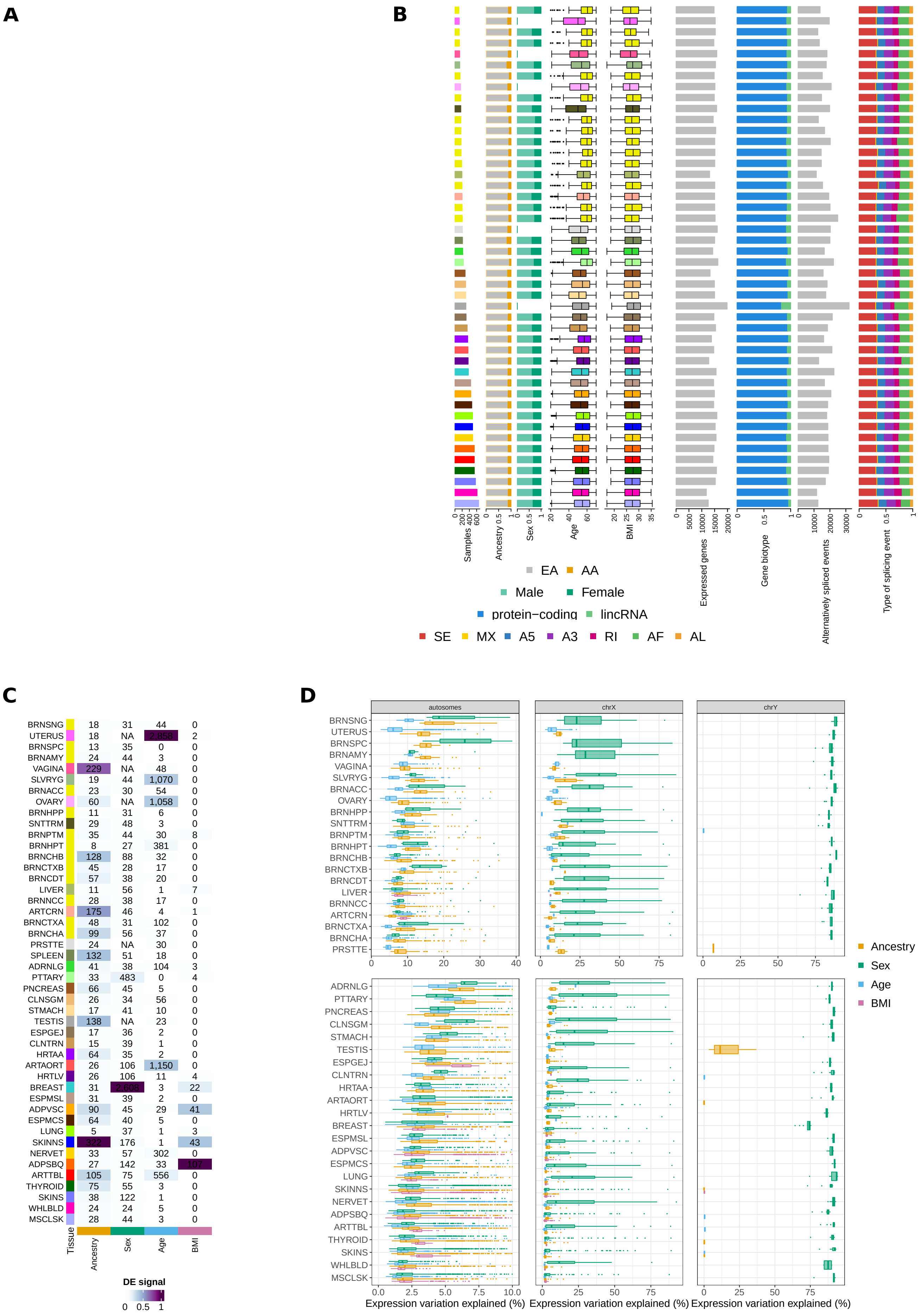
Cell Genomics, Volume *3*

Supplemental information

The landscape of expression and alternative splicing variation across human traits

Raquel García-Pérez, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, Paul Joris Rognon, Salvador Capella-Gutierrez, Miquel Calvo, Ferran Reverter, Roderic Guigó, François Aguet, Pedro G. Ferreira, Kristin G. Ardlie, and Marta Melé



Tissue	Ancestry	Sex	Age	BMI
BRNSNG	18	31	44	0
UTERUS	18	NA	2,858	2
BRNSPC	13	35	0	0
BRNAMY	24	44	3	0
VAGINA	229	NA	48	0
SLVRYG	19	44	1,070	0
BRNACC	23	30	54	0
OVARY	60	NA	1,058	0
BRNHPP	11	31	6	0
SNTTRM	29	48	3	0
BRNPTM	35	44	30	8
BRNHPT	8	27	381	0
BRNCHB	128	88	32	0
BRNCTXB	45	28	17	0
BRNCDT	57	38	20	0
LIVER	11	56	1	7
BRNNCC	28	38	17	0
ARTCRN	175	46	4	1
BRNCTXA	48	31	102	0
BRNCHA	99	56	37	0
PRSTTE	24	NA	30	0
SPLEEN	132	51	18	0
ADRNLG	41	38	104	3
PTTARY	33	483	0	4
PNCREAS	66	45	5	0
CLNSGM	26	34	56	0
STMACH	17	41	10	0
TESTIS	138	NA	23	0
ESPGEJ	17	36	2	0
CLNTRN	15	39	1	0
HRTAA	64	35	2	0
ARTAORT	26	106	1,150	0
HRTLV	26	106	11	4
BREAST	31	2,608	3	22
ESPMSL	31	39	2	0
ADPVSC	90	45	29	41
ESPMCS	64	40	5	0
LUNG	5	37	1	3
SKINNS	322	176	1	43
NERVET	33	57	302	0
ADPSBQ	27	142	33	107
ARTTBL	105	75	556	0
THYROID	75	55	3	0
SKINS	38	122	1	0
WHLBLD	24	24	5	0
MSCLSK	28	44	3	0

