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

Генетическое детерминирование патологий
Частые и редкие (орфанные) патологии

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 SNP 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[28]. Варианты, приходящиеся на сайты связывания CTCF, могут разрушать границы ТАДов и вызывать изменения экспрессии.

2.4. Методы детектирования

Кариотипирование.

CGH.

FISH.

STS.

MLPA.

Секвенирование по Сэнгеру. Метод, позволяющий с высокой точностью анализировать короткий (до 1kb) фрагмент ДНК[20]. 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.

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). WGS также может быть использовано в диагностике микробиома с целью определения источника хронической инфекции, реконструкции путей передачи инфекции, а также выявления антибиотикорезистентных штаммов[1].

Полноэкзомное секвенирование (WES)

Таргетные панели

Hi-C

2.6. Базовая схема обработки результатов секвенирования

Удаление адаптерных последовательностей. В процессе секвенирования к целевым фрагментам ДНК могут пришиваться так называемые адаптерные последовательности. Если целевая ДНК короче длины прочтения, то фрагменты адаптера 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[12].

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[23]). В настоящее время «золотым стандартом» являются утилиты, использующие алгоритм Берроуса–Уиллера[21].

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[22]. 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[5], в которой отражены частоты по популяции Ближнего Востока, ABraOM[25], предоставляющая частоты вариантов среди практически здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, например, пакистанская[27].

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[26].

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[18], WGS[19]) были взяты из публичных источников

Контроль качества — FastQC[13].

Удаление адаптерных последовательностей производилось с помощью cutadapt[12].

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[14] и BWA[15]. 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[16], 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 параллелизация 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[17].

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[11], затем мы мигрировали на ANNOVAR[10].

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®)[7]
2. Online Mendelian Inheritance in Man (OMIM®)[8]
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)[6]
7. 1000 Genomes Project — https://www.internationalgenome.org/
8. Great Middle East allele frequencies (GME)[5]
9. dbNSFP: Exome Predictions[4]
10. dbscSNV: Splice site prediction[9]
11. RegSNPIntron: intronic SNVs prediction[3]

Интерпретация данных и составление отчёта производилось в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики (ACMG) и Ассоциации молекулярной патологии[23].

Пограничным значением pLI было взято 0.9, согласно рекомендациям в оригинальной статье[24].

4. Результаты

5. Обсуждение результатов

6. Предварительные выводы

Список литературы

[1] 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

[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] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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.

[13] 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/

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

[15] 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

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

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

[18] 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

[19] 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

[20] 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

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

[22] 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

[23] 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

[24] 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

[25] 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.

[26] 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

[27] 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

[28] 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