MCP Papers in Press. Published on May 4, 2015 as Manuscript M114.047647

Architecture of a host-parasite interface: complex targeting mechanisms revealed through proteomics

RUNNING TITLE

Proteomic definition of a host-parasite interface

Catarina Gadelha[1,2,*], Wenzhu Zhang[3], James W. Chamberlain[1], Brian T. Chait[3], Bill Wickstead[1], Mark C. Field[4]

[1]School of Life Sciences, University of Nottingham, Nottingham, UK, NG2 7UH

[2]Department of Pathology, University of Cambridge, Cambridge, UK, CB2 1QP

[3]Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, New York, USA, 10021

[4]Division of Biological Chemistry and Drug Discovery, University of Dundee, Dundee, UK, DD1 5EH

*Corresponding author. Email: catarina.gadelha@nottingham.ac.uk

SUMMARY

Surface membrane organization and composition is key to cellular function, and membrane proteins serve many essential roles in endocytosis, secretion and cell recognition. The surface of parasitic organisms, however, is a double-edged sword; this is the primary interface between parasites and their hosts, and those crucial cellular processes must be carried out while avoiding elimination by the host immune defenses. For extracellular African trypanosomes, the surface is partitioned such that all endo- and exocytosis is directed through a specific membrane region, the flagellar pocket, in which it is thought the majority of invariant surface proteins reside. However, very few of these proteins have been identified, severely limiting functional studies, and hampering the development of potential treatments. Here we used an integrated biochemical, proteomic and bioinformatic strategy to identify surface components of the human parasite *Trypanosoma brucei*. This surface proteome contains previously known flagellar pocket proteins as well as multiple novel components, and is significantly enriched in proteins that are essential for parasite survival. Molecules with receptor-like properties are almost exclusively parasite-specific, whereas transporter-like proteins are conserved in model organisms. Validation shows that the majority of surface proteome constituents are *bona fide* surface-associated proteins, and as expected, the majority present at the flagellar pocket. Moreover, the largest systematic analysis of trypanosome surface molecules to date provides evidence that the cell surface is compartmentalized into three distinct domains with free diffusion of molecules in each, but selective, asymmetric traffic between. This work provides a paradigm for the compartmentalization of a cell surface and a resource for its analysis.

INTRODUCTION

The cell surface is the major point of interaction between unicellular parasites and their surroundings, and is the site for many essential functions such as nutrient uptake, host recognition, and environment sensing. This interface, however, also represents the primary target for host immune attack. To evade adaptive immune defenses, many pathogens (including the causative agents of malaria, Lyme disease and AIDS) use some form of antigenic variation – the expression of a series of immunologically-distinct surface proteins (1, 2)). As an exclusively extracellular parasite of the blood, African trypanosomes have made a huge investment in this strategy. In the human-infective species *Trypanosoma brucei*, around ten million copies of a single variant surface glycoprotein (VSG) form a dense surface coat that protects the parasite against complement-mediated lysis. Periodic switching of the single expressed VSG gene from a vast silent library enables trypanosomes to avoid clearance by the host's adaptive immune response, prolonging infection and increasing the chances of transmission. The monoallelic expression of *VSG* is achieved through tight regulation from telomeric expression sites (ES), with only one from about 20 ES being transcriptionally active at any one time.

For the strategy of antigenic variation to work, the African trypanosome surface coat must be kept free of many essential invariant antigens that might otherwise elicit an immune response. Most of these are thought to be sequestered within a specialized region of the surface membrane at the base of the flagellum called the flagellar pocket (FP). This relatively small membrane domain is the sole site for all endocytosis and exocytosis performed by African trypanosomes, and has the highest rate of endocytosis for any system thus far observed (3, 4). Thus, the FP is a crucial interface between the parasite and host. Unsurprising, disruption of FP function by loss of the associated cytoskeleton or endocytic vesicular traffic is lethal (5-7), highlighting the potential of this host-parasite interface as a therapeutic target.

Our current understanding of FP function and its possible exploitation for therapeutic gain have been significantly inhibited by the paucity of data on its molecular composition. Mining the parasite genome for genes encoding simple characteristics of membrane-association is of limited predictive power, as a large proportion of predicted membrane proteins unlikely to be on the cell surface, and in-silico generated datasets are often not amenable to validation studies (for example, the parasite genome is predicted to encode 257 GPI-anchored proteins, 1963 transmembrane proteins, and over 7000 potentially glycosylated proteins, from a pool of only 9202 predicted proteins (genedb.org, v4)). Attempts to purify specific FP components, however, have been hampered by technical difficulties in isolation and thus far none have succeeded in providing validated, high-confidence datasets (8-11).

Only a few validated FP constituents are known to date. The first reported was the heterodimeric transferrin receptor (12) encoded by the ES associated genes 6 and 7 (*ESAG6* and *ESAG7*). Since then, only four more proteins have been specifically localized to the FP of bloodstream-form stages: the haptoglobin-hemoglobin receptor (13), an aquaporin (14), a hypothetical protein identified by proteomics of flagellar fractions (15), and a protein associated with differentiation (16). These components likely represent only a tiny subset of the FP proteome. Here, we address this knowledge gap using a comparative, semi-quantitative approach for the high-confidence identification of cell surface proteins in bloodstream-form *Trypanosoma brucei*. By creating a new genetic toolkit for endogenous locus tagging of membrane proteins, we validate our proteomic set by localization of 25 putative surface molecules of unknown function. As well as demonstrating the location for many novel FP components, we show that individual proteins access different combinations of cell surface membrane domains, and present a bioinformatics analysis of sorting signals. From these data, we propose a new model for the domain organisation of the *T. brucei* surface.

EXPERIMENTAL PROCEDURES

Isolation of surface membrane proteins. We used bloodstream-form *Trypanosoma brucei* Lister 427 expressing VSG221 (BES1/MITat 1.2/VSG427-2/TAR 40), as monitored by immunofluorescence microscopy using an affinity-purified polyclonal antibody anti-VSG221. $5x10^8$ mid-log phase cells were harvested by centrifugation and resuspended at $2x10^8$ cells ml^{-1} in PBS (10mM PO_4, 137mM NaCl, 2.7mM KCl, pH 7.5) plus 20mM glucose. Cells were held on ice while pulsed with 500µM fluorescein-hexanoate-NHS (referred hereafter to as fluorescein) dissolved in DMSO and HPG buffer (20mM HEPES pH 7.5, 140mM NaCl, 20mM glucose). Pulse duration was 15 minutes on ice, during which time cells remained actively motile and morphologically normal (as assessed by light microscopy). Fluorescence microscopy showed fluorescein to be exclusively associated to the parasite cell surface [Figure 1B]. At the end of this period, unreacted fluorescein was blocked by the addition of TBS (25mM Tris-HCl pH 7.5, 150mM NaCl) plus 0.25% w/v glycine, and removed by washing cells in TBS plus 20mM glucose. Fluorescein-labeled cells were lysed with 2% v/v Igepal CA-630 and 2% w/v CHAPS in the presence of protease inhibitors (5µM E-64d, 2mM 1,10-phenanthroline, 50µM leupeptin, 7.5µM pepstatin A, 500µM phenylmethylsulfonyl fluoride, 1mM EDTA, 1mM DTT) and 200µg ml^{-1} DNAse I, and centrifuged at 20,000g for 30 minutes to separate soluble labeled proteins from the insoluble fraction. To increase identification sensitivity towards less abundant surface membrane proteins, we included a VSG-depletion step by affinity chromatography, for which a polyclonal antibody anti-VSG221 was generated (please see below). The soluble fraction was allowed to bind to 8mg polyclonal antibody anti-VSG221 conjugated to protein G-sepharose 4 fast flow (GE Healthcare). Then the soluble fraction partially-depleted of VSG was allowed to bind to 30mg protein G-Dynabeads (Invitrogen) cross-linked to 400µg polyclonal antibody anti-fluorescein for 1 hour, after which period unbound material was collected as flow-through and beads were washed several times in the presence of high salt and detergent (500mM NaCl, 0.02% v/v Tween-20). Bound proteins were

deglycosylated native *on column* with 1000U of PNGase F for 1 hour before acid then basic elutions in 0.2M glycine pH 2.5 and 0.2M triethanolamine pH 11 respectively. To control for non-specific binding to anti-fluorescein column, a parallel isolation was carried out with unlabeled cells. To account for possible cell lysis during the surface labeling step, $1x10^8$ cells were subjected to hypotonic lysis by resuspension in 20mM Hepes pH 7.5 in the presence of the protease inhibitors aforementioned for 30 minutes at room temperature, and then pulsed with fluorescein as above.

Mass spectrometry. Proteins in the final eluate were precipitated with cold acetone, solubilized in Laemmli buffer, and treated with 1M iodoacetamide to alkylate reduced cysteines. Proteins were resolved by SDS-PAGE using pre-cast gels and standard techniques. Post-electrophoresis gels were stained with SyproRuby (Life Technologies) for imaging, or Coomassie blue for band excision. Mass spectrometry analysis of proteins that were digested in-gel was performed on an LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Scientific).

Label-free quantitation of mass spectrometry results. mzXML data files were uploaded onto the Central Proteomics Facilities Pipeline (release version 2.1.1; www.proteomics.ox.ac.uk), which uses Mascot, X!Tandem and OMSSA search engines. The data were searched with the following peptide modifications: fluorescein (K), acetylation (protein N-terminus), carbamidomethylation (C), oxidation (M), and deamidation (N/Q). Note that we expect the majority of peptides, even those derived from fluoresceinated proteins, not to contain the fluorescein modification, as only a few lysines in any given surface protein would be accessible. Precursor mass tolerance was set at 20 ppm, MS/MS fragment ion tolerance at 0.5 Da, and number of missed cleavages permitted at 2. Searches were performed against a custom, non-redundant trypanosome protein sequence database combining predicted protein sequences from TREU927 and Lister 427 genomic data [tritrypdb.org], with the inclusion of ES and VSG sequences (17, 18), and containing in total 20,195 entries. The resulting peptide

identifications from each search engine were validated with PEPTIDEPROPHET and PROTEINPROPHET and lists compiled at the peptide and protein level. IPROPHET was used to combine the identifications from three search engines and further refine identifications and probabilities. Normalized spectral index quantitation (SINQ) was applied to the grouped meta-searches to give protein-level quantitation between labeled samples and controls (19). All lists of peptide and protein identifications were generated with a probability cut-off corresponding to 1% false discovery rate (FDR) relative to a target decoy database. Only proteins identified with 2 or more spectra were considered for further analysis.

Bioinformatics. Signal peptide and anchor sequences were predicted from the first 70aa of each coding sequence by a stand-alone implementation of SignalP v3.0b (20, 21) using the hidden Markov model methodology, 'eukaryotic' settings and thresholds of $p \geq 0.9$. For GPI-anchor prediction, to reduce false positives, proteins were considered only if they were a PredGPI hit (22) with false-positive rate ≤ 0.1 and also had SignalP peptide prediction with $p \geq 0.7$ (since only proteins directed to the endoplasmic reticulum are processed for anchor addition). Transmembrane domains were predicted using TMHMM v2.0c (23, 24).

