Supplementary Information for

Purifying selection on noncoding deletions of human regulatory loci detected using their cellular pleiotropy

David W. Radke[1,2,3], **Jae Hoon Sul**[4], **Daniel J. Balick**[1,2,3], **Sebastian Akle**[1,2,3], **Alzheimer's Disease Neuroimaging Initiative***, **Robert C. Green**[2,3,5], and **Shamil R. Sunyaev**[1,2,3]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, 02115, USA; [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA; [3]Broad Institute of Harvard and MIT, Cambridge, MA, 02142 USA; [4]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, 90095, USA; [5]Ariadne Labs, Boston, MA, 02115, USA; *Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding author: ssunyaev@rics.bwh.harvard.edu

Table of Contents

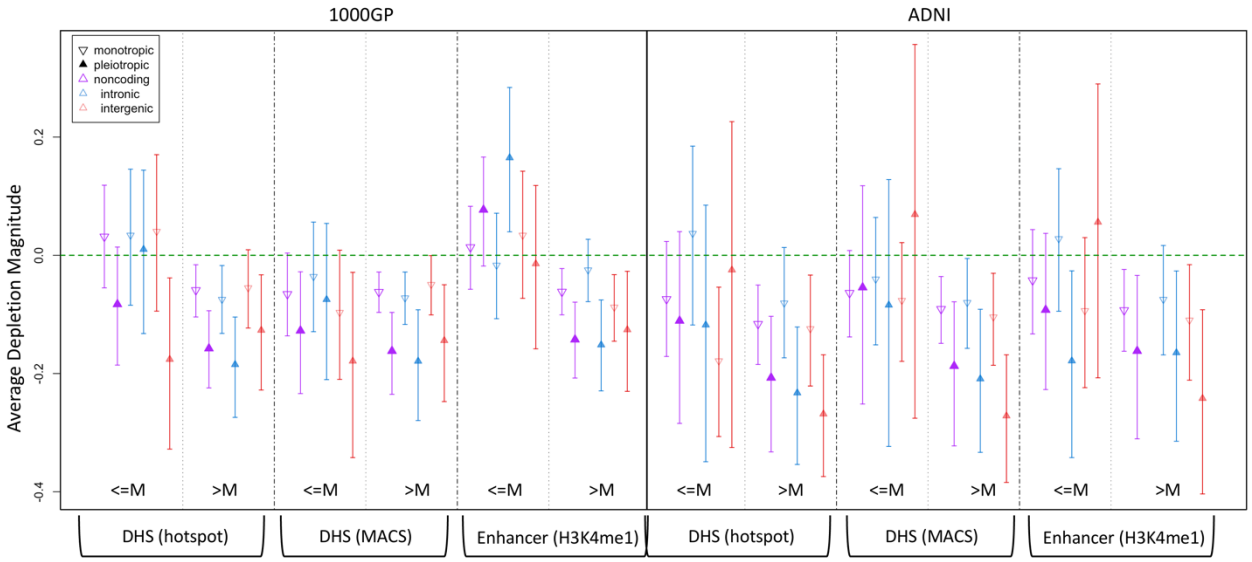

Figure S1. Selective pressure greater on pleiotropic longer deletions. We calculated the depletion magnitude of each deletion in both the 1000GP dataset (left half) and the ADNI dataset (right half) in terms of $PlyRS_{sum\text{-}mono}$ (monotropic) and $PlyRS_{sum\text{-}pleio}$ (pleiotropic). We separated deletions from each dataset into those that had a length of less than or equal to the median (M) length or greater than the median length of the deletion subset. For regulatory annotations (listed on the bottom) we plot the average $PlyRS_{sum\text{-}mono}$ (or $PlyRS_{sum\text{-}pleio}$) depletion magnitude for each deletion subset and generate 95% confidence intervals from bootstrapping (100 trials). A more negative value for the average depletion magnitude indicates a larger depletion of real deletions compared to simulation.

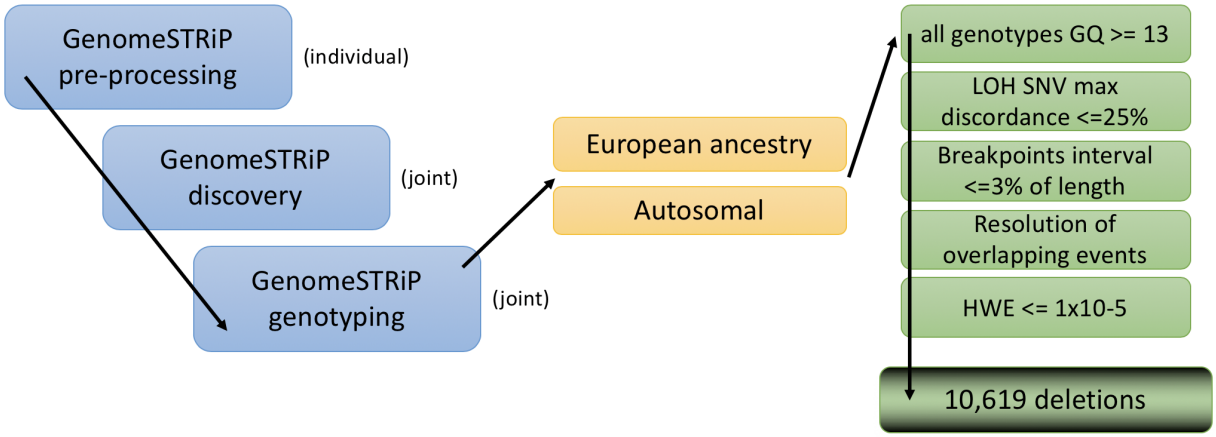

Figure S2. Schematic overview of the ADNI deletion callset generation. GQ = genotype quality. LOH = loss of heterozygosity. HWE = Hardy-Weinberg equilibrium.

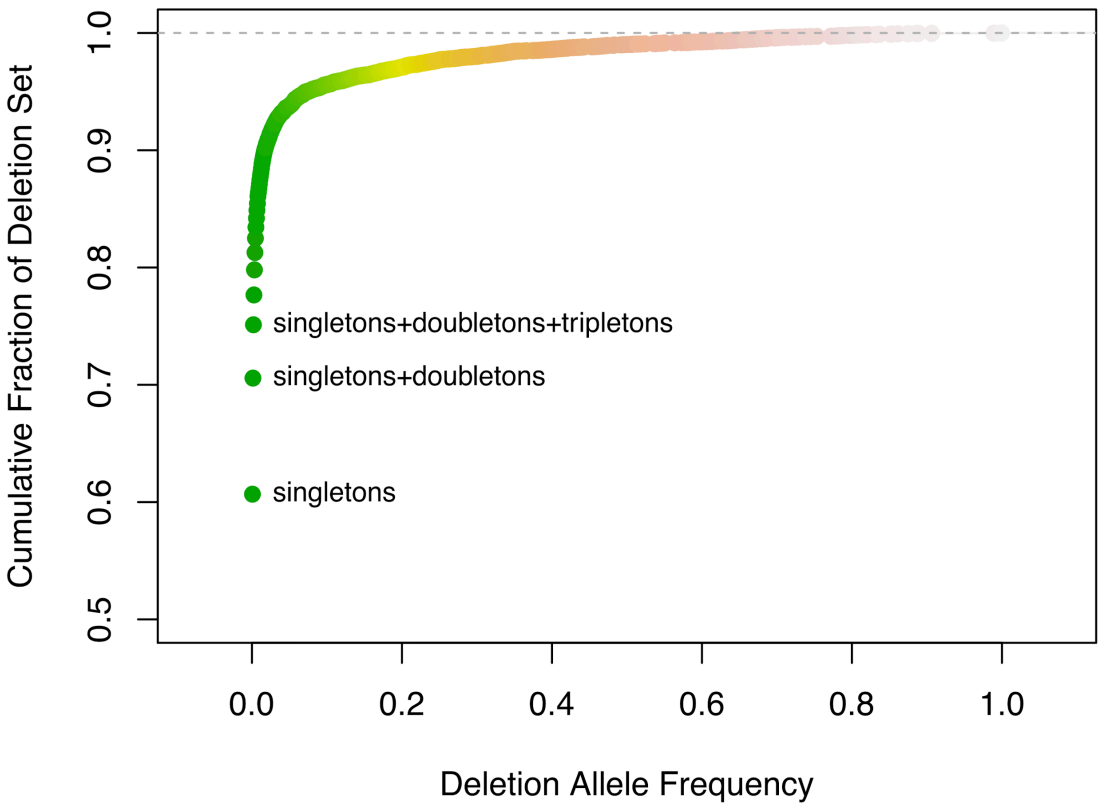

Figure S3. Cumulative fraction of ADNI deletion allele frequency. Deletions in the final quality-controlled and filtered callset are rank-ordered from lowest allele frequency to highest. The color gradient slowly changing from dark green to light grey corresponds to different observed allele frequency value levels throughout the dataset.

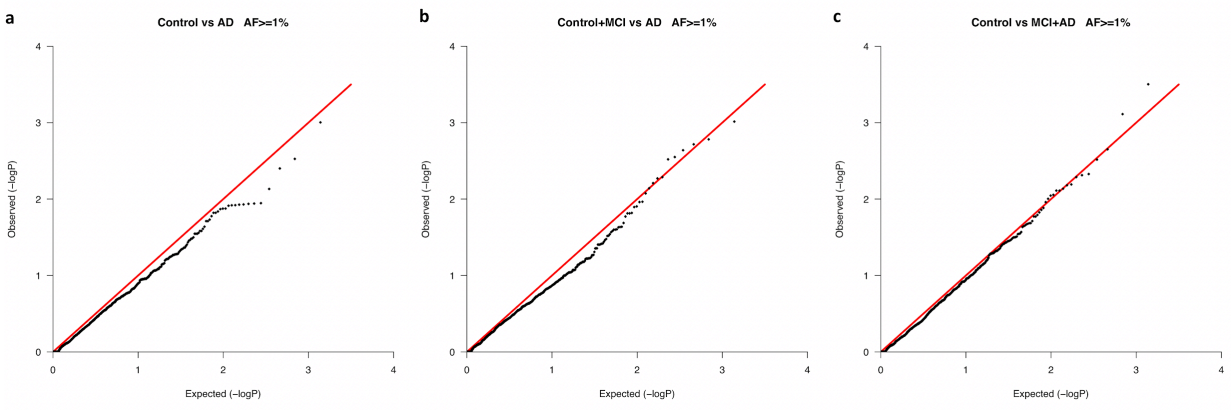

Figure S4. Q-Q plots of phenotype comparisons. Quantile-quantile plots of the quality-controlled and filtered deletion callset, before Hardy-Weinberg Equilibrium filter. P-values of each deletion's association with the phenotype ('Control'=brain-healthy cognition, 'MCI'=mild cognitive impairment, 'AD'=Alzheimer's Disease) were determined from Fisher's exact test. Since it is unclear if MCI better associates with Control or with AD, in **a)** MCI is not included, in **b)** Control+MCI are merged, and in **c)** MCI+AD are merged.

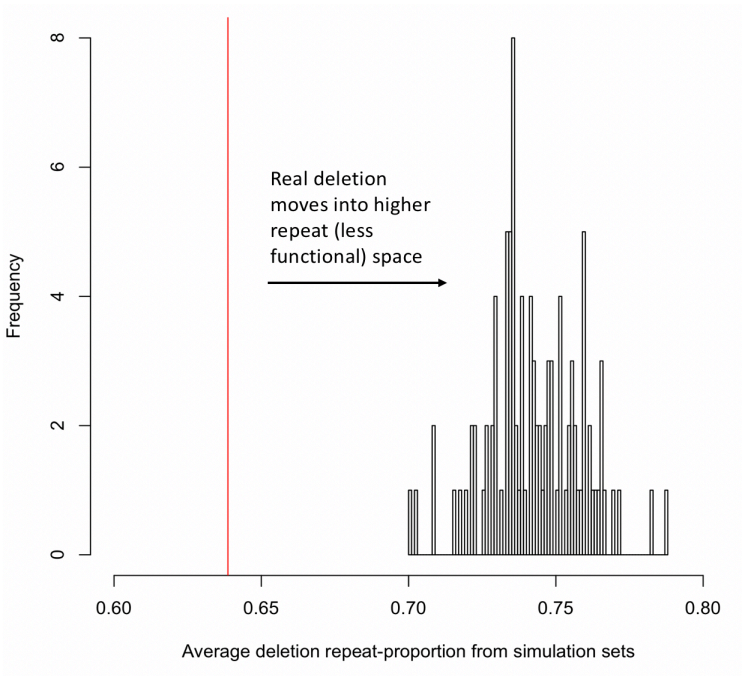

Figure S5. Example in simulations of deletion 'spreading' from more-unique to less-unique regions. Simulation values are based on RepeatMasker annotation.

