

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ОБРАБОТКИ РАЗЛИЧНЫХ ТИПОВ ВАРИАЦИЙ ДНК ЧЕЛОВЕКА

М. С. Галкин, М. Д. Лабукин

Белорусский государственный университет, г. Минск;
mefa2411@gmail.com, mike_labutin@yahoo.com;
науч. рук. – В. В. Скакун, канд. физ.-мат. наук, доц;
С. В. Гилевский, канд. техн. наук, доц.

Разработана модель данных и серверное приложение для хранения, структуризации и отображения в удобном для человека виде данных об однонуклеотидных полиморфизмах ДНК пациентов и связанных с ними диагнозов для более удобной работы с этими данными. Автоматизированы и собраны в единый инструмент подготовка данных с помощью Genome Analysis Toolkit (GATK) и поиск вариаций числа копий генов с помощью инструмента eXome-Hidden Markov Model (XHMM). Обе системы реализованы как серверные приложения, что упрощает пользователю работу с ними.

Ключевые слова: однонуклеотидный полиморфизм; вариации числа копий генов; реляционные базы данных; MySQL; variant call format; XHMM; GATK.

ОБРАБОТКА ОДНОНУКЛЕОТИДНЫХ ПОЛИМОРФИЗМОВ

Данные об однонуклеотидных полиморфизмах поступают в файлах Variant Call Format (VCF), которые представляют из себя текстовые файлы с документированной структурой, содержащей метаданные (пара символов «##») и строки, содержащие информацию о полиморфизмах, разделённые на столбцы. Каждому полиморфизму соответствует отдельная строка (рис. 1).

