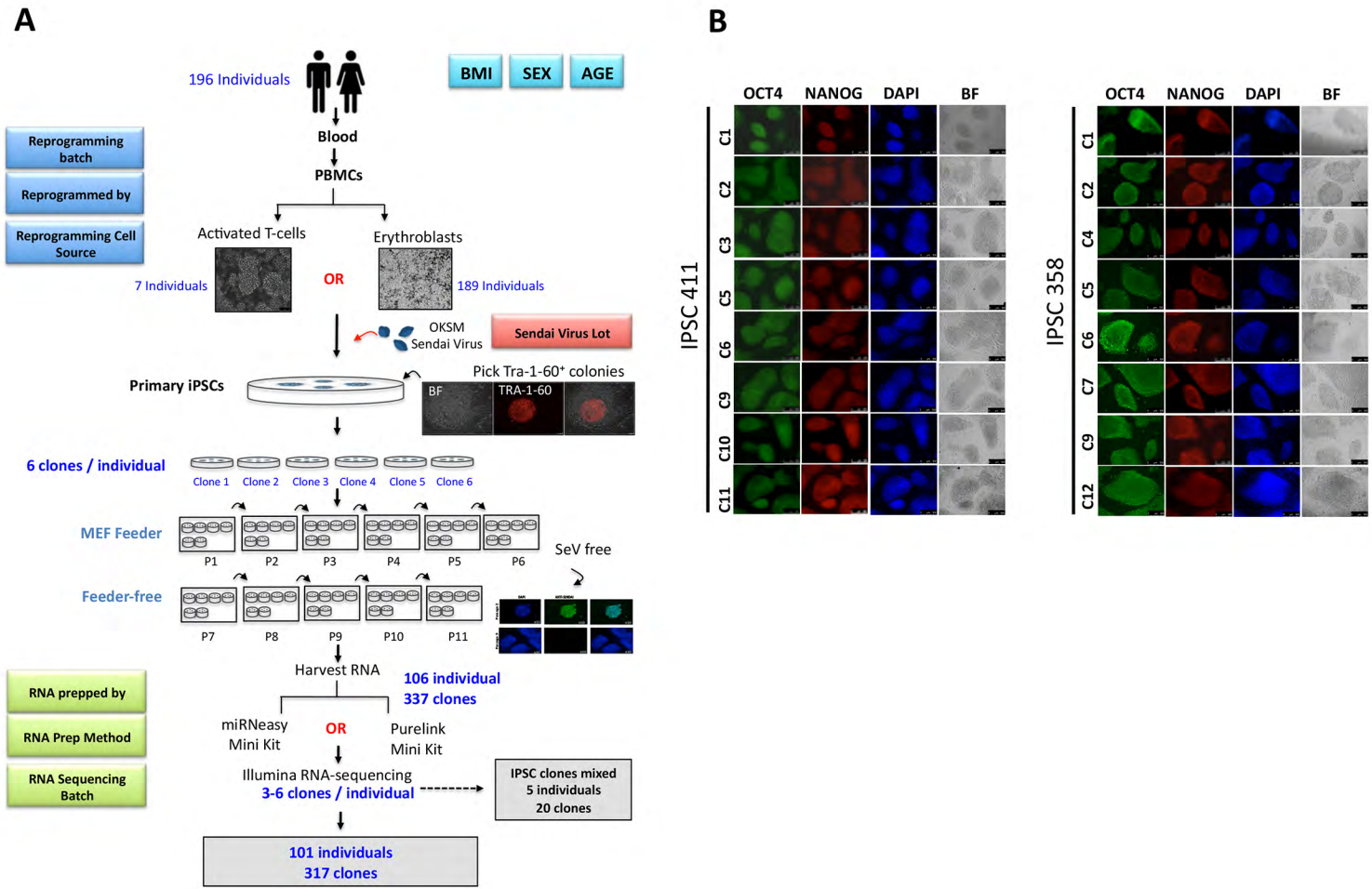
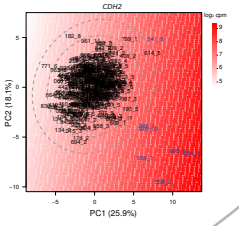
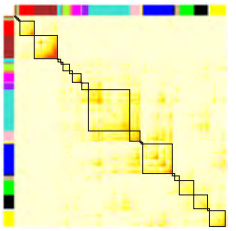
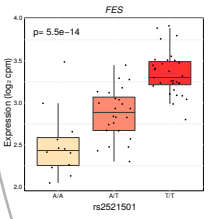

A
196 Individuals
BMI
SEX
AGE
Blood
PBMCs
Reprogramming batch
Reprogrammed by
Reprogramming Cell Source
Activated T-cells
Erythroblasts
7 Individuals
OR
189 Individuals
OKSM Sendai Virus
Sendai Virus Lot
Primary iPSCs
Pick Tra-1-60+ colonies
BF
TRA-1-60
6 clones / individual
Clone 1
Clone 2
Clone 3
Clone 4
Clone 5
Clone 6
MEF Feeder
P1
P2
P3
P4
P5
P6
SeV free
Feeder-free
P7
P8
P9
P10
P11
Harvest RNA
106 individual
337 clones
RNA prepped by
RNA Prep Method
RNA Sequencing Batch
miRNeasy Mini Kit
OR
Purelink Mini Kit
Illumina RNA-sequencing
3-6 clones / individual
IPSC clones mixed
5 individuals
20 clones
101 individuals
317 clones
B
OCT4
NANOG
DAPI
BF
IPSC 411
C1
C2
C3
C5
C6
C9
C10
C11
IPSC 358
C1
C2
C4
C5
C6
C7
C9
C12