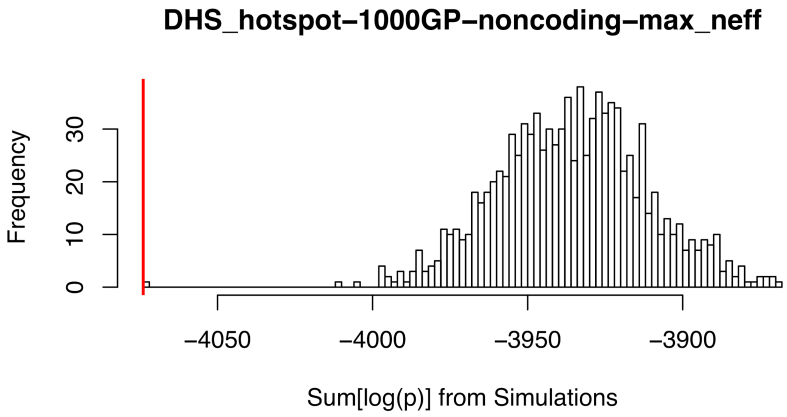

Figure S6. Sum of log(empirical p-value) distribution example. Sum of natural logarithm of empirical p-value from real 1000GP deletion set (red line with its datapoint) compared to simulations for $PlyRS_{max}$. *t*-test is significant at $5.84x10^{-09}$ indicating that the 1000GP deletion set is depleted of $PlyRS_{max}$ compared to expectation.

Table S1. Depletion simulation results for DHS, enhancer, transcribed, polycomb-repressed, or heterochromatin ($PlyRS_{sum}$).

regclass	*dataset*	*gencompartment*	*cohensd*	*cohensd_low*	*cohensd_high*
DHS_hotspot	1000GP	noncoding	6.70	6.40	7.00
DHS_hotspot	1000GP	intronic	5.04	4.81	5.27
DHS_hotspot	1000GP	intergenic	4.31	4.11	4.51
DHS_hotspot	ADNI	noncoding	5.45	5.21	5.70
DHS_hotspot	ADNI	intronic	3.44	3.28	3.61
DHS_hotspot	ADNI	intergenic	4.25	4.05	4.45
DHS_MACS	1000GP	noncoding	7.14	6.82	7.46
DHS_MACS	1000GP	intronic	5.73	5.48	5.99
DHS_MACS	1000GP	intergenic	4.28	4.09	4.48
DHS_MACS	ADNI	noncoding	5.49	5.25	5.74
DHS_MACS	ADNI	intronic	3.69	3.51	3.86
DHS_MACS	ADNI	intergenic	4.08	3.89	4.27
enhancer_H3K4me1	1000GP	noncoding	4.32	4.12	4.52
enhancer_H3K4me1	1000GP	intronic	2.54	2.41	2.67
enhancer_H3K4me1	1000GP	intergenic	3.86	3.68	4.04
enhancer_H3K4me1	ADNI	noncoding	4.87	4.65	5.09
enhancer_H3K4me1	ADNI	intronic	3.06	2.91	3.21
enhancer_H3K4me1	ADNI	intergenic	4.04	3.85	4.23
transcribed_H3K36me3	1000GP	noncoding	-0.70	-0.77	-0.63
transcribed_H3K36me3	1000GP	intronic	-1.19	-1.27	-1.11
transcribed_H3K36me3	1000GP	intergenic	2.19	2.08	2.31
transcribed_H3K36me3	ADNI	noncoding	-1.71	-1.80	-1.61
transcribed_H3K36me3	ADNI	intronic	-1.98	-2.09	-1.87

transcribed_H3K36me3	ADNI	intergenic	0.69	0.62	0.76
polycomb_H3K27me3	1000GP	noncoding	0.26	0.20	0.32
polycomb_H3K27me3	1000GP	intronic	1.58	1.49	1.67
polycomb_H3K27me3	1000GP	intergenic	-0.85	-0.92	-0.77
polycomb_H3K27me3	ADNI	noncoding	2.18	2.06	2.29
polycomb_H3K27me3	ADNI	intronic	1.36	1.27	1.44
polycomb_H3K27me3	ADNI	intergenic	1.74	1.64	1.84
heterochromatin_H3K9me3	1000GP	noncoding	-1.18	-1.26	-1.10
heterochromatin_H3K9me3	1000GP	intronic	-1.49	-1.58	-1.40
heterochromatin_H3K9me3	1000GP	intergenic	-0.48	-0.55	-0.42
heterochromatin_H3K9me3	ADNI	noncoding	-3.40	-3.56	-3.24
heterochromatin_H3K9me3	ADNI	intronic	-2.18	-2.29	-2.07
heterochromatin_H3K9me3	ADNI	intergenic	-2.61	-2.74	-2.48

Table S2. Logistic regression results for DHS, enhancer, transcribed, polycomb-repressed, or heterochromatin ($PlyRS_{sum}$).

regclass	*dataset*	*gencompartment*	*OR_estimate*	*OR_estimate_lower*	*OR_estimate_upper*
DHS_hotspot	1000GP	noncoding	1.07	1.04	1.11
DHS_hotspot	1000GP	intronic	1.09	1.04	1.14
DHS_hotspot	1000GP	intergenic	1.06	1.02	1.10
DHS_hotspot	ADNI	noncoding	1.11	1.01	1.23
DHS_hotspot	ADNI	intronic	1.09	0.94	1.29
DHS_hotspot	ADNI	intergenic	1.12	1.00	1.28
DHS_MACS	1000GP	noncoding	1.09	1.05	1.13
DHS_MACS	1000GP	intronic	1.11	1.06	1.18
DHS_MACS	1000GP	intergenic	1.07	1.03	1.12
DHS_MACS	ADNI	noncoding	1.17	1.05	1.31
DHS_MACS	ADNI	intronic	1.08	0.92	1.29
DHS_MACS	ADNI	intergenic	1.23	1.07	1.43
enhancer_H3K4me1	1000GP	noncoding	1.06	1.02	1.10
enhancer_H3K4me1	1000GP	intronic	1.11	1.05	1.18
enhancer_H3K4me1	1000GP	intergenic	1.04	1.00	1.08
enhancer_H3K4me1	ADNI	noncoding	1.09	0.99	1.20
enhancer_H3K4me1	ADNI	intronic	0.96	0.81	1.14
enhancer_H3K4me1	ADNI	intergenic	1.16	1.04	1.31
transcribed_H3K36me3	1000GP	noncoding	1.02	1.00	1.04
transcribed_H3K36me3	1000GP	intronic	1.07	1.02	1.12
transcribed_H3K36me3	1000GP	intergenic	1.01	1.00	1.03
transcribed_H3K36me3	ADNI	noncoding	1.00	0.97	1.03
transcribed_H3K36me3	ADNI	intronic	1.07	0.96	1.22

transcribed_H3K36me3	ADNI	intergenic	0.99	0.97	1.02
polycomb_H3K27me3	1000GP	noncoding	1.00	0.98	1.03
polycomb_H3K27me3	1000GP	intronic	0.99	0.95	1.03
polycomb_H3K27me3	1000GP	intergenic	1.03	0.99	1.07
polycomb_H3K27me3	ADNI	noncoding	1.02	0.94	1.12
polycomb_H3K27me3	ADNI	intronic	0.94	0.84	1.06
polycomb_H3K27me3	ADNI	intergenic	1.13	1.00	1.28
heterochromatin_H3K9me3	1000GP	noncoding	1.03	1.00	1.06
heterochromatin_H3K9me3	1000GP	intronic	1.04	1.00	1.08
heterochromatin_H3K9me3	1000GP	intergenic	1.02	0.98	1.07
heterochromatin_H3K9me3	ADNI	noncoding	1.06	0.98	1.15
heterochromatin_H3K9me3	ADNI	intronic	1.00	0.90	1.11
heterochromatin_H3K9me3	ADNI	intergenic	1.14	1.02	1.29

Table S3. Depletion simulation results for DHS or enhancer ($PlyRS_{sum\text{-}mono}$).

regclass	*dataset*	*gencompartment*	*cohensd*	*cohensd_low*	*cohensd_high*
DHS_hotspot	1000GP	noncoding	4.46	4.25	4.66
DHS_hotspot	1000GP	intronic	2.95	2.81	3.10
DHS_hotspot	1000GP	intergenic	3.35	3.19	3.51
DHS_hotspot	ADNI	noncoding	4.14	3.94	4.33
DHS_hotspot	ADNI	intronic	1.88	1.78	1.98
DHS_hotspot	ADNI	intergenic	4.01	3.82	4.20
DHS_MACS	1000GP	noncoding	6.22	5.94	6.50
DHS_MACS	1000GP	intronic	4.75	4.53	4.96
DHS_MACS	1000GP	intergenic	3.91	3.73	4.09
DHS_MACS	ADNI	noncoding	4.48	4.27	4.68
DHS_MACS	ADNI	intronic	2.96	2.82	3.10
DHS_MACS	ADNI	intergenic	3.35	3.19	3.50
enhancer_H3K4me1	1000GP	noncoding	3.32	3.16	3.48
enhancer_H3K4me1	1000GP	intronic	2.20	2.08	2.31
enhancer_H3K4me1	1000GP	intergenic	2.66	2.53	2.79
enhancer_H3K4me1	ADNI	noncoding	4.23	4.03	4.43
enhancer_H3K4me1	ADNI	intronic	2.42	2.29	2.54
enhancer_H3K4me1	ADNI	intergenic	3.66	3.49	3.84
transcribed_H3K36me3	1000GP	noncoding	0.68	0.61	0.75
transcribed_H3K36me3	1000GP	intronic	0.19	0.13	0.25
transcribed_H3K36me3	1000GP	intergenic	1.76	1.66	1.86
transcribed_H3K36me3	ADNI	noncoding	-0.56	-0.63	-0.50
transcribed_H3K36me3	ADNI	intronic	-0.71	-0.78	-0.64

transcribed_H3K36me3	ADNI	intergenic	0.38	0.32	0.45
polycomb_H3K27me3	1000GP	noncoding	1.46	1.37	1.54
polycomb_H3K27me3	1000GP	intronic	1.93	1.82	2.03
polycomb_H3K27me3	1000GP	intergenic	0.36	0.30	0.43
polycomb_H3K27me3	ADNI	noncoding	2.04	1.93	2.15
polycomb_H3K27me3	ADNI	intronic	1.34	1.25	1.42
polycomb_H3K27me3	ADNI	intergenic	1.58	1.48	1.67
heterochromatin_H3K9me3	1000GP	noncoding	-0.45	-0.51	-0.38
heterochromatin_H3K9me3	1000GP	intronic	0.49	0.42	0.55
heterochromatin_H3K9me3	1000GP	intergenic	-0.90	-0.97	-0.83
heterochromatin_H3K9me3	ADNI	noncoding	1.59	1.50	1.69
heterochromatin_H3K9me3	ADNI	intronic	-1.58	-1.67	-1.48
heterochromatin_H3K9me3	ADNI	intergenic	2.88	2.74	3.02

Table S4. Depletion simulation results for DHS or enhancer ($PlyRS_{sum\text{-}pleio}$).

regclass	*dataset*	*gencompartment*	*cohensd*	*cohensd_low*	*cohensd_high*
DHS_hotspot	1000GP	noncoding	6.67	6.38	6.97
DHS_hotspot	1000GP	intronic	5.05	4.82	5.28
DHS_hotspot	1000GP	intergenic	4.32	4.12	4.51
DHS_hotspot	ADNI	noncoding	4.96	4.73	5.19
DHS_hotspot	ADNI	intronic	3.61	3.44	3.78
DHS_hotspot	ADNI	intergenic	3.51	3.35	3.68
DHS_MACS	1000GP	noncoding	6.99	6.67	7.30
DHS_MACS	1000GP	intronic	5.44	5.19	5.69
DHS_MACS	1000GP	intergenic	4.19	4.00	4.39
DHS_MACS	ADNI	noncoding	5.22	4.98	5.46
DHS_MACS	ADNI	intronic	3.56	3.39	3.73
DHS_MACS	ADNI	intergenic	3.91	3.73	4.10
enhancer_H3K4me1	1000GP	noncoding	4.44	4.24	4.64
enhancer_H3K4me1	1000GP	intronic	2.99	2.84	3.13
enhancer_H3K4me1	1000GP	intergenic	3.68	3.50	3.85
enhancer_H3K4me1	ADNI	noncoding	4.04	3.85	4.22
enhancer_H3K4me1	ADNI	intronic	2.89	2.75	3.03
enhancer_H3K4me1	ADNI	intergenic	2.95	2.81	3.10
transcribed_H3K36me3	1000GP	noncoding	-1.90	-2.00	-1.80
transcribed_H3K36me3	1000GP	intronic	-1.90	-2.01	-1.80
transcribed_H3K36me3	1000GP	intergenic	0.21	0.15	0.27
transcribed_H3K36me3	ADNI	noncoding	-1.78	-1.88	-1.68
transcribed_H3K36me3	ADNI	intronic	-1.81	-1.91	-1.71

transcribed_H3K36me3	ADNI	intergenic	0.91	0.84	0.99
polycomb_H3K27me3	1000GP	noncoding	0.00	-0.06	0.06
polycomb_H3K27me3	1000GP	intronic	1.49	1.40	1.57
polycomb_H3K27me3	1000GP	intergenic	-0.95	-1.03	-0.88
polycomb_H3K27me3	ADNI	noncoding	1.74	1.64	1.84
polycomb_H3K27me3	ADNI	intronic	1.79	1.69	1.89
polycomb_H3K27me3	ADNI	intergenic	1.00	0.92	1.07
heterochromatin_H3K9me3	1000GP	noncoding	-0.96	-1.04	-0.89
heterochromatin_H3K9me3	1000GP	intronic	-2.47	-2.60	-2.35
heterochromatin_H3K9me3	1000GP	intergenic	0.26	0.20	0.33
heterochromatin_H3K9me3	ADNI	noncoding	-4.72	-4.94	-4.51
heterochromatin_H3K9me3	ADNI	intronic	-2.49	-2.62	-2.37
heterochromatin_H3K9me3	ADNI	intergenic	-3.92	-4.10	-3.74

