Consequence

Сергей Лебедев

Руководитель: Николай Вяххи

http://github.com/bioinf/consequence

4 июня 2012 г.

- aka **S**ingle **N**ucleotide **P**olymorphism,
- точечное отличие в последовательности ДНК представителей одного вида.
- rs334[1], HBB (haemoglobin beta) gene, position 70614:

  `TGACTCCTGAGGAGAAGTCT => TGACTCCTGAGGAGAAGTCT`

- Выявлено значительное количество SNP, влияющих на вероятность развития отдельных заболеваний; примеры: *серповидно-клеточная анемия*, *глаукома*, *болезнь Альцгеймера*, *диабет* – см. SNPedia;
- Крупные проекты по исследованию SNP и др. вариаций: 1000 Genomes, 23andMe, deCODE.

[1] http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=334

- Общая задача: найти в данных ридах все SNP, по сравнению с референсным геномом; когда рефернс отсутствует – в двух наборах ридов;
- Для человека референс имеется, поэтому большинство существующих решений для поиска SNP *de-novo* предполагают на входе риды, выравненные на референсный геном, а затем максимизируют $Pr(genotype|reads)$, примеры: samtools, GATK, SOAP2;
- Часто, требуется проверить наличие *уже известных* SNP, а не искать новые, например для того, чтобы определить предполагаемый фенотип по ридам.

Общеиспользуемый вариант

```
reads -> BAM -> sorted BAM -> raw VCF -> SNPs
```

`reads -> BAM -> SNPs`

- Выравниваем риды на **частичный** референсный геном, из которого исключены участки, не содержащие ни одного SNP из **индекса**.
- После выравнивания, для каждого рида известно его местоположение в *частичном* референсном геноме, из которого можно получить соотв. позицию в полном геноме;
- а затем поискать в *индексе* все базы, не совпавшие с референсом.
- На выходе получаем список SNP с частотой встречаемости ($MAPQ \cdot occ$), для каждого SNP в индексе хранится соотв. множество геномов $\rightarrow$

- Индекс: *Genomic position* $\rightarrow$ 4 *tuple*;
- Индекс поддерживается для каждой хромосомы в отдельности;
- Обновление индекса можно производить в фоне, параллельно для всех новых геномов, которые необходимо добавить.

Пример:

```
                A    C      G                T
                |    |      |                |
1049631 -> ([], NULL, ["HG00096"], [])
                     ^
                     |
               reference base
```

Результаты: Время работы

	E.coli		*HG00096*	
	samtools	consequence	samtools	consequence
Reference indexing	<1m	<1m	191m	69m
Read mapping	20m	16m	42m	31m
BAM sorting	13m	—	30m	—
SNP calling	15m	3m	103m	61m
BCF $\rightarrow$ VCF	<1m	<1m	12m	—
	48m	19m	378m	161m

Таблица: Сравнение[2] времени работы на ридах для E.coli (*?*) и человека (1000 Genomes, HG00096)

[2] Для выравнивания ридов в обоих случаях использовался Bowtie

	Общие	samtools	consequence	Чувствительность
E.coli	23	139	19	14.2%
HG00096[3]	10987	860454[4]	3962	1.2% / 75.8%

Таблица: Количество SNP, найденных каждым из методов

Sensitivity and ~~specificity~~

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

$$specifity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$$

[3] Индекс содержал данные openSNP для пользователя Mark Davis

[4] Реальное количество SNP, присутствующих в индексе – 14488

Результаты: E.coli

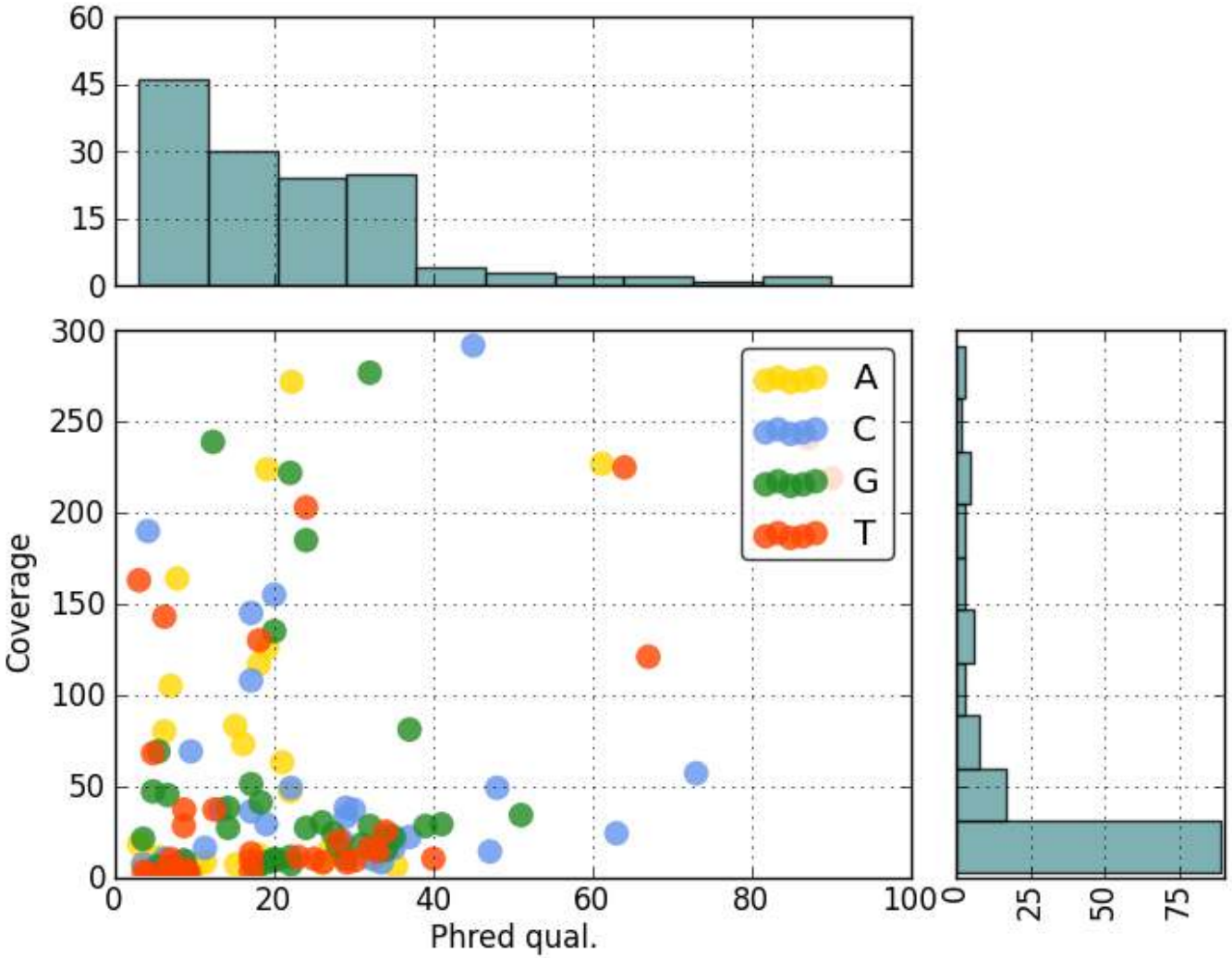

Рис.: Точечный график покрытия **позиций**, в которых не были найдены SNP и **качества** этих SNP, в выводе samtools

Результаты: HG00096

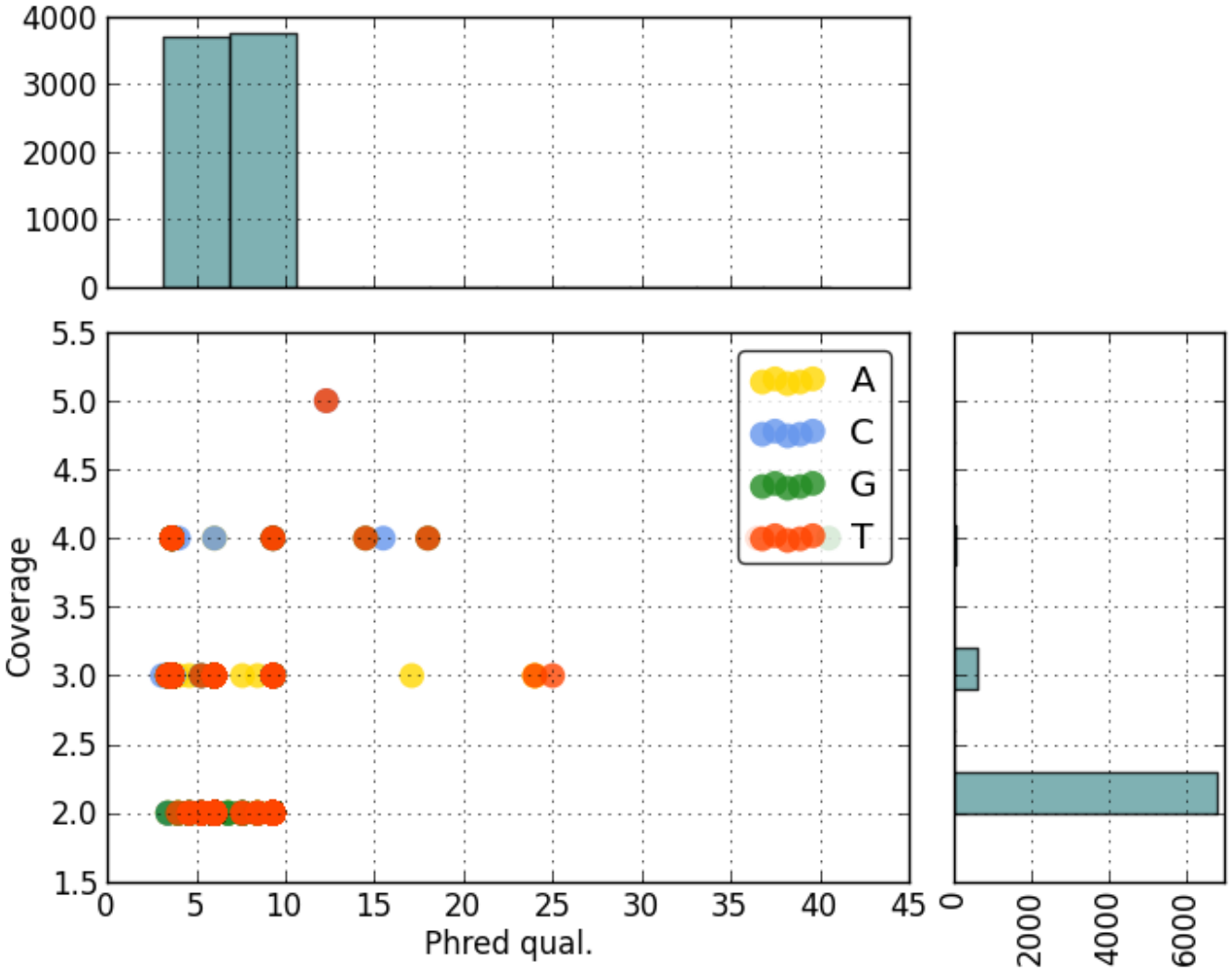

Рис.: Точечный график покрытия **позиций**, в которых не были найдены SNP и **качества** этих SNP, в выводе samtools

- Текущая реализация не обладает достаточной чувствительностью для поиска **всех** SNP в исходных ридах, но даже такой чувствительности достаточно для поиска “родственных” геномов и определения фенотипа.
- Количество false positives свидетельствует о недостаточной точности фильтрации – хотелось бы получить меньше, но **достоверных** SNP.

Планы

- Улучшить точность фильтрации найденных SNP, например добавив к итоговому score для каждого SNP phred. quality из соотв. ридов.
- Реализовать поддержку индексации форматов, доступных на openSNP – phenotype, finally!

Спасибо за внимание!