A

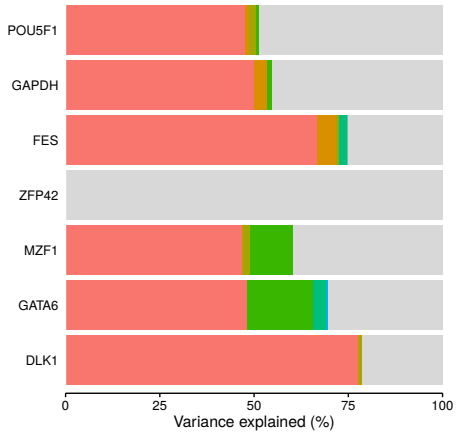
FASTQ
STAR
BAM
gatk
VCF
Imputation
shapeit2
impute2
GWAS genotypes
RNA-SeQC
featureCounts
Quality control
Quantify expression: gene-level counts
Concordance analysis: genotype and RNA-Seq
Identify outlier clones: based on stem cell genes
limma/voom
samtools
Allele specific expression
Normalized expression
MatrixEQTL
eQTL analysis
MDS
WGCNA / coexpp
Co-expression networks
Relationship to external datasets
limma/voom
Differential expression
Hierarchical clustering
variancePartition
Magnitude of variance
Gene Ontology
MSigDB
Gene enrichment analysis
Characterize stem cell heterogeneity
Predictive causal networks



