CellectSeq: *In Silico* Discovery of Antibodies Targeting Integral Membrane Proteins Combining *In Situ* Selections of Phage Displayed Synthetic Antibodies and Next-Generation Sequencing

Abdellali Kelil (abdellali.kelil@utoronto.ca)
University of Toronto
Eugenio Gallo
University of Toronto
Jarrett Adams
University of Toronto
Jason Moffat (j.moffat@utoronto.ca)
University of Toronto https://orcid.org/0000-0002-5663-8586
Sachdev Sidhu (sachdev.sidhu@utoronto.ca)
The Donnelly Centre, University of Toronto

Article

Keywords: In silico, Antibody, Phage-Display, Cellular Selection, NGS, Antibody Discovery, Cell-Surface Receptors, Integral Membrane Proteins, Synthetic Library, Binding Selectivity, Motif-Based Algorithm

Posted Date: August 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-48667/v1

Version of Record: A version of this preprint was published at Communications Biology on May 12th, 2021. See the published version at https://doi.org/10.1038/s42003-021-02066-5.

Abstract

Synthetic antibody (Ab) technologies are efficient and cost-effective platforms for the generation of monoclonal proteomic tools against human antigens. Yet, they typically depend on purified proteins, which exclude from interrogation integral membrane proteins that require the lipid bilayers to support their native form or function. Here, we present a novel Ab discovery strategy, termed CellectSeq, for targeting integral membrane proteins presented on native cells in complex environment. As proof of concept, we targeted the challenging tetraspanin receptor CD151, a target linked to cancer. First, we optimized *in situ* cell-based selections to enrich Ab pools for antigen-specific binders. Then, we designed novel NGS procedures to explore Ab pools diversities and abundances with enhanced accuracies. Finally, we developed novel motif-based scoring and error filtering algorithms for the comprehensive interrogation of NGS data to identify Abs with high diversities and specificities, even at extremely low abundances. We identified highly selective and diversified Abs against CD151 with abundance as low as 0.00009% for which manual sampling or identification using Abs abundances in NGS data would have been impossible. Here we show that CellectSeq enables the rapid discovery of diversified and selective antibodies against CD151, with implications for other integral membrane proteins and cell-surface receptors.

Introduction

The application of antibodies (Abs)[1] for targeting cell surface proteins has prompted the development of synthetic human Abs[2]. By this method, synthetic phage-displayed libraries containing > 10^{10} unique Abs can be constructed to rival the combinatorial diversity of natural *in vivo* immune repertoires and, in many ways, outperform natural repertoires for the production of Abs with high affinities and specificities[2, 3]. Synthetic Ab technologies have also proven amenable to automation to enable high-throughput methods of selection to target large families of soluble antigens[4–6].

However, a major limitation to both *in vitro* and *in vivo* methods for antibody generation is the difficulty of targeting multi-pass integral membrane proteins, which generally cannot be purified in a native form in the absence of a cell membrane. Integral membrane proteins remain a recalcitrant group of critical targets for Ab development due to their inherent association with the lipid bilayer, differential multi-conformational states[7, 8], and interactions with other cell surface proteins[9, 10]. Moreover, multi-pass integral membrane proteins often lack large, structured domains in their extracellular regions[11, 12], and thus, pose a particular challenge for recombinant expression and purification[13]. Given that many essential biological processes and diseases depend on integral membrane proteins, the difficulties in targeting this large subset of the human proteome is a major roadblock in many areas of biological research and drug development[14, 15].

With 33 members in the human proteome, the tetraspanin receptor family (Pfam:PF00335, Transmembrane 4 superfamily) represents a particularly interesting set of potential therapeutic targets,

as many family members are involved in processes implicated in cancer progression, including tumor proliferation, migration, and metastasis[16–18]. In particular, cluster of differentiation 151[19–23] **(CD151; Fig. 1)**, also known as PETA-3 or SFA-1, is a 30kD tetraspanin receptor that is widely expressed in normal cells and tissues (*e.g.* epithelium, endothelium, cardiac muscle, dendritic cells, and hematopoietic cells)[24], and overexpressed in diverse tumor tissues (*e.g.* lung, colon, prostate, pancreas, breast, and skin)[25–27]. Moreover, the elevated expression of CD151 is correlated with cancer patient mortality and enhanced metastasis of tumors[28, 29]. The primary role of CD151 in cancer appears to be its ability to organize the distribution and function of growth factor receptors and integrins[25, 30]. Consequently, CD151 may guide the migratory activity of tumor cells to induce invasiveness and metastasis. CD151 also modulates the pharmacological response of therapeutics that antagonize other cell surface receptors[31], and also appears to synergize and modulate intracellular signal activities in cancer. For example, integrin-associated CD151 may drive HER2 evoked mammary tumor onset and metastasis, and may enhance the activation of HER2 and other receptor tyrosine kinases by regulating dimerization[32–34]. Thus, CD151 is an integral membrane protein that may be a promising target for the development of antibodies that can antagonize the interactions mediated by its extracellular domains. However, the recalcitrant nature of CD151 receptor, due to its diminutive stature protruding only 4–5 nm above the membrane and displaying limited surface exposed regions[35], makes it challenging to target **(Fig. 1)**.

Recently, we reported optimized methods for *in situ* selections with phage-displayed synthetic antibody libraries with native antigen on live cells to develop a large panel of selective antibodies for integrin-α11/β1, a marker of aggressive tumors that is involved in stroma-tumor crosstalk[36]. Manual screening of over one thousand phage clones identified unique Abs with strong and selective binding to cells expressing integrin-α11/β1, and notably, most of these Abs did not recognize the purified antigen, suggesting that cell-based selections were essential for targeting native epitopes[36]. Moreover, next generation sequencing (NGS) analysis of the abundance of unique clones in the selection pools showed that most of the Abs identified by clonal screening were among the most abundant and enriched amongst the NGS sequences, but intriguingly, many other sequences were also identified, suggesting that clonal screening had only isolated a small subset of antigen-specific clones[36].

Here, we have further optimized *in situ* cell-based selection procedures to enrich Ab-phage pools for antigen-specific binders. In addition, we implemented a novel *in silico* analysis to efficiently explore and identify unique antigen-specific clones against CD151 in the enriched pools. In conjunction with rapid and cost-effective gene synthesis and recombinant Ab production strategies, the antigen-specific Ab-phage sequences were purified as Abs for direct assessment of cell-surface antigen recognition. We have collectively termed this methodology “CellectSeq”, which utilizes phage display, *in situ* selections, next-generation sequencing, and motif-based scoring and error filtering algorithms for the comprehensive interrogation of candidate Abs in enriched but highly diverse Ab-phage pools. We used the CellectSeq to target native CD151 displayed on cells and discovered specific anti-CD151 Abs with frequencies as low as one in a million NGS reads. Thus, we show that CellectSeq can identify rare but highly selective and

out the need for screening of individual clones

at the phage level. The technology should be applicable for the generation of Abs targeting many integral membrane proteins that have proven recalcitrant to conventional *in vivo* and *in vitro* methods.

Results

Cell-based in situ **selection for anti-CD151 Abs**

To generate Abs targeting CD151, we used the phage-displayed Library F[37] of synthetic antigen-binding fragments (Fabs) that offers advantageous features for CellectSeq **(Figure S1)**. First, Library F is extremely diverse (> 10^{10} unique members) and precisely designed to ensure that most members are stable and well-displayed on phage[38]. Second, the library proves functional for selections with either purified antigens[37] or cell-surface antigens[36] and has yielded numerous selective Abs per selections[36]. Third, the library was constructed with a single, highly stable human framework resulting in negligible display bias where most library members are presented at similar levels[37]. Also, the abundances of individual clones in pools enriched for target antigens are highly correlated with relative affinities[39]; this property enhances NGS analysis based on enrichment ranking, allowing for the identification of highly selective and high affinity clones. Fourth, the synthetic Abs are diversified at only four complementary determining regions (CDRs; H1, H2 and H3, and L3), which permit standard NGS procedures utilizing primers that anneal to common framework regions in a cost-effective manner **(Figure S2 & S3)**. Fifth, each of the four CDRs is composed of defined amino acid positions with restricted diversities. Therefore, the NGS data quality can be very accurately evaluated by assessing any deviations from the fixed framework or occurrence of unexpected codons at diversified positions. For instance, CDRs H1 and H2 contain only six or eight binary degenerate codons and offer a diversity of 64 and 256 unique sequences, respectively **(Figure S1D)**. Conversely, CDRs L3 and H3 are much more diverse in terms of loop lengths (3–7 or 1–17 degenerate codons, respectively), and in terms of sequence composition (encoded by defined ratios of nine codons encoding nine amino acids). The CDRs L3 and H3 offer a theoretical diversity of the order of 10^7 and 10^{17} unique sequences, respectively **(Figure S1D)**. Ultimately, the four CDRs combined offer a practical diversity approximating 10^{11} unique clones[37]. Thus, the highly diverse Library F, with defined length and chemical diversity encoded in CDRs L3, H1, H2, and H3, permits the precise probabilistic detection and elimination of artifactual CDR sequence combinations from NGS data, such as those derived from PCR sequence amplifications required for the NGS Illumina process[40] **(see Material and Methods)**.

We performed *in situ* selections against cell-surface CD151 on live cells, where CD151 is targeted at its native cell-surface environment. For cell engineering, we selected the HEK293T cell line because it grows rapidly in suspension and exhibits high display of transgenic cell surface proteins[41]. To enrich binders for CD151, we engineered the HEK293T cells to stably overexpress CD151 (HEK293T-CD151+; positive cells) **(Figure S4A)**. Conversely, to deplete non-target selective binders we engineered HEK293T cells that stably expressed a short hair-pin RNA that depleted CD151 mRNA, and consequently, reduced cell-surface display of CD151 (HEK293T-CD151- negative cells) (**Figure S4A**). The strategy of CellectSeq *in situ*

selections utilizes multiple rounds of selection against antigen positive and negative cells, where it aims to produce a positive Ab pool enriched with selective clones for the target antigen, and a negative Ab pool enriched with non-specific clones (background).

To this end, the naïve phage pool representing Library F was subjected to four rounds of selections with the engineered cell lines (Fig. 2). Round 1 consisted of a positive selection on HEK293T-CD151 + cells to enrich for Fab-phage that bound to CD151, followed by elution of bound phage and amplification by passage through *E. coli*. In round 2, we employed a strategy whereby phage pools were exposed to control cells HEK293T-CD151- to deplete clones that bound to other cell-surface antigens, followed by positive selections with HEK293T-CD151 + cells. Round 3 repeated the round 2 process using the amplified phage pool from round 2. For the last round the amplified phage pool from round 3 was split into two pools, and then subjected to a round 4 selection process that involved elution and amplification of phage bound to either HEK293T-CD151 + cells (positive selection) or to HEK293T-CD151- cells (negative selection) (Fig. 2). Thus, the round 4 phage selection output consisted of two pools, a positive and a negative pool. After the four rounds of selection for binding to *in situ* CD151, we manually isolated 96 random Ab-phage clones derived from the round 4 phage output of HEK293T-CD151 + cells (positive pool). We screened all 96 clones by cellular phage ELISA[42], where phage signals were measured for binding to HEK293T-CD151 + cells and compared to control HEK293T-CD151- cells. Here, we identified 49 phage clones, with binding signals 5-fold or greater over controls deemed as positive binders for cellular CD151 **(Figure S5)**. After Sanger DNA sequencing analysis, all 49 clones shared the same sequence of clone CD151-1 **(Table 1)**, indicating the Ab selection enriched for an immune-dominant clone. Accordingly, manual Ab screening failed at deriving multiple unique and diversified CD151 selective clones; consequently, we next performed NGS analysis of the output selection.

NGS enrichment ranking selection for anti-CD151 Abs

To identify unique CD151 specific Fab-phage clones in the round 4 selection output, we performed NGS analysis to explore the output diversity and relative abundance of every Ab clone. Therefore, we deep sequenced the round 4 output derived from the positive and negative pools. This allowed us to obtain CD151 selective sequences (derived from the positive pool), and non-specific background sequences (derived from the negative pool). The phage DNA from the Ab selection output pools were subjected to PCR amplification resulting in amplicons with Illumina NGS adaptor sequences and unique barcode identifiers that flanked the region of CDRs L3 and H3 **(Figure S2 & S3)**. The amplicons from each output pool (positive and negative) were quality controlled for correct size, purified, and quantified, then normalized and pooled, and finally sequenced using an Illumina HiSeq 2500 instrument **(see Materials and Methods)**. Besides the Illumina universal sequencing primers (PE1 and PE2), the NGS runs also included a custom primer that allowed for the complete sequencing of CDRs H1 and H2. Thus, the three primer reads **(PE1, PE2, and custom; Figure S2 & S3)** provided the complete sequence coverage of the four diversified CDRs in Library F[37] **(Figure S1)**. We performed duplicate NGS runs, and each run controlled for high sequence quality scores[43, 44]. The sequences were filtered from instrument sequencing

errors using per base high quality score cut-off of Q = 30, which corresponds to 1:1000 of incorrect base call[43]. Following, all sequencing reads from the duplicate NGS runs were combined and deconvoluted. The three different primer reads (PE1, PE2, and custom) for each clone were transformed into a single sequence to derive the complete synthetic Ab sequence **(see Materials and Methods)**.

The obtained high quality nucleotide sequences were then compared to the designed sequence repertoires of Library F[37] to remove technical errors inherent to Illumina sequencing and PCR amplification. For each Ab clone, the nucleotide sequences were evaluated for codon deviations from the synthetic design of the fixed framework and restricted CDR positions **(Figure S1A-B)**. Any divergent sequences from the synthetic library were discarded. Subsequently, the sequences were filtered for potential PCR-induced artifacts that may arise during the NGS sample preparation and Illumina sequencing process **(see Material and Methods)**. This may occur due to incorrect annealing amalgams (*i.e.* combinations) of different clones[40], which for our case may be driven by the fixed Ab framework coding region (non-CDR). Therefore, for every sequence we obtained the frequencies (*i.e.* number of observations) of CDRs H3 and L3, respectively, since these two CDRs are the most diversified in the synthetic library and drive the majority of affinity interactions with the antigen[37]. We then identified valid L3/H3 pairs by calculating a frequency cut-off to determine a minimal threshold of valid occurrences, with all below-threshold pairs filtered from the selection pool **(see Materials and Methods)**. Thus, we obtained 7,541,189 and 7,250,873 high quality NGS reads for the positive and negative pool, respectively. The reads were then translated into amino acid. This process ultimately yielded 23,671 and 56,352 unique amino acid sequences in the positive and negative pools, respectively.

To perform NGS Ab enrichment ranking selection of potential CD151 selective clones, the unique high-quality sequence reads from each pool were parsed based on CDR sequences and observation counts. For each unique paratope we plotted the counts in the positive pool (x-axis) versus the ratio of its abundance (*i.e.* frequency) in the positive pool relative to the negative pool (y-axis) **(Fig. 3)**. To estimate the number of potential unique CD151 binding clones in the plot, we defined an upper-right quadrant of putative binders. Here, the upper-right quadrant sequences represent observations counts of more than 200 in the positive pool, and more than four-fold enriched relative to the negative pool **(Fig. 3)**. After performing comparative analysis of the unique sequences, the NGS enrichment ranking revealed all upper-right quadrant clones as close homologs of clone CD151-1; all showing more than 80% sequence identity in both L3 and H3 sequences. This finding reveals that the Ab selection is enriched for homolog clones with a potentially similar targeted epitope (immunodominant), where CD151-1 is the most abundant and selective clone.

Motif-based algorithm identifies selective and diversified Abs against CD151

Due to the lack of Ab diversity derived by both manual selection of Ab clones and NGS enrichment ranking, we developed a novel motif-based algorithm to identify highly selective Abs for CD151 from the [illegible] oring CD151 selective Abs is based on exploring

all possible sequence motifs (*i.e.* consensus motifs) in the positive pool and scoring their enrichment over the negative pool **(Fig. 4A)**. This follows the premise that highly selective Abs are enriched with paratope motifs (*i.e.* linear information) that recognize the target antigen, whereas non-selective Abs lack such enrichment[45] **(Figure S6)**. Therefore, for each Ab clone in the positive pool (*i.e.* candidate) **(Fig. 4A1)** we explored the entire space of linear information by exhaustively enumerating all possible motifs matching its CDR sequences[46], and obtained the frequencies (number of matching sequences / total number of sequences) of every motif in the positive and negative pools **(Fig. 4A2)**. According to the premise above, the high enrichment of the motifs in the positive pool relative to the negative pool implies the Ab candidate is potentially highly selective **(Fig. 4A3)**. Thus, we analyzed each Ab in the positive pool for the selective binding to CD151 by scoring the separation between the two distributions of frequencies of the motifs in the positive and negative pools **(see Methods for details)**. To this end, we calculated the *t*-test[47] to score the separation of the two distributions, then we calculate the *p*-value to evaluate the statistical significance of the *t*-test[48–50]. Thus, the lower the *p*-value the higher is the separation between the two distributions, thus, the higher is the selectivity of the candidate Ab. Finally, we applied the stringent *p*-value cut-off of 10^{-10} to identify highly selective Ab clones **(see Materials and Methods)**. Therefore, this motif-based *in silico* strategy allowed us to explore rapidly and exhaustively the selectivity of all Ab clones in the positive pool. We were able to identify potentially selective CD151 binders, regardless of their individual frequencies in the total pool of sequences; thus, bypassing the limitations of standard NGS analyses based solely on enrichment counts of individual clone sequences, which has difficulties for discriminating between selective Ab clones and background.

Filtering PCR-induced sequence artifacts improves the in silico Ab selection results

As previously mentioned, PCR-induced artifacts may arise during the NGS sample preparation and Illumina sequencing process[40]. These artifacts represent invalid amalgams of existing CDRs L1, H1, H2, and H3 sequences, which may be seen as novel Ab clones[40]. These artifacts may significantly bias the frequencies of individual clones that will inevitably affect the *in silico* Ab discovery strategy. Therefore, for both the positive and negative pools, we obtained the frequencies (*i.e.* number of observations) of CDRs H3/L3 pairs, where both CDRs are the most diverse in terms of length and amino acid compositions in Library F[37]. We calculated a frequency cut-off to determine valid L3/H3 pairs utilizing a minimal occurrence threshold, with all invalid pairs filtered from the selection pool as potential PCR and NGS artifacts **(see Materials and Methods)**. We therefore applied the motif-based *in silico* Ab discovery strategy to predict CD151 highly selective binders (*p*-values < 10^{-10}) for both scenarios, before and after filtering. The application of error-filtering to the positive pool Abs reduced their clonal diversity to 80% less unique Abs **(Fig. 4B1-C1-D1)**. Similarly, the application of the error-filtering before the motif-based *in silico* prediction of CD151 clones reduced their diversity to 85% less unique Abs **(Fig. 4B2-C2-D2)**. Interestingly, before error-filtering the *in silico* predicted Abs clustered into 183 distinct families of similar L3/H3 sequences (> 80% identity), whereas after filtering the Abs reduced to only 4 distinct families **(Fig. 4B3-C3-D3)**.

To experimentally assess the validity of predicted antibodies in both scenarios, we selected the Abs with best specificity scores (*p*-values; **Fig. 4A3**) from each of the 4 families predicted after filtering, as well as 23 additional Abs predicted before filtering **(Fig. 4B3-C3-D3 & Table S1)**. Due to the low NGS enrichment of the identified Ab clones, instead of PCR rescue or similar methods[51, 52], all 27 candidate clones were synthesized as Ab DNA sequences into Fab protein expression plasmids. After Fab purification, we tested the activity of each clone by flow-cytometry for binding to HEK293T-CD151 + cells when compared to HEK293T-CD151- cells (control). All four Ab clones predicted after filtering were determined as CD151 binders (**Pass validation; Table S1**), with fluorescence signals of 3-fold or greater than controls. On the other hand, all 23 pre-filtering Abs failed to bind to CD151 **(Table S1)**. The success rate of the motif-based *in silico* Ab discovery before and after filtering is respectively 4:27 (*i.e.* ~15%) and 4:4 (*i.e.* 100%). This difference between the success rates highlights the requirement to filter PCR-induced and NGS artifacts to derive accurately and effectively selective clones. Furthermore, the abundance (enrichment) of the 4 identified clones (based on motif-based *in silico* Ab selection) varied from high (30%) to extremely low. In fact, the clones CD151-2 and CD151-3 have frequencies below 0.01%, and clone CD151-4 possesses the extremely low frequency of 0.00009% **(Table 1)**. These latter clones would be impossible to identify using manual sampling or standard NGS analyses solely based on enrichment.

Characterization of motif-based in silico **identified Abs against CD151**

To demonstrate the advantage of the motif-based *in silico* Ab discovery strategy, termed CellectSeq **(Figure S7)**, we measured all 4 clones (CD151-1 thru – 4) as Fab versions for dose-dependent binding to HEK293T-CD151 + cells. Quantitative flow cytometry displayed tight and saturable binding of each Fab to HEK293T-CD151 + cells **(Fig. 5A)**, with EC50 values in the low-nanomolar range **(Table 1)**. We also used flow cytometry to assess epitope overlap by measuring the ability of immunoglobulin (IgG) versions of each clone to block binding of each Fab to HEK293T-CD151 + cells. As expected, preincubation of HEK293T-CD151 + cells with each IgG reduced subsequent binding of the cognate Fab. Moreover, all IgGs blocked binding of the different Fab clones **(Fig. 5B)**, implying that all four distinct clones share a similar CD151 binding epitope.

Further corroboration of specificity for CD151 was provided by performing immunoprecipitation mass spectrometry (IP-MS) experiments with each Fab for HEK293T-CD151 + cells and HT1080 cells (express native levels of CD151 protein). Tandem mass spectra were searched against a human database to validate MS/MS protein identifications. Protein identifications were accepted if they could be established at greater than 99% probability based on identified peptides. After background filtering to remove keratin, immunoglobulin and cytoplasmic proteins, the highest peptide counts for all four Fabs were for CD151 on both different cell lines **(Fig. 5C and S8)**. The integrin β1 (ITGB1), a receptor identified to associate with CD151[53], also immunoprecipitated with Fabs CD151-1, CD151-3, and CD151-4 **(Fig. 5C)**, adding further validity of the Fabs selectivity for CD151. Taken together, the results show that the four *in silico* Abs recognize cell-surface CD151 with high affinity and specificity, with all different clones likely bind to overlapping epitopes.

Discussion

Multi-domain membrane proteins represent about 70% of current drug targets, especially for their role in the progression and tumorigenesis of numerous cancers[14]. However, the cellular surface component of many integral membrane proteins makes their production and purification extremely difficult for *in vitro* Ab selections[13]. The instability of membrane proteins makes them challenging targets to work with, as many of these proteins depend on the membrane environment for their correct structure. The Ab selection strategy presented in this work, termed CellectSeq **(Figure S7)**, bypasses the need for purified antigens where Ab library selections are performed directly on cell-surface antigens. Moreover, CellectSeq may target difficult receptors, such as those containing minimal loop protrusions and present in complex mixtures *in situ*, as is the case of tetraspanin receptors.

By this method, we targeted the challenging CD151, a cell surface protein that is linked in the disease and progression of tumors. The *in situ* Ab selection against CD151, then followed by conventional manual Ab screening yielded a unique immunodominant clone CD151-1. While NGS enrichment ranking analysis of the same selection identified highly homologous clones to CD151-1, with greater than 80% sequence identity in both CDRs L3 and H3. On the other hand, the motif-based and error filtering *in silico* analysis of CellectSeq yielded multiple diversified clones, with multiple identified at extremely low frequencies in the output pool. Moreover, all four distinct paratopes identified by CellectSeq share a similar target epitope **(Fig. 5B)**; this observation highlights the recalcitrant nature of CD151 receptor[35], which displays limited surface exposed regions that limits the available epitopes **(Fig. 1)**. Furthermore, the advantage of CellectSeq over conventional strategies of NGS derived Abs, such as enrichment ranking[54–56], is the exhaustive analysis of all paratope motifs in the NGS dataset, rather than unique observations of clonal sequence identities, that enables the discovery of low abundant selective Abs. The statistical evaluation of paralogs allows for the successful prediction of representative Abs against CD151, including low abundance clones at observed frequencies as few as 7/7.3 million reads.

We also accredit the success of CellectSeq to the design of the synthetic Ab repertoire itself. In contrast to existing strategies for enhancing sequencing fidelity in Illumina datasets[57–60], as demonstrated in this report the restricted synthetic framework of Library F permits simple and accurate detection of erroneous Illumina reads. Here, the design of positional codon frequencies in the restricted CDRs allows for rapid deconvolution of NGS datasets and the removal of errors and artifacts, whereas natural repertoires prove more random and difficult to assess NGS errors. The synthetic framework also provides a deep analysis of paratope diversities in the NGS data by utilizing motif-based *in silico* strategies that predict infrequent but target-specific Abs, which was demonstrated by the successful prediction of the CD151 diversified and selective Abs with frequencies below 0.01%.

The implementation of NGS analysis facilitates the rapid and successful discovery of Abs, and highlights that membrane associated antigens are accessible to synthetic Abs *in situ*. As demonstrated in this report, the strategy of CellectSeq surpasses standard methods of Ab manual screenings and NGS

b pool. Additionally, because NGS is an

attractive technology for generating meta-data, in regards to costs and access for Ab discovery[61, 62], we foresee the expanded implementation of NGS in conjunction with CellectSeq to identify diversified paratopes and low abundance clones. This is evidenced by recent reports of deep sequencing approaches that characterize human Ab libraries and V-gene repertoires from immunized mice[63, 64]. This combination of Ab cellular selections coupled with NGS analysis and CellectSeq may create a proteomic gateway for modulating the surfaceome.

Materials And Methods

Cell lines and culture practices

Both, the CD151 knockdown (HEK293T-CD151-) and CD151 overexpressing (HEK293T-CD151+) cell lines were gifts from the Dr. Rottapel lab at University of Toronto, Princess Margaret Cancer Centre. Briefly, the HEK293T-CD151- cells were generated using the Tet-pLKO-puro plasmid and the HEK293T-CD151 + cells were generated using the pLX304 plasmid, both as previously described[66]. The HEK293T cell backgrounds were cultured in Dulbecco's Modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS). The human fibrosarcoma H1080 cell line (ATCC; CCL-121) was cultured in Eagle's Minimum Essential Medium (EMEM) with 10% FBS. All cells were cultured at 37ºC in a humid incubator with 5% CO2.

Antibody Selections with Cellular Antigen

Phage pools representing synthetic antibody library-F[37] were cycled through four rounds of binding selections using a HEK293T-CD151- cell line as the background depleting step, and a HEK293T-CD151 + cell line as the target selection step (Fig. 2). The adherent cell lines were suspended using PBS, 10 mM ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich). For round 1, ten million re-suspended HEK293T-CD151 + cells (greater than 90% viability) were incubated with Fab-phage (3×10^{12} cfu) in cell growth media under gentle rotation for 2 hours at 4 °C. For rounds 2 and 3, the Fab-phage were cycled between antigen negative (to remove non-specific phage binders) to antigen positive cells. Here the Fab-phage were first incubated with the HEK293T-CD151- cell line for 2 hours at 4 °C, then the cells were spun down utilizing a chilled centrifuge and Fab-phage supernatant collected. Similarly, the HEK293T-CD151 + cells were spun down utilizing a chilled centrifuge and supernatant discarded. Next, the HEK293T-CD151 + cells were resuspended utilizing the Fab-phage supernatant, and incubated for 2 hours at 4 °C. For round 4, both HEK293T-CD151- and HEK293T-CD151 + cells were independently presented with Fab-phage and incubated for 2 hours at 4 °C. The HEK293T-CD151- and HEK293T-CD151+, cell lines were washed four times with chilled PBS and 1% BSA. For all rounds, after washing the bound phages were eluted from the cell pellet by resuspending the cells in 0.1 M hydrochloric acid and incubating for 10 minutes at room temperature. The cell solutions were neutralized using 11 M Tris buffer (Sigma-Aldrich), cellular debris was removed by high-speed centrifugation, and the eluent was transferred to clean vials. The output phages were amplified by infection and growth in *E. coli* OmniMAX™ cells (Thermo-Fisher). After round 4,

infected E. coli OmniMAX™ cells were plated on 2YT/carbenicillin (Sigma-Aldrich) plates for isolation of single colonies.

Phage ELISAs

Colonies of *E. coli* OmniMAX harboring phagemids were inoculated into 450 µl 2YT broth supplemented with carbenicillin and M13-KO7 helper phage, and the cultures were grown overnight at 37 °C in a 96-well format. Culture supernatants containing Fab-phage were diluted two-fold in PBS buffer supplemented with 1% BSA and incubated for 15 minutes at room temperature. To test binding to native antigen on cells, phages were added directly to the cellular media of HEK293T-CD151- and HEK293T-CD151 + adherent cells (95–100% confluence) in tissue-culture-treated 96-well plates (Thermo-Fisher). After incubation for 45 minutes at room temperature, the plates were washed gently with PBS and the cells were fixed with 4% paraformaldehyde (Sigma-Aldrich). The cells were washed with PBS and incubated for 30 minutes with horseradish peroxidase/anti-M13 Ab conjugate (Sigma-Aldrich) in PBS buffer supplemented with 1% BSA. The plates were washed, developed with TMB Microwell Peroxidase Substrate Kit (KPL Inc.), and quenched with 1.0 M phosphoric acid; the absorbance was determined at a wavelength of 450 nm. Clones were identified as positive if they produced at least three-fold greater signal on wells with HEK293T-CD151 + cells over antigen negative HEK293T-CD151- cells. All positive clones were subjected to Sanger DNA sequence analysis (Genewiz).

Fab Protein Purification

Fab proteins were expressed in *E. coli* BL21 (ThermoFisher), as described[38]. Following expression, cells were harvested by centrifugation and cell pellets were flash-frozen using liquid nitrogen. The cell pellets were thawed, re-suspended in lysis buffer (50 mM Tris, 150 mM NaCl, 1%Triton X-100, 1 mg/ml lysozyme, 2 mM $MgCl_2$, 10 units of benzonase), and incubated for 1 hour at 4 °C. The lysates were cleared by centrifugation, applied to rProtein A-Sepharose columns (GE Healthcare), and washed with 10 column volumes of 50 mM Tris, 150 mM NaCl, and pH 7.4. Fab protein was eluted with 100 mM phosphoric acid buffer, pH 2.5 (50 mM NaH_2PO_4, 140 mM NaCl, 100 mM H_3PO_4) into a neutralizing buffer (1 M Tris, pH 8.0). The eluted Fab protein was buffer exchanged into PBS and concentrated using an Amicon-Ultra centrifugal filter unit (EMD Millipore). Fab protein was characterized for purity by SDS-PAGE gel chromatography and concentration was determined by spectrophotometry at an absorbance wavelength of 280 nm.

IgG Purification

Full-length IgG proteins were expressed in mammalian cells, as described[67]. Briefly, plasmids designed to express heavy and light chains were co-transfected into Expi293 cells (ThermoFisher) using the FuGENE® 6 Transfection Reagent kit (Promega), according to the manufacturer's instructions. After 5 days, cell culture media was harvested and applied to an rProtein-A affinity column (GE Healthcare). IgG protein was eluted with 25 mM H_3PO_4, pH 2.8, 100 mM NaCl and neutralized with 0.5 M Na_3PO_4, pH 8. Fractions containing eluted IgG protein were combined, concentrated, and dialyzed into PBS, pH 7.4. IgG

protein was characterized for purity by SDS-PAGE gel chromatography and concentration was determined by spectrophotometry at an absorbance wavelength of 280 nm.

Flow-cytometry validations and cellular binding titrations

Adherent cells were brought into suspension using PBS supplemented with 10 mM ethylenediaminetetraacetic acid (EDTA; Sigma-Aldrich). The cells were washed with PBS, resuspended in PBS supplemented with 1% BSA, and incubated for 15 minutes at 4 °C. The cells were labelled with 500 nM Fab, or IgG for 30 minutes at 4 °C, then washed with PBS and resuspended in PBS supplemented with 1% BSA. Next, the cells were labelled with anti-Flag (for Fabs) conjugated Alexa-488 secondary Ab (Abcam) according to manufacturer's instructions. Data were collected using a CytoFLEX-S flow-cytometer (Beckman Coulter) using a 488-nm laser with 530/25 nm filter settings. The cells were analyzed in PBS, and all acquired live events were greater than 10,000 cells per sample. Quantitation analysis was carried out using FlowJo v10.2 Software (FlowJo, LLC). For Ab cellular titration analysis, the Abs were added to antigen positive HEK293T-CD151 + cells in triplicate samples from a concentration range of 500 pM to 1 µM. The mean fluorescence signal values were subtracted from the control antigen negative HEK293T-CD151- cells signals, and EC_{50} determined using Graph-pad Prism (GraphPad Software, San Diego, California, USA), where x is the Fab concentration:

$$Y = Y_{max} + \frac{Y_{min} - Y_{max}}{1 + \left(\frac{EC50}{X}\right)^{HillSlope}}$$

Mass Spectrometry

For immunoprecipitation of cell-surface protein, 10^7 lifted and dissociated HEK293T-CD151 + or HT1080 cells were incubated with 500 nM Fab protein in DPBS with calcium and magnesium (Gibco, 0404) for 1 hour at 4 °C. Cells were washed with PBS and lysed using IP lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.0% IGEPAL CA-630, 0.25% Na-deoxycholate, 1 mM EDTA, and protease inhibitor cocktail (Roche)) for 15 minutes at 4 °C and centrifuged at 12000 x g for 5 minutes at 4 °C. The supernatant was incubated with 30 µl of sepharose protein-A beads (GE Healthcare) for 1 hour at 4 °C. The beads were washed three times with lysis buffer, once with PBS, and resuspended in 22 µl 10 mM glycine, pH 1.5. After 5 minutes, the supernatant was collected and neutralized with 2.2 µl 1 M Tris, pH 8.8. DTT was added to a final concentration of 10 mM. The sample was incubated at 40 °C for 1 hour and cooled to room temperature. Iodoacetamide was added to a final concentration of 20 mM, and the sample was incubated at room temperature in the dark for 30 minutes. Trypsin (1 µg, Promega) was added and the sample was incubated overnight at 37 °C. Peptides were purified using C18 tips and analyzed on a linear ion trap-Orbitrap hybrid analyzer (LTQ-Orbitrap, ThermoFisher) outfitted with a nanospray source and EASY-nLC split-free nano-LC system (ThermoFisher).

Tandem mass spectra were extracted, and charge state was deconvoluted and deisotoped by Xcalibur version 2.2. All MS/MS samples were analyzed using PEAKS Studio (Bioinformatics Solutions, Waterloo, ON Canada; version 8.0 (2016-09-08)) and X! Tandem (The GPM, thegpm.org; version CYCLONE (2010.12.01.1)). Samples were searched against the Uniprot Human database (Downloaded May 1st, 2017: 20183 entries) assuming the digestion enzyme trypsin. Carbamidomethyl of cysteine was specified as a fixed modification. Deamidation of asparagine and glutamine were specified as variable modifications. Scaffold (version Scaffold_4.7.5, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95% probability. Peptide Probabilities from PEAKS Studio (Bioinformatic Solutions, Inc.) were assigned by the Peptide Prophet algorithm with Scaffold delta-mass correction. Peptide Probabilities from X! Tandem were assigned by the Scaffold Local FDR algorithm. Protein identifications were accepted if they could be established at greater than 99% probability and contained at least one identified peptide. Protein probabilities were assigned by the Protein Prophet algorithm [42]. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters.

Next-Generation Sequencing Analysis

The Fab-phage output pools from HEK293T-CD151- and HEK293T-CD151 + cell lines were utilized as input templates of PCR reactions using forward and reverse primers that flanked CDRs L3 and H3, respectively. The primers included a 24 base-pair template annealing region followed by a 6–8 base-pair unique nucleotide barcode identifier and an Illumina universal adapter tag (PE1 or PE2 for the reverse or forward primer, respectively). Duplicate PCR amplicons were generated per Fab-phage pool that were then isolated by gel electrophoresis and followed by agarose gel extraction (Qiagen). The duplicate PCR amplicons were combined, and the sample concentrations were determined by spectrophotometry (BioteK). The amplicons for antigen positive and negative Fab-phage pools were normalized, pooled, and sequenced using a HiSeq 2500 instrument (Illumina) with 300 paired-end cycles. Besides PE1 and PE2 Illumina universal primers, the sequencing runs also included a custom primer that allowed for complete sequencing of CDRs H1 and H2. Thus, the three primer reads together provided complete sequence coverage of the four CDRs that were diversified in Library F[37] **(Figure S2)**. We performed duplicate NGS runs and then combined them. Subsequently, the sequencing reads were deconvoluted for each clone, and the three primer reads (PE1, PE2, and custom) were combined into a single sequence to derive the complete sequence. Sequences were filtered from sequencing errors using per base high quality score cut-off of Q = 30, which corresponds to 1:1000 of incorrect base call [43]. High quality nucleotide sequences were obtained, translated into amino acid sequences, and compared to the designed sequence repertoire of Library F[37] to filter out technical errors inherent to sequencing and PCR amplification.

Filtering hybridization errors in NGS selection pools

The diversity, *i.e.* combinatorial possibilities, of our phage-displayed synthetic Ab library[37] is dominated

e theoretical diversity of L3 and H3 are to the

order of magnitude of 6 and 16, while H1 and H2 cover a diversity of 2^6 and 2^8. Thus, we theoretically assumed that the majority of PCR-induced artifacts and bias representing amalgams of existing sequences, *i.e.* hybridization errors, to be present in the pairs of sequences L3/H3. In addition, we assumed that among all possible pairs of sequences L3/H3, the valid pairs are overrepresented compared to the invalid pairs. Thus, for every sequence H3 in the Ab selection pool, we obtained the frequencies (*i.e.* number of observations) of all its paired sequences L3, and we calculated a frequency cut-off according to the maximum interclass inertia method using the *Koenig-Huygens* theorem[68]. The cut-off serves as a minimum frequency threshold to identify valid pairs L3/H3, thus, all Ab sequences with the invalid pairs are filtered from the selection pool.

Enumeration of consensus motifs in the CDR sequences

Consensus motifs, or motifs, are utilized to represent the linear information that is shared among groups of sequences. While certain positions in the motifs are defined (*e.g.* **P** as proline and **R** as arginine in the motif **PXXR**), others do not and are called wildcards (*e.g.* **X** as any amino acid in the motif "**PXXR**"). We utilize here the motifs to explore the linear information in the CDR sequences of each candidate Ab. To this end, we adapted the algorithm DALEL[46] that was first developed to explore the linear information in proteins. To avoid the explosion of the number of motifs, we restricted the number of allowed wildcards in each motif to 55% of its length. In addition, we considered only motifs with wildcards matching more than one amino acid in the matching sequences in the positive pool (*e.g.* wildcard **X** in motif **PXR** matches amino acids **Y** and **S** in the sequences **PYR** and **PSR**). Finally, we restricted the number of motifs by limiting the minimum number of sequences in the positive pool matching every motif to a 100.

Scoring separation between distributions of frequencies in positive and negative selections

The following procedure is performed for every Ab from the positive pool. We enumerate all possible motifs in the CDR sequences, as described earlier. Then, for each motif we obtain the frequencies (*i.e.* number of matching sequences / total number of sequences) in both the positive and the negative pools. According to the premise that selective Abs are enriched with paratope motifs that recognize the target antigen and non-selective Abs are not, it is then possible to assess Ab selectivity by checking whether the frequencies in the positive pool are higher compared to the negative pool. This is performed by scoring the level of separation between the distribution of frequencies in the positive and the negative pools.

To score the separation between the two distributions of frequencies of the motifs in the positive and the negative selection pools we utilized the Welch-Satterthwaite version of the *t*-test in conjunction with the rank transformation[47]. This approach has the advantage to simultaneously counteract the undesirable effects of both non-normality and unequal variances in our distributions of frequencies. First, the two distributions of frequencies are combined and arranged in ascending order, with tied frequencies receiving a rank equal to the average of their positions. Given the two distributions of ranks x_P and x_N that correspond to the frequencies in the positive and the negative selection pools, and $\bar{x}_P$ and $\bar{x}_N$ are

the means, s_P and s_N are the standard deviations, and n_P and n_P are the sizes, all respectively. We calculate the p-value to evaluate the statistical significance of the t-score following the procedure described below[48 – 50].

We first calculate the t-score by:

$$t = \frac{\bar{x}_P - \bar{x}_N}{\sqrt{\frac{s_P^2}{n_P} + \frac{s_N^2}{n_N}}}$$

We then calculate the degree of freedom by:

$$f = \frac{\left(\frac{s_P^2}{n_P} + \frac{s_N^2}{n_N}\right)^2}{\frac{\left(\frac{s_P^2}{n_P}\right)^2}{n_P - 1} + \frac{\left(\frac{s_N^2}{n_N}\right)^2}{n_N - 1}}$$

We finally calculate the p-value by:

$$p = \frac{1}{\sqrt{f\pi}} \frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)} \int_{-\infty}^{t} \frac{1}{\left(1 + \frac{t^2}{f}\right)^{\frac{f+1}{2}}} dt$$

Where t is the t-score, f is the degree of freedom, $\Gamma(.)$ is the Gamma function, and p is the probability that a single observation from the t distribution with f degrees of freedom will fall in the interval $[-\infty, t]$. In other terms, the p is the probability to have by chance any t-score that is equal or below to the t-score t. Thus, the lower is the p-value p, the higher is the significance of the t-score t, and consequently the higher is the separation between the two distributions of frequencies in the positive and the negative selections. We filtered the p-values using a stringent cut-off of 10^{-10} to identify highly specific Ab clones. To complement the p-values, we calculated Cohen's d effect size coefficient[69] to evaluate the difference between the means of the frequencies in the positive and the negative selections, and we kept p-values with huge effect size[70] ($d > 2$).

$$d = \frac{\bar{x}_P - \bar{x}_N}{\sqrt{\frac{\left(n_P - 1\right) s_P^2 + \left(n_N - 1\right) s_N^2}{n_P + n_N - 2}}}$$

Abbreviations

Antibody (Ab), cluster of differentiation 151 (CD151), next generation sequencing (NGS), antigen-binding fragments (Fabs), complementary determining regions (CDRs), immunoprecipitation mass spectrometry (IP-MS), immunoglobulin (IgG), integrin β1 (ITGB1), Dulbecco's Modified Eagle medium (DMEM), Eagle's Minimum Essential Medium (EMEM), fetal bovine serum (FBS), ethylenediaminetetraacetic acid (EDTA), and ethylenediaminetetraacetic acid (EDTA).

Declarations

Acknowledgements

We would like to thank Dax Torti and Tanja Durbic from the Donnelly Sequencing Centre for assistance with NGS sample runs and analysis. In addition, we would like to thank Meghan McLaughlin and Wei Ye for design of primers used in the NGS reads, and improvements to the cellular selection protocol.

Conflict of Interest

The authors have declared no conflict of interest.

Supporting Information

The data and the results supporting the findings of this study are available within the paper and its supplementary information files. Similarly, any additional data related to this study are available from the corresponding author upon reasonable request.

References

1. Weiner, L.M., Surana, R. & Wang, S. Monoclonal antibodies: versatile platforms for cancer immunotherapy. *Nat Rev Immunol* **10**, 317–327 (2010).
2. Miersch, S. & Sidhu, S.S. Synthetic antibodies: concepts, potential and practical considerations. *Methods* **57**, 486–498 (2012).
3. Adams, J.J. & Sidhu, S.S. Synthetic antibody technologies. *Curr Opin Struct Biol* **24**, 1–9 (2014).
4. Miersch, S. et al. Scalable high throughput selection from phage-displayed synthetic antibody libraries. *J Vis Exp*, 51492 (2015).

5. Hornsby, M. et al. A High Through-put Platform for Recombinant Antibodies to Folded Proteins. *Mol Cell Proteomics* **14**, 2833–2847 (2015).
6. Turunen, L., Takkinen, K., Soderlund, H. & Pulli, T. Automated panning and screening procedure on microplates for antibody generation from phage display libraries. *J Biomol Screen* **14**, 282–293 (2009).
7. Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat Rev Drug Discov* **1**, 198–210 (2002).
8. Baker, J.G. & Hill, S.J. Multiple GPCR conformations and signalling pathways: implications for antagonist affinity estimates. *Trends Pharmacol Sci* **28**, 374–381 (2007).
9. Oldham, W.M. & Hamm, H.E. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat Rev Mol Cell Biol* **9**, 60–71 (2008).
10. Guo, W. & Giancotti, F.G. Integrin signalling during tumour progression. *Nat Rev Mol Cell Biol* **5**, 816–826 (2004).
11. Speers, A.E. & Wu, C.C. Proteomics of integral membrane proteins–theory and application. *Chem Rev* **107**, 3687–3714 (2007).
12. Ulmschneider, M.B., Sansom, M.S. & Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59**, 252–265 (2005).
13. Helenius, A. & Simons, K. Solubilization of membranes by detergents. *Biochim Biophys Acta* **415**, 29–79 (1975).
14. Lundstrom, K. Structural genomics and drug discovery. *J Cell Mol Med* **11**, 224–238 (2007).
15. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993–996 (2006).
16. Detchokul, S., Williams, E.D., Parker, M.W. & Frauman, A.G. Tetraspanins as regulators of the tumour microenvironment: implications for metastasis and therapeutic strategies. *Br J Pharmacol* **171**, 5462–5490 (2014).
17. Zoller, M. Tetraspanins: push and pull in suppressing and promoting metastasis. *Nat Rev Cancer* **9**, 40–55 (2009).
18. Hemler, M.E. Tetraspanin proteins promote multiple cancer stages. *Nat Rev Cancer* **14**, 49–60 (2014).
19. Kitadokoro, K. et al. CD81 extracellular domain 3D structure: insight into the tetraspanin superfamily structural motifs. *EMBO J* **20**, 12–18 (2001).
20. Fitter, S., Tetaz, T.J., Berndt, M.C. & Ashman, L.K. Molecular cloning of cDNA encoding a novel platelet-endothelial cell tetra-span antigen, PETA-3. *Blood* **86**, 1348–1355 (1995).
21. Seigneuret, M. Complete predicted three-dimensional structure of the facilitator transmembrane protein and hepatitis C virus receptor CD81: conserved and variable structural domains in the tetraspanin superfamily. *Biophys J* **90**, 212–227 (2006).
22. Bienstock, R.J. & Barrett, J.C. KAI1, a prostate metastasis suppressor: prediction of solvated structure and interactions with binding partners: integrins, cadherins, and cell-surface receptor proteins. *Mol*

Carcinog **32**, 139–153 (2001).

23. Masciopinto, F., Campagnoli, S., Abrignani, S., Uematsu, Y. & Pileri, P. The small extracellular loop of CD81 is necessary for optimal surface expression of the large loop, a putative HCV receptor. *Virus Res* **80**, 1–10 (2001).
24. Sincock, P.M., Mayrhofer, G. & Ashman, L.K. Localization of the Transmembrane 4 Superfamily (TM4SF) Member PETA-3 (CD151) in Normal Human Tissues: Comparison with CD9, CD63, and α5β1 Integrin. *Journal of Histochemistry & Cytochemistry* **45**, 515–525 (1997).
25. Sadej, R., Grudowska, A., Turczyk, L., Kordek, R. & Romanska, H.M. CD151 in cancer progression and metastasis: a complex scenario. *Laboratory investigation; a journal of technical methods and pathology* **94**, 41–51 (2014).
26. Zeng, P. et al. Tetraspanin CD151 as an emerging potential poor prognostic factor across solid tumors: a systematic review and meta-analysis. *Oncotarget* **8**, 5592–5602 (2017).
27. Kumari, S., Devi, G.t., Badana, A., Dasari, V.R. & Malla, R.R. CD151-A Striking Marker for Cancer Therapy. *Biomark Cancer* **7**, 7–11 (2015).
28. Zijlstra, A., Lewis, J., Degryse, B., Stuhlmann, H. & Quigley, J.P. The inhibition of tumor cell intravasation and subsequent metastasis via regulation of in vivo tumor cell motility by the tetraspanin CD151. *Cancer Cell* **13**, 221–234 (2008).
29. Sadej, R. et al. CD151 regulates tumorigenesis by modulating the communication between tumor cells and endothelium. *Mol Cancer Res* **7**, 787–798 (2009).
30. Yang, X.H. et al. CD151 accelerates breast cancer by regulating alpha 6 integrin function, signaling, and molecular organization. *Cancer Res* **68**, 3204–3213 (2008).
31. Yang, X.H. et al. Disruption of laminin-integrin-CD151-focal adhesion kinase axis sensitizes breast cancer cells to ErbB2 antagonists. *Cancer Res* **70**, 2256–2263 (2010).
32. Deng, X. et al. Integrin-associated CD151 drives ErbB2-evoked mammary tumor onset and metastasis. *Neoplasia* **14**, 678–689 (2012).
33. Novitskaya, V. et al. Integrin alpha3beta1-CD151 complex regulates dimerization of ErbB2 via RhoA. *Oncogene* **33**, 2779–2789 (2014).
34. Mieszkowska, M. et al. Tetraspanin CD151 impairs heterodimerization of ErbB2/ErbB3 in breast cancer cells. *Transl Res* **207**, 44–55 (2019).
35. Hemler, M.E. Tetraspanin functions and associated microdomains. *Nature reviews. Molecular cell biology* **6**, 801–811 (2005).
36. Gallo, E. et al. In situ antibody phage display yields optimal inhibitors of integrin α11/β1. *mAbs* **12**, 1717265 (2020).
37. Persson, H. et al. CDR-H3 Diversity Is Not Required for Antigen Recognition by Synthetic Antibodies. *Journal of Molecular Biology* **425**, 803–811 (2013).
38. Sidhu, S.S. et al. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol* **338**, 299–310 (2004).

39. Pollock, S.B. et al. Highly multiplexed and quantitative cell-surface protein profiling using genetically barcoded antibodies. *Proc Natl Acad Sci U S A* **115**, 2836–2841 (2018).
40. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M.F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71**, 8966–8969 (2005).
41. DuBridge, R.B. et al. Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Mol Cell Biol* **7**, 379–387 (1987).
42. Fellouse, F.A., Wiesmann, C. & Sidhu, S.S. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci U S A* **101**, 12467–12472 (2004).
43. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8**, 175–185 (1998).
44. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**, 186–194 (1998).
45. Kelil, A., Levy, E.D. & Michnick, S.W. Evolution of domain-peptide interactions to coadapt specificity and affinity to functional diversity. *Proc Natl Acad Sci U S A* **113**, E3862-3871 (2016).
46. Kelil, A., Dubreuil, B., Levy, E.D. & Michnick, S.W. Exhaustive search of linear information encoding protein-peptide recognition. *PLoS Comput Biol* **13**, e1005499 (2017).
47. Zimmerman, D.W. & Zumbo, B.D. Rank transformations and the power of the Student t test and Welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* **47**, 523 (1993).
48. Welch, B.L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
49. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biometrics bulletin* **2**, 110–114 (1946).
50. Shaw, W. New Methods for Managing" Student's" T Distribution. *Preprint King's College* (2006).
51. Posner, B. et al. A revised strategy for cloning antibody gene fragments in bacteria. *Gene* **128**, 111–117 (1993).
52. Keim, M., Williams, R.S. & Harwood, A.J. An inverse PCR technique to rapidly isolate the flanking DNA of dictyostelium insertion mutants. *Mol Biotechnol* **26**, 221–224 (2004).
53. Baldwin, G. et al. Tetraspanin CD151 regulates glycosylation of (alpha)3(beta)1 integrin. *J Biol Chem* **283**, 35445–35454 (2008).
54. t Hoen, P.A. et al. Phage display screening without repetitious selection rounds. *Anal Biochem* **421**, 622–631 (2012).
55. Ravn, U. et al. Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**, 99–110 (2013).

56. Yang, W. et al. Next-generation sequencing enables the discovery of more diverse positive clones from a phage-displayed antibody library. *Exp Mol Med* **49**, e308 (2017).
57. Quail, M.A. et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005–1010 (2008).
58. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40** (2012).
59. Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182–189 (2009).
60. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *P Natl Acad Sci USA* **108**, 9530–9535 (2011).
61. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133–141 (2008).
62. van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30**, 418–426 (2014).
63. Lee, C.V. et al. High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J Mol Biol* **340**, 1073–1093 (2004).
64. Reddy, S.T. et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* **28**, 965–969.
65. Lefranc, M.P. et al. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res* **27**, 209–212 (1999).
66. Medrano, M. et al. Interrogation of Functional Cell-Surface Markers Identifies CD151 Dependency in High-Grade Serous Ovarian Cancer. *Cell reports* **18**, 2343–2358 (2017).
67. Chen, G. et al. Synthetic antibodies and peptides recognizing progressive multifocal leukoencephalopathy-specific point mutations in polyomavirus JC capsid viral protein 1. *MAbs* **7**, 681–692 (2015).
68. Kelil, A., Wang, S. & Brzezinski, R. CLUSS2: an alignment-independent algorithm for clustering protein families with multiple biological functions. *Int J Comput Biol Drug Des* **1**, 122–140 (2008).
69. Statistical Power Analysis for the Behavioral-Sciences - Cohen,J. *Percept Motor Skill* **67**, 1007–1007 (1988).
70. Sawilowsky, S.S. New effect size rules of thumb. (2009).

Tables

Table 1. Summary of anti-CD151 clones derived from CellectSeq

Summary of identified CellectSeq clones showing the NGS unique sequence counts, the Ab concentrations that result in half-maximal affinities (EC_{50}), and the description of amino acids in the diversified CDR loop regions, as defined by the IMGT nomenclature[65].

Clone ID	Illumina pools		Log2 FC Pos/Neg	Pvalue	EC_{50} (nM)	L3						H1						H2										H3										
	Counts	% total				107	108	109	110	114	115	116	30	35	36	37	38	39	55	56	57	58	59	62	63	64	65	66	107	108	109	110	111	112.1	112	113	114	115
CD151-1	2260484	3E+01	7.2	< 1E-200	25	S	F	F	-	-	P	I	L	S	Y	Y	S	M	S	I	Y	P	S	Y	G	Y	T	Y	S	H	Y	G	V	W	Y	G	A	M
CD151-2	137	2E-03	4.2	< 1E-035	71	S	W	V	Y	S	L	I	I	Y	Y	Y	S	M	S	I	S	P	Y	S	G	Y	T	Y	S	P	Y	A	V	-	F	Y	G	M
CD151-3	74	1E-03	5.7	< 1E-142	36	S	G	W	P	F	L	I	L	S	Y	Y	Y	M	S	I	Y	P	Y	Y	G	Y	T	Y	S	H	Y	G	V	W	Y	G	A	M
CD151-4	7	9E-05	1.8	< 1E-027	177	S	S	Y	-	S	L	I	L	S	Y	S	Y	M	S	I	S	S	S	Y	G	Y	T	S	S	Y	S	G	-	-	-	G	A	L

Supplemental Legends

Table S1. Validation summary of CD151 selective Ab clones derived from NGS enrichment ranking analysis. The antibody sequences were synthesized as Fab protein and assayed for cellular binding on HEK293T-CD151+ cells via flow-cytometry. In-situ validation result “Pass” = fluorescence signal 3-fold or greater than background (HEK293T-CD151- cells).

Figure S1. Library F CDR sequences. (A) Nucleotide sequences are formatted according to IUPAC code, and showing the nucleotide composition of template (parental) CDR sequences utilized to construct library F. Among $3x10^{10}$ unique clones in Library F, diversity was incorporated in approximately 80% of the population in each CDR and the retention of template sequences in the remainders. **(B)** Composition of pattern (variable) nucleotide sequences in library F, where the pattern CDR sequences describe the composition and length diversity introduced to CDRs-H1, H2, H3 and L3 by allowing loop lengths that are found within these regions of natural antibodies. X(3-7) and X(1-17) indicates the insertion of 3 to 7 and 1 to 17 tri-nucleotides from a mixture designed to contain nine different amino acids of the following composition; 25% Tyr, 20% Ser, 20% Gly, 10% Ala, and 5% each of Phe, Trp, His, Pro and Val, all respectively. **(C)** Framework Structure of Fab region showing the CDR loops L1 and L2 (orange), L3 (black), H1 (green), H2 (red), and H3 (blue) as spheres. The figure was generated using PyMOL (http://www.pymol.org/) with crystal structure coordinates (Protein Data Bank entry 1MIM). **(D)** Description of the synthetic antibody library F CDR amino acid sequences, highlighting the CDR diversity by position (shaded in gray are fixed positions). Allowed amino acids are denoted by the single-letter code, where X denotes a mixture of nine amino acids (Y, S, G, A, F, W, H, P or V). The lengths of CDR-L3 and CDR-H3 may vary from 3–9 and 1–19 respectively and the residue numbering is according to the IMGT scheme.

Figure S2. Description of NGS strategy for Illumina read-out of the diversified CDRs. The template DNA Fab region includes a PhoA promoter followed by a ribosome binding site (RBS) and two gene cassettes with signal peptide (SP) sequences and light or heavy antibody regions followed by the M13 gene-3 for display on phage particles. The phagemid template Fab region is PCR amplified utilizing two distinct primers that contain barcoded indexes and Illumina adapter sites. The amplicons are sequenced with three distinct read-out primers that cover the sequences of the diversified regions L3, H1, H2, and H3.

Figure S3. Library F NGS three-read diagram showing annealing regions of primers on Ab framework.

Figure S4. Validation of CD151 transgenic and knock-down cell lines used in this report. (A) Flow-cytometry histograms showing the CD151 surface expression of different HEK293T cell lines. Measurements were performed using an overexpressing CD151 cell line (HEK293T-CD151+), a short hair-pin RNA CD151 knockdown cell line(HEK293T-CD151-) cell line, and the parental HEK293T cell line, and CD151 surface expression measured utilizing a mouse anti-human CD151 PE conjugated IgG (Biolegend; cat. 350408). **(B)** Schematic of a round of selection where amplified Fab-phage is utilized to label antigens at the surface of live mammalian cells. After cellular wash and elution, the Fab-phage are amplified in *E. coli* and isolated for the next round of selection.

Figure S5. Binding measurements of Fab-phage to CD151 expressing cells (HEK293T-CD151+) by single-point cellular ELISA measurements. The fold change signal is measured by taking the ratio of signal from HEK293T-CD151+ cells over HEK293T-CD151- cells. A fold change signal of 5 or greater is deemed as a potential positive CD151 Fab-phage binder.

Figure S6. Linear information of paratope motifs as predictor for binding specificity of antibodies. Representation of the premise stating that highly selective Abs are enriched with paratope motifs that enable specific recognition of target epitopes, whereas non-specific Abs lack such enrichment.

Figure S7. Description of CellectSeq Methodology. Flow-chart of workflow for the CellectSeq Methodology.

Figure S8. IP-MS peptide coverage results of CD151 protein: (A) Fab CD151-1 assayed utilizing HEK293T-CD151+ cells; (B) Fab CD151-1 assayed utilizing HT-1080 cells; (C) Fab CD151-2 assayed utilizing HEK293T-CD151+; (D) Fab CD151-2 assayed utilizing HT-1080 cells; (E) Fab CD151-3 assayed utilizing HEK293T-CD151+ cells; (F) Fab CD151-3 assayed utilizing HT-1080 cells (G) Fab CD151-4 assayed utilizing HEK293T-CD151+ cells; and (H) Fab CD151-4 assayed utilizing HT-1080 cells. NTT - Number of termini consistent with the enzymatic cleavage or tryptic termini; Observed - Mass over charge (M/Z) of the parent or precursor ion measured by the mass spectrometer.; Actual Mass - Peptide mass in Dalton obtained by multiplying the charge to the subtraction of one proton from the observed M/Z; Charge - Peptide charge; Delta Da - (Actual Mass - Theoretical Peptide Mass) in Dalton, where the Theoretical Peptide Mass or Calculated peptide mass, is given by the sum of amino acid residue masses included in the peptide plus a water molecule; Delta PPM - (Actual Mass - Theoretical Peptide Mass) in PPM also referred to in the spectrum as the Parent error. It is calculated by dividing the delta mass expressed in Dalton by the Actual Mass and then multiplied by one million.; Retention Time - Measured in seconds, it is included in the table only if the information is listed in the peak list of the loaded data; TIC - MS/MS Total Ion Current; Start - Peptide start index; and Stop - Peptide stop index.

Figures

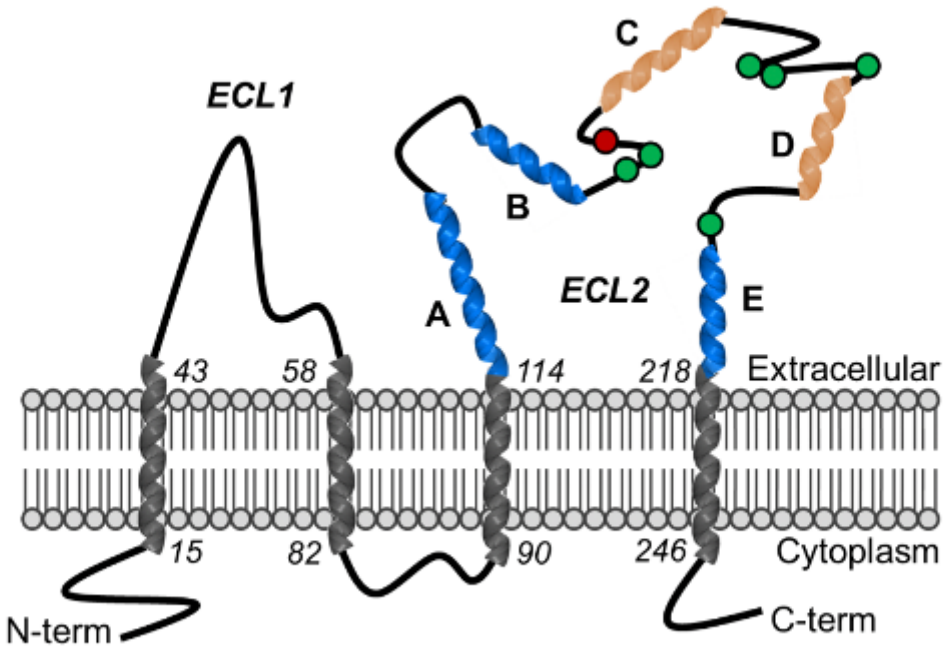


Figure 1

Schematic of CD151 structure CD151 consists of four alpha-helical transmembrane domains, two extracellular loops one short (ECL1) and one long typically 100 amino acid residues (ECL2), and one very short intracellular loop, all flanked by relatively short cytoplasmic N-terminal and C-terminal tails. The EC1 loop displays small stature and low structural organization. The ECL2 loop is composed of five α-helical domains A, B, C, D, and E, forming stalk and head elements of a mushroom-like structure. The A, B and E helices (blue) are forming the constant region and are suggested to mediate homodimerization, while the C and D helices (orange) are forming the variable region and their flanking sequences mediate interactions with other proteins. The six cysteine residues in ECL2 are indicated in green, and the N-glycosylation site in red. The numbering represents the amino acid occurring before or after each transmembrane domain.

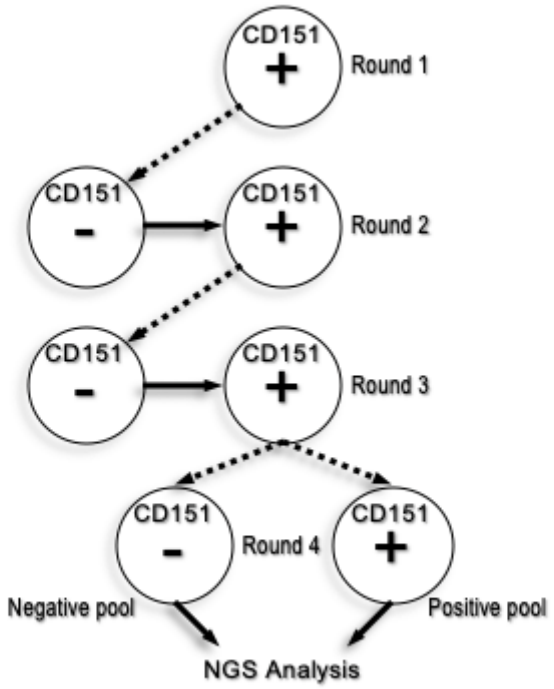


Figure 2

CD151 in situ Ab selection process Circles represent live HEK293T cell lines of either overexpressing CD151 (HEK293T-CD151+; denoted by “+” sign), or CD151 knockdown cells (HEK293T-CD151-; denoted by a “-” sign). For round 1, the Fab-phage (naïve library F) were incubated with HEK293T-CD151+ cells, then eluted and amplified for the next round. For rounds 2 and 3, the Fab-phage were first incubated with HEK293T-CD151- cells, then transferred and incubated with HEK293T-CD151+ cells. For round 4, amplified round 3 Fab-phage were independently incubated with HEK293T-CD151- or HEK293T-CD151+ cells. The dashed lines with arrows indicate that eluted phage from the preceding round were amplified through E. coli prior to incubation with cells in the succeeding round. The solid lines with arrows indicate that unbound phage from the preceding cell line were transferred directly to the succeeding cell line. In round 4, phages from Fab-phage pools HEK293T-CD151+ (Positive pool) and HEK293T-CD151- (Negative pool) were amplified and used as DNA template for NGS analysis.

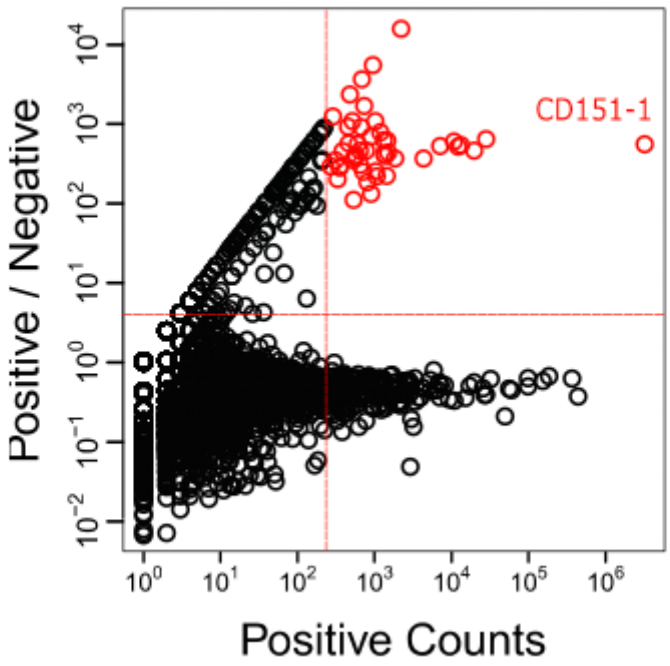


Figure 3

NGS enrichment ranking selection of CD151 in situ selection of Ab clones The abundance of each sequence in Fab-phage pools selected for binding to HEK293T-CD151+ cells (Positive Counts, x-axis) is plotted versus the ratio of the abundance in pools selected for binding to HEK293T-CD151+ cells over pools selected for binding to HEK293T-CD151- (Positive/Negative, y-axis). Each circle represents one unique paratope (i.e. unique combination of CDRs L3, H1, H2 and H3). The dashed red lines define an upper-right quadrant that contains putative CD151 binding clones, defined arbitrarily as those occurring more than 200 times in the positive pool and being greater than four-fold enriched relative to the negative pool. The red circles represent the 100 Ab clones in the top-right quadrant that were selected after the NGS analysis and predicted to bind to CD151. All selected clones are close homologs (>80% sequence identity) of the immunodominant Ab clone CD151-1. The red circle at the far right represents the immunodominant Ab clone CD151-1 that were manually sampled and validated as specific CD151 binding Ab by cellular phage ELISA (Figure S2).

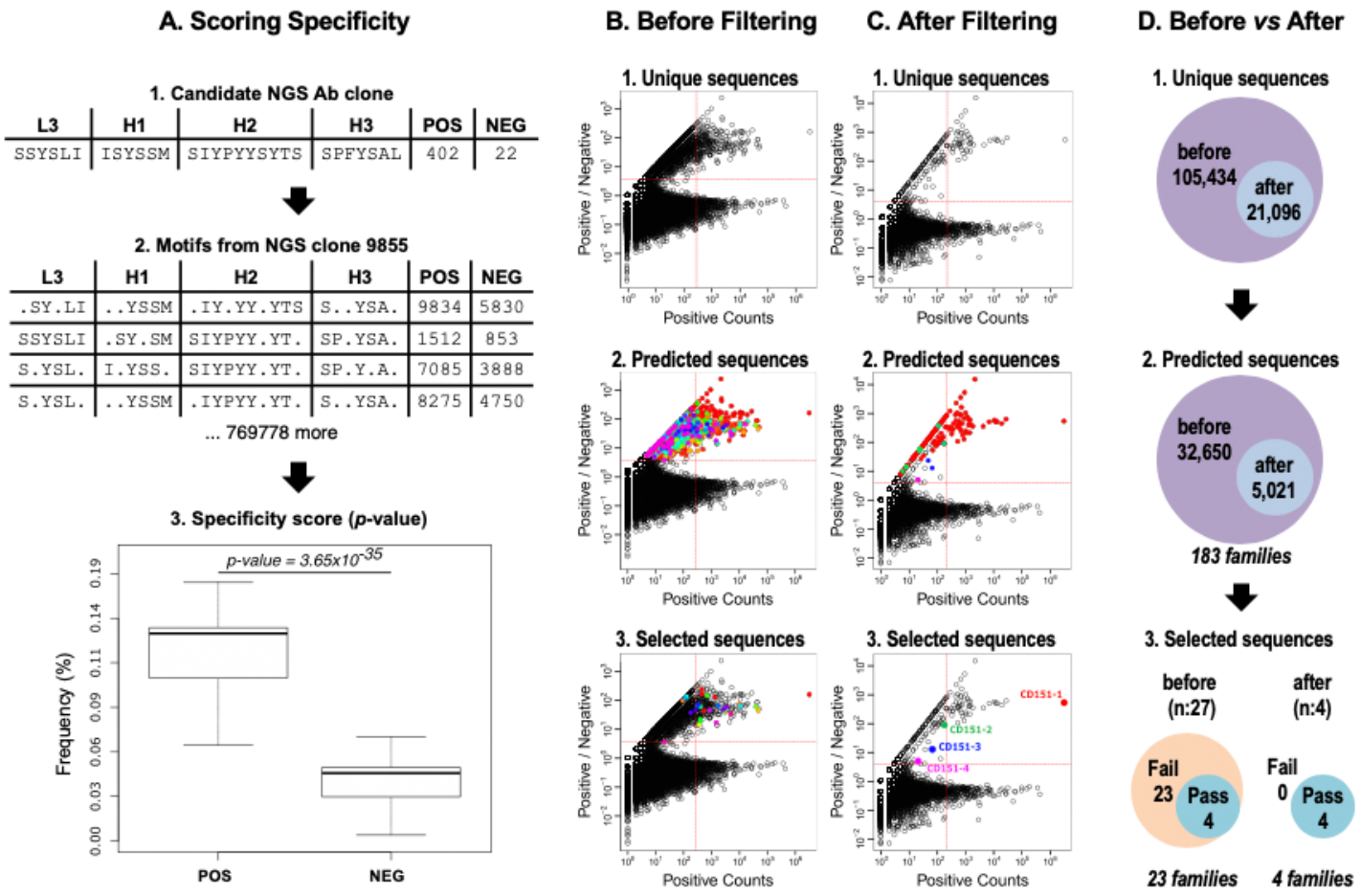
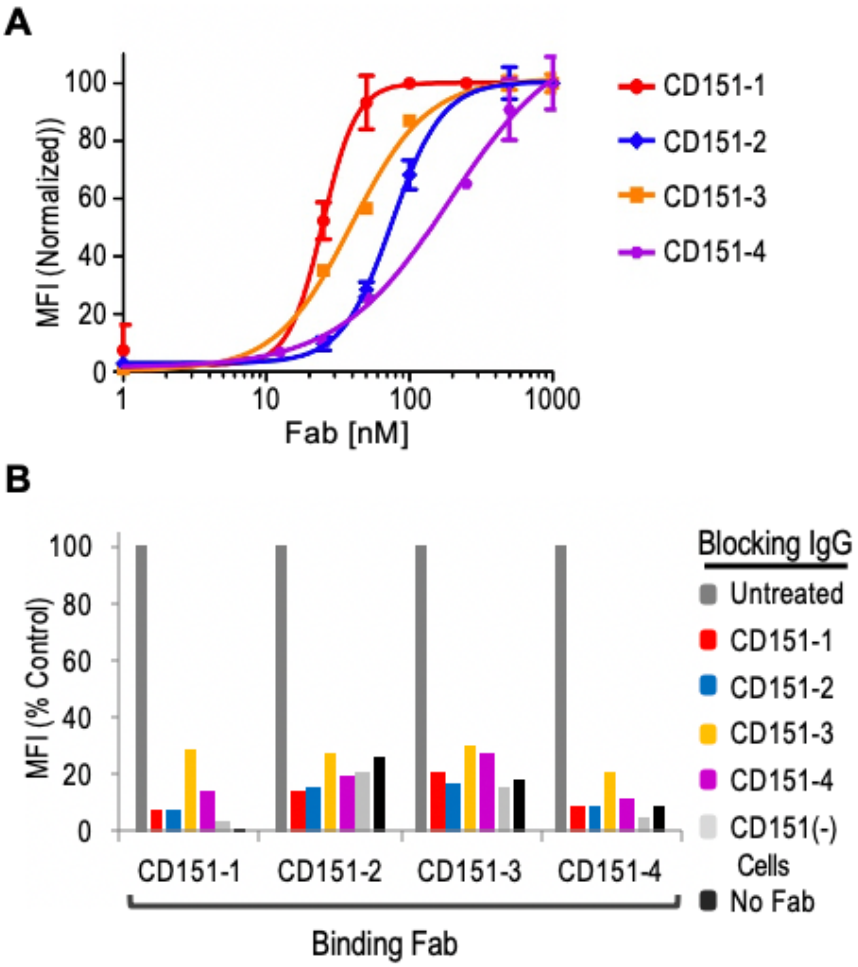


Figure 4

Motif-based in silico Ab discovery strategy of CD151 selective Abs. (A) Scoring Specificity: (1). Consider the candidate antibody NGS clone, the CDR sequences and counts in the positive (POS) and the negative (NEG) selection pools at round 4 are reported. (2). We enumerate all consensus motifs in the CDR sequences. (3). We score the binding specificity of the candidate antibody by assessing the separation between the distributions of the frequencies of the motifs in the positive (POS) and negative (NEG) pools. To this end, we calculate the p-value of the t-test (Methods). (B) and (C). In each figure, the abundance of each sequence in the positive pool (Positive Counts, x-axis) is plotted versus the ratio of the abundance in the positive pool over the negative pool (Positive/Negative, y-axis). Each circle represents one unique Ab clone, colored circles (except black) represent families of homologous sequences (sequence identity > 0.75). The dashed red lines define an upper-right quadrant that contains enriched Ab clones defined as occurring more than 200 times in the positive pool and being greater than four-fold enriched relative to the negative pool. (B) Before Filtering: NGS sequences and predicted sequences before filtering hybridization errors. (1). Distribution of unique sequences in the positive pool and their enrichment over the negative pool. (2). Distribution of predicted sequences with high specificity, colored by family of homologs. (3) Distribution of selected sequences for in situ validation, colored by family of homologs. (C) After filtering: NGS sequences and predicted sequences after filtering hybridization errors. (1). Distribution

of unique sequences in the positive pool and their enrichment over the negative pool. (2). Distribution of predicted sequences with high specificity, colored by family of homologs. (3). Distribution of selected sequences for in situ validation, colored by family of homologs. Colored circles represent Ab clones named CD151-1 to CD151-4 which were selected and validated as specific CD151 binding Abs by cellular phage ELISA. (D) Before vs. After filtering: Difference between number of unique sequences and prediction results before and after filtering hybridization errors. (1). Number of unique sequences in the positive pool before and after filtering. (2). Number of predicted sequences before and after filtering. (3). Number of validated sequences before and after the filtering.



C

Clone ID	HEK293T Cells (CD151+)		HT1080 Cells (CD151)		HT1080 Cells (ITGB1)	
	Total Pep	*% Cov*	*Total Pep*	*% Cov*	*Total Pep*	*% Cov*
CD151-1	61.0	35.0	11.0	18.0	3.0	4.5
CD151-2	17.0	20.0	3.0	9.9	0.0	0.0
CD151-3	3.0	11.0	13.0	30.0	9.0	12.4
CD151-4	16.0	17.0	1.0	3.6	1.0	1.3
Control	0.0	0.0	0.0	0.0	0.0	0.0

Figure 5

Characterization of anti-CD151 Abs derived from CellectSeq (A) Dose response curves for anti-CD151 Fabs assessed by flow cytometry fluorescence (y-axis) using HEK293T-CD151+ cells. The MFI signals were subtracted from background Fab binding to HEK293T-CD151-cells and normalized to the highest concentration value for each sample. Experiment performed in triplicates and error bars indicate SD. (B) -CD151+ cells by indicated IgGs, assessed by

flow cytometry fluorescence (y-axis). Experiment a representative of duplicate experiments. (C) Mass-spectrometry summary table of enriched isolated peptides from immuno-precipitated CD151 cellular lysates from HEK293T-CD151+ or HT1080 cells with anti-CD151 Fabs or control Fab. Percent coverage is the percentage of CD151 protein detected by the total peptides.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementCD15120200724.pptx