication fork. J. Virol. *79*, 569–580.

Kaeppel, C., Beattie, S.G., Fronza, R., van Logtenstein, R., Salmon, F., Schmidt, S., Wolf, S., Nowrouzi, A., Glimm, H., von Kalle, C., et al. (2013). A largely random AAV integration profile after LPLD gene therapy. Nat. Med. *19*, 889–891.

Kaludov, N., Brown, K.E., Walters, R.W., Zabner, J., and Chiorini, J.A. (2001). Adeno-associated virus serotype 4 (AAV4) and AAV5 both require sialic acid binding for hemagglutination and efficient transduction but differ in sialic acid linkage specificity. J. Virol. *75*, 6884–6893.

Kapranov, P., Chen, L., Dederich, D., Dong, B., He, J., Steinmann, K.E., Moore, A.R., Thompson, J.F., Milos, P.M., and Xiao, W. (2012). Native molecular state of adeno-associated viral vectors revealed by single-molecule sequencing. Hum. Gene Ther. *23*, 46–55.

Kashiwakura, Y., Tamayose, K., Iwabuchi, K., Hirai, Y., Shimada, T., Matsumoto, K., Nakamura, T., Watanabe, M., Oshimi, K., and Daida, H. (2005). Hepatocyte growth factor receptor is a coreceptor for adeno-associated virus type 2 infection. J. Virol. *79*, 609–614.

Kawakami, T., Chano, T., Minami, K., Okabe, H., Okada, Y., and Okamoto, K. (2006). Imprinted DLK1 is a putative tumor suppressor gene and inactivated by epimutation at the region upstream of GTL2 in human renal cell carcinoma. Hum. Mol. Genet. *15*, 821–830.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. *40*, e3.

Kotin, R.M., Siniscalco, M., Samulski, R.J., Zhu, X.D., Hunter, L., Laughlin, C.A., McLaughlin, S., Muzyczka, N., Rocchi, M., and Berns, K.I. (1990). Site-specific integration by adeno-associated virus. Proc. Natl. Acad. Sci. U. S. A. *87*, 2211–2215.

Kozarewa, I., and Turner, D.J. (2011). Amplification-free library preparation for paired-end Illumina sequencing. Methods Mol. Biol. Clifton NJ *733*, 257–266.

Krueger, F., Andrews, S.R., and Osborne, C.S. (2011). Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. PLoS ONE *6*.

LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. *42*, 4257–4269.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Laurence, M., Hatzis, C., and Brash, D.E. (2014). Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PloS One *9*, e97876.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., Haurigot, V., Doyon, Y., Li, T., Wong, S.Y., Bhagwat, A.S., Malani, N., Anguela, X.M., Sharma, R., Ivanciu, L., et al. (2011a). In vivo genome editing restores haemostasis in a mouse model of haemophilia. Nature *475*, 217–221.

Li, H., Malani, N., Hamilton, S.R., Schlachterman, A., Bussadori, G., Edmonson, S.E., Shah, R., Arruda, V.R., Mingozzi, F., Wright, J.F., et al. (2011b). Assessing the potential for AAV vector genotoxicity in a murine model. Blood *117*, 3311–3319.

Li, J.-W., Wan, R., Yu, C.-S., Co, N.N., Wong, N., and Chan, T.-F. (2013). ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics *29*, 649–651.

Li, Y., Liu, X., Yang, Z., Xu, C., Liu, D., Qin, J., Dai, M., Hao, J., Feng, M., Huang, X., et al. (2014). The MYC, TERT, and ZIC1 genes are common targets of viral integration and transcriptional deregulation in avian leukosis virus subgroup J-induced myeloid leukosis. J. Virol. *88*, 3182–3191.

Li, Z., Düllmann, J., Schiedlmeier, B., Schmidt, M., von Kalle, C., Meyer, J., Forster, M., Stocking, C., Wahlers, A., Frank, O., et al. (2002). Murine leukemia induced by retroviral gene marking. Science *296*, 497.

Lim, J.-A., Li, L., and Raben, N. (2014). Pompe disease: from pathophysiology to therapy and back again. Front. Aging Neurosci. *6*, 177.

Limberis, M.P., Adam, V.S., Wong, G., Gren, J., Kobasa, D., Ross, T.M., Kobinger, G.P., Tretiakova, A., and Wilson, J.M. (2013). Intranasal antibody gene transfer in mice and ferrets elicits broad protection against pandemic influenza. Sci. Transl. Med. *5*, 187ra72.

Lisowski, L., Dane, A.P., Chu, K., Zhang, Y., Cunningham, S.C., Wilson, E.M., Nygaard, S., Grompe, M., Alexander, I.E., and Kay, M.A. (2014). Selection and evaluation of clinically relevant AAV variants in a xenograft liver model. Nature *506*, 382–386.

Longo, M.S., O'Neill, M.J., and O'Neill, R.J. (2011). Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. PLoS ONE *6*, e16410.

Louis, N., Evelegh, C., and Graham, F.L. (1997). Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. Virology *233*, 423–429.

Lovric, J., Mano, M., Zentilin, L., Eulalio, A., Zacchigna, S., and Giacca, M. (2012). Terminal differentiation of cardiac and skeletal myocytes induces permissivity to AAV transduction by relieving inhibition imposed by DNA damage response proteins. Mol. Ther. J. Am. Soc. Gene Ther. *20*, 2087–2097.

Luckey, J.A., Drossman, H., Kostichka, A.J., Mead, D.A., D'Cunha, J., Norris, T.B., and Smith, L.M. (1990). High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res. *18*, 4417–4421.

MacLaren, R.E., Groppe, M., Barnard, A.R., Cottriall, C.L., Tolmachova, T., Seymour, L., Clark, K.R., During, M.J., Cremers, F.P.M., Black, G.C.M., et al. (2014). Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. Lancet *383*, 1129–1137.

Maguire, A.M., Simonelli, F., Pierce, E.A., Pugh, E.N., Mingozzi, F., Bennicelli, J., Banfi, S., Marshall, K.A., Testa, F., Surace, E.M., et al. (2008). Safety and efficacy of gene transfer for Leber's congenital amaurosis. N. Engl. J. Med. *358*, 2240–2248.