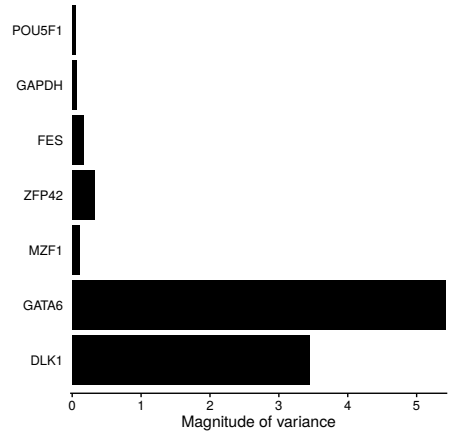




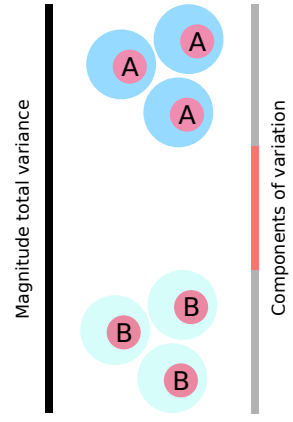
B



C

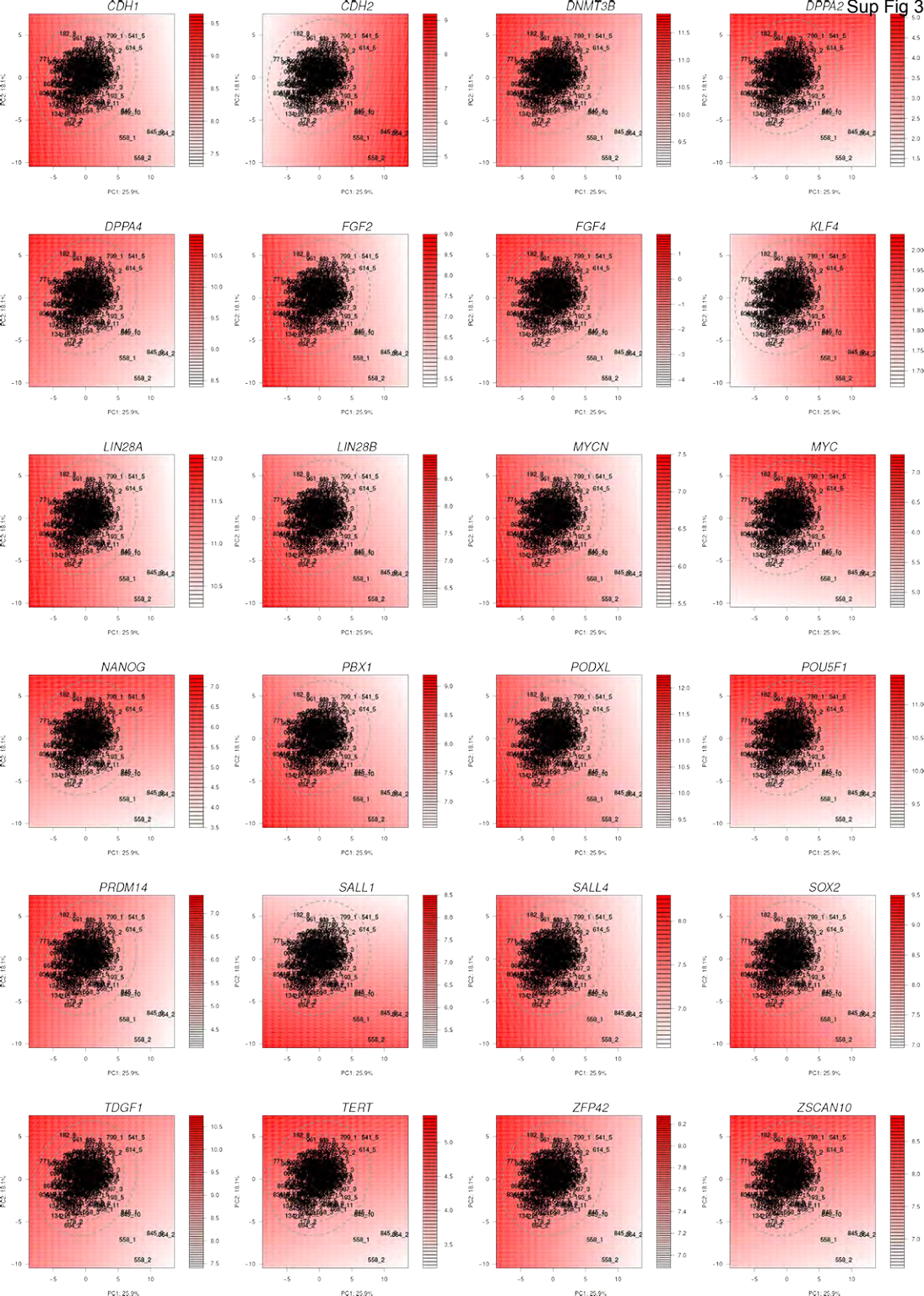


D

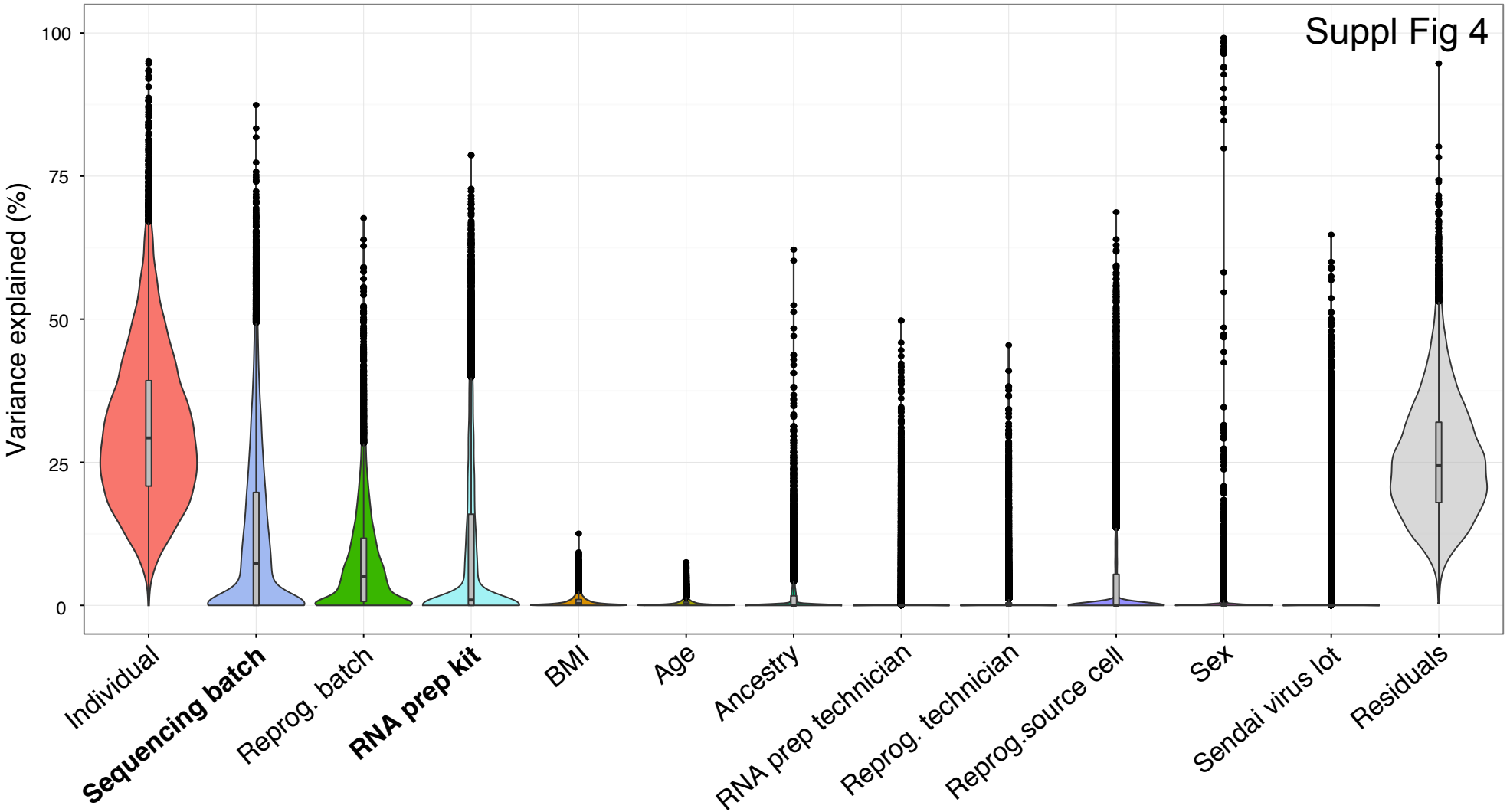


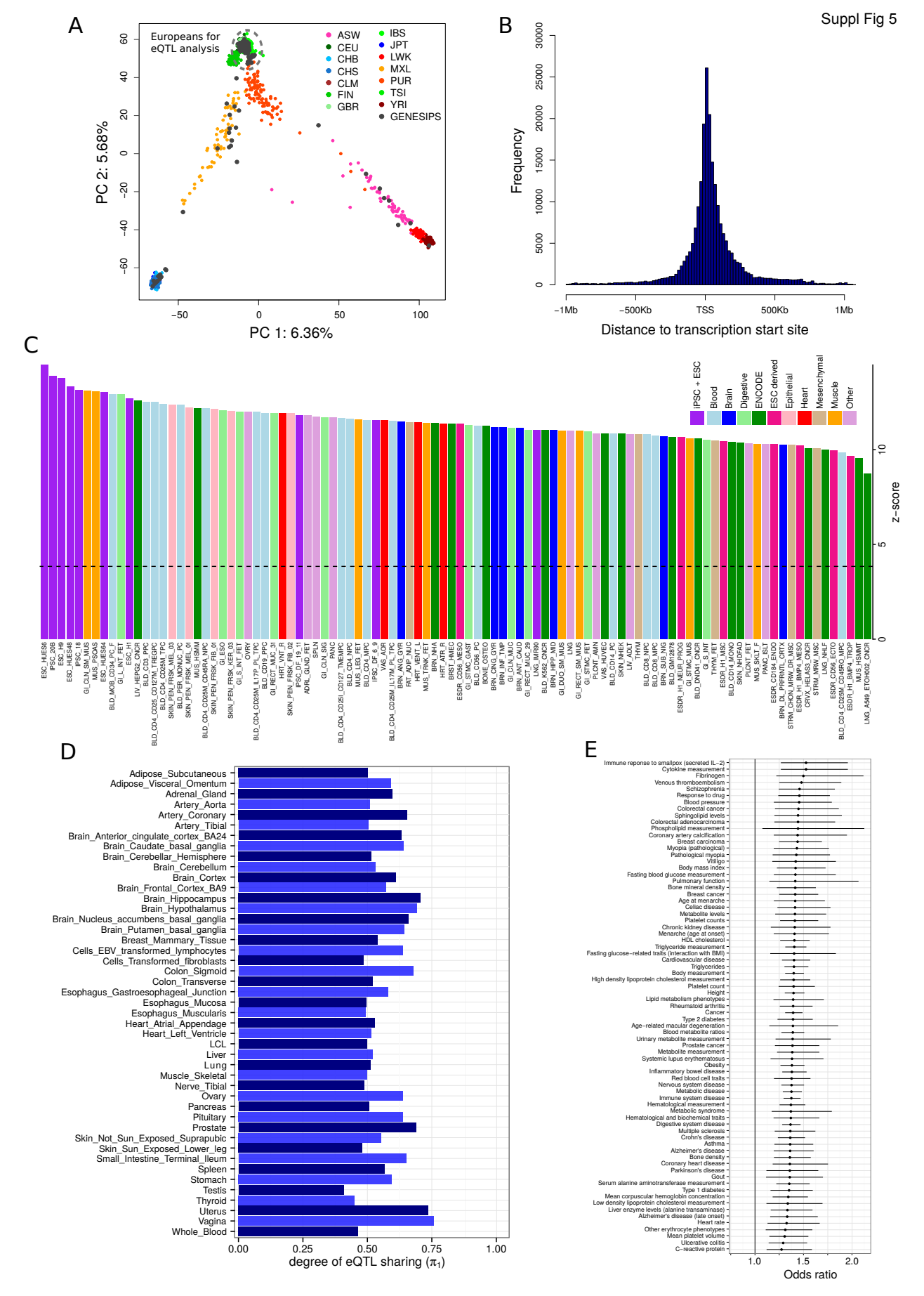
% variance **across** individuals = Magnitude of variance **across** individuals / Magnitude total variance x 100

