Метод анализа наследственных диспергированных и центромерных повторов с использованием данных полногеномного секвенирования и его применение на примере семейных трио пациентов с шизофренией

М.С. Протасова [1,2], Ф.Е. Гусев [1,2], Т.В. Андреева [1,2], А.П. Григоренко[1,2], Е.И. Рогаев [1,2,3].

[1]ИОГен РАН, Москва
[2]Центр нейробиологии и нейрогенетики мозга, ИЦиГ СО РАН, Новосибирск
[3]UMASS, Вустер, США.

Цель

Разработать метод анализа вставок наследственных диспергированных и центромерных повторов, отличных от референсного генома, в данных полногеномного секвенирования.

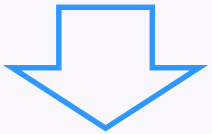
Введение

Вариативность генома:

- SNP, Инделы
- CNV
- Структурные варианты
  - инверсии
  - **инсерции/делеции**

Одна из причин - **мобильные элементы** генома

В геноме человека > 50% составляют повторы



- Влияют на экспрессию генов
- Могут вызывать заболевания

Классы диспергированных повторов и механизмы их интеграции в новые геномные локусы

Количество повторов в геноме человека

>500 000 копий

LINE

6 тпн

SINE

>1,1 млн копий

Alu

300 пн

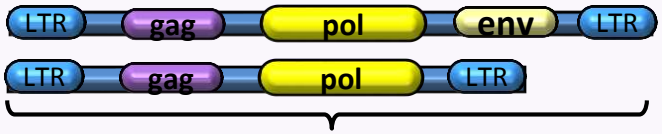
SVA

>2 000 копий

2 тпн

LTR

>280 000 копий

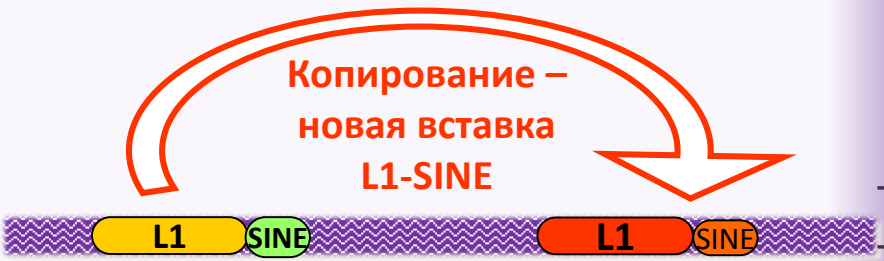


5-7 тпн

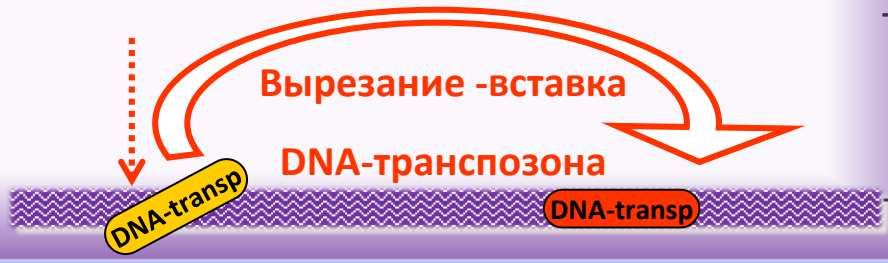
DNA транспозоны

>100 000 копий

2-3 тпн



ретротранспозиция



транспозиция

Сложности и ограничения, возникающие при использовании существующих методов поиска повторов

- Время анализа.
- Вычислительные мощности кластера при анализе геномных данных с высоким покрытием.
- Установка дополнительного программного обеспечения, а также совместимость версий программ, не входящих в основной программный пакет.
- Приведение исходных данных в определенный формат, необходимость использования определённой программы для выравнивания чтений, в том числе повторное выравнивание генома.

