RESPONSE TO REVIEWERS COMMENTS FOR “ANALYSIS OF SENSITIVE INFORMATION LEAKAGE IN FUNCTIONAL GENOMICS SIGNAL PROFILES THROUGH GENOMIC DELETIONS”

RESPONSE LETTER

-- Ref1.1: Introductory comments –--

Reviewer Comment	Built on previous work from the aspect of SNPs (published in 2016), here the authors expand onto structural variants (SVs), and onto functional genomics data such as RNS-seq and ChIP-seq.
Author Response	We sincerely thank the reviewer for the constructive comments, which we believe made our paper stronger. We respond to reviewer’s comments below.

-- Ref1.2: Introductory comments –--

Reviewer Comment	The authors’ analyses provided evidence that private indels and other SVs can be recovered from the raw reads from RNA-seq and ChIP-seq (histone modification) experiments. The deletions discovered from these raw data sets can be cross-linked by malicious attackers to potentially reveal the identity of the individual being sequenced. The authors proposed approaches such as smoothing the reads profile to remove the dips in the signal profile, which can alleviate the potential risk of information leakage.
Author Response	The reviewer’s comments summarize parts of our manuscript. We believe, however, we need to clarify some of it. The reviewer indicates that our analyses provides evidence that the private SVs can be recovered from the **raw reads from RNA-seq and ChIP-Seq experiments**. **This is not in our manuscript**. In fact, our analysis does not provide evidence that raw reads, by themselves, can be used to recover SVs. Our analysis does not use raw reads, at all. We would like to make sure this is very clear: The data from a functional genomics sequencing experiment is a very rich information source. The main purpose of the functional genomics experiment is to understand the differences in the regulation and expression of genes under different conditions, for example among individuals with cancer. However, these data may ostensibly leak variant information at the same time. This is generally not the

Deleted: ¶

Moved down [1]: The authors’ analyses provided evidence that private indels and other SVs can be recovered from the raw reads from RNA-seq and ChIP-seq (histone modification) experiments. The deletions discovered from these raw data sets can be cross-linked by malicious attackers to potentially reveal the identity of the individual being sequenced. The authors proposed approaches such as smoothing the reads profile to remove the dips in the signal profile, which can alleviate the potential risk of information leakage.

Moved (insertion) [1]

	intended purpose of these data. For example, the raw reads from an RNA-sequencing experiment contains nucleotides and these can be used to identify a large number of genetic variant like SNPs, and indels. These variants can be used by an adversary to breach individual's privacy. Thus, the raw reads are almost never shared. There is, however, a great incentive to share the data because they are invaluable resources for disease research. G One way to share the data, is that researchers generate aggregated data files from the raw reads and share these. These are seemingly free of variant information. For example, the read depth signal profiles are one type of data. These profiles are just counts of reads at each position on the genome and they do not have any nucleotide information immediately available. In general, the signal profiles are assumed free of variant information and are safe to share publicly. In fact, THE GTex Consortium generates RNA-sequencing data for hundreds of healthy individuals and the signal profiles for these data are shared publicly through UCSC genome browser. Our manuscript's main focus is this point: We are studying the leakage of variant information from the signal profiles. We show that the signal profiles may be used to detect and genotype small and large genomic deletions and these can be used to identify individuals within a large cohort. G It is important to note that there are other aggregated datasets that can be generated from raw reads and signal profiles. For example, the gene expression levels are computed by averaging the signal profiles over genes. For each gene, the expression level is the average RNA-seq signal that is observed on it. The gene expression levels can be used to detect variant genotypes using eQTLs and these can be used in a linking attack to identify individuals. This has been previously studied and **we are not considering this problem in our manuscript.** Our analysis is focused genotyping small and large deletions using **only the signal profiles from RNA-seq and ChIP-seq data**. Our results provide evidence that the signal profiles can leak enough genotypic information that can be used to pinpoint an individual. We would like to recapitulate this distinction: It is well known that the raw reads contain a very large amount of genotype information in them. It is therefore generally not acceptable to share raw reads from any sequencing experiment. The signal profiles are, however, not very well understood in terms of the privacy risks around sharing them. For example, GTex RNA-seq signal profiles can be found publicly in UCSC genome browser (See below and New

