Succinctus

Proteomic diversity in a prevalent human-infective *Giardia duodenalis* sub-species

Samantha J. Emery-Corbin, Daniel Vuong, Ernest Lacey, Staffan G. Svärd, Brendan R.E. Ansell, Aaron R. Jex

PII: S0020-7519(18)30164-4
DOI: https://doi.org/10.1016/j.ijpara.2018.05.003
Reference: PARA 4080

To appear in: *International Journal for Parasitology*

Received Date: 21 March 2018
Revised Date: 15 May 2018
Accepted Date: 17 May 2018

Please cite this article as: Emery-Corbin, S.J., Vuong, D., Lacey, E., Svärd, S.G., Ansell, B.R.E., Jex, A.R., Proteomic diversity in a prevalent human-infective *Giardia duodenalis* sub-species, *International Journal for Parasitology* (2018), doi: https://doi.org/10.1016/j.ijpara.2018.05.003

Succinctus

Proteomic diversity in a prevalent human-infective *Giardia duodenalis* sub-species

Samantha J Emery-Corbin[a,*], Daniel Vuong[b], Ernest Lacey[b,c], Staffan G Svärd[d], Brendan R E Ansell[d], Aaron R Jex[d,e]

[a] *Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC, Australia*
[b] *Microbial Screening Technologies, Smithfield, NSW, Australia*
[c] *Chemistry and Biomolecular Sciences, Faculty of Science, Macquarie University, North Ryde, NSW, Australia*
[d] Depart*ment of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden*
[e] *Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Melbourne, VIC, Australia*

*Corresponding Author.
Dr Samantha Jane Emery-Corbin, Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia
Tel.: +61-3-9345-2656. *E-mail address*: emery.s@wehi.edu.au

Abstract:

Giardia duodenalis a species complex of gastrointestinal protists, with assemblages A and B infective to humans. To date, post-genomic proteomics are largely derived from Assemblage A, biasing understanding of parasite biology. To address this gap, we quantitatively analysed the proteomes of trophozoites from the genome reference and two clinical Assemblage B isolates, revealing lower spectrum-to-peptide matches in non-reference isolates, resulting in significant losses in peptide and protein identifications, and indicating significant intra-assemblage variation. We also explored differential protein expression between in vitro cultured subpopulations putatively enriched for dividing and feeding cells, respectively. This data is an important proteomic baseline for Assemblage B, highlighting proteomic differences between physiological states, and unique differences relative to Assemblage A.

Keywords: *Giardia duodenalis*, Assemblage B, Label-free proteomics, Protozoology, Comparative proteomics

Proteomics is the study of multiple gene products across spatial, temporal and biological contexts. High-throughput proteomics generates large-scale mass spectra data sets, which are generally matched to predicted peptide sequences derived from a known genome sequence database, known as the 'bottom-up' approach. Parasitology has integrated post-genomic, 'bottom-up' proteomics across multiple clinically relevant parasites, including *Giardia* (Emery et al., 2016). *Giardia duodenalis* (*Giardia lamblia*, *Giardia intestinalis*) is a gastrointestinal protist causing an estimated 200-300 million cases of diarrhoeal disease (giardiasis) annually, disproportionately in young children (Feng and Xiao, 2011). *Giardia duodenalis* taxonomy comprises a complex of host-adapted assemblages (Assemblages A-H), arguably representing morphologically indistinguishable, cryptic species (Thompson and Monis, 2012). Assemblages A and B are both infective to humans and are separated by considerable genetic distance, supporting their reclassification as separate species (Adam et al., 2013). Assemblage A is further defined by molecular classifications into sub-assemblages AI and AII, represented in genome sequences from isolates WB-C6 (Morrison et al., 2007) and DH (Adam et al., 2013), respectively. The genome sequence for Assemblage B is derived from the GS isolate as draft (Franzen et al., 2009) and improved (GS-B) releases (Adam et al., 2013).

To date, post-genomic proteomic data for *G. duodenalis* is biased for Assemblage A isolates (Emery et al., 2016). Compared with Assemblage A, Assemblage B has higher epidemiological prevalence in humans, and is associated with more damaging clinical symptoms and increased incidence of treatment failure (Thompson, 2001; Feng and Xiao, 2011). Comparative genomics has demonstrated the Assemblage B genome also contains both the highest number of unique, and lowest number of shared open reading frames (ORFs) between currently sequenced *G. duodenalis* genomes (Adam et al., 2013), highlighting their considerable genetic distance. Although only a single Assemblage B isolate has been sequenced, comparative genomics of multiple AI sub-assemblage genomes revealed lower single nucleotide polymorphisms (SNP) frequency than observed between AII sub-asssemblage genomes (Nageshan et al., 2011; Ankarklev et al., 2015), with complementary proteomics demonstrating approximate peptide identifications between sub-assemblage AI genome and non-reference isolates, but a 5% decrease in counts when AI peptides were searched against the AII genome sequence (and vice versa) (Emery et al., 2015) (Fig. 1B). As such, spectrum-to-peptide matching metrics in Assemblage A demonstrates that proteomic level data corresponds to intra- and inter-assemblage diversity at the genomic level.

To support further proteomic study of *G. duodenalis* assemblage B, we quantified and characterized proteomes of the infective stage (the 'trophozoite') of the genome reference

isolate (GS/M), and two clinical isolates (BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487). Presently, low intra-assemblage sequence variation in Assemblage A has allowed comparative and quantitative proteomics across and between reference and non-reference (clinical) isolates with negligible impacts on database searching sensitivity, but this has not been confirmed for Assemblage B. We explored the impact of genomic diversity within Assemblage B, including the potential of SNPs, ORFs and chromosomal/structural variants in the clinical isolates, by assessing agreement in spectral mapping from each isolate to peptides predicted from the reference genome (GS) (Fig. 1A). Using this approach, spectrum-to-peptide matching of these isolates to the Assemblage B reference sequence suggests yet undescribed genetic variation within this Assemblage. Further, to explore the effect of physiological differences on spectral matching rates, we compared adhered and non-adhered (motile) sub-populations of trophozoites from in vitro culture flasks, to detect differentially expressed cell viability and cell-cycle markers. This analysis, together with peptide and protein identification rates between genome and clinical isolates, establishes an experimental standard and detailed reference data to support further study of this neglected, human-infective assemblage.

