Rare diseases and big data : biocomputing solutions towards knowledge-guided analyses : applications to ciliopathies

Kirsley Chennen

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

Laboratoire de Génétique Médicale - UMR 1112 et

Complex Systems & Translational Bioinformatics – ICube - UMR 7357

THÈSE présentée par :

Kirsley CHENNEN

soutenue le : **14 Octobre 2016**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Bioinformatique et Biologie des systèmes

Maladies rares et 'Big Data': Solutions bioinformatiques vers une analyse guidée par les connaissances. Applications aux ciliopathies.

(Rare Diseases and Big Data: Biocomputing solutions towards knowledge-guided analyses. Applications to ciliopathies)

THÈSE dirigée par :

Mme DOLLFUS Hélène	Professeur, Université de Strasbourg, Strasbourg
M POCH Olivier	Directeur de recherche, Université de Strasbourg, Strasbourg

RAPPORTEURS :

M BARBRY Pascal	Directeur de recherche, IPMC, Nice
M ROZET Jean-Michel	Directeur de recherche, IHU Imagine, Paris

AUTRES MEMBRES DU JURY :

M LAPORTE Jocelyn	Directeur de recherche, IGBMC, Illkirch
M MOSZER Ivan	Docteur, Institut du Cerveau et de la Moelle Épinière, Paris
Mme DAUCHEL Hélène	Maître de Conférences, Université de Rouen, Rouen

Remerciements

Avant tout, je voudrais exprimer mes plus sincères remerciements à Pascal Barbry, Jean-Michel Rozet, Jocelyn Laporte, Ivan Moszer et Hélène Dauchel pour l'honneur qu'ils me font de juger cette thèse.

Comme Raymond n'a cessé de le rappeler, cette partie est presque toujours la dernière section à être écrite, mais elle n'en demeure pas moins toute aussi importante et sinon la plus difficile à écrire, car son achèvement symbolise aussi celui de l'aventure qu'a été cette thèse.

Je commencerai par remercier les laboratoires qui m'ont accueilli et qui m'ont fourni comme cadres un véritable écrin pour effectuer ma thèse. Ce fut une expérience extrêmement riche, tant du point de vue professionnel que personnel, sans compter la très bonne ambiance qui y règne et la fulgurance de tout un chacun pour le bon mot !

Merci d'abord à l'équipe LGM et aux « filles du 9ème ». Merci à Élodie, Cathy, Malika, Megana, Noëlle, Julia, Yvan et Jean-Marc qui m'ont accueilli à mon arrivée. Merci aussi à Corinne de m'avoir fait découvrir l'importance des aspects techniques et de validations essentielles pour l'analyse d'exomes, ainsi que Véronique et Jean pour les nombreuses discussions pour les aspects bioinformatiqes. Merci à Sophie et Elise pour leurs amitiés et qui ont eu la patience de m'expliquer les nombreuses subtilités existant dans le monde des maladies rares. Merci à Xiangxiang de m'avoir fait découvert la culture chinoise et tous ces bons plats.

Merci à Vincent et Raj, les deux compatriotes qui ont fait vivre un peu de l'île Maurice sur Strasbourg.

Je voudrais ensuite remercier les personnes du CSTB. Merci à Luc, Anne, Odile, et Laetitia pour leur gentillesse et leur bienveillance. Merci Wolfgang, notre Mozart des statistiques, qui m'a fait découvrir les arcanes et toutes les subtilités de la transcriptomique. Merci à Raymond de m'avoir fait découvrir les nombreux aspects de l'informatique (des cartes à trous jusqu'au TCL). Merci à Arnaud pour les nombreuses sessions du PoliKafé, et les jeunes du CSTB pour les nombreuses sorties : Isaac, Raphaël, Pierre, Renaud, Yannis, Hélène, Julio, Salma. Biensûre une pensée spéciale à mes « sparing partners » durant cette thèse, Alexis et Carlos pour leur amitié et leur aide tout au long de thèse. Merci à Julie d'avoir partagé sa culture anglaise non seulement *via* sa grande patience et ses suggestions lors de la relecture de la thèse en anglais, mais aussi en partageant les spécialités culinaires, dont le fameux christmas pudding.

Enfin, il ne reste que deux personnes sans qui cette thèse n'aurait pas été possible et pour qui les mots de remerciement me manquent pour exprimer ma gratitude. Chers Hélène et Olivier, je ne doute pas que mes nombreux prédécesseurs vous ont remercié avec des mots tout aussi choisis pour tout ce que vous avez transmis et fait pour eux scientifiquement et professionnellement parlant. C'est donc mis à part ces aspects évidents que représente l'aboutissement thèse, je vais vous remercier pour le plan humain et personnel tout au long de ces années.

Pour finir, merci à ma famille et surtout vous deux, papa et maman. Cette réussite est aussi la vôtre et rien de tout cela n'existerait si vous n'aviez pas toujours été là pour tant donner et pour m'aider à réaliser mes projets.

Preface

The work realized during this thesis was carried out under the co-direction of Hélène Dollfus, head of the Laboratory of Medical Genetics (LMG), and Olivier Poch, co-head with Pierre Collet of the Complex Systems & Translational Bioinformatics (CSTB) team. The thematic of my thesis is the development of bioinformatics solutions devoted to the deciphering of rare diseases, and notably the ciliopathies. Evolving in this unique translational research environment, I had the opportunity to work on different aspects in the spectrum of problematic around rare diseases, ranging from cohort analysis up to gene identification and functional genomics data analysis.

During the first part of my thesis, I had the opportunity to invest in a better comprehension of genetic and biological causes of rare diseases such as ciliopathies. The work done was realized through a close collaboration with the biologists and clinicians of the Hélène Dollfus team for the 'classical/manual' analysis of whole-exome data, and even, of the 'pre-exome' phase involving the clinical characterization of patient cohorts. In the second part of my thesis, I had the opportunity to complement this experience with an immersion in the CSTB team. Through the fruitful exchanges with the bioinformaticians and bioanalysts of the CSTB team, I could better apprehend the breadth and depth of this data-intensive era in biology, by working and manipulating different high-throughput data (from exomes to transcriptomes…). I was able to grasp not only the limits and bottlenecks of these approaches, but also the importance of the effective integration of large amounts of biological data and biomedical information at all steps of the analysis process, from the raw data analysis up to the validation step. This background helped me to conceive a stratified network view of the biological complexity of life, which I tried to integrate in the developments and bioanalyses performed during my thesis.

In this manuscript, I present three original and robust bioinformatics resources (VarScrut - Chapter 5; Pub*Athena* - Chapter 6; CilioPath - Chapter 7) resulting from this combined biomedical and bioinformatics experience and adapted to the needs of the biologists and clinicians in rare disease research. These resources have already successfully been used to identify three novel ciliopathy genes, two of which have been published (BBIP1/BBS18 as the 18th gene of a very emblematic ciliopathy, the Bardet-Biedl syndrome – see Annexe II:; VSP15, a novel ciliopathy gene linking the process of autophagy to the ciliogenesis – see Annexe III:).

Throughout my thesis, I had the opportunity to work on different rare diseases related projects, which are not presented and discussed in this manuscript notably: i) the transcriptomics analysis for the identification of biomarkers differentiating Alström and Bardet-Biedl syndromes (in collaboration with Vincent Marion from the LMG; currently in the final validation and patenting phase), or ii) the design of a targeted gene sequencing panel for routine molecular diagnosis of Retinopathies (in collaboration with Jean Muller from the LGM; currently in the sequencing and analysis phase of 150 exomes).

Résumé de la thèse

Au cours de la dernière décennie, la recherche biomédicale et la pratique médicale ont été révolutionné par l'ère post-génomique et l'émergence des « Big Data » en biologie, qui s'est maintenant transformée en une science de données de manière intensive avec diverses approches à haut débit (génomique, protéomique, métabolomique, lipidomique ...). En conséquence, les données biologiques sont en expansion et croissent plus rapidement que notre capacité et efficacité à les compiler, les organiser et à les manipuler, limitant ainsi nos efforts pour extraire les connaissances biologiques sous-jacents et à fournir des indications en santé humaine.

Les défis associés à l'ère des « Big Data » sont souvent représentés par le paradigme des quatre Vs qui englobe: i) le Volume, pour l'énorme quantité de données générées, ii) la Vélocité, pour la fréquence à laquelle les données sont générées et mis à jour, iii) la Variété, pour la diversité des sources et des structures de données et iv) la Véracité, pour l'incertitude due à de l'incohérence et des données manquantes. Il y a aussi un cinquième V émergent, qui correspond à la Valeur, pour l'extraction de connaissances significatif. Dans la biomédecine, ce paradigme émergent suppose non seulement, une meilleure intégration des données et des informations au sein et entre la recherche et la médecine, mais aussi, la considération des ressources de calculs et humaines nécessaires pour extraire une connaissance pertinente.

Dans ce contexte biomédical renouvelé, l'étude des maladies rares a pris de l'ampleur car il semble que des analyses axées sur les sous-ensembles mono-géniques fournissent un contexte simplifié pour éclairer les maladies communes telles que l'obésité ou diabète. Il y a environ 7000 maladies mono-géniques rares et affecte sur l'ensemble ~30 millions d'européens. L'origine génétique des maladies rares est établie pour 80% d'entre elles, et les gènes responsables ont été identifiés pour seulement 50% d'entre elles. Paradoxalement, à notre ère des « Big Data », le facteur le plus limitant, pour l'étude des maladies rares est la rareté des données liées au nombre limité de patients atteints par une maladie rare donné. Cela se traduit en une disponibilité dispersée et parcellaire des données génétiques et phénotypiques ainsi que des matériaux biologiques. En outre, il y a aussi un manque de ressources dédiées à la communauté biomédicale pour la collecte et la maintenance des données et des matériaux biologiques.

Dans ce contexte, ma thèse a été consacrée à l'élaboration de solutions bioinformatiques pour améliorer l'étude des maladies génétiques rares. Les objectifs étaient donc à développer: (I) une architecture unifiée pour l'analyse des données d'exomes pour améliorer le diagnostic moléculaire et l'identification de nouveaux gènes (VarScrut), (II) une resource dédiée pour rester à jour avec le flux constant de données entrants et fournir un outil robuste pour la veille de la littérature scientifique et la fouille de textes des articles pertinents, par exemple ceux qui associée à des maladies mendéliennes rares (PubAthena), et (III) le développement d'un prototype de base de connaissances (CilioPath) intégrant les outils développés dans un projet plus global dédié à l'étude des ciliopathies et des retinopathies.

I) VarScrut:

Récemment, l'étude des maladies rares et les diagnostics moléculaires en général, ont été révolutionné par l'avènement des technologies de séquençage de nouvelle génération. Les technologies telles que l'exome, consacré à la partie du génome codant pour les protéines, a démontré sa force pour l'identification de variants et de gènes responsables de maladies génétique. Cette approche a été appliquée avec succès dans plus de 1400 études publiées. Cependant, l'identification de la mutation / gène responsable par séquençage d'exome reste laborieuse et le taux de résolution est d'environ 25%. Il y a plusieurs facteurs qui entravent la résolution d'une étude de l'exome.

La qualité de séquençage est décisive pour la détermination du variant causal. Les stratégies actuelles pour le séquençage d'exome sont capables de capturer seulement environ 90% des exons de protéines connus. Cependant, certaines régions, en particulier les séquences répétitives ou riches en GC, sont sous-représentées. Ainsi, il existe un risque de 10% que même une variation évidente d'une maladie ne soit pas capturée.

Le passage des tests génétiques au séquençage à haut débit a été accompagné par de nouveaux défis, notamment pour l'interprétation et l'analyse des variations de séquence de signification clinique inconnue (VUS). Plusieurs recommandations ont été émises pour améliorer cette analyse, en particulier ceux de l'« American College of Medical Genetics and Genomics » (ACMG). En parallèle, plusieurs logiciels et algorithmes ont été développés pour prédire l'effet des variations de séquence, et ils ont généralement une précision de 65 à 80% lorsqu'ils ont été utilisé sur des données de variations pathogènes connues. Cependant, il est

très difficile d'avoir une interprétation globale et intégrée de l'importance des multiples scores fournis par les différents outils.

Enfin, même avec de bons protocoles de séquençage et d'analyse de qualité, le succès repose sur le fait qu'un certain nombre d'hypothèses doivent être correctes, telles que le mode de transmission ou la pénétrance élevée. Cependant, les cas isolés et les petites familles nucléaires sont les cas les plus fréquents en clinique influençant ainsi l'hypothèse génétique.

Globalement, le taux de diagnostic actuel de 25% de l'exome est lié essentiellement au défi de certifier qu'une variation est effectivement causale. La classification des variations de séquence dans les catégories à risque élevé ou faible est à la fois difficile et critique pour la clarification du statut de causalité. Lorsque la condition de la causalité d'une variation de séquence est indéterminée, l'ACMG recommande des activités de suivi qui pourraient être utiles pour clarifier cette relation et aider à l'évaluation les risques. Cela inclut des procédures pour recueillir des informations supplémentaires, comme la veille périodique de la littérature, des bases de données et de la disponibilité de modèles animaux pertinents pour les maladies humaines.

Dans ce contexte, j'ai d'abord comparés les approches existantes montrant qu'aucun des 40 outils disponibles ne met pleinement en œuvre les recommandations de l'ACMG. Plus, important encore, cette étude a révélé que la majorité des 1400 études d'exomes publiées étaient en fait analysées manuellement par des experts. Cela nous a conduit à concevoir et développer VarScrut, une infrastructure basée sur quatre modules principaux qui considère l'analyse des données d'exome comme un processus unifié allant de l'analyse des données brutes et l'identification / hiérarchisation des variants candidats à l'aiguillage pour l'étape de validation expérimentale. VarScrut est centrée sur une structure qui combine une approche multi-niveau de filtrage d'annotation (15 outils d'annotations), avec des méthodes fonctionnelles étendues axées sur la fusion des connaissances combinant les données phénotypiques, les voies biologiques et de la littérature dans un réseau d'interactions gène-gène. Le rôle potentiel de nouveaux gènes pathogènes est estimé en utilisant l'algorithme « Random Walk with Restart » (RWR) sur le réseau d'interaction des gènes. Pour limiter l'impact d'une information de pedigree incomplète, VarScrut permet l'exploration automatique et simultanée de cinq scénarios génétiques (autosomique récessive, autosomique dominante,

lié à l'X récessif, dominant lié à X, de novo). Ce module de scénario génétique capitalise sur la compilation précédente des règles et commentaires dans la littérature.

Les performances de VarScrut ont été évalués d'une part, sur les données de référence simulée et d'autre part, sur de vrais exomes à travers l'analyse de deux applications cliniques sur des patients atteints du syndrome de Bardet-Biedl et de myopathies. Dans le test de référence, VarScrut a démontré des performances plus élevées que les autres outils largement utilisés (VAAST, PHEVOR et PhenGen) en classant les variants / gènes responsable parmi les 10 premiers dans 61% des cas en utilisant seulement les données génotypiques et 97% des cas en combinant les données de connaissances génotype et phénotype. Sur les cas réels d'applications cliniques, VarScrut a été utilisé avec succès pour identifier le nouveau gène BBS18 qui a été publié dans le Journal of Medical Genetics (Scheidecker et al., 2014) ainsi que deux gènes associés à des ciliopathies - actuellement dans les phases de validation expérimentale.

En outre, pour les exomes non résolus avec une liste de sortie étendue de variants candidats, VarScrut intègre un module de suivi pour la réévaluation automatique et le suivi sur la durée. Le module de suivi fournit i) un rapport des gènes peu ou pas couverts dans les données d'exomes, ii) un suivi des mises à jour quotidiennes des articles Pubmed liés aux gènes de la liste ainsi iii) qu'une mise à jour mensuelle de toutes les annotations pour relancer l'analyse automatiquement. Enfin, à la fin du processus analytique, VarScrut fournit un module de «validation expérimentale» pour faciliter la phase de validation post-exome en fournissant à l'utilisateur des informations supplémentaires telles que la disponibilité de matériels biologiques (ex: les anticorps et les organismes modèles).

I) PubAthena:

L'un des défis majeurs dans la recherche biomédicale est de rester à jour sur les dernières découvertes. Mais avec plus de 100 000 nouveaux articles publiés chaque mois, il est extrêmement difficile de trouver les articles les plus pertinents, sans rien manquer. Même avec un effort dédié pour capturer des informations liées à la maladie dans les bases de données biologiques (ex: OMIM), beaucoup de ces informations reste "verrouillées" dans le texte non structuré de publications biomédicales. Ainsi, il y a un décalage important entre la publication et l'extraction ultérieure de ces informations dans des bases de données. Plusieurs systèmes de fouilles de texte ont été développés, mais la plupart d'entre eux sont consacrés à

des concepts spécialisés tels que le nom de gènes (iHOP), les mutations génétiques (tmpVar) ou les processus biologiques (GoPubMed) et quelques-uns d'entre eux offrent la possibilité de combiner l'analyse des articles avec d'autres concepts (PubTator).

Dans un premier temps, j'ai élaboré et évalué un système de fouille de textes, PubAthena, qui priorise et recommande les articles les plus pertinents sur la base d'un ensemble d'articles d'apprentissage dans la bibliothèque de l'utilisateur et en extrait des concepts annotés telles que les mutations des protéines, les maladies associées et les phénotypes ou les processus biologiques du résumés MEDLINE. PubAthena référence tous les articles disponibles dans PubMed (~22 millions en Juin 2015), avec des mises à jour tous les soirs en utilisant le service e-utils d'Entrez. Ce système comprend une série de modules de prétraitements tels que la détection des phrases et la « tokenisation » en utilisant NLTK, une bibliothèque de langage python. Il comprend également un volet de normalisation lexicale, qui convertit les mots sous forme canonique en utilisant le générateur de variante lexicale (LVG) fournies par la National Library of Medicine (NLM). Le composant d'identification de termes comporte des modules pour reconnaître le nom des protéines, des mutations ponctuelles, des processus ou voies biologiques, les descriptions phénotypiques et le nom des maladies des résumés de PubMed. Les annotations des entités (la protéine / gène, mutation et les maladies) sont récupérées à partir des résumés de PubMed en utilisant le service web de PubTator. Les annotations correspondant aux processus biologiques, les voies biologiques et les descriptions phénotypiques sont obtenues en utilisant un développement en interne complétant les annotations de PubTator. En ligne avec les services de PubTator, qui fournit les identifiants d'une protéine / d'un gène, d'une mutation ou d'une maladie correspondant aux bases Entrez Gene, dbSNP et OMIM respectivement, je développé de nouveaux protocoles pour identifier les processus biologiques et description phénotypique correspondant à des termes GO, KEGG / Reactome et HPO / MPO (pour l'homme et la souris seulement) respectivement. L'identification des articles pertinents est basée sur un module de classification bayésienne qui utilise des articles stockés dans la bibliothèque de l'utilisateur en tant que jeu d'apprentissage et de mots-clés et d'entités normalisées et annotées en tant que paramètres pour évaluer les nouveaux articles. PubAthena a été évaluée sur un recueil de 600 articles associés aux cils et aux ciliopathies et 200 articles associés à des rétinopathies. Les articles publiés avant l'année 2010 ont été utilisé comme jeu d'apprentissage. PubAthena obtient une F-mesure de 88% et 90% respectivement pour la recommandation des articles pertinents publiés après l'année 2010. Les travaux futurs comprendront l'analyse au niveau du

discours pour améliorer les performances d'extraction et l'association protéine-mutation-maladie et de l'indexation intégrale des articles.

IIII) CilioPath:

Enfin, la recherche active fait au cours de la dernière décennie a produit beaucoup de données liées aux cils / ciliopathies. Malheureusement, les données accumulées sont dispersés parmi les diverses ressources hétérogènes qui sont soit des bases de données généralisées (OMIM) ou des bases de données dédiées (Ciliaproteome, Cildb, CiliomeDatabase, Cilia / Centrosome Interactome complexe et Syscilia), et même d'autres de sujets connexes (CepDB). La plupart du temps, ces ressources dédiées ne sont plus à jours (ex : Ciliaproteome), ou fournissent une liste de gènes qui ne représente qu'un aspect spécifique comme, par exemple, les études à haut débit (ex : Cildb).

Dans la croyance que la disponibilité d'un répertoire de référence collectant des preuves expertisées sur des protéines ciliaires constituerait une ressource clé pour la communauté scientifique, j'ai développé CilioPath qui est une base de connaissances pour des analyses exploratoires sur des gènes et maladies liés au cil. CilioPath est composé d'une base de connaissance supportant la recherche de données, le filtrage et le partage.

CilioPath est alimenté semi automatiquement de nouveaux articles proposés par une veille hebdomadaire de PubAthena, qui sont ensuite validés par des experts du domaine. Ceci garantit une mise à jour dynamique du recueil de CilioPath, qui contient actuellement 4736 articles associés au cil. Les connaissances de ce recueil sont extraites afin de compiler tous les gènes ciliaires identifiés dans des études haut débits, de validations ou médicales. Actuellement, CilioPath contient ~ 2700 gènes humains impliqués directement ou indirectement avec les cils dont 41 sont des études à haut débit (19 protéomes, 3 études de xbox, 15 transcriptomes et 4 études de génomique comparatives) et 25 études de validation. Les données ont été cartographiées sur les génomes de l'homme et de la souris à l'aide des relations d'orthologies via InParanoid. CilioPath affiche également les informations sur les maladies et les mutations pour les 92 gènes pathogènes, représentant 18 types de ciliopathies référencés. Les annotations ont été enrichies avec l'intégration des données de plusieurs niveaux tels que la coprexpression (COExpressDB), les voies biologiques (KEGG, Reactome), les processus biologiques (GO), la localisation (GO, Protein Atlas) et les

interactions gène-gène (StringDB, GeneMania, ConsensusPathDB, IntAct et BioGrid). En outre, CilioPath soutient également les informations du réseau en y associant des données phénotypiques issues de HPO, MPO et ainsi que d'une cohorte de 400 patients avec une ciliopathies, du laboratoire de Hélène Dollfus (documentée et normalisée en termes HPO), évaluée sur 86 critères phénotypiques.

Pour conclure, dans le cadre de ma thèse, de nouvelles solutions bioinformatiques ont été développés pour améliorer la compréhension des maladies rares par la collecte, l'analyse et l'intégration dans des bases de connaissances spécialisées de données et d'informations pertinentes et exploitables. Pour être en mesure de saisir toutes les informations pertinentes dans ce flux continu de données, l'outil PubAthena fournit des recommandations précises (65-80% de précision), avec des moyens utiles pour extraire les connaissances encore "verrouillé" dans les articles biomédicaux et absent de bases de données publiques. En outre, pour améliorer la production de nouvelles connaissances par l'identification de nouveaux gènes, VarScrut a été développé pour analyser les exomes afin d'étudier les maladies rares. VarScrut améliore la résolution des exomes en combinant les informations de génotype et phénotype (classement parmi les 10 premiers dans 97% des cas). En outre ces efforts combinés ont été utilisés pour générer une base de connaissances dédiée à l'étude de ciliopathies. Grâce à notre base de connaissances, le chercheur aura la possibilité d'étudier plusieurs aspects de la biologie du cil et de ses ciliopathies. À savoir: (1) l'étude de l'évolution des protéines ciliaires fournira via leur regroupement une prédiction de nouveaux gènes ciliaire, ainsi que leur fonction, (2) le déchiffrement de l'association génotypes-phénotypes aidera à un meilleur diagnostic et une meilleure compréhension des mécanismes sous-jacents communs aux ciliopathies chevauchant, (3) sur l'ensemble, cela améliorera notre compréhension du rôle du cil dans les différents processus biologiques. Plus, généralement, toutes les ressources et les procédures élaborées au cours de cette thèse, bénéficieront également de l'étude d'autres maladies rares comme les rétinopathies.

Thesis summary

This manuscript is divided into four main parts:

The first part consists of an Introduction where I expose the fundamental concept of heredity, which is at the heart of this thesis (Chapter 1), and the challenges associated to rare diseases characterized by the inherent scarcity and scattered resources in the current renewed 'Big Data' biomedical context (Chapter 2). I illustrate the complexity and challenges of rare diseases through the emblematic example of ciliopathies – an emerging class of rare diseases typified through an underlying defective organelle, the cilia.

The second part, Material & Methods, describes the datasets (Chapter 3) and the bioinformatics resources (Chapter 4) used to carry out my work.

The third part, Results and Discussion, presents how the study of rare diseases can be improved by mobilizing data-intensive resources available in biology and medicine. Mainly organized around the manuscripts published or submitted, this part encompasses:

i) a variant analysis tool for whole exome sequencing data to improve the molecular diagnosis and the identification of causative disease-genes (VarScrut; Chapter 5);
ii) a dedicated resource to remain up-to-date within the constant flux of incoming data through the automated literature monitoring and text-mining of relevant articles (PubAthena; Chapter 6);
iii) the development of a prototypal up-to-date knowledge base (CilioPath; Chapter 7) - organized around the central role of the cilia, and coupled to integrative methods to propose a dynamic gold standard of ciliary and ciliopathy genes in the constant flux of incoming data.

The final chapters, Conclusion and Perspectives, discusses some improvements that can be drawn for the deciphering of rare diseases, based on the work carried out during this thesis. On the long winding road from the gene to the phenotype, now starting to be paved with high-throughput approaches and data, I conclude that, although rare diseases represent a challenge through their uniqueness, they potentially represent the new bonanza that can inform and shed light on more complex diseases. Moreover, all the experience and dedicated resources gained from rare diseases might also be re-employed in the coming perspectives of personalized medicine.

Table of content

List of figures

List of tables

Abbreviations

ACMG	American College of Genetics and Genomics
BBS	Bardet-Biedl Syndrome
MoI	Mode of Inheritance
nsSNV	Non-synonyous Sinvgle Nucleotide Variant
RD	Rare Disease
SNV	Single Nucleotide Variant

I. Introduction

Chapter 1 Heredity

1.1 The legacy of the concept of heredity

1.1.1 *The cradle of heredity*

Since their origin, humans have endeavoured to understand the origin of life through the study of the generation of offspring. Even before the initial spur of the concept of '[heredity](#)' in the 17th century, the resemblance between an offspring and its parents had occupied philosophers and scientists since antiquity. The insights on the road to heredity came mainly from a conjunction of a series of hypotheses and discoveries from three distinct disciplines – biological science, agricultural practice, and medicine (Cobb, 2008).

Through the domestication of animals and plants, ancient humans realized that "like breeds like", *i.e.* that each species reproduces according to its kind. The initial germ of thoughts at the foundation of the concept of heredity took root back in antique time. The Greeks already knew that each species produces only offspring like itself and that each animal shows a combination of inherited traits from its progenitors. The physician Alcmaeon of Croton (~510–530 BC) later introduced, through his theory on the physiological nature of semen, that besides the male progenitor, the female progenitor also contributes to heredity *via* the production of an internal and invisible semen in order to explain that children often resembled their mothers (Leitao, 2012). Hippocrates (460-357 BC), the father of Western medicine, believed that the hereditary traits in the semen came from the organs and also, from what he formulated as the four bodily fluids, or "humours" – the blood, phlegm, black bile, and yellow bile. However, the philosopher and scientist Aristotle (384-322 BC) believed that the semen comes from the blood and hence that the invisible female semen was represented by their menstrual blood. His hypothesis was that the mother's semen represents the matrix and the father's semen would provide the shape, *i.e.* transforming the mother's matrix to give an individual.

The idea of heredity only took flesh in the 17th century with the invention of the microscope by Antoni van Leeuwenhoek (1632–1723); reviewed in (Lane, 2015). With the microscope, scientists could observe and describe for the first time the gonad cells (Figure 1-1): oocyte/ovarian follicles by Regnier de Graaf (1641–1673) (Ankum et al., 1996; de

Graaf, 1672); spermatozoa by Antoni van Leeuwenhoek (Lewenhoeck, 1678). Although it was commonly admitted that the fertilization of the female "egg"/semen by the male semen produces a child, there was still a raging debate on the contribution brought by each semen that lasted decades (Cobb, 2006).

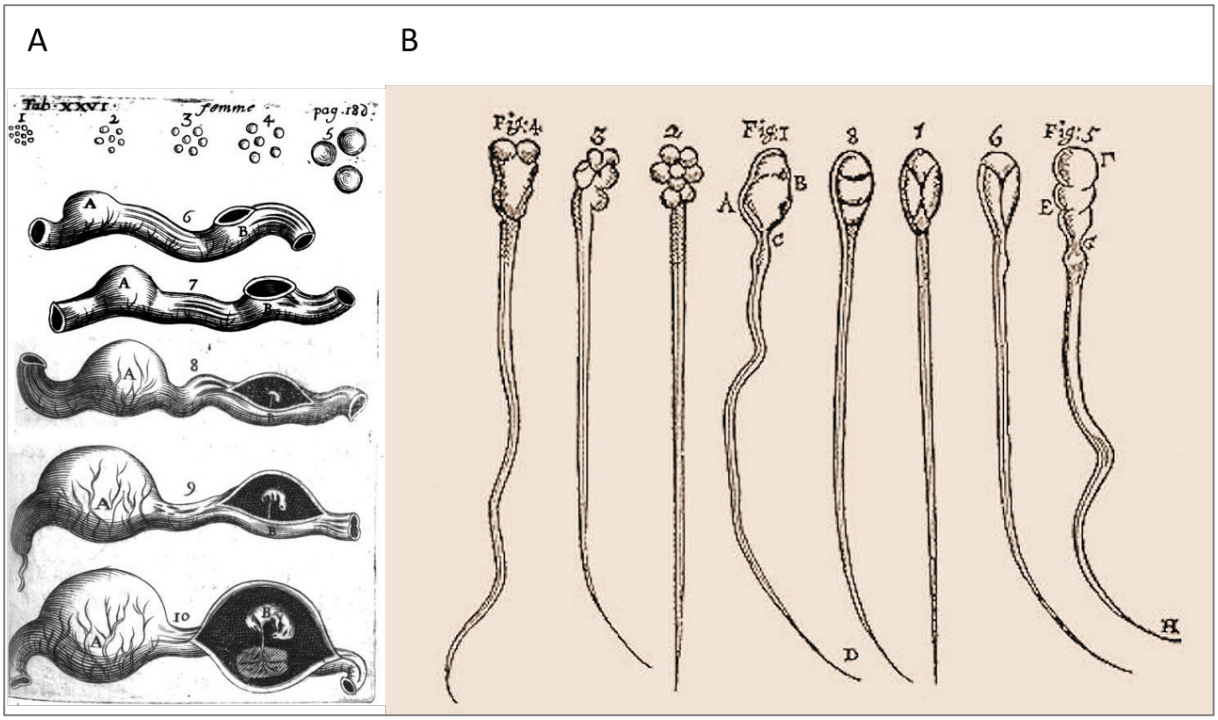


Figure 1-1: First description of gamete cells, the vectors of heredity. (A) First description by Regnier de Graaf of the evolution of rabbit's ovarian follicles, present in the fallopian tube, into a foetus in a time frame of ten days after copulation. (B) First description of spermatozoa from rabbits and dogs by Antoni van Leeuwenhoek (1676–1678), which he later used to formulate his theory of "animalculisme": the foetus is formed by the spermatozoa and there is therefore no need for eggs, but only an environment for fertilization. (Source: Wikimedia commons)

The first extensive work on heredity came from the contribution of French physicians through a series of competitions around the hereditary disorders organized in the 18th century by the Academy of Medicine of Dijon (1748) and the Royal Society of Medicine of France (1788–1790) (López-Beltrán, 1995). Through the description of hereditary "pathological characters" of certain illnesses, various physicians tried to answer the question: "How are hereditary illnesses transmitted?" (López-Beltrán, 1995). These works on what is now known as genetic diseases, contributed not only to the recognition and acceptance of the concept of "heredity" but also to the reflection concerning the traits that could be transmitted through generations by the semen. One famous illustration is the study of polydactyly within familial cases, reported independently by the French naturalists Pierre Louis Moreau de Maupertuis

(1698–1759) (de Maupertuis, 1745) and René-Antoine Ferchault de Réaumur (1683–1757) (de Réaumur, 1751). Through their case reports, they demonstrated the transmission of individual traits in humans, and that some traits had different transmission patterns (*i.e.* dominance of the polydactyl trait) (Figure 1-2) (Sandler, 1983). A century before Mendel, Maupertuis forecasted that the heredity of traits is due to "particles" present in the parents' semen, and that it is the "affinity" between the pairs of "particles" that defines whether the father's or mother's trait will dominate. These conclusions contributed to a wider reasoning on the pool of hereditary traits, *i.e.* are the combination of traits transmitted in different patterns? Can there be possible variations of the transmitted traits between two generations? (de Maupertuis, 1746).

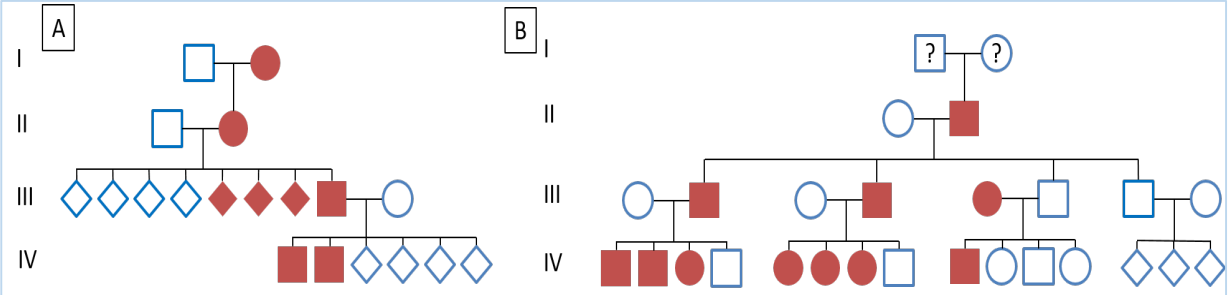


Figure 1-2: Representation of pedigrees reported by de Maupertuis and de Réaumur on familial cases of polydactyly. (A) Representation of the description of polydactyl inheritance in a Ruhe family as reported by de Maupertuis. (B) Representation of the inheritance of polydactyly in the Kellai family of Malta as reported by de Réaumur. The Roman figures indicate the successive generations. The forms represent the gender of the family members: squares represent males; circles represent females; diamonds indicate that the gender is unknown. For the status of polydactyl trait, solid shapes in red indicate affected individuals, open shapes the unaffected, and the question mark indicates that the status is unknown. Adapted from (Pasternak, 2003).

Later, the British breeder Robert Bakewell (1725–1795) exploited these reflections together with the insights already present in agriculture to develop his innovative stockbreeding techniques, which supported the livestock need during the British Industrial Revolution. Among his revolutionary agricultural techniques, Bakewell is renowned for his selective breeding, which he applied successfully on sheep (with his barrel-shaped sheep), cattle and horses (Black Cart horse). Later Bakewell's method was one of the sources of inception that led Charles Darwin to emit his theory on the origin of species, whereby the natural selection of characters that are fixed in a population can be assimilated as an analogy to selective breeding (described as artificial selection) (Darwin, 1859).

Finally, Gregor Mendel really set the foundation for the study of heredity by defining the laws governing the inheritance of characters (Mendel, 1866). Mendel established that

transmissible characters could take several forms, with different and independent patterns of segregation. These laws, resulting mainly from the study of plant heredity (*Pisum sativum*), brought new insights to explain some previous observations performed one century earlier by the French physicians, and stating that some human characters were transmitted from generation to generation, while others skipped a generation.

Thus, it is clear that the lens used by scientists to reveal the concept of heredity has forged the set of fundamental structuring theories for comprehension of life. The understanding of the "generation" of offspring, more generally allowed us to perceive the concept of evolution, the importance of these transmissible characters for the future of a given species or health status.

1.1.2 *An era of molecular insights on heredity*

With the concepts of heredity revealed, the next step was the race to discover what could be the support and mechanisms of heredity. This odyssey was possible with the concomitance of a series of scientific and technological advances that brought each time a leap in the deciphering of heredity. The initial discovery came with the description, by Walther Flemming (1879) (Paweletz, 2001) and Heinrich Wilhelm Waldeyer (1880) (Winkelmann, 2007), of the transmission of structures, later named chromosomes, during cell division. Theodor Boveri and Walter Sutton (Crow and Crow, 2002) later proposed in 1902, that the chromosomes are the physical basis that bear the hereditary factors in accordance with the Mendelian law of heredity. To describe these hereditary factors, Wilhelm Johannsen (1909) (Roll-Hansen, 2014) coined the term 'gene' to define the physical and functional unit of heredity.

Nevertheless, it was really in the 20th century that our molecular comprehension of heredity started to improve. The milestone came from the discovery by Oswald Avery in 1944, that deoxyribonucleic acid (DNA) is the molecule of life (Avery et al., 1944). While most biologists at that time expected proteins to be the molecule of heredity, Avery demonstrated that DNA carries all the genetic information and is responsible for the "transforming principle" of *Pneumococcus*, previously described by Griffith's experiment (Griffith, 1928). This result was later confirmed by the Hershey-Chase experiment (Hershey and Chase, 1952), using phage T2 with radio labelled DNA and proteins to infect cells. The Hershey-Chase experiment showed that it is only the DNA molecule that is transferred to the

infected cells, and that DNA is the only genetic material carrying all the information for the synthesis of proteins and generation of new phages.

Finally, even though not strictly limited to the notion of heredity, other milestones strongly contributed to the understanding of its molecular basis, notably: (i) the discovery by James Watson and Francis Crick in 1953 of the DNA structure encoding all the genetic instructions of heredity (Watson and Crick, 1953); (ii) the formulation of the central Dogma in Biology by F. Crick in 1958, which stated how the genetic information is transmitted in a flow to the cellular effectors (Figure 1-3)(Crick, 1970); (iii) the deciphering of the genetic code containing all the instructions by Matthaei in 1962 (Matthaei and Nirenberg, 1961). All these advances provide the robust foundations for the modern view and description of the different modes of inheritance that shape our current analyses of human genetic diseases.

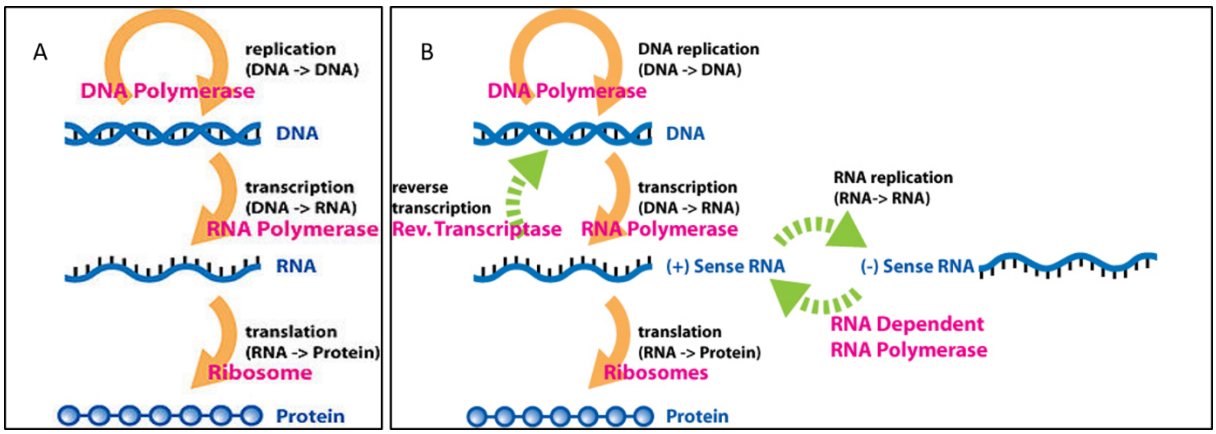


Figure 1-3: The central dogma of molecular biology. (A) Initial dogma established by Francis Crick for the flow of genetic information through three main steps. (B) Current renewed dogma with additional steps of genetic information transfer. (Source: Wikimedia common).

1.2 Gene inheritance and transmission of characters

1.2.1 *Modes of inheritance*

Mendel's studies on the modes of inheritance of characters in pea plants are the core foundation of our current understanding of single gene-based characters. In a diploid organism (such as human beings), each gene has two alleles that can be described as: (i) homozygous (same copy on both alleles) or (ii) heterozygous (different copies on each allele), if the variations occur at the same locus, or (iii) compound heterozigosity if the variations occur at different localizations in the same gene. The modes of inheritance, also known as

Mendelian or monogenic transmission, are extensively studied in human genetic diseases. Pedigree analyses of families with affected individuals are used to determine whether a disease-associated gene is located on an autosome or on a sex chromosome, and whether the related disease phenotype is dominant or recessive.

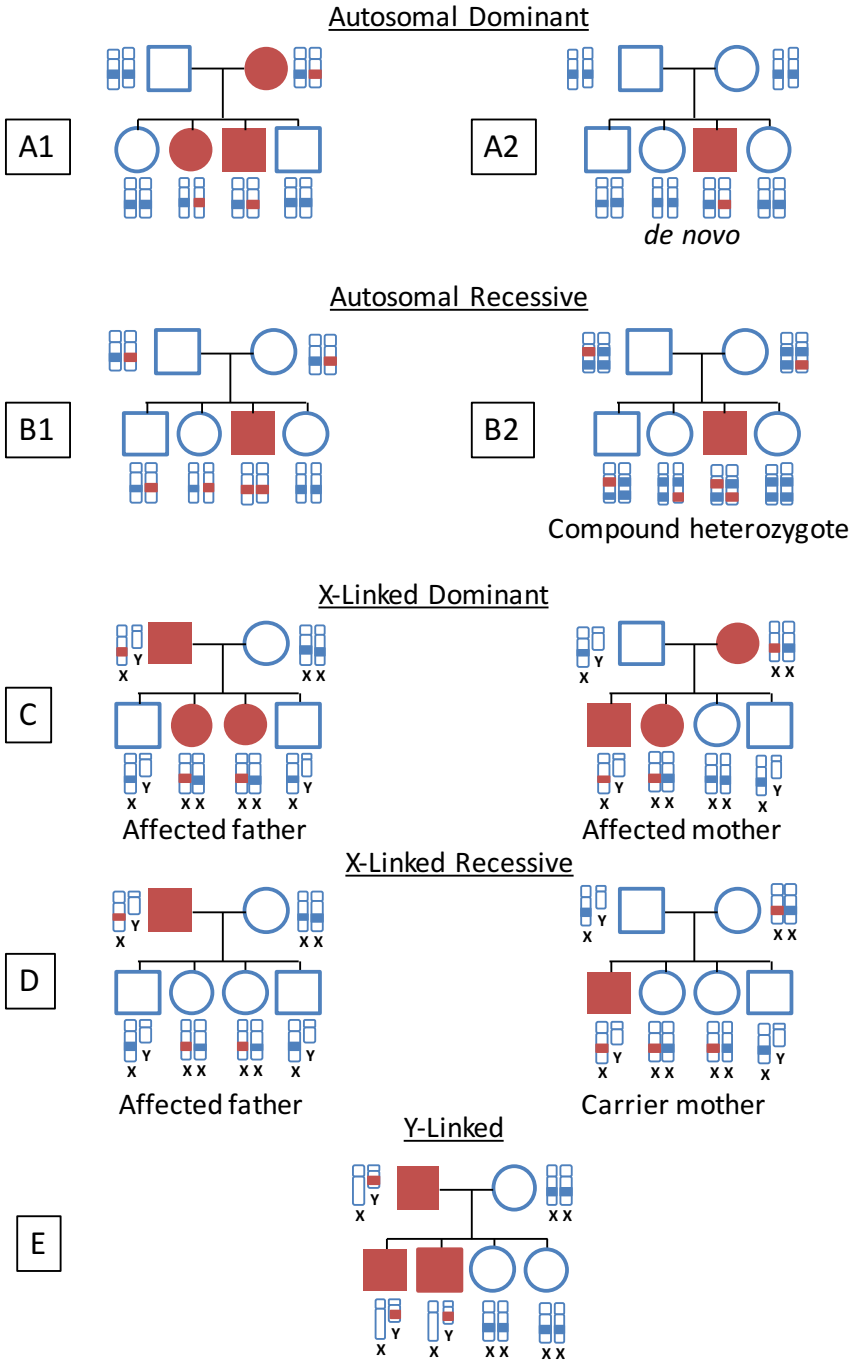


Figure 1-4: Modes of inheritance (MoI).The autosomal mode represents cases where the gene responsible for the phenotype is located on one of the 22 pairs of autosomes (non-sex-determining chromosomes). The X-linked or Y-linked modes represent cases where the gene that encodes for the trait is located on the X or Y chromosome respectively. The disease phenotype is illustrated by the red colour while the normal or carrier phenotype is in white. The forms represent the gender of the family members: squares represent males, and circles represent females.

1.2.1.1 Autosomal dominant diseases

Autosomal dominant single-gene diseases occur in affected individuals who have a single mutant copy of the disease-causing gene. In this case, the presence of a single wild-type copy of the gene is not sufficient to maintain the normal function of the gene and prevent the disease. Individuals can inherit the mutant copy of the disease-associated gene from either an affected mother or an affected father (Figure 1-4 A1). In some rare cases, an un-inherited mutation, absent from the parents (also called a "*de novo*" mutation) can occur in the affected child (Figure 1-4 A2).

Polycystic kidney disease is an example of an autosomal dominant single-gene disease. Autosomal dominant polycystic kidney disease [OMIM: 173900] can be caused by mutations in the polycystic kidney disease 1 (PKD1) or the polycystic kidney disease 2 (PKD2) genes.

1.2.1.2 Autosomal recessive diseases

Autosomal recessive single-gene diseases occur only in individuals with two mutant alleles of the disease-associated gene (Figure 1-4 B1), which can occur either at the same locus, or at different localizations in the case of a compound heterozygosity (Figure 1-4 B2). Individuals with an autosomal recessive single-gene disease inherit one mutant allele of the disease-causing gene from each of their parents.

Phenylketonuria (PKU) is a prominent example of a single-gene disease with an autosomal recessive inheritance pattern [OMIM: 261600]. PKU is associated with mutations in the gene that encodes the phenylalanine hydroxylase (PAH) enzyme. Individuals with mutations in the PAH gene cannot metabolize the amino acid phenylalanine. As a result, phenylalanine accumulates at high concentration in the urine and blood of PKU patients, which eventually causes mental retardation and behavioural abnormalities.

1.2.1.3 X Chromosome–linked dominant diseases

Single-gene diseases that involve genes present on the sex chromosomes have different MoI compared to genes present on autosomes. The reason for these differences lies in the genetic distinction between males (two distinct sex chromosomes: X and Y) and females (two copies of the X sex chromosome). Few dominantly inherited forms of human diseases are X chromosome linked. Females with an X chromosome-linked dominant disease

can inherit the mutant gene from either an affected mother or an affected father (Figure 1-4 C), whereas males always inherit such diseases from an affected mother.

Examples of X chromosome-linked dominant diseases are rare, but several do exist. For example, Rett syndrome [OMIM: 312750], a neurodevelopmental disease, is associated with dominant mutations in the methyl-CpG-binding protein 2 gene (MECP2). Rett syndrome almost exclusively affects females, because male embryos with a dominant mutation in the MECP2 gene rarely survive.

1.2.1.4 X Chromosome–linked recessive diseases

Females with an X chromosome-linked recessive disease must inherit one copy of the mutant gene from an affected father and the second mutant gene copy from their mother, who can be a healthy carrier (heterozygous) or an affected one (homozygous) (Figure 1-4 D). Males have only one copy of the X chromosome received from their mother. Thus, males with an X chromosome-linked disease always receive the mutant gene copy from their mother. Moreover, since males do not have a second copy of the X chromosome to potentially balance the negative effects of X-linked mutations, they are more affected than females by X chromosome-linked recessive diseases.

The blood-clotting disorder haemophilia A [OMIM: 306700] is one of several single-gene diseases that exhibit an X chromosome-linked recessive pattern of inheritance. Males who have a mutant copy of the factor VIII gene (F8) will always have haemophilia.

1.2.1.5 Y Chromosome–linked diseases

Like X-linked dominant diseases, Y chromosome-linked diseases are also extremely rare. Since males have a single Y chromosome, Y-linked single-gene diseases are always passed on from affected fathers to their sons and there is always a manifestation of the associated phenotype, whether the mutation is dominant or recessive (Figure 1-4 E).

One example of a Y-linked disorder is non-obstructive spermatogenic failure [OMIM: 400042], a condition that leads to infertility problems in males. This disorder is associated with mutations in the ubiquitin-specific protease 9Y gene (USP9Y) on the Y chromosome.

1.2.2 *Variability of the genetic information*

Since the introduction of Darwin's theory of descent with modification and the mutation theory by Hugo de Vries (1901), scientists have been aware that the diversity of

phenotypic traits within a species, or even the emergence of a new species, are due to the appearance of variations in the DNA sequence that can be heritable and persist through generations. Today, with access to the genome of many species, we know that what defines the uniqueness of an individual is the complex combination of inherited traits and the accumulation of private variations within the individual. We can thus define the genotype as the genetic makeup that characterizes an individual and consequently, the phenotype as the observable traits of a genotype. In humans, the variation between the genomes of two unrelated individuals is estimated to vary with an average of 0.1% across all ethnic groups. However, the genome is a dynamic entity, which accumulates variation events throughout the life of the individual. As opposed to the variations that occur only in the germ-line / somatic cells, the variations present in the gametes cells are the only ones that participate in the transmission of traits to the offspring and in the species evolution.

1.2.2.1 Causes of variability

Among the main sources of variations in the genome there are:

1. The crossing-over of chromosomes demonstrated by Harriet Creighton and Barbara McClintock in 1931, whereby, during meiotic division, an exchange of genetic material between homologous chromosomes shuffles the allele content between homologous chromosomes (Creighton and McClintock, 1931).
2. Errors introduced during the replication of the DNA during meiosis, despite the high fidelity of most DNA replication and DNA repair enzymes (overall error rate of 10^{-10} per bp) (Kunkel and Erie, 2005). These errors are one of the sources of *de novo* permanent/heritable mutations.

Errors that can occur during the repair of the DNA caused by naturally occurring or induced damages. A well-known example is the *xeroderma pigmentosum* [OMIM: 611153] disease, resulting from the deficiency of the DNA repair mechanisms for damage induced by ultraviolet radiation.

1.2.2.2 Consequences of variability: genetic variations

Depending on the event, the resulting variation can occur either at: (i) a large-scale level (also known as structural variants), usually corresponding to rearrangements of large chromosomal regions that comprise several gene loci and are thought to largely involve alterations in gene dosage; or (ii) a small scale, affecting only few nucleotides (<50 bp).

The small-scale variations include:

1. Point variations that can be classified as either transitions, for substitutions of a purine by another one or a pyrimidine by another one (Adenine⇔ Guanine; Cytosine⇔ Thymine) or as transversions for substitutions of a pyrimidine by a purine or inversely. Point variations are also referred to as Single Nucleotide Variants (SNV), and they represent the most frequent variation events (90% of the variability) with over three million SNV in a human genome with respect to the reference genome (assembly of nucleotide sequence representative of the species).
2. Insertions or Deletions (InDel) – the addition or the removal of one or more nucleotides. These events are usually introduced into the genome by small transposable elements or during DNA replication.

Depending on the regions where these events occur, the variations may have no consequence or may alter the associated phenotypical trait. In genes (protein- or non-protein-coding genes), point variations can alter the phenotype in different ways depending on their localization with respect to the gene and on the type of variation (Figure 1-5):

1. Untranscribed regions of a gene. These regions include promoters, silencers, enhancers ... recognized by transcription factors (TF) regulating gene expression. A variation in these regions could alter the normal expression of the gene by modifying the specific time and space frame of the gene activity (*e.g.* Hox genes).
2. Spliced regions/introns of any gene or UnTranslated Regions (UTR) of a protein-coding gene. Variations in these transcript regions can affect the stability of the transcript and usually result in its rapid decay, with fewer final gene products available. At the disease level, statistically important untranslated regions are the splice site regions (Exonic Splicing Enhancers, Exonic Splicing Silencers), and any variation overlapping them will affect the processing of precursor RNAs into mature RNAs, and hence modify the final gene product.
3. Translated regions of a protein-coding gene. During translation, the encoded information in the mRNA is converted into a sequence of amino acids that will constitute the resulting peptide, using the genetic code (64 combinations of triplets of 4 nucleotides that can be converted into 20 amino acids). The consequences of the variations falling in translated regions can be classified into 3 categories:
    a. Same protein length with identical amino acid sequence – caused by point variations or substitutions that do not affect the final amino acid sequence and are referred to as silent or synonymous variants.

b. Same protein length with varying amino acid sequence – usually correspond to point mutations. Depending on the nature and the position of the amino acid in the protein sequence/structure, it can be tolerated or disrupt the structure and function of the protein.

c. Different protein length with varying amino acid sequence. Variations can result in a difference in the length of the final amino acid sequence, including (i) a nonsense variation that will introduce a premature stop codon, hence shortening the peptide sequence, (ii) a stop lost variant, which will elongate the original peptide sequence or (iii) alternative splicing, frequently resulting from the alteration (gain or loss) of donor/acceptor splice sites. Insertion or Deletion (InDel) variants can also alter the peptide sequence through two types of variants: (i) Inframe InDel – removal/addition of one or more amino acids in the peptide sequence, (ii) Frameshift variant – generates a different peptide sequence by shifting the translation reading frame which frequently introduces a premature stop codon.

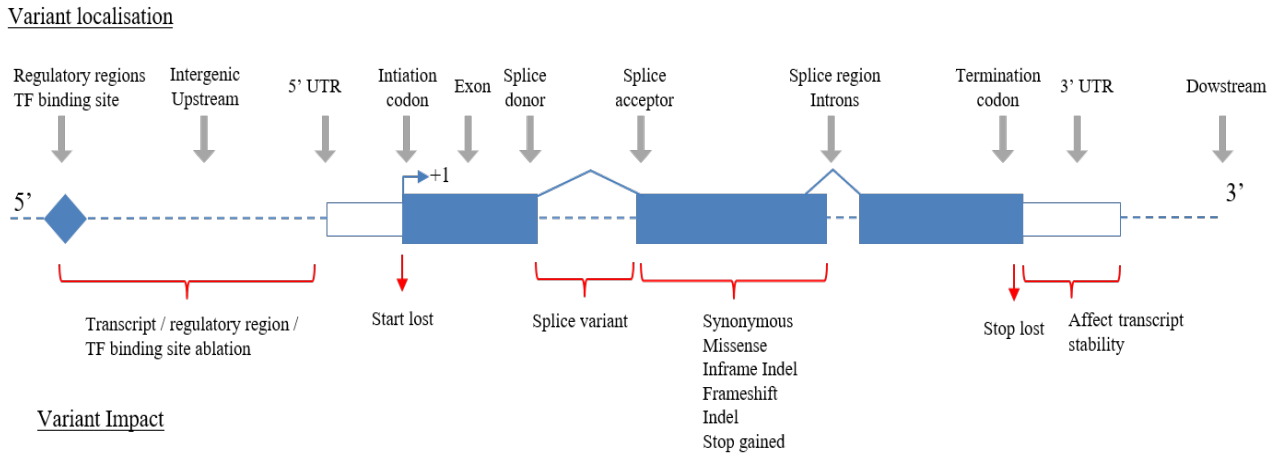


Figure 1-5: Impacts of small-scale variations occurring within a gene structure depending on their localization (Inspired from Ensembl's Variant Effect Predictor documentation)

1.2.2.3 Correlation between genotype and phenotype

The study of the correlation between the genotype and the phenotype is a major entry point to have an insight on the functions of genes and to understand the impact of variations on the resulting phenotype. Nevertheless, characterizing the type and location of a variation is far from sufficient to understand the relationship between an allele and a resulting phenotype and additional information influencing this relationship must be gathered:

1. **Penetrance**

When studying the relationships between genotype and phenotype, it is important to examine the statistical occurrence of phenotypes in association with a particular variant of a gene (allele/genotype). The penetrance represents the proportion of genotypes that actually show expected phenotypes. Incomplete penetrance refers to individuals who have the genotype but do not present the corresponding phenotype. For example, many people with a mutation in the BRCA1 or BRCA2 gene will develop cancer during their lifetime, but some people will not.

2. **Expressivity**

Individuals with the same particular gene variant can also show different degrees of the same phenotype. Expressivity is the degree to which trait expression differs among individuals. Unlike penetrance, expressivity describes individual variability. For example, the features of Marfan syndrome [OMIM: 154700] vary widely; some affected individuals have only mild symptoms, such as being tall and thin with long, slender fingers, whereas other affected individuals experience life-threatening complications involving the heart and blood vessels. All affected individuals with this syndrome have a dominant mutation in the fibrillin1 gene (FBN1), but it turns out that the position of the mutation in the FBN1 gene is correlated with the severity of the Marfan phenotype.

Chapter 2 Rare Diseases

2.1 What is a Rare Disease?

A Rare Disease (RD) is any disorder that affects only a small percentage of individuals compared to the general population (prevalence). The threshold to define rarity differs across countries and regions: the European Union (EU) considers diseases to be rare when no more than 1 individual in 2,000 European citizens are affected (European Parliament and the Council, 2000), while the United States (US) defines a RD as a disease affecting "less than 200,000 persons or about 1 in 1,500 people" (107th United States Congress, 2002). The number of RD patients varies considerably from disease to disease, and most RD have a prevalence fewer than one in 100,000 people affecting only thousands, hundreds, or even fewer than a dozen patients worldwide (Figure 2-1).

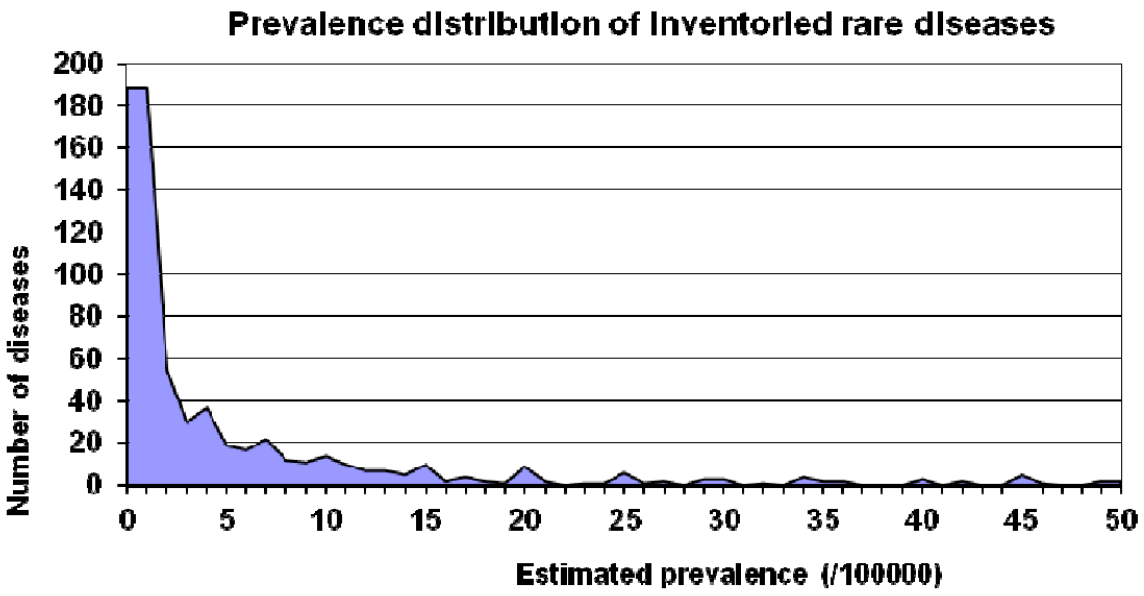


Figure 2-1: The majority of the inventoried RD has a very low prevalence (<1/100000). (Source: Orphanet report 2015)

To date, over ~8,000 different RD have been identified (Amberger et al., 2015). Altogether, it is estimated in 2007 that 6-8% of the population of the 25 EU countries (~27-36 million people) are, or will be concerned by a RD (European Medicines Agency, 2007). Similarly, in the US, the estimation of people affected by a RD is around 25 million Americans (107th United States Congress, 2002), and more generally, the rough estimates go up to 400 million worldwide (Kaplan et al., 2013).