Manno, C.S., Pierce, G.F., Arruda, V.R., Glader, B., Ragni, M., Rasko, J.J., Rasko, J., Ozelo, M.C., Hoots, K., Blatt, P., et al. (2006). Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. Nat. Med. *12*, 342–347.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Mehier-Humbert, S., and Guy, R.H. (2005). Physical methods for gene transfer: improving the kinetics of gene delivery into cells. Adv. Drug Deliv. Rev. *57*, 733–753.

Meldrum, D.R., Evensen, H.T., Pence, W.H., Moody, S.E., Cunningham, D.L., and Wiktor, P.J. (2000). ACAPELLA-1K, a capillary-based submicroliter automated fluid handling system for genome analysis. Genome Res. *10*, 95–104.

Mendell, J.R., Campbell, K., Rodino-Klapac, L., Sahenk, Z., Shilling, C., Lewis, S., Bowles, D., Gray, S., Li, C., Galloway, G., et al. (2010). Dystrophin immunity in Duchenne's muscular dystrophy. N. Engl. J. Med. *363*, 1429–1437.

Meng, L., Ward, A.J., Chun, S., Bennett, C.F., Beaudet, A.L., and Rigo, F. (2015). Towards a therapy for Angelman syndrome by targeting a long non-coding RNA. Nature *518*, 409–412.

Mercier, S., Küry, S., Shaboodien, G., Houniet, D.T., Khumalo, N.P., Bou-Hanna, C., Bodak, N., Cormier-Daire, V., David, A., Faivre, L., et al. (2013). Mutations in FAM111B cause hereditary fibrosing poikiloderma with tendon contracture, myopathy, and pulmonary fibrosis. Am. J. Hum. Genet. *93*, 1100–1107.

Miller, D.G., Petek, L.M., and Russell, D.W. (2004). Adeno-associated virus vectors integrate at chromosome breakage sites. Nat. Genet. *36*, 767–773.

Miller, D.G., Trobridge, G.D., Petek, L.M., Jacobs, M.A., Kaul, R., and Russell, D.W. (2005). Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. J. Virol. *79*, 11434–11442.

Montini, E., Cesana, D., Schmidt, M., Sanvito, F., Ponzoni, M., Bartholomae, C., Sergi Sergi, L., Benedicenti, F., Ambrosi, A., Di Serio, C., et al. (2006). Hematopoietic stem cell gene transfer in a tumor-prone mouse model uncovers low genotoxicity of lentiviral vector integration. Nat. Biotechnol. *24*, 687–696.

Nakai, H., Montini, E., Fuess, S., Storm, T.A., Grompe, M., and Kay, M.A. (2003). AAV serotype 2 vectors preferentially integrate into active genes in mice. Nat. Genet. *34*, 297–302.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. *39*, e90.

Nathwani, A.C., Tuddenham, E.G.D., Rangarajan, S., Rosales, C., McIntosh, J., Linch, D.C., Chowdary, P., Riddell, A., Pie, A.J., Harrington, C., et al. (2011). Adenovirus-associated virus vector-mediated gene transfer in hemophilia B. N. Engl. J. Med. *365*, 2357–2365.

Nathwani, A.C., Reiss, U.M., Tuddenham, E.G.D., Rosales, C., Chowdary, P., McIntosh, J., Della Peruta, M., Lheriteau, E., Patel, N., Raj, D., et al. (2014). Long-term safety and efficacy of factor IX gene therapy in hemophilia B. N. Engl. J. Med. *371*, 1994–2004.

Nicolson, S.C., and Samulski, R.J. (2014). Recombinant adeno-associated virus utilizes host cell nuclear import machinery to enter the nucleus. J. Virol. *88*, 4132–4144.

Nonnenmacher, M., and Weber, T. (2011). Adeno-associated virus 2 infection requires endocytosis through the CLIC/GEEC pathway. Cell Host Microbe *10*, 563–576.

Nowrouzi, A., Penaud-Budloo, M., Kaeppel, C., Appelt, U., Le Guiner, C., Moullier, P., von Kalle, C., Snyder, R.O., and Schmidt, M. (2012). Integration frequency and intermolecular recombination of rAAV vectors in non-human primate skeletal muscle and liver. Mol. Ther. J. Am. Soc. Gene Ther. *20*, 1177–1186.

Opie, S.R., Warrington, K.H., Agbandje-McKenna, M., Zolotukhin, S., and Muzyczka, N. (2003). Identification of amino acid residues in the capsid proteins of adeno-associated virus type 2 that contribute to heparan sulfate proteoglycan binding. J. Virol. *77*, 6995–7006.

Di Pasquale, G., Davidson, B.L., Stein, C.S., Martins, I., Scudiero, D., Monks, A., and Chiorini, J.A. (2003). Identification of PDGFR as a receptor for AAV-5 transduction. Nat. Med. *9*, 1306–1312.

Penaud-Budloo, M., Le Guiner, C., Nowrouzi, A., Toromanoff, A., Chérel, Y., Chenuaud, P., Schmidt, M., von Kalle, C., Rolling, F., Moullier, P., et al. (2008). Adeno-associated virus vector genomes persist as episomal chromatin in primate muscle. J. Virol. *82*, 7875–7885.

Poptsova, M.S., Il'icheva, I.A., Nechipurenko, D.Y., Panchenko, L.A., Khodikov, M.V., Oparina, N.Y., Polozov, R.V., Nechipurenko, Y.D., and Grokhovsky, S.L. (2014). Non-random DNA fragmentation in next-generation sequencing. Sci. Rep. *4*, 4532.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science *238*, 336–341.

Qing, K., Mah, C., Hansen, J., Zhou, S., Dwarki, V., and Srivastava, A. (1999). Human fibroblast growth factor receptor 1 is a co-receptor for infection by adeno-associated virus 2. Nat. Med. *5*, 71–77.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat. Methods *5*, 1005–1010.

Raper, S.E., Chirmule, N., Lee, F.S., Wivel, N.A., Bagg, A., Gao, G., Wilson, J.M., and Batshaw, M.L. (2003). Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. Mol. Genet. Metab. *80*, 148–158.

Reddy, T.B.K., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., and Kyrpides, N.C. (2014). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. gku950.

Rosas, L.E., Grieves, J.L., Zaraspe, K., La Perle, K.M., Fu, H., and McCarty, D.M. (2012). Patterns of scAAV vector insertion associated with oncogenic events in a mouse model for genotoxicity. Mol. Ther. J. Am. Soc. Gene Ther. *20*, 2098–2110.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.

Rutledge, E.A., Halbert, C.L., and Russell, D.W. (1998). Infectious clones and vectors derived from adeno-associated virus (AAV) serotypes other than AAV type 2. J. Virol. *72*, 309–319.

