# Response to reviewers’ Comments for “Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions”

# Response Letter

### -- Ref1.1: Introductory comments --

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | 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: The deletions discovered from these raw data sets can be cross-linked… --

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | The reviewer’s comments summarize parts of our manuscript, but we believe that we need to clarify some of it. The reviewer indicates that our analyses provide evidence that the private SVs can be recovered from the **raw reads from RNA-Seq and ChIP-Seq experiments**. **However, this is** **not in our manuscript**. Our analysis does not provide evidence that raw reads, by themselves, can be used to recover SVs. In fact, 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 rich source of information. The main purpose of our functional genomics experiment is to understand the differences in the regulation and expression of genes under different conditions (e.g., among individuals with cancer). Although the main purpose of the data is not about the detection of variants, these data may provide variant information. For example, the raw reads from an RNA-Seq experiment contain nucleotides, and these can be used to identify a large number of genetic variants, such as SNPs and indels. An adversary could use these variants to breach an individual’s privacy. Thus, the raw reads are almost never shared. There is, however, great incentive to share the data because they provide invaluable resources for disease research.Researchers could generate and share aggregated data files from the raw reads because these are seemingly free of variant information. For example, the read depth signal profiles, which are the central topic of our paper, are one type of such aggregated data. These profiles are just counts of reads at each position on the genome, and they do not have any nucleotide information immediately available (new Supplementary Figure 2a). Thus, these signal profiles are assumed to be free of variant information and are safe to share publicly. In fact the GTEx Consortium generates RNA-Seq data from hundreds of healthy individuals and publicly shares the signal profiles for these data through the UCSC Genome Browser. Our manuscript’s main focus is this point: We are studying the leakage of variant information from signal profiles. We show that we can use signal profiles to detect and genotype small and large genomic deletions, which we can use to identify individuals within a large cohort.It is important to note that other aggregated datasets can be generated from raw reads and signal profiles. For example, gene expression levels are computed by averaging RNA-Seq signal profiles. Gene expression levels and eQTLs can be used to detect variant genotypes, which 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 focuses on genotyping small and large deletions using **only the signal profiles from RNA-Seq and ChIP-Seq data**. Our results provide evidence that these signal profiles can leak enough genotypic information to pinpoint individuals.We would like to recapitulate this distinction: It is well known that the raw reads contain a large amount of genotype information. Therefore, it is generally not acceptable to share raw reads from any sequencing experiment. However, the privacy risks around sharing signal profiles are not well understood. Indeed, the GTEx Consortium publicly shares RNA-Seq signal profiles in the UCSC Genome Browser. (See the figure below Comment 1.3 and New Supplementary Figure S3.) Our manuscript sheds light on this issue. To clarify the above point, we made a new supplementary figure (Supplementary Figure 2a and 2b) to illustrate the leakage from reads, signal profiles, and gene expression levels. We also updated the introduction and discussion sections to clarify that the central theme of our study is signal profiles and not raw reads. |
| Excerpt FromRevised Manuscript | **Introduction:**In this study, we analyzed the leakage of sensitive information from functional genomics data and how an adversary could use it in linking attacks. Functional genomics data, such as those from RNA sequencing (RNA-Seq), is unique in that if the data comes from human subjects the raw reads have genetic variant information. This information could be used to identify individuals (Supplementary Fig. 2b). However, the main purpose of RNA-Seq data is not related to the variants, but rather understanding how the activity of genes changes under different conditions such as cancer. Thus, unlike the variant data, functional genomics datasets have a more complicated “Yin-Yang” aspect with relation to privacy. In addition, functional genomics datasets are sometimes shared with phenotypic information that is potentially of private value (e.g., a particular condition or disease that a person has). This leads to an interesting situation where the data is ostensibly collected and used for non-personal purposes to determine general aspects about a condition. However, the existence of small amounts of residual private information in the data potentially can be revealing about the individual from which they came. Another important factor is the desire to share and study RNA-Seq datasets to help find cures for various diseases. Because of this, there is great incentive to find ways to share functional genomics data without privacy protections. Large-scale privacy protections are an encumbrance on genomic data sharing. These protections do not allow researchers and data owners to share results on the web or use web- or internet-based tools, exerting a great burden on research. Consequently, many consortia, such as the Genotype-Tissue Expression (GTEx) Project, aim to share RNA-Seq information to the maximum extent. Although the raw reads cannot be shared, there is a general belief that other aggregated data computed using raw reads, such as signal profiles and gene-level quantifications, can be shared. Signal profiles simply reflect the overall depth of coverage of the RNA-Seq reads at any given position on the genome. These profiles are computed by counting the number of reads that overlap with each position on the genome (Supplementary Fig. 2a). The profiles ostensibly do not contain variant information. This is why many genomics consortia have decided to openly share RNA-Seq signal profiles in bigwig and wig files. In this study, we focused on leakage from signal profiles. Another commonly shared aggregated data are gene-level quantifications, which are essentially averages of the signal profile over exons. Although overall aggregation and averaging reduces information, private information leakage also decreases. However, private information leakage still occurs from gene expression quantifications through the association of expression levels with variants called expression quantitative trait loci (eQTLs). Although we do not tackle this in the current study, it has been explored elsewhere16,18. …In this study, we analyzed sensitive information leakage from signal profiles of several sequencing-based functional genomics datasets. Signal profiles are currently at the junction between public and private information, and where genomic information has begun to be shared publicly. Hence, it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might be the case that this type of information will not be publicly shareable at all in the future. We emphasize that in this paper we are not trying to look at all sources of leakage from functional genomics data, but just the sources right at the decision boundary of sharing and not sharing.As we introduced earlier, the raw reads from an RNA-Seq experiment contain the nucleotides themselves. We assume that the data owners created the signal profiles and made them publicly available. Several large consortia, for example the Encyclopedia of DNA Elements (ENCODE) project24, the Roadmap Epigenome Mapping Consortium25, and GTEx26,27, publicly share signal profiles (Supplementary Fig. 3).**Discussion:**At this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Seq, and point out the sources that we probed in this paper. First, there is leakage directly from the reads. This is the most obvious leakage, and can be avoided by simply not sharing the raw reads. The next source of leakage is from the signal profile. We address this leakage is in this paper. There is yet another source of leakage, 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 eQTLs and can create substantial leakage. Furthermore, one can envision additional sources of leakage beyond these main areas. For instance, although the eQTLs traditionally have been linked to genes, highly expressed intergenic regions43 may also be linked to eQTLs. In addition, while we consider a particular class of structural variants (i.e., small and large deletions), there may be very large, megabase-scale deletions that affect many genes. This is particularly the case for somatic events in cancer samples. These cases are not addressed in our study.  |