Generation of a genetic toolkit for membrane protein localization. A vector for specific tagging of GPI-anchored protein genes, named pSiG, was created by *de novo* synthesis (MrGene, Invitrogen). pSiG contains an epitope tag and fluorescent protein flanked by a signal peptide and GPI-anchor insertion sequences (derived from VSG221) up- and downstream respectively. A derivative for tagging of transmembrane protein genes, pSiS, was created by replacing the GPI-anchor insertion sequence from pSiG with a stop codon generated by annealing two primers. In these vectors, part of the targeted ORF and its UTR, at either the N- or C-terminus, is cloned in frame with the epitope tag/fluorescent protein, then the plasmid is linearized for transfection and replacement of the endogenous gene fragments. Hence, the sites for targeting the specific locus are supplied by the user along with the site for linearization. The

constructs contain convenient restriction sites on either side of the fluorescent protein/epitope tag for integration of short targeting sequences. Derivatives include 9 different fluorescent proteins, 2 epitope tags and 3 selection markers. These vectors are available from the authors upon request, and their DNA sequences can be found on the authors' webpage (www.catarinagadelha.com/resources).

Endogenous-locus tagging. ESPs and ESAGs predicted to encode transmembrane proteins were tagged at the C-terminus, while those predicted to contain a GPI anchor were tagged at the N-terminus (due to lack of robustness of prediction algorithms). For N-terminal tagging, PCR amplicons containing ~200bp from the 5'-end UTR (untranslated region) and ~200bp from the N-terminal end of the CDS (coding sequence) of interest were cloned together into the XbaI-BamHI sites downstream of the fluorescent protein ORF in pSiG, such that the N-terminal end of the CDS was in frame with the fluorescent protein. For C-terminus tagging, PCR amplicons containing ~200bp from the 3'-end UTR and ~200bp from the C-terminal end of the CDS of interest were cloned together into the HindIII-AvrII sites upstream of the epitope tag sequence in pSiS, such that the C-terminal end of the CDS was in frame with the fluorescent protein. In the same step, a NotI linearization site was introduced between the UTR and CDS. Integration of these constructs at the targeted endogenous locus results in transgenic lines in which one allele of the CDS of interest contains fluorescent protein at its N-/C-terminus, but both 5'- and 3'-UTRs are identical to untagged copy. Vectors (~10µg) were linearized by digestion with NotI restriction endonuclease and transfected into single-marker bloodstream form *T. brucei* (25) using an Amaxa Nucleofector 2b device, followed by selection of stable transformants with 5µg ml^{-1} hygromycin. Correct integration was assessed by diagnostic PCR from genomic DNA of clonal transformants (not shown) and also immunoblotting of cell lysates separated by SDS-PAGE against a mixture of two anti-GFP monoclonals (7.1 and 13.1; Roche)

at 800ng ml^{-1} in 1% w/v skimmed milk in TBS, followed by 80ng ml^{-1} horseradish peroxidase-conjugated goat anti-mouse immunoglobulins.

Analysis of integration into VSG221 expression site. Whole-chromosome-sized DNAs were prepared as described elsewhere (26). Agarose-embedded DNA was digested with SmiI endonuclease and subjected to pulsed-field gel electrophoresis in a contour-clamped homogeneous electric field electrophoresis apparatus (CHEF-DR III; Biorad), loading DNA from $1.7x10^7$ cells per lane. DNA separation was performed in 1% agarose in TB[0.1]E (90 mM Tris-borate, 0.2 mM EDTA, pH 8.2) held at 14°C for 20 hours at 5.2 V cm^{-1} with switching time ramped linearly 2-10 seconds and an included angle of 120°. DNA gels were stained in ethidium bromide and prepared for transfer by UV nicking (80 mJ, 250 nm UV) followed by equilibration in 0.4 M NaOH, 1.5 M NaCl and then transferred to positively-charged nylon membrane by capillary transfer in the same solution. After transfer, membranes were neutralised with 0.5 M Tris-HCl (pH 7) and cross-linked (120 mJ, 250 nm UV). Fluorescein-labelled probes were generated by random priming from unlabelled *GFP*, *HYG* and *VSG221* coding sequences. Denatured template DNA (100 ng) were incubated for 5 hours at 37°C with 0.1 mM dATP, dCTP, dGTP, 0.67 mM dTTP, 0.33 mM Fluorescein-dUTP, 2 µM random heptamers and 5 U Klenow fragment. Hybridisation was performed overnight in 1% w/v SDS, 5% w/v dextran sulfate, 10% v/v blocking solution (Roche), 750 mM NaCl, 75 mM sodium citrate (pH 7) at 60°C. Blots were washed to a stringency of 0.1% SDS w/v, 30 mM NaCl, 3 mM sodium citrate (pH 7) at 62°C. Hybridised probe was detected with anti-fluorescein alkaline phosphatase-conjugated antibody and chemiluminescence. For reprobing, membranes were stripped with hot 0.3% w/v SDS plus 0.3 M NaOH.

Protein localization. For analysis of localization of tagged proteins by native fluorescence, cells were harvested from mid-log phase cultures, washed twice in PBS plus 20mM glucose, allowed to adhere onto derivatized glass slides for 2 minutes (at density of

$2x10^7$ cells ml^{-1}), fixed for 10 minutes in 2.5% w/v formaldehyde, counter-stained with 5µg ml^{-1} concanavalin A (ConA; it binds to α-D-mannose and α-D-glucose moieties associated to VSG and possibly other surface proteins) conjugated to tetramethylrhodamine isothiocyanate (TRITC) for 20 minutes, and mounted in a solution containing DAPI and a photostabilizing agent (1% w/v 1,4-Diazabicyclo[2.2.2]octane, 90% v/v glycerol, 50mM sodium phosphate pH 8.0, 0.25mg ml^{-1} 4',6-diamidino-2-phenylindole).

Generation and purification of polyclonal antiserum. A fragment encoding residues 27-384 of VSG221 (Tb427.BES40.22) was amplified by PCR from *T. brucei* Lister 427 genomic DNA and cloned in frame into the bacterial expression vector pQE-30 (Qiagen) to allow expression of the coding sequence fragment fused to an N-terminal 6xHis tag. Expression of recombinant protein was induced in M15[Rep4] *Escherichia coli* (Qiagen) and protein was subsequently isolated from cleared, sonicated bacterial lysates by nickel-affinity chromatography by standard methods. 200µg of recombinant VSG221 was used as immunogen in rabbits. Reactive antiserum was purified by binding to recombinant protein coupled to CNBr-activated sepharose beads, washed extensively with PBS and eluted with 0.2M glycine pH 2.5 followed by 0.2M triethanolamine pH 11. Affinity-purified polyclonal antibodies were dialysed against PBS and concentrated by ultrafiltration.

Immunoblotting of surface membrane protein isolation fractions. Immunoblots to test the purification procedure (Fig. 1C) were performed with the following polyclonal antisera: anti-ISG65 (kind gift from Mark Carrington, University of Cambridge, UK), anti-TfR (Piet Borst, The Netherlands Cancer Institute, Netherlands), anti-p67 and anti-BiP (James Bangs, University at Buffalo (SUNY), USA).

RESULTS

Chemical modification of the cell surface

A mechanistic understanding of the interface between African trypanosomes and their mammalian host requires the identification and characterization of the FP molecular composition. As the FP membrane is contiguous with the membranes of both the cell body and the flagellum, it is extremely challenging to isolate pocket proteins through classical cell fractionation procedures. To address this problem, we devised a workflow to specifically isolate cell surface proteins and generate a validated dataset of bloodstream-form cell surface constituents of *Trypanosoma brucei*. Our strategy is summarized in Figure 1A and starts with the chemical modification (fluoresceination) of the surface of live cells held at low temperature (0°C). Under these conditions, recycling of the surface coat and endocytosis are blocked, but chemical tags are still able to access proteins at both the plasma membrane (~90% of which being VSG (27, 28)) and also the FP lumen [Figure 1B]. Labeled cells were then solubilized and fluoresceinated surface proteins purified by affinity chromatography. The purification method was optimized by a VSG depletion step to increase sensitivity of detection of less abundant surface proteins [Figure 1A and Supplemental Figure 1], and on-column enzymatic removal of N-glycans to improve mass spectrometry identification of glycosylated surface proteins [Figure 1A and Supplemental Figure 1]. Finally, to allow more efficient solubilization of membrane proteins and increase dynamic range, the sample was resolved by SDS-PAGE [Supplemental Figure 1], and gel regions subjected to tandem mass spectrometry (GeLC-MS-MS). The final eluate was enriched in an invariant surface glycoprotein (ISG65) that localizes to the cell surface, and the low-abundance transferrin receptor (ESAG6 subunit) which is found in the FP [Figure 1C]. High-abundance markers of internal compartments, specifically the abundant luminal ER chaperone BiP and the LAMP-like lysosomal protein p67, were either greatly reduced or undetectable [Figure 1C].

Semi-quantitative comparative mass spectrometry defines a *Trypanosoma brucei* surface proteome

Our surface protein preparation is anticipated to contain many FP proteins as well as those localized more generally to the cell surface and early/recycling endosomes. Many FP components, however, are expected to be present at only tens or hundreds of copies per cell, as seen for the haptoglobin-haemoglobin receptor (13). This necessitates highly sensitive detection, but also the exclusion of inevitable contaminating proteins. To identify proteins specifically enriched in our surface protein preparation, we used a label-free semi-quantitative mass spectrometry approach against two controls: i) to account for non-specific binding to affinity chromatography columns, we carried out parallel isolations with unlabeled cells; and ii) to account for cell lysis during the chemical modification, in which the fluorescein tag would access internal proteins as well as those at the surface, a further control was made by labeling hypotonically lysed cells. We then compared the integrated spectral intensities from mass spectrometry of material isolated from labeled vs. control preparations, allowing for removal of contaminants through testing for signal enrichment in the labeled sample [Figures 1D].

Across all preparations and replicates, we detected 1683 uniquely distinguishable proteins (each being represented by two or more detectable peptides). The full list of hits and their respective integrated spectral intensities is provided in Supplemental Table 1. The most abundant protein in bloodstream-form *T. brucei* cells is VSG (27) and, as expected, VSG221 (MITat 1.2/VSG427-2) expressed from the active ES is detected in all preparations. However, its signal is highly enriched (80x) in labeled samples versus controls [Figure 2A], despite being deliberately depleted in these preparations [Figure 1A and Supplemental Figure 1]. We also observed, ~10^3 times less abundantly, several other VSGs including those in other telomeric ESs, likely representing rare cells in the parasite population which have undergone switch events. Along with these “true” VSGs, a number of VSG-related proteins (transcribed from

chromosome internal locations (29)) are also enriched in the preparation, representing the first evidence that this family of proteins is translated in bloodstream form parasites and that they are surface-associated.

Our procedure enriches for VSG, several ISGs and proteins known to be localized specifically to the cell surface membrane [Figure 2A]. Importantly, low abundance FP proteins, such as HpHbR are also detected in these experiments and are highly enriched (250x) in the labeled preparation. Analysis of the features or annotations of enriched proteins compared to all those detected showed a substantial over-representation of those with predicted signal peptide or glycosylphosphatidylinositol (GPI) anchor, as well those with annotations for “VSG”, “ESAG” or “ISG” [Figure 2B]. Conversely, annotations associated with ribosomes, mitochondrion or cytoskeleton motors are under-represented in the enriched cohort [Figure 2B], as are proteins detected as part of the *T. brucei* flagellar proteome (30) or glycosomal proteome (31). Interestingly, we also find an under-representation of proteins with predicted transmembrane (TM) domains in labeled preparations [Figure 2B]. This may suggest that a significant fraction of TM proteins in trypanosomes are expected to be associated with internal membranes, although this may also reflect less efficient chemical modification of multipass proteins with few extracellular lysines (see Discussion).