Table S5. Logistic regression results for DHS or enhancer ($PlyRS_{sum\text{-}mono}$).

regclass	*dataset*	*gencompartment*	*OR_estimate*	*OR_estimate_lower*	*OR_estimate_upper*
DHS_hotspot	1000GP	noncoding	1.02	1.00	1.04
DHS_hotspot	1000GP	intronic	1.02	0.99	1.05
DHS_hotspot	1000GP	intergenic	1.02	0.99	1.05
DHS_hotspot	ADNI	noncoding	1.04	0.97	1.13
DHS_hotspot	ADNI	intronic	1.03	0.92	1.17
DHS_hotspot	ADNI	intergenic	1.05	0.95	1.17
DHS_MACS	1000GP	noncoding	1.04	1.01	1.07
DHS_MACS	1000GP	intronic	1.04	1.00	1.09
DHS_MACS	1000GP	intergenic	1.04	1.00	1.07
DHS_MACS	ADNI	noncoding	1.13	1.02	1.27
DHS_MACS	ADNI	intronic	1.02	0.87	1.20
DHS_MACS	ADNI	intergenic	1.23	1.06	1.45
enhancer_H3K4me1	1000GP	noncoding	1.02	0.99	1.04
enhancer_H3K4me1	1000GP	intronic	1.00	0.97	1.04
enhancer_H3K4me1	1000GP	intergenic	1.03	1.00	1.06
enhancer_H3K4me1	ADNI	noncoding	1.08	1.00	1.18
enhancer_H3K4me1	ADNI	intronic	1.01	0.89	1.15
enhancer_H3K4me1	ADNI	intergenic	1.14	1.02	1.28
transcribed_H3K36me3	1000GP	noncoding	1.01	1.00	1.02
transcribed_H3K36me3	1000GP	intronic	1.01	0.97	1.04
transcribed_H3K36me3	1000GP	intergenic	1.01	0.99	1.02
transcribed_H3K36me3	ADNI	noncoding	1.00	0.96	1.04
transcribed_H3K36me3	ADNI	intronic	1.01	0.91	1.14

transcribed_H3K36me3	ADNI	intergenic	0.99	0.96	1.04
polycomb_H3K27me3	1000GP	noncoding	1.00	0.98	1.02
polycomb_H3K27me3	1000GP	intronic	0.99	0.96	1.02
polycomb_H3K27me3	1000GP	intergenic	1.02	0.99	1.05
polycomb_H3K27me3	ADNI	noncoding	1.00	0.93	1.08
polycomb_H3K27me3	ADNI	intronic	0.97	0.88	1.09
polycomb_H3K27me3	ADNI	intergenic	1.03	0.93	1.15
heterochromatin_H3K9me3	1000GP	noncoding	1.02	1.00	1.04
heterochromatin_H3K9me3	1000GP	intronic	1.03	1.00	1.07
heterochromatin_H3K9me3	1000GP	intergenic	1.00	0.97	1.04
heterochromatin_H3K9me3	ADNI	noncoding	1.09	0.99	1.20
heterochromatin_H3K9me3	ADNI	intronic	1.11	0.98	1.28
heterochromatin_H3K9me3	ADNI	intergenic	1.06	0.93	1.23

Table S6. Logistic regression results for DHS or enhancer ($PlyRS_{sum\text{-}pleio}$).

regclass	*dataset*	*gencompartment*	*OR_estimate*	*OR_estimate_lower*	*OR_estimate_upper*
DHS_hotspot	1000GP	noncoding	1.04	1.02	1.06
DHS_hotspot	1000GP	intronic	1.06	1.03	1.10
DHS_hotspot	1000GP	intergenic	1.02	1.00	1.05
DHS_hotspot	ADNI	noncoding	1.07	1.00	1.16
DHS_hotspot	ADNI	intronic	1.05	0.96	1.19
DHS_hotspot	ADNI	intergenic	1.08	0.99	1.20
DHS_MACS	1000GP	noncoding	1.04	1.02	1.06
DHS_MACS	1000GP	intronic	1.07	1.04	1.12
DHS_MACS	1000GP	intergenic	1.02	1.00	1.05
DHS_MACS	ADNI	noncoding	1.07	1.01	1.16
DHS_MACS	ADNI	intronic	1.06	0.97	1.21
DHS_MACS	ADNI	intergenic	1.08	1.00	1.21
enhancer_H3K4me1	1000GP	noncoding	1.03	1.01	1.05
enhancer_H3K4me1	1000GP	intronic	1.06	1.02	1.10
enhancer_H3K4me1	1000GP	intergenic	1.02	0.99	1.04
enhancer_H3K4me1	ADNI	noncoding	1.04	0.99	1.11
enhancer_H3K4me1	ADNI	intronic	0.99	0.90	1.11
enhancer_H3K4me1	ADNI	intergenic	1.07	1.00	1.16
transcribed_H3K36me3	1000GP	noncoding	1.01	1.00	1.02
transcribed_H3K36me3	1000GP	intronic	1.05	1.02	1.08
transcribed_H3K36me3	1000GP	intergenic	1.00	1.00	1.01
transcribed_H3K36me3	ADNI	noncoding	1.00	0.99	1.01
transcribed_H3K36me3	ADNI	intronic	1.08	1.00	1.18

transcribed_H3K36me3	ADNI	intergenic	1.00	0.99	1.01
polycomb_H3K27me3	1000GP	noncoding	1.00	0.99	1.02
polycomb_H3K27me3	1000GP	intronic	1.00	0.99	1.02
polycomb_H3K27me3	1000GP	intergenic	1.01	0.99	1.04
polycomb_H3K27me3	ADNI	noncoding	1.02	0.98	1.07
polycomb_H3K27me3	ADNI	intronic	0.99	0.96	1.05
polycomb_H3K27me3	ADNI	intergenic	1.07	1.00	1.17
heterochromatin_H3K9me3	1000GP	noncoding	1.01	1.00	1.03
heterochromatin_H3K9me3	1000GP	intronic	1.01	1.00	1.03
heterochromatin_H3K9me3	1000GP	intergenic	1.02	0.99	1.04
heterochromatin_H3K9me3	ADNI	noncoding	1.03	1.00	1.07
heterochromatin_H3K9me3	ADNI	intronic	1.00	0.96	1.05
heterochromatin_H3K9me3	ADNI	intergenic	1.07	1.02	1.15

Table S7. Depletion simulation results for chromatin loop anchors (binary).

regclass	*dataset*	*gencompartment*	*cohensd*	*cohensd_low*	*cohensd_high*
loops	1000GP	noncoding	2.63	2.50	2.76
loops	1000GP	intronic	1.63	1.54	1.73
loops	1000GP	intergenic	2.10	1.99	2.21
loops	ADNI	noncoding	0.45	0.39	0.52
loops	ADNI	intronic	-0.27	-0.34	-0.21
loops	ADNI	intergenic	0.93	0.86	1.00
loops_noCTCF	1000GP	noncoding	1.62	1.52	1.71
loops_noCTCF	1000GP	intronic	0.80	0.73	0.87
loops_noCTCF	1000GP	intergenic	1.52	1.43	1.61
loops_noCTCF	ADNI	noncoding	-0.36	-0.42	-0.29
loops_noCTCF	ADNI	intronic	-0.54	-0.60	-0.47
loops_noCTCF	ADNI	intergenic	0.06	0.00	0.12
loops_CTCF	1000GP	noncoding	3.76	3.59	3.94
loops_CTCF	ADNI	noncoding	1.37	1.28	1.45

Table S8. Depletion simulation results for chromatin loop anchors ($PlyRS_{max}$).

regclass	*dataset*	*gencompartment*	*cohensd*	*cohensd_low*	*cohensd_high*
loops	1000GP	noncoding	2.39	2.27	2.51
loops	1000GP	intronic	1.45	1.36	1.53
loops	1000GP	intergenic	2.01	1.90	2.11
loops	ADNI	noncoding	0.97	0.89	1.04
loops	ADNI	intronic	0.39	0.32	0.45
loops	ADNI	intergenic	1.00	0.93	1.08
loops_noCTCF	1000GP	noncoding	1.25	1.16	1.33
loops_noCTCF	1000GP	intronic	0.48	0.41	0.54
loops_noCTCF	1000GP	intergenic	1.34	1.25	1.42
loops_noCTCF	ADNI	noncoding	-0.06	-0.12	0.01
loops_noCTCF	ADNI	intronic	-0.34	-0.40	-0.27
loops_noCTCF	ADNI	intergenic	0.28	0.22	0.34
loops_CTCF	1000GP	noncoding	3.81	3.64	3.99
loops_CTCF	ADNI	noncoding	1.20	1.12	1.28

Table S9. Logistic regression results for chromatin loop anchors (binary).

regclass	*dataset*	*gencompartment*	*OR_estimate*	*OR_estimate_lower*	*OR_estimate_upper*
loops	1000GP	noncoding	1.19	1.03	1.39
loops	1000GP	intronic	1.13	0.92	1.40
loops	1000GP	intergenic	1.28	1.03	1.62
loops	ADNI	noncoding	1.66	1.15	2.49
loops	ADNI	intronic	1.64	0.96	2.98
loops	ADNI	intergenic	1.71	1.02	3.04
loops_noCTCF	1000GP	noncoding	1.15	0.98	1.34
loops_noCTCF	1000GP	intronic	1.12	0.91	1.39
loops_noCTCF	1000GP	intergenic	1.19	0.95	1.51
loops_noCTCF	ADNI	noncoding	1.42	0.97	2.15
loops_noCTCF	ADNI	intronic	1.37	0.80	2.49
loops_noCTCF	ADNI	intergenic	1.50	0.89	2.73
loops_CTCF	1000GP	noncoding	2.70	1.35	6.37
loops_CTCF	ADNI	noncoding	7.67	1.76	108.20

Table S10. Logistic regression results for chromatin loop anchors ($PlyRS_{max}$).

regclass	*dataset*	*gencompartment*	*OR_estimate*	*OR_estimate_lower*	*OR_estimate_upper*
loops	1000GP	noncoding	1.46	1.01	2.17
loops	1000GP	intronic	1.20	0.72	2.07
loops	1000GP	intergenic	1.84	1.07	3.31
loops	ADNI	noncoding	4.13	1.45	14.24
loops	ADNI	intronic	3.93	0.82	28.51
loops	ADNI	intergenic	4.44	1.16	24.13
loops _noCTCF	1000GP	noncoding	1.28	0.87	1.93
loops _noCTCF	1000GP	intronic	1.18	0.70	2.08
loops _noCTCF	1000GP	intergenic	1.44	0.82	2.62
loops _noCTCF	ADNI	noncoding	3.02	1.02	10.93
loops _noCTCF	ADNI	intronic	2.19	0.45	15.79
loops _noCTCF	ADNI	intergenic	4.04	0.95	25.41
loops _CTCF	1000GP	noncoding	36.80	3.49	1,277.08
loops _CTCF	ADNI	noncoding	30.11	1.27	8,491.69

Table S11. Filtered deletion callset characteristics for 1000GP and ADNI. The abbreviations 'AF' and 'bp' correspond to allele frequency and DNA base-pairs, respectively.

	ADNI	*1000GP*
Number of Deletions	3,306 (100%)	12,013 (100%)
Number of Intronic Deletions	1,459 (44.1%)	5,896 (49.1%)
Number of Intergenic Deletions	1,847 (55.9%)	6,117 (50.9%)
Singleton AF	2,007 (60.7%)	4,362 (36.3%)
Doubleton AF	327 (9.9%)	1,241 (10.3%)
Tripleton AF	163 (4.9%)	617 (5.1%)
>1% AF	395 (11.9%)	2,165 (18.0%)
Average Length	2,722 bp	1,437 bp
Median Length	1,893 bp	629 bp
Minimum Length	440 bp	50 bp
Maximum Length	23,344 bp	22,648 bp
Average Length Singleton AF	2,760 bp	1,400 bp
Median Length Singleton AF	1,924 bp	448 bp
Average Length >1% AF	3,072 bp	979 bp
Median Length >1% AF	2,297 bp	355 bp

Table S12. Tissues and cell types analyzed from REC.

