## Supplemental Note 1 – Grouping Parameters for MaxQuant and SpectrumMill PSMs Data

### MaxQuant:

Andromeda PSM results were compiled at 1% FDR, and these PSM qValues were used in IDGroup definition with default gpGrouper thresholds. PSMs results that are filtered at 1% PSM FDR will results in only odd numbered IDGroup bins delineated further by spectrum match scores. While Andromeda and Mascot use a similar, positively correlated scoring mechanism for indicating the quality of spectral match (1), the Andromeda and Mascot scores lie on different scales. To determine concrete values for Andromeda score thresholds we adjusted the default IonScore bin cutoffs used by gpGrouper (10, 20, 30) such that the same fraction of PSMs exists below the cutoffs under each search method. For this, we first searched proteomic profiling of 100% human and a 50:50 hs:mm mixture searched through Mascot+Percolator and through Andromeda via MaxQuant. The conversion was performed by first finding the fraction of PSMs under the default Mascot IonScore cutoffs and then by finding the corresponding Andromeda Score that allowed for the same fraction of PSMs. For the human sample with 80,077 PSMs identified at 1% FDR, the percentage of PSMs under the Mascot IonScores of 10, 20, 30 are 6.42%, 21.13%, and 40.68%. This corresponds to Andromeda IonScores of approximately 63, 83, and 101 using the 102,313 identified PSMs at the same FDR level. For the human-mouse mixed sample with 101,091 PSMs identified via Mascot and 94,503 PSMs identified from Andromeda, the corresponding Andromeda bins were found to be 65, 86, and 105. The thresholds are very similar and we suggest that 65, 85, and 105 can be used to calculate IDGroup bind in Andromeda-based searches.

## Spectrum Mill:

The Spectrum Mill protein output was obtained from Huang et al. study through the CPTAC portal (2, also as reference #24 in the main text). As for the MaxQuant cases described above, we calculated the Spectrum Mill PSM score thresholds by finding the scores at the same percentiles as the Mascot Ion Score thresholds of 10, 20, and 30. The raw files corresponding to WHIM11, WHIM04, and WHIM12 were downloaded from https://cptac-data-portal.georgetown.edu/cptacPublic and searched with the PD/Mascot pipeline described above. After filtering for Rank 1 PSMs with a q-value ≤ .05, the percentiles were calculated for each of the three Ion Score thresholds 10, 20, 30 (4, 24, and 51%) corresponding Spectrum Mill Scores (7.5, 9.9, 12.6) and used in gpGrouper. These thresholds were used for all WHIM iTRAQ set to calculate IDGroup and then SRA bins for iTRAQ-gpG results.

To approximate qValues in the Spectrum Mill PSMs output, we subtracted the deltaForwardReverseScore from each PSM score, removing any zeros (which indicate a lack of a decoy match). The qValue for each PSM was then calculated from Gaussian Kernel Density functions for the decoy and PSM match scores. Based on these models, the majority of the target peptides have acceptable values with 89.2% of peptides ≤ 5% FDR. These calculated qValues were used under default numerical thresholds for IDGroup bins.

## Supplemental Note 2 - Comparison with MaxQuant grouping and Quantification Mechanism.

As MaxQuant is a widely used label-free protein identification and quantification tool for MS-based proteomics (3, 4), we decided to benchmark gpGrouper against MaxQuant's inference and quantification algorithm. To highlight the handling of multitaxa species, we used 100% human and equal human/mouse mixtures for this comparison. Specifically, we compared our binning SRA quality mechanism with 1% protein FDR, and quantification based on the "winner-take-all" parsimonious razor distribution of peptides by MaxQuant with the distribution of shared peptide quantities by gpGrouper.

By comparing MaxQuant and gpGrouper results obtained from the same Andromeda search, we find that the Strict/Relaxed/All SRA quality metric serves as an acceptable alternative to protein-level FDR. First, we converted the MaxQuant protein-level output to gene level by mapping protein GIs to gene identifiers. For the human sample, over 97% (5841/5983) of the protein groups mapped to a single GeneID despite of the fact that many protein groups consist of more than one protein (Supplementary Figure 4-A,B). For the human/mouse sample, 77% (6590/8550) of the protein groups mapped to a single GeneID, with the increase in the number of GeneIDs per protein group largely attributable to a split between a human and mouse GeneID (Supplemental Figure 4-C,D). In Supplemental Figure 4-E we see the gene-level overlap between MaxQuant and gpGrouper results. While gpGrouper lists all gene assignment possibilities, by omitting “All” group proteins, which consists of mostly IDSet3 genes (parsimonious subsets) and some proteins with low scoring spectra for unique peptides sets, we see a reduction in the number of proteins identified exclusively by gpGrouper. At the “Relaxed” level for a single taxon, the number of genes reported by only MaxQuant (158) or only gpGrouper (180) is similar, while at the “Strict” level MaxQuant exclusively reports more genes (348) than gpGrouper (81). These genes do not meet the “Relaxed” or “Strict” criteria for gpGrouper due to the absence of high quality unique-to-gene peptides (Supplemental Figure 4-I), while genes with high quality unique-2-gene peptides (PeptideCount_u2g_S) as identified by gpGrouper are only identified in MaxQuant at a protein FDR above 1%. In fact, genes preserved by gpGrouper but not MaxQuant tend to have a lower total number of peptides than genes preserved by MaxQuant, but are enriched for unique-to-gene peptides of high quality (Supplemental Figure 4-J). This demonstrates how gpGrouper preserves identifications that can be excluded by more traditional parsimony methods by taking search engine score into account. Even with this consideration, at the strict level gpGrouper is generally more conservative than 1% protein FDR.

The discrepancy is larger in the human/mouse mixture sample, with many more genes present exclusively in the MaxQuant results than exclusively in the gpGrouper results in both the relaxed (1320 vs 188) and strict (2069 vs 111) levels. Genes are excluded from gpGrouper results at these thresholds for the same reasons as before – a lack of unique-2-gene peptides of high quality and genes exclusive to gpGrouper which do not fall below protein FDR of 1% in spite of possessing unique-2-gene peptides with high Andromeda scores (Supplemental Figure 4-K,L).

Next, we compared the quantitation between gpGrouper and MaxQuant. We examined gene products that were designated as IDSet1 and were the sole member of a given protein group in MaxQuant. For the human sample with 5,533 examined gene products, the two grouping algorithms show very strong positive correlation in gene product quantities with a Pearson correlation of 0.99 while the human/mouse mixed sample with 6,981 examined gene products also showed a strong positive correlation with a Pearson correlation of 0.93, but many gene products have significantly higher quantified values by gpGrouper with respect to MaxQuant as seen by the deviation from the identity line (Supplemental Figure 4-F). These gene products that deviate from the identity line have a lower proportion of razor+unique peptides, which are used by MaxQuant to quantify proteins. The deviation in quantification between gpGrouper and MaxQuant increases as the proportion of razor peptides within a gene group decreases with a Pearson correlation of -0.81 (Supplemental Figure 4-G). It is clear that as the proportion of razor peptides for a given gene product decreases, the difference between areas shifts with higher values reported by gpGrouper. With the winner-take-all assignment of peptide area that MaxQuant uses when razor peptides are included for quantification, gene product area is highly influenced by the number of razor peptides. This is in contrast to gpGrouper, by which a gene product may have relatively high area despite the fact that it has many shared (but not designated razor by MaxQuant) peptides. This phenomenon can similarly be observed for the gene products that are composed of 100% Razor peptides as reported by MaxQuant but contain shared peptides; as the fraction of gpGrouper unique to gene peptides decreases, the difference between areas shifts with higher values reported by MaxQuant (Supplemental Figure N5-H). For these gene products, gpGrouper is assigning a proportion of these shared peptide peak areas to the protein while MaxQuant is assigning the entire peptide peak area. Similar but less pronounced trends can be seen in the human sample comparison (Supplemental Figure 4-M,N).

## Supplemental Note 3 – On Characterization of Shared Peptides in Protein and Gene-Centric Databases.

We explored the impact of shared peptides on single taxon and mixed human and mouse proteomes. The shared peptides within a mouse or human peptidome effect a substantial portion of proteins. Using the human proteome *in silico* peptidome from 71,315 unique UniProt protein sequences (Swiss-Prot + TrEMBL, downloaded on July 21, 2017), we calculate that 45.02% of fully cut tryptic peptides (276,028 out of 613,080 potential peptides in the 7 amino acid to 10kDa range) are shared in 2 or more proteins, with 90.29% of proteins affected by at least 1, and 83.86% by 2 or more shared peptides. In NCBI RefSeq (downloaded on June 24, 2017) with 110,382 records and 79,788 unique proteins sequences, these numbers are 70.52%, 94.53%, and 93.52% (Supplemental Figure 5-A). However, the majority of protein ambiguity in empirical data comes from completely indistinguishable protein isoforms of the same gene due to partial – and often minimal – coverage of protein sequences in all bottom-up experiments. This observation alone suggests that gene-level reporting may in fact be a more fair representation of empirical profiling data. It is also a prudent option for the emerging wealth of proteogenomics projects, where information flows between proteomic, genomic and transcriptomic analysis, and selection of representative isoforms among indistinguishable gene-specific proteins before comparison may artificially increase the difference between multi-omic results.

Once we convert the RefSeq to gene-unique FASTA list (RefProtDB, Methods), the theoretical fraction of shared peptides in human is reduced to 2.98%, with 36.37% and 25.50% of gene products containing at least 1 and 2 shared peptides, respectively (Supplemental Figure 5-A). The issue is still exacerbated in multi-taxa samples, particularly mouse-based PDXs, where the peptidomes of the combined species are similar (Supplemental Figure 5-B). A combined human and mouse RefProtDB peptidome with 1,015,856 peptides would theoretically have 97.39% of proteins and 85.36% of gene products affected by at least 1 shared peptide (Supplemental Figure 5-D).

We then evaluated the shared peptide issue in empirical data. For this, we used proteomic profiling experiments of single and mixed human/mouse (PDX-like) samples. Specifically, we calculated the effect on counts of shared peptide, affected genes and proteins, and the total peak area (per gene product and overall). In empirical single-taxon profiling data, an average 4.91% of peptides are usually shared between 2 or more genes, affecting 2,454 gene products.

However, in terms of total peak area, the effect is more substantial with 18.47% of total peak are being shared across genes (Supplemental Figure 5-C, Human Gene Products Empirical). This is expected as many shared peptides actually originate from different gene products, adding to the argument against discrete parsimonious allocation - at least with regard to protein level estimation. In mixed taxa samples, empirical data shows an average of 44.96% (26,439) of peptides as shared and 70.65% area belonging to taxon-ambiguous peptides (n = 3, standard error < 0.05 for each measurement) (Supplemental Figure 5-D, HS MM Gene Products Empirical). Unsurprisingly, this effect is more pronounced for MS1 peak area than spectral counts, of which the latter is expected to saturate faster than the former (Supplemental Figure 5-E).