Due to their epidemiological definitions, RD represents a heterogeneous class of disorders with different aetiologies: 80% are of genetic origins and 20% are caused by viruses, bacteria or environmental factors (chemicals, food, etc.). Despite a considerable heterogeneity, RD have some common characteristics. Many rare disorders are life threatening (35% of the cases) or chronically debilitating for the patient's physical, mental, behavioural and sensorial abilities. A majority of RD have an early onset (>75%), and almost all RD are without effective treatment or are incurable (Orphanet, 2015). Given their complexity and debilitating nature, living with a RD can be a very isolating experience for the sufferers and their families.

2.2 Selected key issues of RD

Despite the growing public awareness of RD in the last two decades, there are still many gaps in knowledge amenable to their molecular understanding or development of treatment. Indeed, in our era of data-intensive science and medicine, RD are characterized by scarcity at every level: epidemiological data; knowledge on natural history (*i.e.* the progression from the presymptomatic phase to the chronic manifestation phase); data/tools for correct and early diagnosis; appropriate medical care; absence of pharmacology, all of which have an important impact on the burden of many RD.

Among the major key issues characterizing RD, there are:

(i) RD epidemiological data are scarce

Due to the low number of patients and the lack of national or generic registration, databases of patients with genetic and medical information or information on the disease progression are limited. Consequently, the epidemiological data that are available are inadequate for most of the RD to give firm details on the number of patients with a specific RD. Moreover, some disorders are not present or detectable at birth but manifest in the first years of life, at juvenile age or in adulthood. Thus, follow-up studies from birth are needed to assess the true RD prevalence (Alwan et al., 1997). For some specific RD, data are available at regional, national and/or international levels. However, these databases (*e.g.* www.lovd.nl/) are rarely coordinated and frequently, do not use the same codes. Consequently, it is extremely difficult to get reliable epidemiological data on RD and it is hard to find enough patients to gather for proper clinical trials.

(ii) Rare disease expertise is scattered across countries

A consequence of the scattering of patients is that medical expertise on a specific RD is a scarce resource and a heavy logistic has to be put in place to reach the dispersed patients. Hence, fragmented knowledge about RD combined with limited access to research material (biological samples, mice models, etc.) means it is critical that investments in fundamental research go hand-in-hand with investments in dedicated infrastructure and international networks (biobanks, registries, networks of expertise...).

(iii) Lack of an internationally recognized RD classification system

One of the reasons for the lack of registered patients in databases is the lack of a consensus on a RD classification system that could help generate reliable epidemiological data. Such a system would provide a useful basis for further research into the natural history and causes of RD, and enable monitoring of the safety and clinical effectiveness of therapies and assessment of the care quality.

The International Classification of Diseases (ICD) code [1]‡, used by the World Health Organization (WHO), in practice is not convenient for many RD. For example, a specific ICD number is present for the more known RD such as thalassemia, cystic fibrosis, haemophilia, and amyotrophic lateral sclerosis but other rare disorders are summed up as 'other endocrine and metabolic disorders'. Consequently, it is difficult to register RD people on a national or international basis and in a reliable, harmonised way.

(iv) Limited knowledge on RD and diagnostic wandering of RD patients

For over ~1,600 RD [2], the cause (pathway, gene, variation...) of the disease and the pathogenesis/physiopathology are still unknown and there are few insights on the natural history of these diseases (la Paz et al., 2010). No animal models are available, and *in vitro* or *in vivo* studies are limited. Thus, this situation significantly hampers the ability to both diagnose the disease and identify potential pharmacological/therapeutic targets for treatment. For well-documented rare genetic diseases, the diagnosis can be made *via* enzymological methods or molecular biology tools, whereas for other poorly documented RD or unresolved cases, no diagnostic tools exist due to a lack of research on the cause of these diseases. For RD without documented genetic causes, the diagnosis is only clinical.

‡ Square brackets correspond to Webliography references.

Hence, the absence of a diagnosis or late diagnosis may lead to a deterioration of the patient's condition. The time period between the first symptoms of a patient and the final diagnosis varies enormously. A study of the Genetic Interest Group (UK), representing 600 families with a genetic RD, showed that 75% of the diagnoses took on average 6 months (Donnai et al., 2001). In 30% of the cases, diagnosis took more than 2 years and in 15%, more than 6 years. The main problems underlying such a late diagnosis are due to ignorance of the physician, absence of centres of expertise, unavailability of techniques for diagnosis and/or no insight in the natural history of the disease.

Making a disease easy to diagnose at an early stage (notably *via* sequencing technologies, see next section) will allow the development of prevention strategies that, even in the absence of a treatment, can have a significant positive impact on a patient's life. Diagnosis and prevention strategies represent important tools in reducing the burden of RD. Phenylketonuria (PKU) is a classic example where newborn screening allows successful therapeutic intervention through a strict diet or through sapropterin dihydrocloride (Kuvan®) in conjunction with diet that dramatically modify the patient's prognosis.

2.3 Why study RD? / RD: the new bonanza

Initially, RD was receiving very little attention from the different stakeholders. The status and the consideration for RD has changed mainly through the actions of patient associations such as the National Organization for Rare Disorders (NORD) [3] in the USA, the European Organization for Rare Diseases (EURORDIS) [4], or the "Association Française contre les Myopathies" (AFM-Téléthon) in France [5]. These associations advocate for public awareness and recognition of the challenges of RD from the legislative assemblies, researchers and industrialists, with even the instauration of a world RD day on the last day of February every year [6]. These efforts are starting to pay-off as several international initiatives, namely the International Rare Disease Research Consortium (IRDiRC) [7] are emerging to federate research resources around RD such as the RD-Connect, which is an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research (Thompson et al., 2014).

Besides the inherent issues associated to RD, RD patients by their singularity are now starting to be considered by scientists as a future bonanza to unravel novel biomedical knowledge and a means to improve public health for other common diseases. Indeed, biomedical research has historically concentrated on more common diseases, seeking to

benefit the many rather than the few, and as a result RD have often been overlooked. However, a growing body of evidence shows that RD research can yield important insight into more common conditions and fundamental knowledge (Banks et al., 2006; Taylor et al., 2011; Treharne et al., 2009). Additionally, RD with a genetic aetiology are often more extreme and straightforward than their common counterparts, and therefore provide models and gateways to understand common conditions and human physiology. For instance, the Hutchinson-Gilford progeria syndrome (HGPS; OMIM: 176670), a rare disorder characterized by a premature aging starting from the neonatal period, represents a very good model to study cellular aging. Another interesting example is Familial Hypercholesterolemia (FH; OMIM: 143890), which is caused by mutations in the LDL Receptor gene. Biallelic mutations cause the most severe form of disease, and heart attacks. Statins were developed as a cure for FH, and are now the most commonly prescribed drug for high cholesterol, with Pfizer reporting Lipitor sales revenues of $12.4 billion in 2008.

With the advent of novel high-throughput technologies (see section 2.4.2) in the momentum to resolve the genetic aetiology of RD, RD is enlightening the path towards the new era of personalized medicine, which requires a profound understanding of phenotype-genotype relationships.

2.4 Identifying causative genes in rare genetic diseases

About 80% of the RD are of genetic origin, for which the causative genes have been identified for only half of them through research approaches. In the clinical context, the identification of genes responsible for monogenic disorders is essential for molecular diagnosis of affected patients, and prenatal screening. Gene identification leads to a better understanding of the pathophysiological role of the related proteins and disease pathways, which can serve as a starting point for developing therapeutic approaches (Antonarakis and Beckmann, 2006). Over the decades, several conventional techniques used in genetic disease analysis have been applied to resolve the genetic aetiology of RD, with increasing precision.

2.4.1 Conventional methods for disease gene identification

Until six years ago, disease-causing genes, in conventional clinical protocols (Figure 2-2), were identified by Sanger sequencing for a limited number of targeted candidate genes. The candidate genes were targeted based on several assumption criteria, such as: i) functions similar to genes known to be responsible for the studied disease; ii) knowledge-based profiles (tissue or developmental expression, pathways or interaction data, response to drugs…) of candidate genes similar to the knowledge-based profiles of known causative genes implicated in related diseases; iii) positional mapping of genes in a genomic region frequently associated with the studied disease (Botstein and Risch, 2003). The most successful strategy was selecting candidates based on positional mapping techniques that can take advantage of the inheritance patterns observed in affected families and can be applied in an unbiased manner without any prior medical or biological knowledge.

The most frequent genetic positional mapping approaches rely on the identification of genomic features, which are common to all affected individuals and absent from unaffected individuals. Compared genomic features can concern: (i) large genomic regions, such as in the Karyotyping techniques and Copy Number Variant (CNV) analysis, or (ii) small genomic regions used as genomics markers, from microsatellites to Single Nucleotide Polymorphisms (SNP). When available, pedigree data of affected families are used to drastically reduce the chromosomal region to be investigated.

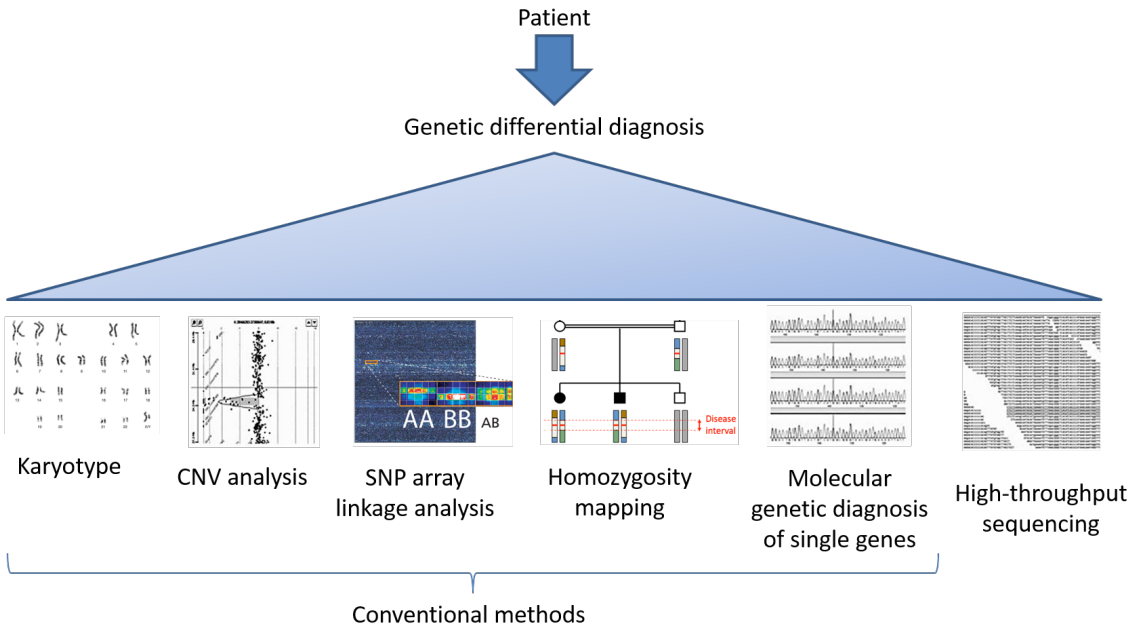


Figure 2-2: Conventional and high-throughput genetics tests used in clinics for molecular diagnosis.

In clinical practices, four main conventional methods have been extensively applied (Figure 2-2):

(i) Karyotype analysis, which involves cytogenetic techniques such as Fluorescence *in situ* hybridization (FISH) for the staining of chromosomes in order to analyse microscopically observable changes in chromosome numbers or structures in affected *versus* unaffected individuals (de Ravel et al., 2007). This technique was used successfully for instance to identify a chromosomal translocation in the 5q35 region of abnormal karyotypes where the gene NSD1 was isolated in patients with Sotos syndrome, a RD characterized by excessive physical growth during first years of life, mild mental retardation, delayed motor, hypotonia and speech impairment (Kurotaki et al., 2001).

(ii) Copy number variation (CNV) analysis, where the CNV in a patient's samples are analysed with different techniques such as fluorescent *in situ* hybridization, array comparative genomic hybridization (aCGH) or single nucleotide polymorphism (SNP) array technologies (Lee et al., 2007). In the aCGH technique, fluorescently labelled DNAs from affected and unaffected individuals are hybridized to the genomic array. Differences in the hybridization signals of some chromosomal regions in affected *versus* unaffected DNA can be detected and can be used for identifying abnormal regions in the genome (Kirchhoff et al., 2001). This technique was used successfully, for instance, in the identification of the gene implicated in CHARGE syndrome, a disorder characterized by hearing defects, retarded growth and development, ear anomalies and genital hypoplasia (Vissers et al., 2005).

(iii) Linkage analysis aims at identifying known marker loci that co-segregate with the disease loci. Historically, marker loci were known microsatellites that were sparse on the human genome and thus, provided a poor resolution power. Today, marker loci are Single Nucleotide Polymorphisms (SNP) grouped on high-throughput arrays (up to 64,000 SNP) and scattered on the human genome thus offering a better coverage. Classically, the LOD score (logarithm of the odds) (Ott et al., 2015) in linkage analysis represents this statistical estimate of the closeness of two loci. A LOD score equal to or higher than 3 means the odds are a thousand to one that the two loci are linked and is generally indicative that loci are sufficiently close to be inherited together. Comparing the data obtained from affected individuals *versus* unaffected relatives of a same family, allows identifying marker loci that co-segregate with the

affected status and to isolate candidate genes found within these regions. This technique was used successfully, for instance, in the identification of the cystic fibrosis gene (Kerem et al., 1989) or of the gene implicated in inflammatory bowel disease, a chronic, relapsing inflammatory disorder of the gastrointestinal tract (Duerr et al., 2006).

(iv) Homozygosity mapping is similar to linkage analysis but generally applied in consanguineous families, to identify genes that cause recessive traits and are specifically located in homozygous regions in affected individuals (Lander and Botstein, 1987). This technique was used successfully, for instance, in the identification of the gene implicated in retinoblastoma, the common intraocular cancer of childhood (Lee et al., 1987).

For over 30 years, these conventional methods have been quite successful with ~3,480 genes discovered to underlie Mendelian phenotypes (Figure 2-3). Although the above-mentioned techniques have been successful in identifying genes, they are now reaching a resolution limit for the remaining Mendelian diseases, which are mainly characterized by sporadic cases or families with a small number of informative relatives. These situations frequently provide extended lists of candidate regions that are difficult and expensive to validate by Sanger sequencing. To tackle these rare sporadic cases, we are now witnessing a major paradigm shift in RD gene-discovery strategies since the year 2010, with the emergence of Next-Generation Sequencing (NGS) technologies; reviewed in (Ng et al., 2010b). The NGS technologies brought novel techniques, such as Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES; discussed in next section) allowing low-cost sequencing of almost all human genes at a glance (Ng et al., 2009). Gene-discovery strategies based on WES and WGS introduced powerful alternatives that were agnostic to both known biology and mapping data and that did not require any prior selection (Brunham and Hayden, 2013; Ng et al., 2010b).

Initially, WES approaches were mainly used in research for novel gene identification. In clinics, which require high standards of stringency, when combined with conventional genetic approaches, WES and WGS have proved to be transforming technologies that have rapidly accelerated the pace of discovery of genes underlying Mendelian phenotypes. Thus, over the last five years with the rapid shift toward the systematic WES/WGS usage, over 650 novel disease genes have been identified (Figure 2-3)(Chong et al., 2015).

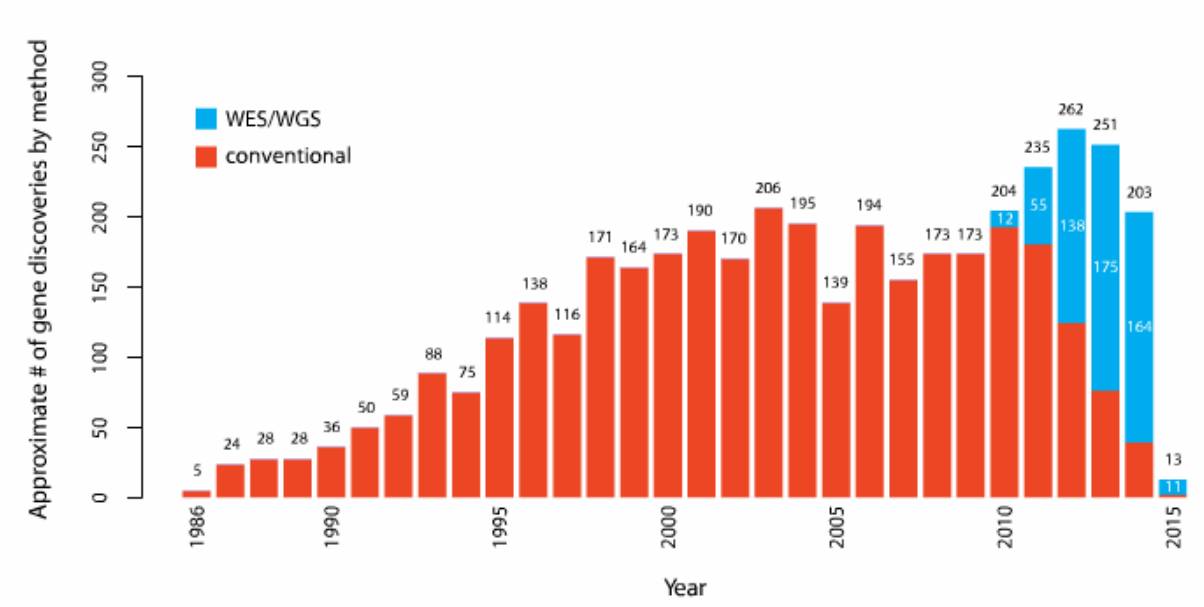


Figure 2-3: Approximate number of gene discoveries made by WES and WGS versus conventional approaches. Source: figure 4 of (Chong et al., 2015).

2.4.2 Whole Exome Sequencing (WES): an essential mean to study RD

Since their advent in 2005, NGS platforms have become widely available, reducing the cost of DNA sequencing by four orders of magnitude relative to the Sanger sequencing technique (Metzker, 2010). Specifically, WES relies on the development of methods coupling the design of probes capturing all the exonic regions of a genome (capture probes) and massively parallel DNA sequencing. This combination ensures a cost-effective access to nearly all the coding variation present in an individual human genome. Since its introduction, WES proved its efficiency by identifying genes responsible for Mendelian disorders in circumstances for which conventional approaches have failed (Ng et al., 2009; 2010a). In this section, our focus, after presenting the principle of WES, is to explain some of the experimental and analytical options for applying exome sequencing as a tool for disease gene discovery and to describe some of the key challenges in using this approach. We review how exome sequencing is being used to identify disease-causing genes.

2.4.2.1 Overview of the WES principle

The exome is defined traditionally as the sequence encompassing all exons of protein-coding genes in the genome, and in human, the WES covers between 1 and 2% of the

genome. Overall, ~85% of the Mendelian disease-causing variants thus far identified are located in protein-coding regions (Botstein and Risch, 2003).

As summarized in Figure 2-4, regardless of the NGS platform used, the sequencing pipeline starts with extraction of DNA of an individual, usually collected from white blood cells or epithelial cells from a saliva sample for humans. The DNA is then shredded into short fragments and amplified using polymerase chain reaction (PCR) or hybridization-based approaches. Classically, in the WES approach, all exons of protein-coding genes are amplified, but the regions that are amplified can be limited to a subset of genes (targeted approach) or extended to functional non-protein-coding elements (*e.g.* microRNA, long intergenic noncoding RNA, etc.) as well as specific candidate loci (Clark et al., 2011). The efficiency of capture probes varies considerably, and some sequences fail to be targeted by capture probe design. As a whole, these limits can represent a substantial fraction of the exome (~5–10%, depending on the kit), which might be poorly covered or missed.

Compared to traditional Sanger sequencing, the WES readouts, often simply referred to as "reads", generated by the platforms are considerably shorter (*e.g.* 75 nucleotides for the SOLiD platform; 150 nucleotides for the Illumina platform; 500 nucleotides for 454 pyrosequencing platform) and contain more sequencing errors. Moreover, each platform introduces sequencing errors that are characteristic for its sequencing workflow. Hence, compared to Sanger sequencing, NGS produces much more sequences, but of much shorter length and inferior quality; this has a tremendous impact on how the resulting readouts have to be processed in a downstream analysis. To address these caveats, the bioinformatics pipeline processing the data comprises five major steps (numbered 1 to 5 in Figure 2-4):

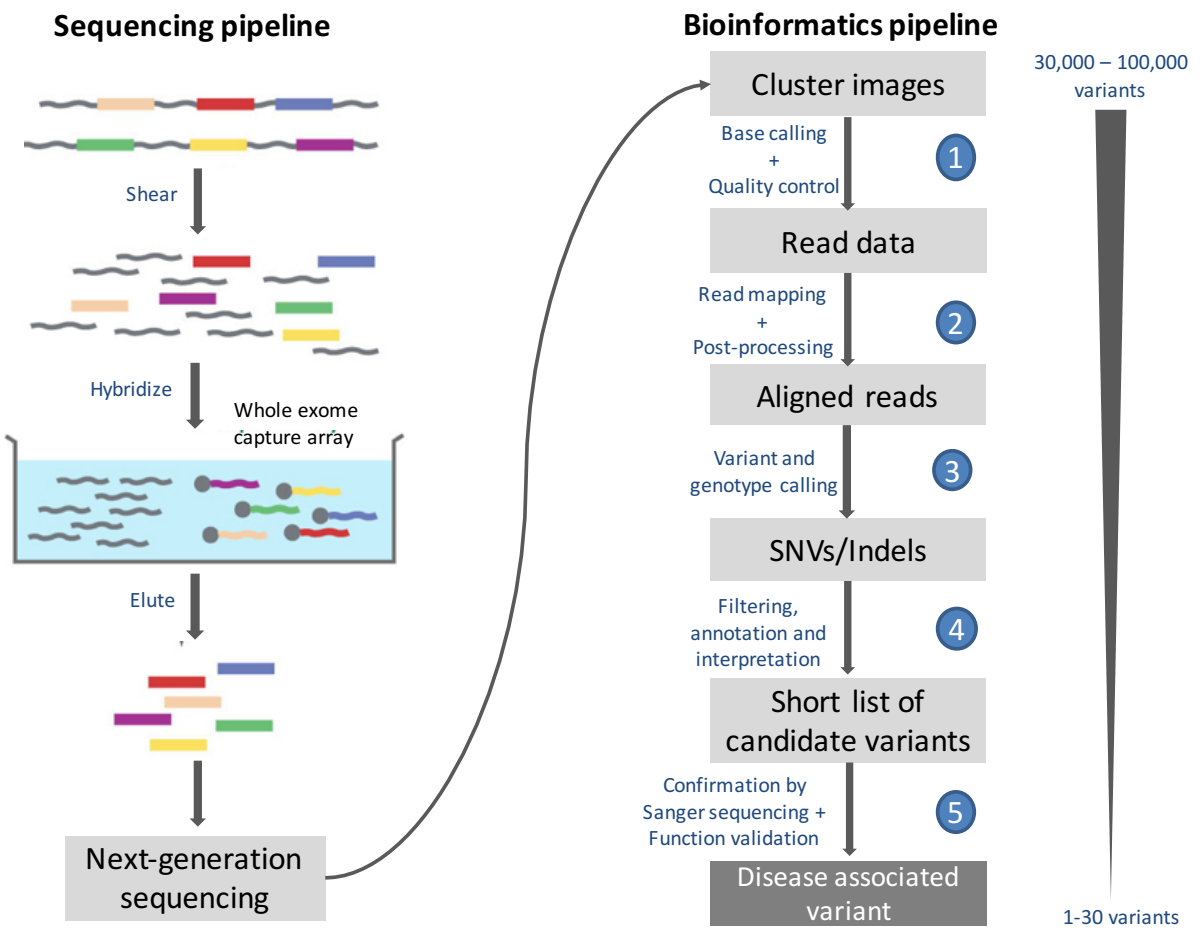


Figure 2-4: Principle of the Whole Exome Sequencing analysis.

1. Base calling and Quality Control of reads

The amplified products obtained after the PCR amplification are sequenced by one of the various types of sequencing technologies (e.g. Illumina's sequencing by synthesis, Life Technologies' sequencing by ligation) to generate millions of short sequence reads. The first step of the pipeline, which also referred to as the acquisition step or base-calling step, involves the detection and evaluation of the signals corresponding to sequenced reads. The signals detected are then processed to obtain the corresponding nucleotide sequence in a read. After the acquisition step, the read sequences are converted into a numeric format (FASTQ)(Cock et al., 2010) , in the form of strings of data representing the order of the DNA nucleotides. Each sequencing platform introduces specific errors corresponding to the underlying sequencing methodology, and each manufacturer then applies a base calling algorithm with a statistical model for the error estimation, expressed as a Phred-like quality score (probability of a missed called base).

Checking the quality of the generated sequence data is an essential step to verify that the sequencing was performed correctly, without any bias. The distribution of the quality scores at each sequence position is one of the most interesting quality parameters for the overall quality of the run.

2. Mapping

In order to interpret the millions of reads generated, the next step consists of aligning the reads to a reference genome/assembly (latest version: GRCh38 for human genome). There are currently several algorithms (Hatem et al., 2013) available for this procedure, and the most used is the Burrows-Wheeler transform (BWT) algorithm (Li and Durbin, 2009) for an efficient data compression. In general, the choice of the alignment tool and the corresponding settings significantly affect the outcome. This is especially true for SNV calling, as wrongly aligned reads may result in artificial deviations from the reference. These deviations in turn may falsely be classified as SNV in the downstream processing.

Moreover, prior to the downstream analysis, a post-processing step is usually required to format the data. This data wrangling step usually corresponds to procedures such as the sorting of the aligned reads with respect to their chromosomal position; the removal of PCR artefacts introduced by the amplifying library and adapters used; the removal of reads with multiple matching alignments; the realigning of reads around small InDel. This step is usually carried out with tools like SAMtools (Li et al., 2009).

3. Variant and genotype calling

The principle of the variant and genotype calling step consists in the identification of nucleotide sites, in one or more individuals, which display possible genomic variations from the reference genome, and the determination of the set of alleles present at that locus. It is usually accompanied by an estimate of variant frequency and some measure of confidence. There are several approaches that have been developed (Siepel et al., 2005) and modern variant calling tools, such as GATK (McKenna et al., 2010), are based on a statistical framework that integrates several information types, such as the abundance of high-quality nucleotides at a single site, or the joint variant calling in multiple individuals at the same time. The resulting SNV or InDel are usually represented in the commonly used Variant Call Format (VCF) [8]. The VCF format records for each identified variant, basic information such as the chromosomal position, the reference base, and the identified alternative base.

Furthermore, information on the quality of the SNV call as well as the amount of sequence data available for the call is stored.

4. Filtering, annotation and interpretation of variant candidates

Finally, the filtering, annotation and interpretation step is essential in order to obtain a short list of candidate variants to be validated experimentally. Filtering is essential in order to reduce the number of false-positive variant calls. Typically, applied filters check for deviations from the Hardy–Weinberg equilibrium (HWE), minimum and maximum read depth, adjacency to InDel, strand bias, etc. Several tools, such as GATK, SAMtools (*via* the script 'vcfutils.pl') and VCFtools, offer several functionalities to manipulate VCF files, such as merging of multiple files and extraction of variants in defined regions. In particular, GATK provides 'best practice' settings for the variant calling pipeline, including SNV candidate filtering. Usually, some of the filtering criteria in the best practices are the hard filtering of variants with a quality score below 30, or a read depth below 5, or SNV within a homopolymer of length 6 and more are discarded.

Next, subsequent annotation of the variants is necessary to further evaluate their potential biological effect and publicly available software and databases can be used for this purpose. At this stage, the variants are evaluated with additional annotating information to interpret the data in the context of the sequencing study. With NGS, now the bottleneck step is no longer the generation of the data but more the downstream analysis phase.

Classically, due to the intrinsic variability of the genome, in healthy or affected individuals, WES methods identify roughly 20,000–25,000 single nucleotide polymorphisms and several thousand base insertions or deletions. Thus, the challenge is to determine which, if any, of these mutations is responsible for a disease. Discrete filtering and prioritization of variants are therefore crucial to the disease-gene identification process (Bamshad et al., 2011). Common filtering criteria in variant analysis protocols are:

i. <u>Allele frequencies</u>

With the increasing number of referenced alleles submitted to the public database of single nucleotide polymorphisms (dbSNP) (Sherry et al., 2001), more than 95% of the 20,000-25,000 variants identified in a WES, have already been reported as being polymorphic, *i.e.* a common allele found in general human populations. In general, variants with a minor allele frequency (MAF) in a control population that is greater than the expected prevalence for the disorder is considered to be a strong support for interpretation as a benign

effect (*e.g.* MAF >5%). For disorders which are fully penetrant at an early age, variants with a MAF < 1% in population databases are considered as potential good candidates for further screening.

ii. Loss-of-function variants (LoF)

Some sequence variations have major deleterious effects on their corresponding protein products. These LoF variants include: i) large deletions removing either the first exon or more than 50% of the protein-coding sequence of the affected transcript, ii) nonsense SNV and iii) small insertion/deletion (InDel) variants disrupting the reading frame of a transcript, iv) mutations that disrupt canonical splice sites and v) missense SNV disrupting the protein structure or a binding site.

The deleteriousness of the variants are assessed on several biological criteria/rules, such as the biochemical nature of the change, or the location of the altered position relative to functionally and/or structurally important domains of the protein. The degree of evolutionary conservation (and therefore, intolerance to change) of an amino acid is also an important predictor of the likelihood of clinical significance of a particular substitution. Several software tools for the *in silico* determination of potential damaging properties have been developed, for examples CADD (Kircher et al., 2014), SIFT (Kumar et al., 2009), PolyPhen-2 (Adzhubei et al., 2013) and GERP++ (Davydov et al., 2010), KD4v (Luu et al., 2012), KD4i (Bermejo-Das-Neves et al., 2014).

Interpreting these different annotations and prediction scores remains a challenging task when evaluating the deleteriousness of a variant. Indeed, various genome-sequencing studies indicate that healthy human genomes can carry many genetic variants predicted to cause LOF, suggesting an unexpected tolerance for gene inactivation. It was estimated that human genomes typically contain around 100 genuine LoF variants with approximately 20 genes completely inactivated (MacArthur et al., 2012). Only a small proportion of these "common" LoF variants are likely to represent severe, recessive disease genes. Others may affect nonessential, redundant genes or other genes with smaller effects on human phenotypic variation and possibly disease risk.

iii. Using pedigree information

For Mendelian phenotypes, the use of pedigree information can substantially narrow the genomic search space for candidate causal alleles. For example, two first cousins share a rare allele that is identical-by-descent in approximately one eighth of the genome. Hence, by

sequencing the two most distantly related individuals with the phenotype of interest, we can substantially restrict the genomic search space.

Moreover, in the case of *de novo* mutations, WES of parents–child trios are a highly effective approach to identify the causative mutations. Although many apparent *de novo* mutations in fact result from sequencing artefacts, those that remain after Sanger validation are highly likely to be functionally relevant. This approach can reveal autosomal dominant diseases amongst individuals with apparently sporadic disease.

iv. Stratification based on function

Candidate genes can be stratified by existing biological information: for example, the predicted role in a biological pathway or interactions with genes or proteins that are known to cause a similar phenotype. This often points directly to the correct gene amongst those filtered out from the sequencing data and facilitates validation of genetic findings by functional, biochemical means.

5. Validation of short-list of candidate genes

Finally, additional wet lab evidence has to be gathered to ascertain the clinical significance of a variant in relation to a given disease. In brief, the assessments try to answer three main questions: (i) Is the variant a truncating variant, which alters/impairs the function of the gene? (ii) Does the gene impairment result in disease phenotype? (iii) Is the associated disease/phenotype relevant for the specific clinical condition of the patient?

The wet lab evidence is gathered at multiple levels. First at the variant and gene level, Sanger sequencing of the mutated region of a patient's cell sample is performed to confirm that the variant is not an artefact and that it segregates with the disease in affected family member sequences. To confirm that the variant impacts the gene either at the transcriptional level or at the translational level, additional wet lab validations include the observation or not of full-length RNA and protein sequences, the correct dosage and localization of the corresponding protein.

At the protein or functional network level, further functional assays are carried out *in vitro,* to confirm that the gene impairment reflects disease-associated cellular mechanisms in relevant cell-lines or dedicated cell models and *in vivo*, to confirm in relevant animal models that the expressed phenotypes correspond to the human disease.

2.5 Ciliopathies: RD associated with ciliary dysfunctions

In this section, I will discuss the case of an emerging class of RD linked to the dysfunction of the cilia: the ciliopathies, which are emblematic of the resurging interest for RD studies. As we will see, this interest is now even supported by several national and international initiatives. Thus, before the description and classification of ciliopathies, the different cilia structures and functions will be described.

2.5.1 *Phylogenetic Distribution and Structure of the Cilia*

The cilium is a specialized organelle, found in all the major eukaryotic clades and was probably present in the last common ancestor of eukaryotes (Cavalier-Smith, 2014). The cilium was also lost several times during evolution and is hence absent in angiosperms, a majority of fungi, amoeba and in some of the stramenopiles (Figure 2-5). In metazoa, cilia are present at the surface of most of the eukaryotic quiescent cells (in G0/stationary phase of the mitotic cell cycle) (Wheatley, 1995) [http://www.bowserlab.org/primarycilia/cilialist.html], with the most important exceptions being the bone marrow–derived cells and the intercalated cells of the kidney collecting duct (Praetorius and Spring, 2005).

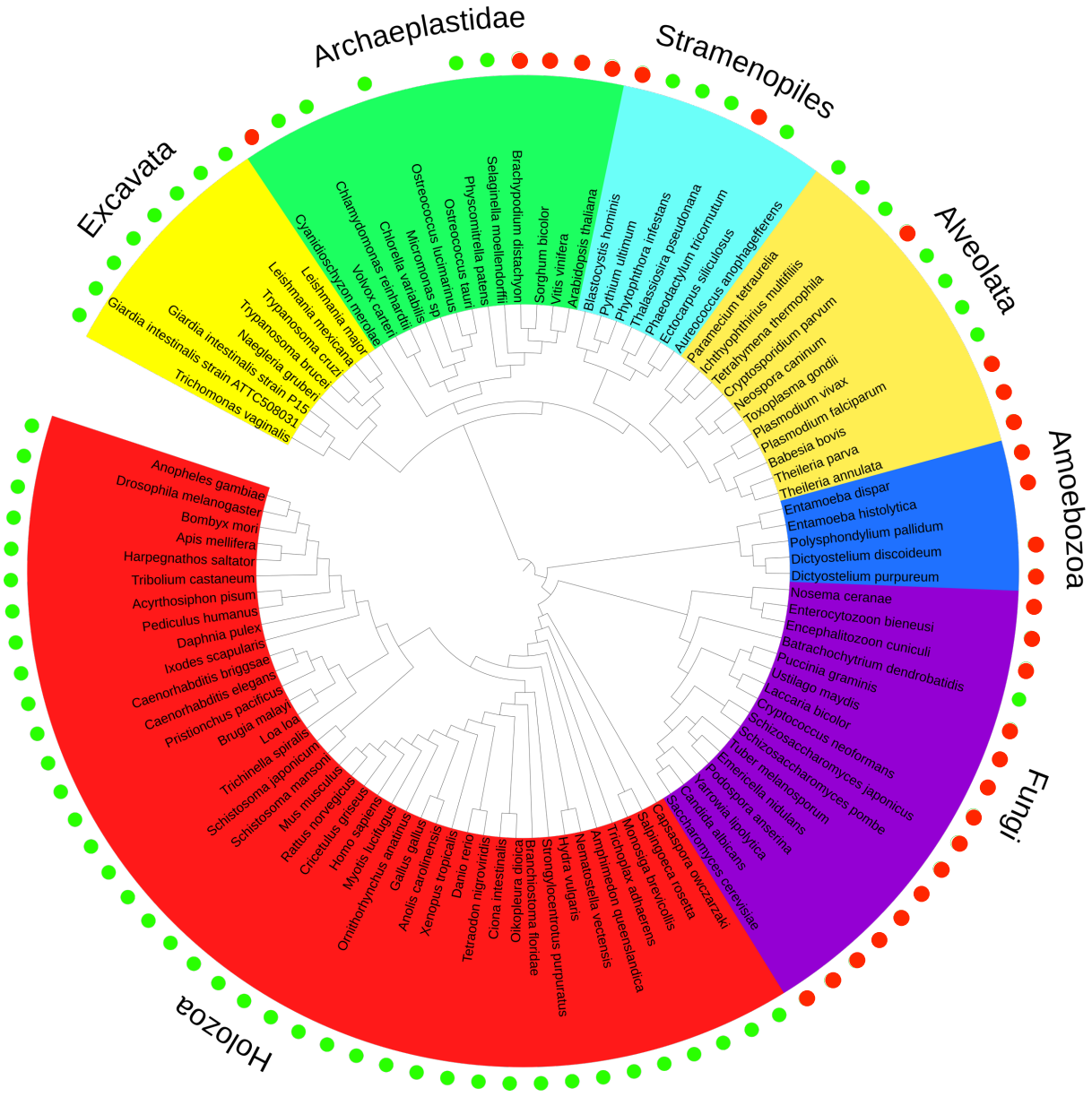
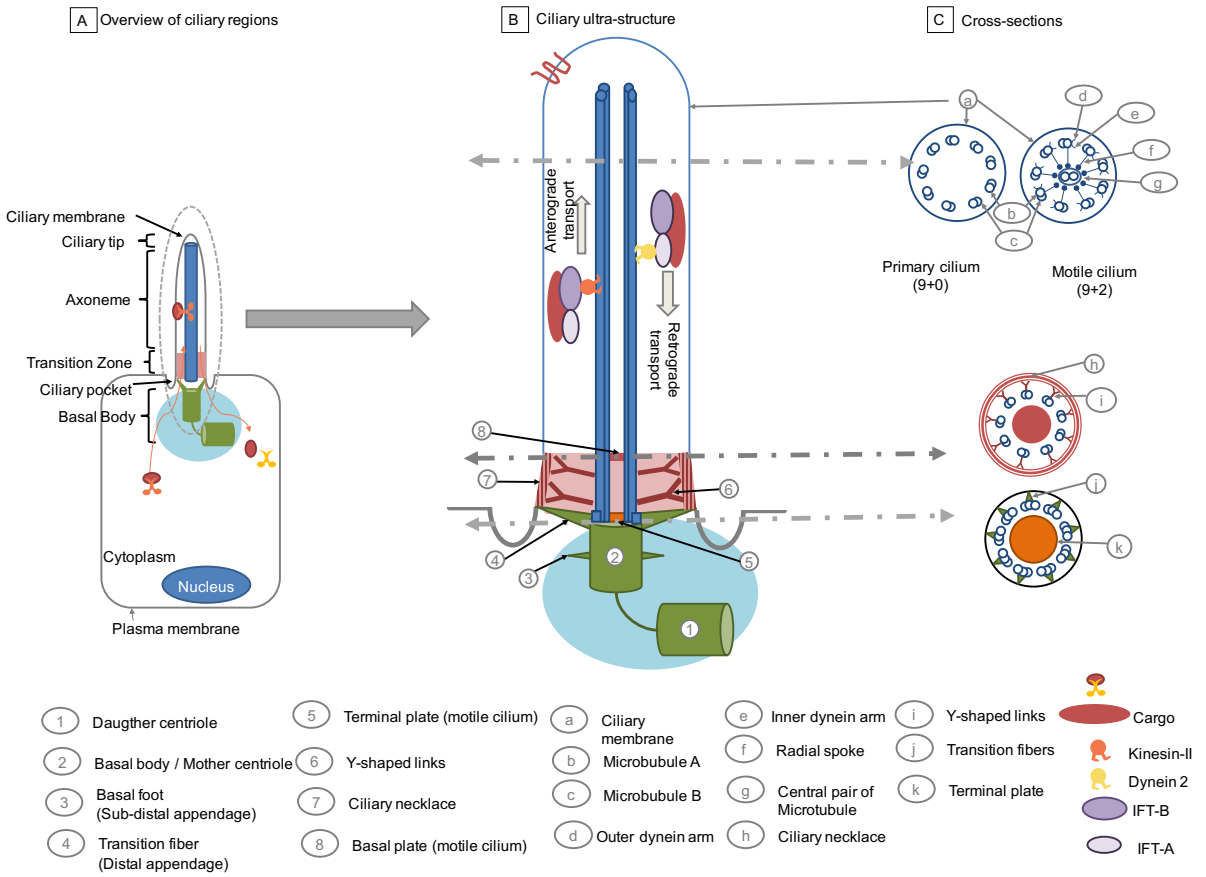


Figure 2-5: Distribution of ciliated organisms across the tree of life.The green dot indicates ciliated organisms; the red dot indicates non-ciliated organisms. (Source: internal communication from Yannis Nevers)

The cilium consists of a polarized microtubule-based structure (the axoneme) covered by a specialized plasma membrane, which emanates from the basal body and extends from the cell surface into extracellular space (1-10 µm) (2-6) (Hoyer-Fender, 2010). From the basal body to the ciliary tip, the cilia have a "9-symmetry" backbone microtubule-based structure, and 4 regions have been identified:



2-6: Ultra-structure of the cilia. A - A cell with a primary cilium. B – Ultra-structure of a typical primary cilium. C- Three cross-sectional views of a primary cilium: at the base, in the transition zone and at the tip.

(i) The basal body

The basal body is a derived modification of the mother centriole (microtubule-organizing centre), which consists of nine triplets of microtubules (A, B and C) (Vertii et al., 2016). In its distal portion, the basal body has transition fibres that link the outermost microtubule (microtubule C) to the apical plasma membrane of the cell body.

(ii) The transition zone

The proximal region of the cilium, ranging from the transition fibres of the basal body to the start of the axoneme, defines a region referred to as the transition zone (TZ)(Gilula and Satir, 1972; Omran, 2010). The TZ comprises at its basal end triplets of MTs and is devoid of central microtubules. The tips of transition fibres, emerging from the basal body, are thought to anchor microtubules to the plasma membrane, although the composition of transition fibres is still largely unknown (Graser et al., 2007; Ishikawa et al., 2005; Kilburn et al., 2007). Other substructures found in the TZ are the so-called "Y-shaped" linkers and the ciliary necklace, characteristic structure of the transition zone. The "Y-shaped" linkers are structures connecting the outer doublets of microtubules to the plasma membrane and the ciliary

necklace. The detailed protein of the Y-shaped linkers structure is as yet uncharacterized and their shape and name is species specific (Gilula and Satir, 1972). The ciliary necklace is a specialized structure that consists of several parallel strands of intramembrane particles and their number is species and cell-specific. The identity of these strands is unknown, but they encircle the ciliary membrane spacing from the plasma/ciliary membrane boundary to the basal plate (Gilula and Satir, 1972). Y-shaped linkers and the ciliary necklace are especially visible in the elongated transition zone structure of connecting cilia in photoreceptors (Fisch and Dupuis-Williams, 2011).

Although observed for many years in TEM (transmission electron microscopy) cross-sections, the function of the TZ remained mysterious, until recent mutagenesis studies suggested a role of "ciliary gate". Mutagenesis studies suggest that proteins present in this region participate in the control of molecules going in and out of the cilia (Basiri et al., 2014; Williams et al., 2011).

(iii) The axoneme

The axoneme is composed mainly of nine peripheral doublets of microtubules (MT). Each peripheral doublet consists of MT A and MT B, built of 13 and 11 protofilaments, respectively; protofilaments are composed of heterodimers of α and β tubulin. Based on the organization of the MT doublets of the axoneme, cilia were initially classified into two main categories: motile and immotile/primary.

Motile cilia comprise of nine peripheral double of MT and a central pair (CP) of MT. The MT A of axoneme comprises an outer (ODA) and inner dynein arms (IDA), which can generate the force needed for motility in various ATP-dependent processes. The peripheral doublets of MT are bound to the CP through radial spokes. In human, this organization of the axoneme, referred to as 9+2 cilia, is found on the apical surface of epithelial cells in the airways (respiratory cilia), female reproductive system (cilia in fallopian tube) or in male reproductive system (sperm flagella).

Immotile/primary cilia possess only the nine peripheral double of MT without the CP nor ODA-IDA or radial spokes. In human, this organization of the immotile cilia axoneme, referred to as 9+0 cilia is present for instance in kidney (renal cilia), bile duct (cholangiocyte cilia), pancreas (cilia in pancreatic duct), bone or cartilage (cilia in osteocyte or chondrocyte) or in the eye (photoreceptor connecting cilia).

However, recent findings suggest a more blurred classification of cilia (Ferkol and Leigh, 2012), which includes exceptions: motile cilia with the 9+0 pattern (they lack CP but still contain ODAs and IDAs) are found in the embryo and are referred to as nodal cilia;

immotile cilia with the 9+2 pattern are found in the inner ear and are named kinocilia/stereocilia (Falk et al., 2015).

(iv) The ciliary tip

The ciliary tip, which corresponds to the distal part of the cilia, is a specialized region involved in the biological functions of the cilia, such as in signalling and sensory function (Gluenz et al., 2010; Goetz and Anderson, 2010). Uptil now, little was known about the ultrastructure or composition of the ciliary tip, but now several work are focusing to characterize these proteins (Chávez et al., 2015; Gluenz et al., 2010). However, it has been shown to be an important centre for the addition or removal of axonemal MT during ciliogenesis or the resorption of the cilia (Marshall and Rosenbaum, 2001). Since no protein synthesis takes place in the cilium, all the precursors and other essential biomolecules for its biogenesis, maintenance, and resorption have to be conveyed and returned to the cytoplasm. This ciliary trafficking, also referred to as IntraFlagellar Transport (IFT), is performed by cargo proteins along microtubules using molecular motors: kinesin for anterograde transport (IFT-B) and dynein for retrograde transport (IFT-A) (Perrone et al., 2003; Snow et al., 2004; Sung and Leroux, 2013).

Additionally, the ciliary tip is also thought to be a specialized compartment of the cilia, which acts as an integrating centre of intra- and extra-cellular signals, by concentrating several receptors and signalling biomolecules (Satish Tammana et al., 2013). Numerous findings showed the implication of the ciliary tip as an integrating centre of intra- and extra-cellular signals in several pathways, such as the recruitment of transcription activators of the Gli family, regulator of Sonic hedgehog pathway, or Wnt pathway (Goetz and Anderson, 2010; Ko, 2012; Liem et al., 2009; Tukachinsky et al., 2010).

2.5.2 *Function of the cilia*

Historically, motile cilia have been studied mostly for their motile function (Figure 2-7). Motile cilia are present on epithelial cells of the trachea, on ependymal cells in the brain, and on cells lining the oviduct and epididymis of the reproductive tracts. Normally concentrated in large numbers on the cell surface, motile cilia beat in an orchestrated wavelike fashion and are involved in fluid and cell movement such as mucociliary clearance in the lung, cerebrospinal fluid movement in the brain, and ovum and sperm transport along the respective reproductive tracts (Yuan and Sun, 2013). Primary cilia are solitary organelles projecting from the surface of cells. These cilia function primarily as either photo- or chemo-

or mechano-sensors, such as in the detection of the fluid flow in kidney cells (Yuan and Sun, 2013) (Figure 2-7). Nodal cilia have been localized to the node in gastrulation-stage embryos. Recent studies have shown that nodal cilia play essential roles in establishing signalling events required for specification of the left-right body axis in mammals (Bonnafe et al., 2004; Murcia et al., 2000; Watanabe et al., 2003) (Figure 2-7).

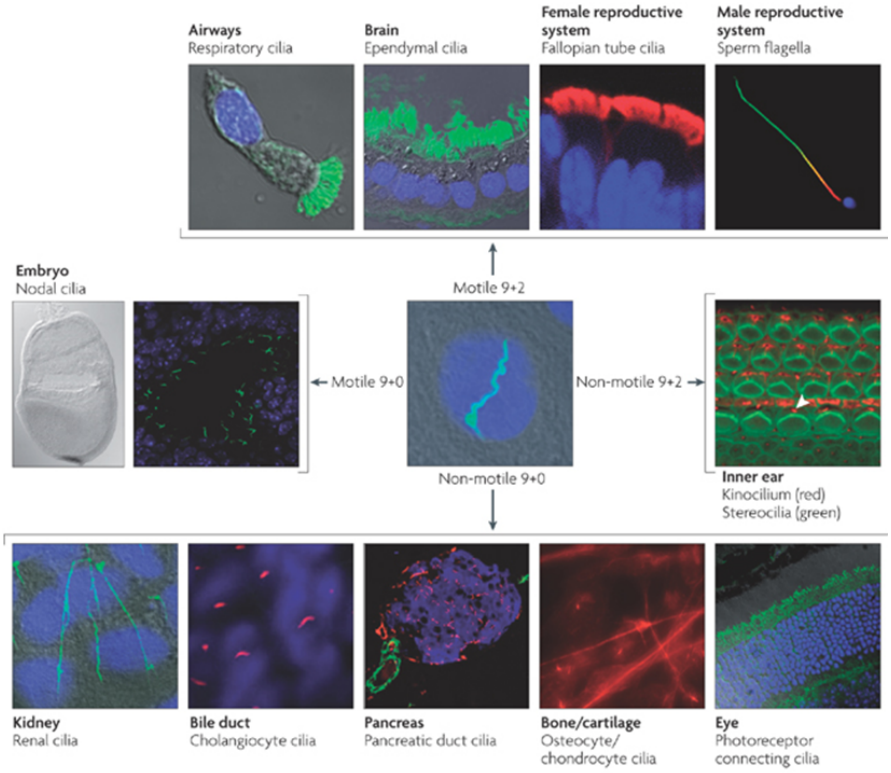


Figure 2-7: Distribution of cilia based on their motile or immotile property.

2.5.3 *Ciliopathies*

The dysfunction of the cilia results in a group of severe clinical manifestations nowadays referred to as ciliopathies. Although ciliopathies are individually rare disorders, an amazing spectrum of what were previously disparate syndromes is now recognized as part of the ciliopathy spectrum(Hildebrandt et al., 2011). Given the ubiquity of primary cilia and variety of specific functions performed by these organelles, it is not surprising that their dysfunction causes multisystem disorders with a plethora of overlapping distinct phenotypes (Figure 2-8) (Hildebrandt et al., 2011).

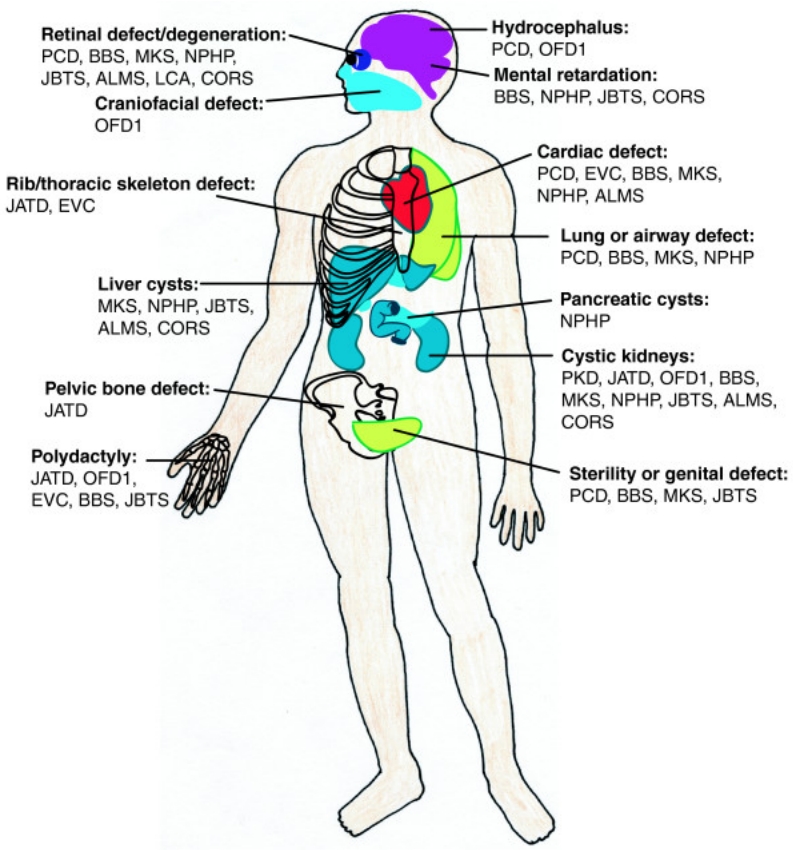


Figure 2-8: Organs affected in human ciliopathies. Numerous pleiotropic human disorders have been attributed to defects in cilia formation. Most ciliopathies have overlapping clinical features in multiple organs. Cystic kidney and retinal defects are frequently observed. Skeletal dysplasia is predominantly seen in JATD, OFD1 and EVC. ALMS, Alström syndrome; BBS, Bardet-Biedl syndrome; CORS, cerebello-oculo-renal syndrome; EVC, Ellis-van Creveld syndrome; JATD, Jeune asphyxiating thoracic dystrophy; JBTS, Joubert syndrome; LCA, Leber congenital amaurosis; MKS, Meckel syndrome; NPHP, nephronophthisis; OFD1, oral-facial-digital syndrome type 1; PCD, primary ciliary dyskinesia; PKD, polycystic kidney disease. (Source: Fig 5 from (Goetz and Anderson, 2010)).

Ciliopathies can be subdivided into 2 main categories: 'motile ciliopathies' and 'non-motile ciliopathies'. Motile ciliopathies comprise a class of disorders displaying prominent *situs inversus* (a condition in which the normal positions of organs are reversed) and recurrent respiratory tract infections, reflecting the abnormal beating of the cilia responsible for mucociliary clearance, and fertility disorders.

Immotile ciliopathies show major but mixed features in several vital organs, including the brain, kidney and liver, and others, such as the eye and digit. These ciliopathies range from largely organ-specific disorders, such as polycystic kidney disease (PKD), to pleiotropic disorders, such as the iconic Bardet-Biedl syndrome (BBS), Jeune asphyxiating thoracic dystrophy (JATD) and Meckel-Gruber syndrome (MKS) (Figure 2-8).

The clinical diagnosis of a particular ciliopathy is based on the presence of a given combination of clinical signs (Baker and Beales, 2009) (Figure 2-9:). Moreover, as illustrated

in Figure 2-9, several ciliopathies are characterized by important overlapping phenotypic features (*e.g.* Joubert, Meckel-Gruber, Jeune), and the differential diagnosis is carried out by considering the combination of additional minor phenotypic features.

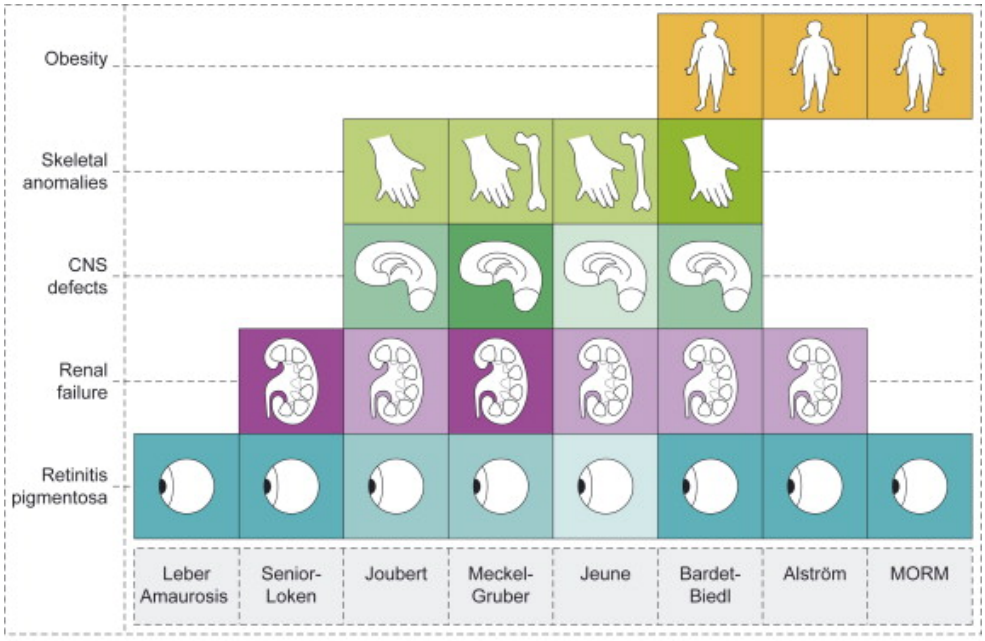


Figure 2-9: The spectrum of the most studied ciliopathies with an ocular affection. Within the broad range of ciliopathies, the number of affected organs varies. To illustrate this variability, here is a schematic representation of the clinical manifestations most often observed (RP, renal failure, central nervous system defects, skeletal anomalies and obesity) and their occurrence in different ciliopathies (Leber congenital amaurosis, SLS, Joubert, Meckel–Gruber, Jeune, Bardet–Biedl, Alström and MORM syndromes). Ciliopathy phenotypes include, on the one hand, isolated retinal dysfunction in LCA and, on the other hand, a multi-organ syndrome in BBS. CNS: central nervous system (Source: Fig 5 of (Mockel et al., 2011)).

Moreover, the ciliopathies are characterized by a major genetic heterogeneity: each clinical syndrome can be caused by bi-allelic mutations in different genes, and mutations of a given gene can cause different clinical profiles of ciliopathies. In parallel with the phenotypic overlap between different ciliopathies, there is also a strong genetic overlap with marked allelism between various ciliopathies. This is particularly well illustrated by the Joubert and Meckel-Gruber syndromes whose genetic overlap is such that these two syndromes may, in fact, be considered as the two extremes of the same pathology (Mougou-Zerelli et al., 2009). Despite this genetic complexity, there is some consolidation in functional module gene products associated with a given ciliopathy. For example, many of the genes causing Joubert or Meckel syndromes encode proteins associated in multi-protein complexes at the transition zone (Sang et al., 2011) and several genes involved in Bardet-Biedl Syndrome encode proteins forming a distinct multi-protein complex named BBSome (Nachury et al., 2007). But

beyond the physical interactions between multiple proteins associated with the same ciliopathy, there are also physical interactions between the functional modules of different components, and the network of protein interactions underlying the ciliopathies (van Reeuwijk et al., 2011).

Finally, as mentioned earlier, the ciliopathies are a recent emerging class of RD since 2003, and are emblematic of the resurging interest for RD studies, with even the creation of a dedicated journal: the BioMed Central Cilia journal [9]. Several international initiatives have already been constituted to decipher cilia and ciliopathies, such as the molecular dissection of the ciliary gate project funded by the NIH [10]; the SYSCILIA project funded by an EU FP7 grant, which aims to apply system biology approaches to model the ciliary system in order to study the associated biological processes and diseases [11]. Additionally, some ciliopathies are already being applied as models to provide insights into common diseases like the emblematic BBS, which has been shown to be a remarkable model for understanding the mechanisms leading to obesity. In the current context of translational medicine and system biology, the study of obesity in a BBS context offers unprecedented means to dissect obesity and new hope to identify novel mechanisms and targets to better fight this modern pandemic (see article in Annexe I).

II. Material and Methods

During my thesis, I developed several bioinformatics resources ranging from an exome variant analysis tool to the design of a knowledge base dedicated to the study of ciliopathies. Chapter 3 presents the common selected datasets agglomerated, and the integration procedures I used for all these applications. I will then present in Chapter 4, the general bioinformatics tools, algorithms, and computational infrastructure used to process them. The original developments and specific uses of these datasets and tools will be further detailed in the Results and Discussion section.

Chapter 3 Datasets

In the context of my different developments (namely VarScrut, Pub*Athena* and CilioPath), I rely on several heterogeneous, asynchronous and disperse datasets. To rationalize and optimize the developments, I set up a common framework, developed during this thesis (referred to as BioData Toolkit; BDT), which agglomerates different datasets and integrates them into a multi-level knowledge framework, through different data-fusion procedures (Figure 3-1).

