genes

MDPI

Article

ADGR: Admixture-Informed Differential Gene Regulation

In-Hee Lee [1,*] and Sek Won Kong [1,2]

[1] Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02215, USA
[2] Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA
* Correspondence: in-hee.lee@childrens.harvard.edu

Abstract: The regulatory elements in proximal and distal regions of genes are involved in the regulation of gene expression. Risk alleles in intronic and intergenic regions may alter gene expression by modifying the binding affinity and stability of diverse DNA-binding proteins implicated in gene expression regulation. By focusing on the local ancestral structure of coding and regulatory regions using the paired whole-genome sequence and tissue-wide transcriptome datasets from the Genotype-Tissue Expression project, we investigated the impact of genetic variants, in aggregate, on tissue-specific gene expression regulation. Local ancestral origins of the coding region, immediate and distant upstream regions, and distal regulatory region were determined using RFMix with the reference panel from the 1000 Genomes Project. For each tissue, inter-individual variation of gene expression levels explained by concordant or discordant local ancestry between coding and regulatory regions was estimated. Compared to European, African descent showed more frequent change in local ancestral structure, with shorter haplotype blocks. The expression level of the Adenosine Deaminase Like (*ADAL*) gene was significantly associated with admixed ancestral structure in the regulatory region across multiple tissue types. Further validations are required to understand the impact of the local ancestral structure of regulatory regions on gene expression regulation in humans and other species.

Keywords: regulatory elements; local ancestry; gene expression; genotype-tissue expression

Citation: Lee, I.-H.; Kong, S.W. ADGR: Admixture-Informed Differential Gene Regulation. *Genes* **2023**, *14*, 147. https://doi.org/10.3390/genes14010147

Received: 17 November 2022
Revised: 15 December 2022
Accepted: 3 January 2023
Published: 5 January 2023

1. Introduction

The known history of human evolution and migration out of Africa, and the recent migration of people across the continents, suggest that the genomes of modern people are a composite admixture of haplotypes from multiple ancestral populations [1]. The admixture of haplotype blocks could locally introduce novel combination of alleles not observed in ancestral populations [2]. Admixture can be thought of as a series of meiotic recombination over multiple generations that contribute to genetic diversity, as successive offspring exhibit admixed genomes with new combinations of alleles [3]. As such, recombination and mutation are the main sources of genetic variation in populations. Studying the genomic composition of admixed individuals across diverse populations provides a lens to infer the recombination rate of recent admixture events estimated by constructing genetic maps using pedigree or linkage disequilibrium (LD) based approaches [4].

Genome-wide association studies (GWASs) have revealed that most disease-associated risk loci lie outside of protein coding genes [5]. Fine-mapping and expression quantitative trait loci (eQTL) analysis demonstrated that risk alleles in the non-protein coding region may alter the regulation of gene expression by modifying the binding affinity of diverse DNA binding and interacting proteins implicated in gene expression regulation [6]. For instance, a single nucleotide polymorphism (SNP) changes the binding affinity of transcription factors and epigenetic regulators, which results in differential efficiency in transcription [7] and mRNA processing [8]. Moreover, the transferability of GWAS findings to other populations is challenging since variant allele frequency of risk alleles varies between populations. To this end, estimating the rate of recent recombination events

allows for the identification of genomic loci that may be associated with disease across populations [1].

Almost 80% of currently available GWASs were performed using DNA samples from European descent, especially from the people of the United States, the United Kingdom and Iceland [9–11]. Therefore, previous eQTL studies used genotype-derived global ancestry and/or top-most principal components of genotypes as covariates, or just focused on minimally admixed European populations [12,13]. Generalizing the findings from these studies to non-European populations is challenging, especially for admixed populations. To this end, Zhong and colleagues incorporated local ancestry information to explain a proportion of variance in gene expression levels between individuals and found polygenic contributions to gene expression variations in admixed individuals [14]. Thus, a new method must be sought out that not only describes the effect of local ancestry on gene expression regulation, independent of population, but also adjusts the model to reduce false positive association between genotype and expression phenotype.

Our approach to describing admixed ancestral structure is focused on transitions—i.e., genomic loci delineating potential recombination events between continent-level populations—that indicate changes in local ancestry from one ancestral population to another. We inferred local ancestry with whole genome sequencing (WGS) data using RFMix, an algorithm that learns from a reference panel of haplotypes and genetic recombination maps to infer the most likely local ancestral structure of a query genome. RFMix uses a genetic map-based approach for estimating local ancestry, which differs from the other algorithms based on linkage disequilibrium (LD) and, therefore, is not bound to the limitation of classifying local ancestry of up to two populations in LD-based approaches [15]. Thus, we could use larger reference panels to predict among several ancestral populations at a time. We propose a method, Admixture-informed Differential Gene Regulation (ADGR), for modeling differences in gene expression, which may be used as a proxy for phenotypic changes associated with disease, due to changes in local ancestry between protein coding and upstream regulatory regions as a result of admixture.

2. Materials and Methods

We collected paired genome-wide variant calls from phased WGS and tissue-wide RNA-seq datasets from the Genotype-Tissue Expression (GTEx) project (release V8) [16,17]. For 838 phased WGS variant call files (VCFs), we used RFMix (version 2) to infer genome-wide local ancestral structure [2]. Reference panels were constructed from combinations of continent-level populations of the 1000 Genomes Project: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and Southeast Asian (SAS). Samples that represented each continent-level population were selected from populations that showed the least degree of admixture: Yoruba (YRI) for AFR, Peru (PEL) for AMR, Han Chinese (CHB) for EAS, Utah Residents with Northern and Western European ancestry (CEU) for EUR, and Italian Telugu (ITU) for SAS. From each of the five populations, we chose 85 individuals with lesser degrees of admixture according to ADMIXTURE results [18]. Three reference panels that consisted of two-, three-, and five-populations were used to assign local ancestry with RFMix. A two-population panel consisted of the 85 individuals from each of YRI and CEU, and a three-population reference panel consisted of individuals from YRI, PEL, and CEU.

For each individual WGS from GTEx, RFMix assigned one of the ancestral populations in the reference panel to each of two alleles along the chromosome. Then, consecutive regions with the same local ancestry formed a haplotype block. A transition point was defined as the genomic locus between two adjacent haplotype blocks with different local ancestries. For further analysis, we focused on two populations—i.e., EUR (N = 715) and AFR (N = 103)—since there were only a small number of individuals from the other populations (AMR N = 2, ASN N = 12 and unknown N = 6) in the GTEx project.