**Supplemental Note 4 - On Definition of the Largest Sequence Coverage.**

For proteins identified by a fully shared pool of peptides, parsimony is often used to choose a master assignment based on the largest sequence coverage. The largest sequence coverage is frequently defined as percent of total FASTA that is covered by found peptides.

Here we reasoned that the largest number of distinct peptide sequences, sans modifications and charge distributions, is the more appropriate definition of the "largest" identification set for mass spectrometry data. This is distinct from the notion that the protein with the highest percent coverage is the most likely observed protein. Therefore, in the case where two proteins are present with an equal number of shared peptides with no uniques, both proteins are reported as equally likely regardless of percent coverage. However, a situation where smaller percent of sequence is represented by larger number of MS peptides can arise. For MS-based identifications, more sequence hits provide more certainty than coverage.

**Supplemental Note 5 - On Two Cases of IDSet2 GPGroup Assignments.**

IDSet 2 assignments follow into two categories.

The first trivial case is an identical set of peptides that maps to multiple gene loci; this results in one GPGroup with multiple indistinguishable GeneID assignments.

The second possibility is a largest gene-specific peptide set, where subsets of peptides map to multiple other genes; this results in a GPGroup with one GeneID assignment but, importantly, no unique-to-gene peptides.

**Supplemental Note 6 - On Distribution of PSM and Peptide Counts.**

PSMs and peptide counts are not distributed for the sequences that map to multiple gene loci. The reason we opted for counting maximum number of PSMs and peptides per gene are three-fold:

(1) These are technically identities, and identity is strictly speaking not dividable. We feel that reporting a total number of identities belonging to the locus gives a more clear picture of MS evidence.

(2) A quantity distribution method is implemented on peptide peak AUC level, where it is appropriate.

(3) Because the distribution procedure is implemented on the peptide peak AUC level, we did not see a need to distribute either PSMs or peptide counts that are semi-quantitative in nature.

## Supplemental Note 7 - On the Distribution and Aggregation of Peak Areas.

gpGrouper implements three fundamental types of gene-level summed areas: (1) maximum assignable area; (2) summed unique-to-gene peptide only area; and (3) distributed area based on the ratios of unique peptides. If there is evidence of unique to gene peptides in multiple loci that also share peptides, the most likely scenario is that the shared peptide amount is a summed amount from multiple loci. Therefore, the actual amount of protein gene product is somewhere between unique-to-gene and maximum area values, with distributed area being the most reasonable estimate we can provide.

Alternative implementations of “top” peptide sums or averages have also been used in the field (5–7). The current version of gpGrouper does not have these options; however, they could be calculated from existing rank annotations in PSMs table. The top-to-bottom ranking is done by gpGrouper, and the top n peptides can be easily found by their sequence ranks (in the PeptRank column).

## Supplemental Note 8 – On the Composition of Species-Indistinguishable Gene Products

A question arose regarding the composition of species-indistinguishable gene products across the human/mouse mixtures. Namely, whether the high level of species-shared peptide area (~70%) is primarily due to highly conserved and expressed housekeeping genes. At the gene level, 30% of shared peptide area is fully indistinguishable between and mouse, with no corresponding unique-to-gene peptides to aid in redistributing the area across genes and species. This 30% of total peptide area cannot be allocated to either species and gpGrouper splits it by taxon ratio. However, we consistently see an additional roughly 40% of shared peptide area that does have corresponding unique-to-human or mouse genes. This area is split as in single species, via unique-to-gene peptide area ratios.

We further examined these two groups of gene products through GO enrichment via the R packages, DOSE and clusterProfiler (8, 9). First we examined human gene products that cannot be discerned between species; selecting those that consistently appear in the mixtures (12/15 runs) and in the human mixture run independently. We find strong enrichment for chromosome and ribosome associated genes – typically associated as housekeeping – and exhibit high homology between mouse and human. We performed this same analysis on those gene products that are species-discernable without finding similar enrichment. Thus, while many genes can be classified as house-keeping, a substantial proportion a substantial proportion of species-discernable genes with various roles are also present.

# References

1. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. a., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805

2. Huang, K.-L., Li, S., Mertins, P., Cao, S., Gunawardena, H. P., Ruggles, K. V, Mani, D. R., Clauser, K. R., Tanioka, M., Usary, J., Kavuri, S. M., Xie, L., Yoon, C., Qiao, J. W., Wrobel, J., Wyczalkowski, M. A., Erdmann-Gilmore, P., Snider, J. E., Hoog, J., Singh, P., Niu, B., Guo, Z., Sun, S. Q., Sanati, S., Kawaler, E., Wang, X., Scott, A., Ye, K., McLellan, M. D., Wendl, M. C., Malovannaya, A., Held, J. M., Gillette, M. A., Fenyö, D., Kinsinger, C. R., Mesri, M., Rodriguez, H., Davies, S. R., Perou, C. M., Ma, C., Townsend, R. R., Chen, X., Carr, S. A., Ellis, M. J., and Ding, L. (2017) Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. *Nat Commun*. **8**, 14864

3. Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J. V., and Mann, M. (2009) A practical guide to the maxquant computational platform for silac-based quantitative proteomics. *Nat. Protoc.* **4**, 698–705

4. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc*. **11**, 2301–2319

5. Braisted, J. C., Kuntumalla, S., Vogel, C., Marcotte, E. M., Rodrigues, A. R., Wang, R., Huang, S.-T., Ferlanti, E. S., Saeed, A. I., Fleischmann, R.

D., Peterson, S. N., and Pieper, R. (2008) The APEX Quantitative Proteomics Tool: Generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics*. **9**, 529

6. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Mol. Cell. Proteomics*. **4**, 1265–1272

7. Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C., and Geromanos, S. J. (2006) Absolute Quantification of Proteins by LCMS $^{E}$. *Mol. Cell. Proteomics*. **5**, 144–156

8. Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. **31**, 608–609

9. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. **16**, 284–7

## Supplemental Figure Legends

**Supplemental Figure 1. Overview of the Input Data and Processing Workflow for gpGrouper Algorithm.**

**A (Input).** A tab separated PSMs file and a fasta file are provided as inputs. The PSMs file can originate from any search engine as long as it contains the necessary information of Sequence, Spectrum Score, FDR q-value, MS1 Peak Area, and MS2 Reporter Ion Intensities (if applicable). The input fasta file critically contains annotations for the GeneID and TaxonID for each protein sequence. **B (Process).** Each PSM record is mapped to and split among all potential GeneIDs. If the experiment is isotopic or isobaric, each PSM is further split on their isotopic/isobaric labels. If the experiment has gene products from more than one taxon, an estimate of the proportion of each taxon is made. For each gene product, if unique peptides are present then shared peptides are distributed by unique peptide ratios. In the absence of unique peptides, all shared peptides are first distributed across species based on the estimated taxon ratios (if applicable), and then divided equally across all species-specific genes. **C (Output).** An annotated PSMs (psms table) and gene product table (experiment-2-gene, “e2g” table) are provided as output.

**Supplemental Figure 2. Key parameters in gpGrouper PSMs output tables.**

Superscripts:

[1]user specified integers, or unspecified = 1; [2]True=1, False=0; [3] integer flags for label free = 0; SILAC (+6) = 1; iTRAQ = 114, 115, 116, or 117; TMT = 126 (for 126C), 1270 (for 127C), 1271 (for 127N), 1280 (for 128C), 1281 (for 128N), 1290 (for 129C), 1291 (for 129N), 1300 (for 130C), 1301 (for 130N), or 131 (for 131C).

**Supplemental Figure 3. Key parameters in gpGrouper experiment-to-gene (e2g) output tables.**

**Supplemental Figure 4. Panels A-D. Characterization of the Protein Groups as Reported By MaxQuant. A/B.** Number of GeneIDs per protein group for human (**A**) and human/mouse (**B**; hs:mm) 1:1 mixed samples. Majority of proteins in each protein group map to a single GeneID in the human data. The rise in the number of multiple GeneIDs per protein group in the hs:mm data is driven by proteins that are indistinguishable between species. **C/D.** Relationship between the number of GeneIDs and proteins within a given protein group for human (**C**) and hs:mm (**D**) samples shows that many protein groups contain multiple proteins that all map to one or few GeneIDs.

**Panels E-H. Comparing The Identification and Quantification of Gene Products After Grouping by gpGrouper and MaxQuant from Human and 1:1 Human Mouse Mixture Samples. E.** Overlap of gene products identified in gpGrouper and MaxQuant at three filtering levels on the gpGrouper data and 1% protein FDR on the MaxQuant data for human and 1:1 human mouse mixture samples. Strict : IDGroup_u2g ≤ 3 and IDSet < 3 (or IDGroup ≤ 3 if no unique peptides), Relaxed : IDGroup_u2g ≤ 5 and IDSet < 3 (or IDGroup ≤ 5 if no unique peptides), All : all potential gene products, including IDSet 3. **F.** Correlation plots of gene products identified in both gpGrouper and MaxQuant for human (left) and hs:mm (right) samples. Gene products were filtered to those that are IDSet 1 as classified by gpGrouper and is only present in one MaxQuant protein group. **G.** A negative correlation between the fraction of peptides in a protein group that are "Razor" as reported by MaxQuant and the log10 transformed gene quantification ratio in the human:mouse mixture sample is observed. **H.** For gene products with 100% "Razor" peptides, a positive correlation between the number of unique peptides as identified by gpGrouper and the log10 transformed gene quantificaiton ratio in the human:mouse mixture sample is observed.

**Supplemental Figure 4, Panels I-J. Characterization of Non-Overlapping GeneIds Between gpGrouper Strict and Relaxed Levels and MaxQuant for Human Profiling Data. I.** Barplots show a lack of strict unique-to-gene peptides for gene products reported by MaxQuant only (left) while gene products reported by gpGrouper have fewer but higher quality peptides as indicated by strict unique-to-gene counts. **J.** Radviz projection of the non-overlapping GeneIDs showing that gpGrouper reports gene products enriched for strict unique-to-gene while MaxQuant reports gene products with higher total number of PSMs.

**Supplemental Figure 4, Panels K-L. Characterization of Non-Overlapping GeneIDs Between gpGrouper Strict and Relaxed Levels and MaxQuant for HS:MM Mixture Profiling Data. K.** Barplots show a lack of strict-unique-to-gene peptides for gene products reported by MaxQuant only (left) while gene products reported by gpGrouper have few peptides but of high quality as indicated by strict unique-to-gene counts. **L.** Radviz projection of the non-overlapping GeneIDs showing that gpGrouper reports gene products enriched for strict unique-to-gene while MaxQuant reports gene products with higher total number of PSMs.