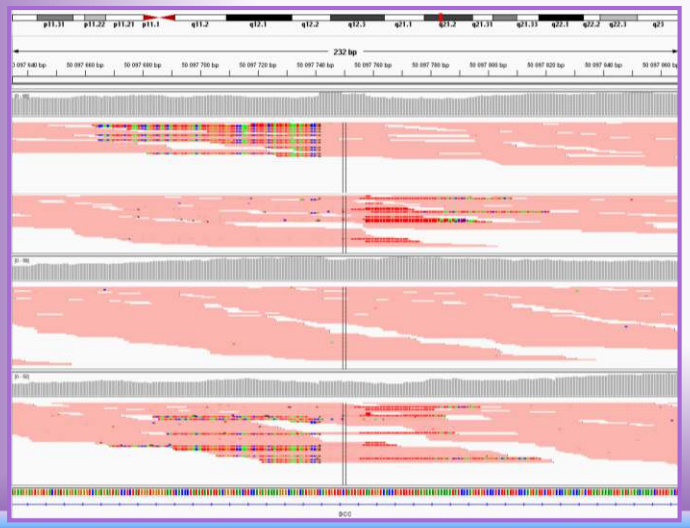
MOSAIK, Spanner, RetroSeq, TEA

Анализ **наследственных** диспергированных и центромерных повторов с использованием данных **полногеномного секвенирования**

Два семейных трио

Полногеномное секвенирование Illumina HiSeq 2000.

Анализ NGS данных.
Определение инсерций повторов.



Анализируемые классы:

- **LINE:**
  - **LINE1**
  - **LINE2**
  - **L3/CR1**
- **SINE**
  - **Alu**
  - **MIR**
- **LTR:**
  - **ERVL**
  - **ERVL-MALRs**
  - **ERV class1**
  - **ERV class2**
- **DNA-transposon:**
  - **hAT-Charlie**
  - **TcMar-Tigger**
- **Satellites**

Метод включает в себя использование следующих биоинформатических программ:

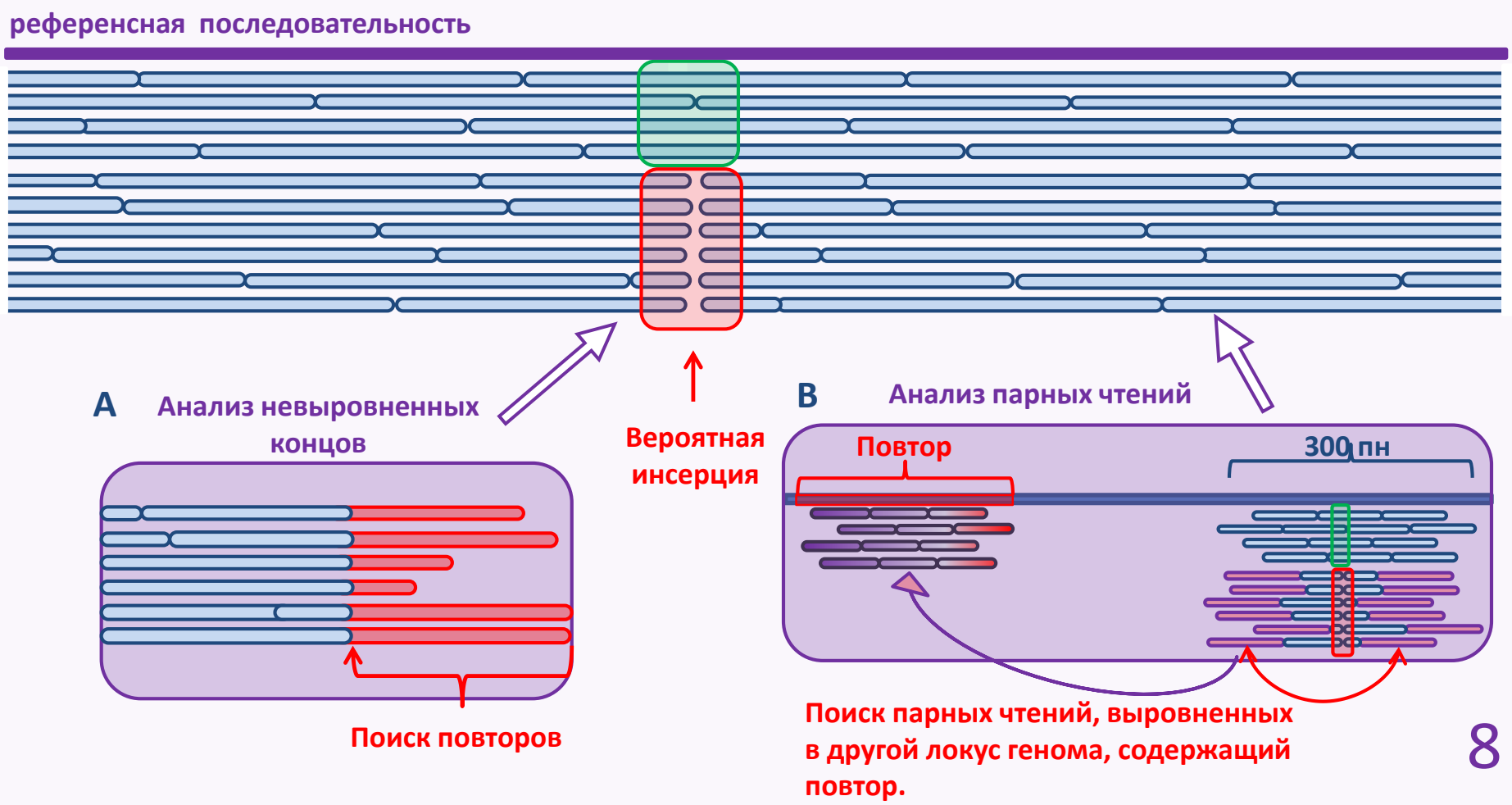
- BWA – выравнивание генома
- Samtools – операции с .bam файлом генома
- RepeatMasker – обнаружение повторов
- Bedtools – дальнейшие операции с полученными позициями инсерций/делеций

