SNP detection for massively parallel whole-genome resequencing

Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang

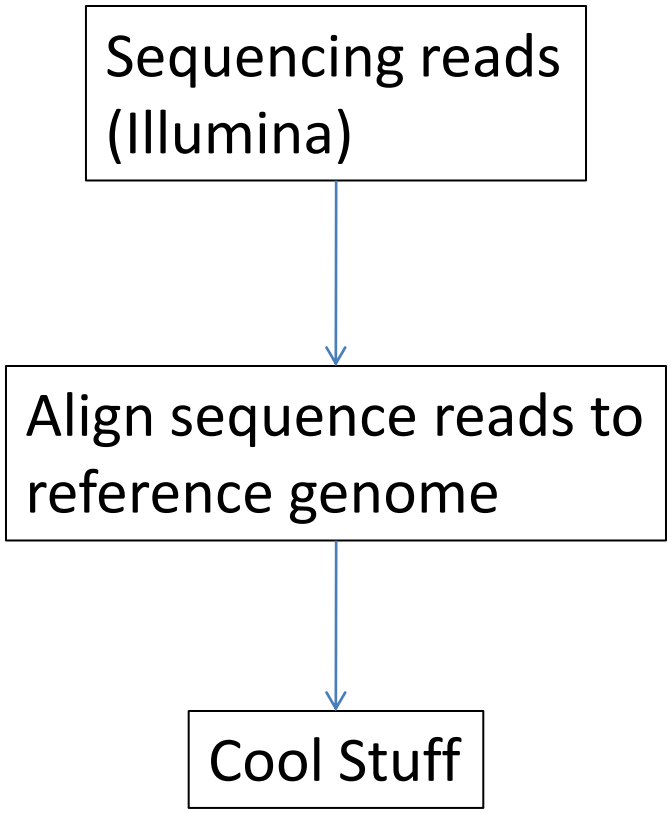
Presenter: Mark Sun

Overview

- Background
- SOAPsnp method
- Validation
- Comments

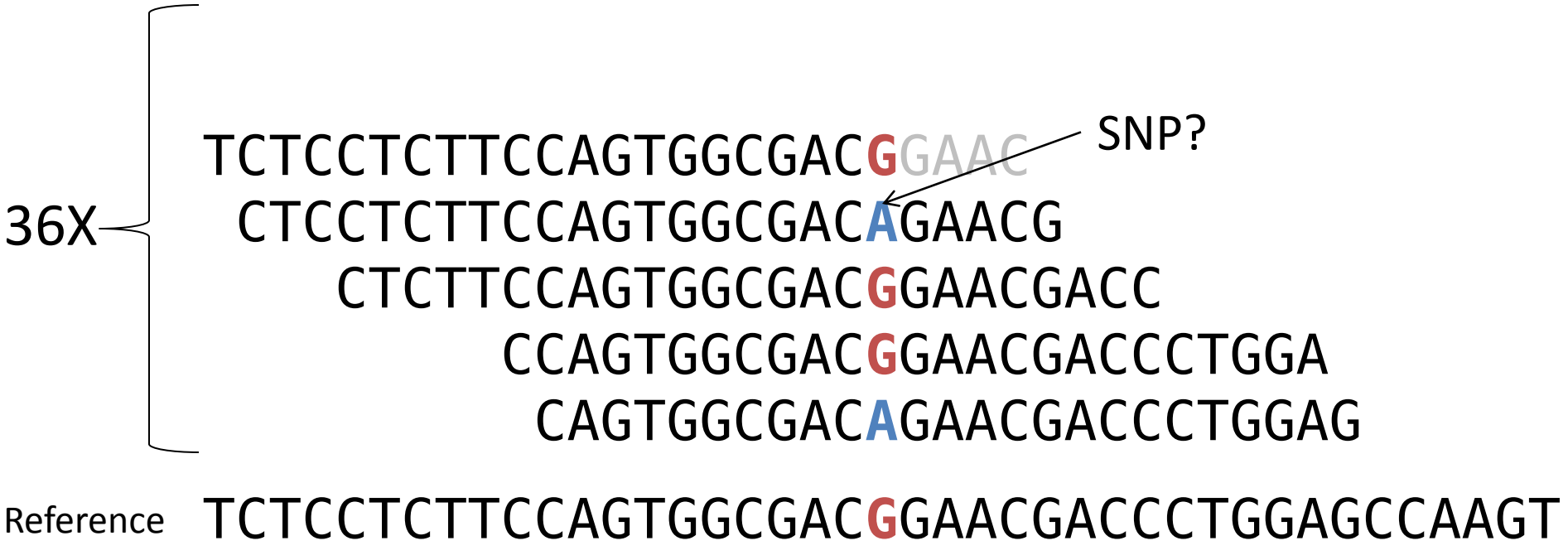
Background

Background – Where we are



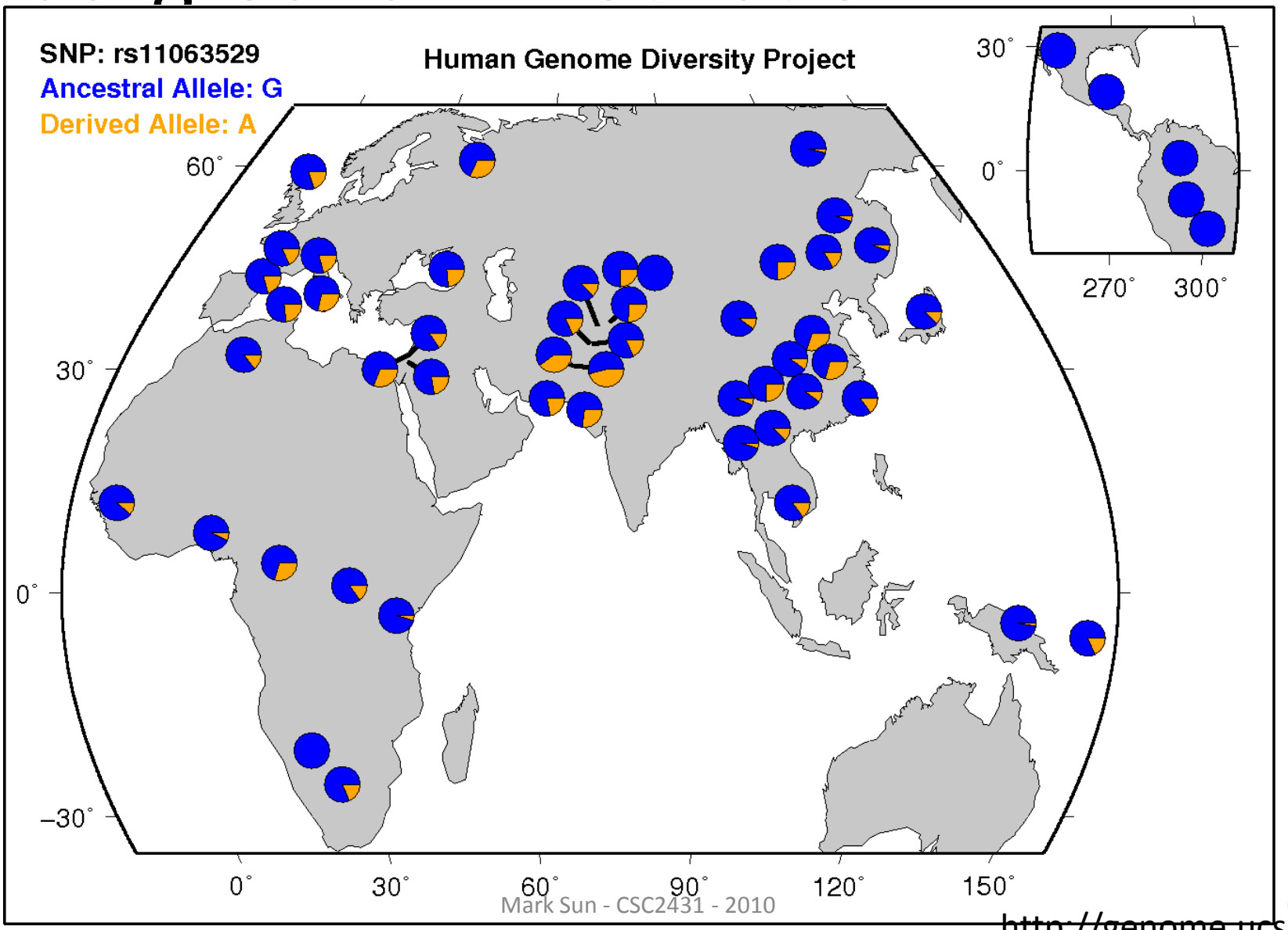
Background – Goal

- Given aligned short reads to a reference genome, is a read position a SNP?