Isolates GS/M, BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487 (Upcroft et al., 1995) were cultured as previously described (Emery et al., 2015) to late-log phase in biological triplicates. For the GS/M isolate, non-adhered trophozoite material was collected from total decanted media and any sediment material, while adhered trophozoites were collected by addition of cold PBS after harvesting decanted media from all three isolates. Cell viability after harvest of enriched trophozoites was verified by a trypan blue exclusion dye assay (0.4%w/v). Protein from pelleted and washed trophozoites was extracted in 5% sodium dodecyl sulfate (SDS) in 100 mM Tris base (Sigma) pH 8, containing 1 mM EDTA and 5% beta- mercaptoethanol, and was heated to reduce and solubilise proteins at 75°C for 10 min. Insoluble material was removed via centrifugation at 0°C at 13,000 *g* for 10 min and then nucleic acids, lipids and detergents removed via methanol-chloroform precipitation. The concentration of protein in each solution was measured by bicinchoninic acid (BCA) assay (Pierce, USA), and cysteines were reduced with 10 mM dithiothreitol followed by alkylation with 15 mM iodoacetamide. Proteins were trypsin digested (Promega, USA) overnight into peptides at a ratio of 100:1 protein to enzyme. Samples were acidified an de-salted via SPE on in-house assembled stage-tips containing polystyrene-divylbenzene (SDB-RPS) (Empore, USA), and 50 ug of peptide material fractionated using a four stage elution on SDB-RPS as previously described (Rappsilber et al., 2007). Samples were vacuum centrifuged to dryness, and reconstituted with 0.1% formic acid, 2% acetonitrile for LC-MS/MS analysis.

LC MS/MS was carried out on a LTQ Orbitrap Elite (Thermo Scientific, USA) with a nanoESI interface in conjunction with an Ultimate 3000 RSLC nanoHPLC (Dionex Ultimate 3000). The LC system was equipped with an Acclaim Pepmap nano-trap column (Dionex-C18, 100 Å, 75 µm x 2 cm) and an Acclaim Pepmap RSLC analytical column (Dionex-C18, 100 Å, 75 µm x 50 cm). The tryptic peptides were injected into the enrichment column at an isocratic flow of 5 µL/min of 3% v/v CH_3CN containing 0.1% v/v formic acid for 6 min before the enrichment column was switched in-line with the analytical column. The eluents were 0.1% v/v formic acid (solvent A) and 100% v/v CH_3CN in 0.l% v/v formic acid (solvent B). The flow gradient was (i) 0-6 min at 3% B, (ii) 6-95 min, 3-20% B (iii) 95-105 min, 20-40% B (iv) 105-110 min, 40-80% B (v) 110-115 min at 80% B (vii) 115-117 min, 80-3% B and (viii) 117-125 min at 3% B. The LTQ Orbitrap Elite spectrometer was operated in the data-dependent mode with nanoESI spray voltage of 1.8 kV, capillary temperature of 250^{o}C and S-lens radio frequency value of 55%. All spectra were acquired in positive mode with full scan MS spectra scanning from m/z 300-1650 in the Fourier transform mode at 240,000 resolution after accumulating to a target value of 10^6. Lock mass of 445.120025 was used. The top 20 most intense precursors were subjected to rapid collision induced dissociation (rCID) with normalized collision energy of 30 and activation q of 0.25. A dynamic exclusion width of 30 s was applied for repeated precursors.

Database searching was performed using MaxQuant software (v1.5.5.1) for label-free quantification (LFQ) (Cox and Mann, 2008), and the re-sequenced GS-B genome (release 5.1; GiardiaDB.org). As MaxQuant LFQ creates a normalised intensity profile relative to search groups and dimensions, isolate triplicates and their respective fractions were searched as GS/M, BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487 separately from GS/M adhered and non-adhered trophozoite LFQ quantitative searches. Default parameters were used for target and decoy searching with a false discovery rate (FDR) of 1% imposed for spectrum-to-peptide matches; the LFQ minimum ratio count was set to 1, and matching between runs set to 'match from and to'. Oxidation of methionine and N-acetylation of protein N-termini were set to variable modifications, and carbidomethylation of cysteine was considered a fixed modification. The 'proteingroups.txt' output file from MaxQuant was imported to Perseus (Tyanova et al., 2016) (v1.5.5.3) and protein groups identified in the reverse database, contaminant database, or only by site, were removed. Low stringency protein groups used in further analyses were filtered to include only those reproducibly identified in biological triplicates within at least one isolate/sample group (termed 'high stringency').

The complete raw files and search results can be accessed via the ProteomeXchange Consortium (Vizcaino et al., 2013) via the PRIDE partner repository with the dataset

identifier **PXD007943.** Supplementary methods can be found in Supplementary Data S1, while Supplementary Data Tables S1-S4 and Supplementary Figs. S1-S4 are hosted on Mendeley (http://dx.doi.org/10.17632/brhg8bhcvm.3) togeher with their respective table and figure legends. All these supplementary files can be downloaded from the Mendeley repository via the DOI website URL above.

Searching the spectra from each isolate against the GS-B reference database for isolates GS/M, BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487 identified a non-redundant total of 3162 reproducibly identified proteins across the three isolates, with 3127 (98.9%) of all proteins identified detected in the GS/M isolate (Fig. 2C; Supplementary Table S1). Substantially fewer peptide matches were observed for the other isolates, with an average 58.8% fewer peptide counts corresponding to 25.5% fewer proteins identified (Fig. 2A). These observed decreases in peptide matches were not due to fewer MS/MS spectra recorded in non-genome isolates, but rather to lower rates of spectrum-to-peptide matching from similar levels of recorded MS/MS spectra (Fig. 2B). Despite lower identification rates in BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487 relative to the reference genome, both isolates had similar peptide counts and protein identifications, with 89.4% of protein identifications in common (Fig. 2C). Only 1.4% of these identifications were not also identified in GS/M. This equivalency between BRIS/91/HEPU/1279 and BRIS/92/HEPU/1487 protein identifications, as well as overlap with GS/M, support the presence of a core, conserved proteome within Assemblage B, with low sequence variation. However, when LFQ abundances of protein identifications common to all three isolates were compared via Principal Component Analysis (PCA) (Supplementary Fig. S1), there was significant distance between clinical isolates and the GS/M isolate on the first component, while the clinical isolates were separated but remained close on both first and second components. This may indicate that there is a closer relationship between these clinical isolates from their shared origins in a sympatric population (Upcroft et al., 1995), but also supports the need to compare more isolates across broader populations to gain a better measure of population variability in Assemblage B.

Differences in spectral matching rates between Assemblage B isolates are more than 10-fold lower than intra-subassemblage matching rates reported in proteomic studies of Assemblage A (Emery et al., 2015), and imply relatively greater intra-assemblage diversity between Assemblage B isolates (Fig. 1). Further to this finding, proteins identified in all three isolates also had higher peptide counts and increased sequence coverage against the GS/M reference (Supplementary Fig. S2), indicating sequence variation within ORFs in the clinical isolates which may prevent spectrum-to-peptide matching to the reference sequence (Fig. 1A).

