ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Институт медицины и психологии В. Зельмана НГУ

КУРСОВАЯ РАБОТА

Валеев Эмиль Салаватович
Группа 12452

Тема работы: «Разработка инструментов для поиска клинически значимых полиморфизмов в геноме человека на основе данных секвенирования 3С-библиотек»

Научный руководитель:
Фишман Вениамин Семенович,
к.б.н., ведущий научный сотрудник,
заведующий Сектором геномных механизмов онтогенеза, ИЦиГ СО РАН

ФИО: _____/_____
«____»_____20____ г.

Оценка: _____

Новосибирск, 2020

Содержание

1. Введение

1.1. Актуальность

1.2. Цели

1.3. Задачи

2. Обзор литературы

Генетические болезни (моногенные заболевания, геномные структурные дефекты, вариации числа копий) — это основная причина смертности детей до 10 лет.

Генетическое детерминирование патологий

Частые и редкие (орфанные) патологии. Генетические патологии делятся на группы по частоте встречаемости в популяции — частые и редкие (орфанные). Определения орфанных заболеваний могут различаться — например, в США, согласно “Health Promotion and Disease Prevention Amendments of 1984”, редкими считаются патологии, поражающие менее 200 тыс. населения страны (примерно 1 : 1630 при текущей численности населения в 326 млн человек)[38]. Европейское Медицинское Агентство определяет границу как 1 : 2000. Систематический анализ показал, что существует более 290 определений, и среднее значение находится в интервале 40–50 на 100 тыс. населения[39].

Некоторые заболевания могут быть орфанными в одной популяции и частыми в другой (эффект основателя). Например, бета-талассемия в Средиземноморье.

Несмотря на то, что каждое из орфанных заболеваний само по себе встречается редко, в сумме они поражают значительный процент населения (предположительно 5–8% европейской популяции). Общее число орфанных болезней неизвестно по причине недостатков стандартизации, наиболее частая оценка — 5000–8000. Около 80% редких болезней имеют генетическую природу и начинаются в раннем детстве[40].

Основные источники информации по орфанным заболеваниям:

1. Global Genes
2. Online Mendelian Inheritance in Man (OMIM)[24]
3. Orphadata

2.1. Механизмы развития патологий

Структура белка.

Эпигенетика. Также патологии могут развиваться из-за изменения экспрессии, вызванных эпигенетическими механизмами, не затрагивающими непосредственно последовательность ДНК генов. К таким механизмам можно отнести, например, метилирование ДНК, ацетилирование гистонов. Кроме того, на экспрессию в значительной степени влияет трёхмерная структура хроматина, регулируемая механизмами loop extrusion, block copolymers, фазовая сепарация. ТАДы, петли, етц.

2.2. Типы генетических аномалий, лежащих в основе патологий

Хромосомные аномалии. Анэуплоидии (изменение числа хромосом), перестройки (крупные делеции, дупликации, инверсии и транслокации).

Вариации числа копий (CNV).

Точечные полиморфизмы (SNV).

Короткие инверсии и делеции (indels).

2.3. Функциональные классы вариантов

Внутригенные варианты могут находиться в:

- Нетранслируемых областях (3’ и 5’ UTR), вовлечённых в регуляцию транскрипции, трансляции и деградации транскрипта.
- Экзонах, непосредственно отвечающих за последовательность белка. SNP могут быть синонимичными (без замены аминокислоты) и несинонимичными — миссенс (замена на другую АК), нонсенс (замена на стоп-кодон) либо сдвиг рамки считывания, приводящий к изменению значительной части белковой молекулы.
- Интронах, которые содержат регуляторные области и сплайс-сайты, необходимые для процессинга транскрипта в готовую мРНК.

Внегенные SNP могут приходиться на различные регуляторные последовательности, например, энхансеры, сайленсеры. Также известно, что за трёхмерную структуру хроматина отвечают в том числе и специфические белки, связывающиеся с ДНК — например, CTCF[36]. Варианты, приходящиеся на сайты связывания CTCF, могут разрушать границы ТАДов и вызывать изменения экспрессии.

