Cryptography for genetic material

Sterling Sawaya

GeneInfoSec LLC, Boulder, Colorado, USA
sterlingsawaya@gmail.com

geneinfosec.com

June 29, 2017

Abstract

Genetic information can be highly sensitive and can be used to identified its source. To conceal genetic information, cryptographic methods can be applied to genetic material itself, concealing sensitive information prior to the generation of sequence data. The cryptographic method described here uses randomly divided subsets of barcodes and random pooling to securely generate pools of genetic material. The privacy obtained by these methods are measured here using differential privacy.

1 Introduction

Genetic data can contain sensitive health information, such as risk of cancer [Gen94] or neurodegenerative disease [DHMB$^+$11, RMW$^+$11]. Genetic data can also be used to identify its source (see [HSR$^+$08, Cla10, JYW$^+$09, VH09, EN14]). Furthermore, when genetic data is combined with other information about an individual, identification of that individual becomes even easier [EN14, HHH$^+$15]. This leads to important ethical considerations for those collecting genetic data for research or diagnostics [KBDV$^+$10, EWG$^+$14].

Consequently, many researchers have turned their attention to protecting genetic privacy. Informed consent can help the participants of a genetic study understand their risks, and frameworks are being developed to help guide scientists who collect genetic information [LCVC08, EWG$^+$14]. To allow genetic data to be shared between researchers, various methods of encryption can help prevent re-identification [CMM13, CCL$^+$, DDC14a, DDC14b, BBDC$^+$11, JWB$^+$17, KBLV13, KJLM08, KL15, LGDM10, Mal04, TJW$^+$16, XKB$^+$14, SSB16].

Those methods are designed for genetic data. The methods described here take a different approach. Here cryptographic methods are applied to genetic material itself, securing genetic information at the molecular level. This approach adds an additional layer of security, allowing genetic material to be sent to untrusted parties for analysis without revealing sensitive information to those parties.

The cryptographic method proposed here utilizes random molecular barcodes, sometimes referred to as unique molecular identifiers or tags. The barcoding of genetic material uses nucleotides as codes. The DNA nucleotides adenine, cytosine, thymine and guanine can be combined in a polymer to form a code. With only four nucleotide bases a large number of codes can be generated. For example, with a nucleotide composed of n or fewer bases, $\sum_i^n 4^i$ possible codes can be generated. Trillions of possible codes can be generated with short barcodes of only $n \leq 20$.

Various biotechnologies have found applications for random barcodes [SNP$^+$16, ZLZ$^+$14, SJSX12, SKJ$^+$16, LLC$^+$16, GBE$^+$15, IZJ$^+$14, KVK$^+$12, BRL$^+$15]. These technologies utilize barcoding at the molecular level to improve genetic sequencing accuracy, or examine unique and rare mutations present within a sample. The use of molecular barcoding has yet to be applied to genetic information security. Here, a cryptographic method using random molecular barcodes on genetic material is described.

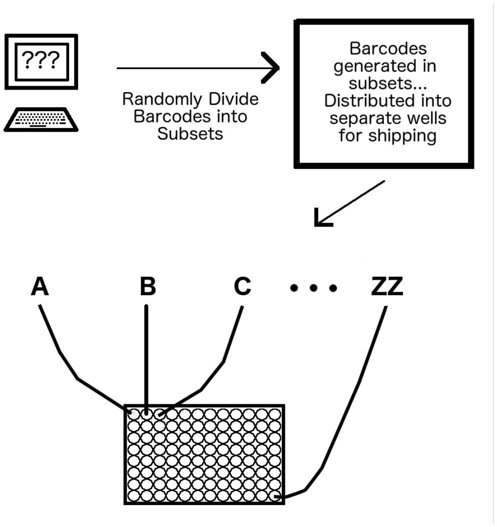
???
Randomly Divide Barcodes into Subsets
Barcodes generated in subsets... Distributed into separate wells for shipping
A
B
C
• • •
ZZ

Figure 1: **A set of barcodes is randomly divided into subsets.** Barcodes are generated in combination with nucleotides used to affix the barcodes to nucleic acids (e.g. adapters). Each subset of barcodes is sealed within a well on a plate. The barcodes present in each subset are only known to the consumer. Barcodes are randomly divided into unique subsets each time they are manufactured.

2 Methods

2.1 Overview

To begin this crypotgraphic method, genetic material is digested to form separate, unlinked molecules. Second, these molecules are each given random barcodes. Third, the barcoded molecules are combined with other genetic material that has been barcoded, which serve as decoys. Genetic information is concealed because the connections between variants are disrupted, hiding diploid genotypes and haplotypes. Also, the use of decoys can obfuscate which variants in the pool belong to the sample. This general method can be applied in a variety of ways, providing the ability to conceal different forms of genetic information with varying levels of security.

Both randomness and secrecy are required when generating the barcodes. This method takes a set of barcodes, and with secrecy, it randomly divides this set into subsets (Figure 1). Each subset is labelled, and a table of barcodes and labels is provided to the consumer (e.g. Table 1). The consumer can then use the subsets of barcodes to label their genetic samples so that only they know which barcodes belong to which sample.

Subset name	# in subset	Barcode sequence
A	1	ATCCCATGGTAGTCCTTAGA
A	2	CTTGGGAGTCTATCACCCCT
A	3	AGGGCCCATATCTGGAAATA
A	4	GACGCCAAGTTCAATCCGTA
A	5	TTCCGACGTACGATGGAACA
...	...	...
B	1	GTGTGGGTGAGACGTGCTTC
B	2	ATTTATACCCTACGCAGGCT
B	3	GGACCGAGGTCCGCAAGGCG
B	4	CGGCGGTGCACAAGCAATTG
B	5	ACAACTAACCACCGTGTATT
...	...	...
...	...	...
Z	9,999	CATTATGGTACCAGGGACTT

Table 1: **Example of randomized subsets of barcodes.** Barcodes are randomly divided into numerous subsets for each consumer. Only the consumer is provided the unique table that can be used to determine which barcodes are in which subsets.

