A Novel Method for Detecting Contaminated Sample Based on Illumina Sequencing Data

Zheng Huang, Qibin Li, Wei Jin, Qijun Liao, and Xiao Sun

***Abstract*—Illumina sequencing platform is widely used in genetics research. Due to the complex andlong-term library construction and DNA sequencing, samples can be contaminated with different sources, which can lead to false-positive SNP calling. To identify the contaminated samples, we built a model of mappability score to quantitatively measurethe accessibility of different parts ofhuman genome. By characterizing the genomic region with high probability of uniqueness and counting the discordant reads against genotypes on the unique region, we could detect outliers as the contaminated samples in a population scale. Totest the effectiveness of our method, we manually mixed the sequencing reads of two clean samples. With the prior knowledge of mixture rate, we concluded that ourmethodis quite sensitive for female samples contaminated even slightly by male samples, accurate for male samples with moderate contamination by female samples and powerful for severe cross-individual contamination with the same gender. This method is easily understood but fairly effective in population-scale sample quality control.**

***Index Terms*—Contamination, mappability score, sample quality control, unique region.**

I. Introduction

Next-generationsequencingis well known for its higher throughput, lower costs and reduced error rates [1], which allows for studying genetic variation that causes human disease and traits. Since the sequencing flow is not totally automated, *i.e*, few robots are used to automate the library preparation, samples may be contaminated by experimenters' DNA. Besides, samples are often pooled together to reach the enough sequencing concentration, resulting in DNA from more than one individual end up in the same well or prepared library. These reasons account for the major classes of within-species contamination, which can greatly reduce the accuracy of genotype calling and lead to false positive SNPs and genes associated with disease.

Unlike cross-species contamination, within-species contamination could not be easily identified by aligning reads to published genomes [2] or searching the database like tRNAdb [3]. Generally among a population, unusually high ratio of heterozygous to homozygous variant genotypes (HET/HOM ratio) across all SNP sitesin a sample suggests cross individual contamination may happen. However, we can't figure out what the HET/HOM ratio should be because populations with recent admixture will skew towards heterozygosity while populations with inbreeding will skew towards homozygosity.

Besides the HET/HOM ratio method, Cibulskis *et al*. [4] uses a Bayesian approach to calculate theposterior probability of the contamination level with known genotypes from genotyping array. And Jun *et al*. [5] demonstrates a likelihood-based method that could detectDNA sample contamination either using sequence data alone or with array-based genotypes. Although both methods are sensitive for estimating levels of contamination as low as 1%~1.5%, neither of them considers the genomic feature of large proportion of repeated sequences and pseudoautosomal (PAR) gene region on sex chromosomes. Reads may not be mapped onto the exact position on these un-unique regions, resulting in an overrated contamination level.

Though unique genomic region are much accessible or mappable than the un-unique region, unique and un-uniqueterminology are too simple to provide a quantitative measurement of the accessibility of different parts ofhuman genome. In this paper, we attempt to design a mappability score that could give us a probabilisticmeasurement of the accessibility of human genome, given current sequencing settings. Once the boundaries of the unique genome were fixed, we could detect cross-gender and within-gender sample contamination.

II. Modeling the Mappability of Human Genome

We firstly introduced a parameter ϵ, which represents the probability a base mis-sequenced when sequencing an individual genome. Suppose no error bias exists, the probability that a base is miscalled to one of the other 3 bases is $\epsilon/3$. The probability that a base is sequenced to itself is$1-\epsilon$.

