Noninvasive, low-cost RNA-sequencing enhances discovery potential of transcriptome studies

Molly Martorella

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Noninvasive, low-cost RNA-sequencing enhances discovery potential of transcriptome studies

Molly Martorella

Transcriptome studies disentangle functional mechanisms of gene expression regulation and may lend key insights into disease mechanisms. However, the cost of RNA-sequencing and types of tissues currently assayed pose major limitations to study expansion and disease-relevant discovery. This thesis develops methods for sampling noninvasive biospecimens for transcriptome studies, investigating their technical and biological characteristics, and assessing the feasibility of using noninvasive samples in transcriptomic and clinical applications. Chapter 1 explores the technical and biological features of four potential noninvasive sample types (buccal swabs, hair follicles, saliva, and urine cell pellets) in a pilot study of 19 individuals whereby four separate collections of each tissue were performed (i.e. 76 samples/tissue, 304 samples in total). From this data, consistency of library preparation, cell type content, replication of GTEx cis-eQTLs, and disease applications were assessed. In all, hair follicles and urine cell pellets were found to be most promising for future applications. Chapter 2 investigates the scaling potential of noninvasive sampling in SPIROMICS, a COPD clinical cohort. To do so, 140 hair follicle and 110 buccal swab samples were collected from seven different clinical sites. Consistency of sample quality was observed to be high for hair follicles, and hair cell type abundance estimates were consistent within SPIROMICS and compared to the 19 subject pilot study. Mapping of cis-eQTLs in hair revealed 339 associations not identified in any prior study. These cis-eQTLs show higher replication in GTEx tissues that share cell types with hair follicles, indicating hair follicles

may indeed capture gene expression regulatory mechanisms found in more invasive tissue types of the body. This thesis suggests future use of noninvasive sampling will facilitate discovery by increasing sample sizes in more diverse populations and in tissues with greater cell type diversity and biological relatedness to disease mechanisms. Moreover, the nature of noninvasive sampling enables complex, longitudinal study designs with greater ability to capture context-dependent mechanisms of genetic regulation not currently able to be interrogated.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work would not be possible without the contribution of many mentors throughout the many years I have spent in academia. I would first like to acknowledge Catherine Oertel, my undergraduate career advisor, without whom I would not have pursued a research career at all. During my undergraduate years I had the fortune to work in Todd Emrick's lab at UMASS, Richard Huganir's lab at JHU, and both Michael Nee and Jan Thornton's labs at Oberlin. Each of these opportunities shaped my future research aspirations and led me to pursue the work I do today.

I would particularly like to thank Christopher Deppmann and Tony Spano at UVA. Chris provided me immense independence during my research gap year before starting the MD/PhD program at Columbia. This experience profoundly impacted my readiness to pursue an independent research project during my PhD. Tony taught me to maintain a healthy level of skepticism, humor, and jadedness in regards to science, as well as let me follow him around while he taught me the ins and outs of turning someone's incomprehensible methods section into a useable protocol for yourself.

During my PhD, there were many actors contributing to my happiness and growth as a scientist (unfortunately all are not able to be mentioned here). First, my amazing friend and fellow MD/PhD masochist, Joana Petrescu, who I could always rely on for emotional support and for trying to figure out what we're supposed to do next with this crazy degree. Fernanda Garcia-Flores became my mentee during the final stretch of my PhD, and I am grateful for her enthusiasm, humor, and hard work that maintained my optimism during that time. Then of course are my fellow (now former) graduate students in the Lappalainen lab – Margot Brandt, Elise Flynn, and Jonah Einsen – with all of whom I have shared the indescribable experience of

being a PhD student. I would be remiss not to mention Alper Gokden, Stephanie Hao, and Bill Mauck for their help troubleshooting unforeseen problems and repairing robots in the wet lab, and Paul Hoffmann for helping me overcome various coding crises throughout the years. And of course, Kristina Buschur, who I have worked hand in hand with on the work of Chapter 2 and who I have commiserated with over the bureaucracies of various unnamed institutions.

I would particularly like to thank two former Lappalainen lab post-docs, Silva Kasela and Stephane Castel. Silva taught me how to write iterative, reusable code on the scale necessary to complete a large and intensive computational project. Her organizational skills cannot be beat, and I am forever grateful her wisdom was imparted in me. The low cost, noninvasive project presented here was started by Stephane, and he certainly helped me finish it. Stephane has a very direct, no-nonsense approach to science, provides excellent feedback, and working with him helped me overcome my hesitancy and lack of confidence in my scientific decision making.

And finally, a special thanks to my thesis advisor, Tuuli Lappalainen, and my thesis committee Molly Przeworski, Peter Sims, and Graham Barr. Throughout my PhD I have been grateful to work in a welcoming and stimulating lab environment. This, of course, is because of Tuuli. As a mentor, Tuuli leads by example, showing her mentees it is possible to be a brilliant, highly successful scientist while maintaining a balanced life and being kind and collaborative. She regularly speaks to issues of inequality, diversity, and inclusion, and this in turn curates a lab space of individuals who are similarly passionate in these regards and who are comfortable being themselves. It is also clear she takes personal, continued interest in her mentee's success long after they leave the lab, and I am excited to continue to work with and learn from her in the future.

Dedication

This thesis is dedicated to my cat Moon, who awkwardly stared at me for most of the analyses presented here and effectively interrupted nearly every zoom meeting I had, and also to my partner Jimmy, whose passion, cooking, wit, and kindness inspire me to be a better version of myself every day.

Introduction

Initial Advances and Difficulties Mapping Human Genetic Variation to Phenotypes

The field of human genetics centrally aims to understand the mechanisms by which genetic variation contributes to phenotypic variation. In the context of medicine, this question historically focused on disorders of Mendelian inheritance, whereby a change at a single locus results in disease. In the past 20 years since the human genome was mapped, SNP arrays and sequencing technology underwent rapid advancements. This not only expedited the genetic mapping of Mendelian disorders and added further complexity as to our understanding of their manifestation, but also enabled genetic interrogation of complex, common diseases whereby many genetic changes contribute to a phenotype of variable presentation. At the outset, this approach was anticipated to yield massive, transformative insights into our understanding of disease risk, mechanisms, and potential treatments.

However, the first genome wide association studies (GWASs) attempting to link sequence changes to human traits revealed the genetic architecture underlying most traits to be non-trivial. For one, there are many more genetic associations of smaller effect size than initially expected, and the majority of trait-associated variants localize to noncoding regions whereby their target gene and mechanism of action are not readily discernible from genetic sequencing alone[1–3]. Even identifying the causal variant is complicated by linkage disequilibrium blocks tagging both causal and nearby sites[1–3]. Extreme Eurocentric bias in early studies further exacerbates this problem and is still needing remedy today[4]. Furthermore, it is increasingly apparent that each trait will not be associated with a separate set of variants but that pleiotropy,

whereby a variant affects multiple traits, is a pervasive issue and adds further complexity to resolving genetic disease risk[1–3,5].

Much of these observed phenomena are now largely recognized as an anticipated product of our evolutionary past and may not be understood without integrating data from multiple sources and contexts. The process of evolution typically involves the co-option of previously existing molecular pathways and structures, thus providing some explanation for the pleiotropic role of many genes[6]. In comparing effect size distributions of variants associated with neutral, directional, or stabilizing selection on some trait, it is known that directional and stabilizing selection result in lower variance[7]. Most traits are under stabilizing selection whereby selection opposes variants resulting in large phenotypic changes, and those that do impart large changes are typically found at much lower frequency in the population[7,8]. Thus the commonly occurring variants discovered via GWAS most often result in small effect sizes on a given trait[6,7]. Further, many phenotypes display a pattern of evolutionary trade-off, in which an adaptation imposes a penalty on some other trait, or environmental mismatch, whereby modern day exposures are at odds with prior evolutionary changes[6]. These observations suggest context may play a substantial role in determining the mechanism and consequence of a given variant and that the variant effect may only be observed under specific conditions. As such, acquisition of an array of intermediate molecular phenotypes across multiple clinical and environmental conditions is necessary to reveal variant consequences and properly interrogate GWAS hits.

Transcriptome studies provide one such area of promising investigation. In comparing the transcriptome across tissues, cell types or states, under different environmental or treatment conditions, or from samples originating from different donors, one may begin to unravel which genes and molecular pathways centrally contribute to various biological and disease-related

processes[9,10]. Characterizing the transcriptome began initially with microarrays, which only quantify a subset of all genes (and only known isoforms), restrict analyses to predetermined sets of probes, or impart high costs per experiment due to probe design[10]. RNA-sequencing enables quantification of many more transcripts, and, over time, advancements in technology have reduced cost, increased throughput, and improved the read length such that information regarding an even greater number of transcripts is obtained[10]. Further advancements in single cell RNA-sequencing will only continue to augment these efforts[10]. However, it should be noted that RNA-sequencing cost still imposes a large barrier for study expansion, and often reductions in cost come at the expense of valuable biological data[11–13].

A subset of transcriptome studies focus specifically on observations of gene expression or splicing changes in the context of genotypic variation. These types of studies, termed molecular quantitative trait loci (molQTL) studies, hold potential to contribute to the elucidation of GWAS findings and provide fruitful discovery in translational applications. For one, overlap between GWAS variants and those that elicit a change in gene expression or splicing may help prioritize which variants could be mechanistically important for a trait and should be further interrogated in laboratory settings[14,15]. Often expression and splicing events are cell-type specific, and thus discovering which cell types and tissues GWAS variants exert their effects may yield insights into disease etiology, pathophysiology and potential treatments[16–19]. This approach also provides opportunity to illuminate context specific genetic regulatory events that may be key to understanding genetic and environmental contributors to disease mechanisms[20,21]. However, there are several obstacles currently preventing transcriptome studies from delivering on their promised potential. Advances, limitations, and future directions of transcriptomics will be the focus of this thesis and covered in detail in the following sections.

Molecular Quantitative Trait Loci

One technique for uncovering the phenotypic consequences of genetic variation is molecular Quantitative Trait Loci (molQTL) mapping. In fields of genetics studying other species, QTLs have long been observed and studied, but only in the last 15 years or so has this approach been applied to human datasets to further resolve disease-associated genetic variation[15,22]. QTL mapping involves comparing levels of an observed trait across different genotypes. It should be noted that genotypic variation may be tested against any trait for that locus to be a potential QTL. However, for the purposes here the focus will primarily be in regards to expression (eQTLs) and splicing QTLs (sQTLs), as these have been most heavily interrogated. Discoveries are typically referred to as eVariant-eGene or sVariant-sGene pairs to indicate a given variant associates with either expression levels or transcripts observed for a given gene. Previous findings support that GWAS hits are enriched for eQTLs and their discovery may provide insights into the regulatory consequences of GWAS variants[5]. Thus far, these sites identify a target gene for approximately 47% of GWAS loci[22].

Typically, QTLs are defined either as local or distant, and in cis or in trans. The following describes consequences of eQTLs, but the same nomenclature applies to sQTLs and transcript abundances. Local eQTLs affect expression of a gene on the same chromosome and are often within 1MB of the gene, whereas distant eQTLs impact expression of genes located anywhere throughout the genome[15,23]. Cis-eQTLs directly modify the expression of a gene, whereas trans-eQTLs affect gene expression via some diffusible factor[15]. Most often, cis-eQTLs act locally and trans-eQTLs affect genes distally, but this is not always the case[15].

In comparing eQTLs to sQTLs, observations so far suggest these loci occur largely independent of one another[24,25]. sQTLs tend to be more concentrated in transcribed regions, the

5’ UTR, and for splice-site acceptors, donors, and regions[23,25]. On the other hand, eQTLs are found proximal to transcription start sites and enriched in transcriptional regulatory elements like transcription factor binding sites, open chromatin and active promoter sequences, and enhancers[24–26]. If gene expression levels and splice junctions used are shared across tissues, then most often sQTLs are also overlapping[27]. However, many genes and splice junctions are cell-type specific, and overall, sQTLs display increased tissue-specificity relative to eQTLs[25,28,29].

trans-eQTLs are the least characterized regulatory loci but are anticipated to contribute greatly to mapping genotypes to phenotypes via revealing larger networks of gene expression regulation. In fact, current observations suggest the majority of gene expression heritability may be explained by trans effects (~60-75%)[15]. However, discovery thus far is limited due to the increased burden of multiple testing corrections as well as the relatively small effect sizes of trans-eQTLs[30,31]. Notably, most trans-eQTLs also act as a cis-eQTL to a nearby gene[21,25]. However, compared to typical cis-eQTLs, the majority of loci are located farther from the TSS and enriched in CTCF binding sites, open chromatin, and promoter-flanking regions, further supporting their role in overarching and distal regulatory effects[25]. Much like sQTLs, trans-eQTLs identified so far display greater tissue specificity relative to cis-eQTLs[25].

In addition to genomic features and location, QTLs may be defined in terms of the context in which they are active. Steady-state QTLs produce a consistent effect on gene expression regardless of changing conditions[21]. Most studies to date capture a single time point or use post-mortem tissue expression, and thus are primarily assumed to discover steady-state QTLs. Some QTLs derived from these studies show a pattern of changing phenotypic variance depending on genotype (vQTLs), and it is anticipated that a gene by environment interaction or other context-dependent effect underlies this pattern[32]. These QTLs are termed dynamic or

interaction QTLs (iQTLs), and their effect size changes depending on the value of some other known variable. The primary difference between dynamic and iQTLs is in regards to experimental design rather than underlying biology. iQTLs typically refer to a static variable and are modeled using an interaction term for the analysis. Thus far, iQTLs have been most heavily studied in regards to sex and accounting for cell type proportions[16–19,33]. Dynamic QTLs, on the other hand, refer to a temporally applied condition or biological event such that the QTL effect may be measured before and after a treatment or at multiple timepoints. For these studies, QTL analysis is performed per each measurement and results are compared afterwards rather than using an interaction term. Studies of immunological responses and developmental processes often interrogate dynamic QTLs[34–36].

Performing Quantitative Trait Loci analyses

Quantitative Trait Loci (QTL) analysis involves linearly modeling a phenotype as a function of genotype with covariates accounted for in the model (Equation 1). Notably, QTL models may be modified to include interaction terms to identify potential iQTLs (Equation 2).

(1) $y \sim \beta_G * G + \beta_C * C + \varepsilon$

where y is phenotype; β_G is the effect size of genotype (G) encoded as 0,1,2; β_C is the effect size of covariate (C); and ε is noise.

(2) $y \sim \beta_G * G + \beta_I * I + \beta_{GxI} * G * I + \beta_C * C + \varepsilon$

where y is phenotype; β_G is the effect size of genotype (G) encoded as 0,1,2; β_I is the effect size of some known potential interacting variable (I), β_{GxI} is the effect size of the interaction between genotype (G) and the variable (I); β_C is the effect size of covariate (C); and ε is noise.

The immensity of data collected in human studies today necessitates computationally efficient algorithms for association testing and statistical approaches that address the enormity of associations tested without overcorrecting. The most commonly used program for QTL analysis, tensorQTL, navigates both of these problems. First, the association testing is done based on efficient matrix operations established by MatrixQTL, and this greatly reduces the computational load[37]. Multiple testing corrections need to be implemented at both the locus and genome-wide level. Permutation testing previously served as the standard approach for significance testing of loci, but the permutation p-value limit is determined by the total number of permutations, and thus p-values less than 10^{-3} remained computationally infeasible. FastQTL introduced modeling the p-value distribution for every locus using a beta distribution[38]. This method defines the extreme tail of the null distribution without explicitly sampling from it, again, greatly reducing the computational burden of p-value calculations. Genome-wide significance is determined using the FDR (false discovery rate) approach described by Storey & Tibshirani[39]. This method provides a data-driven estimate of the expected proportion of null findings (v.s. Bonferroni-Hochberg which assumes this number to be 1), and therefore allows for less-stringent thresholding. The variable pi0 typically represents this estimated value, with pi1 being 1 - pi0 and representing the expected proportion of truly alternative findings[39]. FDR is calculated by dividing the expected number of false positives by the total number of findings (i.e. FP/(FP + TP)) below a given significance level. Significance per loci is calculated using the q-value, which is defined as the expected proportion of false positives when calling a feature as or more extreme significant. This reduced stringency compared to other multiple testing corrections is important because a very large fraction of phenotypes are anticipated to be affected by genetic variation[38,39].

To discover overlaps between GWAS and QTL hits, colocalization approaches are used. The most frequently employed method is coloc[40,41]. This Bayesian method compares the p-value distribution of GWAS and QTL studies and outputs posterior probability estimates for five hypotheses, H0-H4, with H0 being the null. H1 states evidence for a GWAS variant but no QTL effect at that locus, while H2 supports a QTL effect without a GWAS association. H3 supports distinct QTL and GWAS findings at a given locus. Finally, H4 localizes the GWAS association and QTL effect to the same variant. It should be noted that allelic heterogeneity, whereby multiple variants are associated with a trait, still pose significant difficulties for colocalization approaches, and coloc explicitly imposes a single causal variant assumption[40,41]. Stepwise regression and masking allow one to test multiple variants at a given locus using coloc, however, stepwise regression requires an LD matrix and masking tends to bias results towards supporting H3 when H4 is true[41]. In all, colocalization methods should be viewed not necessarily as providing evidence for a causal variant, but rather as providing further evidence for shared or distinct causality.

One emerging method of note is transcript wide association studies (TWAS). These analyses predict gene expression using known cis-regulatory loci and then associate predicted gene expression with traits[42]. More specifically, gene expression weights per variant are determined using a reference transcriptome and elastic net regression to model gene expression as a function of genotypic variation[42,43]. This model is then used to predict gene expression data for a study with only genotyping data available[42,43]. This predicted gene expression is then associated with the trait using regression or non-parametric methods[43]. This approach is still under development, and many limitations in regards to expression co-regulation, correlated expression levels across individuals, and reference panel biases remain challenging[44].

Discoveries from and limitations of QTL studies in unraveling GWAS signals

Questions have been raised as to whether QTL studies truly narrow the scope of possible disease-relevant variants and successfully pinpoint genes and molecular changes causally related to the GWAS trait. Largely, this critique is in regards to our current knowledge of eQTLs. Despite discovering an eVariant for nearly every gene, fewer than half of GWAS variants colocalize with an eQTL and eGenes frequently do not overlap with anticipated disease genes and pathways[21,26,45]. Moreover, only ~11% of complex trait heritability may be explained by cis gene expression regulation[46]. These observations may not merely be explained by a lack of necessary sample sizes and power, as larger studies reveal cis-eQTL discovery saturates at approximately 600 samples[21,25,26]. Notable features of e/sQTLs and limitations in our current approaches may underlie the lack of insight provided by transcriptomics thus far.

One of the most comprehensive studies of e/sQTLs is the Genotype-Tissue Expression (GTEx) project. This study collected 54 tissue samples from 838 healthy, post-mortem donors and performed bulk RNA-sequencing for e/sQTL mapping, and much of our current QTL knowledge derives from this work[25,29]. One such observation is the pattern of QTL activity across different tissues. Tissue sharing of cis-eQTLs follows a bimodal distribution, whereby cis-eQTLs are either found in all tissues or a few tissues[25]. This same pattern is observed for sQTLs and trans-eQTLs, however, these loci are generally more tissue specific[25,29]. Additionally, this data revealed nearly every gene is associated with a variant affecting its expression, and thus nearly every gene is an eGene[23–25,29]. Some eVariants affect the expression of multiple genes, and many genes display allelic heterogeneity whereby multiple variants play a role in their expression regulation[23–25]. The ubiquity and redundancy by which variants affect gene regulation and their broad sharing across many tissues emphasizes the immense pleiotropy present across

the genome and the complexity of disentangling variant-gene-trait associations. These observations suggest, much like GWAS, that many of our observations may not capture signals central to disease biology.

Indeed, e/sQTLs showing activity across a large number of tissues typically show greater GWAS trait pleiotropy[19,47]. Conversely, e/sQTLs with tissue specificity show enrichment for GWAS signals[19,47]. Bulk RNA-sequencing tends to distort or result in loss of signal due to the mixture of cell types present with varying effect sizes[16,17,20]. Accounting for cell type composition in the GTEx tissues increases e/sQTL discovery and further enhances colocalization of variants with GWAS[16,17]. Typically tissues with shared underlying cell types also show a higher degree of QTL capture and effect size sharing, and as such, these tissues also tend to be similarly enriched for GWAS disease relevant signals[18,25]. Therefore, narrowing investigations to tissues suspected to play a central role in disease mechanisms may provide greater insights into variants identified by GWAS. However, tissues involved in disease processes tend to be highly invasive to collect and thus inaccessible.

Instead, many studies rely on blood-related biospecimens for QTL analyses. This approach fruitfully identifies many regulatory mechanisms relevant to blood cell traits, and the large sample sizes enable trans-eQTL network discovery[31]. However, of the cis-eQTLs in linkage disequilibrium with GWAS variants, fewer than half have been found in blood[47]. Moreover, GTEx demonstrated the genetic regulatory mechanisms found in blood differ in effect size or are absent compared to other tissues of the body, making blood an outlier in its genetic regulatory architecture[25,29]. Likely this difference is driven by the unique cell type composition and function of the cell types in blood[18,47]. As stated above, this bears implication for potential future disease-relevant discovery when using blood-related samples. In cases where the cell

types in blood are centrally implicated in the disease or trait of interest, meaningful biological insights may be derived[16–18,47]. Elsewise, the majority of discoveries reflect broad mechanisms of gene regulation that lack association with disease[19,47].

By and large accounting for cell type composition and/or narrowing the focus to disease-relevant tissues likely will not in and of itself resolve the lack of overlap between GWAS variants and cis-eQTL findings. Another consideration is the manner in which samples for e/sQTL studies are obtained. Currently samples for e/sQTL analysis are collected at a single time point, potentially post-mortem, and are thus steady-state e/sQTLs. Most post-mortem samples also originate from reportedly healthy donors, and thus the gene regulatory networks and associated molecular pathways bear little insight into the aberrative biological processes of disease[21]. This issue in study design is further supported by general differences in genomic features seen for cis-eQTLs versus GWAS variants. For one, highly conserved genes and genes involved in many protein-protein interaction networks show depletion for cis-eQTLs[6,7,21,23,24,48]. Generally, eQTLs reside near transcription start sites and are found less frequently at loci with transcription factor activity, but GWAS variants typically localize to more distal loci and to functional annotations under selective constraints[26,48]. Moreover, variants with higher allele frequencies display reduced QTL effect sizes and vice versa[6,7,21]. Jointly, these results indicate the way evolutionary processes may influence the genetic architecture underlying disease processes and suggest current approaches to assaying gene regulation may not capture the most phenotypically important signals. Likely many gene regulatory processes are context-specific, and thus evade selective forces under most circumstances unless the exposure is present[6,7,21]. Historically, mutations affecting LOF intolerant or highly conserved genes tend to bear greater downstream disease-relevant consequences. Serial sampling may reveal these biological

processes via discovery of dynamic e/sQTLs involved in more conserved molecular pathways[21,34–36].

In all, e/sQTL studies at this time largely capture data with tangential relevance to disease and thus struggle to parse the vast number of GWAS variants into discernible mechanisms. This is due to both the tissue i.e. cell-type specificity of disease processes as well as the temporal aspect of environmental exposures and disease progression. Overall, current approaches fail to obtain cell types and contexts necessary for impactful discovery.

Changing the transcriptome study design paradigm

Barriers to wide scale transcriptomic discovery and clinical use fall into two general categories. One regards the limitations of the biology captured, as described above, and this issue is largely propagated by prohibitive financial and logistical costs as well as lack of longitudinal access to pertinent biospecimens[49,50]. The second issue relates to the lack of comprehensive sampling and study of underrepresented and other vulnerable and typically excluded populations[4,44,51].

One way to navigate these problems is by exploring the biology of alternative, less invasive biospecimens. This approach would not only increase the flexibility of experimental designs to enable time course studies, but it would also mitigate cost and access barriers by reducing the specialization of the personnel and institutions required to obtain the samples. Also, depending on the source of the samples, the cell types captured could bear consequences for a much larger swath of diseases compared to blood. Combined, this enhances studies because greater sample sizes may be achieved with the same resources and because more sensical biological tissues and contexts may be captured.

Importantly, the reduced invasiveness of sample collection may result in increased subject enrollment and correct for current biases in transcriptome studies. The reasons for inequitable research recruitment are many. For one, minority populations have been historically subjected to unethical and immoral research studies[52,53], and hesitancy surrounding research involvement understandably still lingers today[54,55]. There are additional structural factors at play resulting in reduced access to healthcare and decreased access to research study enrollment[56]. Because these populations often lack access to preventative care, they present with more severe illness, and for this reason may be excluded from research studies[57,58]. Unfortunately, the low sample sizes yielded from these populations frequently leads to their removal in downstream analyses[59]. These issues require more active recruitment from these communities and rapid scaling of sample sizes because the risk of worsening healthcare disparity due to clinical use of genomic and transcriptomic findings is high[4]. Noninvasive transcriptome sampling provides an avenue to do so. These biospecimens are highly unlikely to exacerbate current health conditions, and thus individuals will not be excluded by study criteria, and collection does not require a treatment or invasive procedure that may evoke discomfort and distrust in potential subjects. The simpler logistics of sample collection may also facilitate sampling in less advantaged healthcare and community settings. It is worth noting that others normally excluded from research studies, like children and pregnant persons, may be more readily enrolled due to the lack of danger posed by noninvasive sampling. This opportunity would provide additional insight into biological processes not currently available and provide better healthcare to these populations not currently represented.

As a final note, some have suggested computational approaches like transcriptome wide association studies (TWAS) may be able to synthetically increase sample sizes by leveraging

datasets with only genotyping data. Arguably, this will not resolve any of the aforementioned issues. The design of TWAS require a reference expression panel, which, as stated above, these panels are largely European and capture steady-state cis-eQTL regulatory processes. The mechanisms of disease across populations may indeed converge on the level of gene expression, however, the exact variants exerting control over gene expression or splicing mechanisms may differ[44]. When performing TWAS, this results in the dropping of many sites and worse gene expression predictions for underrepresented groups[44]. Finally, as previously stated, the majority of trait heritability is expected to be determined by trans regulatory effects. Current expression reference panels lack this information and thus would not be able to fully recapitulate disease mechanisms.

Thesis scope

Contained here are two chapters detailing the piloting and then clinical application of low cost, noninvasive RNA-sequencing methodologies. Chapter one investigates the use of buccal swabs, saliva, hair follicles, and urine cell pellets in traditional transcriptome analyses. An in-house, low-cost library preparation method is prepared in parallel to commercial kits to demonstrate the feasibility of implementing reduced cost approaches for these biospecimens. The quality of the data yielded is compared to results from a cell line and to RNA-sequencing metrics typical of standard RNA-sequencing tissue types. Both biological and technical sources of variability are delineated. And finally, use of these samples in typical transcriptomic and disease-relevant applications is tested.

Chapter two explores the use of buccal swabs and hair follicles in a clinical setting of patients with Chronic Obstructive Pulmonary Disease (COPD). First, the quality of the data yielded is established. The underlying biology of these samples in regards to cell type is

compared to estimates observed in the pilot study as well as to biological replicates within the cohort. Finally, we discover eQTLs in hair follicles and demonstrate replication of these eQTLs in GTEx corresponds with underlying cell type composition. Colocalization of hair eQTLs with variants previously discovered in lung function and imaging GWAS is a future direction of this work.

In all, this thesis aims to address key issues preventing scaling and biological relevancy of QTL discoveries.

Chapter 1: Noninvasive transcriptomics in a 19 subject pilot study[1]

1.1 Introduction

Large scale transcriptomic studies have the potential to disentangle the functional consequences of genetic variation and play a key role in realizing the goals of precision medicine. Prospective applications include biomarker identification of disease risk, onset, prognosis, and treatment response, discovery of potential therapies, and assessing outcomes of *in vitro* perturbations of environmental or pharmacological exposures[60]. One major area of investigation involves the integration of transcriptomic data with genetic information. Despite the identification of tens of thousands of trait-variant associations from thousands of genome-wide association studies (GWAS)[61], genomics alone has failed to unravel the mechanistic underpinnings driving these findings[1,62,63]. Transcriptomic data lends interpretation and prioritization of genetic variants for follow-up experimental and clinical investigation[3,15,64]. In the realm of cancer research, this approach has been fruitful in identifying early diagnostic markers, classifying cancer subtypes, identifying novel drug targets, and optimizing treatment choice[65–67]. In cases of rare, genetic disease, diagnosis is enhanced by inclusion of transcriptomic data due to improved detection and annotation of rare functional variants[68–70]. For common traits and diseases, however, discovery is complicated by trait polygenicity, linkage disequilibrium, small effect sizes of variants, and widespread pleiotropy[5]. Additionally, key pathways and potential molecular targets for a given trait may only be observable in relevant tissues and specific contexts[21]. Finally, the highly Eurocentric sampling of genetic and transcriptomic data results in greatly diminished performance of genomic risk assessments across ancestries[4]. To overcome the inherent limitations of genomic and transcriptomic data, capture meaningful biology, and mitigate impending health disparities, transcriptomic studies must expand to include

large, multi-ancestry sample sizes and sampling methodology that meets the needs of complex experimental designs.

There are several barriers to transcriptomic study expansion. For one, processing sufficient sample sizes for discovery remains prohibitively expensive. In the early 2000s, sequencing costs rapidly declined, but this progress has stalled since 2015[49]. An array of low cost library preparation methods exist[11–13], though nearly all of them sequence only the 3' end of transcripts, losing valuable splicing information in the process. New sequencing technologies have very recently emerged and promise future cost reductions[71], and others propose greater enrollment at lower sequencing depths to increase effective sample sizes with the same monetary resources[72]. Nonetheless, these approaches have yet to be widely implemented.

A less studied area of potential improvement is biospecimen choice. At this time, whole blood and peripheral blood mononuclear cells (PBMCs) are the most readily collected tissue for expanding transcriptomic studies. However, the Genotype-Tissue Expression (GTEx) project, the largest and most comprehensive study of genetic regulation across post-mortem human tissues, found blood is an outlier in its gene expression regulatory mechanisms relative to other tissues of the body[25,29], and the majority of expression quantitative trait loci (eQTLs) in linkage disequilibrium with GWAS variants are not found in whole blood[47]. Previous observations suggest this difference is driven by the unique functions and cell type composition of blood versus other tissues of the body[16–19]. This biological difference poses major limitations to discovery in contexts where the cell types in blood are not centrally implicated in trait development or disease pathogenesis[25,29,47]. Sampling directly from meaningful tissue types, if the tissues of interest are known, provides more biologically relevant data, however, current approaches primarily include surgical extractions, invasive biopsies, and post-mortem donations.

The nature of these study designs results in high cost and complicated logistics[50], low participant enrollment that is often biased against vulnerable and minority populations, and it only captures the steady-state transcriptome observed at a single time point. Dynamic changes in the transcriptome and molecular QTLs (molQTLs), which change over the course of development, disease, or environmental conditions, are suspected to lend even greater insight into the genetic architecture of the genome and are essential for using the transcriptome as a biomarker. However, current study designs do not easily accommodate gathering of longitudinal samples[21,46,50].

Here, we investigate the use of low cost, noninvasive biospecimens in transcriptomic studies as an approach to overcome these barriers. Early studies suggest buccal swabs, hair follicles, nasal swabs, saliva, and urine cell pellets may have potential use in clinical settings and for discovery. Observations thus far show that changing cell type proportion measures or estimates in buccal swabs, saliva, and urine cell pellets relate to current health of the individual[73–76]. For urine in particular, methods for isolating and propagating urine derived cells in the lab are of interest for testing patient and context specific gene regulation and treatment response, however, a consensus approach has not been reached[74,77–79]. At this time, a few groups have explored single cell RNA-sequencing of urine cell pellets and found enrichment of gene networks depending on disease status, though these cohorts were very small[75,76]. More recently, an increasing number of studies are leveraging nasal swabs to characterize the host transcriptome during viral infection and show promising results[80,81]. Previous investigations of hair found differences in expression between males and females using microarray data[82], and one study mapped cis-eQTLs in hair[83]. Despite this progress, further work investigating noninvasive tissue types is needed. No study to date evaluated sample quality and reliability across different library

preparations, and sources of biological and technical variance across donors and collections remains unknown. Thus far, no comparisons to invasive sample types have been made in terms of cell type composition and regulation of gene expression and splicing. All studies before focused on a specific disease application, leaving exploration of the broad use of noninvasive sample types for disease and transcriptomic applications untouched. Here, we address these limitations. We collected buccal swabs, hair follicles, saliva, and urine cell pellets from 19 individuals over 4 timepoints, and we prepared the samples for sequencing using both in-house and commercial library kits. We explore the unique biology of noninvasive sample types and delineate sources of technical variance, identify suitable invasive tissue type proxies, and demonstrate their use in transcriptomic and disease-relevant applications. We conclude hair follicles and urine cell pellets hold great promise for use in future studies.