DE signal

0 0.5 1

Figure S1. Gene expression variation explained by demographic traits, Related to Figure 1. A, Tissues (including 11 distinct brain regions) with more than 100 RNA-seq samples from GTEx v8 genotyped donors. Tissue names, abbreviations, and color coding are indicated for each tissue. **B**, Descriptive summary per tissue of the demographic traits, number of genes expressed and their biotype, and number of alternative spliced events (ASEs) and their type. Tissues are sorted by sample size. EA: European Americans; AA; African Americans; SE: exon skipping; MX: mutually exclusive exons; A5: alternative 5 prime; A3: alternative 3 prime; RI: retained intron; AF: alternative first exon; AL: alternative last exon. **C**, Number of differentially expressed genes (DEGs) per tissue and demographic trait when downsampling to 100 tissue samples. Cell numbers indicate the average number of DEGs from 10 random tissue subsets. Heatmap cell colors are normalized to maximum value per column. **D**, Gene expression variation explained in each tissue by each demographic trait. Demographic traits are color-coded. Tissues are sorted by sample size. Box plots represent the first to third quartiles and the median gene expression variation explained, and the whiskers indicate ± 1.5 x interquartile range (IQR). From left to right, expression variation explained for autosomal, chrX, and chrY genes. Median gene expression variation explained decreases with sample size since larger numbers of samples allow the detection of more subtle but significant expression differences.

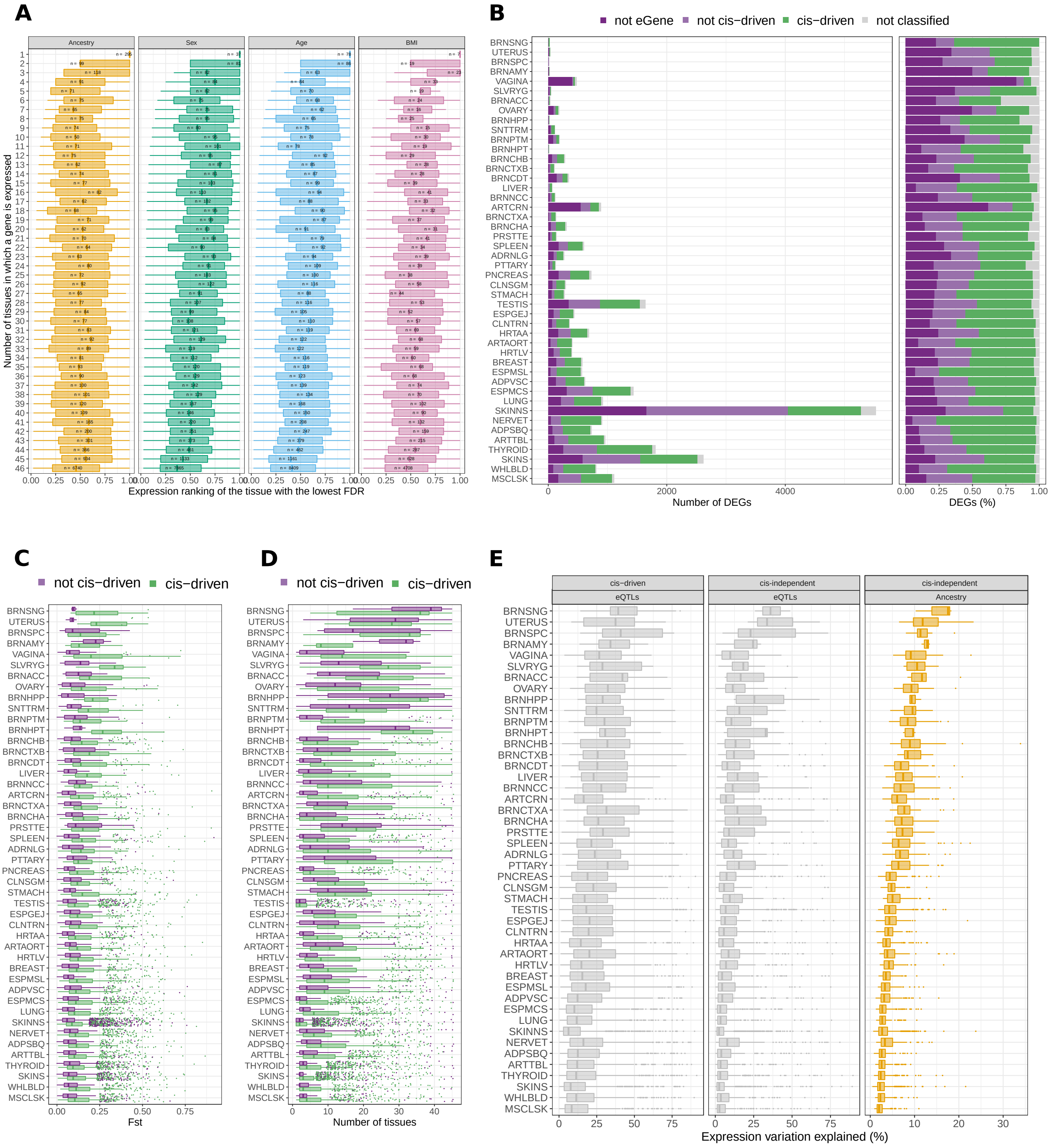
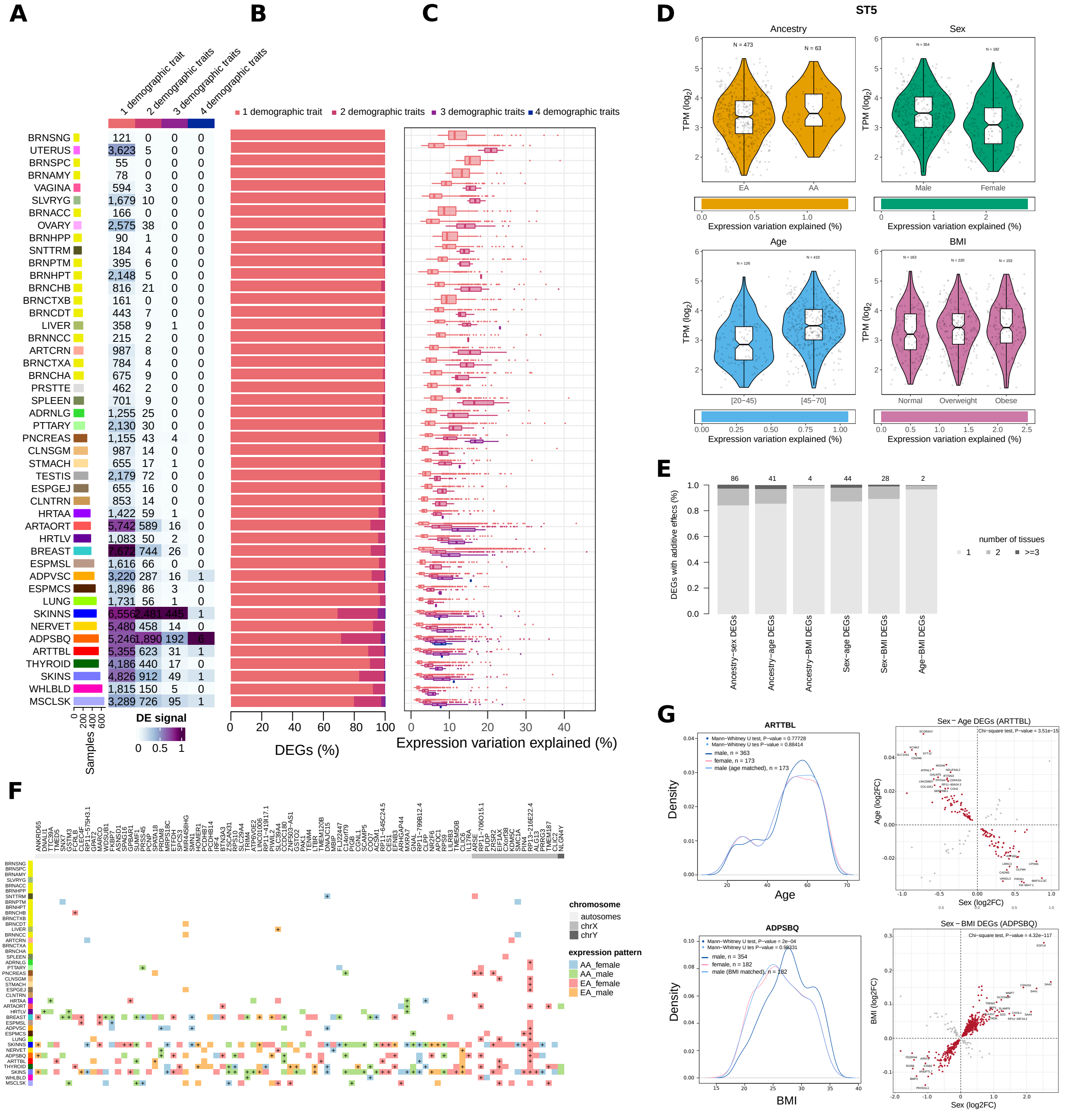