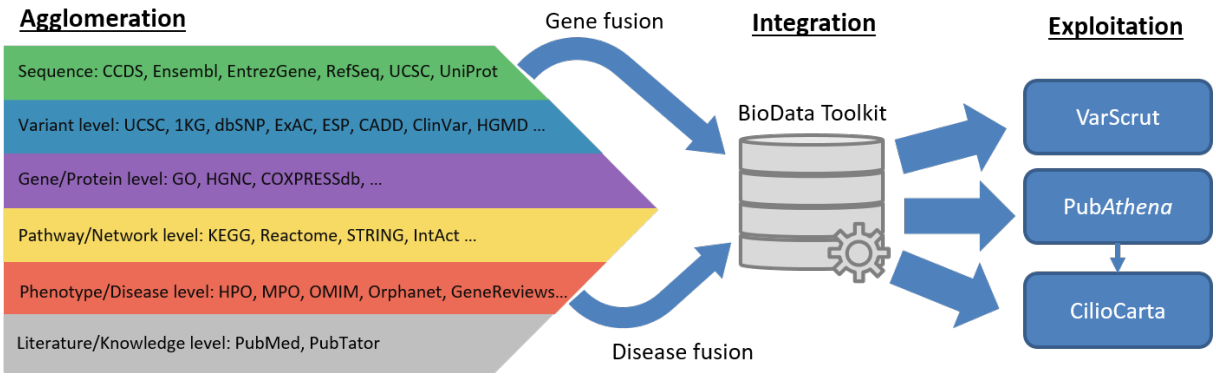


Figure 3-1: Overview of the datasets agglomeration and integration in BioData Toolkit for exploitation.

Data in BioData Toolkit (BDT) are organized according to a star schema, whereby all the information is integrated in a gene-centric or disease-centric (when no disease-causing gene is known) manner. Since each type of dataset has its own sequence reference identifiers (*e.g.* OMIM uses RefSeq; GO uses UniProt…), in order to have a coherent integration, the Ensembl dataset is chosen as the reference sequence dataset. The merging of sequence identifiers (see section 3.1) onto Ensembl was performed according to the rule of reciprocal

cross-links among databases, and a confidence coefficient of 1 is attributed. In case of inconsistencies, for protein coding-genes, the CCDS dataset was used to guide the merging of identifiers (see section 3.1.4). As for the other categories of genes, the merging of identifiers was performed only if Ensembl provides a cross-link to another sequence identifier, and a confidence coefficient of 0.5 is attributed. All the different sources of information (see Figure 3-1) were then cross-referenced to the corresponding Ensembl gene id.

Additionally, for disease datasets, a merging of identifiers was also performed on the reciprocal cross-links principle. For diseases, the merging process was performed in a hierarchical manner starting with the OMIM dataset (see section 3.5.3), then a mapping with GeneReviews (see section 3.5.5) and finally, with Orphanet (see section 3.5.4). At the end of the merging process, a new internal identifier was created for every identifier or singleton, and then the linked phenotype from HPO was inferred to the new disease identifier.

The whole integration process of BDT is performed every month, triggered by updates retrieved by the Jenkins server (see section 4.4.44.4.4).

3.1 Sequence datasets

Reference sequence datasets such as Ensembl, UCSC (University of California Santa Cruz), RefSeq, EntrezGene, CCDS (Consensus CDS database) and UniProt are primary sources of data that provide entry points for all kinds of information: from sequence to associated functional annotations or curated literature references. Moreover, they are also used as references for cross-links to specialized databases.

3.1.1 *Ensembl*

The Ensembl project (Yates et al., 2016), hosted by the European Bioinformatics Institute (EBI), is a system providing the most up-to-date genomic information for key model organisms in numerous kingdoms. Ensembl provides various well-documented information, from curated gene definitions and functional annotations *via* Vega (Wilming et al., 2008) up to different OMICS data. In addition, Ensembl provides several means and tools to access and query this data, such as public MySQL databases, an API web service, or an FTP repository. Finally, with a 3-month release cycle, Ensembl creates a dedicated database for every species with the most updated annotations.

At every release, the database corresponding to the build GRCh37 for *Homo sapiens* is selected for download through the dedicated MySQL database access [http://www.ensembl.org/info/data/mysql.html] to retrieve:

(i) The 'core' schema that contains all the genome features (*e.g.* chromosomes, genes, proteins, RNAs…) and associated annotations of a given species,

(ii) The 'variation' schema that contains all referenced variation events across several public resources. For instance, the release 83 of the variation database for Human ('homo_sapiens_variations_80_37') references more than 152 million variations across 23 data sources.

3.1.2 *UCSC Genome Browser*

The University of California Santa Cruz (UCSC) Genome Browser is a collection of tools for the visualization and analysis of large-datasets. The UCSC genome browser has an integrated view to display various genome related features in a single window in the form of independent tracks. Among the other tools in the UCSC Genome Browser collection, there is a set of popular ones for the manipulation of large genomic datasets, such as BLAT tool for the alignment of sequences (DNA, RNA, or protein) against a genome, or LiftOver for the mapping of genomic positions from one genome to another *via* sequence homology.

The UCSC Genome Browser is also known as a genomics reference data hub (from genomic sequence to annotation features) for over 70 model organisms. For instance, it is the primary source to obtain the reference assemblies of the human genome and its related genomic annotations, such as data from the ENCODE project (Rosenbloom et al., 2013). The data hosted by the UCSC are either accessible on its FTP repository or *via* its public MySQL database.

BDT includes all the UCSC tables mapping to other major gene repositories (*e.g.* 'ensGene' for Ensembl, 'refGene' for NCBI's RefSeq…) [https://genome.ucsc.edu/goldenpath/help/mysql.html]. Additionally, to evaluate the effective impact of a variant in the scope of the development of VarScrut (see Chapter 5), conservation constraint data hosted on the UCSC FTP [ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/] is retrieved, namely those for phyloP, GERP, and PhastCons (see section 3.2.2).

3.1.3 *RefSeq and EntrezGene*

RefSeq (Pruitt et al., 2014) and EntrezGene (Maglott et al., 2011) are two reference sequence databases of the National Center for Biotechnology Information (NCBI).

Refseq, for Reference Sequence database, is a non-redundant collection of sequences derived from GeneBank (Benson et al., 2013). Its objective is to provide a synthetic, curated, annotated, and updated source of sequence information for functional studies. It contains different types of sequences, from the genome to proteins for any given species.

EntrezGene is a gene-specific database derived from RefSeq, only for species with a complete sequenced genome. EntrezGene is a database with unique and stable identifiers, hence representing a stable source of information for various specialized databases such as OMIM (see section 3.5.3) or GeneReviews (see section 3.5.5). Besides the access to EntrezGene's records *via* the NCBI's E-utilities (see section 3.6.1), additional data are available in its FTP repository [ftp://ftp.ncbi.nih.gov/gene/DATA] with daily updates. BDT includes the files:

(i) geneinfo – used for the mapping on gene symbols,

(ii) gene2ensembl – used for the mapping on Ensembl gene identifiers,

(iii) gene2pubmed – used for the mapping of EntrezGene identifiers onto PubMed articles. This file is used when evaluating the performance of the different text-mining tools, such as PubTator (see section 3.6.2).

3.1.4 *Consensus CDS database*

Each of the gene sequence databases available uses its own method for gene annotations. The Consensus CDS database (CCDS) (Farrell et al., 2014), is the first international collaborative initiative which strives for convergence towards a standard set of protein-coding gene annotations that are identical across the major reference sequence databases (EBI, UCSC, NCBI). The CCDS provides on its FTP server [ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/] a core set of protein-coding regions with high quality and consistent annotations for the 2 most studied organisms in biology: human and mouse. The release 15 of CCDS is integrated in BDT, which reported 29,045 curated CDS in Human and 23,093 curated CDS in mouse.

3.1.5 *UniProt*

The Universal Protein resource (UniProt) is the reference database for all information and annotations related to proteins (UniProt Consortium, 2015). UniProt is composed of several components:

(i) UniProt Knowledge Base (UniProtKB), which is the central hub collecting all the functional annotations on proteins.

(ii) The UniProt Reference Clusters (UniRef), which provides different sets of clustered proteins according to their different degree of identity (50%, 90%, and 100%).

(iii) UniProt Archive (UniParc), which is a comprehensive non-redundant database of most of the publicly available protein sequences.

UniProtKB is composed of two sections: (a) UniProtKB/SwissProt, which contains all the proteins records that have been manually curated from the literature by experts and (b) UniProtKb/TrEMBL, which contains the protein records that have only been obtained *via* a computational pipeline.

In BDT, UniProtKB is the major source of annotations for genes and proteins by using the cross links provided by Ensembl. UniProt's controlled vocabulary is also used as an additional source of annotation:

(i) Keywords [http://www.uniprot.org/keywords/] for literature citations.

(ii) Subcellular locations [http://www.uniprot.org/locations/] and disease [http://www.uniprot.org/diseases/] repositories.

3.2 Variant level information

To assess the nature and the status of a variant, different types of annotations are collected across different databases. The first part of this section presents lookup datasets used firstly, to verify if any given variant has already been reported as polymorphic in public population databases (Population datasets) or as pathogenic in a specialized database (Disease datasets) and secondly, in the different variant annotation datasets (Benchmark datasets) to train the VarScrut meta-predictor for nsSNV (see Chapter 5). The second part presents the tools (Variant annotation and predictors) used to train the VarScrut meta-predictor.

3.2.1 Repertories of polymorphic and disease datasets

Dataset type	Name	Purpose	Last update	PMID
Population	dbSNP	"Public archive for genetic variation within and across different species"	09/2015	11125122
Population	1,000 Genome	"A deep catalogue of Human Genetic Variation"	24/06/2014	20981092
Population	ExAC	"Harmonized processing of > 60,000 exome sequencing data"	13/01/2015	-
Disease	ClinVar	"Public archive of relationships among sequence variation and human phenotype."	09/2015	24234437
Disease	HGMD	"Comprehensive database on published human inherited disease mutations."	04/2013	20038494
Disease	SwissVar	Comprehensive collection of polymorphic and disease SNP in UniProt/SwissProt KB	09/2015	20106818
Training	filtHumVar	One of the training set of PolyPhen2 to distinguish common nsSNP (MAF >1%) variants and disease-causing mutations in UniProt	07/2010	20354512
Training	filtExoVar	Training dataset for the logit model to distinguish rare common variants (MAF <1% + at least 1 homozygous genotype in 1KG) from Mendelian disease causing variant in UniProt	06/2012	23341771

Table 3-1: List of the variant datasets used in VarScrut.

3.2.1.1 dbSNP (population dataset)

The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001) is the largest archive of genetic variations for different species, hosted at the NCBI. Initially, the aim of dbSNP was to collect all the polymorphic single nucleotide variants, but over the time, it turned into the repertoire containing all types of variants (SNV, InDel, microsatellites). The term 'Polymorphism' in the dbSNP name might be misleading, since the minor allele frequency in the general population and the associated phenotypic status (neutral or pathogenic) is not always documented. Another caveat is that dbSNP is suspected to contain a high rate of false positive variants (~15-17%) since not all candidate variants have been validated experimentally (Mitchell et al., 2004). Despite all these burdens, dbSNP remains the

central repository of all reported variants and provides reference identifiers (rs ids) that are used by all other databases.

When used with caution, dbSNP remains a useful complementary database in the initial evaluation step of a variant. The VCF version of build 144 of dbSNP from the dedicated FTP server [ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b146_GRCh37p13/] of the NCBI, containing over 85 million human variants is used in the implementation of VarScrut.

3.2.1.2 The 1,000 Genomes (population dataset)

The 1,000 Genome (1KG) (1000 Genomes Project Consortium et al., 2012) project was the first major international project to release the most detailed map of human genetic variants. The initial project collected over 38 million SNPs from 1,092 healthy individuals belonging to 14 populations. Currently, the 1KG dataset represents the largest and most unique set of genetics variants from healthy individuals, while other repositories of genetic variants frequently contain individuals with a non-healthy status (*e.g.* individuals with risk factors for diabetes and cardiomyopathies in NHLBI-GO Exome Sequencing Project; individuals with Schizophrenia or Inflammatory Bowel Disease in ExAC (see section 3.2.1.3).

The 1KG dataset phase 3 [ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/] is used in VarScrut as a lookup dataset to flag polymorphic variants with a MAF (Minor Allele Frequency) > 1%.

3.2.1.3 Exome Aggregation Consortium (population dataset)

The Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org/), is an international coalition of researchers initiated by Daniel MacArthur, which focuses on exome sequencing data [12]. This consortium has aggregated more than 60,000 human exomes corresponding to several previous exome-sequencing projects across the world (*e.g.* 1KG, Venter genome, Exome Variant Server…). Hence, this pooled cohort of exomes comes from healthy individuals as well as, for a none-negligible proportion, from affected individuals. The objective of this consortium is to reprocess all of the ~1 Petabytes of raw data in order to remove any source of artefacts resulting either from the sequencing technologies used, or from the analytical pipelines applied. ExAC provides unprecedented resolution for the analysis of genetic variants at very low frequencies in the general population.

The release 0.3 of VCF file has been downloaded, which contains over 10 million unique variants from its FTP server [ftp://ftp.broadinstitute.org/pub/ExAC_release/current]. The ExAC dataset is used in VarScrut to train the meta-predictor of deleterious nsSNV and to flag polymorphic variants.

3.2.1.4 ClinVar (disease dataset)

ClinVar (Landrum et al., 2014) is a public archive, hosted by the NCBI, that references curated information on the relationship between a human variation and a phenotype, with supporting evidence. The submitter of the variant initially provides the supporting evidence. Since 2013, all the ClinVar variants and novel submissions, are assessed by a panel of experts from the Clinical Genome Resource (ClinGen) project in order to ascertain the status of a variant and its clinical interpretation (Rehm et al., 2015). This clinical significance of a given variant is encoded in the form of a rating system (Table 3-2).

ASN.1 terms	**ClinVar clinical significance**
0 – unknown	Uncertain significance
1 – untested	Not provided (includes the cases where data are not available or unknown)
2 - non-pathogenic	Benign
3 - probable-non-pathogenic	Likely benign
4 - probable-pathogenic	Likely pathogenic
5 – pathogenic	Pathogenic
6 - drug-response	Drug response
7 – histocompatibility	Histocompatibility
255 – other	Other
	Confers sensitivity
	Risk factor
	Association
	Protective

Table 3-2: Representation of the different set of clinical terms used by ClinVar to describe the pathogenic status of a variant.

The VCF version of ClinVar available on the NCBI's FTP server [ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/] is downloaded on a monthly basis. I used the ClinVar dataset in VarScrut to train the meta-predictor of deleterious nsSNV and to flag deleterious variants. As of December 2015, ClinVar contains over 45,000 pathogenic variants, among which over 27,000 variants are pathogenic SNV (CLNSIG=5).

3.2.1.5 Human Gene Mutation Database (disease dataset)

The Human Gene Mutation Database (HGMD) (Stenson et al., 2014), is the first comprehensive collection of disease variants compiled and curated from literature. It was initially a public database, but it is now a commercial database accessible *via* a commercial licenced distributed by BIOBASE®. The system of classification in HGMD gathers concepts different from ClinVar (Table 3-3):

The version of the HGMD Pro database used for the training of VarScrut meta-predictor for deleterious nsSNV is the version V4 from 2013. From this database, ~90,000 variations, of the disease-causing mutations category (Flag=DM), are used for the assessment of pathogenicity of known variants or as a training set to predict new ones.

Category	**Definition**
DM	Disease causing mutation
DM?	Possible pathological mutation
DP	Disease associated polymorphism
FP	Functional polymorphism
DFP	Disease-associated polymorphisms with supporting functional evidence
FTV	Frameshift or truncating variants

Table 3-3: Representation of the system of classification of clinical significance in HGMD.

3.2.1.6 SwissVar (benchmark dataset)

The SwissVar database (Mottaz et al., 2010) is a collection of 31,475 disease variants derived from the literature and curated by experts from UniProt/SwissProt KB [http://www.uniprot.org/docs/humsavar].

3.2.1.7 HumVar (benchmark dataset)

The HumVar dataset is one of the retrievable variant training sets compiled by the authors of PolyPhen2 (Adzhubei et al., 2010) [ftp://genetics.bwh.harvard.edu/pph2/training]. This dataset consists of 22,196 human disease-causing mutations obtained from UniProtKB (positive control data) and 21,151 common human nsSNV (MAF>1%) without known reported involvement in any disease (negative control data).

3.2.1.8 ExoVar (benchmark dataset)

The ExoVar dataset [http://grass.cgs.hku.hk/limx/kggseq/download/ExoVar.xls] was initially compiled in the framework of the development of a logit model to predict the pathogenicity of nsSNV (Li et al., 2013). In this dataset, the positive control data is defined by 5,340 nsSNV from UniProt database with a known implication in a Mendelian disease and a negative control data defined by 4,752 rare nsSNVs (MAF <1% and with at least one homozygous genotype in the 1,000 Genomes Project).

3.2.2 *Variant annotations and predictors*

In order to assess the pathogenicity of a variant, seven annotations and functional prediction scores are agglomerated in the VarScrut meta-predictor (Chapter 5) for deleterious SNV:

3.2.2.1 Sorting Intolerant From Tolerant (SIFT)

The Sorting Intolerant From Tolerant (SIFT) is an algorithm developed by (Ng and Henikoff, 2003) to predict the effect of missense coding variants on protein function based on sequence conservation and physico-chemical similarities between the wild-type and mutant amino acids. Since its release, SIFT has become one of the standard tools for characterizing missense variation. The SIFT annotations (v5.2.2) are retrieved *via* the default option available in Ensembl's Variant Effect Predictor (VEP).

3.2.2.2 PolyPhen 2 (PPH2)

PolyPhen-2 (PPH2) is another algorithm predicting the effect of missense variants (Adzhubei et al., 2010). PPH2 is based on a set of predictive features (8 sequence-based and 3 structure-based predictive features), an alignment pipeline, and a Naïve Bayesian Classification method. The PPH2 algorithm compares the values of the 11 defined parameters between the wild-type (ancestral, normal) allele and the corresponding mutant (derived, disease-causing) allele to predict the impact of the resulting amino acid replacement. The PPH2 predictions (v2.2.2) are retrieved *via* the default option available in VEP.

3.2.2.3 PhastCons

PhastCons (Siepel et al., 2005) is a program in the PHAST (Phylogenetic Analysis with Space/Time models) package. PhastCons is based on a phylogenetic hidden Markov

model (phylo-HMM) to identify the evolutionarily conserved elements in a multiple alignment, given by a phylogenetic tree. As an output, it will generate a base-by-base conservation score, which represents the probability that the allele at that base is the most-conserved one throughout evolution.

The UCSC provides various phastCons' pre-calculated scores. In the framework of the VarScrut Meta-Predictor, the phastCons100way alignment scores are retrieved relying on 99 vertebrate genomes compared to the human genome.

3.2.2.4 phyloP

phyloP (Pollard et al., 2010) forms part of the same package as PhastCons. It allows the detection of sites under negative or positive selection, while allowing changes in evolutionary rates over the branches of the phylogenetic tree (Miller et al., 2007). To date, it represents one of the best approaches to identify individual sites under selection.

The pre-calculated scores of phyloP (phyloP100way) are retrieved from the UCSC.

3.2.2.5 Genomic Evolutionary Rate Profiling (GERP)

The Genomic Evolutionary Rate Profiling (GERP) (Cooper et al., 2005) is a program that identifies the constrained elements in multiple alignments by quantifying the substitution deficits, under a neutral evolutionary rate hypothesis, by using a Maximum Likelihood approach. This deficit in substitutions is referred to as the "Rejected Substitution" (RS), indicating regions under constraints of purifying selection.

The precomputed base-wise RS scores for the human genome compared to 99 other vertebrates are downloaded from UCSC.

3.2.2.6 Database of consensus splice-altering SNV (dbscSNV)

The database of consensus splice-altering SNV (dbscSNV) includes functional prediction and annotation for all of the ~15 million human SNV within splicing consensus regions (−3 to +8 at the 5' splice site and −12 to +2 at the 3' splice site) (Jian et al., 2014)[ftp://dbnsfp:dbnsfp@dbnsfp.softgenetics.com/dbscSNV.zip]. For each potentially altering splicing variant, it combines related functional annotations from 8 *in silico* tools (Position Weight Matrix model (Shapiro and Senapathy, 1987); MaxEntScan (Yeo and Burge, 2004); Splice Site Prediction by Neural Network (NNSplice) (Reese et al., 1997); GeneSplicer (Pertea et al., 2001); Human Splicing Finder (HSF) (Desmet et al., 2009);

NetGene2 (Brunak et al., 1991); GENSCAN (Burge and Karlin, 1997); SplicePredictor (Brendel et al., 2004)) through scores from 2 modes of ensemble learning methods: adaptive boosting (Freund and Schapire, 1997) and random forest (Breiman, 2001).

3.2.2.7 Combined Annotation Dependent Depletion (CADD)

Unlike most of the other variant annotation and scoring tools that tend to exploit only a single type of information (*e.g.* conservation) and/or are restricted in scope (*e.g.* to missense changes), the Combined Annotation Dependent Depletion (CADD) framework combines several annotations and predictors for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome (Kircher et al., 2014). It integrates multiple annotations in a Support Vector Machine (SVM) model to contrast variants that survived natural selection from simulated mutations, in order to provide a single integrative metric (C-Score) for all 8.6 billion possible single nucleotide variants (SNV) of the reference genome and the InDel found in the 1KG dataset. The version v1.2 is used in VarScrut for the training of the meta-predictor [http://cadd.gs.washington.edu/download].

3.3 Gene/Protein level information

In order to have a better comprehension of a gene's function, gene/protein annotation datasets provided by Ensembl cross-references are gathered:

3.3.1 *Gene Ontology (GO)*

The Gene Ontology (GO) (Gene Ontology Consortium, 2015) is a controlled vocabulary that describes the biological role and products of all the genes of a genome. The GO ontology comprises 3 categories of terms: (i) biological processes, (ii) molecular functions and iii) cellular components. Moreover, for each GO term associated with a gene, the ontology also provides the source of this annotation (such as experiments, data mining or predictions). In BDT, I use the GO dataset provided in each Ensembl release database.

3.3.2 *HUGO Gene Nomenclature Committee (HGNC)*

The HGNC is the international organization responsible for attributing a standard nomenclature for the unique official symbols and names for every reported human locus (Gray et al., 2015). A committee of researchers manually curates the data to create unique

gene symbols and names that are acceptable by researchers of the field. HGNC symbols and names are seen as a standard and used in all the major databases that concentrate on human genes or proteins (such as Ensembl, UniProt, GenBank), as well as on human diseases or phenotypes (such as OMIM, ClinVar). HUGO provides a list of 39,000 gene names with their associated aliases, as well as a mapping list of database identifiers used for the curation. In BDT, I integrate the HUGO dataset using the web service [http://www.genenames.org/help/rest-web-service-help] provided by the EBI server.

3.3.3 *Available biological resources*

BioData Toolkit also includes data on the availability of any related biological materials to assess biological functions. Currently, only the existence of an animal model and the availability of antibodies are referenced. This information is particularly used in VarScrut to guide the post-exome analysis step (experimental validation) for the screening of numerous candidate variants (Chapter 5).

3.3.3.1 Antibodies

After an exome analysis, when dealing with a large list of candidate genes to be experimentally screened, the availability of antibodies for protein encoding genes is one of the essential criteria to prioritize the screening order. This information is agglomerated using data from the Human Protein Atlas project (Colwill et al., 2011) [http://www.proteinatlas.org/about/download] and the derived database, Antibodypedia (Björling and Uhlen, 2008) which references the availability of antibodies from 65 providers.

3.3.3.2 Animal models

The availability of an animal model exhibiting genetic, physiological and anatomical similarity to humans is an essential tool in biology for the understanding of a gene function, especially when considering disease models. Two types of animal models are currently referenced:

(i) Mouse model

When available, the data associated with a gene knock-out mouse model is retrieved from the International Mouse Phenotyping Consortium (IMPC) (Rosen et al., 2015). This international initiative aims: i) to systematically generate a knockout strain for the ~20,000

mouse protein-encoding genes and ii) to evaluate the corresponding phenotype. As of October 2015, this project has already generated ~4,900 engineered mice and ~18,500 Embryonic Stem (ES) cells and represents established models for 938 human genetic diseases.

I used the API service [http://www.mousephenotype.org/data/documentation/api-help] provided to retrieve the catalogue of KO gene models available with the associated phenotypes and the corresponding Human disease converted into OMIM identifiers.

(ii) Zebrafish model

The Zebrafish Model Organism Database (ZFIN) is the central resource for genetic and genomic data from zebrafish (*Danio rerio*) research (Ruzicka et al., 2015). ZFIN references all the related information (such as genes, mutants, genotypes, expression patterns, phenotypes, gene product function…) from publications, for each of the ~36,000 zebrafish genes.

I used the portal of ZFIN [http://zfin.org/downloads] to download the list of gene models publicly available. Using orthology information from the InParanoid database, ~15,000 genes were found homologous with human genes, among which 7,430 genes have a corresponding zebrafish model.

3.4 Pathway/Network level information

Besides gene functions and annotations, functional genomics data of different types (such as co-expression patterns, interaction partners, metabolic/signalling pathways…) were collected to identify genes with similar profiles or interacting partners, which might suggest a similar biological role. All these data are included in BDT and are mainly used in VarScrut to build the prioritization module (see Chapter 5).

3.4.1 *Expression datasets*

The COXPRESdb provides gene co-expression relationships for 11 animal species (human, mouse, rat, chicken, fly, zebrafish, nematode, monkey, dog, budding yeast and fission yeast) (Okamura et al., 2015). This database estimates the co-expression relationships between gene profiles obtained through analysis of expression datasets from the NCBI GEO experiments (Barrett et al., 2013). Hence, for every gene, COXPRESdb provides a list of corresponding genes with the most similar expression profiles.

I use COXPRESdb database v6 [http://coxpresdb.jp/top_download.shtml] in VarScrut program for the prioritization of genes (see Chapter 5).

3.4.2 *Kyoto Encyclopedia of Genes and Genomes (KEGG)*

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases initiated in 1995 by the GenomeNet Database Service Kanehisa Laboratory at the University of Kyoto. KEGG is a database that references high-level curated and structured information to represent our knowledge of biological systems (Kanehisa and Goto, 2000). This embraces basic molecular building blocks such as genes and chemical compounds up to their organization and interactions in biological systems such as metabolic or signalling pathways.

KEGG consists of 4 major datasets connected together and acting as entry points:

(i) PATHWAY is the basic knowledge of metabolic pathways drawn manually and non-metabolic pathways generated automatically,

(ii) BRITE is the ontology of all the concepts and knowledge present in KEGG,

(iii) GENES is a gene catalogue of several complete genomes,

(iv) LIGAND is a catalogue of chemicals and reactions involved in the field of life.

BDT integrates the 39,742 entries in the release 77 corresponding to the PATHWAY dataset and all the associated genes for *Homo sapiens* using the Web service provided by KEGG portal [www.kegg.jp/kegg/rest].

3.4.3 *Reactome*

Reactome, is another curated biological pathway database that references human pathways and reactions, originating from collaborations between several international institutions (Fabregat et al., 2016). Reactome further includes extended information such as transport of molecules from one compartment to another and interactions to form a complex as well as the chemical transformations of classical biochemistry. Finally, it should be noted that the Reactome datasets are enriched with data from “cancer” pathways due to the high contribution of cancer research groups.

BDT integrates the complete list of pathways from the portal of Reactome [http://www.reactome.org/pages/download-data/]. As of October 2015, Reactome references 2,395 different pathways in Human.

3.4.4 Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING-db)

STRING-db is one of the most popular databases used in biology to retrieve known and predicted protein-protein interactions (PPI) (Szklarczyk et al., 2015). STRING-db derives its binary interaction associations from 4 different sources: (i) genomics context, (ii) high-throughput experiments, (iii) conserved co-expression experiments, and (iv) text-mining from PubMed. Binary interactions for *Homo sapiens* of the release v10 of STRING-db are downloaded [http://string-db.org/newstring_cgi/show_download_page.pl] from the website and only high-confidence interactions (score > 0.7) are kept in the VarScrut implementation.

3.4.5 GeneMANIA

GeneMANIA (Zuberi et al., 2013) is a database combining several high-throughput resources, like STRING-db. It is a flexible user-friendly web interface to analyse gene lists and to prioritize genes for functional assays. GeneMANIA extends a given list of genes with functionally similar genes that it identifies using available genomics and proteomics data. GeneMANIA also reports weights that indicate the predictive value of each selected data set for the query. GeneMANIA provides data for 7 organisms (*Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus, Homo sapiens, Rattus norvegicus and Saccharomyces cerevisiae*) and hundreds of data sets have been collected from GEO, BioGRID, IRefIndex and I2D, as well as organism-specific functional genomics data sets.

I retrieved the latest dataset of GeneMANIA corresponding to the release of 2014 from its online data archive [http://genemania.org/data/current/], which accounts for around 2,152 association networks containing 537,599,442 interactions for the 166,084 genes from the 7 eukaryotic organisms.

3.4.6 *IntAct*

The IntAct database is a very high-valued source for molecular interaction data, hosted by the EBI (Orchard et al., 2014). The binary interaction information is populated by curated data derived from either literature or direct data deposits. IntAct curation process is based on a sophisticated web-based curation tool utilized by multiple curation teams within the EBI. As of December 2015, IntAct contains approximately 578,879 curated binary interactions from over 14,010 publications. The UniProtKB/Swiss-Prot curators use these annotations. Data from IntAct are retrieved from the FTP server [ftp://ftp.ebi.ac.uk/pub/databases/intact/current] of the EBI in the PSI-MI tabular format in order to have detailed descriptions of the interactions.

3.5 Phenotype/Disease level information

3.5.1 *Human Phenotype Ontology (HPO)*

The Human Phenotype Ontology (HPO) (Köhler et al., 2014) is an ontology that describes over ten thousand phenotypical features (terms) that can be associated with a disease mined from the medical literature or from disease datasets such as Orphanet or OMIM. Moreover, a correlation link between a disease phenotype and a gene curated from literature is also provided. It is an on-going effort for a larger structure with less subjectivity in the description of diseases. The ontology compendium is fed by both manual annotations by HPO teams and automated detection of HPO terms in the "Clinical Synopsis" of OMIM entries. The HPO information is updated on a daily basis and available on its public server Hudson [http://compbio.charite.de/jenkins/].

3.5.2 *Mammalian Phenotype Ontology (MPO)*

The Mammalian Phenotype Ontology is a compendium of ontology terms used to describe the phenotypical traits observed in genetically generated rodent models (rat and mouse) (Smith and Eppig, 2009). The MPO is hosted and generated mainly by the Mouse Genome Informatics Database (MGI), which references all the catalogues of mouse gene KO models. However, unlike the HPO, the terms are not always complete and documented with few details. Among the different formats and outputs available on the FTP server of the MGI [ftp://ftp.informatics.jax.org/pub/reports/index.html], I downloaded the file 'MGI_PhenoGenoMP.rpt', which is a tabular file containing the most comprehensive list of

phenotypic terms (>18,000) with the associated genes and cross links to other databases such as OMIM.

3.5.3 *Online Mendelian Inheritance in Man (OMIM)*

The Online Mendelian Inheritance in Man database (OMIM) is the first and most complete compendium of human genetic disease data initiated by McKusick at the John Hopkins Institute in 1966 (Amberger et al., 2015). It contains curated descriptions of human genes and phenotypes and the relationships between them and is updated on a daily basis through mining and curation of the literature. OMIM is based on the published peer-reviewed biomedical literature and is the reference source for Mendelian diseases. Genes and phenotypes are described in separate entries with a unique, stable six-digit identifier (MIM numbers). OMIM entries have a structured free-text format that provides the flexibility necessary to describe the complex and nuanced relationships between genes and genetic phenotypes in an efficient manner. As of December 2015, OMIM references over 15,000 genes and ~8,100 disorders. I use the API service to retrieve all the data and daily updates from OMIM [http://www.omim.org/help/api].

3.5.4 *Orphanet*

The Orphanet database [http://www.orpha.net/] is originally a French initiative, started in 1997, that has now turned into a European one with new partners such as the EBI. Orphanet provides a structured vocabulary for RD and integrates: i) an ontology for nosology (RD classification), ii) relationships information (gene-disease relations, epidemiological data) and iii) links with other terminologies (MeSH, UMLS, MedDRA), gene databases (OMIM, UniProtKB, HGNC, Ensembl, Reactome) or classifications (ICD-10). Like OMIM, the data in Orphanet are extracted from the literature, but the review process is carried out by the clinicians of the Orphanet team. Besides description of the diseases and related genes, Orphanet also references additional data such as epidemiological data and the list of current drugs assays in Europe and in America.

I download both the Orphanet classification of diseases (‘en_product1.xml’) and the associated epidemiological data (‘en_product2.xml’) from the OrphaData portal (http://www.orphadata.org/cgi-bin/index.php), which is a dedicated platform for all the data gathered by the Orphanet initiative.

3.5.5 *GeneReviews*

GeneReviews (GR) [13] is a curated registry of Mendelian diseases hosted by the NCBI. Among the different sources of disease databases, GR represents the dataset with the highest standards. The information in GR is authored by the corresponding experts of the disease, and the description is also peer-reviewed and presented in a standardized format with information focused on clinically relevant and medically actionable information for the diagnosis. GeneReviews currently comprises 640 merged disease records and the updated information is peer-reviewed at least 7 times by internationally acknowledged subject experts. I use the E-utilities tools of the NCBI to retrieve all the GR data on a daily basis.

3.6 Literature level information

For the development of Pub*Athena* and the literature-tracking module of VarScrut, I focus mainly on scientific literature available *via* PubMed services. In order to enrich the information extracted by text mining from the abstracts, I use PubTator to annotate different concepts (species, gene, disease, mutation, and chemical).

3.6.1 *PubMed*

PubMed is a search engine, developed and maintained by the National Center for Biotechnology Information (NCBI). As of December 2015, PubMed provides access to over 25 million biological or biomedical related documents (articles and books) referenced in the MEDLINE database of the US National Library of Medicine (NLM). The NCBI also provides several programmatic accesses to PubMed for the querying and the retrieval of data in batches, like the API mode or the console Linux mode [14].

The API mode is available through the Entrez Programming Utilities (E-utilities), which comprise nine server-side programs (see Table 3-4) that provide a stable interface for the Entrez system of 38 databases at NCBI. The E-utilities use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data.

E-utility name	Purpose
EInfo	Provides the latest records statistics for each of database resource.
ESearch	Entry point for a text query of a specified database, and provides in response the list of matching UIDs that can then be used by other services.
EPost	Provides a temporary storage of E-utility services for a list of UIDs of a specified database, and the list of UIDs can be later accessed with a query key and a web environment.
ESummary	Provides document summaries for a given list of UIDs for a specific database.
EFetch	Retrieves the corresponding data records in a specified format for a given list of UIDs.
ELink	Provides of a list related UIDs within or across databases for a given list of UIDs.
EGQuery	Provides the number of records across all databases that match a text query.
ESpell	Provides spelling suggestions for a text query in a given database.
ECiteMatch	Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.

Table 3-4: Entrez Programming Utilities provided by the National Centre for Biotechnology Information for the access to the Entrez system and databases.

Pub*Athena* uses an implementation of the ESearch and EFetch services with PubMed to search and retrieve article abstracts. An initial run was performed to obtain a dump copy of the 25 million documents available *via* PubMed (February 2015). This implemented solution is then relaunched on a regular basis to retrieve the latest updates. The retrieval of the article data is done in XML and JSON format in order to obtain the associated metadata as well, such as the list of keywords, authors, or publication details. As of December 2015, altogether the dump of the PubMed abstracts represents ~8.1 GB.

3.6.2 *PubTator*

PubTator (Wei et al., 2013), is a web-based tool to accelerate the manual curation of the literature and helps the reader to process numerous articles. Through a series of automatic text-mining tools (GeneTUKit for gene mentions; GenNorm for gene normalization; SR4GN for species; DNorm for diseases; tmVar for mutations; a dictionary-based lookup approach for chemicals) PubTator provides, the means to extract important biological entities and their relationships in an abstract. PubTator is an annotated interface synchronized with PubMed, which provides daily updated annotations for each PubMed article. These PubTator annotations are available on its FTP server [ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator/], and

through a web service [http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#curl] making it extremely convenient for the mining of a compendium of articles.

In the implementation of Pub*Athena*, the initial dump of PubMed articles was annotated with the data found on the FTP server of PubTator, while the daily updates of PubMed articles are annotated on the fly by using PubTator's web services. Altogether, as of December 2015 the PubTator annotations represent up to ~11 GB.

Chapter 4 Bioinformatics tools and architecture

When dealing with variant analysis from NGS data, several tools are available to perform the variant calling and to generate the list of variants in the VCF format. However, all the algorithms available do not provide the same representation of a variation event (usually InDel) occurring at a given locus. These multiple representations of the same event can be considered as a false positive variant and left unannotated by different annotation tools. Hence, to minimize this effect, VarScrut includes two bioinformatics tools for the pre-processing and annotation of variant information.

4.1 Variant tool (*vt*)

The variant tool package includes a set of utilities for several procedures to pre-process and normalize the representation of variant alleles at a given locus (Tan et al., 2015). Two procedures are implemented in VarScrut:

4.1.1 *Decomposition*

In the VCF format, multiple allelic variant events occurring at the same locus tend to be summarized on the same line. This summarized representation is a problem as alternate alleles are often skipped when screening public databases or annotating using annotation tools. The decomposition procedure reformats the VCF file by separating the multiple alleles into distinct variant records while maintaining the corresponding metric information and genotypes for each sample.

Example of decomposition:

Before:

#CHROM	POS	ID	REF	ALT	FILTER	INFO
1	1001	.	CTCC	CCC, C, CCCC	.	.

After:

#CHROM	POS	ID	REF	ALT	FILTER	INFO
1	1001	.	CTCC	CCC	.	.
1	1001	.	CTCC	C	.	.
1	1001	.	CTCC	CCCC	.	.

4.1.2 *Normalization*

In order to reduce potential interpretation errors when checking whether a variant has already been reported in public databases such as 1,000 Genomes, the variant needs to be standardized. The vt normalization procedure recalculates the alleles into their most left aligned and parsimonious representation.

Example:

Before				After		
POS	REF	ALT	→	POS	REF	ALT
1001	CTCC	CCC	→	1001	CT	C
1001	CTCC	C	→	1001	CTCC	C
1001	CTCC	CCCC	→	1002	T	C

4.2 Variant Effect Predictor (VEP)

In VarScrut, the annotation of variants is based on the Ensembl Variant Effect Predictor (VEP) tool (McLaren et al., 2010). This tool predicts the variation type and its resulting consequences at the transcript or protein levels. Moreover, it is a very versatile tool that can be customized with plugins to add additional functionalities and annotations. The latest version of VEP (v81) is configured to manage the GhRC37 genome build and additional annotations and predictor scores (dbSNP, 1KG, ExAC, ClinVar, PhastCons, phyloP, GERP, dbscSNV, CADD) were integrated *via* plugins (Table 4-1).

```
## Co-located variants flags
check_alleles       1
check_existing      1
check_ref    1
per_gene     1
pick_order   canonical, tsl, biotype, rank

## output annotation flags
allele_number1
biotype      1
canonical    1
ccds   1
domains      1
gmaf   1
hgvs   1
maf_1kg      1
maf_esp      1
numbers      1
polyphen     b
pubmed 1
regulatory   1
sift   b
symbol 1
total_length 1

## Custom annotations
custom /biolo/vep/clinvar_latest_tidy.vcf.gz,CLINVAR,vcf,exact,0,CLNSIG,CLNACC
custom /biolo/vep/1KG_phase3_tidy.vcf.gz,1KG,vcf,exact,0,AF
custom /biolo/vep/dbsnp141_tidy.vcf.gz,dbSNP141,vcf,exact,0
custom /biolo/vep/All_hg19_RS.bw,GERP_RS,bigwig
custom /biolo/vep/hg19.100way.phastCons.bw,PHASTCONS,bigwig
custom /biolo/vep/hg19.100way.phyloP100way.bw,PHYLOP,bigwig
custom /biolo/vep/hg19_fitcons_fc-i6-0_V1-01.bw,FITCONS,bigwig

## plugins
plugin Blosum62
plugin CADD,/biolo/cadd/lite/whole_genome_SNVs.tsv.gz,/biolo/cadd/lite/InDels.tsv.gz
plugin ExAC,/biolo/vep/exac_latest_tidy.vcf.gz## Co-located variants flags
```

Table 4-1: List of parameter settings used in VarScrut

4.3 Algorithms

4.3.1 *Random Walk with Restart (RWR)*

In order to benefit from the extended data network integrated in BDT, the Random Walk with Restart algorithm (RWR) was used to find similarity between genes (Köhler et al., 2008).

In VarScrut, RWR is applied as a prioritization system to rank candidate genes resulting from the analysis of a patient exome, on the basis of their proximity to known disease-causing genes matching the symptoms of the patient. The detailed RWR implementation will be further discussed in the VarScrut section (see Chapter 5). Briefly, VarScrut uses the algorithm NetWalker, an implementation in C++ of a random walker that provides a statistical evaluation of random walk results, as described in (Zhang et al., 2011).

Intuitively, RWR calculates the similarity between two genes, ***i*** and ***j***, on the basis of the likelihood that a random walk through the interaction network starting at gene ***i*** will finish at gene ***j***, whereby all possible paths between the two genes are taken into account. The random walk is initialized for each exome analysis to take into account the most probable disease genes based on the phenotypic description of the patient with HPO terms. Hence, the random walk will start from the most probable known disease-gene family members in order to search additional family members in the linkage intervals. All the evaluated genes in the network are then ranked according to their degree of similarity in the functional interaction network.

Formally, the random walk with restart is defined as:

$$p_{t+1} = (1-r)\, Wpt + rp_0$$

Where W is the column-normalized adjacency matrix of the graph and p_t is a vector in which the i^{th} element holds the probability of being at node ***i*** at time step t.

4.4 Computational infrastructure and software development

All the developments carried out during this thesis were done using a continuous integration approach which involves several steps, from (i) an integrative framework which handles all the continuous incoming flux of heterogeneous data, (ii) the processing and structuring of the data into BDT, and (iii) their integration and usage with the different developed resources (VarScrut, CilioPath, Pub*Athena*).

4.4.1 *Computational resources*

Given the different tools developed (VarScrut, Pub*Athena*, CilioPath) and the various datasets used (BDT), the computational resources play an essential role for the maintenance and performance of the hosted applications. The CSTB laboratory computational power is composed of a cluster of seven Dell PowerEdge R720 servers configured with an average of 24 cores, 128G of memory and 8 terabytes of local disk space per server.

4.4.2 *Software development*

4.4.2.1 Python programing

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. Moreover, Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library for scientific calculations. Python is widely used supported by the bioinformatics community and has even an extension interface with C-language, Cython, for demanding computing performances.

Moreover, Python interpreters are available for installation on many operating systems, allowing Python code execution on a wide variety of systems. Python code can be packaged into stand-alone executable programs for some of the most popular operating systems, allowing the distribution of Python-based software for use on those environments without requiring the installation of a Python interpreter.

All the work carried out during this thesis is coded in Python v2.7.

4.4.2.2 PyCharm

All the coding tasks performed during this thesis were carried out with PyCharm Professional Edition 5.0.4, which is an Integrated Development Environment (IDE) used for programming in Python [15]. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCS), and supports web development with Flask. PyCharm is developed by the company JetBrains.

4.4.2.3 Python libraries

In order, to improve the performances of the applications developed and increase the maintainability, several standard and open source Python libraries were used. For dedicated tasks, external libraries used during the development are listed in Table 4-2.

Library	VarScrut	Pub*Athena*	CilioPath	Usage	Reference
CyVCF	Yes	No	No	A fast Python library for VCF files leveraging Cython for speed.	https://github.com/arq5x/cyvcf
Flask	No	Yes	Yes	A lightweight Python web framework based on Werkzeug and Jinja 2.	http://flask.pocoo.org/
NLTK	No	Yes	Yes	NLTK, is a suite of libraries and programs for statistical natural language processing	http://www.nltk.org/
Peewee	Yes	Yes	Yes	Peewee is a simple and small ORM (Object-Relational Mapper) for databases.	http://docs.peewee-orm.com/
PETL	Yes	Yes	Yes	petl is a general purpose Python package for extracting, transforming and loading tables of data.	https://petl.readthedocs.org/
SciKit	Yes	Yes	Yes	SciKit is an open source machine learning library for the Python programming language	http://scikit-learn.org/stable/

Table 4-2: Python libraries used in software developments.

4.4.3 *Database engine*

All the developments involving the storage of datasets was performed using SQLite. SQLite is a relational database management system contained in a C programming library. In contrast to many other database management systems, SQLite is a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine. SQLite is the most widely deployed database engine in the world.

4.4.4 *Jenkins*

For the continuous integration and processing of the datasets coming from several sources, I set up a Jenkins server (http://lbgi.fr/jenkins/). Jenkins is an open source continuous integration tool written in Java. It is widely used and supported by the software development community. Jenkins provides continuous integration services for software development through a server-based system. It supports code versioning tools as well as arbitrary shell scripts and Windows batch commands.

Jobs can be started by various means, including being triggered by commit in a version control system, by scheduling *via* a cron-like mechanism, *i.e.* sequential complex jobs can be orchestrated with the triggering of a new job when the previous builds have completed. Nodes can be created on multiple servers to distribute the load and scripts can be run.

In this thesis, Jenkins is used to perform several predefined and scheduled tasks in parallel, such as the automatic updating of all the datasets in BDT or the literature tracking on PubMed or the periodic re-launching of variant annotations in VarScrut.

III. Results and Discussion

Chapter 5 VarScrut

As previously described in section 2.4.2, WES is now the method of choice for gene-identification and diagnostics in biomedical research, as it provides in a single run, access to the complete protein-coding exome sequences for screening of potential disease-causing variants and genes. Although WES has successfully been applied in several biomedical projects, the feedback that emerges is that, apart from the identification of well-known disease-causing genes, WES has a diagnostic yield of ~30% for the identification of novel genes (except in specialized and localized diseases; *e.g.* retinal dystrophy) (Table 5-1)

Disease category cohort	Total # of patients	Diagnostic rate (%)	PMID
Prenatal abnormalities	30	10	24476948
Sever intellectual disability	100	16	23033978
Sporadic sever early-onset epilepsy	264	17	23934111
Suspected mitochondrial disorders	102	22	23596069
Childhood neurodevelopmental disease	118	27	22700954
Pediatric-onset ataxia	28	46	24108619
Retinal distrophy	33	76	23105016
Genetics laboratory cohort of children with neurologic phenotypes	250	25	24088041
Genetics laboratory cohort	500	37	25356970

Table 5-1: Diagnostic yield with WES projects according to different types of disease cohorts screened.

Several factors may explain the current resolution yield in RD:

(i) The quality of the sequencing (Bloch-Zupan et al., 2011).

(ii) The low number of affected individuals available for a study (Gilissen et al., 2012).

(iii) The low, or lack of, information available on candidate genes.

In addition, the complexity of the bioinformatics protocols to analyse WES data is likely to introduce discrepancies and limit the final resolution yield. Several recommendations and guidelines have been issued for best practices and standards have emerged for the interpretation of variation data, such as those issued by the American College of Genetics and Genomics (ACMG) (Richards et al., 2015) or the ClinGen consortium (Rehm et al., 2015).

Among the essential points highlighted by these guidelines for a proper interpretation of variation data, there are: (i) normalization of the variants using the standard Human Genome Variation Society nomenclature (HGVS) (Dunnen and Antonarakis, 2000), (ii) screening of the different MoI (Modes of Inheritance), (iii) screening of the literature (will be discussed in the next Chapter), variation databases and use of the latest annotation datasets (*e.g.* GO, OMIM…), (iv) use of multiple robust computational predictors to determine the effect of the variants, (v) use of the patients phenotypical data to prioritize the candidate variants/genes accordingly.

In view of the discussed recommendations and guidelines, and some additional features that I also defined (*e.g.* downloadability or follow-up), I carried out a survey of the different variant analysis solutions currently available to evaluate this complex context (Table 5-2). Schematically, current variant analysis tools can be classified into three main categories: those based on (i) a statistical analysis of variants (*e.g.* VAAST); (ii) SQL-query based tool implying an almost totally manual analysis (*e.g.* GEMINI); (iii) case comparison with predefined filtering methods (*e.g.* EVA, VAAST-Phevor).

Category	Tools	Downloadable	All types of variants	Annotation module	MoI module	Filtering module	Prioritisation module	Follow-up		# of applied citations	Publication Year	PMID
								Update revaluations	Literature tracking			
Statistical methods	eXtasy	🟢	🟡	🟡	🔴	🔴	🟢	🔴	🔴	1	2013	24076761
	VAAST	🟢	🟢	🔴	🟢	🔴	🟢	🔴	🔴	6	2011	21700766
	VariantMaster	🟢	🟢	🔴	🟢	🟡	🟢	🔴	🔴	3	2014	24389049
	DeNovoGear	🟢	🔴	🔴	🔴	🟡	🔴	🔴	🔴	4	2013	23975140
SQL query based	GEMINI	🟢	🟢	🟢	🟡	🟡	🟡	🟡	🔴	6	2013	23874191
	vtools	🟢	🟢	🟢	🟡	🟡	🟡	🔴	🔴	7	2012	22138362
Predefined filtering methods	ExomeAssistant	🔴	🟢	🟡	🟢	🟡	🔴	🔴	🔴	1	2012	23231371
	EVA	🟢	🟡	🟢	🟡	🟡	🔴	🔴	🔴	2	2012	23095660
	FAVR	🟢	🟡	🟢	🔴	🟢	🔴	🟡	🔴	1	2013	23441864
	wANNOVAR	🔴	🟢	🟢	🟡	🟡	🔴	🟡	🔴	14	2012	22717648
	GeneTalk	🔴	🟢	🟡	🟢	🟡	🔴	🔴	🔴	7	2012	22826540
	Famseq	🟢	🟡	🟡	🟢	🔴	🔴	🔴	🔴	1	2013	23426633
	VarSifter	🟢	🟢	🔴	🟢	🟢	🔴	🔴	🔴	10	2012	22210868
	VAR-MD	🟢	🟡	🟡	🟡	🟡	🟡	🔴	🔴	4	2012	22290570
	TREAT	🟢	🟡	🟡	🔴	🟡	🔴	🔴	🔴	4	2012	22088845
	gSearch	🟢	🟢	🟡	🔴	🟡	🔴	🔴	🔴	1	2012	22730434
	SNVerGUI	🟢	🟡	🟢	🟡	🟡	🔴	🔴	🔴	0	2012	23024288
	TrioVis	🟢	🟡	🔴	🟡	🟡	🔴	🔴	🔴	0	2013	23658417
	Annotate-it	🔴	🟡	🟢	🟡	🟡	🔴	🔴	🔴	1	2012	23013645
	SVA	🟢	🟢	🟢	🟡	🟡	🔴	🔴	🔴	8	2011	21624899
	Phen-Gen	🟢	🟢	🟢	🟢	🟢	🟢	🔴	🔴	1	2014	25086502
	VAAST-Phevor	🟢	🟢	🟢	🟢	🟢	🟢	🟡	🔴	0	2014	24702956

Table 5-2: Simplified representation of the survey of variant analysis tools available. The tools are described according to functionalities essential for variant analysis and the colours indicate the level of satisfaction of a given criterion; Green: Full-filled, Yellow: Partially full-filled and Red: Unfilled. All types of variants refer to the detection, normalization and evaluation of all types of variants (e.g.; SNV, InDel) with predictors or meta-predictors. Annotation module refers to the use of updated gene functional information to identify candidate genes (e.g. GO, OMIM, linked phenotypes). MoI module refers to the screening of the five possible Modes of Inheritance (Autosomal recessive and dominant, X-linked recessive and dominant, de novo). Filtering module refers to the possibility to apply multiple filtering criteria to obtain short list of genes. Prioritisation module: refers to the ability to reorder the list of results according to different variant or gene criteria. Follow-up module refers to the possibility to revaluate the list of candidates with updated information. Number of applied citations refers to the number of published cases in which the tool was used for the identification of the causative gene followed by the Publication Year of the tool manuscript and the PubMed Identifier (PMID).

The statistical analysis methods such as VAAST or DeNovoGear, use a statistical model to estimate the likelihood that any given variant is specific to an affected individual compared to a group of healthy individuals. The efficiency of these methodologies is very much dependent on the set of healthy exomes used as control for the training. For instance, VAAST, which was the first statistical tool developed to analyse variation data, requires at least 100 healthy exomes for the training of its model.

SQL-query based tools are centred on a local SQL database which stores all the variation data and they rely on the power of the SQL language to formulate different filtering queries. These tools can propose some predefined queries to filter the results by genetic scenarios (MoI) or by allele frequency, such as GEMINI or vtools. While such tools, like GEMINI, provide great flexibility to customize and refine queries, they remain accessible only to advanced users who have programming skills in SQL and are confortable with Linux commands. Consequently, to generate a complete analysis procedure, users have to write a script that contains the corresponding sequence of commands.

Finally, the category that represents the majority of the variant analysis tools implies predefined procedures and filtering methods. These tools, such as EVA, wANNOVAR, VarSifter or SVA, can perform predefined analyses like the comparison of case *versus* control, or the discovery of *de novo* variants in a trio. However, few of these tools provide a full set of analyses (*e.g.* MoI – filtering – prioritization modules), and most of them are specialized, such as EVA that compares groups of affected *versus* unaffected individuals but does not include the compound heterozygote scenario (see section 1.2.1.2) or TrioVis which is more dedicated to the analysis of trios for *de novo* variants. Moreover, most of these popular tools are either, webservers (wANNOVAR) with few procedures and limited customizable parameters, or they can be downloaded, but can only handle a limited number of samples for the analysis (VarSifter). Recently, a new type of tools with predefined procedures has emerged (*e.g.* VAAST-PHEVOR, Phen-Gen) that verify most of the features defined. It should be highlighted that in addition to the variation data, these tools require as mandatory input, the list of phenotypical features (in HPO terms) of the individuals being analysed. The genes associated with the phenotypes provided are used as seed in a dedicated prioritization module (*e.g.* ontology propagation for PHEVOR, Random Walk with Restart for Phen-Gen). These efficient tools are further discussed in the VarScrut manuscript.

Overall, with respect to the recommended features and best-practice guidelines, few tools integrate all the requirements. Indeed, statistical tools have efficient prioritization modules, but mostly lack annotation and filtering modules and are strongly dependent on the

dataset used to train the model. SQL-query based tools provide the most complete source of annotations for all types of annotations and genetic scenarios, but users have to define their own filtering and prioritization module. Tools with predefined procedures and filters depend on the dedicated purpose. Generally, they provide basic annotation modules, a few provide a customizable filtering module and none of them have a prioritization module (except PHEVOR and Phen-Gen).

Thus, it appears that there is still an unmet need for a complete variant analysis workflow verifying high quality, user-friendly and automated procedures with literature-tracking modules ensuring systematic screening of the literature or follow-up procedures to reprocess the results periodically with the latest annotations.

5.1 VarScrut philosophy

The goal of the VarScrut design was the development of a fully automated and valuable tool, encompassing the 5 *in silico* steps of a WES analysis (see Figure 2-4) and amenable for routine in clinical uses. In collaboration with the clinicians of Hélène Dollfus' team and the bioanalysts of Pierre Collet and Olivier Poch's team and after an initial period of manual analysis of the exome data to understand the depth and breadth of WES analysis, we defined a set of major features for VarScrut:

- **Automatic screening of the MoI (modes of inheritance)** – VarScrut should scrutinize in parallel all the MoI (in the manuscript named inheritance scenarios) (Autosomal Recessive, Autosomal Dominant, X-Linked Recessive, X-Linked Dominant, *de novo*), in order to take into account sporadic cases with unknown or dubious scenarios.
- **Annotation module** – VarScrut should annotate the variants with the latest up-to-date annotations. Moreover, VarScrut should be able to summarize into a single score the deleteriousness of nsSNV.
    - **Meta-predictor –** VarScrut scoring system should be able to distinguish between deleterious nsSNV and rare benign variants, in order to decipher Variants of Unknown Significance (VUS) (MacArthur et al., 2012).
- **Prioritisation module** – VarScrut should limit the mandatory requirements and be able to prioritize the list of candidates both on variant information criteria and on gene information criteria when available. Moreover, in view of personalized

medicine, VarScrut should be able to prioritize the candidates according to the specific phenotypical profiles of the patient when available.

- **Post-analysis module –** VarScrut should be able to provide guidance for most appropriate experimental validation in order to facilitate the screening of the list of candidates.
- **Follow-up module** – VarScrut should be able to re-evaluate the list of variant candidates according to novel information, such as updated gene annotations or novel knowledge in literature (see Pub*Athena* for more details in Chapter 6).

Last but not least, as opposed to the other available tools, VarScrut has been designed to propose all the functionalities in a unifying framework, which can perform automated handling of an exome analysis project from the variant analysis step up to the validation step. This aspect has been specifically conceived for routine clinical uses, which require systematic and robust tools. Additionally, VarScrut re-evaluates the data periodically according to information updates, thus simplifying the follow-up of unresolved WES. This unifying approach mimics the natural logic of a biologist processing in a variant analysis project, thus facilitating the rapid and ergonomic acceptance and usage of VarScrut (see submitted manuscript of VarScrut to the American Journal of Human Genetics).

5.2 Manuscript of VarScrut

VarScrut: an expert system for assessing the causality of exome variants in Mendelian disorders

Kirsley Chennen,[1,2] Alexis Allot,[1] Osorio Abath Neto,[3] Raphaël Schneider,[1,3] Jocelyn Laporte,[3] Raymond Ripp,[1] Arnaud Kress,[1] Laetitia Poidevin,[1] Corinne Stoetzel,[2] Anne Boland,[4] Jean-François Deleuze,[4] Julie D. Thompson,[1] Odile Lecompte,[1] Hélène Dollfus,[2,5] Olivier Poch[1]

1 Complex Systems and Translational Bioinformatics - ICube, CNRS UMR 7357, Federation of Translational Medicine of Strasbourg (FMTS), University of Strasbourg, 11 rue Humann, Strasbourg, France.

2 Laboratory of Medical Genetics, INSERM U1112, Institute of Genetics and Medicine of Alsace (IGMA), Strasbourg Medical School, University of Strasbourg, 11 rue Humann Strasbourg, France

3 Department of Translational Medicine and Neurogenetics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC,) CNRS UMR 7104, INSERM U964, University of Strasbourg, Illkirch Cedex, France.

4 Centre National de Genotypage (CNG), Institut de Génomique, CEA, 91057 Evry, France

5 Institute of Medical Genetics of Alsace, National Reference Centre for Rare Diseases in Ophthalmic Genetics (CARGO), Strasbourg Medical School, 1, place de l'Hôpital, Strasbourg, France.

Contact: kchennen@unistra.fr, olivier.poch@unistra.fr

Abstract

The widespread use of next-generation sequencing (NGS) technologies has revolutionized research and diagnosis in the field of rare diseases. However, the effectiveness of NGS approach is still limited and, overall only ~25% of whole exome

sequencing (WES) experiments succeed in identifying the causative genes. One of the main limits of NGS/WES approaches remains the challenge in obtaining a shortlist of candidate genes and ascertaining the deleteriousness of the associated variants. Here, we present VarScrut, an expert system with a knowledge-guided approach for exome analysis, which automatically screens genotyped variants in 5 inheritance scenarios (autosomal recessive, autosomal dominant, X-linked recessive, X-linked dominant, and *de novo*) combined with a dedicated Logit-based meta-predictor scoring system for nsSNV and a knowledge-driven step for candidate genes prioritization.

