The American Journal of Human Genetics, Volume *111*

Supplemental information

Structural and genetic diversity in the secreted mucins *MUC5AC* and *MUC5B*

Elizabeth G. Plender, Timofey Prodanov, PingHsun Hsieh, Evangelos Nizamis, William T. Harvey, Arvis Sulovari, Katherine M. Munson, Eli J. Kaufman, Wanda K. O'Neal, Paul N. Valdmanis, Tobias Marschall, Jesse D. Bloom, and Evan E. Eichler

Note S1. Using tandem repeats finder, we discovered an 8-mer VNTR in intron 15 of *MUC5AC* with a canonical motif of TCACCCAC in all human haplotypes. Haplogroup 3 (H3) alleles harbor more copies of the repeat and a lower percent motif identity compared to haplogroup 1 and 2 alleles in humans. STREME (Sensitive, Thorough, Rapid, Enriched Motif Elicitation) analysis reveals five 24-mers that are exclusive to H3 alleles for genotyping this haplogroup (ACCATTCACTCACCCATTCACCCATTCACC, ACTCACCCACTCACCCATTCACCCATTCAC, ACTCACCCACTCACTCACTCACCTA CTCAA, CAGTGGGTGATTGAGTGGGTGAATGGGTGA, and TAAGTTGAGTGAGTGA GGTGAGTGAGTGGA). The canonical 8-mer motif is exclusive to human haplotypes, with all NHPs featuring motif lengths of 12-28 nucleotides. Total VNTR copy number is also highly variable among the NHPs, leading to differences in total sequence length for intron 15 across haplotypes (Fig. S5).