Tissue/Cell type Description	*Consolidated Epigenome ID*	*Fetal (F) or Adult (A)*
Primary monocytes from peripheral blood	E029	A
Primary B cells from peripheral blood	E032	A
Primary T cells from cord blood	E033	F
Primary T cells from peripheral blood	E034	A
Primary Natural Killer cells from peripheral blood	E046	A
Primary hematopoietic stem cells G-CSF-mobilized Female	E050	F
Primary hematopoietic stem cells G-CSF-mobilized Male	E051	A
Fetal Adrenal Gland	E080	F
Fetal Brain Male	E081	F
Fetal Brain Female	E082	F
Fetal Heart	E083	F
Fetal Intestine Large	E084	F
Fetal Intestine Small	E085	F
Fetal Kidney	E086	F
Fetal Lung	E088	F
Fetal Muscle Trunk	E089	F
Fetal Muscle Leg	E090	F
Placenta	E091	F
Fetal Stomach	E092	F
Fetal Thymus	E093	F
Gastric	E094	A
Ovary	E097	A
Pancreas	E098	A
Psoas Muscle	E100	A
Small Intestine	E109	A

Table S13. European subject cohort within ADNI. Of the 803 ADNI subjects for which we analyzed whole genome sequencing data, 752 subjects were determined to be of European ancestry using principal components analysis. 'Control' phenotype corresponds to brain-healthy cognition. 'MCI' phenotype corresponds to mild-cognitive impairment. 'AD' phenotype corresponds to Alzheimer's Disease diagnosis. Deletions genotyped within these individuals were selected for further downstream analysis.

Phenotype	*Control*	*MCI*	*AD*
Number of Subjects	233	342	177
Percentage of Cohort	31	45	24

Table S14. ADNI deletion callset characteristics. Deletion allele frequency and length characteristics of the quality-controlled and filtered final deletion callset among 752 European-ancestry ADNI individuals. The abbreviations 'AF' and 'bp' correspond to allele frequency and DNA base-pairs, respectively.

Deletion Callset Characteristic	
Number of Deletions	10,619 (100%)
Singleton AF	6,443 (60.7%)
Doubleton AF	1,051 (9.9%)
Tripleton AF	483 (4.5%)
>=1% AF	1,366 (12.9%)
Average Length	9,285 bp
Median Length	3,114 bp
Minimum Length	440 bp
Maximum Length	853,585 bp
Average Length Singleton AF	10,735 bp
Median Length Singleton AF	3,244 bp
Average Length >=1% AF	6,155 bp
Median Length >=1% AF	3,367 bp

S1 Alzheimer's Disease Neuroimaging Initiative (ADNI)

S1.1 Brief Overview

Whole genome sequencing (WGS) was previously performed on >800 participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al. 2010). For this project, we compiled a deletion callset on 752 ADNI individuals of European ancestry using the published deletion algorithm GenomeSTRiP (Handsaker et al. 2011). We examined re-aligned analysis-ready BAM files (~150 terabytes), which were generated by the Broad Institute using their GATK best practices pipeline. By combining a variety of computational quality-control criteria and filters (see S1.3 Data Processing From BAMs) on the deletion variants identified by GenomeSTRiP, we arrived at a callset of 10,619 autosomal deletions. A schematic overview of our deletion callset generation is shown in Figure S2. We identified that our deletion calls (see S1.4 Technical Validation): a) have the qualitatively expected allele frequency of a robust population dataset; b) have high concordance with 1000 Genomes Project-identified (Sudmant et al. 2015a) common deletions (>=5%) for deletions at or longer than our median length of 3,114 base-pairs (bp); c) have a well-behaved, flat quantile-quantile plot distribution, as would be expected from a dataset such as the ADNI sequencing which is likely underpowered to identify any single variants of large phenotypic effect, indicating high-quality genotyping across subsets of the ADNI data. We provide three possible applications for which these data may be useful to researchers (see S1.5 Possible Applications) and make these data publicly available to registered users of the ADNI (see S1.6 Data Availability).

S1.2 Background

It is estimated that by 2050, the prevalence of Alzheimer's Disease (AD), the most common form of mental deterioration amongst adults, will quadruple, affecting 1 in 85 people worldwide (Brookmeyer et al. 2007), fostering tremendous research interest to understand the

development and progression of AD. The Alzheimer's Disease Neuroimaging Initiative (ADNI) was launched in 2003 in order to address these research challenges. ADNI's goal is to combine clinical imaging and neuropsychological assessments, along with genetic and other biological data, to study and measure the progression of cognitive impairment into early diagnosis of AD. In 2012, high-coverage (>40x average coverage) whole genome sequencing (WGS) on DNA derived from whole blood of 818 ADNI subjects was performed, in order to assess how genomic variants might be contributing to progression toward, and into, AD.

Many studies have identified a variety of single nucleotide variants (SNVs) throughout the genome that are significantly associated with Alzheimer's Disease (Zhang et al. 2013). There has also been interest in analysis of copy number variants, in particular deletions, in regards to their association to AD susceptibility. Genomic deletions (the loss of genetic sequences on loci scattered across the genome) provide a layer of interpretation often not available with SNVs: the loss of function (LoF) of the underlying sequence that a deletion removes (in a heterozygous or homozygous manner). This additional layer of interpretation is especially important in noncoding regulatory regions of the genome, where there is no obvious way to identify LoF SNVs as there is in coding regions. However, most of the previous studies have only been able to examine very long deletion events (>100 kilo-bp [kb]) (Cuccaro et al. 2016), because of the technologies available, thereby missing deletion events of potential association to AD, especially in noncoding genomic regions (for which single nucleotide polymorphism array probes are less dense compared to genic regions) where the majority of AD genome wide association study (GWAS) signals are located (Han et al. 2017). Now, with high-coverage WGS data on ADNI individuals, and the development of higher-resolution deletion algorithms, increased sensitivity to identify shorter deletion variants is possible.

S1.3 Data Processing From BAMs

Whole genome sequencing and alignment

Whole genome sequencing (WGS) on DNA derived from whole blood of 818 ADNI subjects was performed by Illumina's laboratory in 2012-2013, using Illumina HiSeq sequencers, generating 100 bp paired-end sequence reads. In 2014, the Broad Institute donated resources to take the Illumina-generated BAM files and re-process the data using Broad's best practices GATK pipeline (McKenna et al. 2010; Depristo et al. 2011; Van der Auwera et al. 2013), with the goal to create improved SNP and indel call accuracy over that generated from Illumina's CASAVA software. Starting from recovered FASTQ files, the sequence reads were mapped to the human reference genome (GRCh37) using BWA-MEM (Li 2013), and then processed using the GATK pipeline (version GenomeAnalysisTK-3.1-144-g00f68a3.jar), resulting in analysis-ready BAM files. Of the 818 ADNI subjects, nine were deemed to have provided insufficient consent, and one was dropped due to quality control issues during re-processing, leaving 808 subjects on which analysis of their WGS data was subsequently performed. Average genome-wide coverage across the 808 BAM files was ~42x (median: ~41x, range: 33x-81x). The storage size of the 808 BAM files on the disk drive was approximately 150 terabytes.

ADNI Cohort

Data used in this research were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. The ADNI has been a highly successful and collaborative community effort (Saykin et al. 2015; Weiner et al. 2015), with ADNI data being utilized in over 1,800 scientific publications as of early 2018.

SNV and indel variant discovery

Broad Institute's GATK pipeline was run to generate raw single nucleotide variant (SNV) and insertion-deletion (indel) calls. GATK HaplotypeCaller was run on each BAM sample separately, producing single-sample gVCF files. These files were merged into a single gVCF file (using GATK CombineGVCFs). Joint genotyping was then performed (using GATK GenotypeGVCFs) across all 808 samples to produce variant calls. These calls were subsequently filtered (using GATK VariantRecalibrator and GATK ApplyRecalibration) to take into account both sensitivity and specificity. Variant calls that failed the 'Variant Quality Score Recalibration' step of the GATK pipeline were excluded, and genotypes with a genotype quality (GQ) score of <= 20 were set to missing. None of the individuals had a SNV genotype missing rate greater than 1.54%. These procedures produced a total set of approximately 45 million non-monomorphic sequence variants (38,443,567 SNVs and 6,092,213 indels). The transition/transversion (Ti/Tv) ratio of 2.02 for novel SNVs discovered in our callset was comparable with the Ti/Tv ratio of ~2.2 for SNVs catalogued in dbSNP (https://www.ncbi.nlm.nih.gov/snp/). Also, genic SNVs annotated using MapSNPs (part of the PolyPhen-2 software suite (Adzhubei et al. 2010) resulted in a Ti/Tv ratio for coding SNVs of 2.97, comparable with findings in other datasets of ~3 (Emond et al. 2012).

Genotype concordance with microarrays

To ensure the BAM file re-processing procedure was performed at high-quality for the 808 samples, SNV variants discovered in our callset were compared with SNVs identified on Illumina Omni 2.5M microarray data previously performed on the same ADNI subjects in 2013. We compared the 'PASS' SNVs found in WGS sequencing with those previously called using microarrays using the PLINK software suite (Purcell et al. 2007). We examined biallelic non-monomorphic SNVs using only SNVs whose strands could be matched (including by using the '-

-flip' option), while excluding SNVs that had the same position but different reference SNV cluster (rs) IDs and excluding SNVs with a missing rate >10%. Only non-missing genotypes in both datasets were considered. Genotype concordance analysis was performed on the called SNVs (total of 1,987,307 SNVs) and all of the 808 ADNI samples had at least 99.88% genotype concordance (average of 99.95%) between SNVs found in WGS and those found in the microarray data.

Identity by descent (IBD) analysis

To identify any close genetically-related individuals amongst the 808 samples, identity by descent (IBD) estimates called PI_HAT were calculated on bi-allelic SNVs using PLINK. To ensure only high-quality genotypes were used in estimating PI_HAT, stringent quality control filters using PLINK options were employed. These included removal of SNVs with: a) Hardy-Weinberg Equilibrium (HWE) p-value (using the '--hwe' option) of $<1 \times 10^{-5}$ for common variants (defined as SNVs with minor allele frequency (MAF) ≥5%) and $<1 \times 10^{-2}$ for rare variants (defined as SNVs with MAF <5%); b) genotype missing rate of >0.5% (using the '--geno' option); c) genotype concordance rate to microarray data of <99%. Additionally, we performed linkage disequilibrium (LD)-based SNV pruning to obtain a set of independent common SNVs. This procedure involved: a) removal of SNVs with MAF <15%; b) application of variance inflation factor (VIF)-based LD-pruning (using the '--indep 200 5 1.15' option); c) application of pairwise genotypic correlation-based LD-pruning (using the '--indep-pairwise 100 5 0.1' option). Altogether, after the quality control filters and LD-pruning procedures, 54,210 SNVs were used to compute PI_HAT. Five pairs of samples were identified that had PI_HAT between 0.4-0.6, indicating first-degree relatives. There were no pairs of samples with PI_HAT between 0.2-0.4 (second-degree relatives).

Principal components analysis (PCA)

To perform a principal components analysis on the WGS ADNI data, the EIGENSTRAT algorithm (Price et al. 2006) was used along with data from 1000GP as a reference panel (1000G Phase I v3 ShapeIt2 Reference; 2013-09 haplotype: http://csg.sph.umich.edu/abecasis/MACH/download/1000G.2013-09.html) (1000 Genomes Project Consortium et al. 2012). We performed the previously discussed IBD analysis on the 1000GP reference panel dataset and identified 65 individuals with PI_HAT >0.2, indicating relatedness of at-least second-degree. These 1000GP individuals and one random individual of each pair of related ADNI samples were dropped from the PCA analysis. To obtain high-quality SNVs for the PCA, the same stringent quality control filters were used as in the IBD analysis, except that variants with <5% MAF were removed (rather than <15% MAF). The 1000GP and ADNI datasets were then merged using overlapping variants, while removing SNVs with inconsistent strands. Additionally, the same LD-based SNV pruning procedures were performed as in the IBD analysis. Altogether, a set of 92,000 independent SNVs were given as input for EIGENSTRAT to compute principal components. Individuals having a PC1 of <-0.01 were considered to be European ancestry (EU) individuals. Out of the 803 ADNI individuals analyzed, 752 were deemed to be EU (~94%). Among the 752 subjects, roughly 31% (233) were classified as brain-normal controls, roughly 45% (342) were classified as exhibiting mild cognitive impairment (MCI), and roughly 24% (177) were classified as exhibiting Alzheimer's Disease (AD) (Table S13).

Deletion variant discovery

To identify deletion variants >500 bp in length in the ADNI dataset, the published software algorithm GenomeSTRiP (version 1.04.1456) (Handsaker et al. 2011) was run on the 808 ADNI re-processed WGS BAM files. GenomeSTRiP combines three lines of technical sequence evidence for calling deletion candidates: breakpoint-spanning reads (split reads), abnormal read-pair separation, and local variation in read depth of coverage, and was previously found to

be superior compared to other callers in terms of call specificity, sensitivity, and genotype accuracy (Handsaker et al. 2011). Deletions were called with respect to the GRCh37/hg19 version of the human reference genome (the reference genome file used was 'human_g1k_v37.fasta', downloaded from the 1000 Genomes FTP server: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp//technical/reference/human_g1k_v37.fasta.gz). Several genome annotation files are necessary in order for GenomeSTRiP to properly infer genomic deletions from the BAM file data; files compatible with the reference version used were downloaded from the 'svtoolkit' FTP server hosted by the Broad Institute. These included the genome mask: ftp://ftp.broadinstitute.org/pub/svtoolkit//svmasks/human_g1k_v37.mask.100.fasta.gz , the genome ploidy map: ftp://ftp.broadinstitute.org/pub/svtoolkit//ploidymaps/humgen_g1k_v37_ploidy.map , and the genome copy number mask: ftp://ftp.broadinstitute.org/pub/svtoolkit//cn2masks/cn2_mask_g1k_v37.fasta.gz .

GenomeSTRiP deletion detection consisted of three main workflow phases: pre-processing, discovery, and genotyping. Alternative allele alignment was not performed because the underlying data were 100 bp high-coverage WGS BAM files. Default GenomeSTRiP parameters were used. In the pre-processing phase, each ADNI sample was individually processed. BAM file metadata compiled by GenomeSTRiP for each sample was merged for all individuals before the discovery phase. The insert size distribution metadata was merged for all individuals (using the 'org.broadinstitute.sv.apps.MergeInsertSizeDistributions' module). In the discovery phase, all ADNI samples were jointly processed. Minimum and maximum deletion settings were set to 100 and 1,000,000 (using the -minimumSize and -maximumSize options, respectively). To speed GenomeSTRiP computational run time due to the size of the ADNI samples on disk, samples were parallel processed in 5-10 megabase-pair windows (using the -L option), using 1

megabase-pair overlapping windows. The overlapping discovery phase VCF files were merged removing duplicate records using VCFtools (version 0.1.15 using 'vcf-merge' option) (Danacek et al. 2011) and Tabix from htslib (version 1.3.2: http://www.htslib.org/doc/tabix.html). In the genotyping phase, 'PASS' variants from the discovery phase were genotyped across all ADNI samples jointly.

Deletion QC and filtering

Deletions were discovered and genotyped by GenomeSTRiP in all 808 ADNI samples; however, to ensure robust computational quality control (access to original DNA samples not being feasible) and downstream population genetic analysis and inference that relies on limited population demographic parameters, only deletions genotyped in individuals deemed to be of 'European ancestry' (752/803, ~94%) were retained. Additionally, because of population genetic forces potentially differing on the X chromosome necessitating special quality control that would be less reliable with only computational tools, and no Y chromosome calls being generated, only autosomal deletions were selected for further analysis. We additionally removed all deletions that were monomorphic (allele frequency=1), indicating reference genome artifacts or unique insertions.

The remaining deletion calls were then individually screened for properties to compile a high-confidence callset. Criteria were manually set to maximize the total number of deletion calls while also ensuring reasonable quality control of the accuracy in terms of genotypes and genomic coordinates. To ensure high-quality genotyping of the population at each deletion site, deletions were only retained that had a phred-based GQ score for all 752 individuals of >=13 (corresponding to ~95% estimated genotype accuracy). Most individuals in most deletions had a reported GQ of 99.

Deletions in individuals either heterozygous or homozygous for the deletion would be expected to have loss of heterozygosity (LOH) at the genomic coordinates where the deletion resides. Therefore, SNV concordance was measured in relevant individuals within deletion coordinates. Deletions were only retained where all individuals had <=25% SNV discordance. Discordance here is defined as the proportion of SNVs within a deletion's coordinates that are heterozygous in any relevant het-del or hom-del individual using GATK-derived SNV calls as the assumed 'gold-standard' correct genotype. Having a high SNV discordance within a deletion's coordinates across individuals would indicate that either those individuals were not well genotyped for the deletion or that the coordinates of the deletion are largely misspecified. Deletions for which there were no overlapping SNVs were additionally retained.

Because high-quality deletion coordinate localization is important for biological interpretation, only deletions with breakpoint (start coordinate and end coordinate) confidence intervals (as given from GenomeSTRiP output) corresponding to <=3% of deletion length were retained. We use 'average' deletion coordinates to represent final coordinates (not extremes) since we don't want to induce false genomic annotation overlaps in downstream analysis.

Because of the way GenomeSTRiP genotypes candidate deletions, it is possible for a larger deletion overlapping a smaller deletion to be simultaneously genotyped in the same individual. This would be the case, for example, when a longer common-frequency deletion overlaps a shorter deletion that is a singleton/rare-frequency deletion in another individual(s). In situations like this, rule-based filters were used to clarify genotype assignment and collapse redundant calls. These filters were designed to balance remaining deletion counts with also ensuring robust breakpoint accuracy and genotyping accuracy. Filter rules were applied as follows:

- Deletions which extend further in both start and end directions will have corresponding redundant genotypes in shorter, fully-overlapped deletion candidates. Shorter deletion

candidates are collapsed into larger deletions for all matching genotypes. Remaining shorter deletions with now mutually exclusive genotypes are kept as correct calls.

- Overlapping deletions with mutually exclusive genotypes are deemed to be separate deletions.
- If two deletions are completely overlapped or have very close coordinates (>90% overlapping), the deletion genotype with homozygous deletion calls is deemed to be the correct genotype and the other coordinates are ignored.
- Two deletions with very close coordinates (>90% overlapping) that share all genotypes, plus a few additional for one of the deletions, is deemed to be a single deletion event. Since the deletion is likely real and common, the deletion call with the most genotyped individuals is deemed to be the correct deletion genotype and coordinates.
- When two deletions have less than 80% overlapping coordinates, heterozygous and homozygous deletion genotypes are seen as coming from different deletion events (such as when one deletion is common in allele frequency and the other deletion is a singleton).
- When a rare, longer deletion overlaps a rare, shorter deletion, any genotypes shared between the deletions is assigned to the longer deletion since GenomeSTRiP would likely have also assigned the genotype(s) to the shorter one because of the overwhelming technical support given from the longer one.
- When two deletions of similar length have very close overlapping coordinates (>90%), if a singleton/rare call has genotypes also present in the other common deletion, the genotypes for the singleton/rare deletion are retained and removed from the other deletion. In this type of situation, it is likely that GenomeSTRiP added the same genotypes to the common deletion because of the nature of the joint-calling algorithm which may in some circumstances give more weight to common alleles.
- Some deletions cannot be interpreted in light of ambiguous genotyping, such as when one individual genotype is shared between two partially overlapping deletions. When genotypes are not able to be confidently assigned using these filter rules, the corresponding deletions are dropped from further analysis.

Deletions remaining after all prior QC and filtering steps were assessed for violation of Hardy-Weinberg Equilibrium (HWE) (using VCFtools '--hwe' option) in order to identify deletions undergoing obvious selection pressures other than purifying natural selection, or to identify

deletions with low-quality genotyping. The threshold of removal was set to a HWE p-value $<=1x10^{-5}$. Of the remaining deletions, only 0.3% were between the range of $1x10^{-2} < p < 1x10^{-5}$. After application of all QC criteria and filtering steps, a set of 10,619 autosomal deletions genotyped within the 752 ADNI EU individuals remained (see Table S14).

S1.4 Technical Validation

Previous validation has been extensive for the GATK pipeline, widely used in the genomics sequencing community, as well as for the deletion-calling algorithm GenomeSTRiP, employed in multiple consortium efforts including the 1000 Genomes Project (Sudmant et al. 2015a). GenomeSTRiP has been previously found to offer advantages in both sensitivity and specificity in comparison with many other deletion callers (Handsaker et al. 2011). Therefore, we focused our validation efforts on examining the properties of the distributions of sequence variants that we generated, expecting high-quality distributions, qualitatively similar to that observed in previous successful studies using these tools.

Allele frequency spectrum shape

Deletion mutations initially start as a singleton in allele frequency in the population, deriving from a de-novo mutational event in the germline. Deletion mutation recurrence at the same start and end coordinates is extremely unlikely due to chance because of low mutation rates (Kloosterman et al. 2015). Therefore, in considering the shape of the distribution of deletion variant allele frequencies in a population, most events are rare, except a few events that occurred many generations ago (and are therefore present in high frequency across worldwide or broad demographic populations), or have arisen to high frequency due to positive natural selection in favor of the deletion allele. However, since many deletions overlap a functional regulatory element (especially for deletions >10,000 bp), widespread positive selection on deletions is not observed; conversely, negative selection is often observed (Sudmant et al.

2015a). To assess whether our deletion callset matches qualitative expectations in the shape of the allele frequency spectrum, we rank order all deletions by genotype frequency. Figure S3 shows a cumulative fraction plot of the deletion allele frequency across the deletion callset. The smooth curve of the ranking shows the expected pattern of abundant distinct rare variant counts transitioning into infrequent common variant counts (Sudmant et al. 2015a). The dramatic abundance of rare variants in our callset (~75% of our deletion calls are tripleton or lower in allele frequency) compared to other datasets (Sudmant et al. 2015a) is likely at least partially due to enhanced sensitivity in deletion calling available with the high-coverage ADNI WGS data.

Common deletion variant concordance

Rare variant deletion calls, often unique to one particular dataset, are difficult to validate as true calls (some may instead be false positives due to technical artifacts in the underlying data) without experimental evidence in the genotyped samples. However, common deletion variants should persist across datasets when the underlying population samples are from the same broad demographic population. The 1000GP consortium has released a set of breakpoint-resolved deletion calls (Sudmant et al. 2015a Supplemental Table 3), derived from a combination of multiple structural variant callers used in variant discovery and genotyping including GenomeSTRiP, as a part of their goal to characterize common genomic variation in worldwide populations. Taking deletions found to be >=5% allele frequency in the EUR population (broadly of European ancestry) from the 1000GP callset (see S2 1000 Genomes Project Phase 3 [1000GP]), and comparing with ADNI deletions >=1% allele frequency from our quality-controlled and filtered callset, we find that for ADNI deletions greater than or equal to our median call length (3,114 bp), there is a 92.1% (279 ADNI/303 1000GP) concordance rate at a minimum of 90% overlapping coordinates for common 1000GP EUR deletions. This means that for deletions at our median length or longer, we have both high sensitivity and high breakpoint accuracy to correctly identify deletion variants. However, we do note that for lengths below our

median length, we observe deletion sensitivity loss: e.g. there is a 55.6% confirmation rate for ADNI deletions >=1,000 bp at a minimum of 90% overlapping coordinates. This analysis assumes that the 1000GP calls are the correct calls. As with all structural variation callsets (which can have sensitivity loss due to the short-read sequencing technology typically employed, as was the case with the ADNI WGS), the absence of a deletion variant at a particular locus does not indicate that the deletion is not present in the population, rather it indicates that at the particular QC and filter thresholds used (such as by GenomeSTRiP and our downstream workflow), a deletion variant could not be called with statistical confidence. It is likely that at least a portion of this sensitivity loss, compared with 1000GP calls, at shorter deletion lengths arises from deletion calls failing our QC and filtering protocol, as well as deletion calls originating from a different deletion algorithm other than GenomeSTRiP used by the 1000GP in the compilation of their dataset.

Flat quantile-quantile plot

Since the ADNI WGS dataset is likely underpowered from the perspective of identifying common variants with large effect size on the phenotype of interest, consistent with prior association analyses from AD-GWAS (Cuccaro et al. 2016; Han et al. 2017), (our dataset may be useful instead for targeted analyses, see S1.5 Possible Applications), we would expect that a case-control analysis of common variant loci would result in only a few loci that would exhibit at most marginal significance. This case-control comparison can be represented using a quantile-quantile (Q-Q) plot. If the realized distribution of p-values from the actual case-control comparisons is similar to the null expectation distribution of p-values, then the points in the Q-Q plot will lie approximately on the diagonal line, corresponding to y = x. Using Fisher's exact test p-values of deletion counts between cases and controls at each common variant locus, Figure S4 shows the resulting Q-Q plots, for the three possible phenotype groupings (Control vs AD, Control+MCI vs AD, Control vs MCI+AD). The most significant single locus occurs in the Control

vs MCI+AD phenotype grouping comparison, with an uncorrected p-value of 0.00031427. However, when applying multiple-test correction to the result, the corrected p-value is 0.43, in line with expectations from an underpowered dataset. The flatness of the Q-Q plots ('well-behaved'), compared to null expectation, indicates high-quality genotyping across subsets of the ADNI deletion callset. This enables confidence in the accuracy of the genotyping in the callset across the population as a whole, which is useful in other research contexts, since the vast majority of deletions in the callset would not be phenotype-specific.

Nearly all identified deletion variants in the ADNI dataset would be unrelated to phenotype labels; however, if any deletions associated with phenotype are included in the dataset, that could mean the results in our downstream analyses of purifying selection are slightly conservative, since a population cohort with a few disease-associated regulatory deletions would actually slightly bias our depletion and allele frequency spectrum shift results in a direction against our conclusion. Therefore, any significant result in our analysis that remains with the inclusion of a few phenotypically-associated variants (if known) would likely be slightly more significant had these deletions been removed from our analysis. We did not, however, find any statistically significant phenotypically-associated deletion variants in ADNI.

S1.5 Possible Applications

LoF analysis in combination with SNVs and indels

Detecting loss of function (LoF) in the genome from genetic mutations can be difficult, especially in noncoding regions; however, indels (defined as <=50bp) and deletions allow inference of loss of regular biological function (which may sometimes technically be gain of immediate function if the deleted region encodes a suppressor, etc). Joint analysis of indels and deletions may uncover regions of the genome where an individual or group of individuals is homozygous for loss of function at a particular locus, i.e. heterozygous for a deletion indel on one chromosome

and heterozygous for a deletion on the other chromosome. Additional analysis involving SNVs in combination with indels and/or deletions may also uncover interesting biological activity at loci where these 'complex' heterozygotes occur, due to joint interactions from variants on each parental chromosome.

Deletion variant network burden testing

While many cohort-based datasets fail to uncover single variants of large effect, there is tremendous interest in the medical genetics community to understand how the collective burden of a group of variants may be contributing together across a population to influence susceptibility toward the phenotype of interest, or to affect the severity (penetrance) of the trait (Khera et al. 2018). Deletions present in this dataset are on average very rare (75% of all deletions in the callset have an allele frequency of tripleton or lower). These rare variant calls may provide insight into the genetic etiology of Alzheimer's Disease, when taken together as hypothesis-driven sets of related biological units. For example, a set of deletions overlapping genes highly expressed in relevant tissues and cell types, or a set of noncoding deletions overlapping important regulators (such as enhancers or DNase I hypersensitive sites) in relevant tissues and cell types, or other similarly formed sets based upon biological intuition, may provide insight into the contribution of rare variants towards onset of Alzheimer's Disease. This network burden analysis can be especially useful in noncoding regions where the presence of a deletion can be seen as a heterozygous loss of function, and so a set of deletions could be assembled to assess the collective burden imposed by variants overlapping regulatory loci of interest.

Population deletion callset

The vast majority of the deletions identified in this callset would not be related to any particular phenotype, but instead represent a collection of segregating genomic variants identified in an

otherwise healthy population. This is especially the case given that Alzheimer's Disease is a post-reproductive phenotype (all samples were collected from individuals that were at least 50 years old) and natural selection would therefore be expected to be modest or absent on any AD-related variants in the dataset, unless pleiotropic in nature. There is great interest in studying large collections of individuals for analysis of segregating genomic variation to learn about mutational processes and natural selection, amongst other fundamental biological and evolutionary processes. Since this deletion callset was generated from only individuals of European ancestry and also has properties of a robust population dataset, the deletion calls included in this dataset (especially the deletions at the median length or longer, where sensitivity was greatest) could be of use in projects that require these properties for deletion variant analysis.

S1.6 Data Availability

Five data records are deposited to the Laboratory of Neuro Imaging (LONI) Image and Data Archive (IDA) hosted at the University of Southern California, Los Angeles, CA, USA: a) the original GenomeSTRiP deletion callset ("ADNI_deletions_full_merged.vcf.gz") before quality control and filters were applied; b) the quality controlled and filtered deletion callset ("ADNI_deletions_filtered_indgeno.bed") with individual genotypes; c) the quality controlled and filtered deletion callset ("ADNI_deletions_filtered_summarypheno.bed") with summary counts for each phenotype; d) the quality controlled and filtered deletion callset ("ADNI_deletions_filtered_wholegenome-withAF.bed") with summary population allele frequency; e) the final noncoding deletion callset ("ADNI_deletions_extrafilters_noncoding-withAF.bed") with summary population allele frequency used for most analyses in the main paper. Additionally, an ADNI sample info file is deposited ("ADNI_IDinfo.txt"). The original WGS BAM files are available via hard drives from the LONI IDA. The re-aligned WGS sequence data using GATK best practices were processed at the Broad Institute on live disk storage, but due to

the size of the data, were subsequently moved to 'cold' storage offsite. These re-aligned data may no longer be available given the cost of data maintenance of ~150 terabytes. The SNV/indel callset was previously posted to the LONI IDA. All files obtained from the LONI IDA require that investigators download, review, sign, and submit the ADNI WGS Data Use Agreement and be a registered user of ADNI data. More information on obtaining ADNI data access can be found at: http://adni.loni.usc.edu/data-samples/access-data/ .

S2 1000 Genomes Project Phase 3 (1000GP)

S2.1 Consortium Dataset Filtering

We additionally assembled deletion data from the 1000 Genomes Project Consortium Phase 3 callset (1000GP) of breakpoint-resolved deletions which were genotyped in 2,504 individuals from 26 modern human populations (Sudmant et al. 2015a). The 1000GP SV callset was downloaded from the FTP site hosted by EBI: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ . This dataset was derived from running multiple structural variant algorithms on low-coverage (~7x average coverage) whole genome sequencing data of 2,504 individuals from 26 populations. The VCF file (located in subdirectory '/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz' ['original file']) contained the deletion calls with allele frequency (AF) but did not include any Y chromosome calls. Deletion calls passing metrics for fine-resolution of breakpoints (contained in TABLE_3-Breakpoints.xlsx) were additionally downloaded as a text file (from subdirectory 'phase3/integrated_sv_map/supporting/breakpoints/1000GP_phase3_all_bkpts.v5.txt.gz' ['breakpoints file']). These data were generated from a variety of published deletion callers and underwent quality control for accuracy in variant genotyping and fine-resolution of coordinate breakpoints. Since we want to analyze only variant calls which are deletions (i.e. loss) of genomic information relative to the human reference genome, and not analyze the absence of an insertion of genomic information (potentially human reference sequence insertions in the individuals catalogued), only variant calls likely to be true deletions from BreakSeq (Lam et al. 2010) prediction were retained (only 'NAHR', 'NAHR_EXT', or 'NH' MUTMECH designations as specified in the 'breakpoints file'). To arrive at this set with corresponding allele frequency, we extracted AF from the 'original file' containing all the deletion calls and matched deletion name identifiers from this file with that in the 'breakpoints file'. For deletions with the same/synonymous name identifier that had more than one set of breakpoints identified (about

0.2% of deletions), we used the average breakpoint-resolved start coordinate and average breakpoint-resolved end coordinate. The total collection of autosomal deletions with AF gathered in these processes numbered 22,684.

The allele frequency used in the 1000GP dataset was the global AF ('AF=' in the 'original file'). Most deletions are population specific because they are rare (singleton or doubleton, etc.), but common deletions are subject to genetic drift and by taking a global AF, this smoothed-out these population-specific demographic histories for those variants. We were interested in studying purifying selection to preserve regulatory sites broadly during human evolution, and not examining purifying selection on regulatory sites for potential population-specific deletions. Also, besides losing statistical power by breaking the deletion set into population-specific sets, since each demographic population would have a different genetic distance to the human reference sequence (given the underlying construction of the reference sequence being biased toward European variation [Sudmant et al. 2015b]), differences seen in the underlying population-specific deletion sets would be potentially artifactual and likely not representative of biological differences. We additionally chose to examine the 1000GP dataset as a single global population, since with population-specific deletion datasets (and resulting population-specific simulations), we would be re-examining common deletions multiple times (since they would often be shared between most/all populations), which would have introduced statistical confounding in the interpretation of the results.

With the shorter deletion length distribution found in the 1000GP dataset, mobile element insertion could have contaminated our analyses. However, because of our quality and technical filters employed, nearly all predicted mobile element insertion (MEI) variants (identified in the VCF file as 'SVTYPE=DEL_*\t') were removed. Remaining predicted MEI-derived deletions were not removed, as manually removing these variants may have lead to unknown biases in

downstream analysis of purifying selection when those same underlying genomic coordinates would otherwise be allowed in regulatory assays or computational simulations (i.e. removing these deletions manually may have induced artificial depletion in these loci). We were interested in the missing sequence in an individual of any uniquely-mappable sequence in the noncoding human genome, given our other filters. Since some MEIs (which become called as deletions in some individuals) may have epigenomic regulatory annotation, those 'deletion' variants should be left in the dataset because the absence of that sequence in one or more individuals may have functional consequence in those individuals. We don't judge *a priori* the importance of that noncoding sequence space if it is uniquely alignable and is already available to regulatory experimental assays (see Supplemental Note S4.1). For example, if sequence at a locus is marked as having DHS activity, then if a MEI 'deletion' occurs at that locus, that means some humans don't have that open chromatin, which may be functionally important. Ideally, perfectly identified human ancestral sequence would be able to identify true losses of genetic information (derived deletions after last common ancestor) from contemporary deletion data, however reconstructing the human ancestral genome as well as comparison to primate genomes is difficult due to differing reference sequence qualities (Kronenberg et al. 2018).

S2.2 Filters Applied to Both 1000GP and ADNI datasets

Additional filters were applied to ensure later careful examination of the effects of selection on deletions overlapping regulatory elements. For both deletion datasets, we restricted our analyses to noncoding deletions by removing any deletion that overlapped any exon or UTR by one base-pair or more, as exonic deletions have been previously shown to be under strong purifying selection because of their protein-altering effects (Conrad et al. 2010). Genomic coordinates used to identify exonic and genic sites were downloaded from Ensembl Biomart: http://grch37.ensembl.org/biomart/martview/ with dataset 'Human genes (GRCh37.p13)'. We also examined only deletions occurring on autosomes because sex-chromosome functional

elements may involve complex sex-biased regulation (Khramtsova et al. 2019) which might be subject to unique selective properties. This filter removed only 1000GP X chromosome deletions (there were no X chromosome deletions called in ADNI and no Y chromosome deletions called in either 1000GP or ADNI). To mitigate non-uniform (i.e. biased) deletion callability in the noncoding genome which might distort the allele frequency spectrum of the remaining set of deletions, we additionally excluded (using tracks downloaded from the UCSC Genome Table Browser: https://genome.ucsc.edu/cgi-bin/hgTables) deletions overlapping any regions of non-unique mappability ('wgEncodeCrgMapabilityAlign100mer') (see S4.1 Deletion Callability and Need for Unique Coordinates), segmental duplications ('genomicSuperDups'), and centromeres or reference assembly gaps ('gap'). We additionally removed ADNI deletions overlapping regions of B-cell instability using the same regions already excluded by the 1000GP consortium in their released deletion callset (Sudmant et al. 2015a). BEDTools software (version 2.26.0) (Quinlan et al. 2010) was used to remove deletion variants that overlapped loci excluded from analysis. Additionally, deletions longer than 25 kb were removed because downstream analysis with simulated deletions/mock datasets require independent coordinate assessment and deletions longer than this may cause spurious 're-mutation' in the mock datasets (because of simulation procedure rules used, see S4.2 Deletion Simulation Procedure) not representative of deletions in the real datasets. Because of the stringent filters already employed, principally the low mappability filter, only three 1000GP deletions and one ADNI deletion were removed because of this length cutoff. For both deletion datasets, allele frequency was kept as the raw deletion AF with respect to the reference, not minor AF, because we want to analyze the loss of genetic information in the form of deletions, not just the minor allele which might represent the non-deleted state for extremely common deletions, potentially resulting in biases in AF comparison analyses. Deletion sequences are found with respect to the human GRCh37/hg19 reference genome. The resulting deletion datasets remaining after the filtering procedures were applied included 12,013 1000GP deletions and 3,306 ADNI deletions. As

expected, the bulk (>80%) of deletions in our datasets remaining after filtering were rare (below 1% AF). Specific characteristics of the deletion datasets are shown in Table S11.

S3 Regulatory Feature Annotations

S3.1 NIH Roadmap Epigenomics Consortium (REC)

To analyze genomic deletions within regulatory regions, we used regulatory annotation data from the NIH Roadmap Epigenomics Consortium (REC) (Roadmap Epigenomics Consortium et al. 2015) for definition of regulatory breakpoints as well as uniform processing across multiple tissues and cell types. In particular, we used two callsets of chromatin accessibility data (DNase I hypersensitivity 'DHS') and four callsets of histone modification data (H3K4me1 'enhancer', H3K36me3 'transcribed', H3K27me3 'polycomb-repressed', and H3K9me3 'heterochromatin'). DHS annotations are typically associated with sites of open chromatin allowing accessibility for regulator binding and histone annotations are typically associated with sites of specific regulatory activity, as noted. Regulatory element data were downloaded from the supplementary website for the 2015 NIH Roadmap Epigenomics Consortium ('REC') paper (Roadmap Epigenomics Consortium et al. 2015): https://egg2.wustl.edu/roadmap/web_portal/index.html . All data we used were derived from REC 'consolidated epigenomes', which were uniformly reprocessed and standardized epigenomes, designed to eliminate differences between research centers and changes to sequencing technology that occurred over the course of the REC project. Two types of DNase I callsets with bp resolution were used to check for consistency in the downstream analyses: callsets (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/) made from the Hotspot algorithm (John et al. 2011) which uses a binomial distribution model and callsets (https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) made from the MACS algorithm (Zhang et al. 2008) which uses a Poisson distribution model. Both callsets that were generated with an expected false discovery rate (FDR) of 1% were chosen. For histone modification data, callsets

(https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/) with base-pair resolution made from MACS with a 1% expected FDR were chosen.

To ensure reliable regulatory data in the analysis of overlapping genomic deletions, we selected only tissues and cell types for analysis that were primary in nature, i.e. not embryonic stem (ES) cells, induced pluripotent stem (iPS) cells, or ES-derived cells, as non-primary tissues or cell types undergo significant passaging effects from which structural variant artifacts could accumulate in the cells (Funk et al. 2012; Liu et al. 2014). We selected all primary tissues and cell types which were available across both types of DNase I callsets and all four histone modification callsets. We excluded two cell lines, E017 and E124, that appeared to be highly similar (same tissue/cell type and same relative donor age) to E088 and E029, which were retained. Of the 25 tissues and cell types selected, 15 were derived from fetal samples, and 10 were derived from adult (three years or older) samples. Table S12 summarizes the selected tissues and cell types. These data assume that regulatory elements are shared across all human populations, as the assays performed included samples from various ancestries.