**Supplemental Figure 4, Panels M-N. Comparing the Quantification of Gene Products in the Human Profiling Data After Grouping by gpGrouper and MaxQuant. M.** A negative correlation between the fraction of peptides in a protein group that are "Razor" as reported by MaxQuant and the log10 transformed gene quantification ratio in the human:mouse mixture sample is observed. **N.** For gene products with 100% "Razor" peptides, a slight positive correlation between the number of unique peptides as identified by gpGrouper and the log10 transformed gene quantitation ratio is observed.

**Supplemental Figure 5. Characterization of *In Silico* Trypsin/P Digestion of Proteomes Reported by Different Databases.** SwissProt+TrEMBL, NCBI RefSeq, RefSeq after mapping to GeneIDs, and empirical results from profiling data searched and grouped against NCBI RefSeq after mapped to GeneIDs. **A/B.** Number of protein or GeneID that each peptide maps to in human (**A**) and concatenated human/mouse (**B**) databases. **C/D.** For each identifier, the number of peptides shared across multiple proteins or GeneIDs for human (**C**) and concatenated human/mouse (**D**) databases. **E.** Emprical data for total Spectral Counts and AUC for each 1 or more gene in human (left) and human/mouse (right) mixtures. n = 3 technical replicates.

**Supplemental Figure 6. Comparison of Alternative Methods for Handling Species-shared Peptide Peaks for 50% Human Sample to Expected Values via the 100% Human Sample.** (**A**) Assume all species shared peptides belong to human. (**B**) Razor peptide peak assignment by which shared peptide quantities are assigned to the protein with the largest number of unique peptides. (**C**) Random distribution of shared peptide peaks areas.

**Supplemental Figure 7. Metrics for Each Species for all Human/mouse Cell Mixture Data.** Number of Gene Products (**A**) and Gene Product Groups (**B**) that are uniquely identified to originate from either human or mouse, or shared across species. The number of identified gene products increases with the percentage of each species, while the number of species-indistinguishable gene products remains roughly constant across dilutions. **C,D** Number of gene products identified across all dilutions and between each dilution and the 100% human/mouse sample). **E**. RMSE between the expected and measured values for each quantified human and mouse gene product after filtering for gene products identified and quantified across all samples. **F.** RMSE between the expected values (based on 100% human/mouse samples) and the measured values for each quantified human and mouse gene product. Note that only gene products that were identified and quantified both within the dilution and the 100% species sample were included in the counts. **G.** Percentage of total MS1 peak area for PSMs that are shared across multiple genes. PSMs that map to genes which also have unique-to-gene peptides are contained within human or mouse categories, while PSMs with no corresponding unique-to-gene nor unique-to-taxon peptide are designated as fully shared. n = 3 technical replicates.

**Supplemental Figure 8. Area Distribution Schematics. A.** A theoretical scenario in which three gene products (GPs) contain both unique and shared peptides. The calculation of the distributed peptide area sum for each gene product comprises of the sum of unique peptides and shared peptides after weighting by the unique peptide ratio. **B.** Flowchart for peptide splitting in mixed-species samples when no unique peptides exist. First, peptides P1 and P2 are split by taxon ratios (here 75% human and 25% mouse). Second, P1 and P2 are split by the number of mapped genes within each species (here 3 for human and 2 for mouse).

**Supplemental Figure 9. Human/Mouse Ratios and Identifications for WHIM PDX Profiles. A.** Estimated species percentages for each WHIM PDX from data generated and analyzed previously by Huang et al. These are the gpGrouper results based on the PSMs table from the Spectrum Mill search performed as described in the original publication (PSMs data table available at https://cptac-data-portal.georgetown.edu/cptacPublic/). **B.** Estimated species percentages from label free profiling data for later passages of WHIM PDX generated and analyzed in BCM. These data were searched with Proteome Discoverer/Mascot and compiled by gpGrouper.

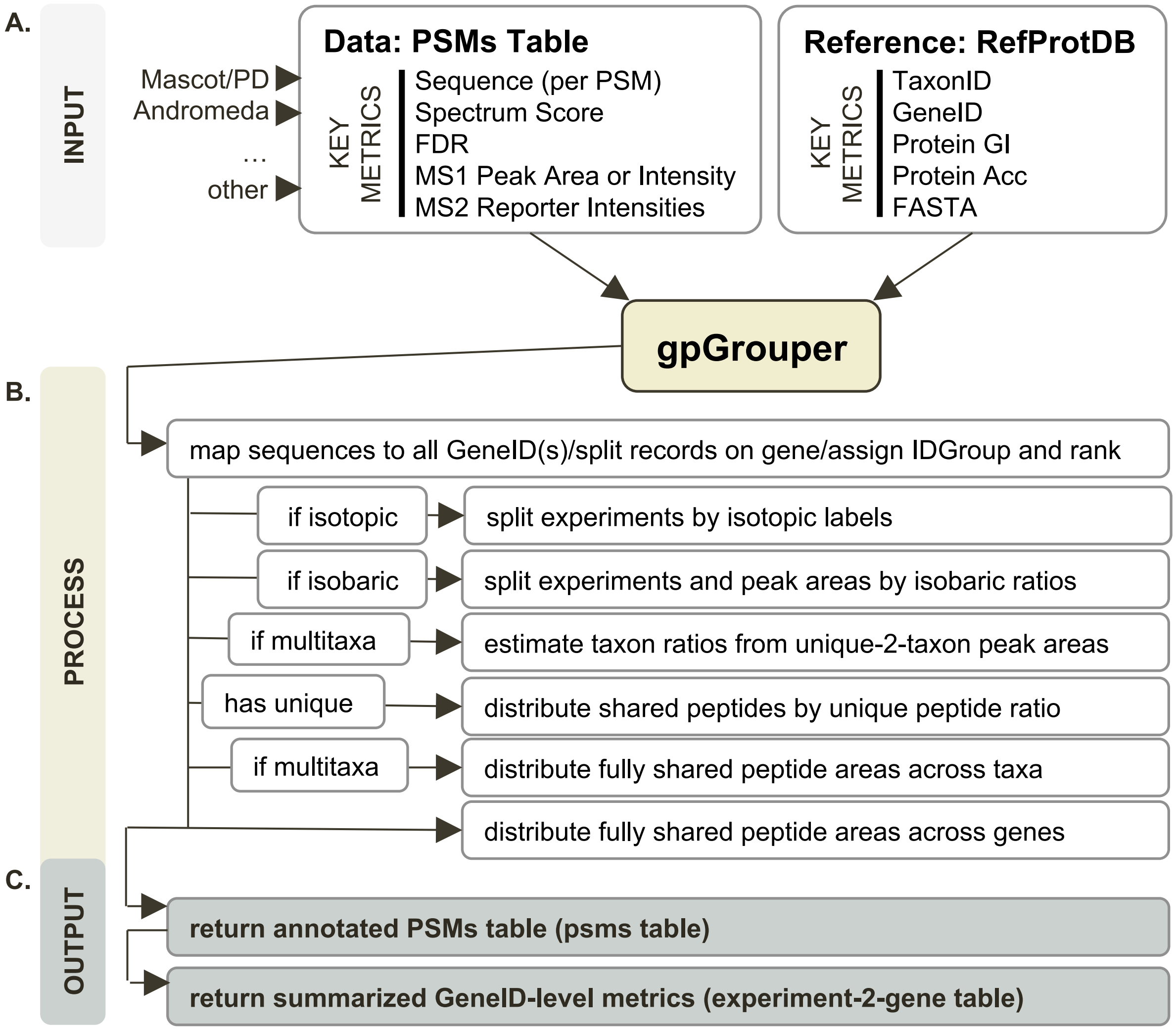
**Supplemental Figure 10. Comparison between Grouping Methods for 4 PDXs of BCM-4913 Tumor with Drastically Different Stromal Contributions.**

Each PDX for Tumor BCM-4913 was grouped using three different approaches for dealing with peptides shared across species: (1) an unbiased approach grouping against a human/mouse concatenated RefSeq and distributing peptide peaks across species as necessary, (2) ignoring species-shared peptides, and (3) assuming all peptides originate from human by grouping with the human RefSeq. Scatterplots of the mutually identified gene products are shown in the bottom panels, with the Pearson correlations and number of gene products in each comparison listed in the upper panels. The diagonal axis labels the grouping approach with the number of quantified human gene products. Full data for each of these grouping methods available in Supplemental Table 6.

**Supplemental Figure 11. Correlation between MS1 Peptide Peak Area and Summed Reporter Ion Intensity for TMT SPS Data Used in Figure 3 of Main Text.**

**Supplemental Figure 12. Human Go Cellular Component enrichment for Species Indistinguishable (A) and Species Discernable (B) gene products consistently observed across human/mouse mixtures.**

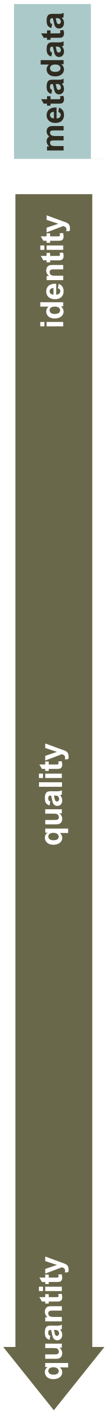
# Supplemental Figure 1: gpGrouper Pipeline


A.
INPUT
Mascot/PD
Andromeda
…
other
Data: PSMs Table
KEY METRICS
Sequence (per PSM)
Spectrum Score
FDR
MS1 Peak Area or Intensity
MS2 Reporter Intensities
Reference: RefProtDB
KEY METRICS
TaxonID
GeneID
Protein GI
Protein Acc
FASTA
gpGrouper
B.
PROCESS
map sequences to all GeneID(s)/split records on gene/assign IDGroup and rank
if isotopic
split experiments by isotopic labels
if isobaric
split experiments and peak areas by isobaric ratios
if multitaxa
estimate taxon ratios from unique-2-taxon peak areas
has unique
distribute shared peptides by unique peptide ratio
if multitaxa
distribute fully shared peptide areas across taxa
distribute fully shared peptide areas across genes
C.
OUTPUT
return annotated PSMs table (psms table)
return summarized GeneID-level metrics (experiment-2-gene table)