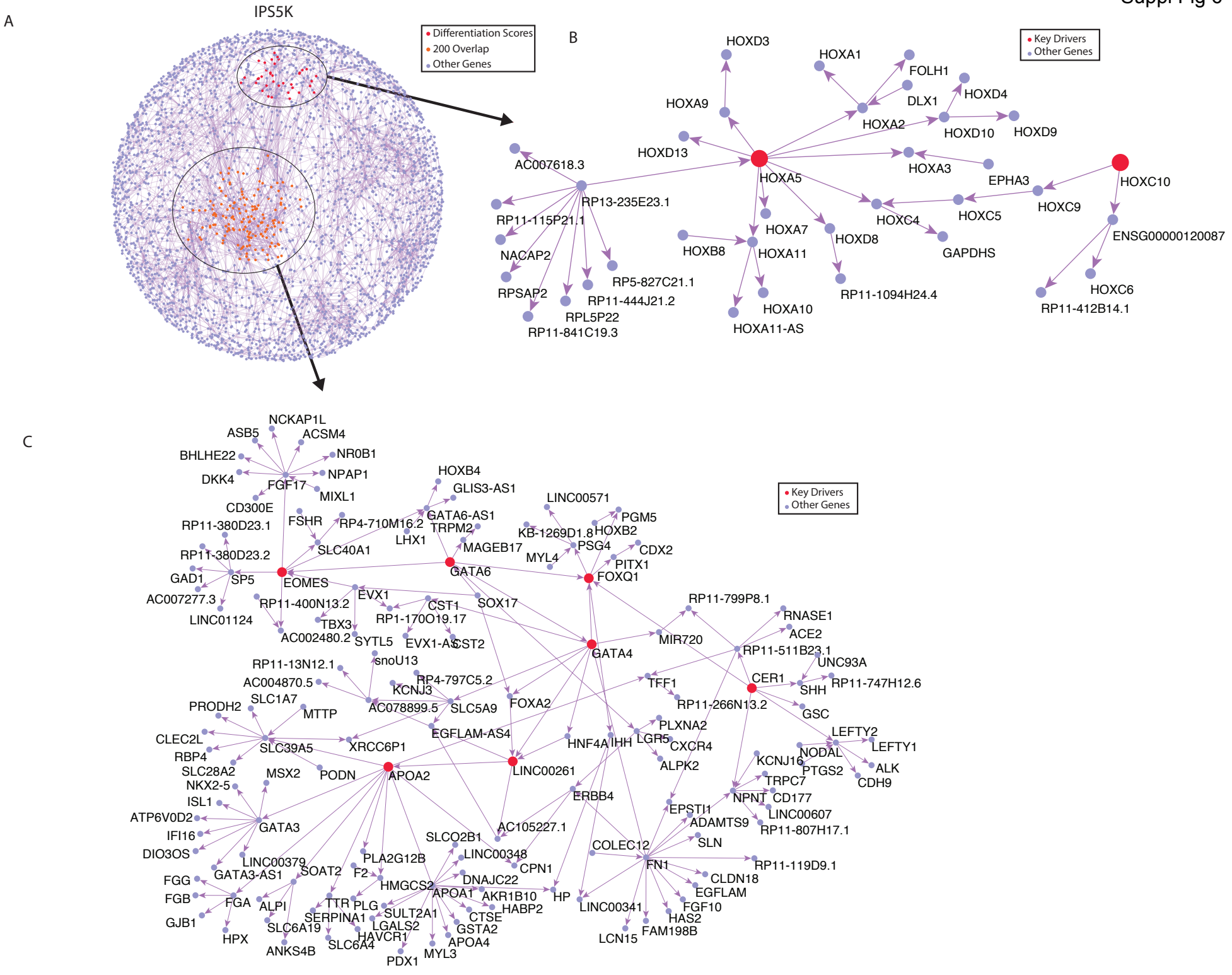
% variance **within** individuals = Magnitude of variance **within** individuals / Magnitude total variance x 100

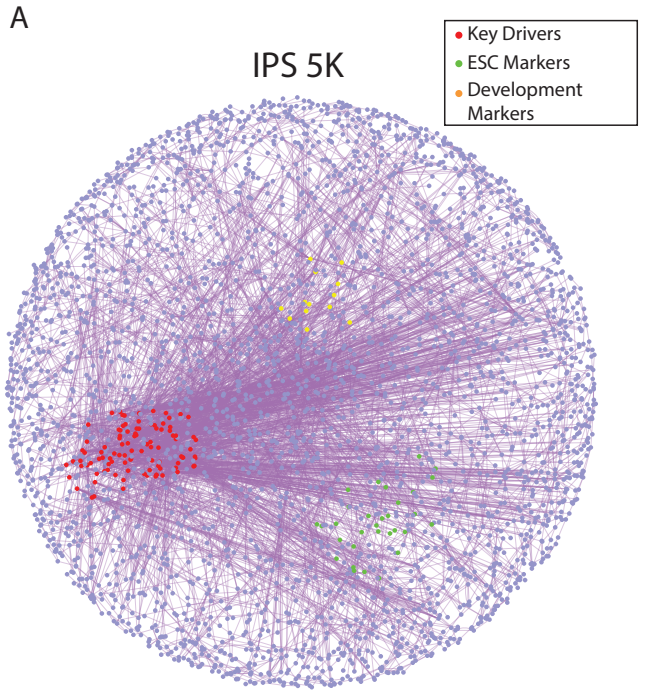

CDH1
CDH2
DNMT3B
DPPA2
DPPA4
FGF2
FGF4
KLF4
LIN28A
LIN28B
MYCN
MYC
NANOG
PBX1
PODXL
POU5F1
PRDM14
SALL1
SALL4
SOX2
TDGF1
TERT
ZFP42
ZSCAN10
PC1: 25.9%
PC2: 18.1%