The barcoding can occur in many ways. Barcodes can be added before or during an enrichment step. Barcoding after or without enrichment can result in a unique barcode for each molecule (Figure 2). If polymerase-chain-reaction (PCR) amplifies the genetic material after it has been barcoded, the resulting molecules would share identical barcodes, indicating they originated from the same molecule and belong to a particular sample (Figure 3). Either way, the barcodes act as random identifiers of the sample. After pooling and analysis, the table of barcodes is required to determine which results belong to which sample.

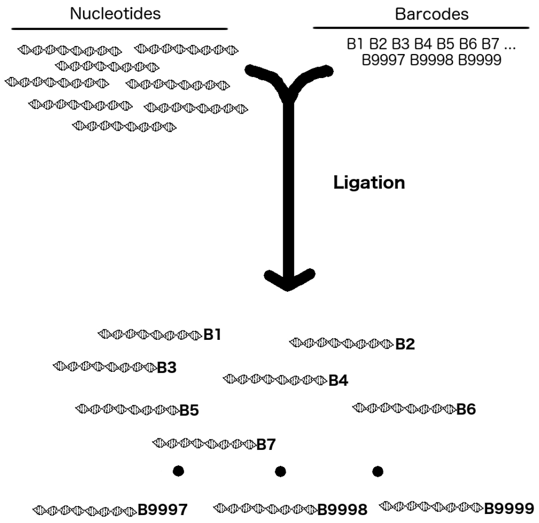

Figure 2: **Ligation of barcodes to nucleic acids.** A subset of barcodes is ligated to target nucleic acids, resulting in a unique identifier for each molecule. Only the consumer has knowledge about the barcodes used for a specific sample.

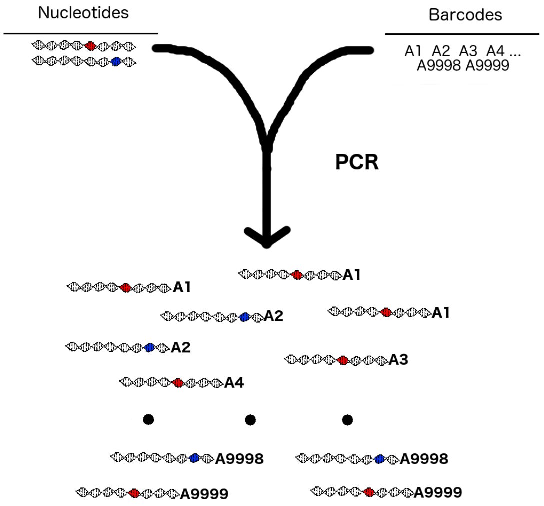

Figure 3: **Barcoding before or during enrichment.** Nucleic acids are amplified along with barcodes, resulting in multiple nucleic acids with identical barcodes. Nucleic acids that share the same barcode will have originated from the same molecule.

The method of pooling, in which decoys are combined with the sample(s), can be simple or complex, depending on the extent to which one choses to conceal their sample. A simple method of pooling is to barcode a group of samples and combine them together in a pool. These different samples would then act as decoys for each other. More advanced pooling methods can combine samples with non-sample decoys, chosen to conceal specific types of genetic information. The privacy that can be obtained with different pooling methods will be examined here.

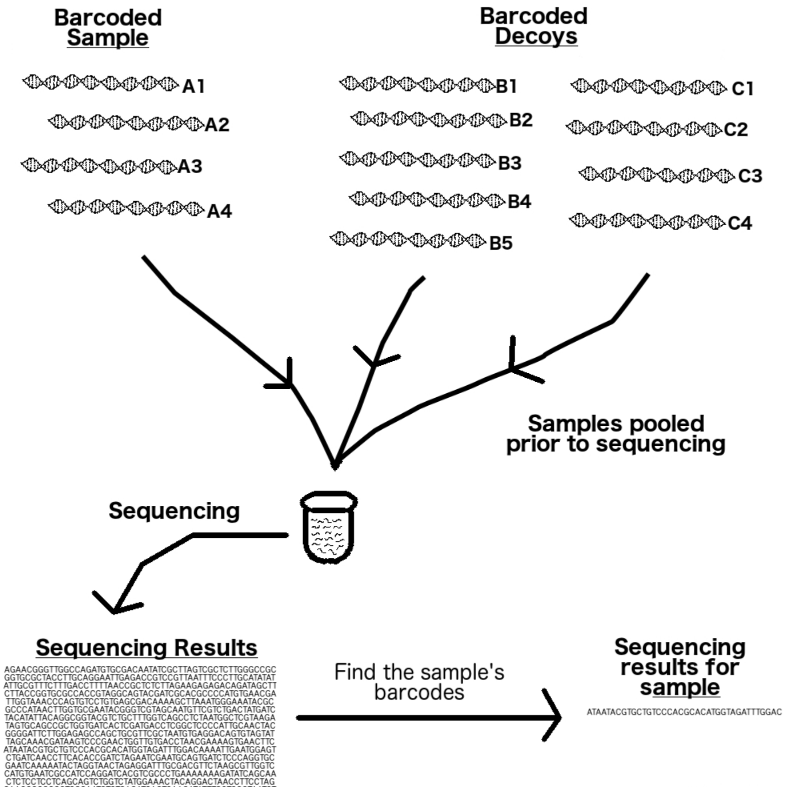
Barcoded Sample
A1
A2
A3
A4
Barcoded Decoys
B1
B2
B3
B4
B5
C1
C2
C3
C4
Samples pooled prior to sequencing
Sequencing
Sequencing Results
AGAACGGGTTGGCCAGATGTGCGACAATATCGCTTAGTCGCTCTTGGGCCGC
GGTGCGCTACCTTGCAGGAATTGAGACCGTCCGTTAATTTCCCTTGCATATAT
ATTGCGTTTCTTTGACCTTTTAACCGCTCTCTTAGAAGAGAGACAGATAGCTT
CTTACCGGTGCGCCACCGTAGGCAGTACGATCGCACGCCCCATGTGAACGA
TTGGTAAACCCAGTGTCCTGTGAGCGACAAAAGCTTAAATGGGAAATACGC
GCCCATAACTTGGTGCGAATACGGGTCGTAGCAATGTTCGTCTGACTATGATC
TACATATTACAGGCGGTACGTCTGCTTTGGTCAGCCTCTAATGGCTCGTAAGA
TAGTGCAGCCGCTGGTGATCACTCGATGACCTCGGCTCCCCATTGCAACTAC
GGGGATTCTTGGAGAGCCAGCTGCGTTCGCTAATGTGAGGACAGTGTAGTAT
TAGCAAACGATAAGTCCCGAACTGGTTGTGACCTAACGAAAAGTGAACTTC
ATAATACGTGCTGTCCCACGCACATGGTAGATTTGGACAAAATTGAATGGAGT
CTGATCAACCTTCACACCGATCTAGAATCGAATGCAGTGATCTCCCAGGTGC
GAATCAAAAATACTAGGTAACTAGAGGATTTGCGACGTTCTAAGCGTTGGTC
CATGTGAATCGCCATCCAGGATCACGTCGCCCTGAAAAAAAGATATCAGCAA
CTCTCCTCCTCAGCAGTCTGGTCTATGGAAACTACAGGACTAACCTTCCTAG
Find the sample's barcodes
Sequencing results for sample
ATAATACGTGCTGTCCCACGCACATGGTAGATTTGGAC