Given a genomic site, its mappability can be defined asthe probability that the read really comes from the site. Let U be the genomic sequence the read came from, C_0be the genomic sequence the read mapped and C_i , $i = 1, 2 \ldots n$ be n high similar copies of C_0 across the human genome, d_i be the mismatch number between C_i and C_0. By Bayesian formula,

$$Mappability = Pr\{U = C_0|R\} = \frac{Pr\{R|U=C_0\}\times Pr\{U=C_0\}}{\sum_{i=0}^{n} Pr\{R|U=C_i\}\times Pr\{U=C_i\}} \quad (1)$$

$\Pr\{R|U = C_i\}$ is the likelihood that read R comes from C_i. Here R is a generic read, *i.e.* the mismatch number between R and C_0 (denoted asυ) can be any value between 0 and its maximal value $max(\upsilon)$. In practice, $max(\upsilon)$ is a common

Manuscript received September 14, 2013; revised November 23, 2013. This work was supported in part by BGI-Shenzhen.

Zheng Huang and Xiao Sun are with the State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, CO 210096, China (e-mail: huangzheng@genomics.cn, xsun@seu.edu.cn).

Qibin Li, Wei Jin, and Qijun Liao are with BGI-Shenzhen, Shenzhen, CO 518083, China (e-mail: liqb@genomics.cn, jinwei@genomics.cn, liaoqijun@genomics.cn).

setting of read aligners which specifies the maximal mismatch number allowed when map short reads to reference genome. The likelihood that read R comes from C_i can be expressed as:

$$Pr\{R|U = C_i\} = \sum_{j=0}^{max\ (v)} Pr\{v = j|U = C_i\} \quad (2)$$

$Pr\{v = j|U = C_i\}$ isthe probability that there are mismatches between R and C_0, given the read comes from C_i. To get mappability, we have to find a way to solve this conditional probability first. The difficulty is sites that differ between C_i and C_0 may overlap with mismatches between R and C_0, so the mismatch number between R and C_i is uncertain. Suppose there are k sites at which both R and C_i differ from C_0, given R comes from C_i, we could get the following 4 items:

1) There are $d_i - k$ sites at which R is same with C_0, but C_i differ from C_0. The base generating probability is $\epsilon/3$.
2) There are $j - k$ sites at which R differ from C_0, but C_i is same with C_0. These sites can be at any positions along the read except d_i fixed sites where C_i differ from C_i. The base generating probability is ϵ.
3) Of k sites where both R and C_i differ from C_0, a of them are identical between R and C_i, the generating probability is $1 - \epsilon$. The remaining $k - a$ bases in R differ from both C_i and C_0, the base generating probability is $2\epsilon/3$.
4) If the read length is L, the remaining $L - d_i - j + k$ bases in R are identical to C_i. The base generating probability is $1 - \epsilon$.

Given above information, we could obtain the value of $Pr\{v = j|U = C_i\}$ when there are k overlaps and a of them in R are identical to C_i, denoted as f_{ka}:

$$f_{ka} = \binom{d_i}{d_i - k}\left(\frac{\epsilon}{3}\right)^{d_i - k}\binom{L - d_i}{j - k}\epsilon^{j-k}\binom{k}{a}(1 - \epsilon)^{a}\left(\frac{2\epsilon}{3}\right)^{k-a}(1 - \epsilon)^{L - d_i - j + k} \quad (3)$$

Here, $k \in (0, min(d_i, j))$. The mismatch number between R and C_i (denoted as t_i) is $d_i + j - k - a$. As aligners always report the best or one of the equal best hits of a sequencing read, so $t_i \geq j$ holds. We could get $a \leq d_i - k$. It is obvious that a also is less than or equal to k, so the upper bound of a is $min(k, d_i - k)$. The lower bound of a is 0. Now, we could express $Pr\{v = j|U = C_i\}$ as:

$$\begin{aligned} &Pr\{v = j|U = C_i\} \\ &= \sum_{k=0}^{min\ (d_i, j)}\ \sum_{a=0}^{min\ (k, d_i - k)} f_{ak} \\ &= \sum_{k=0}^{min\ (d_i, j)}\ \sum_{a=0}^{min\ (k, d_i - k)} \binom{d_i}{d_i - k}\left(\frac{\epsilon}{3}\right)^{d_i - k}\binom{L - d_i}{j - k}\epsilon^{j-k}\binom{k}{a}\left(\frac{2\epsilon}{3}\right)^{k-a}(1 - \epsilon)^{L - d_i - j + k + a} \end{aligned} \quad (4)$$

$Pr\{v = j|U = C_i\}$ is a function of d_i, j. From (2), we could calculate $Pr\{R|U = C_i\}$.