1.2 Materials and Methods

Noninvasive sample collection

IRB approval was obtained for the study. 19 participants were recruited and consented, and four total collections of four tissue types were completed. The first collection occurred 6 months prior to confirm piloted procedures were ready to scale, and the remaining 3 collections were performed within a 2-4 week window per participant. 75 samples of each tissue type were obtained (1 participant provided only 3 collections). A detailed noninvasive sample collection protocol is provided on protocol.io (DOI: dx.doi.org/10.17504/protocols.io.kqdg3pjzzl25/v1).

Library preparation and sequencing

RNA extraction procedures are unique to each tissue type, and thus were performed separately for each tissue. Collections were randomized across extractions. All samples were prepared using our in-house method with 15 samples of each tissue type in duplicate. Takara Bio

SMART-seq v4 and Illumina stranded mRNA prep kits were prepared according to the manual and using a random subset of 12 and 16 samples of each tissue type, respectively. In total, 6 library preparation plates were prepared, and 2 HEK cell samples were included in triplicate on each library prep plate. This resulted in 508 total samples. 485 samples passed yield and size pre-sequencing quality parameters (>2nM yield and <600 bp average size). Samples were pooled by tissue type with HEK cell samples randomized across the library pools. Libraries were sequenced 2x150 bp on a Novaseq 6000 S4 flow cell. A detailed sample preparation protocol for RNA-extraction and our in-house method is provided on protocol.io (DOI: dx.doi.org/10.17504/protocols.io.kqdg3pjzzl25/v1).

Alignment, quantification of gene counts, and quality assessment

Adapter sequences were removed using Trimmomatic 0.36[84]. Sequences were aligned to build 38 of the human genome with Gencode v35 annotation using STAR 2.7.3a[85] and Samtools 1.9[86] set to GTEx mapping parameters[25]. Marking of duplicate reads was done using Picard 2.23.7[87], and gene counts were quantified using featureCounts from Subread 1.6.5[88]. QC results output from FastQC 0.11.3[89], STAR, and RNA-SeQC 2.3.6[90] were consolidated using MultiQC 1.8[91]. QC filtering based on standard sequencing quality metrics or based on protein coding and lncRNA depth thresholds were found to be largely redundant (Supp. Fig. 2), and thus depth of mapped reads was used for its simplicity. Genes were determined to be expressed in a tissue and included in downstream analyses if raw counts >= 8 and TPMs >= 0.1 in >= 20% of samples within a given tissue type.

Unmapped reads analysis

Unmapped reads were output to fastq files during alignment. These reads were remapped using DecontaMiner 1.4[92]. 23,488 bacterial, 21 fungal, and 11,120 viral genome references were

downloaded as suggested in the Installation and User Guide. Default parameters were used to remove low quality, human ribosomal and mitochondrial reads. For BLASTn alignments to the reference databases, bacterial and fungi parameters included minimum length = 50bp and gaps and mismatches = 2bp. Gaps and mismatches were increased to 5bp for viral genome remapping. Organisms were left unfiltered during initial remapping settings (Supp. Fig. 5a). For the analysis, we normalized the results by genomic length of the remapped species and by number of remapped reads per sample. We selected the top 0.5% of remapped species across all tissue types.

Technical and biological sources of unmapped reads were investigated by first removing all Decontaminer remapped reads from the unmapped fastqs, and then using FastQC to identify overrepresented sequences in the remaining reads. These sequences were compiled across all tissues and samples into a list of 707 sequences. Command line tools were used to filter and quantify the overrepresented read counts per sample. Computational and manual comparison to primer and adapter sequences as well as comparisons across preps and tissues were used to delineate the potential sources of the reads (Supp. Fig. 5d).

Downsampling

For analyses involving downsampling, binomial sampling was performed 5 times on the raw, unfiltered counts matrix and then the average of the sampling was taken. Binomial probability of success was set to the (desired depth)/(original depth), number of observations set to the total gene number, and trials set to the gene counts per a given gene. Samples were QCed for protein coding and lncRNA depth prior to downsampling. For comparisons across different library preparations, all samples passing a threshold depth of 1 million reads mapped to the human genome were included and then subsequently downsampled to 1 million (Supp. Fig.

2a,b,c). Loseq analyses were thresholded and downsampled to 2.5 million protein coding depth (Supp. Fig. 2d,e,f), with the exception of the cell type deconvolution analysis which was thresholded and downsampled to 5 million. All GTEx comparisons were made with both Loseq and GTEx thresholded and downsampled to 5 million.

Cross-preparation comparison

Median TPMs per gene were calculated within a given tissue, prep, and replicate group, and the Pearson correlation across replicate groups 1 and 2 for a given tissue and prep was quantified (Supp. Fig. 3). Overlap of gene expression capture was evaluated by taking the median TPMs within a tissue and prep, filtering genes with zero median expression, and determining the gene overlap using ComplexUpset 1.3.1[93,94] in R 3.6.0 (Supp. Fig. 3). Principal component analysis was performed using DESeq2 1.26.0[9] VST normalized counts and by selecting for the top 500 most variable genes. Variance attributable to tissue and prep was done by performing linear regression per PC (PC ~ tissue + prep) followed by ANOVA with p-value correction based on the number of PCs tested (Supp. Fig. 4).

Loseq cross-sample variance assessment

Principle components analysis was done across tissues using DESeq2 1.26.0 VST normalized counts and by selecting for the top 1000 most variable genes. Correlation of technical variables with PCs was investigated via linear regression. VariancePartition 1.21.6[95] was used to identify sources of gene expression variance within and across tissue types. The cross tissue VariancePartition model included collection, extraction, donor id, and tissue as random variables and rRNA rate, mapping rate, duplicate rate, exonic rate, 3’ bias, RNA concentration, a260/280, cDNA size, and GC content as fixed variables. Within tissue models were the same except tissue type was dropped as a variable (Supp. Fig. 6).

Cell type deconvolution of noninvasive samples

Deconvolution was done using GEDIT 1.7[96] and the provided BlueCodeV1.0.tsv reference matrix. For the final analyses, cell types were collapsed into broader umbrella categories by adding the estimated proportions together (Supp. Fig. 7b). Only the top 25% most abundant cell types per tissue were considered when looking across collections, and only donors with all 4 collections passing QC in urine, hair, and buccal were included in the final plot. All donors are plotted in the supplement (Supp. Fig. 7a).

PCA projection of noninvasive samples onto GTEx

For Loseq, the sample with the highest protein coding and lncRNA depth passing the 2.5 million threshold was selected per donor and per tissue type (19 samples per hair and urine, 17 buccal, 5 saliva), and both Loseq and GTEx were downsampled to 5 million. 19 samples of each representative (see xCell section) GTEx tissue were randomly sampled. Counts were VST normalized using DESeq2 1.26.0. Principal components analysis was run on centered and scaled GTEx counts using the top 1000 most variable genes. The resulting PCA loadings were multiplied by GTEx and Loseq centered and scaled counts, and this data was plotted as shown in the main figure.

The splicing PCA was performed in the same manner except using an exon inclusion level matrix generated by rMATS 4.1.2[97] as input. Splicing events with zero inclusion for any sample in a tissue were excluded. The rMATS results were filtered for events found to be significantly different across the select GTEx tissues and with inclusion levels greater than 2 standard deviations beyond the average inclusion (0.3). Again, PCA was run for the top 1000 most variable events in GTEx, and the loadings were multiplied by the GTEx and Loseq centered and scaled inclusion levels.

Cell type enrichment analysis of noninvasive and GTEx samples

75 samples were sampled from each GTEx tissue to approximately match the number of Loseq samples included (75 hair, 63 urine, 25 buccal, 5 saliva). GTEx and Loseq TPMs were deconvolved for enrichment using 64 cell type signatures in xCell 1.1.0[98] in R 3.6.0. Select GTEx tissues were chosen using gene expression clustering of median TPMs per GTEx tissue. GTEx groups were established using k means, and the tissue with the highest sample size per group was selected as representative. Kidney cortex, esophageal mucosa and lung were added based on their proximity to noninvasive tissues we studied. Clustering of cell type enrichment was done by taking the median enrichment score per tissue and then using k means in ComplexHeatmap 2.2.0[99].

GTEx eQTL replication analysis

Participants provided their genotyping data from 23andme (9 donors), Ancestry (2 donors), and Gencove (8 donors) platforms. Array data was imputed using the 1000 genomes phase 3 reference[100] and the Sanger Imputation Service[101]. Eagle 2.4.1[102] and the 1000 genomes reference were used for VCF phasing. Monomorphic alleles, alleles with MAF < 0.05, multiallelic sites and indels were excluded from all analyses. This imputed, phased, and filtered VCF was used for eQTL and ASE analyses.

For the genotyping PCA, LD pruning was performed with PLINK 1.90-b3.29[103] with a window size of 50, shift of 5, and r squared cutoff of 0.2. Only SNPs with a 100% genotyping rate and HWE 1e-5 were included. Ancestry was imputed by merging the donor VCF with the 1000 genomes VCF, excluding any individuals with relatedness >= 0.0625, running smartpca with eigensoft 6.1.3[104], and using k-nearest neighbors to infer ancestry labels of the donors based on genetic similarity. Running smartpca on the donor samples alone revealed genotyping PC1

corresponded with ancestry, while PC2 corresponded with genotyping panel. These were included as covariates in the eQTL analysis, as well as expression PCs with variance explained > 15% that accounted for global changes in gene expression (PC1 for hair and buccal and PCs 1 and 2 for urine). Counts were TMM and inverse normalized and filtered for genes with raw counts >= 6 and TPMs >= 0.1 (as is the GTEx standard). eQTL mapping was done on a per tissue basis using TensorQTL v.1.0.5 with the window set to 1MB (following the GTEx parameters).

To calculate replication, each GTEx tissue was filtered for the top eVariant-eGene pair per gene with MAF >= 0.05, qvalue <=0.05, and with effect size greater than the minimum observed in the lowest powered GTEx tissue (kidney cortex = 0.32). This was intersected with the pairs discovered in the noninvasive dataset, and pi1 was calculated using qvalue 2.18.0[39] and R 3.6.0. Null datasets were of the same size as the overlapping significant GTEx pairs set and were generated by sampling allele-frequency matched eVariant-eGene pairs from the noninvasive data 1000 times. Pi1 was calculated per each dataset. Significance for enrichment was determined based on permutation p-value calculations (Supp. Fig. 9).

Differential expression analysis and FGSEA

Sex-based differential expression analysis was performed on a per tissue basis using edgeR 3.28.1[105] and Limma-Voom 3.42.2[106]. Counts were filtered based on GTEx parameters and TMM normalized prior to analysis. GTEx sex-based differential expression results[33] were retrieved from the GTEx Portal (https://gtexportal.org/home/datasets) for overlap comparisons (Supp. Fig. 10b). For noninvasive tissues, a named list of sex-based differentially expressed genes and their t-scores was input into FGSEA 1.12.0[107]. The Hallmark gene sets file was

obtained from Molecular Signatures Database (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#H)[108] and used for the analysis.

Loss of function detection using ASE

Allele-specific expression was calculated using ASEReadCounter 4.0.1.1[109] and using the imputed, phased, and filtered VCF described in the GTEx replication analyses. Sites with fewer than 16 total counts were filtered from the analysis. The reference counts divided by total counts was assessed across tissues and donors, and one donor was removed due to extreme ratios and thus potential genotyping errors (Supp. Fig. 11a). Ensembl Variant Effect Predictor 5.28.1[110] was used to annotate variant consequences, and the ratio of reference to total allele counts was compared given these annotations (Supp. Fig. 11b).

OMIM Mendelian disease gene overlap

Genes with Mendelian inheritance were downloaded from the OMIM database[111] (https://www.omim.org/). Tissues were filtered were genes meeting minimum GTEx expression thresholds, and then the remaining gene set was overlapped with OMIM genes (Supp. Fig. 13). ComplexUpset 1.3.1 was used to compare genes captured across noninvasive and select GTEx tissues.

OpenTargets evaluation of disease-relevancy

Data was retrieved from the OpenTargets database[112]. Disease ids were selected by choosing the broadest ontological category specific to a given disease (as provided on the OpenTargets platform). For the analysis, the association file incorporating all sources of evidence was used, and disease genes were included only if there were 5 or more sources of evidence (Supp. Fig. 12a). Loseq samples were thresholded and downsampled to a protein coding and lncRNA depth of 5 million, and samples with the highest depth per donor and tissue

were selected. GTEx was similarly downsampled and 19 samples of each GTEx tissue were randomly selected. GTEx tissues were chosen based on their relevance to selected diseases. Genes were filtered based on GTEx parameters, and the median TPMs per tissue was calculated. 10,986 tissue-elevated genes were obtained from the Human Protein Atlas[113] (https://www.proteinatlas.org/humanproteome/tissue/tissue+specific), and each tissue was additionally filtered for tissue-elevated genes. In the end, the top 3,411 most expressed, tissue-elevated genes present in a tissue were analyzed, based on the tissue with the lowest number of genes passing post-expression and HPA filtering (Supp. Fig. 12b).

To calculate the summed evidence score, first, the original target overlap was calculated by intersecting the genes present across all tissues with the disease gene targets. Summing the evidence score for these genes resulted in the total possible score for a given disease. Then, the genes expressed in a particular tissue were overlapped with disease gene targets. The evidence scores for tissue-specific overlapping genes were summed together and then divided by the total possible score for a disease. This normalized score is the summed evidence score reported in the analysis (Supp. Fig. 12c).

1.3 Results

1.3.1 Noninvasive tissues are amenable to low-input library preparations

Buccal swabs, hair follicles, saliva, and urine cell pellets were collected from 19 healthy individuals at four separate time points (Figure 1a). Briefly, participants deposited a saliva sample into an Oragene saliva collection kit, 10 hair follicles were plucked, and participants provided a buccal swab sample (see Methods and DOI: dx.doi.org/10.17504/protocols.io.kqdg3pjzzl25/v1). Saliva samples were stored according to kit instructions, and hair follicles and buccal swabs were flash frozen at -80C. Saliva, hair follicles,

and buccal swabs were all able to be collected and stored within 9 minutes, on average. Urine samples were obtained at any time throughout the day, and underwent serial centrifugation before flash freezing the cell pellet at -80C. Individuals enrolled in the study provided genotyping data from 23andMe or Ancestry SNP arrays, or from low-pass whole genome sequencing provided by Gencove.

Total RNA yielded from noninvasive tissues was typically less than 1ug and thus lower than traditional bulk tissue samples for RNA-sequencing. Buccal swabs and saliva resulted in consistent and sample-specific yields, while hair follicles and urine displayed donor-specific variability, with urine exhibiting a greater degree of variation (Supp. Fig. 1e). This observation agrees with clinically known interindividual differences in cell numbers found in urine[74], while hair follicle output appeared to be due to individual differences in hair texture. All samples were prepared for sequencing using a low-cost in-house library preparation we developed specifically for low-input bulk RNA applications, which uses template-switching oligo (TSO) and tagmentation chemistry and reduces cost by 83% and 68% per reaction compared to the Illumina TruSeq Stranded mRNA Library Prep and SMART-seq V4 kits, respectively (Supp. Table 1, Table 2). We herein refer to this method as Loseq. To validate the consistency of our in-house method, we included 15 technical replicates per tissue in each Loseq library preparation batch. To compare performance across library preparation methods, 12 randomly selected samples of each tissue were prepared using the TakaraBio Smartseq V4 kit, a commercially available kit with similar chemistry to Loseq, and 16 samples were prepared using the Illumina TruSeq Stranded mRNA kit, one of the most frequently used kits in the transcriptomics field (Supp. Fig. 1a). For all preparation methods, 2 HEK293 cell samples were included in triplicate to serve as a quality control standard. Samples passing pre-sequencing quality criteria of > 2nM concentration

and average library size < 600 bp were sequenced on the Illumina NovaSeq S4 platform with a mean depth of 25.6 million total reads per sample (Supp. Fig. 1f, 1g).

Post-sequencing, we evaluated multiple quality control metrics returned from RNA-SeQC[90]. Ultimately, we found protein-coding and lncRNA read depth corresponded with high quality samples and consistent gene expression capture (Supp. Fig. 2). Across all tissues, Illumina prepared samples yielded a lower number of reads, and we thus used a less stringent 1 million protein-coding and lncRNA depth threshold for cross-preparation comparisons. A 2.5 million threshold and a 5 million threshold were used for the Loseq-only and GTEx comparison analyses, respectively.

Looking across tissues and library preparations, we observe most samples pass pre- and post-sequencing quality checks when using low-input methods (Figure 1b). Notably, hair performs well regardless of kit used and meets pre- and post-sequencing standards typical of traditionally used bulk RNA-sequencing samples (Figure 1c, Supp. Fig. 1, 2). Many urine and buccal samples fail pre-sequencing checks for the Illumina kit. Given urine shows excellent performance via traditional RNA-sequencing quality metrics when low-input protocols are used, we suspect this is driven by low and highly variable RNA yield not amenable to traditional bulk kits. Buccal and saliva display higher rates of RNA degradation, as reflected by lower RIN and computationally derived transcript integrity, as well as diminished performance apparent in other RNA-seq QC metrics (Supp. Fig. 1b, 2). The lower quality of these tissue types is likely resulting in the higher rate of failure across the different preparations. We found saliva to be a particularly poor biospecimen for transcriptomic study with few samples passing our thresholds. Looking across saliva collections we observe some donors more consistently pass or fail quality checkpoints, suggesting donor-specific variables may play a larger role in determining the

sample quality relative to other tissue types. Buccal, though not an ideal sample type, performs well enough for use in targeted applications.

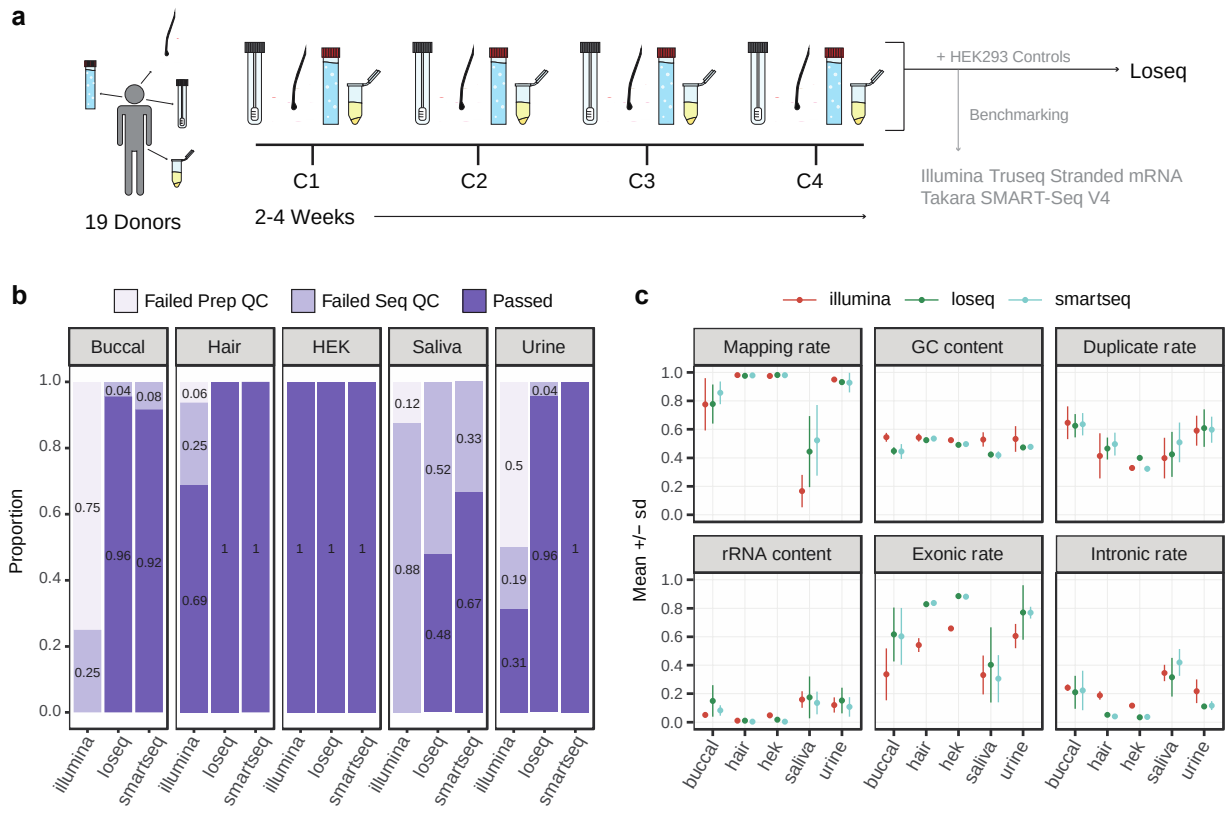


Figure 1. Noninvasive sample study design and processing outcomes. a. Four collections (C1-C4) of four noninvasive tissues were collected from 19 donors over the course of 2-4 weeks per donor. All samples were processed using our in-house method, Loseq, while a subset was prepared using commercially available kits. Two biological replicates of HEK293 cell controls were included in triplicate for all library preparations. **b.** Proportion of samples passing per tissue type and preparation. Failed Prep QC = exceeded 600bp average size or less than 2nM yield. Failed Seq QC = protein-coding and lncRNA depth less than 1 million. **c.** RNA-seq quality metrics for all tissues and library preparations.

To compare gene expression patterns across libraries in an unbiased manner, we first downsampled all QC-passed samples to a depth of 1 million. We found the majority of genes expressed in a tissue were captured regardless of preparation method and there was high agreement in gene expression levels across library preparations (Supp. Fig. 3a, 3d). One difference of note is Loseq tends to capture longer genes with lower GC content relative to the commercial kits (Supp. Fig. 3b, 3c). Principal component analysis (PCA) shows that the

preparation method contributed minimally to variance observed across the samples (Supp. Fig. 4).

1.3.2 Unmapped reads capture tissue-specific microbial signatures

Since we observed a low mapping rate for buccal and saliva and because the noninvasive samples were collected from non-sterile human tissues, we decided to investigate biological and technical sources of unmapped reads in our samples (Supp. Fig. 5a). First, we remapped unmapped reads to microbial genomes using Decontaminer[92]. This process remapped only a small fraction of the unmapped reads (Figure 2a, Supp. Fig. 5b), a somewhat unsurprising result given our library preparation is targeted to capture poly-A mRNA transcripts and most microbial transcripts are not polyadenylated. Nonetheless, when we look at the top 0.5% most abundant remapped species we observe distinct microbial signatures across the noninvasive tissues that support previously known microbiota of the oral cavity, human skin, and genitourinary tract (Figure 2b)[114–117]. In addition, we see high correlation in estimated species abundances across technical replicates suggesting that, despite the limitations in our library preparation approach, we are capturing real, replicable biological signal (Supp. Fig. 5c). Altogether, these findings support noninvasive samples may bear biological utility in follow-up microbiome studies using microbiome-specific or total RNA library preparations. Further investigation of the remaining unmapped reads showed those reads to be largely of technical origin (Supp. Fig. 5d). Overall we were able to account for 58.9%, 56.8%, 17.2%, and 58% of unmapped reads across all buccal, hair, saliva, and urine samples, respectively.

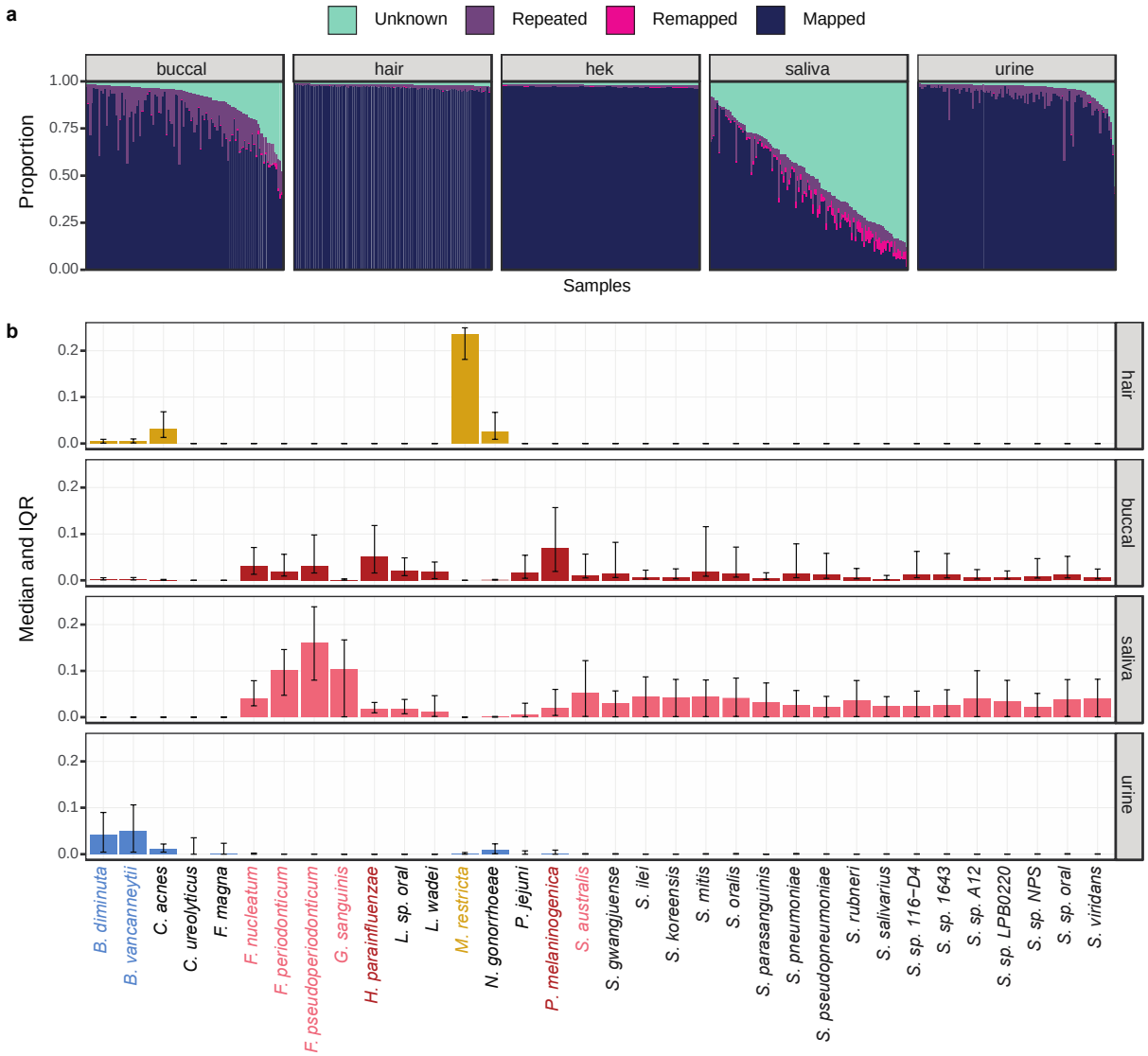


Figure 2. Classification of unmapped reads in noninvasive samples. a. Characterization of proportion of reads per sample. Mapped = aligned to hg38. Remapped = aligned to microbial species using Decontaminer. Repeated = highly abundant reads identified by FastQC. Unknown = reads not mapped or highly abundant. **b.** Normalized proportion of reads remapping per species for each tissue. The top 0.5% most abundant microbes are shown. Highlighted species have a median abundance > 0.05 for that tissue.

1.3.3 Sources of gene expression variance between individuals and noninvasive tissue types

Next, we investigated gene expression variance due to noninvasive tissue type or individual. Saliva samples largely failed to pass post-sequencing quality standards and were excluded from several analyses for this reason. Using a mixed linear model to delineate

biological and technical contributors to gene expression variability, we found tissue type is the main driver of variance across samples (Figure 3a). These results were supported by PCA, which segregated samples primarily by tissue type, with hair forming a separate, distinct cluster from the other tissues (Figure 3b).

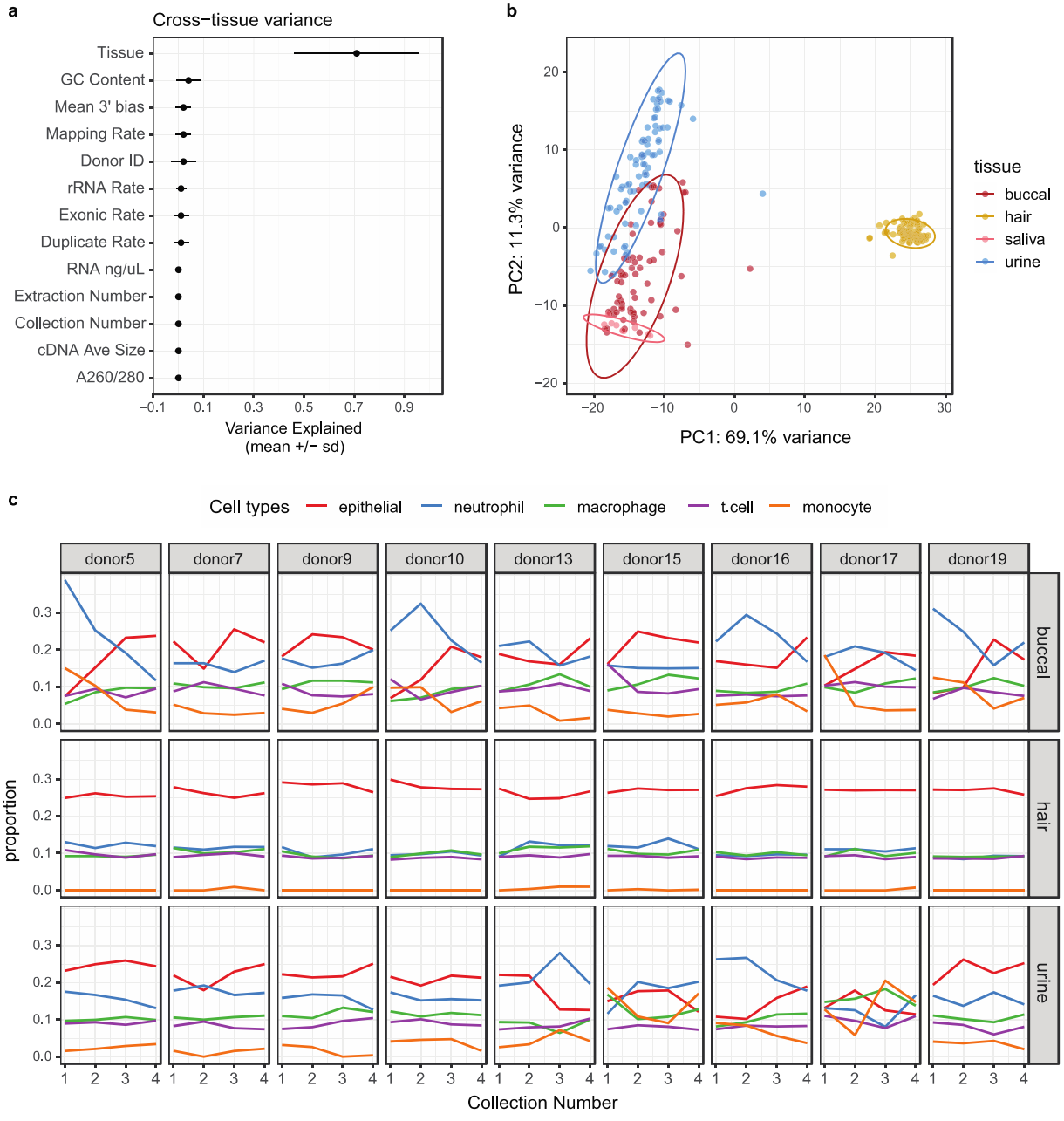


Figure 3. Technical and biological sources of variance in noninvasive samples. a. Factors contributing to variance in gene expression across tissues as determined by mixed linear modeling. **b.** Principal component analysis using DESeq2 normalized counts and the top 1000 most

variable genes. **c.** GEDIT cell type proportion estimates across collections per donor. Only donors with samples passing QC for all collections are displayed here. Niche cell types were collapsed into larger categories, and the top 25% most abundant cell type categories across tissues are shown.