Previously, low genomic diversity in Assemblage A has allowed comparative proteomics (Emery et al., 2016) and transcriptomics (Ansell et al., 2017) between non-genome isolates without isolate-dependent loss of coverage that compromises comparisons. However, given the relatively lower peptide and protein identity between non-reference isolates of Assemblage B *Giardia*, the reference genome for Assemblage B appears less representative of this assemblage overall.

Giardia are tetraploid, with variable karyotypes and many reported chromosomal rearrangements, deletions and duplications, which are hypothesised to account for the expanded, heterologous gene families observed in *Giardia* (Adam et al., 2013). Of the 3127 proteins identified in GS/M, 725 (24.0%) were identified exclusively within the reference isolate, suggesting the presence of unique ORFs between Assemblage B isolates as well as SNP-based sequence variations (Fig. 1A; Supplementary Fig. S2). Comparative proteomics in Assemblage A demonstrated less identity in the products of expanded, *Giardia*-specific gene families, particularly the variant-specific surface proteins (VSPs) encoded in regions of intra-assemblage diversity (Franzen et al., 2013; Emery et al., 2015). Therefore, we compared protein identifications in expanded *Giardia* protein families between Assemblage B isolates, and observed relatively more VSPs in the reference isolate GS/M (Fig. 2D). Specifically, 83.9% more VSPs could be identified in GS/M than in the non-reference isolates, despite the fact that the GS-B genome encodes the largest repertoire (503 full-length genes) of VSPs among the presently sequenced *Giardia* genomes (Adam et al., 2013) (Fig. 2D). Substantial differences have been previously reported for the numbers of expressed VSPs within sub-assemblages and isolates (Emery et al., 2015). However, given the depth of our dataset and observed decreases in spectrum-to-peptide matches in non-reference isolates (Fig. 2), we believe these decreased VSP variant identifications in fact correspond to intra-assemblage diversification of VSP repertoires, or substantial SNPs within existing variants. However, VSPs do not account for all of the 725 proteins identified exclusively in GS/M, with a further 350 (48.3%) of these proteins annotated only as 'hypothetical'. Therefore, while expanded gene families are susceptible to lower rates of spectrum-to-peptide matching, genomic variation appears to extend to a wider range of functionally un-annotated genes.

As the GS-B genome has the highest allelic heterozygosity of sequenced *Giardia* genomes (Adam et al., 2013), and evidence of bi-allelic expression (Franzen et al., 2013), we employed the 'Dependent Peptide' search mode on Maxquant to identify prospective peptides containing deletions, insertions or substitutions from potentially co-expressed polymorphic alleles, or polymorphic paralogues in GS/M. A full description of this search and its logic is provided in Supplementary Data S1. This error-tolerant search is sensitive to peptides with

amino acid substitutions that would otherwise be excluded from traditional database searching. A total of 3479 dependent peptides with putative sequence variations and high positional probability (>0.9) were detected in the GS/M isolate, corresponding to a small increase in peptide counts, averaging 3.48% per replicate (Supplementary Table S2). However, substitutions in these peptides should be considered putative, and require further validation using deep DNA or RNA sequencing, which would also permit construction of a true variant database for proteomic searching. Nonetheless, given the low proportion of high-confidence peptides containing substitutions, we believe that proteome coverage is not compromised by sequence variants in GS/M that would lower sensitivity in this isolate and assemblage.

Cyto-adherence in *Giardia* trophozoite cultures is a correlate of cell cycle synchronicity, viability and fitness, and proteomics and transcriptomics vary in their approach to collecting trophozoites for passage, or for harvest of stress assays, to ensure homogeneity. Database searching of mass spectra from adhered and non-adhered trophozoite fractions from the GS/M isolate (Fig. 3B) revealed a non-redundant total of 3309 reproducibly identified proteins, with an average of 35,528 total peptides in each replicate (Fig. 3A). The overlap between adhered and non-adhered trophozoite proteomes was near complete, with 3221 (97.3%) proteins in common (Supplementary Table S3). To quantify protein abundance, label-free, MS1 intensity values were exported from Maxquant into Perseus, log transformed, and missing values were imputed based on an assumption of normality. Pearson's correlation of protein abundance between adhered and non-adhered trophozoites was high (>0.98) (Supplementary Fig. S3), although the two populations were separated in the first and second principle components (Fig. 3C). Differential protein expression between adhered and non-adhered trophozoites was calculated using paired t-tests, with an FDR corrected significance threshold of 0.1 and an additional effect size threshold of log(0.1). Using this approach, we detected a single statistically significant, differentially expressed (DE) protein between adhered and non-adhered trophozoites (Fig. 3D). The single DE protein (GSB_151618) was annotated as a 'hypothetical protein' with interpro annotations suggesting it may be a putative P-loop containing nucleoside triphosphate hydrolase.

From a technical perspective, the lack of DE proteins detected in this study between trophozoite fractions is fortunate, indicating both suspended and adhered cells may be sufficiently similar to include in experimental protocols requiring large amounts of starting material, such as mass spectrometry of protein post-translation modifications (Emery et al., 2016). However, adherence in *Giardia* trophozoites is likely a function of cell cycle state, and viability, and microscopic examination demonstrated trophozoites in the adherent fraction

were viable (via trypan blue dye exclusion) and homogenous in morphology, while non-adherent trophozoites varied in size and morphology, potentially due to cell growth prior to cell division or decreasing viability (Fig. 3B). The media column containing non-adherent cells also contained background cellular debris. Indeed, analysis of LFQ intensities of 'contaminant' protein identifications from Maxquant searching revealed that although contaminants were a minimal fraction of identifications (Supplementary Fig. S3A), contaminant LFQ intensities were cumulatively higher in non-adhered fractions (Supplementary Fig. S4B) and statistically significantly higher than in the adhered fraction (Supplementary Fig. S4C). Therefore, study designers must be mindful of experimental conditions which could induce loss of trophozoite viability causing contaminant proteins and non-viable trophozoites to accumulate in non-adhered fractions, together with their peptides in downstream mass spectrometry.