# Supplemental Figure 2: PSMs Output Table Features

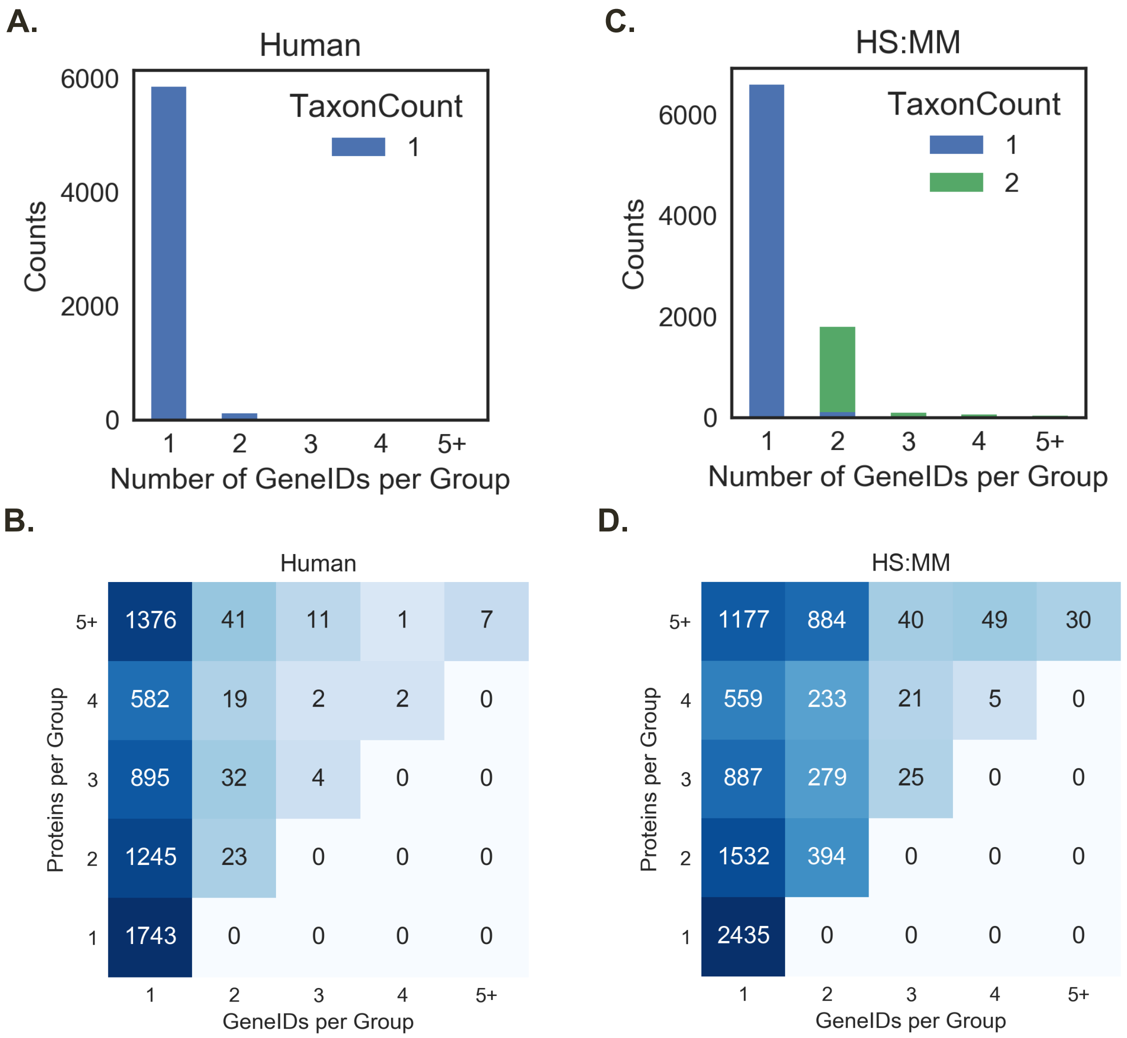
| | | |
|---|---|---|
| EXPRecNo | meta | numerical experiment identifier (defines sample[1]) |
| EXPRunNo | | serial run number (defines mass spectrometry sequencing run[1]) |
| EXPSearchNo | | serial search number (defines search configuration[1]) |
| Sequence | from input | from input: peptide aminoacid sequence |
| PSMAmbiguity | | from input: assignment ambiguity (Default: Unambiguous) |
| Modifications | | from input: list of modifications |
| PrecursorArea | | from input: peak AUC or equivalent intensity |
| q_value | | from input: Percolator q-Value or other equivalent FDR metric |
| PEP | | from input: PEP or other equivalent probability metric |
| IonScore | | from input: Mascot IonScore or other equivalent spectrum score |
| Charge | | from input: precursor ion charge |
| *Reporter Intensities | | from input: isobaric tag reporter intensities |
| oriFLAG | flags | binary flag[2] marking original record rows in input |
| LabelFLAG | | numeric flag[3] marking label type |
| IDG | | IDGroup bin of the PSM |
| SequenceModi | | standardized modification-containing sequence annotation |
| SequenceModiCount | | count of modifications |
| PSM_UseFLAG | | binary flag[2] indicating whether the PSM will be counted for e2g |
| AUC_UseFLAG | | binary flag[2] indicating whether the AUC will be counted for e2g |
| Peak_UseFLAG | | binary flag[2] marking PSM that represents a distinct peak |
| PrecursorArea_split | quantities | precursor are split on reporter ratios |
| SequenceArea | | sum of peak areas for equivalent peptides |
| PrecursorArea_dstrAdj | | distributed SequenceArea of shared peptides |
| PeptRank | | MS1 intensity-based rank of peptides (per GeneID) |
| GeneID | ID mapping | GeneID assignment (only one GeneID) |
| GeneIDs_All | | all GeneIDs in a given reference that the sequence maps to |
| GeneIDCount_All | | value count of GeneIDs_All |
| ProteinGIs | | Protein GIs (per GeneID) that the sequence maps to |
| ProteinGIs_All | | all protein GIs in a given reference that the sequence maps to |
| ProteinGICount_All | | value count of ProteinGIs_All |
| ProteinRefs | | Protein Accessions (per GeneID) that the sequence maps to |
| ProteinRefs_All | | all protein Accessions in a reference that the sequence maps to |
| ProteinRefCount_All | | value count of ProteinRefs_All |
| HIDs | | Homologene IDs |
| HIDCount_All | | value count of HIDs |
| TaxonID | | TaxonID (per GeneID) |
| TaxonIDs_All | | all TaxonIDs in a given reference that the sequence maps to |
| TaxonIDCount_All | | value count of TaxonIDs_All |

# Supplemental Figure 3:
# Experiment-to-gene (e2g) Output Table Features



| | |
|---|---|
| EXPRecNo | numerical experiment identifier |
| EXPRunNo | serial MS sequencing run number |
| EXPSearchNo | serial search configuration number |
| EXPLabelFLAG | numerical identifier for the experiment label channel |
| GeneID | NCBI Gene ID (from RefProtDB) |
| GeneSymbol | NCBI gene symbol (from RefProtDB) |
| GeneDescription | NCBI gene description (from RefProtDB) |
| TaxonID | NCBI taxon identifier (from RefProtDB) |
| PeptidePrint | all identified peptide sequences |
| GPGroup | serial number for distinct gene-based peptide group |
| GPGroups_All | all gene groups for any of the assignable peptides |
| ProteinGIs | all possible protein GIs for any of the assignable peptides |
| ProteinGIs_Count | count of ProteinGIs |
| ProteinRefs | all possible protein accession numbers |
| ProteinRefs_Count | count of ProteinRefs |
| ProteinGI_GIDGroups | list of gene-specific distinguishable protein isoform groups |
| ProteinGI_GIDGroupCount | count of protein isoform groups |
| SRA | Strict/Relaxed/All quality bins |
| IDSet | homology/inference set (unique/indistinguishable/subsets) |
| IDGroup | best peptide quality |
| IDGroup_u2g | best unique-to-gene peptide quality |
| Coverage | average coverage of all gene product isoform (by all peptides) |
| Coverage_u2g | average coverage of isoforms (by unique-to-gene peptides) |
| PSMs | spectral counts |
| PSMs_u2g | unique-to-gene spectral counts |
| PeptideCount | count of distinct peptide sequences |
| PeptideCount_u2g | count of distinct unique-to-gene peptide sequences |
| PeptideCount_S | strict quality distinct peptide sequences |
| PeptideCount_S_u2g | strict quality distinct unique-to-gene peptide sequences |
| AreaSum_u2g_all | sum of all unique-to-gene peptide peak AUCs |
| AreaSum_max | sum of all peptide peak AUCs (shared not distributed) |
| AreaSum_dstrAdj | sum of all peptide peak AUCs (shared distributed) |
| GeneCapacity | theoretical number of tryptic peptide sequences for the gene |
| iBAQ_dstrAdj | iBAQ of AreaSum_dstrAdj |

