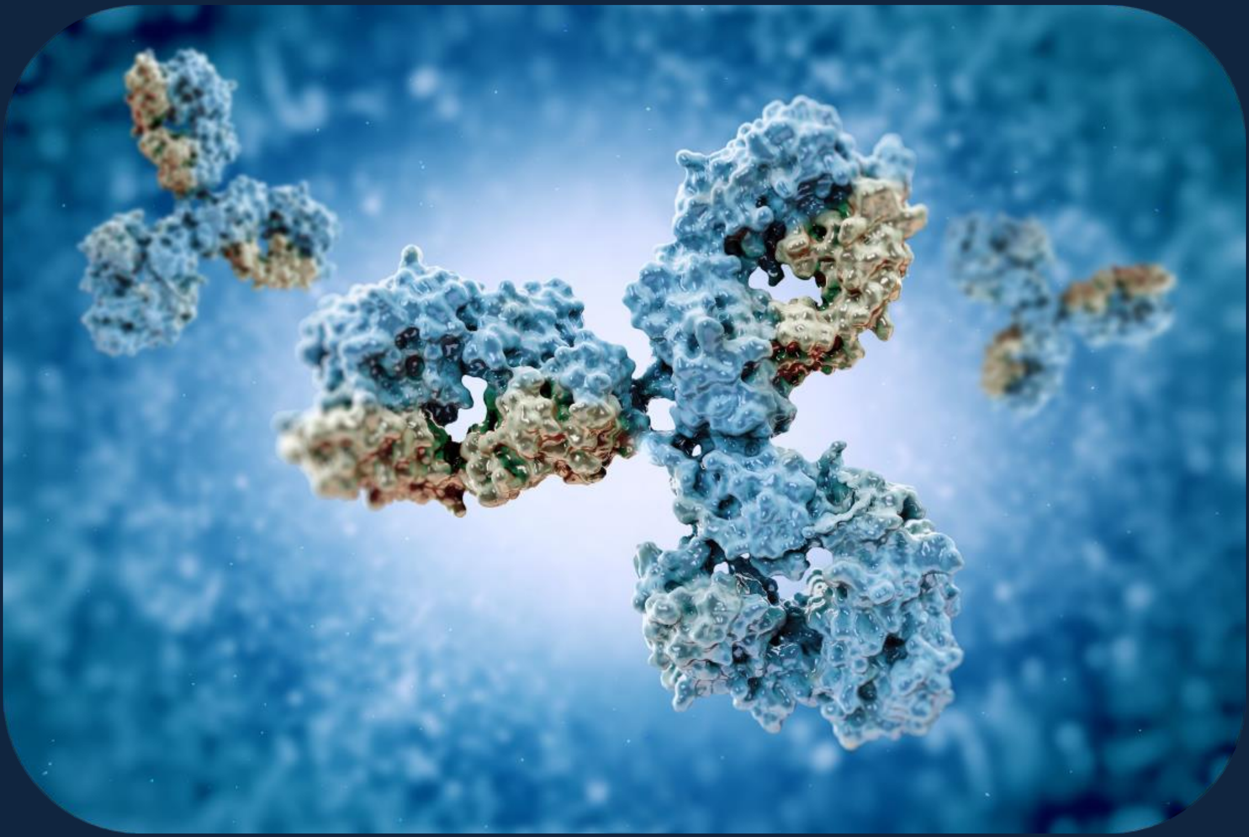
Structurally Primed Phage-display Libraries

Jorge Moura Sampaio



Dissertation presented to obtain the Ph.D degree in Bioengineering

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras,
April, 2022

Structurally Primed Phage-display Libraries

Author: Jorge Moura Sampaio

Supervisor(s): Dr. Ana Paula Batista; Prof. Paula Alves

Thesis Committee: Dr. Stefan Ewert; Dr. Patrik Maurer; Prof. Manuel Carrondo

Dissertation presented to obtain the Ph.D degree in Bioengineering

Instituto de Tecnologia Química e Biológica António Xavier | Universidade Nova de Lisboa

Oeiras, Month, Year

Preface

The work presented in this Thesis is derived from my PhD research, made possible primarily by the PhD fellowship ref. PD/BD/128321/2017, from *Fundação para a Ciência e Tecnologia* (FC&T), *Ministério da Ciência, Tecnologia e Ensino Superior* (MCTES), Portugal. These studies, carried out between 2017 and 2021, were integrated in the MIT-Portugal Program (MPP) in Bioengineering. The first year matched the MPP Executive Masters, a one-year educational program sponsored by the Massachusetts Institute of Technology (MIT) in collaboration with five Portuguese universities, that focused on innovation and technology commercialization across the bioengineering field. The first year was supervised by Prof. Joaquim Sampaio Cabral (*Instituto Superior Técnico, Universidade de Lisboa*).

From 2018 onwards, the work was dedicated to what is presented on this Thesis, which took place at the Novartis Pharma Lab, at the *Instituto de Biologia Experimental e Tecnológica* (iBET) and Instituto de *Tecnologia Química e Biológica António Xavier* (ITQB), under the supervision of Dr. Ana Paula Batista (iBET) and Dr. Paula Alves (iBET). The work stemmed from challenges posed by Dr. Stefan Ewert (Novartis AG, Basel), which kindly provided scientific support throughout this body work. Dr. Stefan Ewert also provided access to his lab in Basel's HQ, where I was able to interact with the Basel's colleagues and bring new methodologies to the Novartis Pharma Lab, namely on the use of semi-automated phage-display protocols using the King Fisher Flex platform.

During my PhD project I had the opportunity to attend several courses and conferences. A Design of Experiments (DoE) course at the University College of London (in 2018), as well as a DoE training workshop by Sartorius (in 2019) at iBET, which allowed me to gather DoE knowledge necessary to undertake responsibilities at an internal project that happened parallelly to the project presented on this thesis.

The project focused on the “cell-free” *in vitro* expression of antibodies, which aimed to scale-down and increase the throughput of candidate testing.

I was also given the opportunity to participate in 3 international conferences (European Antibody Congress Basel 2019, PEGS Lisbon 2019, and European Antibody Congress 2020). This allowed me to present a poster related to the internal “cell-free” project (titled: “*in vitro* expression of single chain variable fragments and their comparison with in vivo e. Coli production”), and a poster on recent advancements we did on NGS (titled: Antibody discovery powered by NGS Retrieving Fab sequences with coupled V_L-V_H information and six CDR assessment by cluster coordinate matching).

From 2020 onwards the PhD project was done amidst the COVID-19 pandemic, which severely impacted laboratory work and cut short other outreach activities. Time out of the laboratory opened the window for more *in silico* approaches to be tested at the lab. Simulations of off-rate experiments done in excel and power-law analysis (PLAS) software’s and were instrumental in future lab work. Sequence probability calculation was explored as an added parameter for candidate selection. A python script for the analysis of mutational co-variance was also explored, and recently taken by the Data Analytics team at iBET for further development.

This Thesis is composed by 6 chapters and deals with all steps of the antibody discovery process and antibody library development. It also comprises innovative approaches to old problems that add to the value of this body of work and that go beyond the core objective of this Thesis.

Acknowledgments

I would like to express my deep gratitude to Dr. Ana Paula Batista for her invaluable guidance and trust, and Dr. André Faustino for his continuous support, lessons given and eagerness to teach. I am thankful for having learnt such a great deal from you both. This thesis would not have been possible without Dr. Stefan Ewert, who challenged us to tackle this project and kindly shared his knowledge and very constructive criticism. Dr. Stefan Ewert and Dr. Patrik Maurer were instrumental for this thesis’ success, both through their meaningful advices as part of my thesis commission, but also by thoughtfully bridging the Novartis Discovery Technologies lab in Basel with our lab at iBET and sharing their expertise. I would like to thank Prof. Manuel Carrondo and Prof. Paula Alves for providing me the golden opportunity to do my thesis at iBET and for always setting the bar high. Finally, a big thank you for my friends and colleagues at the NVS lab, you are the best team I could ask for.

This thesis would also not be possible without the overwhelming support of many people around me. To my bffs André Barateiro and Mariana Dias – we made it, after all these years. I am proud of what we achieved and how we did it. To my love conspiracy gang – Gonçalo Sá, Gonçalo Silva, Rita Gomes, Zé Pereira and Sara Pereira – you stormed into my life when I needed the most, and I am forever grateful for having the luck to meet you all. Chaotic-good vibes only. To my dear friend André Costa, being your friend is a privilege and speaking with you always energized me when motivation was low. To my best-man, Afonso “Coxo” Rodrigues, I know I will always count on you. To my family, Sabina, Jorge, Pedro, Sandra, Joana, and Sara, I know you probably still do not know what I am doing (xD) but I did it! To my grandpa Jorge Moura, I know you would be proud.

To my “chéfinha” Dulce, Artur and Marta, Carlos and Manuela, I will never forget how you embraced me into your family.

I dedicate this thesis to my dear grandma, Maria Amélia Sampaio, my soulmate and the family’s cornerstone that made all of this possible. This thesis is also dedicated to my dear mom and dad, thank you for always being my greatest fans and for your unconditional love and support.

Finally, to the love of my life, Joana Ferreira, thank you for going on this journey with me, for being my light at the darkest times, and for cheering for me at my best. I will forever cherish every second spent with you.

Table of Contents

Resumo

Os anticorpos terapêuticos monoclonais (mAbs) são um dos principais impulsionadores de receitas do mercado farmacêutico. Independentemente da origem e plataformas utilizadas, os anticorpos monoclonais gerados contra um determinado alvo podem ter margem para melhoria. A utilização de bibliotecas *in vitro* de *affinity maturation* visa ultrapassar as limitações das abordagens clássicas de *affinity maturation* por cristalografia de raios X, fornecendo uma abordagem generalizada (ou cega) que pode ser aplicada a muitos candidatos de uma vez. Os métodos generalizados usados actualmente nem sempre asseguram que são encontradas mutações sinergisticas, e podem não respeitar as limitações estruturais da molécula IgG em questão. Idealmente, novos métodos *affinity maturation* deveriam ser generalizáveis para fornecer resultados de alto rendimento, mas com um certo grau de especificidade em relação à estrutura de anticorpos que está a ser considerada. Como tal, é necessário que se preste atenção a regiões específicas, tais como as que podem estar em contacto com o antigénio, ou regiões que influenciam a integridade estrutural dos anticorpos ou a sua viabilidade de produção.

Nesta tese, desenvolvemos uma abordagem de *affinity maturation* semi-cega utilizando bibliotecas primárias estruturalmente preparadas, para se chegar a um compromisso entre generalização e precisão. As bibliotecas têm mutações CDR-null em *hotspots* estruturalmente identificados. Estas posições podem depois ser alvo durante a *affinity maturation* para maximizar a probabilidade de encontrar mutações benéficas que melhorem a afinidade.

Em primeiro lugar, estruturas de anticorpos (FW-κ e FW-λ) foram sujeitas a mutações nas suas sequências germinativas CDR (LCDR1, LCDR2, LCDR3, HCDR1, HCDR2) com o objectivo de reduzir a probabilidade de contactos antigénio-anticorpo nessas regiões. Os resíduos polares e carregados que apontavam para o solvente foram substituídos por serinas e alaninas, e o efeito

dessas mutações foi avaliado em termos de agregação, hidrofobicidade, e estabilidade térmica. As mutações que não levaram a uma destabilização excessiva da estrutura foram combinadas para gerar várias estruturas "CDR-null" distintas. As mutações "CDR-null" introduzidas também foram capazes de alterar a cinética de ligação das estruturas parentais.

Assim, foram geradas bibliotecas primárias com mutações CDR-null (denominadas FW-κN1 e FW-κN2), e diversificadas na região HCDR3. Estas foram então utilizadas em protocolos de *phage-display* contra Herceptina, e avaliadas na sua qualidade global, reprodutibilidade, e diversidade de sequências de HCDR3. Os conjuntos de dados κN1 e κN2 eram também essencialmente diferentes da FW-κ, em termos de diversidade de sequências de HCDR3 e distribuição do comprimento das sequências HCDR3. As bibliotecas κN1 e κN2 também conseguiram produzir candidatos de todos os tamanhos de HCDR3, com grande afinidade para a Herceptina, e distribuição igual de afinidades quando comparadas com a biblioteca κ.

Finalmente, utilizámos diferentes estratégias de *affinity maturation* para os candidatos oriundos de FW-κ, FW-κN1 e FW-κN2 para avaliar se a utilização das bibliotecas primárias κN1 e κN2 conduz a resultados manifestamente melhores do que as bibliotecas baseadas em FW-κ. A biblioteca κN1 demonstrou ter a maior taxa de sucesso entre todas as condições (31,3%), especialmente quando combinada com o método de maturação por afinidade semi-cego recentemente concebido (36,4%). Isto realça o poder da estratégia subjacente a este trabalho, onde um pequeno compromisso na afinidade dos candidatos primários é compensado por um maior ganho de afinidade após a *affinity maturation*. Os candidatos oriundos da FW-κN1 foram também responsáveis pelo maior aumento de afinidade (FI), e muitas vezes reverteram parte dos seus resíduos CDR-null de volta aos resíduos de germinais originais, o que lhes conferiu melhores características para manufactura e desenvolvimento.

Esta tese abre um novo paradigma na descoberta de anticorpos, ao empregar uma abordagem contra-intuitiva na selecções primárias para alcançar melhores resultados nas fases posteriores da descoberta de anticorpos. Além disso, este método foi capaz de extrair mais sequências de HCDR3 da concepção da aleatorização, expandindo efectivamente a diversidade funcional das bibliotecas originais contra o mesmo alvo.

Abstract

Therapeutic monoclonal antibodies (mAbs) are one of the main drivers of revenue of the pharmaceutical market. Regardless of the origin and platform used, monoclonal antibodies generated against a given target may have room for improvement. Using *in vitro* affinity maturation libraries aims to surpass the throughput limitations of classical X-ray crystallography affinity maturation approaches, by providing a generalizable approach (or blind) that can be applied to many candidates. The current blind methods do not always assure that synergistic mutations are found and may not respect the structural constraints of the IgG molecule in question. Ideally, innovative affinity maturation methods should be generalizable to provide high-throughput results while maintaining a certain degree of specificity towards the antibody structure being considered. As such they require attention to be paid to specific regions, such as the ones likely to be in contact with the antigen, or regions that influence the antibodies' structural integrity and overall developability.

In this thesis, we developed a semi-blind affinity maturation approach using structurally primed primary libraries, to reach a compromise between generalization and precision. The primed libraries carry CDR-null mutations on structurally identified hotspots. These positions can then be targeted during affinity maturation to maximize the likelihood of finding beneficial mutations that improve affinity.

Firstly, antibody frameworks (FW-κ and FW-λ) were subjected to mutations to their CDR germline sequences (LCDR1, LCDR2, LCDR3, HCDR1, HCDR2) with the objective of reducing the likelihood of antigen-antibody contacts on those regions. Polar and charged residues pointing outwards towards the solvent were replaced by serines and alanines, and the effect of such mutations on the antibodies' developability was evaluated in terms of aggregation, hydrophobicity, and thermal stability. Mutations that did not lead to an over de-stabilization of the framework were combined to generate several distinct "CDR-null" frameworks. The "CDR-null" mutations introduced were also shown to alter the binding kinetics of the parental frameworks.

Then, primed libraries bearing the CDR-null mutations were generated (named FW-κN1 and FW-κN2) and randomized on HCDR3. These were then used in phage-display pannings against Herceptin, and evaluated for their overall quality, reproducibility, and diversity of HCDR3 sequences. The κN1 and κN2 datasets were also essentially different from FW-κ, in terms of HCDR3 sequence diversity and HCDR3 length distribution of outputs. The primed libraries (κN1 and κN2) were also able to yield candidates of all HCDR3 sizes, with high affinity towards Herceptin, and equal distribution of affinities when compared with the κ library.

Finally, we employed different affinity maturation strategies to FW-κ, FW-κN1 and FW-κN2 candidates to evaluate if using primed libraries κN1 and κN2 leads to manifestly better results than libraries based on FW-κ. The primed framework κN1 was shown to have the biggest rate of success among all conditions (31.3%), specially when combined with the newly designed semi-blind affinity maturation method (36.4%). This highlights the power of the strategy behind this work, where a small compromise in the affinity of primary binders is compensated by a bigger gain in affinity after affinity maturation. FW-κN1 candidates were also responsible for the biggest fold-increase (FI) in affinity, and often reverted part of their CDR-null residues back to the original germile residues, which conferred them more optimal developability characteristics.

This body of work opens a new paradigm in antibody discovery, by employing a somewhat counter-intuitive approach in the primary panning to achieve better outcomes in the later stages of antibody discovery. Additionally, this method was able to extract more HCDR3 sequences out of the randomization design, effectively expanding the functional diversity of the original libraries against the same target.

Keywords

Phage-display

Antibody Libraries

Monoclonal Antibodies

Next-generation sequencing

Affinity Maturation

Palavras-chave

Phage-display

Bibliotecas de anticorpos

Anticorpos Monoclonais

Next-generation sequencing

Affinity Maturation

Chapter 1 – Introduction

1.1. The Principles of Immune Response

The human immune response is governed by two distinct systems that act synergistically with each other, the innate system and the adaptive system. The innate system is comprised by a big variety of general defenses against external agents and is fully encoded in the host's genome. These include physical barriers such as the epithelia and the secreted mucus layers that covers it as well as cell-cell contacts themselves (tight junctions, cadherin-mediated cell interactions). Other examples of innate immunity are the complement proteins (constitutively expressed in biological fluids), and cytokines, chemokines, and reactive free radical species, which are released from the cells upon immune activation.[1] Some membrane-bound receptors that bind molecular patterns expressed on the surfaces of invading microbes are also part of the innate immune system (e.g. mannan-binding lectins).[2] In contrast to the recognition receptors of the innate immune system – which are all encoded in their fully functional form in the germline genome – adaptive immune responses depend on tailor-made receptors that arise from somatic rearrangement (also known as V(D)J recombination) of immunoglobulin (Ig) genes. These Ig genes will form intact T-cell receptors (TCR), and B-cell receptors (BCR). The TCRs locate on the surface of T-lymphocytes, and BCRs on the surface of B-lymphocytes and both will engage in antigen recognition. The latter can also be secreted in a soluble form by B-plasmocytes (B-lymphocyte progeny). BCRs and their soluble forms (IgM, IgD, IgA, IgG, IgE) are colloquially known as antibodies, and are the mediators of humoral adaptive immunity.[3]

1.1.1. The Structure of Antibodies

The basic structure of antibodies consists of two identical heavy-chains (HC) of 50 kDa, and two identical light-chains (LC) of 25 kDa. The HC and LC will be paired together to form two identical "arms" – the Fab domains. The Fab contains one complete LC and two portions of HC: V_H and C_H1. A flexible hinge region connects the C_H1 domain with the remaining HC domains: C_H2 and C_H3. From that hinge region, two disulfide bonds connect both HC to form the Fc-domain, bringing together the Fab domains and giving rise to the classical "Y" shape of antibodies. Besides connecting the two Fab domains, the Fc-region is also responsible for mediating the activity of immunoglobulin molecules, by binding to Fc-receptors and activating the complement system, as well as being involved in the isotype class switch mechanisms.

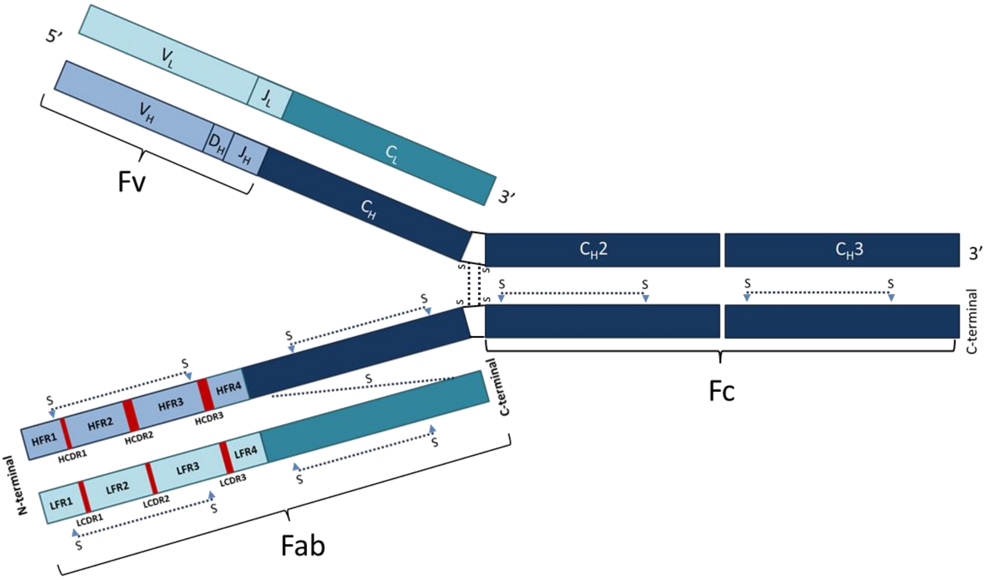


Figure 1 – Schematic Representation of an IgG molecule. On the upper half, the DNA sequences resultant from the V(D)J joining process are highlighted. On the lower half, framework regions (FR) and CDR loops (in red) are highlighted.

The Fab domain consists of two variable domains and two constant domains. The C-terminal half of the Fab domain, which serves as a structural framework, is composed by constant domains from the heavy (C_H1) and the light chain (C_LK or $C_L\lambda$). The N-terminal half of the Fab domain is composed by a variable heavy-chain (V_H) and a variable light-chain (V_L). The juxtaposition of these variable regions makes up the variable fragment (Fv), which varies greatly amongst different antibodies, and will engage on the recognition of antigen molecules. Such variability is a direct consequence of somatic recombination, which shuffles a few hundred germline-encoded immunoglobulin genes into millions of different antigen receptors, each with a potentially unique specificity for a different antigen. Moreover, both V_H and V_L contain three hypervariable sequences, called **C**omplementarity-**D**etermining **R**egions (HCDR1/2/3 and LCDR1/2/3, respectively), which are situated between four stable regions termed framework regions (HFR1/2/3/4 and LFR1/2/3/4, respectively) (Figure 1).[4]

1.1.2. V(D)J recombination and CDR diversity

The hypervariability of these regions is a direct consequence of the somatic recombination of V(D)J genomic segments that occurs in the three immunoglobulin gene *loci*. V stands for variability, D for diversity and J for joining.

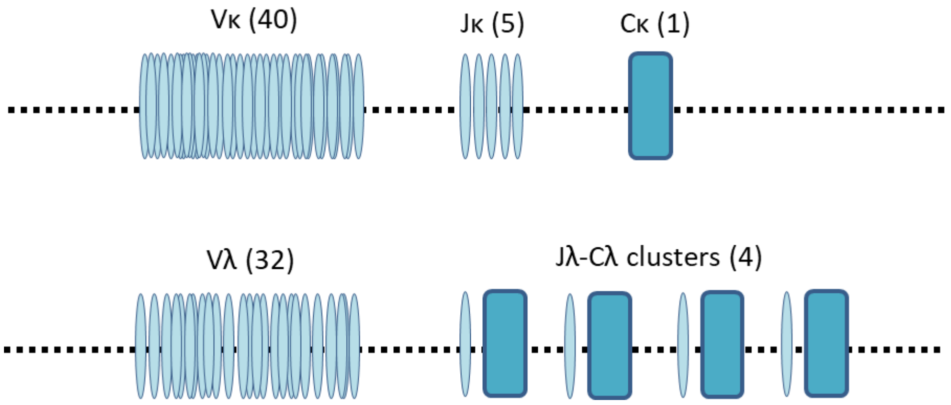


Figure 2 - Schematic representation of V, J and C gene segments located on κ and λ *loci*.

