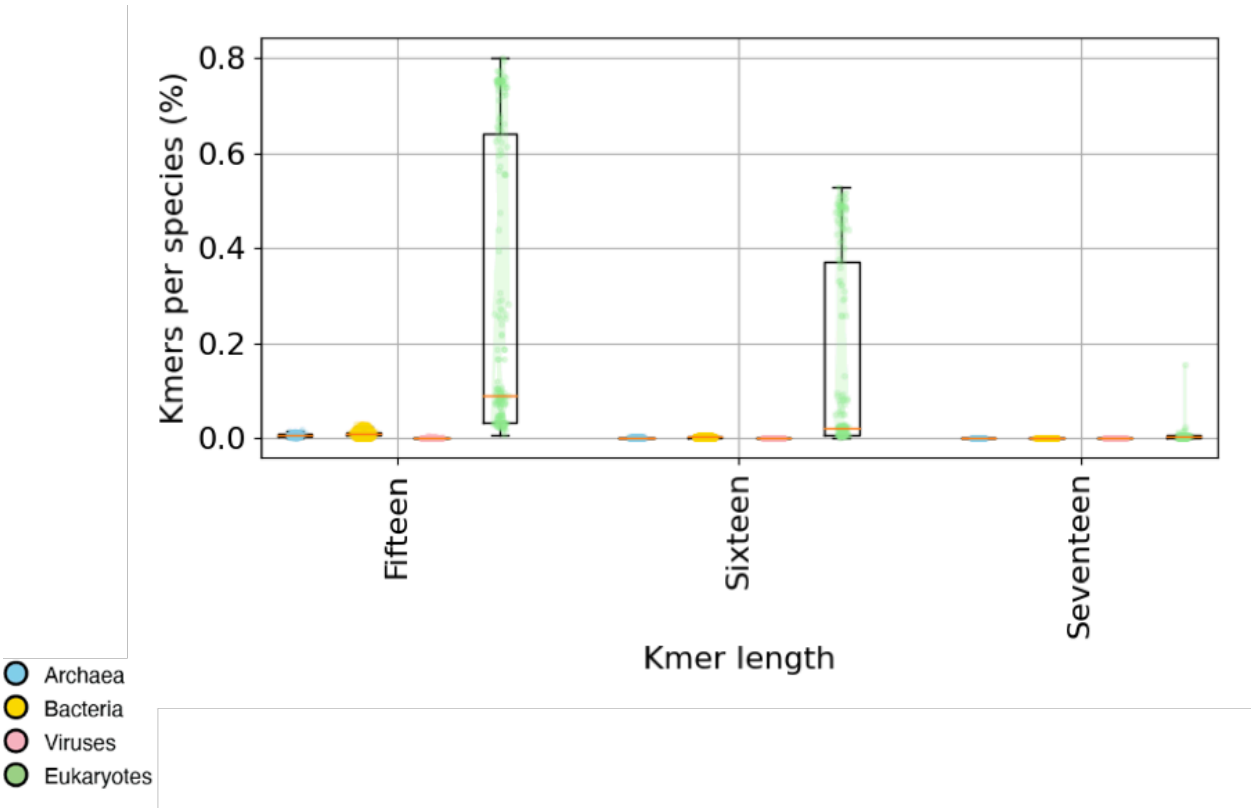
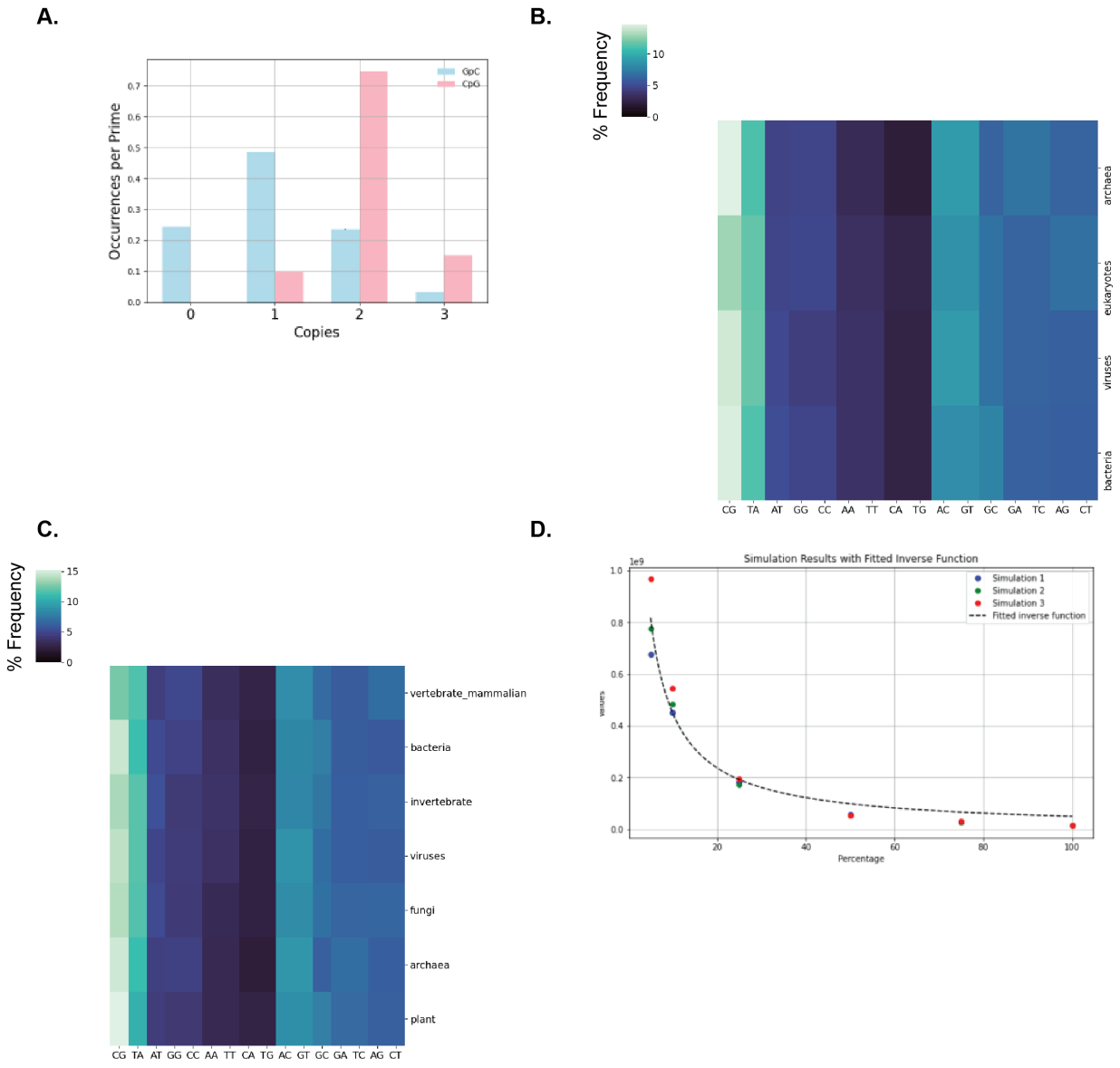
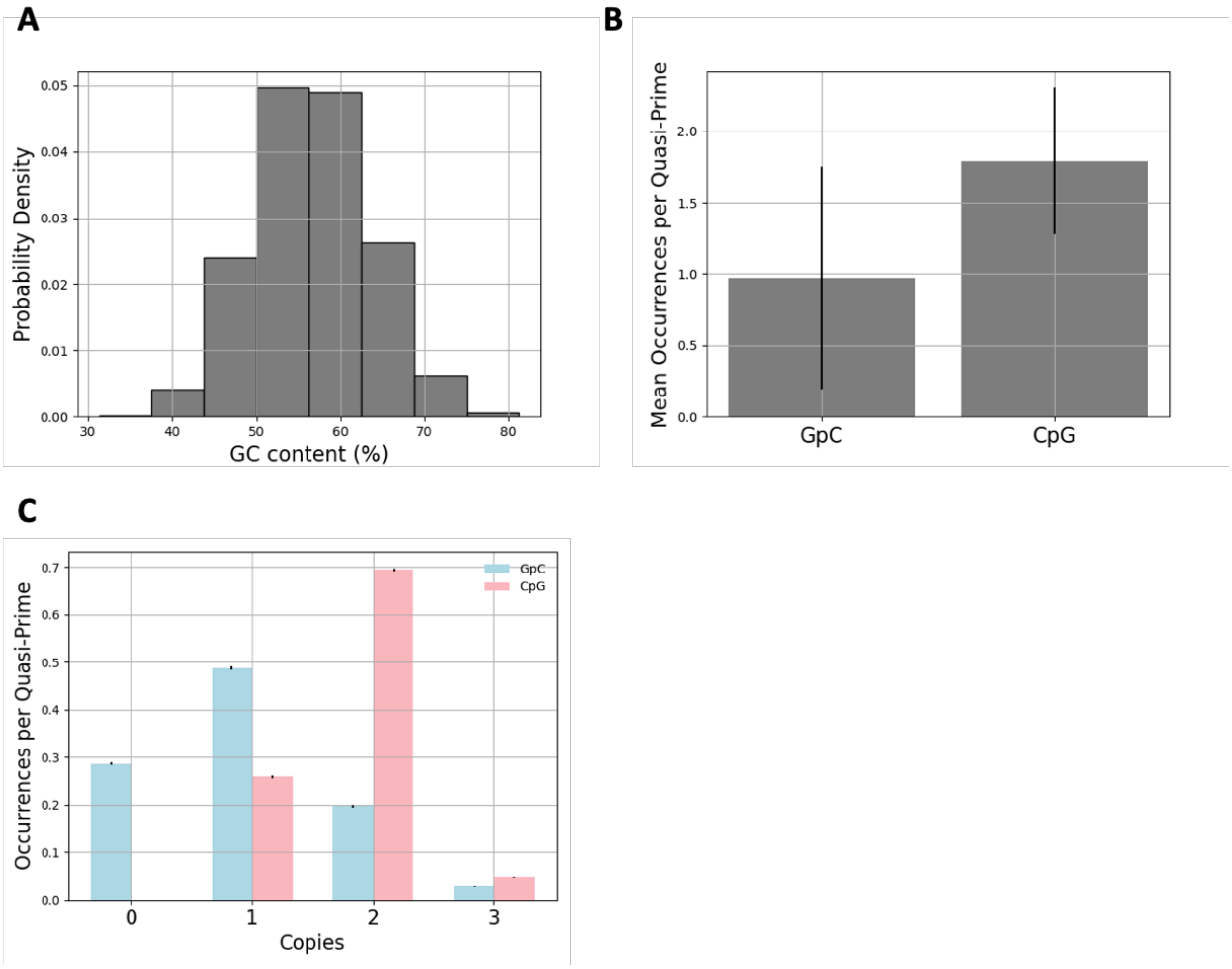
Supplemental Figures Table of Contents