We defined protein coding and upstream regulatory regions according to the GENCODE annotation (version 26) [19,20]. For each protein coding gene, the upstream genomic

region from the transcription start site (TSS) was further partitioned to three regions by distance from TSS: immediate upstream (up to 5 kilobase pairs, kbps), distant upstream (5–50 kbps) and distal (50–500 kbps) regions (Figure 1).

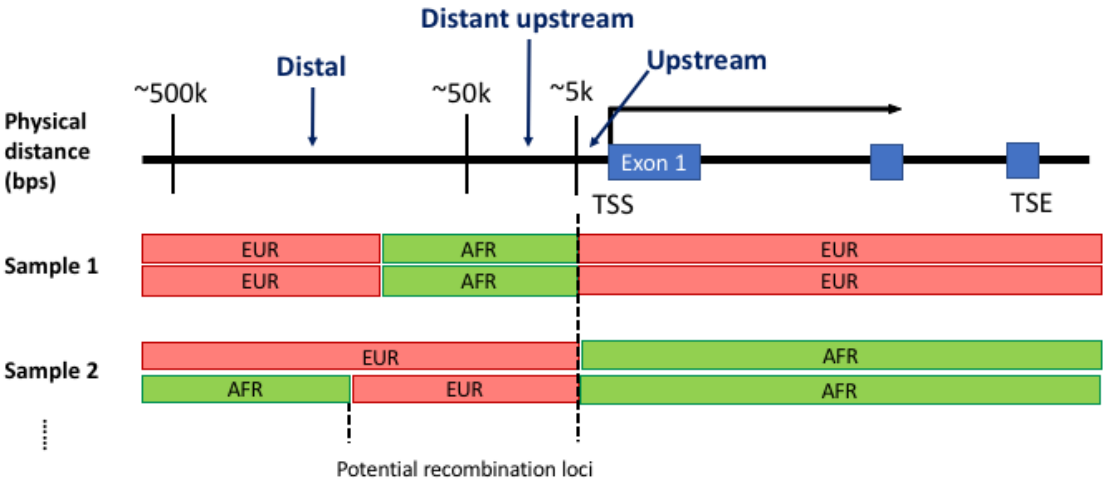

Figure 1. Admixed ancestral structure relative to protein coding gene. Immediate upstream, distant upstream and distal regions are defined by their physical distance from the transcription start site (TSS) for each of protein coding genes defined in the GENCODE annotation. For each individual, a block of genomic region is assigned to one of the continent-level ancestral populations included in the reference panel prepared for RFMix. For the reference panel with two populations—i.e., European (EUR) and African (AFR)—haplotype blocks are assigned to one of the ancestral populations as either heterozygous or homozygous. A transition point between adjacent haplotype blocks suggests ancestral recombination locus.

For each gene *i*, we compared two linear regression models with and without the presence of transition points in upstream regions. The baseline model (M_0) was y_i~*age* + *sex* + *global ancestry* and the alternative model (M_1) was y_i~*trans* + *age* + *sex* + *global ancestry*, where y_i denoted the standardized expression level of gene *i*. For global ancestry, we cross-checked reported information in the GTEx phenotype table and predicted global ancestry derived from WGS—the largest proportion of local ancestries for an individual by RFMix. For prostate, uterus, and ovary, we excluded the variable *sex* from both M_0 and M_1. The independent variable *trans* represents the transition point status upstream of the gene *i*. We used three different approaches to model the local ancestry transition point: (1) dominant model: *trans* = 1 if any of the two alleles contained transition points in the upstream of the gene (otherwise *trans* = 0); (2) additive model: the variable *trans* equals the number of alleles that have transition points; and (3) recessive model: *trans* = 1 only if both alleles have transition points (otherwise *trans* = 0). For each model and three upstream regions, we compared the two models—i.e., M_0 and M_1 using a two-sided χ^2 test—to find the genes for which expression levels were significantly better explained by the presence of transition events in upstream regions.

3. Results

3.1. Local Ancestral Structure of 838 Individuals

The reported global ancestry of each individual from GTEx phenotype data matched with the ancestral population predicted for the largest portion of its genome for all individuals in the current study. Global and local ancestral structure were summarized in three ways. Firstly, we created an ADMIXTURE-style graph to visualize the overall proportion of continent-level populations for all. Secondly, we counted the total number of transition points in each individual. Thirdly, we checked the distributions of putative haplotype block sizes between two transition points. In Figure 2A–C, the two largest populations

in the GTEx project—i.e., African and European—are shown, and the largest proportion of predicted local ancestry from RFMix was concordant with the reported global ancestry from the GTEx phenotype metadata. This observation was consistent regardless of the number of ancestral populations in the reference panels for RFMix: two (AFR and EUR, Figure 2A), three (AFR, AMR and EUR, Figure 2B) or five populations (AFR, AMR, ASN, EUR, and SAS, Figure 2C).

