Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions

Arif Harmanci, Mark Gerstein

Abstract

The functional genomics data is emerging as a valuable resource for personalized medicine. Although one might think that the functional genomics data is safe to share, the extent to which they leak sensitive information is not well studied. Here, we show for the first time that the read depth signal profiles for several functional genomics data types can cause concerns for privacy. A signal profile is generated by counting the number of reads at each genomic position. We show that there is significant leakage from the signal profiles of a number of sequencing based functional assays including RNA-seq, ChIP-Seq, and Hi-C. We demonstrate that an adversary can predict small and large deletions and use those to accurately cross-reference an individual among a large pool of individuals in a linking attack. We also propose a metric to measure the accuracy of genotyping the deletion variants using signal profiles. To show the practicality of linking attacks through signal profiles, we present several outlier based genomic deletion genotyping methods that lead to accurate linking attacks. We finally present a novel and effective anonymization procedure for protection of signal profiles against genotype prediction based linking attacks.

1. Introduction

Individual privacy is emerging as an important aspect of biomedical data science. A deluge of genetic data is being generated with the Cancer Moonshot Project\cite{XXXX}, Precision Medicine Initiative\cite{XXXX}, and UK100K from hundreds of thousands, if not millions, of individuals. Moreover, there is much effort to make genetic data more prevalent in the standard of care \cite{XXXX}. This will increase personal genomic data storage in healthcare providers. Leakage of the genetic information creates many privacy concerns, e.g. genetic predisposition to diseases may bias insurance companies. The initial studies on genomic privacy has focused on protection of single nucleotide polymorphism (SNP) datasets and analysis of privacy of the participants in genetic studies [1, 2]. It is worth noting that cryptographic approaches are also utilized for protecting genetic information \cite{XXXX, XXXX}. The significant increase in available datasets has made genomic linking attacks much more relevant [3–5].

A very relevant example of these attacks took place in the Netflix Prize Competition[3]. In this competition, a training dataset was released by the movie rental company Netflix, which was to be used for training new automated movie rating algorithms. The dataset was anonymized by removing names. Two researcher have shown that this training dataset can be linked to a seemingly independent database of IMDb web site and revealed movie preferences and identities of many Netflix users. We believe this will be a significant route to breaches in individual genomic privacy. Most of the previous studies focus on leakage of single nucleotide polymorphisms (SNPs) genotypes as a source of sensitive

"OFT SHARED"

MIRS

Deleted: Privacy has recently emerged

Deleted: an important aspect of genomic data sharing. The studies on genomic privacy has focused mainly on protection of genomic variants.

Deleted: has been largely assumed

Deleted: recent studies has shown that

Deleted: may

Deleted: substantial amount of individual characterizing information. For example, it has been recently shown that the gene expression quantification datasets can leak significant amount of

Deleted: in the context of linking attacks. ¶
In this paper

Deleted: analyze

Deleted: leakage from

Deleted: of

Deleted: datasets. In particular we focus on

Deleted: datasets

Deleted: analyze

Deleted: predictability

Deleted: from

Deleted: genotype prediction

Deleted: have high

Deleted: accuracy

Deleted: method

Deleted: . These focus on Homer et al's landmark study

Deleted: [1] showed that a simple statistic can be used to detect the existence of an individual within a sample of individuals for which only the allele frequencies are known. Im et al[2] proposed methods for detection of participants of QTL studies using phenotypic information. There is also much effort to develop the privacy preserving analysis methods of genetic data. Cryptographic approaches are the major workhorses, e.g. like homomorphic encryption \cite{XXXX, XXXX, XXXX}, in this area. The increase in the genomic datasets has now made the linking attacks possible where multiple seemingly unrelated datasets can be linked together to reveal sensitive information about the individuals in the datasets. A very relevant example of these attacks took place in the Netflix Prize Competition.