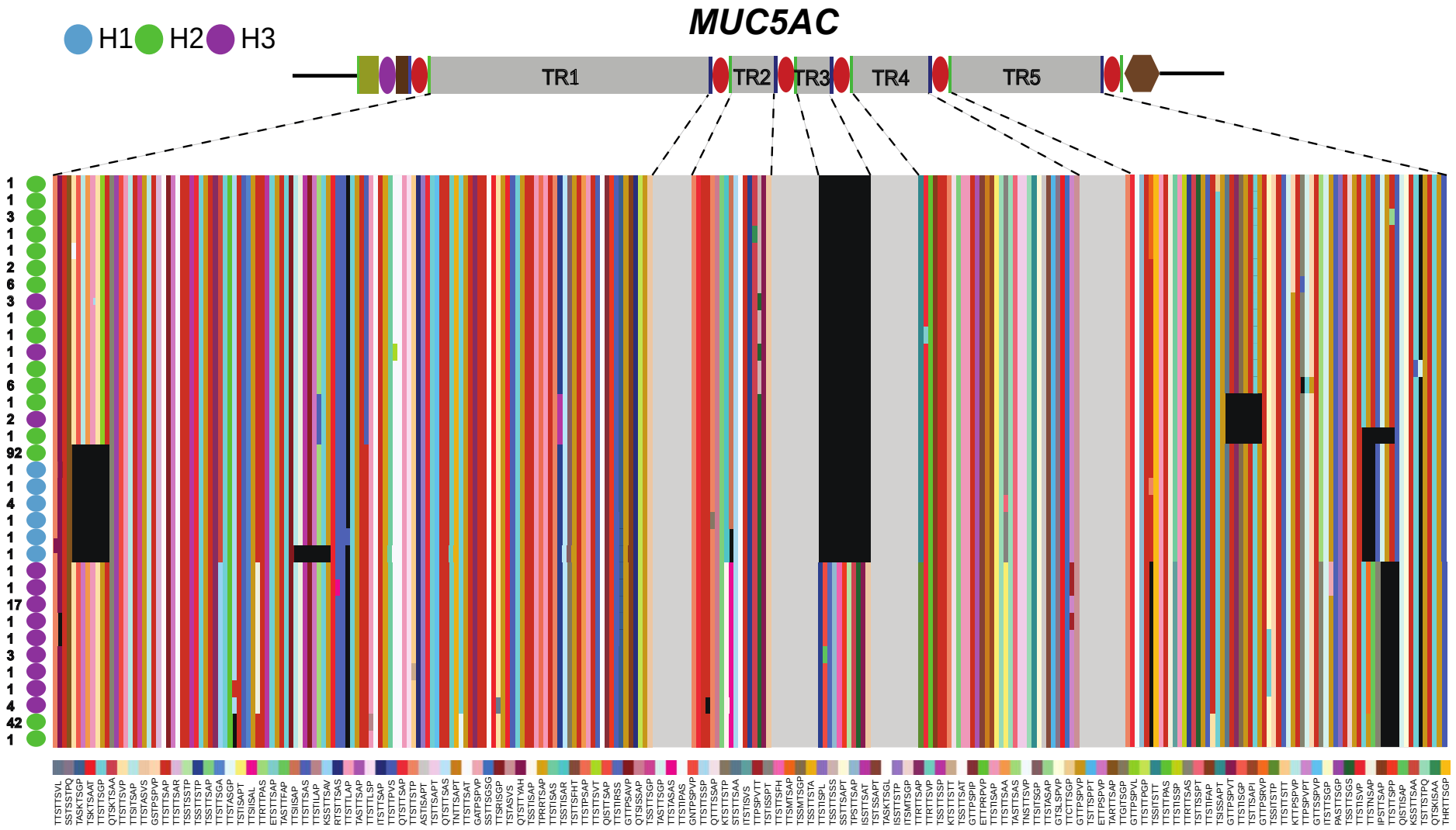
H1
H2
H3
MUC5AC
TR1
TR2
TR3
TR4
TR5
1
1
3
1
1
2
6
3
1
1
1
1
6
1
2
1
92
1
1
4
1
1
1
1
1
17
1
1
3
1
1
4
42
1
TTSITSVL
SSTSSTPQ
TASKTSGP
TSKTSAAT
TTSITSGP
QTSKTSAA
TTSITSVP
TTSITSAP
TTSTASVS
GSTPSPVP
TTSITSAP
TTSITSAR
TSSTSSTP
TSSTTSTP
TSSTTSAP
TTSITSGA
TTSTASGP
TSTISAPT
TTSISAL
TTRTTPAS
ETSTTSAP
TASTTFAP
TTSIISAP
TTSITSAS
TTSTILAP
RTSTTSAS
TTSTTLAP
TASTTSAP
TTSTTLSP
ITSTTSAP
TTSTTPVS
QTSTTSAP
TTSTTSTP
ASTISAPT
TSTTSAPT
QTSTTSAS
TNTTSAPT
TTSTTSAT
GATPSPVP
SSTTSGSG
TTSRISGP
TSTASVS
QTSTTYAH
TSSTISAP
TPRRTSAP
TTSTISAS
TSSTISAR
TSTTSFHT
TTSTPSAP
TTSTTSVT
QISTTSAP
TTSTIRSS
GTTPSAVP
QTSISSAP
TSSTTSGP
TASTTSGP
TTSTIPAS
GNTPSPVP
TTSTTSSP
QTTTSSAP
KTSTTSTP
STSTTSAA
ITSTISVS
TTPSPVTT
TSTISSPT
TTSTTSFH
TTSMTSAP
TSSMTSGP
TSSTTSTA
TTSTIISPL
TSSTTSSS
SSTTSAPT
TPSTTSAP
ISSTTSAT
TSTSSAPT
TASKTSGL
ISSTTSTP
ITSMTSGP
TTRTTSAP
TTRTTSVP
TSSTTSSP
KTSTTSTT
TSSTTSAT
GTTPSPVP
ETTPSPVP
TTSTTSAA
TASTTSAS
TTSITSVP
TTSITSGP
GTSLSPVP
TTCTTSGP
GTTPSPVP
TSTTSPPT
ETTPSPVP
TARTTSAP
TTGITSGP
GTTPSPVL
TTSITPGP
TSSITSTT
TTSITPAS
TTSTISSP
TTRTTSAS
TSTTSSPT
TTSTIFAP
TSISSAPT
GTTPSPVT
TTSTISGP
TSTTSAPI
GTTPSRVP
TSSITSTP
TTSITSTT
KTTPSPVP
TTPSPVPT
GTTSSPVP
ITSTTSGP
PASTTSGP
TSSTTSGS
TTSTISVP
TTSTNSAP
IPSTTSAP
TTSTTSPP
QISTISAP
KSSTTSAA
TSTTSTPQ
QTSKISAA
THRTTSGP

