Genome Biology

SHORT REPORT Open Access

Recovering genotypes and phenotypes using allele-specific genes

Gamze Gürsoy[1,2], Nancy Lu[3,4], Sarah Wagner[5] and Mark Gerstein[1,2,4,5]*

* Correspondence: mark@gersteinlab.org
[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, USA
[2]Molecular Biophysics and Biochemistry, Yale University, New Haven, USA
Full list of author information is available at the end of the article

Abstract

With the recent increase in RNA sequencing efforts using large cohorts of individuals, surveying allele-specific gene expression is becoming increasingly frequent. Here, we report that, despite not containing explicit variant information, a list of genes known to be allele-specific in an individual is enough to recover key variants and link the individuals back to their genotypes and phenotypes. This creates a privacy conundrum.

Background

Owing to the surge in functional genomics data over the past decade, several reports have focused on identifying genomic privacy issues related to molecular phenotype data, such as gene expression levels [1, 2]. These studies exploit the known and publicly available relationship between genotype and molecular phenotypes such as expression quantitative trait loci (eQTLs). That is, given a matrix of gene expression values collected from a cohort of individuals and a list of eQTLs, one can link a genome from a known individual to the gene expression matrix and uncover potentially stigmatizing phenotypes such as HIV status [1].

The increase in personal genomes and functional genomics data allows researchers to investigate the allele-specific activity of the genome. With the surge in large-scale RNA sequencing and genotype efforts such as the Genotype-Tissues Expression (GTEx) project [3, 4], more studies have begun focusing on allele-specific expression (ASE) in the human genome [4–7]. ASE is a characteristic of having an imbalance in the quantity of expressed copy of a gene (maternal or paternal allele) and may lead to phenotypic variation. Between 10 and 22% of human genes show allele-specific regulation of gene expression [8]. ASE can be created in part by underlying biological processes such as imprinting. However, most observed cases are not necessarily due to underlying biological phenomena. There is increasing evidence that ASE could be linked to the predisposition to diseases such as autism spectrum disorder [9], colorectal cancer [10], and tumorigenesis in general [11].

Due to their clinical importance and direct relationship to the phenotype of the organism, there is an incentive to broadly share a list of allele-specific genes, or allele-specific gene expression matrices, belonging to a patient or a study participant. Moreover, ASE information is often shared with the accompanying phenotype of the individual. Many assume that haplotype-level gene expression data, i.e., expression levels of a gene in different alleles, do not contain any identifying information and are safe to share even if the data are derived from individuals who did not provide broad consent [4].