These data show that known surface, FP and flagellum membrane proteins are substantially enriched in our chemically modified preparations. Using enrichment analysis, 307 and 650 uniquely distinguishable protein hits were identified with 50- or 5-fold enrichment in the labeled sample when compared to controls [Figure 1D]. However, these sets are unlikely to represent only genuine membrane-associated proteins. To further improve discrimination between true surface proteins and contaminants, we applied a bioinformatic filter to create sets representing only those proteins with predicted signal-peptide or signal-anchor sequence, GPI-anchor addition sites or TM domains [Supplemental Figure 2]. This is equivalent to intersecting

our enrichment datasets with bioinformatic prediction of membrane-association as used by Jackson et al. (32), and constitute “high-confidence” sets that have support from both methods. This procedure is also analogous to the approach used to analyze the trypanosome nuclear envelope and identify nuclear pore complex components (33), and here identified 82 or 175 uniquely distinguishable putative surface proteins at 50x or 5x enrichment thresholds respectively [Figure 1D]. The full list of these sets is given in Supplemental Table 2. These sets represent hits with a high likelihood of being genuine surface proteins, and identified proteins include known FP components, VSGs and ISGs, as well as proteins with predicted function as transporters (Tb427.04.4830, Tb427tmp.02.0630, Tb427.03.4630, Tb427.08.2380, Tb427.08.3620, Tb427.04.4860, Tb427.08.650, Tb427.08.2160), permeases (Tb427.05.3390) and channels (Tb427.10.11680). We herein refer to the 5x-enriched, high-confidence set of 175 putative surface membrane proteins as the *T. brucei* bloodsteam surface proteome (TbBSP).

Most TbBSP proteins are true parasite cell surface components

Having demonstrated an efficient enrichment of known FP proteins and related annotation in the labeled dataset, we next sought to robustly test our TbBSP dataset by directly interrogating the cellular location of multiple protein hits of unknown localization, and looking for specific signal at the FP. A set of 25 candidates were selected from the high-confidence sets for further characterization using the following criteria: i) they were annotated as “hypothetical” proteins for which no functional data had been previously reported for *T. brucei* at the start of this work; ii) they represented the range of general protein topologies detected, e.g. predicted GPI-anchored proteins, type I and type II TM proteins, and multipass TM proteins; and iii) they included proteins with enrichment ranging from 5 to >6000 times and spanning >3 orders of magnitude of mass spectrometry signal intensity. Figure 3 shows the enrichment and architectures of these candidates, and Supplemental Table 3 provides their accession numbers and predicted features.

Chimeric proteins were created by integration of tagging constructs at endogenous gene loci. Tagging cell surface proteins is potentially complicated by requirements for signaling sequences at both amino and carboxyl termini, and issues with folding of fluorescent proteins targeted through the ER. To overcome these problems, we created two new series of vectors specifically designed for the endogenous-locus tagging of genes encoding GPI-anchored and non-GPI-anchored sequences containing N-terminal signal sequences. These vectors are called the pSiG and pSiS series, respectively [Supplemental Figure 2] and include the incorporation of a 'superfolder' GFP (or derivatives) with improved folding dynamics and greater resistance to the reducing environments encountered in the ER lumen or extracellular space compared with conventional GFP variants (34), plus an epitope tag (HA). The pSiG/pSiS series also include processing signals (trypanosome signal peptide or GPI-anchor addition sequences), providing a means to rapidly and accurately tag surface proteins at either N- or C-terminus [Supplemental Figure 2]. These vectors provide the correct FP localisation of previously analysed proteins, for example, either GPI-anchored or non-anchored subunits of the transferrin receptor [see Figure 7]. Moreover, the toolkit does not force a non-TbBSP protein (ESAG1) onto the cell surface [Supplemental Figure 1].

The 25 selected genes encoding candidate surface-associated proteins, designated as 'enriched in surface-labeled proteome' (ESP) proteins (ESP1-25), were tagged at their endogenous loci using the vectors described above. Correct integration of the tagging construct and expression of fusion proteins was assessed with immunoblotting of whole-cell extracts [Supplemental Figure 3A]. Since genes are tagged by integration at the endogenous loci, it is expected that protein expression levels will be close to those for wild-type protein and, consistent with this, different fusion proteins were expressed at different levels. Two tagged proteins (ESP4 and ESP7) did not show a detectable signal on Western blots, and were not pursued further.

For the 23 fusion proteins with detectable expression, 12 were clearly present at the FP membrane as assessed by native fluorescence [Figure 4 and Supplemental Figure 4]. These 12 proteins localized either exclusively to the FP (ESP1, 6, 10 and 11) or in addition to another surface domain – for example, five proteins localized to the FP and endosomal system (ESP12, 14, 19, 21, and 22), while ESP8 localized to the FP and the junction between the cell body and the flagellum membranes (the flagellum attachment zone). In addition to these 12 FP proteins, ESP13 and 24 were present across the entire cell surface (FP, flagellum and cell body) and a further four ESPs were predominantly localized to endosomes (ESP5, 9, 15, 20). This is expected, since the endosomal membrane is in constant flux with the cell surface and proteins with clear FP function, such are TfR, maintain a steady-state concentration in early/recycling endosomal compartments (35, 36). Likewise, ISGs are equally distributed between endosomes and FP/cell surface (37). Therefore, these four predominantly endosomal ESPs are likely to be transiently present at the FP, albeit at low abundance, and are thus enriched in our chemical modification procedure. ESP17 and ESP18 were found at both the cell body membrane and an intracellular compartment tentatively interpreted as the lysosome. The remaining five proteins (ESP2, 3, 16, 23, 25) localized elsewhere in the cell and may represent contaminants, although mislocalisation due to tagging cannot be excluded [Supplemental Figure 4]. Overall, experimental validation by cellular localization of 23 ESPs shows that we have identified 18 novel membrane proteins on the parasite cell surface, the majority of which reside at the FP (exclusively or in combination with another surface membrane domain).

Diversification of parasite surface architecture

ESPs at the FP may represent promising therapeutic targets due to their exposure and potential roles in modulating essential parasitic processes, but only if those proteins are sufficiently different to host ones. To map the evolutionary distribution of ESPs, we asked if orthologs could be detected in organisms representing a wide taxonomic diversity of

eukaryotes, including humans, and for which complete or near-complete genome sequences were publicly available. Phylogenetic analysis show that most ESPs are specific to African trypanosomes and closely related parasites [Figure 5]. This provides evidence for a lineage-specific architecture for the surface membrane of kinetoplastid cells, reflecting their shared ancestry and biological similarities. Striking, however, was the finding that ESPs predicted to be GPI-anchored are often restricted to *T. brucei*, while type I and II TM proteins tend to be conserved in all kinetoplastids (both intra- and extracellular parasites) [Figure 5]. This distribution suggests specific protein evolution to match distinct selective pressures encountered by these parasites, such as mechanisms of survival, host immune invasion and transmission. In contrast, many of the multipass TM ESPs are from families conserved right across eukaryotes [Figure 5] and, thus, may have arose early in eukaryotic evolution. This likely reflects the expected hierarchy of conservation, with essential transporters being more evolutionarily constrained.

9 out of 12 ESAGs encode surface-associated proteins

The first FP component identified was the TfR previously mentioned, encoded by the expression site-associated genes (*ESAG*s) 6 and 7. There are 12 distinct families of *ESAG*s (*ESAG*1 to 12) that are co-transcribed with the active VSG gene from one of ~20 telomeric expression sites (ESs). Some or all *ESAG*s may be present in a particular ES (17), and most have chromosome-internal paralogs known as genes-related to *ESAG* (*GRESAG*s). Only a few other ESAGs have been characterized in detail in *T. brucei*: ESAG8 is a protein of unknown function that has been localized to the nucleus (38, 39), while ESAG4 is an adenylate cyclase localized to the flagellum membrane (40), and whose activity has been associated with control of parasitaemia (41). GRESAG9 is specifically expressed and secreted by the quiescent 'stumpy' bloodstream-form stage (42). Finally, an ESAG specific to the subspecies *T. b. rhodesiense* – the serum resistance associated gene, or SRA – confers resistance to a

trypanolytic factor associated with the heavy density lipoprotein found in normal human serum (43-46).

Given that ESAGs are co-expressed with the active VSG during infection, they are believed to play roles in parasite survival in the human host. All but two ESAGs (ESAG8 and ESAG12) are predicted to encode a signal peptide sequence, a GPI-anchor insertion site, or a TM domain, suggesting that they may be associated to the surface membrane or secreted proteins, but for the majority this has not been tested. Significantly, seven ESAGs (ESAG2, 4, 5, 6, 7, 10 and 11) are present in our high confidence datasets [Figure 6B], but the remainder were not. We took this finding, and the genetic tools developed here, as an opportunity to investigate the cellular localization of all ESAGs and to test our surface proteome for false negatives (i.e. true surface proteins not detected in our set). We tagged every *ESAG* present in the active ES of bloodstream-form trypanosomes used in this study [BES1, Figure 6A]. Only pseudogenes of *ESAG5* and *ESAG11* are present in this ES, and *ESAG9* and *ESAG10* are absent (17). Since chromosome-internal copies of ESAG5, 10 and 11 were highly enriched in our surface proteome, these were also targeted for protein fusions. GRESAG9 has previously been shown not to be expressed in proliferative bloodstream-form parasites (42), and was not pursued here.

Correct tagging of the active ES copy was confirmed by Southern blotting [Supplemental Figure 5] and ESAG fusion proteins were assessed for correct tagging by immunoblotting [Supplemental Figure 3B], and localized by native fluorescence microscopy [Figure 7]. Significantly, all ESAGs detected in our surface proteome localize to the surface membrane. ESAG6/ESAG7 localized to the FP and ESAG4 localized to the FP and flagellum membranes, as previously described (12, 41). Other surface proteome ESAGs localized to the cell body membrane (ESAG2), cell body and FP (ESAG10), cell body and flagellum (ESAG11), or FP and endosomes (ESAG5). With respect to those ESAGs not detected or not enriched in our surface proteome, ESAG12 was detected in endosomes, consistent with being also at the surface at low

levels and/or recycling through the endomembrane and surface compartments. ESAG8 was expressed at levels close to the limit of detection by immunoblot when tagged at either end of the endogenous ES copy, and was undetectable in localization experiments. Importantly, ESAG1 and ESAG3 – which contain signal sequences suggestive of possible surface-association, but which were not enriched in the surface proteome – did not localize to the cell surface when tagged. These data demonstrate that i) tagging with our vectors does not cause non-TbBSP proteins to mis-localize to the surface and ii) ESAG1 and 3 are unlikely to be surface-associated. Hence, of the 12 ESAG proteins, nine are shown to be surface-associated or secreted, five of which present at the FP membrane, clearly arguing for direct roles in host-parasite interactions by virtue of being exposed to the host environment.

Protein localization suggests distinct functional membrane domains maintained by selective barriers

The trypanosome surface can be conceptually divided into three regions of contiguous membrane: the FP, the flagellum membrane and the cell body. Our localization data, using the same tag with 14 hypothetical proteins and 6 ESAGs that clearly target the cell surface membrane, allowed us probe for the existence of these or other membrane domains with the largest set of trypanosome surface proteins systematically tested to date. Individual proteins in our sets were found to be restricted to any one of these domains or to combinations of them [Figure 4 and Supplemental Figure 4], suggesting that the three regions indeed act as specialized domains of surface membrane, divided by selective barriers. Notably, we found only one example of sub-localization within a region (for ESP8), indicating that most proteins have free diffusion within each of the surface membrane domains.