Figure S1. MUC5AC protein VNTR motif utilization across 206 human haplotypes. MUC5AC protein 8-mer variation plots across the five possible VNTR domains. Domain plot height corresponds to the total number of unique alleles across the large central exon. Numbers per unique allele correspond to the number of haplotypes that are matching in motif composition across all domains in linear sequence space. Black boxes indicate an absence of sequence. Gray represents cys domain sequence between distinct VNTR domains. Alleles are ordered vertically by hierarchical clustering of motif composition. Motifs are shifted in sequence space based on the same identity (indicating linear shifts of individual motifs).

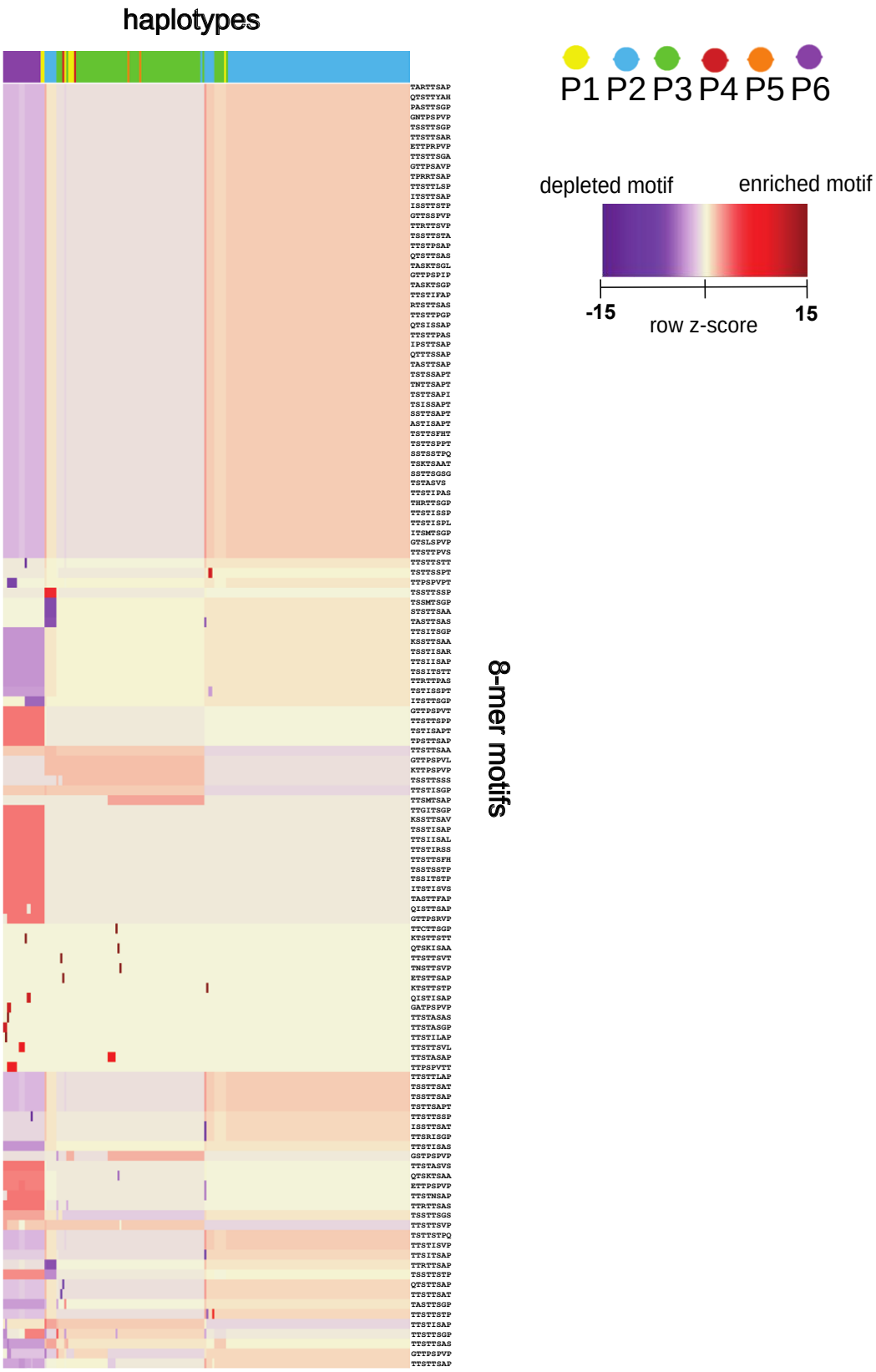

Figure S2. Extended version of Figure 1G.