Deleted: A recent example of a linking was presented in Gymrek et al's study[3] where the authors demonstrated that several participants of the 1000 Genomes Project[4] can be identified within the public database of a popular genealogical genetics services company. Recently, Schadt et al[5] showed that the gene expression levels can be used ...

information. There are two major aspects that are not well addressed in the previous studies. Firstly, although it is well known that the major portion of individual genomic polymorphism is structural variants: deletion, insertion, translocation, and transversion of large chunks of DNA sequence, these did not receive much attention in the debate of genomic privacy[6]. The structural variants can have much more significant effects than SNPs, which may render some of the SVs more predictable compared to SNPs. Secondly, functional genomics data is not in center of the most studies. Especially the newer functional genomics datasets based on sequencing assays, like RNA-Seq[7] and ChIP-Seq[8] are very rich sources of information that can lead to leakage of individual characterizing information. In general, the raw sequenced reads from these experiments are not shared because of privacy concerns. File formats like MRF[9] and tagAlign can enable removing raw sequence information from reads while keeping the information about read mapping intact. These reads can be used to create the genome-wide signal profiles by piling them up along the genome. Indeed, the genome-wide signal profiles are publicly shared by many projects like ENCODE[10], Roadmap Epigenome Mapping Consortium[11], and GTex[12, 13]. It is urgently necessary to evaluate the sensitive and characterizing information leakage from these data types.

In this paper, we analyze the leakage in the signal profiles of several sequencing based functional genomics datasets. By signal profile, we refer to the signal generated by counting the number of reads that overlap with each nucleotide on the genome. Although the signal tracks do not contain any sequence information, an adversary can utilize signal processing techniques to detect the large and small structural variants. The most notable of these variants are the small and large deletions. Many methods have been developed to identify genomic deletions and duplications from the DNA-sequencing read depth signal [14] \cite{XXXX}. On the other hand, detection of structural variants from functional genomics datasets is not well-studied. The main reason for this is the dynamic and non-uniform nature of the signal profiles of functional genomics experiments, unlike DNA-sequencing signal profiles that uniformly cover the genome. For example, RNA-seq[7] and ChIP-seq[8] signal profiles concentrate mainly on the exonic regions and promoters of the genome, respectively. Recently, Hi-C[15] is emerging as a functional genomics assay that is used to detect small and large variants[16]. Moreover, these experiments are generally done in combination. In aggregate, multiple functional genomics assays can be utilized for accurately detecting large genomic variants. We show that this strategy is useful especially for detecting and genotyping small and large deletions because their effect is immediately observable in the signal profiles. We also show that the detected deletions can be used in a successful linking attack.

The paper is organized as following: We propose a new metric for quantifying how correctly genotypes of small and large deletion variants can be estimated. In combination with information content of the deletion variants, we use this new metric for evaluating the extent of characterizing information leakage from functional genomics datasets. We next present several practical instantiations of linking attacks that utilizes deletion variant genotype prediction using outlier signal levels. Finally To protect the signal profiles against linking attacks, we present a novel signal processing methodology for anonymizing the signal profile. We show that it is effective in decreasing the predictability of deletion variant genotypes from signal profiles. The source code for linking attacks and anonymization can be downloaded from privaseq2.gersteinlab.org.

Deleted: In general, the reads from functional genomics experiments can leak much genetic information and are not shared. The signal profiles that are generated from pileup of sequencing reads are, however, assumed safe to be publically shared.

Deleted: In this paper, we analyze the leakage in the signal tracks of several functional genomics datasets. Although the signal tracks do not contain any sequence information, an adversary can utilize signal processing techniques to detect the large and small structural variants. The most notable of these variants are the small and large deletions. Many methods have been developed to identify genomic deletions and duplications from the DNA-sequencing read depth signal [9] \cite{XXXX}. Although it is generally assumed that the structural variants cannot be reliably identified from functional genomics datasets, we show that an adversary can identify enough number of deletion variants in a linking attack. We show that the approaches we present are effective in linking genotype datasets to signal profiles datasets for different functional genomics assays, e.g. RNA-seq, ChIP-seq, and Hi-C experiments. The leakage analysis from the RNA-seq signal profiles is especially important because these data are very broadly shared by many projects like GTex project.¶

Deleted: also

Deleted: also

Deleted: and

2. Results

2.1. Genome-wide Linking Attack Scenario

Figure 1 summarizes the linking attack scenario that we focus. The adversary has access to a leaked structural variation (SV) dataset and another molecular phenotype dataset that contains genome-wide functional genomics signal profiles for example RNA-seq or ChIP-Seq signal profiles. The SV dataset contains identifying sample IDs for each individual and SV genotypes for multiple locations. We assume that the SV dataset comprises different types of variants like deletions, duplications, and translocations. The phenotype dataset also contains very sensitive information, i.e. HIV status, about the individuals. He/She uses the signal profiles to perform SV genotype prediction. He/She then compares the predicted SV genotypes and the leaked genotype dataset. The results are used to link the genotype samples to the phenotype samples and the HIV status of genotype samples are revealed to the adversary.

2.2. Information Content and Correct Predictability of Structural Variant Genotypes

It has been observed that the prediction of SV genotypes from functional genomics signal profiles has relatively low accuracy. In order to assess the predictability of SV genotypes, we propose using a measure named genome-wide predictability of SV genotypes, denoted by π_{GW}, from signal tracks. The predictability measures how accurately an SV genotype can be estimated given the signal profile (Methods Section). Given the genotype of a variant, the predictability is the conditional probability of the variant genotype given the signal profile. By this definition, the predictability only depends on the genomic signal levels of an individual and how well they can be used to predict genotypes. In principle, the genome-wide predictability is computed for each individual separately and independently from other individuals. Because of this fact, the predictability is independent of the population frequency of the variants.

Other than the predictability, an important measure in the linking attacks is the information content each SV genotype supplies. We utilize a previously proposed metric termed individual characterizing information (ICI)[1] to quantify the information content of each SV. This measure gives higher weight to the genotypes that have low population frequency and vice versa. For a given variant genotype, ICI measures how much information it supplies for pinpointing an individual in a population. As we discussed above, the genome-wide predictability is independent of the population frequency of the variants. Therefore the adversary can utilize genome-wide prediction approaches and predict rare variant genotypes to gain high ICI and characterize individuals very accurately. This is one of the major differences between genome-wide prediction approach proposed in this study and the recently proposed sample-wide prediction [4] based approach (Supplementary Fig 1). To compare these two approaches, we computed the sample-wide predictability of all the genomic deletions from the 1000 Genomes Project using the gene expression quantifications from GEUVADIS project [6, 17] (Supplementary Fig 1). ICI versus π_{SW} plot shows that there are not many SVs that have high predictability and high information content. One reason for this is that a significant number of SVs that impact gene expression levels have low population frequency and their sample-wide predictability are rather low. This implies that the gene expression levels can be shared without high risk of individual characterization using SV genotype prediction. However, as we will show later, a large fraction of these low frequency SVs have high genome-wide predictability and they can be used in individual characterization and identification.

Deleted: [6]

Deleted: [10, 11] (Supplementary Fig 1).

2.3. Genome-wide Linking with Short Deletion Prediction from RNA-Seq Signal Profiles

We first focus on predictability of short deletions using RNA-seq signal profiles (Fig 1b). Each deletion is manifested as an abrupt dip in the signal profile. The prediction of a deletion is done by detecting these dips in the signal profiles. The genome-wide predictability (π_{GW}) of the small deletions quantifies how well the adversary can identify the dips from the signal profile (Methods Section).

We computed genome-wide predictability for short deletions in 1000 Genomes Project using the RNA-seq expression signal profiles from the GEUVADIS project. Figure 2a,b show π_{GW} vs ICI for short deletions, for genotyping of known deletions (Methods Section). For both cases, there is a substantial number of deletions that have much higher predictability compared to a randomized dataset where the signal profile is randomized with respect to location of deletions. There are also many more variants with very high ICI (on the order of 5-6 bits) with high predictability. In comparison to sample-wide predictability of genotypes, there are a lot of deletions that provide deletions with very high ICI (higher than 5 bits) with high genome-wide predictability (Supplementary Figure S1).

In order to present practicality of small deletion predictability and information content, we propose an instantiation of a linking attack where we utilize outlier signal levels in the signal profiles for prediction of small deletion genotype prediction (Methods Section). For each individual, the prediction method sorts the short deletions with respect to *deletion-to-neighbor signal ratio* and assigns homozygous genotype to a number of deletions with smallest *deletion-to-neighbor signal ratio* (Methods Section). The adversary then compares the assigned homozygous deletion loci to the genotype dataset and identifies the individual whose deletion genotypes that are closest to the predicted genotypes. Thus, the attacker utilizes the outliers in *deletion-to-neighbor signal ratio* to predict genotypes and identify individuals. In order to minimize the bias on the variant call set, we used the known deletions with minor allele frequency greater than 1% in this analysis. Also, we extended the genotype dataset by re-sampling 1000 Genomes deletion dataset. Figure 2c shows the accuracy of linking versus number of deletions used in linking attack. The linking is perfect when the adversary utilizes more than 40 deletions. The attacker can also perform linking by first predicting *existence of deletion* (Fig 2c) and using the identified deletions to perform linking (Methods Section). When he/she utilizes this criteria, around 60 deletions are required for perfect linking.

We also studied the scenario where the adversary does not have access to the deletion loci but aimed at finding deletions and estimating their genotypes at the same time. This is a harder linking problem because the adversary must also correctly find deletion variants. We call this, linking attack based on joint deletion discovery and genotype prediction. Figure 2d shows that the linking accuracy is maximized (around 60%) when the attacker utilizes the top 50 deletion candidates in linking. If the attacker uses the existence of variant criteria in linking, the linking accuracy decreases.

In the previous analysis, the SV discovery set and RNA-seq sample set are matching. Since this may introduce a bias, we studied linking attack where signal profiles are generated by the GTex Project Consortium [12, 13] and the small deletions are called in the 1000 Genomes Project. This way, the deletions are identified in 1000 Genomes individuals while the linking is performed for the individuals in GTex Project datasets. Moreover we merged the genotype dataset from 1000 Genomes and the genotype dataset from GTex project. We first computed π_{GW} versus ICI for the deletions and observed that there is substantial enrichment of deletions that have high predictability with high ICI compared to randomized datasets, when the known deletions are utilized (Fig 3b). For the case of joint deletion

discovery and genotype prediction, the number of highly predictable and high ICI variants decrease (Fig 3b). When known deletions are utilized in extremity based attack, the linking accuracy is close to 100% for approximately 20 variants (Fig 3c). When the attacker increases the number of variants used in the attack, the linking accuracy decreases. Although the number of variants increase (more ICI), the genome-wide predictability of variants decrease faster. When the attacker predicts existence of deletion, the accuracy is maximized at around 240 variants and decreases when the number of variants in linking is increased (Fig 3d). In addition, the linking accuracy for joint deletion discovery and genotype prediction is low (Results not included), which indicates that joint prediction and genotyping of small deletions does not have enough power to perform linking attacks through RNA-seq signal profiles.

2.4. Genome-wide Linking with Large Deletion Prediction from ChIP-Seq Signal Profiles

We next focused on predictability versus ICI of long deletions, which are longer than 1000 base pairs. In this analyses, we utilize the ChIP-Seq signal profiles. Several recent studies have generated individual level epigenomic signal profiles through ChIP-Seq experiments [18–20]. These studies aimed at revealing how the variants interact with the epigenomic signals, mainly the histone modifications. The histone modifications are especially useful for identifying deletion genotypes because some of them cover a large portion of the genome, which is useful for predicting deletion genotypes. We use these personalized epigenomic signal profiles for quantifying how much characterizing information leakage they provide. For any individual where there are multiple histone mark ChIP-Seq signals, we pool them then compute several features for each large deletion. These are then used for quantifying information leakage (Methods Section).