dbSNP build 130 - rs11063529 (A/G): chr12:527665-527665 against NCBI build 36.1

Background – Motivation



http://genome.ucsc.edu

Background – Challenges

- Sequencing errors
- Misalignments
  - Indels (SOAP issue) . . . GAC**G**GAACTTT
  - Repeats

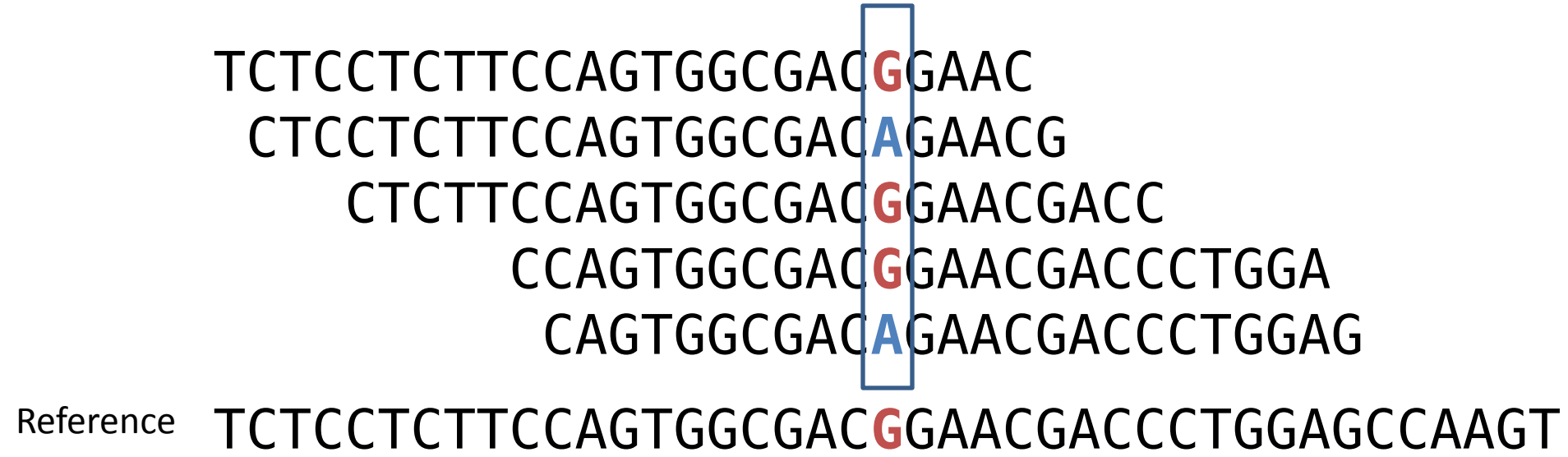
```
TCTCCTCTTCCAGTGGCGACGGAAC
 CTCCTCTTCCAGTGGCGACAGAACG
    CTCTTCCAGTGGCGACGGAACGACC
          CCAGTGGCGACGGAACGACCCTGGA
           CAGTGGCGACAGAACGACCCTGGAG

Reference TCTCCTCTTCCAGTGGCGACGGAACGACCCTGGAGCCAAGT
```

SOAPsnp

SOAPsnp – a solution

- Use Bayesian model to capture Illumina errors
  - T_i: genotype
  - D: data at a locus
  - S: total number of genotypes

