Статистический подход для детекции крупномасштабных вариаций числа копий в данных NGS, полученных с использованием мультиплексной ПЦР

Демидов Герман Михайлович

научный руководитель: Брагин А.Г., к.б.н.

СПб АУ НОЦНТ РАН

16 июня 2015 г.

Варианты в геноме

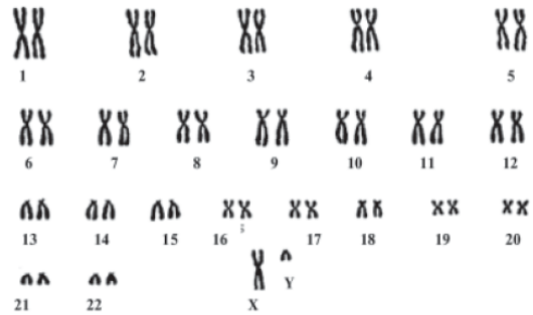
Генетические варианты бывают - 1) точечные мутации и 2) вставки/удаления хромосомных сегментов. Вставки и удаления бывают:

- короткие (до 300 пар оснований);
- крупномасштабные, CNV, ≥ 300 пар оснований.

Как точечные мутации, так и вставки и удаления могут приводить к развитию генетически обусловленных заболеваний.

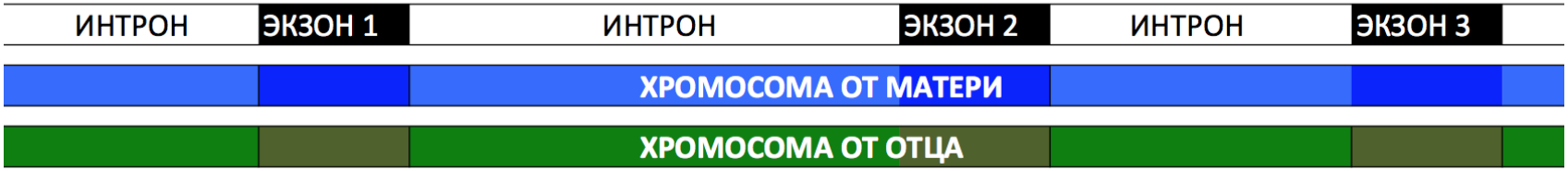
Схема структуры гена

В норме у человека 2 варианта генов – один вариант от матери, другой от отца.