Because underlying cell type composition frequently explains gene expression variance across samples and tissues[16,17], we deconvolved our noninvasive samples using GEDIT[96] and the provided BlueCodeV2 single cell reference. From this we observed hair is primarily composed of epithelial cell types, buccal and urine capture both epithelial and immune cells, and saliva mostly contains neutrophils and monocytes (Supp. Fig. 7b). Looking across tissues, donors, and collections, hair was highly consistent in cell type abundance estimates, both within and across donors (Figure 3c, Supp. Fig. 7a). On the other hand, urine and buccal samples were more variable. In these samples, cell type composition was occasionally consistent within and across donors, but there were also patterns where cell type composition changed across collections for the same donor (e.g. donor 5, buccal), showed a different pattern of abundance compared to other donors (e.g. donor 16, urine), or completely lacked any consistency (e.g. donor 17 urine). When we used linear mixed modeling to identify biological and technical sources of gene expression variance within tissues, the individual donor was the primary contributor in buccal and urine (Supp. Fig. 6). Both of these results indicate that cell type compositions sampled from urine and buccal samples are potentially highly variable and donor specific. For hair, the relative contribution of technical factors and the donor of origin to gene expression variance is similar. As previously described, hair follicle data quality matches gold-standard RNA-sequencing data, which together with the low variance in cell type composition leads to highly consistent gene expression profiles.

1.3.4 Noninvasive tissue characteristics suggest potential invasive tissue type proxies

To investigate the biological similarity of noninvasive tissues to known, invasive sample types, we compared gene expression, splicing, and cell type enrichment patterns to GTEx. To do so, we first selected a single noninvasive sample per donor and per tissue with the highest protein-coding depth. Representative GTEx tissues were chosen based on k means clustering, and 19 samples of each tissue type were randomly selected. Both the noninvasive and GTEx samples were downsampled to 5 million read counts to normalize for differences in total sequencing depth.

Using this data we projected the noninvasive samples onto the GTEx PCA space to observe global patterns of gene expression similarity (Figure 4a). Hair clusters closely with esophageal mucosa and skin, saliva is proximal to spleen, blood, and EBVs, and buccal and urine are intermediaries between these groups. We repeated this analysis using splicing events generated from rMATS[97] (Figure 4b). From this we recapitulate similar clustering patterns observed for expression.

Since these clusters may reflect similarities in underlying cell types, we investigated this question by using xCell[98] to calculate cell type enrichment scores. Here we used xCell because it is the most comprehensive cell type database available, thus enabling analysis of diverse tissues and biospecimens, and we observed high concordance in cell type estimates between GEDIT and xCell for cell types present in both references (Supp. Fig. 8). From this analysis we replicated the same tissue clustering we observed by PCA except using cell type enrichment estimates (Figure 4c). Because cell type sharing is highly predictive of shared gene regulatory mechanisms[26–29], this suggests gene expression regulatory mechanisms present in invasive tissues may be captured noninvasively.

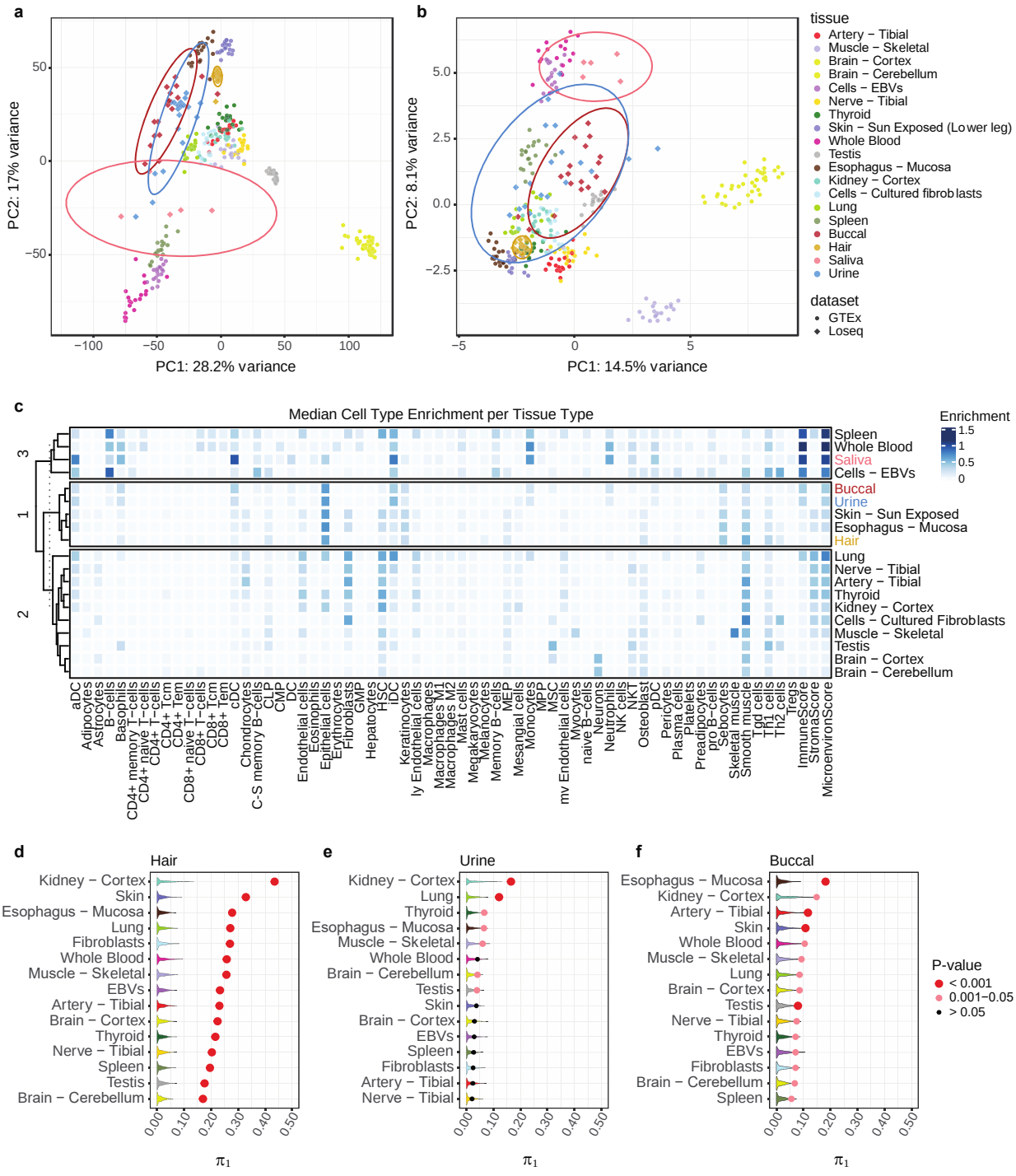


Figure 4. Comparison of noninvasive samples to the GTEx dataset. a. Noninvasive sample types projected onto the GTEx expression PCA space. Counts were normalized using DESeq2, centered and scaled, and the top 1000 most variable genes were used. Ellipses represent 95% confidence intervals. **b.** Noninvasive sample types projected onto the top 1000 most variable rMATS splicing events in GTEx. **c.** xCell cell type enrichment estimates per tissue. Tissues are clustered using k-means clustering. **d. e. f.** GTEx eQTL replication estimates for hair, urine, and

buccal samples. Dots show π_1 calculated by selecting significant GTEx gene-variant pairs from the noninvasive data with sizing indicating permutation p-value significance. Violin plots show null π_1 distributions generated from allele-frequency matched, randomly selected gene-variant pairs. 1000 permutations were performed.

To further explore how noninvasive samples capture genetic regulatory variants discovered in postmortem tissues, we assessed replication of GTEx eQTLs in buccal, hair, and urine samples. With our sample size being insufficient for full eQTL discovery, we looked for enrichment of low p-values in our noninvasive dataset for the eVariant-eGene pairs previously discovered in GTEx. A null distribution was generated by randomly sampling allele-frequency matched eVariant-eGene pairs from our noninvasive data, and we calculated π_1, an estimate of the true positive rate, in both the null datasets and when selecting the significant GTEx eVariant-eGene pairs. In hair, we find significant replication of GTEx pairs across all studied GTEx tissues, with kidney cortex and skin showing the highest degree of replication ($\pi_1 = 0.44$ and $\pi_1 = 0.33$, respectively, Figure 4d). Buccal and urine are less homogenous tissue types, thus further decreasing our power especially as our sample size does not allow highly efficient approaches to correct for latent variation[118]. As such, they showed less clear signal across all tissues (Supp. Fig. 9). However, we did observe most significant enrichment for kidney cortex eVariant-eGene pairs in urine and esophageal mucosa signal enrichment in buccal ($\pi_1 = 0.17$ and $\pi_1 = 0.18$, respectively, Figures 4e and 4f). Of note, kidney cortex has a low sample size relative to other tissues in GTEx and thus little power for discovery of more subtle eQTL effects. Thus, it is unclear whether the high replication of kidney eQTL signal across noninvasive tissues is due to similar biology or an abundance of common and/or high effect size eQTLs in this tissue. In all, our results suggest noninvasive tissues capture cell types and gene expression regulatory mechanisms present in invasive tissue types and may provide insight into disease processes affecting these tissues.

1.3.5 Sex-specific differences in gene expression in noninvasive samples

Because biological and environmental contexts play a major role in expression regulation, we aimed to explore whether RNA-sequencing from noninvasive tissues may be used for this purpose. To this end, we tested for sex-based differential expression and the replication and biological role of the discoveries. Using edgeR[105] and limma-voom[106], we were able to identify 25 and 1032 sex-based differentially expressed genes in hair and urine, respectively (Supp. Fig. 10a). In comparing hair to sun-exposed skin, 8 of the 25 significant hits were previously seen in GTEx and are highlighted in Figure 5a. Looking across all GTEx tissues, 17 of the 25 genes were previously observed (Supp. Fig. 10b). Running FGSEA[107] on the hair results showed significant enrichment for E2F targets and G2M checkpoint pathways in females (Figure 5b), and these are central to regulating the cell cycle and proliferation[119,120]. Though nonsignificant, Wnt signaling, the top hit for males, has previously been reported to play a key role in hair loss prevention[121]. In urine, 33 of the 1032 significant findings were seen in kidney cortex, and, overall, 582 were differentially expressed in any GTEx tissue (Figure 5c, Supp. Fig. 10b). Notably, estrogen response is greatly enriched in females (Figure 5d). Estrogen signaling is central to many physiological processes in the kidney and is considered potentially protective against many renal diseases, though much remains unknown[122]. This analysis demonstrates the potential for noninvasive samples to elucidate underlying biology in a variety of potential contexts and assays.

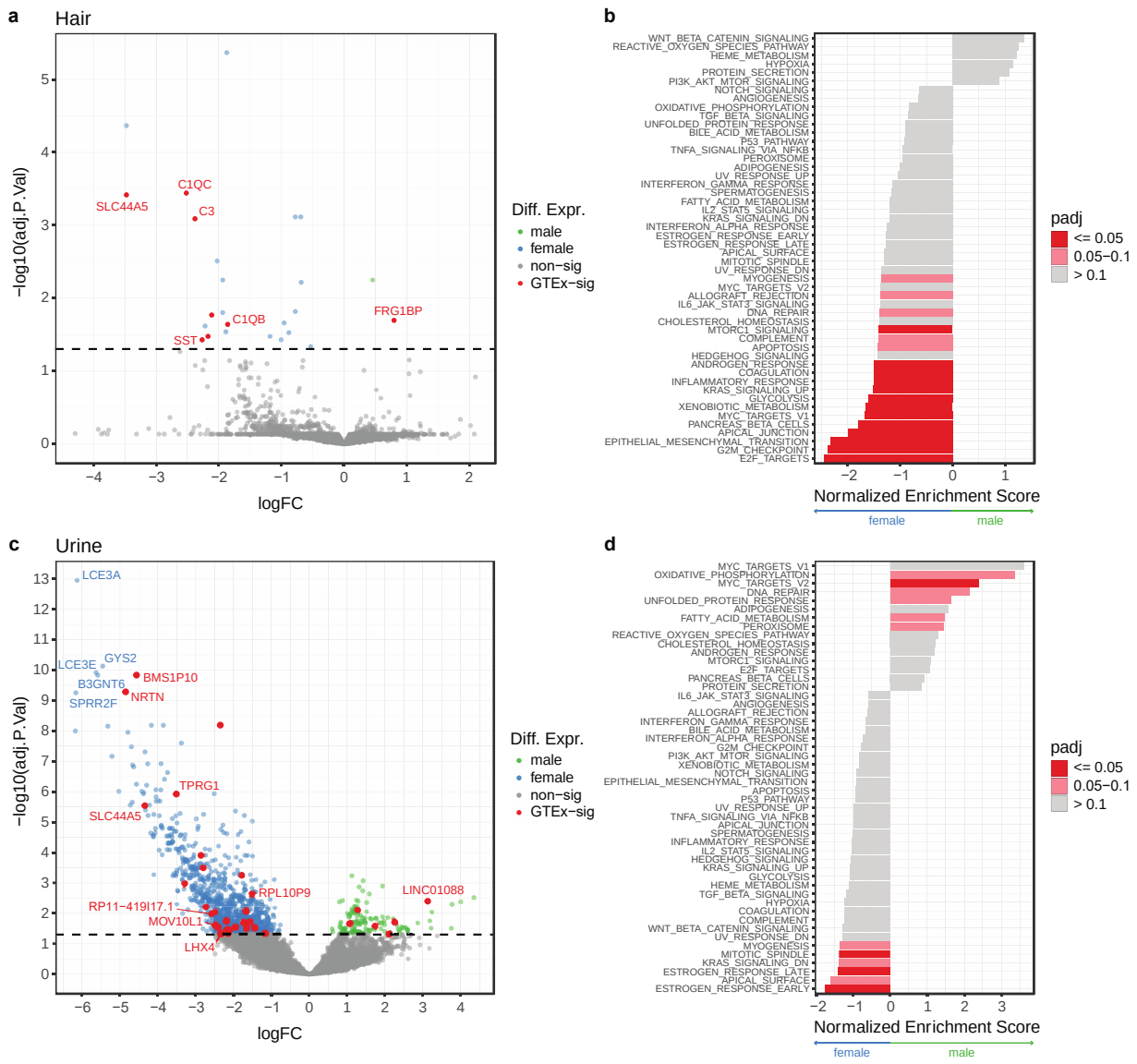


Figure 5. Sex-based expression differences in noninvasive samples. a. Volcano plot of sex-based differentially expressed genes in hair. Genes highlighted in red are replicated sun-exposed skin findings in GTEx. Dotted line indicates significance threshold. **b.** FGSEA of all genes ranked by z-score and using the Hallmark Gene set from MSigDB. **c.** Sex-based differentially expressed genes in urine cell pellets. Genes highlighted in red are replicated kidney cortex findings in GTEx. **d.** FGSEA of all genes ranked by z-score and using the Hallmark Gene set from MSigDB.

1.3.6 Noninvasive samples may be leveraged for disease-relevant applications

Allele specific expression (ASE) analysis compares allelic expression levels within the same individual, and it is an important tool for investigating rare and cis-regulatory variation, nonsense mediated decay, and genomic imprinting[109]. Here we quantified ASE using

heterozygous sites across tissues. From this, we observe anticipated reference allele ratio patterns depending on the SNP annotation (Supp. Fig. 11b), and we show robust nonsense mediated decay for stop-gain variants across all collections for buccal, hair, and urine samples (Figure 6a). This suggests noninvasive sampling may be used to identify gene-disrupting variants that are a common focus of genetic diagnosis in rare disease.

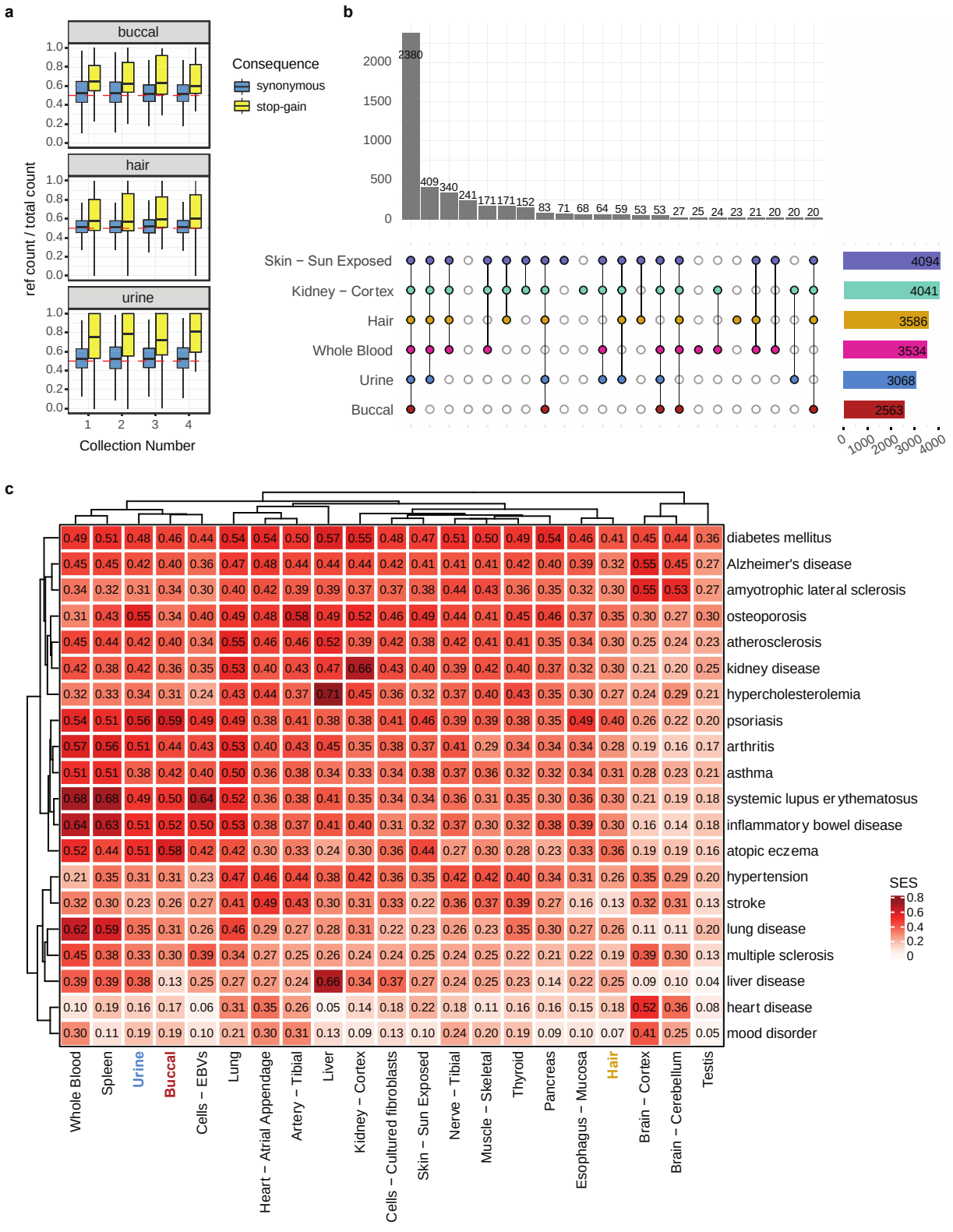


Figure 6. Use of noninvasive samples in disease-relevant analyses. a. ASE for annotated stop-gain variants vs synonymous. Only sites with > 16 total counts were included. **b.** Genes with median expression > 0.1 TPM in a tissue were intersected with the OMIM gene set. Shown is the

intersection of OMIM genes captured across tissues. **c.** Capture of common disease signals in the OpenTargets database. SES = Σ(evidence scores of disease genes expressed in a tissue)/ Σ(evidence scores of disease genes expressed in any included tissue).

Application of noninvasive samples to rare and common disease was evaluated using the OMIM[111] and OpenTargets[112] repositories. For Mendelian disease, our samples captured a median of 80% (hair), 70% (urine), and 55% (buccal) of genes above a 0.1 TPM threshold (Supp. Fig. 13a). This capture is consistently high for hair samples across collections but shows donor-dependent consistency for urine (Supp. Fig. 13b). Clustering samples by median OMIM gene expression with GTEx recapitulated our prior proxy observations (Supp. Fig. 13c). In Figure 6b, we show the overlapping gene sets for Mendelian genes with a median expression greater than 0.1 TPMs in a given tissue after employing GTEx expression thresholds within that tissue. The selected GTEx tissues shown are relatively minimally invasive or identified as most similar by clustering and eQTL replication. We see the vast majority of genes are captured in both noninvasive and invasive tissue types, indicating noninvasive samples may be a suitable biospecimen for studying gene expression and regulatory processes in many rare disease applications. Notably, we observe a subset of genes expressed in noninvasive samples that are not captured in whole blood. This suggests diseases where the primary tissue type affected is more akin to noninvasive samples may benefit more from the use of noninvasive sampling versus whole blood collection. In all, efforts to improve clinical genomics studies using transcriptomics may be further augmented by use of noninvasive samples, especially where invasive surgical sampling of tissues primarily affected is often not possible.

Looking at common disease, we first selected the most general OpenTargets ontology category for every disease included. We filtered for genes with greater than 5 sources of evidence and with tissue elevated specificity from the Human Protein Atlas database[113] (Supp.

Fig. 12). Disease enrichment was calculated by summing together OpenTargets gene evidence scores (summed evidence score, SES) for genes with median expression greater than zero in a given tissue and dividing by the total possible summed evidence score. From this, we found buccal and urine captured diseases with a strong immunological component, much like whole blood and spleen (Figure 6c). Urine additionally showed strong signal for kidney disease (SES = 0.42) and an array of other diseases. Hair performed best for skin-related diseases (psoriasis SES = 0.40), but overall did not show strong enrichment for any particular disease. These results suggest noninvasive samples bear promise for use in disease-relevant studies while providing the advantage of study designs with potentially longitudinal monitoring and greater enrollment.

1.4 Discussion

Discovery from transcriptomic data and its use in precision medicine is considerably limited by cost and access to biologically applicable biospecimens[25,49]. As a result, most transcriptomic studies have lagged behind GWAS in sample size, which now often include hundreds of thousands of individuals. Further, disentangling causation and finding context-specific disease mechanisms is challenging using a single collection time point[21,46]. To address these limitations, we sought to investigate low-cost, noninvasive RNA-sequencing as an alternative approach. From our study we observed hair follicles and urine cell pellets provide the highest quality data and perform best in functional genomics applications.

A primary advantage of noninvasive biospecimens over blood-related specimens to the transcriptomics field is the set of cell types captured. Shared cell type composition corresponds with shared regulation of gene expression and splicing[16–19], and a major limitation of blood-related samples is that they represent a highly tissue-specific set of cell types. In our samples we observed greater similarity by expression, splicing, and genetic regulation to invasive GTEx

tissues relative to GTEx whole blood. Cell type deconvolution analysis suggested that our tissues contain epithelial cells, myocytes, stromal cells, and others, all of which are unable to be captured using blood and play a key role in mechanisms of many diseases.

Several considerations should be taken into account when deciding to use noninvasive tissues. Generally, ease of sample collection and consistency of library preparation performance and quality is biospecimen-dependent. Additionally, the feasibility of clinical use and longitudinal study design varies depending on the tissue type. We observed high failure rates using buccal swabs and saliva, and though the data yielded from samples passing quality standards provides valuable insights and the samples themselves are simplest to collect, we believe use of these biospecimens should be reserved for specialized applications where health status of the oral cavity and/or upper gastrointestinal tract is primarily under study.

Though we found urine cell pellets to yield high quality RNA, the pellet itself may be minimal and difficult to visualize for some donors. This is especially true for healthy donors, who tend to shed fewer cells into their urine[74], and could introduce bias into future study designs if special care is not taken when using this tissue. Similarly, we found urine cell pellet RNA quantity to be variable and donor dependent. Others have aforementioned single cell approaches for urine specimens[75,76], and we anticipate further development of these methods will better control for sampling inconsistency. Here, we showed urine cell pellets capture genetic regulatory mechanisms seen in the kidney as well as gene expression signatures relevant to kidney disease and diseases mediated via kidney functions. Given the enormous health burden kidney disease poses in the US and worldwide, and the central role the kidney plays in many diseases[123], methods for noninvasively monitoring kidney function and enabling early diagnosis could meaningfully improve morbidity and mortality. In addition to the analyses performed here,

others have proposed laboratory protocols for propagating cells collected from urine for use in identifying disease mechanisms and novel treatments, or, in the case of stem cells, developing autologous cell therapies[77–79]. Overall, given further optimization, we expect urine holds the greatest potential for clinical use and discovery.

Hair follicles perform robustly using any library preparation and exhibit low technical variance across collections and donors. It should be noted that hair follicle collection does require additional training of personnel not necessarily needed for the other biospecimens. Also, fine versus coarse hair type played a role in determining the ease of collection, and we do observe slight differences in yield depending on the donor, though this did not impact sample performance. We do foresee the need to explore collection of hair follicles from other parts of the body when head hair is not available, and it is likely necessary in future studies to collect additional information regarding the use of cosmetics and medications applied to the head and skin. Here, we showed hair follicles result in consistent quality, cell types, and expression profiles across collections, and, despite our low sample size, we found significant replication of previously observed eQTLs across all GTEx tissues. Together, these findings suggest hair is a highly robust biospecimen with potentially broad applications, and, because of its consistency, biological perturbations due to disease, treatment, or other environmental exposures will likely be observable in clinical and longitudinal settings.

Notably, across all noninvasive tissues we observed a large majority of Mendelian disease genes were expressed. The ease and decreased invasiveness of noninvasive proxy biospecimens could facilitate greater use of transcriptome analyses in diagnosing rare, genetic disease, however, further work is needed to explore this possibility.

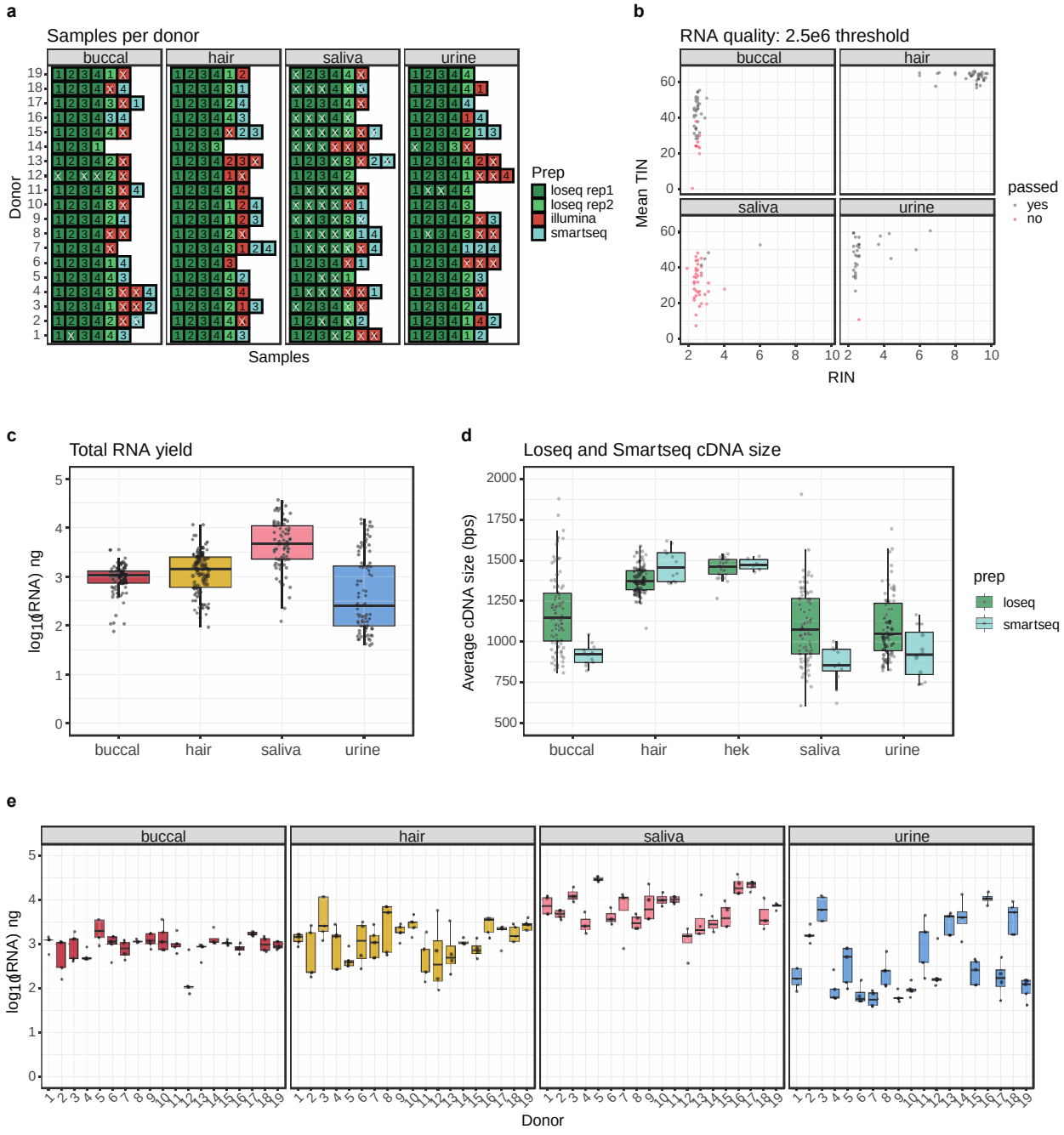
Our findings will need to be validated using larger, more diverse, and clinical cohorts. We expect noninvasive tissues may reduce Eurocentric sampling bias and enable sampling from more vulnerable populations, but this expectation will need to be measured against future study enrollment. Additionally, we used bulk RNA-sequencing, and the performance and features of noninvasive samples using single cell methods will need to be optimized and evaluated.

This study aimed to establish whether noninvasive sampling may be used to scale transcriptomic studies. From our work, we were able to characterize many of the technical and biological features of four possible noninvasive samples, and we showed their potential utility in both transcriptomic and disease-related applications. Overall, we find hair follicles and urine cell pellets to be the most promising biospecimens, and we propose advantages in terms of cost and study designs for pursuing noninvasive sampling. In all, noninvasive RNA-sequencing offers meaningful improvements to current transcriptomic approaches that could enable dramatic scaling in sample size and increased discovery potential. This scaling would bring closer parity with GWAS via transformational increases in power, thus better positioning transcriptomic studies for use in diagnostic and clinical applications.

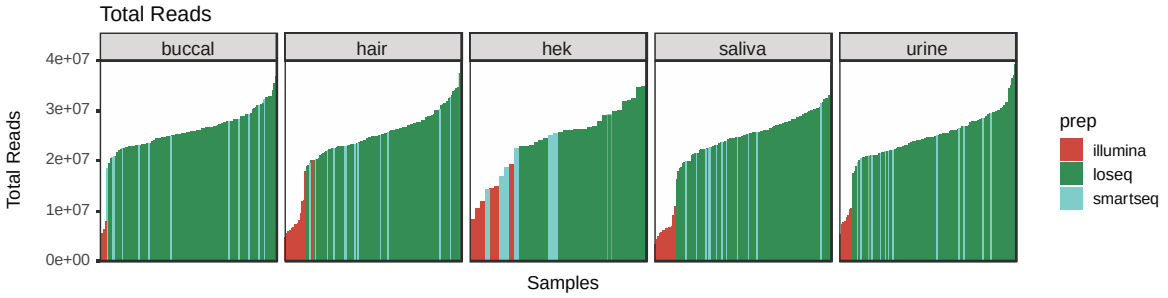
1.5 Supplementary Figures and Tables

Supplementary Figure 1. a. Outcome of library preparation QC per donor, collection, and preparation. The crosses indicate a failed sample, and the numbers correspond to the collection. **b.** Measured RIN vs computationally-derived transcript integrity number (TIN) per sample. Passing is determined by 1 million protein-coding depth threshold. **c.** RNA yield distribution per tissue type. **d.** cDNA average size for Loseq and SmartSeq preparations across noninvasive tissues. **e.** RNA yield per donor and tissue. Each data point is a collection. **f.** Total reads

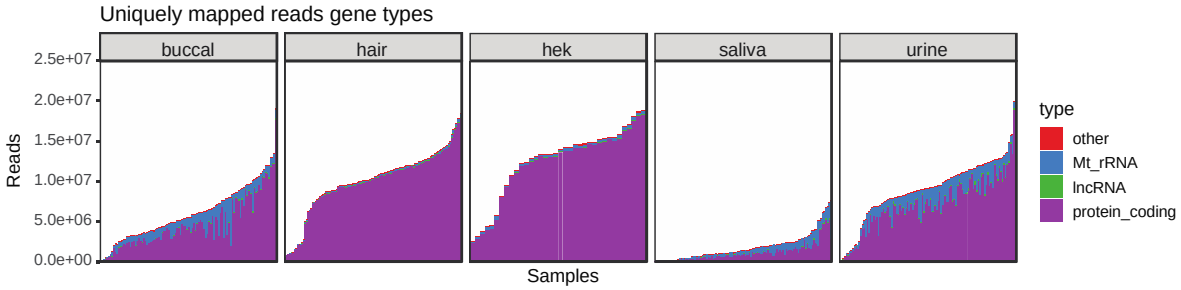
sequenced per sample, colored by prep. **g.** Categorization of gene types for uniquely mapped reads mapping to genes across tissues.