2.4. Методы детектирования

Кариотипирование.

Флуоресцентная *in situ* гибридизация (FISH). Основой является гибридизация содержащих флуоресцентную метку последовательностей с комплементарными ими участками НК. Гибридизация может производиться на ДНК (метафазные или интерфазные хромосомы) и на РНК. FISH позволяет определить количественные характеристики НК и их пространственное расположение в ядре. Метод является «золотым стандартом» в определении хромосомных патологий — как в клетках с врождёнными перестройками, так и в клетках опухолей.

Данные при FISH можно получить как за счёт спектрального анализа сигналов, так и за счёт их отсутствия/присутствия. Всего 7 флюорофоров дают 127 вариантов цветовых меток; это позволяет реализовать, например, спектральное кариотипирование. В методике MER-FISH количество цветовых меток увеличено до 1001. Тем не менее, лимитирующими факторами остаются:

- потребность в хорошо обученном персонале. Протокол FISH зависит от характера пробы и образца, и должен быть настроен эмпирически;
- цена реактивов;

- время гибридизации. Кинетика реакций гибридизации в ядре изучена недостаточно, и требуется достаточно долгое время, чтобы получить сигналы, которые можно измерить и сравнить между собой.

В настоящее время методика FISH значительно усложнилась. Биотехнологические компании предлагают панели олигонуклеотидов для определённых целей, определяющие участки от десятков килобаз до мегабаз, а также олиги с высокой чувствительностью, позволяющие определить сплайс-варианты и даже SNP. Разрабатываются технологии micro-FISH (μFISH), сочетающие FISH с микрофлюидными технологиями. При этом процесс удешевляется, автоматизируется, ускоряется (за счёт уменьшения объёмов, а соответственно, времени гибридизации) и упрощается для использования в обширных исследованиях и для внедрения в клинику[41].

CGH. Как и в случае с FISH, основой метода является флуоресцентная гибридизация. Однако CGH использует два образца генома — тест и контроль, каждый из которых метится флюорофором, а затем гибридизуется 1 : 1. Таким образом в тестовом образце можно обнаружить CNV и перестройки.

В отличие от FISH, CGH проверяет весь геном на наличие перестроек, не требует знаний о целевом регионе и может быть использован на интерфазных клетках. Однако, как и у классических методов, разрешение CGH ограничено 5–10 Мб.

В настоящее время CGH используется в виде array-CGH (aCGH), комбинируя этот метод с микрочипами. <...> [42]

STS.

MLPA.

Секвенирование по Сэнгеру. Метод, позволяющий с высокой точностью анализировать короткий (до 1kb) фрагмент ДНК[29]. В настоящее время используется для подтверждения вариантов, найденных с помощью описанных выше методов.

Хромосомный микроматричный анализ (ХМА)

Микрочиповая гибридизация. Гаплотипы.

NGS. Секвенирование нового поколения (NGS) — это комплекс технологий, позволяющих прочитать за сравнительно небольшое время миллионы коротких последовательностей ДНК. Благодаря этому единовременно можно проанализировать несколько генов, либо весь геном (в отличие от традиционных методов).

Проблемы данных NGS:

- Ошибки секвенирования.
- Неоднородность покрытия генома.

- Ошибки ПЦР и ПЦР-дубликаты.
- Неточное выравнивание инделов и повторяющихся последовательностей (например, поли-А трактов).

В настоящее время NGS используется и для выявления крупных перестроек (метод Hi-C).

2.5. Виды NGS

Полногеномное секвенирование (WGS). Полногеномное секвенирование со слабым покрытием может быть использовано для определения CNV — например, NIPT, когда используется свободная ДНК плода (cfDNA), циркулирующая в крови матери[37]. WGS также может быть использовано в диагностике микробиома с целью определения источника хронической инфекции, реконструкции путей передачи инфекции, а также выявления антибиотикорезистентных штаммов[28].