In comparison with state-of-the-art WES analysis tools, VarScrut outperformed current widely used tools for the identification of the causative gene among the top 10 genes on both simulated disease-exomes (identification in 61-68% of the cases with only genotype information and 91-97% of the cases when combined with phenotypic information) and real disease-exome datasets (identification in five of six cases of heterogeneous myopathies). An important advantage of VarScrut concerns the original 'follow-up procedure' that, to our knowledge, is the first tool to perform periodic re-annotation and literature tracking for the reassessment of candidate variants of unresolved exomes.

Introduction

Over the past 30 years, scientists have attempted to identify the causative genes involved in rare Mendelian diseases through a variety of techniques, ranging from candidate gene approaches to linkage analysis, essentially in familial cases[1]. Recently, the study of rare genetic diseases and molecular diagnostics, in general, has been revolutionized by the advent of the Next Generation Sequencing technologies (NGS). Among the various NGS applications, Whole Exome Sequencing (WES), which is the

targeted sequencing of exonic regions of known protein-coding genes (~1–3% of the human genome), is gradually replacing conventional approaches for the study of rare Mendelian diseases. Briefly, a WES data analysis workflow consists of three main phases: (i) a NGS data processing phase that aligns the raw sequence reads (FASTQ files) on the Human reference genome (BAM files), (ii) a variant calling and genotyping phase that identifies deviations such as Single Nucleotide Variants (SNV) and Insertion-Deletion variants (InDel) from the Human reference sequence (VCF files), and (iii) an integrative variant analysis phase to identify the disease-causing variants, comprising an annotation, a filtering and a candidate gene prioritization step. The efficiency and cost-effectiveness of WES has been demonstrated[2,3] by the identification of over 550 novel causative genes, responsible for rare Mendelian diseases since 2010[4]. However, the identification of the causative gene remains a laborious process with a resolution rate of ~25% for WES analyses[5,6], implying that numerous WES remain unresolved with hundreds of potential genes that cannot be easily screened experimentally. Several medical and technical issues may impair one of the WES processing phases and hamper the identification of the causal gene.

At the medical level, rare diseases can cover disparate situations ranging from sporadic cases to heterogeneous cohorts of patients. Sporadic cases and small nuclear families (with often incomplete trios available for WES) are the most frequent clinical cases encountered[1] and this scarcity in patients influences several assumptions (*e.g.* modes of inheritance (MoI), high penetrance…) needed in the variant analysis phase and may impede WES analyses success.

At the technical level, the quality of the sequenced regions can be decisive for the identification of the causal variant. WES focuses only on variants within the known exonic regions and intronic variants closed to exons, such as the consensus splice

sequence. The yield of WES capture kit reagents is currently about 90%, implying that there is a 10% chance that regions like repetitive or GC-rich sequences are under-represented. Thus, for these poorly covered regions, WES needs to be complemented by other dedicated means to look for the existence of genuine potential deleterious variants[7].

Moreover, the last phase of the WES workflow (integrative variant analysis) is highly dependent on the quality of the variant annotation step. Indeed, with several thousands of variants identified by each WES, the annotation, the prediction of variant impacts and the interpretation of their potential deleteriousness remains the most challenging part, as each healthy individual bears ~5–10% private variants, not catalogued in public databases, as well as ~100 loss-of-function variants[8]. This further complicates the identification of disease-causing variants and the interpretation of Variants of Uncertain Significance (VUS)[9] in the extended list of variants from unresolved exomes. The majority of the VUS is composed of non-synonymous Single Nucleotide Variants (nsSNV). Numerous methods and algorithms have been developed to predict the effect of nsSNV, and in general, these predictor algorithms have an accuracy ranging from 65 to 80% when benchmarked on known disease variants[10,11]. Recently some meta-predictor approaches (CADD[12], FATHMM[13], dbNSFP[14]) combining several predictor scores have emerged, allowing an increase in accuracy up to 90% [15].

Several tools (*e.g.* ANNOVAR[16], VAAST[17], SnpEff[18], or GEMINI[19]) have also been proposed to perform WES variant-data analysis, including annotation, filtering and prioritization of candidate genes according to the variant impact/deleteriousness. Recently, novel approaches have been developed to further prioritize automatically and shorten the list of candidate genes, by combining the variant impact

100 characterizations with additional knowledge such as ontologies like the Human
101 Phenotype Ontology[20] (HPO) (PHEVOR[21]) or disease-gene functional networks
102 (Phen-Gen[22]).
103 In this context, several recommendations and guidelines, including those from the
104 American College of Medical Genetics and Genomics (ACMG)[23,24] covering the
105 major limits associated with the clinical applications of NGS[25,26], have been issued to
106 improve the analysis and interpretation of exome variants. In addition, for unresolved
107 exomes with an extended list of VUS, the ACMG guidelines recommend follow-up
108 procedures to periodically collect updated information in order to re-evaluate the
109 causative status of each VUS. This follow-up procedure generally consists in the
110 gathering of additional information linking variants/genes to diseases like periodic
111 monitoring of the literature, databases screening and availability of animal models
112 relevant to the human disease[25].
113 Here, we present VarScrut, an expert system with a multi-level variant analysis
114 approach specifically designed to simplify and automatize the WES analysis for
115 clinicians and researchers. In line with the current practices, VarScrut considers WES
116 analysis as a unified process ranging from the variant/gene identification up to
117 guidance for experimental validation and the management of unresolved exomes. The
118 variant analysis procedure of VarScrut combines an automated screening of five MoI
119 scenarios (autosomal recessive, autosomal dominant, X-linked recessive, X-linked
120 dominant, and *de novo*), a dedicated meta-predictor scoring system for nsSNV based
121 on an originally trained Logit model and a knowledge-driven procedure for candidate
122 genes prioritization, which based on an integrated gene functional network with
123 mined data from clinical phenotypes, pathways or literature origins. VarScrut proved

to out-performed the most widely used and efficient variant analysis tools on both simulated and real disease exomes.

Finally, besides indicating the existence of biological materials (antibodies, animal models) to facilitate the experimental validation, the post-analysis step includes a follow-up procedure for unresolved exomes resulting in an extended list of variants/genes. This post-analysis step performs an automated relaunching of the variant analysis procedures based a monthly update of the variant/gene annotations and a daily tracking of their associated PubMed records.

Material and Methods

Implementation and Infrastructure

VarScrut is an expert system for WES analysis based on a modular infrastructure organized around an internal SQLite database (v3.8.10). VarScrut encompasses three main steps: a variant processing step (Figure 1A), an optional knowledge-driven gene prioritization step (Figure 1B) and a post analysis step (Figure 1C1-C2). VarScrut is written in python compliant with version 2.7, for the scripting parts to manipulate VCFs and to perform the calculations. For each new exome analysis, a dedicated SQLite database is created to store all the input data and the intermediate results. The monitoring and execution of parallel scheduled tasks, such as the update of datasets, the re-annotation module or the literature-tracking module, are managed with a local Jenkins server (v1.621) for a continuous integration. The datasets used in VarScrut are updated regularly *via* Jenkins jobs: every 3 months for Variant Effect Predictor[27] (VEP) v81 and Ensembl[28] data, every month for all the variants and disease datasets (ClinVar[29], OMIM[30], GeneReviews, HPO[20]) and PubMed articles are tracked on a daily basis. Annotations and computations are executed on a cluster of seven Dell

PowerEdge R720 servers configured with an average of 24 cores, 128G of memory and 8 terabytes of local disk space per server. VarScrut is available online as a public server. For the sake of confidentiality, the variant processing step, including the meta-predictor tool and the multi-scenario module (see Figure 1A) can be downloaded and run on Unix platform. The variant processing step provides an output file that can be submitted online on the VarScrut web server in order to use the rest of the workflow. The knowledge-driven gene prioritization step and the post-processing steps involve an integrated architecture and constantly updated voluminous databases; hence they that are not currently amenable for downloading.

Variant Datasets for VarScrut Meta-Predictor

To train the scoring system of VarScrut meta-predictor, we defined a VarTraining dataset that comprises neutral (Negative set) and deleterious (Positive set) variants (Figure S1).

The Negative set of 92,252 neutral nsSNV variants comprises: i) 39,126 common polymorphism variants from the ExAC project verifying a Minor Allele Frequency (MAF) higher than 10% and a depth coverage higher than 15, in at least 2 different populations and ii) 53,126 rare benign variants from the 1,000 genomes project (1000G) verifying a MAF <1%, with at least 2 homozygous genotypes and a depth coverage >30X.

The Positive set of deleterious variants was initially collected from the ClinVar database[29] (v03/2015) and the Human Gene Mutation Database (HMGD)[31] (V2013.3). After exclusion of non-deleterious variants, 25,918 ClinVar variants with the clinical significance code value (CLNSIG) equal to 5 and 91,952 HGMD variants with the DM (Disease Making) flag were retained. This resulted in 98,000 non-

overlapping deleterious variants from which we excluded the InDel as well as the nsSNV used as a deleterious training set by 3 predictor algorithms (PolyPhen2[32], SIFT[33], MutationAssessor[34]) and by the meta-predictor Combined Annotation Dependent Depletion (CADD)[12]. As the training set of the FatHMM[35] predictor algorithm is not available, nsSNV variants used for the FatHMM[35] training procedures could not be excluded. Finally, the positive set comprises 23,912 'original' deleterious nsSNV (Figure S1).

Similarly, in order to evaluate VarScrut meta-predictor, we defined two evaluation datasets: filtVariBench and filtSwissVar, obtained by excluding from the VariBench[36] dataset and the SwissVar[37] dataset, the variants used in the VarTraining dataset.

Simulated exomes

For the benchmarking of variant analysis tools, two sets of simulated disease exomes were constructed. The first set (SimExome) comprised only simulated disease exomes, while the second set (SimPhenExome) comprised simulated disease exomes with simulated corresponding disease-phenotypes (Figure S2).

SimExome was generated through a three step process. First step, a background exome is randomly selected from the 1000G Project[38] (1,092 exomes from healthy individuals). Second step, a known disease-causing gene is ramdomly selected from the pooled-dataset of ClinVar and HGMD. Third step, randomly selected known deleterious variants, associated to the previously selected gene, is inserted into selected background exome. Depending on the MoI, damaging alleles are inserted in the appropriate copy number (*e.g.* two copies of the same allele for autosomal recessive MoI, one copy for autosomal dominant MoI and one copy of two different alleles affecting the same gene for compound heterozygous autosomal recessive MoI),

and the quality metrics of the closest mapped variant are assigned to it. The whole procedure was repeated 100 times using each time different genes with established disease associations.

SimPhenExome was generated as SimExome with an additional fourth step: For each selected disease-causing gene in the second step, up to 5 associated Human Phenotype Ontology (HPO) terms were randomly selected to simulate a corresponding patient's phenotype.

Real exome datasets

Exomes from six patient cases or families with myopathies with poorly documented phenotypical features were analysed (Table S1). Patients and relatives gave informed consent for the genetic analysis according to the declaration of Helsinki and French legislation (CPP Est IV DC-2012-1693). Genomic DNA was extracted from blood by standard methods. The cohort is composed of different types of myopathies: 3 Centronuclear Myopathy (CNM) families, 1 Central Core Disease myopathy (CCD) family, 1 Reducing Body myopathy (RBM) and 1 Myotubular Myopathy (MTM) family. Exome sequencing was performed at the National Sequencing Centre in Evry, France. Exons of DNA samples were captured using the in-solution SureSelect Target Enrichment System (Agilent, Human All Exon Kits v2), followed by 75bp paired-end sequencing on Illumina HiSeq 2,500. Image analysis and base calling were performed with default parameters of the Illumina RTA v1.14 pipeline. The alignment of clean reads on the human reference genome (hg19/GRCh37) was performed with BWA; the post-alignment processing with Picard and variant calling was performed with GATK (HaplotypeCaller).

Variant Processing Step (Figure 1A)

VarScrut requires two mandatory input files: a VCF containing the list of patient variants and a Pedigree Description (PED) file describing the patient kinship. The VCF file first undergoes a pre-processing procedure to normalise the representation of the variant information. Using the *vt* tool, the original VCF is decomposed so that loci with multiple alleles are expanded into distinct variant records; one record for each reference and alternative allele combination (REF/ALT). The decomposed VCF is then normalized using *vt* so that variants are left aligned and represented using the most parsimonious alleles. Variants are then annotated with gene structure information based on the Variant Effect Predictor (VEP) v81 script with customized parameters and plugins (Table S2). VEP is configured to provide different kinds of information: (1) prediction of variant deleteriousness for each gene isoform using distinct algorithms for InDel (SIFT Indel[39]) and nsSNV (CADD[12], SIFT[33], PPH2[32]), (2) conservation level (phastCons[40] phyloP[41], GERP[42], FitCons[43]), (3) associated frequencies in public databases (1000G[38], dbSNP142[44], EVS6500[45], EXAC v0.3). Additionally, VarScrut indicates whether a given variant is a known disease causing one based on ClinVar[29], HGMD[31], OMIM[30], SwissVar[37] annotations. For variants present in overlapping genes, the prediction for each gene is reported.

VarScrut then imports into a dedicated SQLite database all the variants information, as well as the associated pedigree information describing the kinship among the samples available in the provided Pedigree Description (PED) file.

The next major procedure is the Quality Control (QC) and filtering procedure to reduce the initial list of variants. The QC module filters are applied to each variant according to customizable parameters: (i) a read depth >=10, (ii) a supporting read count >=10, (iii) a ratio of supporting reads >=15%. The QC module is also used to

flag genes that are poorly covered based on a customizable parameter (by default: < 3 QC variants per gene). Next, a polymorphism-filtering module is applied on the variant list using a customizable Minor Allele Frequency (MAF) threshold. The MAF variant frequencies are retrieved from public population databases. By default, the 1000G data are used and common variants with a MAF threshold higher than 1% are ignored for the rest of the analysis process. The polymorphic poorly covered genes are also ignored for the rest of the analysis process. If specified otherwise by the user, MAF from sub-populations (African, American, Asian) can be used with different thresholds. Based on the pedigree information and health status provided in the PED file, the variants are then segregated according to five inheritance scenarios in parallel (autosomal recessive, autosomal dominant, X-linked recessive, X-linked dominant, *de novo*) and are finally ranked according to the associated deleterious score.

Scoring of variants

The VarScrut meta-predictor for deleterious nsSNV is based on a Logit model combining 4 commonly used variant predictor scores (CADD[12], Grantham[46], SIFT[47], PolyPhen-2[48]), and 6 annotation values (phyloP[41], phastCons[40], GERP[42], FitCons[49], the distance to splice site and the haploinsufficiency gene probability). The score generated by the Logit model for each nsSNV is used to calculate a global variant score. The meta-predictor was implemented with the python scikit-learn library[50] v0.16.1.

Finally, to rank the variant, a deleteriousness score for each gene (S_{gv}) (Equation 1) is calculated by retaining the maximum score resulting from the product of the variant class score (C_v) defined as a numerical value and derived from the Sequence

Ontology (Table S3) and the variant score (with S_v = SIFT Indel score for InDels; S_v= VarScrut meta-predictor score for nsSNV).

$$S_{gv} = \max (C_v \times S_v) \quad \text{(Equation 1)}$$

Knowledge-driven Gene Prioritization (Figure 1B)

To complement and improve the variant scoring provided by the variant processing step, VarScrut can integrate additional optional data corresponding to phenotypic data related to the affected individuals and described in normalized HPO terms (Figure 1B). The HPO terms are matched to known human disorders using the Phenomizer approach[51], implemented in VarScrut, in order to estimate the significance of each disease match. Briefly, the Benjamini-Hochberg multiple testing–corrected P values are translated to disease probabilities ($\mathbf{D_{P0}}$) assuming the disease set has a uniform prior (corrected P-values < 0.05). Using the mapping information of disease to genes provided by HPO, each disease significant disease is converted into the list of corresponding genes ($\mathbf{D}_{genes0} = \mathbf{[D_{gi}, D_{gj}, D_{gz}} \ldots.]$) and equal were distributed to each gene, with the sum of the probabilities equal to 1.

We use the NetWalker[52] program to identify potential novel genes not yet associated with a disease but functionally similar to known associated disease-causing genes. NetWalker implements the Random Walk with Restart (RWR) algorithm, previously described by Köhler *et al.*[53], to sweep a functional network of genes, with start probabilities ($\mathbf{D}_{genes0}$), to obtain a list of ranked candidate genes. In VarScrut, the network is a gene functional interaction network, where the nodes represent genes and the edges the functional interactions between genes. The functional interaction network is encoded as an adjacency matrix (W) where the row and columns represent the genes and the cell the potential functional interaction(s) based on: i) similarities of

297 ontology terms (GO[54], HPO[20]), ii) correlated expression (GeneMania[55],
298 COXPRESdb[56]), iii) physical interactions (IntAct[57], HRPD[58], BioGrid[59]), iv) common
299 pathways (KEGG[60], REACTOME[61], NCI-Nature[62]), v) common cellular localizations
300 (Human Protein Atlas[63]), vi) predictive data-mining interactions with high confidence
301 (STRING[64] v9.05 score>=0.7, GeneMania[55]), and vii) common disease groups
302 (OMIM[30] / PhenotypicSeries, GeneReviews). This information is implemented in an
303 undirected graph using the Python library NetworkX v1.9.1.
304 The RWR algorithm calculates a final priority score for each node based on the inertia
305 state probabilities. Random walk with restart is formally defined as the following:

306 $$\mathbf{D}_{genes\,t+1} = (1-r)\,\mathbf{W}\mathbf{D}_{genes\,t} + r\mathbf{D}_{genes\,0} \quad \text{(Equation 2)}$$

307 where r is the probability of a gene to be a starting node (this is equivalent to letting
308 the random walker begin from each of the known disease genes); W is the column-
309 normalized adjacency matrix of the graph network; D_{genest} is a score vector in which
310 the i-th element holds the probability of being at node i at time step t; D_{genes0} is the
311 initial probability vector. The RWR was run, with NetWalker default parameters,
312 until obtaining a steady state P_∞ probability vector, which will be used to calculate
313 the final gene prioritization score (Equation 3). The global gene score is calculated
314 for each gene by combining the scores obtained at each step.
315
316 Formally, the global gene score S_G of gene G is defined as:

317 $$S_G = S_{gv} + D_{genes\,t} \quad \text{(Equation 3)}$$

318

Post-Analysis Step (Figure 1C)

Validation Guidance (Figure 1C1)

Candidate genes associated with an unresolved exome are annotated with information related to the availability of associated biological material for subsequent experimental validations (Figure 1C1). Antibody availability for a candidate protein is obtained from Antibodypedia[65] (20,507 gene products) and Human Protein Atlas databases (for protein evidence in different tissues). Information about the availability of animal models is obtained for: (1) mouse models using the International Mouse Phenotyping Consortium[66] and Mouse Genome Informatics[67] databases, (2) zebrafish models using the Zebrafish Model Organism Database. These data are downloaded monthly using the dump download file available on each website.

Follow-up of Unresolved Exomes (Figure 1C2)

The follow-up module (Figure 1C2) provides additional information concerning the candidate genes associated with an unresolved exome:

1. A literature-tracking module for a daily lookup of articles citing the candidate genes. The detection of the gene symbols in PubMed abstracts is performed using the PubTator[68] service and the resulting lists of articles are filtered according to a user-supplied list of keywords *via* E-utils and gene2pubmed.
2. The polymorphic poorly covered genes are listed if they are associated with the phenotypic traits of affected individuals.
3. For each Ensembl data release (every 3 months), all the variation and annotation databases are updated *via* a Jenkins job.

The user provides an email address that is used in the follow-up module, to push any PubMed news or annotation update that changes the previous ranking of the candidate genes.

Results

In order to build an expert system simplifying and mimicking the clinician/researcher approach for WES analysis, VarScrut integrates not only the two main steps (variant processing and knowledge-driven gene prioritization) of the exome analysis process, but also a post-analysis step to provide experimental validation guidance and follow-up procedures (Figure 1). The expert system VarScrut includes several optimized and evaluated procedures, namely a variant ranking procedure based on a meta-predictor for deleterious nsSNV, or a knowledge-driven gene prioritization procedure, using phenotypic information. Moreover, thanks to its follow-up procedures, VarScrut allows the further processing of unresolved exomes with extended list of candidate genes and/or numerous poorly covered genes, by performing periodic variant re-evaluation using regular variant/gene annotation updates and literature-tracking (PubMed). VarScrut is thus applicable to a wide variety of WES cases, ranging from poorly documented cases with no *a priori* inheritance scenario knowledge, to richly documented exome projects in combination with phenotypic data.

Performance of VarScrut's Meta-Predictor

The performance of VarScrut's meta-predictor for deleterious nsSNV was compared to 4 commonly used predictors (PolyPhen2[48], SIFT[47], MutationAssessor[34], LRT[69]) and 3 meta-predictors (Condel, CADD[12], FATHMM[35]), using filtVariBench and filtSwissVar datasets, which includes rare benign variants (Table 1).

The performances are reported as Area Under the ROC Curve (AUC) values per tool and per dataset. Globally, FATHMM outperforms all tools on filtVariBench with an AUC of 0.938, while the VarScrut Logit predictor is the second-best performing tool with an AUC of 0.735. However, on filtSwissVar, FATHMM shows a severe drop in performance with an AUC of 0.713, while the VarScrut Logit predictor has the best performance with an AUC of 0.759.

Performance of VarScrut's Variant Ranking procedure

Evaluation on SimExome dataset

To evaluate the performance of VarScrut in the case of single exomes with no phenotypical documentation, a benchmark was carried out on SimExome dataset. VarScrut was compared to the VAAST tool for the ranking of the causative gene among the top genes. Since VAAST uses only nsSNV to rank genes, the simulated "disease exomes" have been generated using only deleterious nsSNV. Figure 2 summarizes VarScrut's outperformance over VAAST on 100 single disease exome simulations. As expected, we observed different performances depending on the scenario, and genotype/variant information alone is not sufficiently informative to allow either of the tools to rank the causative gene among the top 5 genes in the case of dominant scenarios. The average rank for dominant and recessive scenarios is respectively the 99^{th} and 64^{th} rank for VAAST and 32^{nd} and 13^{th} rank for VarScrut (Table S4). In the dominant scenario, VAAST was able to rank the causative gene among the top 10 in only 43% of cases compared to 61% of cases for VarScrut. In the recessive scenario, VarScrut also out-performs VAAST with 68% of the cases in the top 10 genes, while VAAST could resolve only 51% of the cases. Overall, VAAST

ranked the causative gene among the top 5 genes in the 2% of the cases, compared to 13% of the cases for VarScrut.

Evaluation on SimPhenExome dataset

Currently, there are only two other available tools that combine automatically both genotype and phenotype information to process variant data: VAAST+PHEVOR and Phen-Gen. PHEVOR uses multiple biomedical ontologies to improve the performance of VAAST by further narrowing the gene list based on ontology terms supplied by the user. For a fair comparison, only Human Phenotype Ontology (HPO) terms were used, as Phen-Gen takes only HPO terms. For each simulated disease exomes in the SimPhenExome dataset, up to 5 random HPO terms associated with the causative gene were selected and submitted with the exomes to each tool, as this is the maximum number of terms taken by PHEVOR's website. The simulations were performed on recessive and dominant scenarios (Figure 3). On 100 simulated cases with the dominant scenarios, by prioritizing VAAST results, PHEVOR could rank the known disease allele among the top 10 genes in 64% of cases, while Phen-Gen and VarScrut ranked 82 and 91% respectively (Table S5). On 100 simulated cases with the recessive scenarios, the causative variant was classified among the top 10 genes in 66% of the cases for VAAST+PHEVOR and 98% and 97% for Phen-Gen and VarScrut respectively. The top 5 ranking for dominant scenarios was reached in 55% of the cases for VAAST+PHEVOR, 76% for Phen-Gen and 78% for VarScrut. However, for the recessive scenarios, almost all of the disease genes identified were among the top 5 when combining variant/genotype–phenotype information. Overall, we conclude that Varscrut provides better results than VAAST+PHEVOR and Phen-Gen for both scenarios.

Analyses of Real Exome Datasets

The performance of VarScrut was also assessed on real datasets, involving exomes from a heterogeneous cohort of 6 myopathy families with poorly documented phenotypic characteristics, as well as unknown inheritance scenarios. For each category of myopathy, at least 5 HPO terms were selected that best characterized the category (Table S6). Overall, for five of the six families (G21660, 5–3829, G14314, G5417, MCMU16032), VarScrut was able to identify the disease-causing gene among the top 10 candidate genes and in it addition, identified the inheritance scenarios, for which no *a priori* parameter was set (Table 2). All the variations have been experimentally validated and the inheritance scenarios confirmed by segregation analysis in the families. Phen-Gen (the other best performing tool on SimPhenExome dataset) was also run on the real exome datasets with default parameters and the identified scenario was specified in parameter. Phen-Gen was able to identify the causative gene among the top 10 genes in only three of the families (G21660, 5–3829, MCMU16032). Phen-Gen also identified the causative gene for a fourth family, but it was ranked at the 28^{th} position. For the remaining families, Phen-Gen failed to propose any solution.

Discussion

We have developed an expert system, VarScrut, to address the unmet needs for a complete WES analysis tool proposing a unified process, ranging from variant filtering and ranking up to the post-analysis step. The post-analysis step includes a validation guidance procedure for resolved exomes and a follow-up procedure for unresolved exomes.

439 The initial step of the exome analysis involves the characterization and annotation of
440 variants in order to obtain a proper interpretation of their potential deleteriousness.
441 For a more reliable characterization of nsSNV, the most common type of variant, we
442 designed a dedicated meta-predictor following the ACMG recommendations to
443 combine the scores of several predictors in order to achieve more confidence in the
444 interpretation[23,24]. Although the strategies implemented in existing predictors can
445 differentiate rare deleterious nsSNV from common neutral variants (MAF >10%),
446 VarScrut's meta-predictor also tries to differentiate rare benign variants (MAF <1%)
447 in its VarTraining dataset. We highlighted the robust performance of the VarScrut
448 meta-predictor compared to other popular predictors and meta-predictors. Currently,
449 there are few predictors and annotations available to evaluate the impact of InDel. In
450 VarScrut, we chose the SIFT-Indel scoring, as it is currently the only downloadable
451 InDel predictor.
452 Furthermore, other factors such as the MoI and its related assumptions also influence
453 the analysis strategy. As single affected individuals and small nuclear families are the
454 most frequently encountered cases in the clinic, this situation can result in an
455 unknown or dubious inheritance scenario. VarScrut thus offers the possibility to
456 screen several inheritance scenarios in parallel when the user does not want to make
457 any genetic assumptions. This procedure takes into account previous
458 recommendations, with an optimized analysis strategy applied for each scenario[26,70].
459 The evaluation of performances on SimExome dataset showed that the VarScrut
460 ranking procedure out-performs VAAST. This is due to the difference in strategy for
461 the evaluation of deleterious nsSNV. VarScrut integrates different evaluation criteria
462 into a single scoring system *via* its meta-predictor, while VAAST's disease-associated
463 probability is based on the observed allele frequency in 1000G and amino-acid

conservation. Globally, these results also show that the variant-level analysis is informative in only 60–70% of the cases when considering the top 10 ranked genes. Hence, additional knowledge is necessary to improve the ranking performance and increase the resolution rate.

More recently, some gene prioritization methods have been introduced that exploit additional data, such as phenotypic data or model organism data[71] to improve the final result. Random-walk analysis of protein–protein interaction data has been shown to be a powerful approach for gene prioritization[53,72]. In VarScrut, we adopted a Random Walk with Restart approach, over a gene functional interaction network and initialized by the most probable disease genes corresponding to the patient's phenotype. When combining phenotypic and genotype/variant data, we showed that this approach substantially improves the top 10 genes ranking from 60-70% to 91–97% of the cases. Both Phen-Gen and VarScrut performed better than VAAST + PHEVOR in both recessive and dominant scenarios. This can be attributed to the fact that both Phen-Gen and VarScrut integrate different functional data on gene functional interactions, such as pathway information or protein-protein interacting data. PHEVOR is currently based solely on ontological data, which do not completely reflect all the different types of biological interactions, and curated data. Globally, VarScrut has a better ranking performance than Phen-Gen on real exome datasets. VarScrut was able to identify the causative gene among the top 10 in six of the seven families, compared to only four of them for Phen-Gen. The functional interaction network of VarScrut benefits from the diversity of sources used and the continuous integration of updates.

WES analysis remains a complex procedure based on several assumptions, where many previously reported pitfalls[26,70] lead to unresolved exomes. To address this

489 issue, the post-analysis step in VarScrut was based on the ACMG's recommendations
490 for the follow-up of unresolved exomes. As for other high-throughput approaches,
491 quality control is a crucial part of the process when analysing the data. For VarScrut,
492 in a clinical context, we chose to work only with VCF. VarScrut takes into account
493 both candidate genes and poorly covered genes in its literature tracking module,
494 which will report any publications linked to any genes on the watch list based on a
495 provided list of biological context keywords. Any article reported is supplied with a
496 list of detected concepts (disease, gene, species, mutation) obtained *via* PubTator to
497 facilitate the screening (data not shown).
498 In addition to the implementation of various robust solutions at each step for an
499 efficient integrated exome analysis, some architectural choices were also made to
500 enhance and maintain the performance of VarScrut, such as the adoption of a
501 continuous integration approach *via* a Jenkins server, to allow a permanent update of
502 all the different annotation sources even with different/asynchronous periodicity. This
503 mode of implementation offers the opportunity to leverage the available
504 computational capacities with the possibility of executing several jobs in parallel,
505 distributed across a cluster of nodes. To store the intermediate and final results, a
506 SQLite relational database is used because of its expressive power and the enhanced
507 portability provided. Thus, a given VarScrut database can be exported as a single file
508 and can be reused without a dedicated database server or additional configuration.
509
510 To conclude, VarScrut represents some significant advances over current approaches
511 for exome sequencing analysis in Mendelian disorders by providing an effective and
512 automated multi-scenario approach associated with an integrative framework
513 combining prior knowledge and sequence data. In this article, we have highlighted its

capacity to resolve exome-sequencing projects on both simulated cases and on real datasets. Moreover, with its post-analysis step including the automatic relaunch and follow-up modules, VarScrut is the only tool to our knowledge that addresses the challenges of unresolved exome projects. Future improvements of VarScrut will include additional scenarios and supplementary annotations such as non-coding RNA and regulatory information for genome analysis, and the possibility of exploring the results in an integrative web interface.

Acknowledgements

VarScrut was developed with the support of ANR Investissement d'Avenir Bioinformatique BIP:BIP (ANR10-BINF03-05) and the sequencing of the myopathy exomes was supported by Fondation Maladies Rares within the Myocapture sequencing project (ANR-10-INBS-09). We thank Dominik Grimm for benchmark variation datasets, Anne Boland and Jean-François Deleuze for the exome sequencing, Chrystel Chéraud, John Rendu, Johann Bohm and Valérie Biancalana for mutation validations, Norma Romero, Elisabeth Ollagnon Roman, Valérie Drouin Garraud, Michel Fardeau, Pascal Laforêt and Pascale Guicheney for patient evaluation and classification.

Web Resources

The URLs for datasets and tools used are as follows:

Antibodypedia: http://www.antibodypedia.com/

E-utilities: www.ncbi.nlm.nih.gov/books/NBK25500

ExAC: http://exac.broadinstitute.org/

gene2pubmed: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz

HPO: http://compbio.charite.de/hudson/job/hpo.annotations.monthly/lastStableBuild/

Human Protein Atlas: http://www.proteinatlas.org/

International Mouse Phenotyping Consortium: http://www.mousephenotype.org/

Jenkins server: https://jenkins-ci.org/

Mouse Genome Informatics: http://www.informatics.jax.org/

OMIM: http://www.omim.org/

PubMed: http://www.ncbi.nlm.nih.gov/pubmed

PubTator: http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator

Scikit-learn library: http://scikit-learn.org/stable/

SwissVar: http://swissvar.expasy.org/

VarScrut: http://varscrut.lbgi.fr

vt tool: http://genome.sph.umich.edu/wiki/Vt

Zebrafish Model Organism Database: http://zfin.org/

References

1. Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. *20*, 490–497.

2. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics *12*, 745–755.

3. Zhang, X. (2014). Exome sequencing greatly expedites the progressive research of Mendelian diseases. Front Med *8*, 42–57.

4. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. The American Journal of Human Genetics *97*, 199–215.

5. Gahl, W.A., Markello, T.C., Toro, C., Fajardo, K.F., Sincan, M., Gill, F., Carlson-Donohoe, H., Gropman, A., Pierson, T.M., Golas, G., et al. (2012). The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet. Med. *14*, 51–59.

6. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A.,

Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. New England Journal of Medicine *369*, 1502–1511.

7. Bloch-Zupan, A., Jamet, X., Etard, C., Laugel, V., Muller, J., Geoffroy, V., Strauss, J.-P., Pelletier, V., Marion, V., Poch, O., et al. (2011). Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. The American Journal of Human Genetics *89*, 773–781.

8. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

9. Cooper, G.M. (2015). Parlez-vous VUS? Genome Research *25*, 1423–1426.

10. Houdayer, C., Caux-Moncoutier, V., Krieger, S., Barrois, M., Bonnet, F., Bourdon, V., Bronner, M., Buisson, M., Coulet, F., Gaildrat, P., et al. (2012). Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. Human Mutation *33*, 1228–1238.

11. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation *32*, 358–368.

12. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics *46*, 310–315.

13. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation *34*, 57–65.

14. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Human Mutation *34*, E2393–E2402.

15. Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation *36*, 513–523.

16. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research *38*, e164–e164.

17. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Research *21*, 1529–1542.

18. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

19. Paila, U., Chapman, B.A., Kirchner, R., and Quinlan, A.R. (2013). GEMINI: integrative exploration of genetic variation and genome annotations. PLOS Comput Biol *9*, e1003153.

20. Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research *42*, D966–D974.

21. Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. The American Journal of Human Genetics *94*, 599–610.

22. Javed, A., Agrawal, S., and Ng, P.C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nature Methods *11*, 935–937.

23. Richards, C.S., Bale, S., Bellissimo, D.B., Das, S., Grody, W.W., Hegde, M.R., Lyon, E., Ward, B.E., Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet. Med. *10*, 294–300.

24. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. (Nature Publishing Group), pp. 405–424.

25. Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M.A., Pugh, T.J., Funke, B.H., Rehm, H.L., and Lebo, M.S. (2013). A systematic approach to assessing the clinical significance of genetic variants. Clin. Genet. *84*, 453–463.

26. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.

27. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.

28. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Research *44*, D710–D716.

649 29. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M.,
650 and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence
651 variation and human phenotype. Nucleic Acids Research *42*, D980–D985.

652 30. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A.
653 (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online
654 catalog of human genes and genetic disorders. Nucleic Acids Research *43*, D789–
655 D798.

656 31. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N.
657 (2014). The Human Gene Mutation Database: building a comprehensive mutation
658 repository for clinical and molecular genetics, diagnostic testing and personalized
659 genomic medicine. Hum. Genet. *133*, 1–9.

660 32. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork,
661 P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting
662 damaging missense mutations. Nature Methods *7*, 248–249.

663 33. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding
664 non-synonymous variants on protein function using the SIFT algorithm. Nature
665 Protocols *4*, 1073–1081.

666 34. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of
667 protein mutations: application to cancer genomics. Nucleic Acids Research *39*, e118–
668 e118.

669 35. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N.M., and Gaunt, T.R. (2013).
670 Predicting the functional consequences of cancer-associated amino acid substitutions.
671 Bioinformatics *29*, 1504–1510.

672 36. Sasidharan Nair, P., and Vihinen, M. (2013). VariBench: a benchmark database
673 for variations. Human Mutation *34*, 42–49.

674 37. Mottaz, A., David, F.P.A., Veuthey, A.-L., and Yip, Y.L. (2010). Easy retrieval of
675 single amino-acid polymorphisms and phenotype information using SwissVar.
676 Bioinformatics *26*, 851–852.

677 38. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D.,
678 DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and
679 McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human
680 genomes. Nature *491*, 56–65.

681 39. Hu, J., and Ng, P.C. (2013). SIFT Indel: predictions for the functional effects of
682 amino acid insertions/deletions in proteins. Plos One *8*, e77940.

683 40. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K.,
684 Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily
685 conserved elements in vertebrate, insect, worm, and yeast genomes. Genome
686 Research *15*, 1034–1050.

687 41. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection
688 of nonneutral substitution rates on mammalian phylogenies. Genome Research *20*,

689 110–121.

690 42. Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing
691 Program, Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity
692 of constraint in mammalian genomic sequence. Genome Research *15*, 901–913.

693 43. Henn, B.M., Botigué, L.R., Bustamante, C.D., Clark, A.G., and Gravel, S. (2015).
694 Estimating the mutation load in human genomes. Nature Reviews Genetics *16*, 333–
695 343.

696 44. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M.,
697 and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic
698 Acids Research *29*, 308–311.

699 45. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S.,
700 McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of
701 rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

702 46. Grantham, R. (1974). Amino acid difference formula to help explain protein
703 evolution. Science *185*, 862–864.

704 47. Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect
705 protein function. Nucleic Acids Research *31*, 3812–3814.

706 48. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect
707 of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*,
708 Unit7.20–7.20.41.

709 49. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for
710 calculating probabilities of fitness consequences for point mutations across the human
711 genome. Nature Genetics *47*, 276–283.

712 50. Hauck, T. (2014). scikit-learn Cookbook (Packt Publishing Ltd).

713 51. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos,
714 C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human
715 genetics with semantic similarity searches in ontologies. The American Journal of
716 Human Genetics *85*, 457–464.

717 52. Zhang, B., Shi, Z., Duncan, D.T., Prodduturi, N., Marnett, L.J., and Liebler, D.C.
718 (2011). Relating protein adduction to gene expression changes: a systems approach.
719 Mol Biosyst *7*, 2118–2127.

720 53. Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the
721 interactome for prioritization of candidate disease genes. The American Journal of
722 Human Genetics *82*, 949–958.

723 54. Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward.
724 Nucleic Acids Research *43*, D1049–D1056.

725 55. Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and
726 Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic Acids

727 Research *41*, W115–W122.

728 56. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and
729 Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species
730 by DNA-microarray and RNAseq-based expression data with multiple quality
731 assessment systems. Nucleic Acids Research *43*, D82–D86.

732 57. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter,
733 F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct
734 project--IntAct as a common curation platform for 11 molecular interaction databases.
735 Nucleic Acids Research *42*, D358–D363.

736 58. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S.,
737 Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009).
738 Human Protein Reference Database--2009 update. Nucleic Acids Research *37*, D767–
739 D772.

740 59. Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S.,
741 Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The
742 BioGRID interaction database: 2015 update. Nucleic Acids Research *43*, D470–
743 D478.

744 60. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and
745 genomes. Nucleic Acids Research *28*, 27–30.

746 61. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw,
747 R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome
748 pathway Knowledgebase. Nucleic Acids Research *44*, D481–D487.

749 62. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and
750 Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids
751 Research *37*, D674–D679.

752 63. Colwill, K., Renewable Protein Binder Working Group, and Gräslund, S. (2011).
753 A roadmap to generate renewable protein binders to the human proteome. Nature
754 Methods *8*, 551–558.

755 64. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-
756 Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING
757 v10: protein-protein interaction networks, integrated over the tree of life. Nucleic
758 Acids Research *43*, D447–D452.

759 65. Björling, E., and Uhlen, M. (2008). Antibodypedia, a portal for sharing antibody
760 and antigen validation data. Mol. Cell Proteomics *7*, 2028–2037.

761 66. Koscielny, G., Yaikhom, G., Iyer, V., Meehan, T.F., Morgan, H., Atienza-
762 Herrero, J., Blake, A., Chen, C.-K., Easty, R., Di Fenza, A., et al. (2014). The
763 International Mouse Phenotyping Consortium Web Portal, a unified point of access
764 for knockout mice and related phenotyping data. Nucleic Acids Research *42*, D802–
765 D809.

766 67. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., Mouse Genome

767 Database Group (2015). The Mouse Genome Database (MGD): facilitating mouse as
768 a model for human biology and disease. Nucleic Acids Research *43*, D726–D736.

769 68. Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining
770 tool for assisting biocuration. Nucleic Acids Research *41*, W518–W522.

771 69. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three
772 human genomes. Genome Research *19*, 1553–1561.

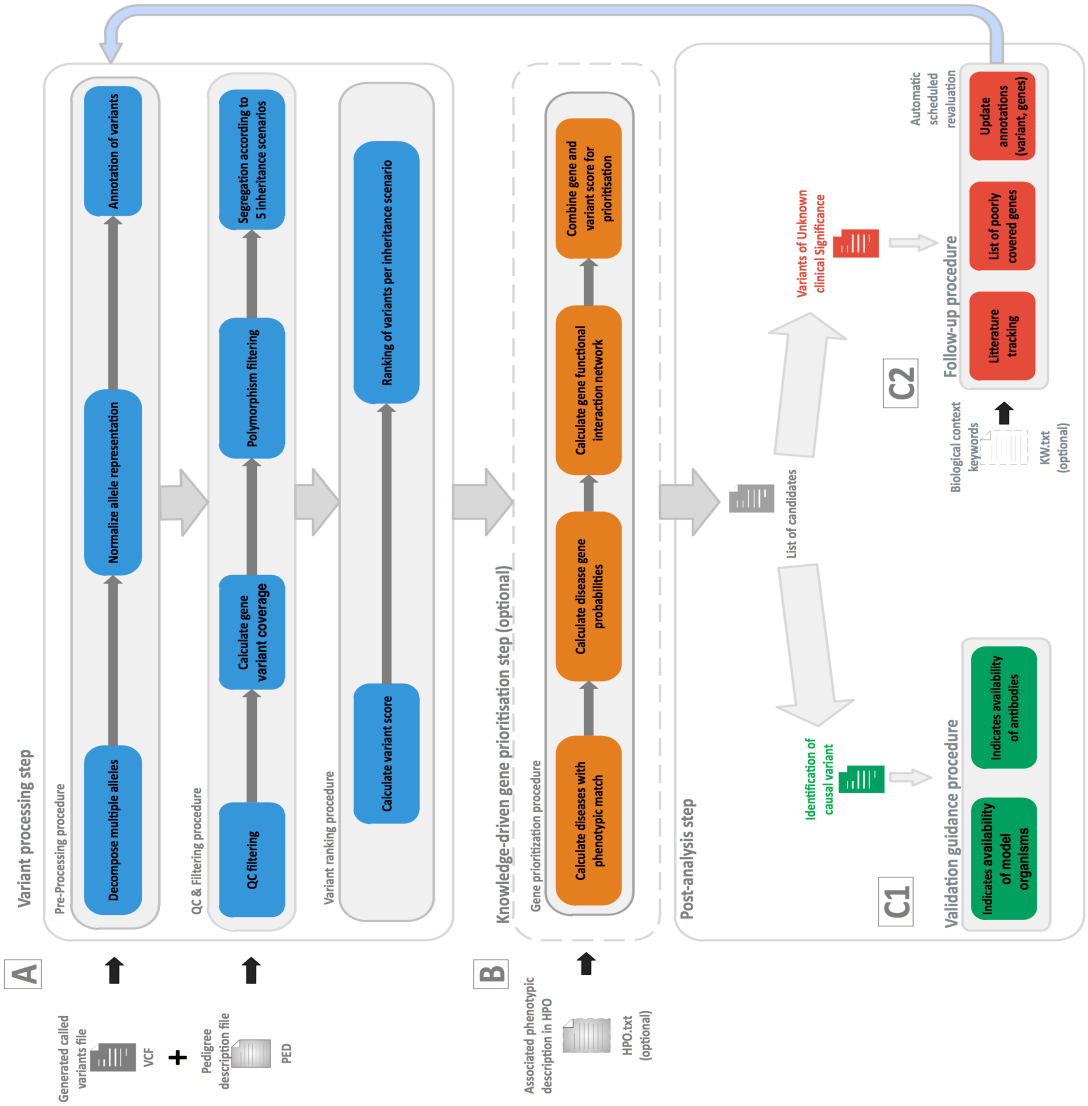
773 70. Ku, C.-S., Naidoo, N., and Pawitan, Y. (2011). Revisiting Mendelian disorders
774 through exome sequencing. Hum. Genet. *129*, 351–370.

775 71. Robinson, P.N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, Wang,
776 K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., et al. (2014).
777 Improved exome prioritization of disease genes through cross-species phenotype
778 comparison. Genome Research *24*, 340–348.

779 72. Navlakha, S., and Kingsford, C. (2010). The power of protein interaction
780 networks for associating genes with diseases. Bioinformatics *26*, 1057–1063.

781

782



783

784 **Figure 1: Workflow of the 3 major steps of VarScrut.** A) The variant processing
785 step (blue) with identification and prioritization procedures according to 5 'potential'
786 inheritance scenarios, takes as input two mandatory files: a VCF format file
787 containing the variants and a PED format file describing the relationship between
788 samples in the VCF. The processing and the filtering procedures are based on variant
789 information only. B) The gene prioritization step (orange) is based on a knowledge-
790 driven approach and takes as input a tabular file containing the affected phenotype in
791 HPO terms. C) The post-analysis step (green) comprises: (1) a validation module
792 which annotates the list of candidate genes according to the availability of the
793 biological material (antibodies, animal models) for each gene, and (2) a follow-up

module (red) which provides support for regular re-evaluation of a given list of variants on the basis of annotation updates and a literature tracking based on a user's-defined list of keywords.

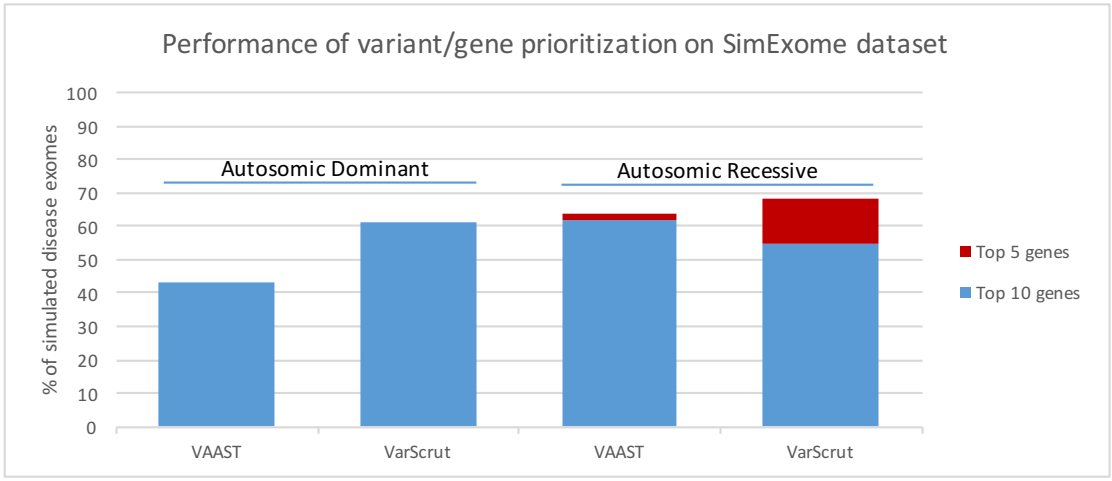


Figure 2: Comparison of VarScrut with VAAST for the prioritization of disease variants using SimExome dataset for recessive and dominant scenarios. The ability of each method to rank the causative gene within the top 5 (red) or top 10 (blue) genes over 100 simulated exomes *per* scenario is indicated by the height of the bars.

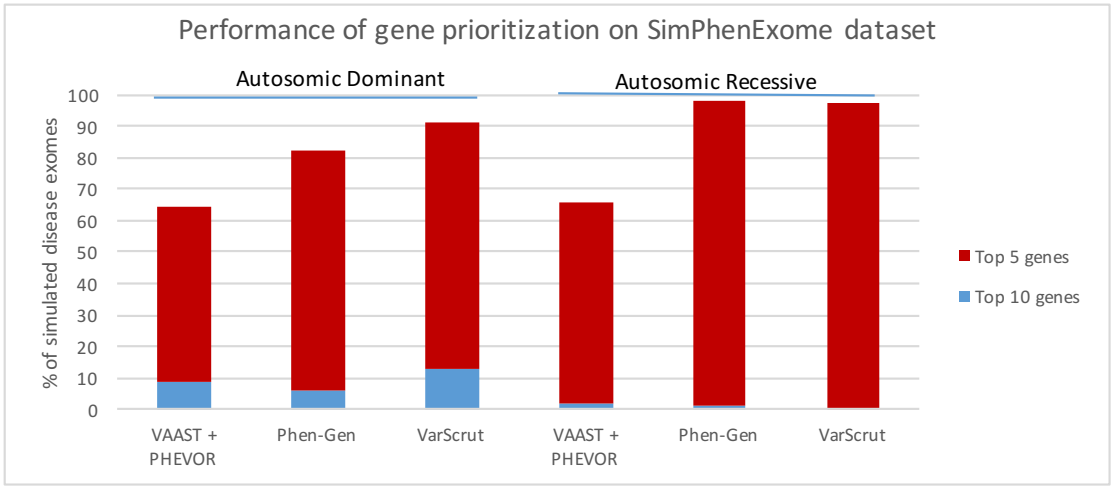


Figure 3. Comparison of VarScrut with VAAST+PHEVOR and Phen-Gen for the prioritization of disease variants using SimPhenExome dataset for recessive

and dominant scenarios. The ability of each method to rank the causative gene within the top 5 (red) or top 10 (blue) genes over 100 simulated exomes *per* scenario is indicated by the height of the bars.

Algorithm		Area under the ROC curve	
Category	**Name**	**filtVariBench dataset**	**filtSwissVar dataset**
Predictor	SIFT	0.6981	0.6826
	PolyPhen2	0.6822	0.7118
	Condel	0.6989	0.7031
	LRT	0.6199	0.6793
Metapredictor	MutationAssessor	0.6969	0.6847
	CADD	0.6634	0.7287
	FATHMM	0.9376	0.7130
	VarScrut Logit	0.7349	0.7587

Table 1. Performance comparison of different predictors of nsSNV deleteriousness on filtVariBench and filtSwissVar datasets. The performance is expressed in terms of the area under the ROC curves (sensitivity v/s 1-specificity).

Family_id	Myopathy Category	Scenario	Gene	VarScrut gene rank	Phen-Gen gene rank
G21660	CNM	Autosomal Recessive	TTN:chr2_179550326_C_T, splice acceptor; TTN:chr2_179444429_G_A, stop gain	Top 5	Top 5
5–3829	MTM	Autosomal Recessive	RYR1:chr19_38934252_C_T, missense; RYR1:chr19_38987595_G_T, splice donor	Top 5	Top 5
G14314	CCD	Autosomal Recessive	RYR1:chr19_39002235_ C_T, stop gain; RYR1:chr19_38954133_T_C, missense	Top 5	NA
G5417	CNM	Autosomal Recessive	DYSF: chr2_71894617_CAGCC_-, frameshift homozygous	Top 10	NA
MCMU16032	CNM	*de novo*	MTM1: chrX_149828138_G_A, misense	Top 5	Top 5
G23031	RBM	X-Linked Dominant	FHL1:chrX_135290014_ G_T, missense	12	28

Table 2. Comparison of VarScrut with Phen-Gen for the analysis of myopathy families with poorly documented phenotypic data. The causative gene and the validated mutations are reported with the corresponding scenarios. The performance of VarScrut and Phen-Gen is reported if the causative gene is among the top 5 genes or the top 10 genes, otherwise the numerical rank is specified. NA indicates that no results were provided by Phen-Gen.

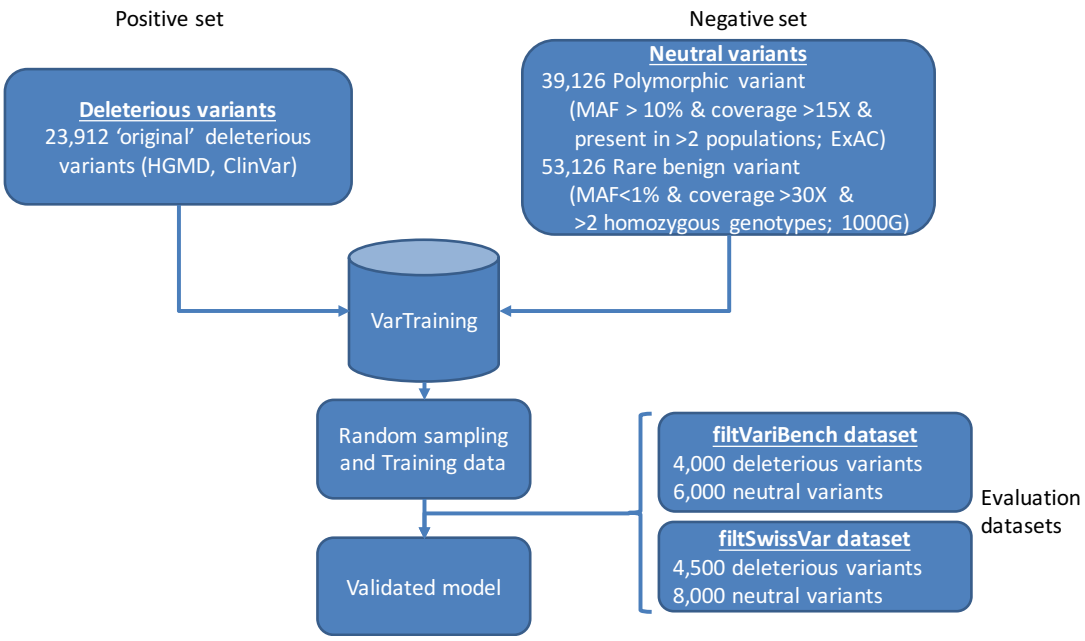


Figure S1: Protocol for training and evaluating the VarScrut meta-predictor for deleterious nsSNV based on a Logit model.

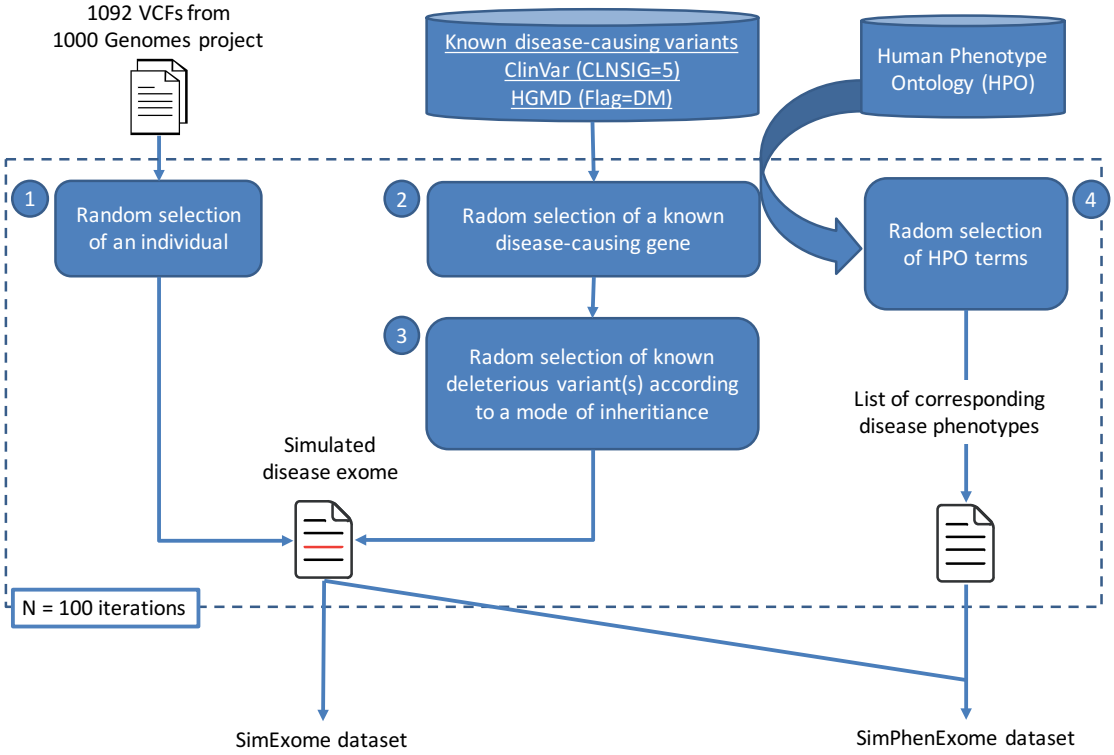


Figure S2: Protocol to simulate disease exome variants for the benchmark analysis.

Family id	Individual id	Family position	Sex	Health status	Myopathy type
5-3829	B00FZKT	Child	Male	Affected	MTM
	B00FZKU	Father	Male	Unaffected	-
	B00FZKV	Mother	Female	Unaffected	-
G23031	B00FZEZ	Mother	Female	Affected	RBM
	B00FZF0	Child	Female	Affected	RBM
G21660	B00FZF3	Father	Male	Unaffected	-
	B00FZF4	Mother	Female	Unaffected	-
	B00FZF5	Child	Male	Affected	CNM
	B00FZF6	Child	Male	Affected	CNM
MCMU16032	B00FZMQ	Child	Female	Affected	CNM
	B00FZMR	Father	Male	Unaffected	-
	B00FZMS	Mother	Female	Unaffected	-
G5417	B00FZMI	Child	Female	Affected	CNM
G14314	B00G45U	Child	Female	Affected	CCD

Table S1. Description of the cohort of myopathy patients used for the evaluation of VarScrut and Phen-Gen as real disease exome dataset. The cohort includes four main types of myopathy: (i) Central Core Disease (CCD), (ii) Centro-Nucleolar Myopathy (CNM), and (iii) Myotubular Myopaphy (MTM), Reducing Body myopathy (RBM).

```
## general features flags
assembly GRCh37
cache 1
dir_cache$vep_dir/cache
dir_plugins $vep_dir/cache/Plugins
fasta $vep_dir/cache/homo_sapiens/79_GRCh37/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa
force_overwrite 1
fork 40
offline 1
species homo_sapiens
terms so
vcf 1
## output annotation flags
allele_number 1
biotype 1
canonical 1
ccds 1
domains 1
hgvs 1
numbers 1
polyphen b
pubmed 1
regulatory 1
sift b
symbol 1
total_length 1
## Co-located variants flags
check_alleles 1
check_existing 1
check_ref 1
flag_pick 1
pick_order canonical, tsl, biotype, rank
biotype 1
## plugins
plugin Blosum62
plugin CADD,$vep_dataset/cadd/lite/whole_genome_SNVs.tsv.gz,$vep_dataset/cadd/lite/InDels.tsv.gz
plugin Carol
plugin Condel,$vep_dir/cache/Plugins/Condel/config,b
plugin ExAC,$vep_dataset/population_db/exac_latest_tidy.vcf.gz
## Custom annotations
custom $vep_dataset/disease_db/clinvar_latest_tidy.vcf.gz,CLINVAR,vcf,exact,0,CLNSIG,CLNACC
custom
        $vep_dataset/population_db/1KG_phase3_tidy.vcf.gz,1KG,vcf,exact,0,AF,EAS_AF,AMR_AF,AFR_AF,EUR_A
F,SAS_AF
custom $vep_dataset/population_db/dbsnp141_tidy.vcf.gz,dbSNP141,vcf,exact,0
custom $vep_dataset/population_db/esp6500_tidy.vcf.gz,ESP6500,vcf,exact,0,MAF
custom $vep_dataset/ucsc_encode/All_hg19_RS.bw,GERP_RS,bigwig
custom $vep_dataset/ucsc_encode/hg19.100way.phastCons.bw,PHASTCONS,bigwig
custom $vep_dataset/ucsc_encode/hg19.100way.phyloP100way.bw,PHYLOP,bigwig
custom $vep_dataset/ucsc_encode/hg19_fitcons_fc-i6-0_V1-01.bw,FITCONS,bigwig
```

Table S2. Variant Effect Predictor configuration.