Supplementary Figure 1: Proportion of k-mers found in each reference genome for kmer lengths of 15bp, 16bp, and 17bp across the taxonomic subdivisions. Each dot represents the proportion of k-mers observed in a reference genome. The majority of k-mers are found in a minority of the species studied across taxonomies. Error bars represent standard deviation.

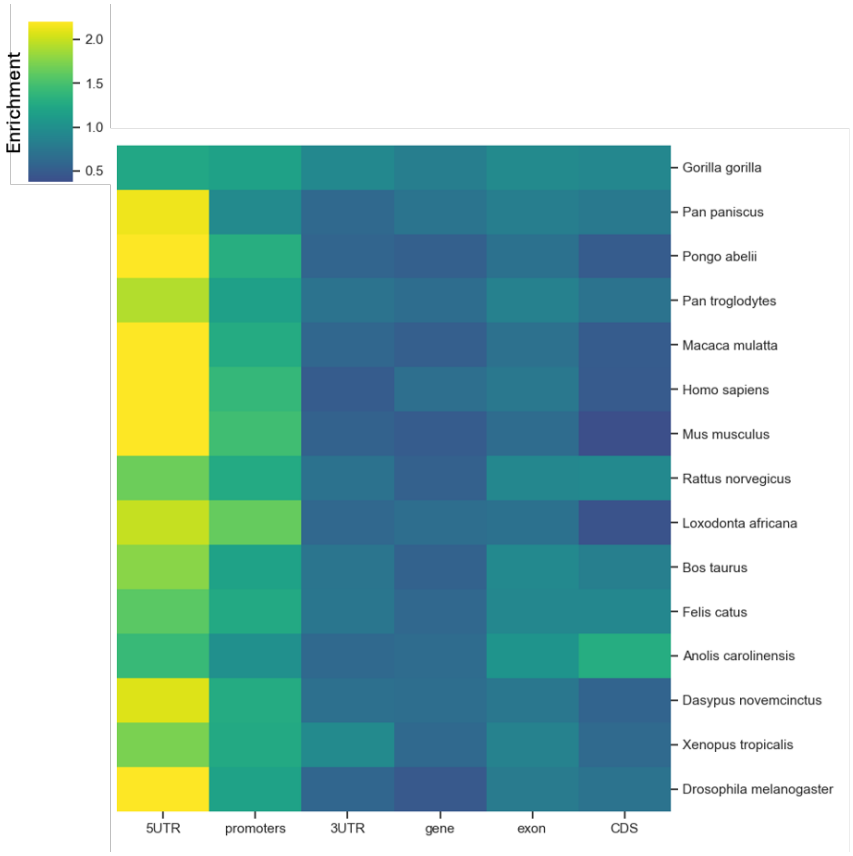


Supplementary Figure 2: Features of nucleic quasi-primes. A. Number of occurrences of GpCs and CpGs for different copy numbers, per nucleic prime sequence. Error bars are derived from bootstrapping with replacement (n=1,000) and represent standard deviation. **B.** Nucleotide composition across three domains of life and virus and **C.** in eukaryotes. **D.** Simulation experiments examining how the number of quasi-primes identified changes as a function of the number of genomes available.

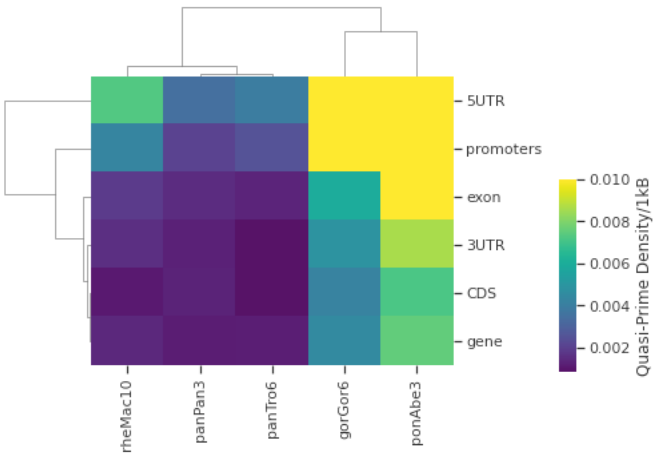


Supplementary Figure 3: GC content distribution of human quasi-prime sequences. A. GC content percentage of human quasi-prime sequences. **B.** Average number of GpC and CpG occurrences per human quasi-prime. Error bars show standard deviation. **C.** Number of occurrences of GpCs and CpGs for different copy numbers, per nucleic quasi-prime sequence. Error bars are derived from bootstrapping with replacement (n=1,000) and represent standard deviation.