Samulski, R.J., Srivastava, A., Berns, K.I., and Muzyczka, N. (1983). Rescue of adeno-associated virus from recombinant plasmids: gene correction within the terminal repeats of AAV. Cell *33*, 135–143.

Samulski, R.J., Zhu, X., Xiao, X., Brook, J.D., Housman, D.E., Epstein, N., and Hunter, L.A. (1991). Targeted integration of adeno-associated virus (AAV) into human chromosome 19. EMBO J. *10*, 3941–3950.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977a). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977b). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687–695.

Sanlioglu, S., Benson, P., and Engelhardt, J.F. (2000). Loss of ATM function enhances recombinant adeno-associated virus transduction and integration through pathways similar to UV irradiation. Virology *268*, 68–78.

Satkunanathan, S., Wheeler, J., Thorpe, R., and Zhao, Y. (2014). Establishment of a novel cell line for the enhanced production of recombinant adeno-associated virus vectors for gene therapy. Hum. Gene Ther. *25*, 929–941.

Schmidt, M., Katano, H., Bossis, I., and Chiorini, J.A. (2004). Cloning and characterization of a bovine adeno-associated virus. J. Virol. *78*, 6509–6516.

Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat. Methods *4*, 1051–1057.

Schmidt, W.M., Uddin, M.H., Dysek, S., Moser-Thier, K., Pirker, C., Höger, H., Ambros, I.M., Ambros, P.F., Berger, W., and Bittner, R.E. (2011). DNA damage, somatic aneuploidy, and malignant sarcoma susceptibility in muscular dystrophies. PLoS Genet. *7*, e1002042.

Schnepp, B.C., Jensen, R.L., Chen, C.-L., Johnson, P.R., and Clark, K.R. (2005). Characterization of adeno-associated virus genomes isolated from human tissues. J. Virol. *79*, 14793–14803.

Schuettrumpf, J., Liu, J.-H., Couto, L.B., Addya, K., Leonard, D.G.B., Zhen, Z., Sommer, J., Summer, J., and Arruda, V.R. (2006). Inadvertent germline transmission of AAV2 vector: findings in a rabbit model correlate with those in a human clinical trial. Mol. Ther. J. Am. Soc. Gene Ther. *13*, 1064–1073.

Schwartz, R.A., Palacios, J.A., Cassell, G.D., Adam, S., Giacca, M., and Weitzman, M.D. (2007). The Mre11/Rad50/Nbs1 complex limits adeno-associated virus transduction and replication. J. Virol. *81*, 12936–12945.

Shan, X., Tashiro, H., and Lin, C.G. (2003). The identification and characterization of oxidized RNAs in Alzheimer's disease. J. Neurosci. Off. J. Soc. Neurosci. *23*, 4913–4921.

Shen, S., Bryant, K.D., Brown, S.M., Randell, S.H., and Asokan, A. (2011). Terminal N-linked galactose is the primary receptor for adeno-associated virus 9. J. Biol. Chem. *286*, 13532–13540.

Shin, J.-H., Pan, X., Hakim, C.H., Yang, H.T., Yue, Y., Zhang, K., Terjung, R.L., and Duan, D. (2013). Microdystrophin ameliorates muscular dystrophy in the canine model of duchenne muscular dystrophy. Mol. Ther. J. Am. Soc. Gene Ther. *21*, 750–757.

Smith, R.H., Levy, J.R., and Kotin, R.M. (2009). A simplified baculovirus-AAV expression vector system coupled with one-step affinity purification yields high-titer rAAV stocks from insect cells. Mol. Ther. J. Am. Soc. Gene Ther. *17*, 1888–1896.

Snyder, R.O., Im, D.S., Ni, T., Xiao, X., Samulski, R.J., and Muzyczka, N. (1993). Features of the adeno-associated virus origin involved in substrate recognition by the viral Rep protein. J. Virol. *67*, 6096–6104.

Song, S., Laipis, P.J., Berns, K.I., and Flotte, T.R. (2001). Effect of DNA-dependent protein kinase on the molecular fate of the rAAV2 genome in skeletal muscle. Proc. Natl. Acad. Sci. U. S. A. *98*, 4084–4088.

Sonntag, F., Bleker, S., Leuchs, B., Fischer, R., and Kleinschmidt, J.A. (2006). Adeno-associated virus type 2 capsids with externalized VP1/VP2 trafficking domains are generated prior to passage through the cytoplasm and are maintained until uncoating occurs in the nucleus. J. Virol. *80*, 11040–11054.

Sonntag, F., Schmidt, K., and Kleinschmidt, J.A. (2010). A viral assembly factor promotes AAV2 capsid formation in the nucleolus. Proc. Natl. Acad. Sci. U. S. A. *107*, 10220–10225.

Strong, M.J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., Fewell, C., Taylor, C.M., and Flemington, E.K. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. *10*, e1004437.

Summerford, C., and Samulski, R.J. (1998). Membrane-associated heparan sulfate proteoglycan is a receptor for adeno-associated virus type 2 virions. J. Virol. *72*, 1438–1445.

Summerford, C., Bartlett, J.S., and Samulski, R.J. (1999). AlphaVbeta5 integrin: a co-receptor for adeno-associated virus type 2 infection. Nat. Med. *5*, 78–82.

Tatum, E.L. (1966). Molecular Biology, Nucleic Acids, and the Future of Medicine. Perspect. Biol. Med. *10*, 19–32.

Tebas, P., Stein, D., Tang, W.W., Frank, I., Wang, S.Q., Lee, G., Spratt, S.K., Surosky, R.T., Giedlin, M.A., Nichol, G., et al. (2014). Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. N. Engl. J. Med. *370*, 901–910.

Thwaite, R., Pagès, G., Chillón, M., and Bosch, A. (2015). AAVrh.10 immunogenicity in mice and humans. Relevance of antibody cross-reactivity in human gene therapy. Gene Ther. *22*, 196–201.

Tratschin, J.D., West, M.H., Sandbank, T., and Carter, B.J. (1984). A human parvovirus, adeno-associated virus, as a eucaryotic vector: transient expression and encapsidation of the procaryotic gene for chloramphenicol acetyltransferase. Mol. Cell. Biol. *4*, 2072–2081.

Urabe, M., Ding, C., and Kotin, R.M. (2002). Insect cells as a factory to produce adeno-associated virus type 2 vectors. Hum. Gene Ther. *13*, 1935–1943.