Deleted: [14–16].

First we computed π_{GW} versus ICI for the large deletions in 1000 Genomes Project. Figure 4a,b show π_{GW} versus ICI for the large deletions from the 1000 Genomes. We use the personal epigenome profiling ChIP-Seq datasets presented in studies by Kasowski et al and Kilpinen et al (Methods Section). Similar to the small deletion analysis, it can be seen that for both datasets there are many large deletions with high predictability and high ICI.

We next performed practical linking attack utilizing the genotyping of known deletions and deletion discovery followed by genotyping. We again utilize a variant of the outlier based genotype prediction in the linking attack. The genotype prediction is done as follows. The average pooled ChIP-Seq signal on each deletion is computed and the variants are sorted with respect to their average signal in increasing order. The deletions with smallest pooled ChIP-Seq signal are assigned homozygous deletion genotype. For the deletions with assigned genotypes, we identified the individual in genotype dataset whose genotypes match closest to the assigned genotypes. We repeated this linking attack with different number of windows and computed the accuracy of linking (Methods Section). Figure 4c shows the linking attack accuracy when known large deletions are used in linking. The linking accuracy reaches 100% with fairly small number of deletions for both datasets. For the scenario where the adversary first discovers deletions then genotypes them, i.e. deletion loci are not given, the accuracy is also very high with small number of identified deletions (Fig 4d).

An interesting question about histone modifications is which combinations of histones leak the highest amount of characterizing information. To answer this question, we studied the individual NA12878, for which there is an extensive set of histone modification ChIP-Seq data from the ENCODE Project [10]. We have evaluated whether different combinations of histone modifications render NA12878 vulnerable against a linking attack among 1000 Genomes individuals, which is illustrated in Fig 4e. In general, we

Deleted: [17].

have observed that NA12878 is vulnerable when the dataset combinations that cover the largest space in the genome. This can be simply explained by the fact that when histone marks cover more space, higher number of deletions can be predicted. For example, H3K36me3 and H3K27me3, an activating and a repressive mark respectively, are mainly complementary to each other and they render NA12878 vulnerable. In addition, H3K9me3, a repressive mark that expands very broad genomic regions, renders NA12878 vulnerable in several combinations with other marks. On the other hand, H3K27ac, an activating histone mark that spans punctate regions do not render NA12878 vulnerable.

2.5. Genome-wide Linking with Large Deletion prediction from Hi-C Matrices

We also asked whether a relatively new data type, Hi-C can be used for identification of genomic deletions. Hi-C is a high throughput method for identifying the long range genomic interactions and three dimensional chromatin structure[15]. It is based on proximity ligation of the genomic sequences that are close-by and high throughput sequencing of the ligated sequences. After sequencing data is processed, it is converted to a matrix where the entry (i, j) represents the strength of interaction between i^{th} and j^{th} genomic positions. To study leakage from Hi-C datasets, we again focused on NA12878 individual for whom Hi-C interaction matrices are generated at different resolutions [21]. In order to convert the matrix into a genomic signal profiles, we summed the interaction matrix along columns and obtained a signal profile along the genome (Fig 5a, Methods Section). Next we simulated an extremity based linking attack using the outliers in this signal profile: For all the large deletions in the 1000 Genomes, we computed the average Hi-C signal. We next sorted the deletions in increasing order and assigned top 1000 windows with homozygous deletion genotype. We next compared the predicted genotypes with all the genotypes in the 1000 Genomes project. NA12878 is vulnerable to this attack when the Hi-C contact matrix resolution (bin length) is 10 kilobases or smaller (Fig 5b).