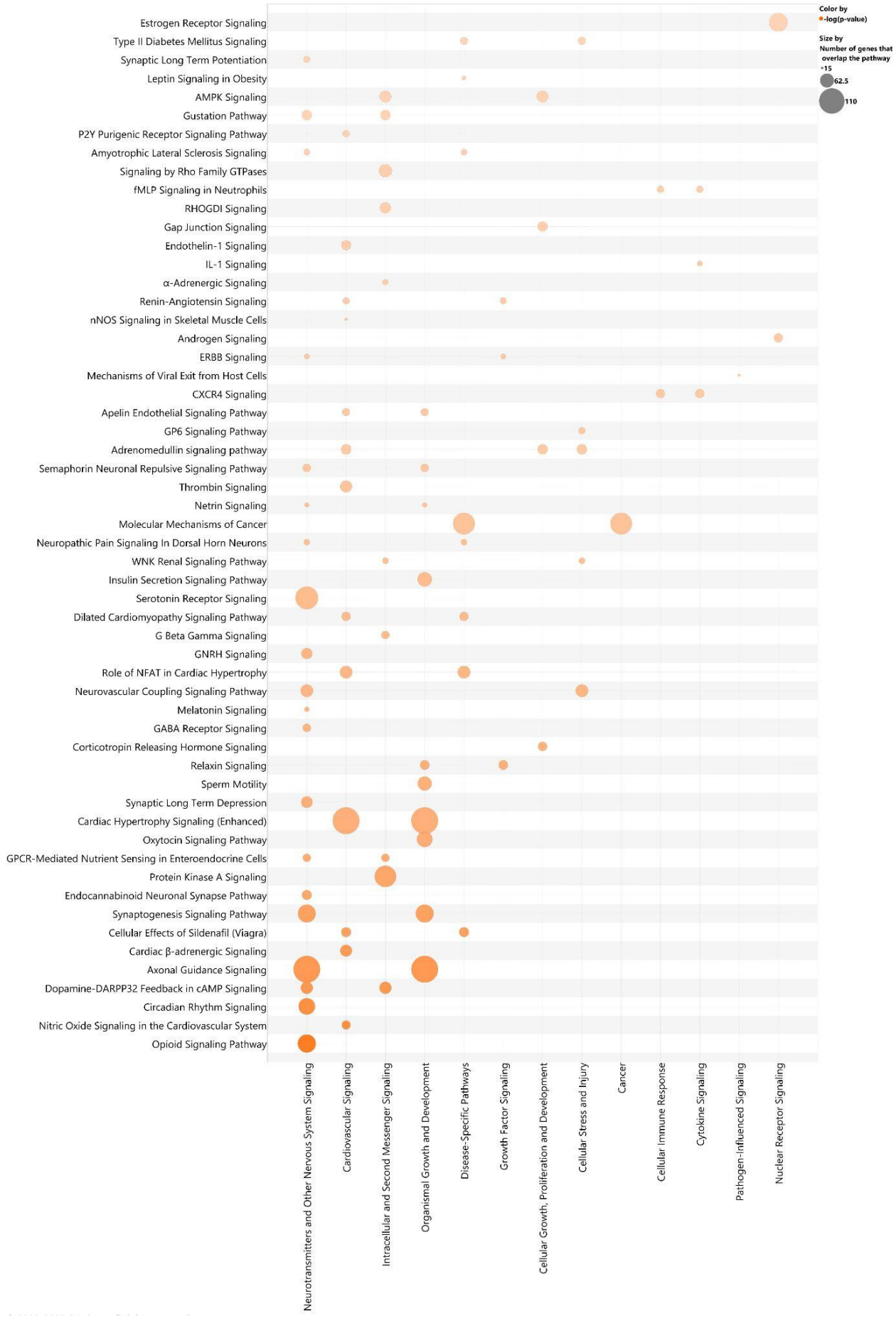
A.



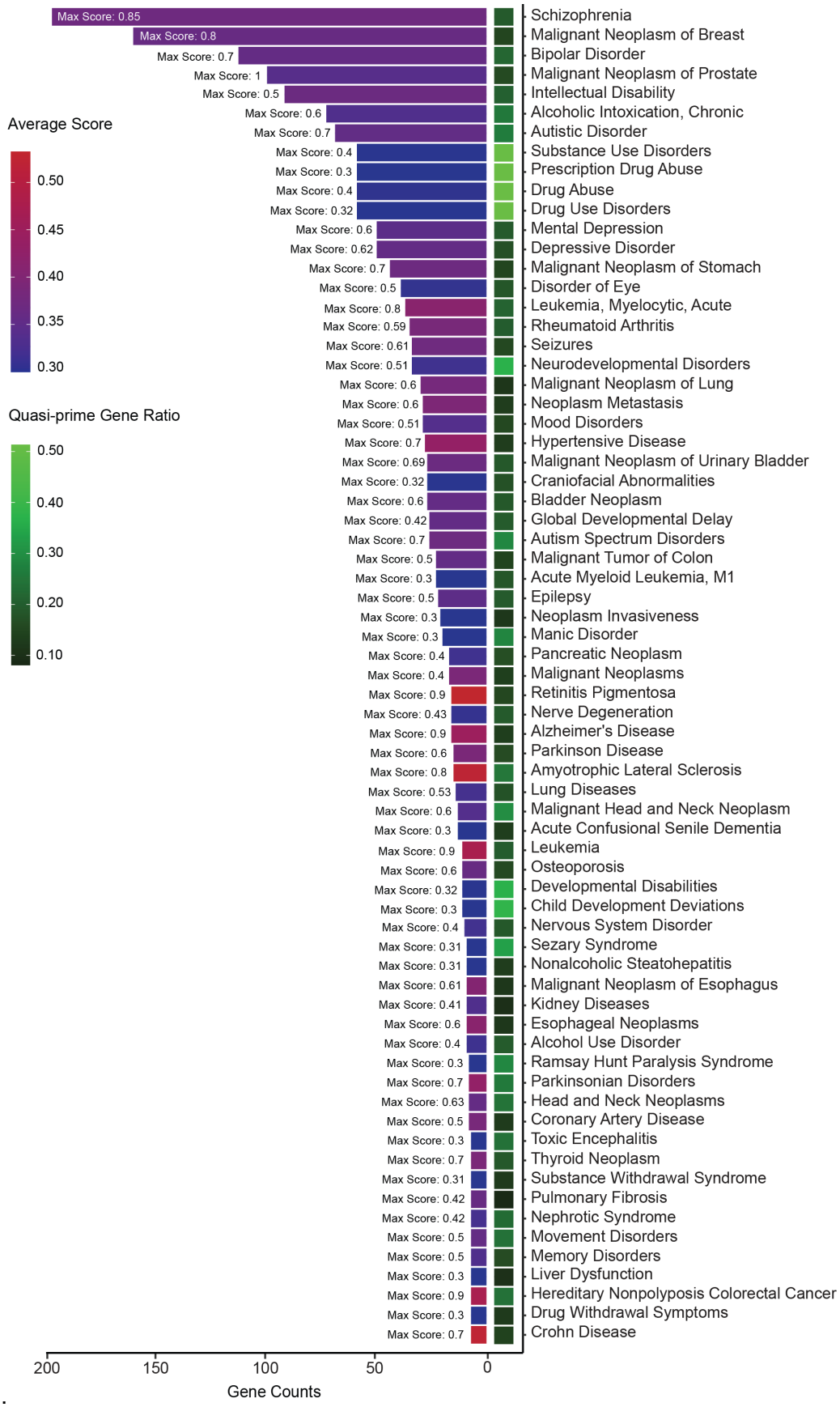
B.