Ustek, D., Sirma, S., Gumus, E., Arikan, M., Cakiris, A., Abaci, N., Mathew, J., Emrence, Z., Azakli, H., Cosan, F., et al. (2012). A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *12*, 1349–1354.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Voit, T., Topaloglu, H., Straub, V., Muntoni, F., Deconinck, N., Campion, G., De Kimpe, S.J., Eagle, M., Guglieri, M., Hood, S., et al. (2014). Safety and efficacy of drisapersen for the treatment of Duchenne muscular dystrophy (DEMAND II): an exploratory, randomised, placebo-controlled phase 2 study. Lancet Neurol. *13*, 987–996.

Walz, C., Deprez, A., Dupressoir, T., Dürst, M., Rabreau, M., and Schlehofer, J.R. (1997). Interaction of human papillomavirus type 16 and adeno-associated virus type 2 co-infecting human cervical epithelium. J. Gen. Virol. *78 (Pt 6)*, 1441–1452.

Wang, P.-R., Xu, M., Toffanin, S., Li, Y., Llovet, J.M., and Russell, D.W. (2012). Induction of hepatocellular carcinoma by in vivo gene targeting. Proc. Natl. Acad. Sci. U. S. A. *109*, 11264–11269.

Wang, Q., Jia, P., and Zhao, Z. (2013). VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. PLoS ONE *8*, e64465.

Weinberg, M.S., Nicolson, S., Bhatt, A.P., McLendon, M., Li, C., and Samulski, R.J. (2014). Recombinant adeno-associated virus utilizes cell-specific infectious entry mechanisms. J. Virol. *88*, 12472–12484.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Wistuba, A., Kern, A., Weger, S., Grimm, D., and Kleinschmidt, J.A. (1997). Subcellular compartmentalization of adeno-associated virus type 2 assembly. J. Virol. *71*, 1341–1352.

Wolff, J.A., Malone, R.W., Williams, P., Chong, W., Acsadi, G., Jani, A., and Felgner, P.L. (1990). Direct gene transfer into mouse muscle in vivo. Science *247*, 1465–1468.

Wu, Z., Miller, E., Agbandje-McKenna, M., and Samulski, R.J. (2006). Alpha2,3 and alpha2,6 N-linked sialic acids facilitate efficient binding and transduction by adeno-associated virus types 1 and 6. J. Virol. *80*, 9093–9103.

Xiao, P.-J., and Samulski, R.J. (2012). Cytoplasmic trafficking, endosomal escape, and perinuclear accumulation of adeno-associated virus type 2 particles are facilitated by microtubule network. J. Virol. *86*, 10462–10473.

Xiao, P.-J., Li, C., Neumann, A., and Samulski, R.J. (2012). Quantitative 3D tracing of gene-delivery viral vectors in human cells and animal tissues. Mol. Ther. J. Am. Soc. Gene Ther. *20*, 317–328.

Xiao, X., Li, J., and Samulski, R.J. (1998). Production of high-titer recombinant adeno-associated virus vectors in the absence of helper adenovirus. J. Virol. *72*, 2224–2232.

Ye, G.-J., Scotti, M.M., Liu, J., Wang, L., Knop, D.R., and Veres, G. (2011). Clearance and characterization of residual HSV DNA in recombinant adeno-associated virus produced by an HSV complementation system. Gene Ther. *18*, 135–144.

Yin, H., Kanasty, R.L., Eltoukhy, A.A., Vegas, A.J., Dorkin, J.R., and Anderson, D.G. (2014). Non-viral vectors for gene-based therapy. Nat. Rev. Genet. *15*, 541–555.

Yu, H., Mehta, A., Wang, G., Hauswirth, W.W., Chiodo, V., Boye, S.L., and Guy, J. (2013). Next-generation sequencing of mitochondrial targeted AAV transfer of human ND4 in mice. Mol. Vis. *19*, 1482–1491.

Zangi, L., Lui, K.O., von Gise, A., Ma, Q., Ebina, W., Ptaszek, L.M., Später, D., Xu, H., Tabebordbar, M., Gorbatov, R., et al. (2013). Modified mRNA directs the fate of heart progenitor cells and induces vascular regeneration after myocardial infarction. Nat. Biotechnol. *31*, 898–907.

Zeh, H.J., Downs-Canner, S., McCart, J.A., Guo, Z.S., Rao, U.N.M., Ramalingam, L., Thorne, S.H., Jones, H.L., Kalinski, P., Wieckowski, E., et al. (2015). First-in-man study of western reserve strain oncolytic vaccinia virus: safety, systemic spread, and antitumor activity. Mol. Ther. J. Am. Soc. Gene Ther. *23*, 202–214.

Zeng, W., Tang, L., Bondareva, A., Honaramooz, A., Tanco, V., Dores, C., Megee, S., Modelski, M., Rodriguez-Sosa, J.R., Paczkowski, M., et al. (2013). Viral transduction of male germline stem cells results in transgene transmission after germ cell transplantation in pigs. Biol. Reprod. *88*, 27.

Zentilin, L., Marcello, A., and Giacca, M. (2001). Involvement of cellular double-stranded DNA break binding proteins in processing of the recombinant adeno-associated virus genome. J. Virol. *75*, 12279–12287.

Zhong, L., Malani, N., Li, M., Brady, T., Xie, J., Bell, P., Li, S., Jones, H., Wilson, J.M., Flotte, T.R., et al. (2013). Recombinant adeno-associated virus integration sites in murine liver after ornithine transcarbamylase gene correction. Hum. Gene Ther. *24*, 520–525.