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB GT:GQ:DP:HQ 0|0:48:1:51,51
```

Рис. 1. Скриншот сокращённого для примера VCF файла

В VCF файлах может храниться более 500 000 строк. А общее количество полиморфизмов даже для десятка пациентов будет достигать нескольких миллионов. При этом при использовании таких файлов крайне сложно соотносить полиморфизмы с диагнозами, а также производить какую-либо сортировку данных, кроме как по пациентам. Поэтому был получен запрос на разработку системы, которая бы упростила работу с данными о полиморфизмах как исследователям, так и врачам.

Было решено начать разработку базы данных для хранения полиморфизмов и их связей с пациентами и диагнозами, а также серверное приложение для работы с ними. Использование базы данных позволит эффективно работать с данными, что критично для больших объемов данных. Было решено использовать реляционную базу данных, так как предметная область, по своей сути – несколько отношений (таблиц), связанных друг с другом. Так, тело variant call format схоже с отношением в реляционной модели данных, а метаданные поясняют то, что находится в теле файла. Из реляционных систем управления базами данных пал на MySQL. Модель данных на рис. 2.

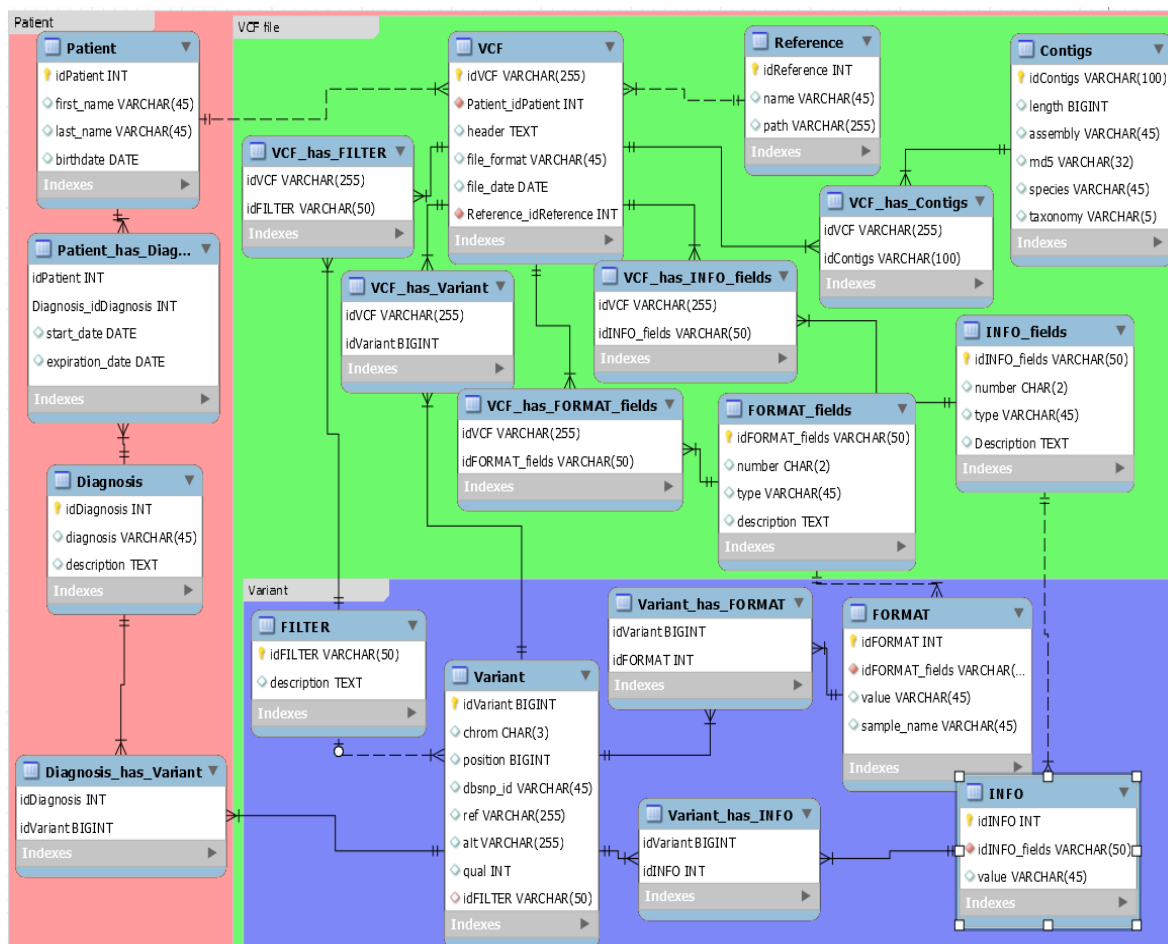


Рис. 2. Модель данных для VCF файла

Для доступа к базе данных и обработки (например, импорта файлов в базу данных) стоял выбор между отдельным приложением для персонального компьютера и серверным приложением с интерфейсом в браузере. Выбор пал на серверное приложение. Преимущество этого подхода в максимальной простоте – пользователю не нужно ничего устанавливать и настраивать, а доступ можно получить с любого устройства, на котором установлен браузер и которое имеет доступ к серверу.

АВТОМАТИЗАЦИЯ ПОИСКА ВАРИАЦИЙ ЧИСЛА КОПИЙ ГЕНОВ

Одним из типов полиморфизмов является вариация числа копий генов (далее – CNV). Этот тип полиморфизмов напрямую влияет на экспрессию гена, который он затронул. Анализируя изменение числа копий генов пациента по отношению к референсному геному можно получить более полную информацию о его заболевании, лучше предсказать течение болезни и её лечение.

Для поиска CNV использовался инструмент eXome-Hidden Markov Model (XHMM). Для обнаружения с его помощью CNV необходимо выровнять полученные фрагменты ДНК пациента на референсный геном, получив в результате BAM файл. Затем определить покрытие генома. Затем полученные данные необходимо собрать в один файл, отфильтровать, использовать метод главных компонент и нормализовать итоговые данные. И только после этого можно детектировать сами CNV. Промежуточные данные вместе с итоговыми DATA.vcf, DATA.xcnv и DATA.aux-xcnv на рис. 3.