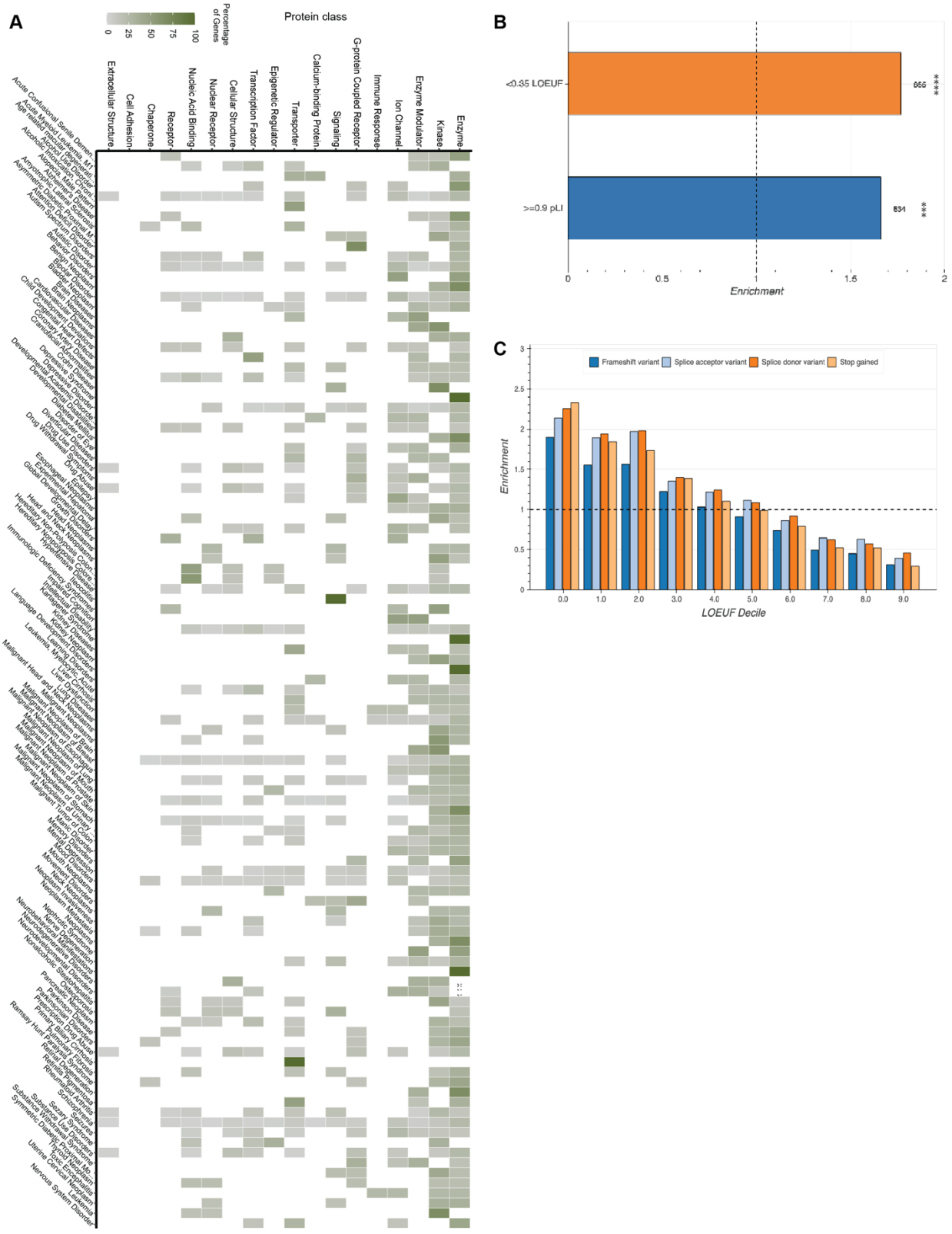
Supplementary Figure 4: Quasi-prime sequences in genomic sub-compartments. A. Density of quasi-primes across genic regions of multiple primate genomes, rodents, and other mammals and invertebrates and **B.** of five non-human primate genomes. NCBI RefSeq annotation was used for all non-human primates.


Color by
-log(p-value)
Size by
Number of genes that overlap the pathway
15
62.5
110
Estrogen Receptor Signaling
Type II Diabetes Mellitus Signaling
Synaptic Long Term Potentiation
Leptin Signaling in Obesity
AMPK Signaling
Gustation Pathway
P2Y Purigenic Receptor Signaling Pathway
Amyotrophic Lateral Sclerosis Signaling
Signaling by Rho Family GTPases
fMLP Signaling in Neutrophils
RHOGDI Signaling
Gap Junction Signaling
Endothelin-1 Signaling
IL-1 Signaling
α-Adrenergic Signaling
Renin-Angiotensin Signaling
nNOS Signaling in Skeletal Muscle Cells
Androgen Signaling
ERBB Signaling
Mechanisms of Viral Exit from Host Cells
CXCR4 Signaling
Apelin Endothelial Signaling Pathway
GP6 Signaling Pathway
Adrenomedullin signaling pathway
Semaphorin Neuronal Repulsive Signaling Pathway
Thrombin Signaling
Netrin Signaling
Molecular Mechanisms of Cancer
Neuropathic Pain Signaling In Dorsal Horn Neurons
WNK Renal Signaling Pathway
Insulin Secretion Signaling Pathway
Serotonin Receptor Signaling
Dilated Cardiomyopathy Signaling Pathway
G Beta Gamma Signaling
GNRH Signaling
Role of NFAT in Cardiac Hypertrophy
Neurovascular Coupling Signaling Pathway
Melatonin Signaling
GABA Receptor Signaling
Corticotropin Releasing Hormone Signaling
Relaxin Signaling
Sperm Motility
Synaptic Long Term Depression
Cardiac Hypertrophy Signaling (Enhanced)
Oxytocin Signaling Pathway
GPCR-Mediated Nutrient Sensing in Enteroendocrine Cells
Protein Kinase A Signaling
Endocannabinoid Neuronal Synapse Pathway
Synaptogenesis Signaling Pathway
Cellular Effects of Sildenafil (Viagra)
Cardiac β-adrenergic Signaling
Axonal Guidance Signaling
Dopamine-DARPP32 Feedback in cAMP Signaling
Circadian Rhythm Signaling
Nitric Oxide Signaling in the Cardiovascular System
Opioid Signaling Pathway
Neurotransmitters and Other Nervous System Signaling
Cardiovascular Signaling
Intracellular and Second Messenger Signaling
Organismal Growth and Development
Disease-Specific Pathways
Growth Factor Signaling
Cellular Growth, Proliferation and Development
Cellular Stress and Injury
Cancer
Cellular Immune Response
Cytokine Signaling
Pathogen-Influenced Signaling
Nuclear Receptor Signaling

Supplementary Figure 5: Bubble plot showing the pathway categories (Y axis) of the enriched canonical pathways (X axis) enriched in the quasi-prime gene set. The color represents the adjusted *p*-value of the enrichment (Increasing orange hue, the lower the *p*-value. The cutoff of *p*-value was set at <0.001). The size of each bubble represents the number of overlapping genes in the pathway.