Полноэкзомное секвенирование (WES) Полноэкзомное секвенирование часто включает тестирование ребёнка и родителей (так называемый трио-тест), что позволяет улучшить интерпретацию вариантов[27].

Таргетные панели. Данный вид тестов позволяет анализировать гены, ответственные за отдельные группы заболеваний — например, существуют таргетные панели для иммунодефицитов, почечных, неврологических болезней, болезней соединительной ткани, сетчатки, а также предрасположенности к отдельным видам онкологических заболеваний. Также таргетные панели позволяют анализировать клетки опухолей — некоторые приспособлены к выявлению общих для многих раковых линий мутаций, другие же разработаны для специфического типа опухолей[27].

Hi-C

2.6. Базовая схема обработки результатов секвенирования

Демультиплексирование. В процессе секвенирования к целевым фрагментам ДНК могут пришиваться так называемые адаптерные последовательности. Эти адаптеры могут содержать так называемые баркоды — последовательности, с помощью которых можно отличить ДНК различных образцов. Процесс сортировки данных секвенирования по баркодам называется демультиплексированием. Чаще всего демультиплексирование производится самим секвенатором, но иногда его приходится производить вручную.

Удаление адаптерных последовательностей. Если целевая ДНК короче длины прочтения, то фрагменты адаптера на 3’ конце могут попасть в готовые данные, и встаёт вопрос об их удалении. Также присутствие адаптера в прочтениях может быть признаком контаминации, и такие прочтения следует исключить из дальнейшего анализа[13].

Картирование. Прочтения необходимо картировать на некую референсную геномную последовательность.

Проблемы картирования:

- Высоковариативные регионы
- Вырожденные (неуникальные) регионы
- Регионы с инделами

Отличие генома образца от референсного генома называется вариантом (синонимичные термины «мутация» и «полиморфизм» не рекомендованы к употреблению[32]). В настоящее время «золотым стандартом» являются утилиты, использующие алгоритм Берроуса–Уиллера[30].

Удаление ПЦР-дубликатов. Тем не менее, было показано, что для WGS-данных удаление ПЦР-дубликатов имеет минимальный эффект на улучшение поиска полиморфизмов — приблизительно 92% из более чем 17 млн вариантов были найдены вне зависимости от наличия этапа удаления дубликатов и использованных инструментов[31]. Учитывая, что удаление ПЦР-дубликатов может занимать значительную часть потраченного на обработку данных времени и ресурсов компьютера, следует взвесить пользу и затраты данного этапа для конкретной прикладной задачи.

Рекалибровка качества прочтений (BQSR). Приборная оценка качества оснований не соответствует эмпирической. Первоочередное влияние на эту разницу оказывают цикл секвенирования и нуклеотидный контекст. Решением является рекалибровка качества, исходя из известных паттернов ковариации.

Поиск точечных полиморфизмов.

2.7. Аннотация, фильтрация и интерпретация результатов

Номер экзона, функциональный класс варианта.

Частота аллеля по основным базам данных. Несмотря на то, что были созданы базы данных для всех рас, очень часто этого недостаточно и необходимо учитывать частоты в популяциях отдельных народов и стран. Такими базами данных являются GME[21], в которой отражены частоты по популяции Ближнего Востока, ABraOM[26], предоставляющая частоты вариантов среди практически здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, например, пакистанская[35].

Loss-of-function. Различные показатели, отражающие устойчивость функции гена, основанные на данных о стоп-кодонах, сдвигах рамки считывания и сплайс-вариантах (pLi).