Results and discussion

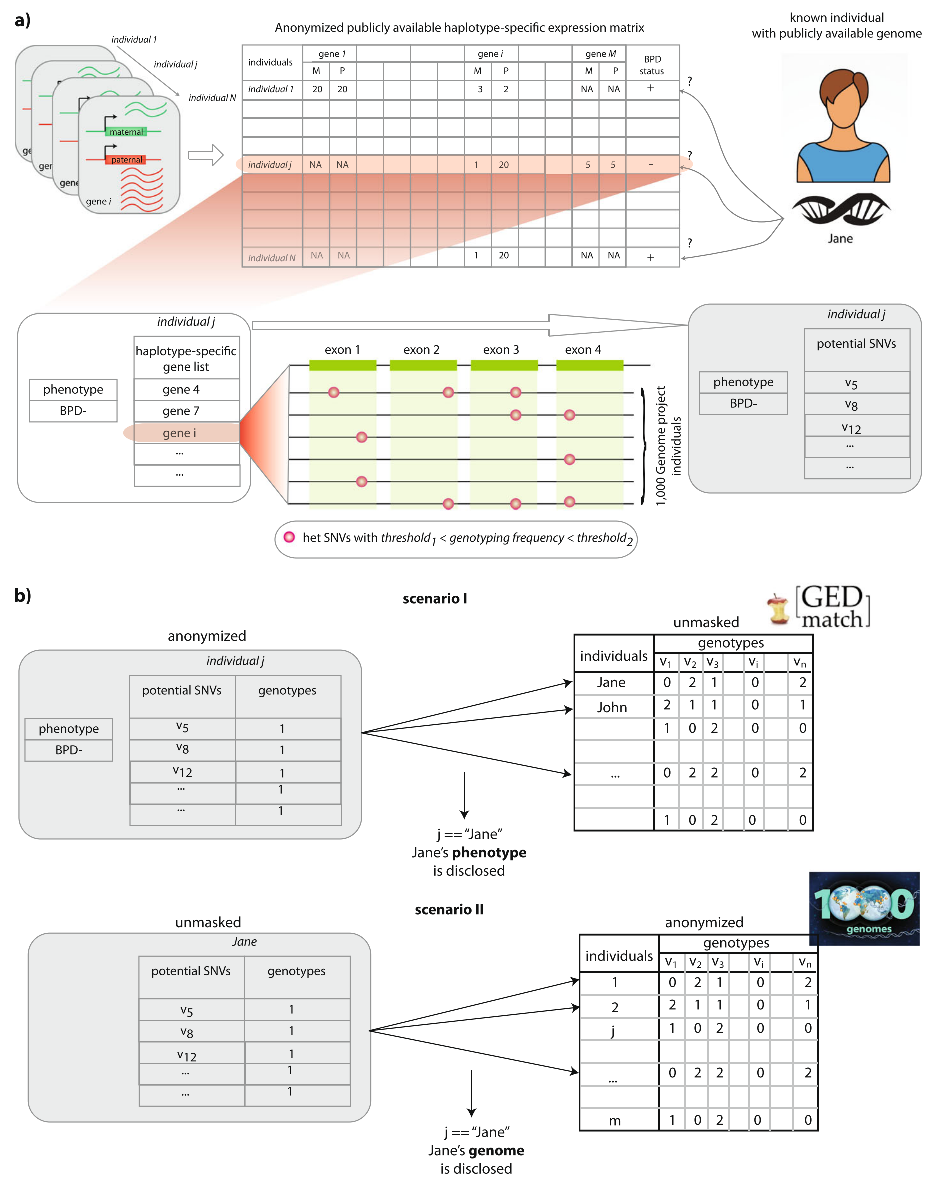
Here, we demonstrate that privacy breaches are possible solely by using a list of genes that are allele-specific in an individual without the knowledge of the underlying genetic variants themselves. As an example, we show these breaches using ASE data from individuals of the 1000 Genomes Project, in which the full genomes of the individuals are broadly shared.

Genomic privacy attacks in the form of linking two datasets together can be categorized differently based on whether the genomic variants observed from the linked datasets are noisy or perfect [12]. However, the privacy attacks we describe here differ in nature from previous linkage attacks [1, 2, 12, 13], as the genomic variants cannot directly be observed from the ASE data (Additional file 1: Fig. S1). Two privacy attacks can be performed using an ASE gene name list: (1) recovering the genome of an individual and (2) inferring the phenotype of an individual.

In the first attack, the adversary obtains a list of ASE gene names of a known individual (perhaps through electronic health records or a friendly conversation). The goal is to link these gene names to an anonymized publicly available genome dataset to recover the genome of the known individual (Fig. 1a).

In the second attack, a research study (e.g., GTEx, PsychENCODE) releases a publicly available anonymized database. This database contains the allele-specific expression of all genes for a number of individuals. It also contains the sensitive phenotypes (e.g., disease status such as HIV) of these individuals. The adversary compiles a comprehensive set of SNP genotypes from known individuals (e.g., those who participated in a research study), by using genetic genealogy databases such as GEDmatch. This database is a look-up table in which the columns correspond to the genotypes of the SNPs of all the individuals (Fig. 1b). The goal of the adversary is to uncover the phenotypes of these known individuals of interest. The allele-specific gene expression data can be summarized as a list of ASE gene names for each individual in the database created by comparing the gene expression between alleles. The adversary uses the genotypes to mine the ASE gene name lists and finally links the genotypes of known individuals to the phenotypes of the anonymized individuals (Fig. 1b).