- Скрипты, написанные на языке Perl

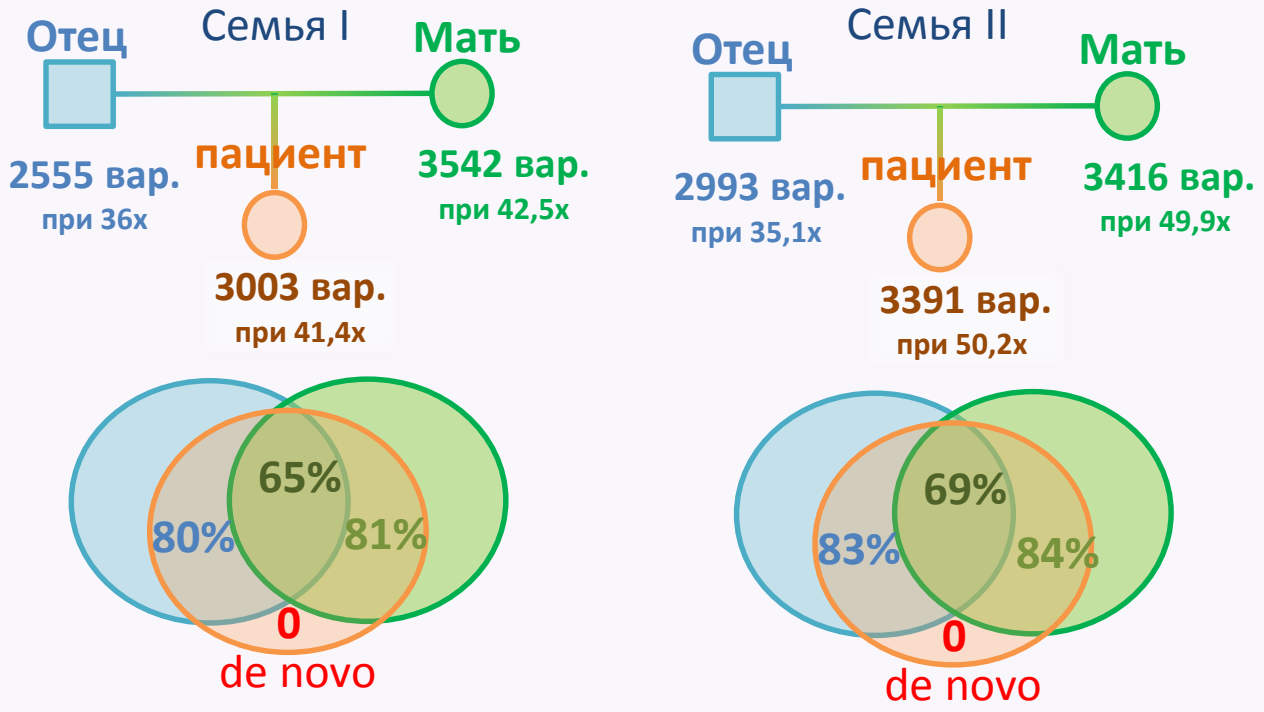
Отбор позиций вставок повторов в данных полногеномного секвенирования семейных трио

1. Выравнивание геномов семейных трио программой BWA на референсный геном человека GRCh37

2. Определение вставок повторов в локусы, отличные от референсного генома, двумя способами (A,B)