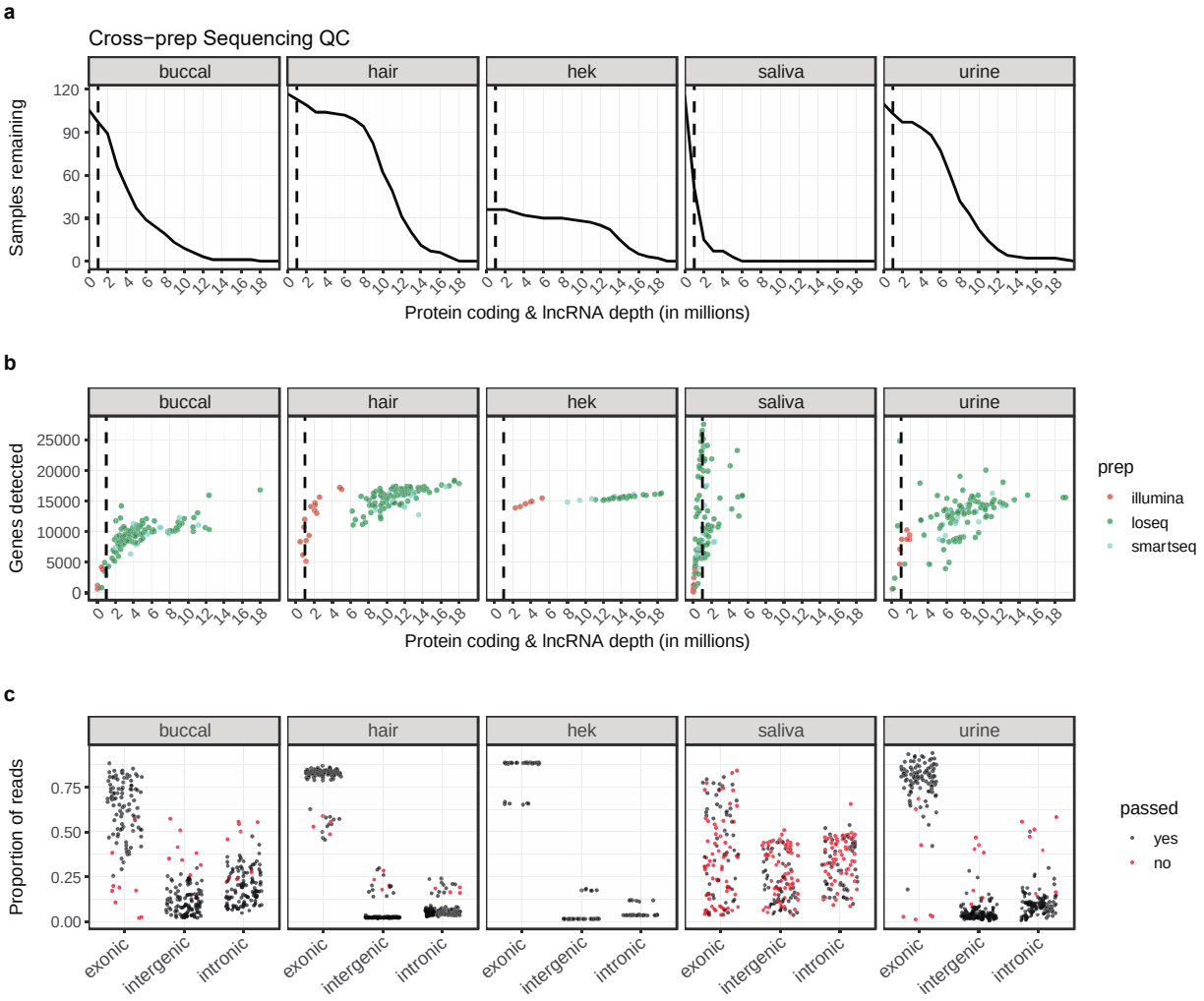
f

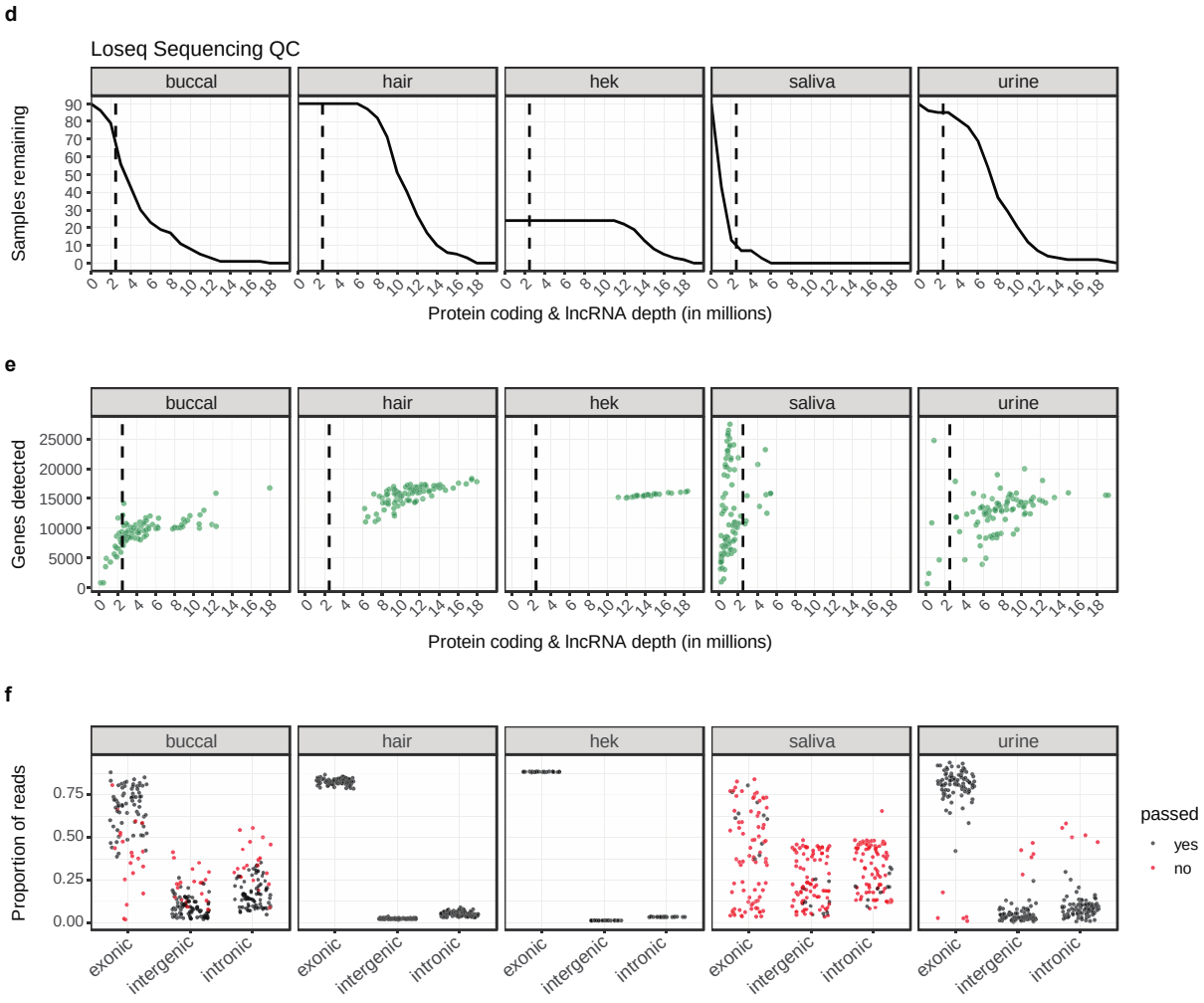

Total Reads
buccal
hair
hek
saliva
urine
4e+07
3e+07
2e+07
1e+07
0e+00
Total Reads
Samples
prep
illumina
loseq
smartseq

g


Uniquely mapped reads gene types
buccal
hair
hek
saliva
urine
2.5e+07
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Reads
Samples
type
other
Mt_rRNA
lncRNA
protein_coding

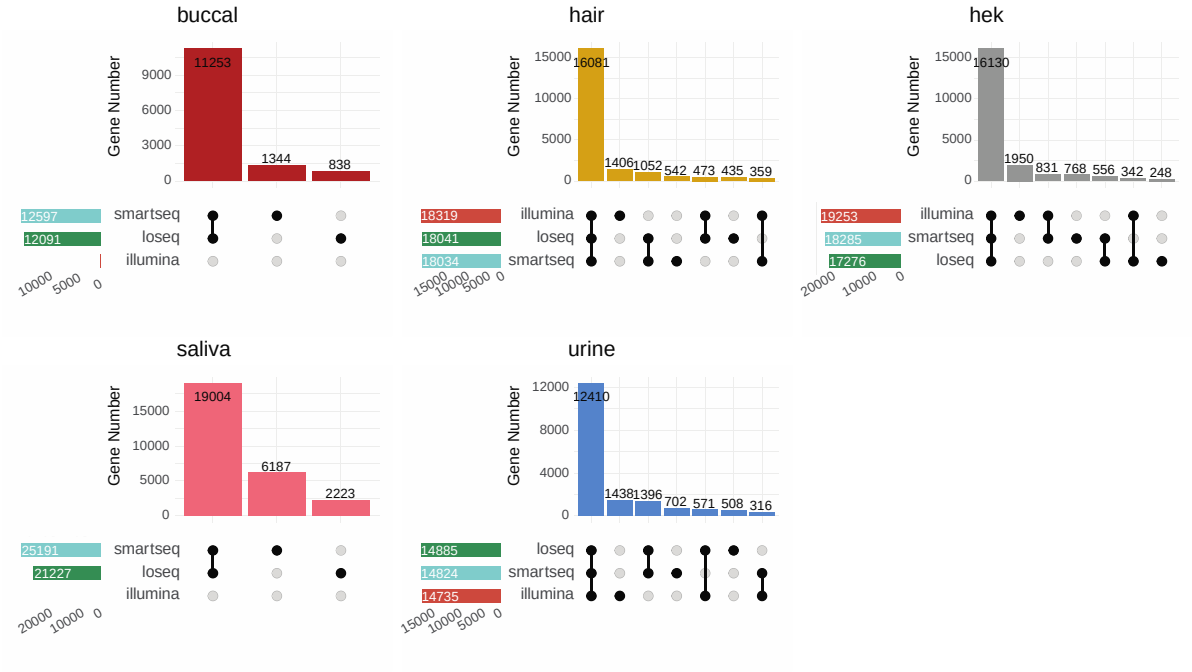
Supplementary Figure 2. a-c quality statistics for all samples across all library preparations: **a. b.** Samples remaining and genes detected depending on depth threshold. Dotted line indicates prep QC threshold of 1 million. **c.** Proportion of reads mapping to various genomic features. Pass indicator is based on prep QC threshold. **d-f** quality statistics for only Loseq samples: **d. e.** Samples remaining and genes detected depending on depth threshold. Dotted line indicates Loseq QC threshold of 2.5 million. **f.** Proportion of reads mapping to genomic features. Pass indicator is based on Loseq QC threshold.



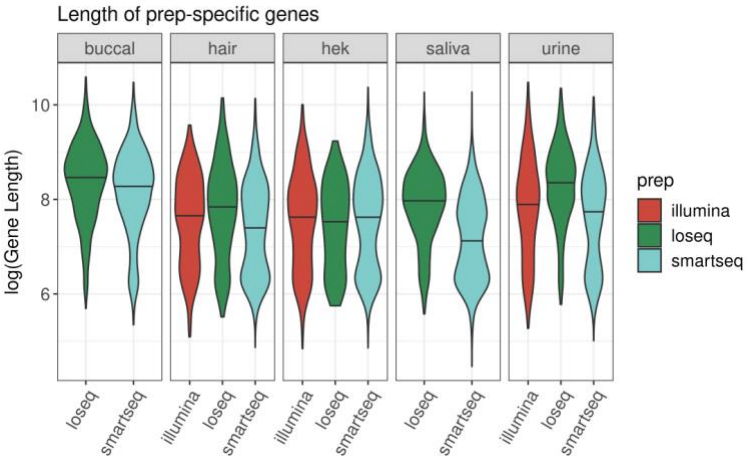


Supplementary Figure 3. a. Intersection of genes with greater than zero median expression across library preparations. **b. c.** Comparison of gene length and GC content of genes uniquely captured in a given preparation. **d.** Median expression for replicate 1 and replicate 2 samples was taken per tissue and preparation. The spearman correlation between these is shown.

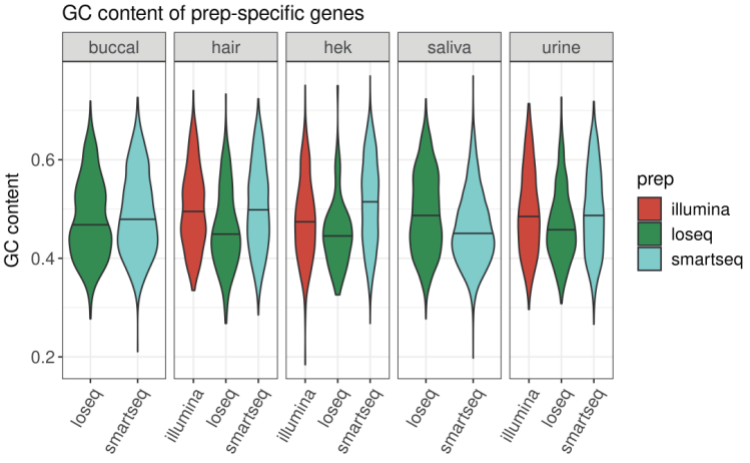
a


buccal
Gene Number
11253
1344
838
12597
12091
smartseq
loseq
illumina
hair
16081
1406
1052
542
473
435
359
18319
18041
18034
illumina
loseq
smartseq
hek
16130
1950
831
768
556
342
248
19253
18285
17276
illumina
smartseq
loseq
saliva
19004
6187
2223
25191
21227
smartseq
loseq
illumina
urine
12410
1438
1396
702
571
508
316
14885
14824
14735
loseq
smartseq
illumina

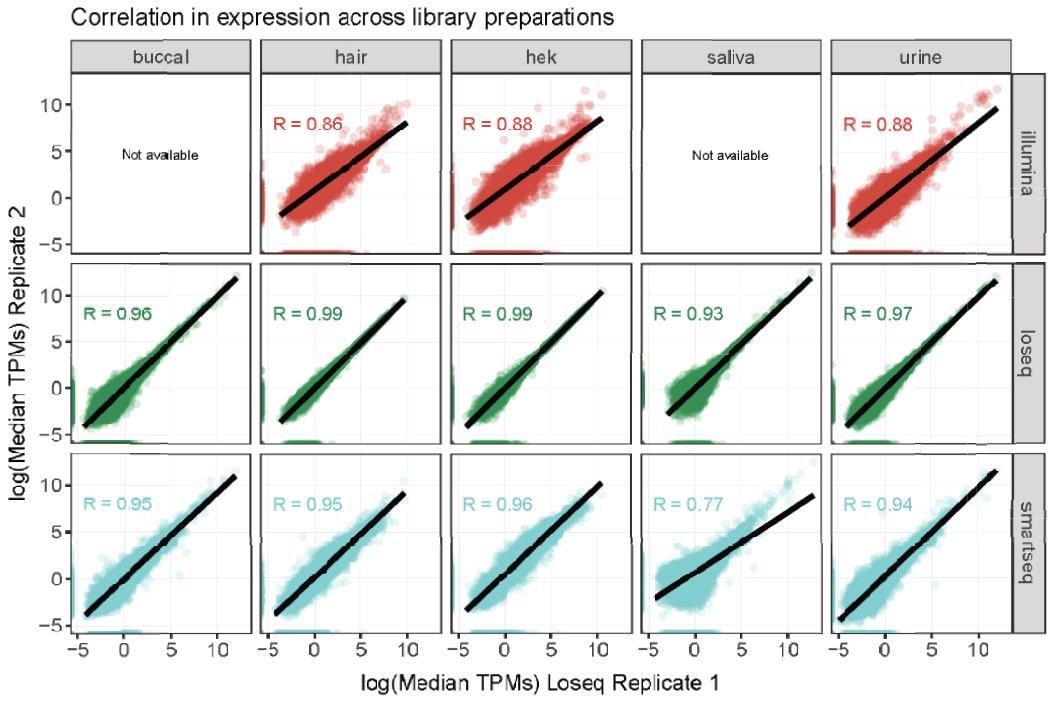
b


Length of prep-specific genes
buccal
hair
hek
saliva
urine
log(Gene Length)
prep
illumina
loseq
smartseq

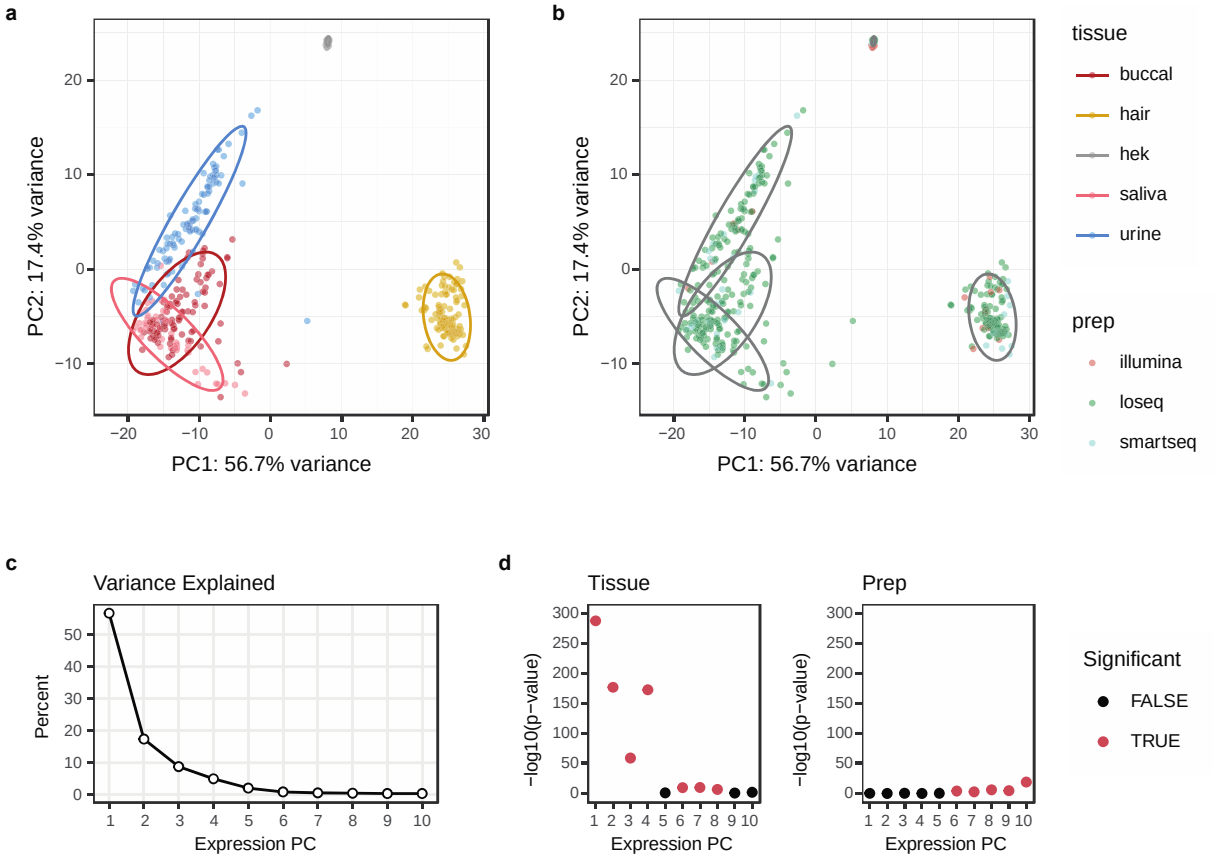
c


GC content of prep-specific genes
buccal
hair
hek
saliva
urine
GC content
prep
illumina
loseq
smartseq

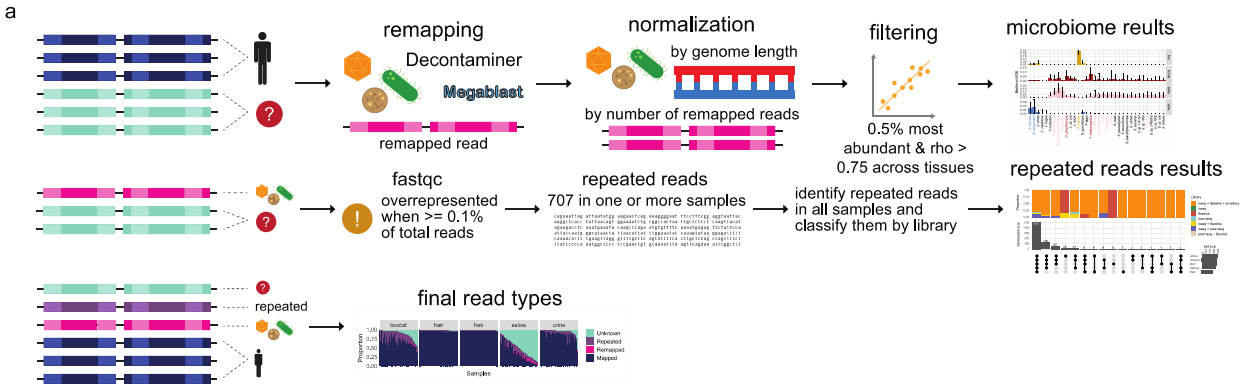
d

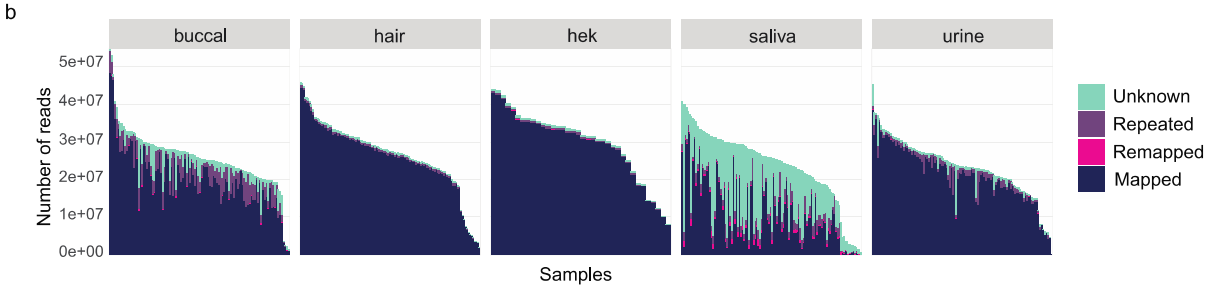

Correlation in expression across library preparations
buccal
hair
hek
saliva
urine
illumina
loseq
smartseq
Not available
R = 0.86
R = 0.88
Not available
R = 0.88
R = 0.96
R = 0.99
R = 0.99
R = 0.93
R = 0.97
R = 0.95
R = 0.95
R = 0.96
R = 0.77
R = 0.94
log(Median TPMs) Replicate 2
log(Median TPMs) Loseq Replicate 1

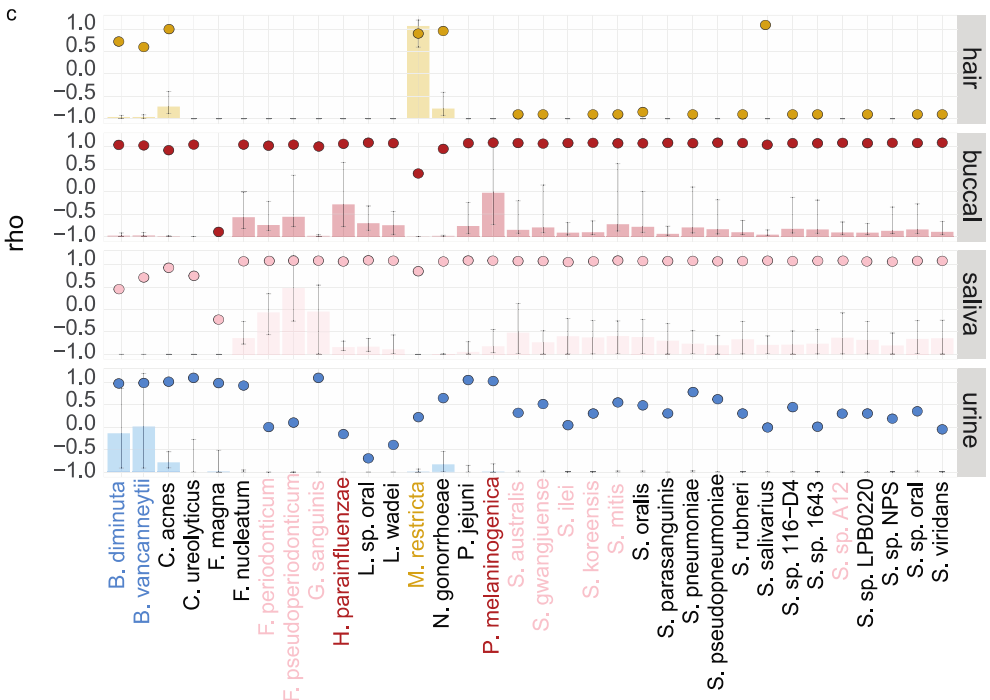
Supplementary Figure 4. a. b. Principal Component Analysis of all samples passing preparation QC thresholds. **c.** Percent variance explained per PC. **d.** ANOVA results for PC ~ tissue and PC ~ preparation. P-values are Bonferroni-corrected for the number of PCs tested (10).

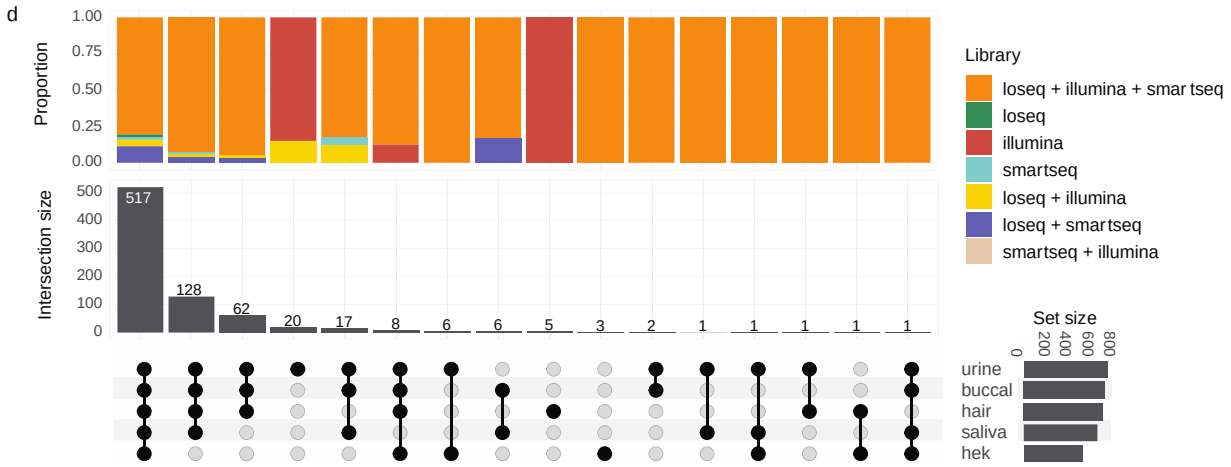


Supplementary Figure 5. a. Schematic of pipeline for assigning unmapped reads. **b.** Total number of reads assigned to each category. Mapped = aligned to hg38. Remapped = aligned to microbial species using Decontaminer. Repeated = highly abundant reads identified by FastQC. Unknown = reads not mapped or highly abundant. **c.** Per species included in the final analysis, spearman rank correlation between replicates is shown with the dots. Bar plot of species abundance with error bars is shown in the background for direct comparison (y-scale of abundance in Figure 2b). **d.** Breakdown of repeated sequence sharing across tissues and library preparations.