In order to compute mappability, we also need to know $Pr\{U = C_i\}$, the prior probability that R comes from U. For simplicity, we use a uniform prior, assuming that R has the same chance to come from each genomic copy, although it is naive.

In practical calculation, we set L to 90 bp and ϵ to0.01 which are canonical parameters for Illumina sequencing platform. Mappabilitycan be computed according to the following steps:

1) Tabulate $Pr\{v = j|U = C_i\}$ with $d_i \in (0, 1, 2, 3)$ and $j \in (0, 1, 2, 3)$.
2) Cut 90 bp short sequences from reference genome with a step of 3 and map them to reference genome allowing 3 mismatches by bowtie [6]. We use bowtie because it can report all genomic hits with 3 or less than 3 mismatches.
3) For each genomic site, we calculate and average mappability scores of all short sequence that cover this site according to (1)to get the mappability score of the site.

Generally for a specific site, if there is another copy present in the genome, the estimated mappability scoreshould be close to 0.5. And a mappability score of 1indicatesthat the site is unique (We call the integration of these sites Mappability-based unique region, Muniqnom). As a result (Table I), we found 85.1%of human genome sites are unique, *i.e*, 85.1%of sites have no more than one copy across the genome. This number is close to the Uniqnom that Koehler *et al.* [7] estimated in 2010 simply usingthe ISAS aligner, which is 87.5%, respectly. Morever, Koehler's Uniqnom shared 98.5% of the unique region detected by our method.

TABLE I: COMPARISON BETWEEN MUNIQNOM AND UNIQNOM

Chromo--some	# Sites except N (Mb)	# Sites in Muniqnom (Mb)	# Sites in Uniqnom (Mb)	# Shared sites (Mb)	Shared rate (%)
1	225.28	191.18	196.85	188.30	98.5
2	238.20	207.15	212.27	204.50	98.7
3	194.80	172.65	176.80	170.33	98.7
4	187.66	164.74	168.62	162.51	98.6
5	177.70	154.59	158.34	152.54	98.7
6	167.40	147.50	151.19	145.43	98.6
7	155.35	129.15	133.43	126.99	98.3
8	142.89	126.20	128.99	124.57	98.7
9	120.14	95.40	98.21	94.04	98.6
10	131.31	112.19	115.35	110.48	98.5
11	131.13	112.97	116.13	111.25	98.5
12	130.48	113.98	117.26	112.17	98.4
13	95.59	85.51	87.43	84.44	98.7
14	88.29	76.89	79.01	75.76	98.5
15	81.69	66.82	69.06	65.74	98.4
16	78.88	64.20	66.46	63.08	98.3
17	77.80	63.22	66.15	61.87	97.9
18	74.66	66.78	68.23	66.01	98.9
19	55.81	43.84	46.50	42.38	96.7
20	59.51	52.76	54.21	52.02	98.6
21	35.11	29.67	30.58	29.25	98.6
22	34.89	27.84	29.11	27.28	98.0
X	151.10	120.80	124.61	118.53	98.1
Y	25.65	9.00	9.79	8.57	95.3
M	0.0166	0.0054	0.0059	0.0047	86.4
Total	2861.34	2435.03	2504.57	2398.07	98.5
Percent*	100%	85.1%	87.5%	83.8%	

*: The overall unique rate with respect to the total length of the non-gapped genome.

