Supplementary Methods for Predicting 102 Novel Public GvL Restricted mHA:

mHA Prediction Algorithm Overview

The minor histocompatibility antigen (mHA) prediction algorithm presented here consists of a stepwise pipeline designed to capture single nucleotide polymorphism (SNP) genetic variation between allogeneic stem cell transplant (allo-SCT) recipients and their HLA-matched donors. It then predicts 'public GvL restricted' mHA peptides based on *in silico* peptide binding to cognate HLA, their parent gene expression in relevant tissues, and their occurrence at optimal population frequency on common HLA.

Capturing Genetic Variation and Patient Cohort Censoring

The mHA prediction algorithm was applied retrospectively to a clinical data set consisting of 139 patients who received allo-SCT for myeloid leukemia (AML, CML, MDS) at M.D. Anderson under an IRB-approved protocol (LAB99-062). This data set included Illumina NS-12 microarray SNP genotype information at 13,917 common coding SNPs for all SCT recipients and their matched donors, with analysis performed at the University of Texas at Houston School of Medicine microarray core facility.[1] From this retrospective data set, 21 patients were censored who had greater than 5% Illumina NS-12 microarray SNP genotype failure or had 100-fold smaller donor-recipient genetic variation than the mean. An additional 17 patients were censored who received either cord blood SCT, allo-SCT with less than 8 for 8 HLA match, repeat SCT from a second donor different than the original, or who did not have HLA-A*02:01. The SNP genotype data from the remaining 101 HLA-matched donor recipient pairs (DRPs) were considered for further analysis. Additional overall survival (OS) curves demonstrate results

typical of a cohort of SCT recipients for myelogenous leukemia (**Figure 2, Supplementary Figure 1**).

Filtering for Coding SNPs

The ENSEMBL Variant Effect Predictor (VEP), used with human genome assembly GRCh37, identified SNPs likely to result in actual peptide sequence differences (coding SNPs or cSNPs) between recipients and their matched donors.[2] Both the VEP and the mHA prediction algorithm more generally accept SNP data in reference SNP ID number (rs number) format as well as whole exome sequencing (WES) data in variant call format (VCF) and mutant annotation format (MAF). Only those SNPs identified by the VEP as likely to result in 'protein coding missense variants' were considered. 11,336 cSNPs of the 13,917 rs numbers tested by the Illumina NS-12 microarray met this criterion. Of these 11,336 cSNPs, only 11,172 cSNPs corresponded unambiguously with a single chromosomal base location, strand, and open reading frame (ORF) within UCSC hg19 annotation and were considered for further analysis.[3]

In Silico Peptide Generation

For every cSNP with appropriate genetic variation within at least one clinical DRP, all possible 8-11, 16, 20, and 24-mer donor and recipient peptides were enumerated based on the human reference genome GENCODE v19 (GRCh37.p13)[4] using custom software. Appropriate genetic variation consisted of recipient allelic variation which was not also found in that recipient's HLA-matched donor. Genetic variation was found in at least one clinical DRP arising from 9,575 cSNPs of the 11,172 cSNPs considered. These 9,575 cSNPs corresponded to 17,111 messenger RNA transcripts. For each identified transcript, every 24, 27, 30, 33, 48, 60, and 72 base pair (bp) sequence that both overlapped an identified cSNP, and that was found in the

appropriate ORF, was translated *in silico* to its corresponding 8-11, 16, 20, and 24-mer peptides for both the donor and recipient alleles. This set of peptides was considered for further analysis.

Predicting Peptide HLA Binding

The set of enumerated peptides were screened *in silico* for predicted binding to cognate DRP HLA using NetMHCpan v2.8 for peptides with 8-11 amino acids (AA) and using NetMHCIIpan v3.0 for peptides with 16, 20, and 24 AA.[5,6] HLA with sensitivity ≥ 0.5 and specificity ≥ 0.7 were considered 'verified' (**Supplementary Figure 2A-D**). In order to predict mHA that could be further investigated biochemically and biologically, we prioritized specificity over sensitivity. When patients only had HLA antigen level typing, we resorted to imputation to the most common subtype to determine allele level sequences. Studies have shown that the MHC binding pocket tertiary structure, electrostatic properties, and disease associations are generally well conserved across HLA subtypes given the same HLA type.[7-9] Our studies using publicly available NetMHC validation data from benchmarking studies[10,11] have shown that the actual peptide binding profiles for different HLA subtypes are similar (**Supplementary Figure 3A-D**). Only peptides binding these verified HLA were considered for further analysis.

Tissue Restriction: Expression Thresholds vs Pearson Gene Inclusion/Exclusion Model

GvL restricted mHA were considered to be those mHA likely to be found in any prospective GvL tissue but unlikely to be found in all tissues commonly affected by GVHD. Gene expression is likely the first order determinant for whether or not a protein's peptide fragments are successfully processed and presented on cell surfaces within a given tissue.[13] Normal bone marrow, testis, and AML sample tissues were considered 'GvL tissues.' Normal skin,

hepatobiliary, and colonic tissues were considered 'GvH tissues.' AML tissue expression values were generated using mean RNAseq levels from 8 AML samples of various subtype with patient leukemia blasts taken at the time of diagnosis as part of an IRB approved protocol at MD Anderson Cancer Center.[12] Normal tissue expression values were taken from Human Protein Atlas RNAseq data developed under EMBL-EBI ArrayExpress / Expression Atlas Experiment E-MTAB-2836, sampling 32 tissues from 122 healthy individuals.[14] mHA peptides were filtered to ensure they had 'high expression' in at least one GvL tissue, defined as mHA parent gene expression greater than 50 transcripts per million (TPM). Similarly, mHA peptides were filtered to ensure they had 'low expression' in all common GVHD tissues, defined as mHA parent gene expression less than 5 TPM. These expression thresholds are similar to the thresholds used by various RNAseq tissue expression databases. The Human Protein Atlas labeled genes greater than 50 FPKM as having 'high expression' and less than 10 FPKM as 'low expression' ($Gene\ Expression_{TPM} = Gene\ Expression_{FPKM} * 10^6 / \sum_{Genes} Expression_{FPKM}$). There was no association between OS or relapse and the number of predicted GVL or total mHA (**Figure 2**)