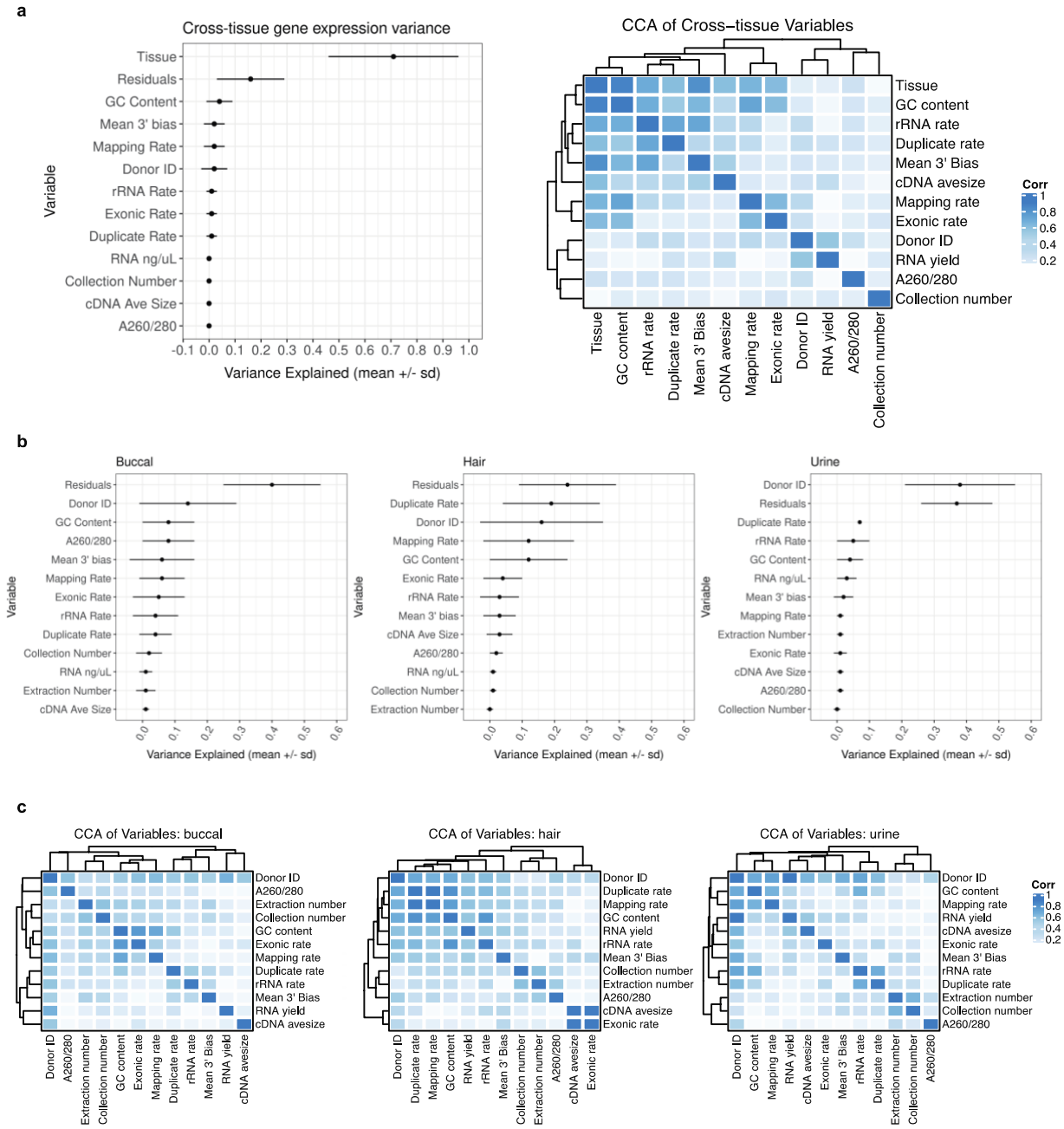

a
remapping
Decontaminer
Megablast
remapped read
normalization
by genome length
by number of remapped reads
filtering
0.5% most
abundant & rho >
0.75 across tissues
microbiome reults
fastqc
overrepresented
when >= 0.1%
of total reads
repeated reads
707 in one or more samples
identify repeated reads
in all samples and
classify them by library
repeated reads results
repeated
final read types
Unknown
Repeated
Remapped
Mapped
Samples


b
buccal
hair
hek
saliva
urine
Number of reads
5e+07
4e+07
3e+07
2e+07
1e+07
0e+00
Samples
Unknown
Repeated
Remapped
Mapped

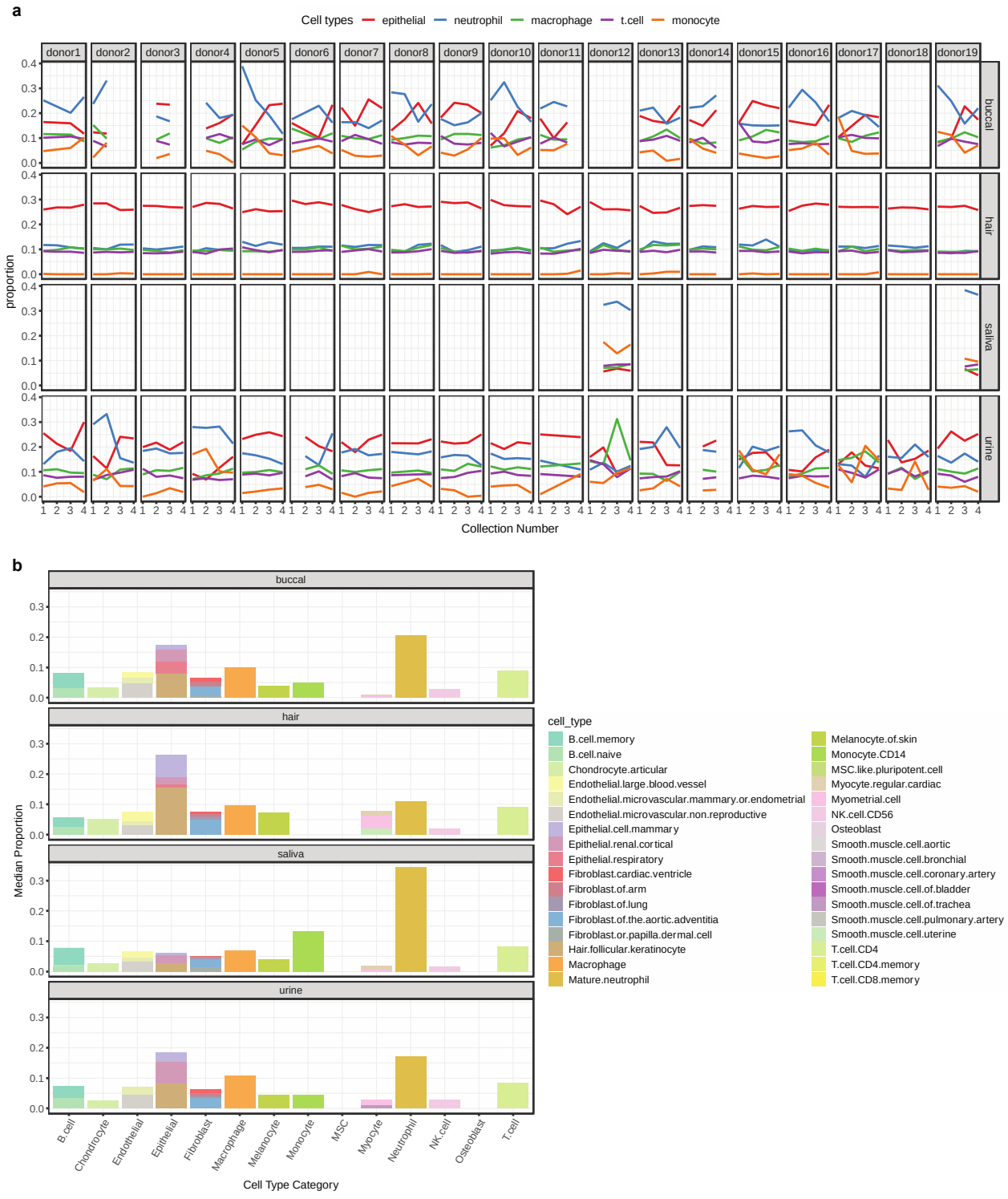

c
rho
hair
buccal
saliva
urine
B. diminuta
B. vancanneytii
C. acnes
C. ureolyticus
F. magna
F. nucleatum
F. periodonticum
F. pseudoperiodonticum
G. sanguinis
H. parainfluenzae
L. sp. oral
L. wadei
M. restricta
N. gonorrhoeae
P. jejuni
P. melaninogenica
S. australis
S. gwangjuense
S. ilei
S. koreensis
S. mitis
S. oralis
S. parasanguinis
S. pneumoniae
S. pseudopneumoniae
S. rubneri
S. salivarius
S. sp. 116−D4
S. sp. 1643
S. sp. A12
S. sp. LPB0220
S. sp. NPS
S. sp. oral
S. viridans


d
Proportion
1.00
0.75
0.50
0.25
0.00
Intersection size
500
400
300
200
100
0
517
128
62
20
17
8
6
6
5
3
2
1
1
1
1
1
Library
loseq + illumina + smartseq
loseq
illumina
smartseq
loseq + illumina
loseq + smartseq
smartseq + illumina
Set size
0
200
400
600
800
urine
buccal
hair
saliva
hek

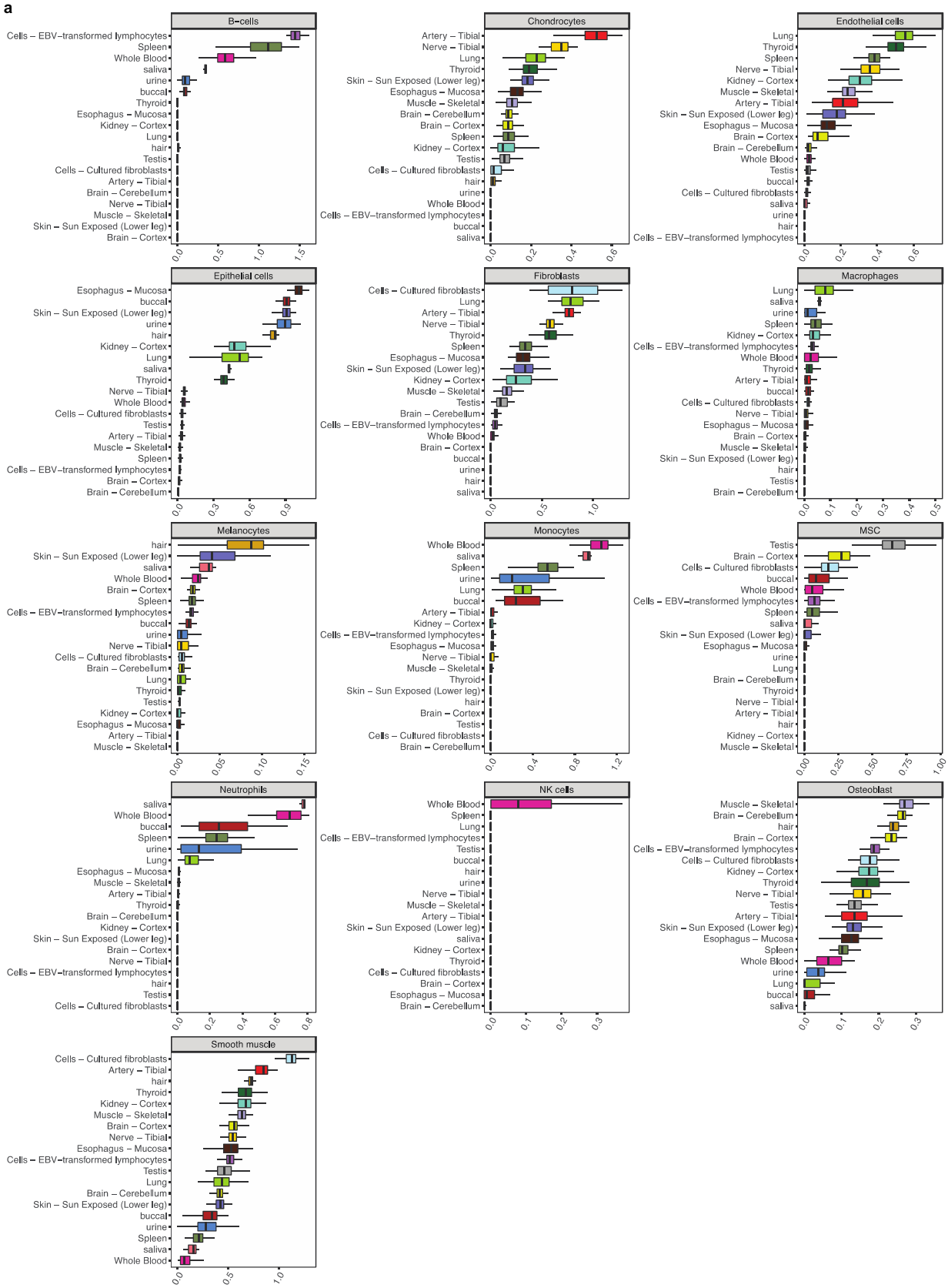
Supplementary Figure 6. a. Variance in gene expression across tissues explained by technical and biological variables. Canonical correlation analysis shows correlation between variables used. **b. c.** Variance in gene expression within each tissue explained by technical and biological variables. Canonical correlation analysis shows correlation between variables used. Only Loseq samples were included in these analyses.

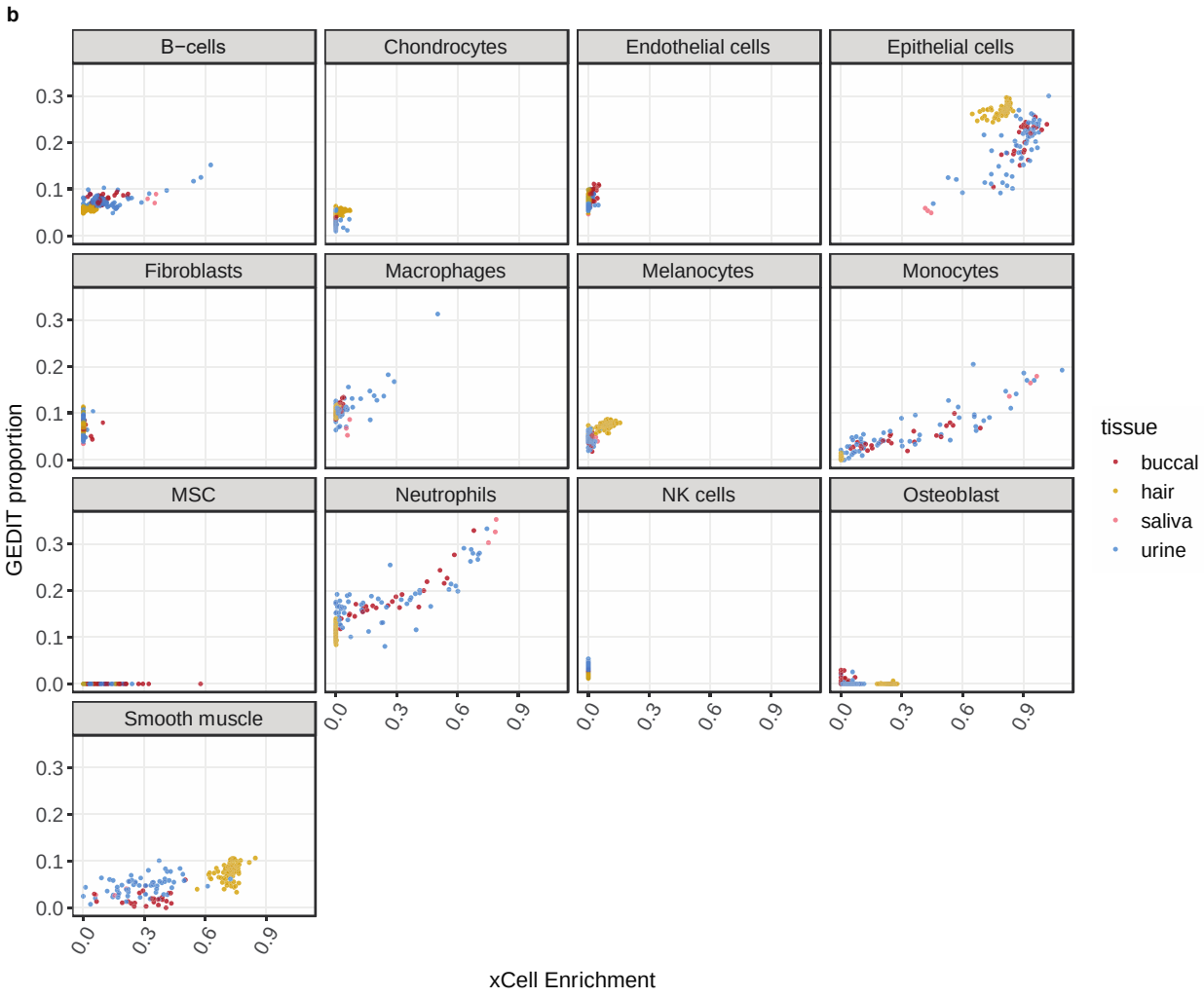


Supplementary Figure 7. a. GEDIT cell type proportion estimates per collection and donor. Top 25% most abundant, condensed cell type categories are shown. **b.** Breakdown of all cell types included in the GEDIT reference. Binning into larger cell type categories is shown.

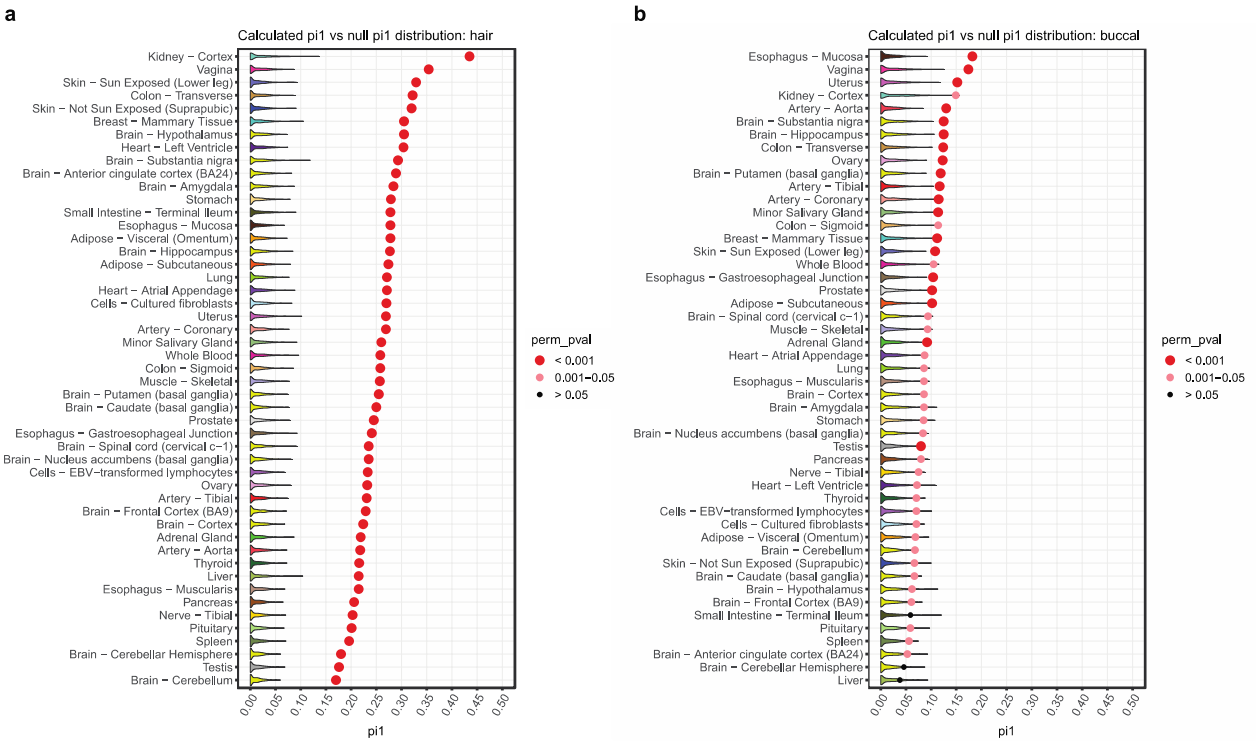


Supplementary Figure 8. a. xCell enrichment scores across the noninvasive and select GTEx tissues for cell types corresponding to the GEDIT collapsed cell type categories. Note that the enrichment score does not correspond to a proportion. **b.** Comparison of xCell enrichment scores and GEDIT proportions for cell types shared by both references. Only noninvasive tissues are shown.

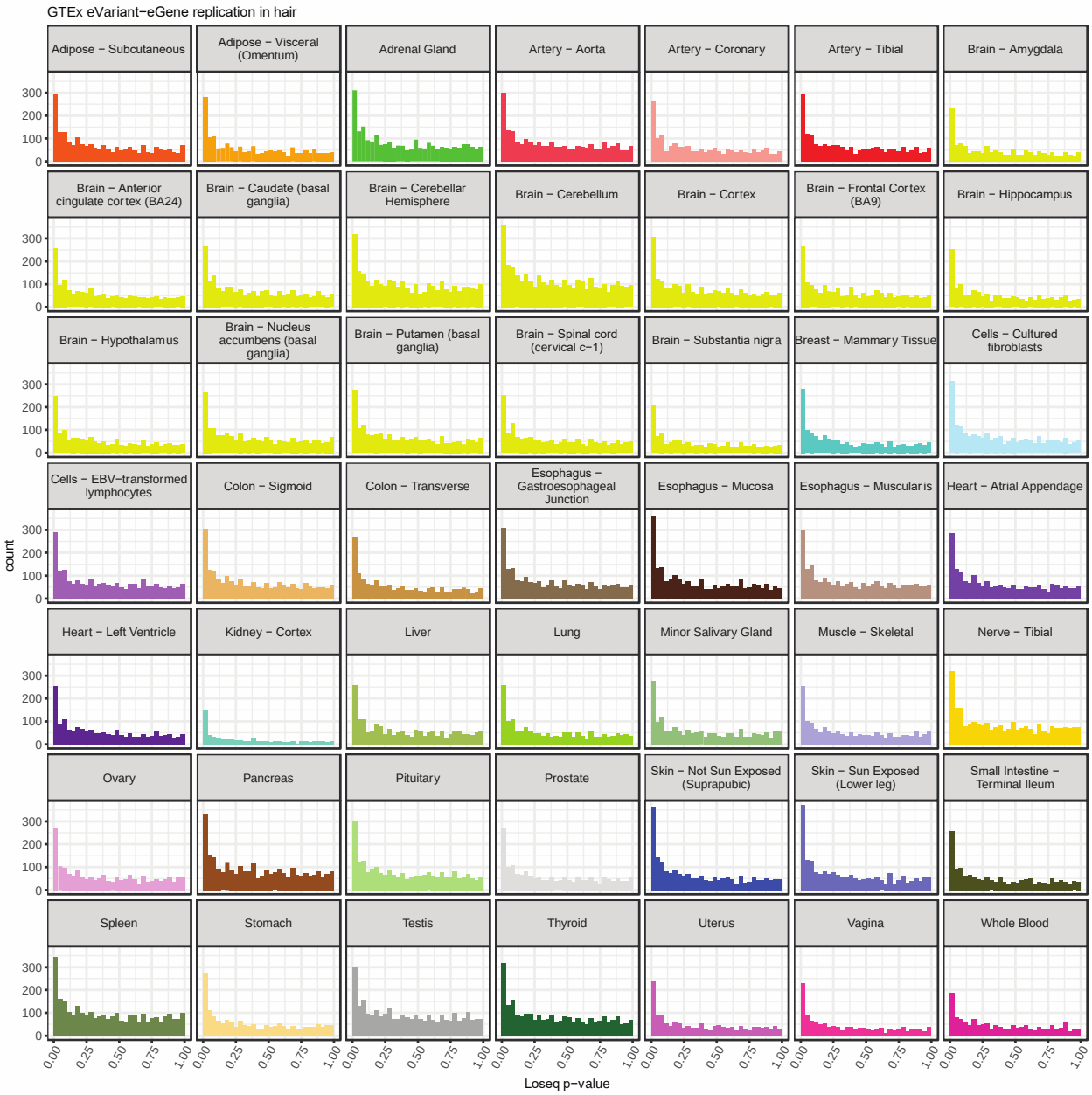

a
B–cells
Cells – EBV–transformed lymphocytes
Spleen
Whole Blood
saliva
urine
buccal
Thyroid
Esophagus – Mucosa
Kidney – Cortex
Lung
hair
Testis
Cells – Cultured fibroblasts
Artery – Tibial
Brain – Cerebellum
Nerve – Tibial
Muscle – Skeletal
Skin – Sun Exposed (Lower leg)
Brain – Cortex
Chondrocytes
Artery – Tibial
Nerve – Tibial
Lung
Thyroid
Skin – Sun Exposed (Lower leg)
Esophagus – Mucosa
Muscle – Skeletal
Brain – Cerebellum
Brain – Cortex
Spleen
Kidney – Cortex
Testis
Cells – Cultured fibroblasts
hair
urine
Whole Blood
Cells – EBV–transformed lymphocytes
buccal
saliva
Endothelial cells
Lung
Thyroid
Spleen
Nerve – Tibial
Kidney – Cortex
Muscle – Skeletal
Artery – Tibial
Skin – Sun Exposed (Lower leg)
Esophagus – Mucosa
Brain – Cortex
Brain – Cerebellum
Whole Blood
Testis
buccal
Cells – Cultured fibroblasts
saliva
urine
hair
Cells – EBV–transformed lymphocytes
Epithelial cells
Esophagus – Mucosa
buccal
Skin – Sun Exposed (Lower leg)
urine
hair
Kidney – Cortex
Lung
saliva
Thyroid
Nerve – Tibial
Whole Blood
Cells – Cultured fibroblasts
Testis
Artery – Tibial
Muscle – Skeletal
Spleen
Cells – EBV–transformed lymphocytes
Brain – Cortex
Brain – Cerebellum
Fibroblasts
Cells – Cultured fibroblasts
Lung
Artery – Tibial
Nerve – Tibial
Thyroid
Spleen
Esophagus – Mucosa
Skin – Sun Exposed (Lower leg)
Kidney – Cortex
Muscle – Skeletal
Testis
Brain – Cerebellum
Cells – EBV–transformed lymphocytes
Whole Blood
Brain – Cortex
buccal
urine
hair
saliva
Macrophages
Lung
saliva
urine
Spleen
Kidney – Cortex
Cells – EBV–transformed lymphocytes
Whole Blood
Thyroid
Artery – Tibial
buccal
Cells – Cultured fibroblasts
Nerve – Tibial
Esophagus – Mucosa
Brain – Cortex
Muscle – Skeletal
Skin – Sun Exposed (Lower leg)
hair
Testis
Brain – Cerebellum
Melanocytes
hair
Skin – Sun Exposed (Lower leg)
saliva
Whole Blood
Brain – Cortex
Spleen
Cells – EBV–transformed lymphocytes
buccal
urine
Nerve – Tibial
Cells – Cultured fibroblasts
Brain – Cerebellum
Lung
Thyroid
Testis
Kidney – Cortex
Esophagus – Mucosa
Artery – Tibial
Muscle – Skeletal
Monocytes
Whole Blood
saliva
Spleen
urine
Lung
buccal
Artery – Tibial
Kidney – Cortex
Cells – EBV–transformed lymphocytes
Esophagus – Mucosa
Nerve – Tibial
Muscle – Skeletal
Thyroid
Skin – Sun Exposed (Lower leg)
hair
Brain – Cortex
Testis
Cells – Cultured fibroblasts
Brain – Cerebellum
MSC
Testis
Brain – Cortex
Cells – Cultured fibroblasts
buccal
Whole Blood
Cells – EBV–transformed lymphocytes
Spleen
saliva
Skin – Sun Exposed (Lower leg)
Esophagus – Mucosa
urine
Lung
Brain – Cerebellum
Thyroid
Nerve – Tibial
Artery – Tibial
hair
Kidney – Cortex
Muscle – Skeletal
Neutrophils
saliva
Whole Blood
buccal
Spleen
urine
Lung
Esophagus – Mucosa
Muscle – Skeletal
Artery – Tibial
Thyroid
Brain – Cerebellum
Kidney – Cortex
Skin – Sun Exposed (Lower leg)
Brain – Cortex
Nerve – Tibial
Cells – EBV–transformed lymphocytes
hair
Testis
Cells – Cultured fibroblasts
NK cells
Whole Blood
Spleen
Lung
Cells – EBV–transformed lymphocytes
Testis
buccal
hair
urine
Nerve – Tibial
Muscle – Skeletal
Artery – Tibial
Skin – Sun Exposed (Lower leg)
saliva
Kidney – Cortex
Thyroid
Cells – Cultured fibroblasts
Brain – Cortex
Esophagus – Mucosa
Brain – Cerebellum
Osteoblast
Muscle – Skeletal
Brain – Cerebellum
hair
Brain – Cortex
Cells – EBV–transformed lymphocytes
Cells – Cultured fibroblasts
Kidney – Cortex
Thyroid
Nerve – Tibial
Testis
Artery – Tibial
Skin – Sun Exposed (Lower leg)
Esophagus – Mucosa
Spleen
Whole Blood
urine
Lung
buccal
saliva
Smooth muscle
Cells – Cultured fibroblasts
Artery – Tibial
hair
Thyroid
Kidney – Cortex
Muscle – Skeletal
Brain – Cortex
Nerve – Tibial
Esophagus – Mucosa
Cells – EBV–transformed lymphocytes
Testis
Lung
Brain – Cerebellum
Skin – Sun Exposed (Lower leg)
buccal
urine
Spleen
saliva
Whole Blood

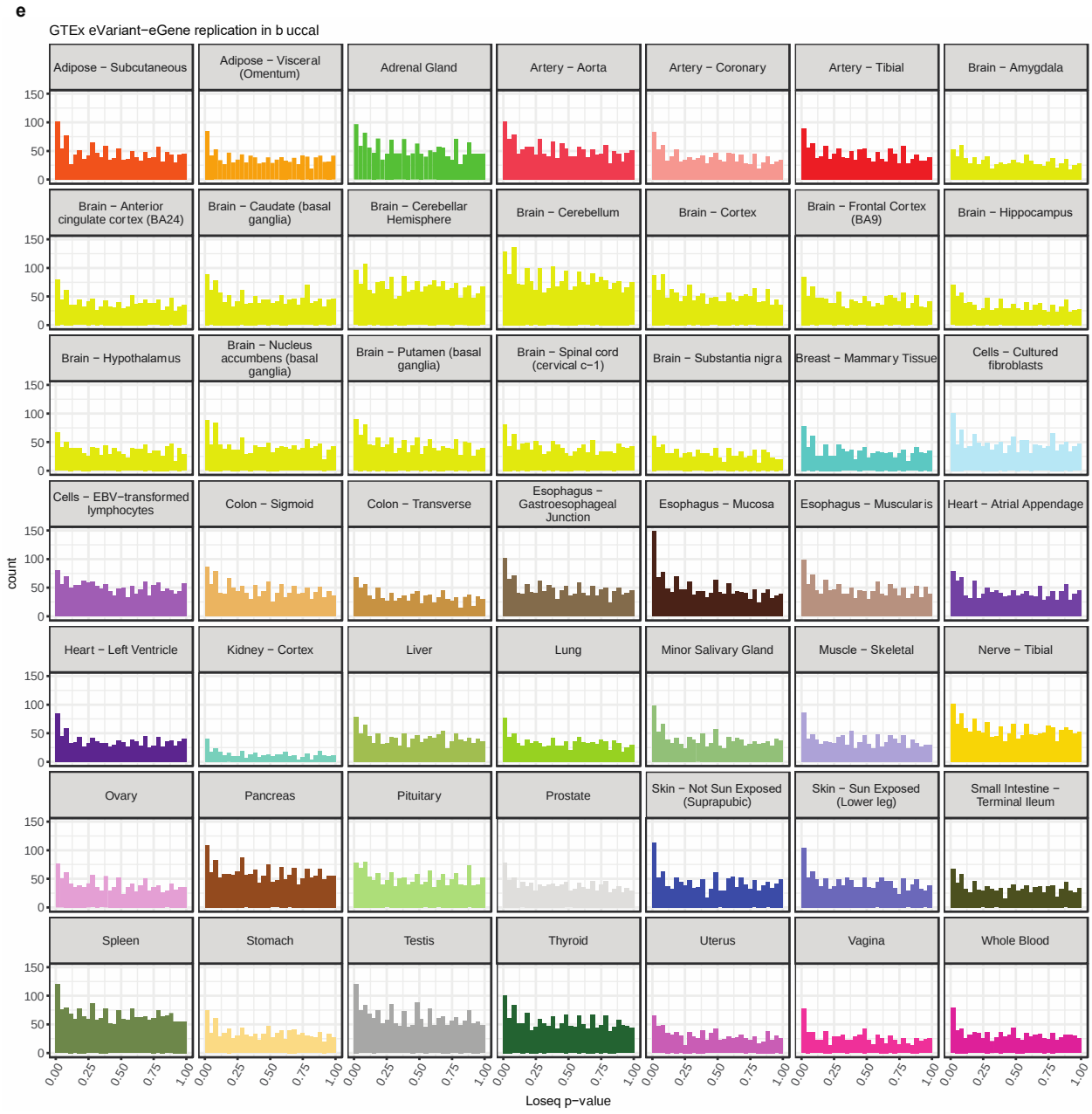


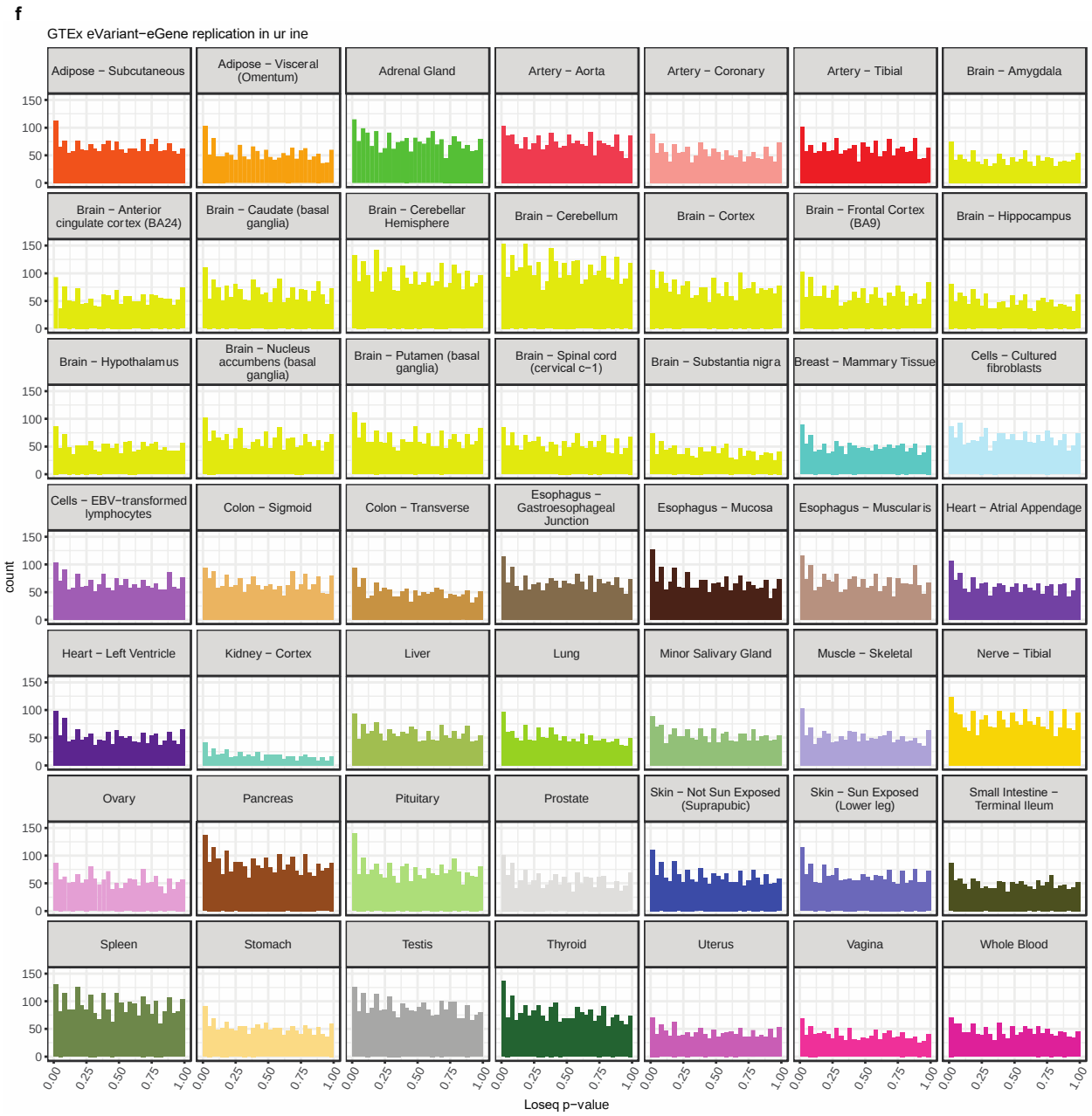
Supplementary Figure 9. a. b. c. Calculated pi1 versus the null pi1 distribution for every tissue in GTEx. Pi1 was calculated by selecting for significant variants (q-value <=0.05) with MAF > 0.05 and minimum effect size greater than the maximum minimum across GTEx tissues (kidney cortex 0.32) that were present in the noninvasive dataset. The null distribution was generated by performing 1000 samples of size equivalent to the number of overlapping gene-variant pairs used for the pi1 calculation for that tissue. **d. e. f.** Histograms of Loseq p-values for gene-variant pairs included in the pi1 calculation.