As mentioned before, there is no explicit genotype information in a gene name. However, if a gene is determined to be allele-specific for an individual, then an accessible heterozygous single-nucleotide polymorphism (SNP) must be present somewhere on the gene body such that researchers were able to assign the gene expression into alleles. By using this information, we overlapped the exon locations of the reported ASE genes with heterozygous SNPs in a database of genomes. This allowed us to generate a candidate SNP/genotype list for each ASE gene (Fig. 1a, see Online Methods for details).



individuals	gene *1*						gene *i*				gene *M*		BPD status
	M	P					M	P			M	P	
individual 1	20	20					3	2			NA	NA	+
individual j	NA	NA					1	20			5	5	-
individual N	NA	NA					1	20			NA	NA	+

potential SNVs	genotypes
v_5	1
v_8	1
v_{12}	1
...	1
...	1

individuals	genotypes						
	v_1	v_2	v_3		v_i		v_n
Jane	0	2	1		0		2
John	2	1	1		0		1
	1	0	2		0		0
...	0	2	2		0		2
	1	0	2		0		0

potential SNVs	genotypes
v_5	1
v_8	1
v_{12}	1
...	1
...	1

individuals	genotypes						
	v_1	v_2	v_3		v_i		v_n
1	0	2	1		0		2
2	2	1	1		0		1
j	1	0	2		0		0
...	0	2	2		0		2
m	1	0	2		0		0

Fig. 1 Schematic representation of using allele-specific genes to de-anonymize individuals. **a** Schematic of going from a list of genes to a list of SNPs. **b** De-anonymizing a list of anonymous ASE genes using publicly available genomes from known individuals and inferring private phenotypes. **b** Recovering the anonymized genome of a known individual by using their ASE gene list

Note that all the SNPs that overlap with the exons are used as long as their genotyping frequency matches the criteria. If a SNP has more than one alternative allele, all alternative alleles are considered to be a candidate SNP. Next, we used a linking approach [12, 14] that weights the SNPs according to their frequency in the database. This approach scored every individual in the database based on the similarity between the genome and the candidate genotypes either by using a best matching linking score approach [12] or by using a probability distribution through entropy calculations [14].

We used lists of allele-specific gene names from 382 individuals [6] and attempted to link them to a database of 2504 individuals that includes these 382 individuals [15]. We were able to link 55% of the individuals to the database using only a list of allele-specific gene names per individual (Fig. 2a). We also calculated the precision as 74% and false-positive rate as 20% (see Additional file 2: SI). The percentage of correctly linked individuals increased to 80% when we relaxed our criteria from the best-matching individual to any individual in the top 20 best matches (Fig. 2b).

Next, we found that highly polymorphic human leukocyte antigen (HLA) genes were in the majority of the individuals' gene lists (Fig. 2c). We first used only HLA genes for the linking and found that we could link only 2.9% of the individuals to the database. We then removed the HLA genes from the original ASE gene list of each individual and found that the percentage of correctly linked individuals increased from 55 to 66%. Lastly, we removed the top 20 genes that were common to the individuals in the database from the gene list. This further increased the percentage of linked individuals to 68% (Fig. 2e, Additional file 2: SI for precision and false-positive rate). This shows that removal of multi-allelic and imprinted genes that are ASE for many individuals from the consideration will enhance the linking attack ability. The genes that are ASE for a smaller number of individuals are the most informative, but, as a trade-off, they distinguish fewer numbers of individuals (see Additional file 2: SI, Fig. 2d).

Our last test was to calculate the improvement in linking when we used auxiliary data. Knowledge of the biological sex of the individuals increased the percentage of correctly linked individuals from 55 to 60%. Adding only the ancestry information of the study participants also increased the percentage of correctly linked individuals to 60%. Although, knowledge of both ancestry and biological sex increased the percentage of correctly linked individuals by only ~1% (Fig. 2f), the percentage of correctly linked individuals with high statistical significance increased by ~10% (Additional file 3: Fig. S2). The reason behind these moderate increases are (a) because the non-European individuals were only 20% of the 382 individuals; hence, no additional information was gained for the majority of the individuals, and (b) the additional information gained by knowing biological sex is relatively low ($-\log_2(0.5)$) compared to the information obtained from a string of rare SNPs. We did not find a significant difference in the number of genes used per individual in the correctly linked and mislinked categories; however, when we weighted the number of genes by their length, we found that correctly linked individuals had longer ASE genes (Additional file 4: Fig. S3a,b). This makes sense as longer genes will contain more SNPs, which will increase the amount of information about the individual. We also did not find a significant difference in the number of candidate SNPs per individual between the correctly linked and mislinked categories (Additional file 4: Fig. S3c).

Conclusions

This study shows that although ASE does not explicitly reveal the location of the SNPs of an individual, using simple and straightforward biological knowledge can enable ASE genes to be linked to the genomes and/or phenotypes of study individuals. We showed the feasibility of this breach with data from individuals who provided broad consent. However, we envision that the same publicly available data could be used to infer private genetic variants of individuals who do not wish to release their genomes broadly. Furthermore, these inferred SNPs can lead to imputation of other genetic markers through linkage disequilibrium, which, in turn, might lead to even bigger privacy issues.

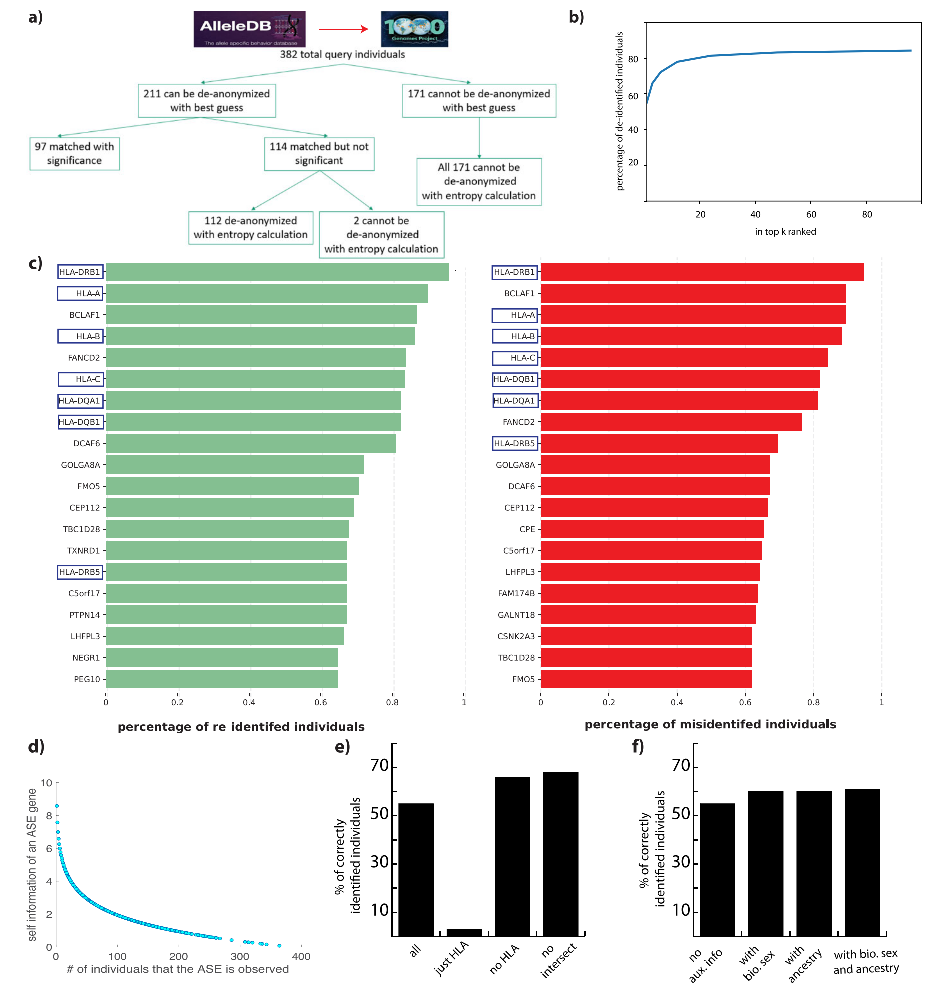


Fig. 2 Linking attack accuracy and impact of auxiliary information on linking ability. **a** The number of individuals that can be linked to their genomes with different statistical techniques. **b** The percentage of individuals that can be linked to their genomes when we relax the criteria from best match to top k ranked. **c** The top 20 genes that are found on the ASE gene list of correctly identified and misidentified individuals. **d** The self-information of ASE genes vs. the number of individuals that they are observed as ASE. **e** The percentage of correctly linked individuals when we used different combinations of ASE genes. **f** The percentage of correctly linked individuals when we used biological sex and/or ancestry as auxiliary information

As is the case with other molecular phenotype and functional genomics data, preventing the public release of ASE genes can hamper biomedical discoveries and clinical studies. Researchers could perform risk assessments of releasing gene names based on their polymorphism and length. Based on this assessment, some genes could be omitted from the list. However, this approach might reduce the utility of the released data as it will bias the list to shorter genes and genes that are commonly allele-specific. More sophisticated data sanitization techniques based on the SNPs might result in better sharing solutions. We believe that the best approach to mediate genomic privacy issues related to hidden information in summary-level functional genomics data is three-pronged: (1) develop clear and detailed informed consent policies, (2) educate participants on the risks and benefits of the study, and (3) establish laws and legislation to prevent bad actors from using genetic information to harm individuals, as noted by earlier genomic privacy studies [16].

Methods

Compiling a list of candidate SNPs from the ASE gene list

We first overlapped the gene names with the genes in the GENCODE comprehensive gene annotation file (release 19, GRCh37.p13) to pinpoint the location of the exons of these genes. For each gene, we found all the SNPs that overlapped with its exons by using a vcf file from 2504 individuals (1000 Genomes database). We first calculated the heterozygous genotyping frequency of these SNPs as $f(\text{genotype}_{\text{SNP}_i} = 1) = \frac{\text{\#of individuals with genotype}_{\text{SNP}i}=1}{\text{total\#of individuals}}$. We then removed the SNPs that had $0.1 < f(\text{genotype}_{\text{SNP}_i} = 1) < 0.5$ from the overlap list (see Supplementary Information for the rationale). We added the remaining SNPs to the candidate SNP list. We repeated this procedure for all of the genes in the list to obtain one final candidate SNP list.

Linking attacks

Let us assume that we have n total SNPs that can be observed in humans (e.g., all of the SNPs observed in the 1,000 Genomes Project). We can represent an individual's genome as a set $S = \{g_1, g_2, \ldots, g_n\}$, where $g_i$is the genotype of the i^{th} SNP. Candidate SNPs obtained using ASE genes become a subset of S, whose genotypes are assumed to be heterozygous ($g_i = 1$), i.e., $S_{can} = \{g_1 = x, g_2 = 1, \ldots, g_n = x\}$, where $g_i = x$ means SNP i is not in the candidate list; hence, its genotype is unknown.

Scenario 1

Let us assume we have an ASE gene list of a known individual. This means we can compile a list of heterozygous SNPs for this known individual. In this case, $S_{can} = \{g_1 = x, g_2 = 1, \ldots, g_n = x\}$ is the set of candidate genotypes for the known individual. The goal is to recover the genotypes for all of the SNPs in the set. Let us assume we have access to a database of anonymized genomes. Each anonymized genome j in the database can be represented as $S_j^D = \{g_1, g_2, \ldots, g_n\}$, where each genotype $g_i$is known.

Best match approach

For each individual j in the database, we find the intersection $S_{can} \cap S_j^D$ and calculate a linking score $L(i, can) = \sum_{t=0}^{t=|S_{can} \cap S_j^D|} \frac{1}{log_2\ f(g_t=1)}$, where $f(g_t = 1)$ is the ratio of individuals whose t^{th} SNP has the heterozygous genotype ($g_t = 1$) to the total number of individuals in D [previously defined in [12]]. To recover the genome for the known individual, we then rank all the $L(i, can)$ scores for all genomes in D in decreasing order. We denote the genome with the highest score as the genome of the known individual with candidate SNPs. To assess the statistical robustness of this prediction, we used our previously defined *gap* measure, which is the ratio between the L(i,can) score of the first-ranked individual (max=L(i,can)$_1$) and that of second-ranked individual (max$_2$=L(i,can)$_2$ and gap=max/max$_2$). We further calculate the statistical significance of *gap* by generating random candidate SNPs (as many as the original candidate SNPs), perform the above attack one thousand times, and compare the real *gap* value against the distribution of random *gap* values.

Entropy approach

The goal of this approach is to assign a probability of correctly linking the ASE list to each genome in D, which allows us to have a distribution. This approach is adopted from Narayanan and Shamtikov [14]. We calculate the probability of linking the candidate SNP list to a genome i in D as $\pi(i, can) = c.e^{\frac{L(i,can)}{\sigma}}$, where c is a constant to satisfy $\sum_i \pi(i, can) = 1$, $L(i,can)$ is the linking score described above, and σ is the standard deviation of the linking scores (Additional file 5: Fig. S4).

Scenario 2

The mathematical formulation of scenario 2 is the same as the first scenario. The only difference is that we have the genome of the known individual and we try to link this known genome to an anonymized ASE gene list, which is connected to a potentially private phenotype.

Identification of the top 20 common genes

After linking 382 ASE gene lists to a genome in D, we calculated the accuracy of the linking. We then separated the gene lists into two categories: (1) lists that led to correct re-identification and (2) lists that led to misidentification. We identified the genes that were shared across many ASE gene lists in both categories. Among the top 20 shared genes, we found that HLA genes were in the lists of >90% of both correctly re-identified and misidentified individuals. We then selectively removed different groups of genes (HLA, and genes at the intersection of both groups) and performed the linking attacks.

Usage of auxiliary data

We added one or two more features to our sets S_j^D (the genotypes of genome j in database D) and S_{can}(the candidate SNP genotype list) such that our new list does not only have genotypes but also includes biological sex and/or ancestry features. $S'_{can} = \{g_1 = x, g_2 = 1, \ldots, g_n = x, \text{sex} = M/F, \text{ancestry} = \text{EUR/AFR/AMR/EAS/SAS}\}$ and $S'^D_j = \{g_1, g_2, \ldots, g_n, \text{sex}, \text{ancestry}\}$ are our new sets and we look for the $S'_{can} \cap S'^D_j$ intersection to calculate the linking scores. Here, M and F are used for biologically male and female individuals, respectively. EUR, AFR, AMR, EAS, and SAS correspond to European, African, Admixed American, East Asian, and South Asian ancestries, respectively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02477-x.

Additional file 1: Figure S1. Different cases of linkage attacks.

Additional file 2. Supplementary Information for Recovering genotypes and phenotypes using allele-specific genes.

Additional file 3: Figure S2. Number of individuals that are correctly identified using different methodologies and different auxiliary information.

Additional file 4: Figure S3. a) Distribution of the number of ASE genes used for correctly identified individuals (in blue) and for misidentified individuals (in orange). b) Same as (a), but the genes are weighted by their length. c)

Distribution of number of candidate SNPs inferred for correctly identified individuals (in blue) and for mis-identified individuals (in orange).

Additional file 5: Figure S4. Examples of entropy-based matching for a correctly identified individual (HG00276) and mis-identified individual (NA20513).

Additional file 6. Review history.