Figure 4: **Pooling barcoded nucleic acids**. The final step to concealing genetic information is the pooling of sample nucleic acids with decoy nucleic acids. A table of barcodes can be used to determine which sequences belong to which samples.

2.2 Measuring Privacy

The privacy of this process can be examined using the differential privacy method of [Dwo06] using the equation:

$$P[K(\mathcal{G}) \in S] \leq exp(\epsilon) \ \cdot \ P[K(\mathcal{G}') \in S] \tag{1}$$

in which ϵ represents the privacy obtained by removing a single individual's genetic material from $\mathcal{G}$ to obtain $\mathcal{G}'$. S is the range of possible outputs from the process K. A group of individuals with genotypes $\mathcal{G}$ is examined. The genetic material from $\mathcal{G}$ is converted into genetic data, $K(\mathcal{G})$, through process K. The probabilities are taken over "coin flips" of K.

This randomization procedure differs from those typically used in differential privacy. Here, one can rely on the randomness of a molecular process to obtain privacy, as well as a computer or coin to direct the randomization procedures. This process can be as simple as labeling a group of samples with random barcodes and pooling them together to be sequenced, or more complicated, mutating and amplifying sample genetic material to conceal specific mutations.

To demonstrate how one can measure privacy using this method, consider a scenario in which a small part of the individuals genomes are examined. Assume that only two variants are present within the population in this region. Denote alleles or haplotypes as "A" for the common variant and "a" for the alternative, less frequent variant. The frequency of the less frequent variant, denoted as p, is thus on $(0, 0.5]$. The frequency within the genetic data, however, is on $[0, 1]$, depending on the genotypes of the individuals being analyzed as well as the process by which the genetic data is generated. Here, the genotypes of the individuals in the pool are assumed to be in Hardy-Weinberg equilibrium.

Differential privacy methods are applied to a wide variety of data-sets, and variety of methods can be used to estimate differential privacy on these different data (e.g. [WLF16, BNS$^+$16, BD14]). Many methods examine the sensitivity of the randomization procedure, the maximum difference in output that can be generated [Dwo08], as:

$$\Delta f = \underset{\mathcal{G},\mathcal{G}'}{max} ||K(\mathcal{G}) - K(\mathcal{G}')||_1. \tag{2}$$

Denote G'_{aa} ($G'_{AA|Aa}$) as the pool of genetic material with the genetic material of an individual's with genotype aa (AA or Aa) removed. For many of the procedures examined here, removal of homozygous rare alleles results in the largest change in the output. That is, it can be shown that

$$||K(\mathcal{G}) - K(\mathcal{G}'_{aa})||_1 \geq ||K(\mathcal{G}) - K(\mathcal{G}'_{AA|Aa})||_1$$

for many procedures K. Consequently, concealing genetic information for individuals with uncommon genotypes is more difficult than concealing that of those with more common genotypes. Therefore, measuring the privacy obtained by removing an individual with the aa genotype determines the value of ϵ. If the frequency of the rare allele is small enough that no aa individuals are present in the pool, privacy is measured with $K(\mathcal{G}'_{Aa})$.

With the above considerations, equation [1] can be restated as:

$$\epsilon \leq log\left(\frac{P[K(\mathcal{G}) \in S]}{P[K(\mathcal{G}'_{aa}) \in S]}\right) \tag{3}$$

for many procedures K. This can lead to a comparison between Kullback-Leibler Divergence (KLD) [Kul97] and differential privacy [WLF16, BNS$^+$16, BD14]. The KLD between two distributions is a measure of the information gained if one distribution is used in place of the other. For two distributions, P and Q:

$$KLD(P||Q) = \sum_i P_i \cdot log\left(\frac{P_i}{Q_i}\right).$$

This can be related to privacy measured in [3], as the $E(\epsilon)$ is the same as $KLD(K(\mathcal{G})||K(\mathcal{G}'_{aa}))$, which has been termed "On-Average KL-Privacy" [WLF16]. Therefore, the differential privacy measured here can be interpreted as the difference in information between the entire pool of genetic material, and the pool with the aa genotype individual's genetic material removed (from the perspective of the original pool). That is, the information lost if $K(\mathcal{G})$ is encoded by $K(\mathcal{G}'_{aa})$.