Variant Class	Variant Impact	Variant Class coefficient
Transcript ablation	HIGH	0.98
Frameshift variant	HIGH	0.95
Splice acceptor variant	HIGH	0.90
Splice donor variant	HIGH	0.90
Stop gained	HIGH	0.90
Stop lost	HIGH	0.75
Transcript amplification	HIGH	0.70
Inframe insertion	MODERATE	0.60
Inframe deletion	MODERATE	0.60
Missense variant	MODERATE	0.60
TFBS ablation	MODERATE	0.60
Regulatory region ablation	MODERATE	0.50
Start lost	LOW	0.40
Splice region variant	LOW	0.30
Incomplete terminal codon variant	LOW	0.30
Stop retained variant	LOW	0.30
Synonymous variant	LOW	0.25
TFBS amplification	MODIFIER	0.20
TF binding site	MODIFIER	0.20
Regulatory region variant	MODIFIER	0.20
NMD transcript variant	MODIFIER	0.15
Non coding transcript variant	MODIFIER	0.15
Regulatory region amplification	MODIFIER	0.15
Mature miRNA variant	MODIFIER	0.10
5 prime UTR variant	MODIFIER	0.10
3 prime UTR variant	MODIFIER	0.10
Non coding transcript exon variant	MODIFIER	0.10
Intron variant	MODIFIER	0.10
Upstream gene variant	MODIFIER	0.10
Downstream gene variant	MODIFIER	0.10
Feature elongation	MODIFIER	0.10
Feature truncation	MODIFIER	0.10
Intergenic variant	MODIFIER	0.10

Table S3. Variant class coefficient scores used in VarScrut to evaluate variants.

		Dominant scenarios			Recessive scenarios	
Tools	Average rank	Top 5 gene (%)	Top 10 gene (%)	Average rank	Top 5 gene (%)	Top 10 gene (%)
VAAST	99	0	43	64	2	64
VarScrut	32	0	61	13	13	68

Table S4: Comparison of VarScrut with VAAST for the prioritization of disease variants using SimExome dataset for recessive and dominant scenarios.

	Dominant scenarios			Recessive scenarios		
Tools	Average rank	Top 5 gene (%)	Top 10 gene (%)	Average rank	Top 5 gene (%)	Top 10 gene (%)
VAAST + PHEVOR	22	55	64	14	64	66
Phen-Gen	13	76	82	5	97	98
VarScrut	11	78	91	3	97	97

Table S5: Comparison of VarScrut with VAAST+PHEVOR and Phen-Gen for the prioritization of disease variants using SimPhenExome dataset for recessive and dominant scenarios.

Myopathy class	HPO term	Term description
MTM	HP:0003700	Generalized amyotrophy
MTM	HP:0001270	Motor delay
MTM	HP:0001319	Neonatal hypotonia
MTM	HP:0002747	Insufficiency due to muscle weakness
CCD	HP:0003324	Generalized muscle weakness
CCD	HP:0001270	Motor delay
CCD	HP:0003198	Myopathy
CCD	HP:0003202	Skeletal muscle atrophy
CCD	HP:0003803	Type 1 muscle fiber predominance
CNM	HP:0003202	Skeletal muscle atrophy
CNM	HP:0006829	Severe muscular hypotonia
CNM	HP:0000467	Neck muscle weakness
CNM	HP:0003324	Generalized muscle weakness
CNM	HP:0002650	Scoliosis

Table S6. List of human phenotype ontology (HPO) terms used to prioritize the genes in the exome evaluation by both VarScrut and Phen-Gen. The HPO terms were randomly selected among all the HPO terms associated to each type of myopathy: Central Core Disease (CCD), Centro-Nucleolar Myopathy (CNM), and MyoTubular Myopaphy (MTM).

5.3 Applications

5.3.1 *Discovery of the 18th BBS gene*

As presented in Chapter 2.5.3, the emblematic ciliopathy, Bardet-Biedl syndrome (BBS) is characterized by pleiotropic clinical symptoms and an extensive genetic heterogeneity. To date, 20 BBS genes have been identified, with the mutations BBS1:M390R and BBS10:C91fsX95, each accounting for about 20-25% of the mutation load in European families. However, analysis of a BBS cohort reveals that around 20% of the BBS patients are still without diagnosis, suggesting that some novel BBS genes are still to be discovered (Muller et al., 2010).

VarScrut was hence used on a sporadic case of BBS (see published manuscript in Annexe II), in which the proband exhibits four major clinical symptoms (retinitis pigmentosa, obesity, kidney failure, cognitive disability) and one minor symptom (brachydactyly). After the exome sequencing of the patient and the application of VarScrut with a suspected scenario of autosomal recessive, VarScrut was able to identify as single top candidate, a null mutation (BBIP1:p.Leu58*) in the gene BBIP1, a binding protein of the BBSome complex, which was later described as the 18th BBS gene (BBS18). The disease-causing status of the novel BBS18 gene was confirmed by validation in a zebrafish model.

5.3.2 *Identification of the gene VSP15 linking autophagy to ciliogenesis*

Primary cilia, which function as sensory and signalling organelles, are generated at basal bodies and require intraflagellar transport (IFT) for elongation, maintenance and function. Autophagy is a conserved catabolic process with an essential function in the maintenance of cellular and tissue homeostasis. It is primarily recognised for its role in the degradation of dysfunctional proteins and unwanted organelles, however in recent years the range of autophagy substrates has also been extended to lipids (for a review see (Ward et al., 2016)). Even though autophagy and ciliogenesis processes both participate actively in the remodelling of the cytoskeleton (Wrighton, 2013), it was not until 2013 that it was proven that there is a functional interaction between autophagy and ciliogenesis (Pampliega et al., 2013; Tang et al., 2013).

VarScrut was used in the familial case of three affected siblings with a ciliopathy phenotype and a high probability of consanguinity. The symptoms exhibited by the affected

individual are prognathism, a retinitis pigmentosa and kidney dysfunction. After exome sequencing of the three siblings and analysis and comparison of the exome data, VarScrut was able to identify among the top 3 candidates, a missense mutation in the VPS15/PIK3R4 gene (VPS15:p.998Q), which turns out to be a novel ciliopathy gene, after subsequent validations in the patient samples and in a zebrafish model (see submitted manuscript in Annexe III). With this identification, VPS15, which is a regulatory sub-unit of the phosphoinositide 3-kinase and plays a role in autophagy, we were able to demonstrate for the first time that the deregulation of autophagy (mislocalization of VPS15 in the cilia, due to the loss of an IFT20 dependent localization) could result in a ciliopathy.

5.4 Discussion

In this section, in complement to the manuscript provided, I discuss complementary aspects of the VarScrut design and notably (i) the technical choice of a SQLite-based architecture; (ii) the choice of VCF as input and how we can still gather some additional information; (iii) some feedbacks on the data scarcity and its impact on the machine-learning steps and what may be some directions for future improvement of VarScrut performance.

5.4.1 *An SQLite-based application provides refined queries*

In its design, VarScrut is organized around a SQLite embedded database. From the beginning, VarScrut stores all the variants initially present in the input file and afterwards, performs SQL-based operations (e.g. the filtering of polymorphic variants with MAF >1%) to perform the successive filtering and prioritization operations. The two most popular and efficient tools (VAAST and Phen-Gen to which VarScrut has been compared in the manuscript) rely on an in-memory approach to perform similar operations. Consequently, Phen-Gen and VAAST tend to apply their filtering operations directly on the initial lists of variants to reduce the list of candidate variants and limit the memory requirements, hence eliminating intermediate results and variants not fulfilling the filtering thresholds.

I could verify that this technical choice had various impacts on the performances of the tools. In the in-memory approach, with the systematic shortening of the variant list in the cascade of operations, these algorithms tend to perform faster (~5-15 min) depending on the inheritance scenario. In comparison, the SQLite-based approach has a slower running time

(~15-45 min). However, for the application on real exomes (see Table 1 in the VarScrut manuscript), Phen-Gen was unable to propose any solutions for the families G14314, G5417, 24,820. Due to the in-memory approach used in Phen-Gen, it eliminates 'good' candidates resulting from variant filtering criteria and provides no result after applying the knowledge filter, which takes into account the phenotypic symptoms. On the contrary, the SQLite-based approach used in VarScrut, allows the conservation of any successive candidate list, thus allowing the presentation of the intermediate list (*e.g.* resulting from the filtering and prioritization based on variant information only) as the final candidate list.

The slow runtime of VarScrut is currently due to the literal representation of the variant information (*i.e.* chromosome, start, stop, reference, alternate allele). A possible future optimization that will boost VarScrut's query performance is to represent the variant information in a bit-wise mode (Layer et al., 2015). In the near future, all SQL-based applications should incorporate bit-wise solutions as a subsequent evolution in order to prepare for the inexorable switch towards genome applications or cohort studies.

5.4.2 *VCF as main source of variant information*

Another crucial technical choice in VarScrut is related to the adoption of VCF format files (described in section 2.4.2.1-3) as sole source of variant information. VCF files contain only called-variants data that are different from the reference genome (~40-100 MB), as opposed to the SAM/BAM file that stores all the aligned sequenced data to the reference genome (~8-18 GB). SAM/BAM files can be handled by the GATK tool for example (McKenna et al., 2010). This technical choice was motivated by the routine experience of biologists or clinicians who rarely manipulate or even have access to the BAM file. Hence, in this initial release of VarScrut, we wanted to fit closely to the practice of our end users.

The major caveat linked to this choice is that variant analysis tools manipulating the BAM files to generate the VCF files must be highly efficient and performed according to the best practices [16]. Recent studies have shown that depending on the preceding applied bioinformatics protocol (Hicks et al., 2011), variant calling tools can have as low as 57% and 27% of concordance for SNV and InDel respectively (O'Rawe et al., 2013). This can have dramatic effects on the analysis process and hence would require extensive validation of the variants to eliminate false assumptions on artefacts. Moreover, as recommended in the ACMG guidelines, it would represent a good practice to check the presence of candidate variants in the BAM file. Currently, this task is performed by visual inspection by the

bioanalysts. VarScrut uses an arbitrary threshold value of 3 variants per gene to determine missed or poorly covered genes. To limit the variation in performances, owing to the quality of the input, the aim of the next release will be to include in VarScrut a standard, best-practice variant calling procedure, which uses SAM/BAM files. Moreover, the sequenced reads information contained in the BAM files can be maximised to scan in depth poorly covered regions and genes, currently done in the follow-up module. With the availability of the BAM files as input, the evaluation of poorly covered regions/genes would be automatically incorporated in the main analysis process for report.

5.4.3 *Gold standard datasets and machine-learning aspects*

5.4.3.1 Publicly available disease exomes

During the development of VarScrut, one difficult task was to gather gold-standard datasets for training and evaluation. During the evaluation of the different variant analysis tools (see VarScrut manuscript, section 5.2 – Benchmark analyses), very few tools (eXtasy, VAAST, PHEVOR, Phen-Gen) provided a benchmark of their performances on either simulated or real exome datasets and almost none provides access to the datasets used for the training or evaluation of their tool. Although some datasets (*e.g.* Kabuki syndrome: SRR063945; Miller syndrome: SRX09441) are deposited in the European Genome-phenome Archive, these data are still under embargo and require several reviews from a board of Data Access Committees and from the data owner before data access can be granted. This is a real hurdle to the reproducibility of previously described methods and the comparison of their results with other tools.

In order to do a fair evaluation of the VarScrut performances and comparison to the other tools, synthetic disease exomes were generated with the same protocol using as initial set, standard exome datasets with known disease causing-genes. I used the 1,000 Genome datasets (see section 3.2.1.2) as a source of healthy exomes modified for the creation of synthetic disease exomes, meaning that the same datasets with the same protocols were used for comparison of VarScrut to other variant analysis tools.

In other words, even though synthetic exomes are effective alternatives, a well-defined freely accessible dataset based on real disease exomes is lacking for the community.

5.4.3.2 Disease-causing variants and next-generation of variant deleterious predictors

Another important point in WES analysis is the use of computational predictors (such as SIFT, PolyPhen2 or GERP…) to annotate and evaluate disease-causing variants. To date, over 40 different variant deleterious predictors are available (Richards et al., 2015). Although the algorithms may differ, all predictors aim at evaluating the effect of the sequence variant on the resulting protein product, *via* determination of the potential variant impact on the primary/alternative gene transcripts, on other genomic elements or on the amino acids. Almost all predictor tools are oriented towards the evaluation of missense or splice variant effects. Indeed, few or no predictors are available to evaluate the impact of non-sense, frameshifts or InDel with the notable exception of SIFT Indel (Hu and Ng, 2013) and KD4i [developed in our laboratory, (Bermejo-Das-Neves et al., 2014)], which are devoted to the evaluation of the InDel impacts. Non-sense and frameshifts are generally considered by default as deleterious on the basic assumption that loss of function will systematically arise from the resulting truncated protein.

Considering available predictors, there are two main categories:

(i) Those that predict how a missense change damages the resulting protein function or structure (*e.g.* SIFT, Mutation Taser, PolyPhen2...). The evaluation of a missense change depends on criteria such as the nucleotide/amino acid evolutionary conservation, the location and context within the protein sequence, the biochemical consequence of the amino acid substitution... Various *in silico* algorithms measure one or a combination of these criteria to assess the predicted impact of the missense change (MacArthur et al., 2014);

(ii) Those that predict the potential splicing effects, such as the creation or loss of splice sites at the exonic or intronic level (*e.g.* GeneSplicer, Human Splicing Effect, MaxEntScan…) (MacArthur et al., 2014);.

Globally, predictors claim to have an accuracy of 65-80% (Thusberg et al., 2011) when examining disease-causing variants. Missense predictor tools have been extensively (Frousios et al., 2013; Thusberg et al., 2011) compared, thus emphasizing their good specificity, for impact prediction of benign and highly deleterious variants while they are less reliable for the in-between variant categories.

To improve predictor accuracy, a novel category of annotation tools, named meta-predictors, has emerged. Based on the scores of several predictors for disease-causing variants and common variants, the meta-predictors train a machine-learning model to predict disease-causing variants. Several meta-predictors are already available such as: Condel, which uses a Weighted Average of the normalized Scores (WAS) based on the PolyPhen2 and SIFT scores for SNV; CADD, which combines ~60 annotation features into a Support-Vector Machine approach (SVM). As meta-predictors combine variant predictors, they tend to generate a single model for the prediction of both splice events and missense mutations.

In VarScrut, a meta-predictor combining 10 annotation features (see VarScrut manuscript, section 5.2 – Benchmark analyses) into a Logit model was developed. The Logit model was initially chosen based on its advantages, such as the tolerance to missing values, which can be extrapolated *via* linear regression model as opposed to other approaches such as SVM or random forest. In comparison to VarScrut meta-predictor, which shows an Area Under the Curve (AUC) of 75%, the other meta-predictors range between 60-73%.

5.4.3.2.1 Good variant predictor needs good training data

For the generation of the VarScrut meta-predictor, several hurdles had to be passed. Firstly, the collection of disease-causing variants had to be compiled and curated. Indeed, disease-causing variants are currently scattered across diverse asynchronous sources, such as dbSNP, ClinVar, HGMD, OMIM, SwissVar or LOVD… Secondly, each source uses its own format, representation form and coding system for deleteriousness. In order to manipulate a homogenous set of variation data, I normalized all the variants according to the Human Genome Variation Society (HGVS) nomenclature using VEP (Ogino et al., 2007). Thirdly, since SwissVar provides only representation of variants at the protein level, the VEP tool was used to provide a DNA description of the variants. When multiple possible DNA-level descriptions were present, the DNA-level description corresponding to previously reported variants was selected; otherwise the description predicted to have the most deleterious impact was selected.

As machine-learning methods are sensitive to the training and control dataset used, rigorous precautions were taken to constitute the different datasets. Because VarScrut's meta-predictor uses the scores of other predictors and meta-predictors, scrupulous attention was taken to remove variants that have already been used by other tools. Since each disease variant database (ClinVar, HGMD, SwissVar) uses different rating systems to classify

deleterious/pathogenic variants, only the most severe consequences were selected (see Manuscript) before the merging of the datasets. Additionally, besides common polymorphic variants (MAF>10%), we took advantage of the mass of variant data available in cohort projects such ExAC, to also include rare benign variants (MAF<1%) to refine our control dataset for the training, resulting in better performance with respect to the other meta-predictors.

Moreover, the importance of the quality of the selected datasets has been shown in the comparison with FATHMM, which showed over-training performances on filtVariBench dataset compared to filtSwissVar dataset (Figure S1 in VarScrut manuscript).

5.4.3.2.2 Towards a next-generation of variant predictors: From machine-learning to deep-learning

Currently, all the predictors and meta-predictors are based on sequence features or rules to evaluate the deleteriousness of a variant, such as the impact of missense or splice variants. Evaluations of the accuracy of the different computational predictors hardly exceed 75-80% (Thusberg et al., 2011), suggesting that these sets of rules and features do not capture all the biological complexity of the events. As presented by MacArthur in his systematic survey of LoF variants on 185 human genomes, 89% of the predicted nonsense-mediated mRNA decay (NMD) SNV had no effect on gene expression when examined by RNA sequencing (MacArthur et al., 2012). These results tend to suggest that predictors based on rules, are far from predicting correctly the different effects of nsSNV (MacArthur et al., 2012).

To overcome these limits, new approaches are emerging, using Deep Learning (DL) on experimental data to directly generate the prediction models. DL is a branch of machine learning that attempts to learn multi-level representations of data, embodying a hierarchy of factors that may explain them. Tools such as SPIDEX (Xiong et al., 2015), trained their computational model of splicing regulation for each of the exons in its training set by estimating all the corresponding sequence features across the RNA-seq data of 16 human tissues from the Illumina Human Body Map 2.0 project NCBI GSE30611). Such approaches, unlike existing ones, do not suffer from the ascertainment biases inherent in databases of disease annotations, and in its evaluation set it claims to have an AUC of 99%.

Finally, the dataset resulting from recent biological assays that uses saturation mutagenesis techniques (Starita et al., 2015) may represent an important opportunity to access

high-resolution experimental data of variant impacts on regulatory (Patwardhan et al., 2009) and protein-coding (Findlay et al., 2014) sequences. These rich datasets of experimental measurements represent a *bona fide* training data for better algorithm development, with pre-computed VUS interpretations, which is currently a key challenge in medical exome analysis (Cooper, 2015).

Chapter 6 PubAthena

The constant innovation in technologies with an ever-increasing throughput has contributed to the prolific production of scientific literature, which is the main diffusion channel for novel findings and hypotheses (Figure 6-1). The screening of the scientific literature is now an essential daily task to keep up-to-date with novel discoveries in any field of interest and is now even part of any analysis procedure in this era of data-intensive biology (*e.g.* exome variant analysis or functional genomics analysis). In this Chapter, I first present an overview of the text-mining approaches available to process large amounts of articles in order to extract novel relevant knowledge. Second, I then present a survey I performed to evaluate the different text-mining solutions currently available. Third, the manuscript of Pub*Athena* is provided in the manuscript. Pub*Athena* is a text-mining tool based on a Naïve Bayesian Classifier model to evaluate and recommend novel relevant articles from the constant flux of new articles. Finally, in complement to the manuscript of Pub*Athena,* I discuss technical and text-mining choices made for the development of Pub*Athena*.

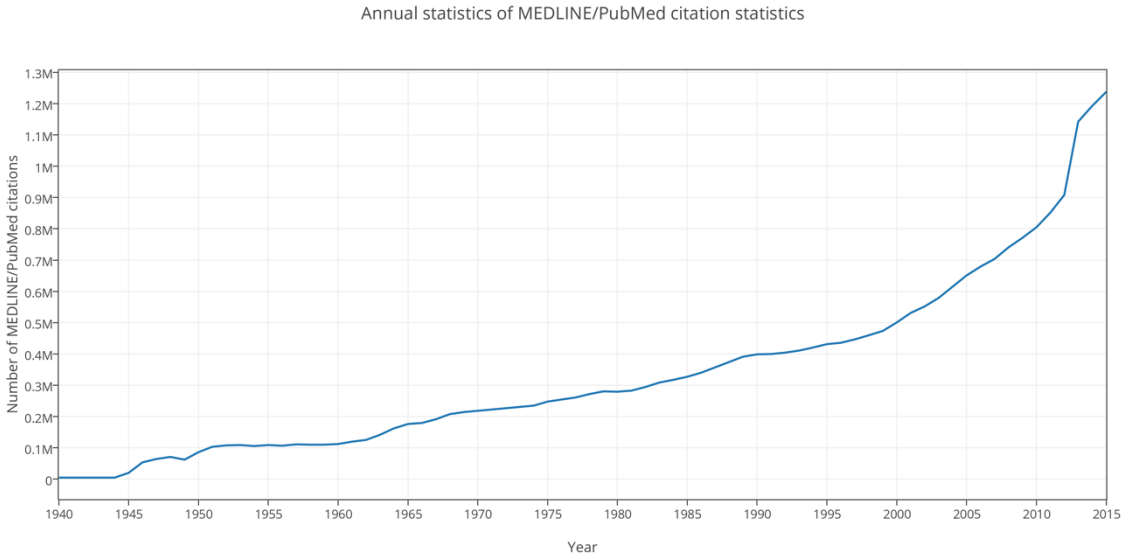


Figure 6-1: Statistics on the annual increase of citations in MEDLINE/PubMed database.

6.1 Pub*Athena* background and philosophy

Information extraction and text data analysis can be particularly relevant and helpful in genetics and biomedical research, in which up-to-date information about complex processes involving genes, proteins and phenotypes is crucial. Even with dedicated effort to capture novel information in biomedical databases, much of this information still remains 'locked' in

the unstructured text of the publications, resulting in a substantial lag between the information present in publications and those present in databases. With ~100,000 new articles published every month; it is thus very challenging to remain up-to-date with all the latest biomedical discoveries. In this context, text mining denotes any procedure that analyses large corpora of natural language text and detects lexical or linguistic usage for knowledge extraction from free text and can be extended to the generation of new hypotheses by combining the extracted information from several publications (Hirschman et al., 2012).

Globally, there are five main steps in text mining approaches (Figure 6-2):

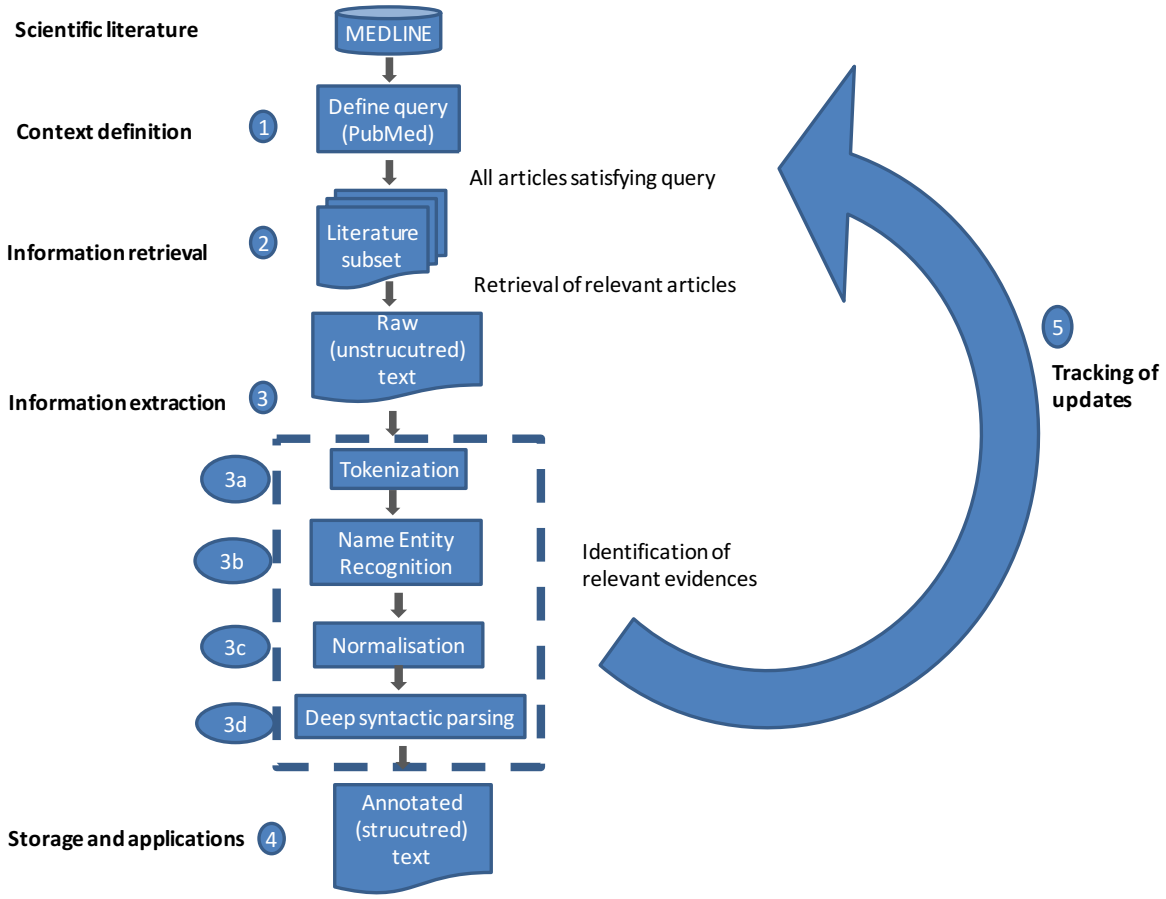


Figure 6-2: Overview of the major steps of text-mining approaches

1. **Context definition of the subject of study** – Every literature-tracking task starts with the definition of the context domain selected for the curation through the appropriate choice of terms/keywords that defines the perimeter of the problem being studied.
2. **Information retrieval** – This step involves the retrieval of relevant abstracts/articles using literature tools such as Entrez e-utils in PubMed, or different query engines such as Google Scholar.

3. **Information extraction** – The text mining approach comprises several computational procedures to automatically help in the identification of mark-up entities, such as genes or diseases, as well as the identification of complex relationships between those entities, such as protein–protein interactions and gene–disease associations. Obtaining such valuable information involves several complex procedures based on defined patterns, machine-learning techniques, statistical analyses or automated reasoning.
    a. **Tokenization** – The first pre-processing step of a document text involves the chunking/segmentation of the original unstructured text into its constituents ('tokens'), such as a list of sentences, single words, digits or punctuation (Kang et al., 2011). Common language or junction words that are generally uninformative (*e.g.* a, the, to, for…) are filtered using a list of "stopwords". Another means to filter uninformative words is to remove words that are too frequent or too rare in a corpus of text according to a user-defined threshold. The processing of the text tokens corresponds to the field of natural language processing (NLP) in computer science and there are numerous libraries such as NLTK [17] available with predefined standard procedures.
    b. **Name Entity Recognitions (NER) –** The generated tokens are then screened through dictionaries (*e.g.* OMIM, EntrezGene, HPO) to identify through unambiguous terms, biological concept entities such as gene or disease names. In the Pub*Athena* manuscript, NER is referred to as Biological Entity Recognition.
    c. **Normalization –** The different forms of the same identified entities (synonyms) are then normalized into unique and preferred terminologies, ontologies (*e.g.* Open Biomedical Ontologies) or database identifiers (*e.g.* Protein => UniProt ID; Disease => OMIM) to improve the downstream analyses.
    d. **Deep syntactic parsing –** The entities identified from the document text can be further parsed to identify structured relationships between pairs of entities, such as descriptions of protein–protein interactions (i.e. activation/inhibition loops), characterizations of gene products in terms of their cellular location, molecular function, biological process or phenotypic effect.
4. **Applications –** this step usually involves the exploitation and valorization of the automatically extracted biological entities and the possible relationships. Among the 'classical' applications of text mining there are usually: (i) co-citation analysis

between two entities to identify for instance interactions between drugs (*e.g.* DrugBank (Tari et al., 2010)); (ii) constitution of knowledge bases [*e.g.* CilioPath - Chapter 7, PharmGKB (Whirl-Carrillo et al., 2012), IntAct (Orchard et al., 2014)]; (iii) constitution of ontology dictionaries (*e.g.* Gene Ontology (Gene Ontology Consortium, 2015), Human Phenotype Ontology (Köhler et al., 2014)); (iv) knowledge discovery using different hypothesis testing (*e.g.* CoPub (Frijters et al., 2008)).

5. **Update tracking –** Finally, with the constant evolution of the knowledge information, tracking literature updates is an essential task to evaluate novel information, such as the change in the association status of a gene to diseases or the update in the functional information of a gene. This involves repeating Step 1-3 periodically to identify the updated information to evaluate the novelty.

Category	Class	Name	Information Retrieval	Information extraction						Update tracking	Clustering	Machine Learning	Source code availablility	Exportable results	Access	Query	Corpus update	Year	PMID
				Name Entity Recognition				Keywords	Deep syntactic parsing										
				Normalization															
				Species	Gene	Disease	Variation												
Annotator	Stand-alone	SR4GN	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🟢	Public	NA	NA	2012	22679507
Annotator	Stand-alone	GeneTUKit	🔴	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🟢	Public	NA	NA	2011	21303863
Annotator	Stand-alone	Dnorm	🔴	🔴	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🟢	Public	NA	current	2013	23969135
Annotator	Stand-alone	tmVar	🔴	🔴	🔴	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🟢	🟢	Public	NA	NA	2013	23564842
Annotator	Dataset-centered	MuGeX	🟡	🔴	🟢	🟢	🟢	🟡	🔴	🔴	🔴	🔴	🔴	🔴	Public	Disease or Gene concept	?	2007	18172928
Annotator	Dataset-centered	Whatizit	🔴	🟢	🟢	🟢	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🟢	Public	NA	NA	2008	18006544
Annotator	Dataset-centered	becas	🔴	🟢	🟢	🟢	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🟢	Public	NA	NA	2013	23736528
Annotator	Dataset-centered	tagtog	🔴	🟢	🟢	🟢	🟢	🟡	🔴	🟡	🟡	🟢	🔴	🟢	Public	Free text	current	2014	24715220
Annotator & Query	Dataset-centered	PubTator	🟢	🟢	🟢	🟢	🟢	🟡	🔴	🔴	🔴	🔴	🟢	🟢	Public	Free text	current	2013	23703206
Query	Dataset-centered	PolySearch2	🟡	🟢	🟢	🟢	🟢	🟢	🟡	🔴	🔴	🔴	🔴	🟢	Public	Concept keywords	current	2015	25925572
Query	Dataset-centered	iHop	🟡	🟢	🟢	🟢	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🔴	Public	Gene concept	?	2004	15226743
Query	Dataset-centered	BioTextQuest+	🟡	🔴	🟢	🟢	🔴	🟢	🟡	🔴	🟢	🔴	🔴	🟢	Public	Concept keywords	current	2014	25100685
Query	Dataset-centered	CoPub	🟡	🔴	🟢	🟢	🔴	🔴	🟡	🔴	🟡	🔴	🔴	🟢	Public	Concept keywords	current	2008	18442992
Query	Dataset-centered	PESCADOR	🟡	🔴	🔴	🔴	🔴	🟢	🟢	🔴	🟢	🔴	🔴	🟢	Public	Free text	current	2011	22070195
Query	Stand-alone	Textpresso	🟡	🔴	🔴	🔴	🔴	🟢	🟡	🔴	🔴	🔴	🟢	🟢	Public	Free text	NA	2008	18949581
Query	Dataset-centered	GoPubMed	🟢	🟢	🟢	🟢	🔴	🔴	🟡	🔴	🟡	🔴	🔴	🟢	Public	Free text	current	2005	15980585
Query	Dataset-centered	Facta+	🟢	🔴	🟢	🟢	🔴	🟢	🟡	🔴	🟢	🔴	🔴	🔴	Public	Free text	2010	2010	21685059
Query	Dataset-centered	PALM-IST	🟢	🔴	🟢	🟢	🔴	🟢	🟡	🔴	🟢	🔴	🔴	🟢	Public	Concept keywords	current	2015	25989388
Query	Dataset-centered	Anne O'Tate	🟢	🔴	🔴	🔴	🔴	🟢	🟡	🔴	🟡	🔴	🔴	🟡	Public	Free text	current	2008	18279519
Query	Dataset-centered	Medline Ranker	🟢	🔴	🔴	🔴	🔴	🟢	🔴	🔴	🟢	🟢	🔴	🟢	Public	Free text	current	2009	19429696
Update tracking	Dataset-centered	PubChase	🟡	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🟡	🟡	🔴	🔴	Commercial	Free text	current	2015	-
Update tracking	Stand-alone	PubCrawler	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🔴	🔴	🟢	🟢	Public	Free text	current	2004	15215341

Table 6-1: Survey of representative text mining solutions available for the exploration of literature. The colours indicate the level of satisfaction of a given criterion, Green: Full-filled; Yellow: Partially full-filled; Red: Unfilled. In the corpus update column, "current" state indicates that the corpus referenced by the tool is up to date; NA: indicates that the criterion is not applicable for a given tool; ?: indicates that the information is unknown.

Many text-mining (TM) resources have been developed to retrieve meaningful information from the MEDLINE corpora, such as iHop (Hoffmann and Valencia, 2005), PolySearch (Cheng et al., 2008), PubTator (Wei et al., 2013), Whatizit (Rebholz-Schuhmann et al., 2008) and others. Based on the previously discussed steps involved in text mining approaches, I defined a set of functionalities essential for the exploration and tracking of literature updates, which were then used to carry out a complete survey of the active TM resources available (Table 6-1).

Basically, there are two main classes of TM resources, those that are stand-alone tools and those that are organized around a dataset or database. Stand-alone resources are downloadable TM tools that can process any user submitted search term query or raw text. Generally, stand-alone TM resources are dedicated to only one step of the TM procedure, such information retrieval for PubCrawler (Hokamp and Wolfe, 2004) or information extraction for Textpresso (Müller et al., 2008). On the other hand, dataset-centred TM resources can process only pre-indexed text documents. Some tools, such as Medline Ranker (Fontaine et al., 2009) or PubTator (Wei et al., 2013), use the most up-to-date and full set version of MEDLINE for the queries, while other resources, such as Facta+ (Tsuruoka et al., 2011), use only a dated version of MEDLINE or queries only a subset of MEDLINE such as PALM-IST (Mandloi and Chakrabarti, 2015) or iHop (Hoffmann and Valencia, 2005). Dataset-centred resources can be dedicated to a specific TM step such as information extraction (*e.g.* Whatizit (Rebholz-Schuhmann et al., 2008)) or update tracking (*e.g.* PubChase) or they can cover a range of TM steps such as PubTator (Wei et al., 2013) (from information retrieval to information extraction), or PALM-IST (Mandloi and Chakrabarti, 2015) (from information retrieval to clustering of articles).

Overall, there are currently no stand-alone programs or data-centred resources that provide full-stack TM procedures necessary for the first 3 steps of TM in batch, and the tracking of the literature for updates (step 5) on a daily basis. Moreover, there is still an unmet need for the screening of relevant articles (using clustering or machine learning approaches) and to retrieve the full set of relevant articles in batch with the associated metadata, such as the list of relevant keywords and biological entities.

In this context, for a meaningful knowledge discovery and sophisticated agglomeration of relevant up-to-date biomedical literature, several objectives were set for the development of the TM tool Pub*Athena*:

- **Lightweight and downloadable** – Pub*Athena* should perform advance literature mining tasks (full range information retrieval and extraction) while being lightweight in storage and computing resources in order to run on laptops and be widely distributable.
- **Multiple biological name entity recognition** – Pub*Athena* should be able to automatically detect several biologically relevant entities (gene, variation, disease, phenotype and biological process) essential in biomedical research, especially for the gene-discovery or case report in the case of RD.
- **Flexible** – Pub*Athena* should be flexible enough to query the PubMed compendium not just by entities or concept keywords but also by free text search in order to define the biological context.
- **Customizable processing and analysis** - Pub*Athena* should be able to store locally intermediate processing results (*e.g.* token identifications, normalized biological entities), in order to allow the reprocessing of the retrieved literature corpus using customizable parameters or external tools, as well as more advanced analyses other than pre-defined classical analyses (*e.g.* text enrichment, co-occurrences).
- **Up-to-date** – In the constant flux of scientific literature production, Pub*Athena* should be able to automatically remain up-to-date with the latest articles available at the time of the query.
- **Literature tracking feature** – Pub*Athena* should provide a dynamic recommendation system, which can be personalized according to the user's interest or query.

To resume, Pub*Athena* is a versatile, lightweight TM framework that can query the full MEDLINE dataset (>25 million published documents), retrieve and extract meaningful information. Pub*Athena* is downloadable and works with a local backed storage system, for the storage of queries and performed analyses, as well as customizable exploitations. Moreover, the stored queries can be reused and the results automatically up-dated with the latest articles published *via* its daemon. A key feature of Pub*Athena* is its Naïve Bayesian Classifier approach as a machine learning recommendation system for relevant articles in the constant flux of published articles.

6.2 Manuscript of Pub*Athena*

Bioinformatics, 2016, 6–1
doi: 10.1093/bioinformatics
Application Note

Data and Text Mining

Pub*Athena*: a Naïve Bayesian Classifier text-mining tool for the tracking of relevant literature updates

Kirsley Chennen[1,*], Alexis Allot[1], Raymond Ripp[1], Arnaud Kress[1], Julie Thompson[1], Odile Lecompte[1], Hélène Dollfus[2,3] and Olivier Poch[1,*]

[1]Department of Complex System and Translational Bioinformatics - ICube, CNRS UMR 7,357, Federation of Translational Medicine of Strasbourg (FMTS), University of Strasbourg, 11, rue Humann, 67,000 Strasbourg, France

[2]Laboratory of Medical Genetics, INSERM U1112, Institute of Genetics and Medicine of Alsace (IGMA), Strasbourg Medical School, University of Strasbourg, 11 rue Humann 67,000 Strasbourg, France

[3]Institute of Medical Genetics of Alsace, National Reference Centre for Rare Diseases in Ophthalmic Genetics (CARGO), Strasbourg Medical School, 1, place de l'Hôpital, 67,091 Strasbourg, France

*To whom correspondence should be addressed.

Abstract

Summary: We present Pub*Athena*, a text-mining framework that can prioritize and recommend the most relevant articles based on a user's defined interest. Pub*Athena* implements text-mining modules, which use Natural Language Programming (NLP) approaches to automatically discover significant biomedical entities (*e.g.* gene, disease, species, variants, chemicals, phenotypes, etc.). The recommendation of relevant articles is based on a Naïve Bayesian Classifier approach, which takes a user's defined list of reference articles as a training set, from which keywords and normalized annotated entities are extracted and used as parameters to evaluate new articles. Pub*Athena* allows automatic periodic screening of the literature for relevant novel articles, which best match the biological interest of the user, and provides automatic extraction of biomedical entities and keywords for further rapid analysis (*e.g.* clustering of articles, co-citation analysis, etc.) and batch curation and rapid exploitation (*e.g.* constitution of a knowledge base).

Availability: PubAthena is implemented in Python and consists of command-line driven programs. PubAthena is freely available at http://lbgi.fr/pubathena, distributed under a GPL license.

Contact: kchennen@unistra.fr, olivierpoch@unistra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the constant increase in the breadth and depth of biomedical literature, the major challenge is to stay up-to-date with all the latest discoveries. With ~100,000 new articles published every month, it is extremely difficult to identify the most recent and relevant articles, using only basic keywords query searches, and without missing any. Even in specialized areas of research, and with dedicated efforts to capture information in biological databases (e.g. OMIM for disease-related studies) (Amberger et al., 2015), it is time-consuming to sift through the tens or hundreds of hits. Moreover, much of this information remains 'locked' in the unstructured text of biomedical publications. Thus, there is a substantial lag between the publication of an article and the subsequent extraction of such information into databases.

Several text mining (TM) resources (Table 1) have been developed for information retrieval (IR; obtaining relevant subset of documents with regards to a query) such as iHop (Hoffmann and Valencia, 2005), PolySearch2 (Liu et al., 2015) and information extraction (IE) from PubMed/MEDLINE's abstracts [http://www.ncbi.nlm.nih.gov/pubmed/]. These rely on several TM procedures: tokenization (identification of the constituents of a text), biological entity recognition (BER; e.g. genes,

diseases, chemicals, pathways and biological processes), entity normalization (mapping of synonyms on preferred terminology or database identifiers), and relation extraction between entities such as BioTextQuest+ (Papanikolaou et al., 2014), PALM-IST (Mandloi and Chakrabarti, 2015)). Additional TM resources are even dedicated to the tracking of literature updates, using either basic queries such as PubCrawler (Hokamp and Wolfe, 2004) or machine learning approaches such as Medline Ranker (Fontaine et al., 2009). However, there is still an unmet need for an automatic solution with full range TM procedures from information retrieval to information extraction and with automatic tracking and recommendation of literature updates.

Here, we present PubAthena, a text-mining framework based on a Naïve Bayesian Classification (NBC) model coupled to a web daemon that automatically tracks the latest relevant articles based on predefined queries in PubMed. PubAthena integrates a complete set of TM resources including information retrieval and information extraction, recognition with of a complete set of biological entities (e.g. species, genes, diseases…) coupled to newly developed tracking and machine learning processes (Table 1). Via a lightweight local storage system (SQLite database), and taking advantage of the NBC model, PubAthena provides two modes of application for literature tracking: (i) a specific mode, which retrieves and ranks novel articles that satisfy the initial query terms (e.g. for the constitution of a dedicated knowledge base); (ii) a discovery mode, which extends the initial query scope/terms and populates biological entities by processing the novel articles and proposing additional articles (e.g. for bibliographical interest or case report tracking). The user-defined queries, as well as the used-validated extension and up-dated NBC, can be stored and the web daemon can be executed at customizable intervals such as daily or weekly.

2 Archicture and implemendation

PubAthena (Figure 1) is a Python-implemented TM framework, which takes as input either a list of keywords or a list of PMIDs, and uses the NCBI's E-utils services to query PubMed via the E-Search service (Figure 1-1) and retrieve a resulting list of article abstracts and associated metadata in XML format via E-utils service (Figure 1-2). PubAthena implements the Natural Language Toolkit library (www.nltk.org) for text-mining procedures (segmentation and tokenization) of the article abstracts (Figure 1-3). BER (genes, diseases, pathways, and variations) is performed using PubTator (Wei et al., 2013) and the Biomedical Concept Annotation System (BECAS) (Nunes et al., 2013) (Figure 1-4). Additionally, an in-house procedure maps other biological entities ontologies, such as biological processes onto Gene Ontology (GO) (Gene Ontology Consortium, 2015); phenotypes or symptoms onto Human Phenotype Ontology (HPO) (Köhler et al., 2014) or Mouse Phenotype Ontology (MPO)(Smith and Eppig, 2009). The detected entities are then normalized (Figure 1-5) on database identifiers such as gene on Entrez Gene; variation on dbSNP (Sherry et al., 2001); disease on Comparative Toxicogenomics Database (Davis et al., 2015) / OMIM (Amberger et al., 2015); biological processes on GO, pathways on KEGG (Kanehisa and Goto, 2000); phenotype on HPO (Köhler et al., 2014) / MPO (Smith and Eppig, 2009) (Figure 1-5). For each query, the resulting articles and the generated metadata are finally stored in an SQLite database. The local SQLite database management system comes with predefined procedures for query and export options (CSV, XML, JSON) for further applications such as knowledge integration, building a knowledge base or co-occurrence analysis. PubAthena includes a procedure to generate a Naïve Bayesian Classifier (NBC) model (Zhang, 2004), using the SciKit library [http://scikit-learn.org/], based on a user defined list of articles used as the training set. The background set is obtained from 10,000 randomly chosen recent abstracts (Figure S1). The generated model is then used to evaluate and rank the significance of novel articles based on a threshold probability score above 0.5 (Figure 1-6). PubAthena includes a daemon, which monitors new PubMed releases and is responsible for downloading novel records and dynamically updating the local index.

PubAthena provides several ways to query relevant articles and to export the associated biomedical entities and keywords for various applications (e.g. clustering of articles, co-citation analysis).

The extracted metadata are used to generate the NBC model in order to identify and retrieve related articles from PubMed/MEDLINE updates.

3 Discussion

PubAthena can be used for various text-mining application scenarios. For instance, one application of PubAthena is the literature tracking of relevant novel biomedical articles. To illustrate this, we searched for relevant novel articles concerning inherited retinal diseases. As a reference/positive dataset, 898 articles on Retinitis Pigmentosa (RP) published before 2010, were selected from the RetNet reference database (https://sph.uth.edu/retnet/) and as a negative dataset, 10,000 recent abstracts (after 2010) were randomly selected from PubMed. The performance of PubAthena was compared to MedlineRanker, which provides the most complete machine learning approach for the recommendation aspect using the same parameters. The NBC model generated was then used to evaluate the recommendations on RP articles published after 2010. PubAthena achieved an F-measure of 92%, while MedlineRanker achieved an F-measure of 67% and provided the results for only the top 150 articles ranked by its models. It should be noted that, as MedlineRanker does not provide information extraction procedures, and the list of 103 discriminative words it provided (data not shown), does not include any biological entities such as genes or diseases.

In conclusion, we presented PubAthena, a pre-configured and customizable framework for text mining of biomedical literature recorded in PubMed. Although our efforts were motivated by a desire to produce initial, non-statistical analyses, we are currently expanding our framework to include a suite of powerful tests for association studies. Our general framework will allow the implementation and comparison of a wide array of analytical methods. Future work will include additional text-mining procedures such as discourse level analysis to allow extracting the gene-mutation-disease associations, the further indexing of full text articles from PubMed Central [http://www.ncbi.nlm.nih.gov/pmc/] to improve the machine-learning NBC recommendation model, and a web interface with API services for simplicity of use for end-users.

Funding

This work was supported of ANR Investissement d'Avenir Bioinformatique BIP:BIP (ANR10-BINF03-05).

Conflict of Interest: none declared.

References

Amberger,J.S. et al. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research, 43, D789–98.

Davis,A.P. et al. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research, 43, D914–20.

Fontaine,J.-F. et al. (2009) MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Research, 37, W141–6.

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Research, 43, D1049–56.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics, 21 Suppl 2, ii252–8.

Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. Nucleic Acids Research, 32, W16–9.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28, 27–30.

Köhler,S. et al. (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research, 42, D966–74.

Liu,Y. et al. (2015) PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Research, 43, W535–42.

Mandloi,S. and Chakrabarti,S. (2015) PALM-IST: Pathway Assembly from Literature Mining--an Information Search Tool. Scientific Reports, 5, 10021.

Nunes,T. et al. (2013) BeCAS: biomedical concept recognition services and visualization. Bioinformatics, 29, 1915–1916.

Papanikolaou,N. et al. (2014) BioTextQuest(+): a knowledge integration platform for literature mining and concept discovery. Bioinformatics, 30, 3249–3256.

Sherry,S.T. et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research, 29, 308–311.

Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 1, 390–399.

Wei,C.-H. et al. (2013) PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research, 41, W518–22.

Zhang,H. (2004) The optimality of naive Bayes. American Association for Artificial Intelligence, 1, 3.

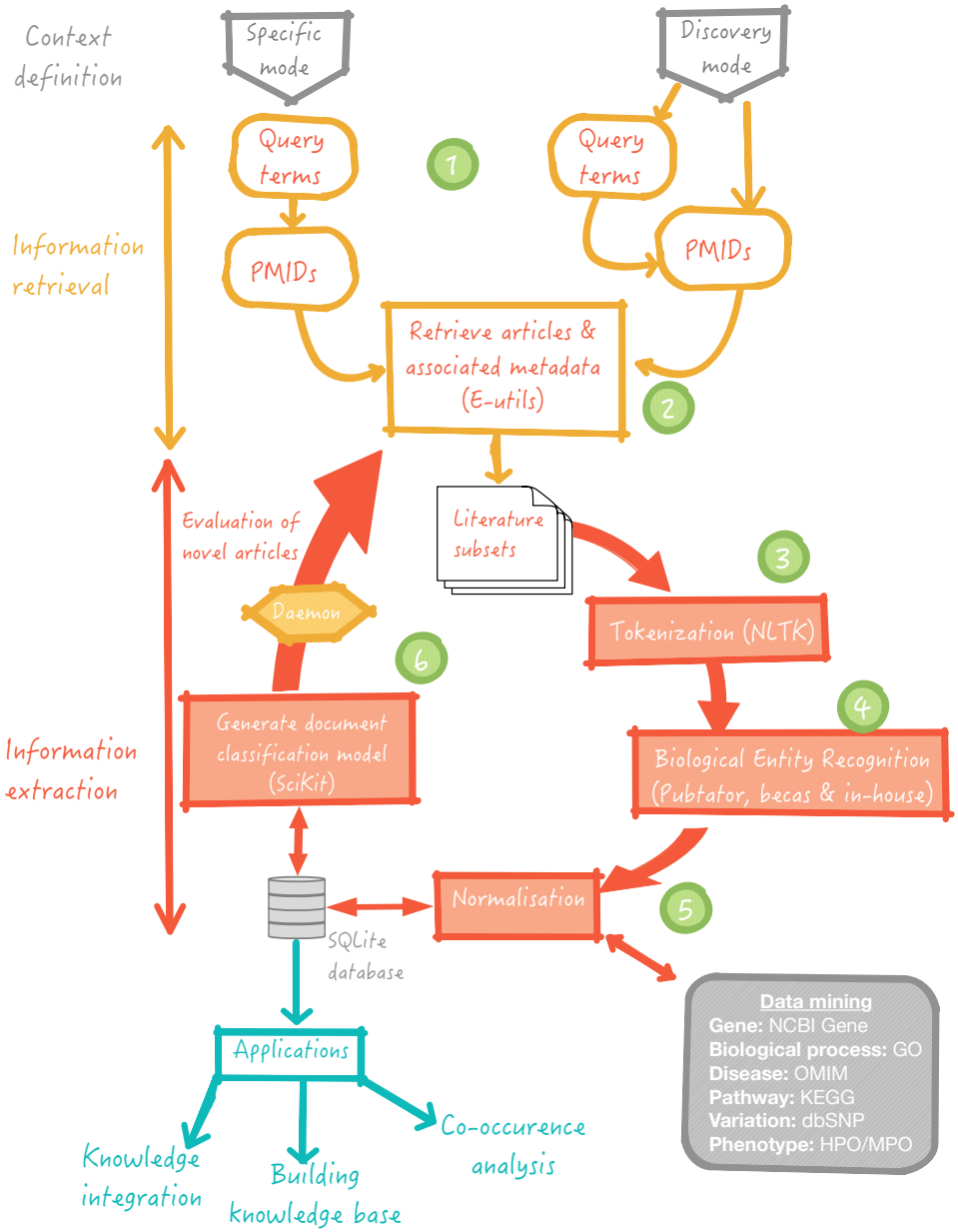


Table 1: Qualitative comparison of the functionalities of text-mining resources for literature mining and tracking. The different features are described in colour codes where a green dot indicates a fully functionality, a yellow dot indicates a partially functionality and a red dot indicates that the function is not available.

Name	Information Retrieval	Information extraction					Update tracking	Machine Learning	Exportable result search	Query	Year	PMID
		Biological Entity Recognition				Keywords						
		Normalization										
		Species	Gene	Disease	Variation							
BioTextQuest+	🟡	🔴	🟢	🟢	🔴	🟢	🔴	🔴	🟢	Concept keywords	2014	25100685
PALM-IST	🟡	🔴	🟢	🟢	🔴	🟢	🔴	🔴	🟢	Concept keywords	2015	25989388
iHop	🟡	🟢	🟢	🟢	🔴	🟢	🔴	🔴	🔴	Gene concept	2004	15226743
PolySearch2	🟡	🟢	🟢	🟢	🟢	🟢	🔴	🔴	🟢	Concept keywords	2015	25925572
PubChase	🟡	🔴	🔴	🔴	🔴	🔴	🟢	🟡	🔴	Free text	2015	-
tagtog	🔴	🟢	🟢	🟢	🟢	🟡	🟡	🟢	🟢	Free text	2014	24715220
Medline Ranker	🟢	🔴	🔴	🔴	🔴	🟢	🔴	🟢	🟢	Free text	2009	19429696
PubCrawler	🟢	🔴	🔴	🔴	🔴	🔴	🟢	🔴	🟢	Free text	2004	15215341
PubAthena	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🔴	🟢	Free text	2016	-

Table 1: Benchmark results of the cascade oscillators model

Supplementary data

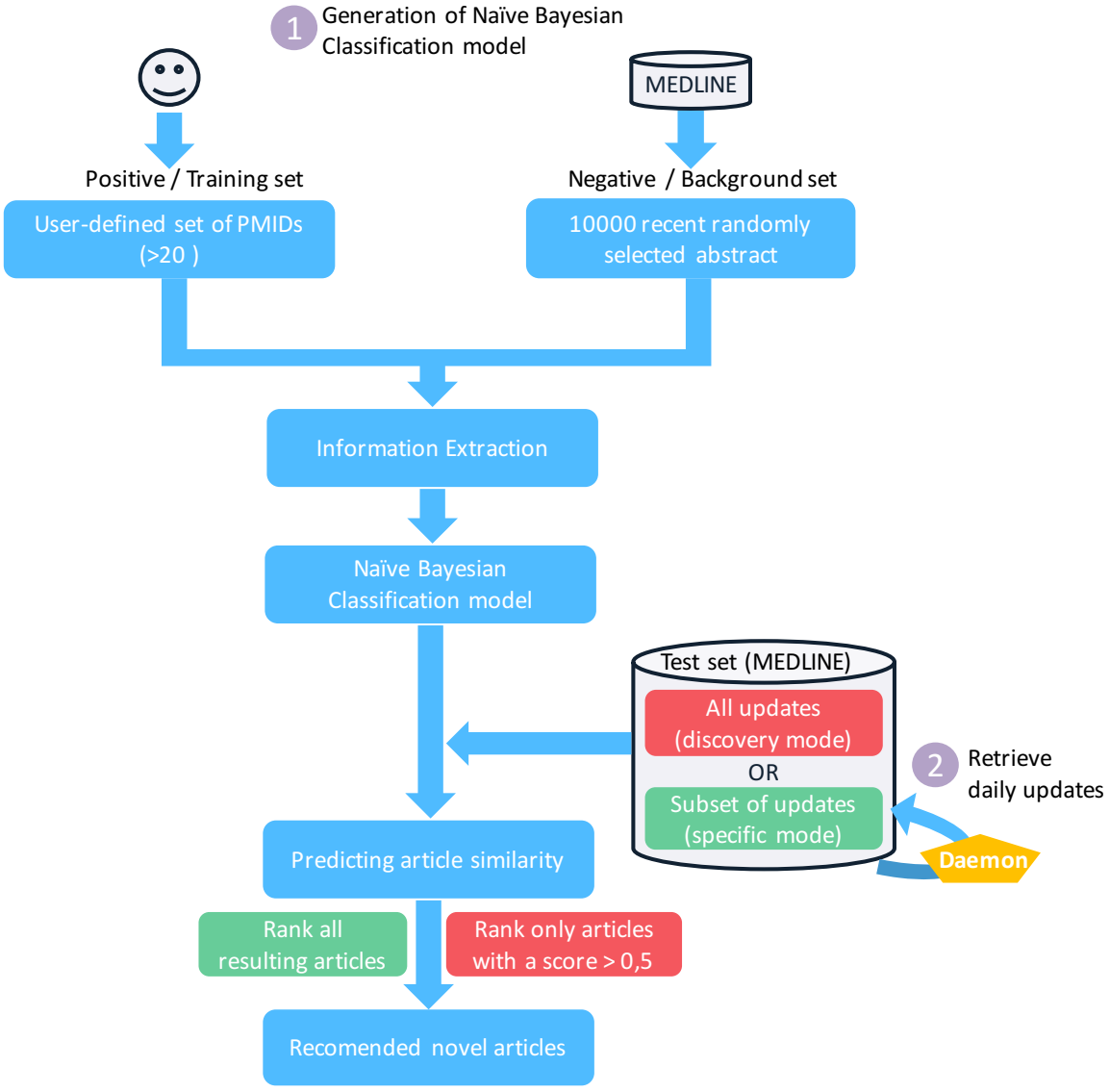


Figure S1: Overview of PubAthena's procedure for the generation of the Naïve Bayesian Classification recommendation model (1) and the daily screening and recommendation of novel relevant articles.

SM1 - Construction of a training set related to retinitis disorders

The website RetNet [https://sph.uth.edu/retnet/], a reference dataset for retinal disorders, was used to set up the training and evaluation datasets. All the reference articles on the website were downloaded and 1060 articles out of 1126 were normalized into PubMed PMIDs. The text of the articles and the associated metadata of the 1060 articles were downloaded using PubMed E-utils services in XML format. The articles were divided into two sets, 898 articles published before 2010 as training set and 162 articles published after 2010 as evaluation set.

6.3 Discussion

In this section, in addition to the manuscript provided, I discuss complementary aspects of the Pub*Athena* design.

Pub*Athena* development is based on a lightweight local framework approach implying that for the information retrieval (IR), only the articles of a query are downloaded and indexed. This contrasts with other data-centred TM resources that have already pre-downloaded and pre-indexed whole, or part of, MEDLINE dataset, hence requiring a heavy file system architecture (*e.g.* for PubTator, ~8 GB of storage for the whole raw MEDLINE dataset, plus >30 GB for the tokens indexation and >10 GB for biological entity annotations). Hence, remaining up-to-date with MEDLINE requires a heavy architecture. For the purpose of literature tracking of updated information on dedicated subjects, the choice of a lightweight architecture thus appears more appropriate.

Another concern of this lightweight architecture is the response reactivity of the system. In Pub*Athena*, the information extraction (IE) is performed on the fly only on the downloaded articles before being stored locally, and depending on the size of the query or the Internet connection, this process can take from a few minutes to thirty minutes (data not shown). Other TM resources, such as MedlineRanker or PolySearch2 perform a pre-indexing of the tokens of all new articles in order to have a quicker response (seconds or minutes). However, for a literature tracking purpose of the same query, that is executed periodically (daily or weekly). To remain lightweight, in Pub*Athena*, on the fly calculation approach was opted, and the IR and IE procedures are scheduled to be executed during off-peak periods at the NCBI (usually at 3 am or during week-ends). Hence, once the calculations have been performed and stored in the local database, the reaction time can be very short (in seconds) for consultation.

Another crucial point for not pre-downloading and pre-indexing the whole MEDLINE is the flexibility in Pub*Athena* to formulate a query event with free-text and to be able to perform at the same time the full range of IE procedures. However, in many cases the user needs to take into account non-standard nomenclature for very specific biological fields. Moreover, non-expert users can have difficulty in providing all the relevant keywords. Currently TM tools uses only partial solutions to prevent the inevitable exploding computing resources needed for a whole indexing of MEDLINE updates. Data-centred TM tools either index only the tokens of all abstracts and can accept free-text search but thus lacks of IE procedures to identify biological entities (*e.g.* MedlineRanker). Others tools index only a

subset of abstract with identified biological entities but restrict the formulation of a query to keywords on indexed biological entities (*e.g.* PolySearch2).

Chapter 7 CilioPath: a compendium of ciliary knowledge

In this chapter, I firstly provide a state-of-the-art of the resources related to the cilia/ciliopathy fields that are currently available. Secondly, I present an overview of the philosophy of CilioPath, a knowledge base prototype for the integrated management and deciphering of ciliopathy data. Thirdly, I give some preliminary results and applications currently underway *via* the use of CilioPath.

7.1 State-of-the-art of ciliary resources

As introduced in Section 2.5.1, the cilia once over-looked as a vestigial organelle, is now gaining resurgent and growing interest with the discoveries of its roles in several essential cellular processes and its involvement in various genetic diseases (Benmerah et al., 2015; Bloodgood, 2009). Initial estimations of the comprehensive set of ciliary genes/proteins (referred to as the ciliome) was evaluated around 250 proteins (Benmerah et al., 2015; Bloodgood, 2009). However, in the lattest years, many High-Throughput Studies (HTS) have been performed based on Mass Spectrometry (MS), expression profiling (TXN), comparative genomics (CG) or transcription factors binding site profiling (XBX). These HTS deeply change our understanding of the ciliary landscape, which are far more complex with thousands of proteins and numerous other molecules (RNA, lipids, ions...) directly or indirectly involved in ciliary function and biogenesis.