# Supplemental Figure 4

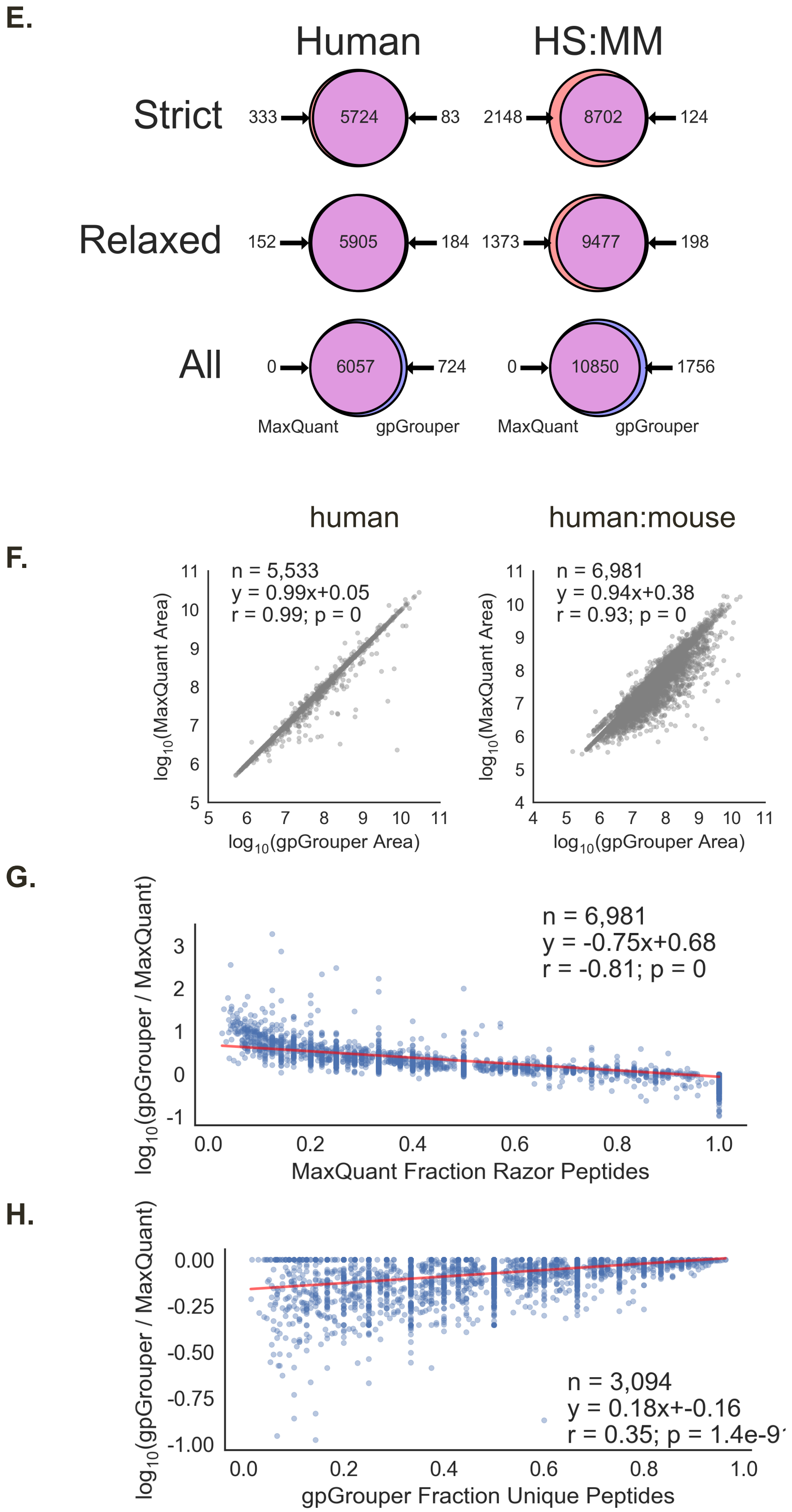


| | 1 | 2 | 3 | 4 | 5+ |
|---|---|---|---|---|---|
| 5+ | 1376 | 41 | 11 | 1 | 7 |
| 4 | 582 | 19 | 2 | 2 | 0 |
| 3 | 895 | 32 | 4 | 0 | 0 |
| 2 | 1245 | 23 | 0 | 0 | 0 |
| 1 | 1743 | 0 | 0 | 0 | 0 |

| | 1 | 2 | 3 | 4 | 5+ |
|---|---|---|---|---|---|
| 5+ | 1177 | 884 | 40 | 49 | 30 |
| 4 | 559 | 233 | 21 | 5 | 0 |
| 3 | 887 | 279 | 25 | 0 | 0 |
| 2 | 1532 | 394 | 0 | 0 | 0 |
| 1 | 2435 | 0 | 0 | 0 | 0 |

continued on next page

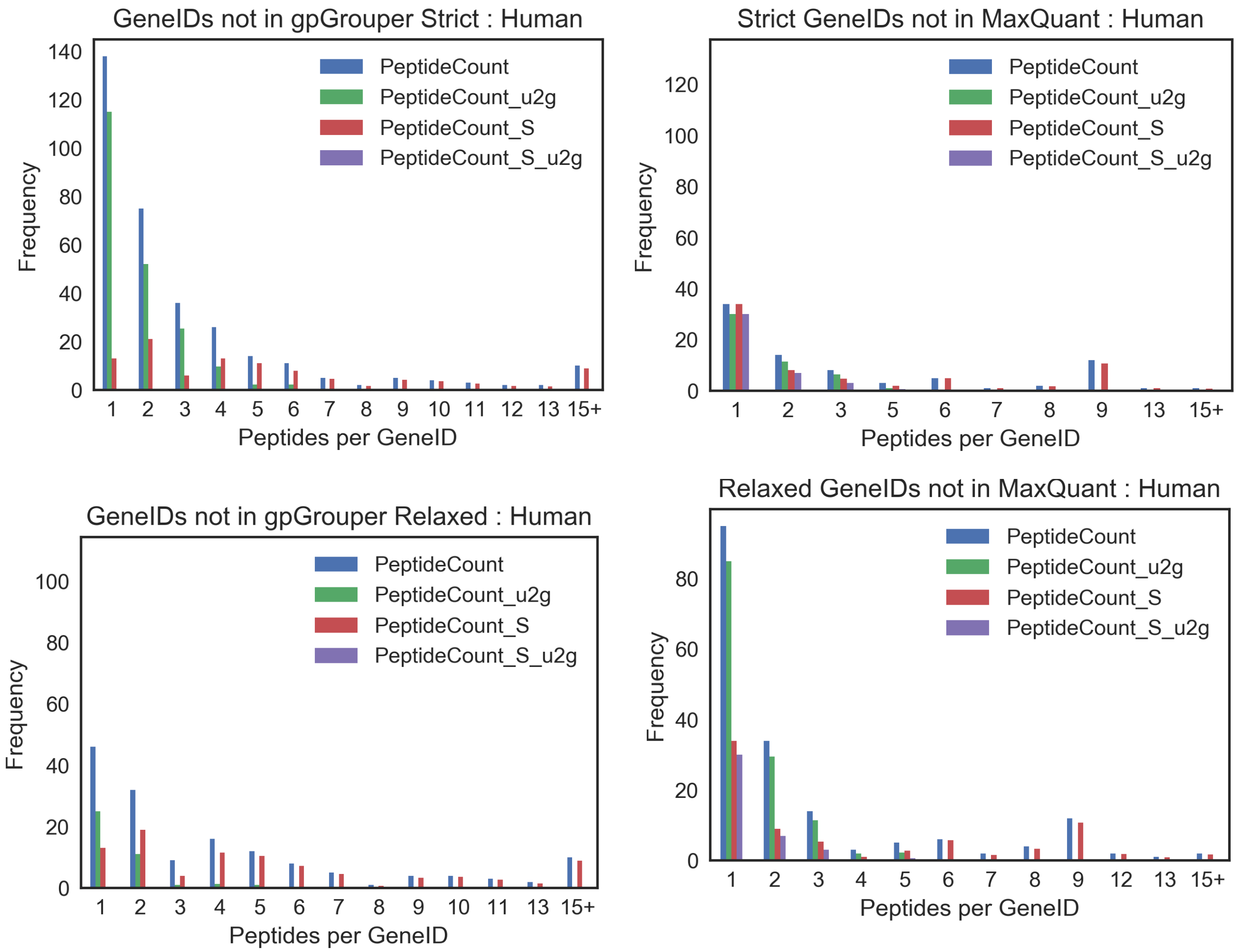
# Supplemental Figure 4



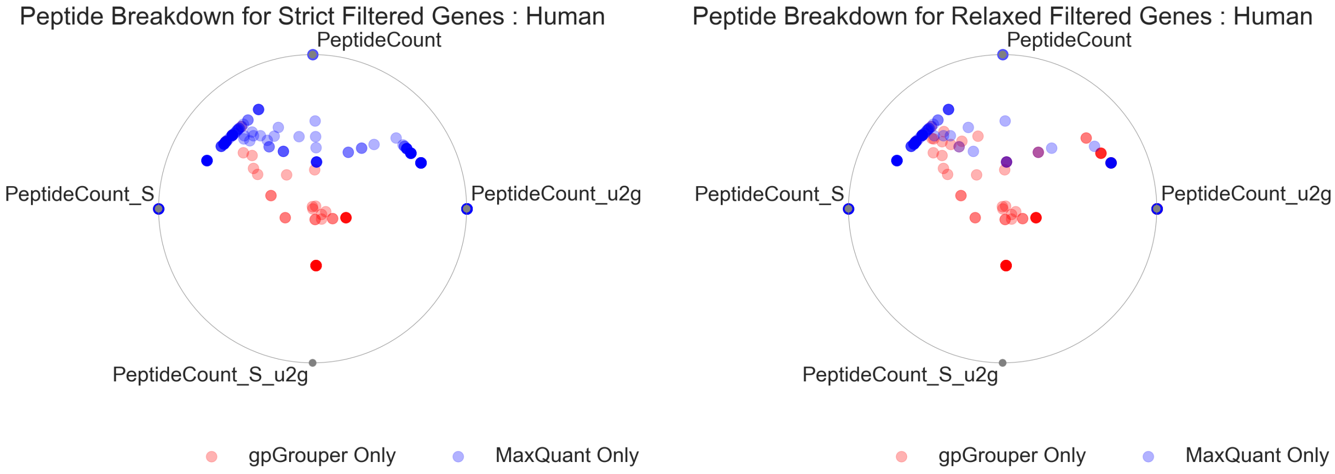
continued on next page

# Supplemental Figure 4

I.