The lack of post-genomic analyses of Assemblage B can also be attributed to insufficient functional annotation of its genome. The number of unannotated, 'hypothetical proteins' in Assemblage A are acknowledged as a significant impediment to 'omics' studies (Emery et al., 2016), reducing the power of pathway analyses and gene set enrichment (GSE) testing. Of the 6166 ORFs annotated in the most recent GS-B release (Aurrecoechea et al., 2009), we generated mass spectrometry evidence of expression for 3309 proteins from trophozoite fractions, representing approximately 53.7% of annotated ORFs, and achieving the highest proteome coverage for any *Giardia* assemblage to date. Gene ontology annotations for 'Cell Component', 'Molecular Function' and 'Biological Process' were downloaded from GiardiadDB.org for the 3309 proteins (Supplementary Table S4), and 1557 proteins (47.1%) had no associated gene ontology (GO) annotations, and of these 645 (19.6%) were 'hypothetical proteins'. GO annotation within our dataset was significantly better compared with the overall Assemblage B genome, where 58.7% of proteins have no GO annotations, and a further 34.0% of these are also 'hypothetical proteins'. Of the 1752 proteins in our dataset with GO annotations, the majority of these possessed 'Molecular Function' annotations (Fig. 3E), and only a small subset (12.4%) had annotations for all three ontologies. For further analysis of GO annotation, there are fewer compatible algorithms comparable to Assemblage A, where *Giardia*-specific accessions link externally to NCBI Entrez, Uniprot and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, whose accessions are compatible with external bioinformatics tools including DAVID (Huang et al., 2007), PANTHER (Mi et al., 2013) and Perseus software. To analyse and group GO annotations, we supplied downloaded GO terms from Giardiadb.org to the Web Gene

Ontology Annotation Plotting (WEGO) tool (Ye et al., 2006) in WEGO native format, and graphs of GO terms are shown in Supplementary Fig. S5.

The *Giardia duodenalis* Assemblage B is a neglected, human-infective parasite with sufficient genomic and biological variation to warrant independent investigation. We believe that the data presented here provides a proteomic baseline to inform experimental design, database searching and functional analyses for future investigations, and will substantially increase the informative value of proteomic comparisons of human-infective *Giardia* assemblages.

Acknowledgements

This work, including the efforts of AJ, was funded by the Australian Research Council (ARC) (LP120200122). SE, BA and AJ are supported by the Victorian State Government Operational Infrastructure Support (Australia) and Australian Government National Health and Medical Research Council (NHMRC) Independent Research Institute Infrastructure Support Scheme. AJ is also supported by a NHMRC Career Development Fellowship (APP1126395). SE and this research are also supported by a Jack Brockhoff Foundation Early Career Grant (Australia) (ID JBF 4184, 2016). Proteomic analysis was performed at the Mass Spectrometry and Proteomics Facility at Bio21 at Melbourne University (Australia) with assistance from Ching-Seng Ang and Shuai Nie. The authors also wish to acknowledge Professor Jacqui Upcroft (Australia) for providing the in vitro *Giardia* isolates used in this study. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References:

Adam, R.D., Dahlstrom, E.W., Martens, C.A., Bruno, D.P., Barbian, K.D., Ricklefs, S.M., Hernandez, M.M., Narla, N.P., Patel, R.B., Porcella, S.F., Nash, T.E., 2013. Genome sequencing of *Giardia lamblia* genotypes A2 and B isolates (DH and GS) and comparative analysis with the genomes of genotypes A1 and E (WB and Pig). Genome Biol Evol 5, 2498-2511.

Ankarklev, J., Franzen, O., Peirasmaki, D., Jerlstrom-Hultqvist, J., Lebbad, M., Andersson, J., Andersson, B., Svard, S.G., 2015. Comparative genomic analyses of freshly isolated *Giardia intestinalis* assemblage A isolates. BMC genomics 16, 697.

Ansell, B.R., Baker, L., Emery, S.J., McConville, M.J., Svard, S.G., Gasser, R.B., Jex, A.R., 2017. Transcriptomics Indicates Active and Passive Metronidazole Resistance Mechanisms in Three Seminal *Giardia* Lines. Front Microbiol 8, 398.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Carlton, J.M., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Morrison, H.G., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert, C.J., Jr., Sullivan, S., Treatman, C., Wang, H., 2009. GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. Nucleic Acids Res 37, D526-530.

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367-1372.

Emery, S.J., Lacey, E., Haynes, P.A., 2015. Quantitative proteomic analysis of *Giardia duodenalis* assemblage A: A baseline for host, assemblage, and isolate variation. Proteomics 15, 2281-2285.

Emery, S.J., Lacey, E., Haynes, P.A., 2016. Quantitative proteomics in *Giardia duodenalis*-Achievements and challenges. Mol Biochem Parasitol 208, 96-112.

Feng, Y., Xiao, L., 2011. Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis. Clin Microbiol Rev 24, 110-140.

Franzen, O., Jerlstrom-Hultqvist, J., Castro, E., Sherwood, E., Ankarklev, J., Reiner, D.S., Palm, D., Andersson, J.O., Andersson, B., Svard, S.G., 2009. Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: is human giardiasis caused by two different species? PLoS Pathog 5, e1000560.

Franzen, O., Jerlstrom-Hultqvist, J., Einarsson, E., Ankarklev, J., Ferella, M., Andersson, B., Svard, S.G., 2013. Transcriptome profiling of *Giardia intestinalis* using strand-specific RNA-seq. PLoS Comput Biol 9, e1003000.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35, W169-175.

Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc 8, 1551-1566.

Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J., Davids, B.J., Dawson, S.C., Elmendorf, H.G., Hehl, A.B., Holder, M.E., Huse, S.M., Kim, U.U., Lasek-Nesselquist, E., Manning, G., Nigam, A., Nixon, J.E., Palm, D., Passamaneck, N.E., Prabhu, A.,

Reich, C.I., Reiner, D.S., Samuelson, J., Svard, S.G., Sogin, M.L., 2007. Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science 317, 1921-1926.

Nageshan, R.K., Roy, N., Hehl, A.B., Tatu, U., 2011. Post-transcriptional repair of a split heat shock protein 90 gene by mRNA trans-splicing. J Biol Chem 286, 7116-7122.

Rappsilber, J., Mann, M., Ishihama, Y., 2007. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2, 1896-1906.

Thompson, A., 2001. Human giardiasis: genotype-linked differences in clinical symptomatology. Trends Parasitol 17, 465.

Thompson, R.C., Monis, P., 2012. Giardia--from genome to proteome. Adv Parasitol 78, 57-95.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., Cox, J., 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13, 731-740.

Upcroft, J.A., Boreham, P.F., Campbell, R.W., Shepherd, R.W., Upcroft, P., 1995. Biological and genetic analysis of a longitudinal collection of *Giardia* samples derived from humans. Acta Trop 60, 35-46.

Vizcaino, J.A., Cote, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., Hermjakob, H., 2013. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 41, D1063-1069.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., Wang, J., 2006. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34, W293-297.