A polarized distribution of ESPs and ESAGs implies intrinsic protein-sorting signals governing location on the cell surface. We therefore analyzed this set for the presence of common sequence motifs or structure which might regulate such sorting; However, no simple

correlation between cellular localization and protein architecture emerged. For example, predicted GPI-anchored proteins were not all restricted to the FP, nor were type I TM proteins restricted to the cell body membrane [Figure 8]. Furthermore, motif elicitation analysis (MEME) detected no common motifs among ESPs and ESAGs with shared localization (data not shown). This suggests that protein topology alone may not be the primary determinant of surface domain segregation in *T. brucei*, and more complex interactions are at play.

DISCUSSION

A surface proteome for African trypanosomes

Here we describe a high-confidence, validated surface proteome for the major host form of African trypanosome parasites. This was achieved through a novel biochemical preparation in which the use of fluorescein was one of several steps optimized to increase both the specificity and sensitivity of our approach. Cell surface proteomic studies of other human pathogens, as well as mammalian cells, have frequently used the biotin-avidin based system to isolate plasma membrane proteins (47-50). In initial experiments we too used sulpho-NHS-biotin to chemically modify the surface of live trypanosomes. However, following affinity chromatography with streptavidin, we found the specificity of the approach was compromised by high background from control (unlabelled) cells, which could not be removed even on extensive washing. This may be a product of the parasite's intrinsic biochemistry: trypanosomatids (except those harboring bacterial endosymbionts) are unable to synthesize biotin (51); but this vitamin is an essential requirement for cell growth (52), and known to be incorporated into endogenous proteins (53). To avoid contamination with endogenously-biotinylated parasite proteins, we abandoned biotin as a chemical tag, and moved to fluorescein labeling combined with an antigen-antibody purification system. Fluorescein is cell-impermeable, ensuring that only surface membrane proteins from intact cells are labeled by covalent modification of accessible lysine residues, and antigen-antibody columns can be washed to high stringency. Fluorescein

also has an advantage that it can be followed visually or by fluorimetry during preparations. Using this approach, we have developed here a strategy for the identification of surface-exposed membrane proteins, which in trypanosomes isolates proteins that, at steady-state, reside at the FP, early/recycling endosomes, flagellum and cell body membranes.

Our surface proteome was extracted from a specific biochemical preparation coupled with comparative semi-quantitative mass spectrometry and bioinformatic filters. The bioinformatic methods used decrease the risk of contaminants in the defined TbBSP in a manner analogous to those used to describe the high-quality set of nucleoporins that compose the trypanosome nuclear pore complex (33). In that study, an initial set of 757 mass spectrometry hits was reduced by removing 448 contaminants on the basis of functionally unrelated sequence homology and gene annotations (e.g. ribosomal, endoplasmic reticulum and cytosolitc proteins). The remaining 309 proteins were informatically filtered for features associated with known nucleoporins (such as functional motifs, molecular weight and predicted secondary structure) (33). Here we filtered our experimental data for sequences that predict targeting to the endoplasmic reticulum and membrane anchoring (either via a GPI anchor or a transmembrane domain).

The contrasting approach of interrogating the entire genome sequence for cell surface localization on the basis of bioinformatic prediction of membrane-association is not applicable to our question because ~15% of the parasite's predicted proteome (1465 proteins) have such features. To define a predicted Cell Surface Phylome, Jackson and colleagues (32) combined this approach with sequence clustering to look more specifically at those putative membrane-associated proteins in multigene families. Of the 50 CSP families present in *T. brucei*, 20 are detected in the surface proteome [Supplemental Figure 6]. Particularly well-represented are Fam10 and Fam79 [Supplemental Figure 6], which comprise proteins of unknown function for which we present the first experimental evidence. For example, of the 7 members in Fam10,

five were detected in our dataset, and we have demonstrated the surface association of one (ESP17). Significantly, however, the majority of proteins (63%) in the surface proteome are not part of multigene families (and hence not part of the CSP), yet are *bona fide* surface-associated proteins according to our validation experiments (12 out of 18 ESPs). This highlights the strength of our joint approach of sensitive, semi-quantitative detection and bioinformatic filtering.

Extent of the surface proteome

The surface proteome includes almost all previously characterized surface proteins for *T. brucei* (albeit rather few in number), as well as hypothetical proteins with predicted function as receptors, transporters, channels and others. These data suggest that the overall coverage of surface proteins in our high confidence set is broad, although it is to be expected that it will not be complete. A natural limitation of our approach is that it only derivatizes surface components with regions of modifiable polypeptide chain exposed to the extracellular space. This excludes proteins solely associated with the cytoplasmic side of the membrane. Hence, proteins modified by N-terminal palmitoyl- or myristoyl-ation(such as for the flagellum calcium-binding protein calflagin (54)) are not expected to be present. Such proteins were not the focus of this work, as our primary objective was to gain knowledge of the molecular components exposed at the host-parasite interface.

A more significant cohort of proteins that may be underrepresented in our surface proteome are those with few exposed extracellular lysine residues. This may explain why aquaporins 2 and 3 (shown to localize to the FP and cell body membranes of *T. brucei* (14)) are not present in the TbBSP. Neither is a putative calcium channel protein (FS179/Tb927.10.2880) localized to the region of flagellum attachment to the cell (15)]. These proteins are multipass TM proteins and are predicted to have limited sequence on the extracellular side of the membrane (e.g. aquaporins have only three lysines predicted to be extracellular, which may or may not be accessible to fluoresceination depending on the folding of the protein). A number of transporters

and channel-like proteins are present in the surface proteome (10/175 proteins in total) – and validation showed that five detected multipass TM proteins are indeed surface-associated – but it is noteworthy that proteins with predicted TM domains were under-represented in our preparations (Figure 2).

Confidence of surface prediction

We believe that a specific strength of the present work is the robust validation. Alongside bioinformatic support, we also developed a genetic toolkit to test a subset of 25 candidates for FP/surface localization. The majority were true surface components (14 out of 23 detectable fusion proteins were present at the cell surface, 4 found predominantly in endosomal compartments that are likely to cycle to the surface in small amounts, while 5 localized elsewhere in the cell). This suggests that our surface proteome contains relatively few false-positives (~22% at the >5x threshold, and likely far fewer at greater enrichment values).

A number of ESAGs were present in our surface proteome and were localized to the cell surface when tagged, compared to only 1 out of 4 (ESAG12) not present in the TbBSP (in spite of containing sequence characteristics that might have suggested surface proteins). Although this is only a small set, it does indicate that the levels of false-negatives in our analysis (i.e. proteins that should have been detected, but were not) is also proportionally low. It is improbable that our surface proteome contains all proteins resident at the parasite surface, but results from localization of hypothetical proteins and ESAGs indicate a high confidence for the 175 proteins identified herein.

One issue with the interpretation of localization data for ESPs and ESAGs is in defining where the cell surface ends. Most of the proteins tested were detectable by light microscopy at locations in the cell consistent with being the FP, flagellum or cell body membranes. However, the plasma membrane is highly dynamic and is in constant exchange with components of the endosomal system. In trypanosomes, TfR, ISG and VSG are all present in endosomes as well

as at the cell surface. It is thus possible that some of the TbBSP proteins not localized to cell surface domains are still molecules that are found transiently or in low abundance at the cell surface. In mammalian cells it is common to find many proteins cycling between the cell surface and early/recycling endosomes, but proteins as 'deep' as those found in lysosomes have also been observed on the surface (55-57). African trypanosomes too have a transport route for newly-synthesized lysosomal membrane glycoproteins to exit the Golgi and reach the lysosome via the FP (58), though the lysosomal marker p67 may take a direct route that bypasses the FP membrane (59). Hence, it may be biologically meaningful that proteins such as ESP15 (a type I TM protein that localized to the lysosome) is in the surface proteome, whereas p67 (also a type I TM glycoprotein) is not.

Membrane domains and domain maintenance

The few surface proteins analyzed to date suggest the existence of at least three biochemically distinct domains across contiguous membranes, and emphasize the idea that individual proteins can access one or more domains on the cell surface. For example, TfR is restricted to the FP and endosomes, the adenylate cyclase encoded by ESAG4 is present at the FP and flagellum, and VSG is distributed across the entire surface membrane and endosomal system. The work here considerably expands these observations, showing that 8 ESAGs and 14 proteins of unknown function localize to one or more of three separate membrane domains: the FP, the flagellar membrane, the cell body. Moreover, these proteins do so in all possible combinations (with the exception of flagellar membrane alone, which was not observed).

Our results support a model whereby trypanosome surface organisation is determined by control of access to any of three membrane domains. The finding that only one surface protein (ESP8) showed evidence of sub-domain localization suggests that diffusion within each domain is essentially free for most components. However, selective diffusion barriers or very rapid transfer systems exist between these domains. Since newly synthesized proteins are delivered

to the FP, most combinations could be produced by the "opening" of symmetrical barriers at either the base of the flagellum (to access the flagellar membrane) or distal end of the FP (to access the cell body membrane). Nonetheless, the existence of proteins that are enriched in at just the cell body (ESP17, ESP18 and ESAG2) or cell body plus flagellar membrane (ESAG11) suggests that for at least some of the surface proteins the barriers or protein movement must be asymmetric.

This model raises major questions with regards to the mechanisms underlying protein sorting and retention in African trypanosomes, and elucidating such mechanisms in any cell type remains a formidable challenge. We considered that common motifs within the primary sequence might be used to target ESPs and ESAGs to their respective domains or enable them to cross specific domain boundaries, but simple common signals were not found in our analyses. It is also clear that gross protein architecture (e.g. GPI-anchor, type I TM, etc.) is not predictive of domain localization, suggesting that the signals are encoded by more complex or protein-specific cues.

The barriers to protein movement on the cell surface are likely to be contained in the structural features described at the boundaries between the domains – the rows of intramembrane particles seen by freeze-fracture electron microscopy forming the ciliary necklace at the junction of the flagellum and FP membranes, and the junction of the FP and neck membrane (60). The molecular identity of these particles remains unknown, but a morphologically similar configuration identified at the base of the mammalian primary cilium requires the GTPase septin for retention of receptors in that organelle (61). Alternatively, lipid composition, particularly that able to accommodate the geometric constraints of highly curved membrane sections (like that at the junction of the flagellum and the FP) could act as barriers to protein movement or as targeting signal. The distribution pattern of membrane probes and GPI-

anchored YFP between the ciliary and plasma membranes are consistent with lipid composition operating in this manner (62).

For two trypanosome membrane proteins lateral movement between surface domains appears to be dependent on protein abundance as well as identity. Over-expression of a membrane-bound acid phosphatase predominantly found in endosomes causes it to re-distribute over the whole cell surface (36). In a similar manner, TfR in excess of normal levels is no longer retained in FP and endosomes, and escapes to the entire cell surface (35). The relevance of such artificial over-expression to endogenous protein targeting is uncertain, but trypanosomes grown in serum with low-affinity transferrin compensate by up-regulating the expression of TfR which, in turn, escapes the FP (35). However, it is clear that surface domain targeting in trypanosomes must be more complex than just a saturable mechanism of FP retention, as has been proposed for TfR, since we observe proteins with localizations specific to each individual domain, and combinations thereof – including proteins excluded from the FP (e.g. ESAG2), from most of the cell body (e.g. ESP8) or from the flagellum membrane (e.g. ESAG10).