ХРОМОСОМЫ

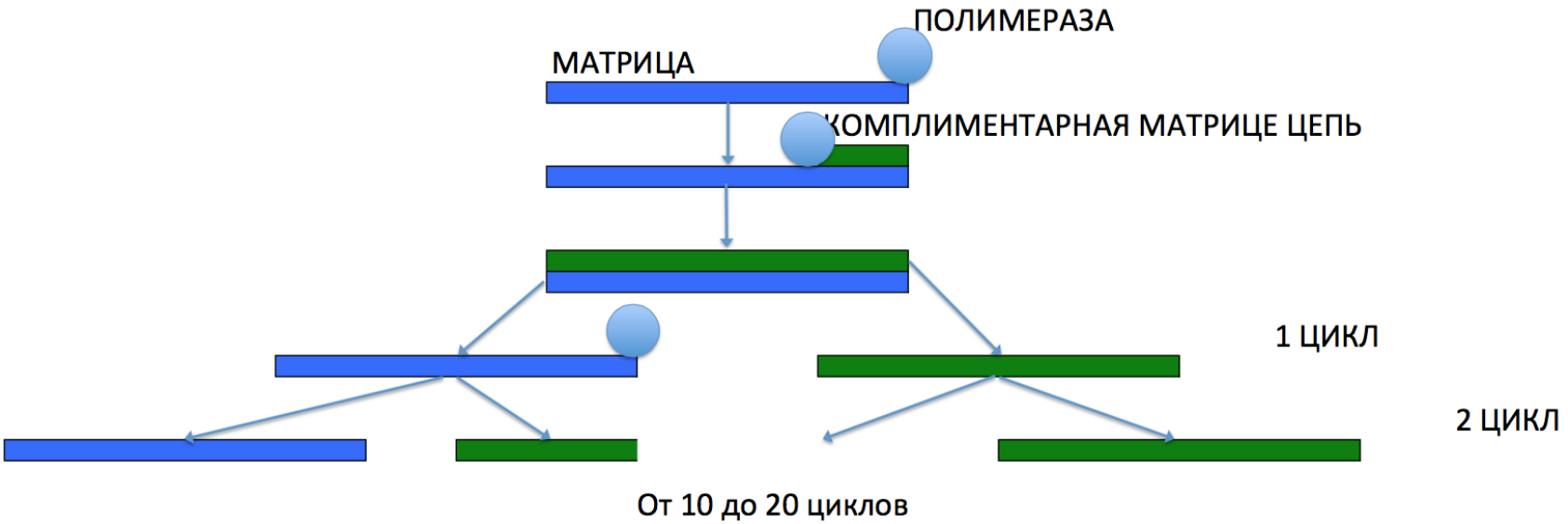
СТРУКТУРА ГЕНА В РЕФЕРЕНСНОМ ГЕНОМЕ



Интроны обычно длиннее экзонов. Экзоны кодируют получающийся белок.

Схема полимеразной цепной реакции (ПЦР)

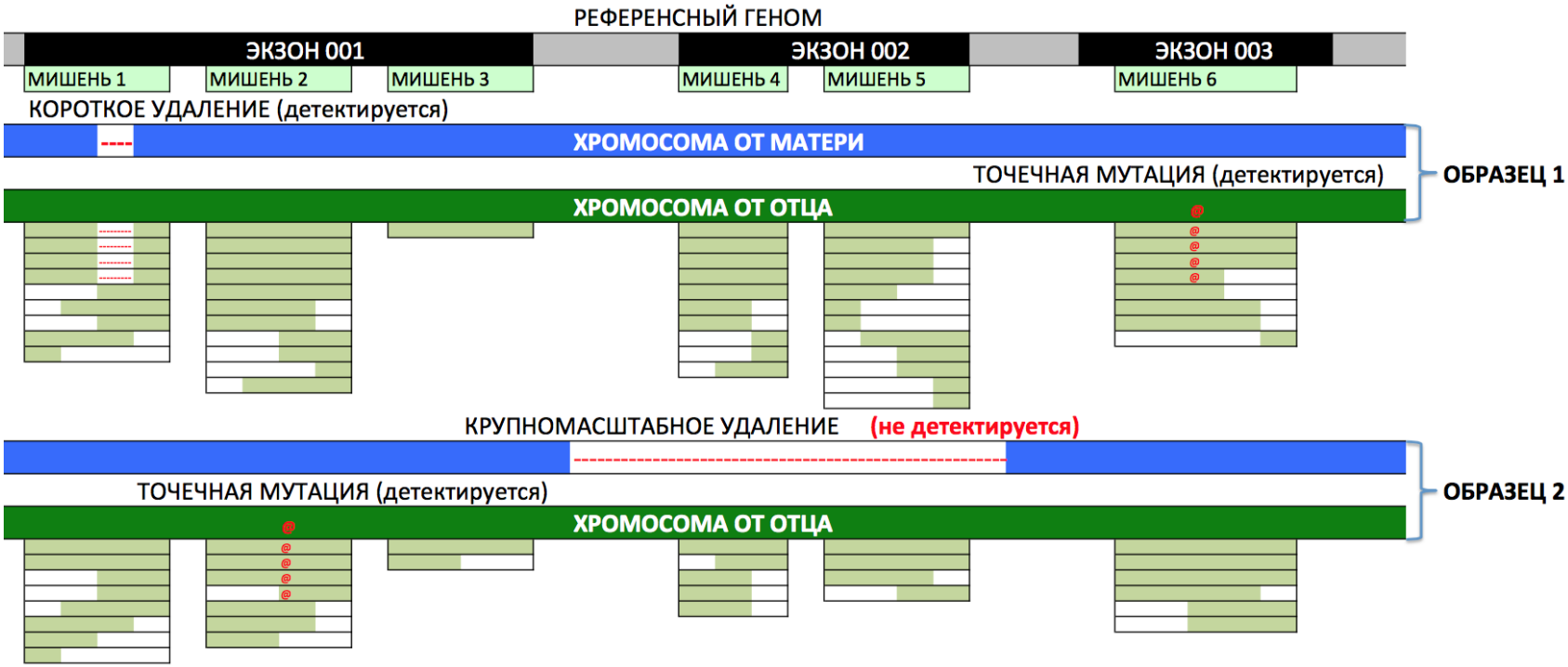
Техника для увеличения концентрации определенных фрагментов ДНК. Можно смоделировать ветвящимся процессом. На каждом шаге с вероятностью p создается комплиментарная копия существующего фрагмента или его части.