Figure S2. Differential expression is independent of expression level and characterization of *cis*-driven regulation of DEGs between populations, Related to Figure 2. A, Differential expression is not correlated with expression level. To compute the relative expression ranking, for each DEG, we ranked the tissues in which it was expressed based on the gene median expression and then selected the ranked value of the tissue where the DEG had the lowest FDR. Gene rankings were normalized taking into account the number of tissues in which a gene is expressed. The number of genes in each box plot corresponds to the number of DEGs with that trait and expressed in the number of tissues specified in the y-axis. Box plots show the distribution of their relative expression ranking. In general, genes are not DE in the tissue where they are expressed the most (median relative expression ranking != 1). **B**, Number (left) and proportion (right) of not eGenes and *cis*-driven, *cis*-independent, and unclassified DEGs between populations (STAR methods). **C**, Box plots show the distribution of the average Fst value of *cis*-eQTLs for *cis*-driven and *cis*-independent DEGs between populations (one-tailed Mann-Whitney U test, FDR < 0.05 in 36 tissues). **D**, Box plots show the tissue sharing distribution for *cis*-driven and *cis*-independent genes DE between populations (one-tailed Mann-Whitney U test, FDR < 0.05 in 23 tissues). **E**, Gene expression variation explained in each tissue by *cis*-eQTLs (in *cis*-driven and *cis*-independent genes) and by ancestry (in *cis*-independent genes). Box plots represent the first to third quartiles and the median gene expression variation explained, and the whiskers indicate ± 1.5 x interquartile range (IQR).



	1 demographic trait	2 demographic traits	3 demographic traits	4 demographic traits
BRNSNG	121	0	0	0
UTERUS	3,623	5	0	0
BRNSPC	55	0	0	0
BRNAMY	78	0	0	0
VAGINA	594	3	0	0
SLVRYG	1,679	10	0	0
BRNACC	166	0	0	0
OVARY	2,575	38	0	0
BRNHPP	90	1	0	0
SNTTRM	184	4	0	0
BRNPTM	395	6	0	0
BRNHPT	2,148	5	0	0
BRNCHB	816	21	0	0
BRNCTXB	161	0	0	0
BRNCDT	443	7	0	0
LIVER	358	9	1	0
BRNNCC	215	2	0	0
ARTCRN	987	8	0	0
BRNCTXA	784	4	0	0
BRNCHA	675	9	0	0
PRSTTE	462	2	0	0
SPLEEN	701	9	0	0
ADRNLG	1,255	25	0	0
PTTARY	2,130	30	0	0
PNCREAS	1,155	43	4	0
CLNSGM	987	14	0	0
STMACH	655	17	1	0
TESTIS	2,179	72	0	0
ESPGEJ	655	16	0	0
CLNTRN	853	14	0	0
HRTAA	1,422	59	1	0
ARTAORT	5,742	589	16	0
HRTLV	1,083	50	2	0
BREAST	7,672	744	26	0
ESPMSL	1,616	66	0	0
ADPVSC	3,220	287	16	1
ESPMCS	1,896	86	3	0
LUNG	1,731	56	1	0
SKINNS	6,556	2,481	445	1
NERVET	5,480	458	14	0
ADPSBQ	5,246	1,890	192	6
ARTTBL	5,355	623	31	1
THYROID	4,186	440	17	0
SKINS	4,826	912	49	1
WHLBLD	1,815	150	5	0
MSCLSK	3,289	726	95	1