For the light-chain *loci* – κ and λ – only V and J segments are present. The κ *locus* has 40 Vk segments, 5 Jk segments, and 1 Ck segment (which gives rise to C_L). The Vκ segments contains LFR1, LCDR1, LFR2, LCDR2, LFR3 and the N-terminal portion of LCDR3. The Jκ element contains the C-terminal portion of LCDR3 and FR4 in its entirety. Combinatorial VJ arrangements in the κ *locus* can generate more than 140 different sequences. Additional diversity arises during the recombination events. Recombination events require the intervention of RAG1/2 enzymes with double-strand break activity, followed by a DNA repair process known as nonhomologous end-joining (NHEJ), which can lead to the loss of 1 to 5 nucleotides upon recombination. This imprecision can be corrected by terminal deoxynucleotidyl transferase (TdT), which adds random N nucleotides until the coding frame is restored.[5] Each codon created by N addition increases the potential diversity of the repertoire by 20-fold. Thus, the initial diversification of the κ repertoire is focused at the VJ junction, where LCDR3 is located. The λ *locus* follows the same principles, with each individual carrying up to 32 Vλ segments that can be arranged with 4 Jλ segments, which are already pre-associated with their respective Cλ segment (Figure 2).[4]

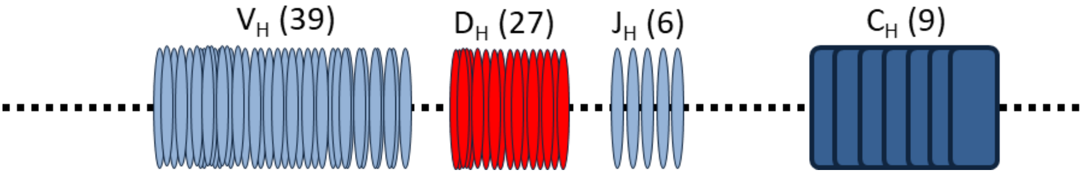


Figure 3 - Schematic representation of V, D, J and C gene segments on immunoglobulin H *locus*.

For the heavy-chain *locus*, D segments are also present. Of the many V_H segments identified, only 39 are functional V_H, and these will recombine with 27 D_H and 6 J_H gene segments (Figure 3). The V_H gene segment contains HFR1, HCDR1, HFR2, HCDR2, HFR3, and the N-terminal portion of HCDR3. The D_H gene segment forms the middle of HCDR3, and the J_H segment contains the C-terminal of HCDR3 and

HFR4 in its entirety (Figure 3). As with the light-chain *loci*, even though the combinatorial pairing of V_H and D_H with J_H generates more than 104 different VDJ combinations, it is the junctional diversity that is a major source of variability of the immune repertoire.[4,6,7] The intersection of V, D and J segments and the junctional diversity that arises from that is the major reason why HCDR3 is the most diverse of all the CDRs.

1.1.3. The CDR loops role in antigen binding

The HCDR3 sequence diversity also means that the HCDR3 will be more structurally diverse and thus lead to a bigger amount of possible paratopes.[8] In fact, while HCDR1/2 and LCDR1/2/3 are not expected to deviate much from a well detailed set of canonical structures – which can be accurately predicted by 3D modeling methods (resolution < 1.0 Å) – it is currently very difficult to reliably predict the HCDR3 loop structure.[9–13] Such variability, together with its advantageous structural position in the center of the binding site,[4,14,15] makes the HCDR3 the biggest contributor for the antigen binding process.

Despite the HCDR3 dominance, all the other five CDR can contribute to antigen binding. The HCDR2 has the biggest median CDR length (14aa versus 11aa in HCDR3) and is often the second biggest contributor in antigen binding.[16] Five different canonical structures were described for HCDR2.[13] LCDR1, LCDR3 and HCDR1 have been found to contribute with the same amount of energetically important residues in antigen binding, and have been identified has having eight, seven and six canonical structures, respectively.[13] LCDR2 is often the loop that least contributes to binding and only two canonical structures have been identified.[13,16] Nonetheless, the relative importance of each CDR will largely depend on the Ab-Ag complex in question. There are cases where one or more CDRs do not contact the antigen at all (HCDR3 included), and/or where the HCDR3 is not the one that contributes with the highest number of antigen-binding residues.[16] Additionally, differences in CDR contribution to binding seem to arise when comparing natural antibodies with those of synthetic origin (for more information on of synthetic mAbs see section 1.3.3.).

In 2016, a study that analyzed non-redundant set of 138 Ab-ag complexes (101 natural, 37 synthetic) revealed that synthetic Abs rely heavily on HCDR3 at the expense of HCDR2 and HCDR1, when compared to natural Abs. The most striking relative change in importance from natural Abs to synthetic Abs occurred in salt-bridge formation (HCDR1: from 11% to 1.6%; HCDR2: from 40% to 16%; HCDR3: from 26% to 61%), which correlated with a decrease in charged amino acids in HCDR2 (E, D, H, K and R). Likewise, changes in H-bond contribution (HCDR1: from 17% to 10%; HCDR2: from 22% to 18%; HCDR3: from 30% to 40%) were also associated with decreased frequency of polar amino acids in HCDR1 (M, N, Q, S, T, W and Y). A slighter change occurred in cation-pi interactions (HCDR1: from 22% to 11%; HCDR2: no change (26%); HCDR3: from 20% to 26%). Here, LCDR3 contribution also changed from 13% to 20%, which further explains the reduction in cation-pi interactions by HCDR1.[17] Since synthetic libraries have historically focused on the diversification of HCDR3 – and the LCDR3 to a certain extent – these observations do not come as a surprise. Hence, the authors question whether heavily focusing on HCDR3 diversification is the best strategy, or if allowing higher diversity in all CDRs simultaneously would allow a more diverse paratope repertoire, and in turn allow the libraries to recognize a greater panel of epitopes.[17] However, careful analysis of other datasets and better integration of other parameters (affinity, stability, hydrophobicity, polyreactivity) should be taken into account before jumping into conclusions. For example, it has been proposed that extensive deviations from the germline are associated with lower library fitness[18] and that the presence of certain residues in specific CDRs may lead to disadvantageous developability profiles.[19] Both of these topics will be explored further below.

1.1.4. The primary immune response and polyreactivity of natural repertoires

A closer look to natural repertoires can help us understand how antibodies bind to the antigens. Intuitively, having a large repertoire of antibodies increases the chance of finding paratopes that bind to the antigen. Recent estimates of naïve repertoires go from around 10^{11}-10^{12} up to 10^{15} - 10^{18} sequences.[18,20–22] However, these number seem unlikely to occur in a single individual, since the total number of cells of all types is 10^{13}, the total number of B cells in the body around 10^{11} and the number of circulating peripheral naïve mature B-cells (CD27−/IgD+) at any point in time never surpasses 10^{9} individual cells.[23,24] It seems that as more individuals are used in the naïve datasets, the probability of finding unique sequences also increases. As pointed out by Briney et al. dataset[21], "largely unique repertoires" were found for each individual studied. This indicates that, rather than each individual having a 10^{15} repertoire, this value is the representation of the sum of the overlapping repertoires within the total human population.[15] A more intuitive size arises from the calculation of the combinatorial possibilities of shuffling the κ and λ gene segments (Figure 2) with the one on the H *locus* (Figure 3). Such estimate gives rise to around 2 x 10^{6} VH-VL pairs. N- and P- junctional diversity have been suggested to increase this value by a factor of 10, giving a diversity >10^{7}.[4,15] This large but finite number of antibody sequences seems far less than the number of epitopes on foreign antigens to which one could be exposed. To had insult to injury, sequences found in circulation are clearly biased to certain subsets of V_H families, and κ an λ families, rather than being an homogenous representation of the total diversity available in the genomic human repertoire.[25–27] Since it is the 3D structure of the antibody that determines binding, not it's sequence *per se*, structural information can help us answer these questions. In fact, it is known that CDRs belonging to the same canonical class (i.e. that have nearly identical structures) can have very different sequences, and that similar HCDR3 sequences may adopt different conformations.[28,29] Thus, rather than looking at unique sequences, the structural diversity should be taken into account if we want to fully understand antibody function.[8] Keeping with that theme, authors explored the notion of “shape space” of antigen epitopes.[30,31] In this model, each individual antibody structure is able to bind to a given structural shape. This means that a single antibody structure may recognize several unrelated epitopes, provided that they present similar shapes. This structural redundancy is most commonly referred to as polyreactivity or polyspecificity, and has been vastly associated with antibodies triggered early in the response (e.g. IgMs) and germline sequences.[15,19,32,33] This mechanism has been recently termed “conformation flexibility hypothesis”. It suggests that germline gene-coded antibodies retain a degree of structural plasticity in their backbone in order to maximize the number of different unrelated antigens that they can recognize. In fact, around 20% of B lymphocytes in the peripheral blood make polyreactive antibodies,[32] and a study of 137 therapeutic mAbs showed that the absence of somatic mutations in germline sequences is a good predictor of polyreactivity.[19] Older studies also report that poly-specific antibodies retain a larger amount of germline sequences than more specific antibodies.[34,35] Hence, germline sequences provide poly-reactive surfaces that can bind to a wide range of structural antigen epitopes with sufficient affinity to initiate an immune response. This allows for a limited diversity repertoire to screen a panel of epitopes that is potentially bigger than its sequence-encoded diversity, in a resource efficient manner.

1.1.5. Affinity maturation in vivo

The polyreactive antibodies that constitute the primary immune response are usually IgM or IgD with relatively weak affinities towards the antigen. But even if binding weakly, these primary binders are sufficient to induce the polyclonal expansion of B-cells, which will then undergo class switching and affinity maturation. These processes take ~2 weeks and give rise to IgG, IgA, and IgE isotypes with higher affinity towards the antigen and different effector functions.[3] These gains in affinity are derived from somatic hypermutation and selection mechanisms, and lead to high specificity towards their cognate antigen, and higher structural rigidity (as opposed to poly-reactivity and structural flexibility).[32,33] Somatic hypermutation is the final mechanism of immunoglobulin diversity. It consists in the apparently random substitution of antigen-binding residues. If these mutations result in loss of affinity for the antigen, the cell loses important receptor mediated growth signals and dies. If, however, the mutations result in increased affinity for the antigen, then the cell producing that antibody has a proliferative advantage in response to antigen and grows to dominate the pool of responding cells.[1,3,4]

1.2. Discovery of monoclonal antibodies

The natural response of any given individual with a healthy immune system is to produce several antibodies that bind to the same target pathogen or antigen. Since these antibodies bind to slightly different epitopes in the antigen and are originated from different B cell progeny, this is also called a polyclonal response.[36] Polyclonal antibodies (pAbs) can be isolated from the serum of an immunized donor and be used in a variety of applications. They are mostly used as a secondary antibody in immunoassays such as ELISA, Western Blotting, Flow Cytometry. Their usage in therapeutic application is avoided due to their tendency to cross-react. In contrast, a monoclonal antibody (mAb) will only bind to one epitope of the antigen with high specificity, and will translate easier into high scales of production.[37] Expectedly, mAb formulations are more homogenous than pAbs and provide more reproducible results across the antibody development pipeline. Additionally, due to the intrinsic nature of pAbs discovery method, which relies on the immunization of individuals, batch-to-batch variations are expected to occur. Due to their high specificity, potency, and robustness, therapeutic mAbs are one of the main drivers of revenues of the pharmaceutical market. The global mAbs market is valued at around 115 billion US dollars and is expected to grow to about 300 billion US dollars until 2025.[38] As of November 2020, 88 mAb products were under late-stage clinical investigation (6 for COVID-19).[39] In May 2021, it was announced that the FDA granted marketing approval for its 100th mAb product.[40]

1.2.1. Early technology development

Monoclonal antibody production was firstly achieved by hybridoma technology, on the seminal paper by Kohler and Milstein. The generation of hybridomas involves immunizing a certain species against a specific epitope on an antigen and obtaining the B-lymphocytes from the spleen of the animal. The B-lymphocytes are then fused (by chemical- or virus induced methods) with an HGPRT-negative immortal myeloma cell line and cultured in vitro in selective medium containing hypoxanthine-aminopterinthymidine (HGPRT) to select for positive fusions.[41] Hybridoma technology was responsible for the first monoclonal antibody to be licensed for therapeutic use in humans. Orthoclone OKT3 (muromonab-CD3) was approved in 1986 for use in preventing kidney transplant rejection.[42] However its use was limited to acute cases due to reports of high immunogenic reactions due to the production of anti-antibodies by the patient after administration. To address these effects, scientists started manipulating antibodies that had been discovered by immunization of mice. The first approach was to develop chimeric recombinant antibodies. This technology was developed in 1984 and involved taking the variable region genes of a mouse antibody-producing myeloma cell line with known antigen-binding specificity and joining them with an human immunoglobulin constant region.[44] The first licensed antibody that came from using this technology was adciximab (ReoPro) in 1994. Notably, it also lead to the development of the tumor necrosis factor (TNF)-specific antibody infliximab (Remicade; Centocor/Merck) which is routinely used to treat rheumatoid arthritis, as well as Crohn’s disease and plaque psoriasis,[2,5] and also rituximab (Rituxan/Mabthera; Genentech/Roche/BiogenIdec), which is used to treat both rheumatoid arthritis and non-Hodgkin’s lymphoma.[42,45]

With the aim of reducing immunogenicity, a more fine-tuned approach was developed in 1986 by Jones et al, which did not rely on replacing the full variable region.[46] Instead, only the CDRs from the mouse antibody were grafted into a human framework, a process latter known as humanization. The first antibody to be licensed through this technology was daciizumab (Zenapax) in 1997, and it has also led to development of the human epidermal growth factor receptor 2 (HER2)-specific antibody trastuzumab (Herceptin; Genentech/Roche), and the vascular endothelial growth factor A (VEGFA)-specific antibody bevacizumab (Avastin; Genentech), both of which are used in the treatment of several types of cancer;[42,45]

To surpass this, two alternatives arose: the development of humanized mice[ref] – which have a fully human immune system, and thus, generate fully human antibodies –, and *in vitro* display platforms, which answer to the limitations of animal immunization approaches.

From an operational standpoint, animal immunization studies require access to a mouse breeding facility, with certified technicians to perform the assays. The turnover of mice available for experimentation is limited to the resources of the laboratory and of the breeding facility, resulting in low throughput. Ethically, animal experimentation has also been contested, and is often cast-off in favor of *in vitro* alternatives. Additionally, by carefully controlling selection and screening conditions – e.g. by the inclusion of competitors to guide the selection towards specific targets or epitopes – *in vitro* display technologies allow the generation of antibodies to defined antigen epitopes, which cannot be done in animal immunization approaches. Finally, popular *in vitro* methods such as the ones based on microbial systems – phage and yeast display – have very high potential regarding parallelization, automation, and miniaturization.

1.2.2. Phage-display: history

The phage-display technique dates back to 1985 when Nobel prize laureate George Smith successfully fused a foreign peptide with the pIII coat protein from filamentous Fd phage. Those phages could then be enriched by *in vitro* phenotypic selective pressure, purified by affinity chromatography, and their DNA extracted to recover the genotypic information encoded inside that same phage at the end of the process.[47] In 1988, George Smith and Stephen Parmley proved that they could purify those phages using antibodies against the cloned gene product from a pool of 10^8 wild-type phages that were not bearing the determinant, and coined the term “panning”, due to the resemblance with the method of finding gold.[48] Their work led to a landmark paper by another Nobel prize laureate, Greg Winter, who generated the first phage-display library. Taking advantage of the small single-chain variable fragments (scFv) that had just been discovered [49], Greg Winter and his colleagues amplified those immunoglobulin variable genes from hybridomas and B cells, and cloned them into phage-display expression vectors. The successfully display of fully functional scFvs meant that the phages carrying the scFv could be selected against their cognate antigen and purified afterwards.[50] Hence, it opened the possibility of finding a rare clone that bearded affinity towards a target of interest among a big pool of non-specific binders, after applying selective pressure.

1.2.3. The phagemid system

The industry standard system for phage display was developed shortly after its discovery, in 1991 by the Deutsches Krebsforschungszentrum (DKFZ) in Heidelberg, Germany. Instead of cloning the expression cassette on the complete phage genome – as was done by MRC Laboratory of Molecular Biology in Cambridge, United Kingdom[50] and at the Scripps Research Institute in La Jolla, USA,[51] – the antibody:pIII fusion gene was cloned into a M13 phagemid system which uncouples the antibody cassette from the process of viral packaging.[52] In this system, in addition to the antibody:pIII cassette, the phagemid carries an antibiotic resistance gene and two origins of replication, one from the filamentous phage – usually f1 ori –, and another one from the bacterial vector, usually *E. coli*. The phagemid will behave as a normal plasmid in the absence of phage proteins, allowing for genetic manipulation and cloning procedures. Because the phagemid cannot produce viral particles alone, it is required the co-infection of a *helper-phage* that encodes all of the remaining structural and morphogenetic proteins necessary for viral replication and assembly. This helper phage will also drive the expression of the f1 origin of the phagemid so that the antibody:pIII cassette can be incorporated inside the phage as single-stranded DNA. After phage packaging, the antibody fragments are displayed as pIII fusion proteins on the surface of M13 phage, and the corresponding antibody gene fragment is packaged in the same phage particle.[52] Due to its smaller size (~5kb) it is also capable of yielding higher transformation efficiencies, when compared with full phage genome approach (~8.5kb). This system became the industry standard, and it is still used nowadays, due to its flexibility, transformation efficiency, and robustness of M13 phages. M13 is a filamentous, non-lytic bacteriophage that is 6–7 nm in diameter and 900 nm long, 12000 kDa, and a member of F positive (F+) phage family. This means that replication of M13 phage initiates through binding of M13 to its receptor on the bacterial cell (bacterial F-pilus). Therefore, bacterial cells that contain pili can be infected with M13 phages.[53] Compared with the other members of its family – Ff, fd, and f1 – M13 is the easiest one to manipulate and purify.[54]

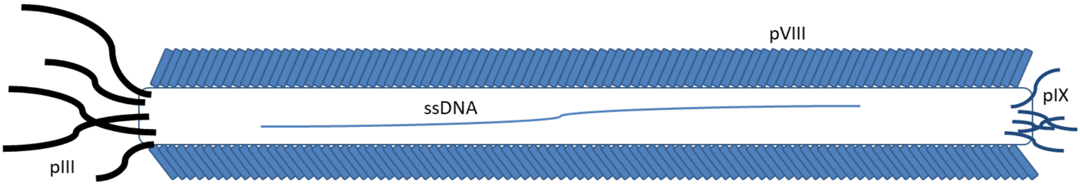


Figure 4 – Schematic representation of a M13 phage.

The protein capsid of filamentous phage comprises several coat proteins, including pIII and pVIII. The pVIII is the major coat protein that wraps around DNA. There are about 3000 copies of pVIII protein per phage particle, depending upon the length of phage genomic DNA. In contrast, the minor coat protein, pIII, is present at only 4–5 copies per phage and plays a crucial role in infection process of the bacterial cell, by binding to the F-pilus of bacteria. The pIII protein interacts with pVIII proteins through its C-terminal domain to maintain attachment to the phage coat, while its N-terminal portion mediates the attachment to the F pilus essential for subsequent infection. Protein or peptide insertions at the N-terminus of pIII typically do not interfere with pIII function or infectivity, as long as they are inserted after the first 3-5 N-terminal residues and the signal sequence responsible for viral assembly is kept.[55] A linker sequence (usually composed by GGGGS repeats) is used to separate the displayed epitope from the rest of the coat protein, allowing for these N-terminal epitopes to be displaced from the rest of the phage particle, making them more accessible for binding interactions with an antigen of interest.[56]

Depending on the helper-phage used, different degrees of display can be achieved on a single phage particle. The most commonly used helper phage is M13KO7.[57] Since the *wild-type* pIII gene from M13KO7 has superior expression levels compared to the phagemid-encoded pIII-antibody fusion gene, the majority of the produced phage population is expressed without a pIII-antibody fusion. This means the population of phages will tend to be composed of phages that have one copy of the pIII-antibody fusion (monovalent display), with the disadvantage of also having phages that do not have the pIII-antibody fusion at all.[58] In contrast, an hyperphage system has been used to ensure that all phages have pIII-antibody fusions, since it uses a helper phage lacking the pIII gene. This will lead to a pIII-antibody polyvalent display, because only the pIII-antibody gene of the phagemid will be expressed.[59] However, concerns that polyvalent display may artificially select weaker binders due to avidity effects may continue to drive the utilization of monovalent display and M13KO7 helper-phages. In the context of this thesis, a M13KO7-based phagemid system was used, with Fab fragments bound to truncated pIII proteins.

1.2.4. Discovery of high-affine antibodies using phage-display

For a phage-display protocol to be successful, the following conditions have to be met: i) The antigen must be produced in sufficient amounts, with high degree of purity and in the correct conformation[60]; ii) The phage-display library needs to have the appropriate size and ensure diversity (further explored on section 1.3.); iii) Selective pressure after exposure to the antigen, by repetitive wash and enrichment cycles;[61] iv) Recovery of positive binders and analysis of results (Figure 5).

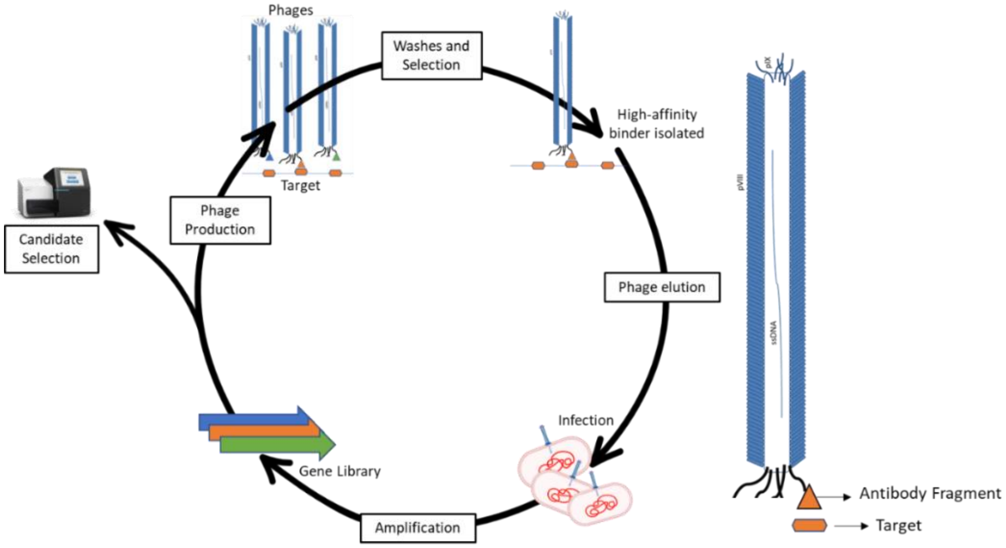


Figure *5* – Schematic representation of *phage-display* process.

Solid panning constitutes the traditional method for affinity screening of phage display libraries, where the antigen molecules are immobilized on solid surfaces, most commonly on polystyrene plates or immunotubes. These provide a simple platform where phages can be easily presented to the antigen and washed to remove unbound or unspecific fraction of phages.[62,63] Since washing is carried out on the same plate where specific binders will be eluted from, several repetitive "batch" operations are required to make sure all unspecific binders are removed. To avoid this shortcoming, chromatographic approaches have been employed. Immobilizing the antigen inside columns not only increases surface area but also allows for washing step to be done in a single continuous operation.[64,65] In any case, upon the adsorption of the antigen molecule, it is of the most importance to block the remaining sites on the surface to prevent non-specific phage binding before incubating with the phages. Bovine serum albumin, milk or other non-serum commercial alternatives are usually used in this step.[66–68] However, the immobilization of antigens in surfaces, be it a plate or a column, may end up concealing epitopes of interest from the paratopes of antibody libraries, and can lead to conformational distortion and/or denaturation of the antigen.[58,69]

Selection of antibodies on the aforementioned conditions will lead to a decrease in paratope diversity and potentially to the accumulation of phages displaying antibodies with affinity to conformations that to not resemble the native antigen. Solid-phase panning strategies have used streptavidin-coated plates to immobilize biotinylated antigens, in an effort to lift the antigen and uncover it's epitopes, with the drawback that additional depletion steps must be added to remove potential streptavidin binders.[58]

Alternative solution-phase panning strategies have been employed, where the antigen and the phages meet in solution, after which the specific binders are selected from. That is usually accomplished by biotinylating the antigen and capturing the biotin-antigen-antibody complexes with magnetic streptavidin/neutravidin beads. The increased surface area available on magnetic beads allows the capture of a much greater amount of antigen. Combined with the greater availability of epitopes in solution, when compared to the immobilized antigen approaches, it is believed that solution-phase panning should recover a more diverse pool of hits.[18,58,70]

Although providing a simple and cheap option, biotinylating antigens can be challenging and impair the assay if not done correctly. The more common chemical biotinylation is not uniform, its difficult to control, and can lead to the formation of heterogenous products. Over-biotinylation of the antigen may heavily decrease the available surface area on the antigen, thereby preventing phage binding, and may also alter the physical and chemical properties of the antigen, possibly leading to undesirable effects, such as antigen aggregation. The biotin-to-antigen ratio can be empirically controlled, but even then the chemical biotinylation will randomly modify any available lysine residues, which may contribute to the heterogeneity of the mixture.[71] Site-specific biotinylation (or enzymatic biotinylation) can be achieved by co-expression of bacterial biotin ligase (BirA) and an exogenously expressed protein of interest that is modified to carry a biotin acceptor peptide.[72,73] This leads to a more controlled process, but also to higher operational costs and longer times of antigen production. Moreover, due to the specificities of each panning campaign, it is highly unlikely to find commercially available site-specific biotinylated antigens that suit the projects needs. As such, unless such a protocol is available *in-house*, most laboratories tend to opt for the simpler chemical biotinylation.

In both solid-phase and solution-phase panning, washing steps are required after incubating the antibody-displaying phages with the antigen. This allows the removal of unbound phages and unspecific phages before moving to the final steps of the panning procedure. The washing steps are critical in every panning procedure since their stringency can be manipulated to achieve many goals. The wash buffer composition can be changed by adding detergents, or by manipulating pH and salt concentration. The duration and vigor of washes can also be controlled, for example, long wash times can be incorporated to specifically select candidates with slow dissociation rates.[62,74] Moreover, the washing steps are gradually increased with every round of panning to increase the stringency in order to isolate higher affinity phage clones.[75]

The elution and recovery of high-affine phages can be achieved through changes in pH, using glycine or citric acid[76,77], or through proteolytic cleavage of cleavage sites between the antibody and pIII protein.[78,79] Eluted phages are used to infect bacterial suspensions, which are then inspected to retrieve information on antibody sequences. This was traditionally done by manual colony picking and Sanger Sequencing[80]. Such allowed for the selection of dominant clones but provided a small snapshot of the total diversity of candidates. While automated colony picking strategies have been implemented to achieve the inspection of up to >10^3 clones per experiment, the usage of next-generation sequencing (NGS) approaches has allowed a much deeper inspection of candidate pools after the final round of panning.[81–83]

1.2.5. Next-generation Sequencing and Antibody discovery

The advent of NGS has increased throughputness and decreased the costs of sequencing large genomic information, when compared with the traditional methods. NGS technologies can provide around 10^7 sequences, a 10.000-fold improvement when compared with common Sanger Sequencing strategies. There are five main NGS platforms available in the market: Illumina, 454, Ion-Torrent, SOLiD and PacBio. These differ in their DNA amplification strategy, chemistry for sequence determination, total number of reads, read length, and error rates, as thoroughly reviewed by Hodkinson and Grice.[84] All have been successfully used for the deep inspection of library diversity before and after panning campaigns. Genes encoded on antibody libraries go from 300bp for single-domain antibody (sdAb) libraries, up to almost 1500bp for Fab libraries. Additionally, some libraries may have diversity encoded to several CDRs at the same time. This means the suitability of each platform will depend on the type of antibody library used, and how far apart the regions of interest are from one another. All NGS platforms will be able to provide big amounts of reads under 100bp, which makes them all compatible to assess the diversity of libraries that focus on a single CDR (for example, HCDR3). As such, the most popular platform among researchers is the Illumina system, due to its cost-effectiveness and larger amount of data generated.[84] However, 454 pyrosequencing was used in the past over the Illumina system when longer sequences were needed (up to 1kb).[25,84,85] Nonetheless, Illumina and Ion-torrent applications can provide good throughputness on the 300-400bp range, and the latest Miseq is able to provide up to 600bp reads through the paired-end sequencing of 300bp reads. This is particularly useful for the simultaneous sequencing of V_H and V_L regions in scFv and Fab sequences.[82] Finally, PacBio applications are able to provide the biggest read length of NGS applications, which comes at the cost of a lower number sequences read.[86] Throughputness also comes at the cost of error susceptibility. There is a probability of 10^{-2} errors per base in the case of Ion Torrent and PacBio applications, and around 10^{-3} per base for Illumina.[87] Besides panning analysis and candidate selection, NGS has also proven to be a useful tool for the quality control of phage-display libraries, of both naïve and synthetic origin, in terms of their CDR length distribution, germline frequencies and clone redundancy.[85,88,89] The implementation of NGS technologies in phage-display applications has provided insights on all stages of antibody discovery: library generation and diversity assessment, quality control, and candidate selection. It is expected that NGS and phage-display advancements synergistically lead to the discovery of more potent antibodies and contribute to a greater amount of phage-display-derived mAbs in the market.

1.2.6. Phage-display and the biopharma market

In 2002, twelve years since phage-display's first implementation, adalimumab (Humira®, AbbVie Inc) became the first therapeutic mAb derived from phage display to be granted a marketing approval. It has since become the most successful mAb on the market, and is currently prescribed for a wide variety of immune-mediated disorders (rheumatoid arthritis, juvenile rheumatoid arthritis, Crohn's disease, psoriatic arthritis, psoriasis, axial spondyloarthritis, ulcerative colitis, uveitis, hidradenitis suppurativa and Behçet syndrome).[90,91,4] To date, a total of 14 antibodies derived from phage display were approved for use in the clinic, and more than 70 have undergone or are undergoing clinical evaluation (Table 1).[62,92] Despite phage-display's notorious advantages – such as bypassing animal immunization, the ability to isolate antibodies against toxic or non-immunogenic antigens and the ability to generate antibodies against specific epitopes by the use of competitors –, the vast majority of the approved therapeutic antibodies are still derived from immunized mice technologies. This dominance is not related to higher affinity against targets, but rather from natural quality-control processes that are imposed by the natural immune system, which enables natural antibodies to have better biophysical attributes when compared to antibodies generated by phage display.[93] However, this gap is expected to shorten as we move towards a better understanding of library design, better selection mechanisms that filter-out biophysical liabilities, higher screening throughputness, and optimization of developability testing.

Table 1 – Phage-display derived antibodies that have been granted marketing approval. Adapted from: Nagano, K. & Tsutsumi, Y. Phage Display Technology as a Powerful Platform for Antibody Drug Discovery. *Viruses* **13**, 178 (2021).[92]

Name		year	Origin
Humira®	Adalimumab	2002	Humanization and guided selection
Lucentis®	Ranibizumab	2006	In vitro affinity maturation
Benlysta®	Belimumab	2011	CAT's library (human naïve scFv)
ABthrax®	Raxibacumab	2012	CAT's library (human naïve scFv)
Cyramza®	Ramucirumab	2014	Dyax's library (human naïve Fab)
Portrazza®	Necitumumab	2015	Dyax's library (human naïve Fab)
Taltz®	Ixekizumab	2016	Lilly Research Laboratories' mice immune library
Tecentriq®	Atezolizumab	2016	Genentech's linbrary (human naïve)
Bavencio®	Avelumab	2017	Dyax's library (semi-synthetic Fab)
Tremfya®	Guselkumab	2017	HuCAL GOLD® (Synthetic Fab library)
Cablivi®	Caplacizumab	2018	Camelidae-nanobody library
Gamifant®	Emapalumab	2018	CAT's library (human naïve scFv)
Lumoxiti®	Moxetumomab	2018	in vitro affinity maturation
Takhzyro®	Lanadelumab	2018	Dyax's library (semi-synthetic Fab)

1.2.7. Developability concerns of phage display-derived antibodies

Specificity and high affinity are not the only attributes that determine the success of therapeutic antibodies. Biophysical attributes, such as solubility, viscosity, expression yield, and thermal and long-term stability are vital to ensure the success of mAb lead candidates in biomanufacturing and clinical trials.[62] For instance, low solubility leads to poor activity, bioavailability, and high immunogenicity [94–96], while high-viscosity has been reported to be caused by mAb self-association [97–101]. Both viscosity and solubility have been shown to impact several downstream processing steps.[102] Expectedly, thermal stability is crucial to maintain structural and functional integrity under different temperatures.[103,104] Moreover, thermally unstable antibodies have shown to express less[105] and thermally stable antibodies have been shown to have lower tendency to aggregate[103,104,106–108]**.** In fact, aggregation is one of the main challenges that’s limit the advancement of therapeutic mAb due to immunogenicity concerns.[93,109,110] The combination of different amino acid sequences directly influences antibodies properties and will lead to different biophysical characteristics. For example, both the absence of somatic mutations to V_H and V_L germlines and the accumulation of positive amino acids (lysine (K), histidine (H), arginine (R)) across CDRs are good predictors of poly-reactivity.[19,111] Individually, the presence of K in the HCDR3 is associated with higher tendency to self-association, the presence of H in HCDR3 with lower expression in HEK293 cell line, and the presence of R in LCDR3 with lower thermodynamic stability.[19] Inversely, the presence of negative amino acids (glutamate (E), aspartate (D)) in HCDR1 and LCDR2 were shown to have a positive effect on thermodynamic stability, but also a reduction in expression titer if D is present in HCDR2. Negative amino acids in HCDR2 are also associated with lower self-association.[19,112] Likewise, aromatic residues (phenylalanine (F), tyrosine (Y), tryptophane (W)) were also shown to drive aggregation and self-association[19,113,114], even though they are key constituents of paratopes.[115–117] All in all, the combination of different residues across V_H and V_L can lead to a plethora of favorable and unfavorable properties. The final antibody’s characteristics will result from the interplay of opposing forces, which may be hard to predict. As such, a common practice in industrial pipelines is to implement extensive developability assessments to determine the biochemical and biophysical features of antibody candidates, as to identify the most favorable ones and avoid difficulties downstream.[118,119] In an effort to maximize the number of positive outcomes from the affinity, specificity, and developability point of view, designing “fail-proof” antibody libraries has been a focus for researchers since their inception. In the context of this thesis, where new antibody libraries will develop, these considerations are highly relevant and played a major role in determining the experimental steps.

1.3. Antibody Libraries

The first phage-display library was developed by Greg Winter and colleagues after amplification of scFv genes from hybridomas and B cells and cloning into phage-display expression vectors. Since then, several types of libraries have been developed throughout the years: immune, naïve, semi-synthetic and synthetic libraries. These libraries vary in the origin of the antibody's variable fragments as well as in their combinatorial variability, sizes, compatible display platforms and practical applications.

1.3.1 Core Principles, Library Size, and Combinatorial Diversity

As pointed out by Bradbury and Marks[61], for a library to successfully lead to the isolation of suitable antibody candidates, four requirements must be fulfilled: Firstly, genotypic diversity must be ensured, which will be a direct consequence of the recombinant antibody repertoire and its construction method. Secondly, genotype-phenotype coupling must be present, which means that there is a physical linkage between the antibody displayed and the DNA encoding for its expression (as is the case with phage particles and yeast cells). Finally, the exposure of recombinant antibody libraries to an antigen followed by repetitive enrichment cycles and/or screening constitutes the final two requirements: selective pressure and amplification.[61]

The probability of finding an high-affinity binder is directly correlated to the library size. The larger the antibody library, the more diverse mAbs that specifically bind random epitopes, and hence, the higher the possibility of selecting the desired molecule. This somewhat intuitive concept was formalized by Perelson in 1979[31], by $P = e^{-Np}$, where P is the probability that a given antibody does not bind randomly, N is the size of the antibody library, and p is the probability that said antibody contacts said epitope with certain affinity. If a $p = 5$ nM is to be achieved with low P values (below 0.01), a library size N of 10^9 individual clones needs to be achieved.

The successful results of larger libraries came to confirm the interrelation between library size and antibody affinity.[120,121]

Unfortunately, the size of phage-display libraries cannot be increased indefinitely. Theoretically, if each position of each of the six CDRs were diversified with the 20 natural amino acids, the corresponding theoretical repertoire would contain ~10^{78} unique antibody variants. Realistically, the bacterial transformation step during library construction limits the size of the library from exceeding 10^{11} antibody variants, as the culture volumes needed to reach higher levels of diversity are impractical.[62,122] Nonetheless, rather than solely judging a library's potential by its diversification possibilities, it should be judged from a functional size standpoint. That is, true library diversity is judged by how many individual functional antibody fragments (and hence, paratopes) are able to identify as many different antigens as possible[62,122,123] This concept was uncovered early on with the advent of semi-synthetic libraries in the 90s, upon the realization that some specific V_H germlines were overly represented in the pool of antigen-specific clones due to their advantageous folding capabilities.[124] This observation implied that only a fraction of the total library size was functional, and that other parameters apart from the theoretical combinatorial diversity could impact the performance of antibody libraries. On that assumption, the first fully synthetic library (HuCal, Morphosys[125]) took into account that there was no significant meaning to the inclusion of poorly folded V_H families, and paved the way for subsequent library designs.

Of the total size of current state-of-the-art libraries, about 85% of it is functional, on average.[122] This 15% margin for improvement stems from intrinsic bottlenecks related with many steps of library generation. Firstly, imprecisions in gene synthesis – either using RT-PCR (Reverse transcription polymerase chain reaction) for naïve libraries or using chemical means for synthetic and semi-synthetic libraries[125] – will influence the functional size of the libraries. Such imprecision can lead to nucleotide sequences with stop codons and frameshifts, which generate truncated fragments devoid of functionality. These are also prone to happen during standard PCR amplification phases. Secondly, synthetic diversity based on degenerated NNK codons or oligonucleotide mixtures (TRIM) are still prone to errors, and can lead to a reduction of diversity at the amino acid level due to the redundancy of the genetic code, unwanted amino acids at targeted positions for diversification and/or unintended length variations in CDR.[126] More recent strategies, such as Slonomics®[127] and Twist Bioscience's silicon-based DNA synthesis platform have been developed to increase library functionality.

Finally, diversity in of itself does not constitute a definitive parameter. As explained previously, the structural positioning of HCDR3 enables optimal antigen detection, which together with its superior diversity, makes it the most important CDR loop for antigen detection. In that line of thought, it makes sense that diversification should be made to regions where the likelihood of contacting with antigen is bigger. Therefore, an hypothetical 10^{10} library randomized exclusively on the HCDR3 is very much likely to succeed more than an hypothetical 10^{10} library randomized exclusively on the LCDR2, for example. Considering the big limitations on library size imposed by the bacterial transformation steps, the push for diversification designs that maximize binding to epitopes within the limited working diversity of 10^{11} is of the utmost importance. Lessons from Janssen Bio's pIX V3.0 library[128] corroborate this vision. Germline scaffolds for pIX V3.0 were firstly chosen based on their high usage in the antibody human repertoire, and on structural considerations. Also, canonical structures with higher propensity of binding to specific types of proteins and/or peptides were prioritized. Then CDRs were diversified in positions frequently found in contact with protein and peptide targets.[129,130] A recent structural study on pIX3.0 determined the structure of the V_H:V_L combinations used in pIX V3.0 and saw that the structural variability between different V_H:V_L pairings provided the libraries with distinct topographies and structural diversity to recognize diverse targets.[131] Such constitutes a striking example on how to leverage structure to maximize outcomes with a limited diversity.

1.3.2. Immune and Naïve libraries

Immune libraries are antibody repertoires that have been generated from an immunized donor, usually a mouse or a human, but also from other species. This allows the isolation of antibodies that have been generated after the host's polyclonal response against a given immunogen. Some species are immunized with the purpose of retrieving such antibodies. Humans, on the other hand, even though they are not purposefully used as immunogen recipients, can also constitute a source of immune antibody repertoires if they are afflicted by a disease of interest. The antibody repertoires of human patients can be profiled to discover antibodies against a broad range of diseases, such as HIV, auto-immune diseases, or cancer.[132]

To generate these libraries, B-cells are isolated either from the bone morrow, spleen and lymph nodes, or peripheral blood. Their DNA is extracted, and separate RT-PCR reactions are used to recover intact V_H and V_L sequences, that arise from *in vivo* V(D)J recombination. Afterwards, the cDNA products are inserted into appropriate display cassettes, usually in scFv or Fab format, but also as sdAb cassettes. The combinatorial pairing of V_H and V_L in each expression cassette usually creates immune libraries of $\leq 10^8$ individual clones.[132] In spite of the limited sizes of these antibody repertoires, due to the natural *in vivo* affinity maturation processes that were activated following the immunization and boosting with a desired antigen, anti-immunogen antibodies with affinities within the nM range can be obtained.[61]. Yet, the high enrichment level of immunogen-specific clones in immune libraries does not eliminate the possibility of isolating antibodies directed against unexpected antigens or self-antigens. Because while enriched for immunogen-specific antibodies, immune libraries still contain the remaining antibodies from the full antibody repertoire of the donor.[132] Moreover, the combinatorial assembly of antibody heavy and light chains may generate V_H:V_L pairings that were not part of the donor's original repertoire. Most importantly, due to their intrinsic nature, immune libraries can only be used for very specific cases and cannot be deployed in a general-use manner in antibody-discovery pipelines.

This is not the case of Naïve libraries, whose construction follows a similar process. Here, the donors are non-immunized (or ones that were not immunized intentionally for the purpose of constructing a library) and are thus not biased towards any specific type of immunogen. However, since the V_H and V_L were not subjected to antigen-driven *in vivo* affinity maturation processes, the library size becomes the most important characteristic for the determination of naïve library's quality.[133] To be truly naïve, the population of B-cells contributing to the library construction should be restricted to cells carrying binding potentials only in IgM/IgD formats. This was the strategy of a recent scFv library called HAL9/10, which used a reverse primer for V regions derived from IgMs, which favored the amplification of genes close to the germline configuration.[134] However, libraries formed by RT-PCR of all possible isotypes were described as equally good: Xoma's state-of-the-art naïve libraries XFab1 and XscFv2 where amplified from all Ig classes and yielded similar performances.[122,135] Interestingly, the Fab and scFv formats returned a similar number of antibodies with similar affinity indicating that the display format did not significantly impact the outcome of the selections. Other notable examples of high-quality naïve libraries are Dyax's libraries [ref] and CAT2.0 (MedImmune)[136], which have yielded two FDA/EMA-approved antibodies each (Table 1), and the more recent KNU-Fab.[137] The major hazard of all naïve libraries are the biases and redundancy of the donor's antibody repertoire, due to its immunological history, polymorphisms or ethnicity. For that reason, all the mentioned naïve libraries used extensive donor pools from a wide variety of ethnic origins. Notably, only a fraction of the existent antibody genes was represented in the pool of sequences, which is consistent with the biased patterns towards preferential frameworks observed in human natural repertoires. The HCDR3 length distribution was also similar across libraries and mirrored the Gaussian distribution typical of the human antibodies.[25,122]

1.3.3. Semi-synthetic and fully synthetic libraries

Naïve libraries do not make assumptions about the diversity of the antibody repertoire. The rationale is that the human antibody repertoire evolved to recognize any target with a reasonable specificity and affinity. Synthetic repertoires on the other hand, try to avoid biases and redundancies of *in vivo* antibodies, by using well expressed and developable scaffolds, targeting specific positions for diversification, and selecting types and frequency of amino acids that facilitate selection of diverse binders to any given target. Additionally, synthetic libraries can also be used to target antigens that are non-immunogenic, toxic, or self-antigens. Two types of synthetic antibody libraries were developed over the years: semi-synthetic libraries that combine natural CDRs with artificial ones, and fully synthetic libraries in which all CDRs are man-made. Due to its lower complexity, semi-synthetic libraries were the first to be generated.[138,139] Even tough the first generation did not lead to any FDA/EMA-approved antibody, they uncovered key-concepts, such as the interrelation between size and affinity, and the concept of functional size [ref]]. Very high-quality semi-synthetic libraries were built from the lessons gained from first generation libraries. Notable examples are the Bioinvent's and Dyax's libraries. The former has many antibodies at different phases in clinical trials and were the first library where 1 framework was combined with natural CDR repertoire. The latter was the first Fab single framework library where HCDR1/2 where synthetically designed to mimic natural diversity, combined with HCDR3 of donors and lead to 2 FDA/EMA-approved antibodies. [ref] As further insights were gained as new results from these libraries came, fully synthetic libraries started being introduced. The first fully synthetic library – HuCAL, was built with 7 V_H and 7 V_L (4 $V_L\kappa$ and 3 $V_L\lambda$) master scaffolds, which yielded 49 antibody sub-libraries when combined. These scaffolds were designed with consensus sequences representing the human Ig genes and cloned into cassettes for easier cloning. Cassettes were optimized for high expression in *E.coli* and only HCDR3 was randomized.[125] HuCAL was subsequently iterated to HuCAL GOLD®[140] and HuCAL PLATINUM®.[141] GOLD® has variability in all six CDRs and introduced cys-display of antibodies to pIII proteins. On top of those modifications, PLATINUM® removed undesirable NXT/S motifs (prone to deamidation) and introduced new HCDR3 sub-designs according to their length. That is, instead of an uniform amino acid distribution regardless of HCDR3 length, the amino acid usage per position was based on a systematic analysis of the sequences across different loop lengths. This innovative HCDR3 design did not lead to an increase in positive clones, but lead to a greater paratope diversity and broader V_H family representation.[141] Another notable example of a state-of-the-art synthetic library is Ylanthia (Morphosys).[85] Instead of arising from the indiscriminate combination of V_H and V_L scaffolds, Ylanthia comprises 36 fixed V_H/V_L framework pairings, from 12 V_H, 12 V_LK and 8 $V_L\lambda$ scaffolds. These fixed pairs were systematically selected from a larger pool of 20 V_H, 12 V_LK and 8 $V_L\lambda$ Ig gene segments which yielded 400 possible combinations. From these combinations, sub-optimal V_H/V_L pairs were filtered out after experimental measurements on expression level, thermal and serum stability, and aggregation propensity. Diversity was only directed towards LCDR3 and HCDR3 and a size of 1.3 x 10^{11} was achieved.[85]

1.3.4. In vitro affinity maturation

Regardless of the origin and platform used, antibodies generated against a given target may have room for improvement. A classical approach for that involves creating a crystal structure of the antigen-antibody complex as to detail which residues contribute to binding and which can be changed. The generation of crystal structures of single proteins is difficult in of itself, which makes antigen-antibody complexes even more difficult to resolve. Additionally, for the successful crystallization of complexes, it requires an extremely stable and somewhat rigid interaction, with very defined binding sites, something usually related with antibodies that already went through affinity maturation steps. Even when having the crystal structure resolved, it is not guaranteed that it provides enough information for a sensible conclusion to be reached about which modifications lead to better characteristics at the end. The lack of throughputness combined with the overall drawbacks of X-ray crystallography means that other methods should be employed.

After the primary immune response *in vivo*, cycles of randomization and secondary selection ensue after the to generate new antibodies with improved affinity towards their cognate antigen. Much like what happens during this secondary immune response *in vivo*, synthetic affinity maturation protocols can be deployed *in vitro* to improve the affinity and developability of any given mAb. This means generating affinity maturation libraries, that retain regions of the parental antibody while randomizing others. These work as traditional antibody libraries but use a single parental antibody as their starting point. In most cases, the HCDR3 and antibody framework are maintained, while the other CDRs are diversified. Diversification methods can be achieved by random mutagenesis (e.g. error-prone PCR), site-specific mutagenesis, or by tailor-made gene synthesis of diversified regions (e.g. TRIM/Slonomics®[127] and Twist Bioscience's technology). The latter is the preferred method to generate diversified antibodies due to its precision. Concrete examples of successful *in vitro* affinity maturation strategies are thoroughly reviewed by Lim *et al.*[142] and will not be fully detailed in this thesis. However, some important considerations must be taken into account.

The first hurdle in *in vitro* affinity maturation is the same as with primary antibody libraries, where the limitations of bacterial transformation limit the size of the library from exceeding 10^{11} antibody variants. If one would sample every possible mutation at all CDR positions (~60) one would get 10^{78} variants. This limits the number of positions and mutations that can be employed in each affinity maturation project and is especially relevant when designing generalized (or blind) *in vitro* maturation approaches. Typical blind approaches involve the diversification of each CDR individually in separate cassettes. These can be used to select several beneficial mutations within each CDR, in parallel. Such approach is suitable for the implementation of generalized affinity maturation protocols. Another generalized *in vitro* maturation approach consists in a pot of sequences with single-point mutations across many CDRs, called Look-through mutagenesis (LTM).[143] These single-point mutated sequences are challenged against a target as a way to identify beneficial mutations. In both cases, mutations selected in the first rounds of selection can be combined to search for synergistic effects. Rather than simply testing combinations of mutations one by one in each antibody molecule, mutations selected in the first phase can be re-sampled into new combinatorial libraries followed by another round of selection. This automatically selects the mutations that have synergistic effects, and removes the ones that don´t. The original LTM paper opted for this strategy, with great success.[143] However, there is no telling whether synergistic mutations were selected out during the first steps of selection, when cassettes were separated from each other and when single-point mutations were being sampled. Such may indicate why sometimes there are no noticeable gains in affinity after such strategies,[144] a phenomena that we also saw in *in-house* results (not shown). This leads us to the second hurdle. While generalized affinity maturation methods allow for a fast and reproducible protocol that can be readily deployed in any project, it may not respect the structural constraints of the IgG molecule and lead to inconsistent results. Ideally, innovative affinity maturation methods should be generalizable to provide high-throughput results while maintaining a certain degree of specificity towards the antibody structure being considered. As such they require attention to be paid to specific regions, such as the ones likely to be in contact with the antigen or regions that influence the antibodies' structural integrity and overall developability.

1.4. Aims and Goals

Synthetic antibody libraries with diversity focused on the HCDR3 are routinely used at our lab, while the other CDRs remain from germline origin. This enables the discovery of HCDR3 sequences that bind to a given antigen, after stringent rounds of panning are employed. However, since the remaining of the CDRs are kept unchanged, they do not suffer any selective pressure. Any successful binding will be guided mostly by HCDR3 and weak interactions from the germline and resulting antigen-antibody interface will be composed of beneficial and detrimental interactions (Figure 6a)

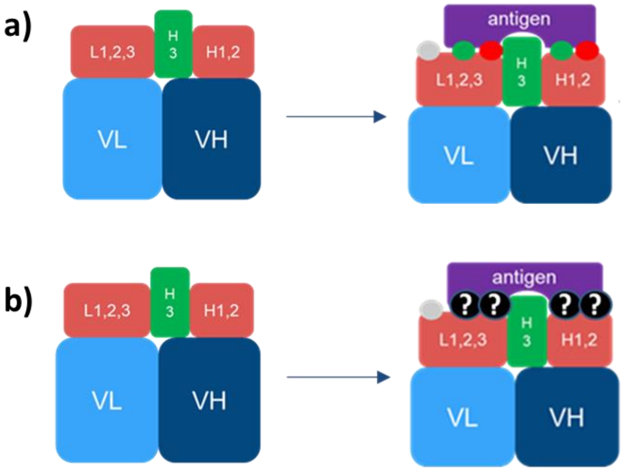


Figure *6* – Scheme of *hypothetical* antibody-antigen interfaces. Beneficial interactions in green. Structural clashes in red.

Primary binders coming from our libraries will usually sit on the 10-100 nM scale regarding binding to their target. Depending on the application, candidates will have to go through a process of affinity maturation to improve their binding and reach the pM scale. This has been achieved with great success at our lab, nonetheless sometimes the gain in affinity is sub-optimal, or leads to increases in antibody cross-reactivity. This might be explained by the fact that we currently have limited information on germline CDR interactions with the antigen (Figure 6b), which hardens the choice between different affinity maturation strategies and mutation selection during analysis. This is a direct consequence of the main drawback of generalized *in vitro* affinity maturation methods. Since we have almost no information about the antigen-antibody interface, we cannot make the best out of the affinity maturation step.

We wondered whether an innovative approach to generalized affinity maturation methods could improve the maturation of antibody candidates. A “semi-blind” affinity maturation process can be proposed as a compromise between generalization and specificity. Residues with a higher likelihood to bind to the antigen – due to their nature or advantageous positions – can be “shaved” into smaller versions to prevent interactions with antigen molecules during the primary selections, an approach hereinafter referred to as “CDR-null” concept. Residues closer to the HCDR3 may engage in (de)stabilizing interactions with the HCDR3, which will also influence antibody-antigen interactions, and will also be considered. This means that the HCDR3 and the remaining germline residues would guide the antibody-antigen interaction on a first phase, with minimal intervention of the CDR-null residues (Figure 7). The CDR-null residues can then be used as diversification “hotspots” that have a higher probability of influencing the antibody-antigen interaction.

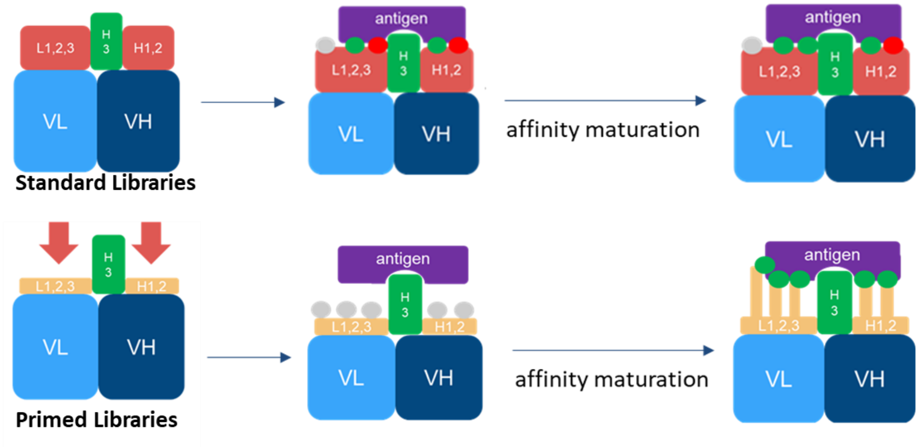


Figure *7* **–** ***Primed*** **libraries and "semi-blind" affinity maturation.** Beneficial interactions in green. Structural clashes in red.

This is expected to improve affinity maturation outcomes, even if at the expense of lower affinity in the early primary discovery stages (Figure 8)

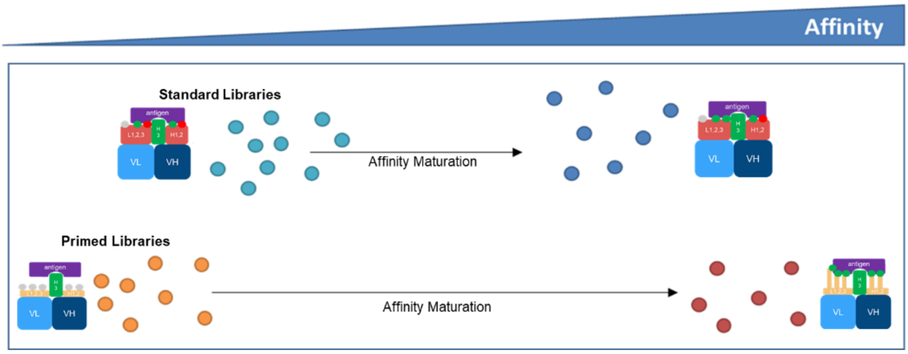


Figure 8 – Expectation of affinity maturation outcomes. Fewer residues interacting with the antigen molecule may compromise the affinity of primary binders, which is then compensated by a bigger gain in the affinity maturation phase.

The aim of this thesis is to build structurally primed antibody libraries, that carry CDR-null mutations on identified hotspots. These positions can then be targeted during a "semi-blind" affinity maturation protocol to maximize the likelihood of finding beneficial mutations that improve affinity. Such endeavor will require several steps:

1 - Identify amino acid residues that have higher likelihood of binding to antigen surfaces. Bioinformatics approaches combined with X-ray crystal structures data from previous crystallography efforts at the host laboratory will be used to identify big residues pointing outwards that tend to bind to the antigen molecule.

2 - Generate primed frameworks that carry mutations on the identified hotspots. Such mutations involve replacing the current amino acid with smaller amino acids that have less likelihood of binding to the antigen, such as alanine or serine. The resulting frameworks will be coined “CDR-null” frameworks, “Null-Frameworks” or “Primed Frameworks”.

3 - Inspect the developability effect of CDR-null mutations on IgGs. Combinations of mutations will be tested to yield frameworks with acceptable developability properties.

4 - Generate and characterize Primed antibody libraries. Combine sampled mutations into a single framework and generate HCDR3-randomized libraries from it.

5 - Discovery of primary binders using standard and CDR-null primary libraries. Perform a primary panning against a target of interest, perform NGS, select, produce, and characterize candidates for their affinity and developability.

6 - Design of a semi-blind affinity maturation method. Diversification of CDR-null positions using a rational amino acid distribution.

7 - Generation of affinity maturation libraries using standard cassette protocols and semi-blind methods. Use two different diversification methods to diversify specific regions of the primary binders identified in 5)

8 - Comparison of affinity maturation panning outcomes.

References of Chapter 1

1. Chaplin, D. D. Overview of the immune response. *J. Allergy Clin. Immunol.* **125**, S3–S23 (2010).

2. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. Receptors of the innate immune system. *Immunobiol. Immune Syst. Health Dis. 5th Ed.* (2001).

3. Bonilla, F. A. & Oettgen, H. C. Adaptive immunity. *J. Allergy Clin. Immunol.* **125**, S33–S40 (2010).

4. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).

5. Lee, S. K., Bridges, S. L., Koopman, W. J. & Schroeder, H. W. The immunoglobulin kappa light chain repertoire expressed in the synovium of a patient with rheumatoid arthritis. *Arthritis Rheum.* **35**, 905–913 (1992).

6. Corbett, S. J., Tomlinson, I. M., Sonnhammer, E. L., Buck, D. & Winter, G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, 'minor' D segments or D-D recombination. *J. Mol. Biol.* **270**, 587–597 (1997).

7. Matsuda, F. *et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J. Exp. Med.* **188**, 2151–2162 (1998).

8. Marks, C. & Deane, C. M. How repertoire data are changing antibody science. *J. Biol. Chem.* **295**, 9823–9837 (2020).

9. Kuroda, D., Shirai, H., Jacobson, M. P. & Nakamura, H. Computer-aided antibody design. *Protein Eng. Des. Sel.* **25**, 507–522 (2012).

10. Almagro, J. C. *et al.* Antibody modeling assessment: Antibody Modeling. *Proteins Struct. Funct. Bioinforma.* **79**, 3050–3066 (2011).

11. Almagro, J. C. *et al.* Second antibody modeling assessment (AMA-II): 3D Antibody Modeling. *Proteins Struct. Funct. Bioinforma.* **82**, 1553–1562 (2014).

12. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917 (1987).

13. North, B., Lehmann, A. & Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. *J. Mol. Biol.* **406**, 228–256 (2011).

14. Sela-Culang, I., Alon, S. & Ofran, Y. A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J. Immunol. Baltim. Md 1950* **189**, 4890–4899 (2012).

15. Rees, A. R. Understanding the human antibody repertoire. *mAbs* **12**, 1729683 (2020).

16. Kunik, V. & Ofran, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng. Des. Sel.* **26**, 599–609 (2013).

17. Burkovitz, A. & Ofran, Y. Understanding differences between synthetic and natural antibodies can help improve antibody engineering. *mAbs* **8**, 278–287 (2016).

18. Mahon, C. M. *et al.* Comprehensive Interrogation of a Minimalist Synthetic CDR-H3 Library and Its Ability to Generate Antibodies with Therapeutic Potential. *J. Mol. Biol.* **425**, 1712–1730 (2013).

19. Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Mol. Immunol.* **112**, 338–346 (2019).

20. Alberts, B. *et al.* The Generation of Antibody Diversity. *Mol. Biol. Cell 4th Ed.* (2002).

21. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).

22. Schroeder, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* **30**, 119–135 (2006).

23. Boyd, S. D. & Joshi, S. A. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. *Microbiol. Spectr.* **2**, (2014).

24. Morbach, H., Eichhorn, E. M., Liese, J. G. & Girschick, H. J. Reference values for B cell subpopulations from infancy to adulthood. *Clin. Exp. Immunol.* **162**, 271–279 (2010).

25. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci.* **106**, 20216–20221 (2009).

26. Hong, B. *et al.* In-Depth Analysis of Human Neonatal and Adult IgM Antibody Repertoires. *Front. Immunol.* **9**, 128 (2018).

27. Soto, C. *et al.* High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).

28. Kovaltsuk, A. *et al.* How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. *Front. Immunol.* **8**, 1753 (2017).

29. Krawczyk, K. *et al.* Structurally Mapping Antibody Repertoires. *Front. Immunol.* **9**, 1698 (2018).

30. Perelson, A. S. & Weisbuch, G. Immunology for physicists. *Rev. Mod. Phys.* **69**, 1219–1268 (1997).

31. Perelson, A. S. & Oster, G. F. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.* **81**, 645–670 (1979).

32. Notkins, A. L. Polyreactivity of antibody molecules. *Trends Immunol.* **25**, 174–179 (2004).

33. Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E. & Meiler, J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. *PLoS Comput. Biol.* **9**, e1003045 (2013).

34. Crouzier, R., Martin, T. & Pasquali, J. L. Heavy chain variable region, light chain variable region, and heavy chain CDR3 influences on the mono- and polyreactivity and on the affinity of human monoclonal rheumatoid factors. *J. Immunol. Baltim. Md 1950* **154**, 4526–4535 (1995).

35. Harindranath, N., Ikematsu, H., Notkins, A. L. & Casali, P. Structure of the VH and VL segments of polyreactive and monoreactive human natural antibodies to HIV-1 and Escherichia coli beta-galactosidase. *Int. Immunol.* **5**, 1523–1533 (1993).

36. Yuseff, M.-I., Pierobon, P., Reversat, A. & Lennon-Duménil, A.-M. How B cells capture, process and present antigens: a crucial role for cell polarity. *Nat. Rev. Immunol.* **13**, 475–486 (2013).

37. Lipman, N. S., Jackson, L. R., Trudel, L. J. & Weis-Garcia, F. Monoclonal Versus Polyclonal Antibodies: Distinguishing Characteristics, Applications, and Information Resources. *ILAR J.* **46**, 258–268 (2005).

38. Lu, R.-M. *et al.* Development of therapeutic antibodies for the treatment of diseases. *J. Biomed. Sci.* **27**, 1 (2020).

39. Kaplon, H. & Reichert, J. M. Antibodies to watch in 2021. *mAbs* **13**, 1860476 (2021).

40. Mullard, A. FDA approves 100th monoclonal antibody product. *Nat. Rev. Drug Discov.* (2021) doi:10.1038/d41573-021-00079-7.

41. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256**, 495–497 (1975).

42. Leavy, O. Therapeutic antibodies: past, present and future. *Nat. Rev. Immunol.* **10**, 297–297 (2010).

43. Kuus-Reichel, K. *et al.* Will immunogenicity limit the use, efficacy, and future development of therapeutic monoclonal antibodies? *Clin. Diagn. Lab. Immunol.* **1**, 365–372 (1994).

44. Morrison, S. L., Johnson, M. J., Herzenberg, L. A. & Oi, V. T. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 6851–6855 (1984).

45. Lonberg, N. Human antibodies from transgenic animals. *Nat. Biotechnol.* **23**, 1117–1125 (2005).

46. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* **321**, 522–525 (1986).

47. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).

48. Parmley, S. F. & Smith, G. P. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. *Gene* **73**, 305–318 (1988).

49. Bird, R. E. *et al.* Single-chain antigen-binding proteins. *Science* **242**, 423–426 (1988).

50. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554 (1990).

51. Barbas, C. F., Kang, A. S., Lerner, R. A. & Benkovic, S. J. Assembly of combinatorial antibody libraries on phage surfaces: the gene III site. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 7978–7982 (1991).

52. Breitling, F., Dübel, S., Seehaus, T., Klewinghaus, I. & Little, M. A surface expression vector for antibody screening. *Gene* **104**, 147–153 (1991).

53. Carmen, S. & Jermutus, L. Concepts in antibody phage display. *Brief. Funct. Genomics* **1**, 189–203 (2002).

54. An overview on application of phage display technique in immunological studies. *Asian Pac. J. Trop. Biomed.* **7**, 599–602 (2017).

55. Kehoe, J. W. & Kay, B. K. Filamentous Phage Display in the New Millennium. *Chem. Rev.* **105**, 4056–4072 (2005).

56. Stricker, N. & Li, M. Phage Display Technologies. in *eLS* (American Cancer Society, 2001). doi:10.1038/npg.els.0000982.

57. Rakonjac, J., Jovanovic, G. & Model, P. Filamentous phage infection-mediated gene expression: construction and propagation of the gIII deletion mutant helper phage R408d3. *Gene* **198**, 99–103 (1997).

58. Ledsgaard, L., Kilstrup, M., Karatt-Vellatt, A., McCafferty, J. & Laustsen, A. Basics of Antibody Phage Display Technology. *Toxins* **10**, 236 (2018).

59. Rondot, S., Koch, J., Breitling, F. & Dübel, S. A helper phage to improve single-chain antibody presentation in phage display. *Nat. Biotechnol.* **19**, 75–78 (2001).

60. Ebersbach, H. & Geisse, S. Antigen generation and display in therapeutic antibody drug discovery -- a neglected but critical player. *Biotechnol. J.* **7**, 1433–1443 (2012).

61. Bradbury, A. R. M. & Marks, J. D. Antibodies from phage antibody libraries. *J. Immunol. Methods* **290**, 29–49 (2004).

62. Alfaleh, M. A. *et al.* Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. *Front. Immunol.* **11**, 1986 (2020).

63. Schirrmann, T., Meyer, T., Schütte, M., Frenzel, A. & Hust, M. Phage Display for the Generation of Antibodies for Proteome Research, Diagnostics and Therapy. *Molecules* **16**, 412–426 (2011).

64. Noppe, W. *et al.* Chromato-panning: an efficient new mode of identifying suitable ligands from phage display libraries. *BMC Biotechnol.* **9**, 21 (2009).

65. Schönberger, N. *et al.* Chromatopanning for the identification of gallium binding peptides. *J. Chromatogr. A* **1600**, 158–166 (2019).

66. Shen, W. *et al.* Blocking Agent Optimization for Nonspecific Binding on Phage Based Magnetoelastic Biosensors. *J. Electrochem. Soc.* **159**, B818 (2012).

67. SuperBlock^{TM} (PBS) Blocking Buffer. https://www.thermofisher.com/order/catalog/product/37515.

68. Menendez, A. & Scott, J. K. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Anal. Biochem.* **336**, 145–157 (2005).

69. Laustsen, A. H., Lauridsen, L. P., Lomonte, B., Andersen, M. R. & Lohse, B. Pitfalls to avoid when using phage display for snake toxins. *Toxicon* **126**, 79–89 (2017).

70. Hawkins, R. E., Russell, S. J. & Winter, G. Selection of phage antibodies by binding affinity. Mimicking affinity maturation. *J. Mol. Biol.* **226**, 889–896 (1992).

71. Kay, B. K., Thai, S. & Volgina, V. V. High-throughput Biotinylation of Proteins. *Methods Mol. Biol. Clifton NJ* **498**, 185–196 (2009).

72. Fairhead, M. & Howarth, M. Site-specific biotinylation of purified proteins using BirA. *Methods Mol. Biol. Clifton NJ* **1266**, 171–184 (2015).

73. Predonzani, A., Arnoldi, F., López-Requena, A. & Burrone, O. R. In vivosite-specific biotinylation of proteins within the secretory pathway using a single vector system. *BMC Biotechnol.* **8**, 41 (2008).

74. Zahnd, C., Sarkar, C. A. & Plückthun, A. Computational analysis of off-rate selection experiments to optimize affinity maturation by directed evolution. *Protein Eng. Des. Sel. PEDS* **23**, 175–184 (2010).

75. Smith, G. P. & Scott, J. K. [15] Libraries of peptides and proteins displayed on filamentous phage. in *Methods in Enzymology* vol. 217 228–257 (Elsevier, 1993).

76. Kang, A. S., Barbas, C. F., Janda, K. D., Benkovic, S. J. & Lerner, R. A. Linkage of recognition and replication functions by assembling combinatorial antibody Fab libraries along phage surfaces. *Proc. Natl. Acad. Sci.* **88**, 4363–4366 (1991).

77. Haque, A. & Tonks, N. K. The use of phage display to generate conformation-sensor recombinant antibodies. *Nat. Protoc.* **7**, 2127–2143 (2012).

78. Ward, R. L., Clark, M. A., Lees, J. & Hawkins, N. J. Retrieval of human antibodies from phage-display libraries using enzymatic cleavage. *J. Immunol. Methods* **189**, 73–82 (1996).

79. Kristensen, P. & Winter, G. Proteolytic selection for protein folding using filamentous bacteriophages. *Fold. Des.* **3**, 321–328 (1998).

80. Selection of human antibody fragments by phage display | Nature Protocols. https://www.nature.com/articles/nprot.2007.448.

81. 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–e308 (2017).

82. Rouet, R., Jackson, K. J. L., Langley, D. B. & Christ, D. Next-Generation Sequencing of Antibody Display Repertoires. *Front. Immunol.* **9**, 118 (2018).

83. Noh, J. *et al.* High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library. *mAbs* **11**, 532–545 (2019).

84. Hodkinson, B. P. & Grice, E. A. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Adv. Wound Care* **4**, 50–58 (2015).

85. Tiller, T. *et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs* **5**, 445–470 (2013).

86. Vollmers, C., Penland, L., Kanbar, J. N. & Quake, S. R. Novel Exons and Splice Variants in the Human Antibody Heavy Chain Identified by Single Cell and Single Molecule Sequencing. *PLOS ONE* **10**, e0117050 (2015).

87. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Gener. Seq. Appl.* **1**, (2014).

88. Zhai, W. *et al.* Synthetic Antibodies Designed on Natural Sequence Landscapes. *J. Mol. Biol.* **412**, 55–71 (2011).

89. Ravn, U. *et al.* Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**, 99–110 (2013).

90. Kaplon, H. & Reichert, J. M. Antibodies to watch in 2019. *mAbs* **11**, 219–238 (2019).

91. Frenzel, A., Schirrmann, T. & Hust, M. Phage display-derived human antibodies in clinical development and therapy. *mAbs* **8**, 1177–1194 (2016).

92. Nagano, K. & Tsutsumi, Y. Phage Display Technology as a Powerful Platform for Antibody Drug Discovery. *Viruses* **13**, 178 (2021).

93. Spencer, S., Bethea, D., Raju, T. S., Giles-Komar, J. & Feng, Y. Solubility evaluation of murine hybridoma antibodies. *mAbs* **4**, 319–325 (2012).

94. Gibson, T. J. *et al.* Application of a High-Throughput Screening Procedure with PEG-Induced Precipitation to Compare Relative Protein Solubility During Formulation Development with IgG1 Monoclonal Antibodies. *J. Pharm. Sci.* **100**, 1009–1021 (2011).

95. Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M. & Popovic, B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Sci. Rep.* **7**, 8200 (2017).

96. Bye, J. W., Platts, L. & Falconer, R. J. Biopharmaceutical liquid formulation: a review of the science of protein stability and solubility in aqueous environments. *Biotechnol. Lett.* **36**, 869–875 (2014).

97. Dobson, C. L. *et al.* Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. *Sci. Rep.* **6**, 38644 (2016).

98. Esfandiary, R., Parupudi, A., Casas-Finet, J., Gadre, D. & Sathish, H. Mechanism of Reversible Self-Association of a Monoclonal Antibody: Role of Electrostatic and Hydrophobic Interactions. *J. Pharm. Sci.* **104**, 577–586 (2015).

99. Guo, Z. *et al.* Structure-Activity Relationship for Hydrophobic Salts as Viscosity-Lowering Excipients for Concentrated Solutions of Monoclonal Antibodies. *Pharm. Res.* **29**, 3102–3109 (2012).

100. Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci.* **111**, 18601–18606 (2014).

101. Yadav, S. *et al.* Establishing a Link Between Amino Acid Sequences and Self-Associating and Viscoelastic Behavior of Two Closely Related Monoclonal Antibodies. *Pharm. Res.* **28**, 1750–1764 (2011).

102. Xu, Y. *et al.* Structure, heterogeneity and developability assessment of therapeutic antibodies. *mAbs* **11**, 239–264 (2018).

103. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci.* **106**, 11937–11942 (2009).

104. Seeliger, D. *et al.* Boosting antibody developability through rational sequence optimization. *mAbs* **7**, 505–515 (2015).

105. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* **114**, 944–949 (2017).

106. Barthelemy, P. A. *et al.* Comprehensive Analysis of the Factors Contributing to the Stability and Solubility of Autonomous Human V $_H$ Domains. *J. Biol. Chem.* **283**, 3639–3654 (2008).

107. Brader, M. L. *et al.* Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies. *Mol. Pharm.* **12**, 1005–1017 (2015).

108. He, F., Hogan, S., Latypov, R. F., Narhi, L. O. & Razinkov, V. I. High throughput thermostability screening of monoclonal antibody formulations. *J. Pharm. Sci.* **99**, 1707–1720 (2010).

109. Lowe, D. *et al.* Aggregation, stability, and formulation of human antibody therapeutics. *Adv. Protein Chem. Struct. Biol.* **84**, 41–61 (2011).

110. Perchiacca, J. M. & Tessier, P. M. Engineering Aggregation-Resistant Antibodies. *Annu. Rev. Chem. Biomol. Eng.* **3**, 263–286 (2012).

111. Rabia, L. A., Zhang, Y., Ludwig, S. D., Julian, M. C. & Tessier, P. M. Net charge of antibody complementarity-determining regions is a key predictor of specificity. *Protein Eng. Des. Sel.* **31**, 409–418 (2018).

112. Dudgeon, K. *et al.* General strategy for the generation of human antibody variable domains with increased aggregation resistance. *Proc. Natl. Acad. Sci.* **109**, 10879–10884 (2012).

113. van der Kant, R. *et al.* Prediction and Reduction of the Aggregation of Monoclonal Antibodies. *J. Mol. Biol.* **429**, 1244–1261 (2017).

114. Wang, X., Das, T. K., Singh, S. K. & Kumar, S. Potential aggregation prone regions in biotherapeutics: A survey of commercial monoclonal antibodies. *mAbs* **1**, 254–267 (2009).

115. Mian, I. S., Bradwell, A. R. & Olson, A. J. Structure, function and properties of antibody binding sites. *J. Mol. Biol.* **217**, 133–151 (1991).

116. Sundberg, E. J. & Mariuzza, R. A. Molecular recognition in antibody-antigen complexes. in *Advances in Protein Chemistry* vol. 61 119–160 (Elsevier, 2002).

117. Local and global anatomy of antibody-protein antigen recognition - PubMed. https://pubmed.ncbi.nlm.nih.gov/29218757/.

118. Bauer, J. *et al.* Rational optimization of a monoclonal antibody improves the aggregation propensity and enhances the CMC properties along the entire pharmaceutical process chain. *mAbs* **12**, 1787121 (2020).

119. Raybould, M. I. J. *et al.* Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci.* **116**, 4025–4030 (2019).

120. Hoogenboom, H. R. Designing and optimizing library selection strategies for generating high-affinity antibodies. *Trends Biotechnol.* **15**, 62–70 (1997).

121. Azriel-Rosenfeld, R., Valensi, M. & Benhar, I. A human synthetic combinatorial library of arrayable single-chain antibodies based on shuffling in vivo formed CDRs into general framework regions. *J. Mol. Biol.* **335**, 177–192 (2004).

122. Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies* **8**, 44 (2019).

123. Schofield, D. J. *et al.* Application of phage display to high throughput antibody generation and characterization. *Genome Biol.* **8**, R254 (2007).

124. Griffiths, A. D. *et al.* Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J.* **13**, 3245–3260 (1994).

125. Knappik, A. *et al.* Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides 1 1Edited by I. A. Wilson. *J. Mol. Biol.* **296**, 57–86 (2000).

126. Shim, H. Synthetic approach to the generation of antibody diversity. *BMB Rep.* **48**, 489–494 (2015).

127. Van den Brulle, J. *et al.* A novel solid phase technology for high-throughput gene synthesis. *BioTechniques* **45**, 340–343 (2008).

128. Shi, L. *et al.* De Novo Selection of High-Affinity Antibodies from Synthetic Fab Libraries Displayed on Phage as pIX Fusion Proteins. *J. Mol. Biol.* **397**, 385–396 (2010).

129. Raghunathan, G., Smart, J., Williams, J. & Almagro, J. C. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J. Mol. Recognit.* **25**, 103–113 (2012).

130. Almagro, J. C. Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit.* **17**, 132–143 (2004).

131. Teplyakov, A. *et al.* Structural diversity in a human antibody germline library. *mAbs* **8**, 1045–1063 (2016).

132. Harel Inbar, N. & Benhar, I. Selection of antibodies from synthetic antibody libraries. *Arch. Biochem. Biophys.* **526**, 87–98 (2012).

133. Benhar, I. Design of synthetic antibody libraries. *Expert Opin. Biol. Ther.* **7**, 763–779 (2007).

134. Kügler, J. *et al.* Generation and analysis of the improved human HAL9/10 antibody phage display libraries. *BMC Biotechnol.* **15**, 10 (2015).

135. Schwimmer, L. J. *et al.* Discovery of diverse and functional antibodies from large human repertoire antibody libraries. *J. Immunol. Methods* **391**, 60–71 (2013).

136. Lloyd, C. *et al.* Modelling the human immune response: performance of a 1011 human antibody repertoire against a broad panel of therapeutically relevant antigens. *Protein Eng. Des. Sel.* **22**, 159–168 (2009).

137. Kim, S. *et al.* Generation, Diversity Determination, and Application to Antibody Selection of a Human Naïve Fab Library. *Mol. Cells* **40**, 655–666 (2017).

138. Marks, J. D. *et al.* By-passing immunization: Human antibodies from V-gene libraries displayed on phage. *J. Mol. Biol.* **222**, 581–597 (1991).

139. de Kruif, J., Boel, E. & Logtenberg, T. Selection and application of human single chain Fv antibody fragments from a semi-synthetic phage antibody display library with designed CDR3 regions. *J. Mol. Biol.* **248**, 97–105 (1995).

140. Rothe, C. *et al.* The Human Combinatorial Antibody Library HuCAL GOLD Combines Diversification of All Six CDRs According to the Natural Immune System with a Novel Display Method for Efficient Selection of High-Affinity Antibodies. *J. Mol. Biol.* **376**, 1182–1200 (2008).

141. Prassler, J. *et al.* HuCAL PLATINUM, a Synthetic Fab Library Optimized for Sequence Diversity and Superior Performance in Mammalian Expression Systems. *J. Mol. Biol.* **413**, 261–278 (2011).

142. Lim, C. C., Choong, Y. S. & Lim, T. S. Cognizance of Molecular Methods for the Generation of Mutagenic Phage Display Antibody Libraries for Affinity Maturation. *Int. J. Mol. Sci.* **20**, 1861 (2019).

143. Rajpal, A. *et al.* A general method for greatly improving the affinity of antibodies by using combinatorial libraries. *Proc. Natl. Acad. Sci.* **102**, 8466–8471 (2005).

144. Chan, D. T. Y. & Groves, M. A. T. Affinity maturation: highlights in the application of *in vitro* strategies for the directed evolution of antibodies. *Emerg. Top. Life Sci.* ETLS20200331 (2021) doi:10.1042/ETLS20200331.

Chapter 2 – Designing CDR-null antibody frameworks

2.1. Summary

In this work two distinct antibody frameworks (FW-κ and FW-λ) were subjected to mutations to their CDR germline sequences (LCDR1, LCDR2, LCDR3, HCDR1, HCDR2) with the objective of reducing the likelihood of antigen-antibody contacts on those regions. Polar and charged residues pointing outwards towards the solvent were replaced by serines and alanines, and the effect of such mutations on the antibodies' developability was evaluated in terms of aggregation, hydrophobicity, and thermal stability. Mutations that did not lead to an over de-stabilization of the framework were combined to generate several distinct "CDR-null" frameworks. After careful analysis of the biophysical parameters, two final CDR-null frameworks were selected (κN1 and κN2, bearing 8 and 11 mutations to CDR positions, respectively). These CDR-null frameworks served as the basis for generating primed libraries (work described in the following chapters of this thesis).

2.2. Introduction

Two different frameworks were used in this work, framework-κ (FW-κ) and framework-λ (FW-λ). These two distinct frameworks share the VH3-23 heavy-chain sequence, which is then paired with different light-chains. VH3–23 has been shown to be one of the most frequent sequences in the human repertoire, with good expression titers and biophysical characteristics.[1] In FW-κ, it is paired with Vκ1-39, which is also one of the most frequent sequences in the human repertoire. Additionally, it also means that FW-κ will have the same V_H/V_L pairing as the industry-standard antibody Herceptin® (trastuzumab). In FW-λ, VH3-23 will be paired with Vλ3-9, which has been shown to have favorable pairing with VH3-23, according to internal data (not shown). A closer look on the amino acid sequence of each CDR loop can be found on Tables 3-4. The framework regions between CDR loops are not shown. The HCDR3 loop is the same as the one found in the Herceptin® commercial antibody (trastuzumab) – WGGDGFYAMDY, for both frameworks.

Table 2 - Gene usage in frameworks.

Framework	VL master gene	VH master gene
FW-κ	Vk1-39	VH3-23
FW-λ	Vλ3-9	VH3-23

Table 3 - CDR-loop sequences for Vk1-39 and Vλ3-9.

LC	LCDR1	LCDR2	LCDR3
Vk1-39	RASQSISSYLN	AASSLQS	QQSYSTPLT
Vλ3-9	GGNNIGSKNVH	RDSNRPS	QVWDSSTVV

Table 4 - CDR-loop sequences for VH3-23.

HC	HCDR1	HCDR2	HCDR3
VH3-23	FTFSSYAMS	AISGSGGSTYYADSVKG	WGGDGFYAMDY

2.3. Results

2.3.1 Design of CDR-null mutations

The CDR-null design aims to replace some of the amino acid residues pointing outwards by smaller ones that are less likely to interact with the antigen epitopes and/or constrain further HCDR3 loop conformations. As such, when designing CDR-null frameworks, we took a look into the crystal structures to understand if there were residues pointing towards the solvent and/or towards the HCDR3. Crystal structures for the two frameworks, FW-κ and FW-λ, in Fab format were resolved by our group (unpublished data) (Figure 4 and 5).

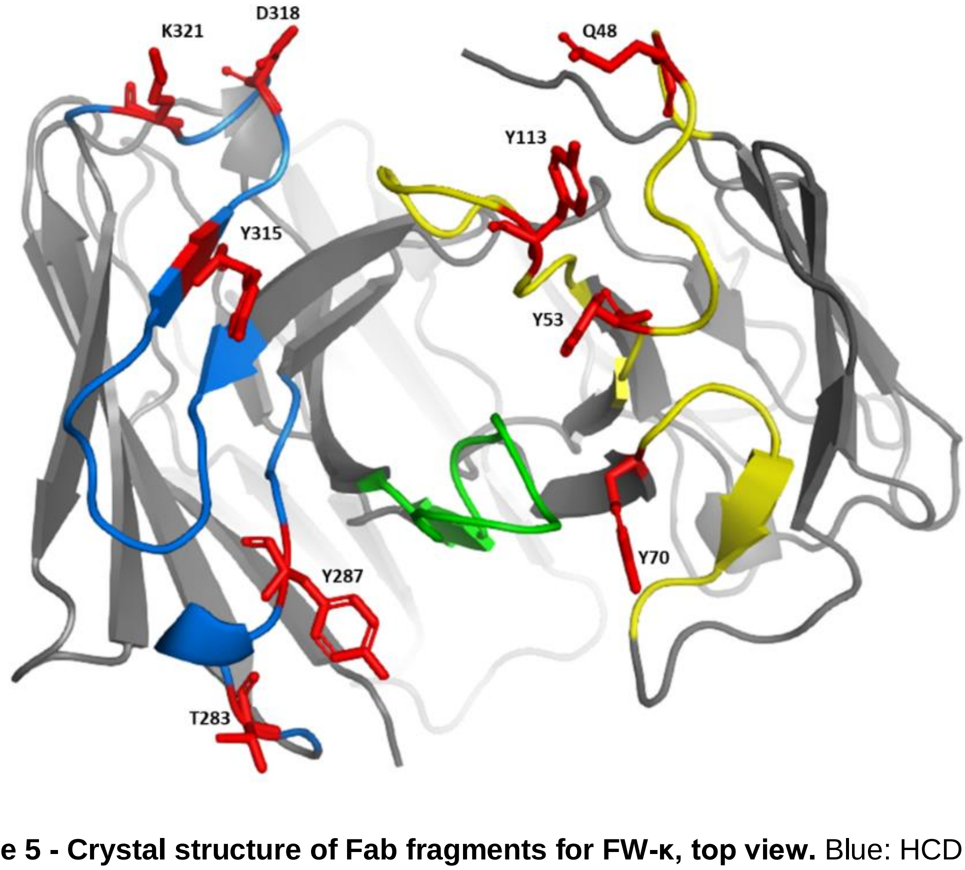


Figure 5 - Crystal structure of Fab fragments for FW-κ, top view. Blue: HCDR1 and HCDR2 residues. Yellow: LCDR1, LCDR2 and LCDR3 residues. Green: HCDR3 residues. Red: CDR-null positions. Grey: Framework-region residues. Resolution = 1.73 **Å.**

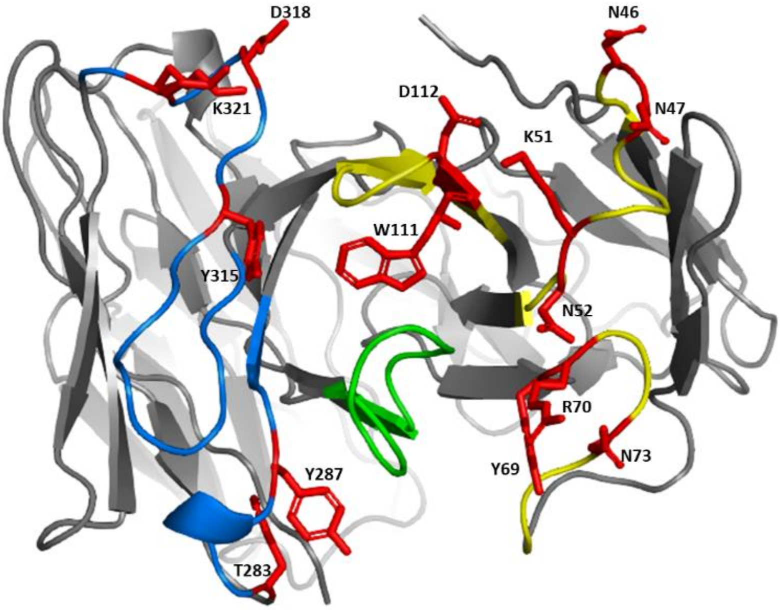


Figure 6 - Crystal structure of Fab fragments for FW-λ, top view. Blue: HCDR1 and HCDR2 residues. Yellow: LCDR1, LCDR2 and LCDR3 residues. Green: HCDR3 residues. Red: CDR-null positions. Grey: Framework-region residues. Resolution = 1.36 **Å.**

2.3.1.1. Choosing CDR-null mutations for VH3-23 (VH of FW-κ and FW-λ)

Mutations on the heavy chain are identical for both FW-κ and FW-λ. A summary of all chosen mutations for VH3-23 can be found in table 5.

Table 5 - Chosen mutations for VH3-23.

Name	Sequence	Mutation
HCDR1	FTFSSYAMS	-
	FAFSSAAMS	T283A, Y287A
	FAFSSDAMS	T283A, Y287D
	FTFSSTAMS	Y287T
HCDR2	AISGSGGSTYYADSVKG	-
	AISGSGGSTSYASSVSG	Y315S, D318S, K321S
	AISGS-GSTSYASSVSG	delG314, Y315S, D318S, K321S
	AISG--GSTSYASSVSG	delS313, delG314, Y315S, D318S, K321S
	AIS---GSTSYASSVSG	delG312, delS313, delG314, Y315S, D318S, K321S
HFR3	RFTISRDNSKNTY	-
	RFTISRDSSKATY	N330S, N333A

In HCDR1, T283 and Y287 are pointing outwards- (Figures 4-5). Threonine and tyrosine typically engage in polar contacts and are thus responsible for the formation of H-bonds with antigen epitopes, making T283 and Y287 ideal candidates for the CDR-null concept. While T283 can be readily substituted by an alanine, different mutations were tested for Y287 due its importance for HCDR3 structural integrity (Table 5). The mechanism by which Y287 exerts its influence is highlighted in Figure 6.

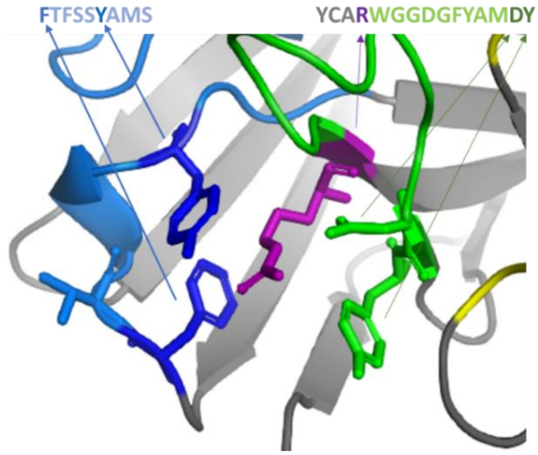


Figure 7 - The importance of F282, Y287, D364 and Y365 for HCDR3 structural integrity. Blue: HCDR1; Green: HCDR3; Grey: HFR2; Purple: R354*.*

Y287 interacts with the last two positions on HCDR3 that are critical for HCDR3 loop structure – D364 and Y365. Together with F282 and Y287 from HCDR1, they lock R354 in place, located immeadiately before the start of the HCDR3 sequence. Locking both extremities on the same place leads to a tie-shaped loop of amino acids pointing outwards, where it will meet the antigen epitopes. As such, Y287 was replaced with an A, D or T (Table 5). Developability characterization will be essential in determining which of these mutations to choose. This will be addressed later on in this chapter – see section 2.3.2. In HCDR2, residues pointing towards the surface such as Y315, D318 and K321 were substituted by serines (Table 5, Figure 4). Charged residues such as aspartate and lysine are usually involved on very relevant salt-bridge interactions with the antigen side, making D318 and K321 relevant picks for substitutions. Furthermore, HCDR2 is only composed of glicines and serines from residues 311 to 316, which ends up creating a bulky one-turn α-helix region that may limit antigen-antibody contacts (Figure 7). Some serines and glycines of this region were deleted with the intent of diminishing its size, which will potentially allow a greater variety of contact positions with antigen targets (Table 5).

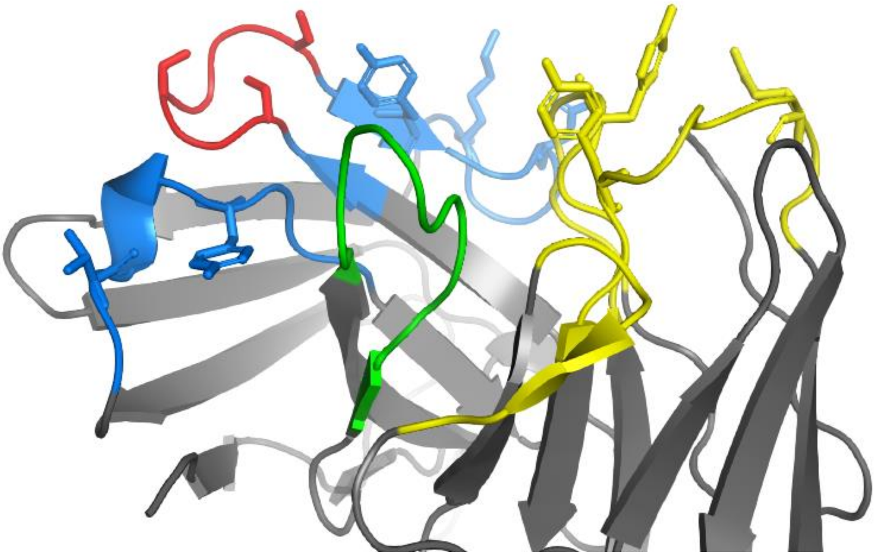


Figure 8 - Crystal structure of Fab fragments for FW-κ, side view. Blue: HCDR1 and HCDR2 residues. Yellow: LCDR1, LCDR2 and LCDR3 residues. Green: HCDR3 residues. Red: SGSGGS. Grey: Framework-region residues. Resolution = 1.73 **Å.**

Additionally, changes were made to HFR3 (Figure 8). This region corresponds to the framework-specific region between the HCDR2 and HCDR3. A portion of HFR3 is also sometimes referred to as "loop 4", since it also points outwards and has been found to interact with antigen epitopes.[2,3] We chose N330 and N333, which were big residues pointing outwards towards the solvent, which could potentially lead to contacts with antigen molecules (Figure 8). Besides that, these two residues also constitute a motif of possible asparagine deamidation, in which asparagines are typically converted into aspartates following the loss of an amide group from their side-chain. In a protein or peptide, these reactions are important because they may alter its structure, stability or function and may lead to faster protein degradation.[1]

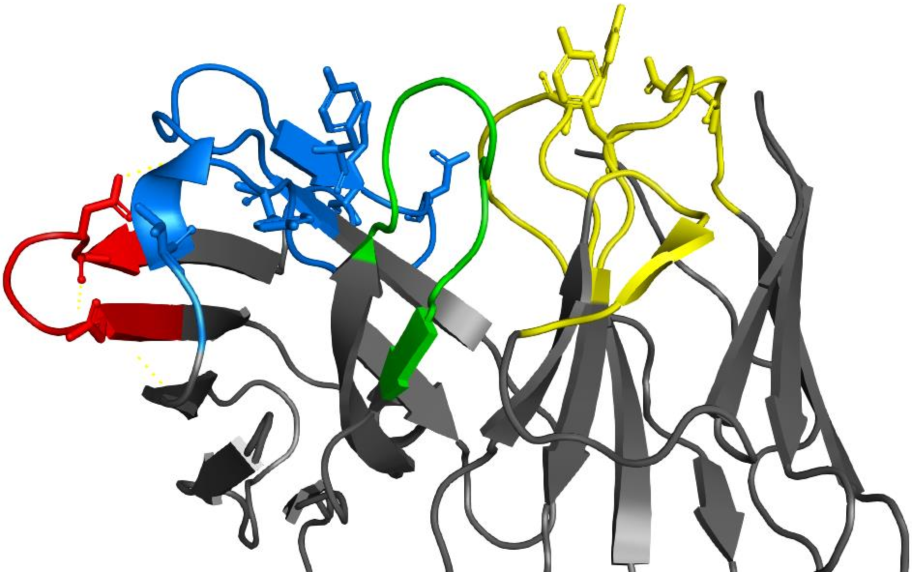


Figure 9 - Crystal structure of Fab fragments for FW-κ, side view. Blue: HCDR1 and HCDR2 residues. Yellow: LCDR1, LCDR2 and LCDR3 residues. Green: HCDR3 residues. Red: HFR3 (loop 4). Grey: Framework-region residues. Resolution = 1.73 **Å**

2.3.1.2. Choosing CDR-null mutations for Vκ1-39 (LC of FW-κ)

Even though the heavy chain is accounted as being the most important part of antigen-antibody binding, the light chain also plays a role in the antigen-antibody interaction. For that reason, CDR-null mutations were also designed for the light chain. Mutations to the Vκ1-39 correspond to the ones made for FW-κ. A summary of all chosen mutations for Vκ1-39 can be found in table 6.