Эффект от мультиплексной ПЦР – разная скорость роста числа копий из разных регионов-мишеней i и j ($p_i \neq p_j$ при $i \neq j$ в общем случае).

Поиск вариантов с помощью NGS

Выровненные прочтения после таргетного NGS с использованием мультиплексной ПЦР. Короткие или содержащие ошибки прочтения теряются.



NGS-анализ используется в диагностике, так как позволяет точно находить точечные мутации и короткие перестройки.

Поиск CNV

Частота CNV среди вариантов некоторых генов, таких как CFTR/PAH – около 2%.

- CNV можно искать биохимически, но долго и дорого.

Поиск CNV

Частота CNV среди вариантов некоторых генов, таких как CFTR/PAH – около 2%.

- CNV можно искать биохимически, но долго и дорого.
- Детекция CNV с помощью NGS сократит время и стоимость анализа. Это может найти применение в диагностике генетически обусловленных заболеваний (муковисцидоз, фенилектонурия, рак, шизофрения и другие).

Цель и задачи

Цель: детекция CNV в данных таргетного NGS, полученных с использованием мультиплексной ПЦР.
Задачи:

- поиск существующих решений/инструментов и определение их характеристик;
- разработка и реализация подхода к поиску CNV;
- оценка чувствительности и специфичности разработанного подхода. Оценка возможности применения в клинической практике.

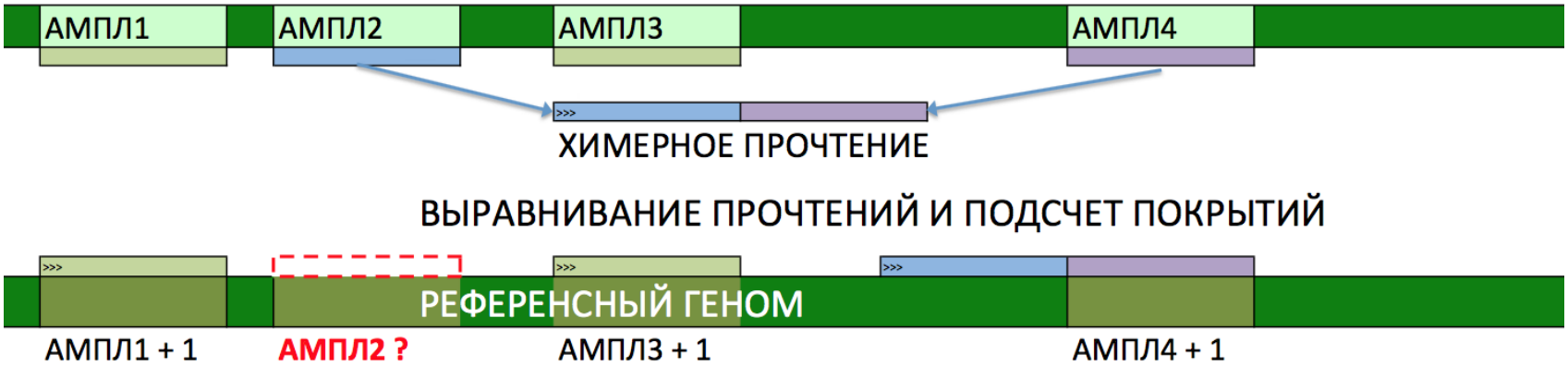
Поиск существующих инструментов

Программы коммерческие, тестирование проводилось в ограниченном режиме, неподходящие требования.