Наследуемые вставки повторов в семейных трио



Характеристика локусов вставок

В уникальной области генома - 22,5% от всех вариантов

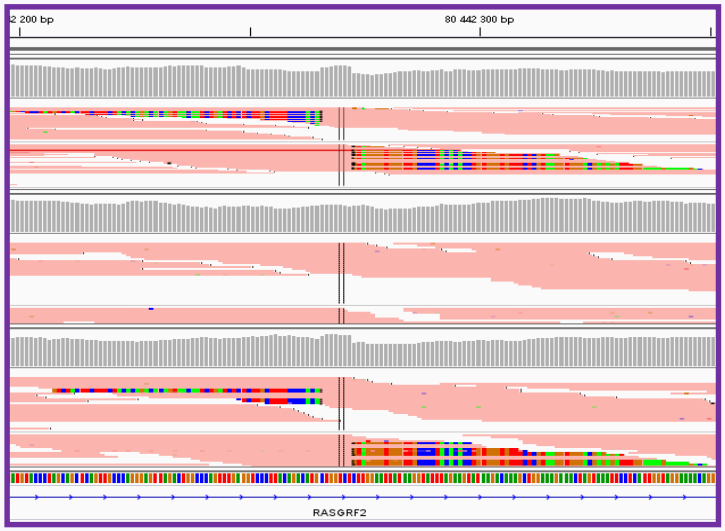
В экзонах генов (гл.обр. UTR3') 2,6% от уникальных вариантов (всего обнаружено 49 позиций вар. во всех анализ. геномах)

46% (25) из вар. в экзонах присутствуют в 1000 Genomes
Из них:
с частотой 0,25-1: **20%**
с частотой 0,02-0,25: **36%**
с частотой ≤0,02: 44%

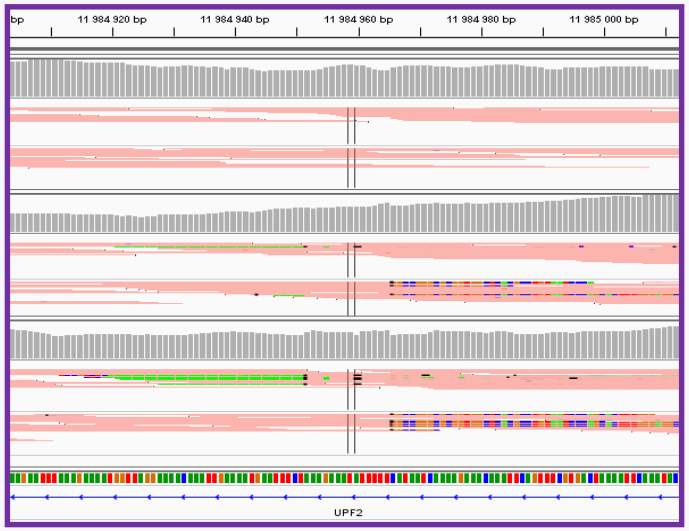
54% из вар. в экзонах отсутствуют в базе 1000 Геномов

