A computational approach to identify enrichment from NGS data across multiple rounds of phage display screening

Tatyana Zamkovaya[1], Jessica Min-DeBartolo[1], James Apgar[1], Jonathan McDaniel[1], Joel Bard[1]

[1] Biomedicine Design, Pfizer, Cambridge, MA 02139, USA.

ABSTRACT

Background:
Phage display selection combined with next-generation-sequencing (NGS) can allow for rapid generation of large cohorts of antibody sequences which may bind specifically to a given target. It can, however, be difficult to identify specific binders based on this NGS data alone.

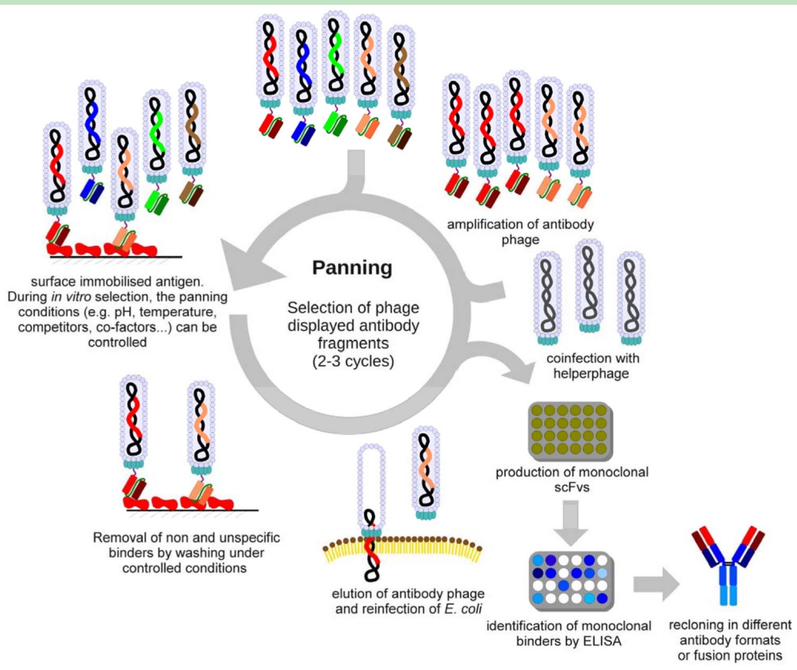
Method:
We developed an NGS-based strategy to identify enriched hits across rounds of selection. Our approach reduces millions of sequences to a manageable number based on their frequency in on-target versus off-target pools. A unique identifier based on key sequence features helps combine and collapse annotated antibody sequences across pools into one table, with each row representing a unique identifier sequence and each column representing its frequency in each pool. Sequence enrichment is then calculated as the normalized change in read frequency between the on- and off-target pools. The enrichment score, ranging between -1 (not enriched) and 1(enriched), can be summed up across multiple comparisons to find the most consistently enriched sequences which are likely to be specific binders.

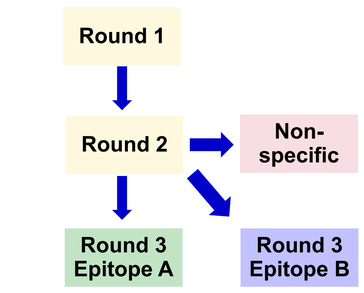
Results:
12E6 antibody sequences from 3 selection rounds and 12 pools were reduced to 18,000 abundant, shared sequences. 2700 sequences were enriched in at least one on-target pool. 288 highly enriched sequences with good manufacturability profiles were selected for recombinant expression and binding assays. 75 percent of the NGS-derived hits demonstrated a strong binding.

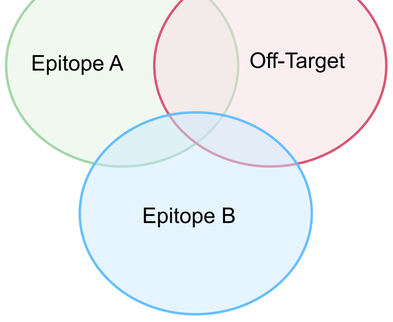
Conclusion:
Our approach enables identification of highly enriched binders from NGS data, saving time and effort in antibody hit selection and allowing examination of a wider range of sequences than traditional in vitro screening alone. Future work using machine learning to predict enrichment will further advance antibody discovery.

PHAGE DISPLAY WORKFLOW FOR ANTIBODY DISCOVERY

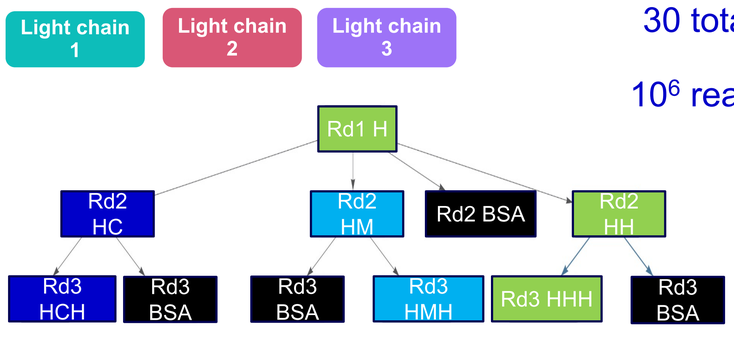
Overall Goal:
Find highly specific and enriched antibodies to target







CASE STUDY: ENRICHMENT ACROSS 3 ROUNDS



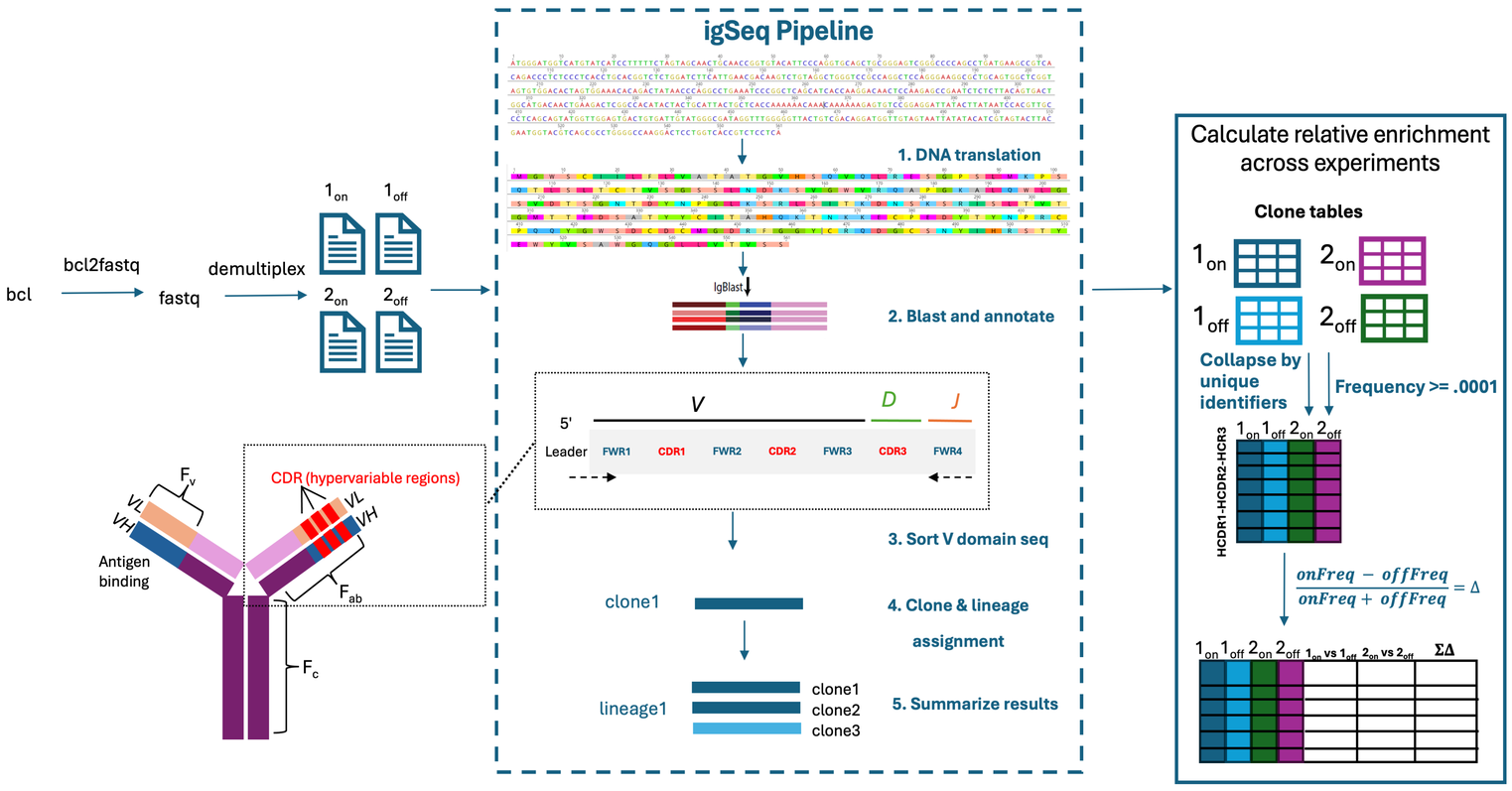
30 total samples

10^6 reads/ sample

H: Human; M: Mouse; C: Complex; BSA: Bovine serum albumin (negative control)

Sample Name	Selection Round	Average Total Reads	Average Distinct Antibody Sequences
Rd2_HC	2	859015	644728
Rd2_HH	2	779831	573612
Rd2_HM	2	824365	552515
Rd2_BSA	2	837144	561503
Rd3_HCH	3	874084	618634
Rd3_HHH	3	808988	560331
Rd3_HMH	3	881734	532372
Rd3_BSA_1	3	837574	467991
Rd3_BSA_2	3	852018	510928
Rd3_BSA_3	3	839294	549624

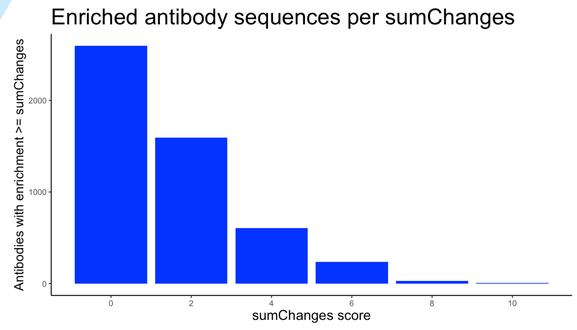
NGS-BASED PIPELINE TO IDENTIFY RELATIVE ENRICHMENT



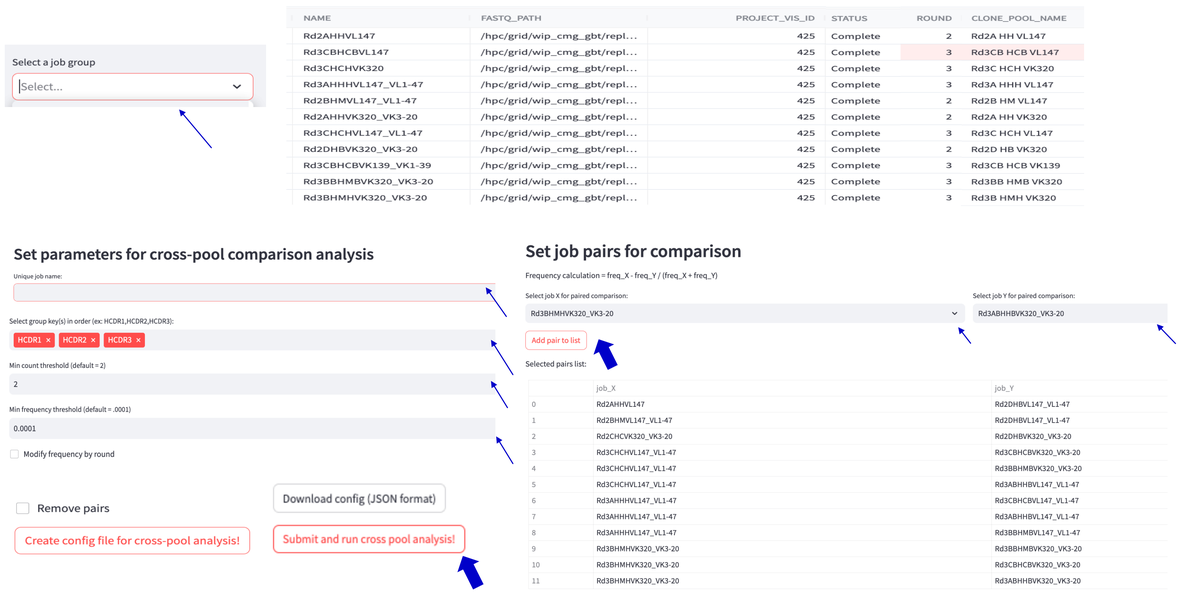
NGS-DERIVED ANTIBODIES ENRICHED FOR TARGET CAN BE FOUND ACROSS MULTIPLE SAMPLES AND SELECTION ROUNDS

Unique LC_vGene_HCDR1_HCDR2_HCD3	Change_Rd2AHH_vs_Rd2DHB	Change_Rd2BHM_vs_Rd2DHB	Change_Rd2_CHC_vs_Rd2_DHB	Change_Rd3_AHHH_vs_Rd3ABHHB	Change_Rd3_AHHH_vs_rd3BBHMB	Change_Rd3_AHHH_vs_Rd3CBHCB	Change_Rd3_BHMH_vs_Rd3ABHHB	Change_Rd3_BHMH_vs_Rd3BBHMB	Change_Rd3_BHMH_vs_Rd3CBHCB	Change_Rd3_CHCHvs_Rd3ABHHB	Change_Rd3_CHCH_vs_Rd3BBHMB	Change_Rd3_CHCH_vs_Rd3CBHCB	ΣΔ (*sum Changes*)
	1	1	0	1	1	1	1	1	1	1	1	1	11
	1	1	0	1	1	1	1	1	1	1	1	1	11
	1	1	0	1	1	1	1	1	1	1	1	1	11
	0	1	0	1	1	1	1	1	1	1	1	1	10
	0	1	1	1	0.713	0.125	1	0.895	0.593	1	0.759	0.224	9.31

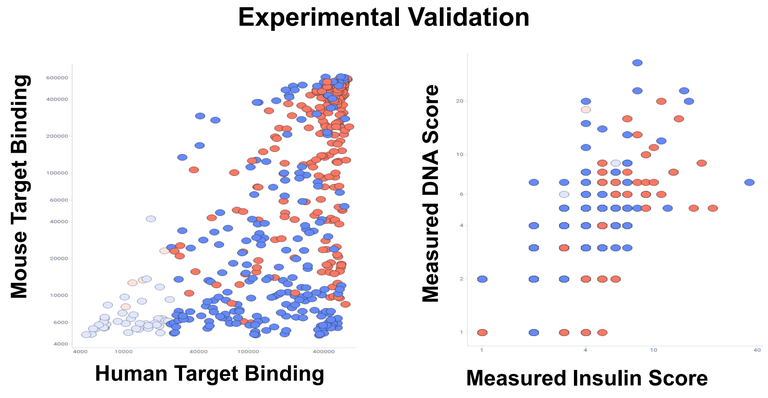
Eliminate low freq. clones
18000 total sequences
Enriched for target
2700 sequences
In silico liability prediction
Select hits for IgG conversion
288 sequences



STREAMLIT APPLICATION AUTOMATES CROSS-POOL COMPARISON CONFIGURATION AND SUBMISSION

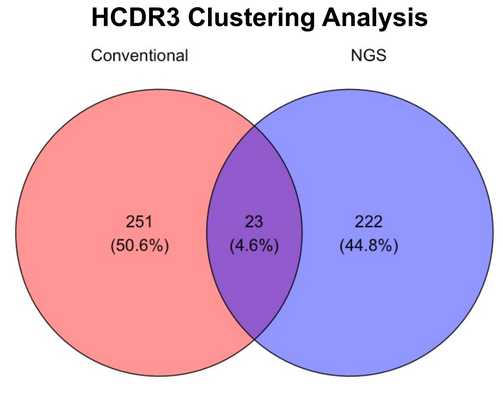


NGS HITS HAVE HIGH RATE OF BINDING AND CROSS-REACTIVE BINDING, INCREASING HIT PANEL QUANTITY & DIVERSITY BY 50 %



75 % of NGS hits show binding to target

NGS hits have low, comparable immunogenicity profiles to conventional hits



~90 % of selected NGS hits are novel HCDR3 not found in conventional screening

New NGS-derived diversity is obtained with minimal time and effort

CONCLUSIONS

New methodology to identify hits from NGS data was established and validated.
NGS clones had a 75% hit rate and increased hit panel by 50 %.
Hit calling from NGS data successfully improved target binding selection

FUTURE GOALS

We plan to use cross-pool NGS results to train ML/AI for target binding prediction and de novo sequence generation. The combination of in-vitro selection, NGS, and AI has significant potential to reduce the time and overall experimental effort needed for antibody lead identification.