1. SeqNext (JSI medical systems GmbH): низкая специфичность (ниже, чем 0.5).
2. NextGENE® CNV Detection (Softgenetics®): анализ проводился разработчиками на наших данных, 4 из 10 предложенных образцов были отвергнуты как слишком зашумленные. Чувствительность и специфичность низки (3 TP, 1 FN, 2 FP из 6 доступных для анализа).
3. Ion Reporter™ Software Copy Number Variation Analysis (Life Technologies™): чувствительность от 0.75 до 0.0, специфичность - от 1.0 до 0.723.

Разработка подхода: подсчет покрытий

Первый шаг детекции - максимально точный подсчет числа прочтений, относящихся к каждому участку-мишени (ампликону).



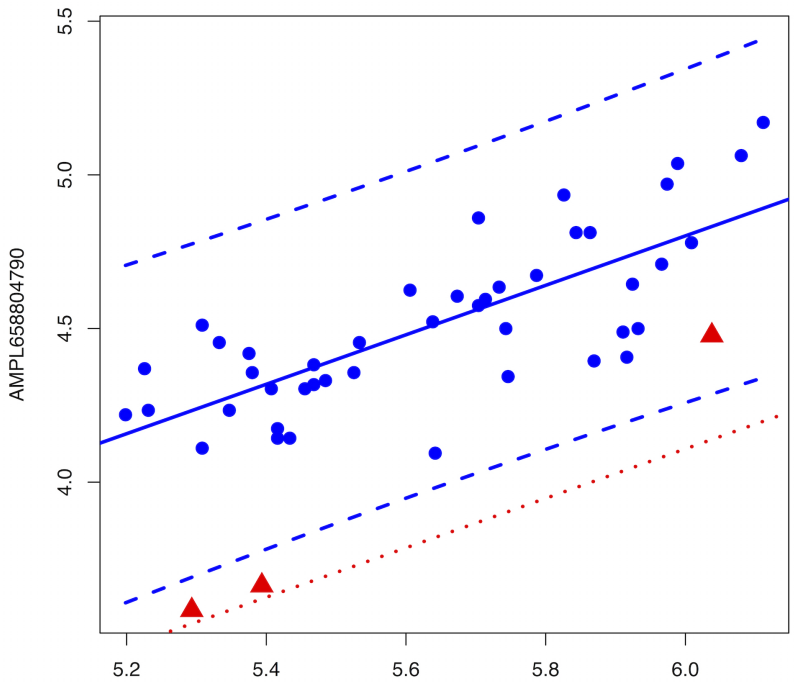
Проблема неправильного подсчета покрытий при образовании химер.

Разработка подхода: идеи

- используем выборку образцов с неизвестным количеством CNV, но предполагаем, что частота CNV в одних и тех же регионах менее 1/5 и не может сильно “испортить” робастные модели;
- для выборки образцов рассмотрим вектора покрытий каждого региона-мишени. Разные вектора могут показывать высокую корреляцию (выше 0.75).