Unraveling the host-parasite interface

With a cell body entirely covered by ten million copies of a single glycoprotein, cellular functions that would normally occur at the plasma membrane of a typical eukaryotic cell are here concentrated at the FP of trypanosomes. The restriction of endocytosis and secretion to a focal point on the parasite surface allows for invariant receptors, channels and transporters, and other signaling molecules to be sequestered in an environment that is protected from the attention of host defenses, while the cell body membrane is mostly denuded of those proteins. Sitting at the critical interface between host and parasite, it is surprising that so few components of the FP have been described prior to this study. The essential nature of receptors such as TfR and HpHbR highlights the FP as an area of vulnerability that could be exploited in a therapeutic

context. Our work has expanded this portfolio to 12 novel FP components with proven localization and identifies a total of 175 in the surface proteome, >50% of which are estimated to also be FP proteins. Importantly, 60% of surface proteome components cause a significant loss-of-fitness when knocked down individually (50/83 genes covered in a large-scale RNAi library screen (63)) compared to 42% for all genes (p=0.001), showing that the TbBSP is notably enriched in genes essential for growth in the bloodstream. Since these proteins are mostly parasite specific and exposed to the extracellular space, our surface proteome is a potential source of drugable targets for disease treatment and control.

The high-confidence surface proteome described here greatly increases our knowledge of the trypanosome surface, and provides a significant resource against which hypothesis about membrane protein sorting and retention might be tested. Moreover, the methods we described are widely applicable to the study of cell membrane composition in human pathogens in general; while the surface compartmentalization is significant for understanding trypanosome biology and an important paradigm for surface organization in other systems.

ACKNOWLEDGEMENTS

This work was supported by the MRC (project grant G0900255 to CG and MCF), the Royal Society (research grant RG130195 to CG), the BBSRC (new investigator award BB/J01477X/1 to BW), the National Institutes of Health (PHS GM103314 and GM103511 grants to BTC and WZ) and the Wellcome Trust (program grant 090007/Z/09/Z to MCF). The authors thank Ben Thomas (University of Oxford) for help with semi-quantitative analysis of mass spectrometry data; Jim Haseloff (University of Cambridge) for superfolder fluorescent protein constructs; and Mark Carrington (University of Cambridge) and Bob Lloyd (University of Nottingham) for access to fluorescence microscopes. We also thank Catherine Lindon (University of Cambridge), Flavia Moreira-Leite (University of Oxford), and David Horn (University of Dundee) for critical reading of the manuscript.

NOTE

During this project, Woods and Oberholzer (and collaborators) reported on the localization of ESP8 (which the authors named FLA3) and ESP10 (therein named FS133), respectively (15, 64).

REFERENCES

1. Morrison L. J., Marcello L. and McCulloch R. (2009) Antigenic variation in the African trypanosome: molecular mechanisms and phenotypic complexity. *Cell Microbiol.* 11, 1724-1734

2. Burton D. R., Poignard P., Stanfield R. L. and Wilson I. A. (2012) Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science* 337, 183-186

3. Engstler M., Thilo L., Weise F., Grunfelder C. G., Schwarz H., Boshart M. and Overath P. (2004) Kinetics of endocytosis and recycling of the GPI-anchored variant surface glycoprotein in *Trypanosoma brucei*. *J. Cell Sci.* 117, 1105-1115

4. Engstler M., Pfohl T., Herminghaus S., Boshart M., Wiegertjes G., Heddergott N. and Overath P. (2007) Hydrodynamic flow-mediated protein sorting on the cell surface of trypanosomes. *Cell* 131, 505-515

5. Allen C. L., Goulding D. and Field M. C. (2003) Clathrin-mediated endocytosis is essential in *Trypanosoma brucei*. *EMBO J.* 22, 4991-5002

6. Garcia-Salcedo J. A., Perez-Morga D., Gijon P., Dilbeck V., Pays E. and Nolan D. P. (2004) A differential role for actin during the life cycle of *Trypanosoma brucei*. *EMBO J.* 23, 780-789

7. Bonhivers M., Nowacki S., Landrein N. and Robinson D. R. (2008) Biogenesis of the trypanosome endo-exocytotic organelle is cytoskeleton mediated. *PLoS Biol.* 6, e105

8. McLaughlin J. (1987) *Trypanosoma rhodesiense*: antigenicity and immunogenicity of flagellar pocket membrane components. *Exp. Parasitol.* 64, 1-11

9. Grab D. J., Webster P., Ito S., Fish W. R., Verjee Y. and Lonsdale-Eccles J. D. (1987) Subcellular localization of a variable surface glycoprotein phosphatidylinositol-specific phospholipase-C in African trypanosomes. *J. Cell Biol.* 105, 737-746

10. Olenick J. G., Wolff R., Nauman R. K. and McLaughlin J. (1988) A flagellar pocket membrane fraction from *Trypanosoma brucei rhodesiense*: immunogold localization and nonvariant immunoprotection. *Infect. Immun.* 56, 92-98

11. Nolan D. P., Jackson D. G., Windle H. J., Pays A., Geuskens M., Michel A., Voorheis H. P. and Pays E. (1997) Characterization of a novel, stage-specific, invariant surface protein in *Trypanosoma brucei* containing an internal, serine-rich, repetitive motif. *J. Biol. Chem.* 272, 29212-29221

12. Salmon D., Geuskens M., Hanocq F., Hanocq-Quertier J., Nolan D., Ruben L. and Pays E. (1994) A novel heterodimeric transferrin receptor encoded by a pair of VSG expression site-associated genes in *T. brucei*. *Cell* 78, 75-86

13. Vanhollebeke B., De Muylder G., Nielsen M. J., Pays A., Tebabi P., Dieu M., Raes M., Moestrup S. K. and Pays E. (2008) A haptoglobin-hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. *Science* 320, 677-681

14. Baker N., Glover L., Munday J. C., Aguinaga Andres D., Barrett M. P., de Koning H. P. and Horn D. (2012) Aquaglyceroporin 2 controls susceptibility to melarsoprol and pentamidine in African trypanosomes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10996-11001

15. Oberholzer M., Langousis G., Nguyen H. T., Saada E. A., Shimogawa M. M., Jonsson Z. O., Nguyen S. M., Wohlschlegel J. A. and Hill K. L. (2011) Independent analysis of the flagellum surface and matrix proteomes provides insight into flagellum signaling in mammalian-infectious *Trypanosoma brucei*. *Mol. Cell Proteomics* 10, M111 010538

16. Dean S., Marchetti R., Kirk K. and Matthews K. R. (2009) A surface transporter family conveys the trypanosome differentiation signal. *Nature* 459, 213-217

17. Hertz-Fowler C., Figueiredo L. M., Quail M. A., Becker M., Jackson A., Bason N., Brooks K., Churcher C., Fahkro S., Goodhead I.et al. (2008) Telomeric expression sites are highly conserved in *Trypanosoma brucei*. *PLoS One* 3, e3527

18. Cross G. A., Kim H. S. and Wickstead B. (2014) Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Mol. Biochem. Parasitol.* 195, 59-73

19. Trudgian D. C., Ridlova G., Fischer R., Mackeen M. M., Ternette N., Acuto O., Kessler B. M. and Thomas B. (2011) Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics* 11, 2790-2797

20. Nielsen H., Engelbrecht J., Brunak S. and von Heijne G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1-6

21. Nielsen H. and Krogh A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6, 122-130

22. Pierleoni A., Martelli P. L. and Casadio R. (2008) PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9, 392

23. Sonnhammer E. L., von Heijne G. and Krogh A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 6, 175-182

24. Krogh A., Larsson B., von Heijne G. and Sonnhammer E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567-580

25. Wirtz E., Leal S., Ochatt C. and Cross G. A. (1999) A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* 99, 89-101

26. Melville S. E., Leech V., Gerrard C. S., Tait A. and Blackwell J. M. (1998) The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. *Mol. Biochem. Parasitol.* 94, 155-173

27. Jackson D. G., Owen M. J. and Voorheis H. P. (1985) A new method for the rapid purification of both the membrane-bound and released forms of the variant surface glycoprotein from *Trypanosoma brucei*. *Biochem. J.* 230, 195-202

28. Grunfelder C. G., Engstler M., Weise F., Schwarz H., Stierhof Y. D., Boshart M. and Overath P. (2002) Accumulation of a GPI-anchored protein at the cell surface requires sorting at multiple intracellular levels. *Traffic* 3, 547-559

29. Marcello L. and Barry J. D. (2007) Analysis of the VSG gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.* 17, 1344-1352

30. Broadhead R., Dawe H. R., Farr H., Griffiths S., Hart S. R., Portman N., Shaw M. K., Ginger M. L., Gaskell S. J., McKean P. G.et al. (2006) Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature* 440, 224-227

31. Colasante C., Voncken F., Manful T., Ruppert T., Tielens A. G., van Hellemond J. J. and Clayton C. (2013) Proteins and lipids of glycosomal membranes from *Leishmania tarentolae* and *Trypanosoma brucei*. *F1000Res.* 2, 27

32. Jackson A. P., Allison H. C., Barry J. D., Field M. C., Hertz-Fowler C. and Berriman M. (2013) A cell-surface phylome for African trypanosomes. *PLoS Negl Trop Dis* 7, e2121

33. DeGrasse J. A., DuBois K. N., Devos D., Siegel T. N., Sali A., Field M. C., Rout M. P. and Chait B. T. (2009) Evidence for a shared nuclear pore complex architecture that is conserved from the last common eukaryotic ancestor. *Mol Cell Proteomics* 8, 2119-2130

34. Pedelacq J. D., Cabantous S., Tran T., Terwilliger T. C. and Waldo G. S. (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* 24, 79-88

35. Mussmann R., Janssen H., Calafat J., Engstler M., Ansorge I., Clayton C. and Borst P. (2003) The expression level determines the surface distribution of the transferrin receptor in *Trypanosoma brucei*. *Mol. Microbiol.* 47, 23-35

36. Engstler M., Weise F., Bopp K., Grunfelder C. G., Gunzel M., Heddergott N. and Overath P. (2005) The membrane-bound histidine acid phosphatase TbMBAP1 is essential for endocytosis and membrane recycling in *Trypanosoma brucei*. *J. Cell Sci.* 118, 2105-2118

37. Leung K. F., Riley F. S., Carrington M. and Field M. C. (2011) Ubiquitylation and developmental regulation of invariant surface protein expression in trypanosomes. *Eukaryot. Cell* 10, 916-931

38. Lips S., Geuskens M., Paturiaux-Hanocq F., Hanocq-Quertier J. and Pays E. (1996) The esag 8 gene of *Trypanosoma brucei* encodes a nuclear protein. *Mol. Biochem. Parasitol.* 79, 113-117

39. Hoek M., Engstler M. and Cross G. A. (2000) Expression-site-associated gene 8 (ESAG8) of *Trypanosoma brucei* is apparently essential and accumulates in the nucleolus. *J. Cell Sci.* 113 (Pt 22), 3959-3968