Figure S3. Differential expression is independent of expression level and breadth and characterization of additive contributions across tissues, Related to Figure 3. A, Number of genes DE with one, two, three, or four demographic traits per tissue. Heatmap cell colors are normalized to the maximum value per column. **B**, Proportion of the total tissue expression variation explained by genes DE with one, two, three, or four demographic traits. **C**, Gene expression variation explained for autosomal genes DE with one, two, three, or four demographic traits. Box plots represent the first to third quartiles and the median gene expression variation explained, and the whiskers indicate ± 1.5 x interquartile range (IQR). **D**, The FGGY gene is DE in muscle with the four demographic traits**.** Gene expression levels are represented as box plots with samples stratified by population, sex, age, or BMI. Violin plots show the gene expression distribution. Points correspond to individual gene expression levels. The number of individuals in each group is shown within the plot. Bars at the bottom indicate the proportion of expression variation explained by each demographic trait. **E**, Proportion of genes DE with any two given demographic traits that are tissue-specific, DE in two tissues or in 3 or more tissues. Numbers on top of the bars indicate the number of genes in the latter category. **F**, Heatmap shows the expression pattern of genes DE between populations and sexes in more than 2 tissues. The colors indicate the directionality of the expression change. Asterisks mark genes that were classified as *cis*-driven ancestry-DEGs. **G**, Biased directionality in genes with additive contributions is not confounded by differences in age or BMI between sexes. Left panels show the age and BMI distribution of males and females in tibial artery and adipose subcutaneous tissue, respectively, before and after matching the distributions. Differential expression analysis using a subset of male and female samples with matching age and BMI distributions replicate our initial findings (see Figure 3).

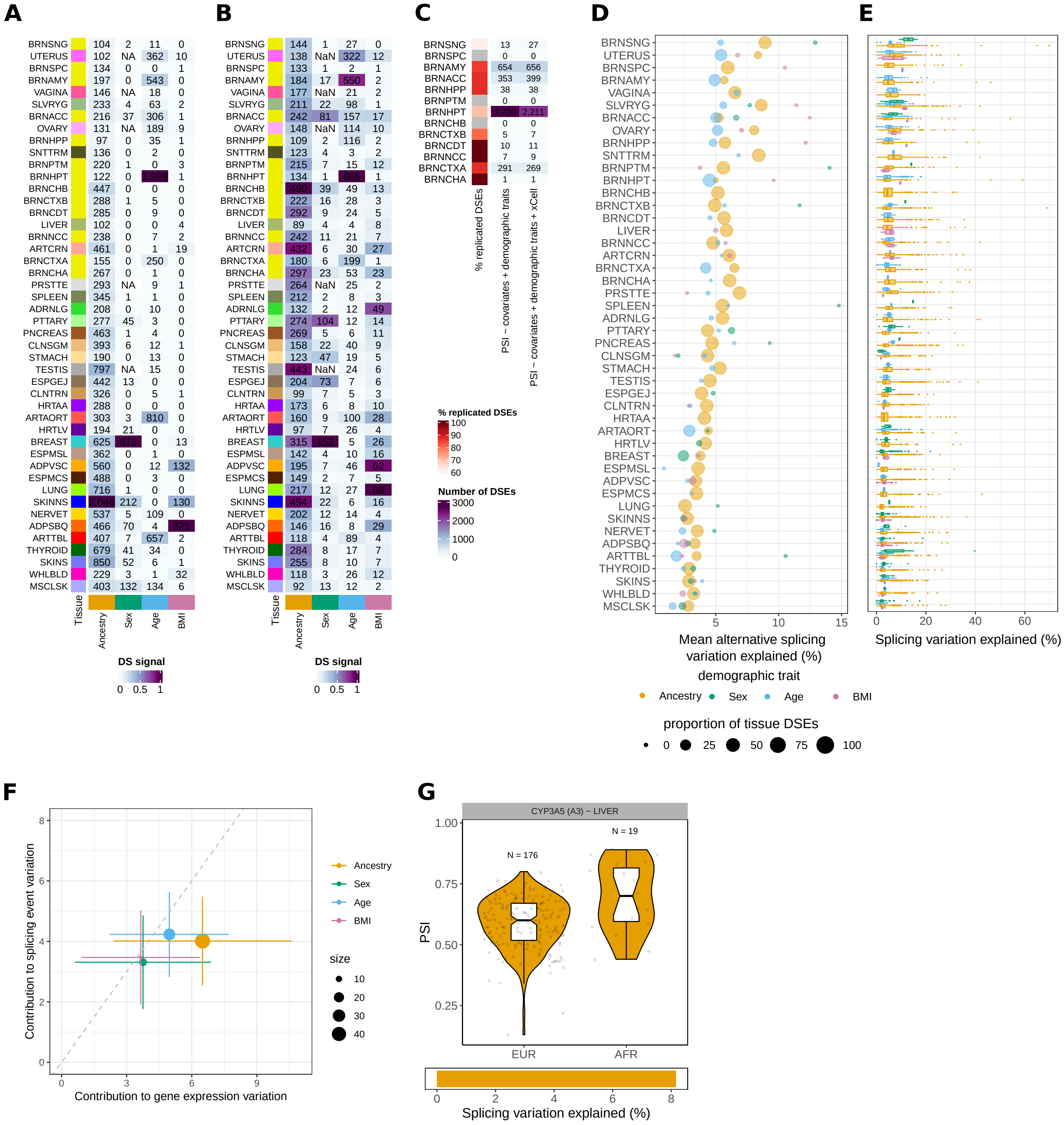


Figure S4. Alternative splicing variation explained by demographic traits, Related to Figure 4. A, Number of genes differentially spliced (DSGs) per tissue and demographic trait. Heatmap cell colors are normalized to maximum value per column. **B**, Number of differentially spliced events (DSEs) per tissue and demographic trait when downsampling to N=100. Cell numbers indicate the average number of DSEs from 10 random tissue sample subsets. Heatmap cell colors are normalized to maximum value per column. **C**, Comparison of the number of DSEs with age in brain regions, after controlling for neuron abundances in the generalized linear model. Most DSEs are replicated. **D**, Mean alternative splicing variation explained by each demographic trait in each tissue. **E**, Alternative splicing variation explained in each tissue by each demographic trait. Demographic traits are color-coded. Box plots represent the first to third quartiles and the median gene expression variation explained, and the whiskers indicate ± 1.5 x interquartile range (IQR). **F**, Comparison of the contribution of each demographic trait to gene expression and alternative splicing variation. For each trait, average value across tissues is plotted. Error bars correspond to the standard deviation. For each demographic trait, we only considered tissues with at least 5 DEGs and 5 DSEs. **G**, Splicing event with a large proportion of its alternative splicing variation explained by ancestry. Violin plots show PSI value distributions stratified by ancestry. Box plots represent the first to third quartiles and the median, and the whiskers indicate ± 1.5 x interquartile range (IQR). Bottom bar indicates the proportion of alternative splicing variation explained by ancestry.