As a result, there have been several attempts to generate bioinformatics resources to collect this information in databases and allow *in silico* analysis of the ciliary genes sets. To better understand the strength and weaknesses of these databases, I performed a comparative analysis of their contents and functionalities (Dutcher, 1995).

Schematically, these databases can be categorized according to their focus, namely the centrosome/basal body or the cilia/flagella (Geremek et al., 2014; Liu et al., 2007). The ciliary related resources include:

1) Two centrosome-centric databases:
   a) CentrosomeDB [http://centrosome.cnb.csic.es],
   b) CepDB [http://www.cebi.sdu.dk/CepDB])
2) Five cilia-centric databases:
   a) Ciliome database [http://www.sfu.ca/~leroux/ciliome_home.htm],

b) Cilia proteome [http://www.ciliaproteome.org],
c) Cildb [http://cildb.cgm.cnrs-gif.fr],
d) SYSCILIA Gold Standard (SCGS) [http://www.syscilia.org/goldstandard.shtml],
e) Cilia/Centrosome Complex Interactome (CCCI) [http://ccci.tigem.it].

Regardless of the disparate set of functionalities and browsing options (*e.g.* downlable, gene query, genome browsing, motif search; see Table 7-1), one of the major caveats of these databases, is that presently most are not up to date (*e.g.* the lattest update is 2014!). Some databases are not even maintained such as Cilia proteome, Ciliome database or CepDB. All these databases integrate mainly HTS ciliary data, such as proteomics or transcriptomics and are thus, limited by the caveats associated to this type of experiments (poor sensitivity, reproducibility, coverage, protocol/threeshold constraints...). In addition, all the ciliary-centric databases integrate the same core set of 8 HTS (Avidor-Reiss et al., 2004; Blacque et al., 2005; Efimenko et al., 2005; Li et al., 2004; Ostrowski et al., 2002; Pazour et al., 2005; Smith et al., 2005; Stolc et al., 2005) implying that potential errors might be propagated in all databases. Currently, Cildb represents the best source of potential ciliary genes with the largest collection of experiments (including some from other databases) and up to 21 proteomics experiments involving several species and, as such, Cildb is probably enriched for common eukaryotic ciliary proteins. Cildb is used as primary source for the SYSCILIA Gold Standard database (SCGS) that represents a set of confirmed ciliary genes (manually reviewed by experts) and for Cilia/Centrosome Complex Interactome (CCCI) that provides potential information on ciliary gene interactions and functional networks. All together these compilations of HTS result in a potential list of ~3000 putative ciliary candidate genes, which may incorporate unrelated genes (biased by the HTS experiments) and still miss genuine ciliary genes due to the HTS approaches (*e.g.* BBIP1/BBS18 gene).

Features			Ciliome database	Ciliaproteome	CentrosomeDB	Cildb	CepDB	SCGS	CCCI
Version			1	3	1	3	1	1	1
Active			🟢	🔴	🟢	🟢	🔴	🟢	🟢
Generation	Date of creation		2006	2006	2008	2008	2011	2013	2014
	Last update		2007	2009	2013	2014	2011	2013	2014
Organism	Organism centric		-	*Homo sapiens*	*Homo sapiens*	*Paramecium tetraurelia*	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*
	# different species		6	6	2	44 (32 cilia specifics)	1	1	1
Source of GCA	External datasets		-	-	MiCroKit, HPRD, GO	-	-	Cildb V2	Cildb V2, STRING
	Validation studies		🔴	🔴	🔴	🔴	🔴	🔴	🔴
	HT studies	# total studies	8	11	1	66	1	-	-
		# of HT ciliary studies	8	11	1	42	1	-	-
		# of HT ciliary experiments	9	11	1	46	1	-	-
		CG	2	2	0	5	0	-	-
		MS	3	6	1	21	1	-	-
		TXN	2	1	0	17	0	-	-
		XBX	2	2	0	3	0	-	-
Ciliopathy genes flag			0	0	0	0	0	97	83
# of human ciliary genes			4898	7855 (1162*)	773	4547‡ (2014*)	177	495‡ (304*)	3540‡
Exportable			🟢	🟢	🟢	🟢	🔴	🟢	🟢
Browsing options	BLAST search		🔴	🟢	🟢	🟢	🔴	🔴	🔴
	Motif serach		🔴	🔴	🔴	🟢	🔴	🔴	🔴
	Free-text search		🟢	🟢	🟡	🔴	🔴	🔴	🔴
	Orthologs search		🟡	🟡	🟡	🟢	🔴	🔴	🔴
	Gbrowse		🔴	🔴	🔴	🟢	🔴	🔴	🔴
PMID			16860433	16940995	18971254	25422781	21399614	23725226	25102769

*Table 7-1: Survey of ciliary databases. The sources of Gene Ciliary Association (GCA) are classified according to (i) External Datasets used, (ii) Validation studies, (iii) studies with High-Throughput (HT) Technologies, CG: Comparative Genomics; MS: Mass spectrometry / Proteomics; TXN: Transcriptome / Expression profile; XBX: Analysis of genes with X-box. The symbol ‡ indicates the number of potential human ciliary genes with low confidence, and * the number of high-confidence human ciliary genes (at least 2 ciliary evidences and at least a human ciliary evidence).*

However, as for the ciliopathies, there is currently no identified resource, which compiles all the related information. I also screened five disease related datasets to assess what was available on ciliopathies. There are two main categories of disease-related datasets:

1) Datasets with an identified "ciliopathy" category:
    a. Orphanet

Orphanet is the European portal for rare disease and orphan drugs. Orphanet generates its own classification system, which is widely used in the scientific and biomedical community. While it does identify a distinct ciliopathy group, the latter does not reference any disease in it. Moreover, when using a simple query search on the portal for "ciliopathy", only 3 retinal ciliopathies can be identified (Bardet-Biedl syndrome, User syndrome, retinal ciliopathy due to RPGR gene).

    b. Disease ontology (DO)

The DO is an open source ontology, which aims to integrate all the biomedical data associated to a human disease. In the "ciliopathies" group of DO only 3 ciliopathies are referenced (Joubert syndrome, Meckel syndrome and Kartagener syndrome). Most of the ciliopathies, such as the iconic Bardet-Biedl Syndrome is referenced in the generic group of autosomal genetic diseases.

2) Datasets without any "ciliopathy" category:
    a. International Classification of Disease (ICD-10)

The ICD-10 is a disease classification system created and maintained by the World Health Organization, which is used by physicians, the health sector and policy-makers as a standard diagnostic tool for clinical purposes, epidemiology, and health management. In the ICD-10, individual ciliopathies are generally not classified and are often regrouped under "Other syndromes" categories. None of the ciliopathies could be obtained when using quering for the keyword "ciliopathy" in any of the disease records.

3) OMIM

OMIM is currently the most comprehensive set of human genes and genetic phenotypes. Even if OMIM has a "Phenotypic Series" category, which tends to regroup diseases corresponding to main syndromes, it does not have proper classification system of diseases and as such, no "Ciliopathy Phenotypic Series". Since each entry is richly documented, querying for the keyword "ciliopathy" can be very helpful to retrieve ciliopathy-

related entries. However, since all the records are gene-centric (149 entries with "ciliopathy" matching term), it is very hard to have the corresponding list of ciliopathies. Moreover, each record has then to be reviewed manually for curation, to see if the "ciliopathy" matching term is used in a classification or general citation context.

4) GeneReviews

GeneReviews (GR) is a highly curated disease dataset with a rigorous editing and peer review process. Each disease record is richly documented by field experts and provides cross-references to all of the corresponding NCBI resources. Using the "ciliopathy" keyword, only 9 ciliopathies are referenced.

7.2 Philosophy of CilioPath

These unmet urges for a resource, with curated and updated resource on ciliary genes and ciliopathies, have motivated the development of a prototypal knowledgebase, CilioPath. CilioPath is a cilia-centric knowledgebase ranging from the list of human ciliary genes and the associated annotated functions, to the compilation of the related ciliopathies and the corresponding phenotypic characteristics.

Currently, literature represents the prime source of the reporting of novel ciliopathies information. For instance, ciliopathies or novel genotype-phenotype relationships are usually reported in case-reports. Similarly, novel ciliary/ciliopathy genes are frequently reported in experimental validation articles that confirm the ciliary function of a specific gene and its potential association to a ciliopathy. To this end, in order to have a semi-automated up-to-date infrastructure, we are developing CilioPath (Figure 7-1) based on the already presented resources, Pub*Athena* (Chapter 6) and the BioData Toolkit (BDT, see Chapter 3 and section 4.4.4).

The development of the CilioPath knowledgebase is organized around a MySQL database. The BDT infrastructure is used to instantiate the database with multi-level datasets (Figure 7-1A). In this prototypal version, only the *Homo sapiens* genome has been included with the version 75 of the Ensembl dataset, which corresponds to the GRCh37/hg19 version of the Human genome assembly. The mapping procedure for gene identifiers (described in section 3.1) is used to integrate multi-level information datasets *via* BDT into the CilioPath database:

i. Variant level information: reported disease-causing variants in ClinVar and OMIM databases.
ii. Gene/Protein level information: gene functional annotations are retrieved from GO, HGNC and UniProt databases. Expression information are retrieved from the Human Protein Atlas dataset. Additionally, biological material availability is retrieved from antibodypedia, the Mouse Genome Database (MGI) and the Zebrafish genome database (ZFIN).
iii. Pathway/Network level: pathway datasets are retrieved from KEGG and Reactome and network datasets are retrieved from IntAct and STRING-db.
iv. Phenotype/Disease level: disease information is retrieved from OMIM, Orphanet and GeneReviews, while the related phenotypic information is obtained *via* the Human Phenotype Ontology (HPO).

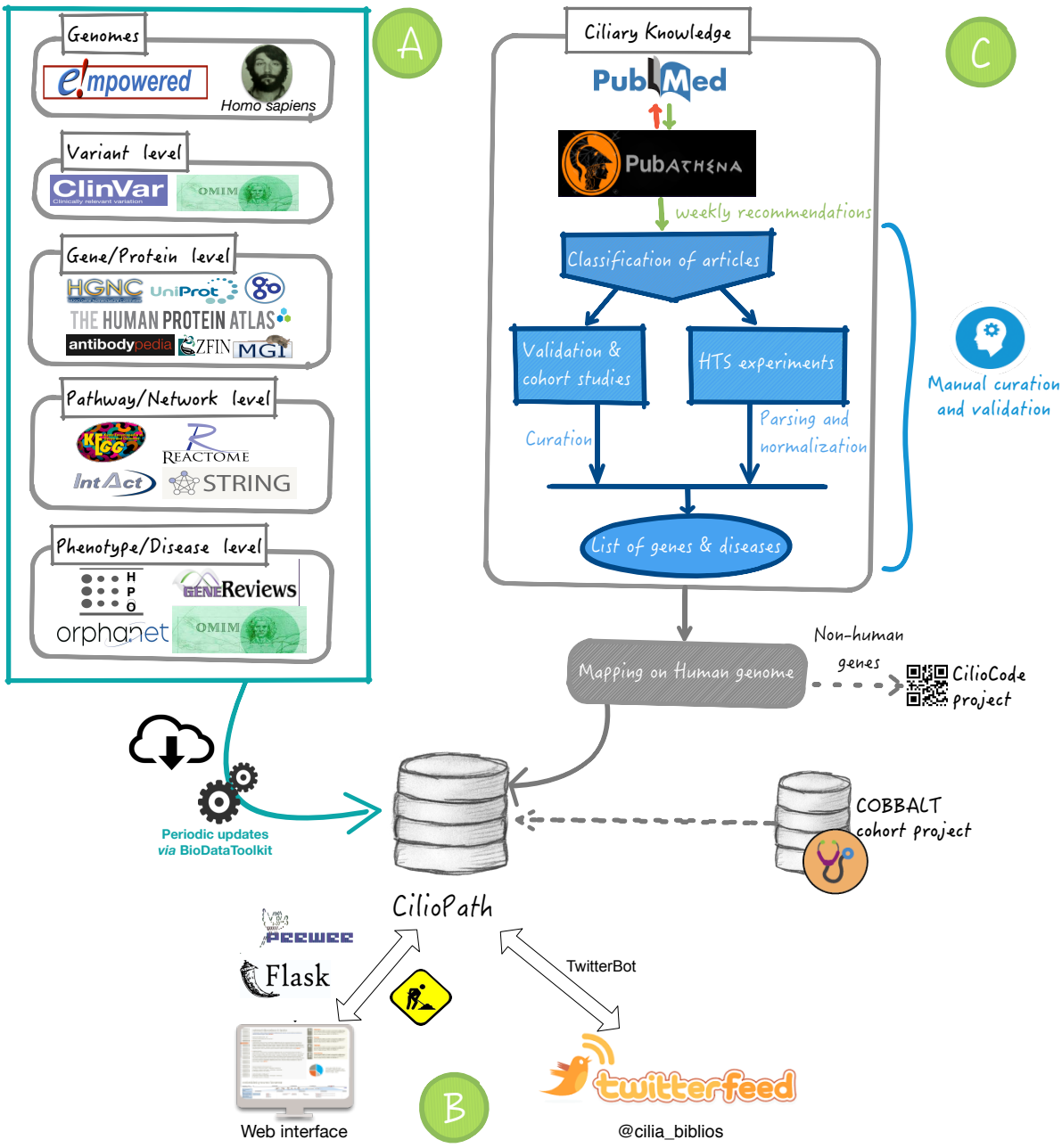


Figure 7-1: Overview of the prototypal architecture of CilioPath. The current design of CilioPath comprises three main parts: (A) – An automated module for the retrieveal of updated multi-level generalist data (from genomes to phenotypes), (B) – Interfaces (graphic user interface and web services) to query the database. (C) – An expert curation process of weekly cilia/ciliopathy related articles recommended by PubAthena.

Moreover, the BDT infrastructure allows the automatic synchronisation of each dataset at every release (monthly basis) in order to have the latest updated information, except for the daily update of OMIM.

Currently, the interface of CilioPath is under active development (Figure 7-1B). The adopted strategies for the access to CilioPath knowledge base are *via*:

1) **A web portal.** The final web portal is currently under development, in order to browse the data with: (i) custom queries, such as gene, ciliopathy or phenotype (ii) advanced queries such as the querying of the ciliome network or by biological function/pathway. Currently the technology chosen is a Python interface, which uses the Peewee v2.7 library to interact with the MySQL database and the Flask v1 framework for the web interface. The developed CilioPath Python package also uses the library SciPy for the scientific/statistical calculation and to generate the graphics to be displayed.
2) **Web services**. The developed CilioPath Python package with the Flask library is implemented in order to have programmatic access to database to submit multiple distant queries and batch downloads. For instance, a prototypal version of a TweetBot (also known as Twitter robot) has been development to provide automatic live feeds on updated ciliary information. The Twitter account of CilioPath TweetBot, @cilia_biblios, has already tweeted >150 messages on novel ciliary and ciliopathy related articles since March 2016. While, this prototypal TweetBot is still confidential (without public communication), it already has around 22 followers, some of whom are researchers or even institutional organisation such as RareConnect.org / EURODIS (@RareConnect) or the South African organization for Rare Diseases (@RarediseasesSA)

As previously stated, in order to have the most comprehensive set of ciliary/ciliopathy related knowledge and to remain up-to-date with the latest advances, our main effort has been related to the extraction of ciliary/ciliopathy related information from PubMed literature using Pub*Athena.* To achieve this objective, we define a Pub*Athena* recommendation model using a training set of 234 selected articles (published before 2013) with well-established ciliary genes and ciliopathies data and a neutral/negative dataset composed of 10,000 random articles published in 2013. This Pub*Athena* instance is configured to provide a list of recommended ciliary/ciliopathy related articles with the associated meta-data (*e.g.* extracted Biological Entities – genes, diseases; see Chapter 6) on a weekly basis.

In the current prototypal version, all the recommended articles are reviewed manually for classification (based on a field experience) and validation (Figure 7-1C). Based on the abstract, each article undergoes a multi-level classification/tagging according a one-letter code system defined as follows (Table 7-2):

1) **Archetype classification** - based on the context on the abstract. Five main exclusive types of articles have been identified:
    a) Bioinformatics (B) – articles concerning *in silico* protocols/methodologies applied to study the ciliary/ciliopathies function/network or bioinformatics resources such as ciliary databases (*e.g.* Cildb)
    b) Cohort (C) – articles presenting a case report with patient phenotype description or the characterisation of a ciliopathy related cohort. This type of articles usually provides information on the genotype-phenotype relationship of a given ciliopathy.
    c) High-throughput study (H) – articles based on a high-throughput approach to study the cilia. This type of article usually provides insights on the composition and function of ciliome according to different ciliary context.
    d) Review (R) – review articles on the cilia and ciliopathies. This can be very useful in order to obtain comprehensive lists of ciliary genes or ciliopathies with the canonical set of phenotypes/symptoms or to have an extended comprehensive update on a given ciliopathy.
    e) Validation study (V) – articles dedicated to a single gene and wet lab validation experiments. This type of article validates/establishes the biological or disease related function of a gene.
2) **Biological information** – the type of biological information treated in the article. An article can be multi-labelled by different types of information based on the automated biological entities provided by Pub*Athena* and the manual validation process. Five types of information have been defined:
    a) Disease (D) – articles about any given disease or phenotype.
    b) Function (F) – articles providing information on the role of a gene.
    c) Mechanism (M) – articles providing information on a pathway/mechanism leading to a biological/pathological response or the characterisation of interactions between genes/proteins.
    d) Species organism (O) – articles in which information discussed or the results obtained comes from an organism other than Human.
    e) Phylogeny (P) – articles discussing about the evolutionary or functional information that can be inferred through the use of a phylogenetic approach.

3) **Novel ciliary information** – correspond to novel ciliary/ciliopathy information/association that has not been previously reported/established in any other databases.
    a) Novel gene ciliary association (X) – correspond to an article reporting a gene for which its ciliary function/association has never been established before.
    b) Novel gene-disease association (Y) – correspond to an article reporting a novel establishment of the association of gene to a known ciliopathy.
    c) Novel ciliopathy (Z) – correspond to an article reporting the novel classification of a disease as a ciliopathy.

Depending on the archetype classification, some articles undergo a further manual processing. Notably, Validation (V) and Cohort (C) studies are curated in order to validate or reject the article based on the pertinence of the article. For instance, excluded articles comprise out-of-subject articles like cohort studies on smoking citing cilia in the abstract, biophysics articles on the ciliary beating or general articles on the Ciliates phylogenetic clade but without discussion about the cilia. Additionally, type H articles undergo a full manual processing of the complete article and supplementary data in order to extract the reported gene lists and their subsequent normalization into gene Ensembl identifiers.

At the end of this manual curation and validation process, all the validated articles with the extracted meta-data (*e.g.* genes or diseases) are inserted into a dedicated table of CilioPath. Additionally, the list of non-human genes is pre-processed to map most of the genes on its best corresponding human orthologs, using InParanoid (Sonnhammer and Östlund, 2015), in order to flag potential human ciliary genes in the CilioPath database (Figure 7-1C). Genes without human 'true' orthologs are flagged in the database to be processed in another in-house project, CilioCode (discussed in the discussion section). Moreover, the ciliopathy information is also curated from OMIM, GeneReviews and Orphanet records to identify the latest set of established ciliopathies (currently 19 identified).

Classification categories	Code	Name	Number of tagged articles	Description
Article type	B	Bioinformatics	7	Bioinformatics study on cilia
	C	Cohort	57	Cohort or case report studies
	H	High-Throughput study	28	Studies using high-throughput methods to study cilia
	R	Review	253	Review or methodological articles related to cilia or ciliopathies
	V	Validation	1014	Experimental studies validating the ciliary function or relationships of a gene
Article information	D	Disease	456	Articles concerning a disease or phenotype
	F	Function	1055	Articles inferring a gene function
	M	Mechanism	270	Articles with pathways, mechanisms or interactions information
	O	Model organism	630	Articles describing results obtained from non-human organisms
	P	Phylogeny	27	Comparative genomics studies
Novel ciliary information	X	Novel gene ciliary association	134	Novel association of a gene to a ciliary function
	Y	Novel gene disease association	75	Novel association of a gene to a ciliopathy
	Z	Novel ciliopathy	5	Novel classification as a ciliopathy

Table 7-2: Classification code used to curate and to categorize PubAthena recommended articles.

Overall, since the setup of CilioPath in 2013, 1362 cilia/ciliopathies articles have been curated and classified, while 120 articles have been rejected using the previously described coding system (Figure 7-2). The rejected articles correspond mainly to articles discussing about Ciliates organisms citing the "cilia" in the abstract, without direct reference to the study of the organelle and the second most rejected type of articles correspond to biophysics articles discussing about ciliary beating. Using this approach, we identify 134 GCA (Gene Ciliary Associations), 75 GDA (Gene Ciliary Disease Association) (See Annexe IV), and identify 5 novel reported ciliopathies (See Annexe IV).

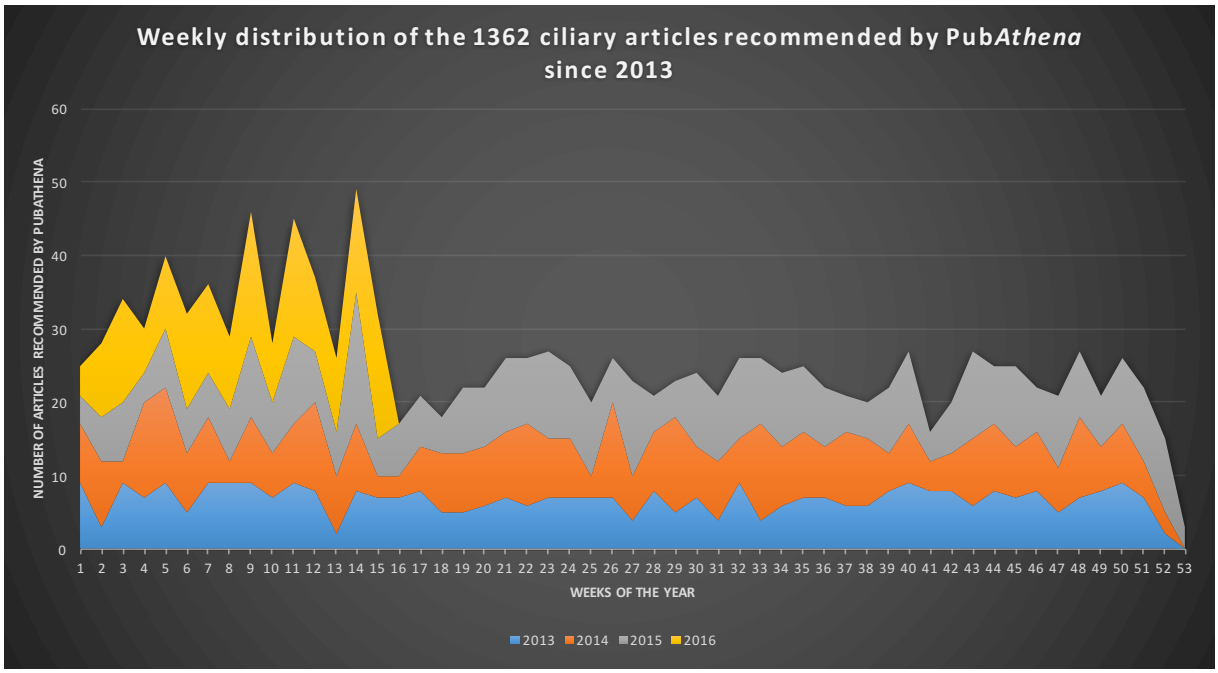


Figure 7-2: Distribution of curated ciliary related articles recommended by PubAthena since 2013.

Using this weekly process, CilioPath is currently the sole up-to-date database with a curated ciliary gene set compared to the other main ciliary resources (Cildb, SYSCILIA) and is currently the only resource that provides the most comprehensive compilation of ciliopathies (**Error! Reference source not found.**). When comparing with other curated resources, for GCA, the last update of SCGS (the sole curated cilia-related database) was in 2013. For GDA, and novel ciliopathy identification, CilioPath is the only comprehensive resource on ciliopathies. Other resources, such as OMIM, Orphanet, GeneReviews or UniProt, have a 4-6 month lag compared to CilioPath to record novel GCA or GDA. Hence, through this semi-automated procedure, CilioPath provides an exclusive list of 42 confirmed ciliary genes not provided in any other resource (See Annexe IV).

Disorder name	Synonyms	Causative genes
Alström syndrome (ALMS)		ALMS1
Bardet-Biedl syndrome (BBS)		BBS1-19, CCDC28B, AZI1
Birt-Hogg-Dubé syndrome (BHD)		FLCN
Cranioectodermal dysplasia (CED)	Sensenbrenner syndrome	IFT122, IFT43, WDR19, WDR35
Cystic Fibrosis Related Disorders (CFRD)	Cystic Fibrosis	CFTR
Ellis-van Creveld syndrome (EVC)		EVC, EVC2
Jeune asphyxiating thoracic dystrophy (JATD)	Jeune syndrome	DYNC2H1, IFT80, IFT172, IFT140, TTC21B, WDR19, WDR34, WDR60
Joubert Syndrome and Related Disorders (JSRD)	Joubert Syndrome, COACH syndrome, Acrocallosal syndrome	JBTS1-24, B9D1, C2CD3, IFT172, POC1B, TALPID3, TCTN2
Leber Congenital Amaurosis (LCA)		LCA1-18, IQCB1
McKusick-Kaufman syndrome (MKKS)		MKKS
Meckel syndrome (MKS)		MKS1-12, FLNA, SYNE2, TMEM107
MORM syndrome (MORM)		INPP5E
Nephronophthisis (NPHP)		NPHP1-16, CCDC41, XPNPEP3?
Oral-facial-digital syndrome (OFD)	Oral-facial-digital syndrome, Mohr-Majewski syndrome	OFD1, TCTN3, C2CD3, C5orf42, DDX59
Polycystic Kidney Disease (PKD)	Autosomal dominant or recessive form	PKD1, PKD2, PKHD1, STK36?
Primary Ciliary Dyskinesia (PCD)	Kartagener Syndrome	CILD1-30, AK7, CCDC164, DNAH1, DNAH8, DYX1C1, MCIDAS, RPSH3
Retinitis Pigmentosa (RP)		RP1, RP2, RPGR, ARF4, ARL3, C2orf71, C8orf37, CNGB1, FAM161A, MAK, NEK2, POC1B, TOPORS
Senior-Loken syndrome (SLSN)		SLSN1-8
Short-Rib Polydactyly Syndrome (SRPS)	Short-Rib Polydactyly Syndrome, Majewski Syndrome	C2CD3, DYNC2H1, DYNC2H1, IFT80, NEK1, WDR19, WDR35, WDR60
Usher syndrome (USH)		USH1B, USH1C, USH1D, USH1E, USH1F, USH1G, USH1J, USH1K, USH2A, USH2C, USH2D, USH3A, USH3B, CEP250, DISC1, NINL

Table 7-3: Non-exhaustive list of main ciliopathies referenced in CilioPath.

Another major aspect of CilioPath is related to the phenotypic information associated to each gene from genotype-phenotype relationship studies. Like in most databases, it includes information from HPO database to infer phenotype (Figure 7-1A), however, a special care is taken for the inference of the gene-disease-phenotype relationships, using the literature as a validation. The automatic HPO terms assigned to a gene or disease are weighted according to the information extracted and validated in reference datasets such as GeneReviews or reported in R articles.

Indeed, in HPO, all the phenotypes of a given disease/syndrome are generally inferred to every gene of that disease. However, in the case of ciliopathies, such a direct and total transfer can be error-prone. Moreover, disorders such as ciliopathies (*e.g.* Bardet-Biedl and Alström syndromes) do not only overlap phenotypically, but distinct ciliopathies (*e.g.* Bardet-Biedl, Joubert, Meckel, Senior-Loken syndromes) can also share common genes (usually genes involved in the transition zones), such as CEP290, which can produce >3 different ciliopathies. For instance, considering the Bardet-Biedl Syndrome, which is usually related to a post-axial polydactyly, it could be misleading to infer this information to the 21 BBS genes. For instance, the gene LZTFL1/BBS17 is specifically associated to a meso-axial polydactyly compared to the other BBS genes.

Hence, in order to evaluate the degree of relationship between genotype and phenotype provided in the literature and public databases, we collect normalized phenotypical data (using HPO terms) from patient cohorts (COBBALT, Figure 7-1E). We have already started to collect data from the Hélène Dollfus' COBBALT cohort of 400 individuals, which includes patients from two closely phenotypically related ciliopathies: the Bardet-Biedl and Alström syndromes. Patients in this cohort were evaluated other 80 phenotypical criteria, by the clinicians of the CARGO (in collaboration with Dr Elise Schaefer and Pr Hélène Dollfus). The complete set of symptoms in the clinical report of each patient is exported into an Excel file. The standardized format of the Excel file in our protocol starts with the first column corresponding to the patient ID, the next 5 columns to his meta-data, namely the age, gender, health status, pedigree position, molecular diagnosis (identified causative mutation), ethnic origin and finally each assessed symptom is distributed in a single column. The status of the assessed symptom is codified into a single digit code with 0 for "unknown", 1 for "absent" and 2 for "present". With an in-house script, each patient record is then automatically normalized in its corresponding HPO term using matching terms into a phenotypical thesaurus based on HPO. Any eventual none normalized symptoms are then manually

reviewed in collaboration with the clinicians to identify the corresponding synonymous HPO term. All the data are stored in a MySQL database, that can be used in-house by the clinicians to create and update a patient records. The COBBALT database can be queried for the statistical analysis using the DataGrid software interface. This preliminary work has helped to setup a semi-automated protocol to normalize/digitalize all the novel patient records, which is currently under evaluation.

This project represents the first initiative to normalize a cohort of patients with more than 10 criteria and by using an ontological approach, namely with HPO. Moreover, the COBBALT cohort represents the world's largest compiled and documented cohort of BBS patients. In the current prototype of CilioPath, we started to focus on the iconic ciliopathy BBS, which comprises 400 patients, for which we identified the causative mutation in 64% of the cases. For instance, preliminary statistical analysis show that the criteria of Beales (at least 3 Major symptoms + 2 minor symptoms) used to diagnose a BBS patient applies to less than half of the BBS patients (Figure 7-3). The data collected already suggest that the phenotypical data present in public databases need to be assessed carefully for RD with overlapping symptoms and genetic aetiology like ciliopathies and that the importance of knowledge based approach, such as CilioPath highly curated data are needed for proper study of genotype-phenotype relationship.

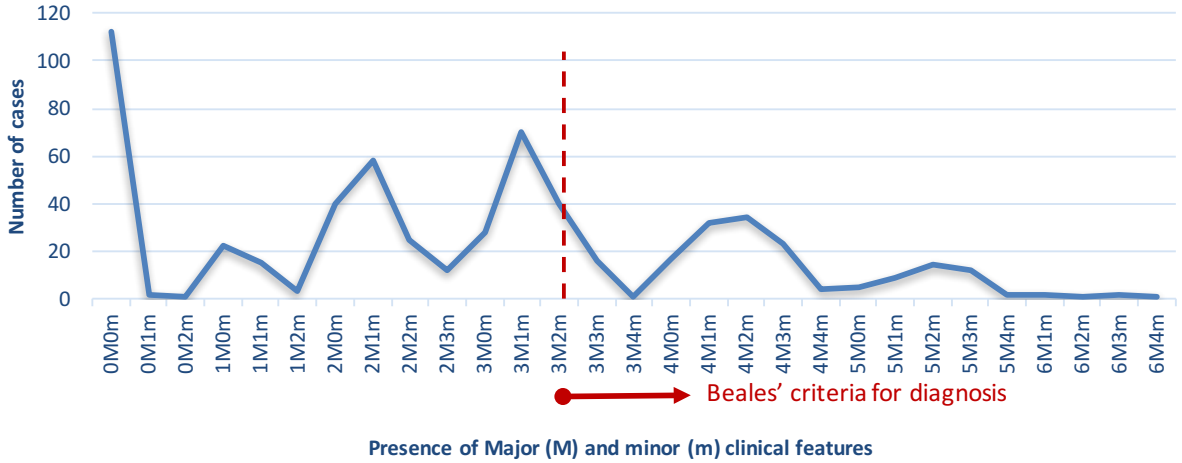


Figure 7-3: Distribution of Representation of the COBBALT cohort according to Beale's criteria for Bardet-Biedl syndrome. The cohort of ~400 suspected Bardet-Biedl syndrome (BBS) patients reported in this graphic are reported according to Beales' Major (M) and minor (m) clinical features used to establish a diagnostic. Beales' criteria diagnose a BBS in the presence of at least 3 Major clinical features and 2 minor clinical features or 4 Major clinical features.

7.3 Discussion

As previously discussed, the major objective of CilioPath is to compile, through recommended and curated ciliary articles, novel GCA, GDA and ciliopathies. The semi-automated procedure, to recommend articles with annotations and meta-data *via* Pub*Athena* (Chapter 6), has already proved to ease the curation process and to quickly identify the latest novelty compared to other curated resources (*e.g.* OMIM, UniProt). Compared to best existing ciliary resources such as Cildb or CCCI, CilioPath is already able to provide a dynamic gold standard (compared to SCGS) with the most up-to-date list of ciliary genes involving 42 exclusively confirmed ciliary genes (See Annexe IV).

Moreover, CilioPath is the first compendium of 1362 cilia/ciliopathy related articles that was generated with Pub*Athena* recommendation model, with the initial goal to track literature updates. The rejection of 120 articles, after manual curation, from the recommended articles by the current model has enabled the identification of key factors influencing the performance of Pub*Athena*. Besides specific keywords used for ciliary related articles, additional training data is needed to refine the recommendation model in order to exclude for instance biophysics articles on the cilia, or general articles on Ciliates. Future directions, for the improvement of the recommendation model, will include the automatic screening and evaluation of the articles published before 2013, in order to constitute a more comprehensive training set to generate a new model. The strategy will be to develop a self-improving recommendation model, which could be periodically and automatically regenerated by integrating novel curated articles into the CilioPath database. Additional machine learning models will be trained in order to automatically tag the different sub-categories of articles, so as to automatize the classification of articles step in the ciliary knowledge module (Figure 7-1B). The aim is to be able to automatically dispatch novel articles among members of a curation board, for simple validation with most of the biological concept and ciliary information automatically extracted.

For instance, identification of novel pathways/interactions and biological processes associated to the cilia in type F, M and V articles by expert biologists, extraction of novel ciliary gene list in type H articles with dedicated scripts by bioinformaticians or the assessing of cohort studies and novel phenotype association studies (type C/D/R articles) by clinicians to update information on the cilia and ciliopathies.

In the prototypal version of CilioPath, only ciliary genes related to humans are included. However, novel non-human ciliary genes as well as the phenotypical data from non-

human model organisms (*i.e.* mouse, zebrafish or the round worm; classically tagged O and X in CilioPath) are kept and tagged for future CilioPath developments. In addition, these non-human ciliary genes feed another in-house project carried out by a PhD student, Yannis Nevers, and named CilioCode (Figure 7-1E). This project aims to identify, *via* knowledge extraction of comparative genomics and 'omics' data, discriminative relationships between ciliopathy modules or phenotypes and ciliary gene patterns inferred from conservation, phylogenetic distribution, expression profiles, functions....

At the cohort characterization level, the semi-automated protocol for the normalization of patient symptoms into HPO terms is currently applied to the cohort of 600 potential ciliopathy patients from the Laboratory de Diagnostique Génétique at the Nouvelle Hospital Civil of Strasbourg in collaboration with Dr Jean Muller. Nevertheless, we are still characterizing the COBBALT cohort data to update and better identify the phenotypical distribution and relationships in the BBS patients. Aside from the statistical genotype-phenotype relationship analysis (*e.g.* comparison with Beale's criteria for BBS diagnosis), our aim is to study the potential relationships between some phenotypes and ciliary functions/modules as inferred from the data compiled in CilioPath. For instance, the study of the distribution and severity of major BBS symptoms such as obesity or kidney failure, with regards to the type of mutated ciliary gene/function (*e.g.* implied in the core BBSome v/s other networks).

Finally, it should be noted that while CilioPath was initiated in 2013, it recently came up that the SYSCILIA consortium is creating a project named also CilioCarta and linked to the SCGS dataset. Hence, in order to avoid any misguidance/confusion between the two projects, we are planning to change the name CilioCarta to CilioPath. Moreover, we wish to collaborate with the SYSCILIA consortium in order to offer to the scientific and biomedical community, the access to a unified resource on gene-centric and ciliopathy-centric data.

IV. Conclusion and Perspectives

Chapter 8 Conclusion

The discovery of the heredity's laws in 1866 and later, of the structure of the main molecule for heredity, the DNA, provided major insights to understand life and fuelled innumerable advances in biological and biomedical research. Nowadays, in the context of the post-genomic era, high-throughput biotechnologies (genomics, proteomics, transcriptomics…) have emerged, turning biology into a data-intensive science, and revolutionizing by the way the biomedical research and medical practice. In biology, the exponential accumulation of data has enabled the emergence of a networked and multi-level understanding of the biological processes. In medicine, with the access to the individual human genome, high-throughput technologies are inexorably spreading towards clinical applications and are starting to be part of the routine medical practice to provide better diagnosis, prognosis and therapeutics.

In this renewed biomedical landscape, rare diseases studies, which were restricted by the limited number of patients and the scarcity of data, have gained a resurgent interest and momentum. As a whole, rare genetic diseases (RD) concern ~300 millions of people worldwide and the genetic aetiology of ~50% of these diseases is still unknown. Central to the renewed interest for RD are the NGS technologies, particularly the Whole Exome Sequencing that has become the method of choice to identify the causative variant/gene and study genetic rare diseases. Thus, thanks to the easiest identification of causative variant/genes, RD are now considered as unique opportunities to understand the aetiology of many common diseases (obesity, diabetes...) by combining the simplified cause of the disease (frequently, a single gene variation) and the high-throughput biotechnology openings.

In this context, this manuscript presents the major projects carried out during this thesis, which represent the first initiatives towards the integration of data-intensive/big-data approaches into translational bioinformatics resources dedicated to the deciphering of thematic RD, such as ciliopathies.

Exome variant analysis

The first part of this manuscript addressed the issues of the current limitations in exome variant analysis, which results in a global resolution rate of ~25% of Whole Exome Sequencing (WES) experiments. As discussed, several factors impede the variant analysis, such as the quality of the sequencing data, the number of affected cases (influencing the

hypothesis on the modes of inheritance) or the current state of our knowledge on a given gene or disease.

To alleviate the low-resolution rate and facilitate the use and exploitation of WES experiments by clinician/researcher, we developed VarScrut as a strongly automated WES analysis software that follows the user concerns and practices from the variant analysis step up to the post-analysis phase. The variant analysis procedure has been designed as a multi-level approach with original features such as an automatic screening of the 5 genetics scenarios (autosomal recessive, autosomal dominant, linked to the recessive X, linked to the dominant X, *de novo*) and a particularly efficient dedicated logit model for the scoring of nsSNV variant. Combine with our knowledge-driven gene prioritization procedure, which integrates patients' phenotypes, this strategy proved to be particularly efficient since VarScrut clearly out-performed the currently widely used WES analysis tools (VAAST, Phen-Gen). Nevertheless, as most of the biocomputing and biotechnological improvements, these higher performances are probably temporary and we assume that the next tools will clearly out-performs VarScrut, hopefully by improving some of our original propositions.

More importantly, VarScrut is currently the sole software that provides post-analysis facilities to guide the validation step *via* the identification of biological materials (antibodies, mouse strains...) that might be helpful for the subsequent experimental validation steps and to provide follow-up procedures for unresolved exomes. These developments are a direct outcome of my stage in the Hélène Dollfus's laboratory and in the initial 'manual' WES analysis, which allows me to better understand the clinical practices, needs and bottlenecks. Besides the resulting WES post-analysis procedures, these developments emphasize that in the future, biocomputing systems will probably have to better take advantage of the current Big Data biological landscape to achieve integrated solutions.

Furthermore, in order to prepare for the inexorable future shifts towards routine genome sequencing in clinics, we yet identify some technical adaptations and improvements for a more efficient processing of variants. Firstly, VarScrut will be optimized to perform bit-wise operations to boost its query performances. Secondly, novel scoring model will be generated to take into account novel types of variants available with the genome, such as (deep intronic) splice variants or regulatory variants. Thirdly, in order to have a systematic and comparable performance across different sequencing projects, future developments will integrate also the BAM files as input files, in order to have the same variant calling procedure and have dedicated analysis of poorly covered regions. Fourthly, VarScrut will integrate a web-interface, which will facilitate its integration and use in routine clinics. Finally, an

instance of VarScrut will be dedicated to accompany a routine molecular diagnosis of retinopathies through the targeted sequencing of a panel of genes, currently under design and validation at the medical platforms in Strasbourg

Finally, one can pinpoint that VarScrut was successfull in identifying the novel BBS18 gene and the novel ciliopathy gene VPS15, which resulted in two publications that are not discussed in the current thesis manuscript (provided in annexes)

Text-mining tool for the tracking of literature updates

The second part of this thesis addresses the challenge of remaining up-to-date in the exponentially increasing flux of incoming data and in the specific context of rare genetics diseases where scientific literature is a primary source of information. In this scope, Pub*Athena* was developed to provide an original solution, as a lightweight text-mining tool, which uses a Naïve Bayesian Classifier machine-learning model to provide personalized recommendation of the most relevant articles and automatically extracts the associated meta-data (*e.g.* Biological Entities…). When compared to other text-mining tools, Pub*Athena* showed the most complete set of text-mining functionalities and demonstrated the highest performances for both the recommendation of relevant articles and literature updates.

Future works are needed to improve Pub*Athena* performances. In the perspective of a server edition of Pub*Athena*, Elastic Search will be implemented so as to boost the query search performance and distribute the query load over a cluster of servers. Moreover, a Jenkins server will also be setup so as to automatize the regeneration of the recommendation models based on literature updates inserted into the database of Pub*Athena.* This sever edition of Pub*Athena* will also integrate the possibility to generate multiple machine-learning models for the multi-tagging of an article.

Another possible lead of improvement is the indexing of all possible tokens or biological entities like MESH. While generating a complete index of a PubMed/MEDLINE will inevitably boost the query performances and improve the generation of recommendations models, this approach however will raise computational challenges to have a maintable solution (*e.g.* storage issues).

Towards a knowledge base on cilia and ciliopathies

Finally, the third part of the manuscript introduces an approach where all the developed resources are federated to construct a prototypal knowledge base, CilioPath, dedicated to the deciphering of cilia and ciliopathies. TheCilioPath integrates automatic updates of multi-level information from public databases using the BioData Toolkit infrastructure and uses Pub*Athena*, to semi-automatically integrate the most up-to-date and

relevant ciliary/ciliopathy information. Preliminary analyses show that this first compendium of ciliary/ciliopathy articles can be used to quickly extract novel exclusive ciliary/ciliopathy information when compared to other cilia-related resources based solely on high-throughput studies. Thus, CilioPath represents the sole source that can be used as a dynamic gold standard for ciliary data. Additionally, CilioPath integrates data from the first and largest worldwide cohort of BBS patients phenotypically normalized in HPO terms (COBBALT cohorts).

As a prototype, CilioPath needs numerous improvements. Notably, the finalisation of the web interface to allow the browsing and the open query of the data *via* its web services before its publication/communication. Secondly, CilioPath will also include additional data such as ciliary data for model organisms (mouse, zebrafish, round worm), and the generation of a machine-learning model to predict potential novel ciliary genes based on multi-level information (function, expression profile, homology, pathway, interaction network and genic intolerance to truncating variations).

For the COBBALT cohort, further characterization with a statistical analysis of the distribution and association of the different phenotypes within BBS patients is underway. Moreover, these data will also be explored to identify if there is a genotype-phenotype relationship in the observed variability of the major BBS symptoms and/or in the ciliary function/localisation of the mutated genes. The results of the characterization and normalization of this world largest cohort of BBS, with potential update in the definition of the BBS diagnosis criteria, will be integrated in CilioPath and probably give rise to a publication.

Chapter 9 Future Perspectives

Finally, as discussed in this thesis, biology has long been studied through phenotypes and heredity (transmitted phenotype). The identification of the underlying molecule, the DNA, and recently, the access to the complete genome has long given the illusion that all the phenotypes could be deciphered and explained. This is reinforced by the advent of multi-level high-throughput approaches that paved the way from a gene to a phenotype and are providing insights to decipher these relationships. However, this *Big Data* era has since provided indications that there is a long winding and complex road from a gene to a phenotype.

In genetic diseases, most of the HTS analysis approaches have been devoted to the assessing of deleterious protein-coding variants, often resulting in the central challenge of

having a biomedical interpretation of VUS. Inexorably, future approaches should also assess other types of variants (*e.g.* regulatory variants, epigenetic effects, spatial and structural constraints) and also additional levels of information. At the simplest level, one can easily anticipate that an approach integrating analysis of the variome data (DNA sequencing) together with transcriptomics (*e.g.* RNA-seq), proteomics or metabolomics data may enhance the understanding of the penetrance of the deleterious effects within the different tissues/organs of an individual. However, such integrative and costly approaches, which are currently experienced in some animal models, are not yet implemented in clinical practices and more importantly, had to be deeply evaluated with respect to effective improvement of patients' wellness. With the continuous drop of the computational capacities and high-throughput technologies, one rising issue will be the usage and privacy of such data. For instance, for the advancement in knowledge and processing of the data, data have to be accessible to the scientific community but even gold standard exomes or genomes, though published and widely referenced too, are still under 'embargo' with very limited access. These caveats are strongly associated to the possible traceability back to the identity of the patient with the famous case of the FDA against the 23andMe company. These contradictory concerns are further emphasized by the recent opportunities allowing access not only to the whole current biological identity and state of an individual but also to his possible future becoming (*e.g.* the incidentalome). Based on machine-learning approaches (and soon deep-learning) that can integrate and process the mass of *Big Data,* current models can provide generalist prediction models. But with the increasing accumulation of individuals' data/cases, future models should be able to look forward and take into account the uniqueness of each individual context and thoroughly provide "integrated", probabilistic medical practice. Could we still investigate personal 'omics' data only for the advent of a clinically suspected disease and still ignoring the forthcoming of another unsuspected disease? How 'probalistic medicine' could be disclosed and benefit to the patients? All these open questions clearly pave the way of the next personnalized medicine.

With respect to RD, in this era of *Big Data*, we will probably still facing the paradox of constant and voluminous flux of data with scarcity of data. But RD now tends not only to represent the new bonanza to study common diseases but also to modify our understanding and approach of a disease.

Indeed, monogenetic subsets of rare diseases are considered as a unique configuration to have insights on the underlying networks and mechanisms of RD and common diseases. More importantly, as the limited number of patients allows a personalized clinical and

familial follow-up, one can ambition to define the framework for a precise grasp of the natural history of the disease. The interest for such formalisation and depiction of the disease natural histories is increasing notably with the recent NIH funded initiative to study "the course that a disease takes in affected individuals from the time immediately prior to its inception, progressing through a presymptomatic phase and different clinical stages, to a final outcome in the absence of treatment".

In this context, ciliopathies are emblematic of RD. With the ubiquitous presence of the cilia, ciliopathies are frequently pleiotropic as in BBS and affect a broad spectrum of cell-lines and organs. Ciliopathies and BBS, hence represent ideal models to study common diseases such as ocular disorders of genetic origin or obesity (modern pandemic). More broadly, the recent emergence of the notion of ciliopathies has introduced a new perspective. Indeed, while previous diseases were mainly defined by symptoms and organ descriptions, for the first time, these diseases are categorized according to another level of resolution, an organelle.

This appeals to wider questions, such as: Will the use of high-throughput approaches and dedicated bioinformatics approaches leverage the reclassification of disorders into novel organellar/networked categories? Does this 'organellar viewpoint' of ciliopathies, is applicable to other diseases and prefigures the emergence of mitopathies, golgipathies, peroxisopathies? Above all, to what extent, the enhanced understanding of the perturbed homeostasis of a system, network or compartment, can effectively improve the care and wellness of the patients and help to design future therapies?

Glossary

Ciliome	The comprehensive set of ciliary genes/proteins.
Compound heterozygote	An individual carrying two different mutant alleles at a particular gene locus, one on each chromosome of a pair.
De novo variation	A new variation that is not inherited from either parent of the carrier.
Downloadability	The property of being downloadable.
Exome	The part of the genome comprising the majority if not all coding exons in all genes expressed as protein or other functional gene products.
Genome	The complete DNA sequence of an organism.
Homozygosity mapping	A gene mapping method used in rare recessive disorders often in consanguineous families. It is based on the assumption that an allele responsible for disease is likely to occur from a common ancestor and looks for shared regions of inherited DNA between subjects.
Incomplete penetrance	A phenomenon whereby not all individuals carrying a disease-associated variant will express the associated trait.
Locus heterogeneity	The appearance of phenotypically similar characteristics resulting from variations at different genetic loci.
Mendelian disorders	Genetic disorders that occur due to alterations or mutations in a single gene. Their pattern of inheritance can be recessive, dominant, or X-linked.
Next Generation Sequencing	Sequencing technology able to sequence thousands or millions of DNA regions at once—also referred to as second-generation sequencing.
Nonsense variation	A point variation in a coding DNA sequence that results in a premature stop codon.
Nonsynonymous variation	A point variation in a coding DNA sequence that changes the amino acid encoded at that position.
Penetrance	The proportion of individuals carrying a particular variant who also express an associated trait (phenotype).
Pleiotropy	Influence of one gene or variant on multiple phenotypic traits.
Single Nucleotide Variant	A single variation of one base at a particular site of DNA.
Stand-alone program	A program that runs as a separate computer process and not as an add-on to an existing process.

Synonymous mutation	A point mutation in a coding DNA sequence that does not result in amino acids change in the encoded protein.
Translational medicine	An interdisciplinary branch of the biomedical field that combines disciplines, expertise, and techniques to promote scientific advances to develop enhancements in prevention, diagnosis, and therapies.
Updateability	The act of updating or the property of something to be updated

Annexes

Annexe I: Bardet-Biedl syndrome: cilia and obesity - from genes to integrative approaches

Annexe II: Exome sequencing of Bardet-Biedl syndrome patient identifies a null mutation in the BBSome subunit BBIP1 (BBS18)

Annexe III: A mutation in VPS15 (PIK3R4) affects IFT20 cilia trafficking and causes a ciliopathy

A mutation in *VPS15* (*PIK3R4*) causes a ciliopathy and affects IFT20 release from the cis-Golgi.

Corinne Stoetzel[1,7], Séverine Bär[2,7], Johan-Owen De Craene[2], Sophie Scheidecker[1], Christelle Etard[3], Johana Chicher[4], Jennifer R. Reck[2], Isabelle Perrault[5], Véronique Geoffroy[1], Kirsley Chennen[1], Uwe Strähle[3], Philippe Hammann[4], Sylvie Friant[2,8], Hélène Dollfus[1,6,8]

1. Medical Genetics Laboratory, INSERM U1112, Institute of Medical Genetics of Alsace, University of Strasbourg, Strasbourg Medical School, 67000 Strasbourg, France
2. Department of Molecular and Cellular Genetics, UMR7156, Centre National de Recherche Scientifique (CNRS), Université de Strasbourg, 67084 Strasbourg, France.
3. Institut für Toxikologie und Genetik, Campus Nord, Karlsruher Institut für Technologie, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein Leopoldshafen, Germany
4. Institut de Biologie Moléculaire et Cellulaire (IBMC), Plateforme Protéomique Strasbourg - Esplanade, FRC1589, 67084 Strasbourg, France.
5. Laboratory of Genetics in Ophthalmology, INSERM UMR1163, Institut Imagine, Université Paris Descartes Sorbonne Paris Cité, Hôpital Necker, 75015 Paris, France.
6. Centre de Référence pour les affections rares en génétique ophtalmologique, CARGO, Filière SENSGENE, Hôpitaux Universitaires de Strasbourg, 67091 Strasbourg, France.
7. These authors contributed equally to this work.
8. Correspondence should be addressed to S.F or H.D. (e-mail: s.friant@unistra.fr or dollfus@unistra.fr)

Abstract

Ciliopathies are a group of diseases that affect kidney and retina among other organs. Here, we identify a missense mutation in *PIK3R4* (phosphoinositide 3-kinase regulatory subunit 4, named *VPS15*) in a family with a ciliopathy phenotype. Besides being required for trafficking and autophagy, we show that VPS15 regulates primary cilium length in human fibroblasts, as well as ciliary processes in zebrafish. Furthermore, we demonstrate its interaction with the golgin GM130 and its localisation to the Golgi. The VPS15-R998Q patient mutation impairs Golgi trafficking functions in humanized yeast cells. Moreover in VPS15-R998Q patient fibroblasts the intraflagellar transport protein IFT20 is not localised to vesicles trafficking to the cilium but is restricted to the Golgi. Our findings suggest that at the Golgi, VPS15 and GM130 form a protein complex devoid of VPS34 to ensure the IFT20-dependent sorting and transport of membrane proteins from the cis-Golgi to the primary cilium.

Introduction

The primary cilium is a non-motile polarized microtubule-based hair-like structure protruding from the surface of many eukaryotic cells, with exception of higher plants and fungi. Primary cilia are sensing extracellular molecules and environmental stimuli. By dynamic exchanges of signalling molecules they convert extracellular signals to intracellular responses [1]. They have a pivotal role in different signalling cascades: the Sonic Hedgehog, Wnt, Notch signalling and mTOR pathways. In humans, cilia dysfunctions can result in developmental defects such as *situs inversus*, polydactyly or central nervous system abnormalities as well as progressive organ alteration, including, retinal degeneration or cystic kidneys and form a class of clinically and genetically heterogeneous diseases termed ciliopathies [2, 3]. Ciliopathies are due to disease-causing mutations in genes encoding ciliary proteins and their study has revealed various actors of ciliary biogenesis and function, as well as unexpected links to previously characterized cellular pathways [2].

Early stages of ciliogenesis are characterized by the maturation of the mother centriole into a basal body as it migrates from its perinuclear position to the plasma membrane, followed by cilium elongation after extension of the axoneme [4]. Axoneme extension and cilia signalling largely rely on the evolutionary conserved mechanism of intraflagellar transport (IFT) via the IFT-A/B protein complexes that ensure bidirectional trafficking along the ciliary axoneme, by using kinesin-2 (anterograde) and dynein (retrograde) molecular motors. Among these IFTs, only the IFT20 protein has a dual localization at Golgi and basal body. At the Golgi, IFT20 ensures the sorting of ciliary cargo from the cis-Golgi to the base of the cilium [5, 6]. Ciliary membrane proteins are either delivered from the Golgi to the plasma membrane before moving towards the basal body or they are directly transported from the Golgi to the cilium [4]. At the base of the cilium, a complex termed BBSome formed by a subset of BBS proteins recognizes ciliary cargoes [7, 8]. The BBSome shares structural features with clathrin and coatomer coats (COPI and COPII) forming a coat on liposome membranes once recruited by Arl6-GTP [7, 8].

Here, we report the identification of a mutation in the *PIK3R4/VPS15* gene (MIM 602610, NM_014602.2: c.2993G>A, R998Q) in three affected siblings with a ciliopathy phenotype. VPS15 (Vacuolar Protein Sorting 15) encodes the phosphoinositide 3-kinase regulatory subunit 4 required for the synthesis of the lipid phosphatidylinositol 3-phosphate (PtdIns3P). VPS15 was first identified and characterized in yeast *Saccharomyces cerevisiae* after its isolation in different genetic screens performed to identify proteins required for vacuolar protein sorting (VPS) [9]. VPS15 binds to and regulates the class III PtdIns3P lipid kinase VPS34/PIK3C3 converting PtdIns to PtdIns3P. VPS15 in association with VPS34 is involved in two well-studied protein complexes conserved from yeast to mammals, the UVRAG/Beclin1 and Atg14L/Beclin1 complex required for membrane

trafficking and autophagy respectively [10, 11]. In metazoan *VPS15* is an essential gene since drosophila homozygous deletion mutant and whole body *Vps15*-deficient mice (*Vps15*$^{-/-}$) display early L3 larval stage or early-embryonic lethality a phenotype similar to the one observed for *Vps34*$^{-/-}$ mice [11, 12]. The VPS15 protein is composed of an N-terminal protein kinase, a HEAT repeat, a coiled coil and seven C-terminal WD40 domains.

In the present work, we describe for the first time patients with a ciliopathy phenotype carrying a mutation in *VPS15* as well as the link between VPS15 and cilia trafficking by using three model organisms (human, zebrafish and yeast). We show that the VPS15-R998Q patient mutation affects the primary cilium length in human fibroblasts, induces axial curvature and kidney cysts in zebrafish embryos and impairs the dominant-negative Golgi to vacuole trafficking phenotype of hVPS15 in yeast cells. We also show that in human fibroblasts a pool of VPS15 localizes to the Golgi and interacts with the golgin GM130 ensuring efficient loading of IFT20 in vesicles addressed from the cis-Golgi to the basal body of the cilium. This work suggests that VPS15 and GM130 function together at the Golgi to ensure the IFT20-dependent sorting and transport of proteins addressed to the ciliary membrane.

Results

A mutation in the *VPS15* gene in a family with a ciliopathy

Three siblings (patients II.1, II.3 and II.5 in Fig. 1A) born from Algerian parents (originating from the same village and with high probability of distant consanguinity) were referred for early onset retinal degeneration and progressive renal disease (CARGO, rare eye diseases center, Strasbourg University Hospital) (Fig. 1B). The proband (II.1) was reported with night-blindness as well as progressive visual impairment since the age of 5. An unrecordable electroretinogram (ERG) at age 11 confirmed the diagnosis of retinitis pigmentosa. Fundus examinations revealed progressive widespread retinal dystrophy associated with pigment migration (Fig. 1C). He also developed progressive renal failure with glomerular and tubular dysfunctions followed by progressive kidney atrophy reminiscent of nephronophthysis requiring dialysis at age 24. General clinical examination showed a small stature (-3SD) due to growth hormone deficiency, overweight (BMI 28 kg/m^2), prognathism (similar to his father suggesting an independent dominant trait), enlarged hands with brachydactyly and clinodactyly of the fifth digit, and mild learning difficulties (Fig. 1B). The younger sister (II.3) at age 6 presented growth delay (-2SD) and was diagnosed with retinitis pigmentosa. Besides her history of brachial plexus palsy at birth and dysmenorrhea at age 19, she presented the same physical features as her brother (prognathism, enlarged hands, bradymetacarpy and brachymesophalangy, and clinodactyly of the fifth finger). She was diagnosed with kidney manifestations at age 14 and had moderate renal dysfunction at the age 19. Abdominal magnetic

resonance imaging (MRI) revealed reduced size kidneys (9 cm axial length for the right and 8 cm for the left; normal axial length 12 cm) filled with multiple corticomedullary microcysts, and a larger cyst (3 cm) in the left kidney (Fig. 1D). The kidney biopsy showed a diffuse tubulo-interstitial nephropathy. The youngest sibling (II.5) had a history of congenital hypotonia, brachial plexus palsy at birth, language delay and mild learning difficulties (Fig. 1B). Fundus examination performed at age 4 revealed retinitis pigmentosa, confirmed by an unrecordable ERG. General clinical examination revealed the same gestalt as probands II.1 and II.3. Proteinuria appeared at age 4 and he is now developing a progressive nephropathy. Thus, 3 out of 5 siblings of this family exhibited typical ciliopathy features of retinitis pigmentosa, renal dysfunction and developmental anomalies prompting for molecular investigations to identify the causal mutation.