40. Paindavoine P., Rolin S., Van Assel S., Geuskens M., Jauniaux J. C., Dinsart C., Huet G. and Pays E. (1992) A gene from the variant surface glycoprotein expression site encodes one of several

transmembrane adenylate cyclases located on the flagellum of *Trypanosoma brucei*. *Mol. Cell. Biol.* 12, 1218-1225

41. Salmon D., Vanwalleghem G., Morias Y., Denoeud J., Krumbholz C., Lhomme F., Bachmaier S., Kador M., Gossmann J., Dias F. B.et al. (2012) Adenylate cyclases of *Trypanosoma brucei* inhibit the innate immune response of the host. *Science* 337, 463-466

42. Barnwell E. M., van Deursen F. J., Jeacock L., Smith K. A., Maizels R. M., Acosta-Serrano A. and Matthews K. (2010) Developmental regulation and extracellular release of a VSG expression-site-associated gene product from *Trypanosoma brucei* bloodstream forms. *J. Cell Sci.* 123, 3401-3411

43. Rifkin M. R. (1978) *Trypanosoma brucei*: some properties of the cytotoxic reaction induced by normal human serum. *Exp. Parasitol.* 46, 189-206

44. Hajduk S. L., Moore D. R., Vasudevacharya J., Siqueira H., Torri A. F., Tytler E. M. and Esko J. D. (1989) Lysis of *Trypanosoma brucei* by a toxic subspecies of human high density lipoprotein. *J. Biol. Chem.* 264, 5210-5217

45. Xong H. V., Vanhamme L., Chamekh M., Chimfwembe C. E., Van Den Abbeele J., Pays A., Van Meirvenne N., Hamers R., De Baetselier P. and Pays E. (1998) A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*. *Cell* 95, 839-846

46. Vanhamme L., Paturiaux-Hanocq F., Poelvoorde P., Nolan D. P., Lins L., Van Den Abbeele J., Pays A., Tebabi P., Van Xong H., Jacquet A.et al. (2003) Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* 422, 83-87

47. Roesli C., Borgia B., Schliemann C., Gunthert M., Wunderli-Allenspach H., Giavazzi R. and Neri D. (2009) Comparative analysis of the membrane proteome of closely related metastatic and nonmetastatic tumor cells. *Cancer Res.* 69, 5406-5414

48. de Miguel N., Lustig G., Twu O., Chattopadhyay A., Wohlschlegel J. A. and Johnson P. J. (2010) Proteome analysis of the surface of *Trichomonas vaginalis* reveals novel proteins and strain-dependent differential expression. *Mol. Cell Proteomics* 9, 1554-1566

49. Niehage C., Steenblock C., Pursche T., Bornhauser M., Corbeil D. and Hoflack B. (2011) The cell surface proteome of human mesenchymal stromal cells. *PLoS One* 6, e20399

50. Biller L., Matthiesen J., Kuhne V., Lotter H., Handal G., Nozaki T., Saito-Nakano Y., Schumann M., Roeder T., Tannich E.et al. (2014) The cell surface proteome of *Entamoeba histolytica*. *Mol. Cell Proteomics* 13, 132-144

51. Klein C. C., Alves J. M., Serrano M. G., Buck G. A., Vasconcelos A. T., Sagot M. F., Teixeira M. M., Camargo E. P. and Motta M. C. (2013) Biosynthesis of vitamins and cofactors in bacterium-harbouring trypanosomatids depends on the symbiotic association as revealed by genomic analyses. *PLoS One* 8, e79786

52. Roitman C., Roitman I. and de Azevedo H. P. (1972) Growth of an insect trypanosomatid at 37C in a defined medium. *J. Protozool.* 19, 346-349

53. dos Santos A. L., Abreu C. M., Batista L. M., Alviano C. S. and de Araujo Soares R. M. (2001) Cell-associated and extracellular proteinases in Blastocrithidia culicis: influence of growth conditions. *Curr. Microbiol.* 43, 100-106

54. Emmer B. T., Souther C., Toriello K. M., Olson C. L., Epting C. L. and Engman D. M. (2009) Identification of a palmitoyl acyltransferase required for protein sorting to the flagellar membrane. *J. Cell Sci.* 122, 867-874

55. Lippincott-Schwartz J. and Fambrough D. M. (1986) Lysosomal membrane dynamics: structure and interorganellar movement of a major lysosomal membrane glycoprotein. *J. Cell Biol.* 102, 1593-1605

56. Harter C. and Mellman I. (1992) Transport of the lysosomal membrane glycoprotein lgp120 (lgp-A) to lysosomes does not require appearance on the plasma membrane. *J. Cell Biol.* 117, 311-325

57. Saftig P. and Klumperman J. (2009) Lysosome biogenesis and lysosomal membrane proteins: trafficking meets function. *Nat. Rev. Mol. Cell Biol.* 10, 623-635

58. Brickman M. J. and Balber A. E. (1994) Transport of a lysosomal membrane glycoprotein from the Golgi to endosomes and lysosomes via the cell surface in African trypanosomes. *J. Cell Sci.* 107 (Pt 11), 3191-3200

59. Alexander D. L., Schwartz K. J., Balber A. E. and Bangs J. D. (2002) Developmentally regulated trafficking of the lysosomal membrane protein p67 in *Trypanosoma brucei*. *J. Cell Sci.* 115, 3253-3263

60. Gadelha C., Rothery S., Morphew M., McIntosh J. R., Severs N. J. and Gull K. (2009) Membrane domains and flagellar pocket boundaries are influenced by the cytoskeleton in African trypanosomes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 17425-17430

61. Hu Q., Milenkovic L., Jin H., Scott M. P., Nachury M. V., Spiliotis E. T. and Nelson W. J. (2010) A septin diffusion barrier at the base of the primary cilium maintains ciliary membrane protein distribution. *Science* 329, 436-439

62. Vieira O. V., Gaus K., Verkade P., Fullekrug J., Vaz W. L. and Simons K. (2006) FAPP2, cilium formation, and compartmentalization of the apical membrane in polarized Madin-Darby canine kidney (MDCK) cells. *Proc. Natl. Acad. Sci. U.S.A.* 103, 18556-18561

63. Alsford S., Turner D. J., Obado S. O., Sanchez-Flores A., Glover L., Berriman M., Hertz-Fowler C. and Horn D. (2011) High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Res* 21, 915-924

64. Woods K., Nic a'Bhaird N., Dooley C., Perez-Morga D. and Nolan D. P. (2013) Identification and characterization of a stage specific membrane protein involved in flagellar attachment in *Trypanosoma brucei*. *PLoS One* 8, e52846

FIGURE LEGENDS

Figure 1. Workflow of biochemical, semi-quantitative mass spectrometry and bioinformatic methods used to identify putative cell surface proteins. A) Scheme illustrating key steps in purification. B) Micrograph of cells following chemical modification with fluorescein (live at 0°C). Native fluorescence at plasma membrane is predominantly derived from fluoresceinated VSG (which makes up ~90% of proteins at the parasite surface). DNA has been counter-stained with DAPI (magenta); the FP is indicated by yellow arrowhead. C) Immunoblots showing isolation of known surface proteins (ISG65, found on the cell surface and TfR, found in the FP) in the final purified eluate. Note faster migration of deglycosylated ISG and TfR in eluate. Common contaminants from the ER (BiP) and lysosome (p67) are highly depleted in final eluate. D) Schematic showing enrichment analysis (for exclusion of contaminants by comparison of labelled samples with controls) and bioinformatic filters (for prediction of membrane proteins

features) applied to protein identification to produce “high-confidence” sets. The numbers of unique proteins present in each set are shown in red. The high-confidence set of 175 putative surface membrane proteins enriched 5x in labelled samples is herein referred to as the *T. brucei* bloodstream surface proteome (TbBSP). Experimental replicates of protein isolation from fluorescein-labeled live cells (“Labeled”), unlabeled cells (“Unlabeled”), and fluorescein-labeled material from lysed cells (“Dead”) are indicated between brackets. See Experimental Procedures for details of protein feature prediction, and Supplemental Table 2 for bioinformatics filter abbreviations.

Figure 2. Identification of surface proteins by comparative label-free semi-quantitative mass spectrometry. A) Enrichment analysis for 1683 unique proteins (integrated spectral intensity) in labeled samples versus unlabeled and osmotically-lysed controls (see Experimental Procedures for more information). Points represent log10-transforms of total intensity (all replicates, samples and controls) against the ratio of intensity in samples versus summed controls. Points representing signal from VSGs (BES copies or from elsewhere in the genome), VSG-related proteins, ISGs, and proteins previously localized to the FP or flagellar membrane are highlighted. B) Representation of proteins with select predicted features (SignalP peptide prediction p≥0.9; PredGPI false-positive rate ≤0.1; ≥1 predicted TM domain), annotation (word match in description) or those detected in either the *T. brucei* flagellar proteome (30) or glycosomal proteome (31). Representation is the ratio of the number of hits in enriched sets versus all uniquely detected proteins.

Figure 3. Enrichment analysis showing proteins of unknown function taken for validation by localization. Dataset as in Figure 2, with points representing 25 ESPs highlighted according to predicted protein architecture.

Figure 4. Localisation of surface proteome components at the FP. ESPs were localized by tagging the gene at the endogenous locus with an ORF encoding superfolder-GFP. Images are

representative of the signal distribution observed for each cell line. Yellow: native fluorescence from superfolder-GFP; Blue: concanavalin A counterstain (ConA); Magenta: DAPI. Nuclear (n) and mitochodrial (mt) DNA contents, and FP (yellow arrowhead) are indicated.

Figure 5. The distribution of ESPs across eukaryotes. Conservation was investigated by analysis of BLAST hits in the predicted proteomes of model species from a wide range of eukaryotic lineages. Spot size represents the strength of BLAST hit (e-value). Red shows reciprocal best-BLAST hits between genomes; gray shows non-reciprocating hits.

Figure 6. Identification and validation of ESAG proteins in the TbBSP. A) Structure of the VSG221 expression site (BES1), which is active in cells used in this study. B) Distribution of ESAG proteins in the comparative label-free semi-quantitative mass spectrometry. Dataset as in Fig. 2, with points representing members of the 12 ESAG families (including predicted GRESAGs) highlighted. C) Protein architectures and localization of the ESAGs. All localization data are from ES copies, except for ESAG5, 10 and 11, which are not present in BES1) for which a detected GRESAG was used. See Experimental Procedures for details of protein feature prediction.

Figure 7. Most ESAGs encode surface-associated proteins. 9 ESAGs were localized by tagging the respective gene at the active ES (except for ESAG5, 10 and 11, for which a surface proteome GRESAG was used) with an ORF encoding superfolder GFP and imaged using native fluorescence microscopy. Signal from superfolder-GFP is shown in yellow. Cells have been counterstained with concanavalin A (ConA, blue) and DAPI (magenta). The FP is indicated by yellow arrowhead.

Figure 8. Domain architecture for surface-associated proteins in *T. brucei*. Names of proteins which were localized as part of this study are emboldened (red: localized for the first time; blue: also localized in other studies). Data from proteins that have been previously localized by equivalent tagging methods (only) are also shown for comparison (references

included in Supplemental Table 3). See Experimental Procedures for details of protein feature prediction.