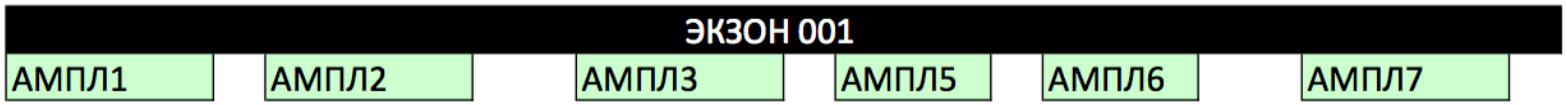
Алгоритм, использующий обучение без учителя

Построение устойчивых к выбросам линейных моделей для наиболее скоррелированных *регионов-мишеней* и поиск отклонений. Учет ошибок в предикторах в виде “голосования” различных моделей.



Алгоритм, использующий обучение с учителем

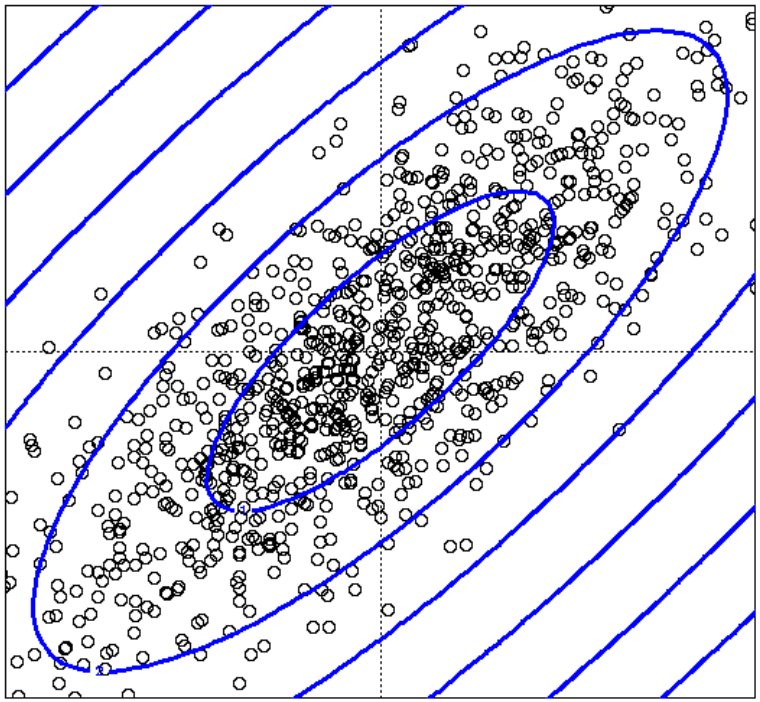
- Метки классов получаем согласно выводу предыдущего алгоритма: *нет свидетельств CNV – в контрольную выборку, есть – в тестовую*.
- Переходим от понятия *регион-мишень* к понятию *CNV-сайт*, содержащему ≥ 1 мишеней.



- Ищем статистические выбросы в многомерных случайных величинах.

Алгоритм, использующий обучение с учителем

Чтобы учесть различные дисперсии в каждой компоненте CNV-сайта, используем устойчивое к выбросам регуляризованное расстояние Махаланобиса.
По осям x, y: $(\xi_x, \xi_y), \xi_i \sim \mathcal{N}(0, \sigma_i^2), \sigma_y < \sigma_x, cor(x, y) > 0$ и доверительные эллипсы с разным α.



Верификация

804 образца: 744 не содержат CNV, 60 содержат.

	Чувствительность	Специфичность
Без учителя	0.95	0.954
С учителем	0.883	0.969

4 CNV были впервые обнаружены в ходе тестирования разработанным инструментом (ввиду недостатоков молекулярно-биологических методов), были предсказаны и подтверждены.

Результаты работы

- Решения для детекции найдены, но они недостаточно чувствительны и специфичны для применения в диагностике.
- Разработан метод детекции CNV, реализован в пакете программ CONVector.
- Установлены аналитические свойства метода. Диагностические свойства требуют оценки в клинических испытаниях.

Благодарности

- **Parseq Lab**: Брагин Антон, Павлов Александр, Внучкова Юлия, Симакова Тамара.
- **СПбАУ РАН**: Коробейников Антон, Шлемов Александр.

Спасибо за внимание!