As we know, repeats comprise at least 50% of the human

genome [8], so unique region rate seems to be far more less than what we estimated here. This misunderstanding can be explained by the different definition of similarity. For example, two Alu elements is similar despite of a dozen mismatches while they can be unique since only 3 mismatches are tolerated in a read length in our definition.

When defining the boundary of the unique region, we only considerthose pieces whose length is larger than the read length. Thus we obtain a 9.00 Mb unique region of Ychromosome and 120.80 Mb unique region of Xchromosome, which will be used in detecting contamination afterwards.

III. The Principle of Contamination Detection

Among cross-individual contamination, a female contaminated by a male sample is relatively easy to detect since no Y chromosome in female genome. By calculating reads mapped onto the unique region of Y, we could easily distinguish normal female sample from contaminated ones. However, this method is not applicable in cases of male sample contaminated by female ones and within-gender contamination since no new chromosomes added. Luckly, there are still signs to be tracked. Unlike HOM/HET ratio which takes all SNPs into account and may not be so sensitive to slight contamination, we only consider homogenous genotypes. If we have high confidence about the homozygosity of a SNP in unique region, then reads discordant with the genotype might be result from contamination allowing for the sequencing error.

In detail, since male has only one X chromosome, SNPs on X are homozygous. To test whether a male sample A is contaminated by female one, e.g. sample B, we denote A_i and B_i be the underling genotypes at avariable site i on X chromosome, A_i can take value M and m, while B_i can take value MM, Mm and mm. Among which M stands for the reference allele and m is the mutant allele. If sample A is contaminated by B and A_i is M and B_i is mm or Mm, for a sequencing read covering this site, we have a higher probability to observe an m in contrast with that A is not contaminated. By calculating the discordant reads number of all the SNPs on unique region of X chromosome, we could easily obtain the discordant fraction F of a given male sample using the equation below:

$$F = \frac{\sum_i N_i(m|M)+N_i(M|m)}{\sum_i D_i} \quad (5)$$

where N_i is the discordant reads number of variable site i and D_i is the total reads number covering the site. F is then the estimated contamination rate. For population-scale sequencing, samples show heavy deviation of F should be contaminated. However, the power of this method decreases with the reduced sample size. In this case, we recommend to take the sequencing error rate as the referred baseline, which is 1% for Illumina sequencing.

Besides male X chromosome, homozygous SNP sites in unique region of autosomes can be used in detecting within-gender contamination. The principle is quite similar except that A_i and B_i can take value MM, Mm and mm, corresponding to homozygote of reference allele, heterozygote and homozygote of mutant allele. And the equation to calculate F is:

$$F = \frac{\sum_i N_i(m|MM)+N_i(M|mm)}{\sum_i D_i} \quad (6)$$

The only problem is to guarantee the homozygosity of SNPs. Assuming a sample that has been genotyped at thousands of or more SNP sites of autosomes, we can calculate the fraction of reads that disagree with known genotypes at homozygous sites(F) and classify the outliers as contaminated samples.

IV. Detecting Female Sample Contamination on Y Chromosome Unique Region

For a normal female sample, few reads should be mapped onto the Y chromosome except the long homologous regions between X and Y chromosome. The higher the abnormal reads rate is, the severer the contamination will be. And the severest condition is the gender mistake. As the abnormal reads rate is too small to be compared among samples, we useread counts which were scaled with the total reads number across the wholegenome normalized to 10 Mb. If we observed that the normalized reads counts per 10 kb (denoted as CY) of females are much smaller than that of males, which coincided with our expectation, it can prove the feasibility of this method.

As our method is based on reads count, low coverage sequencing may induce reduced power of contamination detection, so we sequenced two (a male and a female) exomes (Agilent v2, 44Mb) into a deep depth of 36X and 41X. To exclude the probability that these samples may be contaminated, we performed genotyping of 18Kb sites of the two samples and compared with genotypes called by samtools [9] using sequencing reads. As a result, both samples reached a fairly high genotype concordance of 99.88%, indicating little contamination during library preparation and Illumina sequencing.