S3.2 Chromatin Loop Anchors

We additionally used regulatory data that demarcate chromatin loop anchors (Rao et al. 2014), which enclose local genomic regions associated with physically interacting regulatory activity. Chromatin loop anchor data were downloaded from GEO accession GSE63525 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525), data generated using in situ Hi-C from Rao & Huntley et al. 2014. Genomic coordinates corresponding to loop boundaries were extracted from the files (with names ending in '*_HiCCUPS_looplist_with_motifs.txt.gz'). To have consistency with the DNase I hypersensitivity and histone modification callsets data, only human tissues and cell types with relatively normal karyotypes (i.e. non-cancerous) were

selected for analysis. The five callsets selected were derived from biosamples designated as GM12878, NHEK, IMR90, HUVEC, and HMEC.

S3.3 ENCODE Uniform CTCF TF Peaks

To overlay CTCF transcription factor binding sites within chromatin loop anchors, we used CTCF locations from the same tissues and cell types examined for loops. CTCF ChiP-seq callset data, processed in a uniform pipeline from the ENCODE project March 2012 data freeze (The ENCODE Project Consortium 2012), were downloaded from the UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform ('Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis' track). Each tissue or cell type had more than one callset, each originating from a different analysis center. For each tissue or cell type separately, all CTCF calls across files from the various analysis centers were merged using the BEDTools (Quinlan et al. 2010) 'merge' option. Then, only CTCF sites from a given tissue or cell type which overlapped chromatin loop anchors in the same tissue or cell type were extracted and retained. In this way, misspecification in either loop anchors or CTCF callsets are then not contaminated between analyses. This is especially important in downstream pleiotropic analyses where concordance between a particular CTCF and a particular loop is assumed to be co-occurring in the same tissue(s) or cell type(s). This criteria limits counting transient CTCF binding sites across the tissues and cell types.

S4 Deletion Simulation Schema

S4.1 Deletion Callability and Need for Unique Coordinates

Our analysis depends on accurate deletion genotyping and accurate regulatory locus annotation. In regions of the genome where sequence is non-unique, (i.e. for any 100 bp stretch, that sequence appears in more than one location in the reference human genome), sequence read coverage may be missing, averaged across all sites in the genome, or over-represented depending on alignment algorithm parameters used. This can present problems for deletion calling as well as regulatory element peak calling. In addition, if these sequences are, on-average, less functional (due to repeat sequences), then purifying selection may be operating in a relaxed manner in these regions, biasing or confounding analyses in identifying a shift in the deletion allele frequency spectrum. Additionally, if simulations are performed where deletions are randomly placed along the genome, if non-unique regions are available as potential random placement locations for deletions, a functional-to-less functional 'spreading' will occur. See Figure S5 for an example of 1000GP noncoding deletion spreading in simulations from real (more-unique) coordinates to mock (less-unique) coordinates (using RepeatMasker annotations developed by Smit, AFA, Hubley, R & Green, P., *RepeatMasker Open-4.0*, 2013-2015 [http://www.repeatmasker.org]). This is because in the real deletion callset, non-unique genomic regions presented less-confident deletion evidence, on average, than unique regions of the genome, and are therefore likely underrepresented in the deletion callset. Therefore, more deletions calls were made in unique regions compared to non-unique (all other things being equal) and so in a simulation framework where deletions are randomly placed throughout the genome, there will be a 'migration' from unique-to-non-unique, at a modest noticeable extent. This effect is very hard to control for in matched simulations, given the covariance of non-unique regions with GC content, recombination rate, and other genomic features. This migration of deletion calls can bias analyses that depend on overlap with regulatory annotations

which are predominantly located in more unique regions of the genome given the same issues encountered for ChIP-seq assay sequence read mapping. Therefore, to ensure robust genomic coordinates for our analyses, we restricted to only unique sites in the genome (i.e. for any 100bp stretch, that sequence appears only in that one location in the human genome).

S4.2 Deletion Simulation Procedure

Deletion length can be a confounder in analyses of deletions overlapping regulatory features and associated cellular pleiotropy measures, because longer deletions have a greater chance than shorter deletions to randomly hit sparse genomic annotations such as regulatory elements. Most overlap statistics that could be computed (such as binary association, bp affected, or PlyRS measures [see S5 Pleiotropy Ratio Score (PlyRS) Calculated Measures]) would then have an inherent length bias, making analyses confounded. Additionally, longer deletions are typically easier for deletion callers to identify because missing sequence coverage over a longer length appears more statistically significant. Therefore, to ensure reliable interpretation of deletion overlap within regulatory regions in a length-controlled manner, we developed a deletion simulation strategy. For each real deletion, we placed 1,000 mock deletion copies of the same length randomly along the genome using only allowable genomic coordinates, keeping the mock copies on the same chromosome but not necessarily the same chromosome arm/locus band (which if forced might introduce non-independence of the simulations) and in the same genomic compartment space (intronic or intergenic), to approximate local context-dependent effects. We also tracked the associated AF label in downstream analysis. Using this simulation strategy, we thereby created 1,000 mock deletion datasets each with a random distribution of deletion locations. To later compare significance of overlap associations, we created an additional 1,000 mock deletion copies for each deletion thereby creating an additional 1,000 mock deletion datasets. Using this length-matched simulation framework, we are able to appropriately analyze both horizontal and vertical axes on which purifying selection

may be operating on deletions. We randomized deletions, as opposed to regulatory elements, utilizing the fact that the deletions are mutations (relative to the human reference genome) while assuming that regulatory elements are essentially fixed (i.e. consistent) components in the modern human genome across populations.

When subsetting the deletions overlapping chromatin loop anchors into those that overlap loop annotation but not coincident CTCF annotation, we used a separate allowable genome space including the full genomic coordinates allowed excluding coordinates with CTCF sites within chromatin loop anchor annotations. We do not make a distinction between a loop annotation for which CTCF sites are excluded within its coordinates and a loop annotation for which no CTCF site was originally observed. Using this separate allowable genome space for simulations ensured that the depletion values we observe for $loops_{noCTCF}$ are reflective of the exclusionary criteria we used when conditioning on the real set of deletions which also exclude coordinates with CTCF sites. For all analysis involving CTCF, we analyzed only CTCF sites that are identified in the same tissue or cell type as the coincident chromatin loop anchor annotation.

S5 Pleiotropy Ratio Score (PlyRS) Calculated Measures

We derived a set of PlyRS summary measures for use in downstream analyses to examine how noncoding deletions can potentially remove regulatory function at a genomic locus.

$PlyRS_{sum}$: corresponds to the total cellular pleiotropy (for a specific regulatory feature) of a deletion, encompassing both the horizontal and vertical axes along which purifying selection may be operating on the deletion. It is calculated by summing together all PlyRS values found along the length of the deletion. This is the same result as multiplying the number of overlapped regulatory base-pairs by the average PlyRS value found along the deletion. This value correlates strongly with the total number of bp (and also with the number of regulatory bp) of the deletion, since the horizontal axis often forms the bulk of the sum because sites of activity specific to a particular tissue or cell type make up roughly a quarter of all regulatory sites in our callsets used. $PlyRS_{sum}$ is equal to the sum of two other component sums, $PlyRS_{sum\text{-}mono}$ and $PlyRS_{sum\text{-}pleio}$, described next.

$PlyRS_{sum\text{-}mono}$: includes the sum of PlyRS values of each deleted bp for which that bp is only associated with regulatory activity specific to a particular tissue or cell type. The count at this bp is not 1, however, because the count is adjusted by the correlation between all the tissues and cell types being analyzed.

$PlyRS_{sum\text{-}pleio}$: includes the sum of PlyRS values of each deleted bp for which that bp is associated with cellularly pleiotropic regulatory activity (i.e. activity in more than one tissue or cell type).

$PlyRS_{max}$: corresponds to the maximal PlyRS value found at any bp from examining all bp along the length of a deletion. This value has a maximum of 1, representing 100% cellular pleiotropy across the tissues and cell types analyzed. This measure is more stable than $PlyRS_{sum}$ with regulatory annotations that have less precision on boundaries (such as chromatin loop anchors).

S6 Depletion Significance Calculation

We have developed the following quantitative procedure to detect reduction in deletion variation. For each real (i.e. observed) deletion, we compare, one at a time, a measure of interest (e.g. $PlyRS_{max}$ [see S5 Pleiotropy Ratio Score (PlyRS) Calculated Measures]) of the real deletion's value, to each mock deletion copy's (i.e. expected) value of the measure of interest. Each time that the real deletion has a lower or equal value than a mock copy, which indicates overlap in the real deletion compared to simulation had the same amount or less, we assign that instance to a counter and perform over all 1,000 mock deletion copies. We include the 'equal value' so that information from the simulation is utilized, otherwise all deletions with no real overlap will always receive an empirical p-value of 0.001 ([no counts + 1] / 1000), but including the 'equal value' means that the number of simulation no-overlaps will be included in the empirical p-value. This information is especially useful when considering deletions of different lengths with no real overlap. At the end of this process, we calculate the one-sided empirical p-value of this analysis, taken as the counter of instances plus 1 since our real result can be considered an instance of observation (unless counter=1000 at which point we ignore our observation) divided by 1,000 tests. Therefore, for each deletion, there is an empirical p-value of that deletion's measure of interest versus 1,000 mock copies. We additionally perform this same process for each of the 1,000 mock deletion sets against 1,000 additional matched mock deletion sets. This means that for each original mock deletion set, there is an empirical p-value of every deletion's measure of interest versus that of the 1,000 additional mock copies. To compute significance of the depletion results, we calculate the sum of the natural log of empirical p-value for every deletion in the real dataset, and additionally calculate this sum for every original mock deletion dataset. Using the distribution of this sum for the mock deletion datasets, we perform a *t*-test of where the real deletion dataset sum resides amidst the mock distribution. We can use a *t*-test because the mock distribution from the sum of log(empircal p-value) is approximately normal (see Figure S6 for an example). From the *t*-test, we can also derive the effect size of the result (calculated

by Cohen's *d*, in units of standard deviation), standardizing the interpretation when comparing results across regulatory features, each of which may have been composed of a different sample size of overlapping deletions. The effect size, *d*, can be calculated using the formula *d* = (mean of real observed log(empirical p-value) minus mean of mock distribution of log(empirical p-value) divided by the mock distribution standard deviation. The confidence interval around the effect size is then calculated by finding the 95% confidence interval from the upper and lower noncentrality parameters around two noncentral *t* distributions. We solved for the noncentrality parameter upper and lower bounds using the SciPy Python package (scipy.special.nctdtrinc). A tutorial consulted describing these concepts is presented by David C. Howell, Univ. of Vermont: https://www.uvm.edu/~statdhtx/StatPages/ConfIntEffectSize/Confidence%20limit%20on%20effect%20size.html .

Depletion significance calculation steps in outline format:

- Create 1,000 mock deletion copies for each real deletion
- For each real deletion, compare a measure of interest (e.g. $PlyRS_{max}$) to each deletion copy
- Each time that the real deletion has a value <= a mock copy value (indicates depletion) assign that instance to a counter
- Calculate the one-sided empirical p-value of this analysis [(counter+1)/1,000]
- Perform over all 1,000 mock copies
- Create 1,000 mock deletion copies again
- Repeat analysis of steps above for each original mock copy compared to its own 1,000 new mock copies
- Calculate the sum of the log(empirical p-value) of all deletions in the real dataset, and the sum for all deletions in all original 1,000 mock datasets
- Use the distribution of this sum for the mock datasets to perform a *t*-test of where the real deletion dataset sum resides
- Convert this into effect size in units of standard deviation (Cohen's *d*) with confidence intervals

We measure depletion relative to the deletion, not relative to a percentage of a regulatory element, since the exact boundary of an element can be uncertain across multiple tissues and cell types. Also, longer deletions in our datasets have the ability to potentially overlap multiple elements and we want to capture that information in our simulation experiments.

S7 Logistic Regression

S7.1 Depletion Magnitude Calculation

To get meaningful odds ratio interpretation of magnitude of depletion in logistic regression tests, we use the ratio of proportional difference between real deletions and length-matched simulations (sim) (calculated as: [$PlyRS_{measure}$ - average sim $PlyRS_{measure}$] / average sim $PlyRS_{measure}$) for chromatin accessibility and histone modification annotations. This means that a real deletion for which no regulatory overlap occurs will be considered to be 100% depleted (-1.0) of PlyRS measure in relation to the deletion simulation average of the PlyRS measure. We use the raw difference between real deletions and simulations ($PlyRS_{measure}$ - average sim $PlyRS_{measure}$) for chromatin loop anchor and CTCF annotations. We don't use raw difference for chromatin accessibility and histone modification features because otherwise we would be length-biasing the values since deletions can potentially overlap more than one of these regulatory loci; however because of the length dynamics of our deletion callsets, deletions can only overlap at most one chromatin loop anchor boundary. The odds ratio in these tests means that for a one unit change in the difference (either proportional or raw) compared to simulations, with all other covariates held steady, there is an odds increase or decrease (with 1 as baseline) of the deletion set depletion magnitude being positively associated with allele frequency (i.e. deletions more depleted from simulation average are more likely to be common). Confidence intervals on the odds ratio are calculated as profile likelihood based confidence intervals, as we are not able to depend on the assumption of normality for the estimator.