Review history

The review history is available as Additional file 6.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

GG and MG designed the experiments. GG developed the methods. GG and NL acquired the data and performed the experiments. GG, NL, and SW analyzed the results. GG and MG wrote the manuscript. The authors read and approved the final manuscript.

Funding

This study was supported by grants from the NIH (R01 HG010749 to M.G. and K99 HG010909 to G.G.). This work is also supported by AL Williams Professorship Fund.

Availability of data and materials

The code used in this study can be found at https://github.com/gersteinlab/privaseq4 [17] and at https://doi.org/10.5281/zenodo.5218684 [18] under MIT License. 1000 Genomes vcf files are downloaded from ftp sites in https://www.internationalgenome.org. The processing steps and example files can be found in the data folder of the github page. A gene list for each individual is taken from http://alleledb.gersteinlab.org and can be found in the data folder of the github page. The locations of the exons for these genes are also provided in the data folder of the github page.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, USA. [2]Molecular Biophysics and Biochemistry, Yale University, New Haven, USA. [3]Molecular, Cellular, and Developmental Biology, Yale University, New Haven, USA. [4]Statistics and Data Science, Yale University, New Haven, USA. [5]Computer Science, Yale University, New Haven, USA.

Received: 14 April 2021 Accepted: 23 August 2021
Published online: 07 September 2021

References

1. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat. Methods. 2016;13(3):251–6. https://doi.org/10.1038/nmeth.3746.
2. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nature Genetics. 2012;44(5):603–8. https://doi.org/10.1038/ng.2248.
3. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreservation Biobanking. 2015;13(5):307–8. https://doi.org/10.1089/bio.2015.29031.hmm.
4. Castel SE, Aguet F, Mohammadi P, et al. A vast resource of allelic expression data spanning human tissues. Genome Biol. 2020;21:234. https://doi.org/10.1186/s13059-020-02122-z.
5. Rozowsky J, et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7(1):522. https://doi.org/10.1038/msb.2011.54.
6. Chen J, Rozowsky J, Galeev T, et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016;7:11101. https://doi.org/10.1038/ncomms11101.
7. Onuchic V, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science. 2018;361(6409):eaar3146. https://doi.org/10.1126/science.aar3146.
8. Zhang K, Li J, Gao Y, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nat Methods. 2009;6:613–8. https://doi.org/10.1038/nmeth.1357.
9. Lee C, Kang EY, Gandal MJ, Eskin E, Geschwind DH. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. Nat. Neurosci. 2019;22(9):1521–32. https://doi.org/10.1038/s41593-019-0461-9.

10. Valle L, Serena-Acedo T, Liyanarachchi S, Hampel H, Comeras I, Li Z, Zeng Q, Zhang HT, Pennison MJ, Sadim M, Pasche B, Tanner SM, de la Chapelle A. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. Science. 2008;321(5894):1361–5. https://doi.org/10.1126/science.1159397.
11. Yan H, Dobbie Z, Gruber SB, Markowitz S, Romans K, Giardiello FM, Kinzler KW, Vogelstein B. Small changes in expression affect predisposition to tumorigenesis. Nat Genet. 2002;30(1):25–6. https://doi.org/10.1038/ng799.
12. Gürsoy G, et al. Data sanitization to reduce private information leakage from functional genomics. Cell. 2020;183(4):905–17. https://doi.org/10.1016/j.cell.2020.09.036.
13. Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat. Commun. 2018;9(1):2453. https://doi.org/10.1038/s41467-018-04875-5.
14. Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets. 2008 IEEE Symposium on Security and Privacy (sp 2008). 2008. https://doi.org/10.1109/sp.2008.33.
15. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
16. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321–4. https://doi.org/10.1126/science.1229566.
17. Gürsoy G, Lu N, Wagner S, Gerstein M. Recovering genotypes and phenotypes using allele-specific genes. GitHub. https://github.com/gersteinlab/privaseq4.
18. Gürsoy G, Lu N, Wagner S, Gerstein M. Recovering genotypes and phenotypes using allele-specific genes. https://doi.org/10.5281/zenodo.5218684.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.