a
Calculated pi1 vs null pi1 distribution: hair
Kidney – Cortex
Vagina
Skin – Sun Exposed (Lower leg)
Colon – Transverse
Skin – Not Sun Exposed (Suprapubic)
Breast – Mammary Tissue
Brain – Hypothalamus
Heart – Left Ventricle
Brain – Substantia nigra
Brain – Anterior cingulate cortex (BA24)
Brain – Amygdala
Stomach
Small Intestine – Terminal Ileum
Esophagus – Mucosa
Adipose – Visceral (Omentum)
Brain – Hippocampus
Adipose – Subcutaneous
Lung
Heart – Atrial Appendage
Cells – Cultured fibroblasts
Uterus
Artery – Coronary
Minor Salivary Gland
Whole Blood
Colon – Sigmoid
Muscle – Skeletal
Brain – Putamen (basal ganglia)
Brain – Caudate (basal ganglia)
Prostate
Esophagus – Gastroesophageal Junction
Brain – Spinal cord (cervical c–1)
Brain – Nucleus accumbens (basal ganglia)
Cells – EBV–transformed lymphocytes
Ovary
Artery – Tibial
Brain – Frontal Cortex (BA9)
Brain – Cortex
Adrenal Gland
Artery – Aorta
Thyroid
Liver
Esophagus – Muscularis
Pancreas
Nerve – Tibial
Pituitary
Spleen
Brain – Cerebellar Hemisphere
Testis
Brain – Cerebellum
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
pi1
perm_pval
< 0.001
0.001–0.05
> 0.05
b
Calculated pi1 vs null pi1 distribution: buccal
Esophagus – Mucosa
Vagina
Uterus
Kidney – Cortex
Artery – Aorta
Brain – Substantia nigra
Brain – Hippocampus
Colon – Transverse
Ovary
Brain – Putamen (basal ganglia)
Artery – Tibial
Artery – Coronary
Minor Salivary Gland
Colon – Sigmoid
Breast – Mammary Tissue
Skin – Sun Exposed (Lower leg)
Whole Blood
Esophagus – Gastroesophageal Junction
Prostate
Adipose – Subcutaneous
Brain – Spinal cord (cervical c–1)
Muscle – Skeletal
Adrenal Gland
Heart – Atrial Appendage
Lung
Esophagus – Muscularis
Brain – Cortex
Brain – Amygdala
Stomach
Brain – Nucleus accumbens (basal ganglia)
Testis
Pancreas
Nerve – Tibial
Heart – Left Ventricle
Thyroid
Cells – EBV–transformed lymphocytes
Cells – Cultured fibroblasts
Adipose – Visceral (Omentum)
Brain – Cerebellum
Skin – Not Sun Exposed (Suprapubic)
Brain – Caudate (basal ganglia)
Brain – Hypothalamus
Brain – Frontal Cortex (BA9)
Small Intestine – Terminal Ileum
Pituitary
Spleen
Brain – Anterior cingulate cortex (BA24)
Brain – Cerebellar Hemisphere
Liver
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
pi1
perm_pval
< 0.001
0.001–0.05
> 0.05
c
Calculated pi1 vs null pi1 distribution: urine
Kidney – Cortex
Lung
Brain – Substantia nigra
Stomach
Liver
Colon – Transverse
Adipose – Visceral (Omentum)
Prostate
Thyroid
Skin – Not Sun Exposed (Suprapubic)
Esophagus – Mucosa
Vagina
Muscle – Skeletal
Minor Salivary Gland
Artery – Coronary
Ovary
Brain – Hippocampus
Heart – Left Ventricle
Heart – Atrial Appendage
Esophagus – Muscularis
Pancreas
Brain – Frontal Cortex (BA9)
Uterus
Whole Blood
Brain – Cerebellum
Breast – Mammary Tissue
Testis
Adrenal Gland
Skin – Sun Exposed (Lower leg)
Brain – Nucleus accumbens (basal ganglia)
Pituitary
Brain – Caudate (basal ganglia)
Adipose – Subcutaneous
Brain – Cerebellar Hemisphere
Brain – Putamen (basal ganglia)
Brain – Hypothalamus
Colon – Sigmoid
Brain – Cortex
Cells – EBV–transformed lymphocytes
Spleen
Cells – Cultured fibroblasts
Artery – Tibial
Nerve – Tibial
Brain – Spinal cord (cervical c–1)
Small Intestine – Terminal Ileum
Artery – Aorta
Esophagus – Gastroesophageal Junction
Brain – Amygdala
Brain – Anterior cingulate cortex (BA24)
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
pi1
perm_pval
< 0.001
0.001–0.05
> 0.05

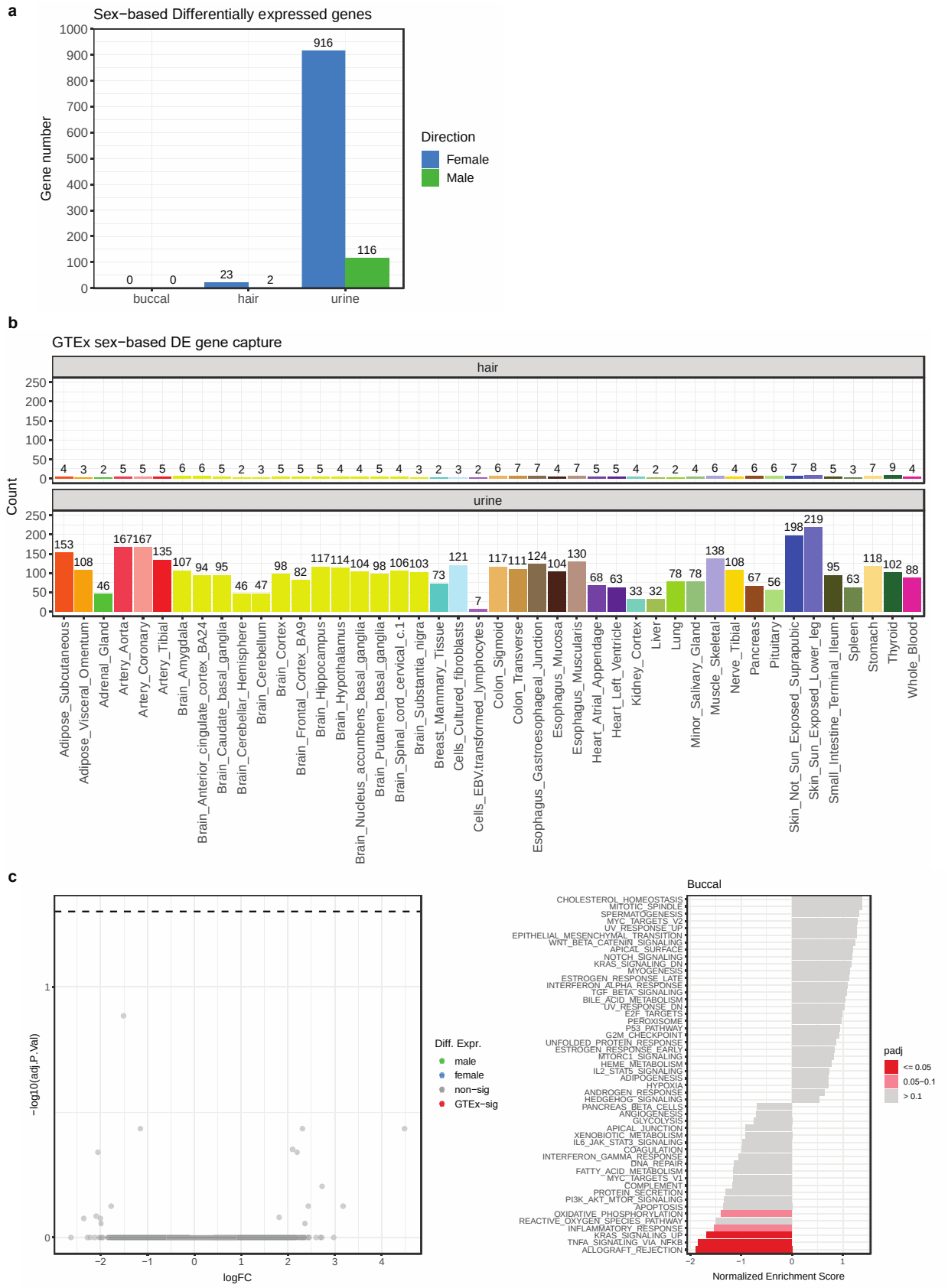
d


GTEx eVariant−eGene replication in hair
Adipose − Subcutaneous
Adipose − Visceral (Omentum)
Adrenal Gland
Artery − Aorta
Artery − Coronary
Artery − Tibial
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Caudate (basal ganglia)
Brain − Cerebellar Hemisphere
Brain − Cerebellum
Brain − Cortex
Brain − Frontal Cortex (BA9)
Brain − Hippocampus
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Breast − Mammary Tissue
Cells − Cultured fibroblasts
Cells − EBV−transformed lymphocytes
Colon − Sigmoid
Colon − Transverse
Esophagus − Gastroesophageal Junction
Esophagus − Mucosa
Esophagus − Muscularis
Heart − Atrial Appendage
Heart − Left Ventricle
Kidney − Cortex
Liver
Lung
Minor Salivary Gland
Muscle − Skeletal
Nerve − Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin − Not Sun Exposed (Suprapubic)
Skin − Sun Exposed (Lower leg)
Small Intestine − Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
count
0
100
200
300
0.00
0.25
0.50
0.75
1.00
Loseq p−value


e
GTEx eVariant−eGene replication in b uccal
Adipose − Subcutaneous
Adipose − Visceral (Omentum)
Adrenal Gland
Artery − Aorta
Artery − Coronary
Artery − Tibial
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Caudate (basal ganglia)
Brain − Cerebellar Hemisphere
Brain − Cerebellum
Brain − Cortex
Brain − Frontal Cortex (BA9)
Brain − Hippocampus
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Breast − Mammary Tissue
Cells − Cultured fibroblasts
Cells − EBV−transformed lymphocytes
Colon − Sigmoid
Colon − Transverse
Esophagus − Gastroesophageal Junction
Esophagus − Mucosa
Esophagus − Muscularis
Heart − Atrial Appendage
Heart − Left Ventricle
Kidney − Cortex
Liver
Lung
Minor Salivary Gland
Muscle − Skeletal
Nerve − Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin − Not Sun Exposed (Suprapubic)
Skin − Sun Exposed (Lower leg)
Small Intestine − Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
count
0
50
100
150
0.00
0.25
0.50
0.75
1.00
Loseq p−value

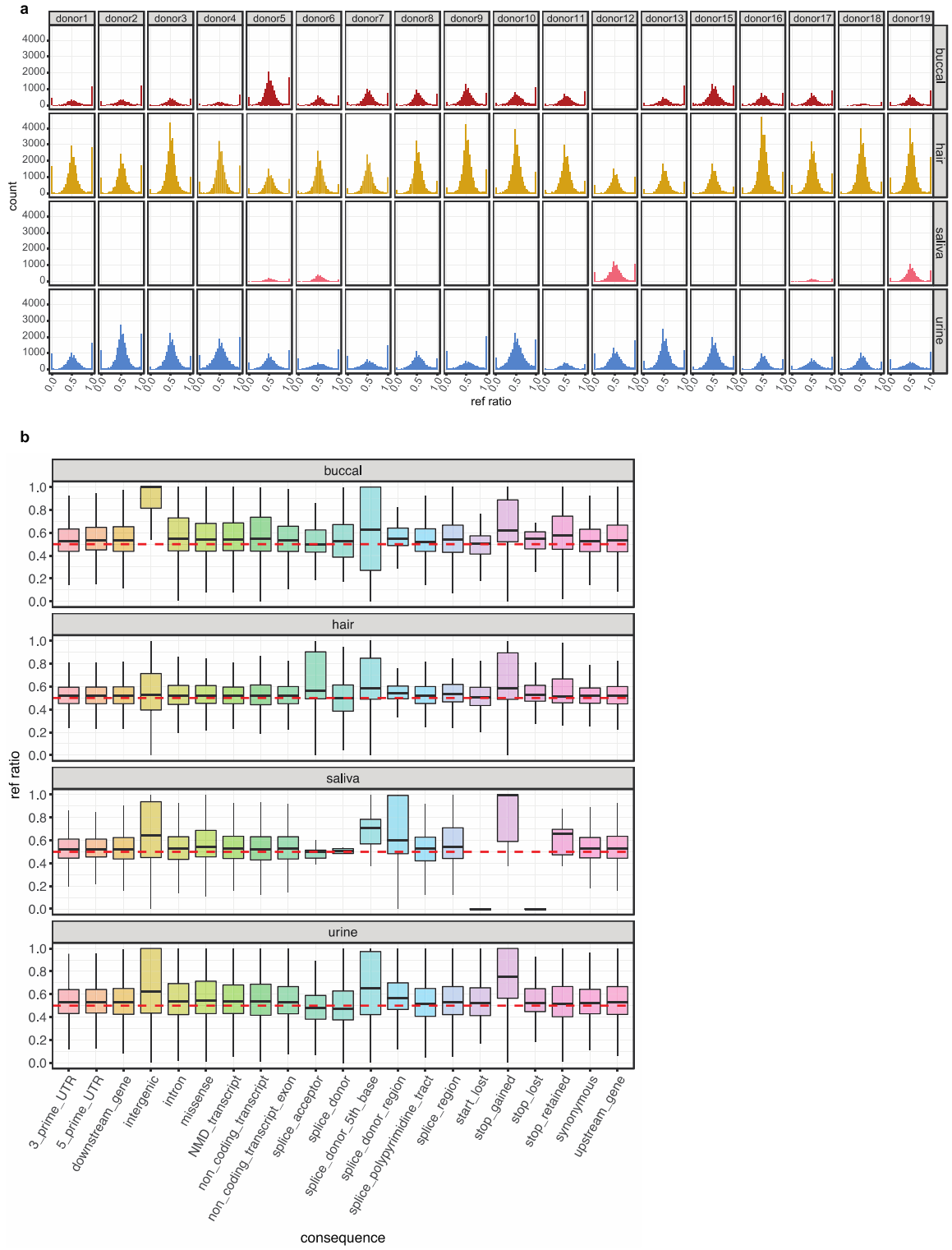

f
GTEx eVariant−eGene replication in ur ine
Adipose − Subcutaneous
Adipose − Visceral (Omentum)
Adrenal Gland
Artery − Aorta
Artery − Coronary
Artery − Tibial
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Caudate (basal ganglia)
Brain − Cerebellar Hemisphere
Brain − Cerebellum
Brain − Cortex
Brain − Frontal Cortex (BA9)
Brain − Hippocampus
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Breast − Mammary Tissue
Cells − Cultured fibroblasts
Cells − EBV−transformed lymphocytes
Colon − Sigmoid
Colon − Transverse
Esophagus − Gastroesophageal Junction
Esophagus − Mucosa
Esophagus − Muscularis
Heart − Atrial Appendage
Heart − Left Ventricle
Kidney − Cortex
Liver
Lung
Minor Salivary Gland
Muscle − Skeletal
Nerve − Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin − Not Sun Exposed (Suprapubic)
Skin − Sun Exposed (Lower leg)
Small Intestine − Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
count
0
50
100
150
0.00
0.25
0.50
0.75
1.00
Loseq p−value

Supplementary Figure 10. a. Number of sex-based upregulated genes per noninvasive tissue type. **b.** Per tissue overlap between significant DE genes in the noninvasive dataset with genes previously found to be significant in the GTEx dataset. **c.** Sex-based differential expression for buccal samples, with no significant genes. FGSEA shows some rank-based gene category enrichment for females.

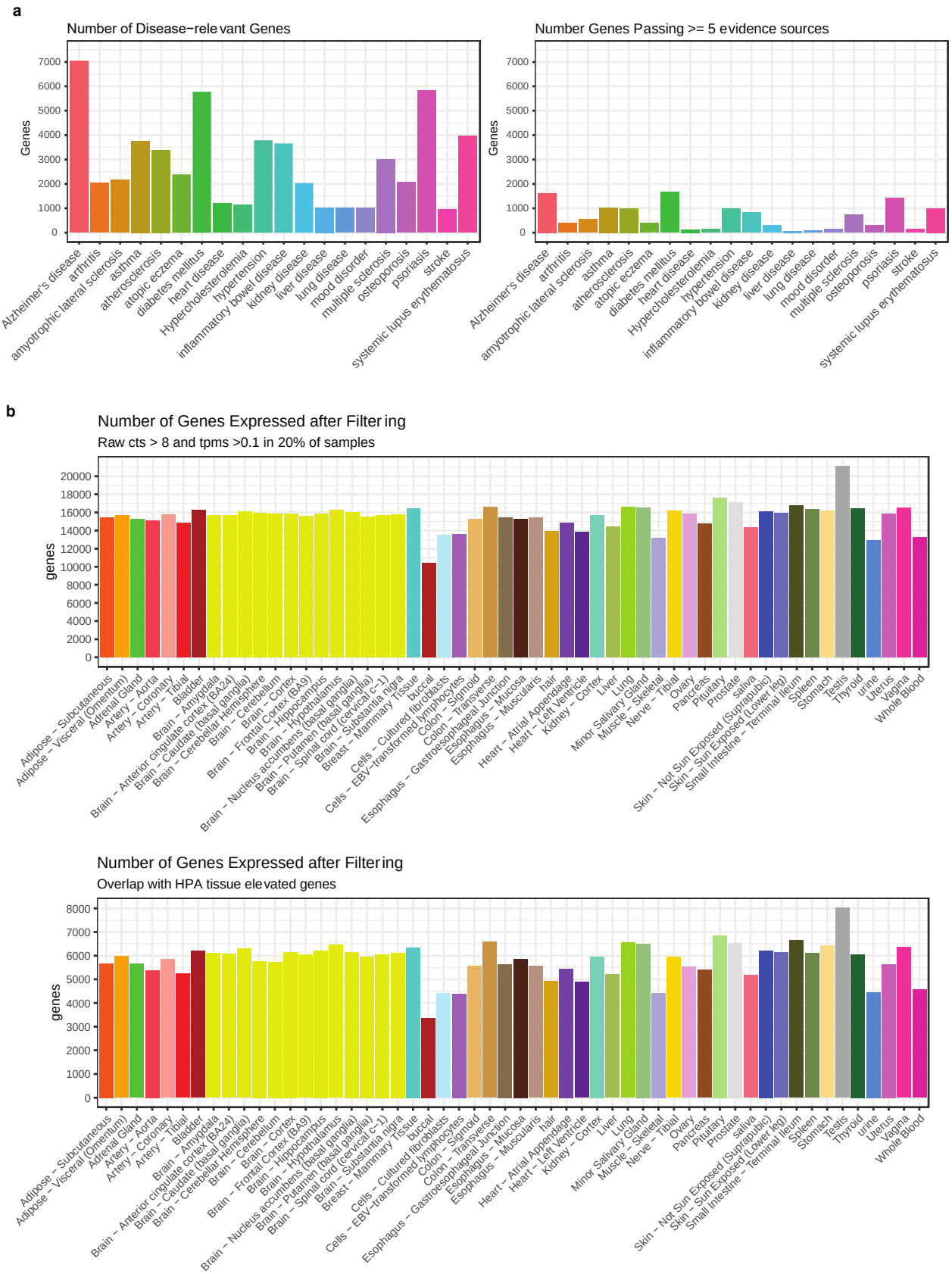

a
Sex−based Differentially expressed genes
Gene number
1000
900
800
700
600
500
400
300
200
100
0
916
116
23
2
0
0
buccal
hair
urine
Direction
Female
Male
b
GTEx sex−based DE gene capture
hair
urine
Count
250
200
150
100
50
0
4 3 2 5 5 5 6 6 5 2 3 5 5 5 4 4 5 4 3 2 3 2 6 7 7 4 7 5 5 4 2 2 4 6 4 6 6 7 8 5 3 7 9 4
153 108 46 167 167 135 107 94 95 46 47 98 82 117 114 104 98 106 103 73 121 7 117 111 124 104 130 68 63 33 32 78 78 138 108 67 56 198 219 95 63 118 102 88
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
Brain_Amygdala
Brain_Anterior_cingulate_cortex_BA24
Brain_Caudate_basal_ganglia
Brain_Cerebellar_Hemisphere
Brain_Cerebellum
Brain_Cortex
Brain_Frontal_Cortex_BA9
Brain_Hippocampus
Brain_Hypothalamus
Brain_Nucleus_accumbens_basal_ganglia
Brain_Putamen_basal_ganglia
Brain_Spinal_cord_cervical_c.1
Brain_Substantia_nigra
Breast_Mammary_Tissue
Cells_Cultured_fibroblasts
Cells_EBV.transformed_lymphocytes
Colon_Sigmoid
Colon_Transverse
Esophagus_Gastroesophageal_Junction
Esophagus_Mucosa
Esophagus_Muscularis
Heart_Atrial_Appendage
Heart_Left_Ventricle
Kidney_Cortex
Liver
Lung
Minor_Salivary_Gland
Muscle_Skeletal
Nerve_Tibial
Pancreas
Pituitary
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Small_Intestine_Terminal_Ileum
Spleen
Stomach
Thyroid
Whole_Blood
c
−log10(adj.P.Val)
1
0
−2
−1
0
1
2
3
4
logFC
Diff. Expr.
male
female
non−sig
GTEx−sig
Buccal
CHOLESTEROL_HOMEOSTASIS
MITOTIC_SPINDLE
SPERMATOGENESIS
MYC_TARGETS_V2
UV_RESPONSE_UP
EPITHELIAL_MESENCHYMAL_TRANSITION
WNT_BETA_CATENIN_SIGNALING
APICAL_SURFACE
NOTCH_SIGNALING
KRAS_SIGNALING_DN
MYOGENESIS
ESTROGEN_RESPONSE_LATE
INTERFERON_ALPHA_RESPONSE
TGF_BETA_SIGNALING
BILE_ACID_METABOLISM
UV_RESPONSE_DN
E2F_TARGETS
PEROXISOME
P53_PATHWAY
G2M_CHECKPOINT
UNFOLDED_PROTEIN_RESPONSE
ESTROGEN_RESPONSE_EARLY
MTORC1_SIGNALING
HEME_METABOLISM
IL2_STAT5_SIGNALING
ADIPOGENESIS
HYPOXIA
ANDROGEN_RESPONSE
HEDGEHOG_SIGNALING
PANCREAS_BETA_CELLS
ANGIOGENESIS
GLYCOLYSIS
APICAL_JUNCTION
XENOBIOTIC_METABOLISM
IL6_JAK_STAT3_SIGNALING
COAGULATION
INTERFERON_GAMMA_RESPONSE
DNA_REPAIR
FATTY_ACID_METABOLISM
MYC_TARGETS_V1
COMPLEMENT
PROTEIN_SECRETION
PI3K_AKT_MTOR_SIGNALING
APOPTOSIS
OXIDATIVE_PHOSPHORYLATION
REACTIVE_OXYGEN_SPECIES_PATHWAY
INFLAMMATORY_RESPONSE
KRAS_SIGNALING_UP
TNFA_SIGNALING_VIA_NFKB
ALLOGRAFT_REJECTION
−2
−1
0
1
Normalized Enrichment Score
padj
<= 0.05
0.05−0.1
> 0.1

Supplementary Figure 11. a. Ratio of (reference allele count)/(total count) per donor and tissue.

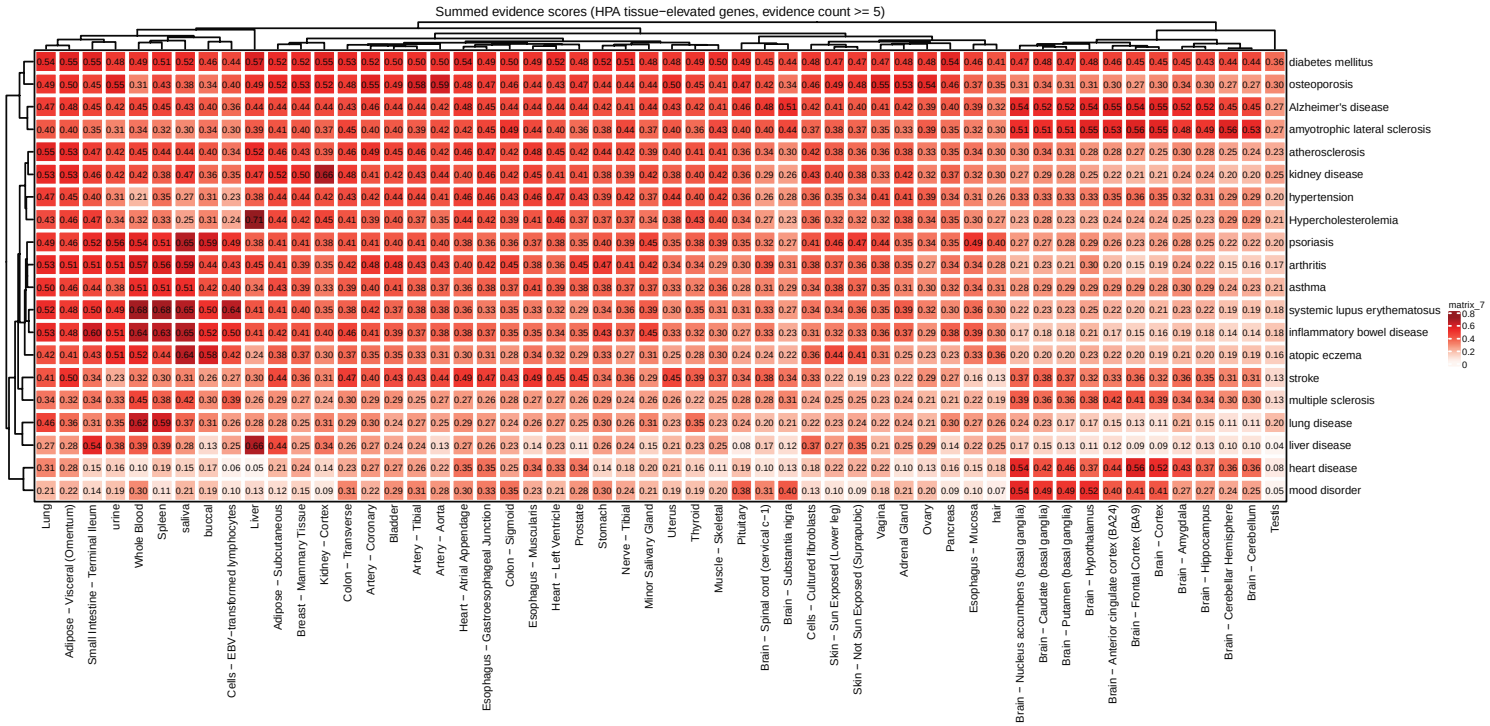
b. Reference ratio breakdown per VEP annotation.



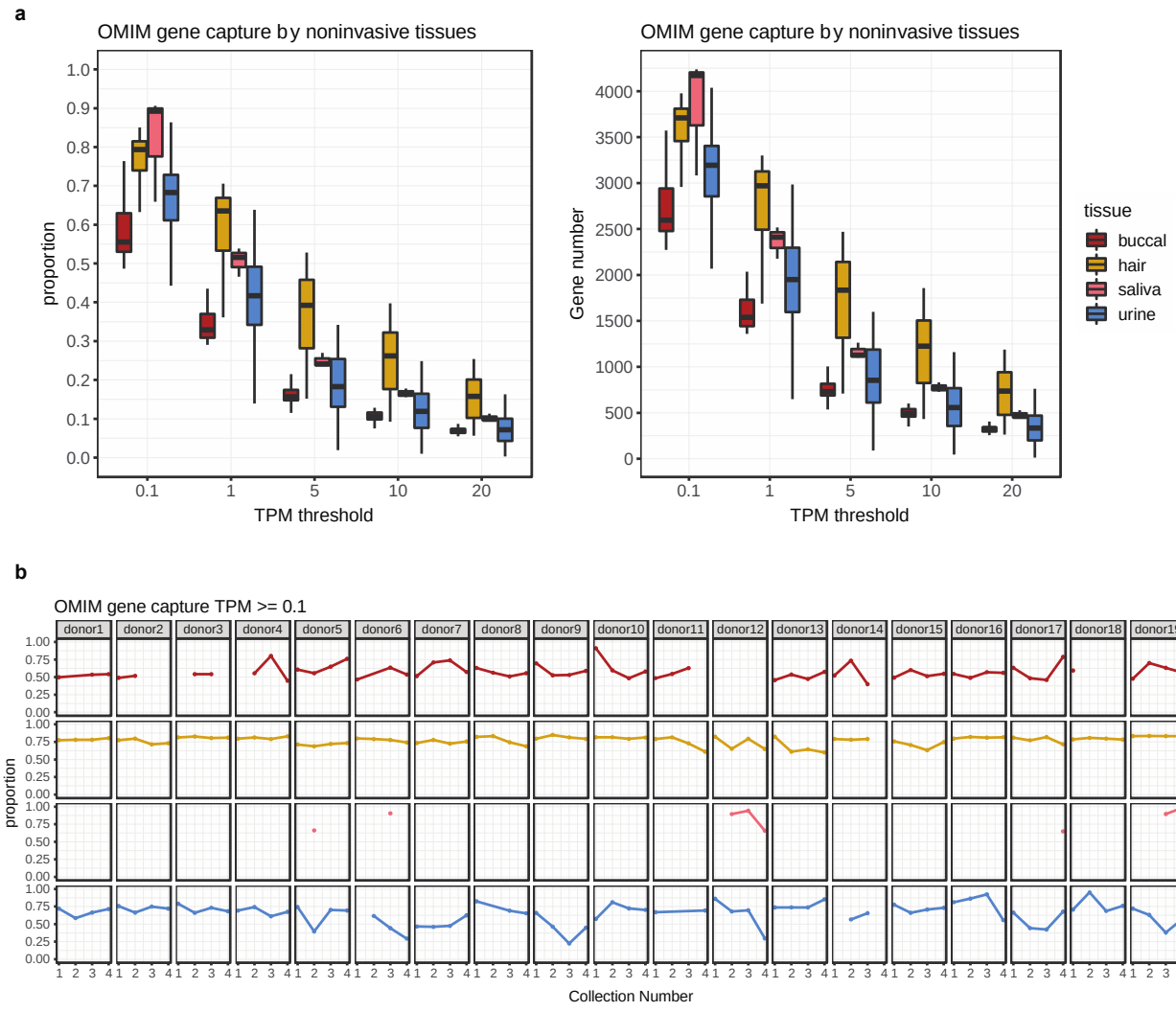
Supplementary Figure 12. a. Total number of disease-relevant genes per OpenTargets ontology category. Genes with >= 5 separate sources of evidence were included in the final analysis. **b.** Number of genes per tissue following minimum expression level thresholding and overlap with the HPA tissue-elevated gene list. The top 3,411 most expressed genes per tissue were included in the analysis. **c.** Summed evidence scores (SESs) for all GTEx and noninvasive tissues.


a
Number of Disease−relevant Genes
Genes
Alzheimer's disease
arthritis
amyotrophic lateral sclerosis
asthma
atherosclerosis
atopic eczema
diabetes mellitus
heart disease
Hypercholesterolemia
hypertension
inflammatory bowel disease
kidney disease
liver disease
lung disease
mood disorder
multiple sclerosis
osteoporosis
psoriasis
stroke
systemic lupus erythematosus
Number Genes Passing >= 5 evidence sources
b
Number of Genes Expressed after Filtering
Raw cts > 8 and tpms >0.1 in 20% of samples
genes
Adipose − Subcutaneous
Adipose − Visceral (Omentum)
Adrenal Gland
Artery − Aorta
Artery − Coronary
Artery − Tibial
Bladder
Brain − Amygdala
Brain − Anterior cingulate cortex (BA24)
Brain − Caudate (basal ganglia)
Brain − Cerebellar Hemisphere
Brain − Cerebellum
Brain − Cortex
Brain − Frontal Cortex (BA9)
Brain − Hippocampus
Brain − Hypothalamus
Brain − Nucleus accumbens (basal ganglia)
Brain − Putamen (basal ganglia)
Brain − Spinal cord (cervical c−1)
Brain − Substantia nigra
Breast − Mammary Tissue
buccal
Cells − Cultured fibroblasts
Cells − EBV−transformed lymphocytes
Colon − Sigmoid
Colon − Transverse
Esophagus − Gastroesophageal Junction
Esophagus − Mucosa
Esophagus − Muscularis
hair
Heart − Atrial Appendage
Heart − Left Ventricle
Kidney − Cortex
Liver
Lung
Minor Salivary Gland
Muscle − Skeletal
Nerve − Tibial
Ovary
Pancreas
Pituitary
Prostate
saliva
Skin − Not Sun Exposed (Suprapubic)
Skin − Sun Exposed (Lower leg)
Small Intestine − Terminal Ileum
Spleen
Stomach
Testis
Thyroid
urine
Uterus
Vagina
Whole Blood
Number of Genes Expressed after Filtering
Overlap with HPA tissue elevated genes

c



Supplementary Figure 13. a. Proportion and total OMIM gene capture per noninvasive tissue type, depending on minimum median TPM threshold. **b.** Proportion of OMIM gene capture per collection, donor, and tissue using a minimum expression threshold of 0.1 TPMs. **c.** Clustering of noninvasive and select GTEx tissues based on median OMIM gene expression.


a
OMIM gene capture by noninvasive tissues
proportion
TPM threshold
OMIM gene capture by noninvasive tissues
Gene number
TPM threshold
tissue
buccal
hair
saliva
urine
b
OMIM gene capture TPM >= 0.1
donor1
donor2
donor3
donor4
donor5
donor6
donor7
donor8
donor9
donor10
donor11
donor12
donor13
donor14
donor15
donor16
donor17
donor18
donor19
buccal
hair
saliva
urine
proportion
Collection Number
c
Median OMIM gene expression
Whole Blood
Muscle − Skeletal
Skin − Sun Exposed (Lower Leg)
Hair
Cells − Cultured Fibroblasts
Artery − Tibial
Cells − Ebv−Transformed Lymphocytes
Nerve − Tibial
Lung
Thyroid
Testis
Brain − Cerebellum
Brain − Cortex
Kidney − Cortex
Urine
Saliva
Spleen
Esophagus − Mucosa
Buccal

Supplementary Table 1.1. Breakdown of reaction cost (as of August 2022) per reagent using our in-house method (Loseq) versus TruSeq Stranded mRNA Library Prep and Takara SMART-seq V4 commercial kits. The cost of Ampure XP beads (cat# A63881) is excluded, but it is notably lower per reaction for Loseq and SMART-seq preparations due to smaller volume requirements.

Catalog Number	Reagent Name	Cost per reaction
Loseq		
50-196-5299	KAPA HiFi HS reaction mix	$1.38
30281-2	NxGen® RNAse Inhibitor	$0.69
EP0752	Maxima H Minus RT	$0.48
R0192	dNTP set (10mM each)	$0.14
	3' RT Primer (anchor) HPLC purified	$0.28
	drop-TSO HPLC Purification	$0.09
	drop-PCR HPLC Purification	$0.04
20027213	Nextera UD indexes	$1.75
FC-131-1096	Nextera XT DNA Library Preparation Kit	$8.95
61012	Dynabeads mRNA Direct Purification kit	$3.75
AM8170G	DNase I Buffer (10x)	$0.12
AM2224	DNase I (RNase-free)	$0.03
Illumina		
20020595	TruSeq Stranded mRNA	$96.46
20020591	TruSeq UD indexes	$5.71
SMART-seq		
634891	TakaraBio SMART-seq V4	$55.92

Supplementary Table 1.2. Comparison of total cost per reaction by library preparation. Loseq percent cost reduction calculated as: (1-(Loseq/Commercial kit))*100

Library Preparation	Total Cost per Reaction	Loseq Cost Reduction
Loseq	$17.70	
Illumina	$102.17	82.68% (-$84.47)
SMART-seq	$55.92	68.35% (-$38.22)

Chapter 2: Using noninvasive transcriptomics in a COPD clinical cohort

2.1 Introduction

SPIROMICS is a multicenter observational study of chronic obstructive pulmonary disease (COPD) intended to facilitate discovery and optimization of treatments[124]. Broadly, COPD involves persistent respiratory symptoms and airflow obstruction, and frequently cited etiological factors include smoking, pollutant exposure, and abnormal lung development[125]. Worldwide, 10% of people over 40 years of age have COPD, and it consistently ranks as a leading cause of death in the US[125]. Current clinically recognized major subtypes of COPD include emphysema, chronic bronchitis, and chronic obstructive asthma, which are generally defined in terms of structural changes, chronic cough, and chronic inflammation, respectively[125]. Diagnosis is determined on the basis of respiratory symptomatology, spirometric evidence of airflow obstruction (FEV/FVC < 0.7 post-bronchodilation), and a lack of alternative diagnosis[125]. At this time, there are no treatments shown to reduce mortality or disease progression, and most efforts are directed towards monitoring and symptom management[126]. Headway in this regard is complicated by the heterogeneous disease presentation. SPIROMICS aims to characterize COPD subtypes and biomarkers using both genetic and clinical data. The SPIROMICS study design is described in detail in Couper et al. 2014[124]. Briefly, 3200 subjects have been enrolled, which includes persons with varying severity of COPD as well as non-smoking controls. Various lung function, imaging, and exercise tolerance metrics were collected over four separate exams. In addition, RNA-sequencing of blood and whole genome sequencing was performed for all participants.

For our work, a subset of 1000 individuals will undergo collection and RNA-sequencing of hair follicle and buccal swab samples. COPD involves not only immune cell types in its etiology and pathophysiology, but the epithelial and endothelial cells of the lung play a major role in determining airway lumen size and thickness[127]. These features importantly determine the degree of airway obstruction that ultimately defines COPD and its severity[128]. Notably, reduced skin elasticity has been associated with worse spirometric measures of lung function, emphysematous pathology, and increased inflammatory biomarkers typical of COPD[129], suggesting the changes seen in the lung may be observed in other tissues across the body. Here, we hypothesize data collected from hair follicles and buccal swabs may capture epithelial and endothelial cell types and thus provide insight into genetic regulatory processes occurring in the lung tissue. This thesis examines the RNA-sequencing results from a pilot batch of 140 hair follicle and 110 buccal swab samples collected from seven clinical sites of the SPIROMICS cohort.

2.2 Results

2.2.1 Quality of clinically collected noninvasive samples

Results from the library preparation largely reflect observations from the 19 subject noninvasive sampling pilot. Figure 1a shows the distribution of unique reads mapping to protein coding and lncRNA genes per sample from each tissue type. The sequencing quality threshold of 10 million mapped reads is indicated. The validity of this approach in discriminating high and low quality samples is demonstrated in comparing RNA-sequencing metrics[90] of failed and passed samples (Figure 1b). Overall, buccal displays poor performance, with only 29 samples remaining after library preparation (28 failed) and sequencing (53 failed) quality filters (Figure 1c). Notably, the hair and buccal samples from the Columbia University (CU) site were prepared

using the Illumina Truseq V2 kit, whereas all other samples were prepared using the in-house, low input, low cost library preparation described in *Chapter 1*. Despite showing similar results regardless of preparation for the pilot study, hair performs robustly and consistently for the low input preparations and quite poorly using kits with higher starting material requirements. This is supported by the observation that failed samples tend to yield less RNA (Figure 1d). In terms of RNA quality based on RIN, it appears there is no passing trend depending on RIN score but that only a minimum quality is required for samples to pass downstream filtering (Figure 1d). It should also be noted that collection procedures at CU underwent further clarification after processing this pilot batch. During RNA extraction it was seen that some CU samples contained fewer than 5 follicles despite a 10 follicle minimum indicated in the collection procedure, and this is reflected in the highly variable and lower RNA yield, particularly for failed samples, from the CU site (Figure 1e). Of the 43 hair samples originating from CU, only 4 ultimately pass all quality filters (Figure 1c). For the remaining 97 hair samples from other sites, only 3 fail to pass quality standards. In all, 29 buccal and 98 hair samples are included in downstream analyses.

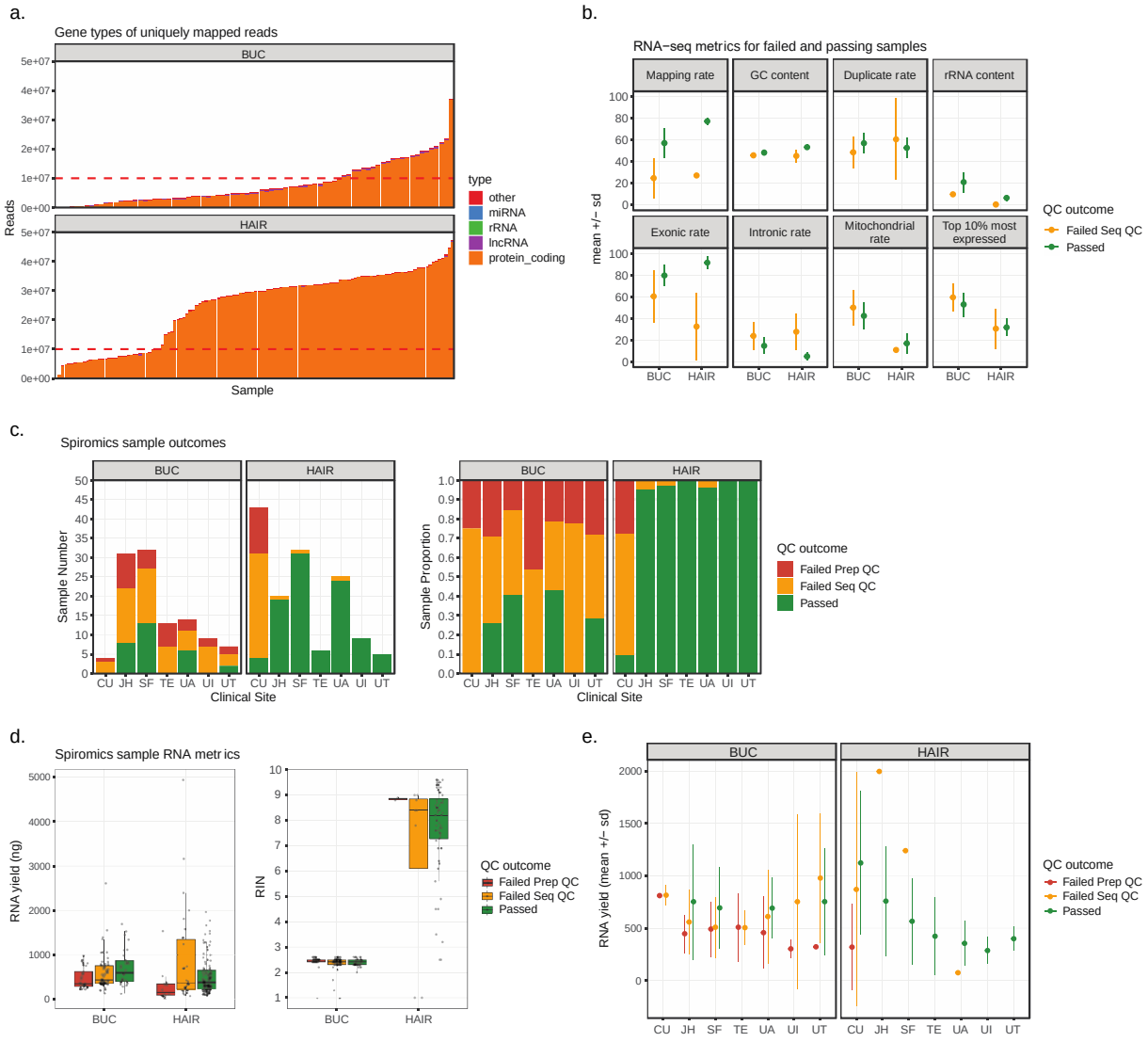


Figure 1. Hair shows high quality and performance across multiple clinical sites if low input RNA-sequencing library preparations are used. a. Number of uniquely mapped reads aligning to each genomic feature per sample as referenced in Gencode v35. The QC cutoff is indicated by the dashed red line. **b.** RNA-SeQC metrics for samples passing and failing sequencing QC. **c.** Total number (left) and proportion (right) of samples passing, failing sequencing QC, or failing library preparation QC per each clinical site. **d.** Summary of RNA yield and quality per sample using a BioAnalyzer. **e.** Differences in RNA yield across sites and its impact on sample performance.

2.2.2 Cell type deconvolution estimates show strong replication

Using GEDIT and the provided BlueCodeV1.0.tsv reference matrix[96], cell type proportion estimates were generated for buccal and hair follicle samples. The cell types

contained in this reference were collapsed into broader cell type categories (Figure 2a). From this, it is apparent buccal swabs express neutrophil cell type signatures most abundantly, and hair is predominantly composed of epithelial cells (Figure 2b). Buccal swabs generally show evidence for a greater abundance of immune related cell types, whereas hair shows a mix of immune and stromal cells. Overall, cell type proportion estimates for buccal swabs show greater variance, particularly for the most prevalent cell types, compared to hair. The consistency in cell type estimates for hair is further demonstrated by comparing biological replicates of the hair samples, collected from the same individual (Figure 2c). Notably, capture of cell type expression signatures is consistent for both highly and lowly present cell types. Unfortunately, insufficient buccal samples passed quality thresholds in order for biological replication to be assessed.

This same approach was used to deconvolve the noninvasive samples from the 19 subject pilot. As such, proportion estimates may be compared across datasets to assess whether the biology captured in noninvasive tissues is consistent. Significant deviations could indicate either a systematic difference in collection and processing of the samples or a difference in biological signals relating to the health of the individuals recruited. Here, we see strong agreement across the datasets, demonstrating global patterns of gene expression and cell type deconvolution estimates from this metric remain consistent for noninvasive tissues (Figure 2d). For buccal, we observe similarly high variance in neutrophil and epithelial cell estimates in both datasets. Hair is remarkably consistent. These results suggest that clinically scaling noninvasive sample collections, even across many clinical sites, will robustly capture similar biology.

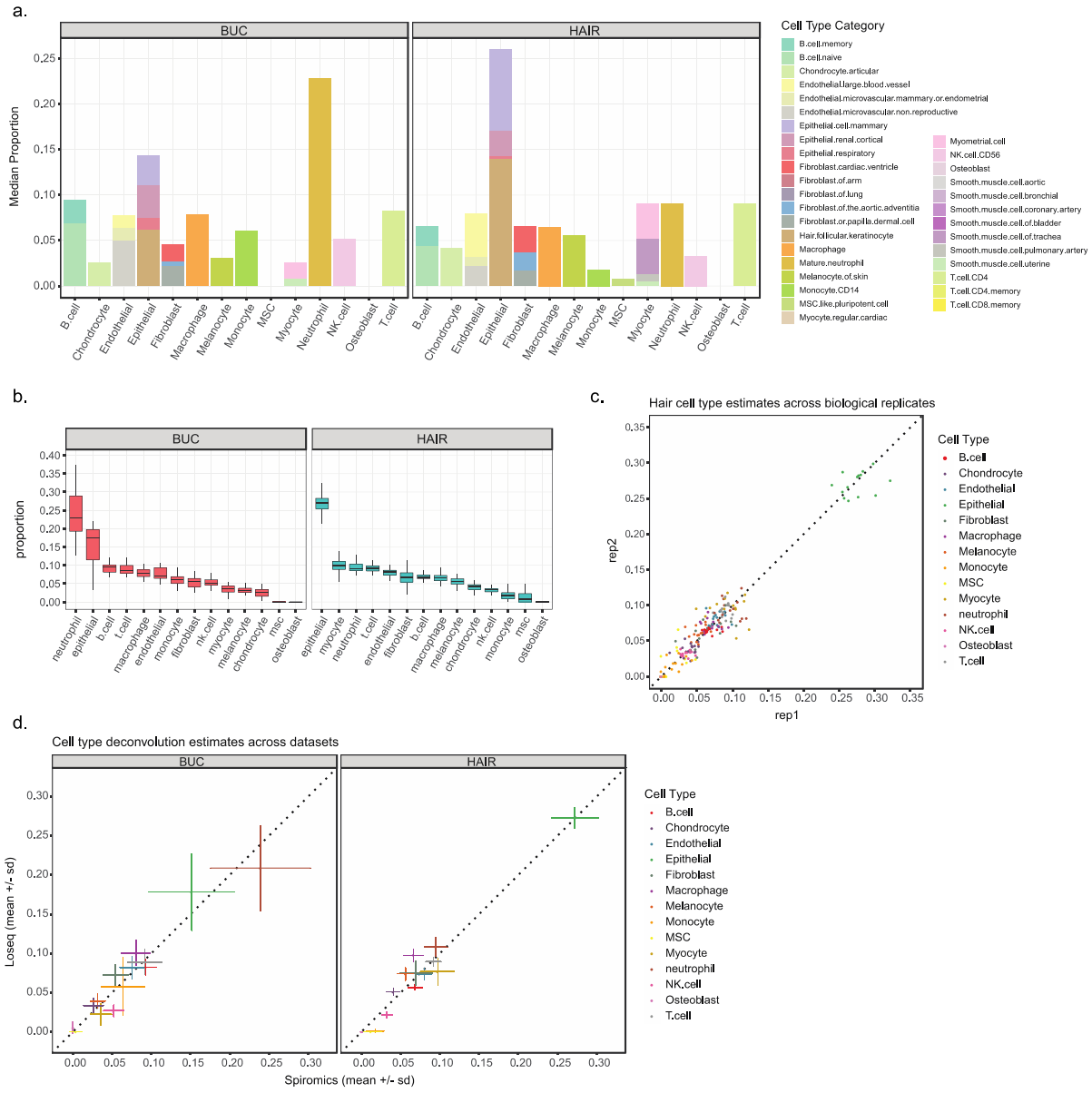


Figure 2. Buccal and hair follicles contain different cell type signatures that replicate within and across datasets. a. Highly specific BlueCode v2 cell type categories are collapsed into broad categories per sample. The median proportion across all samples per tissue type is shown. **b.** Distribution of cell type abundances per each noninvasive sample type. **c.** Comparison of cell type proportion estimates for samples collected from the same individual. The x = y line is shown. **d.** Comparison of cell type proportion estimates between the SPIROMICS and 19 subject pilot study (Loseq).

2.2.3 Discovery of cis-eQTLs in hair

Next, we aimed to investigate genetic variants affecting gene expression in cis in hair and buccal samples. First, genotyping data underwent standard processing. This includes removing monomorphic alleles, alleles with MAF <= 0.05, multiallelic sites and indels. From this, approximately 6.2 million variants were retained for the analysis. GTEx, which imposes the same filters except uses a MAF threshold of 0.01, keeps approximately 10 million variants for QTL testing. To establish ancestry covariates necessary for the analysis, we merged our genotyping data with 1000 genomes, performed LD pruning with PLINK[103], and ran smartpca using eigensoft. As anticipated, we observed our samples contain individuals of both European and African ancestries (Figure 3a). We see that genotyping PC1 correlates strongly with these superpopulations in 1000 genomes, and we see PC3 corresponds with a subpopulation (Figure 3b). From this result, we included genotyping PCs 1-3 as covariates in our model.

To prepare the expression data for eQTL analysis, the data was split by tissue and all replicates were removed, leaving 27 buccal and 80 hair samples. Next, the expression matrix was TMM normalized[105]. Counts were then filtered for genes with raw counts >= 6 and TPMs >= 0.1 in at least 20% of samples for a given tissue and inverse normal transformed per gene (as is the GTEx standard). To account for unknown batch effects in the expression data, we performed PCA on this processed expression data (Figure 3c). The Buja and Eyuboglu (BE) algorithm was then used to determine the number of PCs that explain more variance in the data than expected by chance[130,131]. In all, we included 3 genotype PCs, 7 expression PCs, sex, and RNA-sequencing library preparation method as covariates for hair, and reduced the number of expression PCs to 3 for buccal. Notably, expression PC2 in the hair data shows 4 samples segregating from the rest of the samples, and these are the 4 samples from the CU clinical site.

Thus expression PC2 and library preparation are likely collinear. This may result in wider confidence intervals for the eQTL effect size estimates that could affect significance calculations. eQTL mapping was performed on a per tissue basis using TensorQTL v.1.0.5[38] with the window set to 1MB (following the GTEx parameters).

This analysis resulted in 339 significant eVariant-eGene pairs for hair and no significant hits for buccal. Similar to the 19 donor pilot study, the buccal samples show much more variability across samples and the sample size used is relatively small, thus limiting our power to detect eQTLs. Even though we find 339 pairs for hair, this number is strikingly low in comparison to tissues of similar sample size. For instance, there are 89 kidney cortex samples in GTEx and from this 1260 significant eVariant-eGene pairs were discovered. This result remains consistent and captures nearly the same set of eGenes even if fewer covariates are included or PEERs are used instead of expression PCs. To ensure there are no issues with the genotyping data, we compare the number of heterozygous and homozygous individuals carrying the alternative allele to GTEx tissues. We find no meaningful difference, suggesting low sample sizes of the alternative allele are not causing more noise and less power relative to GTEx (Figure 3d). It is possible including the 4 samples from the CU site affects the normalization of the gene expression data. TMM and inverse normalization should remove the effects of sample depth and RNA composition differences, but the expression PCA shows a lot of structure remaining in the data (Figure 3c). Repeating normalization procedures and the PCA when excluding the CU site samples results in data with less structure and may be necessary for final analyses (Figure 3e). Because we do observe consistent and overlapping results regardless of our approach, we believe the eGenes we do find to be real signals but lingering problems with data processing are impacting our power to detect more eGenes.

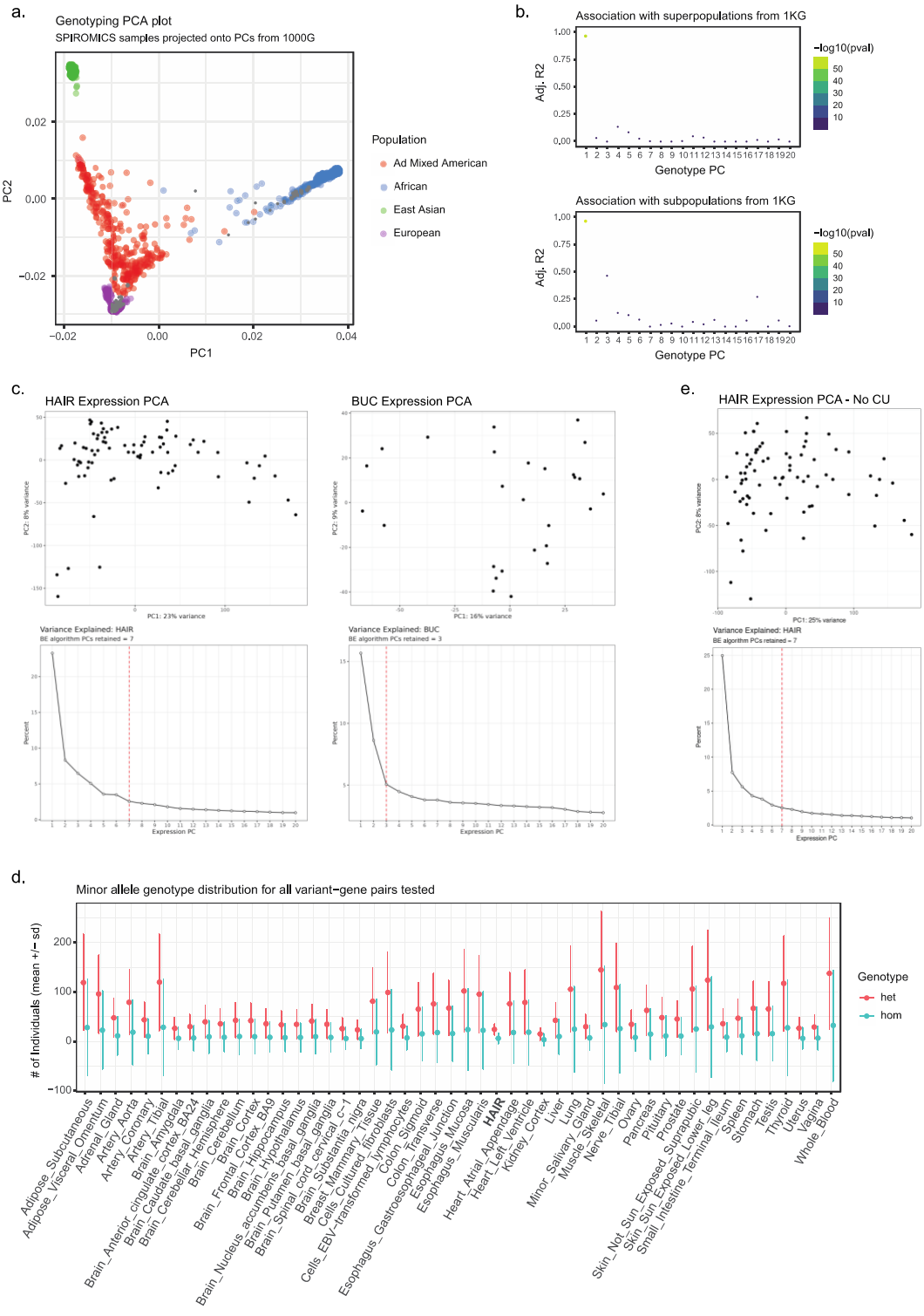


Figure 3. Establishing covariates for hair follicle and buccal eQTL analysis. a. Genotyping data from SPIROMICS samples projected onto 1000 genomes post-processing. **b.** A linear model (PC ~ populations) was used to assess association of top 20 genotyping PCs with 1000 genomes populations. **c.** PCA of TMM and inverse normalized expression data prior to eQTL analysis. The number of PCs retained as covariates

is indicated by the red dashed line on the scree plot. **d.** Comparison of number of heterozygous and homozygous individuals in GTEx tissues vs SPIROMICS hair. **e.** Repeating PCA of TMM and inverse normalized expression data with CU outlier samples removed.

2.2.4 Replication of hair cis-eQTLs in GTEx

Using our current findings, we investigate replication of hair cis-eQTLs in GTEx tissues. First, we observe ~80% of our eVariant-eGene pairs overlap with pairs found in GTEx (Figure 4a). Differences in the number of overlaps likely reflects differences in gene expression filtering such that certain genes are not tested in a given tissue. There is not full overlap because the variants used in each dataset are likely slightly different.

To determine replication of our signal, we use the π_1 statistic[39]. This measure approximates the proportion of true positives in the data. Because the null p-value distribution is uniformly distributed, the proportion of anticipated truly null findings, i.e. π_0, may be estimated by dividing the number of observed p-values above a threshold by the theoretically expected number of p-values exceeding that threshold. For a truly null distribution, π_0 will be equal to 1, and any leftward skewing of the p-value distribution towards significance will result in a smaller π_0 estimate. π_1 is calculated by taking 1-π_0. To calculate π_1 for a given GTEx tissue, the GTEx data is filtered for the significant eVariant-eGene pairs found in hair (Figure 4a). Because the GTEx p-value distribution is skewed towards significance, we calculate a null π_1 distribution by randomly sampling the overlapping number of variant-gene pairs 1000 times for each tissue (i.e. 271 pairs sampled 1000 times for kidney cortex, 246 pairs for whole blood, etc.). Figure 4b and 4c show that selecting for significant findings in hair results in much greater enrichment for significance compared to random sampling. We see the highest rates of replication in tissue types expected to be most similar to hair, like skin, esophageal mucosa, and muscle.

Because we observe replication in tissues likely containing similar cell types to hair, we test this hypothesis by comparing π_1 to cell type enrichment estimates in the GTEx tissues. In the prior section, we observe hair to be most abundant for epithelial cells and myocytes. Here, we see tissues enriched for epithelial cells, keratinocytes, skeletal muscle, sebocytes, and adipocytes all show higher replication of cis-regulatory effects, and tissues more enriched for immune cells or neurons demonstrate weaker replication (Figure 4d).

It is possible we see higher replication in certain tissue types for reasons unrelated to their biological similarity to hair. For instance, GTEx samples with higher sample size are more likely to capture a greater number of cis-regulatory effects and thus incidentally recapitulate our findings in hair. Indeed, as sample size increases we do see an increase in π_1 as well (Figure 4e). However, GTEx showed whole blood to be an outlier in its cis-regulatory architecture, and despite having the largest sample size in GTEx, whole blood does not result in the highest replication. This suggests hair shares regulatory mechanisms with tissues of the body on the basis of their underlying cell type compositions and not as a consequence of sample size alone.

Looking at effect size concordance, we observe generally underwhelming results. There does appear to be greater agreement for tissues with higher replication, but this is overall not a strong relationship (Figure 4f).

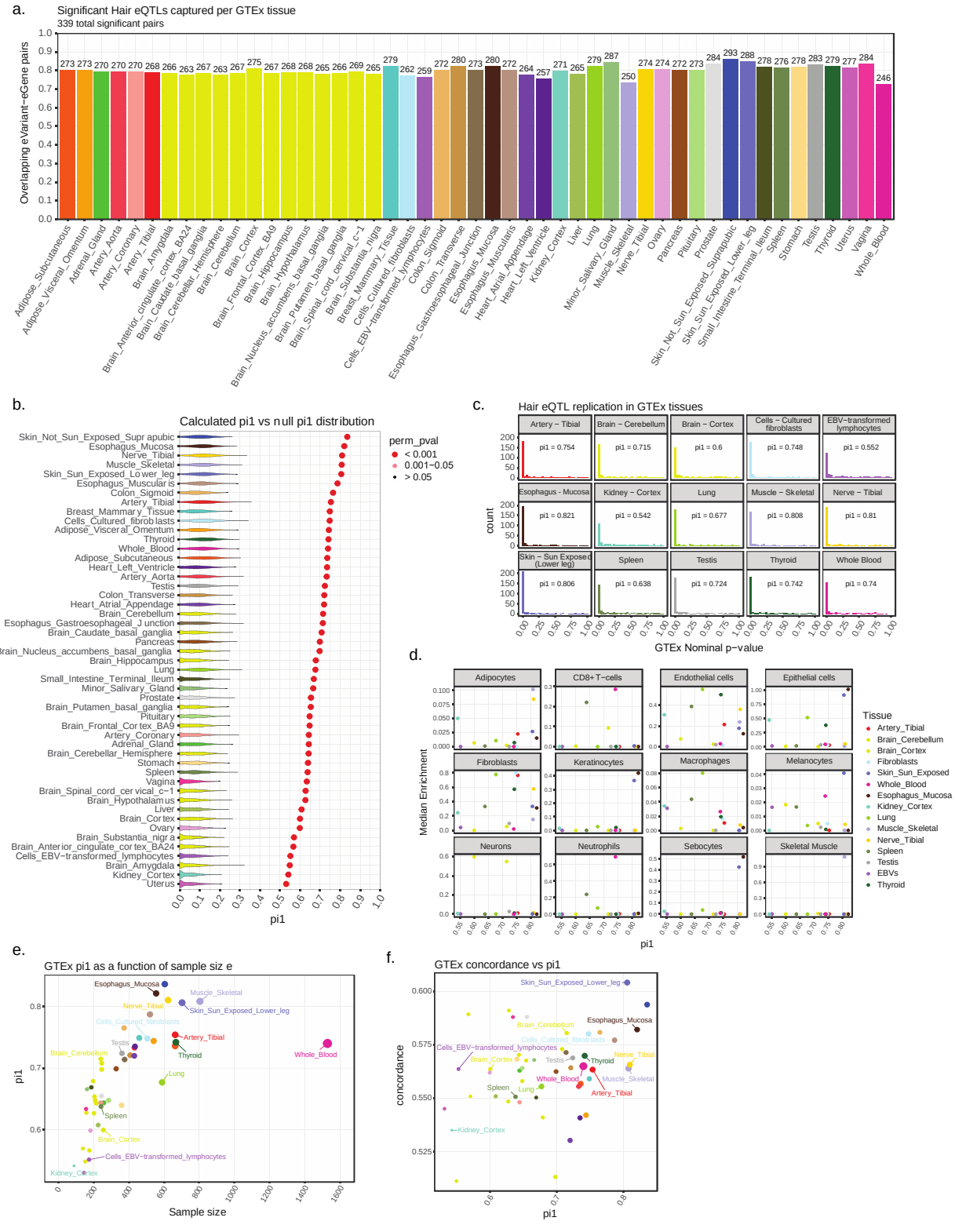


Figure 4. Hair eQTLs replicate strongly in GTEx tissues abundant for cell types found in hair. a. Significant hair eQTLs overlapping gene-variant pairs in GTEx tissues. **b.** pi1 results per GTEx tissue indicated by the datapoint with significance indicated by color and size (all are

significant). The null distribution is indicated by the violin plot. **c.** Histograms of GTEx p-values for significant hair cis-eQTLs for a subset of GTEx tissues. **d.** Median xCell enrichment scores per select GTEx tissues is compared to the pi1 replication metric. **e.** Relationship between pi1 and GTEx tissue sample size. **f.** Relationship between effect size concordance for hair cis-eQTLs with GTEx and pi1.

2.3 Discussion

Here, we sought to determine whether buccal swabs and hair follicles perform robustly and capture cell types and gene regulatory mechanisms relevant to the lung and COPD. Regarding quality, buccal swabs fail processing often and do not show promise for future scaling across many clinical sites. Hair follicles, on the other hand, show a high degree of consistency and promise for continued use if low input library preparations are used. Additionally, we not only observe consistency in quality but in the cell types captured from hair follicles, both within and across datasets. Epithelial cell types are predominantly present in hair follicles, and as stated at the outset, these cells may play a key role in COPD pathophysiology and thus hair may be useful for studying this disease. Regarding gene expression regulation, we discover 339 eQTLs in hair follicles. These regulatory mechanisms replicate strongly across all GTEx tissues, but especially so in tissues enriched for cell types found in hair follicles. However, we do not observe particularly strong concordance in effect size with GTEx tissues.

Next steps will primarily focus on further optimizing eQTL discovery in hair follicles. Given our sample size, the number of eQTLs discovered is notably low, and this suggests underlying issues in discovery power potentially unrelated to the sample size. Most likely, the expression normalization is affected by the inclusion of outlier samples from the CU site, but this hypothesis will require further investigation.

Thus far applications to lung tissue and COPD remain largely unexplored. We first intend to investigate colocalization between hair and whole blood eQTLs with COPD GWAS hits and

compare these results to lung eQTL colocalization. This approach would help decipher whether hair identifies eQTLs with greater relevance to lung tissue and COPD in comparison to whole blood. It would also confirm whether hair may be used as a suitable proxy for studying genetic regulation in the lung. Other analyses include looking into whether COPD imaging phenotypes and severity correspond with differentially expressed genes in hair tissue, as well as exploring whether the microbial abundance in unmapped buccal swab reads relates to COPD exacerbations or severity.

In all, the results here confirm our expectations from the 19 donor pilot study in regards to quality and scaling of hair follicles and the potential limitations of buccal swabs. Here we apply these noninvasive tissue types to studying lung disease, however, there may be other diseases directly involving the skin or oral mucosa that may show even greater benefit. At this time, there is sufficient evidence that hair follicles contain cell types with potential relevance to lung tissue, however, further investigation of its application to COPD is required.

Epilogue

Conclusions

This thesis begins to address the various shortfalls of transcriptomic and genomic research by proposing alternative sampling and processing methodologies of RNA-sequencing biospecimens. As stated at the outset, discovery remains limited by the complex genetic features underlying traits and by the current transcriptome study paradigm. Here, noninvasive, low-cost RNA-sequencing is proposed as a potential solution for augmenting current efforts to understand the genetic code. For one, the cost of noninvasive sampling relative to other procedures is greatly reduced by the lack of highly trained healthcare professionals and resources required. This change would enable massive scaling of studies, particularly in non-European populations that are disproportionately understudied and face the greatest disparities in healthcare. Biologically, noninvasive sampling may capture cell types not assayed by current collection methods, i.e. blood, which may lend insight into invasive tissues composed of non-blood related cell types. At this time, serial sampling of pertinent tissues remains infeasible due to the types of samples collected (surgical biopsies and post-mortem donations). This results in a loss of valuable, context specific genetic regulatory information that may be key to understanding the genesis of traits and disease. In all, noninvasive sampling may provide a means for capturing highly relevant biology and for closing the gap in research access across communities. This thesis aimed to interrogate the validity of this hypothesis.

Chapter 1 investigated four potential noninvasive tissue types (buccal swabs, hair follicles, saliva, and urine cell pellets) as well as a low cost library preparation method. This section explored general features and technical considerations for each tissue, the feasibility and consistency of repeated sampling, cell types contained in each tissue and their similarity to

invasive tissues, their ability to capture genetic regulatory mechanisms found invasively, and finally, general application to disease-relevant analyses.

Of the tissues studied, hair follicles and urine cell pellets bear the greatest promise for future use. This finding was demonstrated largely by library preparation outcomes. Buccal swabs and saliva, for the most part, result in poor performance due to microbial contamination. However, the microbiome is increasingly recognized to potentially play a large role in human health, and the ease of buccal swab and saliva collections may provide a future avenue for associating microbial changes with disease status or severity. Hair shows successful processing and high quality RNA-sequencing metrics regardless of preparation method, and urine is largely successful if low input methods are used. Downstream analyses showed hair to be extremely uniform in cell type composition estimates across collections and individuals. This low variability contributes to its replication of GTEx signals despite a very low sample size. For Mendelian disease, hair captures gene sets not found in blood but that are found in other invasive tissues. Also, hair robustly captures stop-gain allele specific expression. This suggests hair follicle collection may be a useful tool for providing genetic diagnoses of rare disease and may reduce wait times as well as the invasiveness of procedures required to establish a diagnosis. For urine, resolving cell type composition is a potential caveat for its future use because it is highly variable depending on collection and the individual sampled. Nonetheless, the ease of incorporating urine collections into clinical settings provides a strong impetus for its future use. Urine shows replication for kidney cortex eQTLs, among other tissues, as well as enrichment for kidney disease genes found in OpenTargets. Due to the high morbidity and mortality of kidney disease, and the role of the kidney in many other disease processes, urine holds great promise for yielding clinically relevant discovery. Overall, the findings from chapter 1 suggest noninvasive

sampling may be a promising approach for future scaling and discovery in transcriptome studies, but tissue-specific considerations may need to be explored and accounted for prior to their use. All of the tissues studied here were able to be collected by personnel, i.e. me, without a high degree of medical training beyond the specifics of the tissue collection procedures. It should be noted that buccal swabs, saliva, and urine cell pellets require the least time and training for collection, and thus more closely meet the original goal of reduced specialization and training of collection procedures. Hair collections vary depending on hair texture and strand thickness, and the ancestry populations studied in this pilot were largely European and Asian. Thus it is not able to be concluded whether hair follicle collections are readily scalable across all populations from this study alone. To summarize, chapter 1 met many of the initial goals of our original hypothesis and established both technical and biological features of prospective noninvasive samples that will guide future use of these biospecimens as well as a framework for investigating other sample types.

Chapter 2 explores both the feasibility of scaling noninvasive sampling and their potential disease applications by testing buccal swab and hair follicle sample collections in a COPD clinical cohort. Much of what was noted in the pilot study was recapitulated in this work. First, buccal swab samples show low success in processing and in sample quality. From this study, it seems clear the ease of collection does not outweigh the lost cost of processing failed samples. Even though buccal swabs passing quality thresholds do seem to provide meaningful data regarding tissues of the gastrointestinal tract, broader use of this tissue is not cost effective and their use should be reserved for curated scientific aims. Despite seeing robust processing of hair regardless of library preparation in the pilot, we do observe a very high rate of failure for kits requiring higher initial RNA input in this clinical study. However, for the low input preparations

we observe exceptionally high success rates across all clinical sites. This suggests future studies should largely use low input methods for processing noninvasive samples to ensure success because the quantity of material and thus yield of RNA are typically lower than traditional tissue types. This also suggests that despite any perceived difficulty in pursuing hair follicle collections, in practice, this method is able to be scaled and executed across many clinical sites without compromising final sample quality. The SPIROMICS clinical cohort contains individuals from a multitude of ancestries compared to the pilot study in Chapter 1, and therefore differences in hair type likely do not affect sample collection to a meaningful degree. Furthermore, the cell type signatures captured in hair remain consistent across replicates within the study and also across studies. Importantly, the larger sample size in Chapter 2 allowed us to discover cis-eQTLs and investigate their replication in invasive tissue types. We did observe higher replication of these signals in invasive tissues that share cell types with the noninvasive samples, and this lends support to our original hypothesis that noninvasive samples may facilitate discovery of invasive tissue processes by capturing shared cell types. Future investigations will interrogate whether these loci primarily localize to disease-relevant variants and gene pathways. Altogether, these findings lend support that noninvasive sampling does scale, may indeed be used to narrow sampling gaps across populations in genomics, and are amenable to transcriptomic analyses for disease applications.

Future Directions

An immediate potential benefit provided by noninvasive sampling is the improved rate and ease of Mendelian disease diagnosis. Using exome sequencing or gene panels results in a diagnosis only ~50% of the time[132]. Whole genome sequencing is often uninformative because, much like GWAS, variant consequences remain unknown. Transcriptomics has thus been put

forth as a potential option for the remaining individuals without a diagnosis. However, similar to common disease, blood samples often do not capture cell types and thus gene expression necessary to determine a diagnosis[132,133]. In our pilot study, hair and urine biospecimens largely capture the same Mendelian gene sets as more invasive tissue types. We also observe hair captures genes not found in blood and vice versa. Others have similarly observed higher capture of disease-relevant genes in skin biopsies and other samples more closely related to the tissues affected by disease[132,133]. However, expanding the use of transcriptomics in Mendelian diagnosis is limited by lack of access to affected tissues (and therefore invasive biopsies are often required), and clinical pipelines for collection, processing, and analysis of this data have yet to be established[133]. Hair and urine cell pellets provide a considerable advantage by allowing access to pertinent cell types without the need for invasive testing, and these samples are amenable to low cost library preparations that remove financial barriers when attempting to bring clinical testing to scale. Overall, transcriptomic diagnosis of Mendelian disease may be facilitated by use of noninvasive sampling and could improve diagnosis rates and reduce diagnostic delay.

In the chapters of this thesis, longitudinal sampling is proposed and the consistency of sample collections is quantified. However, none of our analyses were able to directly capitalize on the full utility and benefit of multiple sample collections. Much like other clinical metrics collected over the course of care, like bloodwork and imaging, noninvasive sampling provides another data point by which many of the most basic biological processes of the body, e.g. development, pregnancy, and aging, may be understood. If the cell types relevant to a disease are directly sampled, then differences across individuals during these time periods may illuminate gene pathways central to disease processes emerging at that time or during a different stage of life. Even so, diseases often manifest systemically, and even if the tissues and cell types most

afflicted are not sampled from and thus the data gleaned may not reveal primary disease mechanisms, there is opportunity to identify perturbations that may still segregate patients according to genetic predisposition, clinical subtypes, or treatment responsiveness. Generally, continued transcriptome sequencing and follow-up in individuals suffering from progressive, currently untreatable diseases could reveal potential treatment options by providing greater information as to which gene pathways and bodily processes are cause and consequence of disease. Noninvasive transcriptomics could also be used in the context of known, high-risk environmental exposures to better understand protective and predisposing factors. As stated in Chapter 2, pollutants are a major risk factor and cause of COPD[125], and COPD development and progression may be better understood by sampling from the nasopharynx in individuals with varying pollutant exposure. And finally, our understanding of pharmacological treatment success or failure could be improved by sampling before and after the initiation of intervention and during treatment. This same approach could be used to monitor the effectiveness of lifestyle changes. In all, noninvasive sampling provides flexibility to study designs such that these and other questions of this type may be interrogated.

Throughout, this thesis has alluded to the potential use of other noninvasive biospecimens. Box 1 provides a subset of future sample types and applications. An immediately obvious set of samples for follow-up study includes cervical swabs and cervical brushings. In general, female reproductive healthcare lacks adequate diagnostic and treatment tools. For instance, pelvic pain affects 6-25% of reproductive age females, but the diagnostic toolkit and treatment options are astoundingly feeble[134]. Blood work typically only excludes other causes and many current biomarkers (i.e. CA-125) are nonspecific to a single diagnosis. Imaging studies are typically only diagnostic if disease is severe, and for long-standing, severe pelvic pain or for

suspected endometriosis (which affects 10% of females), definitive diagnosis is only achievable via exploratory laparoscopic surgery[134]. Treatment options are similarly limited and typically include over the counter analgesics and birth control[134]. This dearth of clinical options for females also exists in regards to recurrent miscarriages, polycystic ovarian syndrome, and other undetermined causes of infertility. Thus collection of a sample type directly from this region of the body may lend key insights into the unique features of diseases affecting reproductive organs that allow for the development of both improved diagnostic and treatment tools. Notably, the cells collected from cervical brush samples are already preserved in such a way to maintain their integrity for laboratory analysis, and this could easily be adapted for downstream single cell RNA-sequencing. Similar to hair follicles and urine cell pellets, the cell types collected could lend opportunity for use in contexts unrelated to female reproductive health. Glandular cells and squamous and columnar epithelial cells exist across many tissue types of the body and play a role in a multitude of diseases but are not assayed in our current approaches. In the future, it may be beneficial to simultaneously collect multiple, different noninvasive sample types in order to capture the broadest array of cell types available and deepen our understanding of genetic mechanisms underlying disease processes.

Box 1. Potential noninvasive biospecimens and their applications.

Noninvasive Sample	Anticipated Cell Types	Potential Disease Applications
Nasal Brushings	Epithelial (Squamous), Immune	Chronic lung disease[127]
Cervical Brushings	Glandular, Epithelial (Columnar and Squamous)	Female reproductive health[134]
Cervical Swabs	Epithelial, Immune	HPV host-response[135–137]
Menstrual blood	Endometrial stem cells, Epithelial, Immune	Mesenchymal stem cell therapy (Duchenne's, Acute liver or lung injury)[138]
Semen	Spermatozoa (and progenitors), Epithelial, Immune	Male reproductive health[139]
Fecal	Epithelial, Immune	Colorectal cancer, IBD[140]
Breast milk	Breast milk and Mesenchymal stem cells, Epithelial, Smooth muscle, Immune[141]	Breast cancer[142]

Despite the aforementioned potential utility of noninvasive samples in clinical research or monitoring, the use and benefit of genomic and transcriptomic data in clinical applications remains underutilized and largely unproven. Thus far, -omics data has provided gains in regards to Mendelian disease[132], pharmacogenetics[143], and for diseases where inheritance of specific alleles carry a large genetic risk. Current limitations for Mendelian applications are delineated above. Pharmacogenomics faces major challenges for implementation due to a lack of evidence-based treatment algorithms[143]. There are very few randomized controlled trials testing the

outcomes of pharmacogenetically guided clinical decision-making, and without this information, proper utilization of this genetic data requires expertise in the field. Furthermore, no standard collection, processing, and reporting format has been created, and this prevents easy incorporation of pharmacogenetic testing into clinical pipelines[143]. Because of these barriers, very few healthcare professionals use this technology despite its anticipated benefits[143]. For diseases with high risk alleles, like Celiac's disease, genetic testing is often conclusive if symptoms are present, but in asymptomatic individuals it is nondiagnostic and generally uninformative because the majority of carriers do not have the disease[144]. Thus genetic testing is often reserved for situations where a diagnosis is suspect but cannot be achieved through other means, and it suggests using genomic data preemptively for complex disease risk prediction using low effect size alleles may not usefully stratify patients. Given the complications for clinical use arising from genomics data for relatively uncomplicated disease and treatments, it seems unlikely that genomic and transcriptomic data will provide clinical benefits for complex disease diagnosis and treatment unless research goals incorporate practical considerations necessary for their eventual use.

The pitfalls of prior efforts suggest there are several changes necessary in order to support effective clinical use of genomic and transcriptomic data. First, if -omics data will continue to be directly leveraged to inform diagnostic and treatment criteria, it should be generated such that minimal sample processing and low-pass sequencing is sufficient. To do so, improved, standardized, low-cost sequencing technologies that are easily operable by technicians outside of genomics must be developed and tested appropriately. As demonstrated by the shortcomings observed in clinical pharmacogenetics, a lack of access and standardization in these regards prevents wide scale implementation[143]. It should be said, there is additional risk of

worsening healthcare disparities if affordable avenues for accessing this data is not concurrently provided. At this time, only the wealthiest hospitals are equipped with the necessary resources to store, transport, and potentially process samples for RNA and DNA-sequencing. Thus increased use of polygenic or transcriptomic risk scores or machine learning models that rely on genomic or transcriptomic data to tailor diagnosis or treatment may further healthcare disparities because underfunded and under-resourced hospital systems are very unlikely to have continued access to genomic and transcriptomic sequencing, in their current form, for the patient populations they serve. To further ensure equitable access and broad implementation, the final outcome of -omics testing should be readily interpretable without expert consultation, and its use should be tested in sufficient RCTs such that the results are rigorously reliable and healthcare workers and hospitals are provided with necessary guidelines. In ideal circumstances, large scale -omics data will be reserved for research settings and the findings will be distilled into narrow gene pathways and/or more easily measurable biomarkers for clinical assessment. In all, if precise identification of causal disease mechanisms is not possible, then future -omics research should anticipate potential barriers to clinical use and prioritize clinical applications and technology that will ensure care is equitably provided to the most underserved populations.

To conclude, the field of genomics has undergone massive, transformative change in the last 20 years. Despite the immense progress made thus far, the connection between genotype and phenotype is deeply complex and requires a multitude of data types and flexibility of study designs to further our current understanding. Further, the cost of these analyses must be kept low to enable scaling across currently understudied populations. This thesis explored low cost, noninvasive RNA-sequencing as one potential solution to these limitations and demonstrated its promise in addressing key issues faced by the field.

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
3. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
4. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
5. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
6. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* (2021) doi:10.1038/s41576-020-00305-9.
7. Hill, M. S., Zande, P. V. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215 (2021).
8. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. 1–23 (2019) doi:10.1101/588020.
9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 31–21 (2014).
10. Van den Berge, K. *et al.* RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.* **2**, 139–173 (2019).

11. Alpern, D. *et al.* BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* 1–15 (2019) doi:10.1186/s13059-019-1671-x.
12. Kamitani, M., Kashima, M., Tezuka, A. & Nagano, A. J. Lasy-Seq: a high-throughput library preparation method for RNA-Seq and its application in the analysis of plant responses to fluctuating temperatures. *Sci. Rep.* **9**, 7091 (2019).
13. Gaio, D. *et al.* Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microb. Genomics* **8**, 000744 (2022).
14. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).
15. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Publ. Group* **16**, 197–212 (2015).
16. Donovan, M. K. R., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 955–14 (2020).
17. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, (2020).
18. Jagadeesh, K.A., et. al. *Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics*. 1–86 https://www.biorxiv.org/content/10.1101/2021.03.19.436212v1.full.pdf (2021).
19. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).

20. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2016).
21. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
22. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
23. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
24. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
25. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
26. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. 2022.05.07.491045 Preprint at https://doi.org/10.1101/2022.05.07.491045 (2022).
27. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).
28. Olivieri, J. E. *et al.* RNA splicing programs define tissue compartments and cell types at single-cell resolution. *eLife* **10**, e70692 (2021).
29. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

30. Shan, N., Wang, Z. & Hou, L. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics* **20**, 126 (2019).
31. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
32. Marderstein, A. R. *et al.* Leveraging phenotypic variability to identify genetic interactions in human phenotypes. **357**, 444–36 (2020).
33. Oliva, M. *et al.* The impact of sex on gene expression across human tissues. *Science* **369**, eaba3066 (2020).
34. Kim-Hellmuth, S. *et al.* Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* 1–10 (2017) doi:10.1038/s41467-017-00366-1.
35. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
36. Aygün, N. *et al.* Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am. J. Hum. Genet.* (2021) doi:10.1016/j.ajhg.2021.07.011.
37. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
38. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
39. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445 (2003).
40. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet.* **10**, e1004383 (2014).

41. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, e1008720 (2020).
42. Barbeira, A. N. *et al.* Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
43. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
44. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
45. Connally, N. *et al.* The missing link between genetic association and regulatory function. *medRxiv* 2021.06.08.21258515 (2021) doi:10.1101/2021.06.08.21258515.
46. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
47. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
48. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
49. DNA Sequencing Costs: Data. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.
50. GTEx Standard Operating Procedures Library | Programs | BBRB. https://biospecimens.cancer.gov/resources/sops/library.asp.
51. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 1–25 (2019) doi:10.1038/s41586-018-0579-z.

52. Shavers-Hornaday, V. L., Lynch, C. F., Burmeister, L. F. & Torner, J. C. Why are African Americans under-represented in medical research studies? Impediments to participation. *Ethn. Health* **2**, 31–45 (1997).

53. Das, N. K. & Sil, A. Evolution of Ethics in Clinical Research and Ethics Committee. *Indian J. Dermatol.* **62**, 373–379 (2017).

54. Woodall, A., Morgan, C., Sloan, C. & Howard, L. Barriers to participation in mental health research: are there specific gender, ethnicity and age related barriers? *BMC Psychiatry* **10**, 103 (2010).

55. George, S., Duran, N. & Norris, K. A Systematic Review of Barriers and Facilitators to Minority Research Participation Among African Americans, Latinos, Asian Americans, and Pacific Islanders. *Am. J. Public Health* **104**, e16–e31 (2014).

56. Bailey, Z. D. *et al.* Structural racism and health inequities in the USA: evidence and interventions. *Lancet Lond. Engl.* **389**, 1453–1463 (2017).

57. Britton, A. *et al.* Threats to Applicability of Randomised Trials: Exclusions and Selective Participation. *J. Health Serv. Res. Policy* **4**, 112–121 (1999).

58. Luebbert, R. & Perez, A. Barriers to Clinical Research Participation Among African Americans. *J. Transcult. Nurs.* **27**, 456–463 (2016).

59. Dawson, S., Campbell, S. M., Giles, S. J., Morris, R. L. & Cheraghi-Sohi, S. Black and minority ethnic group involvement in health and social care research: A systematic review. *Health Expect.* **21**, 3–22 (2018).

60. Zeggini, E., Gloyn, A. L., Barton, A. C. & Wain, L. V. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science* **365**, 1409–1413 (2019).

61. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

62. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

63. Price, A. L., Spencer, C. C. A. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc. Biol. Sci.* **282**, 20151684 (2015).

64. Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021).

65. Supplitt, S., Karpinski, P., Sasiadek, M. & Laczmanska, I. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *Int. J. Mol. Sci.* **22**, 1422 (2021).

66. Wang, M., Herbst, R. S. & Boshoff, C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat. Med.* **27**, 1345–1356 (2021).

67. Docking, T. R. *et al.* A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.* **12**, 2474 (2021).

68. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, (2017).

69. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).

70. Mohammadi, P. *et al.* Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **366**, 351–356 (2019).

71. Almogy, G. *et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform*. http://biorxiv.org/lookup/doi/10.1101/2022.05.29.493900 (2022) doi:10.1101/2022.05.29.493900.
72. Schwarz, T. *et al.* Powerful eQTL mapping through low-coverage RNA sequencing. *Hum. Genet. Genomics Adv.* **3**, 100103 (2022).
73. Theda, C. *et al.* Quantitation of the cellular content of saliva and buccal swab samples. *Nat. Publ. Group* **8**, 6944 (2018).
74. Bondue, T. *et al.* Urine-Derived Epithelial Cells as Models for Genetic Kidney Diseases. *Cells* **10**, (2021).
75. Cheung, M. D. *et al.* Single-Cell RNA Sequencing of Urinary Cells Reveals Distinct Cellular Diversity in COVID-19-Associated AKI. *Kidney360* **3**, 28–36 (2022).
76. Latt, K. Z. *et al.* Urine Single-Cell RNA Sequencing in Focal Segmental Glomerulosclerosis Reveals Inflammatory Signatures. *Kidney Int. Rep.* **7**, 289–304 (2022).
77. Oliveira Arcolino, F. *et al.* Human Urine as a Noninvasive Source of Kidney Cells. *Stem Cells Int.* **2015**, (2015).
78. Bradley, M. S. *et al.* Urine RNA Processing in a Clinical Setting. *Female Pelvic Med. Reconstr. Surg.* (2017) doi:10.1097/spv.0000000000000525.
79. Manaph, N. P. A., Al-Hawaas, M., Bobrovskaya, L., Coates, P. T. & Zhou, X.-F. Urine-derived cells for human cell therapy. *Stem Cell Res. Ther.* 1–12 (2018) doi:10.1186/s13287-018-0932-z.
80. Ng, D. L. *et al.* A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood. *Sci. Adv.* **7**, (2021).

81. Ziegler, C. G. K. *et al.* Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell* **184**, 4713–4733 (2021).

82. Kim, S. J. *et al.* Gene expression in head hair follicles plucked from men and women. *Ann. Clin. Lab. Sci.* **36**, 115–126 (2006).

83. Herrera-Rivero, M., Hochfeld, L. M., Sivalingam, S., Nöthen, M. M. & Heilmann-Heimbach, S. Mapping of cis-acting expression quantitative trait loci in human scalp hair follicles. *BMC Dermatol.* **20**, 16 (2020).

84. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

85. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).

86. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

87. Picard Toolkit. (2019).

88. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).

89. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

90. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinforma. Oxf. Engl.* **28**, 1530–1532 (2012).

91. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

92. Sangiovanni, M., Granata, I., Thind, A. S. & Guarracino, M. R. From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics* **20**, 168 (2019).

93. Krassowski, M., Arts, M. & Lagger, C. krassowski/complex-upset: v1.3.3. (2021) doi:10.5281/zenodo.5762625.

94. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).

95. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

96. Nadel, B. B. *et al.* The Gene Expression Deconvolution Interactive Tool (GEDIT): accurate cell type quantification from gene expression data. *GigaScience* **10**, (2021).

97. Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci.* **111**, E5593–E5601 (2014).

98. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 1–14 (2017) doi:10.1186/s13059-017-1349-1.

99. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

100. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

101. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

102. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

103. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

104. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

105. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).

106. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

107. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

108. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

109. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 1–12 (2019) doi:10.1186/s13059-015-0762-6.

110. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

111. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-798 (2015).

112. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

113. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

114. Deo, P. N. & Deshmukh, R. Oral microbiome: Unveiling the fundamentals. *J. Oral Maxillofac. Pathol. JOMFP* **23**, 122–128 (2019).

115. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat. Rev. Microbiol.* **16**, 143–155 (2018).

116. Qin, J. *et al.* Characterization of the Genitourinary Microbiome of 1,165 Middle-Aged and Elderly Healthy Individuals. *Front. Microbiol.* **12**, (2021).

117. Ryan, M. P. & Pembroke, J. T. Brevundimonas spp: Emerging global opportunistic pathogens. *Virulence* **9**, 480–493 (2018).

118. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

119. Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).

120. Ji, S., Zhu, Z., Sun, X. & Fu, X. Functional hair follicle regeneration: an updated review. *Signal Transduct. Target. Ther.* **6**, 1–11 (2021).

121. Choi, B. Y. Targeting Wnt/β-Catenin Pathway for Developing Therapies for Hair Loss. *Int. J. Mol. Sci.* **21**, 4915 (2020).

122. Ma, H.-Y., Chen, S. & Du, Y. Estrogen and estrogen receptors in kidney diseases. *Ren. Fail.* **43**, 619–642 (2021).

123. Xie, Y. *et al.* Analysis of the Global Burden of Disease study highlights the global, regional, and national trends of chronic kidney disease epidemiology from 1990 to 2016. *Kidney Int.* **94**, 567–581 (2018).

124. Couper, D. *et al.* DESIGN OF THE SUBPOPULATIONS AND INTERMEDIATE OUTCOMES IN COPD STUDY (SPIROMICS). *Thorax* **69**, 492–495 (2014).

125. King Han, M., Dransfield, M. T. & Martinez, F. J. Chronic obstructive pulmonary disease: Definition, clinical manifestations, diagnosis, and staging - UpToDate. https://www.uptodate.com/contents/chronic-obstructive-pulmonary-disease-definition-clinical-manifestations-diagnosis-and-staging?search=copd&source=search_result&selectedTitle=1~150&usage_type=default&display_rank=1.

126. Hurst, J. R. *et al.* Understanding the impact of chronic obstructive pulmonary disease exacerbations on patient health and quality of life. *Eur. J. Intern. Med.* **73**, 1–6 (2020).

127. Sauler, M. *et al.* Characterization of the COPD alveolar niche using single-cell RNA sequencing. *Nat. Commun.* **13**, 494 (2022).

128. Oelsner, E. C. *et al.* Prognostic Significance of Large Airway Dimensions on Computed Tomography in the General Population. The Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study. *Ann. Am. Thorac. Soc.* **15**, 718–727 (2018).

129. O'Brien, M. E. *et al.* Loss of skin elasticity is associated with pulmonary emphysema, biomarkers of inflammation, and matrix metalloproteinase activity in smokers. *Respir. Res.* **20**, 128 (2019).

130. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivar. Behav. Res.* **27**, 509–540 (1992).

131. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).

132. Gonorazky, H. D. *et al.* Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).

133. Yépez, V. A. *et al.* Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med.* **14**, 38 (2022).

134. Tu, F. F. & As-Sanie, S. Chronic pelvic pain in adult females: Evaluation - UpToDate. https://www.uptodate.com/contents/chronic-pelvic-pain-in-adult-females-evaluation?search=gynecology%20workup&source=search_result&selectedTitle=7~150&usage_type=default&display_rank=7#H1123814595.

135. Andralojc, K. M. *et al.* Targeted RNA next generation sequencing analysis of cervical smears can predict the presence of hrHPV-induced cervical lesions. *BMC Med.* **20**, 206 (2022).

136. del Pino, M. *et al.* mRNA biomarker detection in liquid-based cytology: a new approach in the prevention of cervical cancer. *Mod. Pathol.* **28**, 312–320 (2015).

137. Virtanen, S. *et al.* Vaginal Microbiota Composition Correlates Between Pap Smear Microscopy and Next Generation Sequencing and Associates to Socioeconomic Status. *Sci. Rep.* **9**, 7750 (2019).

138. Lv, H., Hu, Y., Cui, Z. & Jia, H. Human menstrual blood: a renewable and sustainable source of stem cells for regenerative medicine. *Stem Cell Res. Ther.* **9**, 325 (2018).

139. Fedder, J. Nonsperm Cells in Human Semen: With Special Reference to Seminal Leukocytes and their Possible Influence on Fertility. *Arch. Androl.* **36**, 41–65 (1996).

140. Ryan, L. *et al.* Coprocytobiology: A Technical Review of Cytological Colorectal Cancer Screening in Fecal Samples. *SLAS Technol. Transl. Life Sci. Innov.* **26**, 591–604 (2021).

141. Witkowska-Zimny, M. & Kaminska-El-Hassan, E. Cells of human breast milk. *Cell. Mol. Biol. Lett.* **22**, 11 (2017).

142. Anstey, E. H. *et al.* Breastfeeding and Breast Cancer Risk Reduction: Implications for Black Mothers. *Am. J. Prev. Med.* **53**, S40–S46 (2017).

143. Hippman, C. & Nislow, C. Pharmacogenomic Testing: Clinical Evidence and Implementation Challenges. *J. Pers. Med.* **9**, 40 (2019).

144. Charlesworth, R. P. Diagnosing coeliac disease: Out with the old and in with the new? *World J. Gastroenterol.* **26**, 1–10 (2020).