	Supplementary Figure S5.) Our manuscript sheds light on this issue. It is very crucial to make this distinction because our manuscript. To clarify the above point, we have made a new supplementary figure, Supplementary Figure 6, to illustrate the leakage from reads, signal profiles, and gene expression levels. We also updated the introduction and discussion sections to make it clear that the central theme of our study is the signal profiles and not the raw reads.
Excerpt From Revised Manuscript	**Introduction:** In this study, we analyze the leakage of sensitive information from the functional genomics data and how they can be used by an adversary in linking attacks. There are a number of motivating key points related to functional genomics data and privacy. First of all, functional genomics data, such as RNA sequencing data, is unique, in that if the data comes from human subjects, the raw reads have genetic variant information, which may be used to identify individuals. However, the main purpose of RNA-Seq data is not related to the variants; main purpose is more related to understanding dysregulation of genes under different conditions, such as cancer. Consequently, there's a great desire to share and study RNA-Seq datasets, to enable helping to find cures for various diseases. Because of this, there is great incentive to make ways of sharing functional genomics data without privacy protections. Large-scale privacy protections are a great encumbrance on genomic data sharing. They do not allow researchers and data owners to share results on the web, use web and internet-based tools, and they exert a great burden on research. Consequently, many consortia, such as GTEx, aim at sharing RNA-Seq information to the maximum extent. The raw reads obviously cannot be shared, as they contain variant information. However, there's belief that the signal files and the gene-level quantifications can be shared. The signal files simply reflect the overall depth of coverage of the RNA-Seq reads at any given point. Ostensibly, they're do not contain variant information. Many of the genomics consortia have decided to openly share RNA-Seq signal information. We show that there is a high degree of private information leakage in the function genomics signal profile data. The gene-level quantifications essentially are averages over the signal profile over exons. Although the overall averaging reduces information, private information leakage. However, there is also private information leakage through the association with variants called eQTLs. It is important to note that this is tackled in the current study, but is looked at elsewhere[16, 18]. … In this study, we analyze the sensitive information leakage from the signal profiles of several sequencing based functional genomics datasets. By

1-2A

GRAM ?

1-2B

BETTER WORD FOR MAIN TEXT

SEE 1-3 BELOW

signal profile, we refer to the genome-wide signal computed by counting the number of reads that overlap with each nucleotide on the genome. The signal profiles are just one type of aggregated data that is generated from raw reads. Another type of aggregated data is gene expression quantifications, which are averages of RNA-seq signal profiles over genes. The leakage of information from the gene expression quantifications has been previously studied[16, 18]. Rather, we are only considering whether the signal profiles have any genotypic information leakage from them. We show that signal profiles do leak a large amount of genotype information for small and large genomic deletions.

As discussed earlier, the raw reads from an RNA sequencing experiment contain the nucleotides themselves. It is well established that the raw reads must not be released publicly (Supplementary Figure 6) because given the raw reads, and adversary can identify a large number of private SNPs and indels. We therefore assume that the raw reads are not publicly shared and that the adversary does not have access to the raw reads. Rather, we assume that the data owners created the signal profiles and made these publicly available. The adversary gains access to these signal profiles. Regarding the signal profiles, it is generally assumed that the signal profiles are mostly void of sensitive information. Several large consortia, for example ENCODE Project[25], Roadmap Epigenome Mapping Consortium[26], and GTex[27, 28] publicly share signal profiles (Supplementary Figure 5)

Discussion:

Overall, at this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Sequencing, and point out the sources that we probed in this paper. First, there is the leakage directly from the reads. This is the most obvious leakage, and this leakage is avoided with by simply not sharing the raw reads. Next source of leakage is from the signal profile. This leakage is addressed in this paper. There is yet another source of leakage though, when one averages over the signal file, and produces quantifications in particular regions such as genes. These quantifications can be subtly connected with variants through the notion of eQTLs. This is not addressed in this paper, and there can be substantial leakage from these quantifications.

Furthermore, one can envision additional sources of leakage beyond that, in these main areas. For instance, one can imagine complex and subtle correlations between the levels of gene expression of many genes within pathways and networks. Although there has been interest in identifying these higher order QTLs, these are not yet extensively studied[28]. Complex machine learning techniques, such as deep learning, can reveal subtle correlations of gene expression at the network level with variants. Also, eQTLs traditionally have been linked to genes; ostensibly, one