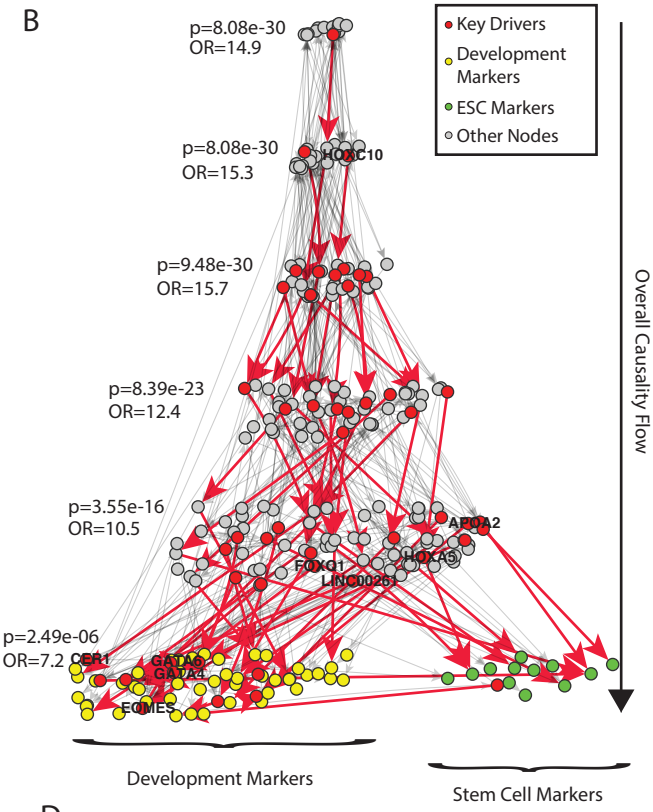
Sup Fig 3

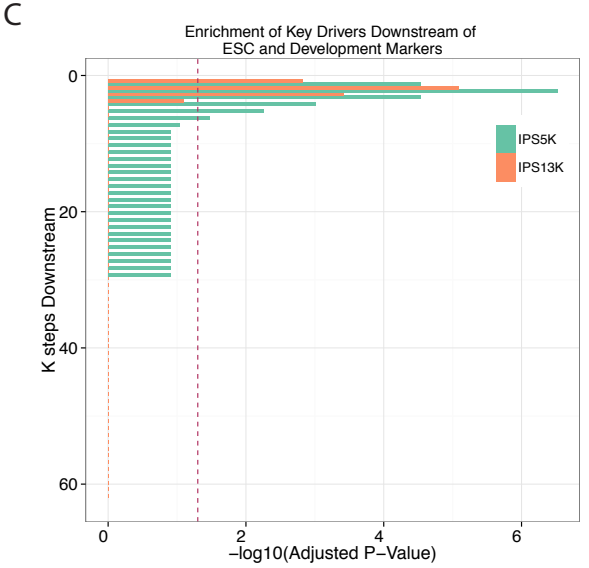

Suppl Fig 4
Variance explained (%)
100
75
50
25
0
Individual
Sequencing batch
Reprog. batch
RNA prep kit
BMI
Age
Ancestry
RNA prep technician
Reprog. technician
Reprog.source cell
Sex
Sendai virus lot
Residuals

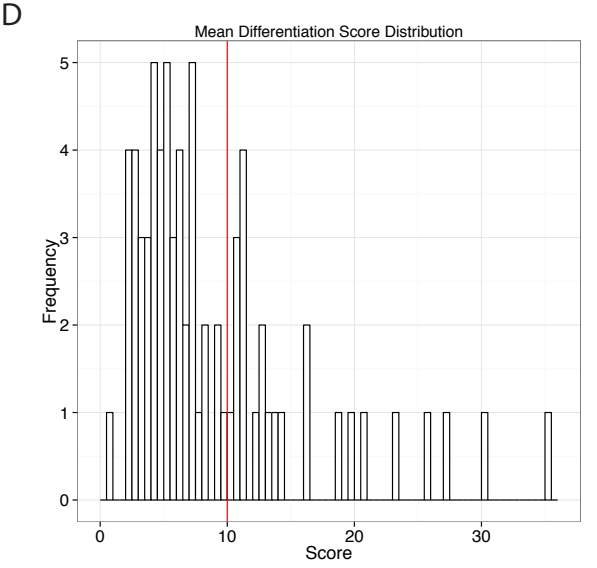

A
Europeans for eQTL analysis
ASW
CEU
CHB
CHS
CLM
FIN
GBR
IBS
JPT
LWK
MXL
PUR
TSI
YRI
GENESIPS
PC 2: 5.68%
PC 1: 6.36%
B
Frequency
Distance to transcription start site
−1Mb
−500Kb
TSS
500Kb
1Mb
C
iPSC + ESC
Blood
Brain
Digestive
ENCODE
ESC derived
Epithelial
Heart
Mesenchymal
Muscle
Other
z−score
D
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
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
Breast_Mammary_Tissue
Cells_EBV−transformed_lymphocytes
Cells_Transformed_fibroblasts
Colon_Sigmoid
Colon_Transverse
Esophagus_Gastroesophageal_Junction
Esophagus_Mucosa
Esophagus_Muscularis
Heart_Atrial_Appendage
Heart_Left_Ventricle
LCL
Liver
Lung
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
degree of eQTL sharing (π_1)
E
Odds ratio