Determining 'GvL restricted genes,' (ie the set of genes from which GvL restricted mHA peptides arise) from tissue gene expression alone gives results similar to using the more complicated MHC associated peptide (MAP) gene inclusion/exclusion model developed by Pearson et al (Pearson model).[14] The Pearson model takes into account many factors beyond gene expression, such as the potential for nonsense mediated decay as well as protein structure. For genes with 'high' and 'low' expression, the Pearson model largely reduces to a question of gene expression. When the Pearson model is applied to the clinical data set, there is only 1.1% disagreement regarding 4,410 'source' GvL genes and only 3.9% disagreement regarding 1,111 'non-source' GvH genes compared to relying on expression thresholds alone. Integrating the

Pearson model into the mHA prediction algorithm presented here yielded similar results (**Supplementary Figure 7A-F**) as when considering tissue expression alone, with the potentially notable exception of a newly statistically significant relationship between the number of GvH mHA and GVHD for MUD that we consider to be a preliminary finding (**Supplementary Figure 7F**).

Predicting 'Public' GvL restricted mHA

For a given predicted mHA to be considered 'public,' the genetic condition for a minor mismatch (gMM) giving rise to a mHA peptide, as well as its cognate HLA, would both need to occur with relatively high frequency in the general population. The vast majority of predicted mHA are 'private.' Most HLA occur at low population frequency and most genetic variation is rare; less than 5% of known cSNPs have the appropriate allele frequency to result in 'public' mHA. Even with an optimal gMM allele frequency, only 25% of MUD SCT would actually contain the gMM, with the rate being slightly depressed for MRD SCT. Assuming optimal gMM allele frequency and HLA-A*02:01, which occurs with a high population frequency of 24%, an 'ideal' public mHA would be found in approximately 6% of allo-SCT. By filtering the set of predicted GvL restricted mHA for HLA that occur with a population frequency of at least 10% and for gMM recipient allele frequencies between 10% and 70% (resulting in a gMM population frequency of at least 10%), the resulting 'public' GvL restricted mHA should occur in at least 1% of allo-SCT.[15,16] Of the 13,917 genetic locations tested across our patient cohort, 102 cSNPs were predicted to lead to novel 'public' GvL restricted mHA peptides, variously presented on 7 high frequency HLA (HLA-A*02:01, A*03:01, A*11:01, A*24:02, B*07:02, B*35:01, and B*44:03). These public GvL restricted mHA are common in SCTs (**Supplementary Figure 8**).

Supplementary Methods for Confirming UNC-GRK4-V

Tissue Expression of GRK4 mRNA

Tissue specific RNA was purchased (Ambion, Applied Stem Cell) or purified from de-identified primary human AML cells obtained from the UNC Hematologic Malignancies Tissue Procurement Facility (UNC-TPF) using the Qiagen RNeasy kit. RT-PCR was performed on 100 ng of each RNA sample using the SuperScript III One-Step RT-PCR kit (Ambion) with the following primers: left = 5'-ATCAACTTCAGGAGGCTGGA, right = 5'-AGACACACCCGGTAGCAAAC. Thermocyler conditions for RT-PCR were RT: 50°C × 30min; PCR: 94°C × 30s, 60°C × 30s, 68°C × 60s (40 cycles); final extension 68 °C × 5min.

Protein expression of GRK4

Human testis protein lysate was purchased from AbCAM, and protein lysates were prepared from primary human AML cells using RIPA buffer. 50μg of each lysate were loaded onto 4-15% pre-cast polyacrylamide gels (BioRad) and run in Tris/Glycine/SDS buffer. Proteins were transferred to a PVDF membrane. GRK4 was probed using mouse anti-GRK4 (clone A-5, Santa Cruz) and detected using ECL. β-Actin was probed using anti-β-actin (Sigma).

Isolation of class I peptide epitopes

U937 cells transfected with HLA-A*02:01 (U937.A2) were genotyped at rs1801058 by Sanger sequencing of a PCR product containing the cSNP: primers left = 5'-GCGTTTCATTCTTGGGAACT, right = 5'-TCCTTACAGTAAACGGCATGA, thermocyler conditions: preheat: 94°C × 2min, PCR: 94°C × 30s, 60°C × 30s, 68°C × 3min (40 cycles), final extension: 68 °C × 7min. The cell line was found to be heterozygous. Roughly 2×10^8 U937.A2

cells were grown in RPMI 1640 media, lysed and cleared by centrifugation. Anti-class I HLA antibody (W6/32) was added to the cleared lysate and incubated on a rocker at 4°C overnight. Protein G separose resin was added to the lysate for immunoprecipitation over 4 hours. The resin was washed and loaded onto columns. Peptide/HLA complexes were released by washing with 10% acetic acid. Glacial acetic acid was then added dropwise to achieve a pH of ~2.5. The mixture was filtered through 5 kDA filters to isolate the ~1 kDA peptide epitopes.[17,18]

DIMS-MS/MS targeted mass spectrometry for UNC-GRK4-V

Pure UNC-GRK4-V peptide was diluted to 100 nM in water/acetonitrile/formic acid (50/50/0.1%) prior to nanoelectrospray-DIMS-MS. The tandem mass spectrometry (MS/MS) spectrum and E_C for optimal transmission of UNC-GRK4-V were recorded.[19,20] The peptide epitope pool was desalted using a ZipTip (C_{18}, 0.6µL) and directly nano-electrosprayed into the DIMS-MS with the E_C set for optimal transmission (86 V/cm). The MS/MS spectrum of the UNC-GRK4-V peptide from the epitope pool was compared to that of the pure UNC-GRK4-V peptide, and a Fit score was calculated based on the agreement between the reference and experimental UNC-GRK4-V MS/MS spectra using Bruker DataAnalysis. The Fit score was calculated as the sum of the product of peak intensities of the experimental unknown and library spectrum squared, divided by the product of the sum of the squared peak intensities in the library spectrum and the sum of the squared peak intensities in the experimental spectrum where the library spectrum has peak intensity greater than zero.

$$Fit\ score = 1000 * \frac{(\sum u * l)^2}{\sum l^2 * \sum_{l>0} u^2}$$

*Synthesis of UNC-GKR4-V/HLA-A*02:01 tetramers*

HLA-A*02:01 containing a biotinylation site and β_2 microglobulin were overexpressed in E. coli and purified by FPLC as previously reported.[21] The HLA-A*02:01, β2 microglobulin and UNC-GRK4-V peptide were combined to induce folded UNC-GRK4-V/HLA-A*02:01 monomers. Folded monomers were purified by FPLC. The purified monomers were biotinylated by BirA ligase. The biotinylated monomers were then complexed with APC-avidin to produce UNC-GRK4-V/HLA-A*02:01 tetramers.