Consider a pool of genetic material from N individuals that has been amplified X times, resulting in $2NX$ alleles in the pool. Simply pooling together genetic material and sequencing a portion of that pool is a randomization procedure that offers privacy. Sampling from a well mixed pool results in data for each allele that follows a hypergeometric distribution. If a sequencing method provides y total sequences, then the probability of z sequences with the "a" variant, $P_z(a)$, from a pool of size $2NX$ is:

$$P_z(a) = \frac{\binom{y}{z}\binom{2NX-y}{2NXp-z}}{\binom{2NX}{2NXp}}. \tag{4}$$

To measure privacy in this method, compare (4) with the probability of sequence results from the same pool with one individual with aa genotype removed, resulting in:

$$P'_z(a) = \frac{\binom{y}{z}\binom{2NX-2X-y}{2NXp-2X-z}}{\binom{2NX-2X}{2NXp-2X}}. \tag{5}$$

Measuring the KLD between these two distributions provides the expected KL-privacy for aa genotypes, setting the bound ϵ:

$$\epsilon \equiv KLD(P_z(a)||P'_z(a)) = \sum_{z}^{y} P_z(a) \cdot log\left(\frac{P_z(a)}{P'_z(a)}\right). \tag{6}$$

The variables here have significance in a genetic sequencing analysis, and thus their values must be chosen carefully so that the proposed analysis would be reasonable. The number of times an allele from an individual is sequenced in a pool, $P_z(allele)$, is also a hypergeometric function determined by the values of X and y chosen for a pool of N individuals,

$$P_z(allele) = \frac{\binom{y}{z}\binom{2NX-y}{X-z}}{\binom{2NX}{X}}. \tag{7}$$

Sequencing analyses are designed so that each allele from every individual achieves an appropriate number of reads. The average number of times an individual's allele will be sequenced is $y/2N$. This expectation, however, may not be likely in many designs, and a more appropriate design would use (7) to ensure that each individual receives sufficient coverage with an appropriate probability.

2.3 Advanced methods

First, consider that the amplification amount, X, can be a random variable. The randomness can be directed by a computer or by flips of a coin, randomly selecting an amount for each sample, X_i, to be added to the pool. Now, consider that the amplification occurs by PCR, and this process is intrinsically random. The randomness of amplification can be further randomized by a computer. For example, a computer can provide a random number of cycles of PCR by which each sample is amplified, or a random quantity of the various PCR ingredients to further vary the amount by which each sample is amplified.

Any random amplification procedure results in the X becoming a random variable. Here, the random amplification is applied to each individual sample, such that each allele in the sample receives the same amplification. Consequently, the privacy measured in (6) must then be measured over the possible values of X. The addition of randomness to the process in the amplification can increase privacy provided by this method. In fact, due to the imprecision of aliquoting genetic material, as well as the randomness that occurs in PCR, one may consider X to always be random. To help estimate the privacy the randomness of PCR can be modeled, e.g. [JK03, Pia04, YY09, LJJ05].

Mutation can also be utilized for privacy. Mutations randomly occur during PCR amplification [KZ15, PO17], with some polymerases having higher mutation rates, for various different types of mutation [PO17]. Furthermore, site directed mutagenesis, e.g. [Car86, HAT$^+$89, HHH$^+$89, LGH90, WCS$^+$94, KM97], can be utilized to obtain specific mutations from the sample. That is, a proportion of the sample can be (randomly) mutated to become decoys with specific variations to be added to the pool. Importantly, the mutated genetic material must be labelled uniquely and combined with uniquely labelled, non-mutated material, so that one can determine which sequencing results belong to the non-mutated genetic material. Methods using mutation can add additional sequencing costs because the sequencing of mutated genetic material usually does not provide useful information to the consumer. However, as sequencing costs continue to decrease, processes that include mutation will become increasingly cost effective.

Applying a random mutation step alters the equations used to estimate privacy. The total size of the pool of genetic material can be generalized, represented here by the variable Z. The pool Z can then be divided into the quantities of the separate alleles, here $Z = Z_A + Z_a = 2\sum_1^N X$. The number of specific alleles in the pool is the sum of the contribution from each genotype. For Z_a:

$$Z_a = 2\sum_{i=1}^{Np^2} X_i + \sum_{j=1}^{2Np(1-p)} X_j \tag{8}$$

and similarly for Z_A:

$$Z_A = 2\sum_{k=1}^{N(1-p)^2} X_k + \sum_{j=1}^{2Np(1-p)} X_j. \tag{9}$$

The hypergeometric (4) then becomes:

$$P_z(a) = \frac{\binom{y}{z}\binom{Z-y}{Z_a-z}}{\binom{Z}{Z_a}}. \tag{10}$$

Denote Z_a' as the number of "a" alleles in a pool from which an individual with genotype aa has been removed:

$$Z_a' = 2\sum_{i=1}^{Np^2-1} X_i + \sum_{j=1}^{2Np(1-p)} X_j. \tag{11}$$

The total resulting pool size for this pool, Z' is simply $Z = Z_A + Z_a' = 2\sum_1^{N-1} X$. With the new variables for the size of the altered pool and quantity of "a" alleles in the pool, (5) then can be generalized as:

$$P_z'(a) = \frac{\binom{y}{z}\binom{Z'-y}{Z_a'-z}}{\binom{Z'}{Z_a'}}. \tag{12}$$

If mutations are considered, then the final pool Z can be modeled as a mixture of alleles that have been replicated and mutated from an original pool of samples. Denote the mutations between alleles as:

$$A \underset{\mu_a}{\overset{\mu_A}{\rightleftharpoons}} a,$$

If the mutation between the two variants are equal ($\mu_a = \mu_A$), then the number of "a" alleles in the pool is:

$$Z_a = 2\sum_{i=1}^{Np^2} X_i(1-\mu) + \sum_{i=1}^{2Np(1-p)} X_j + 2\sum_{i=1}^{N(1-p)^2} X_k\mu \tag{13}$$

and the number of "A" alleles is:

$$Z_A = 2\sum_{i=1}^{N(p-1)^2} X_k(1-\mu) + \sum_{i=1}^{2Np(1-p)} X_j + 2\sum_{i=1}^{Np^2} X_i\mu. \tag{14}$$