To identify the genetic mutation responsible for the disease, we screened a panel of 30 ciliopathy-related genes and we found no mutations. The genome wide homozygosity mapping combined to whole exome sequencing of the three affected siblings (II.1, II.3, II.5) and the healthy sister (II.4) revealed only 4 homozygous regions specific to the three affected individuals: 2 on chromosome 3, and 1 on chromosomes 6 and 20 (Supplementary Fig. 1). Sequencing data processing and variant calling (SNP and InDels) with the Genome Analysis Toolkit revealed 60,612 to 63,887 genetic variants per proband (Supplementary Table 1). Variant filtering using stringent criteria reduced the number of genetics variants to respectively 1085, 793, 1124 and 2009 variants per proband (Supplementary Table 1). To identify variants consistent with autosomal recessive disease inheritance, we kept only compound heterozygous or homozygous variants, reducing the number of variants to 2 compound heterozygous variants (in the *AK9* and *RECQL4*/*MFSD3* genes) and 1 homozygous variant (in the *PIK3R4* gene). Using manual curation based on the known biological, physiological or functional relevance to the disease of the candidate genes, we were left with a single homozygous variant in the *PIK3R4* gene, named herein *VPS15*: p.Arg998Gln, c.2993G>A, Chr3:(GRCh37):g.130422672C>T. The variant and its cosegregation with the phenotype in the family was confirmed by Sanger sequencing. Overall, deleterious or potentially pathogenic homozygous or compound heterozygous variants were not found in 382 patients with presumed ciliopathy and yet unknown molecular origin (our databases and the databases of three laboratories specialized in ciliopathies and other rare genetic diseases) and in 7360 patients screened for other rare diseases (Institut IMAGINE). The PolyPhen-2 (polymorphism phenotyping v2) software, a tool analysing impact of amino acid substitutions on protein structure and function [13], predicts that this R998Q substitution is “probably damaging” with a HumVar score 0.994. The residue Arg998 is in the first WD40 repeat of the VPS15 protein (Fig. 1E), a domain known to be important for protein-protein interactions. Indeed, the WD40 domain of the yeast Vps15 protein is sufficient to bind Atg14 [14], bridges the Vps38/Atg6 heterodimer to form the Vps15-Vps34 complex II [15] and in

human cells the NRBF2 protein interacts with the VPS15 WD40 domain [16].

Cilia are shorter in VPS15-R998Q patient fibroblasts

Some ciliopathies present with impaired ciliogenesis or shorter cilia, while others are characterized by longer cilia. To investigate whether the symptoms observed in patients were associated with a ciliary phenotype, skin fibroblasts from patients II.1, II.3 and II.5 and age-matched control fibroblasts were grown to confluence in medium containing 10% foetal calf serum (FCS), then washed and starved by serum deprivation for 24 hours (-FCS) to induce growth arrest and cilium formation. Primary cilia were labelled with an antibody directed against acetylated α-tubulin highlighting the axoneme (Fig. 2). Percentage of ciliated cells was similar in control (80%) and patient cells (73%). However, measuring cilium length in the age-matched healthy controls and in the patient fibroblasts showed that in cells from patients II.1, II.3 and II.5, cilia were about 50% shorter than in the corresponding control cells (Fig. 2). Fibroblasts from a patient carrying a homozygous *BBS4* deletion were used as positive control [17]. As expected the cilium length of *BBS4*$^{-/-}$ cells was reduced, however to a lesser extent compared to the VPS15-R998Q patient cells. This result shows that skin fibroblasts from patients carrying the VPS15-R998Q mutation present a ciliary phenotype characterized by shorter primary cilia. To confirm that this phenotype is due to the missense mutation in VPS15, skin fibroblasts from patients were transfected with a plasmid expressing 3xHA tagged VPS15 cDNA. The proper expression of VPS15-HA from this plasmid was checked by western blot after transfection of the retinal pigment epithelial cell line hTERT-RPE1 (Supplementary Fig. 2A), then by immunofluorescence in control fibroblasts grown in complete medium or serum deprived to confirm the Golgi localisation (Supplementary Fig. 2B).
Finally, the patient fibroblasts were either mock transfected or transfected with the VPS15-HA construct 6 hours before serum deprivation. 24 hours post serum deprivation, cells were fixed and stained as described above and cilium length determined. The VPS15-HA protein was able to rescue the short-cilium phenotype in all three patients confirming that the VPS15-R998Q mutation is responsible for this defect. Interestingly, while transfection with VPS15-HA and subsequent serum deprivation was deleterious to all the cells and resulted in about 70% mortality, the surviving cells appeared to be those with the lower expression of VPS15-HA, suggesting a negative effect of VPS15 overexpression.

VPS15-RQ associated with ciliopathy phenotypes in zebrafish

Zebrafish *Danio rerio* is a well-established model for human ciliopathies [18]. Antisense morpholino injection was used to elucidate a putative ciliary function of Vps15 in zebrafish by searching for ciliopathy-related phenotypes (curvature of body axis and kidney cysts). The zebrafish

zvps15/pik3r4 gene (ENSDARG0000000469) maps to LG16 and has two predicted splice variants encoding two putative proteins of 1340 and 1347 amino acids. Sequencing and BLAST analysis indicated that only the shorter variant was present (data not shown), encoding a protein with a high degree of sequence similarity to mammalian VPS15 (82% identity to hVPS15) with conservation of the R998 residue at position 976 of the zVps15 protein. Furthermore, the *zvps15* transcript was detected in zebrafish embryos from the 8-cell stage (1.25 hours post-fertilization (hpf)) to the protruding-mouth stage (72 hpf) (Supplementary Fig. 3A and 3B).

To reveal the spatial expression pattern of *zvps15*, we performed whole-mount RNA *in situ* hybridization at 48 hpf. Expression of *zvps15* was observed in the head as well as in the pronephric duct (Supplementary Fig. 3C). Morpholinos directed against the *zvps15* start codon (*vps15*-mo) and control morpholino (*vps15cont*-mo) with 5-base mismatch were synthesized. The *vps15*-mo inhibited specifically protein synthesis of a zVps15-GFP fusion protein (Supplementary Fig. 3D). At 56 hpf, 50% of the *vps15* morphants presented a severe curvature of the body axis (Fig. 3A and 3B). Co-injection of *vps15*-mo or *vps15cont*-mo with a morpholino directed against p53 mRNA (*p53*-mo) did not rescue the phenotype (Fig. 3B), showing that the phenotype was not due to morpholino off-targeting mediated through p53 activation [19]. Moreover, 90% of the *vps15* morphants showed hydrocephalus (Fig. 3A). We also observed the presence of kidney cysts as well as a cystic dilatation of the region slightly posterior to the ear at the level of the pectoral fin in 61% of *vps15* morphants (Fig. 3D, 3E and Supplementary Fig. 3E), a phenotype not due to activation of p53-mediated apoptosis (Fig. 3E). In most morphants, we could detect 2 bilateral pronephric cysts (Fig. 3D).

To assess specificity of the morpholino effects and test the patient VPS15-R998Q missense mutation, we determined whether the *vps15*-mo phenotypes associated with cilia dysfunctions (body axis curvature and kidney cysts) could be rescued by co-injection of a synthetic *zvps15* or *zvps15-R976Q* mutant mRNA (Fig. 3C and 3F). Co-injection of *zvps15* wild-type mRNA caused a statistically significant rescue of both phenotypes (body axis curvature and kidney structure), while co-injection of *zvps15-R976Q* mutant mRNA did not ameliorate the phenotypes displayed by the *vps15* morphants (Fig. 3C and 3F). These results show that the *zvps15-R976Q* mutant does not complement the ciliopathy phenotypes induced by *zvps15* depletion.

In yeast VPS15-R998Q is linked to trafficking defects

To better understand the origin of the ciliopathy linked to the hVPS15 mutation, we studied the heterologous expression of hVPS15 and hVPS15-R998Q mutant in yeast. Human and yeast VPS15 proteins share only 33% identity and the R998 residue is not conserved in yeast ScVps15 protein [15]. Although a specific amino acid is frequently not conserved in yeast, the structure and the function

of yeast and human proteins as a whole are closely related. Humanization of yeast cells by replacing the yeast gene by its human counterpart is a potent approach used to gain insights into the cellular role of the human protein [20, 21]. In humanized yeast cells, we analysed the VPS15-dependent cellular functions such as growth, intracellular trafficking from the Golgi to the vacuole (VPS pathway) and autophagy. Wild-type (WT) or *vps15***Δ** mutant yeast cells were transformed by plasmids bearing either the wild type hVPS15 or mutant hVPS15-R998Q cDNA, allowing either expression (CEN plasmid) or overexpression (2µ plasmid) of the human cDNA, as controlled by western-blot (Supplementary Fig. 3F). We observed growth delays upon overexpression of hVPS15 suggesting that hVPS15 might hijack important cellular functions (Supplementary Fig. 3H). Autophagy was investigated by monitoring the maturation of the aminopeptidase Ape1, a common read-out of yeast autophagy [22]. We observed that as previously described autophagy is blocked in the *vps15***Δ** yeast mutant cells compared to the wild-type cells, indeed the aminopeptidase (Api) is not matured (Fig. 3G). However, although autophagy was impaired in *vps15***Δ** cells (Fig. 3G), this defect was not rescued by expression of hVPS15 wild type or R998Q mutant (Fig. 3G). Transport to the vacuole along the VPS pathway was then analysed by using the CMAC fluorescent probe that normally only labels the lumen of the yeast vacuole (Supplementary Fig.3G). However, *vps15***Δ** cells show an additional CMAC-positive compartment (CPC) (white arrows, Supplementary Fig. 3G) that was not rescued by the expression of hVPS15-R998Q mutant as efficiently as by the wild type hVPS15 (Fig. 3H), indicating a default in transport to the vacuole. These yeast data suggest that the R998Q mutation might impair an *in vivo* function of hVPS15 in intracellular trafficking from the Golgi to the vacuole.

VPS15 and VPS15-R998Q bind to their partners and to GM130

Having shown that VPS15-R998Q leads to ciliary phenotypes, we next investigated whether observed defects were due to a loss of protein interaction. Indeed, VPS15 is known to be involved in two distinct cellular functions when interacting with different proteins, the UVRAG/Beclin1 complex involved in endosomal trafficking and the Atg14L/Beclin1 complex required for autophagy (Fig. 4A) [10, 11]. Endogenous VPS15 or VPS15-R998Q was immunoprecipitated from either control or patient fibroblasts in rich (+FCS) or serum-deprivation conditions (-FCS) to induce cilium formation and co-immunoprecipitation of VPS34 and Beclin1 proteins assessed by western blot (Supplementary Fig. 4A). The VPS15 protein was detected in the input of control and patient cells (in rich and starved conditions) indicating that the R998Q substitution does not hamper the stability of VPS15-R998Q. Moreover, VPS34 and Beclin1 proteins interacted with VPS15 and VPS15-R998Q proteins in both growth conditions (+ and -FCS), suggesting that the VPS15-VPS34 complex is functional and active to produce PtdIns3P. Indeed, PtdIns3P positive structures were

detected in the control and patient fibroblasts transfected with the 2XFYVE-GFP probe (Supplementary Fig. 4B) that is specific for PtdIns3P [23].

To further analyse the VPS15 protein complexes, VPS15 immunoprecipitation was done on control and patient cells in rich conditions (+FCS), proteins specifically retained on the VPS15-magnetic beads and not on naked beads (IPneg) were determined by mass spectrometry (Fig. 4B and Supplementary Fig. 4C). As previously shown, high amounts of VPS34 and UVRAG were found interacting with VPS15 (control) and VPS15-R998Q (patients II.1, II.3 and II.5), and so were Beclin1 and Atg14L, two known interactors of VPS15 (Fig. 4B and Supplementary Fig. 4C). We also found NRBF2 (Supplementary Fig.4C), a protein recently identified as belonging to the Atg14L-Beclin1 complex [24]. Protein abundance estimated from the number of mass spectrometry spectra (Supplementary Fig. 4C) shows that between the two controls and the three patients, the efficiency with which the different proteins (VPS34, UVRAG, Beclin1, NRBF2 and Atg14L) were immunoprecipitated with VPS15 are similar (Fig. 4C). This suggests that the R998Q substitution does not destabilize specifically one of these protein interactions. Interestingly, the cis-Golgi protein GM130 (GOLGA2) was also found among the strongest interactors of wild type and mutant VPS15 (Fig. 4C and Supplementary Fig. 4C). GM130 has not been previously reported as interacting with VPS15 (http://thebiogrid.org). However, this new interaction is highly relevant because the cis-Golgi is involved in transport to the cilium and GM130 is required for Golgi morphology and ciliogenesis via its interaction with the centrosomal protein AKAP450 [25]. To confirm our mass spectrometry data, we tested whether GM130 can be detected by western blot after immunoprecipitation of VPS15 from control and patient (II.5) cells grown in complete medium (+FCS) or deprived of serum (-FCS) (Fig. 5A). We also immunoprecipitated VPS34 from the same protein extracts to determine whether VPS34 was interacting with GM130 (Fig. 5A). As controls VPS15, VPS34, UVRAG and Atg14L were detected by western blot in the same samples (+FCS) (Fig. 4B). The results show that only VPS15 is specifically interacting with GM130, whereas in the same conditions VPS34 does not interact with GM130 but retains binding to VPS15, UVRAG and Atg14L (Fig. 4B and 5B). To further confirm this result we did the reverse immunoprecipitation experiment and detected VPS15 by western blot in control and patient fibroblasts after GM130 immunoprecipitation (Supplementary Fig. 5C).

To determine the intracellular localization of VPS15 in ciliary conditions, control, VPS15-R998Q patient II.5 and *BBS4*$^{-/-}$ patient fibroblasts were grown in medium without FCS, fixed and labelled for VPS15 (red) and acetylated α-tubulin (green) prior to observation (Supplementary Fig. 5A). The primary cilium is present after serum deprivation and there is no colocalisation between VPS15 and acetylated α-tubulin in patient and control cells (Supplementary Fig. 5A). Indeed VPS15 is rather concentrated around the nucleus, suggesting Golgi localization. The intracellular localization of

VPS15 was also analysed by staining VPS15 (red) and GM130 (green) (Fig. 5B). In confocal microscopy, this staining shows that the pool of VPS15 present at the Golgi colocalises with the golgin GM130 in control and patient cells. Since IFT20 required for ciliary assembly is anchored at the cis-Golgi by the golgin GMAP210 [5], immunofluorescence staining against VPS15 and GMAP210 was done on control (ctrl) and patient II.5 fibroblasts grown in complete medium (+FCS) and after serum deprivation (-FCS) (Fig. 5C). This staining confirms the localization of VPS15 at the Golgi in both rich and ciliary conditions. In contrast, and as already suggested by the interaction results (Fig. 4B and 4C), VPS34 did not colocalise with GM130 at the Golgi (Supplementary Fig. 5B). These results indicate that the PtdIns3P kinase VPS34 does not belong to the GM130-VPS15 cis-Golgi localized protein complex. We decided to further investigate the cilium-related role of VPS15 at the Golgi.

IFT20 release from the Golgi is reduced in VPS15-R998Q cells

Some ciliary cargoes are transported from the cis-Golgi to the base of the cilia in IFT20-dependent vesicles. The IFT20 protein, belonging to the IFT-B complex, is anchored at the cis-Golgi by the golgin GMAP210 [5, 6] which is redundant with GM130 for cis-Golgi cargo delivery [26]. Moreover, partial depletion of IFT20 induces a shorter length of the primary cilia [27], a phenotype similar to the one observed in VPS15-R998Q cells (Fig. 2). Therefore, we investigated the localization of IFT20 by immunofluorescence in control and patient fibroblasts upon cilium induction (-FCS) or not (+FCS). In cells grown in complete medium (+FCS), IFT20 (red) was detected in vesicles distributed throughout the cytoplasm and at the cis-Golgi where it colocalised with GM130 (green). In these conditions, IFT20-positive vesicles (red) were observed in patient and control cells (Fig. 6A and Supplementary Fig. 6). However upon serum deprivation, while the distribution of IFT20 was similar to the complete medium conditions in control cells (vesicles and cis-Golgi), significantly less non-Golgi (vesicular) IFT20 was observed in the patient fibroblasts (Fig. 6A). Indeed, in the latter, IFT20 localization seems restricted to the Golgi (Fig. 6A and Supplementary Fig. 6A and B). We confirmed this Golgi localisation in patient cells by performing an immunoprecipitation of IFT20 followed by a western blot against GMAP210. In both, control and patient fibroblasts, the interaction between IFT20 and its anchor GMAP210 [5] was observed (Supplementary Fig. 6C), confirming that the R998Q missense mutation does not disrupt the IFT20-GMAP210 interaction. Thus, the defect observed in patient fibroblasts seems to be localized at the level of the formation and/or release of IFT20 positive vesicles from the cis-Golgi. The difference between control and patient fibroblast is small but significative, and even a slight decrease in transport of cargo to the cilium may result in a decreased cilium growth and/or altered signalling.

Overall, our results support a new role for VPS15, independently of VPS34, in the

formation/development of the primary cilium, presumably by altering the IFT20-dependent Golgi to cilium vesicular transport as this pathway is affected in cells from patients bearing the R998Q mutation in *VPS15*.

Discussion

Here, we report a missense mutation in the *VPS15* gene identified in a unique family presenting with early onset retinal degeneration, late childhood kidney failure associated to mild skeletal developmental features with moderate intellectual disability. Overall, this clinical presentation, compatible with a ciliopathy, showed some overlap mainly with the well-known Senior-Loken and Bardet-Biedl syndromes [2]. More than a thousand proteins are probably required for the biogenesis and function of the vertebrate primary cilium and more than a hundred genes are now identified as mutated in ciliopathies (http://www.omim.org). Over the past ten years, classical gene identification optimized now by next generation sequencing strategies have allowed the identification of ciliary genes carrying mutations within different families such as for example BBS syndrome (mutations in *BBS1* or *BBS10* accounting each for 20% of the families) [28]. Nowadays, new ciliopathy causative genes are very likely to be identified in very few if not unique families. One explanation could be that the gene involved may code for an essential protein for which only a very limited number of mutation sites may be tolerated, the remaining mutational sites being probably lethal. Here, we report the first family with a missense mutation in the *VPS15/PIK3R4* gene. In metazoan (mouse and Drosophila) *VPS15* is an essential gene [11, 12]. The R998Q substitution is predicted to be probably damaging by the PolyPhen-2 program [13]. Indeed, arginine is a positively charged polar amino acid, frequently involved in salt-bridges and important for protein binding sites. Thus a change from an arginine to a glutamine that is polar and uncharged is certainly not neutral. The recently solved crystal structure of the yeast Vps15-Vps34 complex required for trafficking shows that the Vps15 kinase domain interacts with the Vps34 kinase domain to regulate its activity [15]. The WD40 repeat domain of Vps15 is engaged in interactions that bridge the Vp15/Vps34 heterodimer with the Atg6/Vps38 sub-complex [15]. Based on this structure, the kinase domain of Vps15 is probably inactive once bound to Vps34 because its ATP binding site is not accessible [15], however in the human VPS15-GM130 complex this kinase domain might be functional.

The link between VPS15 and cilia was not obvious since VPS15 was mostly shown to be involved in membrane trafficking and autophagy (Fig. 4). However it has long been hypothesized that VPS15 may have additional roles, because yeast results indicate that some VPS15 acts independently of the two UVRAG-Beclin1 and ATG14L-Beclin1 complexes [22]. Now with the clear ciliopathy phenotypes from human patients and zebrafish knockdown shown in this work, an additional function in ciliogenesis or cilia function is suggested. Functional assays in patient cells show that

the VPS15-R998Q mutation affects primary cilia length and IFT20-dependent trafficking from the cis-Golgi. The retinal degeneration could be due to a trafficking defect at the level of photoreceptor cells that are ciliated sensory cells. Indeed, phototransduction proteins such as rhodopsin are transported through the connecting cilium structure to the outer segment by vesicular trafficking involving the IFT machinery [29]. IFT20 is required for assembly and maintenance of the photoreceptor outer segment and likely participates both as component of IFT particle and as a cis-Golgi specific effector independently of canonical IFT machinery [30]. Based on retinal phenotypes in patients, we hypothesize that VPS15 is a novel actor in the photoreceptor transport machinery. Further investigations are necessary to define the role of VPS15 compared to other effectors involved in this trafficking. VPS15 could also be a novel regulator of the ciliary transport machinery in the kidney tubules, since a conditional knockout for IFT20 in mice shows rapid development of kidney cysts [31].

Here, we have identified a novel protein complex localized at the cis-Golgi and encompassing VPS15 and the golgin GM130. This new VPS15-GM130 complex could be involved in IFT20-dependent trafficking to the cilium (Fig. 6B). Indeed, a pool of IFT20 is associated with the cis-Golgi where it colocalizes partially with GM130 with increased IFT20 release from the Golgi in control fibroblast than in patient cells. Golgi-associated IFT20 is required for cilia assembly [6]. Interestingly, in mammalian cells strong IFT20 knockdown results in a lack of cilia assembly [6] and weaker IFT20 knockdown in shorter primary cilia [27] and impaired transport of cargo to the cilia [6]. Thus, even a moderate decrease in the amount of IFT20 involved in the cis-Golgi to cilium transport could have a tremendous effect on the primary cilium assembly and function.

Golgins are key Golgi effectors since they participate to the specificity of intracellular trafficking by capturing vesicles of different origins [32]. At the cis-Golgi level, the golgin GMAP210 serves as an anchor for IFT20[5] and knockdown of GMAP210 leads to a defect in cilia formation [5]. However, knockdown of GM130 has been described as not affecting cilia assembly [33]. Interestingly, recent data show that GMAP210 is redundant with GM130 for cargo delivery to the cis-Golgi [26]. Indeed, among the ten widely conserved golgins tested, only GM130 and GMAP210 were shown to be specific for tethering vesicles arriving at the cis-Golgi from the endoplasmic reticulum [32]. Thus more redundancy in the function of these two proteins might exist, especially at the level of primary cilium formation/function. Another argument playing in favour of a role of GM130 in ciliogenesis is its interaction with the centrosomal protein AKAP450, an interaction required for ciliogenesis and Golgi integrity [25]. Thus, GM130 role may be less direct than GMAP210 and could involve additional proteins. Here we show that the VPS15 kinase interacts with GM130 and that in VPS15-R998Q patient cells IFT20 is partially retained in the Golgi. Thus, the VPS15-GM130 protein complex could serve as platform to form IFT20 positive vesicles targeted from the cis-Golgi to the

cilium. Mass spectrometry analyses aimed at unravelling the molecular mechanism of VPS15 function in cilium formation indicate that indeed, although the VPS15-GM130 interaction is conserved between control and patient fibroblasts, a GM130-IFT20 interaction was detected in control fibroblasts but never in patient cells (by three independent experiments). Additional research along this line of work is needed to confirm this hypothesis.

Misbalanced trafficking to the cilia has dramatic consequences on cilia assembly and function, resulting in ciliopathies. Among the trafficking effectors involved, the BBSome, a complex formed by a subset of BBS proteins, is altered in a significant number of patients with Bardet-Biedl syndrome. The cilia trafficking pathway is still poorly understood, yet very important for proper signalling functions [4]. During evolution and to ensure intracellular trafficking, different coat complexes have assembled, among them the clathrin, COPI/COPII and BBSome coat that share a common overall structure [8]. The VPS15-VPS34 platform is crucial for trafficking and this assembly platform is well conserved during evolution. VPS15 could also be involved at the cis-Golgi to serve as a platform for ciliary-targeted membrane proteins. Indeed, GMAP210 and GM130 golgins are not specific for ciliary cargoes, thus other proteins are involved to control this specificity and VPS15 or an interaction partner could be good candidates (Fig. 6B).

Overall, our results support a new role for VPS15, independent of VPS34, in the IFT20-dependent Golgi to cilium vesicular transport as this pathway is affected in patients bearing the VPS15-R998Q mutation. However, VPS15 does not directly interact with IFT20 and its interaction with GM130 is not hampered by the presence of the point mutation either. Therefore subsequent work will be needed to analyse the interactions between IFT20 and the VP15-GM130 complex in order to confirm VPS15 dependent IFT20-GM130 interaction. Thus the potential involvement of partners of the VPS15-GM130 complex whose function might be hampered by the patient mutation in VPS15 will need further investigations.

Methods

Ethical Approval

After informed consent of the patient and his/her representative according to the French legislation, peripheral blood samples were obtained from the affected children and their parents as well as control individuals. DNA from all collected samples was extracted according to standard procedures. The objectives and the aim of the study were clearly explained to the patients and this study was approved by the local ethics committee at Hôpitaux Universitaires de Strasbourg (Strasbourg University Hospital).

Cell cultures

Fibroblasts of patients and age-matched healthy control individuals were obtained by skin biopsy as previously described [34]. Primary skin fibroblasts from VPS15-R998Q patients and age-matched healthy control skin fibroblasts as well as the hTERT-RPE1 cell line were grown in DMEM supplemented with 10% fetal calf serum (FCS) and 1% Penicillin-streptomycin-glutamin (PSG). To induce primary cilium formation, cells were deprived of serum by growth for 24 hrs in DMEM with 1% PSG but only 0.1% FCS (conditions -FCS).

250K Affymetrix Array and Sanger sequencing

To identify the genetic mutation responsible for the disease, we screened a panel of 30 ciliopathy-related genes by targeted exon-capture strategy coupled with multiplexing and high-throughput sequencing [35], and we found no mutations. Then, we performed a genome wide homozygosity mapping using GeneChip® Human 250K SNP Affymetrix (Supplementary Figure 1) combined to whole exome sequencing (Agilent SureSelect All Exon XT2 50 Mb kit) of the three affected siblings (II.1, II.3, II.5) and the healthy sister (II.4), using the facilities provided by IntegraGen (Evry, France). Sequencing data processing and variant calling (SNP and InDels) with the Genome Analysis Toolkit revealed 60,612 to 63,887 genetic variants per proband (Supplementary Table 1). Variant filtering was performed with the VaRank program [36] using stringent criteria excluding (i) non-pathogenic variants defined in dbSNP 138, (ii) variants represented with an allele frequency of more than 1% in dbSNP, the Exome Variant Server (EVS), the Thousand Genomes Project Catalog and the ExAC database, (iii) variants found in the homozygous state or more than once in the heterozygous state in 70 control exomes, (iv) variants into 5'UTR, 3'UTR, downstream, upstream or intron locations without local splice effect prediction, (v) synonymous variants without local splice effect prediction. This step reduced the number of genetics variants to respectively 1085, 793, 1124 and 2009 variants per proband (Supplementary Table 1). The variant and its cosegregation with the phenotype in the family was confirmed by Sanger sequencing. Sanger sequencing was performed by way of PCR amplification with 50 ng of genomic DNA template. The primers were designed with Primer 3 (http://frodo.wi.mit.edu/primer3) and are detailed in Supplementary Table 2. Bidirectional sequencing of the purified PCR products was performed by GATC Biotech. Segregation analysis ruled out the mutations in the RECQL4/MFSD3 gene as they were both carried in cis by a maternal chromosome. The very closely related intronic AK9 insertions of 6 or 4 nucleotides (NM_001145128.2:c.5315+104_5315+105insAGAGAG and NM_001145128.2:c.5315+106_5315+107insAGAG) in a repetition of an AG motif in intron 38 are present in the three patients and in the healthy sister and thus not specific to the disease.

Zebrafish experiments

Fish were bred and raised at 28.5°C as described previously [37]. The AB wild-type line (University of Oregon, Eugene) was used for all the experiments. Sexually mature fish were crossed in couples, and eggs were collected after being laid. For experiments, fertilized eggs were raised in 1× Instant Ocean salt solution (Aquarium Systems, Inc.) supplemented with 200 μM 1-phenyl 2-thiourea (PTU) to suppress melanogenesis.

The *pik3R4/vps15* full-length sequence was amplified with the pikwt-forw and pikwt-rev primers (Supplementary Table 2) and cloned into pCS2+GFP with EcoRI-XhoI. pGEM T-easy vector containing the 1100 pb fragment used for whole mount was cut with EcoRI, and the resulting fragment cloned into pCS2+GFP to give the zVps15-GFP plasmid. For *zvps15-RQ* mutant we amplified the wild-type sequence with pikmut-forw/pikwt-forw and pikmut-rev/pikwt-rev (Supplementary Table 2). The 2 obtained fragments were then amplified together with a mix of pikwtforw/pikwtrev and finally cloned into pCS2+GFP with EcoRI-XhoI.

RT-PCR was carried out following standard protocol. Total RNA was isolated from 24 to 72 hpf embryos using Tri-reagent (Invitrogen, Carlsbad, CA). For *zvps15* in situ hybridization, we used as probe a fragment of 1100 bp amplified by PCR with the following primers pik3r4WM-forw and pik3r4WM-rev and cloned into pGEM-T-easy vector (Promega). For RNA probe we used NcoI and SP6 RNA polymerase. RT-PCR was done with the same primer pairs.

Whole-mount in situ hybridization was performed as previously described [38]. To prevent pigmentation for expression analysis after 24 hpf, embryos were transferred to water containing 0.2 mM 1-phenyl-2-thiourea at 20 hpf and fixed at appropriate stages.

For injections, zebrafish eggs were collected shortly after being laid. Cleaned eggs were transferred to a petri dish with a minimal amount of water. Embryos were injected (FemtoJet; Eppendorf) through the chorion into the yolk at the one-cell stage with 12 nl of solution. Injection needles were pulled from borosilicate glass capillary tubes with filament (Warner Instruments) using a micropipette puller (Sutter Instrument Co). Morpholinos (Gene Tools, LLC) were injected at the following concentrations: *vps15*-mo: AGTTGGTTCCCCATCTCACTGGATC (0.4 mM); p53-mo: GCGCCATTGCTTTGCAAGA-ATTG (0.4 mM); *vps15cont*-mo: AGTaGcTT CCCgATCTCAgTcGATC (0.4 mM). All dilutions were made in distilled water. Phenol red was added to the samples before injection (0.1% final concentration). zVps15-GFP plasmid was injected at the final concentration of 40 ng/μl. *zvps15* wild type and *zvps15-R976Q* mRNA were injected the final concentration of 40 ng/μl.

Plasmids, strains, media and methods for yeast cells

The R998Q mutation was introduced into the hVPS15 cDNA by polymerase chain reaction (PCR) with Phusion High-Fidelity DNA polymerase (Thermo Scientific) on the pDONR223 entry vector

bearing hVPS15 cDNA (Addgene 23488). The resulting pDONR223-hVPS15-R998Q or pDONR223-hVPS15 were cloned by the Gateway® LR reaction (Invitrogen) into yeast destination vectors (Addgene plasmid numbers 14196 and 14252, [39]) to obtain the pSF194 (pAG423-promGPD-hVPS15-R998Q) and pSF196 (pAG413-promGPD-hVPS15-R998Q) or pSF198 (pAG423-promGPD-hVPS15) and pSF199 (pAG413-promGPD-hVPS15) plasmids. All plasmids sequences were verified by sequencing (GATC Biotech).

S. cerevisiae strains used in this study are BY4742 WT (*MATα leu2Δ0 ura3Δ0 his3Δ0 lys2Δ0*), *vps15Δ* (BY4742 *vps15*::*kanMX*) and *atg5Δ* (BY4742 *atg5*::*kanMX*). The indicated yeast strain were grown at 30°C to mid-exponential growth phase in rich medium (YPD): 1% yeast extract, 2% peptone, 2% glucose or in synthetic medium (SD): 0.67% yeast nitrogen base (YNB) without amino acids, 2% glucose and the appropriate dropout mix. Autophagy was induced by incubation for 4 hrs in SD-N medium: 0.17% (YNB) without ammonium sulfate, 2% glucose. Yeast cells were transformed using the modified lithium acetate method.

Living cells expressing hVPS15 or hVPS15-R998Q were harvested at an OD_{600nm} 0.5-1 and resuspended in synthetic complete yeast medium before visualization. For CMAC (Invitrogen) staining, the indicated yeast strain was harvested by a 500xg centrifugation for 1min, resuspended in SD medium and stained with CMAC (33 µM) for 10 min at 30°C prior two washing with PBS. Observation was performed with 100X/1.45 oil objective (Zeiss) on a fluorescence Axio Observer D1 microscope (Zeiss) using DAPI filter and DIC optics. Images were captured with a CoolSnap HQ2 photometrix camera (Roper Scientific) and treated by ImageJ (Rasband W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/). For western-blot analysis, total yeast extracts were obtained by NaOH lysis followed by TCA precipitation as previously described [40]. The equivalent of 1.5 OD_{600nm} unit of yeast cells were resuspended in 50 µl of 2X Laemmli buffer plus Tris Base. Samples were incubated 5 min at 37°C and analyzed by 10% SDS-PAGE followed by immunoblotting with anti-hVPS15 (Novus Bio, NBP1-30463) or anti-Ape1 (also termed Api, kind gift from Daniel Klionsky) using standard procedures. Images were acquired with the ChemiDoc Touch Imaging System (Bio-Rad).

Immunofluorescence

Primary fibroblasts from patients and control individuals were grown in Nunc™ Lab-Tek™ chamber slides (Thermo Scientific) or on spot slides, deprived of FCS for 24 hrs (or not) and fixed with 4% paraformaldehyde (PFA). After incubation with 0.5% Triton-X100 for 10 min, and blocking in PBS-20% FCS, cells were incubated for 1 hr with primary antibodies, washed 3 times in PBS, incubated for 1 hr with secondary antibodies and DAPI, washed again in PBS and mounted in Elvanol No-Fade™ mounting medium prior observation. Primary cilia were labeled with an

antibody directed against acetylated α-tubulin highlighting the axoneme [41]. Primary antibodies against acetylated α-tubulin (Abcam, ab24610), VPS15 (Novus Bio, NBP1-30463), VPS34 (Cell signaling, 4263), GM130 (Abcam, ab169276), GMAP210 (Thermo scientific, MA1-23294), IFT20 (Proteintech, 13615-1-AP) and HA (Roche, 1867423) were used. Secondary antibodies were goat anti-mouse Alexa Fluor coupled (either 488 or 568) IgG (Invitrogen), goat anti-rat Alexa Fluor 488 coupled IgG (Abcam, ab150157), donkey anti-rabbit IgG-FITC (Santa Cruz sc-2090) and donkey anti-mouse IgG (H+L) FITC conjugate (Thermo Scientific A16012). Cells were observed on a fluorescence (Zeiss Axio Observer D1) or confocal (Zeiss LSM700, 40X objective, Plateforme Microscopie et Imagerie (IBMP, Strasbourg, France)) microscope and images processed with ImageJ. The cilia length from the base to the tip was measured using the ImageJ program on confocal images, and the length of each primary cilia was determined for 120 to 200 cells per sample (ctrl1, ctrl2, ctrl3, II.1, II.3, II.5 and BBS4$^{-/-}$).

Transient transfection for GFP-2XFYVE^{HRS} and VPS15-HA plasmids

The fibroblasts cells were cultured to be 60% confluent the day of transfection in Nunc™ Lab-Tek™ chamber slides (Thermo Scientific). 500 ng of plasmid was transfected using the DNA transfection reagent (jetPEI™, Polyplus transfection) according to the protocol for adherent cells. After 24 hrs, fibroblast cells were washed three times in PBS, then the cells were fixed for 30 min at room temperature with 4% PFA, rinsed three times in PBS for 5 min each, stained with DAPI and confocal fluorescence microscopy was done. Confocal microscopy was performed on a Zeiss LSM700 microscope. The GFP-2XFYVE^{HRS} plasmid was a kind gift from Harald Stenmark. The VPS15-HA plasmid was obtained by cloning the pDONR223-hVPS15 described above into the human destination vector pCSf107mT-GATEWAY-3'-3HA (Addgene, 67616) by the Gateway® LR reaction (Invitrogen). Plasmid was verified by sequencing.

Coimmunoprecipitation and mass-spectrometry analysis

Cells were grown in DMEM (Catalog no.: 31885; Gibco Invitrogen), 10% FBS and Penicillin and Streptomycin (P/S) to full confluence. To induce primary cilia, cells were grown to confluence in DMEM containing 10% FCS then washed with PBS and starved by serum deprivation for 24 hrs with DMEM 0.1% FCS. Prior lysis, cells were rinsed with PBS at 4°C, resuspended in non-denaturing lysis buffer (20mM Tris HCL pH 8; 137 mM NaCl; 1% Nonidet P-40; 2mM EDTA) with a protease inhibitor cocktail (Roche 06538282001) and incubated on ice for 15 min under gentle shaking. The samples were then centrifuged at 12,000X g for 20 min at 4°C, protein concentration was measured using Qubit® Protein Assay Kit (Life Technologies). 500 µg of cell lysates were incubated with VPS15 antibodies (Novus Biologicals NBP1-30463) on a rocker shaker

overnight at 4°C. The immunocomplexes were captured by protein G sepharose beads (Dutcher 17-0618-05) for 2 hrs at 4°C under gentle shaking. Sepharose G beads were washed 6X5min with non-denaturing lysis buffer with a protease inhibitor cocktail, resuspended in 2X Laemmli buffer and boiled 10 min at 95°C to dissociate the immunocomplexes from the beads. For Western blot, the same primary antibodies as for immunofluorescence were used, and anti-GMAP210 (Thermo Scientific, #MA1-23294), anti-Beclin1 (Cell signaling, #3495), anti-Atg14L (Cell Signaling, #5504), anti-UVRAG (Abcam, #ab174550) and anti-GAPDH (Abcam, #ab181602) antibodies were also used. For the VPS15-HA control experiment, immunoprecipitation was done with the anti-HA antibody used for immunofluorescence and western blot detection with a mouse anti-HA antibody (Abcam,ab130275). Uncropped images of the western blots are shown in Supplementary Fig.7 and Supplementary Fig. 8.

For mass-spectrometry analyses, endogenous VPS15 immunoprecipitation was carried out with μMACS Protein A/G microbeads (Miltenyi Biotec) and VPS15 antibody (Novus Biologicals NBP1-30463), according to the manufacturer's protocol. Each protein sample was split in half. The second halves were used as negative controls, omitting antibodies during the immunoprecipitations (IPneg). Proteins complexes were eluted out of the magnetic stand with the SDS gel-loading buffer from the kit. Co-IP experiments were carried out in replicates for all samples (two healthy controls and the three patients). Samples were prepared for mass-spectrometry analyses as previously described [42]. Briefly, eluted proteins were precipitated with 0.1 M ammonium acetate in 100% methanol. After a reduction-alkylation step (Dithiothreitol 5 mM - Iodoacetamide 10 mM), proteins were digested overnight with 1/25 (W/W) of modified sequencing-grade trypsin (Promega, Madison, WI) in 50 mM ammonium bicarbonate. Resulting peptides were vacuum-dried in a SpeedVac concentrator and re-suspended in water containing 0.1% FA (solvent A) before being injected on nanoLC-MS/MS (NanoLC-2DPlus system with nanoFlex ChiP module; Eksigent, ABSciex, Concord, Ontario, Canada, coupled to a TripleTOF 5600 mass spectrometer (ABSciex)). Peptides were eluted from the C-18 analytical column (75 μm ID x 15 cm ChromXP; Eksigent) with a 5%-40% gradient of acetonitrile (solvent B) for 90 min.

Data were searched against the complete Human proteome set from the SwissProt database (released 2013/01/09; 43.964 sequences). Peptides were identified with Mascot algorithm (version 2.2, Matrix Science, London, UK) through the ProteinScape 3.1 package (Bruker). They were validated with a minimum score of 30, a p-value<0.05 and a decoy database strategy was employed to validate Mascot identifications at FDR < 1%. For this study, selected protein partners were considered according to the following rule: presence in the five coIP, absence from all the negative controls. They are sorted by decreasing average number of spectra.

Antibodies

Detailed information on antibodies used in this study as well as their dilution are indicated in Supplementary Table 3.

Data availability

The data and sequences that support the findings of this study are available from the corresponding authors upon request.

The authors declare no competing financial interests.

Author contributions

H.D, S.F, J.O.D.C. and S.B. provided direction for the project, conceived and designed the experiments; C.S., S.B. and J.O.D.C. performed cell biology experiments (human and yeast cells) and data analyses; J.R.R. constructed the VPS15-3xHA expression plasmid; S.S. and H.D. gathered data from patients and performed clinical investigations; C.E. and U.S. designed and performed the zebrafish experiments and data analyses; J.C. and P.H. designed and performed the mass-spectrometry experiments and data analyses; I.P., V.G. and K.C. gathered sequencing data and performed analyses; H.D, S.F, J.O.D.C. and S.B. analysed the data and wrote the paper. S.S., C.S., U.S., P.H. and C.E. contributed to manuscript writing.

Acknowledgments

We wish to warmly thank the members of the family for their participation. We would like to thank JL Mandel and the financial support of the Centre Régional de Génétique Médicale de Strasbourg and the Caisse d'Assurance Retraite et de la Santé au Travail Alsace-Moselle. We also thank RETINA France (100 Exomes Program) and FORMICOEUR for their constant and strong support. We wish to thank the teams of Pr N Katsanis and Dr E Davies (Duke University, USA); Pr F Hildenbrandt (Howard Huges Medical Institute and Harvard Medical School, USA) and Dr JM Rozet (Fondation Imagine, France) for sharing data from their ongoing sequencing programs. We thank D Klionsky (University of Michigan, USA), H Riezman (University of Geneva, Switzerland) and H Stenmark (Oslo University Hospital, Norway) for sharing strains and antibodies. We thank B Rinaldi, D Stuber and L Kuhn (CNRS/UDS, Strasbourg, France) for technical help and R. Poncelet-Bach for schematic representation of cilia trafficking. This work was funded by INSERM (to HD and SB), CNRS (to SF), Université de Strasbourg (HD, SF and IDEX 2015 Attractivité to SB), Agence Nationale de la Recherche (ANR-13-BSV2-0004 to SF) and AFM-Téléthon (AFM-SB/CP/2013-0133/16551 to SF). The mass spectrometry instrumentation was granted from

Investissement d'Avenir program (NetRNA ANR-10-LABX-36). The zebrafish work (C.E. and U.S.) was funded by Deutsche Forschungsgemeinschaft, (DFG Str439/5-1 and DFG RO2173/5-1) and Bundesministerium fur Bildung und Forschung (01ZX1407A).

References

1. Gerdes JM, Davis EE, Katsanis N. The vertebrate primary cilium in development, homeostasis, and disease. *Cell* **137**, 32-45 (2009).

2. Tobin JL, Beales PL. The nonmotile ciliopathies. *Genet Med* **11**, 386-402 (2009).

3. Oh EC, Katsanis N. Cilia in vertebrate development and disease. *Development* **139**, 443-448 (2012).

4. Sung CH, Leroux MR. The roles of evolutionarily conserved functional modules in cilia-related trafficking. *Nat Cell Biol* **15**, 1387-1397 (2013).

5. Follit JA*, et al.* The Golgin GMAP210/TRIP11 anchors IFT20 to the Golgi complex. *PLoS Genet* **4**, e1000315 (2008).

6. Follit JA, Tuft RA, Fogarty KE, Pazour GJ. The intraflagellar transport protein IFT20 is associated with the Golgi complex and is required for cilia assembly. *Mol Biol Cell* **17**, 3781-3792 (2006).

7. Nachury MV*, et al.* A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. *Cell* **129**, 1201-1213 (2007).

8. Jin H*, et al.* The conserved Bardet-Biedl syndrome proteins assemble a coat that traffics membrane proteins to cilia. *Cell* **141**, 1208-1219 (2010).

9. Herman PK, Stack JH, DeModena JA, Emr SD. A novel protein kinase homolog essential for protein sorting to the yeast lysosome-like vacuole. *Cell* **64**, 425-437 (1991).

10. Volinia S*, et al.* A human phosphatidylinositol 3-kinase complex related to the yeast Vps34p-Vps15p protein sorting system. *The EMBO journal* **14**, 3339-3348 (1995).

11. Lindmo K*, et al.* The PI 3-kinase regulator Vps15 is required for autophagic clearance of protein aggregates. *Autophagy* **4**, 500-506 (2008).

12. Nemazanyy I*, et al.* Defects of Vps15 in skeletal muscles lead to autophagic vacuolar myopathy and lysosomal disease. *EMBO molecular medicine* **5**, 870-890 (2013).

13. Adzhubei IA*, et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).

14. Heenan EJ, Vanhooke JL, Temple BR, Betts L, Sondek JE, Dohlman HG. Structure and function of Vps15 in the endosomal G protein signaling pathway. *Biochemistry* **48**, 6390-6401 (2009).

15. Rostislavleva K*, et al.* Structure and flexibility of the endosomal Vps34 complex reveals the basis of its function on membranes. *Science* **350**, aac7365 (2015).

16. Cao Y, Wang Y, Abi Saab WF, Yang F, Pessin JE, Backer JM. NRBF2 regulates macroautophagy as a component of Vps34 Complex I. *The Biochemical journal* **461**, 315-322 (2014).

17. Mykytyn K*, et al.* Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. *Nat Genet* **28**, 188-191 (2001).

18. Sun Z, Amsterdam A, Pazour GJ, Cole DG, Miller MS, Hopkins N. A genetic screen in zebrafish identifies cilia genes as a principal cause of cystic kidney. *Development* **131**, 4085-4093 (2004).

19. Robu ME*, et al.* p53 activation by knockdown technologies. *PLoS Genet* **3**, e78 (2007).

20. Amoasii L*, et al.* Phosphatase-dead myotubularin ameliorates X-linked centronuclear myopathy phenotypes in mice. *PLoS Genet* **8**, e1002965 (2012).

21. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921-925 (2015).

22. Kihara A, Noda T, Ishihara N, Ohsumi Y. Two distinct Vps34 phosphatidylinositol 3-kinase complexes function in autophagy and carboxypeptidase Y sorting in Saccharomyces cerevisiae. *J Cell Biol* **152**, 519-530 (2001).

23. Gaullier JM, Simonsen A, D'Arrigo A, Bremnes B, Stenmark H, Aasland R. FYVE fingers bind PtdIns(3)P. *Nature* **394**, 432-433 (1998).

24. Lu J*, et al.* NRBF2 regulates autophagy and prevents liver injury by modulating Atg14L-linked phosphatidylinositol-3 kinase III activity. *Nature communications* **5**, 3920 (2014).

25. Hurtado L, Caballero C, Gavilan MP, Cardenas J, Bornens M, Rios RM. Disconnecting the Golgi ribbon from the centrosome prevents directional cell migration and ciliogenesis. *J Cell Biol* **193**, 917-933 (2011).

26. Roboti P, Sato K, Lowe M. The golgin GMAP-210 is required for efficient membrane trafficking in the early secretory pathway. *J Cell Sci* **128**, 1595-1606 (2015).

27. Kim S*, et al.* Nde1-mediated inhibition of ciliogenesis affects cell cycle re-entry. *Nat Cell Biol* **13**, 351-360 (2011).

28. Novarino G, Akizu N, Gleeson JG. Modeling human disease in humans: the ciliopathies. *Cell* **147**, 70-79 (2011).

29. Sung CH, Chuang JZ. The cell biology of vision. *J Cell Biol* **190**, 953-963 (2010).

30. Keady BT, Le YZ, Pazour GJ. IFT20 is required for opsin trafficking and photoreceptor outer segment development. *Mol Biol Cell* **22**, 921-930 (2011).

31. Jonassen JA, San Agustin J, Follit JA, Pazour GJ. Deletion of IFT20 in the mouse kidney causes misorientation of the mitotic spindle and cystic kidney disease. *J Cell Biol* **183**, 377-384 (2008).

32. Wong M, Munro S. Membrane trafficking. The specificity of vesicle traffic to the Golgi is encoded in the golgin coiled-coil proteins. *Science* **346**, 1256898 (2014).

33. Asante D*, et al.* A role for the Golgi matrix protein giantin in ciliogenesis through control of the localization of dynein-2. *J Cell Sci* **126**, 5189-5197 (2013).

34. Scheidecker S*, et al.* Exome sequencing of Bardet-Biedl syndrome patient identifies a null mutation in the BBSome subunit BBIP1 (BBS18). *J Med Genet* **51**, 132-136 (2014).

35. Redin C*, et al.* Targeted high-throughput sequencing for diagnosis of genetically heterogeneous diseases: efficient mutation detection in Bardet-Biedl and Alstrom syndromes. *J Med Genet* **49**, 502-512 (2012).

36. Geoffroy V*, et al.* VaRank: a simple and powerful tool for ranking genetic variants. *PeerJ* **3**, e796

(2015).

37. Westerfield M. *The Zebrafish Book; A Guide for the Laboratory Use of Zebrafish (Brachydanio rerio)*. University of Oregon Press, Eugene,2nd edition (1993).

38. Oxtoby E, Jowett T. Cloning of the zebrafish krox-20 gene (krx-20) and its expression during hindbrain development. *Nucleic Acids Res* **21**, 1087-1095 (1993).

39. Alberti S, Gitler AD, Lindquist S. A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast* **24**, 913-919 (2007).

40. Volland C, Urban-Grimal D, Geraud G, Haguenauer-Tsapis R. Endocytosis and degradation of the yeast uracil permease under adverse conditions. *J Biol Chem* **269**, 9833-9841 (1994).

41. Piperno G, LeDizet M, Chang XJ. Microtubules containing acetylated alpha-tubulin in mammalian cells in culture. *J Cell Biol* **104**, 289-302 (1987).

42. Chicher J*, et al.* Purification of mRNA-programmed translation initiation complexes suitable for mass spectrometry analysis. *Proteomics* **15**, 2417-2425 (2015).

Figure legends

Figure 1: Pedigree and clinical presentation of the family affected by VPS15-R998Q mutation. **A.** Pedigree of the family with affected members in black and results of Sanger sequencing confirming co-segregation of the mutation with the disease. **B.** Main clinical features of II.1, II.3 and II.5 affected individuals. **C.** Fundus photographs (respectively right eye, left eye) of II.3 and II.5 individuals showing on the left sides a central view of the retina with the posterior pole of each right eye of the patients with mild narrowing of the retinal vasculature and moderate visible changes at the macula (best seen for II.3). On the right sides peripheral views show high level of pigment epithelium heterogeneity/atrophy with pigment migrations typical for retinitis pigmentosa (arrows). **D.** Abdominal MRI (SPAIR fat-suppression technique, Achieva PHILIPS 3 T) of II.3 and of a normal individual (Ctrl) for comparison (long and short axis are drawn on the normal kidney view). Coronal view of the kidneys showing the very reduced length of patient II.3 kidneys, the right and left kidneys being 9.1 cm and 8.2 cm on long axis and 4.3 cm and 4.1 cm on short axis, compared to the mean 11 cm (long axis) and 5 cm (short axis) length for the normal kidneys. The MRI shows also corticomedullary microcysts as well as larger cysts (arrows), the one in the left kidney with 3 cm in diameter. **E**. Shematic representation of the gene, mRNA and VPS15 protein with the different protein domains and the location of the mutation indicated in red.

Figure 2: Patient fibroblasts have shorter primary cilia than control cells, a feature which can be rescued by expression of wildtype VPS15. Patient (II.1, II.3, II.5 and *BBS4*$^{-/-}$) and control (ctrl1, ctrl2, ctrl3) skin fibroblasts were transfected or not with VPS15-HA, serum deprived for 24

hrs, fixed and stained with DAPI (cyan) and acetylated α-tubulin antibodies (red). Length of cilia was measured using the ImageJ analysis program and mean of 20-200 measurements determined for each experiment. Data shown are the mean of 3 independent experiments and error bars represent s.d.. The control Ctrl is the mean of ctrl1, 2 and 3. Statistical significance was determined using the Student t-test, *: $p<0.05$, **: $p<0.005$. Significance is determined relative to ctrl for non transfected cells and relative to mock transfected cells for VPS15-HA transfected cells. n.d. : not done, n.s. : non significant.

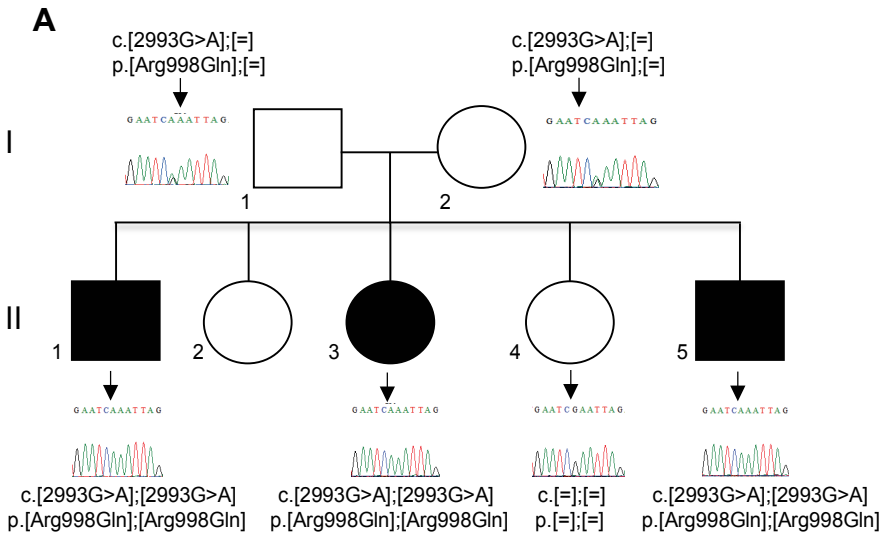
Figure 3: Zebrafish and yeast models to determine the VPS15-R998Q defects. A. *vps15* morphants injected with vps*15-mo* (b-d) show different degrees of body axis curvature and hydrocephaly (black arrows und zoom) compared to the control (*vps15cont*-mo)(a). For quantification, only phenotypes c and d were considered curved. 56 hpf embryos were used. **B.** Percentage of curved embryos among the population injected with different combinations of morpholinos (Morpho). **C.** Percentage of curved embryos upon co-injection with vps*15-mo* plus zVPS15 wild-type (VPS15) or zVPS15-RQ mRNA. In B and C n indicates the number of injected embryos that were counted. Statistical significance was determined using the Student t-test, n.s.: non significant, *: $p<0.05$, **: $p<0.01$ **D.** Pronephric cysts (white stars) were observed in *vps15* morphants. Black line: pronephric tubule; grey dot: glomerulus; grey line: pronephric duct; pf: pectoral fin. **E.** Percentage of embryos forming kidney cysts upon injection with different combinations of morpholinos. **F.** *vps15* morphants co-injected with VPS15 wild-type or VPS15-RQ mRNA. In E and F, n indicates the number of injected embryos that were counted. Statistical significance was determined using the Student t-test, n.s.: non significant, **: $p<0.01$ **G.** The wild-type (WT), *atg5Δ* and *vps15Δ* yeast cells bearing ctrl, hVPS15 or hVPS15-RQ plasmid, were grown under nitrogen deprivation for 4 hrs to induce autophagy and then collected. Western blot was performed to show the immature (proApi) and mature (mApi) forms of the vacuolar protease Api (Aminopeptidase 1). **H.** *vps15Δ* yeast cells were transformed with the empty (ctrl), the wild type hVPS15 or the mutant hVPS15-R998Q plasmid and the percentage of cells displaying an additional CMAC positive compartment (CPC) was determined. At least 100 cells per experiment and transformant were counted. Graph shows mean of three experiments. Statistical significance was determined using the Sutdent t-test, n.s.: non significant, *: $p<0.05$, **:$p<0.01$.
All the data shown in the figure are from at least 3 independent experiments and error bars represent s.d..

Figure 4: Identification of VPS15 protein complexes. A. The VPS15 protein complexes are similar in their organization and function in yeast and mammals. **B.** Immunoprecipitations with

VPS15 or VPS34 antibodies were done on cell lysates from control (ctrl) or patient (II.5) fibroblasts and known VPS15 and VPS34 interaction partners (VPS34, VPS15, UVRAG and ATG14) were detected by western blot. **C.** Mass spectrometry data of VPS15 interaction partners were analyzed, using the number of spectra found for VPS15 as a reference (100%). The number of spectra found for VPS34, GM130, UVRAG, Beclin1, Nrbf2 and Atg14L in the different samples (ctrl 1 and 2 and patient II.1, II.3 and II.5) was expressed relatively to the number of spectra found for VPS15 and percentages plotted on a histogram.

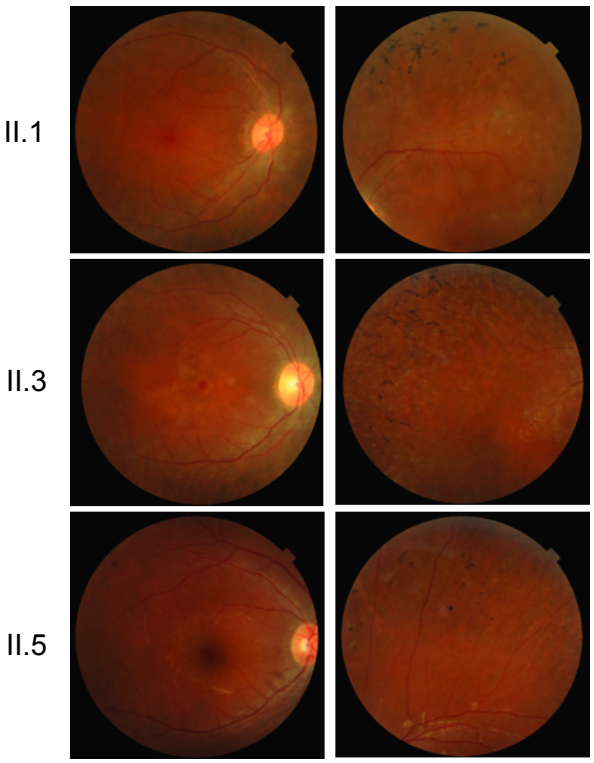
Figure 5: GM130 interacts with VPS15 but not VPS34. **A.** Control and patient fibroblasts were grown in complete medium (+FCS) or deprived of serum (-FCS) for 24 hrs. Immunoprecipitations with VPS15 or VPS34 antibodies were done on cell lysates and GM130 was detected by western blot. GAPDH was used as a loading control. **B.** Control (ctrl) and patient (II.5) fibroblasts were grown and deprived of serum for 24 hrs. Immunofluorescence against VPS15 (red) and GM130 (green) and DAPI staining (blue) was performed and cells were observed with a confocal microscope. **C.** Control (ctrl) and patient (II.5) fibroblasts were grown in complete medium (+FCS) or deprived of serum (-FCS) for 24 hours. Immunofluorescence against VPS15 (green) and GMAP210 (red) and DAPI staining (blue) was performed and cells were observed with a confocal microscope.

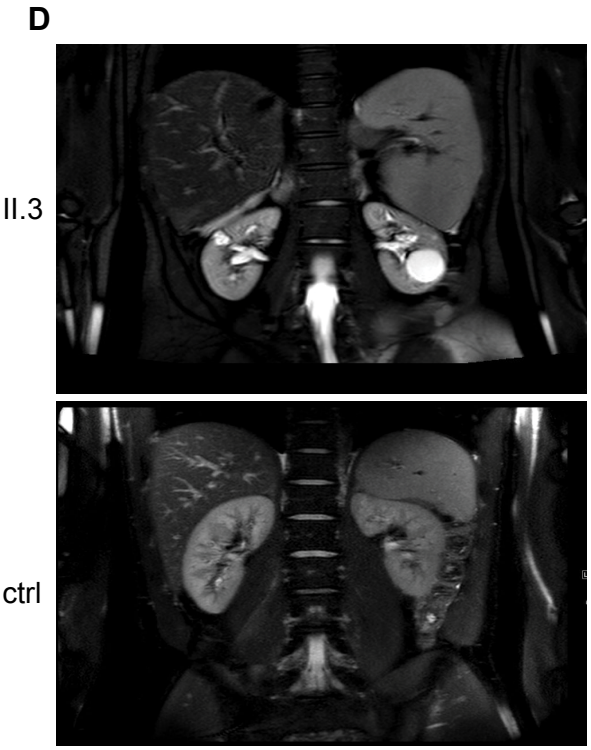
Figure 6: IFT20 is retained in the Golgi in patient fibroblasts. A. Control (ctrl) and patient (II.5) fibroblasts were grown in complete medium (+FCS) or deprived of serum for 24 hrs (-FCS) and fixed. Immunofluorescence against IFT20 (red) and GM130 (green) and DAPI staining (blue) was performed and the cells were observed on a confocal microscope, only the merge (IFT20, GM130 and DAPI) is shown. Amount of red (IFT20) fluorescence in the cytoplasm was determined using ImageJ to measure total amount of IFT20 fluorescence and substract fluorescence at the Golgi (green, GM130). Measures were done on 10 non serum deprived cells and 26 and 28 cells for ctrl and patient cells in serum deprived conditions respectively and error bars represent s.d.. Mean fluorescence was calculated and statistical significance determined using a Student t-test, **: $p<0.001$. **B.** Schematic representation of intracellular trafficking pathways to deliver cargo proteins to the ciliary base. The role of the VPS15 protein at the cis-Golgi is based on our data.