2.6. Anonymization of Signal Profiles against Linking Attacks

An important aspect of the genomic privacy is risk management and protection of datasets. For protection, anonymization of the datasets is the most effective way to share the data publicly in a safe manner. The most effective way to protect against linking attack scenario is to ensure that the deletion genotypes are not predictable from the signal tracks. We believe RNA-seq signals are currently the most vulnerable against the linking attacks and protection of these datasets against prediction of deletion variants is most immediate. As we showed in previous sections, the small deletions are major source of leakage of genetic information from RNA-seq signal profiles. We propose systematically removing the dips in the signal profiles as a way to anonymize the RNA-seq signal profiles against prediction of small deletions. Specifically, we propose smoothing the signal profile using median filtering (Methods Section). We have observed that median filtering removes the dips in the signal that indicate deletions very effectively and while conserving the signal structure fairly well. To evaluate the effectiveness of this method, we applied signal profile anonymization to the RNA-seq signal profiles generated from the datasets generated by GEUVADIS Project consortium and the GTex Project Consortium. After application of the signal profile anonymization, we observed that the large fraction of the leakage is removed for GTex datasets (Supplementary Figure 6). For GEUVADIS datasets, there is still some leakage but the genome-wide predictability of the variants are decreased substantially (Fig 2a). We also observed that the extremity based linking attack proposed in the previous section is ineffective in characterizing individuals such that no individuals are vulnerable for GTex project and at most 1% of the individuals are vulnerable for GEUVADIS dataset. The anonymized signal profiles for GTex and GEUVADIS individuals can be downloaded from privaseq2.gersteinlab.org/Anonymized_Signal_Profiles

Deleted: [18].

Deleted: [19].

Deleted: This signal processing method has been applied previously for genomic signal correction in the multi-mappable regions of the genome [20].

3. Discussion

Our results show that an adversary can perform fairly accurate linking attacks for characterizing individuals by prediction of structural variants using functional genomics signal profiles. In addition, we also showed that the linking can be done by predicting fairly small number of variants (generally less than 100 variants). Although the functional genomics assays do not reveal the full spectrum of structural variants, our results show that these data leak enough information for individual characterization among a fairly large set of individuals. This can be rather problematic because several large consortia are offering these signal tracks publicly. For example GTex signal profiles are publicly viewable and downloadable through the UCSC Genome Browser. In addition ENCODE RNA-Seq and ChIP-Seq signal profiles for several personal genomes (NA12878 and HeLa-S3) are downloadable through the UCSC Genome Browser and ENCODE Project's portal.

We also proposed a new metric for measuring the predictability of deletions from signal profiles. It is worth discussing that the genome-wide predictability measure is complementary to a sample-wide genotype predictability measure, π_{SW}, proposed earlier [4]. π_{SW} represents a sample-wide predictability measure that is computed when genotypes are predicted from sample-wide datasets, for example sample-wide gene expression profiling datasets. Sample-wide predictability is suitable when adversary utilizes sample-wide phenotypic measurements to predict genotypes in a linking attack (Supplementary Fig 1). This scenario is meaningful when the variant genotype is of high frequency and affects phenotype among samples, e.g. quantitative trait loci. For the rare variants sample-wide predictability will not be effective because variant genotypes do not show much variation among samples. For these variants, genome-wide predictability can be computed for each individual separately (Supplementary Fig 1).

It has been shown that the sample-wide predictability is related to the population frequency (and to the information content) of the variant genotypes. For example, the genotypes that have high population frequency are easier to predict than lower frequency genotypes. This is, however, not true for the genome-wide predictability of variant genotypes because it is totally independent of population frequencies. In fact, genome-wide predictability is estimated for each sample separately while estimation of sample-wide predictability requires a sample of measurements from multiple individuals.

Deleted: [6]

4. Methods

Deleted: [[Praise the signal profile anonymization against deletion predictions?]]¶

4.1. Genome-wide Predictability and Individual Characterizing Information

4.2. Prediction of Small Indels from RNA-Seq Signal Profiles

4.3. Prediction of Large Indels from ChIP-Seq Signal Profiles and Hi-C Matrices

4.4. Extremity based Genotype Prediction and Instantiation of Linking Attacks

4.5. Anonymization of Signal Profiles

This signal processing method has been applied previously for genomic signal correction in the multi-mappable regions of the genome [22].