Figure Legends

Fig. 1. Effects of genome and sequence variation on bottom-up proteomics spectrum-to-peptide matching. A) Sources of intra-assemblage variation causing isolate-dependent losses in peptide matches and protein identifications. For the genome reference isolate (grey), open reading frame (ORF) complement and sequences match, however in non-reference isolates (blue), heterologous regions of the chromosomes (insertions or deletions) lead to changes in gene repertoires (often *Giardia* variable gene families (Adam et al., 2013)), with ORFs from the reference isolate potentially lost, and divergent ORFs gained, which will be absent in database searches of the reference genome. Non-synonymous changes (single nucleotide polymorphisms, SNPs) in gene sequence also cause discrepancies between in silico predicted peptides from the reference sequence and MS/MS experimental spectra from non-reference isolates, leading to decreasing spectrum-to-peptide matches. B) Intra-subassemblage variation within multiple genomes from Assemblage A reveal low SNP frequency between sub-assemblage isolates (left), translating to only a 5% decrease in peptide-spectrum matches (Emery et al., 2015) when spectra from each sub-assemblage is searched against alternate subassemblage genomes (right). VSP, Variant-specific Surface Protein.

Fig. 2. Spectrum-to-peptide matching metrics between clinical and genome reference isolates of Assemblage B *Giardia duodenalis.* A) Complete summary of peptide and protein identification data of *G. duodenalis* proteins from the three Assemblage B isolates. B) MS/MS and peptide match rates between the three Assemblage B isolates. The MS/MS recorded in all three isolates is equivalent, however lower rates of peptides are matched in non-reference isolates compared with the reference (GS/M). C) Proportional Venn diagram of reproducibly identified proteins in the three isolates. D) Numbers of reproducible identified proteins from the four variable gene families of *Giardia* for the three Assemblage B isolates. R.I., Reproducibly Identified; HCMP, High Cysteine Membrane Protein; NEK, NimA related Kinase; VSP, Variant-specific Surface Protein.

Fig. 3. Proteome and gene annotation analyses of trophozoite fractions within the *Giardia duodenalis* Assemblage B genome reference isolate, GS/M A) Complete summary of peptide and protein identification data of *G. duodenalis* proteins from enriched adhered and non-adhered trophozoite fractions from GS/M. B) Microscopy images of enriched fractions of adhered and non-adhered trophozoites. C) Principle component analysis (PCA) chart of plots of principal component scores plot in the space of the first two principal components generated for the whole dataset from label-free quantitation (LFQ) intensity values between

adhered (Ad) and non-adhered (NAd) trophozoite fractions. D) Volcano plot of differential protein abundance between adhered and non-adhered trophozoite fractions. The Y-axis shows the negative log 10 FDR-corrected *P* value; the X-axis represents the log 2 difference between the means of the two fractions as a measure of change in abundance. Proteins above the curve (red) meet statistical significance as well an effect size > log2(0.1). E) Upset plot showing the coverage and type of gene ontology (GO) annotation recorded for 1750 of the 3309 reproducible identified proteins from trophozoite fractions. F) Proportional Venn diagram of the coverage and overlap between proteins with GO and InterPro annotations from the 3309 reproducible identified proteins from trophozoite fractions.

A)

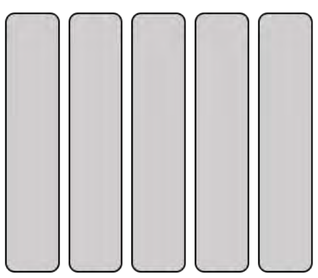
ISOLATE

CHROMOSOME

ORFs

SEQUENCE

*Genome Reference



CORE GENES

ORF_01

ORF_02

ORF_03

...

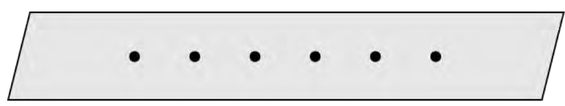
VARIABLE GENOME

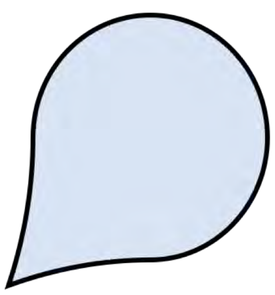
VSP_01

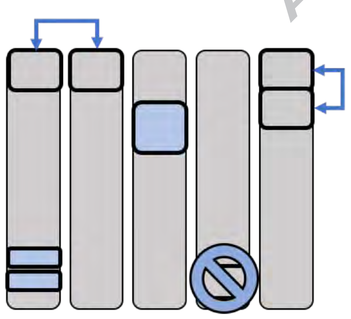
VSP_02

VSP_03

...







CORE GENES

ORF_01

ORF_02

ORF_03

...

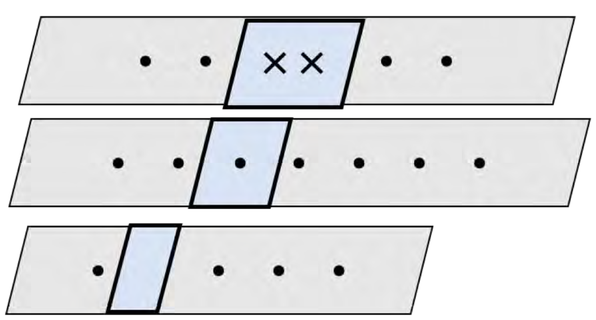
VARIABLE GENOME

VSP_01

~~VSP_02~~

VSP_04

...



B)