Average Score
0.50
0.45
0.40
0.35
0.30
Quasi-prime Gene Ratio
0.50
0.40
0.30
0.20
0.10
Max Score: 0.85
Max Score: 0.8
Max Score: 0.7
Max Score: 1
Max Score: 0.5
Max Score: 0.6
Max Score: 0.7
Max Score: 0.4
Max Score: 0.3
Max Score: 0.4
Max Score: 0.32
Max Score: 0.6
Max Score: 0.62
Max Score: 0.7
Max Score: 0.5
Max Score: 0.8
Max Score: 0.59
Max Score: 0.61
Max Score: 0.51
Max Score: 0.6
Max Score: 0.6
Max Score: 0.51
Max Score: 0.7
Max Score: 0.69
Max Score: 0.32
Max Score: 0.6
Max Score: 0.42
Max Score: 0.7
Max Score: 0.5
Max Score: 0.3
Max Score: 0.5
Max Score: 0.3
Max Score: 0.3
Max Score: 0.4
Max Score: 0.4
Max Score: 0.9
Max Score: 0.43
Max Score: 0.9
Max Score: 0.6
Max Score: 0.8
Max Score: 0.53
Max Score: 0.6
Max Score: 0.3
Max Score: 0.9
Max Score: 0.6
Max Score: 0.32
Max Score: 0.3
Max Score: 0.4
Max Score: 0.31
Max Score: 0.31
Max Score: 0.61
Max Score: 0.41
Max Score: 0.6
Max Score: 0.4
Max Score: 0.3
Max Score: 0.7
Max Score: 0.63
Max Score: 0.5
Max Score: 0.3
Max Score: 0.7
Max Score: 0.31
Max Score: 0.42
Max Score: 0.42
Max Score: 0.5
Max Score: 0.5
Max Score: 0.3
Max Score: 0.9
Max Score: 0.3
Max Score: 0.7
Schizophrenia
Malignant Neoplasm of Breast
Bipolar Disorder
Malignant Neoplasm of Prostate
Intellectual Disability
Alcoholic Intoxication, Chronic
Autistic Disorder
Substance Use Disorders
Prescription Drug Abuse
Drug Abuse
Drug Use Disorders
Mental Depression
Depressive Disorder
Malignant Neoplasm of Stomach
Disorder of Eye
Leukemia, Myelocytic, Acute
Rheumatoid Arthritis
Seizures
Neurodevelopmental Disorders
Malignant Neoplasm of Lung
Neoplasm Metastasis
Mood Disorders
Hypertensive Disease
Malignant Neoplasm of Urinary Bladder
Craniofacial Abnormalities
Bladder Neoplasm
Global Developmental Delay
Autism Spectrum Disorders
Malignant Tumor of Colon
Acute Myeloid Leukemia, M1
Epilepsy
Neoplasm Invasiveness
Manic Disorder
Pancreatic Neoplasm
Malignant Neoplasms
Retinitis Pigmentosa
Nerve Degeneration
Alzheimer's Disease
Parkinson Disease
Amyotrophic Lateral Sclerosis
Lung Diseases
Malignant Head and Neck Neoplasm
Acute Confusional Senile Dementia
Leukemia
Osteoporosis
Developmental Disabilities
Child Development Deviations
Nervous System Disorder
Sezary Syndrome
Nonalcoholic Steatohepatitis
Malignant Neoplasm of Esophagus
Kidney Diseases
Esophageal Neoplasms
Alcohol Use Disorder
Ramsay Hunt Paralysis Syndrome
Parkinsonian Disorders
Head and Neck Neoplasms
Coronary Artery Disease
Toxic Encephalitis
Thyroid Neoplasm
Substance Withdrawal Syndrome
Pulmonary Fibrosis
Nephrotic Syndrome
Movement Disorders
Memory Disorders
Liver Dysfunction
Hereditary Nonpolyposis Colorectal Cancer
Drug Withdrawal Symptoms
Crohn Disease
200
150
100
50
0
Gene Counts
.

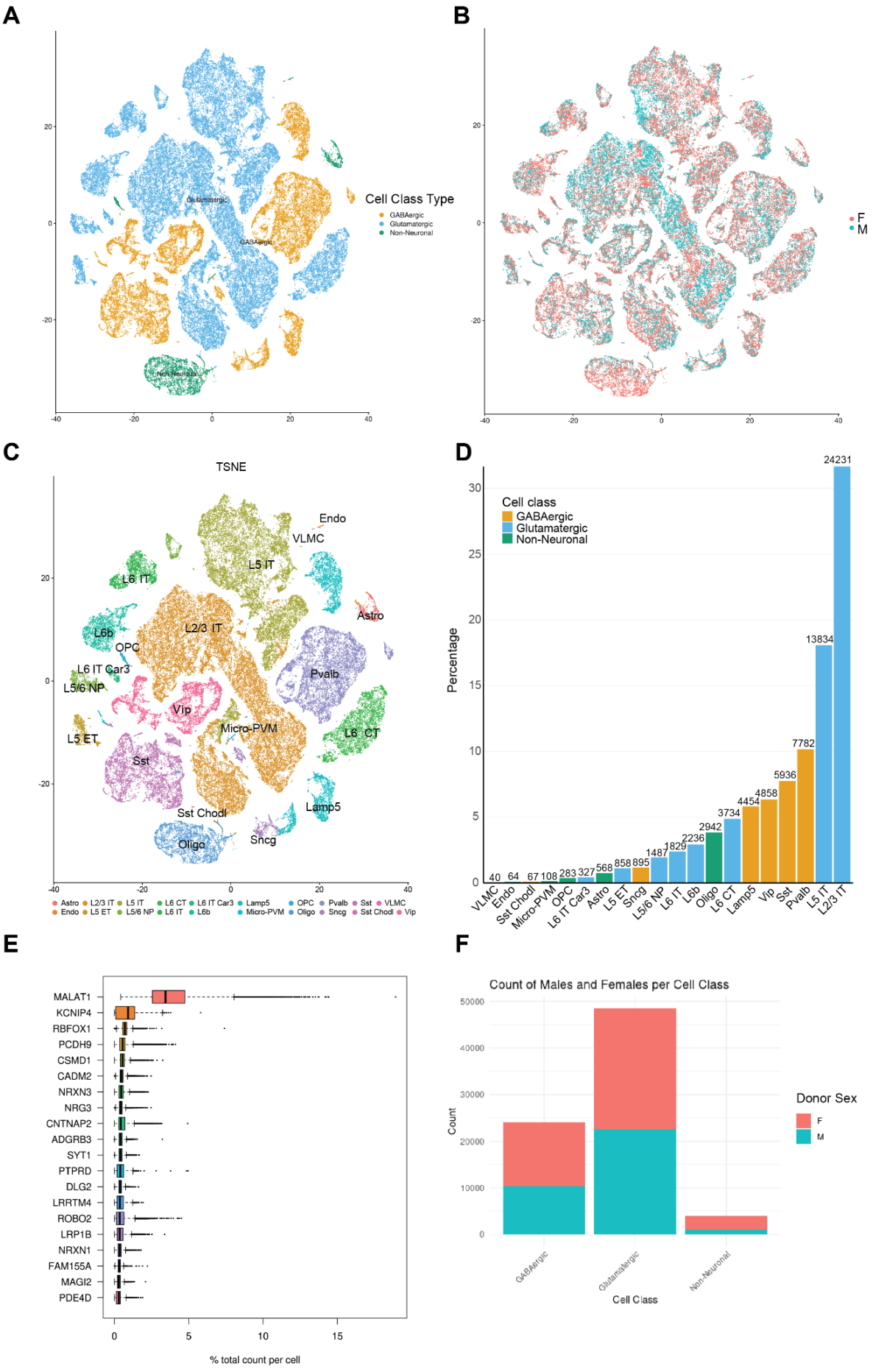
Supplementary Figure 6: Disease association for quasi-prime genes with DisGeNET
Full gene disease bar plot presenting the count of associated genes across different disease conditions for all diseases with more than 6 associated genes. The average disease association score for all quasi-prime genes for a given disease within “ALL” DisGeNET databases is visualized in the color of the bar. The max score for all genes within a disease is displayed in text on figure. The number of quasi-prime genes out of the total genes annotated in the database is represented as a ratio shown in the heatmap tile.