5. Datasets

[[The source and accession numbers of the datasets]]

[[Dataset processing:

GEUVADIS, GTex, 1000 Genomes]]

REFERENCES

1. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson J V., Stephan DA, Nelson SF, Craig DW: **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. *PLoS Genet* 2008, **4**.

2. Im HK, Gamazon ER, Nicolae DL, Cox NJ: **On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy**. *Am J Hum Genet* 2012, **90**:591–598.

3. Narayanan A, Shmatikov V: **Robust de-anonymization of large sparse datasets**. In *Proceedings - IEEE Symposium on Security and Privacy*; 2008:111–125.

4. Harmanci A, Gerstein M: **Quantification of private information leakage from phenotype-genotype data: linking attacks.** *Nat Methods* 2016, **13**:251–256.

5. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference.** *Science* 2013, **339**:321–4.

6. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, et al.: **An integrated map of structural variation in 2,504 human genomes**. *Nature* 2015, **526**:75–81.

7. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.

8. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6**:S22–S32.

9. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: A modular framework to analyze RNA-Seq data using compact, anonymized data summaries.** *Bioinformatics* 2011, **27**:281–283.

10. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.

11. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G: **Epigenomics: Roadmap for regulation**. *Nature* 2015, **518**:314–316.

Formatted: Heading 1, Outline numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0" + Indent at: 0.25"

Moved down [1]: Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference.** *Science* 2013, **339**:321–4.¶

Deleted: 4. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M: **A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.** *Nat Commun* 2016, **7**:11101.¶
5. Schadt EE, Woo S, Hao K: **Bayesian method to predict individual SNP genotypes from gene expression data**. *Nature Genetics* 2012:603–608.¶
6

Moved (insertion) [1]

Moved down [2]: 7. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**:57–63.¶
8. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6**:S22–S32.¶
9.

Moved down [3]: Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing**. *Genome Res* 2011, **21**:974–984.¶

Deleted: 10

Deleted: 11.

Moved (insertion) [2]

Moved (insertion) [4]

12. Consortium TG: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45**:580–5.

13. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, et al.: **The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans**. *Science (80-)* 2015, **348**:648–660.

14. Abyzov A, Urban AE, Snyder M, Gerstein M: **CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing**. *Genome Res* 2011, **21**:974–984.

15. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-C: a method to study the three-dimensional architecture of genomes.** *J Vis Exp* 2010, **6**:1869.

16. Korbel JO, Lee C: **Genome assembly and haplotyping with Hi-C.** *Nat Biotech* 2013, **31**:1099–1101.

17. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.

18. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK: **Identification of genetic variants that affect histone modifications in human cells.** *Sci (New York, NY)* 2013, **342**:747–749.

19. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, Udin G, Thurnheer S, Hacker D, Core LJ, Lis JT, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET: **Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.** *Science* 2013, **342**:744–7.

20. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek D V, Li J, Xie D, Olarerin-George A, Steinmetz LM, Hogenesch JB, Kellis M, Batzoglou S, Snyder M: **Extensive variation in chromatin states across humans.** *Science (New York, NY)* 2013:750–752.

21. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping**. *Cell* 2014, **159**:1665–1680.

22. Harmanci A, Rozowsky J, Gerstein M: **MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework.** *Genome Biol* 2014, **15**:474.

Moved down [5]: Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans.** *Nature* 2013, **501**:506–11.¶

Moved (insertion) [3]

Moved (insertion) [6]

Moved (insertion) [5]

Deleted: 15

Deleted: 16

Moved up [4]: Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.¶

Deleted: 18.

Moved up [6]: van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-C: a method to study the three-dimensional architecture of genomes.** *J Vis Exp* 2010, **6**:1869.¶

Deleted: 17.

Deleted: 19

Deleted: 20