AI

AI

= 7.5/100,000

AII

AII

= 350/100,000

AI

AII

Intensity

MS1 m/z

Intensity

MS2 m/z

Intensity

MS1 m/z

Intensity

MS2 m/z

AII Reference Genome

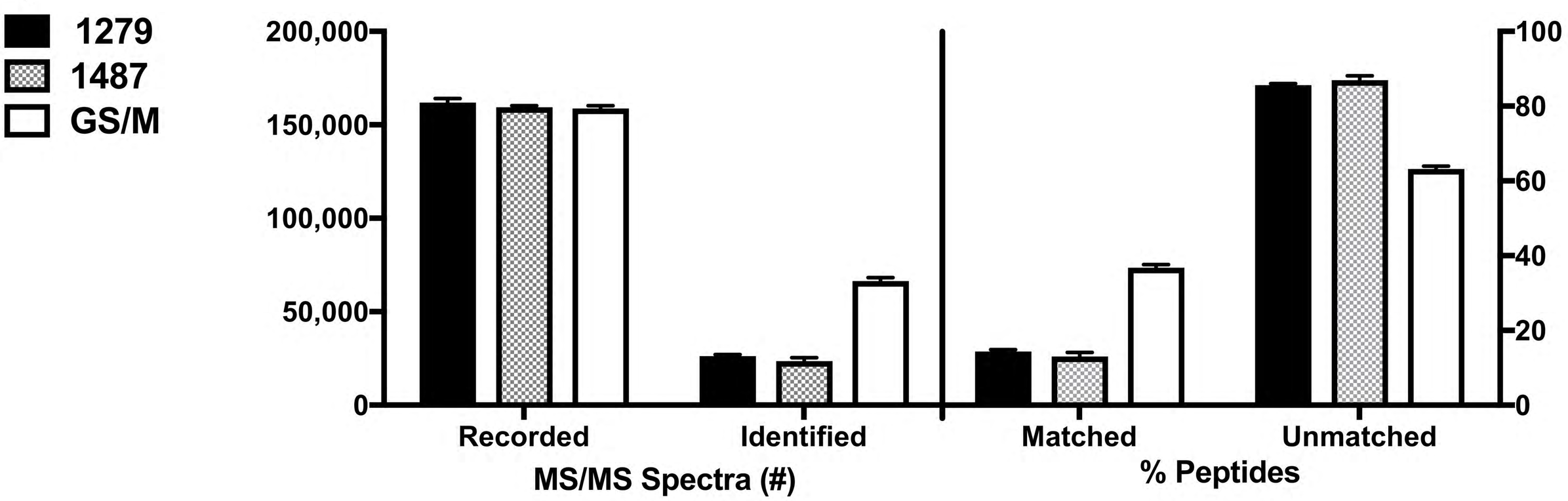
-5% spectrum-peptide matches

AI Reference Genome

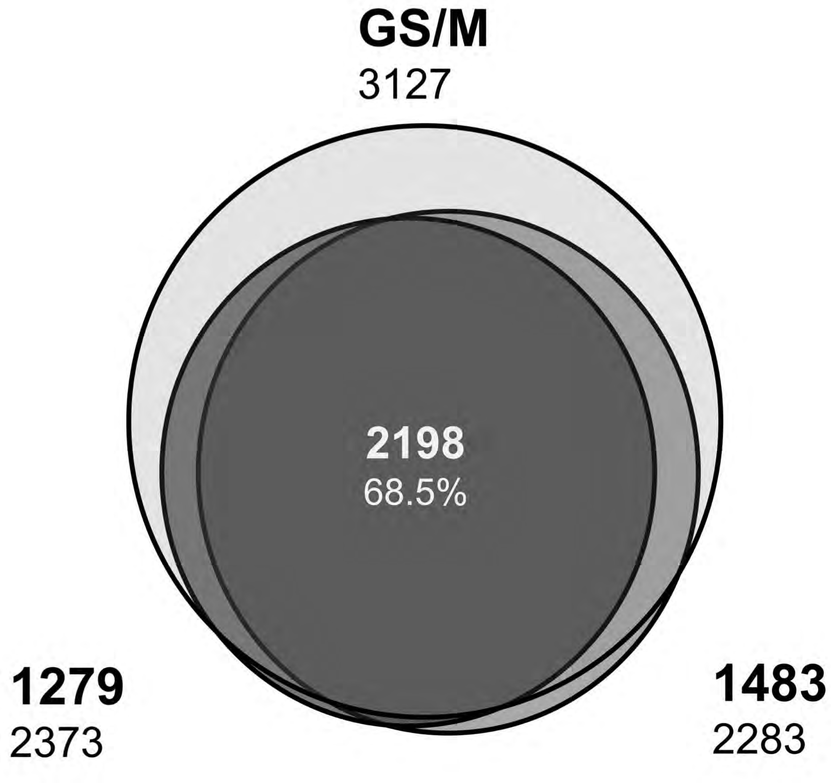
A)

Isolate	Low stringency peptide count			Average number of peptides (±RSD%)	Number of R.I. proteins common to three replicates	R.I. Protein FDR (%)	R.I. Peptide FDR (%)
	Replicate 1	Replicate 2	Replicate 3				
BRIS/91/HEPU/1279	13710	13334	13886	13643.3 ± 1.7%	2373	0.27%	1.01%
BRIS/92/HEPU/1487	13522	12936	12880	13112.7 ± 2.2%	2283	0.29%	2.2%
GS/M	33274	33165	33418	33285.7 ± 0.3%	3127	0.35%	0.07%

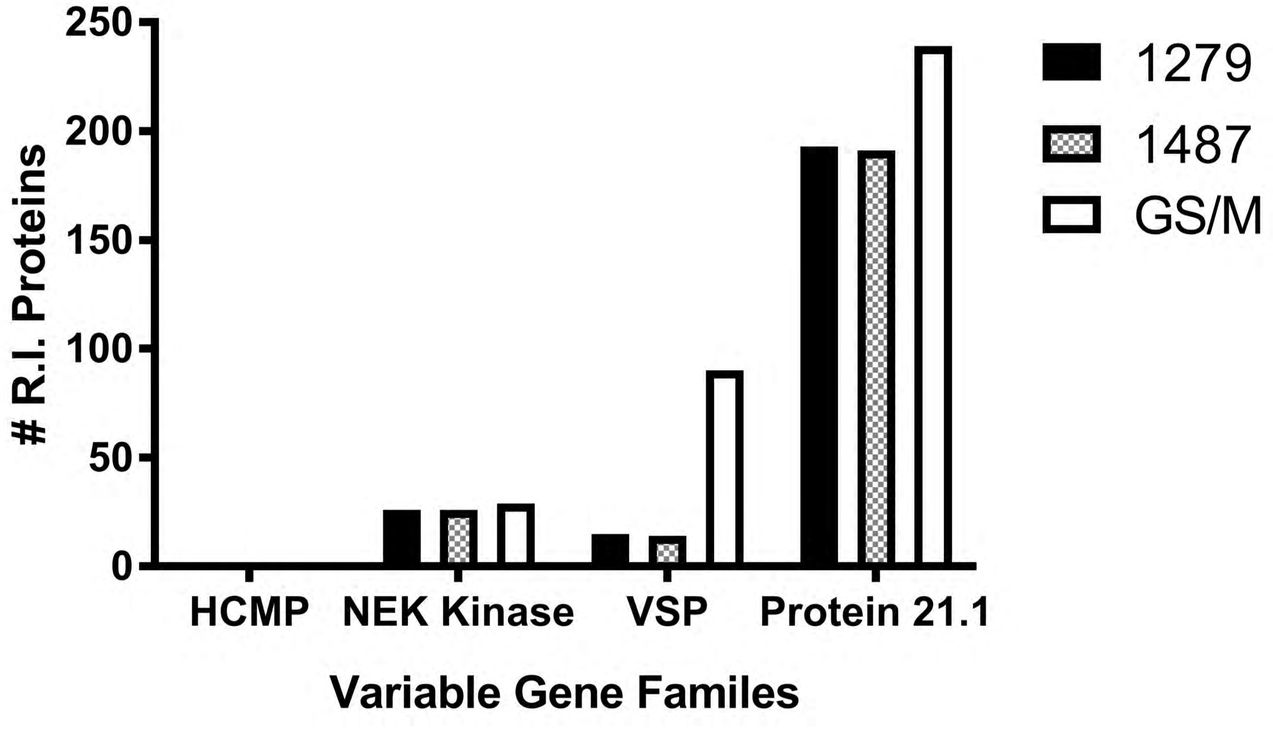
B)