Основные проблемы pLI:

- Плохо приспособлен к распознаванию AR вариантов (из-за того, что частота повреждающих вариантов в популяции может быть высокой) и XR вариантов (из-за наличия в популяции здоровых гетерозиготных носителей).
- Плохо приспособлен к распознаванию вариантов в генах, ответственных за патологии, не влияющие на взросление и воспроизводство. Их частота в популяции также может быть высокой. К таким относятся варианты в генах BRCA1-2.
- Сплайс-варианты рассматриваются как повреждающие, несмотря на то, что вариант в сплайс-сайте может не иметь эффекта на сплайсинг, либо приводить к появлению изоформы белка без потери функции.
- Высокая частота распространения заболевания в контрольной группе. Пример — шизофрения.
- К миссенс-вариантам pLI применять следует с осторожностью, и без функциональной пробы следует исключить из анализа.
- Также следует отнестись с осторожностью к нонсенс-вариантам и сдвигам рамки считывания в последнем экзоне либо в С-терминальной части предпоследнего. Такие транскрипты избегают нонсенс-индуцированного разложения РНК и могут в результате как не привести к каким-либо функциональным изменениям, так и привести к образованию мутантного белка, обладающего меньшей активностью по сравнению с исходным, либо токсичного для клетки.
- В некоторых случаях соотношение pLI с гаплонедостаточностью конкретного гена в принципе сложно объяснить[34].

Таким образом, высокое значение pLI можно считать хорошим показателем LoF, низкое — с осторожностью.

Анализ и предсказание функционального эффекта in silico.

Клинические данные из бд и статей.

Семейный анализ, анализ de novo вариантов.

2.8. Когортный анализ

Помимо того, что когортный анализ необходим для получения информации о частоте аллеля, существует необходимость детекции систематических отклонений покрытия и артефактов выравнивания, связанных с конкретными районами генома и/или особенностями приготовления библиотек. Также анализ нескольких родственных образцов помогает определить зиготность варианта либо импутировать район с недостаточным покрытием.

2.9. Случайные находки

2.10. Exo-C: суть метода

3. Материалы и методы

Данные секвенирования клеточной линии K562 (Hi-C[3], WGS[4][5]) были взяты из публичных источников.

Контроль качества — FastQC[14].

Удаление адаптерных последовательностей производилось с помощью cutadapt[13].

Для картирования был взят геном GRCh37/hg19. Из него были удалены так называемые неканоничные хромосомы, что позволило улучшить качество выравнивания и значительно упростить работу с готовыми данными.

Картирование производилось с помощью инструментов Bowtie2[15] и BWA[16]. BWA показал лучшие показатели; кроме того, он значительно лучше работает с химерными ридами, что немаловажно для метода Exo-C.

Сбор статистики производился с помощью samtools flagstat.

Так как мы использовали данные экзомного секвенирования, а количество образцов у нас было относительно небольшим и мы были заинтересованы в максимально качественной подготовке данных, в пайплайн был включён этап удаления ПЦР-дубликатов. Удаление дубликатов — MarkDuplicates от Picard[17], интегрированный в GATK. Оптимальные показатели скорости MarkDuplicates достигаются при запуске Java с параллелизацией сборщиков мусора и количеством сборщиков мусора равным двум[2].

Рекалибровка qual’ов

Для обучения модели требуются вариации в VCF формате (для человеческого генома - dbSNP >132). Нативная база данных с NCBI требует перепарсинг - другие контиги, а также удаление точек в Ref/Alt. Обжать базу нужно bgzip.

Далее выполняется индексирование (и одновременно проверка на пригодность).

Рекалибровка. В [2] было показано, что оптимальные показатели скорости BaseRecalibrator достигаются, как и в случае с MarkDuplicates, запуском Java с двумя параллельными сборщиками мусора; кроме того, BaseRecalibrator поддаётся внешнему распараллеливанию путём разделения картированных ридов на хромосомные группы. Хромосомные группы формировались вручную для используемой сборки генома, каждая запускалась с помощью bash-скрипта. Нам удалось усовершенствовать данный этап — запуск BaseRecalibrator производился с помощью библиотек Python3 subprocess, а паралле-

лизация осуществлялась библиотекой multiprocessing, таким образом, можно было делить файл с картированными прочтениями по хромосомам и обрабатывать их отдельно, так как multiprocessing автоматически распределяет процессы по имеющимся потокам. Всего для генома GRCh37/hg19 удалось достичь максимально возможное ускорение — в 10 раз (по сравнению с запуском на одном потоке).