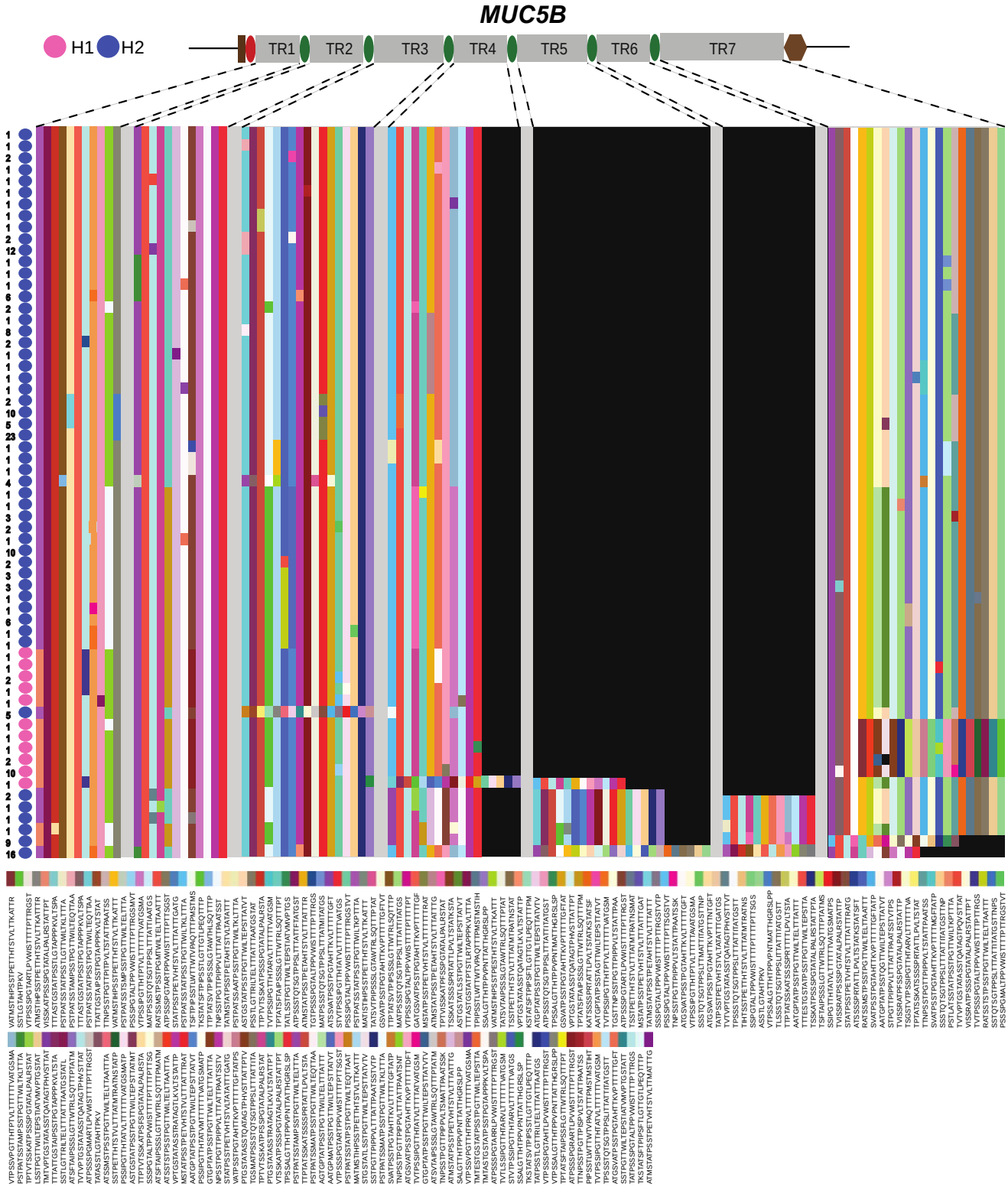
MUC5B
H1
H2
TR1
TR2
TR3
TR4
TR5
TR6
TR7

Figure S3. MUC5B protein VNTR motif utilization across 206 human haplotypes. MUC5B protein 29-mer variation plots across the five possible VNTR domains. Domain plot height corresponds to the total number of unique alleles across the large central exon. The numbers per unique allele correspond to the number of haplotypes that are matching in motif composition across all domains in linear sequence space. Black boxes indicate an absence of sequence. Gray represents cys domain sequence between distinct VNTR domains. Alleles are ordered vertically by hierarchical clustering of motif composition. Motifs are shifted in sequence space based on the same identity (indicating linear shifts of individual motifs).