C)



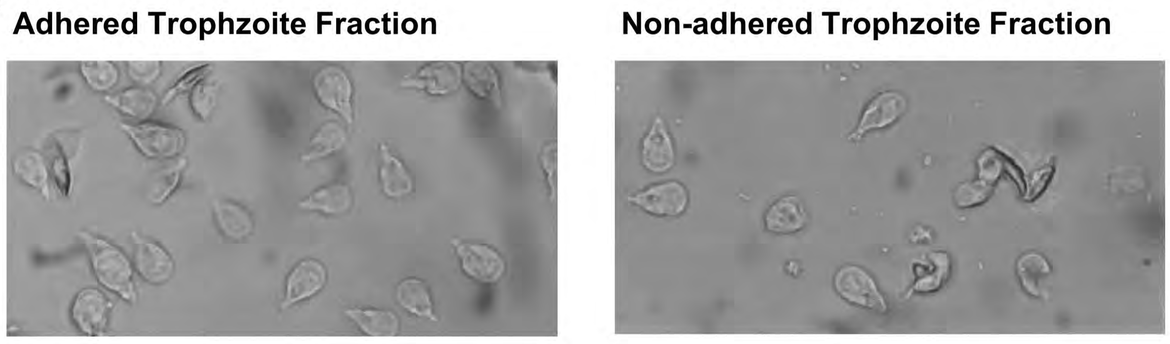
D)



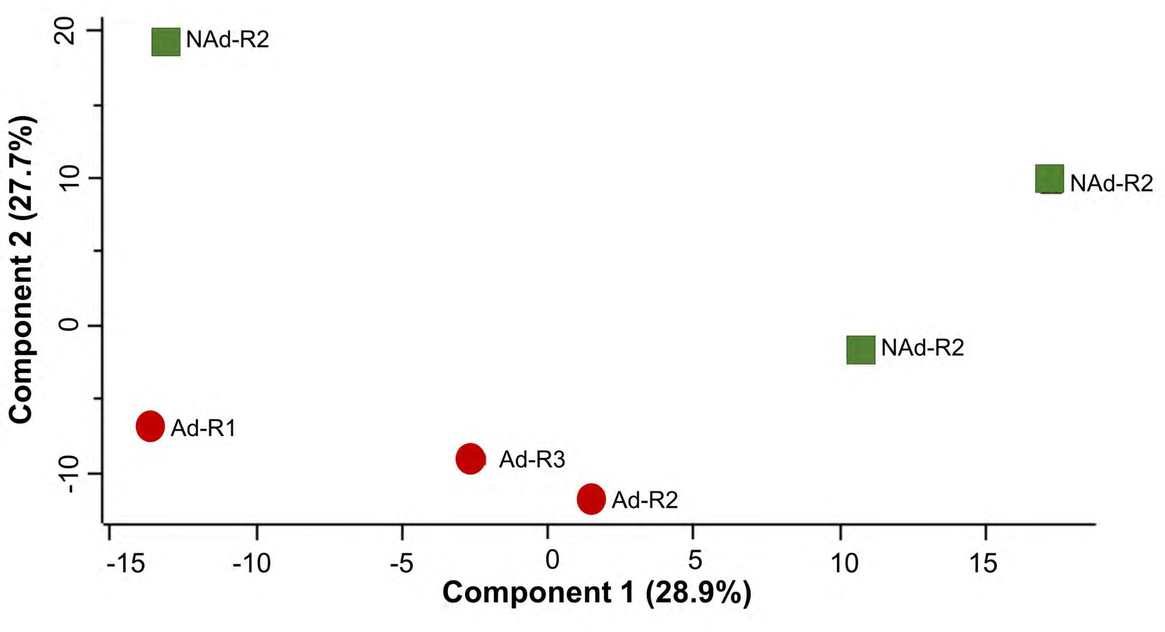
A)

Trophozoite fraction	Low stringency peptide count			Average number of peptides (±RSD%)	Number of R.I. proteins common to three replicates	R.I. protein FDR (%)	R.I. peptide FDR (%)
	Replicate 1	Replicate 2	Replicate 3				
Adhered Trophozoites	35632	35903	36059	36 059 ± 1.2%	3279	0.10%	0.70%
Non-adhered Trophozoites	35589	34890	34512	34 997 ± 1.3%	3251	0.10%	0.68%

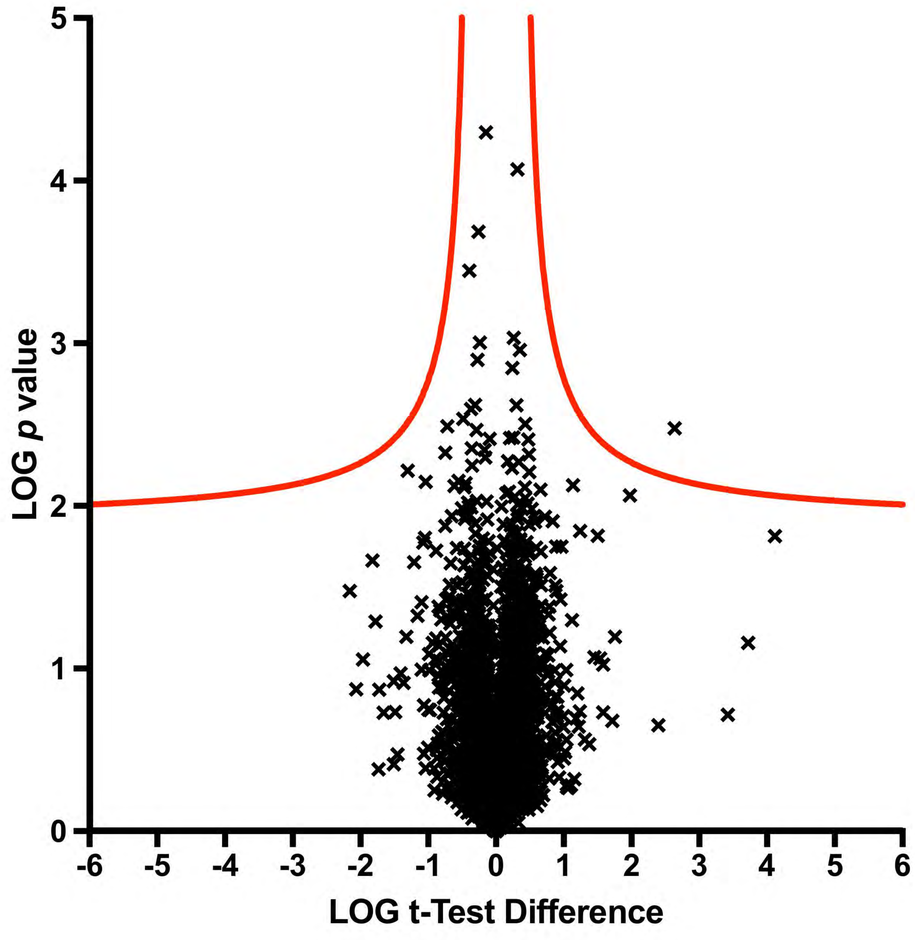
B)