Покрытие и обогащение в экзоме оценивалось с помощью скрипта на основе bedtools[18].

Поиск вариантов производился с помощью GATK HaplotypeCaller. Как и в случае с BaseRecalibrator, HaplotypeCaller поддаётся внешнему распараллеливанию[2]. Мы также осуществили параллелизацию с помощью сочетания subprocess и multiprocessing, достигнув 10-12-кратного ускорения по сравнению с запуском на одном потоке.

Аннотация вариантов производилась вначале с помощью инструмента Ensembl VEP[12], затем мы мигрировали на ANNOVAR[11].

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®)[23]
2. Online Mendelian Inheritance in Man (OMIM®)[24]
3. GeneCards®: The Human Gene Database — https://www.genecards.org/
4. ClinVar — https://www.ncbi.nlm.nih.gov/clinvar/
5. dbSNP — https://www.ncbi.nlm.nih.gov/snp/
6. Genome Aggregation Database (gnomAD)[22]
7. 1000 Genomes Project — https://www.internationalgenome.org/
8. Great Middle East allele frequencies (GME)[21]
9. dbNSFP: Exome Predictions[20]
10. dbscSNV: Splice site prediction[25]
11. RegSNPIntron: intronic SNVs prediction[19]

Интерпретация данных и составление отчёта производилось в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики (ACMG) и Ассоциации молекулярной патологии[32].

Пограничным значением pLI было взято 0.9, согласно рекомендациям в оригинальной статье[33].

4. Результаты

5. Обсуждение результатов

6. Предварительные выводы

A. Data Accession Codes

Name	Article	Type	Reads, M	Source	Accession Codes
GSM1551618_HIC069	Rao et al. [3]	Hi-C	456.8	GEO	SRR1658693
GSM1551619_HIC070	Rao et al. [3]	Hi-C	591.9	GEO	SRR1658694
GSM1551620_HIC071	Rao et al. [3]	Hi-C	79.9	GEO	SRR1658695 SRR1658696
GSM1551621_HIC072	Rao et al. [3]	Hi-C	79.6	GEO	SRR1658697 SRR1658698
GSM1551622_HIC073	Rao et al. [3]	Hi-C	77.4	GEO	SRR1658699 SRR1658700
GSM1551623_HIC074	Rao et al. [3]	Hi-C	80.8	GEO	SRR1658702 SRR1658701
ENCSR025GPQ	Zhou et al. [4]	WGS	130.0	ENCODE	ENCFF574YLG ENCFF921AXL ENCFF590SSX
ENCSR053AXS	Zhou et al. [4]	WGS	796.2	ENCODE	ENCFF004THU ENCFF066GQD ENCFF313MGL ENCFF506TKC ENCFF080MQF
ENCSR711UNY	Zhou et al. [4]	WGS	449.7	ENCODE	ENCFF471WSA ENCFF826SYZ ENCFF590SSX

Невыравненные:

Name	Article	Type	Reads, M	Source	Accession Codes
SRX3358201	Dixon et al. [5]	WGS	366.3	GEO	SRR6251264
GSE148362_G1	Wang et al. [6]	Repli-seq	24.8	GEO	SRR11518301
GSE148362_S1	Wang et al. [6]	Repli-seq	30.9	GEO	SRR11518302
GSE148362_S2	Wang et al. [6]	Repli-seq	45.4	GEO	SRR11518303
GSE148362_S3	Wang et al. [6]	Repli-seq	49.8	GEO	SRR11518304
GSE148362_S4	Wang et al. [6]	Repli-seq	44.1	GEO	SRR11518305
GSE148362_S5	Wang et al. [6]	Repli-seq	38.4	GEO	SRR11518306
GSE148362_S6	Wang et al. [6]	Repli-seq	35.2	GEO	SRR11518307
GSE148362_G2	Wang et al. [6]	Repli-seq	33.0	GEO	SRR11518308
INSITU_HS1	Ray et al. [7]	Hi-C	86.3	GEO	SRR9019504
INSITU_HS2	Ray et al. [7]	Hi-C	127.1	GEO	SRR9019505
INSITU_NHS1	Ray et al. [7]	Hi-C	86.4	GEO	SRR9019506
INSITU_NHS2	Ray et al. [7]	Hi-C	128.5	GEO	SRR9019507
PD_STABLE_REP1	Moquin et al. [8]	Hi-C	67.2	GEO	SRR5470535 SRR5470534
PD_STABLE_REP2	Moquin et al. [8]	Hi-C	52.9	GEO	SRR5470536 SRR5470537
PD_TRANSIENT	Moquin et al. [8]	Hi-C	81.3	GEO	SRR5470539 SRR5470538
PDDE_TRANSIENT	Moquin et al. [8]	Hi-C	55.2	GEO	SRR5470541 SRR5470540
GSM2588815_R1	Belaghzal et al. [9]	Hi-C	72.9	GEO	SRR5479813
GSM2536769_WT	Banaszak et al. [10]	WES[1]	39.2	GEO	SRR5345331
GSM2536770_WT_TF	Banaszak et al. [10]	WES[1]	49.4	GEO	SRR5345332
GSM2536771_MT2	Banaszak et al. [10]	WES[1]	42.0	GEO	SRR5345333
GSM2536772_MT3	Banaszak et al. [10]	WES[1]	43.7	GEO	SRR5345334
GSM2536773_MT4	Banaszak et al. [10]	WES[1]	39.9	GEO	SRR5345335
GSM2536774_MT5	Banaszak et al. [10]	WES[1]	40.8	GEO	SRR5345336

[1]Варианты в гене DNMT3A были исключены из выборки.

Список литературы

[1] Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

[2] Heldenbrand JR, Baheti S, Bockol MA, et al. Recommendations for performance optimizations when using GATK3.8 and GATK4 [published correction appears in BMC Bioinformatics. 2019 Dec 17;20(1):722]. BMC Bioinformatics. 2019;20(1):557. Published 2019 Nov 8. doi:10.1186/s12859-019-3169-7

[3] Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping [published correction appears in Cell. 2015 Jul 30;162(3):687-8]. Cell. 2014;159(7):1665-1680. doi:10.1016/j.cell.2014.11.021

[4] Zhou B, Ho SS, Greer SU, et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. Genome Res. 2019;29(3):472-484. doi:10.1101/gr.234948.118

[5] Dixon JR, Xu J, Dileep V, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018;50(10):1388-1398. doi:10.1038/s41588-018-0195-8

[6] Yuchuan Wang, Yang Zhang, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. bioRxiv 2020.03.09.982967; doi: https://doi.org/10.1101/2020.03.09.982967

[7] Ray J, Munn PR, Vihervaara A, et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. Proc Natl Acad Sci U S A. 2019;116(39):19431-19439. doi:10.1073/pnas.1901244116

[8] Moquin SA, Thomas S, Whalen S, et al. The Epstein-Barr Virus Episome Maneuvers between Nuclear Chromatin Compartments during Reactivation. J Virol. 2018;92(3):e01413-17. Published 2018 Jan 17. doi:10.1128/JVI.01413-17

[9] Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods. 2017;123:56-65. doi:10.1016/j.ymeth.2017.04.004

[10] Banaszak LG, Giudice V, Zhao X, et al. Abnormal RNA splicing and genomic instability after induction of DNMT3A mutations by CRISPR/Cas9 gene editing. Blood Cells Mol Dis. 2018;69:10-22. doi:10.1016/j.bcmd.2017.12.002

[11] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. doi:10.1093/nar/gkq603

[12] McLaren, W., Gil, L., Hunt, S.E. et al. The Ensembl Variant Effect Predictor. Genome Biol 17, 122 (2016). doi: 10.1186/s13059-016-0974-4

[13] MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200.