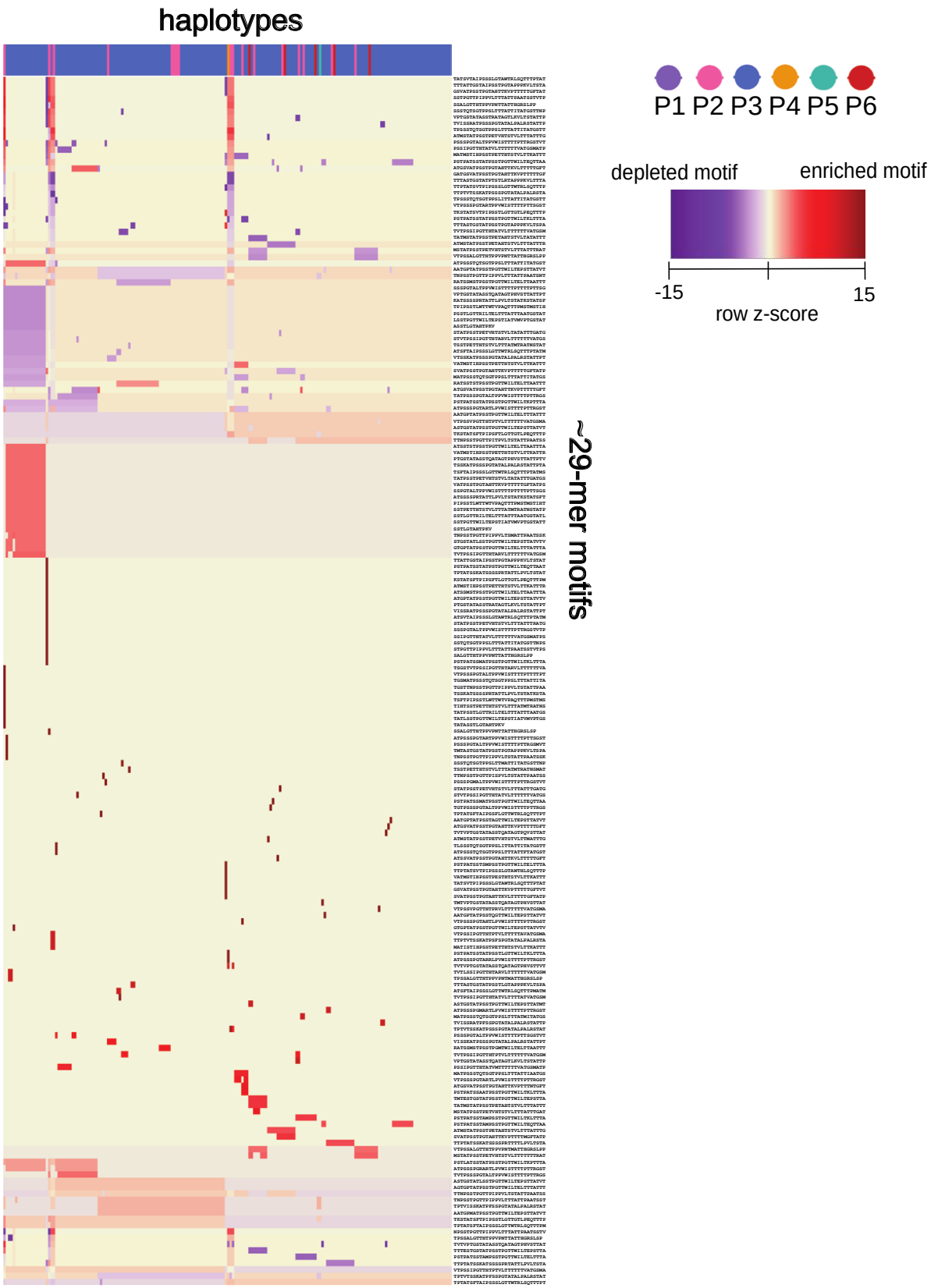

Figure S4. Extended version of Figure 2G.

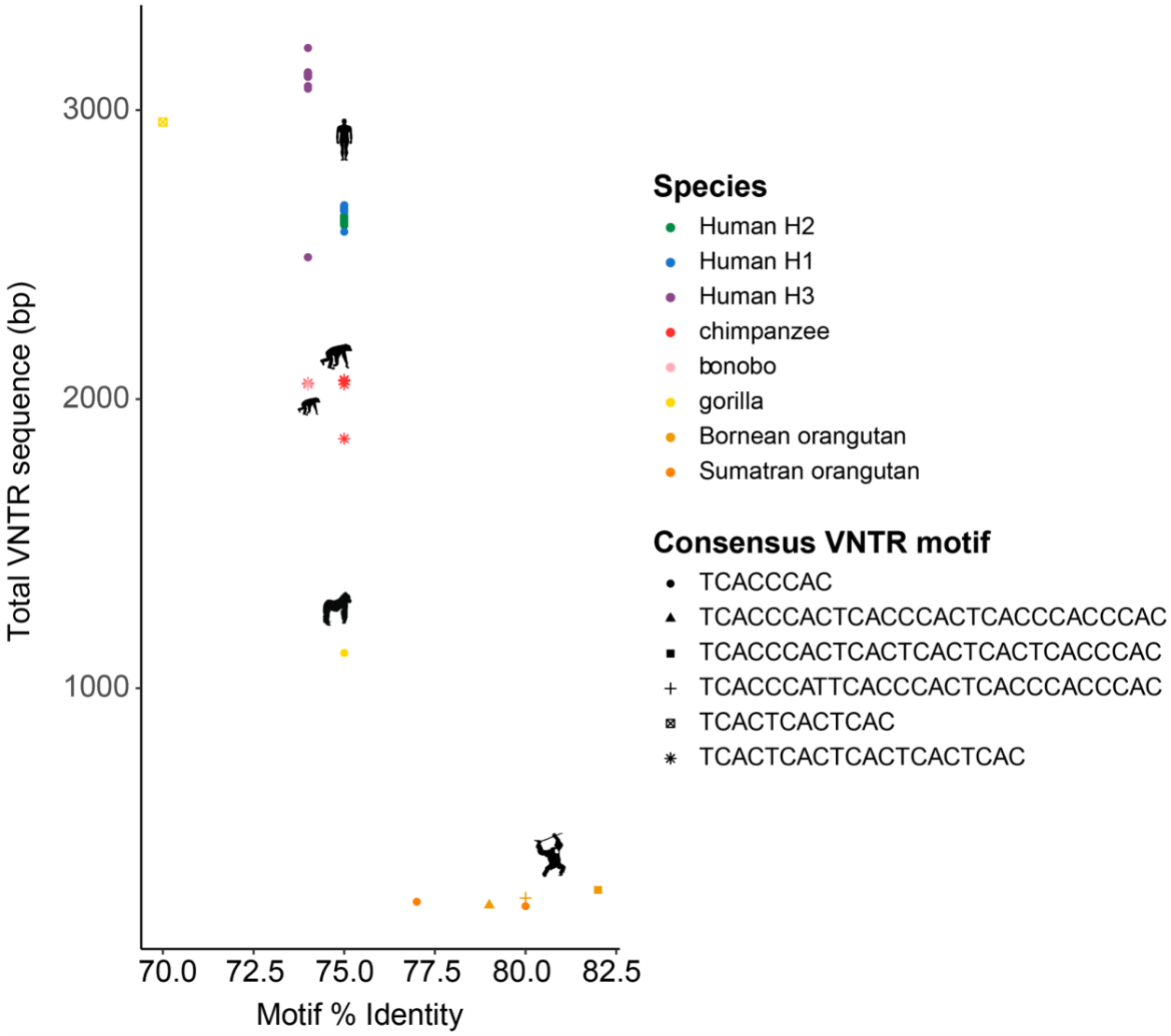

Figure S5. Intron 15 VNTR sequence characterization in *MUC5AC* for humans and nonhuman apes. Comparison of total VNTR sequence length (in base pairs) and canonical motif percent identity across 206 human haplotypes and at least two haplotypes per nonhuman ape species. Species/haplogroup identity coded by color and shape coded by motif.