Валидация обнаруженных редких полиморфных вставок/делеций ретроэлементов

Мать отец Пациент

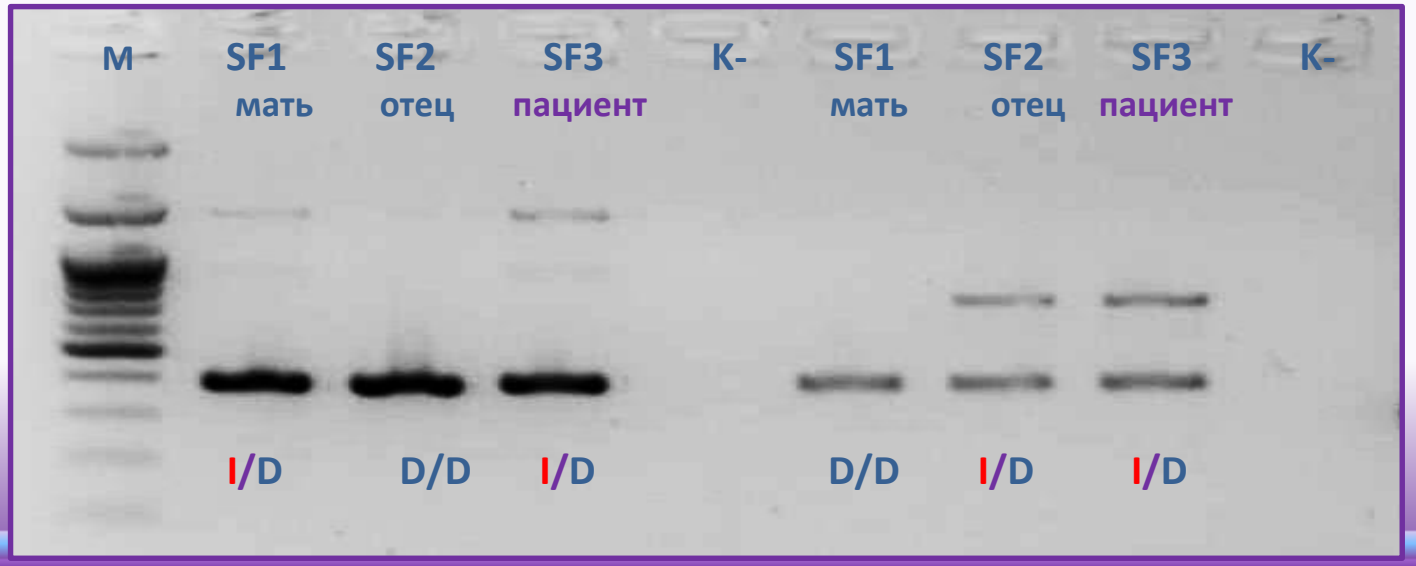


LTR5 Hs в интроне 17-18 гена RASGRF2



AluYa5 в интроне 17-18 гена UPF2

Мать отец Пациент



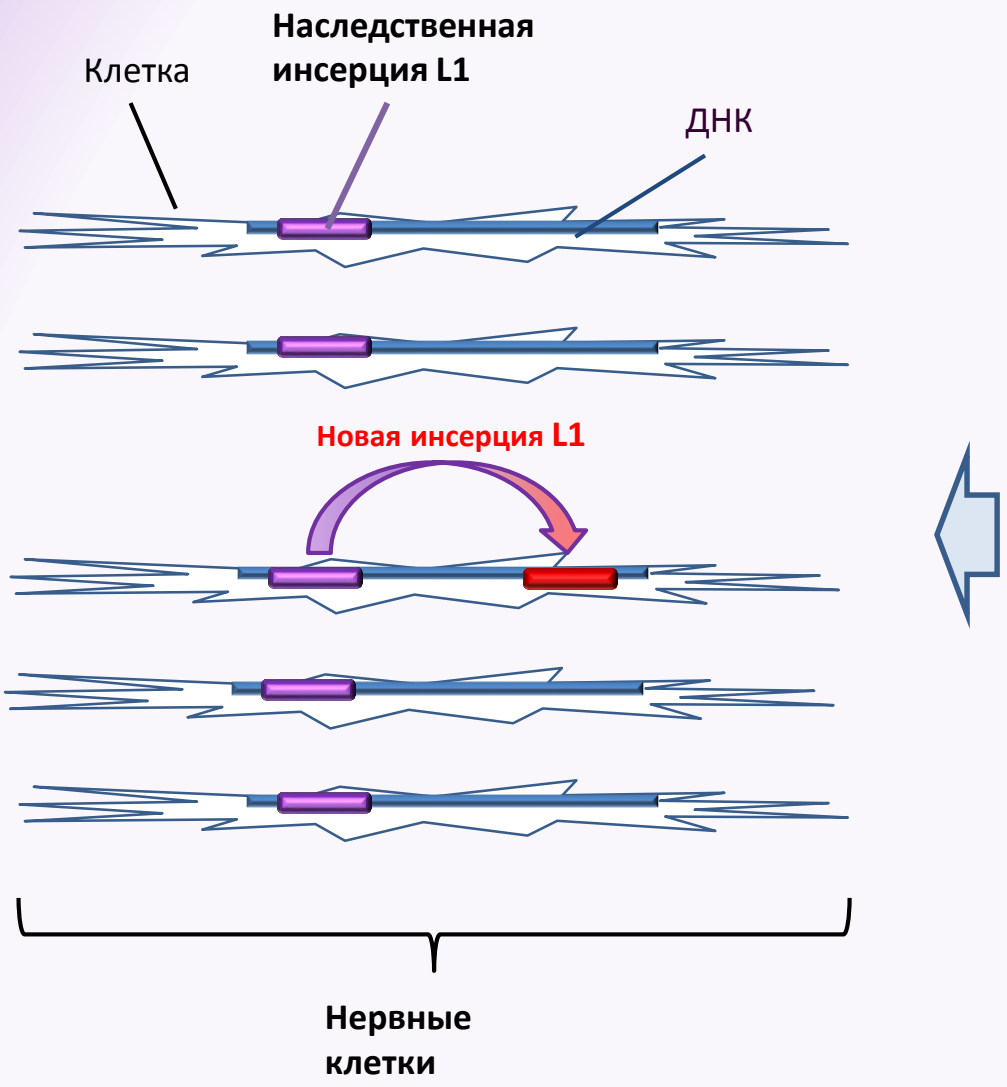
Особенности метода :

- Данный метод адаптирован для исходных геномных данных в формате .bam, полученном после стандартного выравнивания при помощи BWA. Дополнительного выравнивания генома не требуется. Таким образом происходит сокращение времени анализа и вычислительных ресурсов.

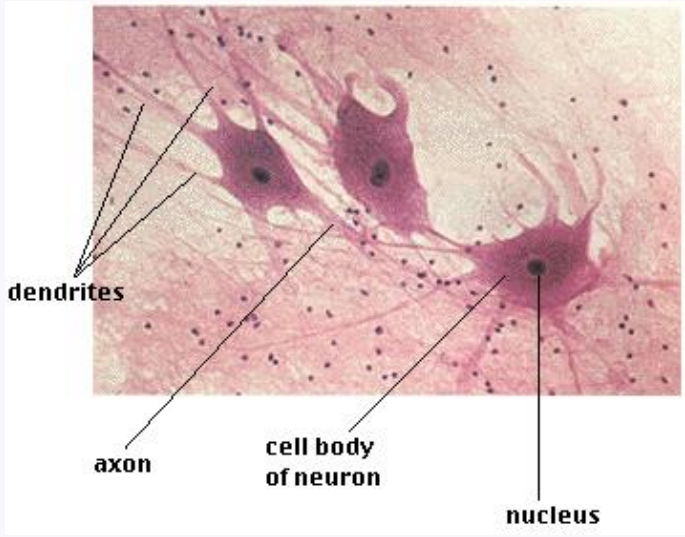
- Благодаря тому, что анализ повторов происходит только в чтениях, потенциально содержащих вставки повторов, происходит сокращение времени на этапе поиска повторов в невыровненных фрагментах чтений с помощью Repeat Masker.

- В процессе ретротранспозиции часто происходит копирование и вставка группы повторов. Метод позволяет определить, какими семействами повторов представлена данная группа ретротранспозиции.

За счет активности мобильных элементов происходит формирование соматической вариабельности



В гиппокампе на 1000 копий L1 больше чем в сердце или печени

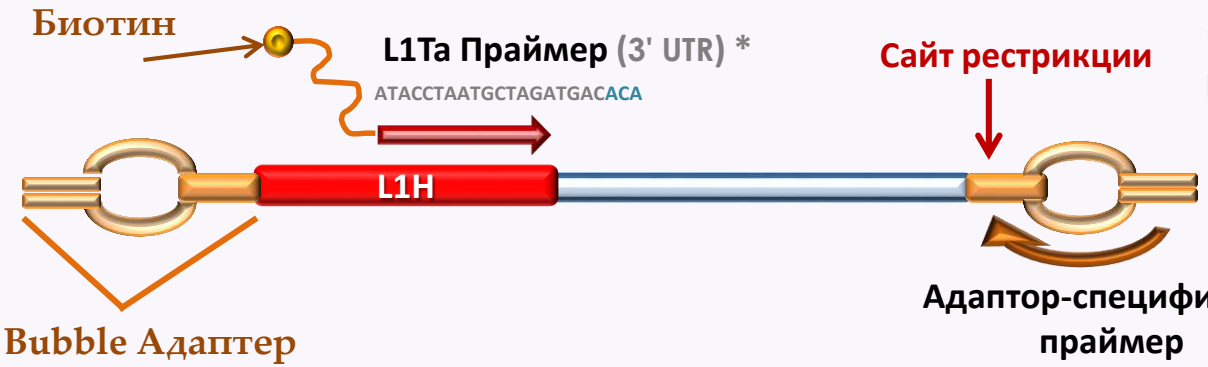


L1 вызывают мутации, изменяют экспрессию генов, могут приводить к заболеваниям

Метод анализа соматических инсерций L1Hs элементов

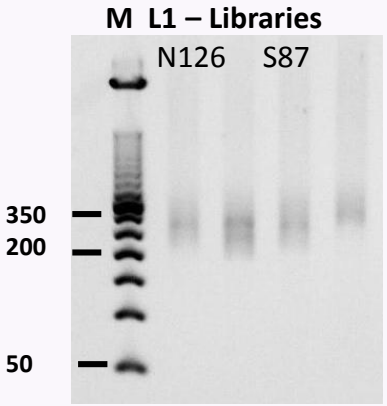
1. Обогащение сайтов L1 инсерций

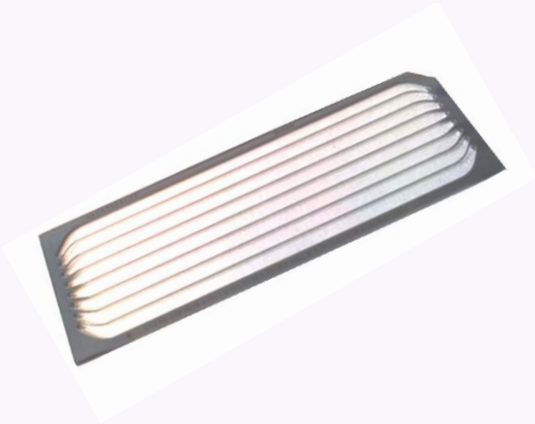
Геномная ДНК → Ультразвуковая Фрагментация → Лигирование Bubble-Адапторов → L1-специфичная Bubble PCR