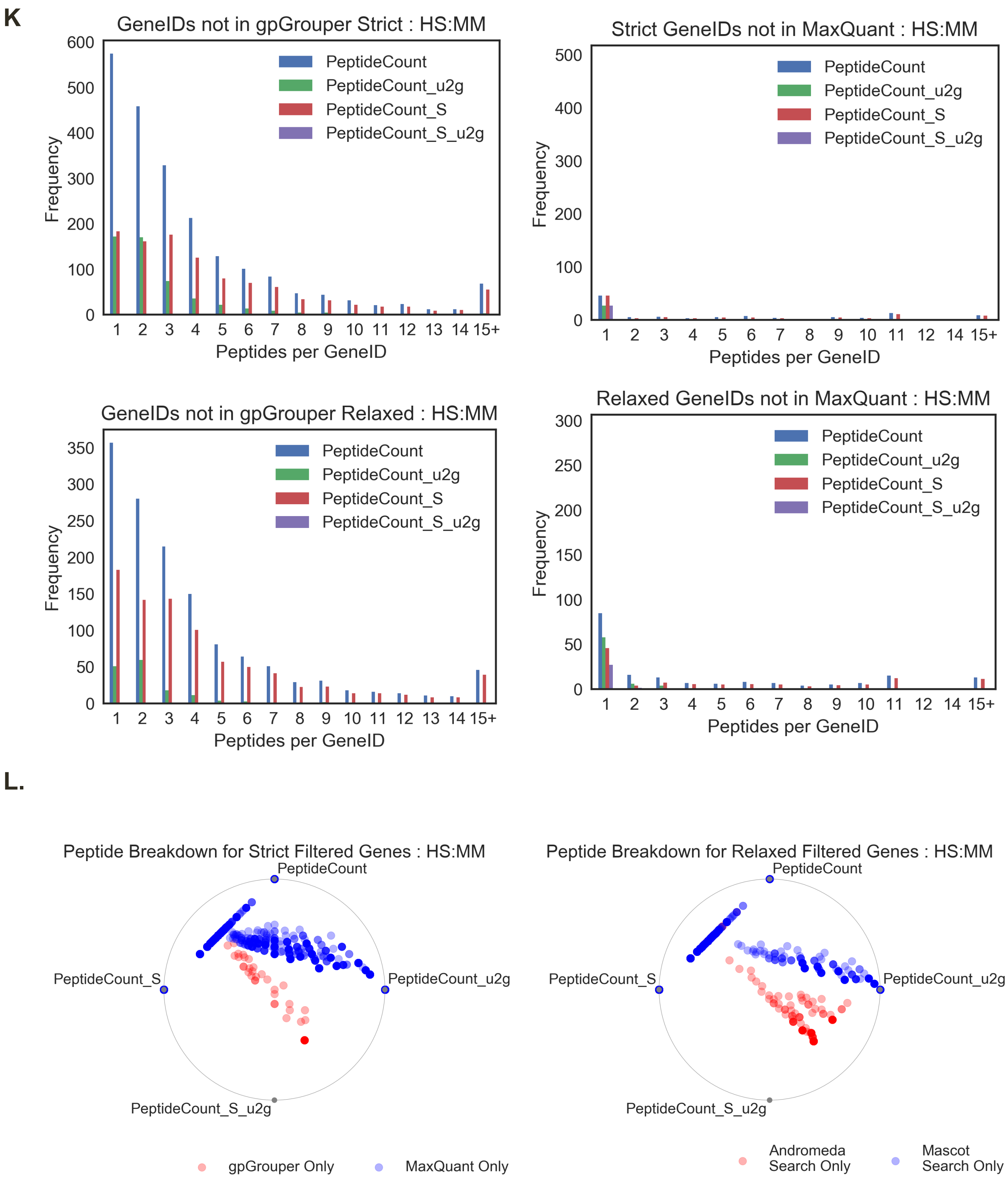


J.