Detection of UNC-GRK4-V specific $CD8^+$ T cells in post-SCT samples

Post-SCT peripheral blood mononuclear cell (PBMC) samples from HLA-A*02:01 expressing patients and donors were genotyped for the rs1801058 cSNP, under IRB approved protocols Lab 99-062 for M.D. Anderson samples and LCCC-0824 for UNC samples. The PBMC samples were thawed, washed in PBS and split so $CD8^+$ T cell avidity to both a negative tetramer and the UNC-GRK4-V/HLA-A*02:01 tetramer could be measured. The washed PBMC were incubated in DPBS + 0.5% BSA at 4°C for 30 min with the following antibodies and stained: DAPI live/dead stain, CD4-FITC, CD14-FITC, CD16-FITC, CD19-FITC (dump channel), CD8-PE, tetramer-APC (either negative tetramer or UNC-GRK4-V/HLA-A*02:01 tetramer). All samples were analyzed on a MacQuant analytical flow cytometer, and data were analyzed using FlowJo software.

References:

1. Armistead PM, Liang S, Li H, et al. Common minor histocompatibility antigen discovery based upon patient clinical outcomes and genomic data. *PLoS One*. 2011;6(8):e23217.
2. McLaren W, Gil L, Hunt S, et al. The Ensembl Variant Effect Predictor. *Genome Biology*. 2016;17:122.
3. Tyner C, Barber GP, Casper J, et al. The UCSC Genome Browser database: update 2017. *Nucleic Acids Res*. 2017;44(D1):D626-D634. hgdownload.cse.ucsc.edu/goldenPath/hg19/database/snp138.txt.gz. Accessed 02 June 2016.
4. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760-1774.
5. Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One*. 2007;2(8):e796.
6. Karosiene E, Rasmussen M, Blicher T, et al. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*. 2013.
7. Kangueane P, Sakharkar MK, Rajaseger G, et al. A framework to sub-type HLA supertypes. *Front Biosci*. 2005;10:879-886.
8. Harjanto S, Ng LF, Tong JC. Clustering HLA class I superfamilies using structural interaction patterns. *PLoS One*. 2014;9(1):e86655.
9. Khan MA. Polymorphism of HLA-B27: 105 subtypes currently known. *Curr Rheumatol Rep*. 2013;15(10):362.
10. Kim Y, Sidney J, Buus S, et al. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics*. 2014;15:241.
11. Wang P, Sidney J, Kim Y, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinformatics*. 2010;11:568.
12. Zhang M, Sukhumalchandra P, Enyenihi AA, et al. A Novel HLA-A*0201 Restricted Peptide Derived From Cathepsin G Is An Effective Immunotherapeutic Target in Acute Myeloid Leukemia. *Clin Cancer Res*. 2013;19(1):247-257.
13. Pearson H, Dauda T, Granados DP, et al. MHC class I–associated peptides derive from selective regions of the human genome. JCI. 2016;126(12):4690-4701.
14. Uhlén M et al, 2015. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. www.proteinatlas.org/about/download/rna_tissue.csv.zip. www.ebi.ac.uk/arrayexpress/files/E-MTAB-2836/E-MTAB-2836.idf.txt. Accessed 25 Jan 2016.
15. Armistead PM, Liang S, Li H, et al. Common minor histocompatibility antigen discovery based upon patient clinical outcomes and genomic data. PLoS One. 2011;6(8):e23217.
16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
17. Hunsucker SA, McGary CS, Vincent BG, et al. Peptide/MHC Tetramer-Based Sorting of CD8+ T Cells to a Leukemia Antigen Yields Clonotypes Drawn Nonspecifically from an Underlying Restricted Repertoire. *Cancer Immunol Res*. 2015;3(3):228-235.
18. Zhang M, Sukhumalchandra P, Enyenihi AA, et al. A novel HLA-A*0201 restricted peptide derived from cathepsin G is an effective immunotherapeutic target in acute myeloid leukemia. *Clin Cancer Res*. 2013;19(1):247-257.

19. Dharmasiri U, Isenberg SL, Glish GL, Armistead PM. Differential ion mobility spectrometry coupled to tandem mass spectrometry enables targeted leukemia antigen detection. *J Proteome Res*. 2014;13(10):4356-4362.
20. Isenberg SL, Armistead PM, Glish GL. Optimization of peptide separations by differential ion mobility spectrometry. *J Am Soc Mass Spectrom*. 2014;25(9):1592-1599.
21. Kerry SE, Buslepp J, Cramer LA, et al. Interplay between TCR affinity and necessity of coreceptor ligation: high-affinity peptide-MHC/TCR interaction overcomes lack of CD8 engagement. *J Immunol*. 2003;171(9):4493-4503.

Supplementary Figure 1: Additional Patient Cohort Survival Outcomes. (A) There was a statistically significant difference in OS according to age ($p = 0.0027$). (B-D) There was no significant difference in OS based on SCT source (peripheral blood vs bone marrow) or patient sex (female vs male) or DRP gender mismatch (female donor into male recipient vs all other combinations).

Supplementary Figure 2: NetMHCpan Performance Testing. (A, B) There was no statistically significant correlation between the number of training peptides during NetMHCpan performance testing and the sensitivity or specificity with which NetMHCpan predicted previously validated experimental peptide HLA binding affinity, though the data does trend toward more precision and accuracy with an increased number of test peptides. (C, D) There was a statistically significant correlation (and negative correlation, respectively) between the percent of test peptides experimentally proven to bind stronger than 500nM and the sensitivity and specificity of the NetMHCpan algorithm for predicting binders. For HLA with training sets having a higher percentage of peptide HLA 'binders,' the NetMHCpan algorithm had increased sensitivity and decreased specificity ($p = 5.7e-8$ and $p = 6.1e-17$, respectively).