A
IPS5K
Differentiation Scores
200 Overlap
Other Genes
B
Key Drivers
Other Genes
HOXD3
HOXA1
FOLH1
DLX1
HOXD4
HOXA9
HOXA2
HOXD10
HOXD9
HOXD13
AC007618.3
HOXA5
HOXA3
EPHA3
HOXC10
RP13-235E23.1
HOXC9
RP11-115P21.1
HOXA7
HOXC4
HOXC5
HOXD8
ENSG00000120087
NACAP2
HOXB8
HOXA11
GAPDHS
RP5-827C21.1
RPSAP2
RP11-444J21.2
RP11-1094H24.4
HOXC6
RPL5P22
HOXA10
RP11-412B14.1
RP11-841C19.3
HOXA11-AS
C
Key Drivers
Other Genes
NCKAP1L
ASB5
ACSM4
BHLHE22
NR0B1
DKK4
NPAP1
FGF17
MIXL1
CD300E
HOXB4
GLIS3-AS1
LINC00571
RP11-380D23.1
FSHR
RP4-710M16.2
GATA6-AS1
PGM5
TRPM2
KB-1269D1.8
HOXB2
LHX1
RP11-380D23.2
SLC40A1
MAGEB17
PSG4
CDX2
MYL4
PITX1
GAD1
SP5
EOMES
GATA6
FOXQ1
AC007277.3
EVX1
RP11-400N13.2
CST1
SOX17
RP11-799P8.1
LINC01124
TBX3
RP1-170O19.17
RNASE1
AC002480.2
SYTL5
EVX1-AS1
SST2
ACE2
MIR720
GATA4
RP11-511B23.1
RP11-13N12.1
snoU13
UNC93A
AC004870.5
RP4-797C5.2
TFF1
CER1
RP11-747H12.6
KCNJ3
SHH
PRODH2
SLC1A7
AC078899.5
SLC5A9
FOXA2
RP11-266N13.2
GSC
MTTP
PLXNA2
CLEC2L
EGFLAM-AS4
LEFTY2
SLC39A5
XRCC6P1
HNF4A
IHH
LGR5
CXCR4
LEFTY1
RBP4
NODAL
SLC28A2
APOA2
LINC00261
ALPK2
KCNJ16
ALK
MSX2
PODN
PTGS2
NKX2-5
TRPC7
CDH9
ERBB4
ISL1
NPNT
CD177
EPSTI1
ATP6V0D2
LINC00607
ADAMTS9
GATA3
RP11-807H17.1
IFI16
AC105227.1
SLCO2B1
SLN
DIO3OS
COLEC12
LINC00379
PLA2G12B
LINC00348
CPN1
FN1
RP11-119D9.1
GATA3-AS1
SOAT2
F2
CLDN18
FGG
HMGCS2
DNAJC22
EGFLAM
FGB
APOA1
AKR1B10
HP
FGA
ALPI
TTR
PLG
HABP2
LINC00341
FGF10
GJB1
SERPINA1
SULT2A1
CTSE
HAS2
SLC6A19
LGALS2
GSTA2
FAM198B
HPX
HAVCR1
APOA4
LCN15
ANKS4B
SLC6A4
MYL3
PDX1


A
IPS 5K
Key Drivers
ESC Markers
Development Markers


B
Key Drivers
Development Markers
ESC Markers
Other Nodes
p=8.08e-30
OR=14.9
p=8.08e-30
OR=15.3
HOXC10
p=9.48e-30
OR=15.7
p=8.39e-23
OR=12.4
p=3.55e-16
OR=10.5
APOA2
HOXA5
FOXQ1
LINC00261
p=2.49e-06
OR=7.2
CER1
GATA6
GATA4
EOMES
Overall Causality Flow
Development Markers
Stem Cell Markers


C
Enrichment of Key Drivers Downstream of
ESC and Development Markers
IPS5K
IPS13K
K steps Downstream
0
20
40
60
0
2
4
6
−log10(Adjusted P−Value)


D
Mean Differentiation Score Distribution
Frequency
0
1
2
3
4
5
0
10
20
30
Score

SUPPLEMENTAL LEGENDS

Figure S1. Outline for Generation and Characterization of Human iPSCs, Related to STAR methods. A) T-cells and erythroblasts expanded from PBMC cultures were infected with Sendai virus containing the Yamanaka factors and grown on MEFs for 12 days. TRA1-60 positive colonies (6 per subject) were further expanded for 6 passages (P1-P6) on MEF monolayer and subsequently in feeder-free conditions. Cells were harvested between passages P8-P11 for stock preparation and RNA extraction for RNAseq analysis. During iPSC characterization, 20 clones from 5 individuals were disqualified while 317 clones from 101 individuals were included in the final analysis. Contribution of biological factors (donors' age, BMI, sex) and technical factors (reprogramming batch and technician, reprogramming source, Sendai virus lot, RNA isolation and technician, RNA sequencing batch) were considered for assessing the transcriptional variability in iPSC lines. **B**) Immunostaining with human anti-Oct4 (green) and anti-Nanog (red) antibodies to assess pluripotency in iPSC lines. DAPI (blue) was used for nuclear staining. BF indicates brightfield images. Data of two representative iPSC lines (iPSC 411 and 358) is shown.