Zhu, J., Huang, X., and Yang, Y. (2007). Innate immune response to adenoviral vectors is mediated by both Toll-like receptor-dependent and -independent pathways. J. Virol. *81*, 3170–3180.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

CORRESPONDENCE

NGS library preparation may generate artifactual integration sites of AAV vectors

To the Editor:

The treatment of monogenic diseases using recombinant adeno-associated virus (rAAV) vectors has recently gained maturity with the approval of Glybera (AAV1-LPL^{S447X}) by the European Medicines Agency[1] for the treatment of lipoprotein lipase (LPL) deficiency. This milestone in gene therapy paves the way for future development of other rAAV products against a wide range of human diseases. rAAV vectors are generally considered safe and efficient, but some studies have raised concerns about their potential genotoxicity and highlighted the need for further investigation regarding their possible role in cancer development[2,3]. In this context, Kaeppel *et al.*[4] recently published an analysis of rAAV integration sites (ISs) in human DNA following Glybera administration. They first applied linear amplification–mediated (LAM) PCR on human biopsies and mouse samples following intramuscular and intravenous injections of AAV1-LPL^{S447X}. As previously described in preclinical samples[5], they found random integration into genomic DNA[5], but they also identified probable ISs in mitochondrial DNA (mtDNA). In the second part of the article, the authors developed a new strategy to specifically assess the ability of rAAV to integrate into mtDNA. The principle of this method substantially differed from that of LAM-PCR and was based on mitochondria enrichment from fresh tissue followed by next-generation sequencing (NGS). We have performed a closely related protocol and have gathered strong evidence that this approach generates artifactual junctions between rAAV and cellular genomes that are indistinguishable from true insertion events (IEs). We think that the lack of appropriate controls led Kaeppel *et al.*[4] to misinterpret the data from NGS regarding integration of the rAAV genome into mitochondrial DNA and to formulate erroneous conclusions with major scientific and clinical implications.

Several characteristics of rAAVs (for example, genome recombination, strong secondary structures and episomal persistence) are major hurdles for IS analysis by PCR-based methods because they can result in partial and eventually skewed retrieval of rAAV IEs. Therefore, we developed a new approach avoiding the use of vector-specific primers and limiting the PCR amplification steps. As direct sequencing of DNA extracted from rAAV-injected tissues would not provide enough vector coverage, we took advantage of targeted sequencing to enrich rAAV genomes in our samples[6]. Briefly, we gave four mice rAAV 2/8 RSVp-eGFP (**Fig. 1a**) by intramuscular injection at a dose of 2.4×10^{13} vector genomes per kilogram (within the effective range for preclinical studies in small and large animal models for muscular dystrophies but one log higher than what was used for Glybera in humans). We recovered muscles 15 d after injection and extracted the DNA. We also prepared a negative control that did not contain rAAV–host DNA junctions. This control consisted of noninjected mouse DNA spiked with *in vitro*–synthesized rAAV circular forms (**Supplementary Methods**). Sample and control DNAs were first sheared and prepared for NGS by ligating barcoded adaptors. Then, we used several overlapping biotinylated RNAs complementary to the whole rAAV genome to capture fragments containing vector sequences. Following PCR amplification, we sequenced these fragments on an Illumina MiSeq platform.

NGS data analyses confirmed the high efficiency of the targeted sequencing with approximately 50% of total sequences from the injected mouse samples mapping on the rAAV genome (**Table 1**). Although these reads came predominantly from rAAV episomes, we could still detect numerous rAAV–host DNA junctions distributed in the mouse nuclear and mitochondrial genomes (**Supplementary Table 1**). We estimated the overall integration frequency to be ~1×10^{-3} IEs per rAAV and also noticed that junction breakpoints occurred along the entire rAAV genome (**Fig. 1b** and **Supplementary Fig. 1**). These findings closely resemble those described by Kaeppel *et al.*[4] following mitochondria enrichment. However, we obtained a similar IS frequency distribution over chromosomes and scattered rAAV breakpoints in our negative control (**Fig. 1c** and **Supplementary Fig. 1**), indicating that chimeric junctions were created during the processing of the DNA before sequencing.

Linkage of unrelated DNA fragments during library preparation is a phenomenon known to occur at a relatively high frequency (1–10% of reads)[7]. In our protocol, chimeric junctions between rAAV and host DNA were probably generated during the adaptor ligation step, although we cannot exclude template switching during PCR[8]. Among these chimeras, CREST and VirusFinder algorithms—with a principle based on multiple retrieval of the same IE—failed to detect ISs in our data set[9,10]. Further analysis confirmed that rAAV–host DNA junctions were unique after removal of PCR duplicates (**Supplementary Methods**).

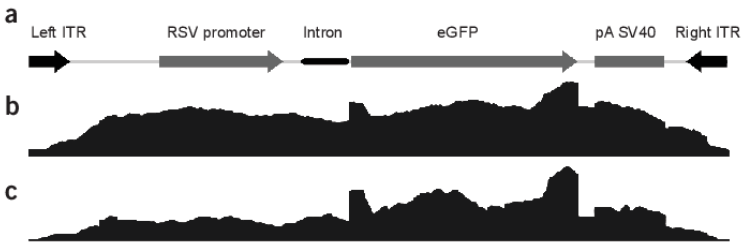


Figure 1 Distribution of junction breakpoints along rAAV genomes. (**a**) Annotated map of rAAV2/8 RSVp eGFP SV40-pA genome. (**b**,**c**) The coverage of reads mapped along rAAV having a mate pair in the mouse genome (GRCm38/mm10) for the rAAV-injected mouse muscles (**b**) and the two negative control replicates (**c**) ITR, inverted terminal repeat.

Table 1 Summary of alignment results for controls and mouse samples

Samples	Alignment results				Junctions retrieved	
	Total reads	rAAV	Mouse genome	Unmapped	rAAV–mouse genome junctions	Junctions per rAAV
Control 1	1,150,854	16.22%	81.17%	2.54%	722	3.88×10^{-3}
Control 2	819,949	8.48%	88.73%	2.75%	373	5.44×10^{-3}
Mouse 1	907,195	57.98%	39.16%	2.69%	1,583	2.96×10^{-3}
Mouse 2	1,307,099	47.26%	50.06%	2.57%	1,525	2.44×10^{-3}
Mouse 3	1,476,965	37.08%	60.60%	2.20%	1,781	3.22×10^{-3}
Mouse 4	1,045,559	58.48%	38.98%	2.37%	1,763	2.83×10^{-3}

NGS data were aligned using a custom bioinformatics pipeline (FindMyVirus) against both the rAAV genome (reference sequence in **Supplementary Methods**) and the latest mouse reference genome (Genome Reference Consortium GRCm38, UCSC version mm10). Read pairs spanning junctions and chimeric reads were selectively extracted. Ultimately, the frequency of junctions per AAV was determined. Data were obtained from a single MiSeq sequencing run. Controls 1 and 2 are independent technical replicates. Mice 1–4 identify muscle samples from four biological replicates.

Targeted sequencing has been applied successfully with other viruses and vectors[11,12] in cell lines and cancer cells. In such clonal samples, each viral IS can be recovered multiple times and subsequently identified in NGS data by filtering out single-occurrence artifacts. Regarding rAAV vectors, ISs are unique in terminally differentiated tissues, explaining the impossibility of distinguishing them from artifacts in our data. Hence, targeted sequencing turns out to be inappropriate for *in vivo* rAAV integration studies. A possible technical solution to bypass this limitation would be to amplify all potential junctions before starting NGS library preparation either by vector-specific PCR or more comprehensively by random whole-genome amplification (**Supplementary Fig. 2**).

Compared with the protocol developed by Kaeppel *et al.*[4], our protocol did not deplete nuclear DNA by mitochondria enrichment but rather by rAAV genome capture before sequencing (**Supplementary Fig. 3**). Thus, our strategy did not favor mtDNA over genomic DNA but instead led to a sharp enrichment in fragments containing rAAV sequences. In both cases, no DNA amplification was initially performed and NGS libraries were prepared similarly. Given the protocol published by Kaeppel *et al.*[4], we feel that there is a strong likelihood that they generated the same artifacts during their NGS library preparation. Because they did not include a negative control in their analysis, they may have misinterpreted the presence of rAAV–mitochondrial DNA junctions. It is noteworthy that another group[13] recently failed to retrieve rAAV ISs in mtDNA with an approach similar to the one developed by Kaeppel *et al.*[4].

Our results do not call into question the LAM-PCR findings published by Kaeppel *et al.*[4] suggesting that rAAV may integrate in mtDNA. However, we think that the conclusions based on mitochondria enrichment experiments deserve further investigation and should be reconsidered unless additional data supporting these claims are available. Finally, new methods should be developed to identify rAAV ISs more extensively and reliably, and appropriate controls must be included to avoid potentially erroneous conclusions that could have a major impact on the field of gene therapy.

Raw Fastq data and BAM files obtained by FindMyVirus are available from the Dryad Digital Repository: http://doi.org/10.5061/dryad.81dg9.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nm.3578).

ACKNOWLEDGMENTS

This study has been supported by funds from the Association Française contre les Myopathies and the French Regional Council of Pays-de-la-Loire. We thank the genomics and bioinformatics core facilities of Nantes (Biogenouest) for their expert services. We are also grateful to M. Haskins for his critical reading.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nm.3578).

Benjamin Cogné[1–3], Richard Snyder[1,4,5], Pierre Lindenbaum[2,3,6,7], Jean-Baptiste Dupont[1–3], Richard Redon[2,3,6,7], Philippe Moullier[1–4] & Adrien Leger[1–3]

[1]INSERM, UMR 1089, Nantes, France. [2]University of Nantes, Nantes, France. [3]Nantes University Hospital, Nantes, France. [4]Department of Molecular Genetics and Microbiology, College of Medicine, University of Florida, Gainesville, Florida, USA. [5]Center of Excellence for Regenerative Health Biotechnology, University of Florida, Alachua, Florida, USA. [6]INSERM, UMR 1087, L'Institut du Thorax, Nantes, France. [7]CNRS, UMR 6291, Nantes, France.
e-mail: adrien.leger@inserm.fr

1. Ylä-Herttuala, S. *Mol. Ther.* **20**, 1831–1832 (2012).
2. Donsante, A. *et al. Science* **317**, 477 (2007).
3. Rosas, L.E. *et al. Mol. Ther.* **20**, 2098–2110 (2012).
4. Kaeppel, C. *et al. Nat. Med.* **19**, 889–891 (2013).
5. Nowrouzi, A. *et al. Mol. Ther.* **20**, 1177–1186 (2012).
6. Gnirke, A. *et al. Nat. Biotechnol.* **27**, 182–189 (2009).
7. Quail, M.A. *et al. Nat. Methods* **5**, 1005–1010 (2008).
8. Kanagawa, T. *J. Biosci. Bioeng.* **96**, 317–323 (2003).
9. Wang, Q., Jia, P. & Zhao, Z. *PLoS ONE* **8**, e64465 (2013).
10. Wang, J. *et al. Nat. Methods* **8**, 652–654 (2011).
11. Ustek, D. *et al. Infect. Genet. Evol.* **12**, 1349–1354 (2012).
12. Duncavage, E.J. *et al. J. Mol. Diagn.* **13**, 325–333 (2011).
13. Yu, H. *et al. Mol. Vis.* **19**, 1482–1491 (2013).

Kaeppel *et al.* reply:

We acknowledge the interest of Cogné *et al.*[1] in our recent publication on AAV1-LPL^{S447X} persistence and integration in samples from lipoprotein lipase (LPL)-deficient patients and wild-type mice[2]. Cogné *et al.*[1] focus their correspondence on the direct use of next-generation sequencing (NGS) technologies and associated library preparations that may lead to artificial false-positive integrations, questioning the rare integration events into mitochondrial DNA detected in our study and promoting their own NGS-based strategies for adeno-associated virus (AAV) integration analyses.

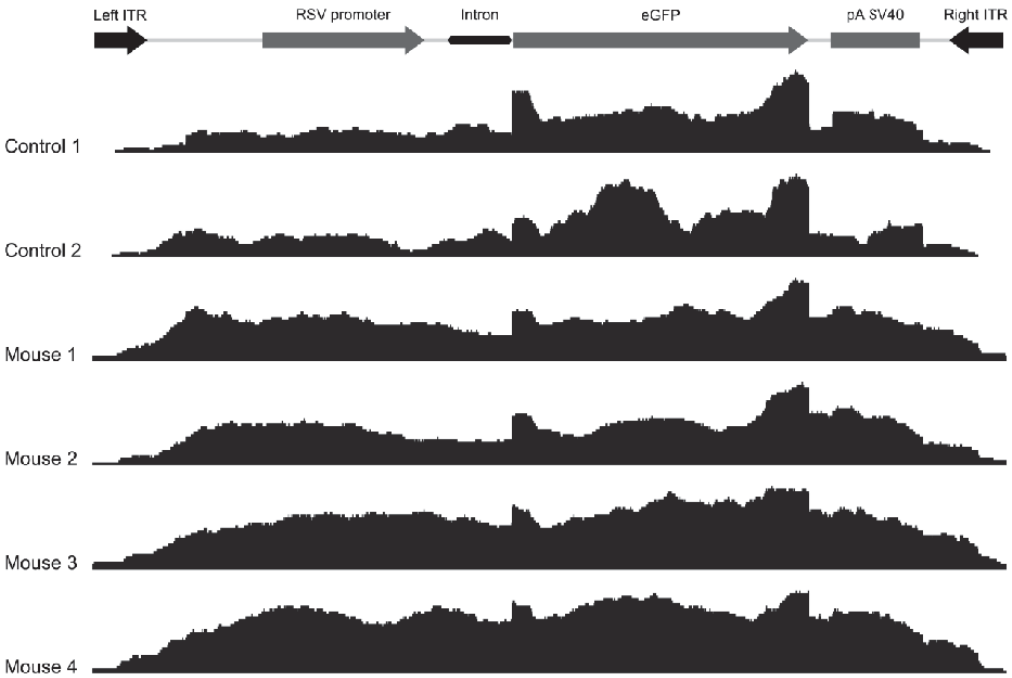
For readers who have not read our manuscript in detail, the impression may arise that our findings were based solely on the analysis of one mouse DNA sample that has been enriched for mitochondrial DNA. However, our findings and conclusions are based on the highly sensitive linear amplification–mediated PCR (LAM-PCR) followed by NGS performed on 25 patient muscle biopsies and 28 mouse samples. In our protocol, we make use of an additional PCR step to directly add the NGS-specific barcode and adaptor sequences before sequencing. We accompanied these studies by extensive positive (AAV1-LPL^{S447X} plasmid and viral vector controls within a complex DNA background in

Supplementary data

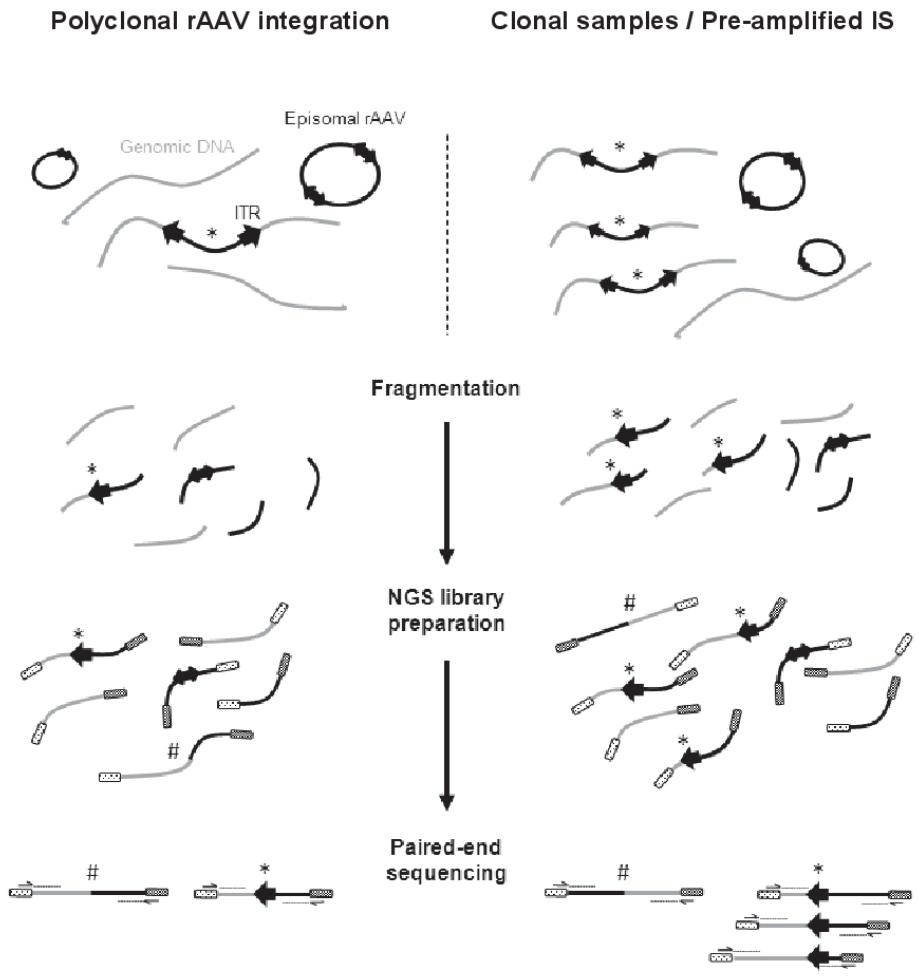
	Controls (%)		Mouse samples (%)			
Chr 1	5.08	4.94	6.43	5.74	6.22	5.32
Chr 2	9.75	9.87	9.39	9.98	11.02	10.23
Chr 3	4.81	3.9	4.66	4.24	3.73	4.17
Chr 4	4.12	5.97	4.66	4.24	5.65	4.06
Chr 5	5.08	4.68	5.42	5.87	5.88	5.49
Chr 6	7.55	6.75	6.24	5.09	4.47	5.2
Chr 7	6.18	6.75	9.39	7.24	8.31	7.2
Chr 8	3.98	3.38	3.78	3.91	3.05	5.09
Chr 9	6.87	6.23	6.36	6.26	7.74	7.2
Chr 10	4.12	4.16	4.41	4.5	4.13	5.32
Chr 11	5.08	4.94	4.35	4.76	4.86	4.4
Chr 12	4.12	3.64	4.22	3.78	3.34	4.8
Chr 13	3.85	3.38	3.78	4.11	3.11	3.77
Chr 14	4.26	4.42	5.61	6.72	5.54	5.6
Chr 15	2.75	2.86	2.02	3.46	3.22	2.86
Chr 16	9.62	4.68	5.48	6.07	4.75	4.97
Chr 17	2.2	2.34	2.9	3.33	3	2.52
Chr 18	2.34	4.16	3.02	2.15	2.66	3.03
Chr 19	1.37	2.34	1.83	2.15	1.53	1.83
Chr X	4.12	5.97	3.53	3.85	3.39	3.2
Chr Y	2.2	3.64	1.83	2.22	3.17	2.86
Mitochondria	0.55	1.04	0.69	0.33	1.24	0.86

Supplementary Table 1: Genomic distribution of rAAV-host DNA junctions. Percentages of junctions mapped on each mouse chromosome (Chr) and mitochondrial genome are represented for the two controls and the four mouse samples.

Nature Medicine doi:10.1038/nm.3578

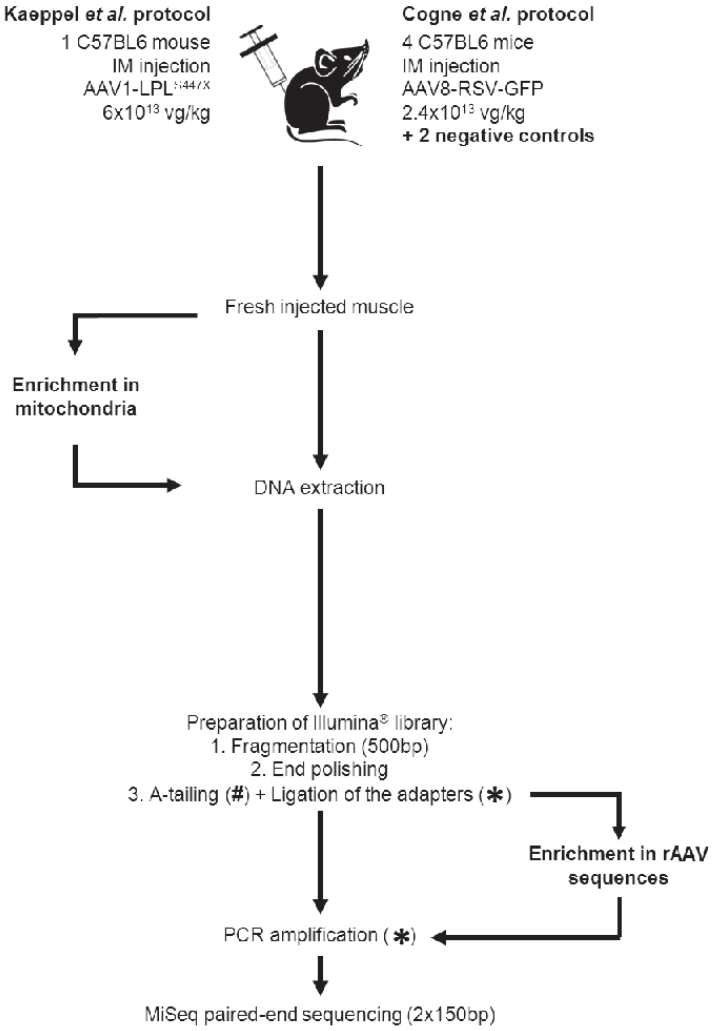


Supplementary Figure 1: Distribution of junction breakpoints along rAAV genomes. The upper panel represents an annotated map of rAAV2/8 RSVp eGFP SV40-pA genome. The coverage of reads mapped along rAAV having a mate pair in the mouse genome (GRCm38/mm10) is figured below, for negative control replicates and experimental samples. A maximum of 119, 61, 201, 213, 188 and 172 overlapping reads were obtained for Control 1, Control 2, Mouse 1, Mouse 2 Mouse 3 and Mouse 4, respectively.



Supplementary Figure 2: Schematic representation of the paired-end sequencing protocol applied by Kaeppel *et al.* and in our study (left) compared to clonal or pre-amplified rAAV IS (right). When rAAV IS are only represented once, the bioinformatic analyses cannot differentiate them from artefacts (left). When IS are represented several times, the multiple retrieval of the same event by bioinformatics allows to correctly identify the IS among the artefacts (right). ITR: inverse terminal repeat, #: chimeric fragment generated during adapter ligation step, *: rAAV integration site.

Nature Medicine doi:10.1038/nm.3578



Supplementary Figure 3: Comparison of the sequencing protocols used by Kaeppel *et al.* (left) and by our laboratory (right), from DNA extraction to the preparation of the library. The standard procedure for library preparation is indicated in the center and specific steps are in bold under the corresponding protocol. IM: intramuscular; vg/kg: vector genomes per kilogram; (#): In the protocol used by Kaeppel *et al.* this step was replaced by a direct blunt ligation of the adapters. (*): Steps prone to generating artificial chimeric reads

Nature Medicine doi:10.1038/nm.3578

Vu, le Président du jury,

Virginie FERRE

Vu, le Directeur de thèse,

Philippe MOULLIER

Vu, le Directeur de l'UFR,

UNIVERSITÉ DE NANTES **2015**

Cogné Benjamin

Caractérisation des virus adéno-associés recombinants par séquençage haut débit : limites et perspectives

---

La thérapie génique consiste à traiter des maladies en transférant des acides nucléiques. Ces nouveaux médicaments permettent d'envisager le traitement de maladies génétiques rares aujourd'hui incurables. Le système le plus efficace pour transférer des acides nucléiques consiste à utiliser des virus modifiés ou vecteurs viraux. Les vecteurs dérivés des virus adéno-associés (AAV) ont récemment prouvé leur efficacité clinique dans le traitement de l'hémophilie B ou encore de l'amaurose congénitale de Leber. L'utilisation de ces nouveaux médicaments implique toutefois de nouvelles problématiques de biosécurité, qui doivent être précisément étudiées avant leur utilisation à plus large échelle.

Dans ce contexte, nous avons développé de nouvelles méthodes utilisant le séquençage haut débit (NGS) et des analyses bio-informatiques afin d'étudier deux aspects importants de la biosécurité des vecteurs AAV : (1) l'évaluation de leur génotoxicité en analysant les sites d'intégration génomique, (2) l'étude des contaminants ADN accidentellement encapsidés pendant la production de ces vecteurs et administrés aux patients (ex: gènes de résistance aux antibiotiques). Cette thèse présente les aspects techniques de préparation des échantillons pour le NGS (plateformes Illumina), les différentes problématiques inhérentes au NGS auxquelles nous nous sommes confrontés et les applications futures de cette technologie dans le domaine de la thérapie génique.

MOTS CLÉS : THERAPIE GENIQUE, SEQUENCAGE HAUT DEBIT, GENETIQUE, BIO-INFORMATIQUE, GENOTOXICITE, AAV

JURY

PRÉSIDENT : **Mme Virginie FERRE**, Pharmacien, Professeur de Virologie
Faculté de Pharmacie de Nantes

ASSESSEURS : **M Philippe MOULLIER**, Médecin, Directeur de recherche
Inserm UMR1089, Nantes
M Stéphane BEZIEAU, Pharmacien, Professeur de Génétique
Faculté de Médecine de Nantes
Mme Marianne COSTE-BUREL, Pharmacien, Praticien Hospitalier,
CHU de Nantes
M Adrien LEGER, PhD, Ingénieur de recherche
Inserm UMR1089, Nantes

Adresse de l'auteur : ***11 rue Flandres Dunkerque, 44100 NANTES***