Статистический подход для детекции CNV в данных NGS полученных с использованием мультиплексной ПЦР

Студент: Герман Демидов, СПбАУ, Parseq Lab
Руководитель: Антон Брагин, Parseq Lab

Зачем диагностировать CNV

- исследования;
- клиника;
- планирование семьи.

Примеры заболеваний

Муковисцидоз, **галактоземия**, адреногенитальный синдром, болезнь кленового сиропа, дефицит биотинидазы, гемоглобинопатия, **фенилкетонурия**, врожденный гипотироз и т.д..

Критерии заболеваний для неонатального скрининга

- без своевременного начала лечения однозначно приводят ребенка к глубокой инвалидности.
- для предотвращения инвалидизации ребенка вследствие этих болезней имеются эффективные методы лечения.
- встречаются не так уж редко - их частота выше, чем 1 на 10000 новорожденных.
- для них ***разработаны точные биохимические методы лабораторной доклинической диагностики***. (http://www.genetik.med.cap.ru)

Почему ДНК-диагностика?

Потовый тест: ионофорез с пилокарпином. Повышение хлоридов более 60 ммоль/л — вероятный диагноз; концентрация хлоридов более 100 ммоль/л — достоверный диагноз. При этом разница в концентрации хлора и натрия не должна превышать 8—10 ммоль/л. Потовый тест для постановки окончательного диагноза должен быть положительным не менее трёх раз. Потовую пробу необходимо проводить каждому ребёнку с хроническим кашлем.

Химотрипсин в стуле: проба не стандартизована — нормативные значения разрабатываются в конкретной лаборатории.

Определение жирных кислот в стуле: в норме менее 20 ммоль/день. Пограничные значения — 20—25 ммоль/день. Проба положительна при снижении функции поджелудочной железы не менее чем на 75 %.

ДНК-диагностика наиболее чувствительная и специфическая. **Ложные результаты получают в 0,5—3 % случаев.**

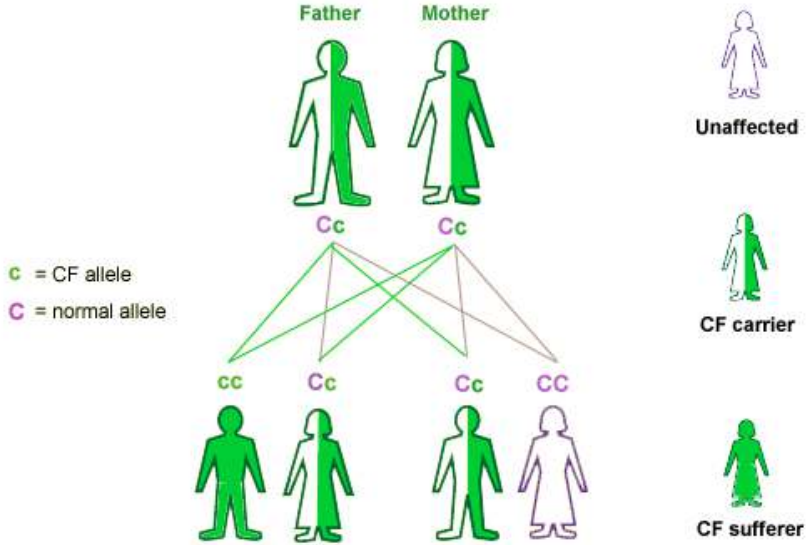
Что можно диагностировать по ДНК

Существуют “простые” наследуемые заболевания, для которых определены отвечающие за них гены.

Выбраны **муковисцидоз, галактоземия, фенилкетонурия**. Максимальная частота носительства: ~1/24 (Amish jews). В России: ~1/50.

ДНК-диагностика

Для этих заболеваний найдены гены CFTR, GALT, PAH соответственно.