$$P(T_i|D) = \frac{P(D|T_i)P(T_i)}{\sum_{x=1}^{S} P(D|T_x)P(T_x)}$$

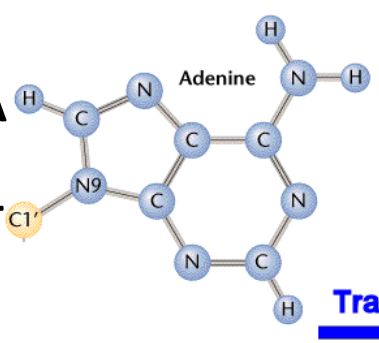


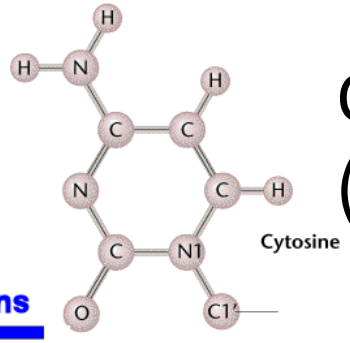
SOAPsnp – Setting the priors

- Haploid, SNP rate = 0.001

A

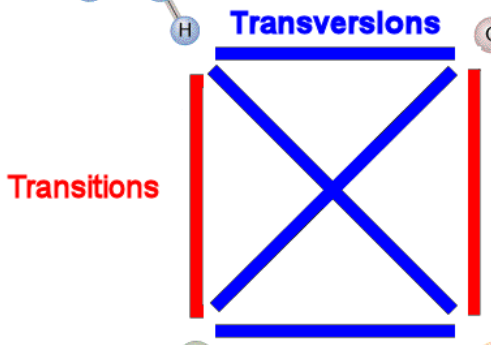
(4/6)x0.001





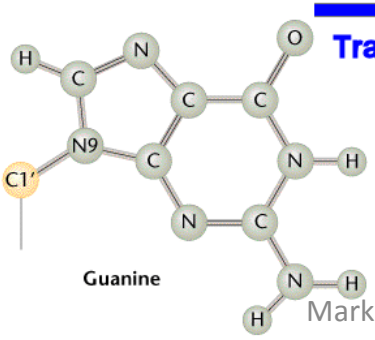
C

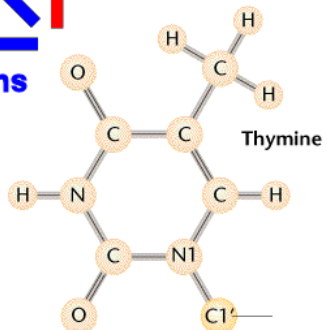
(1/6)x0.001



G

0.999





T

(1/6)x0.001

http://www.mun.ca/biology/scarr/

SOAPsnp – Setting the priors

- Diploid
  - Heterozygous SNP rate = 0.001
  - Homozygous SNP rate = 0.0005

	A	C	G	T
A	3.33×10^{-4}	1.11×10^{-7}	6.67×10^{-4}	1.11×10^{-7}
C		8.33×10^{-5}	1.67×10^{-4}	2.78×10^{-8}
G			0.9985	1.67×10^{-4}
T				8.33×10^{-5}