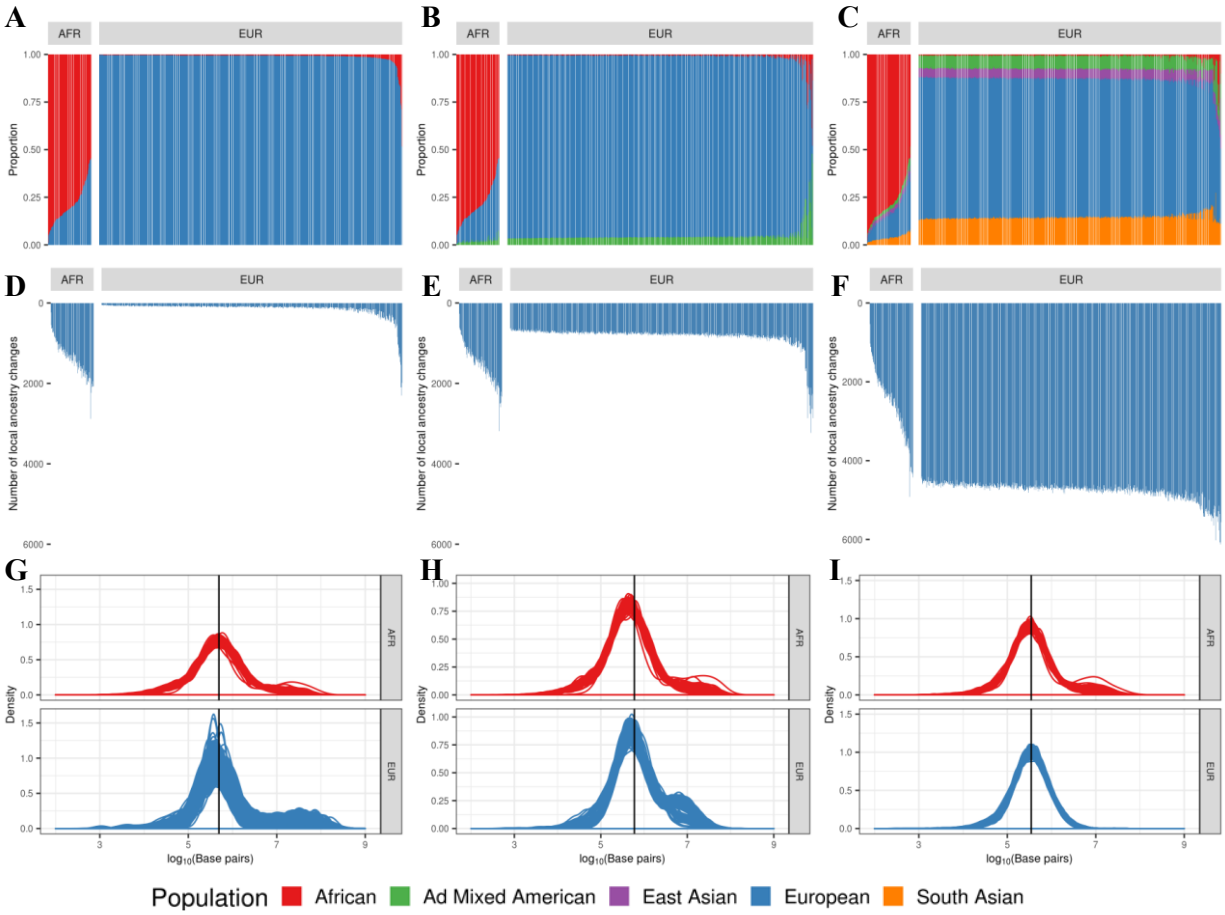

Figure 2. An overview of global and local ancestral structures of whole-genome sequencing data from the Genotype-Tissue Expression project. The overall proportion of continent-level ancestral populations within each subject as predicted by RFMix using the reference panel of (**A**) two populations, (**B**) three populations and (**C**) five populations. The number of transition points within each subject when using two-, three- and five-population panels (**D**–**F**, respectively). The distribution of haplotype block lengths for African and European individuals with two-, three- and five-population panels (**G**–**I**, respectively): African individuals (**top**) and European individuals (**bottom**).

Using the five-population reference panel, African individuals had 2592 transition points on average in their genome (standard deviation (SD) 846, range from 530 to 4914) compared to the average of 4758 transition points found in Europeans (SD 213, range from 4404 to 6120). However, with the two-population reference panel including only AFR and EUR, the average number of transition points in Europeans was significantly reduced compared to reference panels with three or five populations (Figure 2D–F). European individuals showed a higher increase in the number of transition points because about 20% of genomes, which were mapped to EUR with the two-population reference panel (AFR and EUR), were mapped to non-EUR populations as the reference panel changed (right panels in Figure 2A–C). However, for African individuals, the majority of the genome was consistently mapped to AFR and only small portion (~5%) of the genome was mapped differently as the reference panel changed (left panels in Figure 2A–C).

The distribution of block lengths is shown in Figure 2G–I. Specifically, the blocks shown here are for the stretches of putative local ancestry that were concordant with the reported global ancestry in each individual. In both population groups, the average block length predicted by RFMix was approximately 350 kbps. In the five-population panel, African subjects on average have longer stretches of concordant ancestry, as shown by the heavier tail on the right side of the distribution. In the two-population, on the other hand, Europeans showed a similar trend, with a right-side heavy tail in the distribution.

3.2. Transition of Local Ancestral Structure and Gene Model

We checked the location of potential transitions between ancestral blocks in relation to the gene definitions according to the GENCODE annotation. Here, we focused on the results generated using the two-population reference panel. Out of 1413.5 (SD 423.92) total transitions per individual, African individuals had 830.3 (SD 246.65) transitions in the intergenic region. European individuals had 82.8 (SD 103.83) transitions in the intergenic region out of 137.7 (SD 177.87) total transitions on average (Figure 3).