Supplementary Figure 7: Disease association for quasi-prime genes

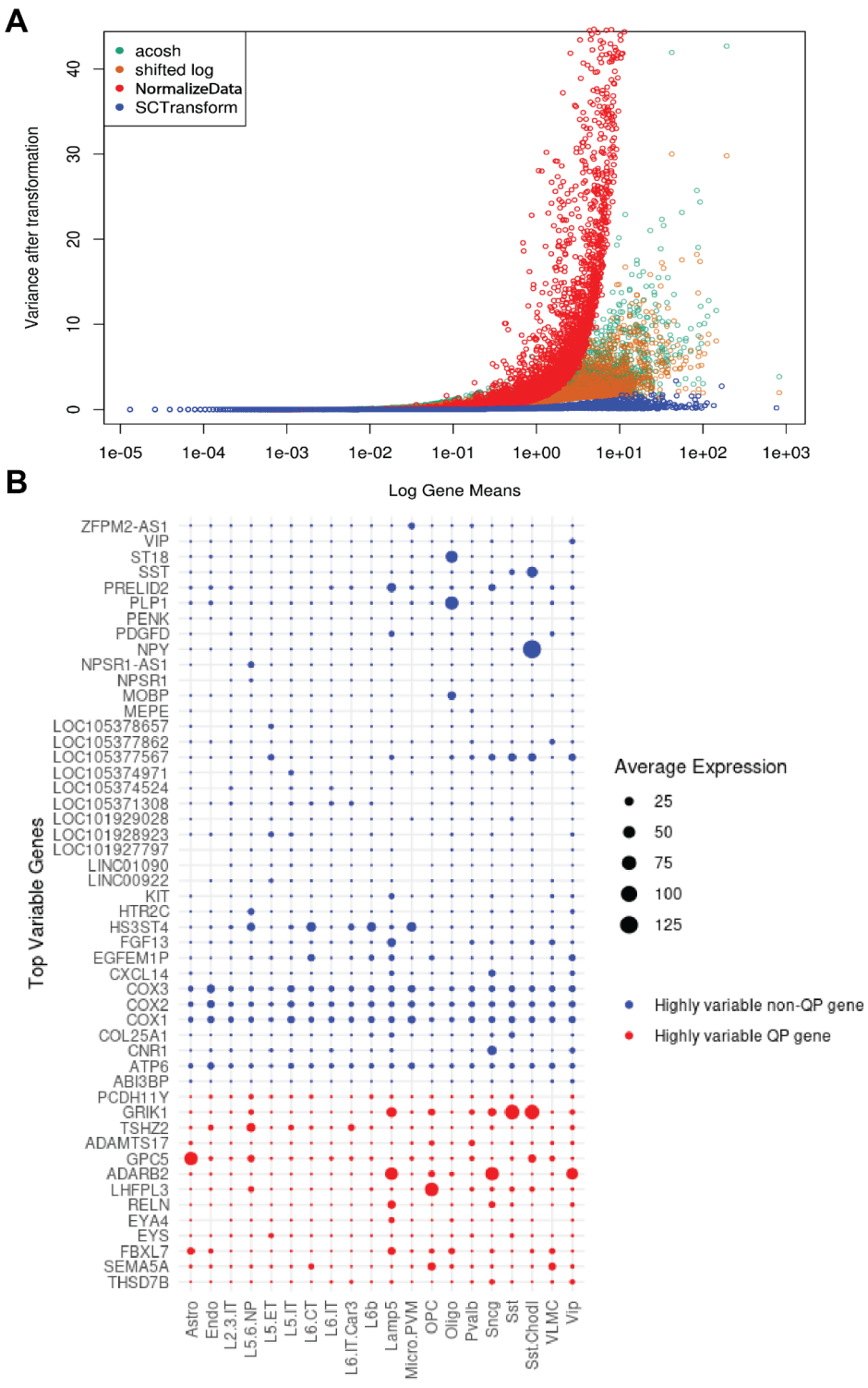
A. Protein class heat map. The color scale represents the percentage of quasi-prime genes subset within the DisGeNET "ALL" databases' query for each disease belonging to each protein

class. **B-C.** gnomAD pLOF constraint gene analysis. **B.** The odds ratio enrichment of highly constrained human quasi-prime genes (with pLI >= 0.9 and LOEUF < 0.35) is displayed as a bar plot. The enriched gene set was tested for statistical significance with a hypergeometric test. Significance levels are indicated as follows: * p < 1E-25, ** p < 1E-35, *** p < 1E-45, and **** p < 1E-55. **C.** pLOF variant type enrichment across different LOEUF Decile bins for human quasi-prime genes.