We next selected reads randomly from the two samples and mixed together, manually making a contaminated sample with known mixture rate.By aligning reads onto the unique region of Y chromosome and normalized to 10 Mb, we counted the CY value of unique region. As Fig.1a shows, only 3out of 10 Mbreads can be mapped onto 10 kb unique Y chromosomeregion of the female sample (mixture rate is 0). While formale sample (mixture rate is 100%), this statistic is as high as888, indicating a good differentiation degree of CY.As for mixture samples, the CY value increases with the contamination rate increase. (Pearson correlation coefficient, PCC=0.9997). If samples were extremely contaminated, one can suspect the real gender of the sample. *i.e*, this method could also be used to discriminate genders.

As for application, we sequenced 2,000 exomes (1000 males and 1000 females, unpublished) and performed the method. The result (Fig.1b) shows most female samples are not contaminated by males. If we set cutoff of contamination level as 0.05, that isCY=45, we got 3 outliers. Specifically, CY of the extremlycontaminated sample was quite close to the males and we suspect the real gender of this sample. As for the other two outliers, the estimated conmination rates are

0.4 and 0.1.

V. Detecting Male Sample Contamination on X Chromosome Unique Region

In order to detect male sample contaminated by female one, we did similar mixture to generate the in-silico contaminated sample with female mixture rate ranges from 0.01 to 0.9. For each sample, we did calculation at 1,959 HapMap SNP sites (CEU dataset, since sample used above are Europeans.) in X unique regions, requiring the sequencing depth $\geqslant$ 8X. Restricting on HapMap high coverage sites guarantees its polymorphism. The average number of reads that disagreewith the consensus base generated by samtools [9] (the so-called sequence calls) are calculated and plotted in Fig. 1 (c). Different from the result of female sample contamination method, the estimated contamination level is not linearly associated with the mixture rate. In fact, the method is sensitive for low level of contmination but weak to detect severve contamination with rate more than 0.5. The reason that the estimated contamination level was greater than 0 on clean samples can be explained by the sequencing error. We also applied this method on the 1000 male exomes and found the fraction of reads that disagree with sequence calls of most samples are closed to 0.5% (Fig. 1 (d)), which is the average error rate of the sequencing on our machine. The only one exception reaches a rate as high as 0.04, indicating a contamination level > 0.2 of this sample.