Figure 1

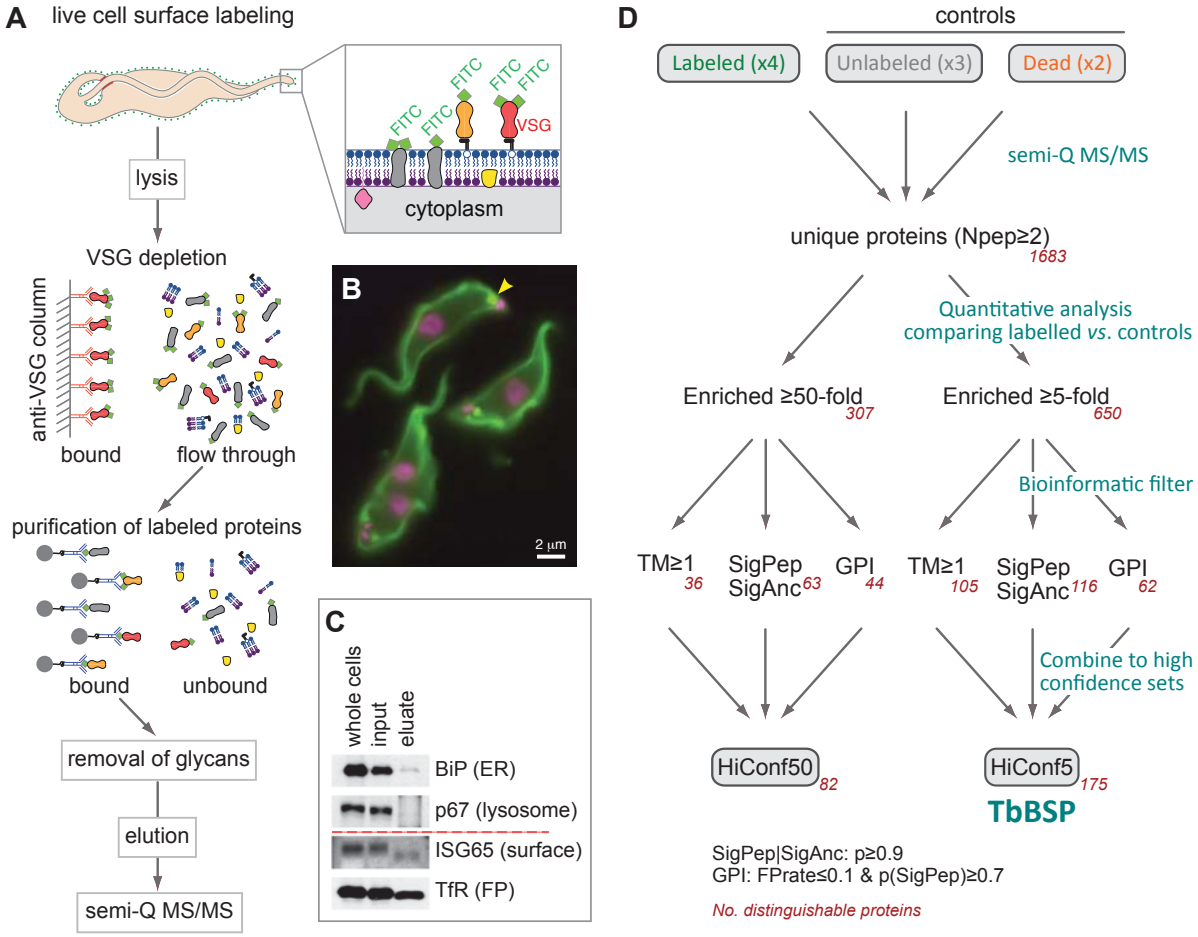
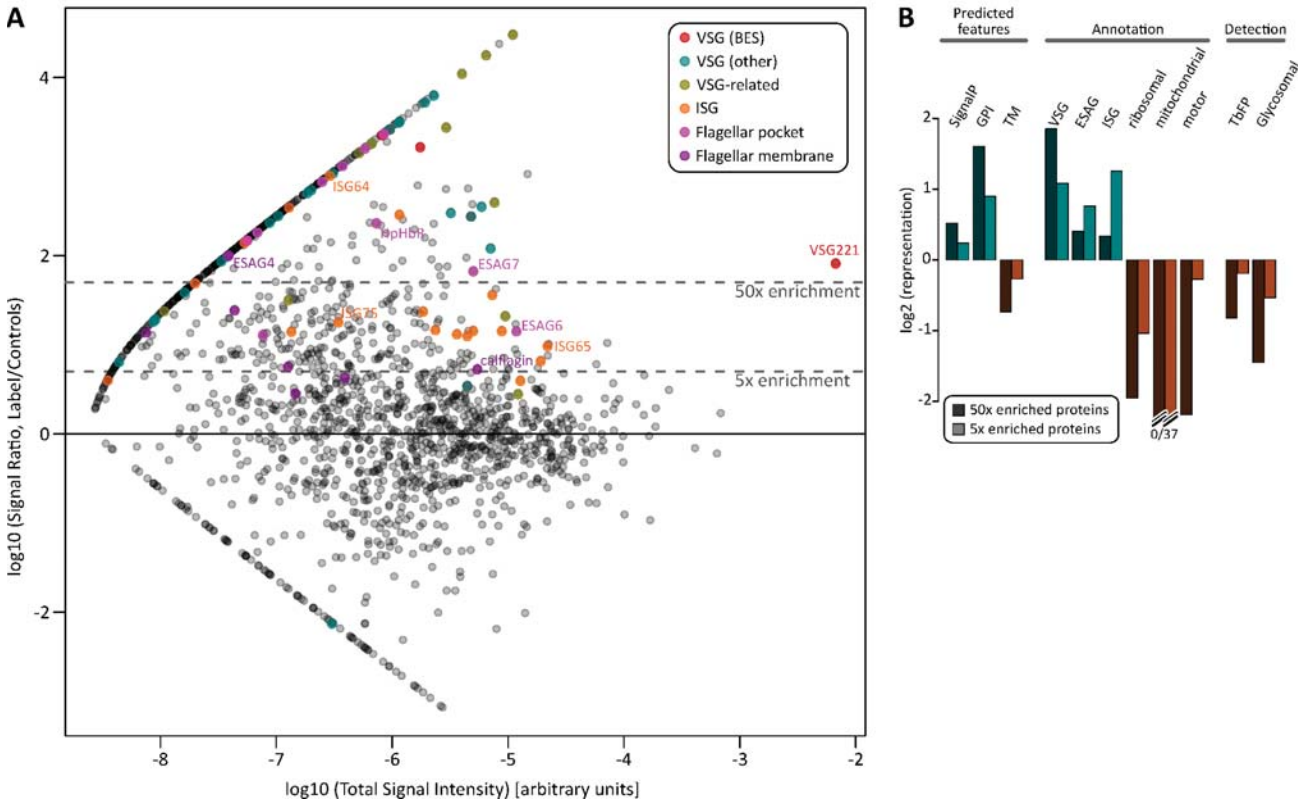


Figure 2



ASBMB
MOLECULAR & CELLULAR PROTEOMICS
MCP

Figure 3

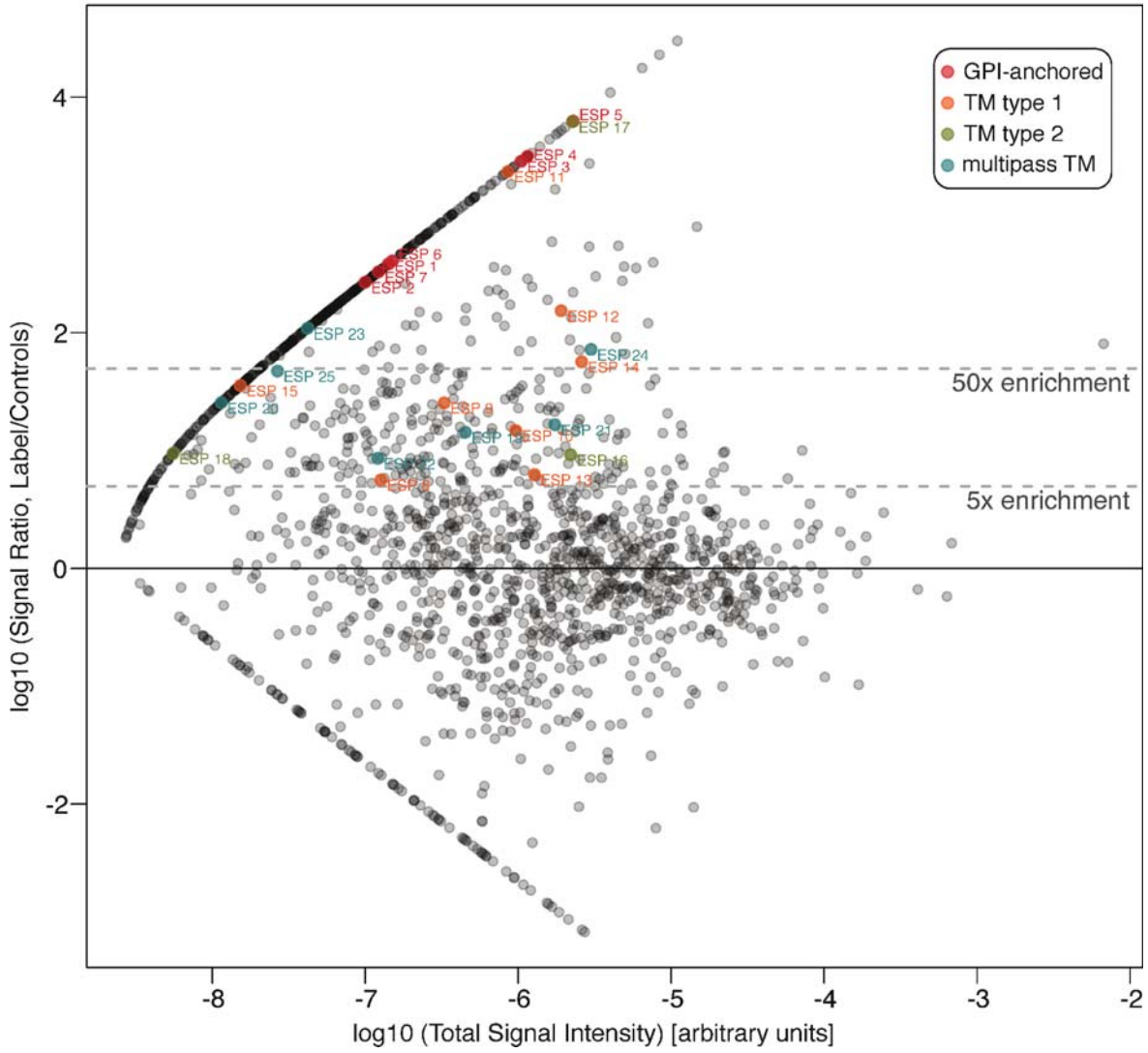
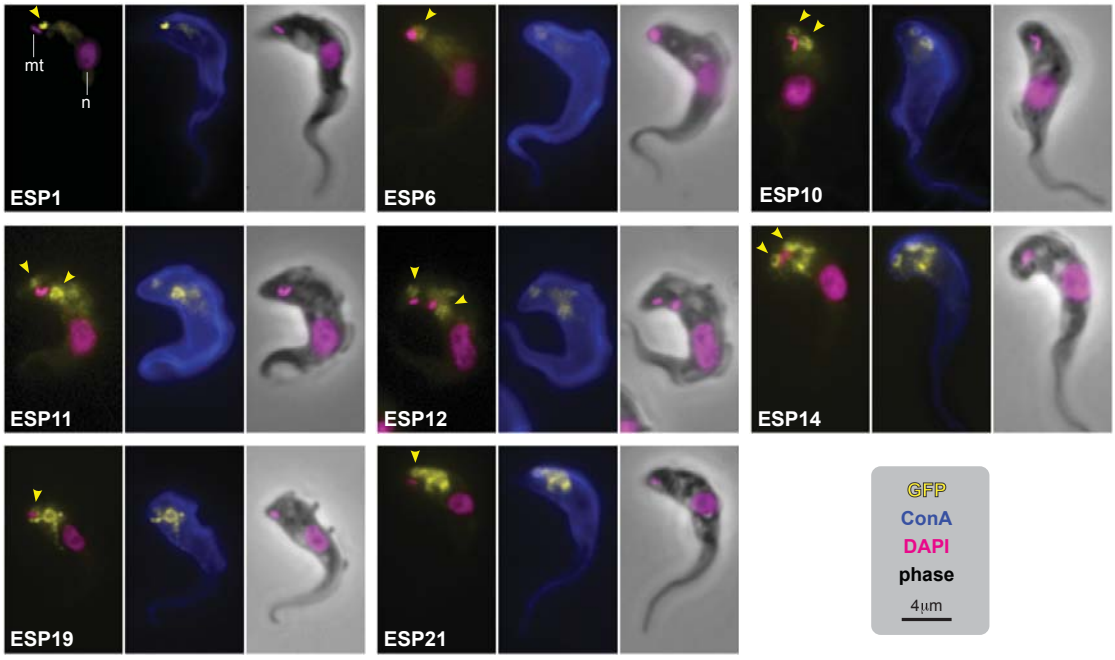