### -- 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 --

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | We thank the reviewer for this insightful comment. Although the reviewer is concerned that deletions may not affect gene expression as much as eQTLs and splice site mutations, we believe that we need to clarify the setup of the attack: In an attack scenario regarding RNA-Seq data, we assume that the attacker uses the signal levels to find small deletions in the signal profile, which manifest themselves as small but noticeable dips in the RNA-Seq signal profiles. We present a hypothetical example of how a small deletion manifests itself on an RNA-Seq signal profile on the left panel of Figure 1d and a real example in the new Supplementary Figure 3 (see simplified version below). This figure shows the screenshot of a UCSC Genome Browser signal track of GTEx whole blood RNA-Seq signal profiles, which are publicly available for viewing and downloading. The two base pair deletion (rs34043625) in 3 of the GTEx individuals can be seen easily by eye. Another important aspect is that the signal in the dips is much smaller than the changes in the gene expression caused by the eQTLs and sQTLs. The eQTLs generally cause changes in the total signal in the signal profile of a gene, whereas small non-eQTL deletions create smaller localized changes in the signal profile.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, they 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 to show that the small deletions that leak from RNA-Seq signal profiles can be used to link individuals correctly.We believe that the confusion stems partly from the fact that the setup of the problem is not clear, so we will review it here for clarity. RNA-Seq datasets are rich sources of information, and there is currently a great desire to generate and share these data. Whereas the purpose of DNA sequencing of genomes is often to identify variants that lead to disease, such as driver mutations in cancer, the main purpose of RNA-Seq data is more related to understanding the differences in gene activity between different conditions, such as healthy versus disease states. Although detection of variants is not the main purpose of RNA-Seq data, they still contain genetic variant information. This is what makes these data problematic in terms of individual privacy because the raw reads from RNA-Seq experiments contain nucleotides and because an adversary can use these to find a large number of variants. In order to share these data, people have developed several aggregated formats, such as the RNA-Seq signal profile that is the center of our study. 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. However, our study shows that the dips in signal profiles can reveal small and large genomic deletions. We show that an adversary can predict enough of the small deletions to identify individuals. The aims of our current study were to demonstrate that leakage from genome-wide signal profiles can cause privacy concerns and to present a way to close this leakage as much as possible so that linking cannot be done reliably. Another type of shared aggregated data files are gene expression matrices. We agree that if an attacker used gene expression levels, they could identify eQTLs and sQTLs; these are out of the scope of the attack that we are considering, but our 2016 study (Harmanci and Gerstein, Nature Methods, 2016) focuses on this scenario of linking eQTL genotypes to gene expression levels. To clarify the types of leakage from functional genomics data, we made a new supplementary figure (Supplementary Figure 2b). This figure illustrates the fact that raw reads leak full genotypic information, signal profiles leak deletions (the focus of our current study), and gene expression matrices leak genotype information through eQTLs and sQTLs.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 2b) to clarify the types of leakages from functional genomics data. |
| Excerpt FromRevised Manuscript | Introduction:In this study, we analyzed the leakage of sensitive information from functional genomics data and how an adversary could use it in linking attacks. Functional genomics data, such as those from RNA sequencing (RNA-Seq), is unique in that if the data comes from human subjects the raw reads have genetic variant information. This information could be used to identify individuals (Supplementary Fig. 2b). However, the main purpose of RNA-Seq data is not related to the variants, but rather understanding how the activity of genes changes under different conditions such as cancer. Thus, unlike the variant data, functional genomics datasets have a more complicated “Yin-Yang” aspect with relation to privacy. In addition, functional genomics datasets are sometimes shared with phenotypic information that is potentially of private value (e.g., a particular condition or disease that a person has). This leads to an interesting situation where the data is ostensibly collected and used for non-personal purposes to determine general aspects about a condition. However, the existence of small amounts of residual private information in the data potentially can be revealing about the individual from which they came. Another important factor is the desire to share and study RNA-Seq datasets to help find cures for various diseases. Because of this, there is great incentive to find ways to share functional genomics data without privacy protections. Large-scale privacy protections are an encumbrance on genomic data sharing. These protections do not allow researchers and data owners to share results on the web or use web- or internet-based tools, exerting a great burden on research. Consequently, many consortia, such as the Genotype-Tissue Expression (GTEx) Project, aim to share RNA-Seq information to the maximum extent. Although the raw reads cannot be shared, there is a general belief that other aggregated data computed using raw reads, such as signal profiles and gene-level quantifications, can be shared. Signal profiles simply reflect the overall depth of coverage of the RNA-Seq reads at any given position on the genome. These profiles are computed by counting the number of reads that overlap with each position on the genome (Supplementary Fig. 2a). The profiles ostensibly do not contain variant information. This is why many genomics consortia have decided to openly share RNA-Seq signal profiles in bigwig and wig files. In this study, we focused on leakage from signal profiles. Another commonly shared aggregated data are gene-level quantifications, which are essentially averages of the signal profile over exons. Although overall aggregation and averaging reduces information, private information leakage also decreases. However, private information leakage still occurs from gene expression quantifications through the association of expression levels with variants called expression quantitative trait loci (eQTLs). Although we do not tackle this in the current study, it has been explored elsewhere16,18.…In this study, we analyzed sensitive information leakage from signal profiles of several sequencing-based functional genomics datasets. Signal profiles are currently at the junction between public and private information, and where genomic information has begun to be shared publicly. Hence, it is particularly important to probe the leakage from the signal profile representation of functional genomics data. It might be the case that this type of information will not be publicly shareable at all in the future. We emphasize that in this paper we are not trying to look at all sources of leakage from functional genomics data, but just the sources right at the decision boundary of sharing and not sharing.As we introduced earlier, the raw reads from an RNA-Seq experiment contain the nucleotides themselves. We assume that the data owners created the signal profiles and made them publicly available. Several large consortia, for example the Encyclopedia of DNA Elements (ENCODE) project24, the Roadmap Epigenome Mapping Consortium25, and GTEx26,27, publicly share signal profiles (Supplementary Fig. 3)Discussion:At this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Seq, and point out the sources that we probed in this paper. First, there is leakage directly from the reads. This is the most obvious leakage, and can be avoided by simply not sharing the raw reads. The next source of leakage is from the signal profile. We address this leakage is in this paper. There is yet another source of leakage, 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 eQTLs and can create substantial leakage. Furthermore, one can envision additional sources of leakage beyond these main areas. For instance, although the eQTLs traditionally have been linked to genes, highly expressed intergenic regions43 may also be linked to eQTLs. In addition, while we consider a particular class of structural variants (i.e., small and large deletions), there may be very large, megabase-scale deletions that affect many genes. This is particularly the case for somatic events in cancer samples. These cases are not addressed in our study. 2.3. Linking Attacks using RNA-Seq Signal ProfilesWe first focused on the predictability of small deletions using RNA-Seq signal profiles. Figure 1d illustrates a hypothetical example of how small deletions in RNA-Seq signal profiles can be detected as small and sudden dips in the signal. As an example showing the relevance of small deletions in RNA-Seq signal profiles, we include a screenshot of signal profiles around a small deletion for six individuals in the GTEx Project (Supplementary Fig. 3). The two base pair deletion, rs34043625, can be easily detected for three of the individuals shown. An important aspect of the effect of small deletions on the signal profile is the extent to which they affect the total expression of a gene. It is clear from Supplementary Figure 3 that the total signal in the small dips in the RNA-Seq signal is much smaller than the perturbations caused by other genetic factors like eQTLs and splicing QTLs. In general, an eQTL is associated with a global change in the total signal on a RNA-Seq signal profile of a gene. However, a small deletion affects a localized position on the RNA-Seq signal profile with a relatively smaller effect on the total expression of the gene, assuming the small deletion is not an eQTL.  |

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

### -- 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”. --

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | We agree with the reviewer’s insightful comments. We need to clarify the statement to express 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 have added the Stranger *et al.* reference and updated the text to reflect the reviewer’s remarks.  |
| Excerpt FromRevised Manuscript | **Introduction:**In this study, we explored whether an adversary could use signal profiles of functional genomics signals to detect and genotype genomic deletions and use them to pinpoint individuals in a large genotype dataset in a linking attack. Most previous studies on genomic privacy focus on single nucleotide polymorphisms (SNPs). This is well justified because the estimated regulatory effect of SNPs on gene expression is much larger than structural variants21. However, the major portion of genomic variation, in terms of the number of nucleotides that are affected, is caused by SVs22,23, as shown by The 1,000 Genomes Project. Since an SV affects a much larger portion of the genome than a SNP, we expect a phenotype caused by an SV to be very obvious. For example, homozygous deletion of a gene will cause the total disappearance of its expression. |

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

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | 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. We believe that it raises a source of leakage that must be tackled in the near future because Hi-C experiments are getting very prevalent in genome sequencing. We therefore decided to keep this analysis in the manuscript. |
| Excerpt FromRevised Manuscript |  |

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

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | We thank the reviewer for making the important point that RNA-Seq data is much more common than ChIP-Seq data. This is one of the main points that supports our study. As we have explained in the Section 2.6, this is exactly why we are focusing on the anonymization of RNA-Seq signal profiles (i.e., RNA-Seq data are much more common, especially in the clinical setting; thus, it is urgent to protect RNA-Seq signal data). However, we 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 are also partly confused by the reviewer’s comment that the RNA-Seq and ChIP-Seq data were derived from 1000 Genomes and similar consortia projects. We would like to point out that the 1000 Genomes project does not currently have any functional genomics data. The RNA-Seq datasets that we used are from the GTEx and GEUVADIS consortia. The GEUVADIS RNA-Seq data were generated from lymphoblastoid cell lines of 462 individuals whose genotypes are available in the 1000 Genomes Project. The diverse GTEx datasets contain many tissue cell lines. In our study, we focus on data from cell lines generated from whole blood of participants of the GTEx project. We updated Section 2.6 (Anonymization of Signal Profiles) to clarify the above points. |
| Excerpt FromRevised Manuscript | 2.6. Anonymization of RNA-Seq Signal ProfilesPersonal RNA-Seq datasets are currently by far the most abundant functional genomic datasets. For example, RNA-Seq signal profiles are being publicly shared from the GTEx project, although 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-Seq. Altogether, these factors make the protection of RNA-Seq data urgent. |