[14] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[15] Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). doi: 10.1038/nmeth.1923

[16] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324

[17] ”Picard Toolkit.” 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute

[18] Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

[19] Lin, H., Hargreaves, K.A., Li, R. et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. Genome Biol 20, 254 (2019). doi:10.1186/s13059-019-1847-4

[20] Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235-241. doi:10.1002/humu.22932

[21] Scott EM, Halees A, Itan Y, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat Genet. 2016;48(9):1071-1076. doi:10.1038/ng.3592

[22] Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). doi:10.1038/s41586-020-2308-7

[23] Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136(6):665-677. doi:10.1007/s00439-017-1779-6

[24] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789-D798. doi:10.1093/nar/gku1205

[25] Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet Med. 2014;16(7):497-503. doi:10.1038/gim.2013.176

[26] Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, Bozoklian D, Sandberg TOM, Brito LA, Lazar M, Bernardo DV, Amaro E Jr, Duarte YAO, Lebrão ML, Passos-Bueno MR, Zatz M. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. Hum Mutat. 2017 Jul;38(7):751-763. doi: 10.1002/humu.23220.

[27] Yohe S, Thyagarajan B. Review of Clinical Next-Generation Sequencing. Arch Pathol Lab Med. 2017 Nov;141(11):1544-1557. doi: 10.5858/arpa.2016-0501-RA

[28] Balloux F, Brønstad Brynildsrud O, van Dorp L, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. Trends Microbiol. 2018;26(12):1035-1048. doi:10.1016/j.tim.2018.08.004

[29] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463

[30] Burrows M, Wheeler DJ. Technical report 124. Palo Alto, CA: Digital Equipment Corporation; 1994. A block-sorting lossless data compression algorithm.

[31] Ebbert MT, Wadsworth ME, Staley LA, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016;17 Suppl 7(Suppl 7):239. Published 2016 Jul 25. doi:10.1186/s12859-016-1097-3

[32] Richards S, Aziz N, Bale S, et al. 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. Genet Med. 2015;17(5):405-424. doi:10.1038/gim.2015.30

[33] Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-291. doi:10.1038/nature19057

[34] Ziegler A, Colin E, Goudenège D, Bonneau D. A snapshot of some pLI score pitfalls. Hum Mutat. 2019 Jul;40(7):839-841. doi: 10.1002/humu.23763

[35] Saleheen D, Natarajan P, Armean IM, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017;544(7649):235-239. doi:10.1038/nature22034

[36] Wutz G, Várnai C, Nagasaka K, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. EMBO J. 2017;36(24):3573-3599. doi:10.15252/embj.201798004

[37] Yu D, Zhang K, Han M, et al. Noninvasive prenatal testing for fetal subchromosomal copy number variations and chromosomal aneuploidy by low-pass whole-genome sequencing. Mol Genet Genomic Med. 2019;7(6):e674. doi:10.1002/mgg3.674

[38] Herder M. What Is the Purpose of the Orphan Drug Act?. PLoS Med. 2017;14(1):e1002191. Published 2017 Jan 3. doi:10.1371/journal.pmed.1002191

[39] Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, Hughes DA; International Society for Pharmacoeconomics and Outcomes Research Rare Disease Special Interest Group. Rare Disease Terminology and Definitions-A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. Value Health. 2015 Sep;18(6):906-14. doi: 10.1016/j.jval.2015.05.008

[40] The Lancet Neurology. Rare neurological diseases: a united approach is needed. Lancet Neurol. 2011 Feb;10(2):109. doi: 10.1016/S1474-4422(11)70001-1. Erratum in: Lancet Neurol. 2011 Mar;10(3):205.

[41] D. Huber, L. Voith von Voithenberg, G.V. Kaigala. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? Micro and Nano Engineering, Volume 1, 2018, Pages 15-24, ISSN 2590-0072. doi: 10.1016/j.mne.2018.10.006.

[42] Theisen, A. (2008) Microarray-based Comparative Genomic Hybridization (aCGH). Nature Education 1(1):45