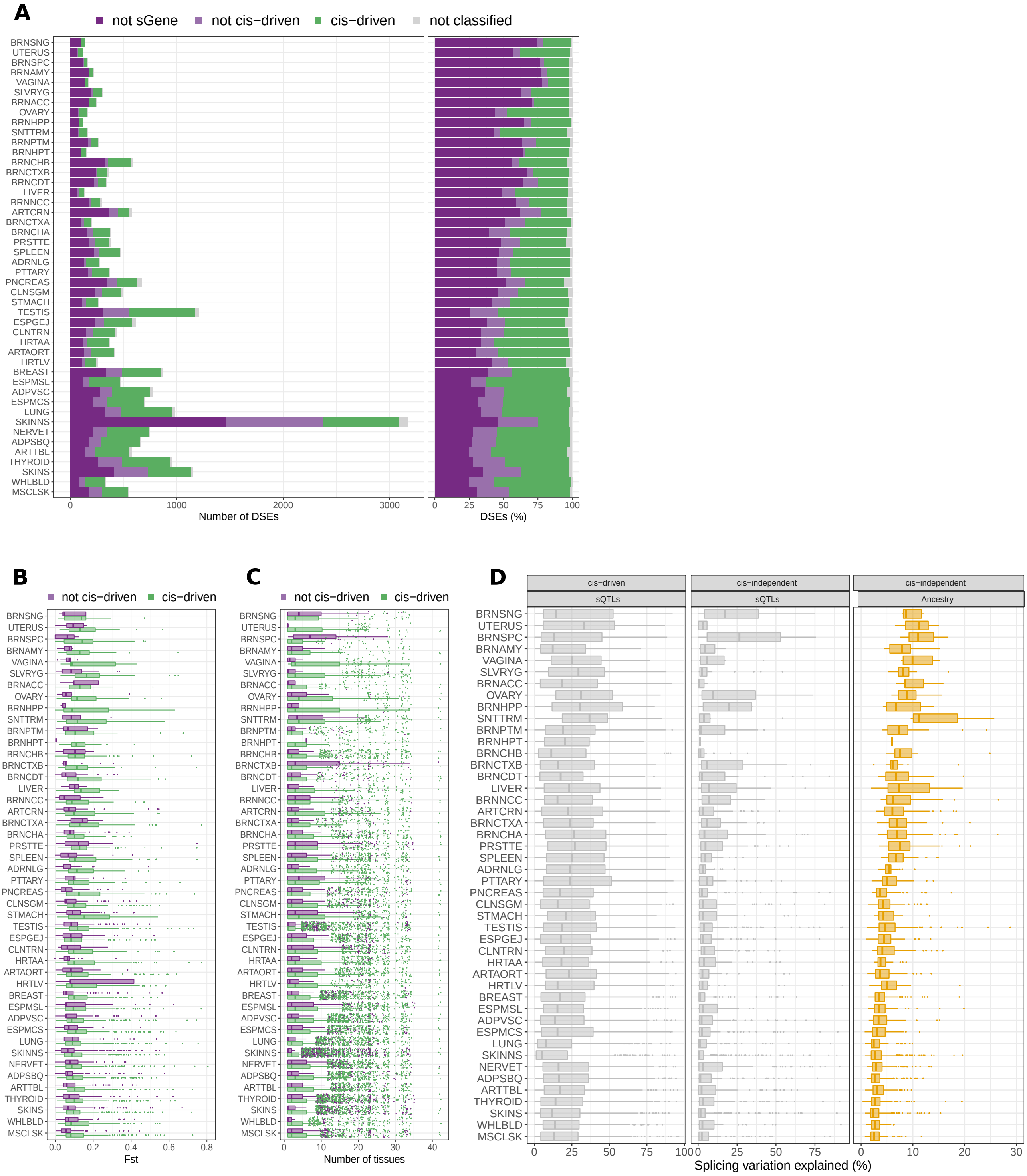


Figure S5. Characterization of *cis*-driven regulation of DSEs between populations, Related to Figure 5. A, Number (left) and proportion (right) of DSEs in not sGenes and *cis*-driven, *cis*-independent, and unclassified DSEs between populations (STAR methods). **B**, Box plots show the distribution of the average Fst value of *cis*-sQTLs for *cis*-driven and *cis*-independent DSEs between populations (one-tailed Mann-Whitney U test, FDR < 0.05 in 36 tissues). **C**, Box plots show the tissue sharing distribution for *cis*-driven and *cis*-independent DSEs between populations (one-tailed Mann-Whitney U test, FDR < 0.05 in 26 tissues). **D**, Alternative splicing variation explained in each tissue by *cis*-sQTLs (in *cis*-driven and *cis*-independent DSEs) and by ancestry (in *cis*-independent DSEs). Box plots represent the first to third quartiles and the median gene expression variation explained, and the whiskers indicate ± 1.5 x interquartile range (IQR).