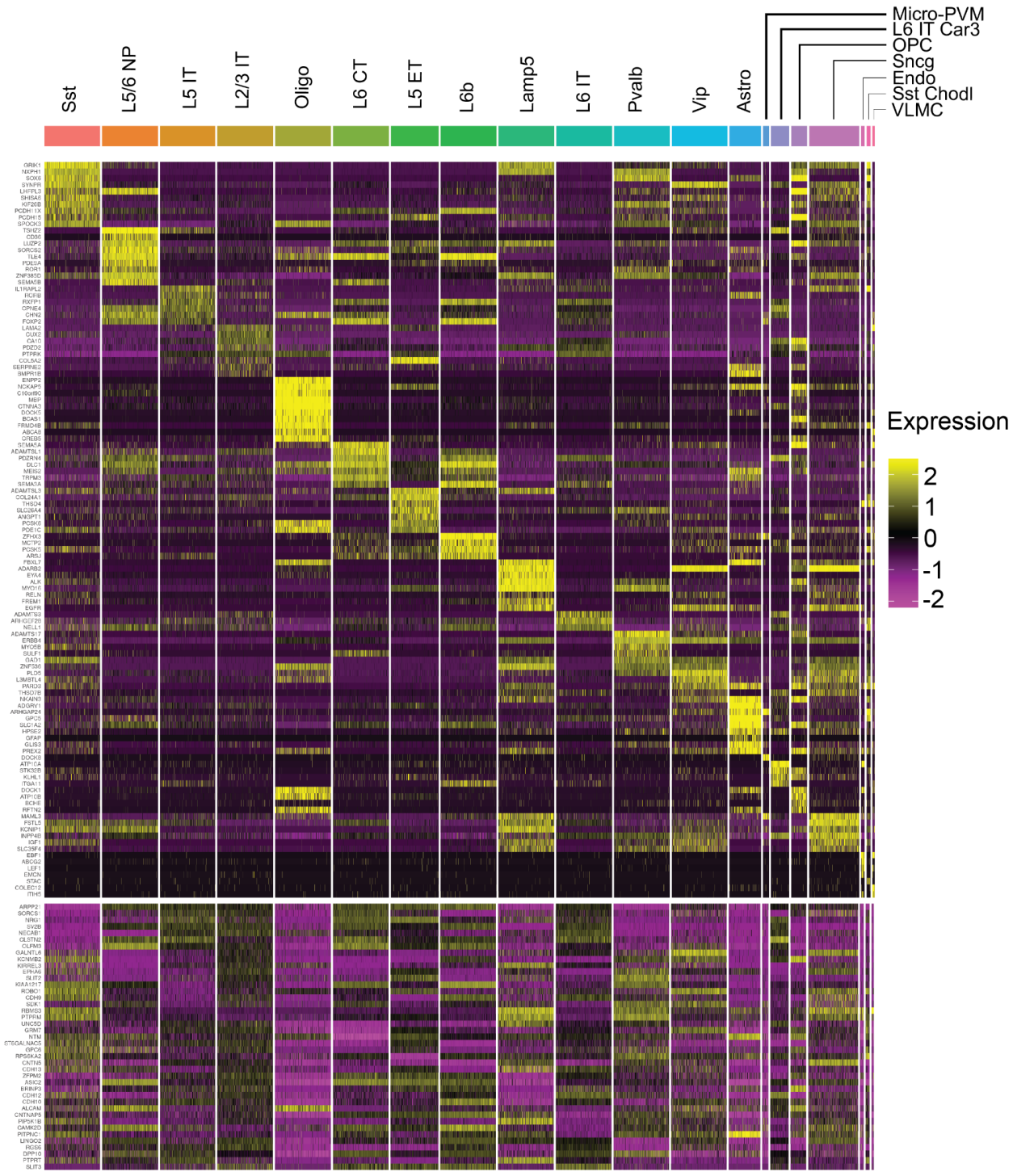

A
Cell Class Type
GABAergic
Glutamatergic
Non-Neuronal
B
F
M
C
TSNE
Endo
VLMC
L5 IT
L6 IT
Astro
L6b
L2/3 IT
OPC
L6 IT Car3
Pvalb
L5/6 NP
Vip
Micro-PVM
L6 CT
L5 ET
Sst
Lamp5
Sst Chodl
Sncg
Oligo
Astro L2/3 IT L5 IT L6 CT L6 IT Car3 Lamp5 OPC Pvalb Sst VLMC
Endo L5 ET L5/6 NP L6 IT L6b Micro-PVM Oligo Sncg Sst Chodl Vip
D
Cell class
GABAergic
Glutamatergic
Non-Neuronal
Percentage
40 64 67 108 283 327 568 858 895 1487 1829 2236 2942 3734 4454 4858 5936 7782 13834 24231
VLMC Endo Sst Chodl Micro-PVM OPC L6 IT Car3 Astro L5 ET Sncg L5/6 NP L6 IT L6b Oligo L6 CT Lamp5 Vip Sst Pvalb L5 IT L2/3 IT
E
MALAT1
KCNIP4
RBFOX1
PCDH9
CSMD1
CADM2
NRXN3
NRG3
CNTNAP2
ADGRB3
SYT1
PTPRD
DLG2
LRRTM4
ROBO2
LRP1B
NRXN1
FAM155A
MAGI2
PDE4D
% total count per cell
F
Count of Males and Females per Cell Class
Count
Donor Sex
F
M
GABAergic
Glutamatergic
Non-Neuronal
Cell Class

Supplementary Figure 8: Single Cell Analysis Metrics

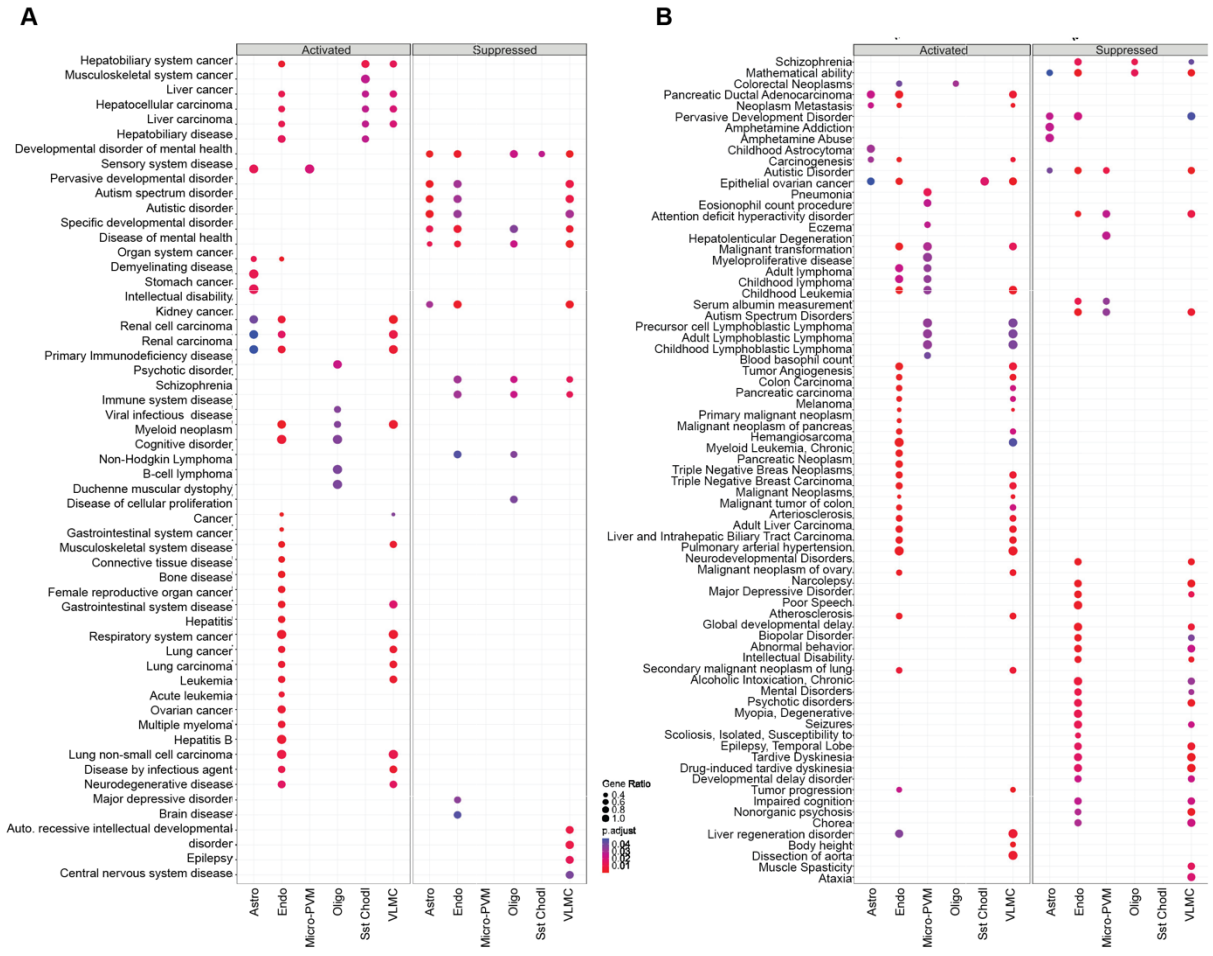
Single cell analysis metrics of M1 Primary Motor Cortex from Single Cell Brain Atlas. **A.** t-SNE plot showing GABAergic, Glutamatergic and non-neuronal cell labels. **B.** t-SNE plot showing cells colored by gender. **C.** t-SNE plot showing cell types found in the human brain atlas (Bakken et al. 2021). **D.** Bar plot shows the percentage of total cells and the count above each bar for each cell type. Bars are colored with cell class. **E.** Box plot sorted by the percentage of genes count expression per cell. **F.** Bar plot representing the differences in counts of cells by donor sex per cell class.



Supplementary Figure 9: Single Cell Transformation and Variance Analysis. A. Variance after transformation using four methods, acosh, shifted log, NormalizeData (Seurat), and SCTransform (Seurat). **B.** Dot plot representing the average expression across cell types between quasi-prime and non-quasi prime genes for the top 25 genes in the dataset.

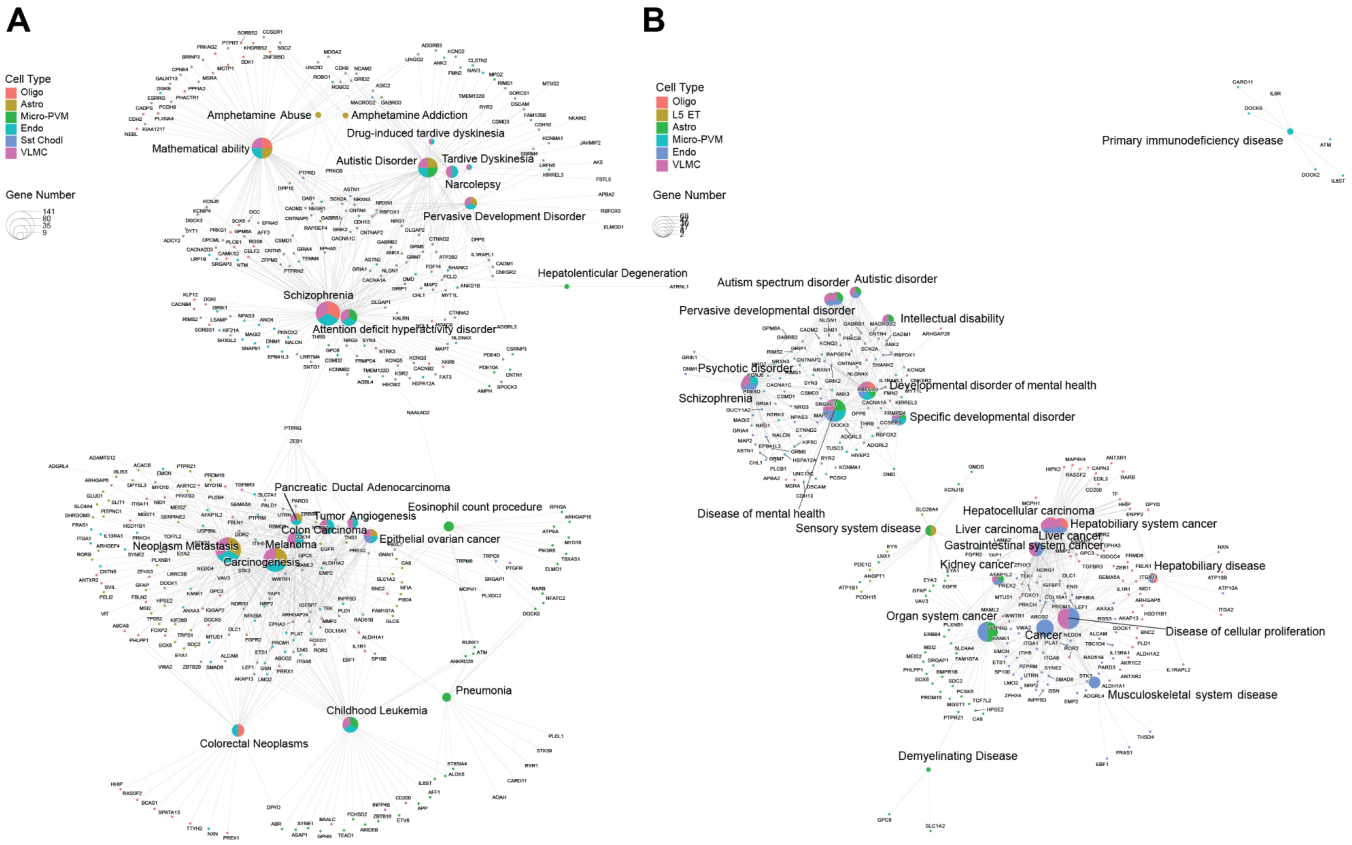


Supplementary Figure 10: Differential Expression based on quasi-prime genes among cell types. A. Differentially expressed quasi-prime genes (p value < 0.05) were sorted by log2 fold change. The top 10 upregulated genes (log2 fold change > 0) and the top 10 downregulated genes (log 2 fold change < 0) from each cell type were selected and visualized in this heatmap. Upregulated and downregulated genes are separated in heatmap by row space.

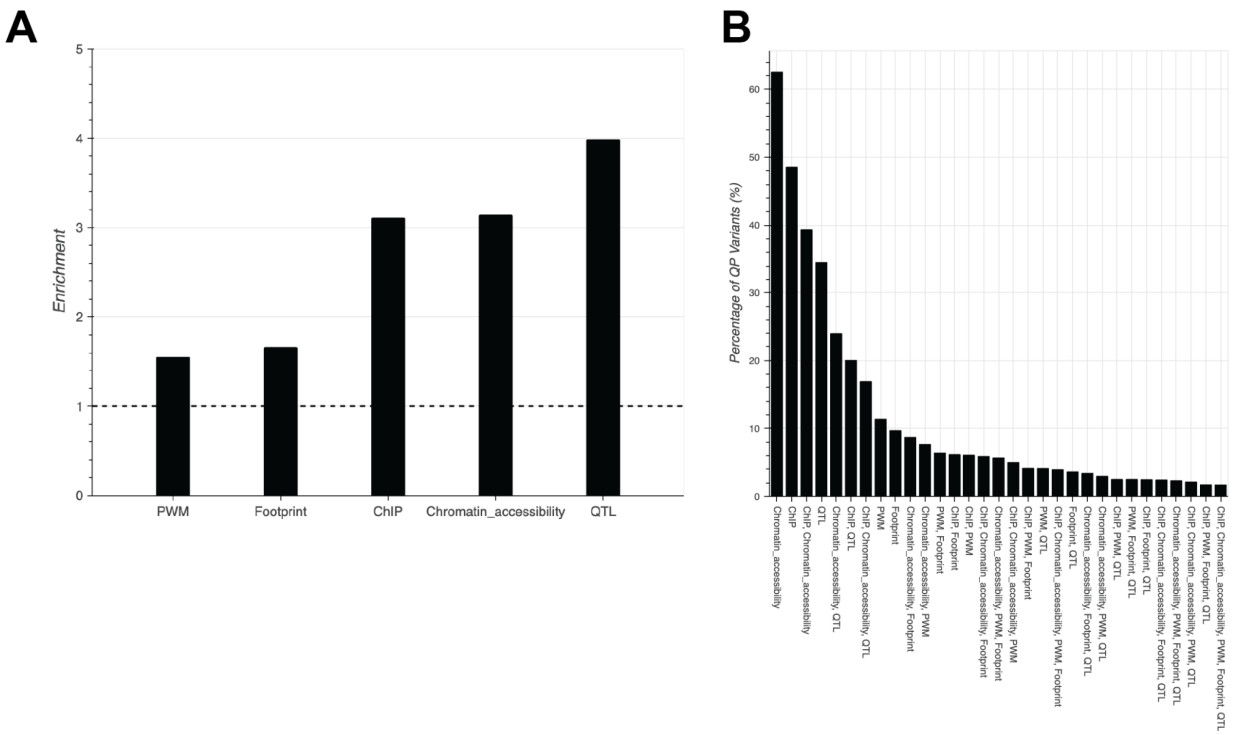


Supplementary Figure 11: Gene set enrichment analysis of genes associated with quasi-prime sequences among cell types found in the human single-cell primary motor cortex brain atlas

A-B. Significant (p-value < 0.05) differentially expressed genes with an absolute $\log_2$ fold change > 1 amongst cell types were used in GSEA analysis as represented by a dot plot. **A**. Disease ontology database. **B**. DisGeNET "ALL" database.



Supplementary Figure 12: Gene set enrichment analysis of differentially expressed genes associated with quasi-prime sequences display several disease associations among cell types found in the human single-cell primary motor cortex brain atlas. Significant (p value < 0.05) differentially expressed genes with an absolute log2 fold change > 1 amongst cell types were used in GSEA analysis as represented by a network graph. **A.** DisGeNET "ALL" database. **B.** Disease ontology database. Different cell types are color coded and pie charts reflect the cell types at which each the diseases shown were associated.



Supplementary Figure 13: Variant Characterization of Human Quasi-primes. A. Enrichment (odds ratio) of quasi-prime regions as compared to simulated controls within the regulomeDB database displayed as bar plot (Binomial test p=0.0) **B.** The percentage of quasi-prime variants for different combinations of how a variant is characterized or where it is located from regulomeDB is displayed as a bar plot.