Supplementary Figure 3: NetMHCpan HLA Verification. (3A,B) HLA-A*02:01 has a similar experimental and predicted peptide binding profile to HLA-A*02:02 and (3C, D) to HLA-A*02:03. (3E) HLA-A*02:01, one of the most well studied alleles, has predictably high accuracy. (3F) HLA-B*51:01 barely met inclusion criteria but still retained good accuracy. (3G) HLA-B*15:02 was one of the few Class I HLA alleles that did not meet inclusion criteria and has significantly worse accuracy.

Supplementary Figure 4: GRK4 Expression Across Tumor Types in TCGA. This violin plot compares GRK4 expression (RSEM upper quartile normalized log transformed read counts) across tumor types in TCGA. GRK4 is expressed at a high level in AML ("LAML") compared to many other cancer types. Glioma and glioblastoma tumors also had relatively high expression of GRK4 compared to other tumor types.

ACC	**Adrenocortical carcinoma**
BLCA	**Bladder Urothelial Carcinoma**
BRCA	**Breast invasive carcinoma**
CESC	**Cervical squamous cell carcinoma and endocervical adenocarcinoma**
CHOL	**Cholangiocarcinoma**
COAD	**Colon adenocarcinoma**
DLBC	**Lymphoid Neoplasm Diffuse Large B-cell Lymphoma**
ESCA	**Esophageal carcinoma**
GBM	**Glioblastoma multiforme**
HNSC	**Head and Neck squamous cell carcinoma**
KICH	**Kidney Chromophobe**
KIRC	**Kidney renal clear cell carcinoma**
KIRP	**Kidney renal papillary cell carcinoma**
LAML	**Acute Myeloid Leukemia**
LGG	**Brain Lower Grade Glioma**
LIHC	**Liver hepatocellular carcinoma**
LUAD	**Lung adenocarcinoma**
LUSC	**Lung squamous cell carcinoma**
MESO	**Mesothelioma**
OV	**Ovarian serous cystadenocarcinoma**
PAAD	**Pancreatic adenocarcinoma**
PCPG	**Pheochromocytoma and Paraganglioma**
PRAD	**Prostate adenocarcinoma**
READ	**Rectum adenocarcinoma**
SARC	**Sarcoma**
SKCM	**Skin Cutaneous Melanoma**
STAD	**Stomach adenocarcinoma**
TGCT	**Testicular Germ Cell Tumors**
THCA	**Thyroid carcinoma**
THYM	**Thymoma**
UCEC	**Uterine Corpus Endometrial Carcinoma**
UCS	**Uterine Carcinosarcoma**
UVM	**Uveal Melanoma**

The TCGA study abbreviations can be found at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations.

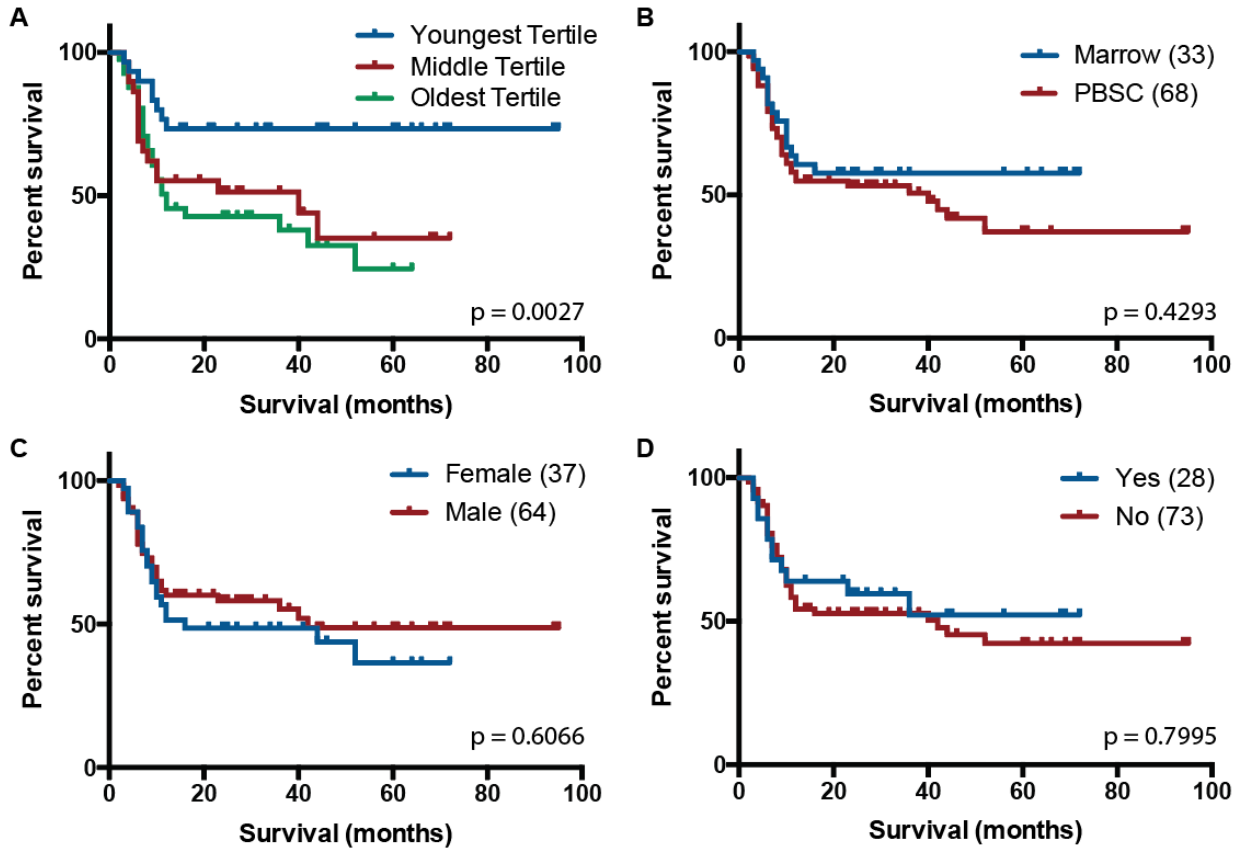
Supplementary Figure 5: Patient outcomes according to mHA number. (A-B) There was no significant difference in OS based upon the number of predicted mHA in MRD transplants (A) or MUD transplants (B). (C-D) There was also no significant difference in relapse rate based on the number of predicted GVL mHA in MRD transplants (C) or MUD transplants (D). (E) There is no association between number of total Class I GvL mHA and relapse for MRD or MUD. (F) There is no association between number of total Class I GvH mHA and GVHD.