Table 6 – Chosen CDR-null mutations for Vκ1-39

Name	Sequence	Mutation	obs.
LCDR1	ASQSISSYLN	-	ref sequence
	ASTSISSALN	Q48T, Y53A	
FR2 + LCDR2	LLIY AASSLQS	-	ref sequence
	LLIA AASSLQS	Y70A	
	LLIA TASTLQS	Y70A, A71T, S74T	
	LLIV AASSLQS	Y70V	
LCDR3	QQSYSTPLT	-	ref sequence
	QQSASTPLT	Y113A	

LCDR1 and LCDR3 were considered first because they are structurally nearby, and thus probably influencing each other. On LCDR1, we detected that Q48 and Y53 were pointing outwards (Figure 4). Since both these residues are able to establish H-bonds due to their polar nature, these were substituted by threonine and alanine, respectively (Table 6). Similarly, on LCDR3 loop Y113 was substituted by an alanine for the same reasons explained above.

On Vκ1-39, LCDR2 seems to have few amino acids that were worth to be mutated. However, we found that the last amino acid of LC Framework region 2 (LFR2) might also be of interest. LFR2 is immediately before LCDR2 and its last residue consists of an Y that seems to be pointing towards the HCDR3, as well as outwards to the antigen-side (Figure 4). Changing Y70 into a shorter residue such as alanine was tested, but this generated an AAA patch that may not be advantageous. A combination of this with two threonines on A71 and S74 was tested. A valine replacement was also tested (Table 6).

´

2.3.1.3. Choosing CDR-null mutations for Vλ3-9 (LC of FW-λ)

Mutations to the Vλ3-9 framework correspond to the ones made for FW-λ. A summary of all chosen mutations for Vκ1-39 can be found in Table 7.

Table 7 - Chosen CDR-null mutations for Vλ3-9

Name	Sequence	Mutation	obs.
LCDR1	GGNNIGSKNVH	-	ref sequence
	GGASIGSKSVH	N46A, N47S, N52S	-
	GGTSIGSKSVH	N46T, N47S, N52S	-
	GGASIGSTSVH	N46A, N47S, K51T, N52S	with W111A, D112T
FR2 + LCDR2	LVIY RDSNRPS	-	ref sequence
	LVIA SDSARPS	Y69A, R70S, N73A	Remove big residues pointing outwards and to CHDR-H3
	LVIV SDSTRPS	Y69V, R70S, N73A	
	FVIA SDSARPS	L66F, Y69A, R70S, N73A	
LCDR3	QVWDSSTVV	-	ref sequence
	QVADSSTVV	W111A	-
	QVATSSTVV	W111A, D112T	with N46A, N47S, K51T, N52S

On LCDR1, three asparagines were identified as significant residues pointing outwards (Figure 5). Asparagines are frequently responsible for establishing H-bonds with the antigen due to their polar nature, so they were replaced for smaller residues (Table 7). As for Vκ1-39, LCDR1 and LCDR3 likely affect each other on Vλ3-9 and will be considered before LCDR2. Interestingly, we saw that K51 from LCDR1 and D112 from LCDR3 might interact due to their structural proximity and opposing charges. These two residues may also form salt-bridges or H-bonds with the antigen. Following the same principles highlighted before, both of them were replaced by a threonine. On LCDR3, W111 was shown to be pointing towards the HCDR3. Since the hypothesis being tested also relies on minimal interaction of the HCDR3 with the remaining loops, we chose to replace W111 with an alanine, which greatly reduces the space occupied while maintaining a certain degree of hydrophobicity. Regarding the LCDR2, and similarly to Vκ1-39, Y69 of LFR2 was also replaced by an alanine. However, in Vλ3-9, the LCDR2 also has an arginine and an asparagine of interest that can also be replaced by smaller residues, since these are known to interact strongly with antigen residues. The residue L66 also caught our attention. In the crystal structure, there is space around it, and near the HCDR3, that could fit a bigger residue. For that reason, the L66F mutation will be tested. In the above-mentioned cases, all residues were pointing towards the HCDR3 (Figure 5).

2.3.2. Developability of CDR-null IgGs

Following the selection of CDR-null mutations, a thorough analysis was performed to evaluate their impact on antibody developability. Combinations of the aforementioned mutations were used to produce the 34 different IgGs summarized on Tables 8 and 9. Two FW-κ and FW-λ un-mutated references were also produced to provide a benchmark. All IgGs had the same HCDR3 (WGGDGFYAMDY), so that they can be compared to each other.

The IgGs were produced in HEK293T cells in batch reaction over 4 days after transiently transfected and purified with protein A columns as described in the methods section (see section 2.5.). Their production titer was assessed by analytical affinity ligand chromatography (ALC) to ensure that the CDR-null mutations allow for IgG production. Following production, the IgGs were analyzed for their aggregation profile by size-exclusion chromatography (SEC) and for their hydrophobic profile by hydrophobic interaction chromatography (HIC). Finally, their thermodynamic stability was measured by differential scanning fluorometry (DSF), as described in the methods section. All of the relevant data regarding IgG production can be found in Tables 10 and 11.

Table 8 – Panel of FW-κ candidates with selected mutations. CDR-null mutations highlighted in red. Deletions indicated with an hyphen.

	Kappa Germline						
	LCDR1	**LFR2 + LCDR2**	**LCDR3**	**HCDR1**	**HCDR2**	**HFR3**	**HCDR3**
κ ref	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 1	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	F**A**FSS**A**AMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 2	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	F**A**FSS**D**AMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 3	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSS**T**AMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 4	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGST**S**YA**S**SV**S**G	RFTISRDNSKNTY	WGGDGFYAMDY
κ 5	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGS-GST**S**YA**S**SV**S**G	RFTISRDNSKNTY	WGGDGFYAMDY
κ 6	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISG--GST**S**YA**S**SV**S**G	RFTISRDNSKNTY	WGGDGFYAMDY
κ 7	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AIS---GST**S**YA**S**SV**S**G	RFTISRDNSKNTY	WGGDGFYAMDY
κ 8	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGST**S**YA**S**SV**S**G	RFTISRD**S**SK**A**TY	WGGDGFYAMDY
κ 9	AS**T**SISS**A**LN	LLIY AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 10	ASQSISSYLN	LLI**A** AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
k 11	ASQSISSYLN	LLI**A** **T**AS**T**LQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 12	ASQSISSYLN	LLI**V** AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 13	ASQSISSYLN	LLIY AASSLQS	QQS**A**STPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 14	ASQSISSYLN	LLIY AASSLQS	QQSYSTPLT	F**A**FSS**A**AMS	AISGSGGST**S**YA**S**SV**S**G	RFTISRD**S**SK**A**TY	WGGDGFYAMDY
κ 15	AS**T**SISS**A**LN	LLI**A** AASSLQS	QQS**A**STPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
κ 16	AS**T**SISS**A**LN	LLI**A** AASSLQS	QQS**A**STPLT	F**A**FSS**A**AMS	AISGSGGST**S**YA**S**SV**S**G	RFTISRD**S**SK**A**TY	WGGDGFYAMDY

Table 9 –Panel of FW-λ candidates with selected mutations. CDR-null mutations highlighted in red. Deletions indicated with an hyphen.

	Lambda Germline						
	LCDR1	**LFR2 + LCDR2**	**LCDR3**	**HCDR1**	**HCDR2**	**HFR3**	**HCDR3**
λ ref	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 1	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FAFSSAAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 2	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FAFSSDAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 3	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSTAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 4	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTSYASSVSG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 5	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGS-GSTSYASSVSG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 6	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISG--GSTSYASSVSG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 7	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AIS---GSTSYASSVSG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 8	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTSYASSVSG	RFTISRDSSKATY	WGGDGFYAMDY
λ 9	GGASIGSKSVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 10	GGTSIGSKSVH	LVIY RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 11	GGNNIGSKNVH	LVIA SDSARPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 12	GGNNIGSKNVH	LVIV SDSTRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 13	GGNNIGSKNVH	FVIA SDSARPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 14	GGNNIGSKNVH	LVIY RDSNRPS	QVADSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 15	GGNNIGSKNVH	LVIY RDSNRPS	QVWDSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	RFTISRDSSKATY	WGGDGFYAMDY
λ 16	GGASIGSKSVH	LVIA SDSARPS	QVADSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 17	GGASIGSTSVH	LVIY RDSNRPS	QVATSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	RFTISRDNSKNTY	WGGDGFYAMDY
λ 18	GGASIGSKSVH	LVIA SDSARPS	QVADSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	RFTISRDSSKATY	WGGDGFYAMDY

.

Table 10 – Affinity-ligand chromatography (ALC) Size-exclusion chromatography (SEC), hydrophobic-interaction chromatography (HIC) and Differential Scanning Flurometry (DSF) data for FW-κ CDR-null IgGs. Total production of IgG in 35 mL of culture in shown in mg. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. An Ammonium Sulfate (AS) concentration below 0.8M is also indicated in yellow. ΔTm2 is calculated by subtracting the tm2 values from the reference antibody that was not mutated. ΔTm2 values above 1°C are indicated in yellow.

	Vk1-39			VH3-23			ALC	SEC		HIC	DSF		
	LCDR1	LFR2+ LCDR2	LCDR3	HCDR1	HCDR2	HFR3	IgG (mg)	HMWS (%)	MP (%)	[AS] (M)	Tm1 (°C)	Tm2 (°C)	ΔTm2 (°C)
κ ref	-	-	-	-	-	-	3.8	**0.6**	**99.3**	**0.76**	**69.0**	**85.0**	**0.0**
κ 1	-	-	-	T283A, Y287A	-	-	3.9	**1.0**	**98.9**	**0.84**	**69**	**83.5**	**-1.5**
κ 2	-	-	-	T283A, Y287D	-	-	3.4	**0.9**	**96.1**	**0.86**	**69**	**80.5**	**-4.5**
κ 3	-	-	-	Y287T	-	-	4.7	**0.9**	**99.0**	**0.83**	**69**	**83.0**	**-2.0**
κ 4	-	-	-	-	Y315S, D318S, K321S	-	4.0	**1.1**	**98.9**	**0.79**	**69**	**83.5**	**-1.5**
κ 5	-	-	-	-	delG314, Y315S, D318S, K321S	-	4.2	**1.1**	**98.8**	**0.86**	**69**	**83.5**	**-1.5**
κ 6	-	-	-	-	delS313, delG314, Y315S, D318S, K321S	-	3.2	**0.8**	**99.2**	**0.79**	**69**	**79.5**	**-5.5**
κ 7	-	-	-	-	delG312, delS313, delG314, Y315S, D318S, K321S	-	3.5	**0.5**	**99.5**	**0.77**	**69**	**76.5**	**-8.5**
κ 8	-	-	-	-	Y315S, D318S, K321S	N330S, N333A	2.5	**0.6**	**99.4**	**0.82**	**69**	**78.0**	**-7.0**
κ 9	Q48T, Y53A	-	-	-	-	-	2.5	**1.0**	**98.9**	**0.97**	**69**	**85.5**	**0.5**
κ 10	-	Y70A	-	-	-	-	2.8	**0.8**	**99.2**	**0.84**	**69**	**84.5**	**-0.5**
k 11	-	Y70A, A71T, S74T	-	-	-	-	1.9	**0.6**	**99.3**	**0.87**	**69**	**84.5**	**-0.5**
κ 12	-	Y70V	-	-	-	-	1.9	**0.5**	**99.4**	**0.89**	**69**	**84.5**	**-0.5**
κ 13	-	-	Y114A	-	-	-	1.9	**0.5**	**99.4**	**0.79**	**69**	**84.5**	**-0.5**
κ 14	-	-	-	T283A, Y287A	Y315S, D318S, K321S	N330S, N333A	1.6	**0.5**	**99.5**	**0.89**	**69**	**75.5**	**-9.5**
κ 15	Q48T, Y53A	Y70A	Y114A	-	-	-	2.2	**0.9**	**99.1**	**1.02**	**69**	**82.5**	**-2.5**
κ 16	Q48T, Y53A	Y70A	Y114A	T283A, Y287A	Y315S, D318S, K321S	N330S, N333A	3.6	**0.4**	**99.6**	**0.97**	**69**	**72.0**	**-13.0**

Table 11 – Affinity-ligand chromatography (ALC), Size-exclusion chromatography (SEC), hydrophobic-interaction chromatography (HIC) and Differential Scanning Flurometry (DSF) data for FW-λ CDR-null IgGs. Total production of IgG in 35 mL of culture in shown in mg. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. A AS concentration below 0.8M is also indicated in yellow. ∆Tm2 is calculated by subtracting the tm2 values from the reference antibody that was not mutated. ∆Tm2 values above 1°C are indicated in yellow.

	Vλ3-9			VH3-23			ALC	SEC		HIC	DSF		
	LCDR-1	LCDR-2	LCDR-3	HCDR-1	HCDR-2	FR-3	mAb (mg)	HMWS (%)	MP (%)	[AS] (M)	Tm1 (°C)	Tm2 (°C)	∆Tm2 (°C)
λ ref	-	-	-	-	-	-	3.2	9.2	90.7	1.07	69	81.00	0.00
λ 1	-	-	-	T283A, Y287A	-	-	3.0	7.0	92.8	1.07	68	79.50	-1.50
λ 2	-	-	-	T283A, Y287D	-	-	2.6	3.2	96.7	1.08	68	76.50	-4.50
λ 3	-	-	-	Y287T	-	-	2.0	6.0	93.7	1.08	69	79.50	-1.50
λ 4	-	-	-	-	Y315S, D318S, K321S	-	2.1	4.3	95.5	1.02	69	78.50	-2.50
λ 5	-	-	-	-	delG314, Y315S, D318S, K321S	-	2.8	3.5	96.2	1.04	69	78.50	-2.50
λ 6	-	-	-	-	delS313, delG314, Y315S, D318S, K321S	-	2.3	3.9	95.9	1.01	68	74.50	-6.50
λ 7	-	-	-	-	delG312, delS313, delG314, Y315S, D318S, K321S	-	1.4	2.0	97.7	1.00	68	71.50	-9.50
λ 8	-	-	-	-	Y315S, D318S, K321S	N330S, N333A	2.1	1.9	97.8	1.03	68	73.00	-8.00
λ 9	N46A, N47S, N52S	-	-	-	-	-	1.3	4.6	95.4	1.06	68	80.50	-0.50
λ 10	N46T, N47S, N52S	-	-	-	-	-	2.2	4.7	95.0	1.06	69	80.50	-0.50
λ 11	-	Y69A, R70S, N73A	-	-	-	-	1.6	3.6	96.1	0.91	69	79.00	-2.00
λ 12	-	Y69V, R70S, N73A	-	-	-	-	1.3	3.5	96.4	1.04	68	78.50	-2.50
λ 13	-	L66F, Y69A, R70S, N73A	-	-	-	-	1.3	1.8	98.1	0.94	68	76.00	-5.00
λ 14	-	-	W111A	-	-	-	2.0	4.0	95.9	1.06	68	80.50	-0.50
λ 15	-	-	-	T283A, Y287A	Y315S, D318S, K321S	N330S, N333A	1.4	1.7	98.0	1.05	68	71.00	-10.00
λ 16	N46A, N47S, N52S	Y69A, R70S, N73A	W111A	-	-	-	1.4	3.5	96.3	0.95	69	78.00	-3.00
λ 17	N46A, N47S, K51T, N52S	-	W111A, D112T	-	-	-	2.0	3.8	96.0	0.95	69	80.50	-0.50
λ 18	N46T, N47S, N52S	Y69A, R70S, N73A	W111A	T283A, Y287A	Y315S, D318S, K321S	N330S, N333A	1.5	1.8	98.1	0.99	69	70.00	-11.00

2.3.2.1. Production Titers

The differences in production outcomes between FW-κ and FW-λ do not come as a surprise since it is well documented how framework choice impacts the developability potential of an antibody molecule.[1,4] As shown on the tables 10 and 11, FW-κ candidates had an average production amount of 3.02±0.90 mg, while FW-λ candidates had and average production amount of 1.97±0.58 mg. A previous study compared the IgG production titers of 20 V_H families combined with 12 Vκ and 8 Vλ families, and showed that there is no statistically significant difference in production when using either Vκ or Vλ [1]. However, Vλ3-9 was not one of the chosen Vλ families, as it did not pass on pre-screening tests for frequent germline usage in the natural human antibody repertoire. The lower usage of Vλ3-9 in natural repertoires may be a consequence of sub-optimal folding conformations that impact its expression and stability, which have been described as a major factor impacting the relative representativeness of frameworks in antibody repertoires.[5–8]

2.3.2.2. Aggregation and Hydrophobicity

Characterization of aggregation propensity and hydrophobicity is of extreme relevance for assessing the developability potential of antibodies. SEC is an invaluable tool for biopharmaceutical process development and will provide immediate detection of contaminant proteins, aggregation intermediates (also referred to as high-molecular weight species) and degradation products (also referred to as lower molecular weight species). Aggregates are the most commonly observed product-related impurities, due decreases in therapeutic potency, and potential to cause to adverse immunogenicity-driven reactions on the patient [9–11].

However, SEC will not provide information on how a given protein might behave in the long-term, in storage or in the therapeutic setting. Inversely, HIC can be used as a predictor of mAb behavior across many stages of pharmaceutical development, since the hydrophobic profile often impacts aggregation, viscosity and solubility of mAbs. [11–13]. It has been shown that mAbs with greater retention times in HIC assays tend to exhibit increased aggregation and precipitation [12,14–18].

HIC has also been shown to correlate with the mAbs' tendency to self-associate upon injection into a patient, a critical factor to take into consideration for the development of therapeutic mAbs, which are typically delivered in high concentrations. [19,20]. This is of extreme relevance in a time where the industry is trending towards higher concentration formulations [21]. Subcutaneous injection of such drug products will be detrimentally impacted by higher viscosity, which as been associated with low patient compliance and to shorter *in vivo* half-life.[22,23] Increases in product viscosity have been reported to be caused by mAb self-association either through electrostatic interactions, through intermolecular interaction driven by CDRs' hydrophobic patches or combinations of both. [24–28]. The impact of hydrophobic patches and electrostatic interactions also holds true for solubility, and likewise, the delivery of high concentration pharmaceuticals will also be adversely impacted by products of lower solubility, which lead to poor activity, bioavailability, and high immunogenicity [29,30]. Both viscosity and solubility have been shown to impact several filtration steps, fill and finish, shipping and storage. [11]

Some methods for the measurement of viscosity (Cannon-Fenske Routine viscometer, Taylor Cone plate method) and solubility (direct measurement by gradual increment of concentration) require plenty of protein material that may not be available during screening phases.[11] Other methods, such as Dynamic Light Scattering (DLS) and PEG-induced precipitation can be used to measure viscosity [31] and solubility [29], respectively, without the need for abundant protein material. Nonetheless, it is likely that mAbs with smaller HIC retention times will have a lower tendency to aggregate, lower viscosity and higher solubility. Combined with its high reproducibility, low technical complexity and low protein requirements, HIC provides a common ground and solid evidence for the evaluation of many critical parameters.

FW-κ candidates (Table 10) had average purities of 99% and 1% of high molecular weight species (HMWS; which can be aggregation products and/or contaminants), across all samples, demonstrating that CDR-null mutations do not seem to impact the aggregation of FW-κ candidates. Inversely, candidates from FW-λ showed an improvement in their aggregation profile when compared to the FW-λ reference (without mutations). The reduction in surface exposure may be an explanation to this, considering that big residues pointing outwards towards the solvent were substituted by residues with smaller side-chains, which are less prone to drive aggregation.[32,33]

Regarding the HIC data, the results are also encouraging. The reference IgG for FW-κ eluted from the column at a 0.76 M of Ammonium Sulfate (AS), which indicates that it has a slightly higher hydrophobicity than the defined threshold for our internal standards (0.8 M). Most of the CDR-null mutations greatly improved the hydrophobic profile of the candidates, anticipating the elution in the HIC column up to 1 M of AS (Table 10). Inversely, the hydrophobic profile of FW-λ candidates got slightly worse when comparing to the FW-λ reference. Still, not only did all of the candidates stay above the desired threshold of 0.8 M, but they also stayed consistently above the best results for FW-κ, showing that CDR-null mutations also generate FW-λ IgGs with advantageous hydrophobic profiles (Table 11).

Even so, rather than being used as a cut-off, information coming from HIC data needs to be overlapped with a myriad of other biophysical assays performed during antibody development, since it may not strongly correlate with antibody precipitation if analyzed exclusively on its own, as shown by a wide study on multiple biophysical metrics of developability against a panel of clinical-stage antibodies.[34]

2.3.2.3. Thermal Stability

One additional factor to take into consideration is the thermodynamic stability. Thermodynamically stable proteins are able to maintain their structural and functional integrity under different temperature environments[15,35]. As such, it can influence the mAb product's characteristics during manufacturing and storage. High thermal stability of a mAb candidate indicates a well-packed structure that requires more energy to unfold, and thus serves as a good predictor of robustness to destabilizing factors such as temperature, pH and pressure. Indeed, it was shown that stably folded antibodies have a lower tendency to aggregate [15,35–38], a phenomenon which is most likely explained by the fact that more stable protein populations will decrease the percentage of aggregation-prone intermediates in solution, and therefore improve long-term storage. Additionally, mAbs with worse thermal stabilities were reported to be poorly expressed.[34,35] Most interestingly, the stability of antigen-binding Fab domains seems to play a crucial role in the overall stability of the IgG, and improving it may be a suitable strategy for a longer colloidal stability [34,37]. Judging by its variable nature, it is likely that the Fab domain can greatly impact the melting profile of an IgG1 and its overall stability. Generally, the first event of IgG1 unfolding starts at the CH2 domain of the Fc-region, which has been consistently reported to happen around 68-71°C.[39] This is followed by the unfolding of the Fab domain and, finally, by the denaturation of the more stable CH3 domains at higher temperatures near 90 °C. The variability coming from the CDR loops on the Fab domain can alter the chain of events that show up on a typical melting curve. For example, if the combination of mutations destabilizes the Fab fragment, the first transition represents the unfolding of the Fab fragment and the CH2 domain, while the second transition represents CH3 domain unfolding. In a case where a Fab domain is very stable, the first transition corresponds to the CH2 domain unfolding, and the second transition represents the unfolding of the Fab fragment and the CH3 domain concomitantly. Ideally, for an IgG, the melting profile may present three transitions, with the Fab unfolding occurring at distinct temperatures compared to the melting of the CH2 and CH3 domains.[40,41]. This observation is in agreement with the DSF data obtained throughout this work, and will be discussed later on this chapter. Thermodynamic stability can be analyzed by Differential Scanning Calorimetry (DSC) – which relies on the enthalpy of transitions –, or by DSF – which relies on the exposure of hydrophobic residues to a fluorescent dye. Both methods measure the folding-unfolding transitions of proteins (and molecules, in general) across an incremental range of temperature, in a given buffer solution. While DSC is the "golden standard" to analyze thermal denaturation of proteins[11,15,35,37,42,43], DSF is much easier to handle and equally good to find Tm values while using much lower sample amounts [38,44,45]. Due to its simplicity and higher throughputness, DSF was the approach employed in this study. As shown in tables 10 and 11, the denaturing temperature of the CH2 domain (Tm1) remained unaltered throughout all conditions. After the denaturation event of CH2 occurs, a second transition of higher enthalpy ensues at higher temperatures, which is normally associated with the concomitant denaturation of the Fab and CH3 domains (Tm2), as explained before. All mutations done to the heavy chain of both frameworks – VH3-23 – had a negative effect on the thermal stability of the IgGs, as measured by ΔTm2. Most notoriously, deletions to the G312, S313, and G314 residues and mutations in the N330 and N333, seem very detrimental and will be avoided in the future. This is shown both in FW-κ and FW-λ by how Tm2 dramatically changes when comparing with the reference and other mutations.