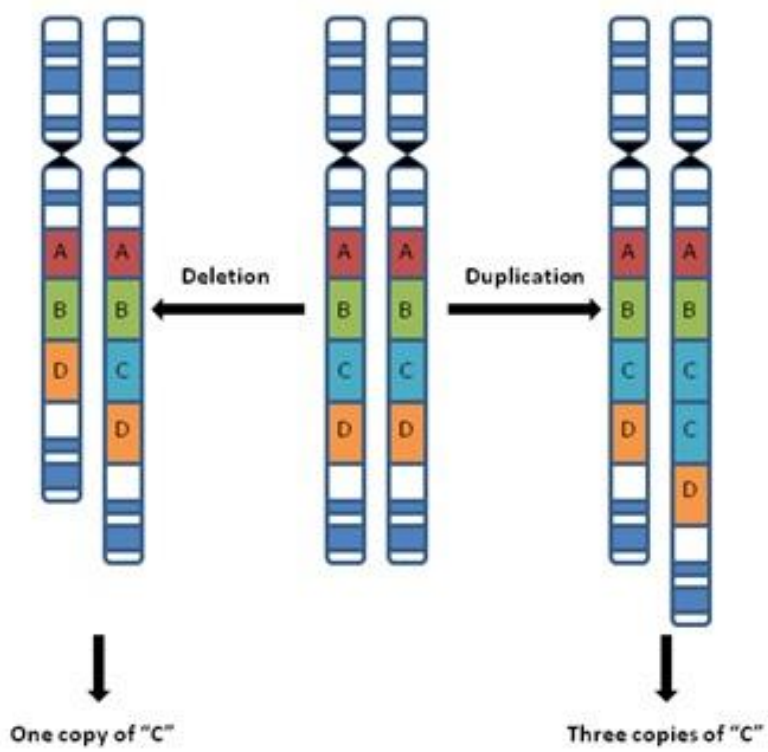


ДНК-диагностика

Parseq Lab разработали решение для клинической диагностики патогенных SNP и коротких indel.

Но кроме того бывают еще CNV. Они довольно редки. Для наших заболеваний - до 3% среди всех случаев в зависимости от популяции.

CNV


A
B
D
A
B
C
D
Deletion
A
B
C
D
A
B
C
D
Duplication
A
B
C
D
A
B
C
C
D
One copy of "C"
Three copies of "C"

Секвенирование (упрощенно)

Разделение на 2 пула.

Мультиплексная ПЦР.

(Lalam et al., 2004)

$$\mathrm{P}(Y_{n+1,i} = 2 \mid \mathcal{F}_n) = p(N_n),$$

$$\mathrm{P}(Y_{n+1,i} = 1 \mid \mathcal{F}_n) = 1 - p(N_n),$$

$$m(N_n) = 1 + p(N_n),$$

$$\sigma^2(N_n) = p(N_n)(1 - p(N_n)),$$

A. Reverse gene-specific primer with universal sequences (non-labeled)

B. Forward gene-specific primer with universal sequences (non-labeled)

C. Multiplex PCR with labeled universal forward sequences

D. Multiplex PCR product

Диагностика CNV

По покрытию. Больше дополнительной информации нет.

ВХОДНЫЕ ДАННЫЕ ДЛЯ CNVDetector:

Пациентов <= 48. Ампликонов 126. <= 6144 тестов.

	Пациент 1	Пациент 2
AMPL0001	539	128
AMPL0002	990	671

Диагностика CNV (проблемы)

- зашумленные вследствие разных причин данные;
- пропущенные значения;
- разный масштаб (у одного пациента 20.000, у другого 150.000).

Диагностика CNV (аналоги)

Tool 1, Germany:

взять контрольных пациентов, взять контрольные ампликоны, провести какие-то отсечки (из головы).

Достоинства: наверное, работает хорошо на однородных данных; работает даже для 2х пациентов.

Недостатки: много ручной работы, необоснованный подход, платный тул, низкая специфичность, чувствительность “как повезет”.

Диагностика CNV (аналоги)

Tool 2, US:

взять контрольных пациентов, сравнить между собой, натренировать HMM.