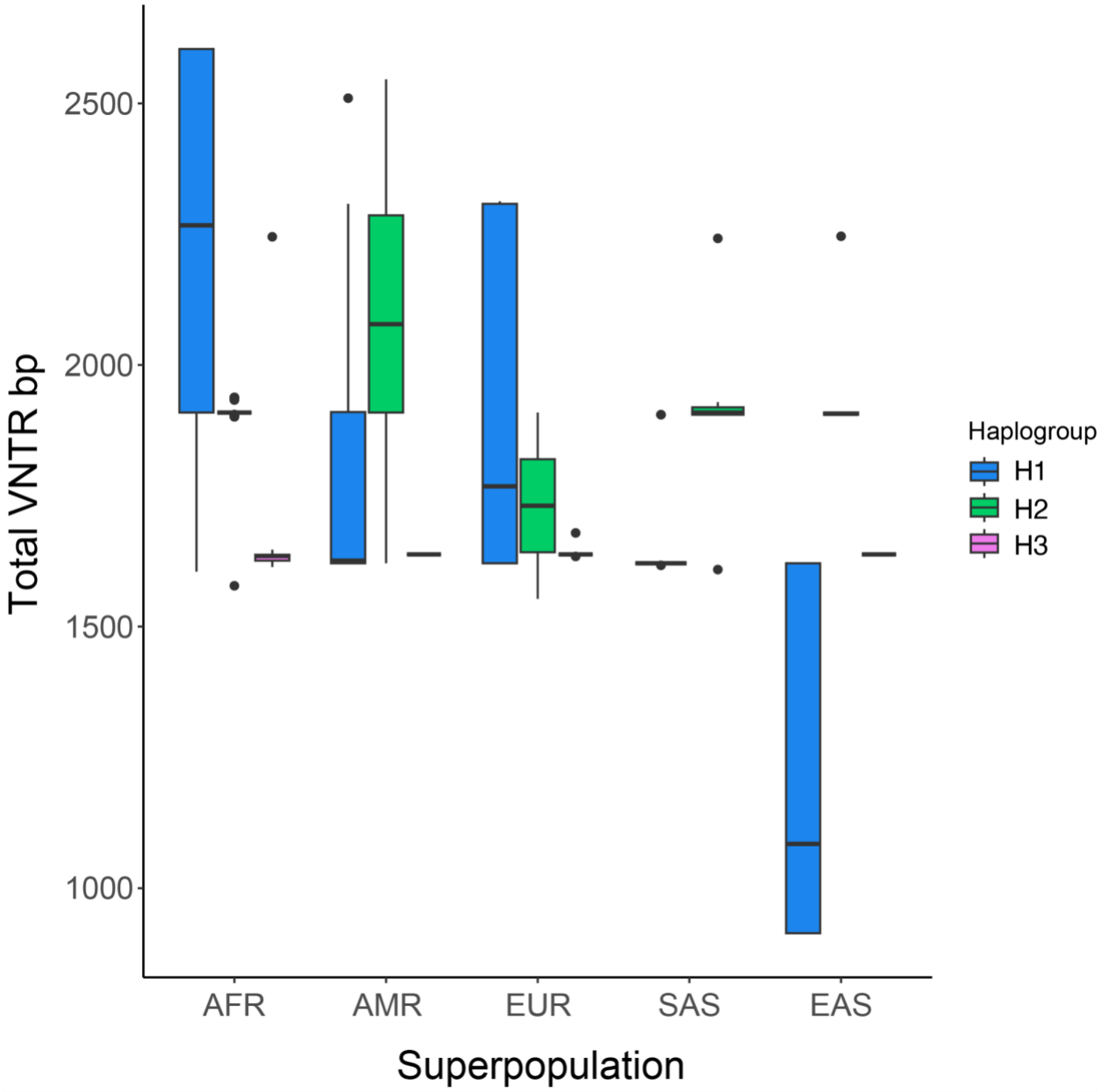

Figure S6. *MUC5AC* VNTR enhancer polymorphism across human super populations in the HGSVC/HPRC sample set, stratified by haplogroup identity.

Total VNTR bp indicates the longest continuous track of degenerate 8-mer VNTR sequence in the enhancer region of *MUC5AC.* Haplogroups correspond to phylogenetic clades of the *MUC5AC* locus.