Supplementary Figure 6: Characteristics of Class II mHA with Patient Outcomes. (A) There is a large difference in the number of peptides that can bind with high affinity to class I HLA (0.09 to 1.32 per gMM) and class II HLA (12.4 to 27.3 per gMM). (B) MUD SCT is associated with roughly twice as many class II mHA compared to MRD SCT. (C) There is no association between number of Class II GvL mHA and relapse for MRD or MUD. (D) There is no association between number of Class II GvH mHA and GVHD.

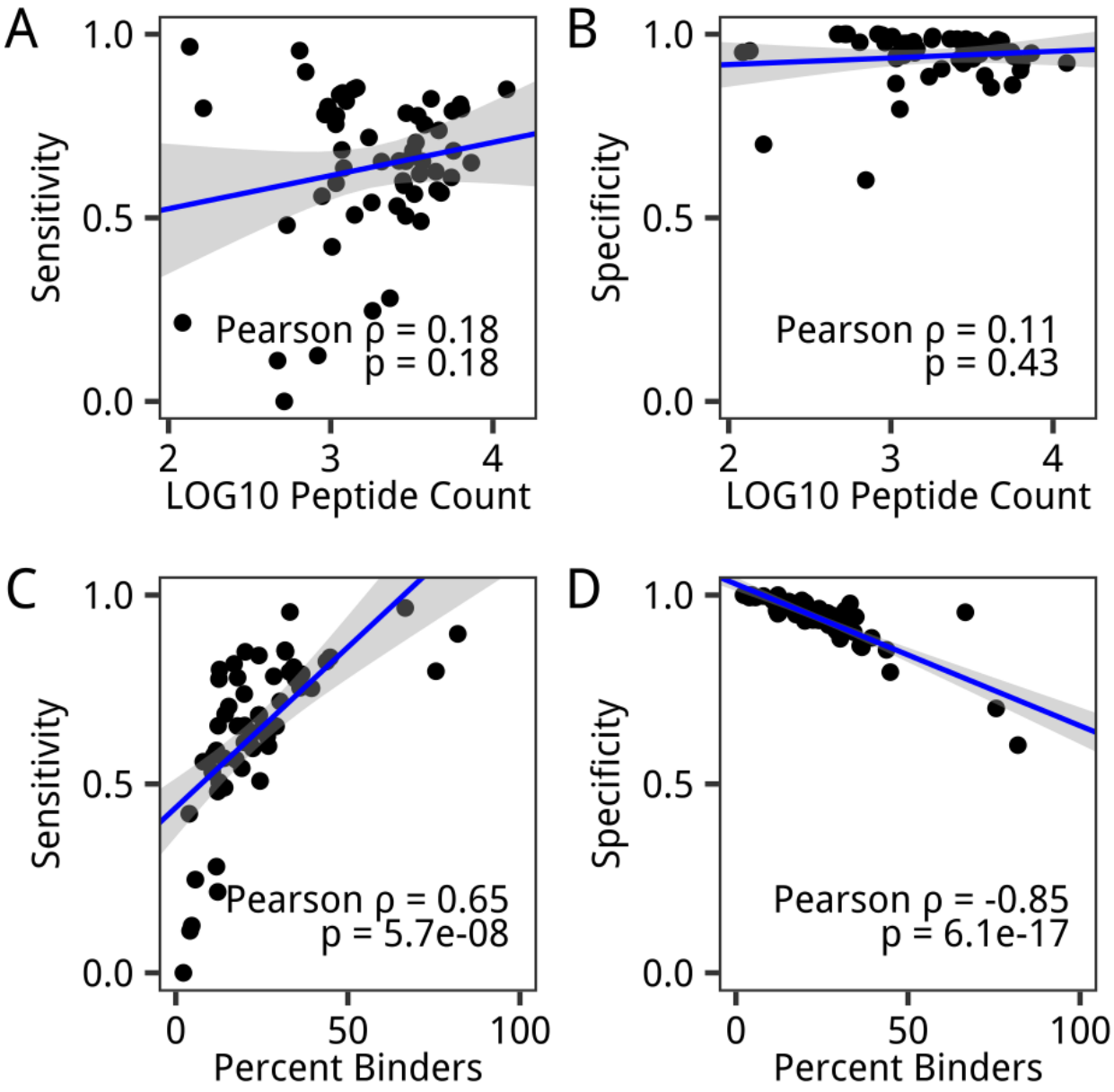
Supplementary Figure 7: Characteristics of GvL and GvH mHA using the Pearson Model. (A, B) The distributions for Class I and Class II GvL mHA determined with the Pearson model closely mirror the distributions of the underlying unrestricted mHA. (C) The number of GvL mHA correlates more tightly with the number of GvH mHA when determined with the Pearson model than with expression thresholds alone (Main Figure 3D). (D) The PFS curve does not significantly depend on the number of GvL mHA when determined with the Pearson model or with expression thresholds alone. (E, F) There is a significant association between the number of GvH mHA and GVHD in MUD when GvH mHA are determined with the Pearson Model ($n = 29$, $p = 0.01$) that is not seen with expression thresholds alone (Main Figure 3F).

Supplementary Figure 8: Coverage of Study Population by Public GvL mHA Targets. Considering the most frequent GvL restricted mHA predicted in our study population first, the cumulative fraction of study DRPs that had at least one predicted GvL restricted mHA target rapidly approached 100% after considering only 12 of the 102 predicted public GvL restricted mHA. This curve approaches 100% somewhat faster than would a random sample of SCT DRPs as everyone in the study population possessed HLA-A*02:01 which is associated with a disproportionate fraction of the public GvL restricted mHA targets predicted by this study. Of the 102 predicted public GvL restricted mHA, 53 mHA had at least one peptide associated with HLA-A*02:01, and 92 mHA had at least one peptide associated with HLA-A*03:01, A*11:01, A*24:02, B*07:02, B*35:02, or B*44:03).