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

|  |  |
| --- | --- |
| ReviewerComment | 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. |
| AuthorResponse | The reviewer makes 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 genomics sequencing experiment, first-level data comprise the raw reads. Therefore, 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. Hence, 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 read depth signal profiles, which are in the main focus of our manuscript. Another layer of aggregation over the signal profiles is 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 eQTLs and sQTLs. This leakage is not addressed in our study but it has been studied in previous papers. We have added a new figure (Supplementary Figure 2b) to illustrate these leakages.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 further, 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 source of leakage consists of megabase length deletions, which affect many genes. These variants are prevalent in the case of somatic events in cancer. These are other sources of leakage that we did not address here.So, to emphasize, we focused on a particular type of leakage of private information in RNA-Seq data, related to signal profiles. While there are many other sources of information, signal profiles require attention because currently they are at the boundary where genomic data is beginning to be shared publicly. Hence, we think it is particularly important to measure the leakage at this level. 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 (specifically the signal profiles) that are believed to be safe to share.As the reviewer rightfully points out, our study does not consider leakage from other sources such as the complicated mechanisms comprising complex genetic pathways. We have clarified this statement. In summary, we would like to convey that our anonymization procedure 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 leakages. We rather point to a major source of leakage and aim at closing it.However, we believe that it would be fair if 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 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 2b 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 FromRevised Manuscript | **Discussion:** Sequencing-based functional genomics assays provide a large amount of biological information for understanding the dynamic nature of gene activity and epigenetic regulation. This information is extremely valuable for understanding genetic mechanisms behind disease initiation and progression. Thus, data producers and owners want to share these data as openly as possible. At the same time, genomic data can contain variant genotype information within the raw reads that may cause concerns for privacy. These two competing factors, the incentive to share and privacy concerns, make it necessary to carefully evaluate the sharing mechanisms of functional genomics data. To decrease genetic variant leakage in sequencing data, aggregate data formats have been widely used. Two examples are signal profiles and gene expression quantifications. Unlike raw reads, these data do not immediately reveal variant information and are generally accepted to be safe for public data sharing. However, gene expression levels have been shown to leak enough genotype data to be used in accurate linking attacks16,18. In this study, we evaluated the possible privacy concerns around sharing signal profiles.…We note that the anonymization method that we presented does not close all sources of leakage. The procedure aims to close the leakages caused by the genotyping of genomic deletions using the dips in the signal profile. These leakages are accessible to an adversary and can be detected directly from the signal profiles. Thus, we believe that they must be urgently closed. For other types of data, additional sources of genotype information leakage could be present after the anonymization is applied. For example, gene expression levels can be used to infer genotype information, which was demonstrated in earlier studies16,18. In addition, the effects of variants on the activity levels of pathways are not well known yet. Complex machine learning frameworks, such as deep learning and neural networks, have great potential to reveal the correlations between variants and activity levels of pathways. Although there has been interest in identifying these higher-order QTLs, these are not yet extensively studied27.At this point, it is useful to review all the sources of information leakage from functional genomics experiments, such as RNA-Seq, and point out the sources that we probed in this paper. First, there is leakage directly from the reads. This is the most obvious leakage, and can be avoided by simply not sharing the raw reads. The next source of leakage is from the signal profile. We address this leakage is in this paper. There is yet another source of leakage, 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 eQTLs and can create substantial leakage. Furthermore, one can envision additional sources of leakage beyond these main areas. For instance, although the eQTLs traditionally have been linked to genes, highly expressed intergenic regions43 may also be linked to eQTLs. In addition, while we consider a particular class of structural variants (i.e., small and large deletions), there may be very large, megabase-scale deletions that affect many genes. This is particularly the case for somatic events in cancer samples. These cases are not addressed in our study. 2.6. Anonymization of RNA-Seq Signal ProfilesImportantly, this procedure can be used for anonymizing not only RNA-Seq signal profiles but also other signal profiles against attacks based on small deletion genotyping. However, the anonymization is not as effective for large deletions. This is not a major concern for RNA-Seq signal profiles, as we observed that large deletions were not easily genotyped using RNA-Seq data. However, as we showed in the previous section, linking attacks can be successful when they use large deletions that are genotyped using ChIP-Seq datasets. |