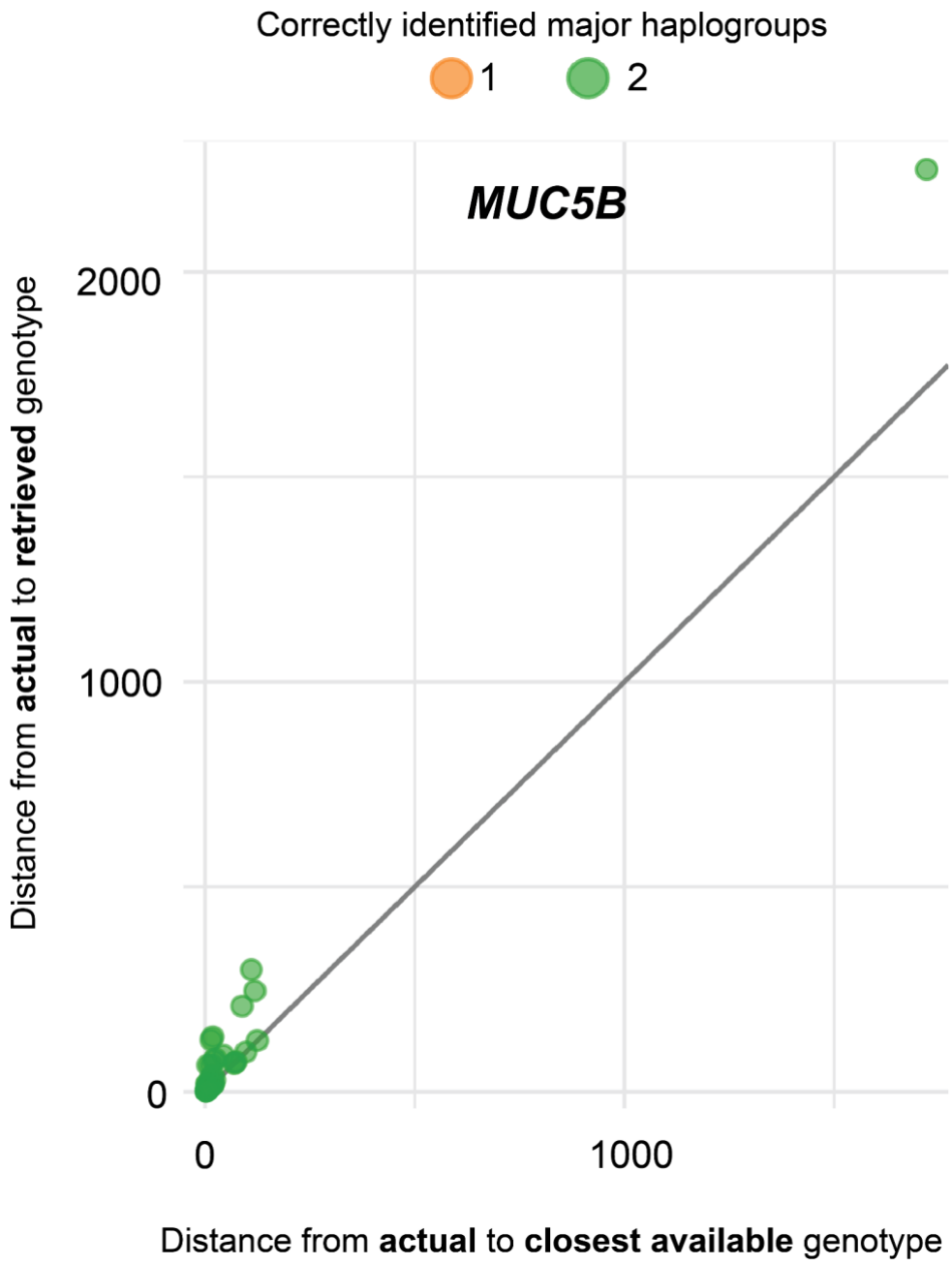

Figure S7. Genotyping accuracy of *MUC5B* haplogroups with Locityper. Locityper leave-one-out results comparing edit distances between actual and retrieved genotype (predicted from genotyper) versus edit distances between actual and closest possible genotype (best possible reference genotype from multiple sequence alignment with true genotype) for *MUC5B.* Dot color corresponds to the number of haplotypes in diploid sample sets that were correctly genotyped.