Looking at the Vκ1-39 light-chain, mutations on LCDR1, LCDR2, and LCDR3 seem mostly innocuous if done separately. However, if all three LCDRs are mutated simultaneously, then the de-stabilization is exacerbated and is bigger than the sum of the effect of the individual mutations (Table 10 – κ15 versus κ9-12). Similarly, mutations on LCDR1 and LCDR3 of Vλ3-9 light-chain seem mostly innocuous (Table 11 – λ9-10 and λ14). Moreover, even when LCDR1 and LCDR3 are mutated at the same time, stability remains mostly unaltered (Table 11 – λ17). But contrarily to Vκ1-39, mutations to the LCDR2 of Vλ3-9 cause a destabilizing effect of ΔTm2 =

-2.5 °C (Table 11 - λ11-13). All in all, the resulting ΔTm2 of cumulative mutations across all three LCDRs for both Vκ1-39 and Vλ3-9 is approximately the same (ΔTm2 $_{k15}$ = -2.5 °C and ΔTm2 $_{\boldsymbol{\lambda 16}}$ = -3.0 °C, respectively). On the other hand, if all mutations from the light-chain and heavy-chain are combined on the IgG, their thermal stability drops significantly, as shown for λ18 and κ16.

Taking all of that into account, the combination of mutations of all the CDRs was revised carefully to select additional candidates (see section 2.3.3). It is most interesting to see that, while CDR-null mutations had both negative and positive on the aggregation and hydrophobicity of IgG molecules, none of these mutations was able to improve the thermal stability on any of the frameworks. This hints that the germline sequences have evolved to maintain stability, and that deviations to the germline tend to destabilize the molecule. It was also interesting to detect a certain degree of addictiveness between mutations on different CDRs for the ΔTm2 value. This means that, when choosing which mutations were to be combined to generate a full CDR-null framework, the following criteria needed to be fulfilled: i) Mutations should not lead to a significant increase in aggregation; ii) Mutations should not significantly increase IgG hydrophobicity,; iii) Mutations should lead to the lowest thermal stability drop possible; iv) If two different mutations for the same CDR fulfill the above criteria, then we chose the mutation that best fits the CDR-null concept – i.e. the smallest amino acid.

2.3.3. Combining advantageous CDR-null mutations into null-frameworks

By looking at Tables 10 and 11, we can see that CDR-null mutations to the light chain do not decrease Tm2 by a great degree. On FW-κ, the combination of Q48T, Y53A, Y70A, and Y114A only leads to a decrease in 2.5 °C and improves the HIC profile to 1.02 M of AS (Table 10 – k15). Similarly, for FW-λ, the combination of N46A, N47S, N52S, Y69A, R70S, N73A, and W111A decreases thermal stability by 3 °C. Even though they do not change the HIC profile significantly, these mutations on FW-λ light chain seem to improve the SEC aggregation profile, reducing the HMWS to 3.5% (Table 11 – λ16). Regarding the heavy-chain, the combination of HCDR1 + HCDR2 + HFR3 mutations were very detrimental (Table 10 - ∆**Tm2** κ14 = -9.5°C), but we believed that this was mainly due to the HFR3 mutations being very de-stabilizing. HCDR1 and HCDR2 mutations alone only decrease thermal stability by 1.5 degrees each, while HCDR2 + HFR3 decreased the thermal stability by 7 degrees (table 10 – k1, k4 versus k8). Taking all these into account, we combined the aforementioned mutations with each other as shown in the tables 12 and 13 and performed the same biophysical characterization as before. It is important to note that some additional conditions were tested, with most of them constituting variations to the LCDR2 mutations. The main reason for such addition was to test different Y70 mutations. When replacing Y70 for an alanine, we end up with three consecutive alanines unless we change some of them, which was exactly what we did on A71T and S74T. A non-hydrophobic substitution of Y70 was also tested, by replacing it with a serine. We decided to test this and see if it yielded better biophysical properties.

Table 12 – Final CDR-null framework possibilities based on FW-κ. CDR-null mutations are indicated in red.

Fw	CDR-L1	CDR-L2	CDR-L3	CDR-H1	CDR-H2	HCDR3
κ	ASQSISSYLN	Y AASSLQS	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	WGGDGFYAMDY
κN1	ASTSISSALN	Y AASSLQS	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
κN2	ASTSISSALN	A TASTLQS	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
κN3	ASTSISSALN	S AASSLQS	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
κN4	ASTSISSALN	Y AASSLQS	QQAASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
κN5	ASTSISSALN	S AASSLQS	QQAASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY

Table 13 - Final CDR-null framework candidates based on FW-λ. CDR-null mutations are indicated in red.

Fw	CDR-L1	CDR-L2	CDR-L3	CDR-H1	CDR-H2	HCDR3
λ	GGNNIGSKNVH	Y RDSNRPS	QVWDSSTVV	FTFSSYAMS	AISGSGGSTYYADSVKG	WGGDGFYAMDY
λN1	GGASIGSTSVH	Y RDSNRPS	QVATSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
λN2	GGASIGSTSVH	A SDSARPS	QVATSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
λN3	GGASIGSTSVH	A SDSNRPS	QVATSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
λN4	GGASIGSTSVH	Y SDSNRPS	QVATSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY
λN5	GGASIGSTSVH	Y SDSARPS	QVATSSTVV	FAFSSAAMS	AISGSGGSTSYASSVSG	WGGDGFYAMDY

The IgGs were produced in HEK293T cells, and the results coming from SEC, HIC and DSF for these frameworks after production and characterization are summarized in Tables 14 and 15. In line with the previous experiments, the HIC profile of FW-κ improved while aggregation remained mostly unaltered. For FW-λ, as before, the CDR-null mutations improved the aggregation profile, and increased the hydrophobicity of the molecules, though this did not lead to values below the defined threshold of 0.8 M AS. Regarding the thermal stability, the behavior was also within expected. From our data, we concluded that Fw_κN1 and Fw_λN1 are the best candidates among their group.

Table 14 - Combination of null mutations on FW-κ. Total production of IgG in 35 mL of culture in shown in mg. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. An Ammonium Sulfate (AS) concentration below 0.8M is also indicated in yellow. ∆Tm2 is calculated by subtracting the Tm2 values from the reference antibody that was not mutated.

ID	ALC	SEC		HIC	DSF		
FW	IgG (mg)	HMWS %	MP %	[AS] (M)	Tm1	Tm2	∆Tm2
κ	3.8	0.6	99.3	0.76	69.0	85.0	0.0
κN1	8.7	1.8	98.2	0.95	69.0	79.0	- 6.0
κN2	8.3	1.9	98.0	0.95	69.0	77.0	- 8.0
κN3	1.6	2	98.0	0.94	69.0	77.0	- 8.0
κN4	2.8	2.2	97.7	0.97	69.0	79.0	- 6.0
κN5	1.8	2.1	97.9	0.98	69.0	77.5	- 7.5

Table 15 - Combination of null mutations on FW-λ. Total production of IgG in 35 mL of culture in shown in mg. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. An Ammonium Sulfate (AS) concentration below 0.8M is also indicated in yellow. ∆Tm2 is calculated by subtracting the Tm2 values from the reference antibody that was not mutated.

ID	ALC	SEC		HIC	DSF		
FW	IgG(mg)	HMWS %	MP %	[AS] (M)	Tm1	Tm2	∆Tm2
λ	3.2	9.2	90.7	1.07	69.0	81.0	0.0
λN	8.4	2.4	97.5	0.93	69.0	76.5	- 4.5
λN2	8.2	3.1	96.8	0.83	69.0	75.5	- 5.5
λN3	1.5	2.1	97.9	0.84	69.0	75.0	- 6.0
λN4	1.5	2.2	97.8	0.89	69.0	76.5	- 4.5
λN5	2.2	2.5	97.5	0.89	69.0	76.0	- 5.0

2.4. Discussion

When designing a synthetic library from scratch, before diving into the randomization of CDR sequences, a very important decision has to be made regarding which antibody framework to use. Formally, a framework is any final sequence that can occur after recombination of V(D)J segments within V_H and V_L loci, which then can be modified to fit the purpose. Several strategies have been used to guide the selection of V_H and V_L frameworks with advantageous folding capabilities and good biophysical characteristics, such as selecting frameworks based on their natural representation, generation of master genes based on their sub-family, or by directly testing fixed V_H:V_L pairs for their aggregation propensity, expression level, and thermal stability.[1,8,46] After that, usually a round of finer optimization ensues by removal of potential post-translational modification sites (PTMs), nucleotide sequence codon optimization and non-paratope changes. This culminates in frameworks that seek to have optimal biophysical characteristics and that can accommodate diversity to their CDRs without resulting in many clones with compromising biophysical characteristics.

CDR-null frameworks were generated from two distinct frameworks: FW-κ and FW-λ. The newly designed CDR-null frameworks kept their framework regions unaltered but differ in specific residues along their germline CDR sequences, which have been modified to minimize potential contacts with antigen molecules. To examine if the CDR-null mutations had an impact on the developability of the IgGs generated from these frameworks, analytical SEC and HIC were performed, as well as a DSF assay. For both frameworks, we were able to establish CDR-null frameworks with good aggregation propensities and hydrophobicity, and with acceptable thermal stability profiles with a clear separation between the CH2 denaturation at 69°C and the Fab denaturation at temperatures >75°C (Table 14-15).

In this work its was shown how simple deviations from the germline sequence can alter the biophysical characteristics of antibody frameworks. More specifically, our results indicate that these mutations can improve the aggregation and hydrophobic profiles, but always lead to a decrease in the overall thermal stability of the framework. Apart from two cases, all of the remaining mutations were done to CDR-loop sequences. Since the majority of CDR-loop residues are in contact with the solvent, it is most unexpected to see how changing such residues impacts the remaining IgG molecule as a whole, specially concerning its thermal stability. This raises important questions about the design of antibody libraries and their expected quality and effective/functional clonal diversity.

As previously referred, extensive tests are done to antibody frameworks to ensure they can yield clones with good biophysical properties. But as thorough as these approaches may be, they do not consider how the randomization of CDR-loops affects the stability of clones generated by the randomization. In the same way that less stable frameworks appear to be selected less frequently[5–8], we postulate that extensive simultaneous randomization of several CDR-loops may generate very unstable clones that will not be selected during phage-display or other *in vitro* mAb-discovery procedures. This is critical as it may play a big role in the difference between the theoretical versus the actual diversity of the library. This is consistent with observations from other authors that HCDR3 randomization is sufficient to derive nanomolar affinity binders to multiple protein and hapten targets, with or without the assistance of a diversified LCDR3.[47] Additionally, it was observed that HCDR3-only libraries were dominant over LCDR3+HCDR3 libraries when pooled together. Such behavior was attributed to a reduction in the overall stability of LCDR+HCDR3-randomized libraries, derived from the higher likelihood of generating dysfunctional clones and inter-CDR structural clashes, as opposed to the HCDR3-randomized library, that is always paired with the more stable germline LCDR3.[47] In line with those observations, our results suggest that germline sequences are optimized towards stability, and we stress the importance that frameworks used to generate antibody libraries must be intrinsically stable enough that they can accommodate oscillations in the final melting temperature of each individual clone. More specifically, our observations increase the relevance of libraries that rely on fewer randomized CDRs in the primary screening phases, as opposed to many synthetic libraries that rely on randomization of several CDR loops at the same time.

As with any deviation from germline sequences, the diversification of HCDR3 sequences can also impact the IgG's developability, which means that the frameworks they are inserted in must be robust enough to accommodate these effects and still lead to a developable IgGs. It has been shown that candidates selected from stable frameworks closely preserved the biophysical features that were characteristic of the parental frameworks, with slight biophysical deviations attributed to the influence of certain HCDR3 sequences.[1] It has also been shown that longer HCDR3 sequences may lead to higher hydrophobicity and tendency to aggregate, possibly due to their bigger contribution in surface exposed residues.[48] As such, after careful analysis we decided not to follow up with null-frameworks derived from FW-λ, as they do not seems to be able to provide enough robustness for the current mutational loads. Even though their hydrophobicity and aggregation profiling presented encouraging results, their expression levels and thermal stability greatly underperformed (Tables 10 and 15). In fact, it was been observed in our lab that FW-λ has more developability issues than FW-κ (data not shown). Such tendencies will certainly be exacerbated by the added thermal instability of CDR-null mutations and lead to poor expression of candidates. On the other hand, FW-κ typically yields more robust candidates, and retains a good thermal stability after the CDR-null mutations. With all the information gathered after the rational design and biophysical characterization, we will proceed with the κN1 and κN2 CDR-null frameworks throughout the rest of this thesis work (described in the following chapters).

2.5. Materials and Methods

FW-κ and FW-λ structure

FW-κ preliminary structure was determined by molecular replacement methods using Phaser Molecular Replacement (MR) program from the CCP4i software package. A Fab (PDB: 3SOB) that exhibited the high sequence homology with the framework FW-κ (88.8% identity with light-chain, 79.8% with heavy-chain) was used for the search model. After retrieving Rotation and Translation solutions for the model, the AutoBuild program (PHENIX package) was used for model building and obtaining improved maps for FW-κ, and ArpWarp was used for FW-λ. These maps were subsequently inspected and manually built in COOT. Structure refinement was carried out with programs REFMAC5. For the purposes of finding positions that were pointing towards the solvent side and towards the HCDR3, the crystal structures were inspected using Pymol, and mutated as stated in section 2.3.

IgG expression

The expression plasmids were ordered from ThermoFisher's GeneArt platform. The Light-chain (LC) and Heavy-chain (HC) of each IgG were ordered separately and transfected simultaneously (in a 1:1 ratio) with Polyethylenimine (PEI, in a 4:1 ratio with DNA) into 100 x 10^6 human embryonic kidney-293T (HEK- 293T) cells in 18 mL of FreeStyle^{TM} 293 Expression Medium (Life Technologies®). After 4 hours, an additional 20 mL of medium are added to the cells for a final cell concentration of 2.5 x 10^6 cells/mL. Transiently transfected cell cultures were incubated for 4 days in humidified atmosphere of 5% CO_2, 37°C and 140 rpm. After 4 days in culture, transfected cells are centrifuged at 300g for 10 minutes, and their supernatant collected, and vacuum filtered using 0.22 µm pore Steriflips (FisherScientific). The supernatant can be stored at 4°C for a week or at -20°C for extended periods.

IgG purification

IgG purification was performed by Affinity Ligand Chromatography, on Tecan Freedom EVO 200 (equipped with a Liquid Handling arm with 8 stainless steel tips, syringes of 1 mL and TeChrom, to enable fast IgG purification) using MabSelect Sure RoboColumns (Repligen; Ref.: PN 01050408R. Total Column Volumn (CV) = 200 µL). Phosphate Saline Buffer (PBS, pH 7.0) was used as the equilibration buffer. Samples were loaded 1 mL at a time, for a total final load of 35 mL. Retrieval of IgGs was achieved by isocratic elution using 5 CV of 50 mM Citrate-NaCl pH 3.0, for a final eluted volume of 1mL. The pH is neutralized by the addition of 150 µL of 1M Tris-HCL pH 9.0. The sample is then filtered through a 0.22 µm filter pore using a syringe and stored at -20°C. Final volume = 1.15 mL.

IgG quantification

IgGs were quantified via HPLC Affinity Ligand Chromatography (HPLC-ALC), using a POROS™ CaptureSelect™ CH1-XL Affinity HPLC Column 2.1 x 30 mm, coupled to an Agilent 1260 Infinity II (Agilent Technologies). Separation of protein species was achieved using a flow rate of 2 mL/min and detection at 210 nm. Samples are injected directly without any previous dilution (injection volume = 50 µL), and the following method on Table 16 is employed for each individual injection:

Table 16 – ALC-HPLC method. Mobile Phase A: 10 mM NaH_2PO_4, 150 mM NaCl, pH 7.5; Mobile Phase B: 10 mM HCl, 150 mM NaCl, pH 2.0;

Time after injection (in minutes)	Mobile Phase A (in %)	Mobile Phase B (in %)
0	100	0
1.87	100	0
1.88	0	100
4.38	0	100
4.39	100	0

mAb peaks are manually integrated to calculate the Peak Area. Antibody concentration is calculated according to Equation 1.

Equation 1:

$$C_A = Peak\ Area_A \times \left(\frac{C_{IS}}{Peak\ Area_{IS}}\right) \times \left(\frac{1}{\frac{RRF_A}{RRF_{IS}}}\right)$$

An internal standard (IS) IgG with known concentration was used to generate an internal response factor (RRF_{IS} = Peak Area $_{IS}$/ Concentration $_{IS}$). Each sample concentration (C_A) was calculated as shown in Rome, K. & McIntyre, A. (2012)[1], by taking into account the concentration of IS (C_{IS}) and by comparing the sample's RRF (RRF_A) with the RRF of IS (RRF_{IS}). (Equation 1)

Size-exclusion chromatography

150 mM Potassium Phosphate pH 6.5 was used to dilute IgG samples to a final concentration of 1 mg/mL. Each candidate was analyzed by size exclusion chromatography on a TSKgel G3000SWXL column (Tosoh Biosciences) using an Agilent 1260 Infinity II HPLC system, equipped with a multi-wavelength detector. A total run time of 35 minutes per sample was employed, after a 10 µg injection of each sample. The mobile phase was 150mM Sodium Phosphate pH 6.0 + 400 mM NaCl. Separation of protein species according to their molecular weight was achieved by applying an isocratic elution using a flow rate of 0.4 mL/min and detection at 210 nm. Peak integration of IgG monomers was done at a retention time around 20 minutes; these are referred to as "main peaks". Peaks and/or shoulders before the "main peak" are indicative of aggregation and referred to as "high molecular weight species" (HMWs). Peaks and/or shoulders after the main peak are indicative of fragmentation of the IgG monomer and designated "low molecular weight species (LMWs).

Hydrophobic-interaction chromatography

The hydrophobic profile of each candidate was analyzed by hydrophobic-interaction chromatography (HIC) in a TSKgel Butyl-NPR column (4.6 mm ID x 35 mm L) (Tosoh Biosciences). PBS was used to dilute the samples to 1 mg/mL. The mobile phase A was composed by 20 mM His/HCl, pH = 6.0 containing 1.5 M AS. Gradient elution of protein species was achieved by a gradual buffer replacement of mobile phase A with 20 mM His/HCl, pH 6.0 (mobile phase B). The gradient is 20 CV in length and has a slope of – 0.103 M AS per minute. A calibration curve was employed, where the retention time of reference standards was plotted against concentration of AS to calculate the hydrophobicity of the protein molecules.

Differential Scanning Fluorometry

Differential Scanning Fluorometry was performed in BioRad CFX96. Samples were diluted to 0.3 mg/mL (Vf = 50uL) in PBS, to which SYPRO orange (previously prepared) was added. Sypro orange preparation was done by diluting the 5000x stock, by pippeting 1.4 uL from the stock solution into 1 mL of H20. The reaction was performed with a temperature increment of 0.5 ºC/min, from 25 °C to 100 °C.

2.6. Acknowledgements

The author – Jorge Sampaio – was responsible for all the work in this chapter. The author would like to thank Ana Maria Gonçalves for her work on solving the structure of FW-κ and FW-λ, which paved the way for this work. The author also acknowledges André F. Faustino for valuable input when evaluating the structural data and Susana L. Lobo and Joana A. Gomes for providing support on the HPLC methodologies.

2.7. References of Chapter 2

1. Tiller, T. *et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs* **5**, 445–470 (2013).

2. Henry, K. A. *et al.* Role of the non-hypervariable FR3 D-E loop in single-domain antibody recognition of haptens and carbohydrates. *J. Mol. Recognit. JMR* **32**, e2805 (2019).

3. Kelow, S. P., Adolf-Bryfogle, J. & Dunbrack, R. L. Hiding in plain sight: structure and sequence analysis reveals the importance of the antibody DE loop for antibody-antigen binding. *bioRxiv* 2020.02.12.946350 (2020) doi:10.1101/2020.02.12.946350.

4. Singer, I. I. *et al.* Optimal humanization of 1B4, an anti-CD18 murine monoclonal antibody, is achieved by correct choice of human V-region framework sequences. *J. Immunol.* **150**, 2844–2857 (1993).

5. Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. *Antibodies* **8**, 44 (2019).

6. Schofield, D. J. *et al.* Application of phage display to high throughput antibody generation and characterization. *Genome Biol.* **8**, R254 (2007).

7. Griffiths, A. D. *et al.* Isolation of high affinity human antibodies directly from large synthetic repertoires. *EMBO J.* **13**, 3245–3260 (1994).

8. Knappik, A. *et al.* Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides 1 1Edited by I. A. Wilson. *J. Mol. Biol.* **296**, 57–86 (2000).

9. Rosenberg, A. S. Effects of protein aggregates: An immunologic perspective. *AAPS J.* **8**, E501–E507 (2006).

10. Singh, S. K. *et al.* An Industry Perspective on the Monitoring of Subvisible Particles as a Quality Attribute for Protein Therapeutics. *J. Pharm. Sci.* **99**, 3302–3321 (2010).

11. Xu, Y. *et al.* Structure, heterogeneity and developability assessment of therapeutic antibodies. *mAbs* **11**, 239–264 (2018).

12. Kohli, N. *et al.* A novel screening method to assess developability of antibody-like molecules. *mAbs* **7**, 752–758 (2015).

13. M, H., S, M., M, S. & A, A. Separation of mAbs molecular variants by analytical hydrophobic interaction chromatography HPLC: overview and applications. *Mabs* **6**, 852–858 (2014).

14. Chennamsetty, N., Helk, B., Voynov, V., Kayser, V. & Trout, B. L. Aggregation-Prone Motifs in Human Immunoglobulin G. *J. Mol. Biol.* **391**, 404–413 (2009).

15. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. *Proc. Natl. Acad. Sci.* **106**, 11937–11942 (2009).

16. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Prediction of Aggregation Prone Regions of Therapeutic Proteins. *J. Phys. Chem. B* **114**, 6614–6624 (2010).

17. Lee, C. C., Perchiacca, J. M. & Tessier, P. M. Toward aggregation-resistant antibodies by design. *Trends Biotechnol.* **31**, 612–620 (2013).

18. Perchiacca, J. M. & Tessier, P. M. Engineering Aggregation-Resistant Antibodies. *Annu. Rev. Chem. Biomol. Eng.* **3**, 263–286 (2012).

19. Hebditch, M., Roche, A., Curtis, R. A. & Warwicker, J. Models for Antibody Behavior in Hydrophobic Interaction Chromatography and in Self-Association. *J. Pharm. Sci.* **108**, 1434–1441 (2019).

20. Lilyestrom, W. G., Yadav, S., Shire, S. J. & Scherer, T. M. Monoclonal Antibody Self-Association, Cluster Formation, and Rheology at High Concentrations. *J. Phys. Chem. B* **117**, 6373–6384 (2013).

21. Garidel, P., Kuhn, A. B., Schäfer, L. V., Karow-Zwick, A. R. & Blech, M. High-concentration protein formulations: How high is high? *Eur. J. Pharm. Biopharm.* **119**, 353–360 (2017).

22. Allmendinger, A. *et al.* Rheological characterization and injection forces of concentrated protein formulations: An alternative predictive model for non-Newtonian solutions. *Eur. J. Pharm. Biopharm.* **87**, 318–328 (2014).

23. Baek, Y. & Zydney, A. L. Intermolecular interactions in highly concentrated formulations of recombinant therapeutic proteins. *Curr. Opin. Biotechnol.* **53**, 59–64 (2018).

24. Dobson, C. L. *et al.* Engineering the surface properties of a human monoclonal antibody prevents self-association and rapid clearance in vivo. *Sci. Rep.* **6**, 38644 (2016).

25. Esfandiary, R., Parupudi, A., Casas-Finet, J., Gadre, D. & Sathish, H. Mechanism of Reversible Self-Association of a Monoclonal Antibody: Role of Electrostatic and Hydrophobic Interactions. *J. Pharm. Sci.* **104**, 577–586 (2015).

26. Guo, Z. *et al.* Structure-Activity Relationship for Hydrophobic Salts as Viscosity-Lowering Excipients for Concentrated Solutions of Monoclonal Antibodies. *Pharm. Res.* **29**, 3102–3109 (2012).

27. Sharma, V. K. *et al.* In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proc. Natl. Acad. Sci.* **111**, 18601–18606 (2014).

28. Yadav, S. *et al.* Establishing a Link Between Amino Acid Sequences and Self-Associating and Viscoelastic Behavior of Two Closely Related Monoclonal Antibodies. *Pharm. Res.* **28**, 1750–1764 (2011).