Имя файла	Размер	Тип файла	Последнее изменение
..			
group1.DATA.sample_summary	244 B	sample_s...	05/14/19 04:38:23
group1.DATA.sample_statistics	6.9 KB	sample_st...	05/14/19 04:38:23
group1.DATA.sample_interval_summary	25.6 MB	sample_in...	05/14/19 04:38:23
group1.DATA.sample_interval_statistics	5.4 KB	sample_in...	05/14/19 04:38:23
DATA.xcnv	15.8 KB	xcnv-файл	05/14/19 04:39:09
DATA.vcf	30.2 KB	vCard	05/14/19 04:39:33
DATA.same_filtered.RD.txt	12.5 MB	txt-файл	05/14/19 04:38:36
DATA.posteriors.DUP.txt	7.8 MB	txt-файл	05/14/19 04:39:10
DATA.posteriors.DIP.txt	7.8 MB	txt-файл	05/14/19 04:39:09
DATA.posteriors.DEL.txt	7.8 MB	txt-файл	05/14/19 04:39:10
DATA.filtered_centered.RD.txt.filtered_targets.txt	62.1 KB	txt-файл	05/14/19 04:38:29
DATA.filtered_centered.RD.txt.filtered_samples.txt	0 B	txt-файл	05/14/19 04:38:29
DATA.filtered_centered.RD.txt	12.3 MB	txt-файл	05/14/19 04:38:29
DATA.aux_xcnv	1.5 MB	aux_xcnv-...	05/14/19 04:39:09
DATA.RD_PCA.PC_SD.txt	51 B	txt-файл	05/14/19 04:38:30
DATA.RD_PCA.PC_LOADINGS.txt	109 B	txt-файл	05/14/19 04:38:30
DATA.RD_PCA.PC.txt	10.8 MB	txt-файл	05/14/19 04:38:30

Рис. 3. Промежуточные и итоговые файлы

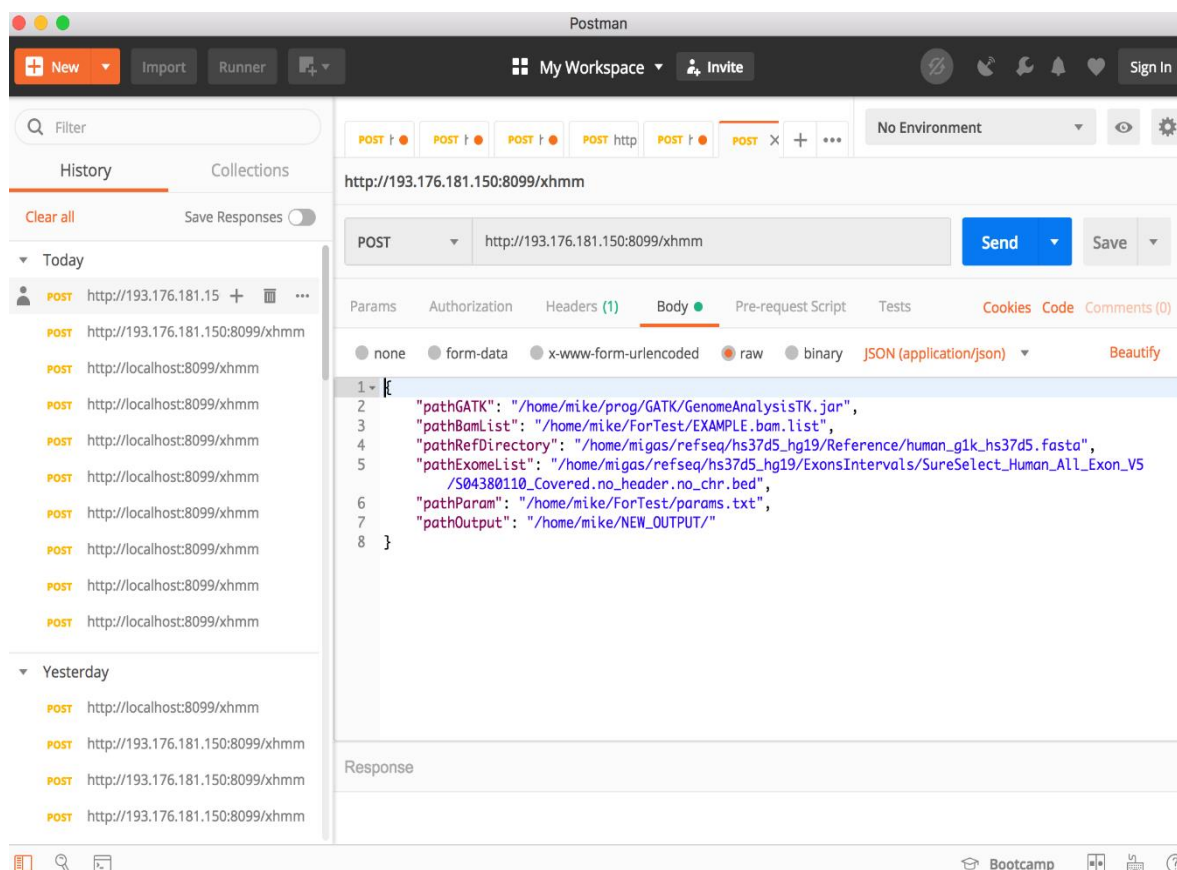


Рис. 4. Передача параметров приложению

Таким образом, для получения результата необходимо проделать множество различных шагов, большинство из которых – очень ресурсоёмкие и занимают много времени. Однако все эти шаги легко автоматизировать, указав в начале исходные файлы и желаемые параметры. Для этого было разработано приложение, сочетающее в себя вызовы различных инструментов, bash команд и скриптов. Приложение работает на сервере, ожидая запроса, содержащего параметры для старта обработки (рис. 4).

Таким образом, вместо последовательного вызова множества инструментов из консоли (как правило сервера) и ожидания выполнения каждой из них пользователю достаточно послать приложению один запрос с конфигурацией и получить окончательный результат. С помощью таких запросов приложение поддерживает множество пользователей.

Библиографические ссылки

1. Селко Дж. Стиль программирования Джо Селко на SQL. СПб., 2006.
2. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives / Min Zhao [et al.] // BMC Bioinformatics. 2013. № 14. Art. №: S1.