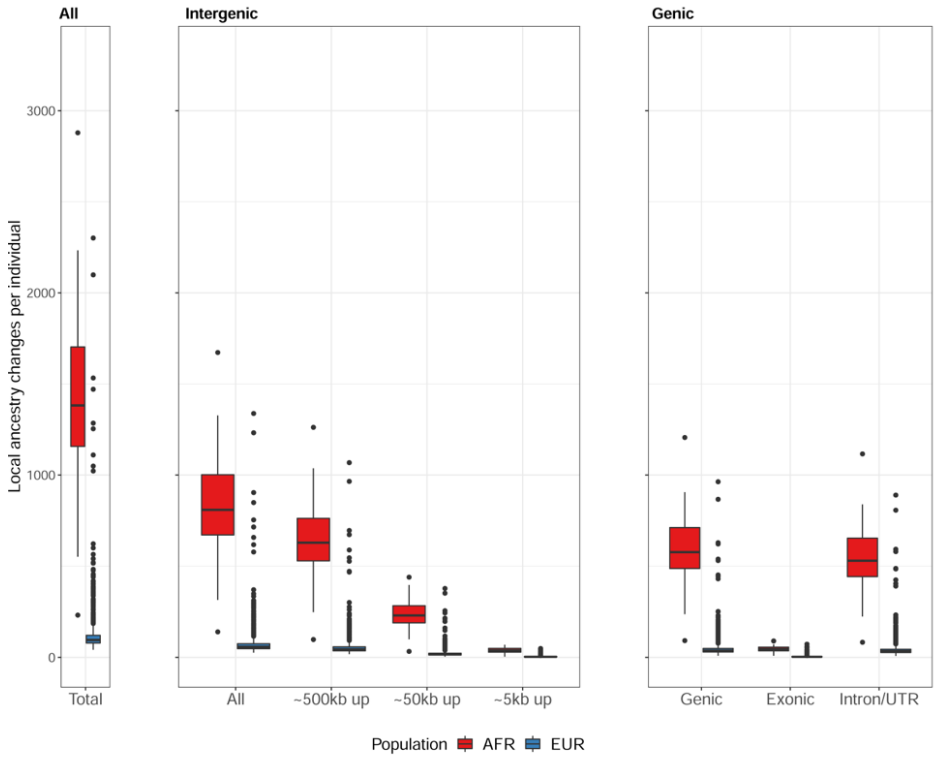

Figure 3. Predicted transition loci relative to the GENCODE gene models: the total number of transition loci from an individual (**left**), the number of transition loci in the intergenic region (**center**) and in the genic region (**right**). African individuals have a significantly larger number of admixed events on average compared to European individuals when the two-population reference panel is used for RFMix analysis for local ancestry inference with transitions.

In both groups, ~60% of total transitions were observed in the intergenic region and 40% in the genic region. Most transitions in the intergenic region were within 500 kbps from TSS, with the largest number between 50 kbps and 500 kbps (center box in Figure 3). The rightmost box in Figure 3 shows that most transitions in the genic region were within introns or the untranslated region (UTR), leaving only a few in exons: 45.6 (SD 14.23) transitions in exons out of 583.2 (SD 179.51) in the genic region among African individuals, and 4.5 (SD 5.96) transitions in exons out of 54.9 (SD 74.43) in the genic region among European individuals.

3.3. Gene Expression Levels Associated with Admixed Ancestral Structure in the Regulatory Region

Next, we used RFMix predictions using the two-population reference panel and standardized gene expression levels to rank the genes whose expression levels could be explained by the presence of transition events in the upstream regions. The gene expression matrices for available tissue types were downloaded from GTEx single-tissue cis-eQTL data in the GTEx portal (https://gtexportal.org, accessed on 26 August 2019). As shown in Table 1, most of the candidate associations were found for genes with transitions in 50~500 kbps upstream. The list of candidate genes varied by tissue types; however, the *ADAL* gene encoding the adenosine aminase like protein (ADAL) was significantly associated with admixed ancestral structure in the regulatory region, especially in 50~500 kbps upstream, across multiple tissue types for both dominant and additive models (Table 1).

Table 1. The list of candidate genes across tissue types. The distance ranges are from the transcription start site of gene model to transition points in the upstream. The three models of transition events (i.e., dominant, additive, or recessive) are used for linear regression analysis. False discovery rate is calculated within each model of transition event, distance range, and tissue type.

Model	Distance from Transcription Start Site	Tissue Type	Gene	False Discovery Rate
Dominant	Less than 5 kbps	Small Intestine, Terminal Ileum	*SLC17A9*	0.047
	5~50 kbps	Brain, Cerebellar Hemisphere	*HLA-DMA*	0.018
	50~500 kbps	Adipose, Subcutaneous	*ADAL*	0.00029
			C10orf107	0.0049
			HLA-DQB2	0.016
		Artery, Aorta	*ADAL*	8.4×10^{-6}
			PSORS1C2	0.007
		Artery, Tibial	*ADAL*	6.6×10^{-7}
		Brain Cerebellum	*HLA-A*	0.01
		Breast, Mammary Tissue	*ADAL*	0.0035
		Colon, Transverse	*ADAL*	0.00083
		Esophagus, Muscularis	*C10orf107*	0.02
			ADAL	0.02
		Heart, Atrial Appendage	*C10orf107*	0.00068
			PLEK2	0.04
			ALOX12	0.04
		Lung	*PCDHGA6*	0.029
			PSORS1C2	0.029
			STEAP2	0.03
		Muscle, Skeletal	*HLA-DQB2*	0.025
			COL8A2	0.025
		Nerve, Tibial	*ADAL*	2.4×10^{-6}

Table 1. *Cont.*

Model	Distance from Transcription Start Site	Tissue Type	Gene	False Discovery Rate
		Ovary	*ADAL*	0.0029
		Skin, Not Sun Exposed Suprapubic	*ADAL*	0.035
		Skin, Sun Exposed Lower leg	*ADAL*	0.00068
		Spleen	*ADAL*	0.00055
		Stomach	*ADAL*	0.00074
		Thyroid	*WDR87*	5.5×10^{-5}
			ADAL	0.00012
			ZSCAN31	0.0053
		Whole Blood	*MISP3*	0.033
Recessive	Less than 5 kbps	Uterus	*SH3GLB1*	0.043
Additive	Less than 5 kbps	Small Intestine, Terminal Ileum	*SLC17A9*	0.047
	5~50 kbps	Brain, Cerebellar Hemisphere	*HLA-DMA*	0.018
	50~500 kbps	Adipose, Subcutaneous	*HLA-DQB2*	6.6×10^{-6}
			ADAL	0.00014
			C10orf107	0.0033
		Adipose, Visceral Omentum	*HLA-DQB2*	0.0092
		Artery, Aorta	*ADAL*	2.1×10^{-5}
		Artery, Tibial	*ADAL*	6.6×10^{-7}
		Breast, Mammary Tissue	*ADAL*	0.0074
		Colon, Transverse	*ADAL*	0.00083
		Esophagus, Muscularis	*C10orf107*	0.02
		Heart, Atrial Appendage	*HLA-DQB2*	3.3×10^{-5}
			C10orf107	0.00034
		Heart, Left Ventricle	*HLA-DRB5*	0.015
		Lung	*PCDHGA6*	0.032
			TLDC1	0.032
		Muscle, Skeletal	*HLA-DQB2*	3.1×10^{-7}
		Nerve, Tibial	*ADAL*	3.7×10^{-6}
		Ovary	*ADAL*	0.0053
		Skin, Not Sun Exposed Suprapubic	*ZNF347*	0.022
			ADAL	0.022
		Skin, Sun Exposed Lower leg	*ADAL*	0.00068
		Spleen	*ADAL*	0.00055
		Stomach	*ADAL*	0.00074
		Thyroid	*WDR87*	5.9×10^{-5}
			ADAL	0.00012
			ZSCAN31	0.0014
		Vagina	*CTNNA2*	0.029
		Whole Blood	*ZFP57*	0.0029

For the individuals with transitions in 50~500 kbps upstream of *ADAL*, expression levels of this gene were significantly lower compared to the individuals without transition in the upstream (Figure 4). Interestingly, *ADAL* was differentially expressed between African Americans and European Americans with colorectal cancer [21].

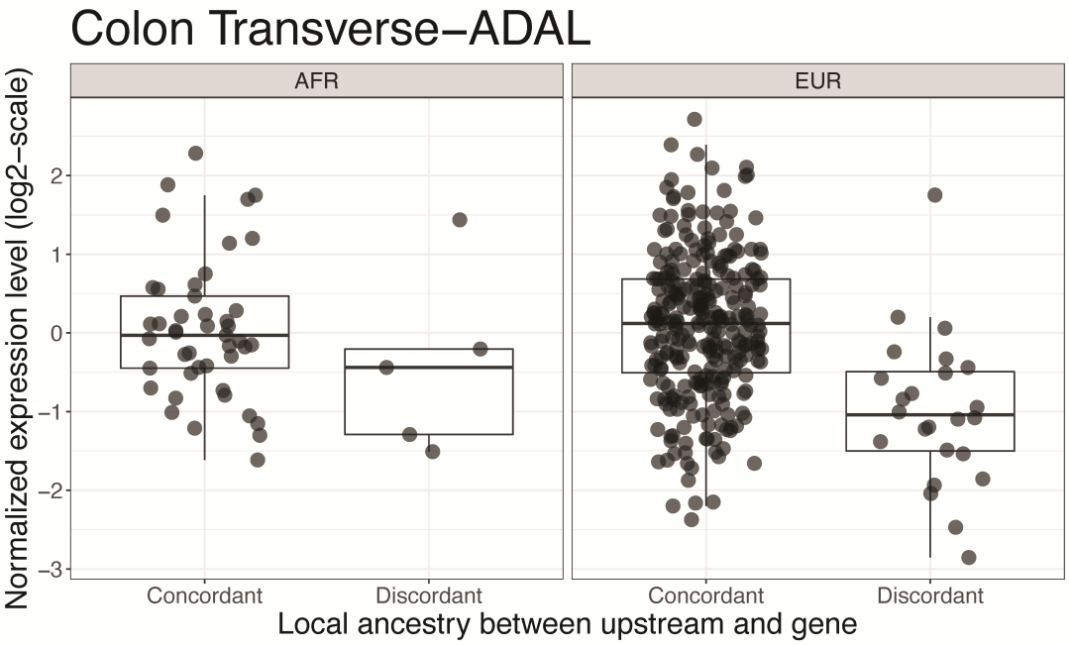

Figure 4. The expression level of the Adenosine Deaminase Like (*ADAL)* gene in transverse colon among the individuals without transition in the upstream ("Concordant") and among the individuals with transition in the upstream ("Discordant"). The expression level (*y*-axis) shows the value from the gene expression matrix for transverse colon in the GTEx single-tissue *cis*-eQTL data, after normalizing with age, sex, and global ancestry. The left panel shows expression levels for African individuals and the right panel for European individuals. The expression levels are lower among "Discordant" individuals.

Compared to dominant or additive models, we found significantly smaller numbers of candidate genes with recessive models, which suggested that most of the transitions in the upstream were heterozygous. Interestingly, *ADAL* was consistently found significant in multiple tissue types for both dominant and additive models. We also observed that all the candidate genes from the additive model had transitions in only one of alleles (heterozygous). Since transition was, after all, infrequent, we expected that homozygous transition would be very rare, which made it difficult to find candidate genes using a recessive model. We found only one candidate gene with the recessive model (*SH3GLB1* in uterus).

3.4. Gene Expression Levels in Chromosome 8q24 Associated with Local Ancestral Transition between Africans and Europeans

Prostate cancer is one of the most common malignancies among men in the U.S., and the incidence among African Americans is ~1.6-fold higher compared to European Americans. Freedman and colleagues used a whole-genome admixture scan to discover susceptibility loci for prostate cancer in African Americans and found chromosome 8q24 as a significant risk locus for prostate cancer, especially for African descent. However, candidate genes in 8q24 were not identified [22]. We focused on the transition events and genes in chromosome 8q24 to find the candidate genes that were differentially regulated by admixed ancestral structure in regulatory regions. We found that more significant associations were from recessive models, in contrast to the whole genome analysis in the previous section (Table 1). Trafficking protein particle complex 9 (*TRAPPC9*) and pyrroline-

5-carboxylate reductase (*PYCRL*) were significantly associated with ancestral admixture in the regulatory region across multiple tissue types (Table 2). *TRAPPC9* is implicated in tumorigenesis through the NF-kB signaling pathway [23]. *PYCRL* plays a role in proline biosynthesis [24] and was significantly associated with prostate proliferation in a murine model of prostate cancer [25]. We checked the locations of the genes and their regulatory regions in Table 2 and did not find any overlap of distal regulatory regions between the genes that were significant in a tissue type (Figure 5).

Table 2. Significant genes associated with transition events in the regulator region on chromosome 8q24. False discovery rate is calculated within each model of transition event, distance range, and tissue type.

Model	Distance from Transcription Start Site	Tissue Type	Gene	False Discovery Rate
Dominant	5~50 kbps	Skin, Not Sun Exposed Suprapubic	*ANXA13*	0.015
			TRAPPC9	0.025
	50~500 kbps	Brain, Spinal Cord Cervical C1	*GPT*	0.035
		Whole Blood	*ZNF572*	0.041
Recessive	Less than 5 kbps	Brain, Anterior Cingulate Cortex	*ZNF623*	0.032
		Brain, Frontal Cortex BA9	*ZNF623*	0.03
		Colon, Transverse	*LYNX1*	0.0015
	5~50 kbps	Brain, Caudate Basal Ganglia	*PYCRL*	0.02
		Brain, Cerebellar Hemisphere	*PYCRL*	0.037
			TRAPPC9	0.037
		Brain, Hypothalamus	*PYCRL*	0.024
			TRAPPC9	0.024
		Skin, Not Sun Exposed Suprapubic	*TRAPPC9*	0.049
Additive	5~50 kbps	Skin, Not Sun Exposed Suprapubic	*ANXA13*	0.015
			TRAPPC9	0.025
	50~500 kbps	Brain, Spinal Cord Cervical C1	*GPT*	0.031
		Skin, Sun Exposed Lower Leg	*ZFP41*	0.025

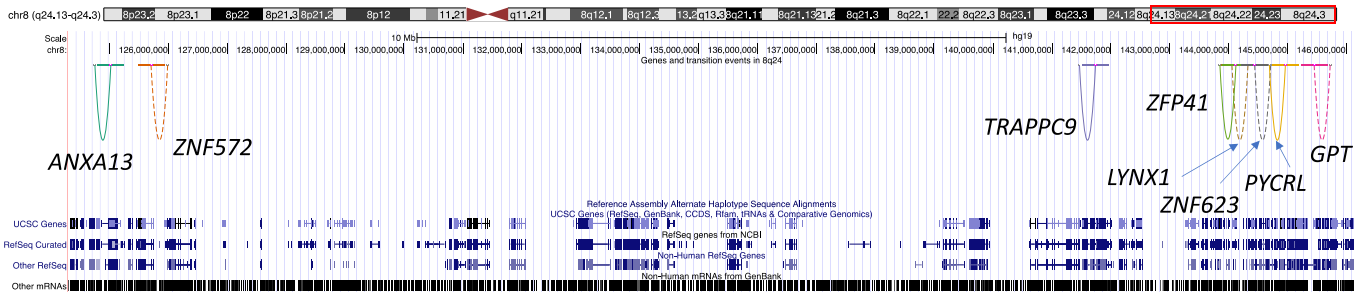

Figure 5. Genes in chromosome 8q24 locus and their upstream regions considered for ancestral structure. Only the genes with significant associations are shown. For each gene, the transcription start site (TSS) that is denoted as a single base position and its upstream region (denoted as a horizontal line in the left/right of TSS) are connected by arcs. For visibility, only the 50~500 kbps upstream regions are shown. The other upstream regions considered during the analysis are located between end points of the arcs. Each gene is represented by different colors and the dashed arcs represent those on the positive strand.

4. Discussion

The proportion of phenotype variance explained by genotype is relatively small for many human traits, including diseases [26]. Gene expression could be used as an endophenotype that is a mediator between genotype and phenotype. Indeed, genetic variants have larger effects on the variance in CpG DNA methylation and gene expression levels compared to effect sizes on phenotypic variance [27]. Positive findings from eQTL analysis and fine-mapping of GWAS results, as well as statistical methods such as PrediXcan [28] and fusion [29], suggest that inter-individual variation of tissue-specific gene expression could be explained by from a single SNP to genome-wide genotype of an individual [30]; however, it is likely that multiple genetic variants in the regulatory region contribute to differential gene expression regulation across individuals [31].

On average, the number of putative transition points was smaller in European descent compared to African descent when a two-population reference panel was used. This observation is consistent with previous reports regarding greater genetic diversity in Africans with shorter LD-block sizes [32–34]. However, we found more frequent transition points in Europeans with three- or five-population reference panels, which was likely due to the limitation of the algorithm in assigning local ancestry to one of the populations in a population reference panel. RFMix performs better with individuals from complex admixed populations compared to the other methods [35]; however, the subjects enrolled in the GTEx project were not necessarily from complex admixed populations. Therefore, the results with two-population reference panel were consistent with previous reports as to the number of transition points and block sizes in Europeans and Africans.

The association between the local ancestry transition In the distal upstream (50~500 kbps from TSS) of *ADAL* and its expression level was recurrently observed across 13 tissue types. ADAL has an important role in the metabolism of mRNA across cell types in multiple species. N^6-methyl adenine (m^6A) is the most abundant posttranscription modification of mRNA, and m^6A is turned over to generate N^6-mAMP. ADAL is an evolutionary conserved catalytic enzyme that hydrolyzes N^6-methyl-AMP (N^6-mAMP) to produce inositol monophosphate (IMP) and methylamine [36]. Therefore, differential regulation of *ADAL* could have an impact on mRNA stability and metabolism [37]. In our analysis, *ADAL* was significant in multiple tissue types, which was not solely due to the ubiquitous expression of *ADAL*. Indeed, 77% of all tested genes (N = 13,556) were quantitatively measured in 40 or more tissue types in GTEx. Nonetheless, *ADAL* and *HLA-DQB2* were two genes that were significantly associated with the local ancestry of regulatory regions in diverse tissue types at a study-wide statistical threshold of $1.87 \times 10^{-8} = 0.05/2{,}670{,}793$ (the number of all tests for all genes across available tissues, models of transition loci (additive/dominant/recessive), and distance between transition loci and TSS (immediate/distant/distal)). Given the sample size in GTEx data (N = 838), however, it requires replication study with a larger sample size for validation.

Gene expression levels are influenced by both genetic and environmental factors. Moreover, environmental factors such as lifestyle and diet are often linked with an individual's global ancestry. In the current study, we aimed to delineate the effect of genetic variants in regulatory regions, in aggregate, on the inter-individual variation of gene expression levels. For some genes, mean expression levels could be different between populations due to environmental factors and gene-environment interactions. In the current study, gene expression levels were residualized for global ancestry (along with sex and age) to estimate the variance explained by the change in local ancestry between regulatory and coding regions. Therefore, the genes that were significantly differentially expressed between populations might have been missed in our analysis.

Although our approach identified candidate genes that may be differentially expressed due to the discordant local ancestry of the regulatory region compared to coding regional structure, there are several technical challenges that make interpretation difficult. Firstly, the potential impact of allele-specific expression was not explored [38]. Most transition events were heterozygous in our dataset. Thus, one of two alleles with discordant local ancestry in

regulatory regions could have a differential effect on gene expression. The next generation sequencing technique used to generate WGS and RNA-seq data for the GTEx project has limitations in resolving the haplotype of regulatory regions relative to coding regions. Third generation long-read sequencing techniques, such as 10× linked-reads sequencing and Oxford Nanopore, would enable the generation of an accurate allele specific map of regulatory and coding regions [39]. Secondly, transition events in upstream might have different impacts across cell types, which it was not possible to analyze using bulk RNA-seq data from the GTEx project. Thirdly, there was a lack of reliable reference haplotype data in the latest human genome build GRCh38. RFMix requires prior information from a human genome-wide recombination map and a reference panel of different ancestral populations matching the target population. Therefore, local ancestry prediction with RFMix is dependent on the quality and size of the required materials. However, publicly available genomic data are biased with European populations [9,40], limiting our ability to investigate individuals of non-European ancestry.

DNA double strand break sites—i.e., sites of meiotic recombination—are often determined by PR domain-containing protein 9 (PRDM9) in the human [41]. Interestingly, different ancestral populations have distinct recombination hotspots. Moreover, *PRDM9* alleles and DNA sequence motif binding PRDM9 show difference between Europeans and African Americans [34]. High resolution genetic maps for diverse ancestral populations are not readily available yet. As such, we found significant differences in the number of transitions and the distribution of size of haplotype blocks between the results using three different reference panels. Further refinement of local ancestry prediction methods would improve the statistical power to detect gene expression variation explained by admixed ancestral structure in the regulatory region.

5. Conclusions

In the current study, we illustrated an intuitive way to estimate the impact of local ancestry on gene expression levels in the two populations (i.e., AFR and EUR) using GTEx WGS data. A total of 61 significant candidate genes were discovered across 24 tissue types. After multiple testing correction for each tissue, *ADAL* was recurrently identified for the additive and dominant models across multiple tissues. We used a paired WGS and RNA-seq dataset generated from autopsy samples in the current study to illustrate a proof-of-concept. Our approach can be applied to study genetic basis of traits (e.g., transcriptome, proteome, and other phenotype of interests) for animals and plants for which more accurate recombination maps could be generated [42,43]. For instance, molecular mechanisms of breed-defining traits have been characterized in livestock animals by genotyping germline mutations in coding and regulatory sequences [44]. Furthermore, the current approach could be refined to understand how genetic variants in regulatory elements lead to various human phenotypes.

Author Contributions: Conceptualization, I.-H.L. and S.W.K.; formal analysis, I.-H.L.; writing—original draft preparation, I.-H.L.; writing—review and editing, S.W.K.; funding acquisition, S.W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the US National Center for Advancing Translational Sciences, grant number U01TR002623, and by the National Institute of Health Common Fund Bridge2AI, grant number 1OT2OD032720.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund (https://commonfund.nih.gov/GTEx, accessed on 26 August 2019) of the Office of the Director of the National Institute of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analysis described in this manuscript were obtained from the GTEx Portal (https:

//www.gtexportal.org, accessed on 26 August 2019) for normalized gene expression matrices and dbGaP for whole genome sequencing data by the accession number phs000424.v8.p2 (https://www.ncbi.nlm.nih.gov/gap/, accessed on 29 July 2019).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Benton, M.L.; Abraham, A.; LaBella, A.L.; Abbot, P.; Rokas, A.; Capra, J.A. The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **2021**, *22*, 269–283. [CrossRef]
2. Maples, B.K.; Gravel, S.; Kenny, E.E.; Bustamante, C.D. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **2013**, *93*, 278–288. [CrossRef]
3. Kong, A.; Thorleifsson, G.; Gudbjartsson, D.F.; Masson, G.; Sigurdsson, A.; Jonasdottir, A.; Walters, G.B.; Jonasdottir, A.; Gylfason, A.; Kristinsson, K.T.; et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **2010**, *467*, 1099–1103. [CrossRef] [PubMed]
4. O'Reilly, P.F.; Balding, D.J. Admixture provides new insights into recombination. *Nat. Genet.* **2011**, *43*, 819–820. [CrossRef]
5. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef] [PubMed]
6. Fish, A.E.; Crawford, D.C.; Capra, J.A.; Bush, W.S. Local ancestry transitions modify snp-trait associations. *Acific Symp. Biocomput.* **2018**, *23*, 424–435.
7. Montgomery, S.B.; Dermitzakis, E.T. From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* **2011**, *12*, 277–282. [CrossRef] [PubMed]
8. Hentze, M.W.; Castello, A.; Schwarzl, T.; Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 327–341. [CrossRef] [PubMed]
9. Martin, A.R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B.M.; Daly, M.J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **2019**, *51*, 584–591. [CrossRef]
10. Mills, M.C.; Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* **2020**, *52*, 242–243. [CrossRef]
11. Sollis, E.; Mosaku, A.; Abid, A.; Buniello, A.; Cerezo, M.; Gil, L.; Groza, T.; Gunes, O.; Hall, P.; Hayhurst, J.; et al. The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Res.* **2022**. [CrossRef] [PubMed]
12. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585.
13. Lappalainen, T.; Sammeth, M.; Friedlander, M.R.; Hoen, P.A.; Monlong, J.; Rivas, M.A.; Gonzalez-Porta, M.; Kurbatova, N.; Griebel, T.; Ferreira, P.G.; et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **2013**, *501*, 506–511. [CrossRef] [PubMed]
14. Zhong, Y.; Perera, M.A.; Gamazon, E.R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am. J. Hum. Genet.* **2019**, *104*, 1097–1115. [CrossRef] [PubMed]
15. Baran, Y.; Pasaniuc, B.; Sankararaman, S.; Torgerson, D.G.; Gignoux, C.; Eng, C.; Rodriguez-Cintron, W.; Chapela, R.; Ford, J.G.; Avila, P.C.; et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **2012**, *28*, 1359–1367. [CrossRef]
16. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **2015**, *348*, 648–660. [CrossRef]
17. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **2020**, *369*, 1318–1330. [CrossRef]
18. Genomes Project, C.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.
19. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774. [CrossRef]
20. Frankish, A.; Diekhans, M.; Ferreira, A.M.; Johnson, R.; Jungreis, I.; Loveland, J.; Mudge, J.M.; Sisu, C.; Wright, J.; Armstrong, J.; et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **2019**, *47*, D766–D773. [CrossRef]
21. Jovov, B.; Araujo-Perez, F.; Sigel, C.S.; Stratford, J.K.; McCoy, A.N.; Yeh, J.J.; Keku, T. Differential gene expression between African American and European American colorectal cancer patients. *PLoS ONE* **2012**, *7*, e30168. [CrossRef] [PubMed]
22. Freedman, M.L.; Haiman, C.A.; Patterson, N.; McDonald, G.J.; Tandon, A.; Waliszewska, A.; Penney, K.; Steen, R.G.; Ardlie, K.; John, E.M.; et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14068–14073. [CrossRef] [PubMed]
23. Zhang, Y.; Liu, S.; Wang, H.; Yang, W.; Li, F.; Yang, F.; Yu, D.; Ramsey, F.V.; Tuszyski, G.P.; Hu, W. Elevated NIBP/TRAPPC9 mediates tumorigenesis of cancer cells through NFkappaB signaling. *Oncotarget* **2015**, *6*, 6160–6178. [CrossRef]

24. Liu, W.; Hancock, C.N.; Fischer, J.W.; Harman, M.; Phang, J.M. Proline biosynthesis augments tumor cell growth and aerobic glycolysis: Involvement of pyridine nucleotides. *Sci. Rep.* **2015**, *5*, 17206. [CrossRef] [PubMed]
25. Xu, Q.; Majumder, P.K.; Ross, K.; Shim, Y.; Golub, T.R.; Loda, M.; Sellers, W.R. Identification of prostate cancer modifier pathways using parental strain expression mapping. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 17771–17776. [CrossRef] [PubMed]
26. Thomas, D. Gene–environment-wide association studies: Emerging approaches. *Nat. Rev. Genet.* **2010**, *11*, 259–272. [CrossRef] [PubMed]
27. Wu, Y.; Zeng, J.; Zhang, F.; Zhu, Z.; Qi, T.; Zheng, Z.; Lloyd-Jones, L.R.; Marioni, R.E.; Martin, N.G.; Montgomery, G.W.; et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* **2018**, *9*, 918. [CrossRef] [PubMed]
28. Gamazon, E.R.; Wheeler, H.E.; Shah, K.P.; Mozaffari, S.V.; Aquino-Michaels, K.; Carroll, R.J.; Eyler, A.E.; Denny, J.C.; Consortium, G.T.; Nicolae, D.L.; et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **2015**, *47*, 1091–1098. [CrossRef] [PubMed]
29. Gusev, A.; Ko, A.; Shi, H.; Bhatia, G.; Chung, W.; Penninx, B.W.; Jansen, R.; de Geus, E.J.; Boomsma, D.I.; Wright, F.A.; et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **2016**, *48*, 245–252. [CrossRef] [PubMed]
30. Liang, Y.; Aguet, F.; Barbeira, A.N.; Ardlie, K.; Im, H.K. A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction. *Nat. Commun.* **2021**, *12*, 1424. [CrossRef]
31. Pai, A.A.; Pritchard, J.K.; Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.* **2015**, *11*, e1004857. [CrossRef] [PubMed]
32. Reich, D.E.; Cargill, M.; Bolk, S.; Ireland, J.; Sabeti, P.C.; Richter, D.J.; Lavery, T.; Kouyoumjian, R.; Farhadian, S.F.; Ward, R.; et al. Linkage disequilibrium in the human genome. *Nature* **2001**, *411*, 199–204. [CrossRef] [PubMed]
33. Zhu, X.; Yan, D.; Cooper, R.S.; Luke, A.; Ikeda, M.A.; Chang, Y.P.; Weder, A.; Chakravarti, A. Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: Findings from the family blood pressure program. *Genome Res.* **2003**, *13*, 173–181. [CrossRef]
34. Hinch, A.G.; Tandon, A.; Patterson, N.; Song, Y.; Rohland, N.; Palmer, C.D.; Chen, G.K.; Wang, K.; Buxbaum, S.G.; Akylbekova, E.L.; et al. The landscape of recombination in African Americans. *Nature* **2011**, *476*, 170–175. [CrossRef]
35. Uren, C.; Hoal, E.G.; Moller, M. Putting RFMix and ADMIXTURE to the test in a complex admixed population. *BMC Genet.* **2020**, *21*, 40. [CrossRef] [PubMed]
36. Chen, M.; Urs, M.J.; Sanchez-Gonzalez, I.; Olayioye, M.A.; Herde, M.; Witte, C.P. m(6)A RNA Degradation Products Are Catabolized by an Evolutionarily Conserved N(6)-Methyl-AMP Deaminase in Plant and Mammalian Cells. *Plant Cell* **2018**, *30*, 1511–1522. [CrossRef] [PubMed]
37. Wu, B.; Zhang, D.; Nie, H.; Shen, S.; Li, Y.; Li, S. Structure of Arabidopsis thaliana N(6)-methyl-AMP deaminase ADAL with bound GMP and IMP and implications for N(6)-methyl-AMP recognition and processing. *RNA Biol.* **2019**, *16*, 1504–1512. [CrossRef] [PubMed]
38. Zhabotynsky, V.; Huang, L.; Little, P.; Hu, Y.J.; Pardo-Manuel de Villena, F.; Zou, F.; Sun, W. eQTL mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. *PLoS Genet.* **2022**, *18*, e1010076. [CrossRef]
39. Dreau, A.; Venu, V.; Avdievich, E.; Gaspar, L.; Jones, F.C. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat. Commun.* **2019**, *10*, 4309. [CrossRef]
40. Bentley, A.R.; Callier, S.; Rotimi, C.N. Diversity and inclusion in genomic research: Why the uneven progress? *J. Community Genet.* **2017**, *8*, 255–266. [CrossRef]
41. Paigen, K.; Petkov, P.M. PRDM9 and Its Role in Genetic Recombination. *Trends Genet.* **2018**, *34*, 291–300. [CrossRef] [PubMed]
42. Wang, Y.; Dai, M.; Cai, D.; Shi, Z. Proteome and transcriptome profile analysis reveals regulatory and stress-responsive networks in the russet fruit skin of sand pear. *Hortic. Res.* **2020**, *7*, 16. [CrossRef] [PubMed]
43. Zhu, Y.; Shao, J.; Zhou, Z.; Davis, R.E. Genotype-specific suppression of multiple defense pathways in apple root during infection by Pythium ultimum. *Hortic. Res.* **2019**, *6*, 10. [CrossRef] [PubMed]
44. Georges, M.; Charlier, C.; Hayes, B. Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* **2019**, *20*, 135–156. [CrossRef]