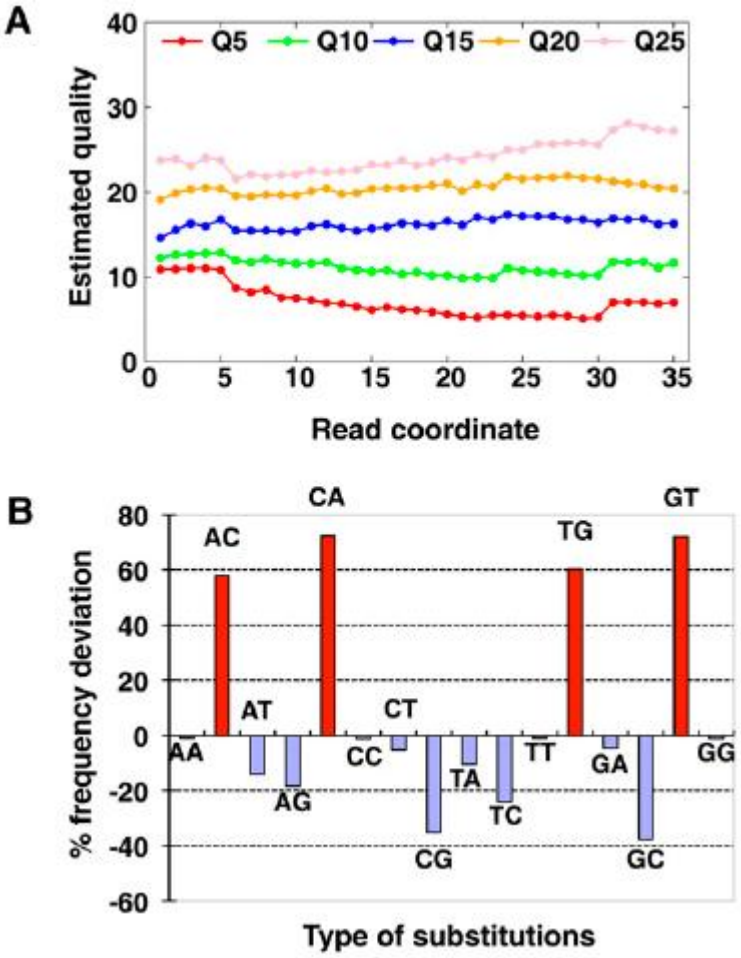
SOAPsnp – Likelihood

$$P(D|T) = \prod_{k=1}^{n} P(d_k|T)$$

$$P(d_k|T)$$
$$= P(o_k, q_k, c_k|T)$$
$$= P(o_k, c_k|q_k, T)P(q_k|T)$$

T: genotype (GG/GA/AA)
o: observed allele type
q: quality score
c: cycle

SOAPsnp – Likelihood



$$P(d_k|T) = P(o_k, q_k, c_k|T) = P(o_k, c_k|q_k, T)P(q_k|T)$$

T: genotype

o: observed allele type

q: quality score

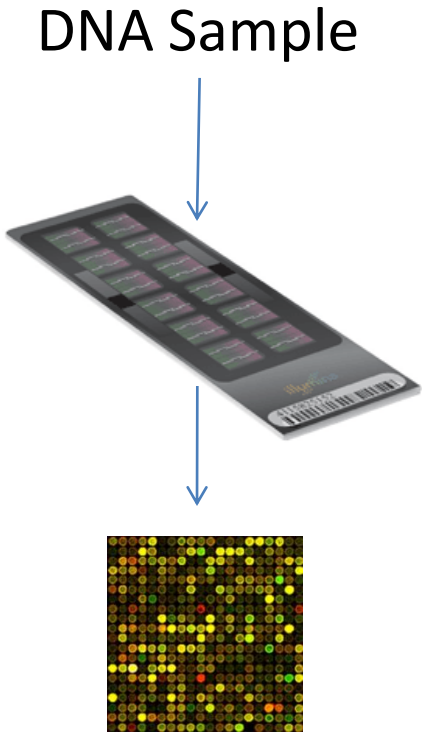
c: cycle

Recalibrated quality scores using mismatch rate from the alignment of Han Chinese genome (Only uniquely aligned reads used and not in dbSNP).

Validation

Validation

- Compare SOAPsnp's genotype calls using an Illumina 1M BeadChip (microArray)
- False Negative: heterozygote site, where one allele missing in sequencing data
- False Positive: incorrect genotype call



$$P(T_i|D) = \frac{P(D|T_i)P(T_i)}{\sum_{x=1}^{S} P(D|T_x)P(T_x)}$$

http://www.illumina.com
http://bioinfoworld.files.wordpress.com

Validation

Table 2. Coverage and accuracy of the Illumina 1M BeadChip genotyped sites of the called consensus sequence

Illumina 1M genotype	Genotyped sites	Covered in assembly	Agreed	FP	FN
Chr X					
HOM reference	27,196	98.654%	99.996%	0.004%	—
HOM mutant	10,737	98.491%	99.887%	0.113%	—
Total	37,933	98.608%	99.965%	0.035%	—
Autosome					
HOM reference	540,878	99.109%	99.956%	0.044%	—
HOM mutant	208,436	98.790%	99.806%	0.194%	—
HET	250,667	94.811%	99.609%	0.017%	0.374%
Total	999,981	97.965%	99.840%	0.069%	0.091%

Sanger sequencing on 57 false SNP loci showed 49 (86%) had allele types consistent with Illumina sequencing.

Comments

Comments

- Filters
  - quality
  - 2 reads for haploid / 4 reads for diploid
  - depth less than 100
  - flanking region copy number < 2
  - at least one paired-end read
  - SNPs must be least 5bp away
- Different priors for low depth (use dbSNP)

Questions