	might imagine by averaging over various intergenic regions, some of the more highly expressed region to signal profile might also show correlations. This is another source of information not studied in this work. Finally, an additional source of information is, while we do look at calling of particular types of structural variants, such as small and large deletions, there may be very large-scale, megabase-scale deletions, which affect many genes. This is particularly the case for somatic events in cancer samples. This case is also not covered by our procedure. Finally, we would like to emphasize that we focused on a particular type of leakage of private information in functional genomics data, such as RNA-Seq data, such that the leakage stems from the signal profile. There are many other sources of information, however the signal file is currently at the junction between public and private information, and is where genomic information is begun to be shared publicly. Hence, we believe it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might unfortunately be the case that this type of information is not able to be shared publicly in the future, perhaps only sharing gene-level quantifications, or even worse, nothing at all. We wish to emphasize that, in this paper, we are not trying to look at all sources of leakage from functional genomics data, but just the sources of leakage right at the decision boundary of sharing and not sharing.

Formatted: Justified

Formatted: Normal

-- Ref1.3: I am doubtful that RNA-seq data is equally useful since the expression level of a gene can be influenced by a single nucleotide SNV (e.g. eQTL), or mutations (SNPs) in splice junction sites –--

Reviewer Comment	I like the concept introduced by the author "predictability of the SV genotype based on the observed signal profile". Figure 1C showed one nice example in which the absence of histone ChIP-Seq data is used to infer a genomic deletion event. I can imagine that histone modification data measured by ChIP-seq is useful in this regard, however I am doubtful that RNA-seq data is equally useful since the expression level of a gene can be influenced by a single nucleotide SNV (e.g. eQTL), or mutations (SNPs) in splice junction sites. I would like the authors to comment on these other confounding factors.
Author Response	We thank the reviewer for the insightful comment. We understand that the reviewer is concerned that deletions may not affect the gene expression as much as eQTLs and splice site mutations. Although we understand the reviewer's concern, we believe that the setup of the attack needs to be clarified: In attack scenario regarding RNA-seq data, we assume the attacker uses the signal

levels to find small deletions in the signal profile. These deletions manifest themselves as small but noticeable dips in the RNA-seq signal profiles. This is illustrated in a hypothetical example shown on the left panel of Fig 1d. A real example of how a small deletion manifests itself on an RNA-seq signal profile is shown in Supplementary Figure 5. A simplified version is included below for reference. This figure shows the screenshot of a UCSC genome browser signal track of GTex whole blood RNA-seq signal profiles. The 2 base pair deletion (rs34043625) in 3 GTex individuals can be seen even by eye easily. It is also worth noting that these tracks are publicly available for viewing and download. Another important aspect of these dips is that the signal in the dips are much smaller compared to the changes in the gene expression caused by the eQTL and sQTLs. The eQTLs generally cause changes in the total signal in the signal profile of a gene while the small deletions create localized changes in the signal profile and these are relatively smaller compared to the effects of eQTLs and sQTLs, assuming that the deletion is not an eQTL itself.

The sensitive information leakage is caused by the fact that these dips reveal small deletions (i.e., shorter than 10 bps) to the attacker. When the attacker identifies these dips, she (assuming the attacker is female) can use those to link the RNA-seq signal profile to the genotype data. One could argue that there may not be enough small deletions in the transcriptome, i.e., the regions of the genome where RNA-seq signal is present. This is why we performed the linking attack and showed that the small deletions that leak from RNA-seq signal profiles can be used to link individuals correctly.

We believe that the confusion stems from the fact that the setup of the problem is not made clear. We review it here for clarity. RNA-sequencing datasets is a very rich information source. There is currently a great desire to generate and share these data. But unlike DNA sequencing of genomes, the RNA-seq data is different in the sense that the main purpose of the data is not finding variants. The main purpose is identifying which genes are more active in a certain condition compared to another condition. Although it is not the main purpose of the data, there is genetic variant information in RNA-seq data. This is what makes this data problematic in terms of privacy. Because the raw reads from an RNA-seq experiments contain the nucleotides and an adversary can use these to find a very large number of variants. These variants will cause concerns for individual privacy. In order to share these data, several aggregated formats have been developed and shared. For example, the RNA-seq signal profile, which is in the

Deleted: This is illustrated in a hypothetical example shown on the left panel of Fig 1d. The

BETTER REFERRED TO IN 1-2B

	center of our study, is one aggregated type. The signal profile is generated by counting the number of reads that overlap with each position on the genome. This profile does not immediately reveal any nucleotide information and is generally assumed to be free of variant information. Our study shows that this is not really the case because the dips in the signal profiles can reveal small and large genomic deletions. We show that an adversary can predict enough of the small deletions and use these to identify individuals. The aim of our current study is to demonstrate that the leakage from the genome-wide signal profiles can cause privacy concerns and present a way to close this leakage as much as possible so that the linking cannot be done reliably. There is another type of aggregated data files that are shared, which are the gene expression matrices. We agree that if the attacker used the gene expression levels, she could identify eQTLs and sQTLs but these are out of the scope of the attack that we are considering. In fact, our 2016 (Harmanci, Gerstein, Nature Methods, 2016) study focuses on exactly this scenario of linking eQTL genotypes to gene expression levels. To clarify the types of leakage that our manuscript studies, we made the supplementary figure 6. This figure illustrates the fact that the raw reads leak the full genotypic information, the signal profiles leak the genotype of deletions and gene expression levels can leak genotype information through eQTLs and sQTLs. Our current study deals with the signal profiles that leak deletions. We have clarified the main text (Section 2.3) about RNA-seq signal profiles and added a paragraph explaining that there can be other sources of leakage from RNA-seq signal profiles. We also added a supplementary figure (Supplementary Figure 5) to demonstrate how the small deletions affect RNA-seq signal profiles. We have included a simplified version of this figure below for reference. We also included a new Supplementary Figure (Supp. Figure 6) to clarify the types of leakages from functional genomics data.
Excerpt From Revised Manuscript	***Introduction*** In this study, we analyze the leakage of sensitive information from the functional genomics data and how they can be used by an adversary in linking attacks. There are a number of motivating key points related to functional genomics data and privacy. First of all, functional genomics data, such as RNA sequencing data, is unique, in that if the data comes from human subjects, the raw reads have genetic variant information, which may be used to identify individuals. However, the main purpose of

Deleted: current

Deleted: The aim of our current study is to demonstrate the leakage from the genome-wide signal profiles and close this leakage as much as possible so that the linking cannot be done reliably.

Deleted: updated the Discussion Section and

Deleted: This figure shows the screenshot of a UCSC genome browser signal track of GTex whole blood RNA-seq signal profiles. The 2 base pair deletion (rs34043625) in 3 GTex individuals can be seen even by eye easily. It is also worth noting that these tracks are publicly available for viewing and download.

Deleted: the

RNA-Seq data is not related to the variants; main purpose is more related to understanding dysregulation of genes under different conditions, such as cancer. Consequently, there's a great desire to share and study RNA-Seq datasets, to enable helping to find cures for various diseases. Because of this, there is great incentive to make ways of sharing functional genomics data without privacy protections. Large-scale privacy protections are a great encumbrance on genomic data sharing. They do not allow researchers and data owners to share results on the web, use web and internet-based tools, and they exert a great burden on research. Consequently, many consortia, such as GTEx, aim at sharing RNA-Seq information to the maximum extent. The raw reads obviously cannot be shared, as they contain variant information. However, there's belief that the signal files and the gene-level quantifications can be shared. The signal files simply reflect the overall depth of coverage of the RNA-Seq reads at any given point. Ostensibly, they're do not contain variant information. Many of the genomics consortia have decided to openly share RNA-Seq signal information. We show that there is a high degree of private information leakage in the function genomics signal profile data. The gene-level quantifications essentially are averages over the signal profile over exons. Although the overall averaging reduces information, private information leakage. However, there is also private information leakage through the association with variants called eQTLs. It is important to note that this is tackled in the current study, but is looked at elsewhere[16, 18].

…

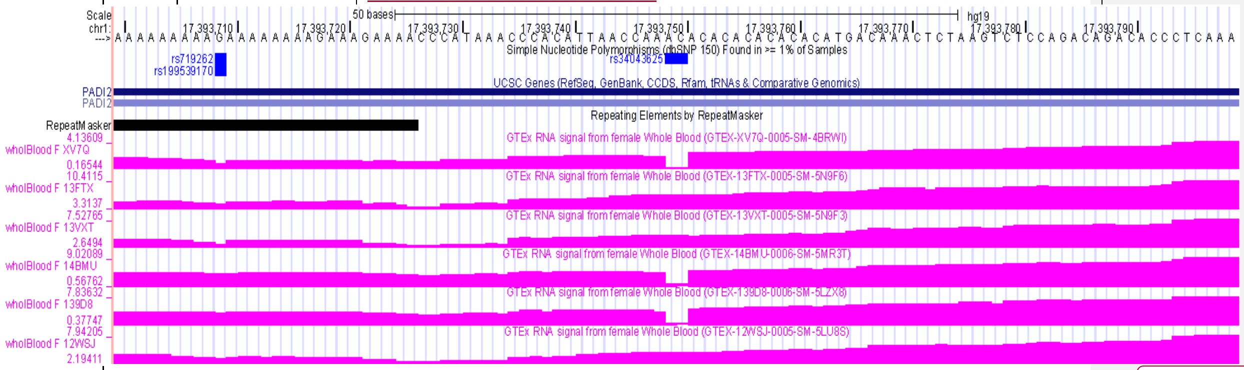
In this study, we analyze the sensitive information leakage from the signal profiles of several sequencing based functional genomics datasets. By signal profile, we refer to the genome-wide signal computed by counting the number of reads that overlap with each nucleotide on the genome. The signal profiles are just one type of aggregated data that is generated from raw reads. Another type of aggregated data is gene expression quantifications, which are averages of RNA-seq signal profiles over genes. The leakage of information from the gene expression quantifications has been previously studied[16, 18]. Rather, we are only considering whether the signal profiles have any genotypic information leakage from them. We show that signal profiles do leak a large amount of genotype information for small and large genomic deletions.

As discussed earlier, the raw reads from an RNA sequencing experiment contain the nucleotides themselves. It is well established that the raw reads must not be released publicly (Supplementary Figure 6) because given the raw reads, and adversary can identify a large number of private SNPs and indels. We therefore assume that the raw reads are not publicly shared and that the adversary does not have access to the raw reads. Rather, we assume that the data owners created the signal profiles and made these publicly available. The adversary gains access to these signal profiles. Regarding the signal profiles, it is generally assumed that the signal profiles are mostly void of sensitive information. Several large consortia, for example ENCODE Project[25], Roadmap Epigenome Mapping Consortium[26], and GTex[27, 28] publicly share signal profiles (Supplementary Figure 5)

2.3. Linking Attacks using RNA-Seq Signal Profiles

We first focus on the predictability of short deletions using RNA-seq signal profiles. Fig 1d illustrates a hypothetical example of how the small deletions in RNA-seq signal profiles can be detected as small and sudden dips in the signal. In order to show an example and represent the relevance of small deletions in the RNA-seq signal profiles, we included a screenshot of signal profiles around a small deletion for 6 individuals in the GTex Project (Supp. Figure 5). The 2 base pair deletion, rs34043625, can be easily detected for three of the individuals that are shown. An important aspect of the effect of small deletions on the signal profile is the extent that they affect the total expression of a gene. It is clear from Supplementary Figure 5 that the total signal in the small dips in the RNA-seq signal is much smaller than the perturbations caused by the other genetic factors like eQTLs and sQTLs. In general, an eQTL is associated with a global change in the total signal in the RNA-seq signal profile of a gene. However, a small deletion affects a localized position on the RNA-seq signal profile of the gene with relatively smaller effect on the total expression of the gene, assuming that the small deletion is not an eQTL. It is also worth noting that these signal profiles are publicly available from the UCSC Genome Browser.

Formatted: Justified


Scale
50 bases
hg19
chr1:
17,393,710
17,393,720
17,393,730
17,393,740
17,393,750
17,393,760
17,393,770
17,393,780
17,393,790
---> A A A A A A A A A G A A A A A A A G A A A G A A A A C C C A T A A A C C C A C A T T A A C C A A A C A C A C A C A C A C A C A T G A C A A A C T C T A A G T C T C C A G A C A G A C A C C C T C A A A
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
rs719262
rs199539170
rs34043625
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PADI2
PADI2
RepeatMasker
Repeating Elements by RepeatMasker
4.13609
wholBlood F XV7Q
0.16544
GTEx RNA signal from female Whole Blood (GTEX-XV7Q-0005-SM-4BRWI)
10.4115
wholBlood F 13FTX
3.3137
GTEx RNA signal from female Whole Blood (GTEX-13FTX-0005-SM-5N9F6)
7.52765
wholBlood F 13VXT
2.6494
GTEx RNA signal from female Whole Blood (GTEX-13VXT-0005-SM-5N9F3)
9.02089
wholBlood F 14BMU
0.56762
GTEx RNA signal from female Whole Blood (GTEX-14BMU-0006-SM-5MR3T)
7.83632
wholBlood F 139D8
0.37747
GTEx RNA signal from female Whole Blood (GTEX-139D8-0006-SM-5LZX8)
7.94205
wholBlood F 12WSJ
2.19411
GTEx RNA signal from female Whole Blood (GTEX-12WSJ-0005-SM-5LU8S)

The screenshot of UCSC Genome Browser's GTex Signal Profile Hub at the location chr1:17,393,700-17,393,799

Deleted: *<object>*¶

-- Ref1.4: I don't agree with the statement that “it is well known that the major portion of the genomic variation is caused by SVs”. –--

Reviewer Comment	I don't agree with the statement that “it is well known that the major portion of the genomic variation is caused by SVs”. Are the authors referring to the total number of nucleotides in the SV regions, or the impact of SVs versus SNPs to gene expression? Earlier work by Barbara Stranger and colleagues had shown that SNP cause more than 80% if the gene expression phenotype (Stranger Science 2007). It is probably true that an individual SV could have greater phenotypic effect than a SNV but SVs are obviously much less common.
Author Response	We agree with the reviewer’s concern. We believe we need to clarify the statement to express exactly that we are referring to the total number of bases that are affected by variants and not to the total effect size on gene expression. We also agree that this statement must be clarified according to the insightful comments of the reviewer. We have added the reference and updated the text to clarify it and reflect the reviewer’s remarks.
Excerpt From Revised Manuscript	***Introduction*** In this work, we are studying whether an adversary can use small and large genomic deletions for performing linking attacks. We study whether the adversary can use signal profiles of functional genomics signals to detect and genotype genomic deletions and use them to pinpoint individuals in a large genotype dataset. Most of the previous studies on genomic privacy focus on the single nucleotide polymorphisms (SNPs). This is well justified because the estimated regulatory effect of SNPs on gene expression is much larger than the structural variants[22]. On the other hand, it is known that the major portion of the genomic variation, in terms of the number of nucleotides that are affected, is caused by SVs[23, 24] as shown by 1000 Genomes Project. Since an SV affects a much larger portion of the genome (in number of nucleotides) than a single nucleotide variant does, its effect on a phenotype is expected to be very obvious, if not more than a SNP. For example, homozygous deletion of a gene will cause the total disappearance of its expression.

Formatted: Justified

-- Ref1.5: I think the part on Hi-C doesn’t really add much to the work. –--

Reviewer Comment	I think the part on Hi-C doesn’t really add much to the work, the results are less convincing than the those of RNA-Seq and ChIP-seq and there are more confounding factors. I suggest to have it removed from the manuscript.
Author Response	The reviewer recommends removing the Hi-C analysis because it is not as convincing. Although we agree that Hi-C analysis does

	not conform to the rest of the RNA-seq and ChIP-Seq analysis, we still think it is valuable to demonstrate the possibility of an attack using this data. Therefore, we moved the Hi-C analysis to the Supplementary Text and we included references to this analysis in the main text.
Excerpt From Revised Manuscript	[[I am not sure if we should do what I am saying above]]

-- Ref1.6: The RNA-seq and chromatin modification data described in this work were derived from 1000 Genome and similar consortia projects. –--

Reviewer Comment	The RNA-seq and chromatin modification data described in this work were derived from 1000 Genome and similar consortia projects, where were mostly transformed lymphoblastoid cell lines instead of primary cell or tissue cell lines. While the observations were interesting and convincing, in practice RNA-seq data is probably more common than ChIP-seq data, especially in a clinical setting.
Author Response	We thank the reviewer for making a strong point that supports the urgency of protecting RNA-seq data. We agree with the reviewer's comment. We are, however, confused by reviewer's comment that the RNA-seq and chromatin data were derived from 1000 Genomes and similar consortia projects. We would like to point out that The 1000 Genomes project currently does not have any functional genomics data. The RNA-seq datasets are from GTex and GEUVADIS consortia. GEUVADIS RNA-seq data is generated from lymphoblastoid cell lines of 462 individuals whose genotypes are available in 1000 Genomes Project. The GTex contains a much more diverse set of data with many tissue cell lines. In our study, we focus on the data from cell lines generated from whole blood of participants of the GTex project. The reviewer also makes an important point that the RNA-seq data is much more common than ChIP-Seq data. This argument supports our study very well: As we have explained in the manuscript (Section 2.6), this is exactly the reason why we are focusing on anonymization of RNA-seq signal profiles, i.e., RNA-seq is much more common data type especially in the clinical setting and it is realistically more urgent to anonymize RNA-seq signal data. We, however, still believe that the leakage analysis from ChIP-Seq data is important as ChIP-Seq is becoming more common in large scale functional genomics projects. We updated the Section 2.6 (Anonymization of Signal Profiles) to clarify above points.

Deleted: on Anonymization of Signal Profiles,

Deleted: on

Deleted: emphasize the clinical relevance of RNA-seq

Excerpt From Revised Manuscript	***2.6. Anonymization of RNA-Seq Signal Profiles*** The personal RNA-seq datasets are currently by far the most abundant datasets compared to other functional genomic datasets. For example, the RNA-seq signal profiles are being publicly shared from the GTex project while the genotypes are not in public access. In addition, RNA-seq is becoming commonly used in the clinical settings and new RNA-seq based assays are being developed to probe gene expression, for example single cell RNA-sequencing. Altogether these make protection of RNA-seq data urgent. We therefore focus on protection of the RNA-seq datasets.

-- Ref2.1: The major concern is that they presume they can anonymize and thus fully understand the system behind the signal data. –--

Reviewer Comment	The major concern is that they presume they can anonymize and thus fully understand the system behind the signal data. They write they "present an effective anonymization procedure for protection of signal profiles against genotype prediction based attacks". The reviewer views this as incorrect overstatement given their manuscript, as functional data have impacts across many genes and networks - many unseen or still to be discovered. In the end, they present one rather ad-hoc method for a linkage attack built on dips & also present how one can protect against that ad-hoc approach. Still, there are many, many more that could also be described and suggesting that they have developed an anonymization approach that is generalization is premature. For example, a basis of much of biology is that DNA level events impact not just the gene that is deleted but entire complex pathways, leaving complex signatures. The reviewer can think of dozens of ways a deletion of a gene that negatively regulates a pathway would lead to downstream upregulation of other genes (not a dip). Beyond this, one can see ways deep neural networks can be trained, and deduce using hidden network via emerging Artificial Intelligence algorithms. The problem with suggesting that one can anonymize the data presumes that new knowledge won’t be gained allowing one to infer laying on complex pathway information within a linkage attack.
Author Response	The reviewer is making a valid point regarding our anonymization procedure. Our statement that the proposed anonymization method is effective for full protection of signal profiles may be viewed as an overstatement. At this point, we believe it is important to systematically clarify the sources of leakage and which leakage our study analyzes:

In any functional sequencing experiments, the generated data is the raw reads. Therefore, the leakage directly from the reads is the main source of leakage from the raw read data. The raw reads contain nucleotide information and an adversary can immediately identify variants from the reads. Therefore, the raw read data is almost always stored away from public access. In order to make data available publicly, several aggregate file formats are used. One of these formats is the read depth signal profiles, which are in the main focus of our manuscript. Another layer of aggregation over the signal profiles is the gene expression quantifications. In these quantifications, for each gene, the average signal over the gene is computed. The gene expression quantifications can leak variant information because they are correlated with expression QTLs and splicing QTLs. This leakage is not addressed in our study but it has been studied in previous papers. We have added a new figure (Supplementary Figure 6) to illustrate these leakages.

2-1A

As the reviewer points out, one can envision additional sources of leakage beyond these aggregated formats. For instance, there can be complex and subtle correlations between variant genotypes and the aggregate expression levels genes within pathways and networks. These are not currently explicitly studied, but they could be detected through complex pattern-matching and machine learning techniques, such as deep learning. Even more so, although eQTLs have traditionally been linked to genes, there may be eQTLs whose variant genotypes are correlated with the expression of intergenic and intronic elements. Finally, another additional source of leakage is, while we do look at calling of particular types of structural variants, such as small and large deletions, there may be very large, megabase-scale deletions, which affect many genes. This is particularly relevant for the case of somatic events in cancer. These are other sources of leakage that we did not address here.

IS THIS FROM THE MAIN TEXT ?