Figure S2. Summary of analysis workflow with multiple computational methods, Related to Figure 1, Figure 4 and STAR methods. A) Computational workflow for RNA-seq and genotype array processing and analysis. **B**) Illustration of how variation in gene expression is attributed to multiple variables in the study design using the variancePartition method. The contribution of each variable to variation in each gene is summarized with a percentage. **C**) The magnitude of variance can be quantified for each gene. This perspective on variation is complementary to the results of variancePartition. **D**) Illustration of the relationship between contribution to variance and magnitude of variance. Cartoon shows 3 iPSC lines from individuals A and B. Variability can be described in terms of 1) the magnitude of variance in the dataset or 2) the fraction (or contribution) of variance attributable to variation across and within these individuals. While this illustration considers a single variable, the variancePartition method considers multiple variables in the study design.

Figure S3. Principal Components Analysis of 24 Key Pluripotency Related Genes, Related to Figure 1. The expression gradient is shown for each gene.

Figure S4. VariancePartition Analysis of RNA-seq Data Illustrates Biological and Technical Variation at Multiple Levels, Related to Figure 1. Expression variation across sequencing batches and RNA preparation kits is substantial and these sources of variation are labeled in bold. All other analyses were performed on the expression residuals after the effect of these two variables were removed.

Figure S5. Analysis of eQTL results indicates genomic enrichments, Related to Figure 2. A) Principal components analysis of genotype data from the current study and the 1000 Genomes Project identifies individuals of European ancestry. Individual from the current study are show in black and most cluster with European individuals colored in green. The dashed ellipse indicates the individuals (n=81) that were included in the eQTL analysis. **B**) Genome-wide significant cis-eQTLs are centered around the transcription start site of their respective genes. **C**) eQTLs show highest enrichment in promoters in iPSCs and ESCs identified by the Roadmap Epigenomics Consortium (Roadmap Epigenomics Consortium, 2015). Z-scores indicate the degree of enrichment in enhancers represented in cells and tissues samples from (Roadmap Epigenomics Consortium, 2015). Bars are colored based on tissue origin. The dashed line indicates the Bonferroni cutoff for multiple testing. **D**) Overlap of eQTLs discovered in the current dataset with eQTLs from GTEx using Storey's π_1. The π_1 statistic indicated the fraction of eQTLs discovered in iPSCs that are shared by these external datasets. E) Enrichment of eQTLs detected in this dataset with GWAS hits from the GWAS Catalog.

Figure S6. 5K Sub-networks Downstream of Key Driver Genes of Interest Contribute to the iPSC Variability, Related to Figure 6 and 7. A) Causal molecular mechanism among 5,000 most varying genes; The sub-networks 2 steps away from the key drivers of interest highlighted in b) and c) are shown in red and orange respectively; **B**) Sub-network 2 steps away from the key driver genes of the top 500 differentially expressed genes for the differentiation score to endothelial cells; **C**) Sub-network 2 steps away from the key driver genes of the 200 genes shared between the top 500 genes most varying within individual and the top 500 genes most varying across individuals.

Figure S7. Key Driver Genes Validation and Distribution of Differentiation Scores, Related to Figure 6 and 7. A) Causal molecular networks covering the 5000 most varying genes. The key drivers genes are highlighted in red, the stem cell markers in green and the development markers in orange. **B**) This Eiffel Tower plot shows the overall causality flow (top to bottom) from any stem cell (green) or development (yellow) markers to any upstream causal gene in the 5K network. It shows the enrichment of key driver genes (red) at every step upstream of the markers with the level-associated Fisher's test p-value on the left. **C**) Barplot of the –log10(enrichment p-value) of the key driver genes at each step downstream of the stem cell and development markers for both networks. **D**) Histogram of the distribution of the differentiation scores from iPS to endothelial cells. The red line represents the cutoff for the differential expression analyses.

Table S1. Patient Demographic Data and iPSC Generation Information, Related to STAR methods

Table S2. Variance Partition and Magnitude of Variance Results, Related to Figure 1 and 4

Table S3. GO enrichments for modules, GSEA enrichments for modules, Related to Figure 4 and 5

Table S4. GO enrichments for fig.4B, GSEA enrichments for fig. 4E, DE gene list for differentiation score, list of 500 most varying genes within patients, list of 500 most varying genes across patients, Related to Figure 4, 6 and 7

Table S5. BN 13K, BN 5K, eQTLs used as prior for BN, Seeding gene list BN 5K, Seeding gene list BN 13K, Coexpression network, KDs BN 5K, KDs BN 13K, Related to Figure 6 and 7

Table S6. Quantification of Endothelial Differentiation, Related to Figure 6

Table S7. List of 197 iPSC/ESC pluripotency and developmental markers, Related to Figure 7