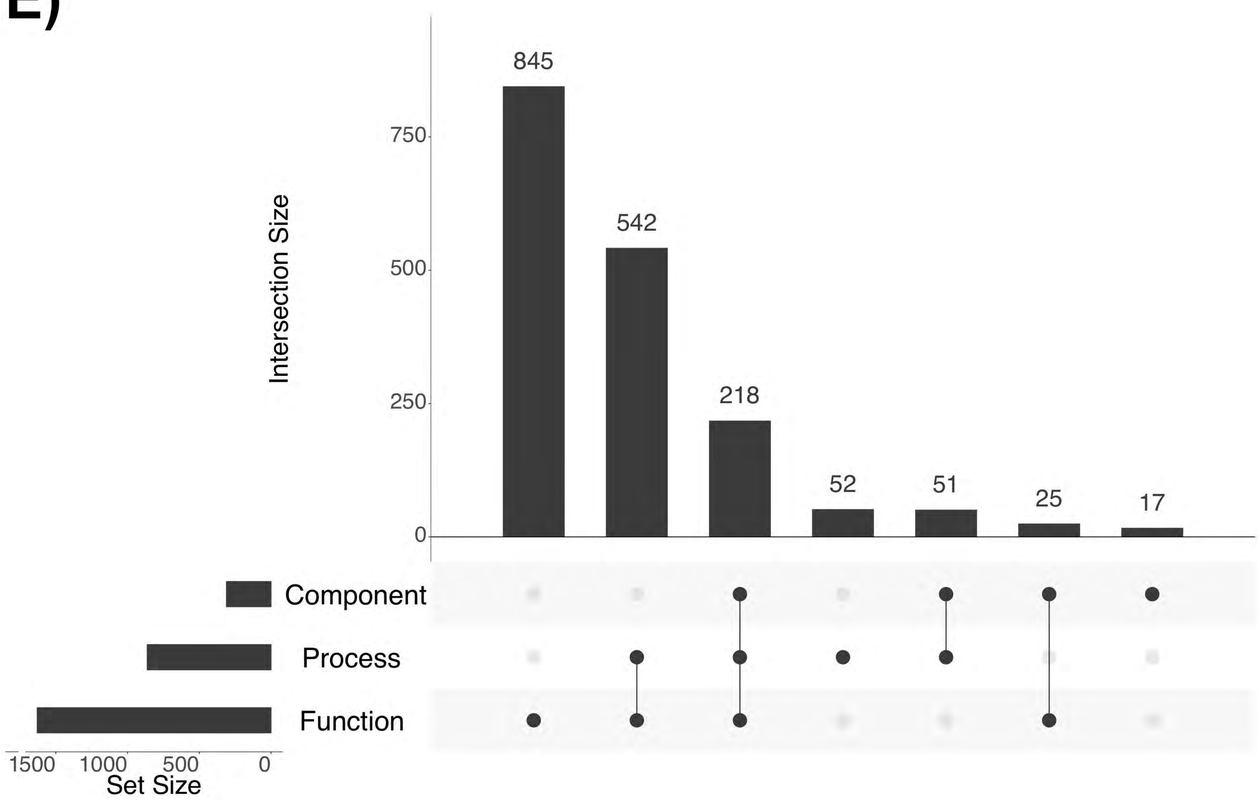
C)



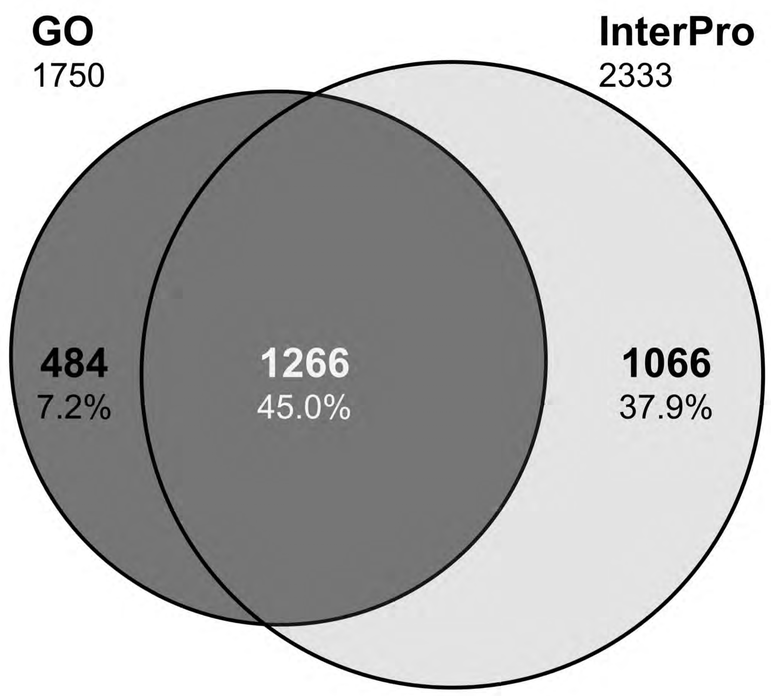
D)



E)



F)



Highlights

- Proteomic data are limited for *Giardia duodenalis* Assemblage B, affecting our understanding of parasite biology.
- Variation between Assemblage B isolates impacts peptide matching in non-reference isolates.
- Peptide counts of non-reference isolates were reduced by >50%, protein identifications down 25%.
- The Assemblage B genome reference is only representative of genome strains.
- We analyzed adhered and non-adhered sub-populations from in vitro culture flasks.