Figure 4



MCP MOLECULAR & CELLULAR PROTEOMICS ASBMB

Figure 5

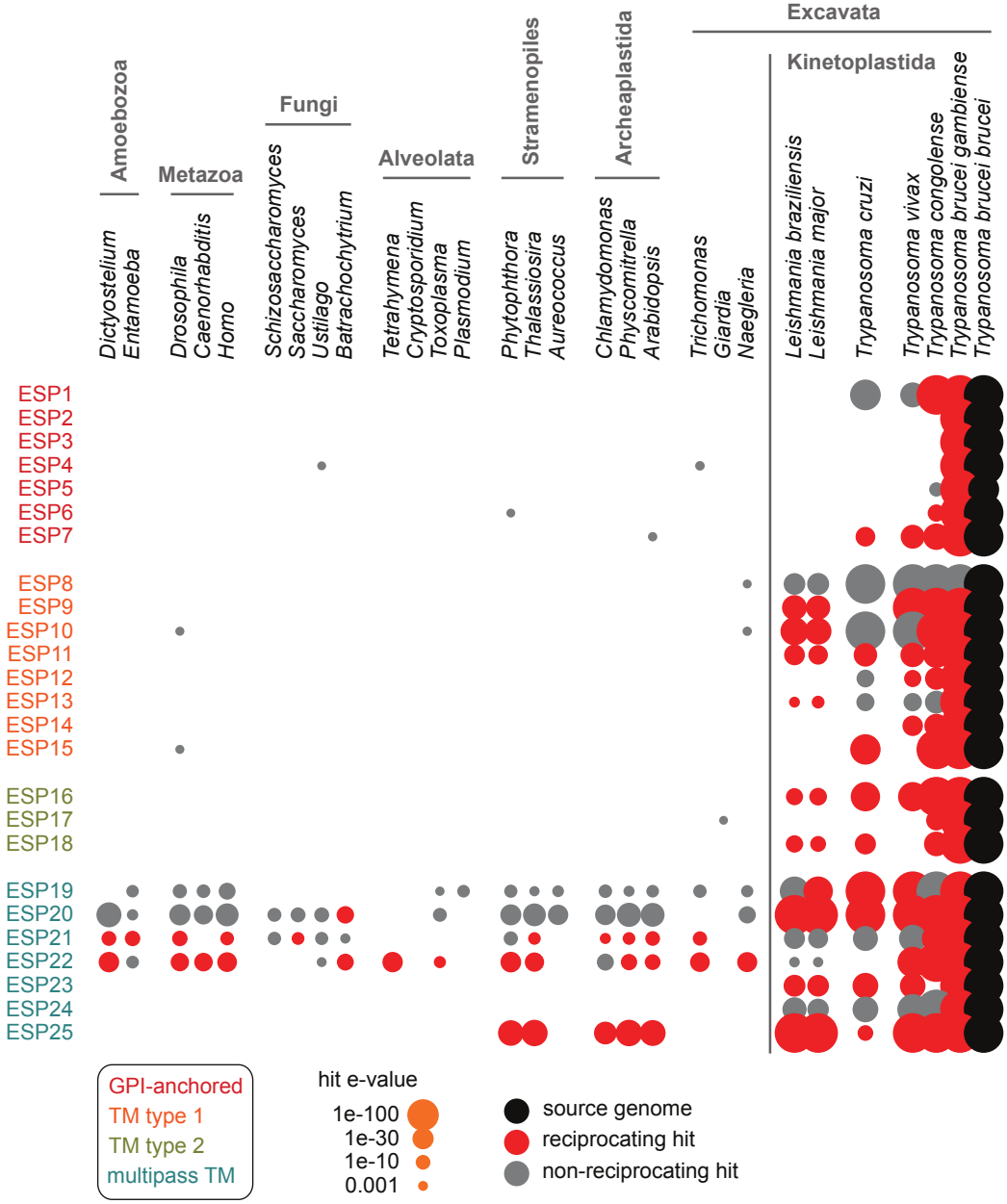


Figure 6

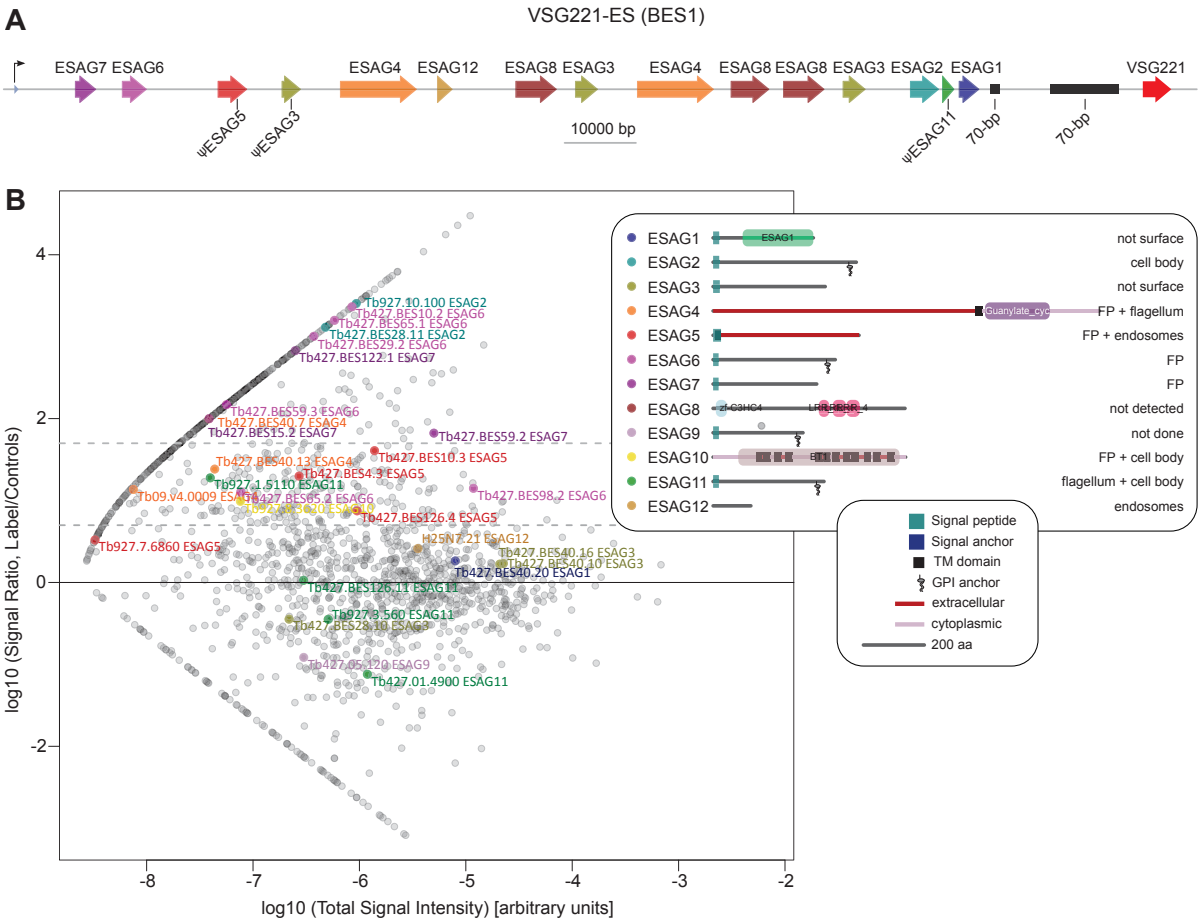


Figure 7

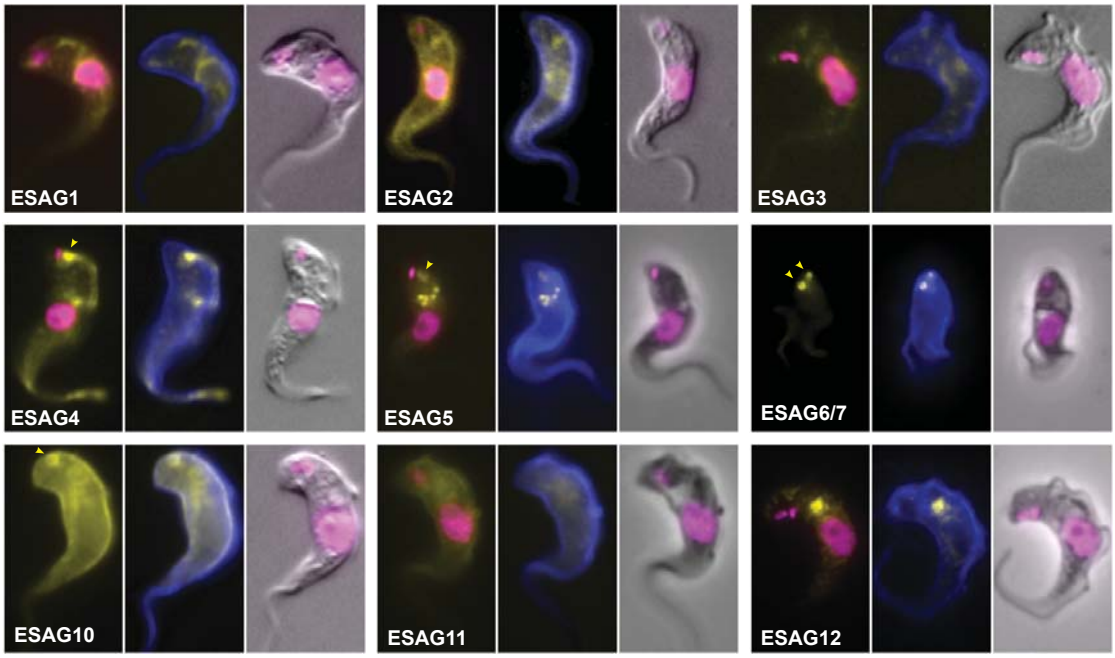


Figure 8