S7.2 Genomic Covariates

For a regulatory element feature, to test whether PlyRS measure depletion magnitude depends on deletion allele frequency, we use logistic regression on rare (AF<=1%) or common (AF>1%) allele frequency in the presence of genomic covariates. Since the vast majority of deletions in

our datasets are rare (~52% of 1000GP deletions and ~76% of ADNI deletions are tripleton or lower in allele frequency), multivariate regression performed directly on allele frequency would create a response variable that does not 'behave well' in terms of its resulting distribution. Therefore, conclusions reached from p-value interpretation would be unreliable. Using logistic regression, we collapse all rare deletions (AF<=1%) into a single class, thereby creating a binary response variable of rare/common, from which multivariate regression can be performed with confidence in the p-value interpretation. We use <=1% as an AF cutoff for rare deletions as it minimizes technical artifacts that might be present from just examining singletons alone (where calling artifacts might predominantly reside in the allele frequency spectrum).

For genomic covariates of each deletion, we choose average regional measures 50 kb upstream of the start coordinate and downstream of the end coordinate, because correlation between 50 kb sides values and within-deletion values is extremely significant ($p<10^{-16}$). To calculate SNV nucleotide diversity (pi), we used VCFtools (with '--remove-indels' option on the 1000GP data, downloaded from server: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) examining only sites in the individuals from which the deletion genotypes were derived (2,504 individuals in 1000GP and 752 individuals in ADNI). Recombination rate was taken from the HapMap-derived (downloaded from: http://www.well.ox.ac.uk/~anjali/AAmap/) 'Combined_LD' column for 1000GP and 'CEU_LD' column for ADNI. We chose the HapMap data since it covered more of the genome than the other measures. For distance to the nearest transcription start site, we used coordinate information downloaded from Ensembl Biomart: http://grch37.ensembl.org/biomart/martview/ with dataset: 'Human genes (GRCh37.p13)'. The Ensembl data was used instead of UCSC data because the Ensembl data appeared to contain more transcripts. GC content proportion is calculated directly from the GRCh37/hg19 version of the human reference genome.

S8 PlyRS Main Scripts

[make_config_database.py]

```
# Gather all annotation configurations
import numpy as np
configurations = {}

f_in1 = open("filepath.../chr1_vectors.txt")
f_in2 = open("filepath.../chr2_vectors.txt")
f_in3 = open("filepath.../chr3_vectors.txt")
f_in4 = open("filepath.../chr4_vectors.txt")
f_in5 = open("filepath.../chr5_vectors.txt")
f_in6 = open("filepath.../chr6_vectors.txt")
f_in7 = open("filepath.../chr7_vectors.txt")
f_in8 = open("filepath.../chr8_vectors.txt")
f_in9 = open("filepath.../chr9_vectors.txt")
f_in10 = open("filepath.../chr10_vectors.txt")
f_in11 = open("filepath.../chr11_vectors.txt")
f_in12 = open("filepath.../chr12_vectors.txt")
f_in13 = open("filepath.../chr13_vectors.txt")
f_in14 = open("filepath.../chr14_vectors.txt")
f_in15 = open("filepath.../chr15_vectors.txt")
f_in16 = open("filepath.../chr16_vectors.txt")
f_in17 = open("filepath.../chr17_vectors.txt")
f_in18 = open("filepath.../chr18_vectors.txt")
f_in19 = open("filepath.../chr19_vectors.txt")
f_in20 = open("filepath.../chr20_vectors.txt")
f_in21 = open("filepath.../step1example-chr21_DHS_hotspot.vectors.head10000.txt")
f_in22 = open("filepath.../chr22_vectors.txt")

f_in_chroms =
[f_in1,f_in2,f_in3,f_in4,f_in5,f_in6,f_in7,f_in8,f_in9,f_in10,f_in11,f_in12,f_in13,f_in14,f_in15,f_in16,f_in17,f_in18,f_in19,f_in20,f_in21,f_in22]

for i in f_in_chroms:
   for j in i:
      line = j.rstrip()
      words = j.split("\t")
      coordinate = words[0]
      configuration = words[1].rstrip()
      if configuration in configurations:
         configurations[configuration] += 1
      else:
         configurations[configuration] = 1

print(len(configurations))


# Find total database frequency of an annotation
total_annotations = 0
total_positions = 0
for i,j in configurations.items():
   key = i
   value = j
   ann_count_config = key.count("1")
   total_ann_count = ann_count_config*value
   total_annotations += total_ann_count
   total_pos_config = len(key)*value
   total_positions += total_pos_config

database_freq = total_annotations/total_positions
print(database_freq)
```

```
totalsites = 0
for key,value in configurations.items():
   totalsites += value
f_out = open("step2example-DHS_hotspot-configurations_database.txt","w")
f_out.write(str("database-frequency=")+str(database_freq)+"\n")
f_out.write(str("totalsites=")+str(totalsites)+"\n")
for k,v in configurations.items():
   config = k
   number_config = v
   f_out.write(str(config)+"\t"+str(number_config)+"\n")
f_out.close()
```

[calc_n-eff.py]

```
import sys

config_range_highest = sys.argv[1]
bottom_range = int(config_range_highest) - 10000 #batch in chunks of 10000
top_range = int(config_range_highest)

filename = "step2example-DHS_hotspot-configurations_database.txt"
f_in = open(filename)

configurations = {}

for i in f_in:
   line = i.rstrip()
   if line.startswith("database"):
      words = line.split("=")
      database_freq = float(words[1])
      continue
   if line.startswith("totalsites"):
      words = line.split("=")
      totalsites = int(words[1])
      continue
   words = line.split("\t")
   config = words[0]
   occurrences = int(words[1])
   configurations[config] = occurrences
f_in.close()

specific_configurations = {}

f_in2 = open(filename)

counter = 0
for i in f_in2:
   line = i.rstrip()
   if line.startswith("database"):
      continue
   if line.startswith("totalsites"):
      continue
   counter += 1
   if (counter > bottom_range and counter <= top_range): #break up configurations scored into smaller chunks
      words = line.split("\t")
      config = words[0]
      occurrences = int(words[1])
      specific_configurations[config] = occurrences
   else: continue
```

```
# Define shared tissues value for each configuration

def vect_prepare(input_vector):
    count = -1
    tissue_shared = set()
    tissue_notshared = set()
    for i in range(0,len(input_vector),1):
        count += 1
        if input_vector[i] == "1":
            tissue_shared.add(count)
        if input_vector[i] == "0":
            tissue_notshared.add(count)
    results = [tissue_shared,tissue_notshared] #this is a list of two sets
    return results

configurations_shared = {}

configurations_notshared = {}

for key,value in configurations.items():
    vector = str(key)
    vect_shared = vect_prepare(vector)[0] #this is a set of reg elem-based annotations from the vector
    vect_notshared = vect_prepare(vector)[1] #this is a set of non reg elem-based annotations from the vector
    configurations_shared[vector] = vect_shared
    configurations_notshared[vector] = vect_notshared

share_sites_dict = {}

notshare_sites_dict = {}

def share_sites_count(vector):
    shared_sites = 0
    shared_site_tissues = configurations_shared[vector] #this is a set
    for key,config_set in configurations_shared.items():
        if shared_site_tissues <= config_set: #sites that minimally share all in the given vector
            shared_sites += configurations[key]
        else: continue
    share_sites_dict[vector] = shared_sites

def notshare_sites_count(vector):
    notshared_sites = 0
    notshared_site_tissues = configurations_notshared[vector] #this is a set
    for key,config_set in configurations_notshared.items():
        if notshared_site_tissues <= config_set: #sites that minimally don't share all in the given vector
            notshared_sites += configurations[key]
        else: continue
    notshare_sites_dict[vector] = notshared_sites

for key in specific_configurations.keys(): #only examine the configurations in the reduced chunk specified earlier
    share_sites_count(key)
    notshare_sites_count(key)

n_eff_share_dict = {}

n_eff_notshare_dict = {}

def n_eff_share_calc():
    for key,value in share_sites_dict.items():
        numerator = value-1 #subtract 1 to remove the observation site
        denominator = totalsites-1 #subtract 1 to remove the observation site
        fraction = numerator/denominator
        n_eff = np.log10(fraction)/np.log10(database_freq) #using mathematical relation to solve for n_eff
        n_eff_share_dict[key] = n_eff

def n_eff_notshare_calc():
```

```
    for key,value in notshare_sites_dict.items():
        numerator = value-1 #subtract 1 to remove the observation site
        denominator = totalsites-1 #subtract 1 to remove the observation site
        fraction = numerator/denominator
        n_eff = np.log10(fraction)/np.log10(1-database_freq) #using mathematical relation to solve for n_eff
        n_eff_notshare_dict[key] = n_eff

import numpy as np

n_eff_share_calc()
n_eff_notshare_calc()

f_out_share_notshare_filename = "neff_share_notshare_DHS_hotspot-group"+str(top_range)+".txt"
f_out_share_notshare = open(f_out_share_notshare_filename,"w")

for k in share_sites_dict.keys():
#     if (k in n_eff_share_dict) and (k in n_eff_notshare_dict):
    n_share = n_eff_share_dict[k]
    n_notshare = n_eff_notshare_dict[k]
    n_ratio = n_share/(n_share+n_notshare) #divide by total, so ratio is always between 0 and 1
    f_out_share_notshare.write(str(k)+"\t"+str(n_share)+"\t"+str(n_notshare)+"\t"+str(n_ratio)+"\n")

f_out_share_notshare.close()
```

Supplementary References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65. doi:10.1038/nature11632

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**: 248–249. doi:10.1038/nmeth0410-248

Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. 2007. Forecasting the global burden of Alzheimer's disease. *Alzheimer's Dement*. **3**: 186–191. doi:10.1016/j.jalz.2007.04.381

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712. doi:10.1038/nature08516

Cuccaro D, De Marco EV, Cittadella R, Cavallaro S. 2016. Copy number variants in Alzheimer's disease. *J. Alzheimer's Dis*. **55**: 37–52. doi:10.3233/JAD-160469

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158. doi: 10.1093/bioinformatics/btr330

Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*. **43**: 491–501. doi:10.1038/ng.806

Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA, et al. 2012. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. *Nat. Genet*. **44**: 886–889. doi:10.1038/ng.2344

Funk WD, Labat I, Sampathkumar J, Gourraud PA, Oksenberg JR, Rosler E, Steiger D, Sheibani N, Caillier S, Stache-Crain B, et al. 2012. Evaluating the genomic and sequence integrity of human ES cell lines; comparison to normal genomes. *Stem Cell Res*. **8**: 154–164. doi:10.1016/j.scr.2011.10.001

Han Z, Huang H, Gao Y, Huang Q. 2017. Functional annotation of Alzheimer's disease associated loci revealed by GWASs. *PLoS One* **12**: e0179677. doi:10.1371/journal.pone.0179677

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet*. **43**: 269–276. doi:10.1038/ng.768

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet*. **43**: 264–268. doi:10.1038/ng.759

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet*. **50**: 1219–1224. doi:10.1038/s41588-018-0183-z

Khramtsova EA, Davis LK, Stranger BE. 2019. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet*. **20**: 173–190. doi:10.1038/s41576-018-0083-1

Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A, Lameijer EW, Moed MH, Koval V, Renkens I, et al. 2015. Characteristics of de novo structural changes in the human genome. *Genome Res*. **25**: 792–801. doi:10.1101/gr.185041.114

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360**: eaar6343. doi:10.1126/science.aar6343

Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. 2010. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol*. **28**: 47–55. doi:10.1038/nbt.1600

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997.

Liu P, Kaplan A, Yuan B, Hanna JH, Lupski JR, Reiner O. 2014. Passage Number is a Major Contributor to Genomic Structural Variations in Mouse iPSCs. *Stem Cells* **32**: 2657–2667. doi:10.1002/stem.1779

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**: 1297–1303. doi:10.1101/gr.107524.110

Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack Jr CR, Jagust WJ, Shaw LM, Toga AW, et al. 2010. Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* **74**: 201–209. doi:10.1212/WNL.0b013e3181cb3e25

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*. **38**: 904–909. doi:10.1038/ng1847

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. **81**: 559–575. doi:10.1086/519795

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. doi:10.1093/bioinformatics/btq033

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680. doi:10.1016/j.cell.2014.11.021

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–329. doi:10.1038/nature14248

Saykin AJ, Shen L, Yao X, Kim S, Nho K, Risacher SL, Ramanan VK, Foroud TM, Faber KM, Sarwar N, et al. 2015. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. *Alzheimer's Dement*. **11**: 792–814. doi:10.1016/j.jalz.2015.05.009

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81. doi:10.1038/nature15394

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761. doi:10.1126/science.aab3761

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. doi:10.1038/nature11247

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma*. **43**: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Cedarbaum J, Donohue MC, Green RC, Harvey D, Jack Jr CR, et al. 2015. Impact of the Alzheimer's Disease Neuroimaging Initiative, 2004 to 2014. *Alzheimer's Dement*. **11**: 865–884. doi:10.1016/j.jalz.2015.04.005

Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, et al. 2013. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**: 707–720. doi:10.1016/j.cell.2013.03.030

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChiP-Seq (MACS). *Genome Biol*. **9**: R137. doi:10.1186/gb-2008-9-9-r137