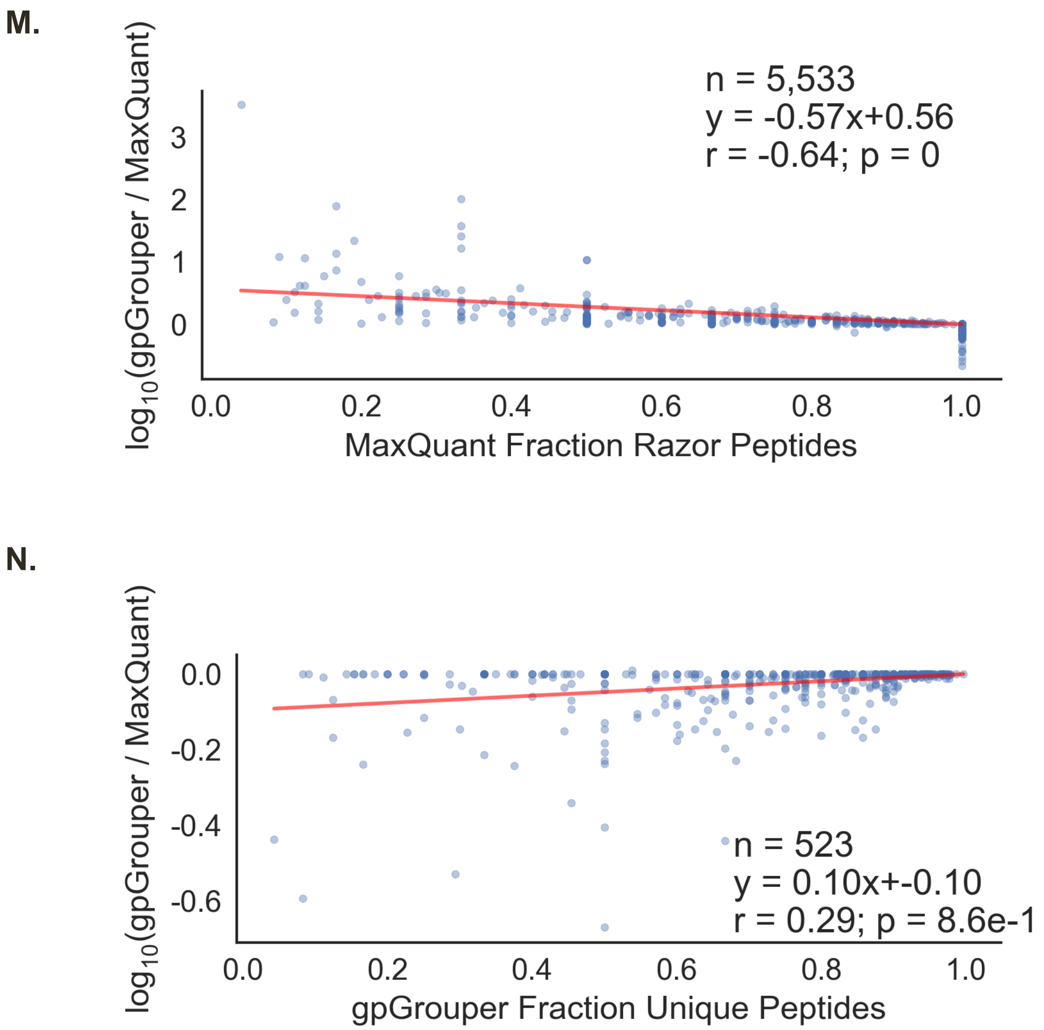
continued on next page

# Supplemental Figure 4



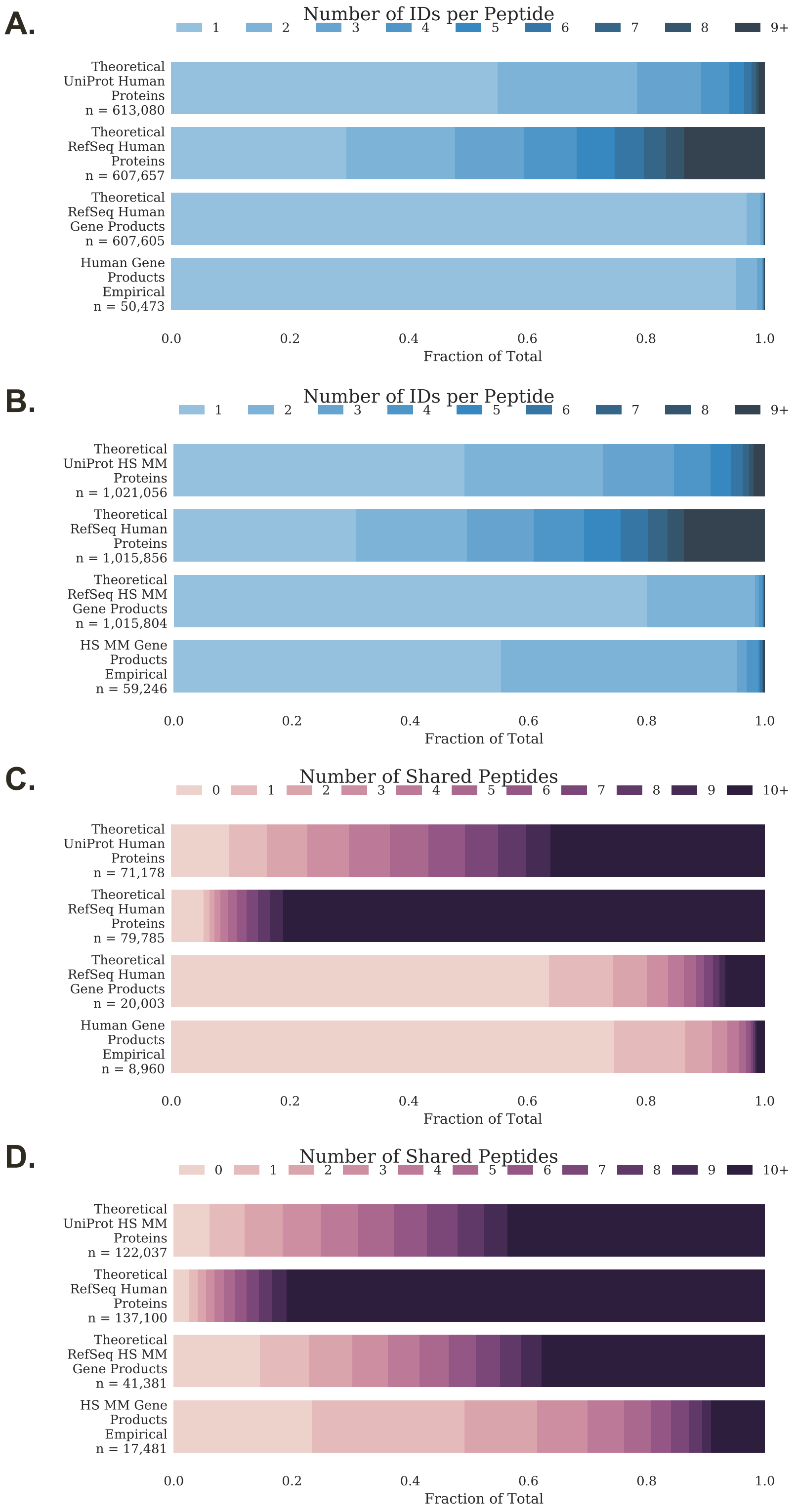
**continued on next page**

# Supplemental Figure 4



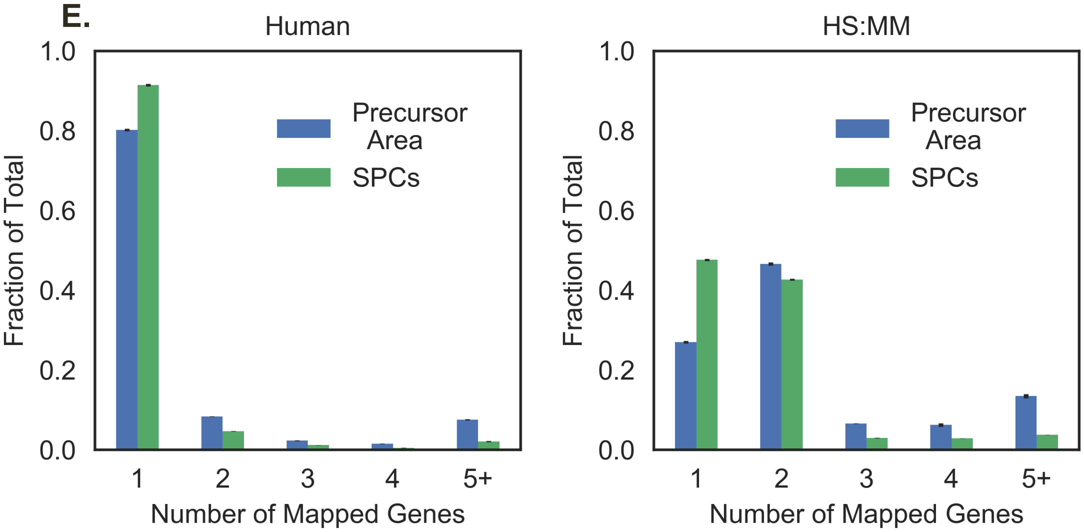
**last page of Supplemental Figure 4**

# Supplemental Figure 5



continued on next page

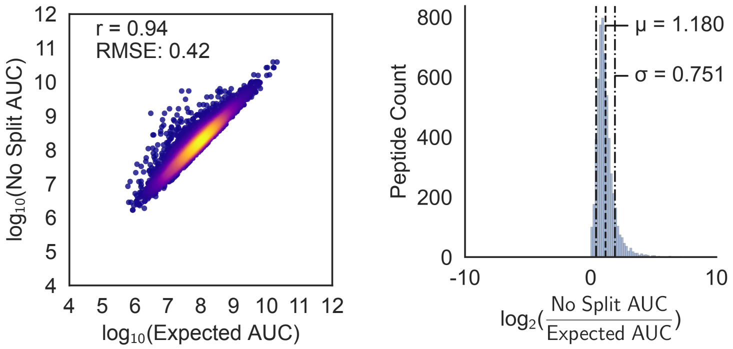
# Supplemental Figure 5



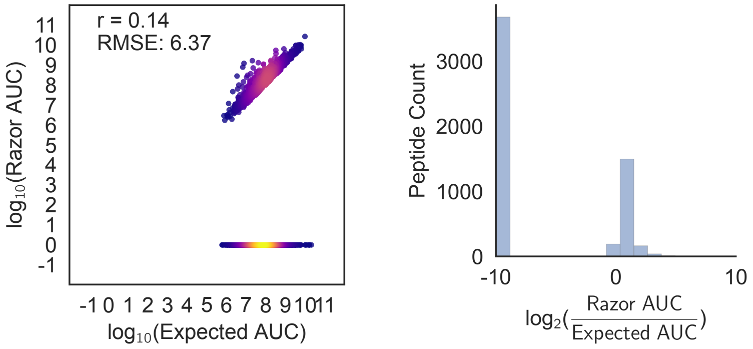
**last page of Supplemental Figure 5**

# Supplemental Figure 6

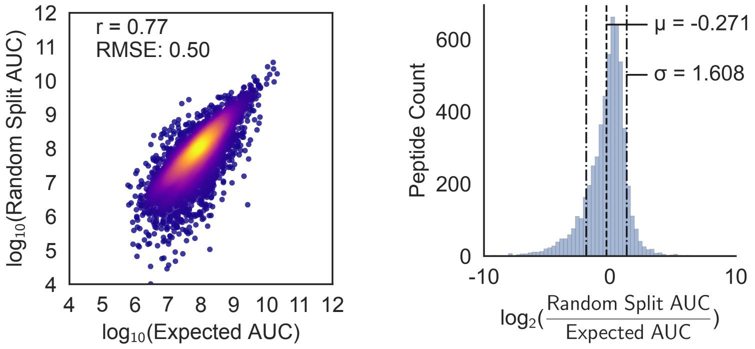
**A.** no split: **estimated AUC** with whole value assigned to human origin



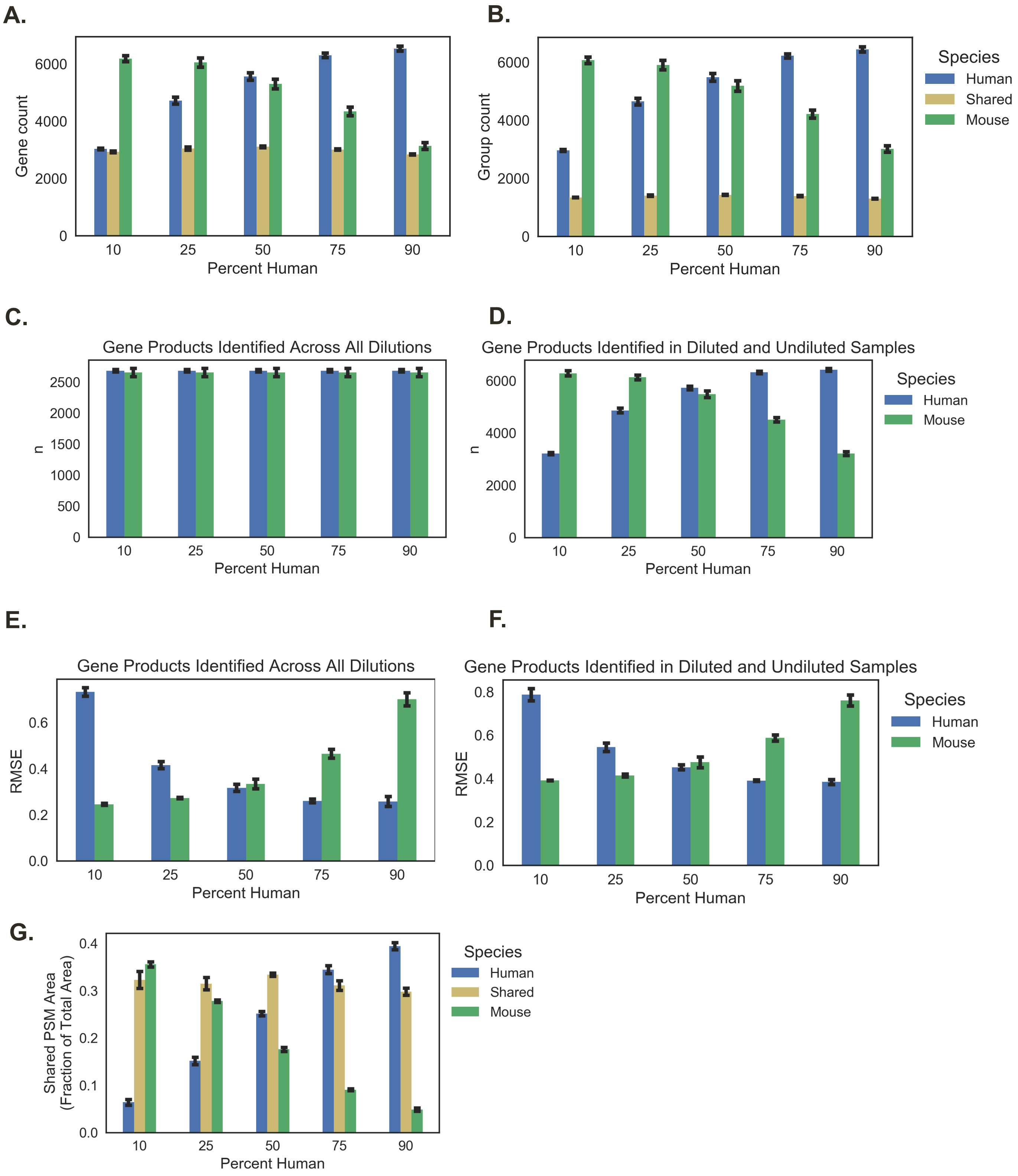
**B.** razor: **estimated AUC** based on razor peptide assignment



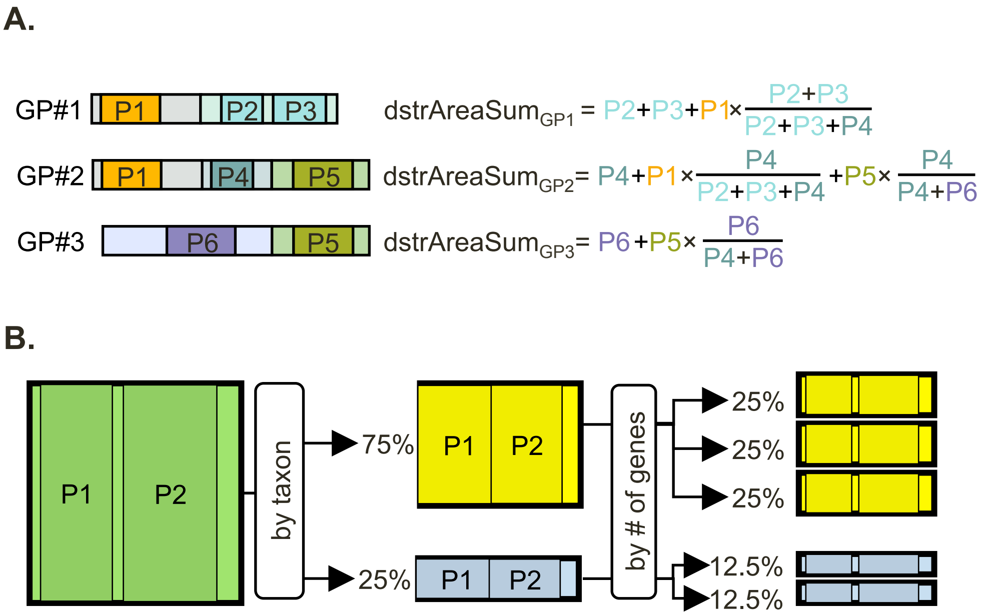
**C.** random: **estimated AUC** by random splitting of peaks (x1000)