Достоинства: отлично детектирует длинные делеции (согласно документации, нет данных для проверки)

Недостатки: много ручной работы, закрытые детали реализации, платный тул, размер детектируемых вариантов больше, чем нужно (минимум 10 ампликонов), иногда Spec/Sens очень низки.

Диагностика CNV (аналоги)

MLPA:

особый вид ПЦР.

Достоинства: золотой стандарт. Всё валидируется этим методом.

Недостатки: действительно много ручной работы, дорого, несмотря на то что является “золотым стандартом” [редко] допускает ошибки обоих родов.

Диагностика CNV (наш метод)

CNV detector.

Пайплайн состоящий из многих частей, каждая из которых либо уменьшает шум, либо детектирует CNV, используя обоснованный метод.

Недостатки: все еще присутствуют ошибки, но их значительно меньше, чем у аналогов. Показано, что ошибки являются следствием пробоподготовки. Не рассматривает некоторые образцы и ампликоны (из-за непредсказуемого поведения).

CNV detector v. 1.08.

В пайплайне тулы:

-pipeline.py

-chimeric_solver.py

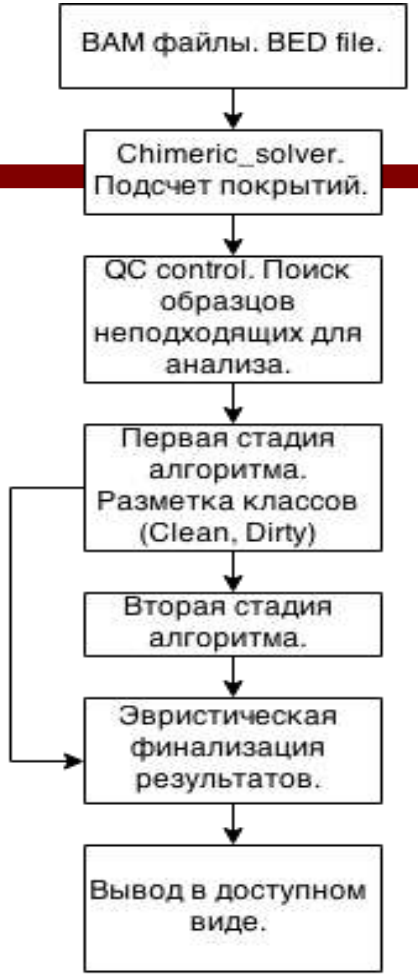
-parallelLocalAligner (java)

-qc.py

-linearDiagnost (java)

-finalizer.py

SAMTools - зависимость.



Chimeric solver

Считает покрытия, разрешает химеры и длинные ампликоны.

Генерирует свой референс по мишеням с учетом ошибок IonTorrent и SNP/indel из пациента.

Parallel Local Aligner

Модифицированный || алгоритм Smith-Waterman со специально подобранными штрафами.

100-3000 химер в нормальных данных, длина химеры > 30 bp, 126 мишеней, каждый длины 180-300.

Задача: отнести химерный кусок к одной из мишеней.

QC control

Идея: посмотреть на соотношения долей покрытий мишеней внутри каждого гена.

Предположение: CNV в 2х генах сразу - очень маловероятно.

QC control

Предположение: доли ампликонов можно аппроксимировать нормальным распределением.

Первая стадия алгоритма

Главная идея, отличающая от аналогов:

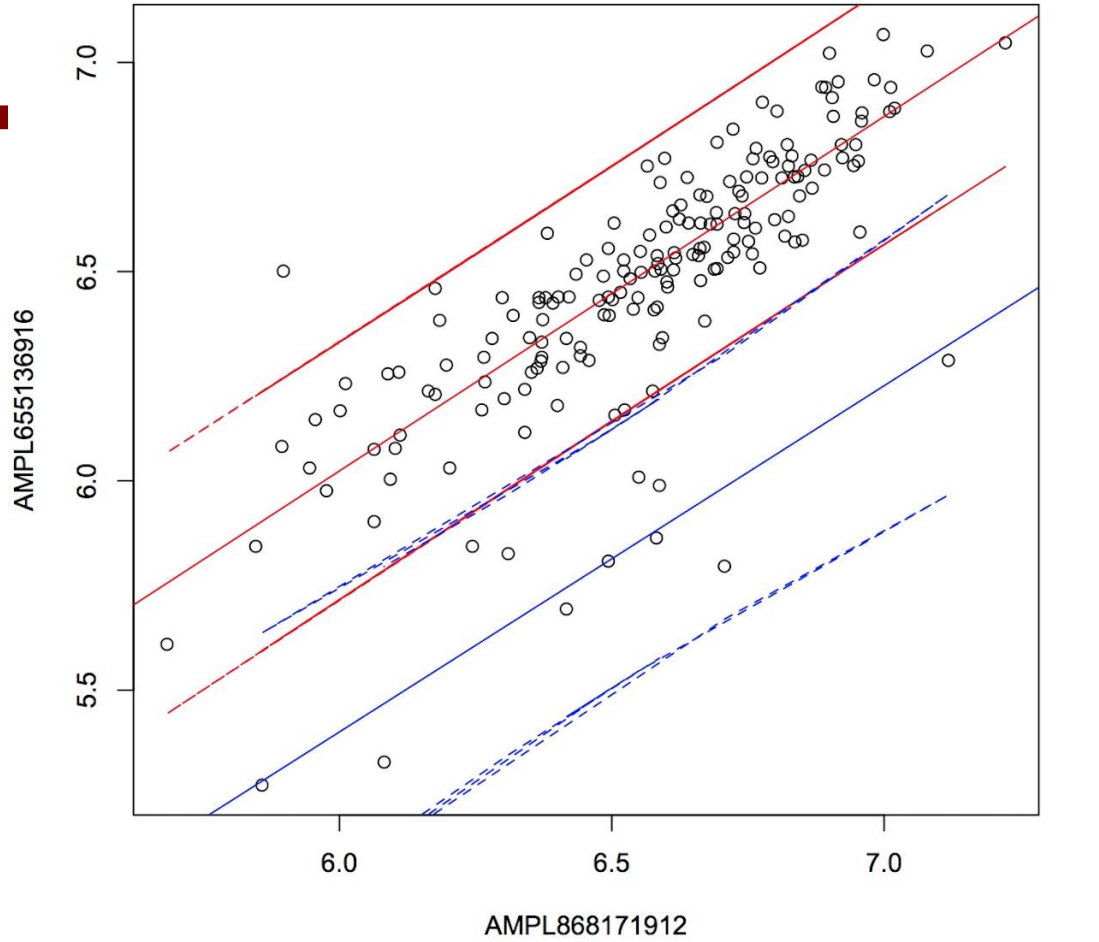
Для диагностики используется ран (<=48 пациентов). Некоторые ампликоны похожи по строению, а значит, амплифицируются похожим образом.

Формализуя математически: параметры случайного процесса будут похожими.

Первая стадия алгоритма

Раз они похожи, то итоговые значения случайных величин, заданных случайными процессами, будут как-то зависеть.

Логарифмы покрытий будут зависеть линейно (branching process, пересечение интервалов для лямбда в Box-Cox transformation).


AMPL655136916
7.0
6.5
6.0
5.5
6.0
6.5
7.0
AMPL868171912

Методы

-Theil-Sen estimator;

-S_n standard deviation estimator;

-S_n correlation;

-externally studentized residuals;

-отсечки выбранные как квантили t-распределения Стьюдента;

-голосование вместо использования TLS regression.

Вторая стадия алгоритма

У нас проставлены метки классов.

Можно улучшить точность алгоритма.

Вторая стадия алгоритма

Преимущества второй стадии:

- даже если у нас большое количество CNV в датасете, вторая стадия найдет их (первая может промазать при частоте ~20%);
- выше специфичность.

Недостатки: плохо работают с CNV длиной в один ампликон.

- чувствительность на таких CNV ниже;
- чувствительность на CNV длиннее выше.

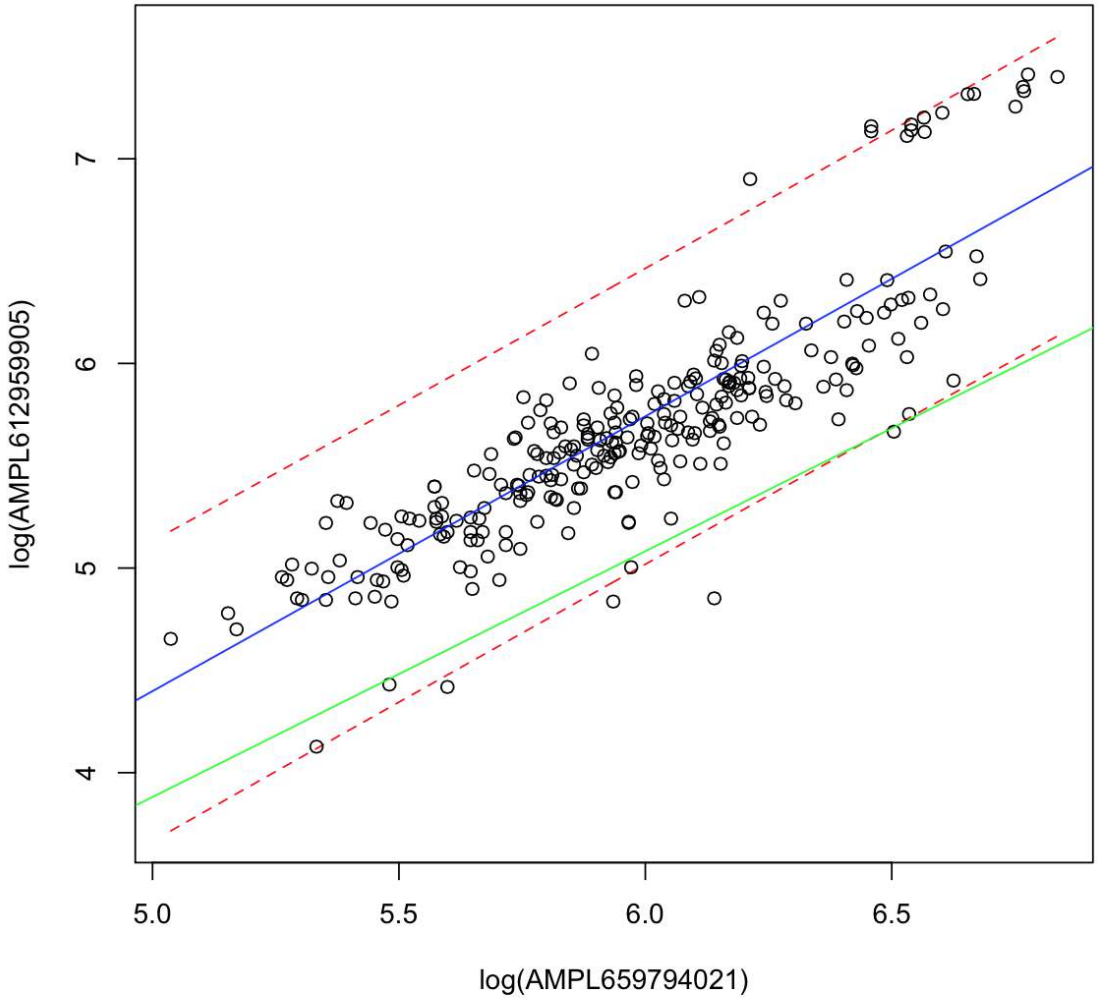
Результаты

Тренировочный датасет: 272 пациента, 6 ранов. 8 пациентов отфильтрованы по качеству.

Тестовый датасет: 236 пациентов, 6 ранов. 22 пациента отфильтровано.

Пояснение

Классификатор тренируется каждый раз по новому датасету. По train set подбирались уровни квантилей распределений.
Тренировать классификатор на одном датасете и тестировать с той же матрицей ковариаций на другом нельзя.


log(AMPL612959905)
4
5
6
7
5.0
5.5
6.0
6.5
log(AMPL659794021)

Тренировочный датасет. 1st stage

Чувствительность: 0.909. Специфичность: 0.939.

Всего было 22 CNV. 2 не нашлось. Это делеции длиной в 1 ампликон и они произошли в “проблемном” регионе PAH5 (смешение prediction intervals у линейных моделей).

РЕЦЕПТ: редизайн панели, навешивание фланкирующих ампликонов (сделано).

Тренировочный датасет. 2nd stage

Чувствительность: 0.772. Специфичность: 0.956.

Всего было 22 CNV. 5 не нашлось. Это делеции длиной в 1 ампликон и они произошли в “проблемном” регионе PAH5 (смешение prediction intervals у линейных моделей).

РЕЦЕПТ: редизайн панели, навешивание фланкирующих ампликонов (сделано).

Тестовый датасет. 1st stage

Чувствительность: 0.944. Специфичность: 0.903.

Всего было 18 CNV. 1 не нашлось. В одном ране было 6 одинаковых делеций. Робастности моделей не хватило.

РЕЦЕПТ: в реальной популяции такая или хуже ситуация произойдет с вероятностью меньшей 10^{-7} (для Amish Jews). Для России - еще меньшая вероятность.

Тестовый датасет. 2nd stage

Чувствительность: 0.833. Специфичность: 0.922.

Всего было 18 CNV. 3 не нашлось. Все они длиной в 1 ампликон.

РЕЦЕПТ: редизайн панели.

Результаты вместе

1я стадия.

Чувствительность: 0.925. Специфичность: 0.922.

2я стадия.

Чувствительность: 0.8. Специфичность: 0.94.

Большинство ошибок (~70%) происходят в регионах, покрытых одним ампликоном.

Porto, Portugal

189 пациентов. 40 отфильтровано.

9 образцов с CNV.

1я стадия. Sens = 1.0, spec = 0.921.

2я стадия. Sens = 0.889, spec = 0.957.

Liverpool, UK

189 пациентов. 9 отфильтровано.

10 образцов с CNV.

1я стадия. Sens = 0.9, spec = 0.935.

2я стадия. Sens = 0.9, spec = 0.971.

Предсказания de novo

Найдено и валидировано:

-3 делеции de novo. CFTRdele4-6, CFTRdele9, CFTRdele2-9.

Уточнили местонахождение 3х делеций CFTRdele9 (по отчетам из другой лаборатории CFTRdele8).

Ложно-положительные результаты по гену PAH пока не проверялись, так что там могут быть и истинно-положительные.

Сравнение

Американский тул.

Прогнали на 2х ранах. 1й ран - выбрали контрольные образцы (10 штук), но нашли всего 3 из 4 “длинных” делеций.

2й ран “новый”, выбрали контрольные образцы из предыдущих ранов (в инструкции допускается такое использование). 13 ложно-положительных результатов из 37, ни одного “попадания в цель” (должно быть 4).

Наш результат: 3 образца детектировано, 1 отфильтрован. 6 FP результатов.

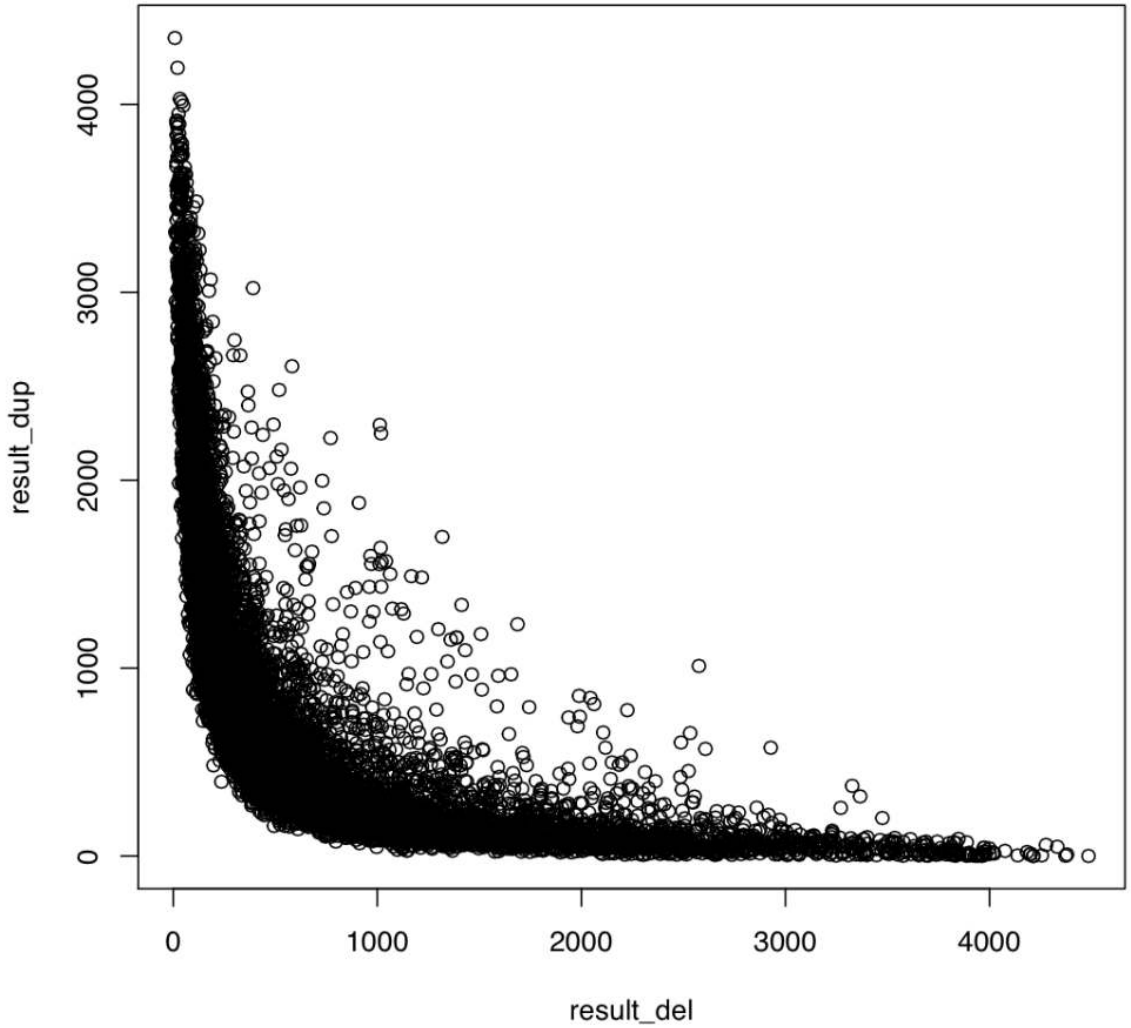
Сравнение

Немецкий тул.

Триал лицензию нам не дали. Симулировали алгоритм на “чистом датасете”. 4715 возможных тестов.

10.000 симуляций работы алгоритма (выбор 2х контрольных пациентов, 3х контрольных ампликонов, алгоритм). Лучший результат из всех симуляций - 631 ложно-положительный результат. Наш результат - 45 ложно-положительных.

В 14 раз меньше.


result_dup
0
1000
2000
3000
4000
0
1000
2000
3000
4000
result_del

Сравнение с MLPA

Тестировали (пока что) 15 образцов.

Где-то половину образцов пришлось переделывать.

Это - длительное время, большие затраты, не гарантированный (хотя очень специфичный и чувствительный) результат.

Подводя итог

Разработана теория, алгоритмы, тулы для решения задачи диагностики. Тул работает в одну кнопку и требует разовой калибровки.

Тул применяется на практике. Тул работает на данных, полученных из разных лабораторий.

Недостатки (ложно-положительные и ложно-отрицательные результаты) решаются изменением условий для эксперимента. Эти условия мы нашли и с их помощью биологи сделали новый протокол.

Вопросы?

(Artwork designed by giftsforawareness)