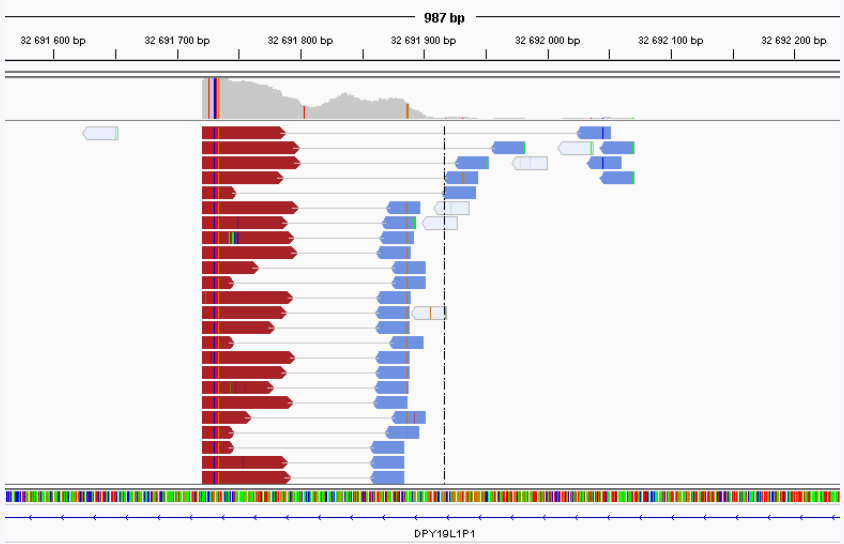
*L1Ta из статьи Iskow RC, et al., 2010

2. Приготовление библиотек и параллельное масштабное секвенирование

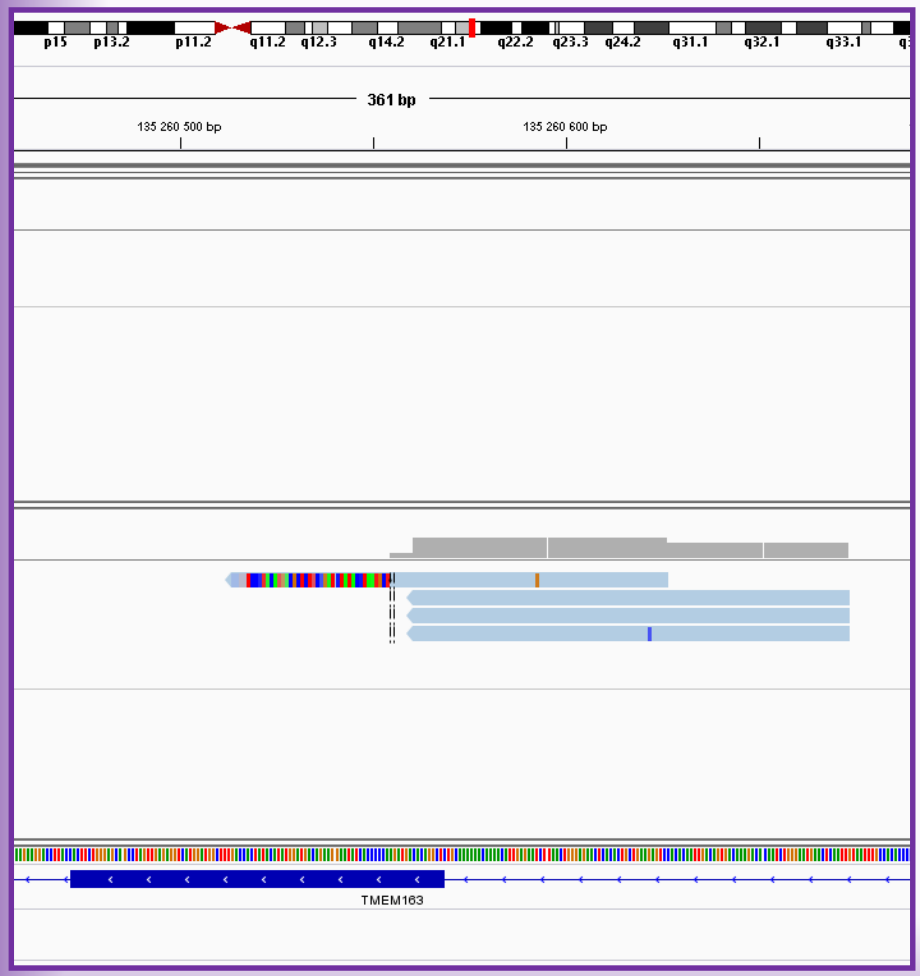




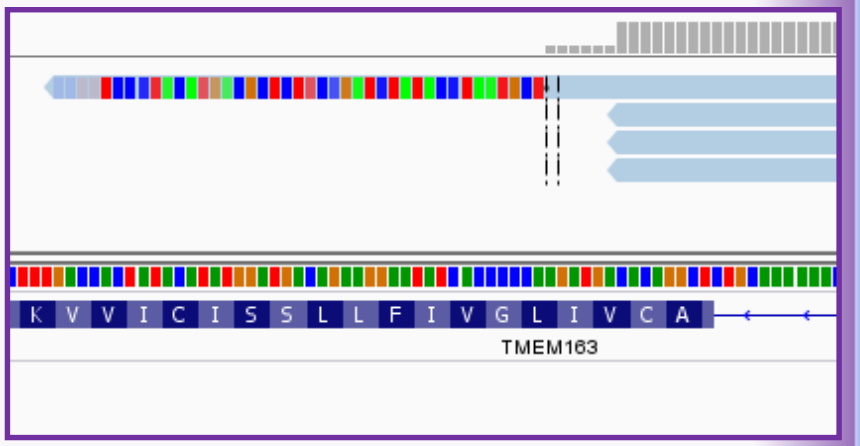
3. Биоинформатический анализ



Пример соматической вставки в CDS



Соматическая вставка в ядрах нейронов:
Экзон 5 гена *TMEM 163*
transmembrane protein 163



Выводы:

- Разработанный метод удобен в работе с данными полногеномного секвенирования с высоким покрытием и позволяет выявлять врожденные полиморфные вставки.
- С помощью данного метода проведен анализ вставок различных классов повторов в данных полногеномного секвенирования образцов ДНК двух семейных трио пациентов с шизофренией.
- Выявлены множественные наследуемые вариации в геномах пациентов с шизофренией и их родителей, обусловленные вставками ретротранспозонов.
- Показано отсутствие *de novo* вставок ретротранспозонов у детей в исследованных семьях.
- Разработан биоинформационный подход для детекции редких соматических мутационных инсерций в данных секвенирования библиотек, обогащенных L1 локусами генома человека. Показано, что соматические ретротранспозиции в кодирующие участки генов в клетках мозга являются редкими событиями, и большинство ретропоцизий происходит в некодирующие участки генома.

Благодарности

- *Научному руководителю Рогаеву Е.И.*

- *Сотрудникам лаборатории эволюционной геномики ИОГен РАН: Андреевой Т.В., Григоренко А.П., Гусеву Ф.Е., Тяжеловой Т.В., Гольцову А.Ю., Гейко А., Кузнецовой И.Л., Золотаревой О., Лисенковой А., Алисейчик М., Манахову А.*

Благодарю за внимание