Analyse lexicale des métadonnées de transcriptomique: Une application au cas du cancer de la prostate

Etude réalisée par : *Décembre 2021*

Ludovic Moisan — Faculté des Langues

Jingwen Liao — Faculté des Langues

Roland Mondjehi — Faculté des Sciences Economiques et de Gestion

Etude supervisée par : **Bruno Kieffer** — ESBS Strasbourg

Introduction

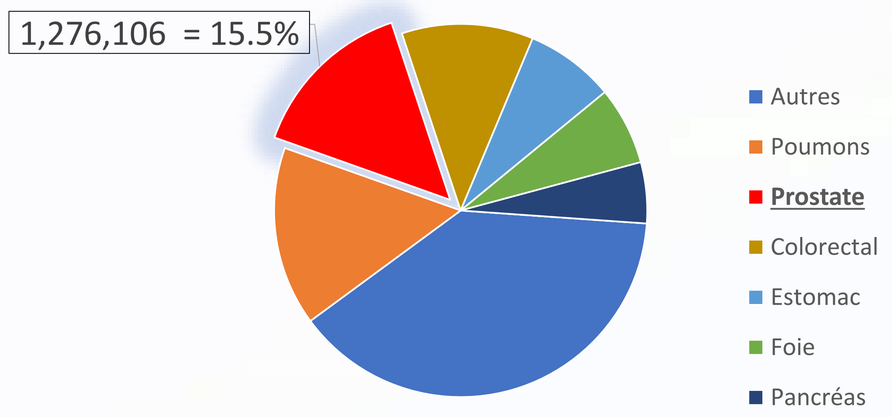
Dans le cadre d'un projet interdisciplinaire du parcours « Approche Interdisciplinaire en Sciences des Données » (AISD), nous avons mis en œuvre des techniques de fouille de données pour analyser les travaux de recherche sur le cancer de la prostate.

En analysant conjointement les métadonnées des données de séquençage des expériences de transcriptomique et les publications associées, nous pouvons observer comment les techniques de séquençage à haut débit (NGS) ont rapidement transformé la recherche fondamentale dans le domaine du cancer.

Qu'est-ce que la transcriptomique ?

La transcriptomique regroupe un ensemble de techniques permettant une analyse quantitative relative des ARN (comparaison des transcriptomes entre différentes conditions expérimentales) et une analyse qualitative par la caractérisation de variant d'épissage, de polymorphismes etc. Cela permet notamment de tester comment réagit les cellules face à un certain type de médicament, leur réaction dans l'expression des gènes face à ces molécules, afin de déterminer les meilleurs outils pour combattre les cellules contaminées par un cancer.

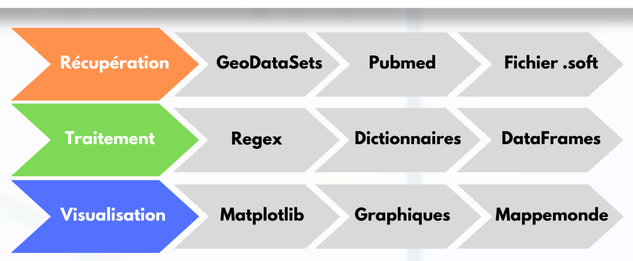
Incidence de cancer pour les hommes en 2018
Dans le monde



Source : wcrf.org

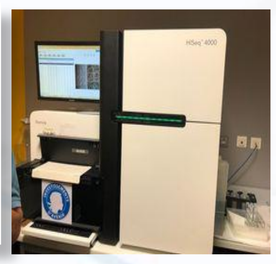
Méthodologie

La méthodologie de ce projet repose entièrement sur le langage Python. Nous avons utilisé le module BioPython pour l'analyse des entrées de séquençages dans les bases de données GeoDataSets et pour la récupération des publications sur PubMed à l'aide des mots clés : « Androgen Receptor & Prostate ». Nous avons ensuite nettoyé et analysé les métadonnées par Regex pour les stocker sous la forme de DataFrames dans Pandas préalablement aux différentes représentations graphiques.



Afin de mieux comprendre le contexte et les technologies mises en œuvre, nous avons réalisé une visite de la plateforme de séquençage de l'IGBMC.

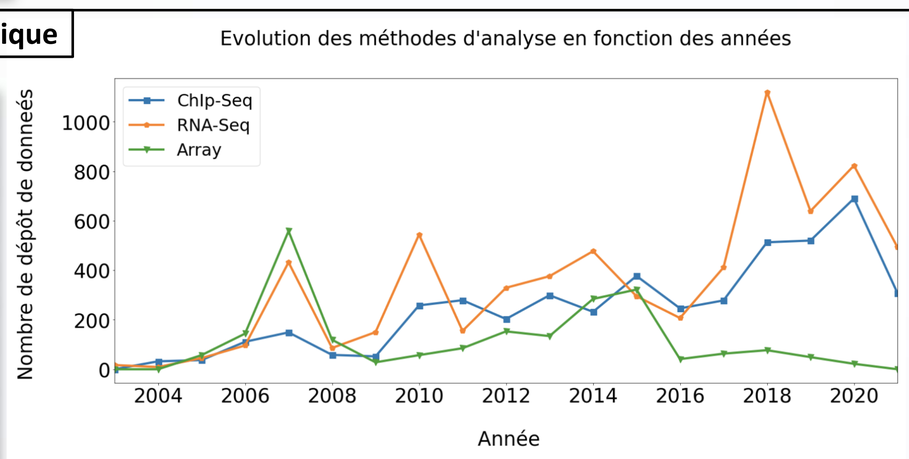
À gauche, une puce pour le séquençage haut débit (technologie Illumina).
À droite, la machine de séquençage haut débit réalisant la récolte de données

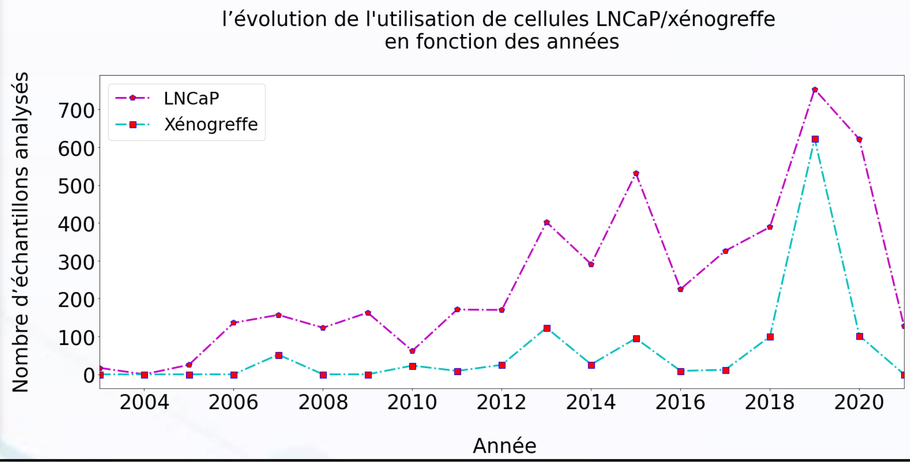


Evolutions technologiques en transcriptomique

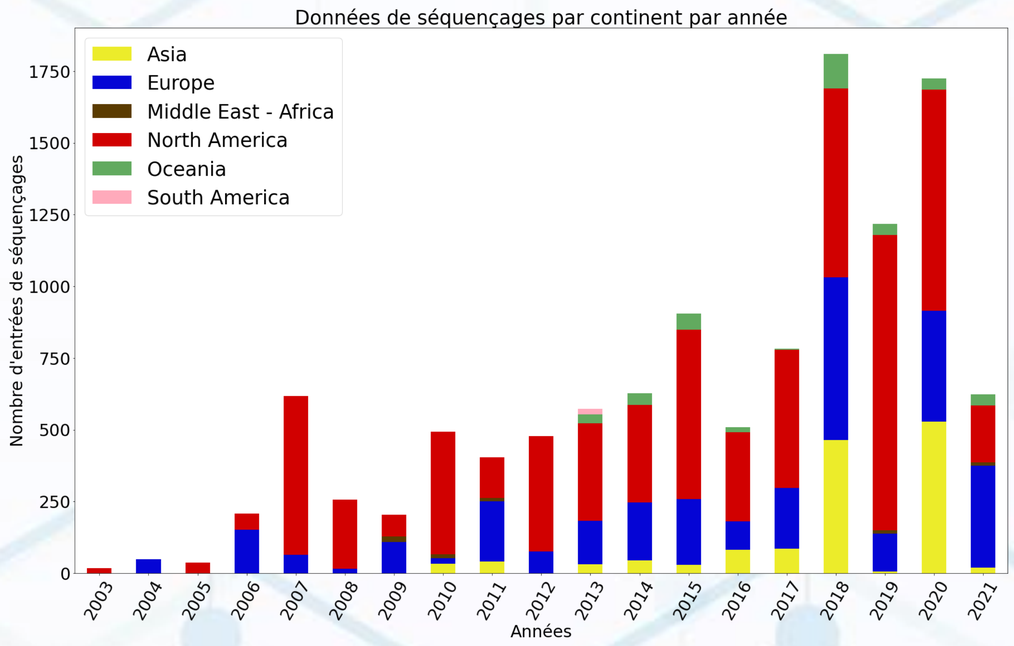
Nous analysons ici les évolutions technologiques des différentes méthodes de séquençage ADN/ARN utilisées dans la recherche grâce aux métadonnées. La méthode la plus ancienne, par Puces à ADN (ou « Array ») disparaît au profit de méthodes de séquençage haut- débit (« RNA-Seq »), souvent associées aux données de ChIp-Seq (sites de fixation des facteurs de transcription).

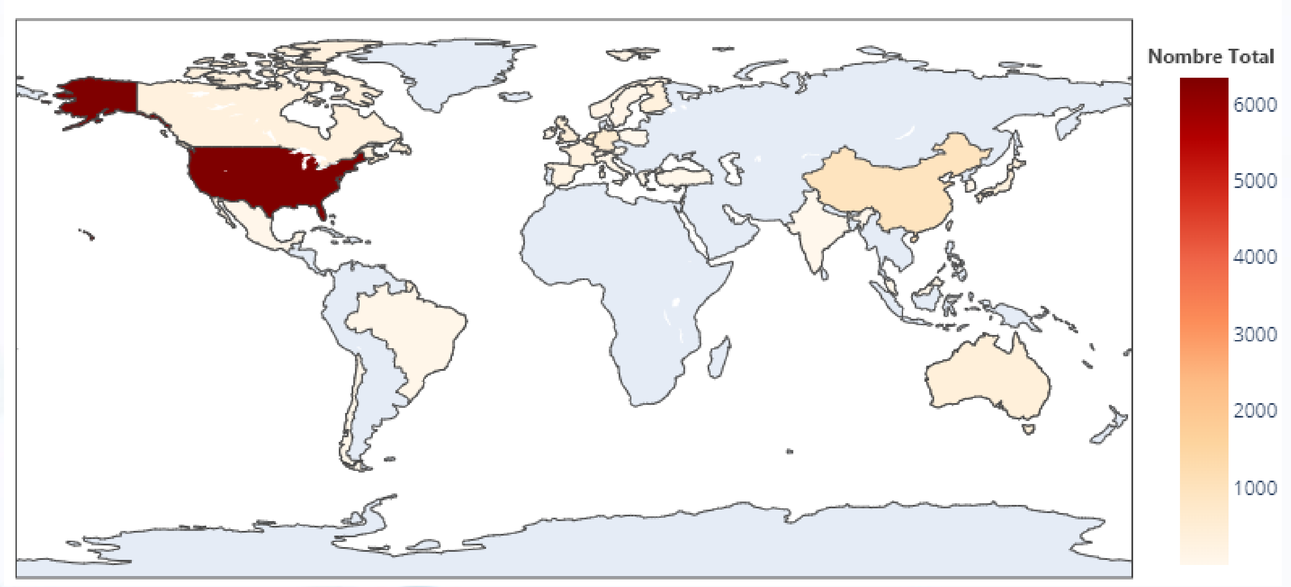
Le modèle cellulaire le plus étudié est le modèle LNCaP, nous voulions donc la comparer avec une autre source de cellules plus innovante, les xénogreffes. Bien que souffrant d'une utilisation peu fréquente dans ce domaine précis de la recherche, cette dernière à gagné un regain d'attention avant l'arrivé de la Pandémie Mondiale de 2020.





Répartition géographique des efforts de recherche contre le cancer de la prostate





Nous analysons ici la différence de répartition géographique entre les publications médicales et les données de séquençages pour la recherche. Nous pouvons observer une nette disparité au niveau mondiale, avec un nombre de publications européennes égale ou supérieure à celles du continent Nord Américain, alors que ce dernier fournit l'essentiel des données. Les pays Asiatiques, majoritairement représentés par la République Populaire de Chine, restent minoritaires dans les publications, bien que présents dans ces dernières années dans le domaine du séquençage. Enfin, la visualisation par pays nous montre également une disparité au sein même des continents, avec les Etats-Unis d'Amérique et le Royaume-Uni particulièrement présents dans le domaine des publications de recherches sur le cancer de la prostate, tandis que les pays Africains ou Asiatique restent en marge.

CONCLUSION

L'analyse des métadonnées de recherches nous montre une claire disparité au niveau mondial. L'analyse des métadonnées, surtout pour la construction des bases de données d'entrées de séquençages, nous a permis également de mettre en lumière un manque de normalisation des contenus. Certaines métadonnées semblent redondantes ou très peu utilisées. Néanmoins, notre analyse nous a permis de voir une nette évolution dans le temps, tant sur les technologies que sur les pays engagés dans la recherche contre le cancer de la prostate. La proportion des métadonnées utilisées reste encore marginale tant dans ce projet quand dans les sites de référencements, il serait intéressant par exemple d'analyser les acteurs humains et les centres de financements dans une prochaine étude.