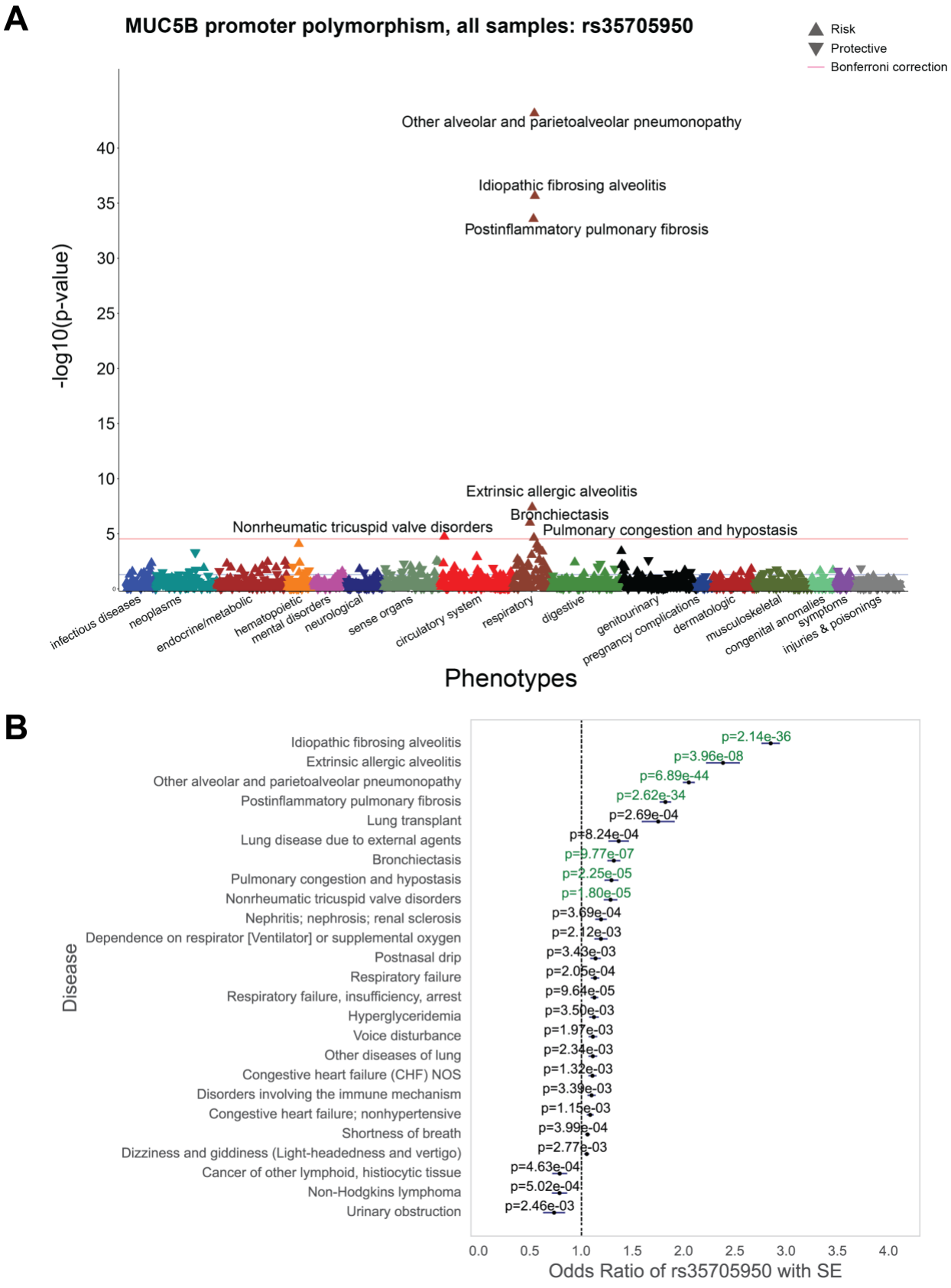
A
MUC5B promoter polymorphism, all samples: rs35705950
Risk
Protective
Bonferroni correction
-log10(p-value)
Other alveolar and parietoalveolar pneumonopathy
Idiopathic fibrosing alveolitis
Postinflammatory pulmonary fibrosis
Extrinsic allergic alveolitis
Bronchiectasis
Nonrheumatic tricuspid valve disorders
Pulmonary congestion and hypostasis
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & poisonings
Phenotypes
B
Idiopathic fibrosing alveolitis
Extrinsic allergic alveolitis
Other alveolar and parietoalveolar pneumonopathy
Postinflammatory pulmonary fibrosis
Lung transplant
Lung disease due to external agents
Bronchiectasis
Pulmonary congestion and hypostasis
Nonrheumatic tricuspid valve disorders
Nephritis; nephrosis; renal sclerosis
Dependence on respirator [Ventilator] or supplemental oxygen
Postnasal drip
Respiratory failure
Respiratory failure, insufficiency, arrest
Hyperglyceridemia
Voice disturbance
Other diseases of lung
Congestive heart failure (CHF) NOS
Disorders involving the immune mechanism
Congestive heart failure; nonhypertensive
Shortness of breath
Dizziness and giddiness (Light-headedness and vertigo)
Cancer of other lymphoid, histiocytic tissue
Non-Hodgkins lymphoma
Urinary obstruction
Disease
p=2.14e-36
p=3.96e-08
p=6.89e-44
p=2.62e-34
p=2.69e-04
p=8.24e-04
p=9.77e-07
p=2.25e-05
p=1.80e-05
p=3.69e-04
p=2.12e-03
p=3.43e-03
p=2.05e-04
p=9.64e-05
p=3.50e-03
p=1.97e-03
p=2.34e-03
p=1.32e-03
p=3.39e-03
p=1.15e-03
p=3.99e-04
p=2.77e-03
p=4.63e-04
p=5.02e-04
p=2.46e-03
Odds Ratio of rs35705950 with SE

Figure S8. PheWAS of *MUC5B* promoter polymorphism rs35705950 in *All of Us*.

A. Phenome-wide association study (PheWAS) Manhattan Plot of phenotype groupings and associations with the *MUC5B* promoter polymorphism rs35705950 in all populations (i.e., all samples). Only phenotypes that passed Bonferroni correction are noted by name.
B. Odds ratios of the top 25 diseases (based on ordered p-values) for associations with rs35705950. Green indicates phenotypes that passed Bonferroni correction. SE = standard error (horizontal bars corresponding to each p-value).