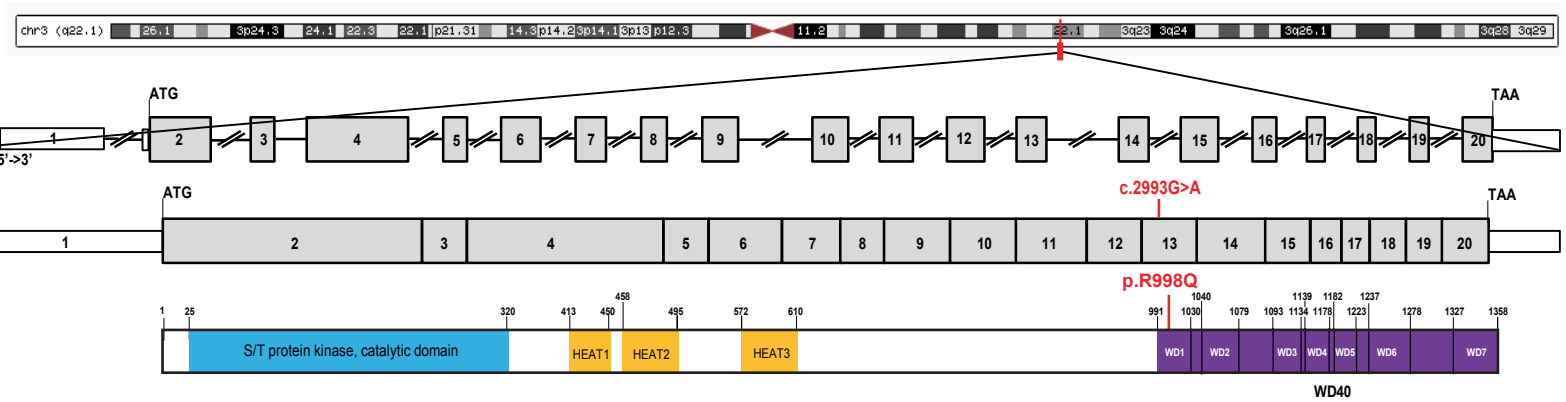


B

	II.1	II.3	II.5
Retinitis pigmentosa	+	+	+
Renal failure	+	+	+
Learning disability	+	+	+
Facial dysmorphism	+	+	+
Hand anomalies	+	+	+







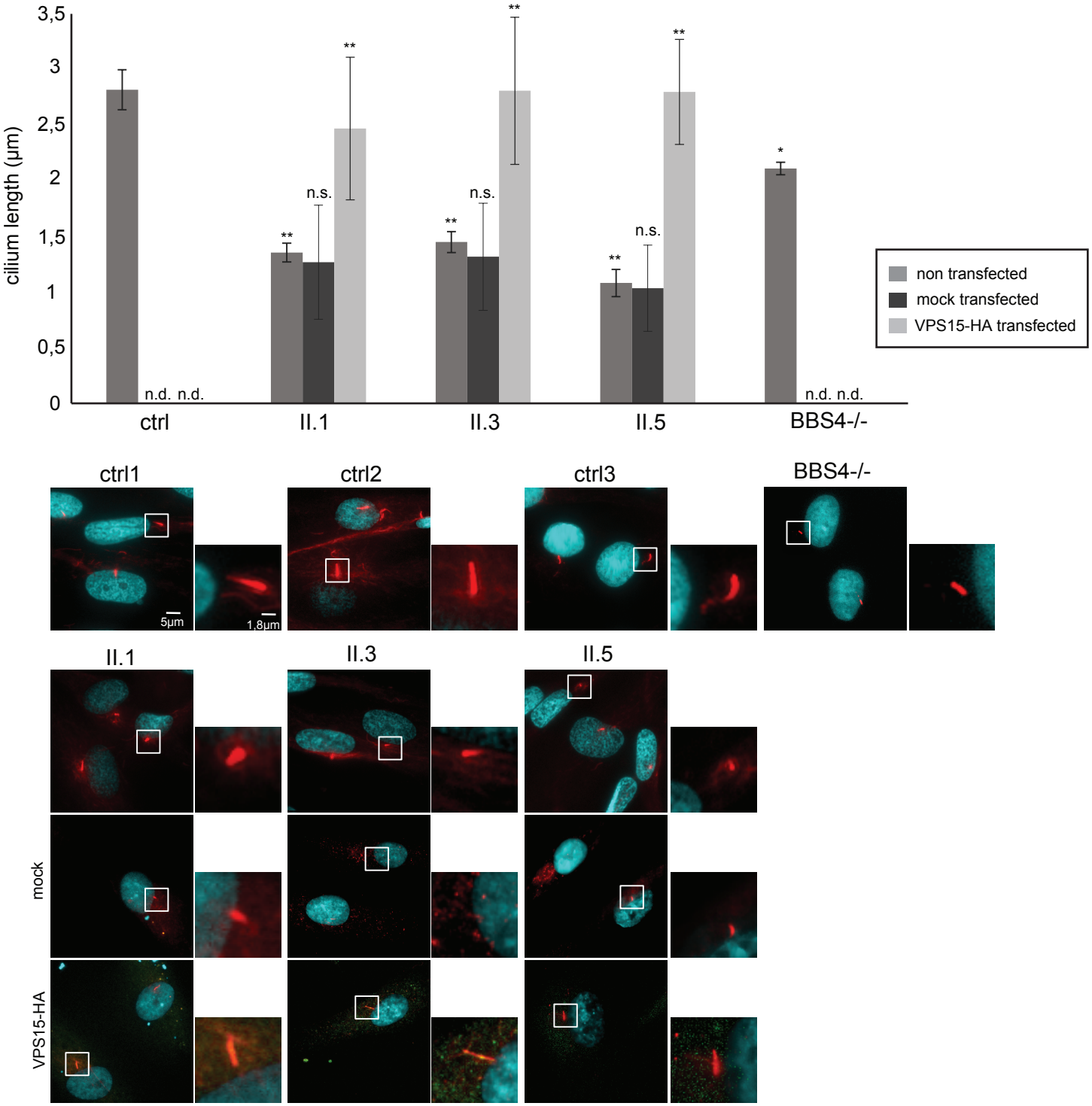
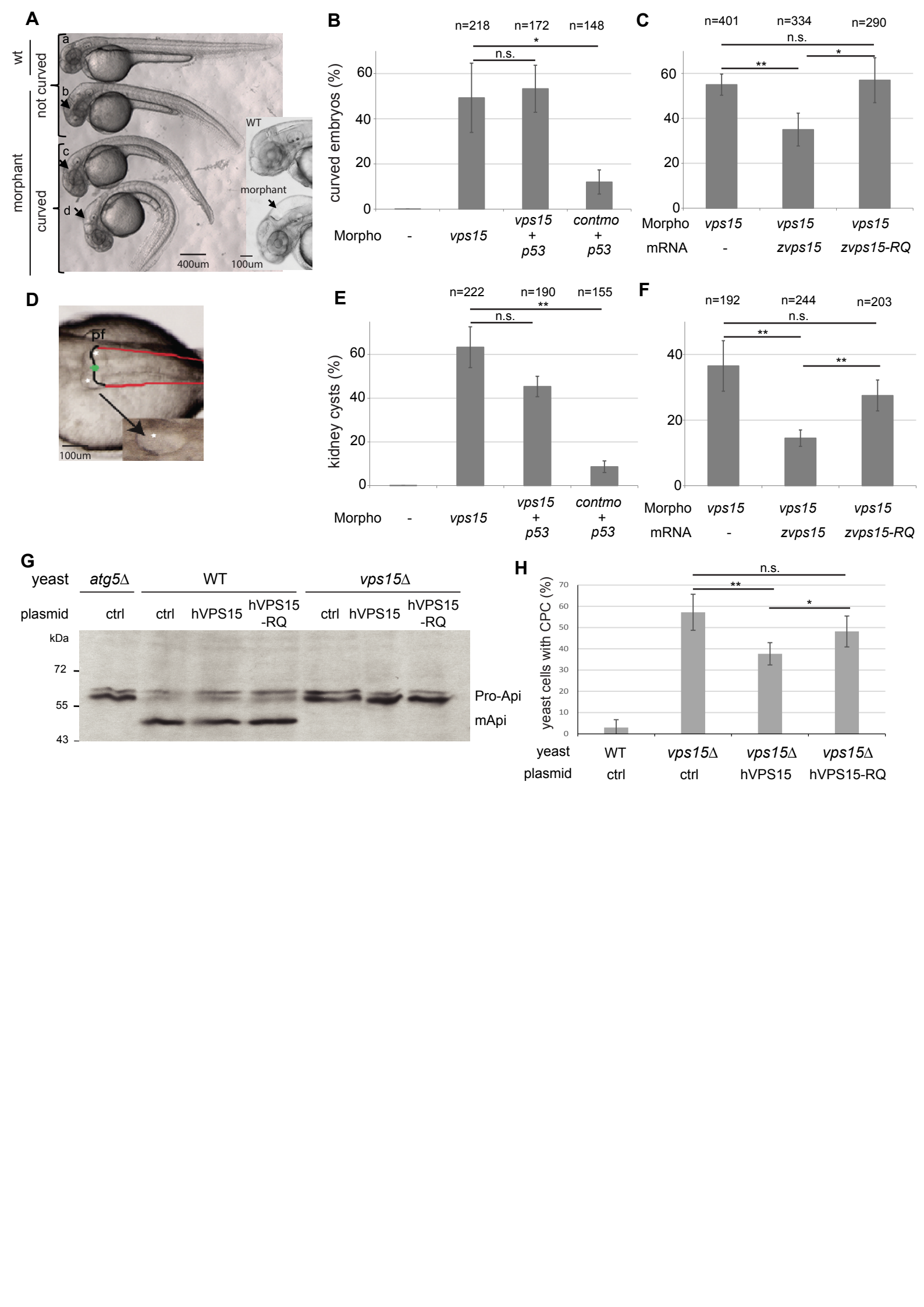
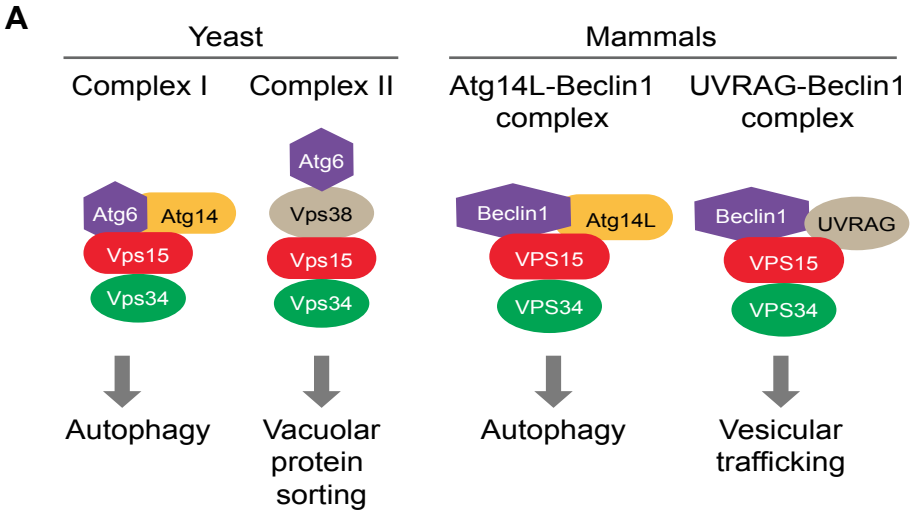
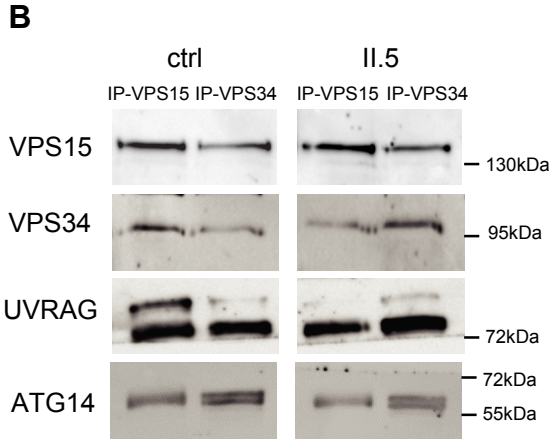
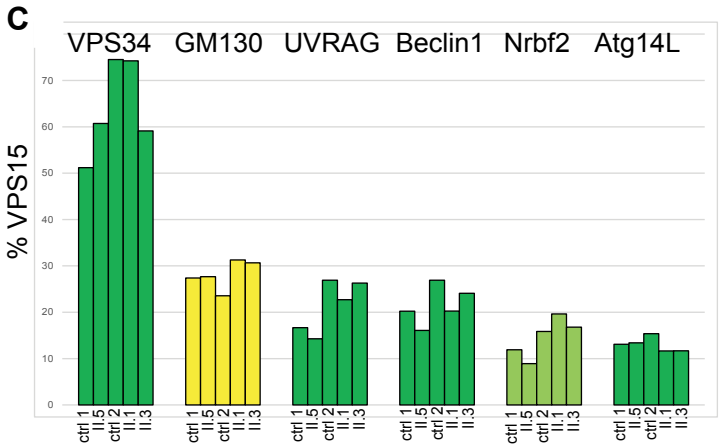


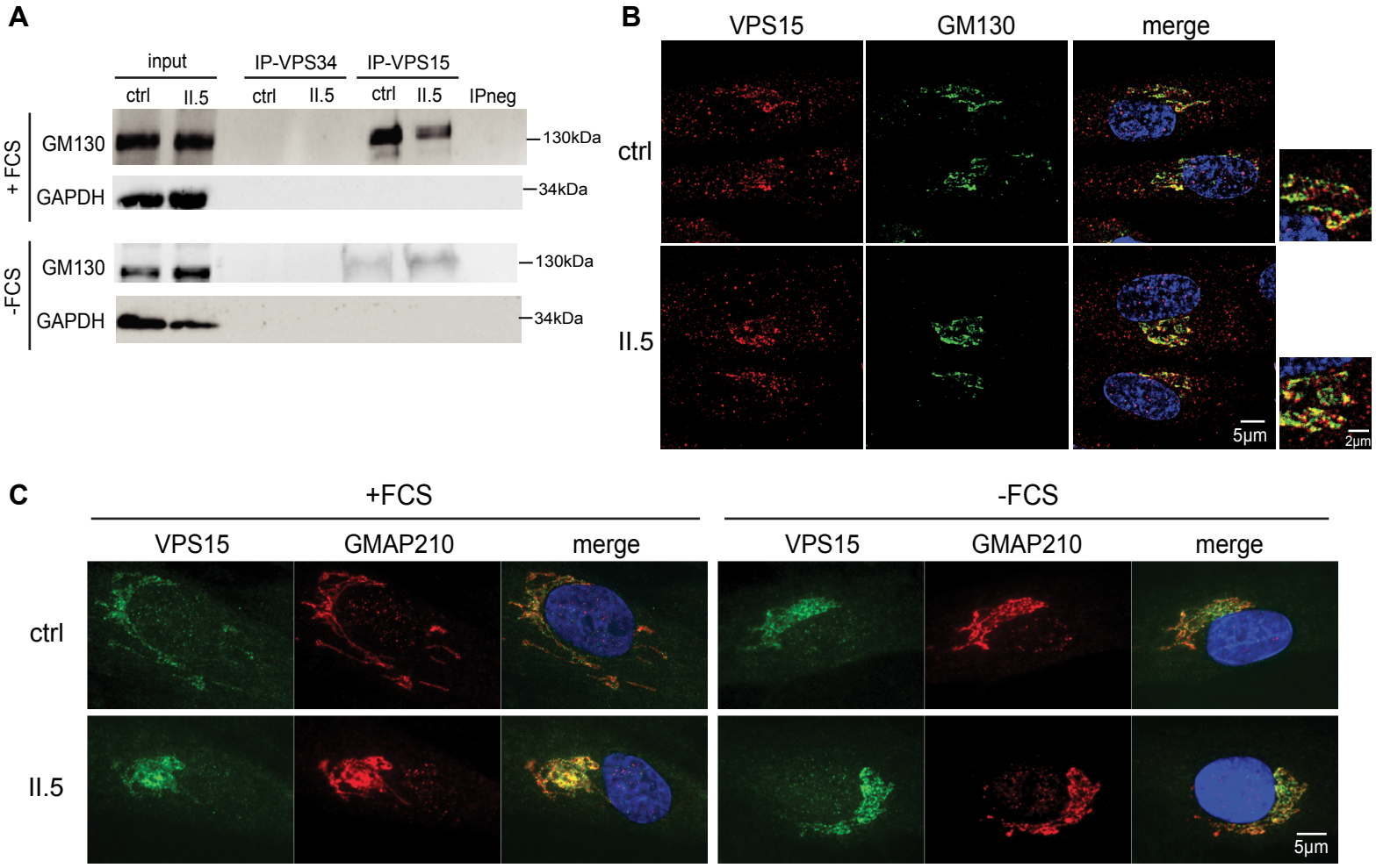
Figure 2 (Friant)


A
wt
not curved
morphant
curved
a
b
c
d
WT
morphant
400um
100um
B
n=218
n=172
n=148
*
n.s.
curved embryos (%)
60
40
20
0
Morpho
-
vps15
vps15 + p53
contmo + p53
C
n=401
n=334
n=290
n.s.
**
*
60
40
20
0
Morpho
vps15
vps15
vps15
mRNA
-
zvps15
zvps15-RQ
D
pf
100um
E
n=222
n=190
n=155
**
n.s.
kidney cysts (%)
60
40
20
0
Morpho
-
vps15
vps15 + p53
contmo + p53
F
n=192
n=244
n=203
n.s.
**
**
40
20
0
Morpho
vps15
vps15
vps15
mRNA
-
zvps15
zvps15-RQ
G
yeast
atg5Δ
WT
vps15Δ
plasmid
ctrl
ctrl
hVPS15
hVPS15 -RQ
ctrl
hVPS15
hVPS15 -RQ
kDa
72
55
43
Pro-Api
mApi
H
n.s.
**
*
yeast cells with CPC (%)
70
60
50
40
30
20
10
0
yeast
WT
vps15Δ
vps15Δ
vps15Δ
plasmid
ctrl
ctrl
hVPS15
hVPS15-RQ

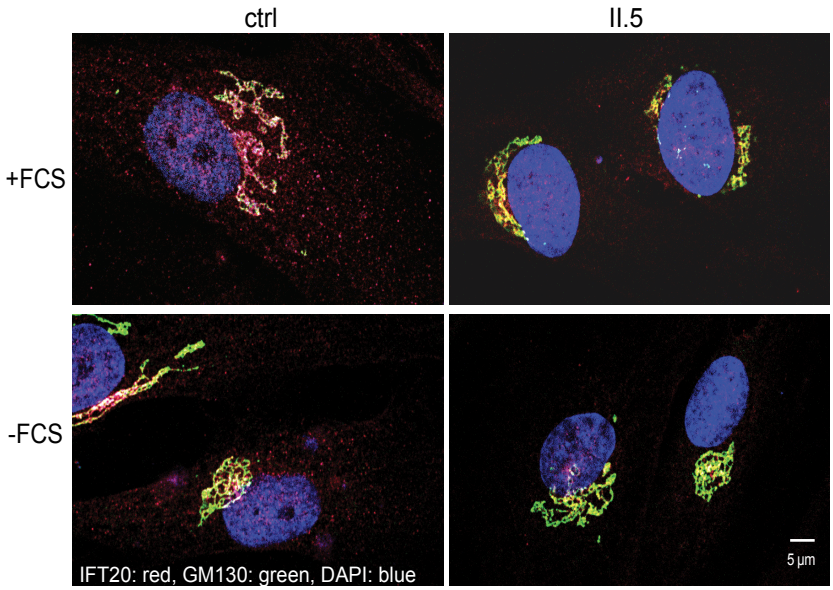

A
Yeast
Mammals
Complex I
Complex II
Atg14L-Beclin1 complex
UVRAG-Beclin1 complex
Atg6
Atg14
Vps15
Vps34
Atg6
Vps38
Vps15
Vps34
Beclin1
Atg14L
VPS15
VPS34
Beclin1
UVRAG
VPS15
VPS34
Autophagy
Vacuolar protein sorting
Autophagy
Vesicular trafficking

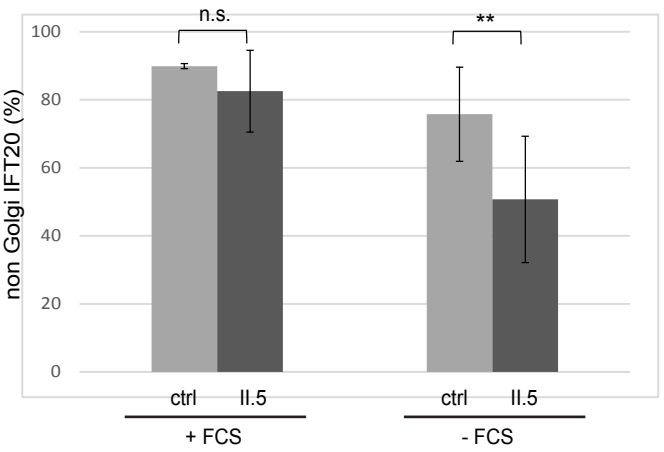

B
ctrl
II.5
IP-VPS15 IP-VPS34
IP-VPS15 IP-VPS34
VPS15
VPS34
UVRAG
ATG14
130kDa
95kDa
72kDa
72kDa
55kDa


C
VPS34
GM130
UVRAG
Beclin1
Nrbf2
Atg14L
% VPS15
0
10
20
30
40
50
60
70
ctrl 1
II.5
ctrl 2
II.1
II.3

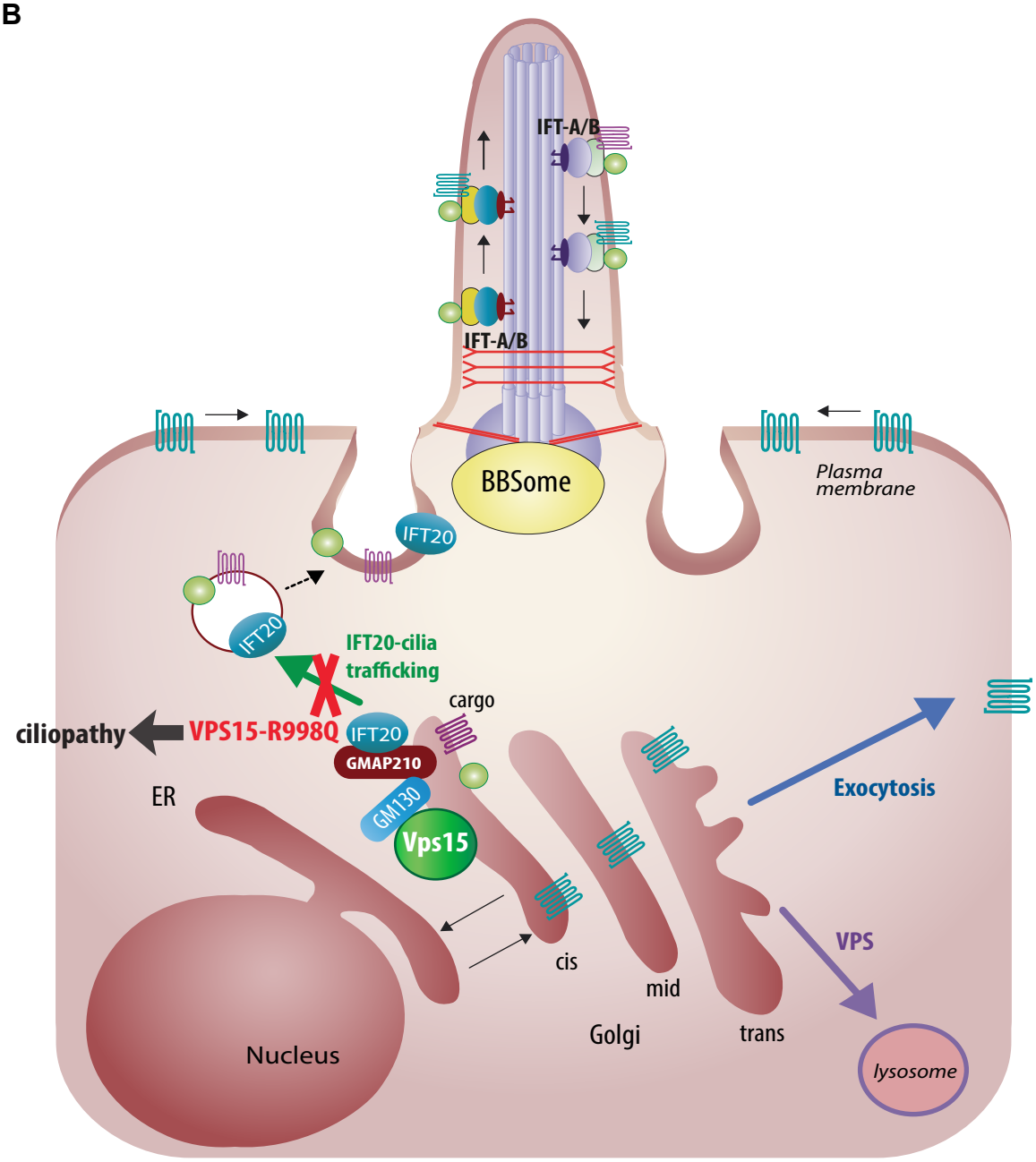

A
input
IP-VPS34
IP-VPS15
ctrl
II.5
ctrl
II.5
ctrl
II.5
IPneg
+ FCS
GM130
GAPDH
130kDa
34kDa
-FCS
GM130
GAPDH
130kDa
34kDa
B
VPS15
GM130
merge
ctrl
II.5
5µm
2µm
C
+FCS
-FCS
VPS15
GMAP210
merge
VPS15
GMAP210
merge
ctrl
II.5
5µm

A


ctrl
II.5
+FCS
-FCS
IFT20: red, GM130: green, DAPI: blue
5 µm


n.s.
**
100
80
60
40
20
0
non Golgi IFT20 (%)
ctrl
II.5
ctrl
II.5
+ FCS
- FCS

B


IFT-A/B
IFT-A/B
BBSome
Plasma membrane
IFT20
IFT20
IFT20-cilia trafficking
cargo
ciliopathy
VPS15-R998Q
IFT20
GMAP210
GM130
Vps15
ER
Exocytosis
VPS
cis
mid
Golgi
trans
Nucleus
lysosome

Annexe IV: CilioPath supplementary data

References

107th United States Congress (2002). An Act to amend the Public Health Service Act to establish an Office of Rare Diseases at the National Institutes of Health, and for other purposes. (RDA).

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature Methods *7*, 248–249.

Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet *Chapter 7*, Unit7.20–7.20.41.

Alwan, A., Modell, B., Bittles, A.H., Czeilel, A., and Hamamy, H. (1997). Community control of genetic and congenital disorders. Alwan, a., Modell, B., Bittles, a.H. <Http://Researchrepository.Murdoch.Edu.Au/View/Author/Bittles, Alan.Html>, Czeilel, a. and Hamamy, H. (1997) Community Control of Genetic and Congenital Disorders. World Health Organisation. Office for the Eastern Mediterranean.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research *43*, D789–D798.

Ankum, W.M., Houtzager, H.L., and Bleker, O.P. (1996). Reinier De Graaf (1641-1673) and the fallopian tube. Hum. Reprod. Update *2*, 365–369.

Antonarakis, S.E., and Beckmann, J.S. (2006). Mendelian disorders deserve more attention. Nature Reviews Genetics *7*, 277–282.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. *79*, 137–158.

Avidor-Reiss, T., Maer, A.M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C.S. (2004). Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell *117*, 527–539.

Baker, K., and Beales, P.L. (2009). Making sense of cilia in disease: The human ciliopathies. American Journal of Medical Genetics Part C: Seminars in Medical Genetics *151C*, 281–295.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics *12*, 745–755.

Banks, R.E., Tirukonda, P., Taylor, C., Hornigold, N., Astuti, D., Cohen, D., Maher, E.R., Stanley, A.J., Harnden, P., Joyce, A., et al. (2006). Genetic and epigenetic analysis of von

Hippel-Lindau (VHL) gene alterations and relationship with clinical variables in sporadic renal cancer. Cancer Res. *66*, 2000–2011.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Research *41*, D991–D995.

Basiri, M.L., Ha, A., Chadha, A., Clark, N.M., Polyanovsky, A., Cook, B., and Avidor-Reiss, T. (2014). A migrating ciliary gate compartmentalizes the site of axoneme assembly in Drosophila spermatids. Curr. Biol. *24*, 2622–2631.

Benmerah, A., Durand, B., Giles, R.H., Harris, T., Kohl, L., Laclef, C., Meilhac, S.M., Mitchison, H.M., Pedersen, L.B., Roepman, R., et al. (2015). The more we know, the more we have to discover: an exciting future for understanding cilia and ciliopathies. Cilia *4*, 5.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013). GenBank. Nucleic Acids Research *41*, D36–D42.

Bermejo-Das-Neves, C., Nguyen, H.-N., Poch, O., and Thompson, J.D. (2014). A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). BMC Bioinformatics *15*, 111.

Björling, E., and Uhlen, M. (2008). Antibodypedia, a portal for sharing antibody and antigen validation data. Mol. Cell Proteomics *7*, 2028–2037.

Blacque, O.E., Perens, E.A., Boroevich, K.A., Inglis, P.N., Li, C., Warner, A., Khattra, J., Holt, R.A., Ou, G., Mah, A.K., et al. (2005). Functional genomics of the cilium, a sensory organelle. Curr. Biol. *15*, 935–941.

Bloch-Zupan, A., Jamet, X., Etard, C., Laugel, V., Muller, J., Geoffroy, V., Strauss, J.-P., Pelletier, V., Marion, V., Poch, O., et al. (2011). Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. The American Journal of Human Genetics *89*, 773–781.

Bloodgood, R.A. (2009). From central to rudimentary to primary: the history of an underappreciated organelle whose time has come. The primary cilium. Methods Cell Biol. *94*, 3–52.

Bonnafe, E., Touka, M., AitLounis, A., Baas, D., Barras, E., Ucla, C., Moreau, A., Flamant, F., Dubruille, R., Couble, P., et al. (2004). The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification. Mol. Cell. Biol. *24*, 4417–4427.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics *33 Suppl*, 228–237.

Breiman, L. (2001). Random forests Machine Learning 45: 5–32 (Figure Titles and Legends described only once and is …).

Brendel, V., Xing, L., and Zhu, W. (2004). Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics *20*, 1157–1169.

Brunak, S., Engelbrecht, J., and Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. Journal of Molecular Biology *220*, 49–65.

Brunham, L.R., and Hayden, M.R. (2013). Hunting human disease genes: lessons from the past, challenges for the future. Hum. Genet. *132*, 603–617.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology *268*, 78–94.

Cavalier-Smith, T. (2014). The neomuran revolution and phagotrophic origin of eukaryotes and cilia in the light of intracellular coevolution and a revised tree of life. Cold Spring Harb Perspect Biol *6*, a016006–a016006.

Chávez, M., Ena, S., Van Sande, J., de Kerchove d'Exaerde, A., Schurmans, S., and Schiffmann, S.N. (2015). Modulation of Ciliary Phosphoinositide Content Regulates Trafficking and Sonic Hedgehog Signaling Output. Dev. Cell *34*, 338–350.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., and Wishart, D.S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Research *36*, W399–W405.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. The American Journal of Human Genetics *97*, 199–215.

Clark, M.J., Chen, R., Lam, H.Y.K., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. Nature Biotechnology *29*, 908–914.

Cobb, M. (2006). Heredity before genetics: a history. Nature Reviews Genetics *7*, 953–958.

Cobb, M. (2008). Generation (Bloomsbury Publishing USA).

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research *38*, 1767–1771.

Colwill, K., Renewable Protein Binder Working Group, and Gräslund, S. (2011). A roadmap to generate renewable protein binders to the human proteome. Nature Methods *8*, 551–558.

Cooper, G.M. (2015). Parlez-vous VUS? Genome Research *25*, 1423–1426.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Research *15*, 901–913.

Creighton, H.B., and McClintock, B. (1931). A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. Proc. Natl. Acad. Sci. U.S.a. *17*, 492–497.

Crick, F.H. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Crow, E.W., and Crow, J.F. (2002). 100 years ago: Walter Sutton and the chromosome theory of heredity. (Genetics Society of America).

Darwin, C. (1859). The Origin of Species (London, Albemarle Street: John Murray).

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLOS Comput Biol *6*, e1001025.

de Graaf, R. (1672). Regneri de Graaf De mulierum organis generationi inservientibus (Leiden, The Netherlands).

de Maupertuis, P.L.M. (1745). Venus physique (La Haye).

de Maupertuis, P.L.M. (1746). Dissertation physique à l'occasion de Nègre Blanc (Haye: Jean Martin Husson).

de Ravel, T.J.L., Devriendt, K., Fryns, J.-P., and Vermeesch, J.R. (2007). What's new in karyotyping? The move towards array comparative genomic hybridisation (CGH). Eur. J. Pediatr. *166*, 637–643.

de Réaumur, R.A.F. (1751). Pratique de l"art de faire eclorre et d"élever en toute saison des oiseaux domestiques de toutes espèces (Paris: Imprimerie Royale).

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research *37*, e67–e67.

Donnai, D., Elles, R., and Ravine, D. (2001). Integrated regional genetic services: Current and future provision/Commentary. British Medical ….

Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science *314*, 1461–1463.

Dunnen, J.T.D., and Antonarakis, S.E. (2000). Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. Human Mutation *15*, 7–12.

Dutcher, S.K. (1995). Flagellar assembly in two hundred and fifty easy-to-follow steps. Trends Genet. *11*, 398–404.

Efimenko, E., Bubb, K., Mak, H.Y., Holzman, T., Leroux, M.R., Ruvkun, G., Thomas, J.H., and Swoboda, P. (2005). Analysis of xbx genes in C. elegans. Development *132*, 1923–1934.

European Medicines Agency (2007). Orphan drugs and rare diseases at a glance. (London: Press Office).

European Parliament and the Council (2000). Regulation (EC) 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products. (16 December 1999).

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B.,

Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Research *44*, D481–D487.

Falk, N., Lösl, M., Schröder, N., and Gießl, A. (2015). Specialized Cilia in Mammalian Sensory Systems. Cells *4*, 500–519.

Farrell, C.M., O'Leary, N.A., Harte, R.A., Loveland, J.E., Wilming, L.G., Wallin, C., Diekhans, M., Barrell, D., Searle, S.M.J., Aken, B., et al. (2014). Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Research *42*, D865–D872.

Ferkol, T.W., and Leigh, M.W. (2012). Ciliopathies: the central role of cilia in a spectrum of pediatric disorders. J. Pediatr. *160*, 366–371.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Fisch, C., and Dupuis-Williams, P. (2011). Ultrastructure of cilia and flagella - back to the future! Biol. Cell *103*, 249–270.

Fontaine, J.-F., Barbosa-Silva, A., Schaefer, M., Huska, M.R., Muro, E.M., and Andrade-Navarro, M.A. (2009). MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Research *37*, W141–W146.

Freund, Y., and Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences *55*, 119–139.

Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., Polman, J., and Alkema, W. (2008). CoPub: a literature-based keyword enrichment tool for microarray data analysis. Nucleic Acids Research *36*, W406–W410.

Frousios, K., Iliopoulos, C.S., Schlitt, T., and Simpson, M.A. (2013). Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy. Genomics *102*, 223–228.

Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Research *43*, D1049–D1056.

Geremek, M., Ziętkiewicz, E., Bruinenberg, M., Franke, L., Pogorzelski, A., Wijmenga, C., and Witt, M. (2014). Ciliary genes are down-regulated in bronchial tissue of primary ciliary dyskinesia patients. Plos One *9*, e88216.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2012). Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. *20*, 490–497.

Gilula, N.B., and Satir, P. (1972). The ciliary necklace. A ciliary membrane specialization. J. Cell Biol. *53*, 494–509.

Gluenz, E., Höög, J.L., Smith, A.E., Dawe, H.R., Shaw, M.K., and Gull, K. (2010). Beyond 9+0: noncanonical axoneme structures characterize sensory cilia from protists to humans. Faseb J. *24*, 3117–3121.

Goetz, S.C., and Anderson, K.V. (2010). The primary cilium: a signalling centre during vertebrate development. Nature Reviews Genetics *11*, 331–344.

Graser, S., Stierhof, Y.-D., Lavoie, S.B., Gassner, O.S., Lamla, S., Le Clech, M., and Nigg, E.A. (2007). Cep164, a novel centriole appendage protein required for primary cilium formation. J. Cell Biol. *179*, 321–330.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org: the HGNC resources in 2015. Nucleic Acids Research *43*, D1079–D1085.

Griffith, F. (1928). The Significance of Pneumococcal Types. J Hyg (Lond) *27*, 113–159.

Hatem, A., Bozdağ, D., Toland, A.E., and Çatalyürek, Ü.V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14*, 184.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. *36*, 39–56.

Hicks, S., Wheeler, D.A., Plon, S.E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Human Mutation *32*, 661–668.

Hildebrandt, F., Benzing, T., and Katsanis, N. (2011). Ciliopathies. New England Journal of Medicine *364*, 1533–1543.

Hirschman, L., Burns, G.A.P.C., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., Wu, C.H., Chatr-Aryamontri, A., Dowell, K.G., Huala, E., et al. (2012). Text mining for the biocuration workflow. Database (Oxford) *2012*, bas020–bas020.

Hoffmann, R., and Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics *21 Suppl 2*, ii252–ii258.

Hokamp, K., and Wolfe, K.H. (2004). PubCrawler: keeping up comfortably with PubMed and GenBank. Nucleic Acids Research *32*, W16–W19.

Hoyer-Fender, S. (2010). Centriole maturation and transformation to basal body. Semin. Cell Dev. Biol. *21*, 142–147.

Hu, J., and Ng, P.C. (2013). SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. Plos One *8*, e77940.

Ishikawa, H., Kubo, A., Tsukita, S., and Tsukita, S. (2005). Odf2-deficient mother centrioles lack distal/subdistal appendages and the ability to generate primary cilia. Nat. Cell Biol. *7*, 517–524.

Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Research *42*, 13534–13544.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research *28*, 27–30.

Kang, N., van Mulligen, E.M., and Kors, J.A. (2011). Comparing and combining chunkers of

biomedical text. J Biomed Inform *44*, 354–360.

Kaplan, W., Wirtz, V.J., Mantel-Teeuwisse, A., Stolk, P., Duthey, B., and Laing, R. (2013). Priority medicines for Europe and the World 2013 Update (Geneva: World Health Organization Press).

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073–1080.

Kilburn, C.L., Pearson, C.G., Romijn, E.P., Meehl, J.B., Giddings, T.H., Culver, B.P., Yates, J.R., and Winey, M. (2007). New Tetrahymena basal body protein components identify basal body domain structure. J. Cell Biol. *178*, 905–912.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics *46*, 310–315.

Kirchhoff, M., Rose, H., and Lundsteen, C. (2001). High resolution comparative genomic hybridisation in clinical cytogenetics. Journal of Medical Genetics *38*, 740–744.

Ko, H.W. (2012). The primary cilium as a multiple cellular signaling scaffold in development and disease. BMB Rep *45*, 427–432.

Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics *82*, 949–958.

Köhler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C.M., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research *42*, D966–D974.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols *4*, 1073–1081.

Kunkel, T.A., and Erie, D.A. (2005). DNA mismatch repair. Annu. Rev. Biochem. *74*, 681–710.

Kurotaki, N., Harada, N., Yoshiura, K., Sugano, S., Niikawa, N., and Matsumoto, N. (2001). Molecular characterization of NSD1, a human homologue of the mouse Nsd1 gene. Gene *279*, 197–204.

la Paz, de, M.P., Villaverde-Hueso, A., Alonso, V., János, S., Zurriaga, O., Pollán, M., and Abaitua-Borda, I. (2010). Rare diseases epidemiology research. Adv. Exp. Med. Biol. *686*, 17–39.

Lander, E.S., and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science *236*, 1567–1570.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and

Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research *42*, D980–D985.

Lane, N. (2015). The unseen world: reflections on Leeuwenhoek (1677) “Concerning little animals.” Philos. Trans. R. Soc. Lond., B, Biol. Sci. *370*, 20140344–20140344.

Layer, R.M., Kindlon, N., Karczewski, K.J., and Quinlan, A.R. (2015). Efficient genotype compression and analysis of large genetic-variation data sets. Nature Methods *13*, 63–65.

Lee, C., Iafrate, A.J., and Brothman, A.R. (2007). Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nature Genetics *39*, S48–S54.

Lee, W.H., Bookstein, R., Hong, F., Young, L.J., Shew, J.Y., and Lee, E.Y. (1987). Human retinoblastoma susceptibility gene: cloning, identification, and sequence. Science *235*, 1394–1399.

Leitao, D.D. (2012). The Pregnant Male as Myth and Metaphor in Classical Greek Literature (Cambridge University Press).

Lewenhoeck, A. (1678). Observationes de Anthonu Lewenhoeck, de natis e semine genitali animalculis. Phil Trans Roy Soc *12*, 1040–1043.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., et al. (2004). Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell *117*, 541–552.

Li, M.-X., Kwan, J.S.H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., and Sham, P.C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet. *9*, e1003143.

Liem, K.F., He, M., Ocbina, P.J.R., and Anderson, K.V. (2009). Mouse Kif7/Costal2 is a cilia-associated protein that regulates Sonic hedgehog signaling. Proc. Natl. Acad. Sci. U.S.a. *106*, 13377–13382.

Liu, Q., Tan, G., Levenkova, N., Li, T., Pugh, E.N., Rux, J.J., Speicher, D.W., and Pierce, E.A. (2007). The proteome of the mouse photoreceptor sensory cilium complex. Mol. Cell Proteomics *6*, 1299–1317.

López-Beltrán, C. (1995). « Les maladies héréditaires » : 18th century disputes in France/Les maladies héréditaires : controverses au XVIIIe siècle en France. Revue D'histoire Des Sciences *48*, 307–350.

Luu, T.-D., Rusu, A., Walter, V., Linard, B., Poidevin, L., Ripp, R., Moulinier, L., Muller, J., Raffelsberger, W., Wicker, N., et al. (2012). KD4v: Comprehensible Knowledge Discovery System for Missense Variant. Nucleic Acids Research *40*, W71–W75.

MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature *508*, 469–476.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research *39*, D52–D57.

Mandloi, S., and Chakrabarti, S. (2015). PALM-IST: Pathway Assembly from Literature Mining--an Information Search Tool. Scientific Reports *5*, 10021.

Marshall, W.F., and Rosenbaum, J.L. (2001). Intraflagellar transport balances continuous turnover of outer doublet microtubules: implications for flagellar length control. J. Cell Biol. *155*, 405–414.

Matthaei, J.H., and Nirenberg, M.W. (1961). Characteristics and stabilization of DNAase-sensitive protein synthesis in E. coli extracts. Proc. Natl. Acad. Sci. U.S.a. *47*, 1580–1588.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.

Mendel, G. (1866). Versuche über Pflanzenhybriden (Brünn: Verhandlungen des naturforschenden Vereines in Brünn).

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nature Reviews Genetics *11*, 31–46.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., et al. (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research *17*, 1797–1808.

Mitchell, A.A., Zwick, M.E., Chakravarti, A., and Cutler, D.J. (2004). Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics *20*, 1022–1032.

Mockel, A., Perdomo, Y., Stutzmann, F., Letsch, J., Marion, V., and Dollfus, H. (2011). Retinal dystrophy in Bardet-Biedl syndrome and related syndromic ciliopathies. Prog Retin Eye Res *30*, 258–274.

Mottaz, A., David, F.P.A., Veuthey, A.-L., and Yip, Y.L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics *26*, 851–852.

Mougou-Zerelli, S., Thomas, S., Szenker, E., Audollent, S., Elkhartoufi, N., Babarit, C., Romano, S., Salomon, R., Amiel, J., Esculpavit, C., et al. (2009). CC2D2A mutations in Meckel and Joubert syndromes indicate a genotype-phenotype correlation. Human Mutation *30*, 1574–1582.

Muller, J., Stoetzel, C., Vincent, M.C., Leitch, C.C., Laurier, V., Danse, J.M., Hellé, S., Marion, V., Bennouna-Greene, V., Vicaire, S., et al. (2010). Identification of 28 novel mutations in the Bardet-Biedl syndrome genes: the burden of private mutations in an extensively heterogeneous disease. Hum. Genet. *127*, 583–593.

Murcia, N.S., Richards, W.G., Yoder, B.K., Mucenski, M.L., Dunlap, J.R., and Woychik, R.P. (2000). The Oak Ridge Polycystic Kidney (orpk) disease gene is required for left-right axis determination. Development *127*, 2347–2355.

Müller, H.-M., Rangarajan, A., Teal, T.K., and Sternberg, P.W. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. Neuroinformatics *6*, 195–204.

Nachury, M.V., Loktev, A.V., Zhang, Q., Westlake, C.J., Peränen, J., Merdes, A., Slusarski, D.C., Scheller, R.H., Bazan, J.F., Sheffield, V.C., et al. (2007). A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. Cell *129*, 1201–1213.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research *31*, 3812–3814.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature Genetics *42*, 790–793.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010b). Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics *42*, 30–35.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med *5*, 28.

Ogino, S., Gulley, M.L., Dunnen, den, J.T., Wilson, R.B., Association for Molecular Patholpogy Training and Education Committtee (2007). Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J Mol Diagn *9*, 1–6.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Research *43*, D82–D86.

Omran, H. (2010). NPHP proteins: gatekeepers of the ciliary compartment. J. Cell Biol. *190*, 715–717.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research *42*, D358–D363.

Orphanet (2015). «Prevalence of rare diseases: Bibliographic data », Orphanet Report Series, Rare Diseases collection, July 2015, Number 2 : Diseases listed by decreasing prevalence, incidence or number of published cases.

Ostrowski, L.E., Blackburn, K., Radde, K.M., Moyer, M.B., Schlatzer, D.M., Moseley, A., and Boucher, R.C. (2002). A proteomic analysis of human cilia: identification of novel components. Mol. Cell Proteomics *1*, 451–465.

Ott, J., Wang, J., and Leal, S.M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics *16*, 275–284.

Pampliega, O., Orhon, I., Patel, B., Sridhar, S., Díaz-Carretero, A., Beau, I., Codogno, P., Satir, B.H., Satir, P., and Cuervo, A.M. (2013). Functional interaction between autophagy and ciliogenesis. Nature *502*, 194–200.

Pasternak, J.J. (2003). Génétique moléculaire humaine (De Boeck Supérieur).

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nature Biotechnology *27*, 1173–1175.

Paweletz, N. (2001). Walther Flemming: pioneer of mitosis research. (Nature Publishing Group).

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J. Cell Biol. *170*, 103–113.

Perrone, C.A., Tritschler, D., Taulman, P., Bower, R., Yoder, B.K., and Porter, M.E. (2003). A novel dynein light intermediate chain colocalizes with the retrograde motor for intraflagellar transport at sites of axoneme assembly in chlamydomonas and Mammalian cells. Mol. Biol. Cell *14*, 2041–2056.

Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Research *29*, 1185–1190.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research *20*, 110–121.

Praetorius, H.A., and Spring, K.R. (2005). A physiological view of the primary cilium. Annual Review of Physiology *67*, 515–529.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Research *42*, D756–D763.

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., and Jimeno, A. (2008). Text processing through Web services: calling Whatizit. Bioinformatics *24*, 296–298.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. *4*, 311–323.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the Clinical Genome Resource. New England Journal of Medicine *372*, 2235–2242.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. (Nature Publishing Group), pp. 405–424.

Roll-Hansen, N. (2014). Commentary: Wilhelm Johannsen and the problem of heredity at the turn of the 19th century. Int J Epidemiol *43*, 1007–1013.

Rosen, B., Schick, J., and Wurst, W. (2015). Beyond knockouts: the International Knockout Mouse Consortium delivers modular and evolving tools for investigating mammalian genes. Mamm. Genome *26*, 456–466.

Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., et al. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Research *41*, D56–D63.

Ruzicka, L., Bradford, Y.M., Frazer, K., Howe, D.G., Paddock, H., Ramachandran, S., Singer, A., Toro, S., Van Slyke, C.E., Eagle, A.E., et al. (2015). ZFIN, The zebrafish model organism database: Updates and new directions. Genesis *53*, 498–509.

Sandler, I. (1983). Pierre Louis Moreau de Maupertuis—A precursor of Mendel? J Hist Biol *16*, 101–136.

Sang, L., Miller, J.J., Corbit, K.C., Giles, R.H., Brauer, M.J., Otto, E.A., Baye, L.M., Wen, X., Scales, S.J., Kwong, M., et al. (2011). Mapping the NPHP-JBTS-MKS protein network reveals ciliopathy disease genes and pathways. Cell *145*, 513–528.

Satish Tammana, T.V., Tammana, D., Diener, D.R., and Rosenbaum, J. (2013). Centrosomal protein CEP104 (Chlamydomonas FAP256) moves to the ciliary tip during ciliary assembly. J. Cell. Sci. *126*, 5018–5029.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Research *15*, 7155–7174.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research *29*, 308–311.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in

vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034–1050.

Smith, C.L., and Eppig, J.T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine *1*, 390–399.

Smith, J.C., Northey, J.G.B., Garg, J., Pearlman, R.E., and Siu, K.W.M. (2005). Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila. J. Proteome Res. *4*, 909–919.

Snow, J.J., Ou, G., Gunnarson, A.L., Walker, M.R.S., Zhou, H.M., Brust-Mascher, I., and Scholey, J.M. (2004). Two anterograde intraflagellar transport motors cooperate to build sensory cilia on C. elegans neurons. Nat. Cell Biol. *6*, 1109–1113.

Sonnhammer, E.L.L., and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Research *43*, D234–D239.

Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M., Parvin, J.D., Shendure, J., and Fields, S. (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics *200*, 413–422.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.

Stolc, V., Samanta, M.P., Tongprasit, W., and Marshall, W.F. (2005). Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes. Proc. Natl. Acad. Sci. U.S.a. *102*, 3703–3707.

Sung, C.-H., and Leroux, M.R. (2013). The roles of evolutionarily conserved functional modules in cilia-related trafficking. Nat. Cell Biol. *15*, 1387–1397.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research *43*, D447–D452.

Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204.

Tang, Z., Lin, M.G., Stowe, T.R., Chen, S., Zhu, M., Stearns, T., Franco, B., and Zhong, Q. (2013). Autophagy promotes primary ciliogenesis by removing OFD1 from centriolar satellites. Nature *502*, 254–257.

Tari, L., Anwar, S., Liang, S., Cai, J., and Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics *26*, i547–i553.

Taylor, A.M., Boyde, A., Wilson, P.J.M., Jarvis, J.C., Davidson, J.S., Hunt, J.A., Ranganath, L.R., and Gallagher, J.A. (2011). The role of calcified cartilage and subchondral bone in the initiation and progression of ochronotic arthropathy in alkaptonuria. Arthritis Rheum. *63*,

3887–3896.

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I.G., Hansson, M.G., 't Hoen, P.-B.A., Patrinos, G.P., Dawkins, H., et al. (2014). RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med *29 Suppl 3*, S780–S787.

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation *32*, 358–368.

Treharne, K.J., Cassidy, D., Goddard, C., Colledge, W.H., Cassidy, A., and Mehta, A. (2009). Epithelial IgG and its relationship to the loss of F508 in the common mutant form of the cystic fibrosis transmembrane conductance regulator. FEBS Lett. *583*, 2493–2499.

Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., and Ananiadou, S. (2011). Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics *27*, i111–i119.

Tukachinsky, H., Lopez, L.V., and Salic, A. (2010). A mechanism for vertebrate Hedgehog signaling: recruitment to cilia and dissociation of SuFu-Gli protein complexes. J. Cell Biol. *191*, 415–428.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Research *43*, D204–D212.

van Reeuwijk, J., Arts, H.H., and Roepman, R. (2011). Scrutinizing ciliopathies by unraveling ciliary interaction networks. Hum. Mol. Genet. *20*, R149–R157.

Vertii, A., Hung, H.-F., Hehnly, H., and Doxsey, S. (2016). Human basal body basics. Cilia *5*, 13.

Vissers, L.E.L.M., Veltman, J.A., van Kessel, A.G., and Brunner, H.G. (2005). Identification of disease genes by whole genome CGH arrays. Hum. Mol. Genet. *14 Spec No. 2*, R215–R223.

Ward, C., Martinez-Lopez, N., Otten, E.G., Carroll, B., Maetzel, D., Singh, R., Sarkar, S., and Korolchuk, V.I. (2016). Autophagy, lipophagy and lysosomal lipid storage disorders. Biochim. Biophys. Acta *1864*, 269–284.

Watanabe, D., Saijoh, Y., Nonaka, S., Sasaki, G., Ikawa, Y., Yokoyama, T., and Hamada, H. (2003). The left-right determinant Inversin is a component of node monocilia and other 9+0 cilia. Development *130*, 1725–1734.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research *41*, W518–W522.

Wheatley, D.N. (1995). Primary cilia in normal and pathological tissues. Pathobiology *63*, 222–238.

Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., and Klein, T.E. (2012). Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. *92*, 414–417.

Williams, C.L., Li, C., Kida, K., Inglis, P.N., Mohan, S., Semenec, L., Bialas, N.J., Stupay, R.M., Chen, N., Blacque, O.E., et al. (2011). MKS and NPHP modules cooperate to establish basal body/transition zone membrane associations and ciliary gate function during ciliogenesis. J. Cell Biol. *192*, 1023–1041.

Wilming, L.G., Gilbert, J.G.R., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J.L. (2008). The vertebrate genome annotation (Vega) database. Nucleic Acids Research *36*, D753–D760.

Winkelmann, A. (2007). Wilhelm von Waldeyer-Hartz (1836-1921): an anatomist who left his mark.

Wrighton, K.H. (2013). Cytoskeleton: Autophagy and ciliogenesis come together. Nature Reviews Molecular Cell Biology *14*, 687–687.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806–1254806.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Research *44*, D710–D716.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yuan, S., and Sun, Z. (2013). Expanding horizons: ciliary proteins reach beyond cilia. Annu. Rev. Genet. *47*, 353–376.

Zhang, B., Shi, Z., Duncan, D.T., Prodduturi, N., Marnett, L.J., and Liebler, D.C. (2011). Relating protein adduction to gene expression changes: a systems approach. Mol Biosyst *7*, 2118–2127.

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic Acids Research *41*, W115–W122.

Webliography

[1] World Health Organization, 'International Classification of Diseases (ICD)', *WHO*. [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed: 17-Jul-2015].

[2] 'OMIM - Online Mendelian Inheritance in Man'. [Online]. Available: http://www.omim.org/. [Accessed: 01-Dec-2015].

[3] 'Home - NORD (National Organization for Rare Disorders)', *NORD (National Organization for Rare Disorders)*. [Online]. Available: http://rarediseases.org/. [Accessed: 11-Dec-2015].

[4] 'Eurordis.org', *EURORDIS*. [Online]. Available: http://www.eurordis.org. [Accessed: 11-Dec-2015].

[5] 'L'association AFM-TÉLÉTHON', *AFM-Téléthon*. [Online]. Available: http://www.afm-telethon.fr/home. [Accessed: 21-Feb-2016].

[6] 'Rare Disease Day 2016 - 29 Feb', *Rare Disease Day - 29 Feb 2016*. [Online]. Available: http://www.rarediseaseday.org/. [Accessed: 21-Feb-2016].

[7] 'IRDiRC – International Rare Diseases Research Consortium'. .

[8] 1000 Genomes, 'Variant Call Format Specifications', *Variant Call Format Specifications*, 08-Dec-2015. [Online]. Available: http://www.1000genomes.org/wiki/Analysis/variant-call-format. [Accessed: 21-Apr-2016].

[9] 'Cilia'. [Online]. Available: http://ciliajournal.biomedcentral.com. [Accessed: 21-Feb-2016].

[10] J. Hu, 'Molecular dissection of the ciliary gate', *Grantome*.

[11] 'Syscilia'. [Online]. Available: http://syscilia.org/. [Accessed: 21-Feb-2016].

[12] Exome Aggregation Consortium, 'Analysis of protein-coding genetic variation in 60,706 humans | bioRxiv'. [Online]. Available: http://biorxiv.org/content/early/2015/10/30/030338. [Accessed: 08-Dec-2015].

[13] R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, A. Amemiya, L. J. Bean, T. D. Bird, C.-T. Fong, H. C. Mefford, R. J. Smith, and K. Stephens, Eds., *GeneReviews(®)*. Seattle (WA): University of Washington, Seattle, 1993.

[14] NCBI, 'E-utilities Quick Start - Entrez Programming Utilities Help - NCBI Bookshelf'. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK25500/. [Accessed: 09-Dec-2015].

[15] 'PyCharm', *JetBrains*. [Online]. Available: https://www.jetbrains.com/pycharm/. [Accessed: 21-Feb-2016].

[16] ‘GATK | GATK Best Practices’. [Online]. Available: https://www.broadinstitute.org/gatk/guide/best-practices.php. [Accessed: 21-Feb-2016].

[17] ‘Natural Language Toolkit — NLTK 3.0 documentation’. [Online]. Available: http://www.nltk.org/. [Accessed: 04-Apr-2016].

UNIVERSITÉ DE STRASBOURG

Kirlsey CHENNEN

Ecole Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Maladies rares et 'Big Data': Solutions bioinformatiques vers une analyse guidée par les connaissances. Applications aux ciliopathies.

Résumé

Au cours de la dernière décennie, la recherche biomédicale et la pratique médicale ont été révolutionné par l'ère post-génomique et l'émergence des « Big Data » en biologie. Il existe toutefois, le cas particulier des maladies rares charactérisées par la rareté, allant de l'effectif des patients jusqu'aux connaissances sur le domaine. Néanmois, les maladies rares représentent un réel intérêt, car les connaissances fondamentales accumulées en temps que modèle d'études et les solutions thérapeutique qui en découlent peuvent également bénéficier à des maladies plus communes. Cette thèse porte sur le développement de nouvelles solutions bioinformatiques, intégrant des données Big Data et des approches guidées par la connaissance pour améliorer l'étude des maladies rares. En particulier, mon travail a permis (i) la création de Pub*Athena*, un outil de criblage de la littérature pour la recommendation de nouvelles publications pertinentes, (ii) le développement d'un outil pour l'analyse de données exomique, VarScrut, qui combine des connaissance multi-niveaux pour améliorer le taux de résolution.

Mots-clés : Maladies génétiques rares, ciliopathies, Exomes, Text-mining

Summary

Over the last decade, biomedical research and medical practice have been revolutionized by the post-genomic era and the emergence of Big Data in biology. The field of rare diseases, are characterized by scarcity from the patient to the domain knowledge. Nevertheless, rare diseases represent a real interest as the fundamental knowledge accumulated as well as the developed therapeutic solutions can also benefit to common underlying disorders. This thesis focuses on the development of new bioinformatics solutions, integrating Big Data and Big Data associated approaches to improve the study of rare diseases. In particular, my work resulted in (i) the creation of Pub*Athena*, a tool for the recommendation of relevant literature updates, (ii) the development of a tool for the analysis of exome datasets, VarScrut, which combines multi-level knowledge to improve the resolution rate.

Keywords: Rare diseases, Exome sequencing, Cilia, Ciliopathies, Text-mining