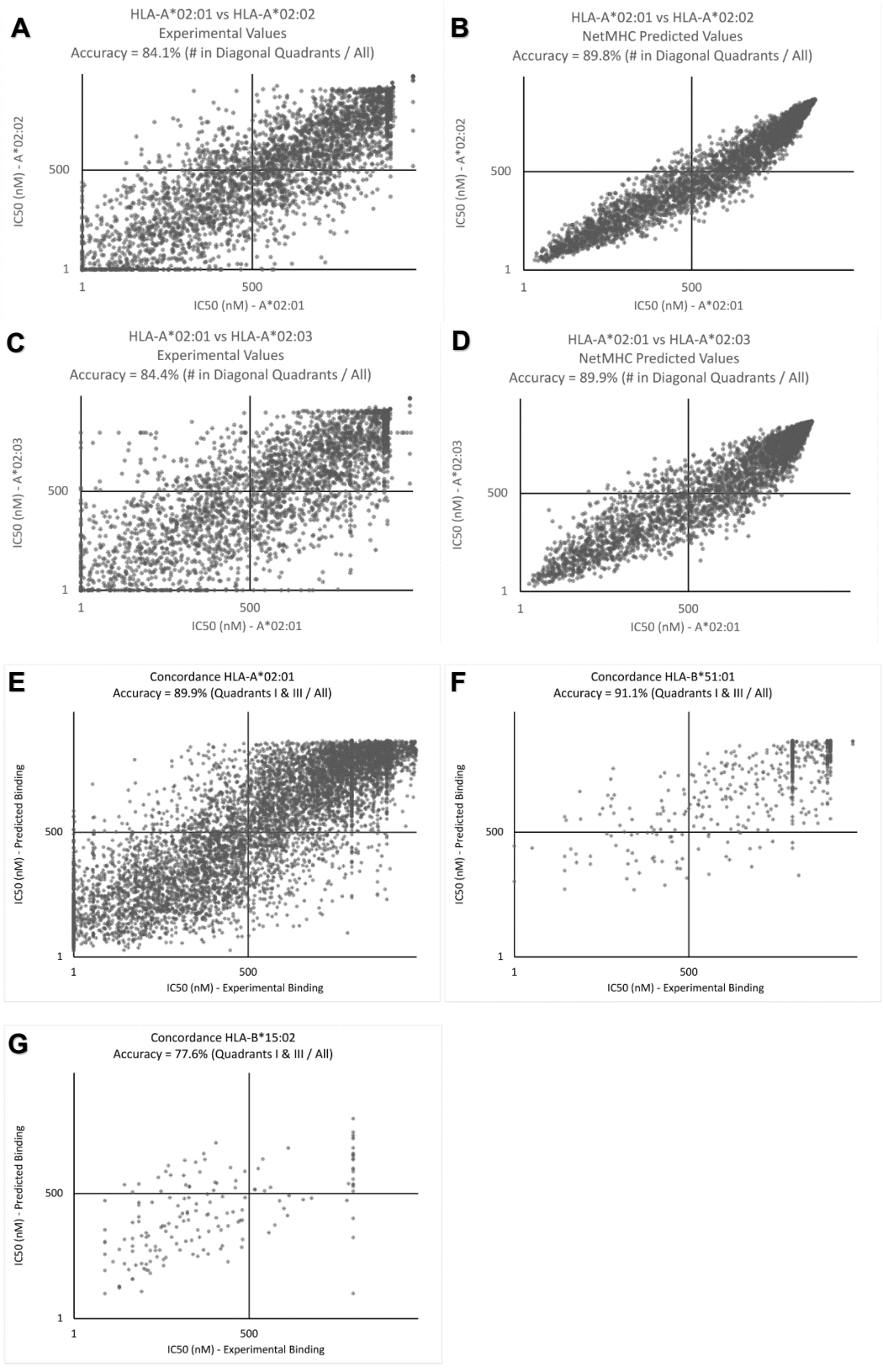
Supplementary Figure 1:



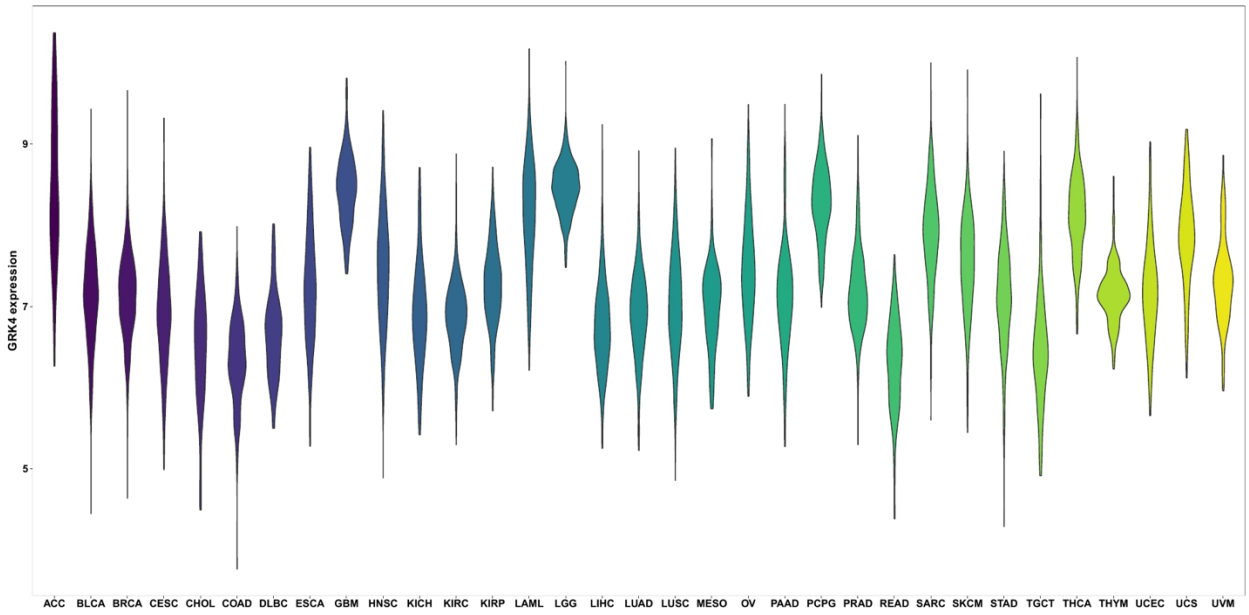
Supplementary Figure 2:



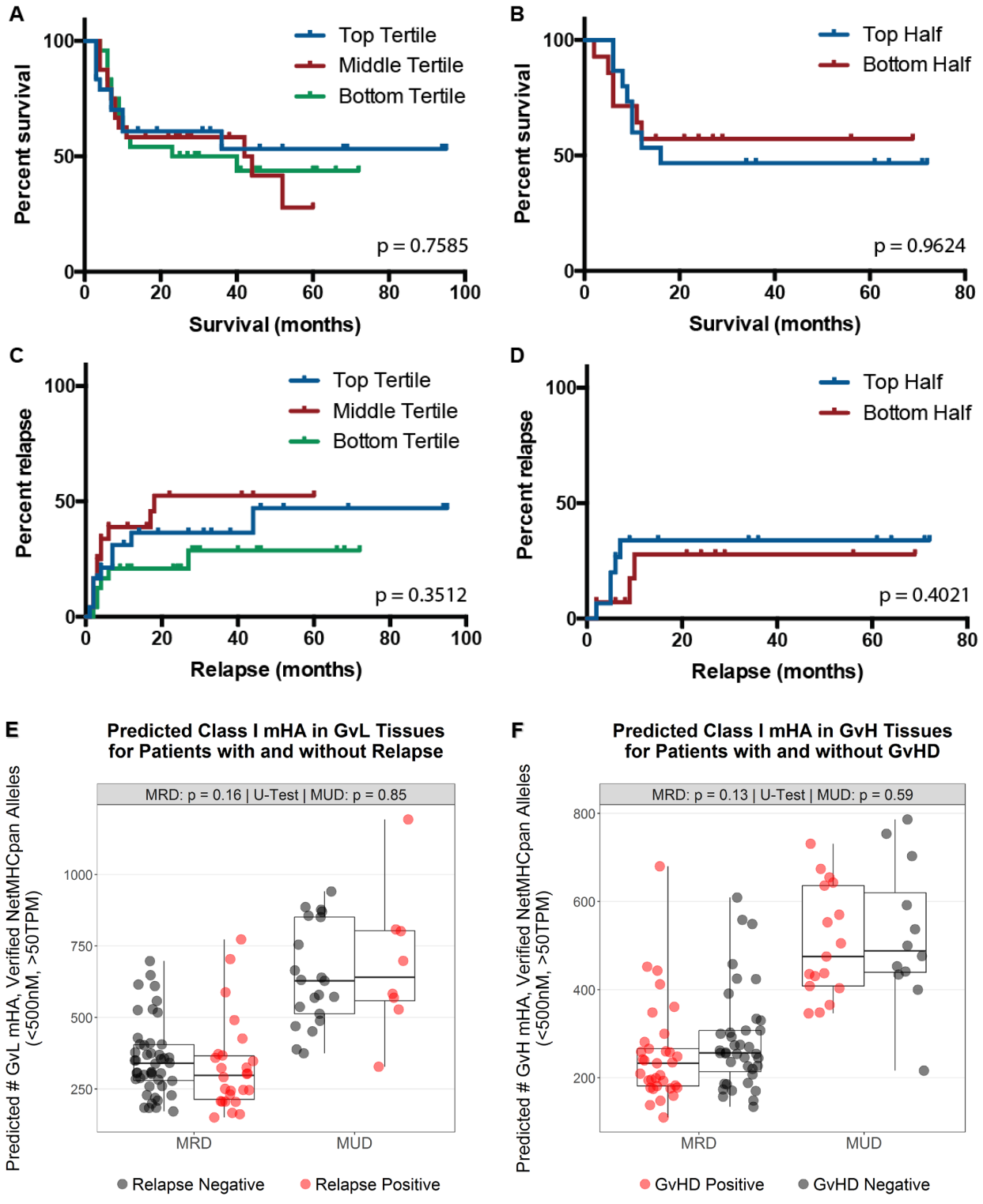
Supplementary Figure 3:



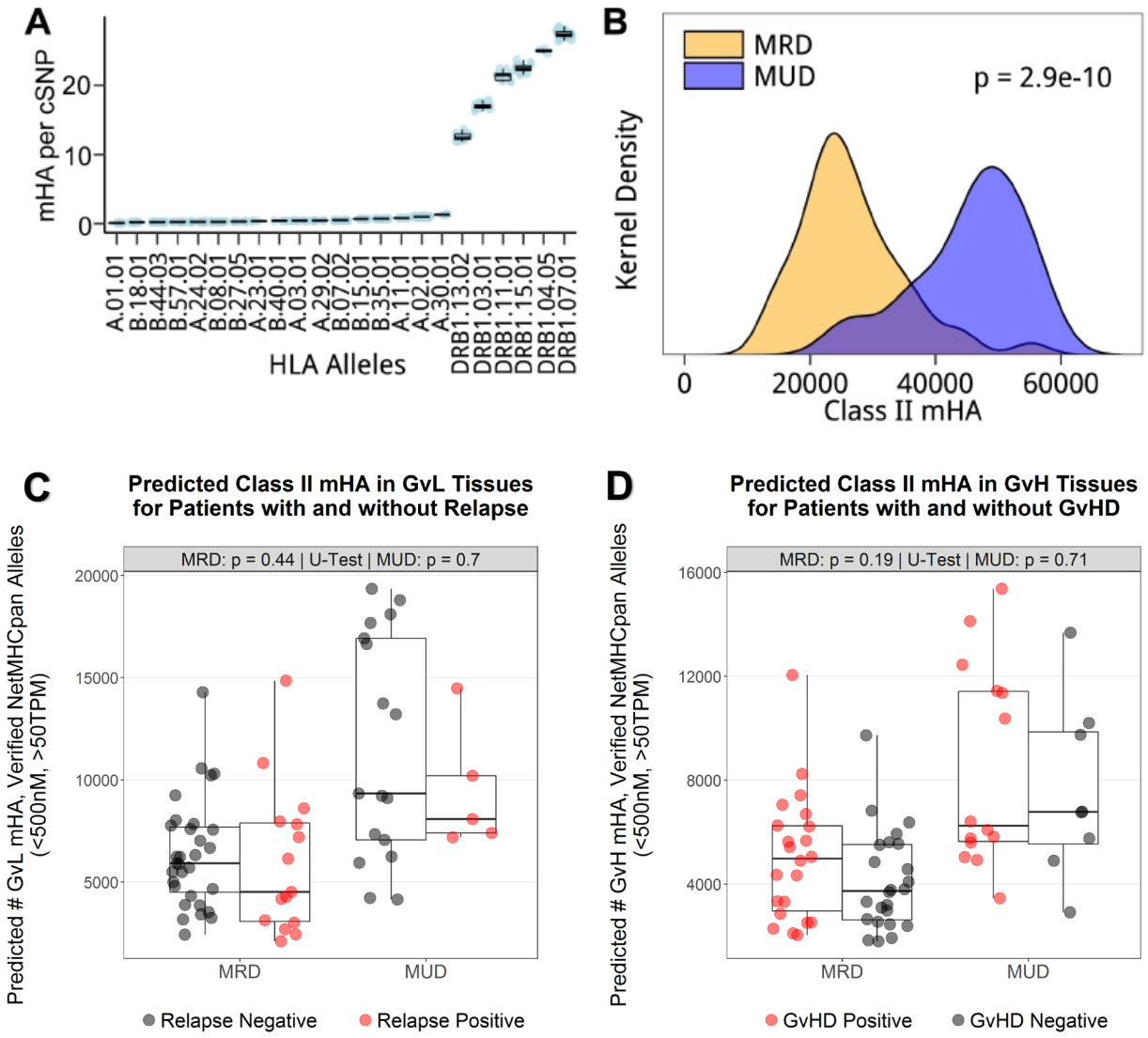
Supplementary Figure 4:



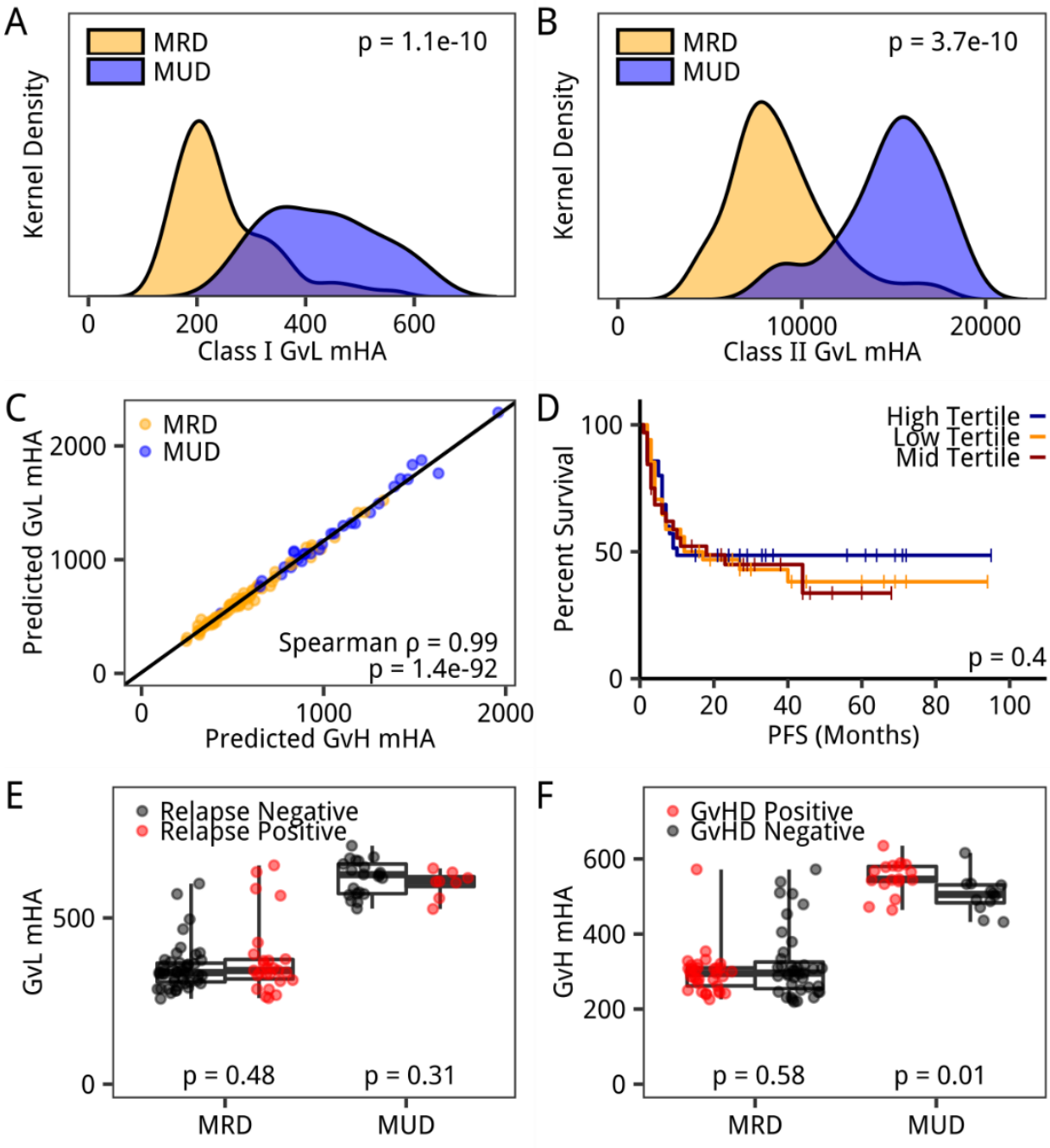
Supplementary Figure 5:



Supplementary Figure 6:



Supplementary Figure 7:



Supplementary Figure 8: