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

RESPONSE LETTER

-- Ref1: Introductory comments –--

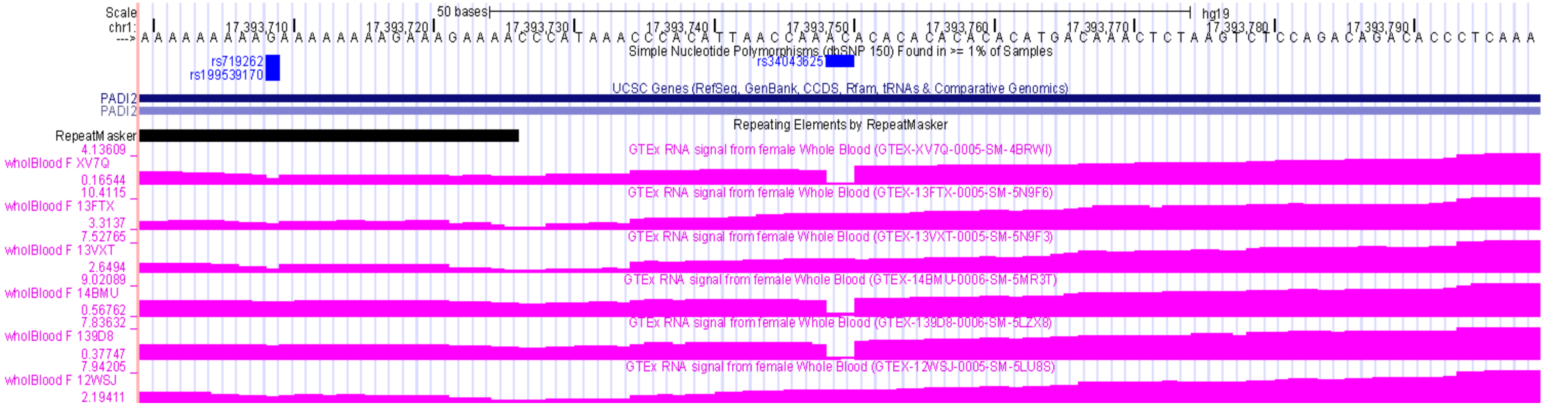
Reviewer Comment	Built on previous work from the aspect of SNPs (published in 2016), here the authors expand onto structural variants (SVs), and onto functional genomics data such as RNS-seq and ChIP-seq. The authors' analyses provided evidence that private indels and other SVs can be recovered from 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	We sincerely thank the reviewer for the constructive comments, which we believe made our paper stronger. We respond to reviewer's comments below.
Excerpt From Revised Manuscript	

NOT RAW READS.

-- Ref1: 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 AGREE [illegible]	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. This is illustrated in a hypothetical example shown on the left panel of Fig 1d. The 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 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 current attack. In fact, our 2016 (Harmanci, Gerstein, Nature Methods, 2016) study focuses on exactly this scenario of linking eQTL genotypes to gene expression levels. 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. We have clarified the main text about RNA-seq signal profiles and updated the Discussion Section and added a paragraph explaining that there can be other sources leakage from RNA-seq signal profiles. We also added a figure to demonstrate how the small deletions affect RNA-seq signal profiles. 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. We have included the figure below for reference.
Excerpt From Revised Manuscript	


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
Simple Nucleotide Polymorphisms (dbSNP 150) Found in >= 1% of Samples
rs719262
rs199539170
rs34043625
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
PADI2
PADI2
Repeating Elements by RepeatMasker
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

-- Ref1: 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 reflect the reviewer's remarks.
Excerpt From Revised Manuscript	

Cit "1000G"

-- Ref1: 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	

-- Ref1: 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 WRONG	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. As we have explained in the Section on Anonymization of Signal Profiles, this is 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 on Anonymization of Signal Profiles to emphasize the clinical relevance of RNA-seq.
Excerpt From Revised Manuscript	

-- Ref2: 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. 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 need to clarify 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.
Excerpt From Revised Manuscript	