So, to emphasize, we focused on a particular type of leakage of private information in RNA-Seq data, related to the signal profile. There are many other sources of information, however the signal file is currently the juncture between public and private information, and is begun to be shared publicly. Hence, we think it's particularly important to measure the leakage at this level. It might unfortunately be the case that this type of information is not able to be shared publicly in the future, and one will have to move up the stack, perhaps only sharing gene-level quantifications, or even worse, nothing at all. We wish to emphasize that we are not, in this paper, trying to look at all sources of RNA-Seq variant information,

	but just the source of leakage for the data formats that are believed to be safe to share. As the reviewer rightfully points out, the current study does not consider the leakage from the much more complicated mechanisms comprising of complex genetic pathways. We have clarified this statement as following: "We have developed an anonymization procedure, which is effective at closing a major source of genetic information leakage that is caused by the dips in the signal." As this new statement reflects, we do not claim to close all the leakage but we demonstrate to a major source. We, however, believe that it would be fair when we state that the leakage from the signal dips that is presented in our study is a major source of the leakage that must urgently be closed. The leakage from the higher order effects of a variants on pathways can be studied separately. We have updated the Signal Profile Anonymization and Discussion Sections to stress and clarify the above points. We also added Supplementary Figure 6 to illustrate the types of leakage from different data formats used in functional genomics and clarify the leakage we are tackling in this paper.
Excerpt From Revised Manuscript	**Discussion** The sequencing based functional genomics assays provide very large amount of biological information. Within this, much of the variant genotype information is within the raw reads (Supplementary Figure 6). In fact an adversary that gains access to the raw reads can easily call SNPs, indels, and structural variants. This is why raw reads are always protected from public access. The gene expression levels have also been shown to leak enough genotype data that can be used in linking attacks[16, 18]. The privacy concerns around sharing signal profiles are not well studied yet. … It is worth noting that the anonymization method that we presented does not close all the sources of leakage. The anonymization procedure aims to close the leakages caused by the genotyping of genomic deletion using the dips in the signal profile. These leakages are very accessible to and adversary and we believe that they must be urgent closed because they can be detected directly from the signal profiles. Given other types of data, there can still be other sources of genotype information leakage after the anonymization is applied. For example, the gene expression levels can be used to infer genotype information, which was demonstrated in earlier studies[16, 18]. In addition, the effects of variants on the activity levels of pathways are not well known yet. The complex machine learning

Deleted: need to clarify

frameworks, such as deep learning methods, have great potential to reveal the correlations between variants and activity levels of pathways. The leakage from the pathway level activity can be analyzed by using deep learning based approaches.

Overall, at this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Sequencing, and point out the sources that we probed in this paper. First, there is the leakage directly from the reads. This is the most obvious leakage, and this leakage is avoided with by simply not sharing the raw reads. Next source of leakage is from the signal profile. This leakage is addressed in this paper. There is yet another source of leakage though, when one averages over the signal file, and produces quantifications in particular regions such as genes. These quantifications can be subtly connected with variants through the notion of eQTLs. This is not addressed in this paper, and there can be substantial leakage from these quantifications.

Furthermore, one can envision additional sources of leakage beyond that, in these main areas. For instance, one can imagine complex and subtle correlations between the levels of gene expression of many genes within pathways and networks. Although there has been interest in identifying these higher order QTLs, these are not yet extensively studied[28]. Complex machine learning techniques, such as deep learning, can reveal subtle correlations of gene expression at the network level with variants. Also, eQTLs traditionally have been linked to genes; ostensibly, one might imagine by averaging over various intergenic regions, some of the more highly expressed intergenic regions might also show correlations with variants. This is another source of information not studied in this work. Finally, an additional source of information is, while we do look at calling of particular types of structural variants, such as small and large deletions, there may be very large-scale, megabase-scale deletions, which affect many genes. This is particularly the case for somatic events in cancer samples. This case is also not covered by our procedure.

Finally, we would like to emphasize that we focused on a particular type of leakage of private information in functional genomics data, such as RNA-Seq data, such that the leakage stems from the signal profile. There are many other sources of information, however the signal file is currently at the junction between public and private information, and is where genomic information is begun to be shared publicly. Hence, we believe it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might unfortunately be the case that this type of information is not able to be shared publicly in the future, perhaps only sharing gene-level quantifications, or even worse, nothing at all. We wish to emphasize that, in this paper, we are not trying to look at all sources of leakage from functional genomics data, but just the sources of leakage right at the decision boundary of sharing and not sharing.

2.6. Anonymization of RNA-Seq Signal Profiles

It is worth noting that this procedure can be used anonymizing not only RNA-seq signal profiles but also other signal profiles against attacks that are based on small deletion genotyping. The anonymization is, however, not as effective for large deletions. This is not a major concern for RNA-seq signal profiles as we observed that large deletions are not easily genotyped using RNA-seq data. However, as we showed in the previous section, the linking attacks can be successful when they use the large deletions that are genotyped using ChIP-Seq datasets.

Formatted: Justified