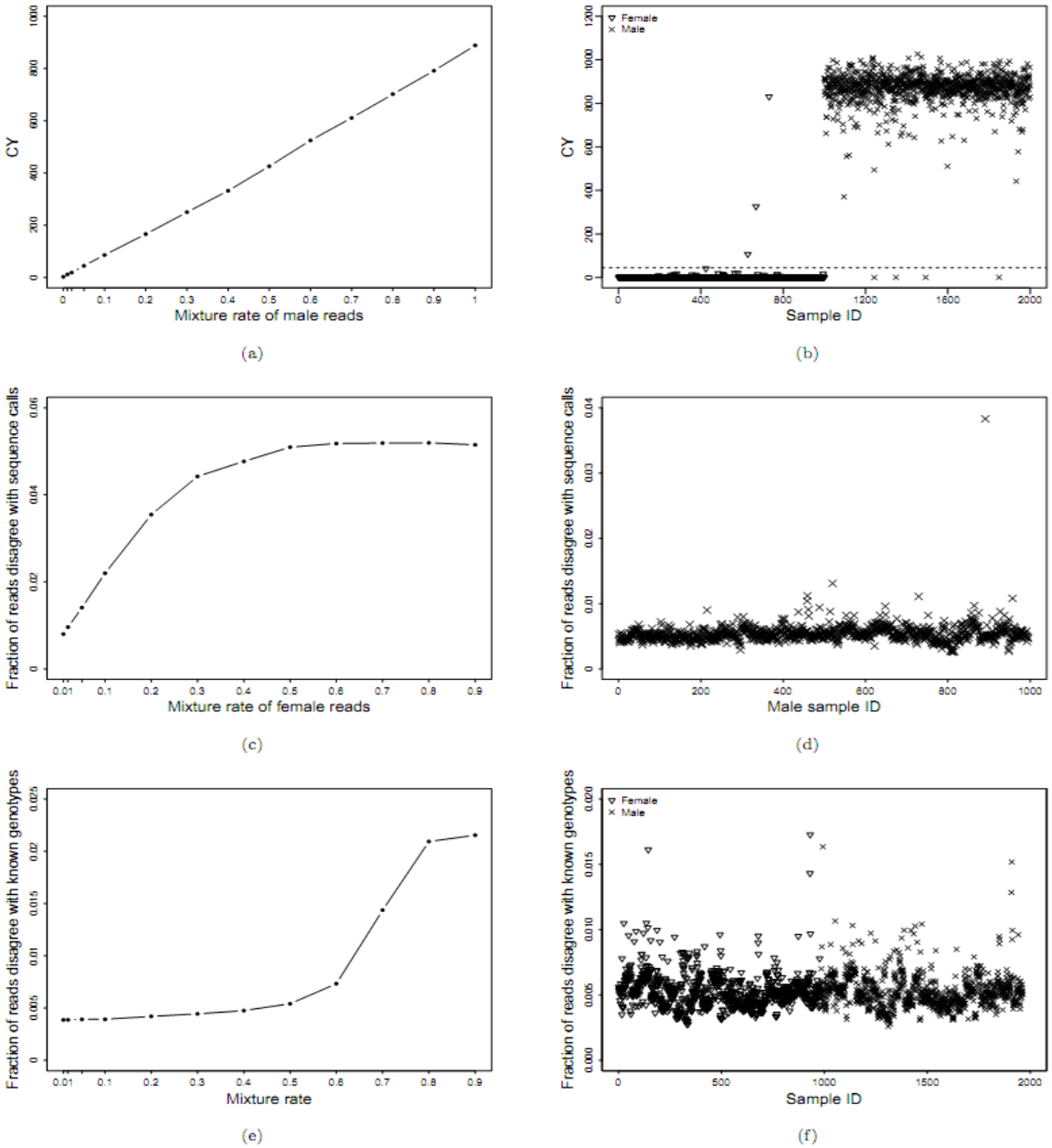

Fig. 1. Simulation and application result of three method for contamination detection. (a) Reads count on unique regions of Y chromosome (CY) of contaminated samples.Each dot represents a manually-made contaminated sample, with mixture rate as the percent of reads from the male sample. Specifically, the sample with 0 mixture rate is the sequenced female sample while the 100% one is the sequenced male sample. (b) CY of 1,000 female and 1,000 male samples. The dashed line is corresponding to the contamination level of 0.05. (c) Fraction of readsdisagree with the consensus bases on known HapMap SNP sites of the X chromosome unique region of contaminated samples. (d) Fraction of reads disagree with sequence calls on 1,000 male samples. (e) Fraction of readsdisagree with known homozygous genotypes on autosomes unique region of contaminated samples. (f) Fraction of readsdisagree with known genotypes on 2,000 samples.

VI. Detecting Male and Female SampleContamination on Autosome Unique Region

With the same principle described above, but not limited to male samples, we did calculation at 7,219 homozygous SNPsites in unique regions of 22 autosomes. This method could be performed on samples without gender information, yet reliable genotype data is needed. The SNP set was selected strictly from the 18 Kb genotyping sites. Reads that disagree with these known homozygous genotypes were counted and the fraction was plotted (Fig. 1 (e)). SNP sites with sequencing depth < 8X were skipped.