Now consider the pool in which an individual with genotype aa has had their genetic material removed. Denote this comparison pool as $Z' = 2\sum_1^{N-1} X = Z_A + Z'_a$ and its quantity of "a" alleles as Z'_a, then:

$$Z'_a = 2\sum_{i=1}^{Np^2-1} X_i(1-\mu) + \sum_{i=1}^{2Np(1-p)} X_i + 2\sum_{i=1}^{N(1-p)^2} X_i\mu. \tag{15}$$

As before, privacy is measured by comparing (10) and (12) using (6).

3 Results

3.1 Non-random pooling

Privacy is measured for pools of varying numbers of individual samples, each sample pooled with equal proportions (Figure 5). Populations, N, of 400, 800 and 1,600 individuals, among a range of allele frequencies are examined, with the amplification $X = 10^6$. The number of reads is 8,000, 16,000 and 64,000 for 400, 800 and 1,600 individuals, respectively, providing a $P_{z\geq 5}(allele) > 0.97$, with an average of 10 reads per allele.

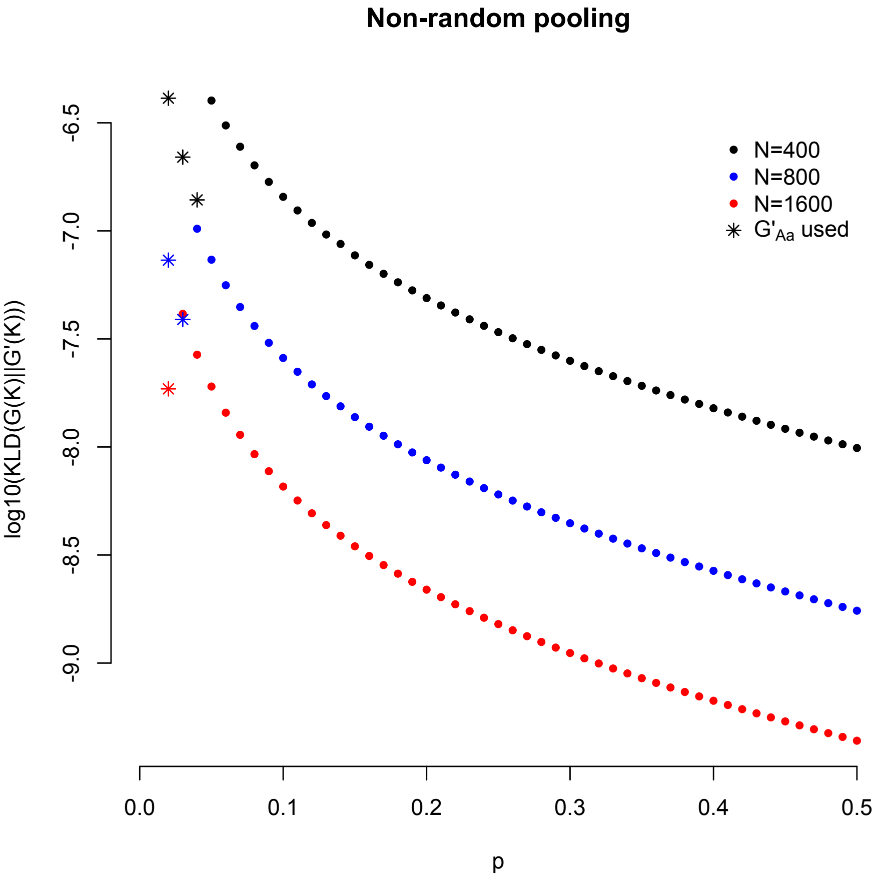

Figure 5: **Privacy obtained by pooling samples together**. The log (base 10) values of ϵ for different pool sizes and different allele frequencies. Samples are pooled together in equal proportions for N=400 (black), N=800 (blue), and N=1600 (red). Samples were amplified to have 10^6 molecules prior to pooling, and the pool was sequenced $20 \cdot N$ such that the average number of sequences per allele is 10. If the allele frequency is small enough that zero "aa" genotypes are expected in the pool, then privacy is measured by removing an "Aa" genotype individual from the pool instead of aa individuals (points plotted with $*$).

3.2 Random pooling with mutation

To examine random pooling and mutation, a population of 100 samples is used (Figure 6). For comparison, non-random pooling of each 100 samples is measured (black points). For random pooling, each individual has 1,000 of each of their alleles added to the pool, and then, for each sample, a coin is flipped two times, and an additional 1,000 of each allele is added for every flip that landed heads (red points). The resulting quantity of each allele for each sample in the pool follows a binomial distribution. The same random amplification method then is applied, but with 20% (orange points) and 40% (blue points) of the alleles mutated to the other variant. Random pooling provides more privacy than non-random pooling, and privacy is further increased if a mutation step is applied.

A random selection of 4,000 reads is then obtained from the pool, resulting in an average of 20 reads per individual when a mutation step is not applied. With a mutation step, some individuals receive far fewer reads and some receive far more (Figure 7). Furthermore, the mutated alleles, which have been uniquely tagged to indicate they are mutants, do not necessarily provide useful information about the sample. Consequently, pooling procedures which apply mutations result in fewer informative reads.

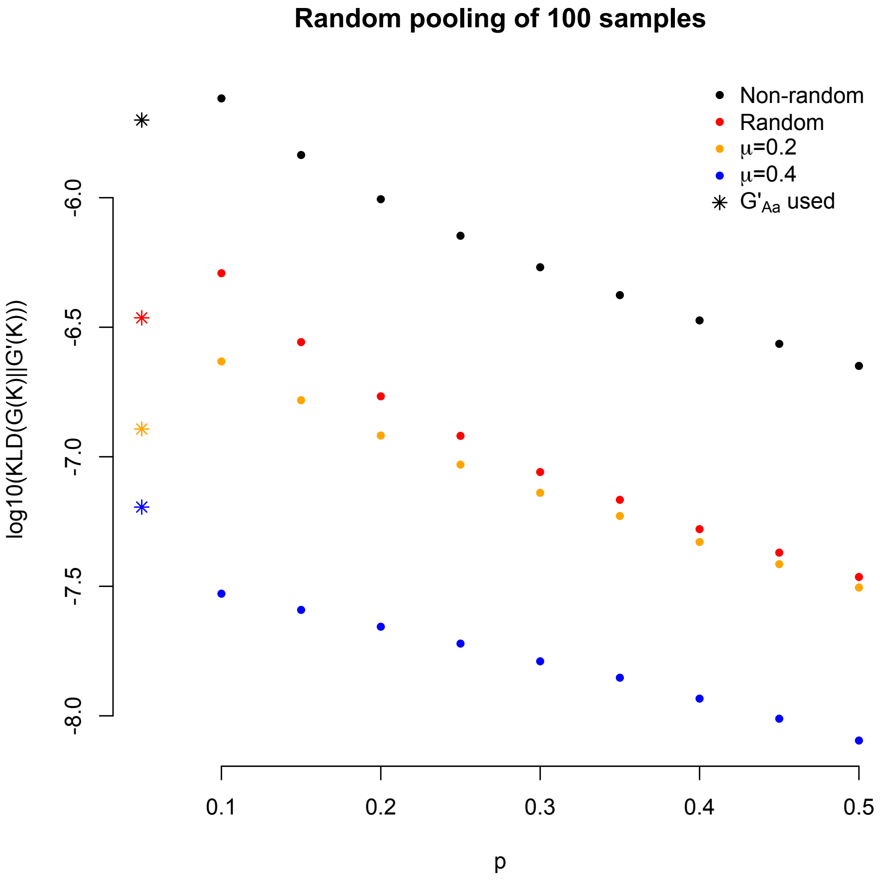

Figure 6: **Privacy obtained by randomly pooling samples together with mutation**. The log (base 10) values of ϵ for different pool sizes, different allele frequencies, for different pooling methods. Black points represent pooling samples of 100 individuals in equal, non-random proportions. Red points represent a random pooling of 100 individuals (see text for details). Orange (blue) points represent a random pooling of 100 individuals with a mutation rate, μ, of 0.2 (0.4). If the allele frequency is small enough that zero "aa" genotypes are expected in the pool, then privacy is measured by removing an "Aa" genotype individual from the pool instead of aa individuals (points plotted with $*$).

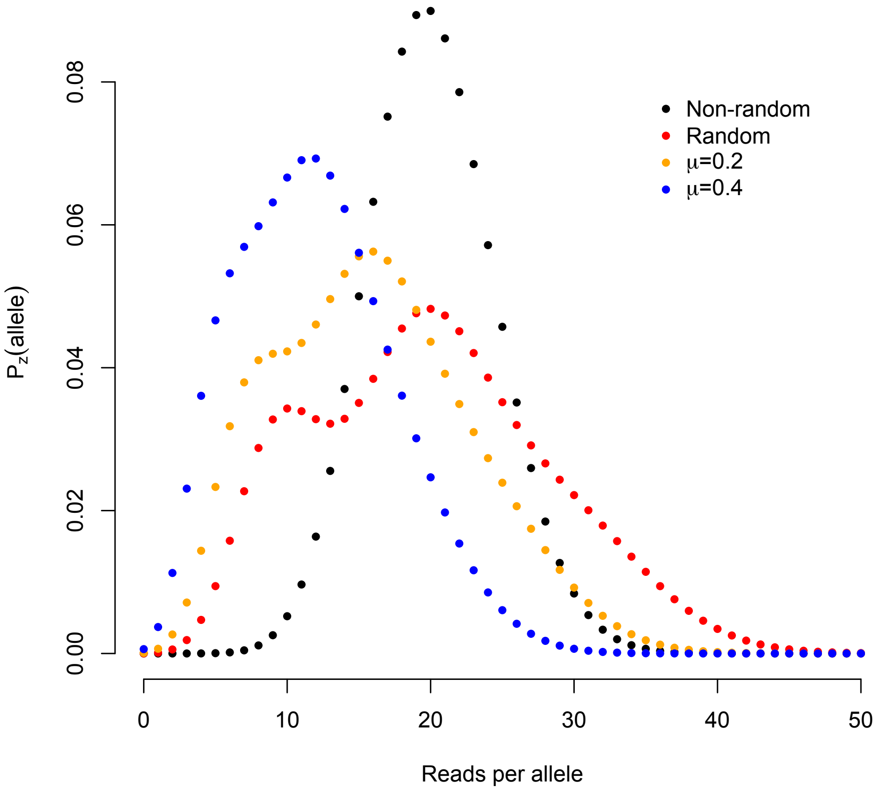

Figure 7: **Probability of read number for sample alleles for different pooling methods**. The number of reads for non-random pooling (black points) approximately follows a binomial distribution. Red points represent a random pooling of 100 individuals. Orange and blue points are random amplification, with a mutation step applied. The reads of mutated alleles are not counted, as they are not typically informative.

4 Discussion

The statistics presented here demonstrate that genetic information can be concealed with the methods proposed. The proposed methods have a wide range of potential applications in genetics, and are not limited to those examined in the results section. For example, the methods can be applied to more than one genetic loci, or to genetic loci that that have more than two states, such as tandem repeats. The use of differential privacy to measure the expected privacy of these methods can also be applied to these other applications. Alternatively, detailed models of adversarial knowledge can also be used to estimate privacy.

Any application of these methods requires careful attention to how information exists within target genetic material, and how it can be concealed. Because this method does not necessarily conceal all information present within genetic material, the consumer needs to decide which information they conceal, and how they conceal it. For example, the basic application of these methods does not conceal the allele frequency in the samples. Advanced applications can conceal the allele frequencies, but may require more sequencing reads to obtain the same amount of information. Importantly, due to correlations between genetic variants, applying these methods to multiple loci requires special considerations to estimate differential privacy (see [CYXX16, KM11] for estimating differential privacy with correlated data).

The advanced methods presented in the results section can be applied with readily available technology, a coin to randomly determine the quantity of genetic material to add to the pool and a controlled quantity of mutated alleles obtained through site directed mutagenesis. However, one can consider all molecular genetics lab work to have a degree of randomness (even when randomness is not desired). Consequently, all applications of this method will have more randomness, and thus usually more privacy, than the ideal applications measured here. Modeling PCR randomness to estimate privacy requires a detailed understanding of genetic information, DNA/RNA replication, as well as mutation rates. However, for many applications of this technology non-random pooling may be sufficient, allowing for relatively straightforward measures of privacy.

Ultimately, the details of the cryptographic method used by any consumer of this technology can be kept secret, and can be altered when applied to different groups of samples. This allows the consumer to control how they conceal their genetic information, further inhibiting potential adversaries from extracting useful data from the sequencing results. The appropriate use of securely generated random barcodes allows sensitive genetic information to be concealed within genetic material, securing it at its source. With hope, these methods will permit the collection of genetic information without jeopardizing sensitive information.

Competing interests

The author is a named inventor on a pending patent application directed to the methods described in this article

Acknowledgement

Thanks to Dr. Dylan Albrecht for helpful discussions on some of the statistics.

References

[BBDC$^+$11] Pierre Baldi, Roberta Baronio, Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Countering gattaca: efficient and secure testing of fully-sequenced human genomes. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 691–702. ACM, 2011.

[BD14] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[BNS$^+$16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

[BRL$^+$15] Erik Borgström, David Redin, Sverker Lundin, Emelie Berglund, Anders F Andersson, and Afshin Ahmadian. Phasing of single dna molecules by massively parallel barcoding. *Nature communications*, 6, 2015.

[Car86] Paul Carter. Site-directed mutagenesis. *Biochemical Journal*, 237(1):1, 1986.

[CCL$^+$] Gizem S Cetin, Hao Chen, Kim Laine, Kristin Lauter, Peter Rindal, and Yuhou Xia. Private queries on encrypted genomic data.

[Cla10] David Clayton. On inferring presence of an individual in a mixture: a bayesian approach. *Biostatistics*, page kxq035, 2010.

[CMM13] Christopher A Cassa, Rachel A Miller, and Kenneth D Mandl. A novel, privacy-preserving cryptographic approach for sharing sequencing data. *Journal of the American Medical Informatics Association*, 20(1):69–76, 2013.

[CYXX16] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy under temporal correlations. *CoRR*, abs/1610.07543, 2016.

[DDC14a] George Danezis and Emiliano De Cristofaro. Fast and private genomic testing for disease susceptibility. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 31–34. ACM, 2014.

[DDC14b] George Danezis and Emiliano De Cristofaro. Simpler protocols for privacy-preserving disease susceptibility testing. In *14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy (GenoPri'14). Amsterdam, The Netherlands*, 2014.

[DHMB$^+$11] Mariely DeJesus-Hernandez, Ian R Mackenzie, Bradley F Boeve, Adam L Boxer, Matt Baker, Nicola J Rutherford, Alexandra M Nicholson, NiCole A Finch, Heather Flynn, Jennifer Adamson, et al. Expanded ggggcc hexanucleotide repeat in noncoding region of c9orf72 causes chromosome 9p-linked ftd and als. *Neuron*, 72(2):245–256, 2011.

[Dwo06] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[Dwo08] Cynthia Dwork. *Differential Privacy: A Survey of Results*, pages 1–19. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[EN14] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.

[EWG$^+$14] Yaniv Erlich, James B Williams, David Glazer, Kenneth Yocum, Nita Farahany, Maynard Olson, Arvind Narayanan, Lincoln D Stein, Jan A Witkowski, and Robert C Kain. Redefining genomic privacy: trust and empowerment. *PLoS Biol*, 12(11):e1001983, 2014.

[GBE$^+$15] Mark T Gregory, Jessica A Bertout, Nolan G Ericson, Sean D Taylor, Rithun Mukherjee, Harlan S Robins, Charles W

Drescher, and Jason H Bielas. Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic acids research*, page gkv915, 2015.

[Gen94] Susceptibility Gene. Breast and ovarian cancer susceptibility gene brca1. *Science*, 266:7, 1994.

[HAT+89] Anne Hemsley, Norman Arnheim, Michael Dennis Toney, Gino Cortopassi, and David J Galas. A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic acids research*, 17(16):6545–6551, 1989.

[HHH+89] Steffan N Ho, Henry D Hunt, Robert M Horton, Jeffrey K Pullen, and Larry R Pease. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, 77(1):51–59, 1989.

[HHH+15] Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. De-anonymizing genomic databases using phenotypic traits. *Proceedings on Privacy Enhancing Technologies*, 2015(2):99–114, 2015.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167, 2008.

[IZJ+14] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, 2014.

[JK03] Peter Jagers and Fima Klebaner. Random variation and concentration effects in pcr. *Journal of Theoretical Biology*, 224(3):299–304, 2003.

[JWB+17] Karthik A Jagadeesh, David J Wu, Johannes A Birgmeier, Dan Boneh, and Gill Bejerano. Revealing the causative variant in mendelian patient genomes without revealing patient genomes. *bioRxiv*, 2017.

[JYW$^+$09] Kevin B Jacobs, Meredith Yeager, Sholom Wacholder, David Craig, Peter Kraft, David J Hunter, Justin Paschal, Teri A Manolio, Margaret Tucker, Robert N Hoover, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature genetics*, 41(11):1253–1257, 2009.

[KBDV$^+$10] Jane Kaye, Paula Boddington, Jantina De Vries, Naomi Hawkins, and Karen Melham. Ethical implications of the use of whole genome methods in medical research. *European Journal of Human Genetics*, 18(4):398–403, 2010.

[KBLV13] Liina Kamm, Dan Bogdanov, Sven Laur, and Jaak Vilo. A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics*, 29(7):886–893, 2013.

[KJLM08] Murat Kantarcioglu, Wei Jiang, Ying Liu, and Bradley Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on information technology in biomedicine*, 12(5):606–617, 2008.

[KL15] Miran Kim and Kristin Lauter. Private genome analysis through homomorphic encryption. *BMC medical informatics and decision making*, 15(5):S3, 2015.

[KM97] Song-Hua Ke and Edwin L Madison. Rapid and efficient site-directed mutagenesis by single-tube 'megaprimer'pcr method. *Nucleic acids research*, 25(16):3371–3372, 1997.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[Kul97] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[KVK$^+$12] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[KZ15] Justus M Kebschull and Anthony M Zador. Sources of pcr-induced distortions in high-throughput sequencing data sets. *Nucleic acids research*, 43(21):e143–e143, 2015.

[LCVC08] Jeantine E Lunshof, Ruth Chadwick, Daniel B Vorhaus, and George M Church. From genetic privacy to open consent. *Nature Reviews Genetics*, 9(5):406–411, 2008.

[LGDM10] Grigorios Loukides, Aris Gkoulalas-Divanis, and Bradley Malin. Anonymization of electronic medical records for validating genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107(17):7898–7903, 2010.

[LGH90] Olfert Landt, Hans-Peter Grunert, and Ulrich Hahn. A general method for rapid site-directed mutagenesis using the polymerase chain reaction. *Gene*, 96(1):125–128, 1990.

[LJJ05] Nadia Lalam, Christine Jacob, and Peter Jagers. Estimation of the pcr efficiency based on a size-dependent modelling of the amplification process. *Comptes Rendus Mathematique*, 341(10):631–634, 2005.

[LLC$^+$16] David F Lee, Jenny Lu, Seungwoo Chang, Joseph J Loparo, and Xiaoliang S Xie. Mapping dna polymerase errors by single-molecule sequencing. *Nucleic acids research*, page gkw436, 2016.

[Mal04] Bradley Malin. *Protecting dna sequence anonymity with generalization lattices*. Carnegie Mellon University, School of Computer Science [Institute for Software Research International], 2004.

[Pia04] Didier Piau. Immortal branching markov processes: averaging properties and pcr applications. *Annals of probability*, pages 337–364, 2004.

[PO17] Vladimir Potapov and Jennifer L Ong. Examining sources of error in pcr by single-molecule sequencing. *PloS one*, 12(1):e0169774, 2017.

[RMW+11] Alan E Renton, Elisa Majounie, Adrian Waite, Javier Simón-Sánchez, Sara Rollinson, J Raphael Gibbs, Jennifer C Schymick, Hannu Laaksovirta, John C Van Swieten, Liisa Myllykangas, et al. A hexanucleotide repeat expansion in c9orf72 is the cause of chromosome 9p21-linked als-ftd. *Neuron*, 72(2):257–268, 2011.

[SJSX12] Katsuyuki Shiroguchi, Tony Z Jia, Peter A Sims, and X Sunney Xie. Digital rna sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–1352, 2012.

[SKJ+16] Anders Ståhlberg, Paul M Krzyzanowski, Jennifer B Jackson, Matthew Egyud, Lincoln Stein, and Tony E Godfrey. Simple, multiplexed, pcr-based barcoding of dna enables sensitive mutation detection in liquid biopsies using sequencing. *Nucleic acids research*, page gkw224, 2016.

[SNP+16] Kamran Shazand, Jing Ning, Anthony Popkie, Egon Ranghini, and John Paul Jerome. High efficiency detection of low frequency alleles in cell-free dna, 2016.

[SSB16] Sean Simmons, Cenk Sahinalp, and Bonnie Berger. Enabling privacy-preserving gwass in heterogeneous human populations. *Cell Systems*, 3(1):54–61, 2016.

[TJW+16] Haixu Tang, Xiaoqian Jiang, Xiaofeng Wang, Shuang Wang, Heidi Sofia, Dov Fox, Kristin Lauter, Bradley Malin, Amalio Telenti, Li Xiong, et al. Protecting genomic data analytics in the cloud: state of the art and opportunities. *BMC medical genomics*, 9(1):63, 2016.

[VH09] Peter M Visscher and William G Hill. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet*, 5(10):e1000628, 2009.

[WCS+94] Michael P Weiner, Gina L Costa, Warren Schoettlin, Janice Cline, Eric Mathur, and John C Bauer. Site-directed mutagenesis of double-stranded dna by the polymerase chain reaction. *Gene*, 151(1):119–123, 1994.

[WLF16] Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

[XKB+14] Wei Xie, Murat Kantarcioglu, William S Bush, Dana Crawford, Joshua C Denny, Raymond Heatherly, and Bradley A Malin. Securema: protecting participant privacy in genetic association meta-analysis. *Bioinformatics*, page btu561, 2014.

[YY09] Andrei Y Yakovlev and Nikolay M Yanev. Relative frequencies in multitype branching processes. *The annals of applied probability*, pages 1–14, 2009.

[ZLZ+14] Zongli Zheng, Matthew Liebers, Boryana Zhelyazkova, Yi Cao, Divya Panditi, Kerry D Lynch, Juxiang Chen, Hayley E Robinson, Hyo Sup Shim, Juliann Chmielecki, et al. Anchored multiplex pcr for targeted next-generation sequencing. *Nature medicine*, 20(12):1479–1484, 2014.