# Supplemental Figure 7



# Supplemental Figure 8

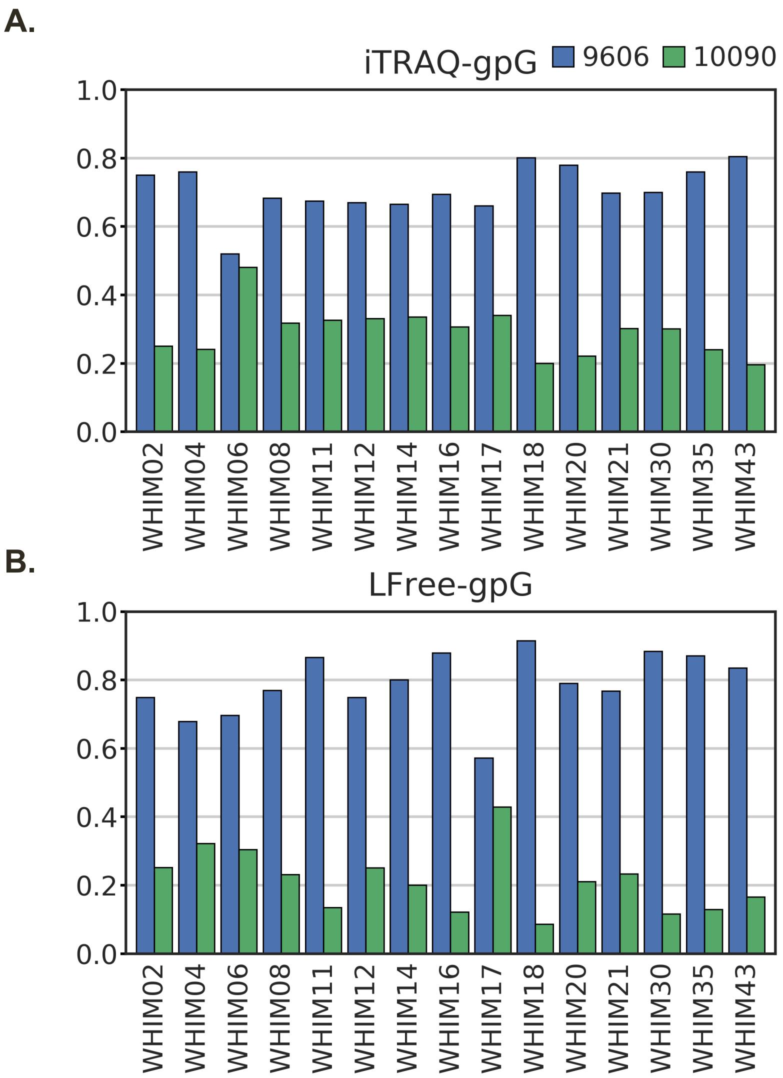


$$\mathrm{dstrAreaSum}_{GP1} = P2+P3+P1\times\frac{P2+P3}{P2+P3+P4}$$

$$\mathrm{dstrAreaSum}_{GP2} = P4+P1\times\frac{P4}{P2+P3+P4}+P5\times\frac{P4}{P4+P6}$$

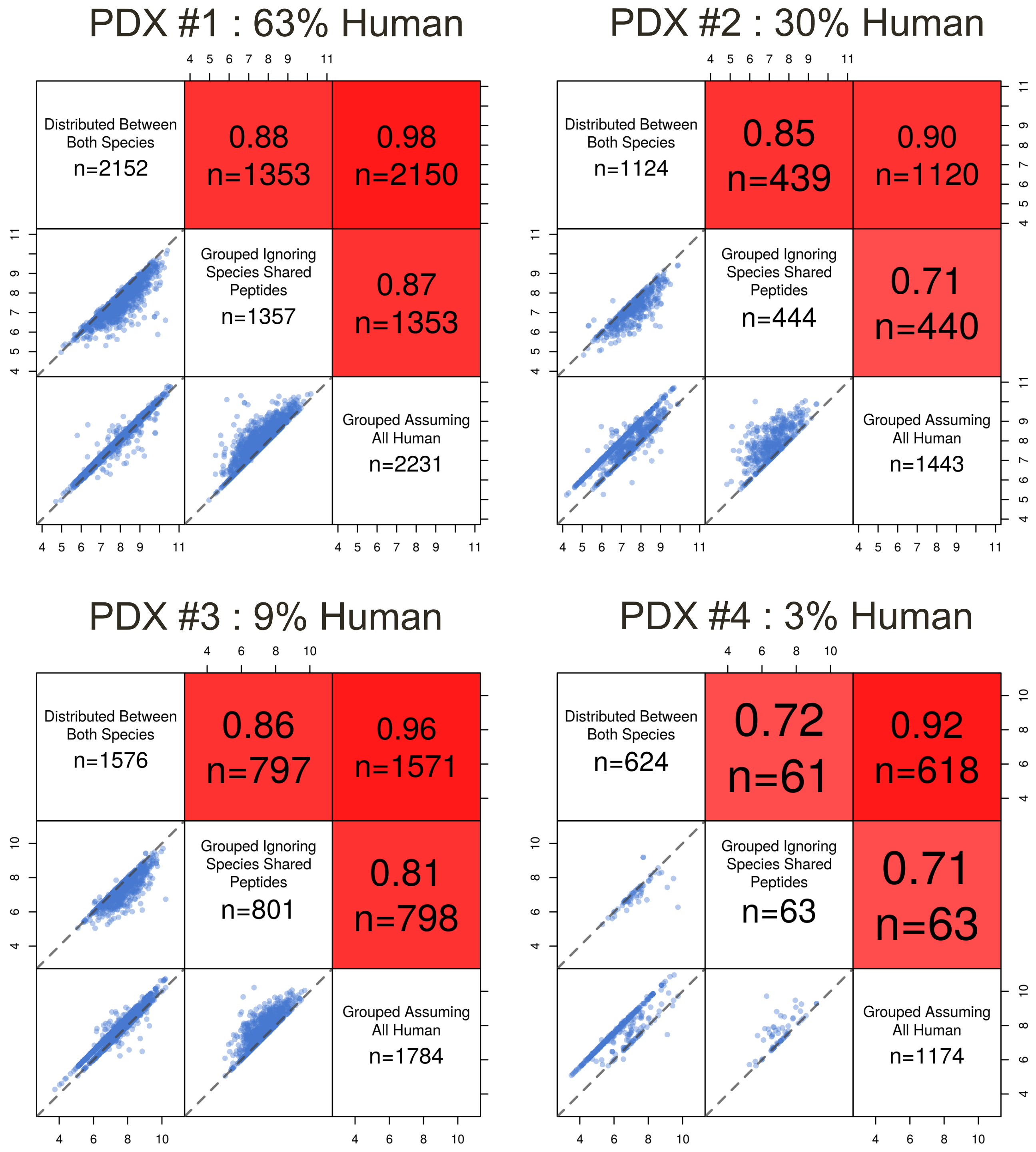
$$\mathrm{dstrAreaSum}_{GP3} = P6+P5\times\frac{P6}{P4+P6}$$

# Supplemental Figure 9



# Supplemental Figure 10

Tumor BCM-4913

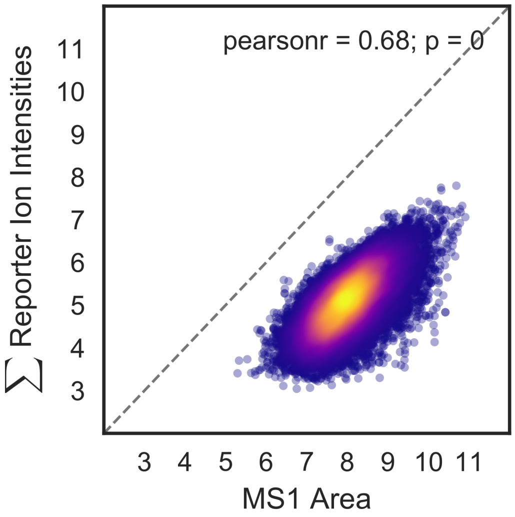

PDX #1 : 63% Human

PDX #2 : 30% Human

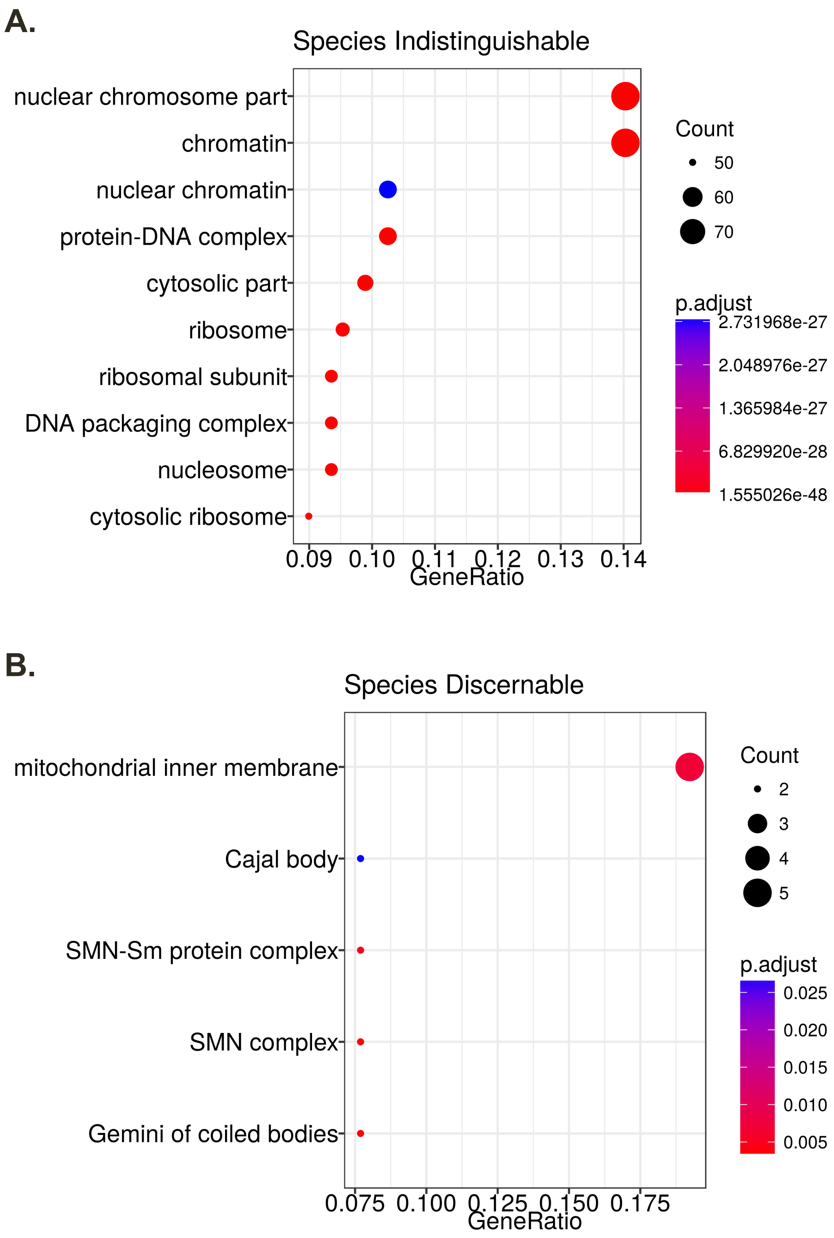
PDX #3 : 9% Human

PDX #4 : 3% Human

# Supplemental Figure 11



# Supplemental Figure 12


A.
Species Indistinguishable
nuclear chromosome part
chromatin
nuclear chromatin
protein-DNA complex
cytosolic part
ribosome
ribosomal subunit
DNA packaging complex
nucleosome
cytosolic ribosome
GeneRatio
Count
p.adjust
2.731968e-27
2.048976e-27
1.365984e-27
6.829920e-28
1.555026e-48
B.
Species Discernable
mitochondrial inner membrane
Cajal body
SMN-Sm protein complex
SMN complex
Gemini of coiled bodies
GeneRatio
Count
p.adjust