Although this method can be applied to all samples without prior knowledge of the gender, it has two main limitations. The first one is the referred known accurate genotypes. And the second one is from the method itself. As seen from Fig. 1 (e), we could conclude that the method is quite sensitive to severe pollution of the sample but less so when the samples is slightly polluted. The lack of power of the method applied to autosomes can be easily explained as follows: the genetic difference between humans are rather small except on sex chromosomes. Besides the homologous regions, X chromosome and Ychromosome are enormously differed, so male DNAcontamination can be easily detected from female samples, even at extremely amount.Similarly female DNA contamination can bring in largeheterogeneity to male X chromosome, making it a clear signof contamination. However, pollution between samples withsame sex is the most difficult to detect. Only severecontamination with high heterogeneity can be caught by our method. Even so, when applying our method on the 2000exomes, the resultswere better than had been looked for. Allcontaminated female samples detected by method 1 andcontaminated male samples detected by method 2 areevidently outliers in Fig. 1f, indicating that our method 3 is practical. Besides, two more outliers which cannot be detected by method 1 and 2 are identified, probably result from within-gender contamination.

VII. Conclusion

Taking advantage of the deep sequencing, our method is powerful to detect contaminated samples and ensure the accuracy of genotype calling. However, limitations still exist. The biggest one is that focusing on sequencing reads rather than genotypes makes it possible to detect slight contamination but can be biased estimation of the contamination rate when sequencing depth is low. To avoid the bias, we restricted the smallest covered reads count on a SNP site as 8. Another limitation is the relatively low detection power of contamination with same gender. To solve this problem, we strongly recommendedto apply our method in a population scale.

Although much to be done to the improvement of our ethod, the application can be expanded in addition to sample contamination detection. For example, the mappability score can be applied to the SNP quality control. Low mappability score suggests more than one copy on the genome and SNPs at this site might be the result of mis-alignment. As seen above, reads count on unique region of Y chromosome can be used to sex determination. Actually, we also counted reads on unique region of X chromosome and found the normalized count was similar to that of Y chromosome for male samples, consistent with the fact that male has one X chromatid and one Y chromatid. While for female samples, the reads count on X chromosome was double.

Within-species contamination is harder to detect and can result in greatly reduced genotype quality for sequencing studies, making it a great force to the improvement of sample quality control method.

Acknowledgment

I would like to thank Dr. Qibin Li for useful discussions about the original idea. I also appreciated Wei Jin for his guidance on modeling and Qijun Liao on programme debugging. I am also grateful to Prof. Xiao Sun for his many suggestions and clarifications on manuscript improvement.

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature biotechnology*, vol. 26, pp. 1135-1145, 2008.
[2] W. Langdon, "Mycoplasma contamination in the 1000 genomes project," *RN*, vol. 13, pp. 10, 2013.
[3] F. Jühling *et al.*, "tRNAdb 2009: compilation of tRNA sequences and tRNA genes," *Nucleic acids research*, vol. 37, pp. D159-D162, 2009.
[4] K. Cibulskis *et al.*, "ContEst: estimating cross-contamination of human samples in next-generation sequencing data," *Bioinformatics*, vol. 27, pp. 2601-2602, 2011.
[5] G. Jun *et al.*, "Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data," *The American Journal of Human Genetics*, 2012.
[6] B. Langmead *et al.*, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol*, vol. 10, pp. R25, 2009.
[7] R. Koehler *et al.*, "The uniqueome: A mappability resource for short-tag sequencing," *Bioinformatics*, vol. 27, pp. 272-274, 2011.
[8] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.
[9] H. Li *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078-2079, 2009.

Zheng Huang was born in August 1988 and admitted to the Department of Biomedical Engineering, which is belonging to the seven year system, in Southeast University in 2007. She was sent to BGI-Shenzhen in her junior year to study bioinformatics. There she was supervised by Dr. Qibin Li, who is the director of Population & Human Disease Research department. At the same time, she kept in touch with her master tutor Prof. Xiao Sun in the school.