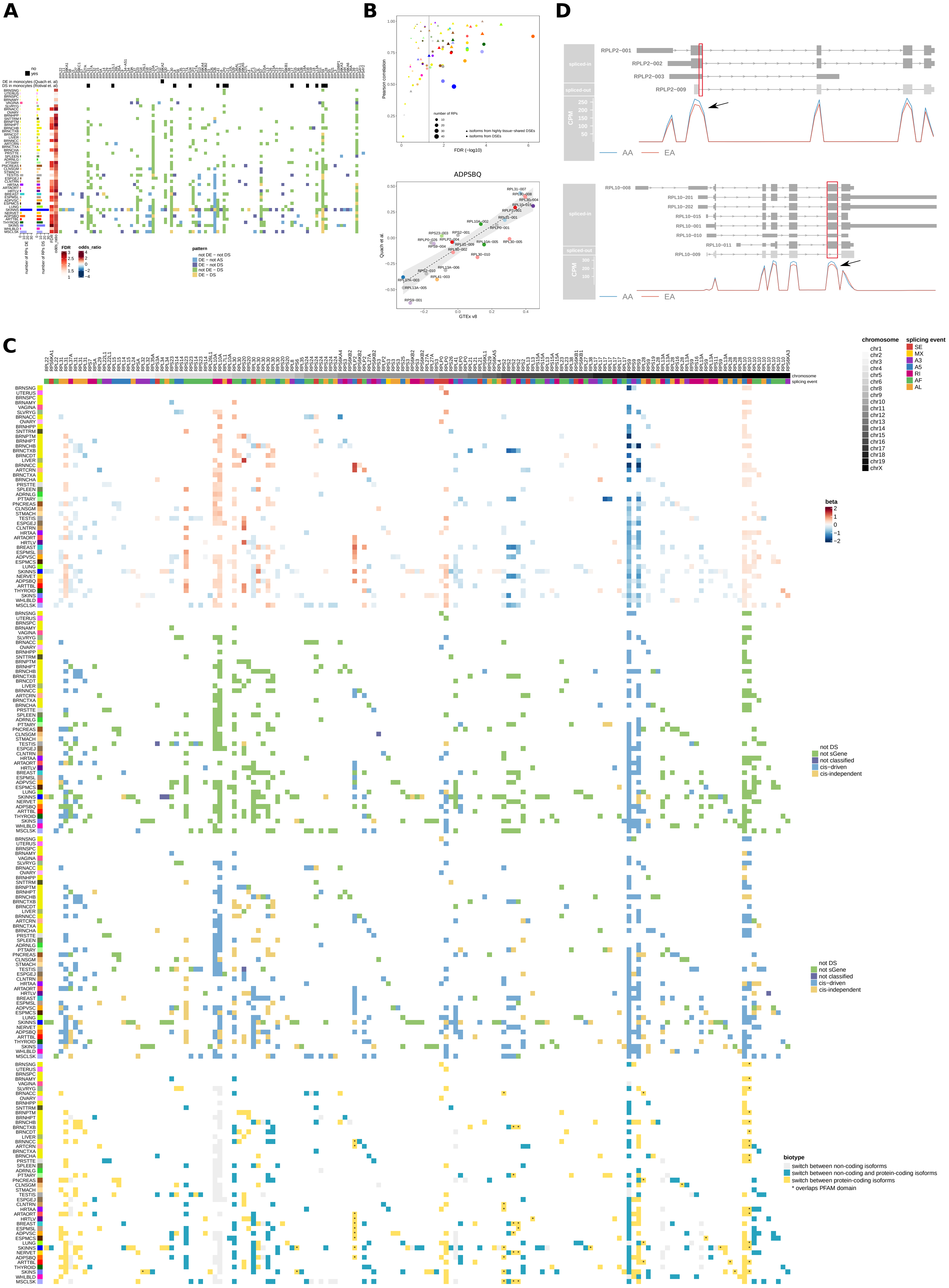
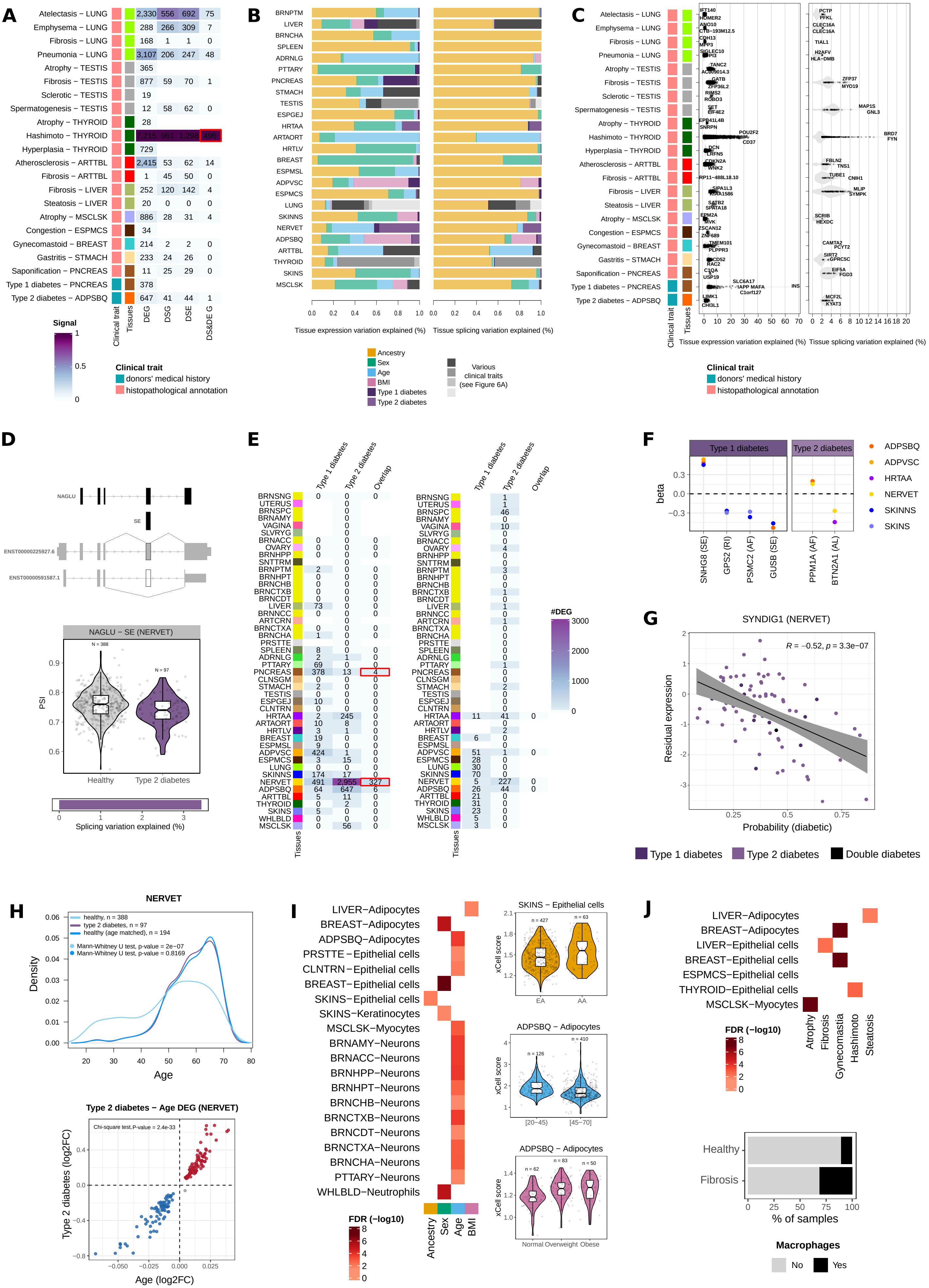

A
no
yes
DE in monocytes (Quach et. al)
DS in monocytes (Rotival et. al)
BRNSNG
UTERUS
BRNSPC
BRNAMY
VAGINA
SLVRYG
BRNACC
OVARY
BRNHPP
SNTTRM
BRNPTM
BRNHPT
BRNCHB
BRNCTXB
BRNCDT
LIVER
BRNNCC
ARTCRN
BRNCTXA
BRNCHA
PRSTTE
SPLEEN
ADRNLG
PTTARY
PNCREAS
CLNSGM
STMACH
TESTIS
ESPGEJ
CLNTRN
HRTAA
ARTAORT
HRTLV
BREAST
ESPMSL
ADPVSC
ESPMCS
LUNG
SKINNS
NERVET
ADPSBQ
ARTTBL
THYROID
SKINS
WHLBLD
MSCLSK
number of RPs DE
number of RPs DS
FDR
odds_ratio
pattern
not DE − not DS
DE − not AS
DE − not DS
not DE − DS
DE − DS
B
Pearson correlation
FDR (−log10)
number of RPs
10
20
30
40
isoforms from highly tissue−shared DSEs
isoforms from DSEs
ADPSBQ
Quach et al.
GTEx v8
RPL31−007
RPL30−008
RPL30−004
RPL10−015
RPLP2−001
RPL31−001
RPL10A−002
RPLP0−001
RPS23−003
RPS2−001
RPLP0−026
RPLP2−004
RPS9−004
RPL31−009
RPL30−002
RPL10A−005
RPL30−005
RPL30−010
RPL13A−006
RPS2−010
RPL37A−003
RPL41−003
RPL13A−005
RPS9−001
C
chromosome
splicing event
chr1
chr2
chr3
chr4
chr5
chr6
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX
SE
MX
A3
A5
RI
AF
AL
beta
not DS
not sGene
not classified
cis−driven
cis−independent
biotype
switch between non-coding isoforms
switch between non-coding and protein-coding isoforms
switch between protein-coding isoforms
* overlaps PFAM domain
D
RPLP2−001
RPLP2−002
RPLP2−003
RPLP2−009
spliced-in
spliced-out
CPM
AA
EA
RPL10−008
RPL10−201
RPL10−202
RPL10−015
RPL10−001
RPL10−010
RPL10−011
RPL10−009

Figure S6. Characterization of differential splicing patterns between populations in ribosomal proteins, Related to Figure 5. A, Patterns of differential expression and splicing between human populations in ribosomal proteins across tissues. Bars on the left indicate the number of ribosomal proteins differentially expressed and spliced in each tissue. Row annotations show the odds ratio and FDR of a two-tailed fisher's exact test used to test if genes differentially spliced in a tissue were enriched in ribosomal proteins. Top annotations indicate the ribosomal proteins differentially expressed [S1] or differentially spliced in monocytes [S2]. Cell color indicates if a ribosomal protein is differentially expressed (DE), differentially expressed but not alternatively spliced (DE - not AS), not differentially expressed but differentially spliced (not DE - DS), or both (DE - DS). **B**, Differences in isoforms expression between European and African American across GTEx tissues are correlated with those observed in monocytes [S1] (left). We computed the differences in expression (effect sizes) between African Americans and European Americans for the isoforms contributing to the tissue DSEs and highly tissue-shared DSEs in ribosomal proteins. Scatter plots show, for each tissue, the Person correlations coefficient versus the adjusted p-value. Isoform expression differences between populations in adipose tissue (tissue with the lowest adjusted p-value) are correlated with those observed in monocytes [S1] (right). **C**, Characterization of splicing patterns in events differentially spliced between populations in ribosomal proteins. In the top panel, cell colors represent the beta values in the generalized regression models: positive betas correspond to events more included in African Americans; negative betas to splicing events more included in European Americans. Note how the beta values show the same directionality across tissues. The middle top panel indicates if the splicing events were classified as *cis*-driven or *cis*-independent. Note here that this analysis was run considering the genes reported as sGenes in the GTEx v8 main paper, and most ribosomal proteins were not sGenes (in green). The middle bottom panel indicates if the splicing event was classified as *cis*-driven or *cis*-independent using a generalized linear model that accounts for closeby variants (see STAR methods). The bottom panel indicates if the splicing event is associated with a switch between non-coding isoforms, a switch between a protein-coding and a non-coding isoform, or a switch between two functional protein-coding isoforms. For the latter, asterisks indicate that the inclusion or exclusion of the splicing event affects a PFAM domain. **D,** Read coverage plots spanning the highly tissue-shared events in ribosomal proteins RPLP2 (top) and RPL10 (bottom). Isoform models for the isoforms that include the exonic region (in dark gray) or exclude the exonic region (in light gray) are shown. The exonic regions that are either spliced-in or spliced-out are highlighted with a red box. The average normalized read coverage per population (AA: African American; EA: European American) in SKINS (top) and in SKINNS (bottom) is shown. CPM (per bin) = number of reads per bin / number of mapped reads (in millions). Please note that the isoforms in the plots can also participate in other splicing events involving adjacent exonic and intronic regions.



A
Atelectasis − LUNG
Emphysema − LUNG
Fibrosis − LUNG
Pneumonia − LUNG
Atrophy − TESTIS
Fibrosis − TESTIS
Sclerotic − TESTIS
Spermatogenesis − TESTIS
Atrophy − THYROID
Hashimoto − THYROID
Hyperplasia − THYROID
Atherosclerosis − ARTTBL
Fibrosis − ARTTBL
Fibrosis − LIVER
Steatosis − LIVER
Atrophy − MSCLSK
Congestion − ESPMCS
Gynecomastoid − BREAST
Gastritis − STMACH
Saponification − PNCREAS
Type 1 diabetes − PNCREAS
Type 2 diabetes − ADPSBQ
Clinical trait
Tissues
DEG
DSG
DSE
DS&DE G
Signal
1
0.5
0
Clinical trait
donors' medical history
histopathological annotation
B
Tissue expression variation explained (%)
Tissue splicing variation explained (%)
Ancestry
Sex
Age
BMI
Type 1 diabetes
Type 2 diabetes
Various clinical traits (see Figure 6A)
C
Tissue expression variation explained (%)
Tissue splicing variation explained (%)
D
NAGLU
SE
ENST00000225927.6
ENST00000591587.1
NAGLU − SE (NERVET)
PSI
Healthy
Type 2 diabetes
N = 388
N = 97
Splicing variation explained (%)
E
Type 1 diabetes
Type 2 diabetes
Overlap
Tissues
#DEG
3000
2000
1000
0
F
Type 1 diabetes
Type 2 diabetes
beta
SNHG8 (SE)
GPS2 (RI)
PSMC2 (AF)
GUSB (SE)
PPM1A (AF)
BTN2A1 (AL)
ADPSBQ
ADPVSC
HRTAA
NERVET
SKINNS
SKINS
G
SYNDIG1 (NERVET)
R = −0.52, p = 3.3e−07
Residual expression
Probability (diabetic)
Type 1 diabetes
Type 2 diabetes
Double diabetes
H
NERVET
healthy, n = 388
type 2 diabetes, n = 97
healthy (age matched), n = 194
Mann-Whitney U test, p-value = 2e−07
Mann-Whitney U test, p-value = 0.8169
Density
Age
Type 2 diabetes − Age DEG (NERVET)
Chi-square test P-value = 2.4e-33
Type 2 diabetes (log2FC)
Age (log2FC)
I
LIVER−Adipocytes
BREAST−Adipocytes
ADPSBQ−Adipocytes
PRSTTE−Epithelial cells
CLNTRN−Epithelial cells
BREAST−Epithelial cells
SKINS−Epithelial cells
SKINS−Keratinocytes
MSCLSK−Myocytes
BRNAMY−Neurons
BRNACC−Neurons
BRNHPP−Neurons
BRNHPT−Neurons
BRNCHB−Neurons
BRNCTXB−Neurons
BRNCDT−Neurons
BRNCTXA−Neurons
BRNCHA−Neurons
PTTARY−Neurons
WHLBLD−Neutrophils
FDR (−log10)
Ancestry
Sex
Age
BMI
SKINS − Epithelial cells
xCell score
EA
AA
ADPSBQ − Adipocytes
[20−45)
[45−70]
Normal
Overweight
Obese
J
LIVER−Adipocytes
BREAST−Adipocytes
LIVER−Epithelial cells
BREAST−Epithelial cells
ESPMCS−Epithelial cells
THYROID−Epithelial cells
MSCLSK−Myocytes
Atrophy
Fibrosis
Gynecomastia
Hashimoto
Steatosis
FDR (−log10)
Healthy
Fibrosis
% of samples
Macrophages
No
Yes

Figure S7. Differential gene expression and alternative splicing with clinical traits and differential cellular abundance, Related to Figure 6. A, Number of DEGs, DSEs, DSGs, and genes both DE and DS per tissue and clinical trait. The red square highlights the only tissue-clinical trait pair where the number of genes DE and DS is significantly larger than expected by chance (two-tailed Fisher's exact test, p-value = 0.0011). **B**, Proportion out of the total tissue expression (left) and splicing (right) variation explained by each demographic and clinical trait. **C**, Distribution of expression (left) and splicing (right) variation explained per gene/splicing event, tissue, and clinical trait. Labels indicate the name of the top 2 genes (top 5 for type 1 diabetes). **D**, Example of a DSE with type 2 diabetes with functional consequences. (Top) Protein domain (NAGLU) affected when excluding an exon (SE) included in the most abundant isoform in healthy individuals (ENST00000225927.6), but excluded in the most abundant isoform of diabetic individuals (ENST00000591587.1). Violin plots (bottom) show the PSI distribution in each population. Points correspond to individual PSI values. **E**, Number of DEGs (left) and DSGs (right) per tissue with type 1 diabetes, type 2 diabetes and with both types of diabetes (overlap). The red squares highlight the tissues with significantly larger than expected DEGs with both diabetes (two-tailed Fisher's exact test, FDR < 0.05). **F**, DSEs with type 1 or 2 diabetes in two or more tissues. The y-axis corresponds to the beta estimate between healthy and diseased individuals per tissue. Tissues are color-coded as in Figure S1. **G**, SYNDIG1 residual expression in diabetic individuals is significantly correlated with their probability of being classified as diabetic by our classifier. **H**, Biased directionality in genes with additive contributions from age and type 2 diabetes are not confounded by differences in age between healthy and diabetic individuals. Left panel shows the age distribution of healthy and diabetic individuals in the tibial nerve, before and after matching age distributions between groups. The p-values correspond to the comparisons between each healthy age distribution against the diabetic age distribution. Differential expression analysis using the subset of age matched samples (right panel) replicates our initial findings (Figure 6H). **I**, Changes in tissue cell type composition associated with demographic traits. Tissue-cell type pairs with significant differences (FDR < 0.05) with at least one demographic trait (left). Three examples of changes in cell type abundance (right). Cell values are proportional to the statistical significance. **J**, Changes in tissue cell type composition associated with clinical traits (top). Example of changes in cell type abundance with type 1 diabetes (middle). Changes in the presence of macrophages in individuals with fibrosis in the lung (two-tailed Fisher's exact test, p-value = 3.045e-07).

REFERENCES

S1. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell *167*, 643–656.e17. 10.1016/j.cell.2016.09.024.

S2. Rotival, M., Quach, H., and Quintana-Murci, L. (2019). Defining the genetic and evolutionary architecture of alternative splicing in response to infection. Nat. Commun. *10*, 1671. 10.1038/s41467-019-09689-7.