29. Gibson, T. J. *et al.* Application of a High-Throughput Screening Procedure with PEG-Induced Precipitation to Compare Relative Protein Solubility During Formulation Development with IgG1 Monoclonal Antibodies. *J. Pharm. Sci.* **100**, 1009–1021 (2011).

30. Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M. & Popovic, B. Rapid and accurate in silico solubility screening of a monoclonal antibody library. *Sci. Rep.* **7**, 8200 (2017).

31. He, F. *et al.* High-throughput dynamic light scattering method for measuring viscosity of concentrated protein solutions. *Anal. Biochem.* **399**, 141–143 (2010).

32. Mishra, A., Ranganathan, S., Jayaram, B. & Sattar, A. Role of solvent accessibility for aggregation-prone patches in protein folding. *Sci. Rep.* **8**, 12896 (2018).

33. Tartaglia, G. G., Cavalli, A., Pellarin, R. & Caflisch, A. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Sci.* **13**, 1939–1941 (2004).

34. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* **114**, 944–949 (2017).

35. Seeliger, D. *et al.* Boosting antibody developability through rational sequence optimization. *mAbs* **7**, 505–515 (2015).

36. Barthelemy, P. A. *et al.* Comprehensive Analysis of the Factors Contributing to the Stability and Solubility of Autonomous Human V $_H$ Domains. *J. Biol. Chem.* **283**, 3639–3654 (2008).

37. Brader, M. L. *et al.* Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies. *Mol. Pharm.* **12**, 1005–1017 (2015).

38. He, F., Hogan, S., Latypov, R. F., Narhi, L. O. & Razinkov, V. I. High throughput thermostability screening of monoclonal antibody formulations. *J. Pharm. Sci.* **99**, 1707–1720 (2010).

39. Vermeer, A. W. P. & Norde, W. The Thermal Stability of Immunoglobulin: Unfolding and Aggregation of a Multi-Domain Protein. *Biophys. J.* **78**, 394–404 (2000).

40. Ionescu, R. M., Vlasak, J., Price, C. & Kirchmeier, M. Contribution of Variable Domains to the Stability of Humanized IgG1 Monoclonal Antibodies. *J. Pharm. Sci.* **97**, 1414–1426 (2008).

41. Schaefer, J. V., Sedlák, E., Kast, F., Nemergut, M. & Plückthun, A. Modification of the kinetic stability of immunoglobulin G by solvent additives. *mAbs* **10**, 607–623 (2018).

42. Kayser, V. *et al.* Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies. *Biotechnol. J.* **6**, 38–44 (2011).

43. Nemergut, M. *et al.* Analysis of IgG kinetic stability by differential scanning calorimetry, probe fluorescence and light scattering: Kinetic Stability Analysis of IgG. *Protein Sci.* **26**, 2229–2239 (2017).

44. Lavinder, J. J., Hari, S. B., Sullivan, B. J. & Magliery, T. J. High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering. *J. Am. Chem. Soc.* **131**, 3794–3795 (2009).

45. Shi, S., Semple, A., Cheung, J. & Shameem, M. DSF Method Optimization and Its Application in Predicting Protein Thermal Aggregation Kinetics. *J. Pharm. Sci.* **102**, 2471–2483 (2013).

46. Shi, L. *et al.* De Novo Selection of High-Affinity Antibodies from Synthetic Fab Libraries Displayed on Phage as pIX Fusion Proteins. *J. Mol. Biol.* **397**, 385–396 (2010).

47. Mahon, C. M. *et al.* Comprehensive Interrogation of a Minimalist Synthetic CDR-H3 Library and Its Ability to Generate Antibodies with Therapeutic Potential. *J. Mol. Biol.* **425**, 1712–1730 (2013).

48. Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Mol. Immunol.* **112**, 338–346 (2019).

Chapter 3 – Grafting anti-Herceptin HCDR3 to CDR-null antibody frameworks

3.1. Summary

In this work, HCDR3 sequences from antibodies discovered using libraries based on FW-κ were grafted into FW-κN1 and FW-κN2, with the objective of evaluating the effects that CDR-null mutations have on binding. The binding kinetics of the de-trained antibodies was measured by bio-layer interferometry (BLI) on Octet Red96, and the role of V_L and V_H in binding isolated. CDR-null mutations were shown to be sufficient to disrupt the binding kinetics of the three parental antibodies tested. This opens the door to further exploration of the CDR-null frameworks and to employ their use in panning campaigns to achieve different outcomes of those achieved by FW-κ.

3.2. Introduction

As shown in the previous chapter, CDR-null mutations do not constitute changes to the full germline sequences, but rather point mutations across a specific germline pair (in this case Vκ1-39/VH3-23). However, altering amino acid residues on the germline sequences was shown to have a significant effect on the biophysical characteristics of the tested IgG molecules (see chapter 2, section 2.3). Thus, we wondered if CDR-null mutations could have an equally remarkable effect on binding kinetics.

As with any other protein, changes to the amino acid residues will modify bonding arrangements within the Fab structure, which in turn leads to alternative conformations with different characteristics. The Fab domain structural packing puts HCDR3 loops in a central position relative to the other CDRs (see Figure 4-5 in chapter 2).[1–3] Hence, changes to germline residues will favor different HCDR3 sequences that are better suited to the new conformations made possible.

In this work, the impact of the CDR-null mutations on the binding kinetics will be experimentally assessed by “de-training” three anti-Herceptin antibodies, by grafting their HCDR3 into CDR null frameworks: κN1 and κN2 (Figure 1) and see if they retain the ability to bind to their cognate antigen. To further investigate the full effect of CDR-null mutations, we also tested antibodies that carried the CDR-null mutations exclusively on the heavy chain (e.g., VH3-23_null, Figure 1) or on the light chain (e.g., Vκ1-39_null1/2, Figure 1).

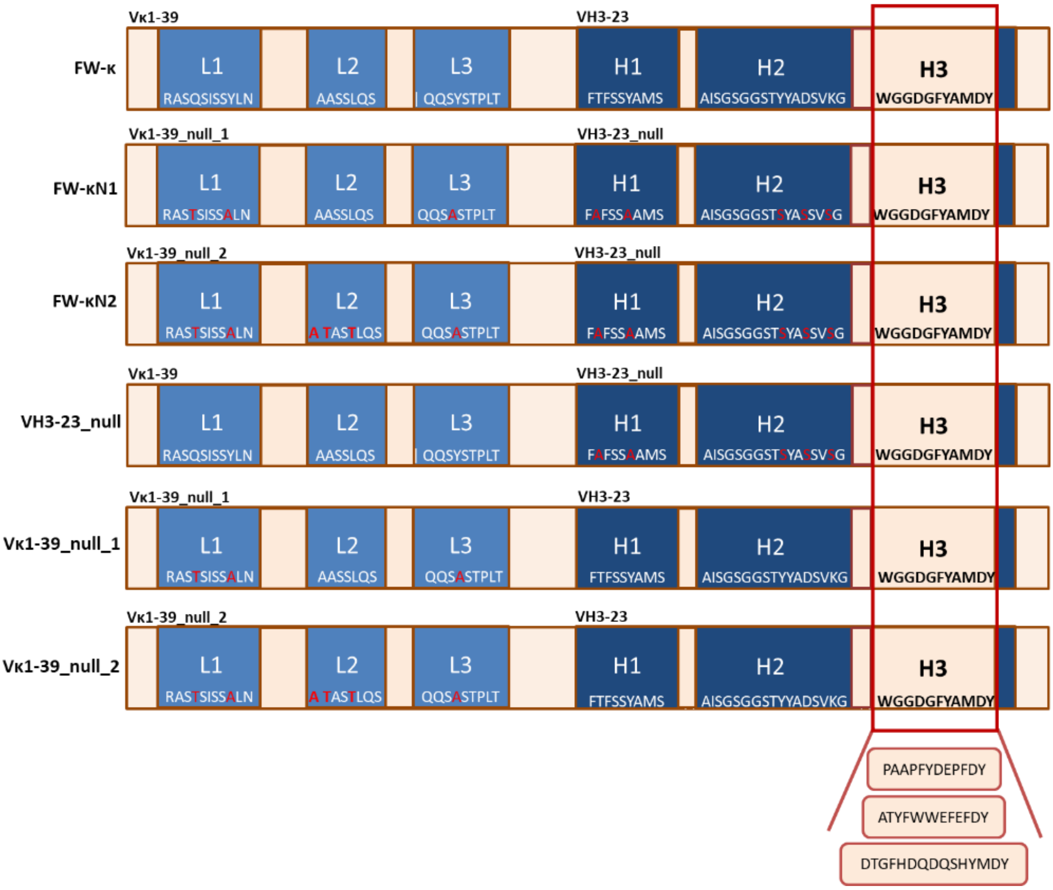


Figure 1 - Schematic representation of HCDR3 grafting into CDR null-frameworks.

3.3. Results

The three case-studies will be hereinafter identified by their HCDR3 sequences: i) PAAPFYDEPFDY; ii) ATYFWWEFEFDY; iii) DTGFHDQDQSHYMDY. The affinity of the three parental anti-Herceptin antibodies towards their cognate antigen was measured by bio-layer interferometry (BLI) on Octet Red96 (for experimental details see section 3.5). Their measured affinity was 3.42 ± 2.5 nM, 2.46 ± 1.2 nM and 9.25 ± 8.7 nM, respectively. After calculating the affinity of the parental antibodies against Herceptin, their HCDR3 were cloned from the parental FW-κ into five different frameworks: Fw_κN1, Fw_κN2, VH3-23_null, Vk1-39_null_1, Vk1-39_null_2. (Figure 1). These were produced as described in section 3.5, analyzed for their affinity towards Herceptin (Tables 1-3), and for their biophysical characteristics (Table 4).

3.3.1. Binding kinetics of de-trained antibodies

For the PAAPFYDEPFDY case-study, we can see that the HCDR3 requires the heavy chain’s support to drive the binding to Herceptin. As shown in Table 1, VH-null mutations produced the biggest impact in terms of affinity. If only Vκ1-39 is mutated, PAAPFYDEPFDY is still able to drive binding towards Herceptin, despite a 10-fold loss in affinity in comparison with the parental antibody. For Vk1-39_null_1 and_2, association curves are fast and specific, and slow dissociation curves are observed. Vk1-39_null_1 and _2 have minimal differences, with both cases reaching affinities around 30 nM (Table 1). On the other hand, when VH3-23 is mutated (VH-null), the antibody loses its ability to bind to Herceptin, regardless of the light-chain pairing used, highlighting the importance of the VH3-23 germline (Table 1).

For the ATYFWWEFEFDY case-study, we can see a very interesting cooperation between both the heavy-chain and light-chain in the stabilization of HCDR3 binding against Herceptin. When both chains are mutated, the IgG loses the ability to bind to the antigen completely. On the other hand, when either VH3-23 or Vk1-39 are not mutated, some binding ability is retained. Mutations to VK1-39 lead to slow association rates, fast dissociation, and lower overall affinity (210 nM and 260 nM, Table 2), indicating that VH3-23 is not sufficient to help ATYFWWEFEFDY maintain high-affinity binding to Herceptin. Inversely, when VH3-23 is mutated, fast and specific association rates are still observed, and affinity is around 54 nM. But while such observations provide evidence that the LCDR1/2/3 loops are positively impacting the association phase, the resulting off-rates of mutating V_H are very fast. This means that even though VH3-23 does not improve on-rates, it still has an important role in stabilizing binding (Table 2).

For the DTGFHDQDQSHYMDY case-study**,** we were able to isolate a very interesting effect for LCDR2. Once again, if both VK1-39 and VH3-23 germlines are mutated, no binding occurs. When VH3-23 is mutated, a certain degree of association rate is maintained but the off-rates are considerably fast, which highlights the role of HCDR1/2 in stabilizing HCDR3 binding to the antigen, and/or that LCDR1/2/3 loops are positively impacting the association phase. The most interesting effect is seen on the difference between VK1-39_null mutations. While VK1-39_null_1 leads to similar results to VH3-23_null, VK1-39_null_2 abrogates binding to Herceptin almost completely. Vk1-39_null_2 replicates mutations from vk1-39_null_1 but is also mutated on the last FR2 position and on LCDR2, while Vk1-39_null_1 is not (Figure 1). These mutated residues seem to play a big role in helping DTGFHDQDQSHYMDY binding to Herceptin (Table 3).

Table 1 - Sensogram curves for PAAPFYDEPFDY case-study. The parental Vκ1-39/VH3-23 condition was challenged against 75 nM, 37.5 nM, 18.8 nM, 9.4 nM, 4.7 nM, 2.3 nM, and 1.2 nM of Herceptin. The remaining samples were assayed against 750 nM, 375 nM, 188 nM, 94 nM, 47 nM, 23 nM, and 12 nM.

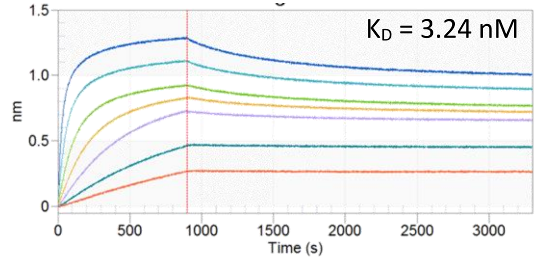
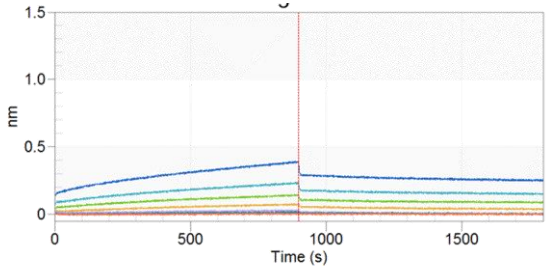
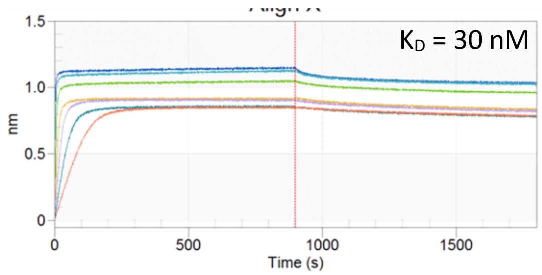
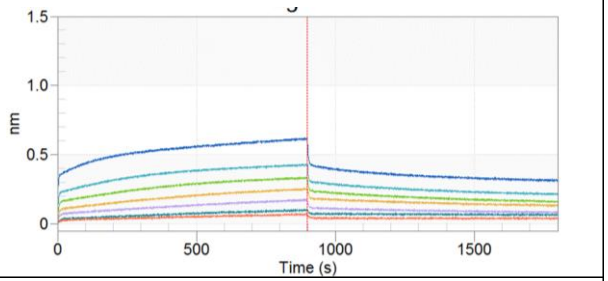
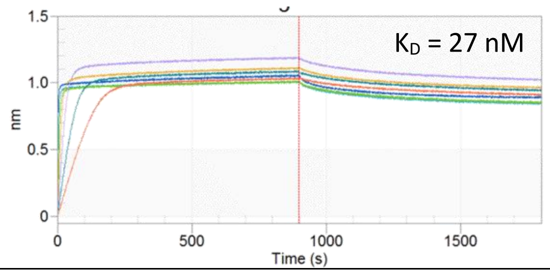
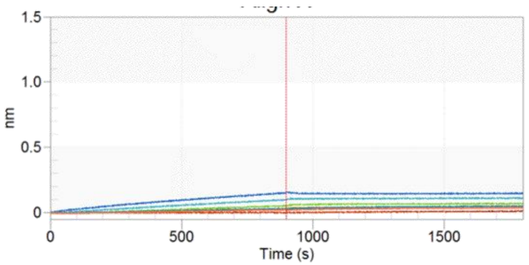
		Heavy-Chain	
		VH3-23	**VH-null**
Light-Chain	**Vκ 1-39**		
	Vκ-null_1		
	Vκ-null_2		

Table 2 - Sensogram curves for ATYFWWEFEFDY case-study. The parental Vκ1-39/VH3-23 condition was challenged against 75 nM, 37.5 nM, 18.8 nM, 9.4 nM, 4.7 nM, 2.3 nM, and 1.2 nM of herceptin. The remaining samples were assayed against 750 nM, 375 nM, 188 nM, 94 nM, 47 nM, 23 nM, and 12 nM.

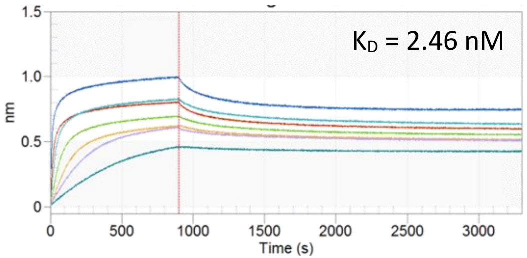
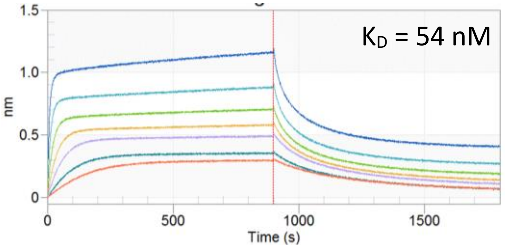
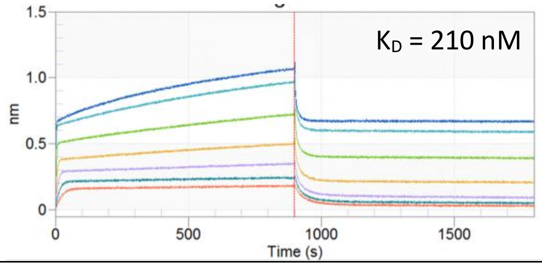
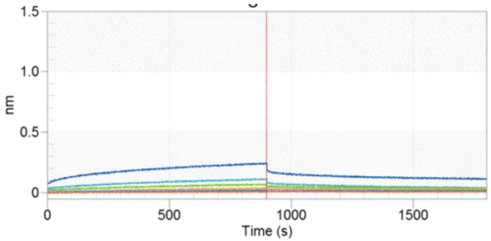
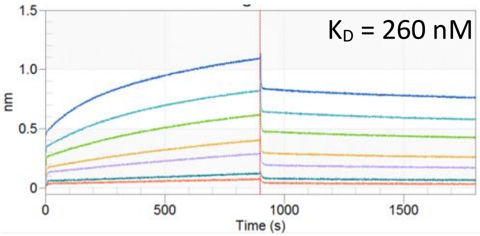
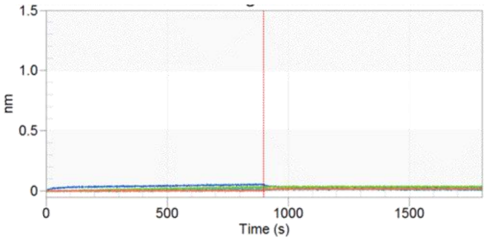
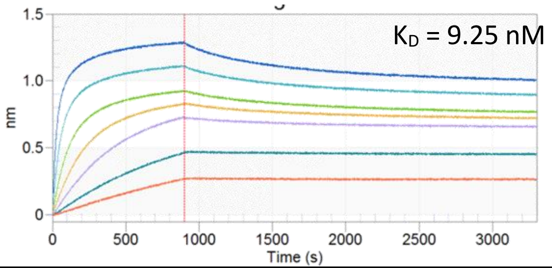
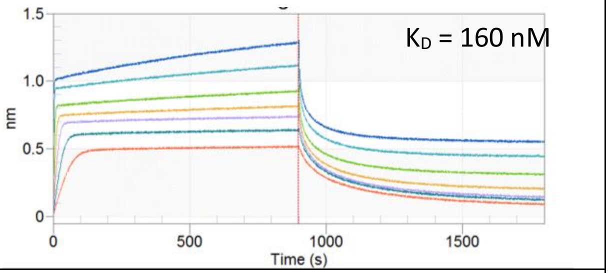
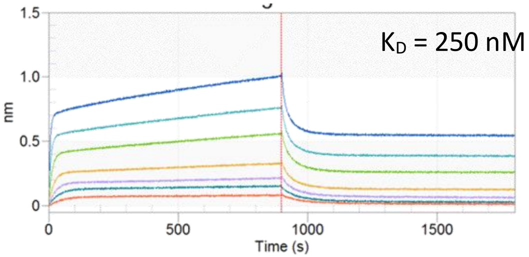
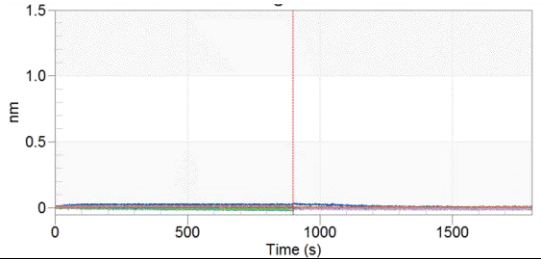
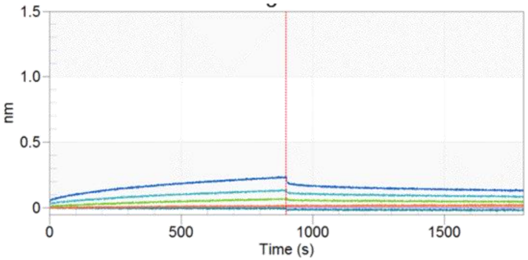
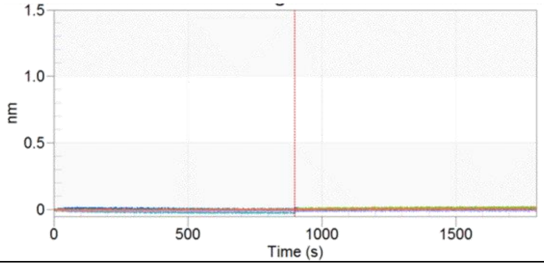
		Heavy-Chain	
		VH3-23	**VH-null**
Light-Chain	**Vκ 1-39**	 K_D = 2.46 nM	 K_D = 54 nM
	Vκ-null_1	 K_D = 210 nM	
	Vκ-null_2	 K_D = 260 nM	

Table 3 - Sensogram curves for DTGFHDQDQSHYMDY case-study. The parental Vκ1-39/VH3-23 condition was challenged against 75 nM, 37.5 nM, 18.8 nM, 9.4 nM, 4.7 nM, 2.3 nM, and 1.2 nM of herceptin. The remaining samples were assayed against 750 nM, 375 nM, 188 nM, 94 nM, 47 nM, 23 nM, and 12 nM.

		Heavy-Chain	
		VH3-23	**VH-null**
Light-Chain	**Vκ 1-39**		
	Vκ-null_1		
	Vκ-null_2		

3.3.2. Developability of de-trained antibodies

De-trained antibodies where also analyzed for their biophysical characteristics. On Table 4, the effects of both frameworks (FW-κN1 and FW-κN2) and HCDR3 sequences on IgG's developability are highlighted. On one instance, it shows how IgG's possessing the same HCDR3 retain similar characteristics. On the other hand, it shows how all IgG´s respond in a similar fashion to framework changes. For instance, all IgG´s that retained the VH3-23 framework closely match the thermal stability and hydrophobicity of their parental antibodies. On the other hand, the substitution with null-framework sequences in either V_L or V_H lead to obvious decreases in thermal stability for all IgG´s (Table 4).

Table 4 - Developability of de-trained antibodies. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. A AS concentration below 0.8M is also indicated in yellow. ΔTm2 is calculated by subtracting the Tm2 values from the reference antibody that was not mutated. ΔTm2 values above 1°C are indicated in yellow.

VL	VH	HCDR3	HMWS (%)	MP (%)	Tm1	Tm2	[AS] (M)
Vκ1-39	VH3-23	PAAPFYDEPFDY	1.2	98.8	68.5	76.5	1.08
		ATYFWWEFEFDY	0.2	99.8	68.5	77.0	0.45
		DTGFHDQDSHYMDY	1.3	98.7	68.5		1.11
Vκ1-39_null_1	VH3-23_null	PAAPFYDEPFDY	0.8	99.2	68.5		1.01
		ATYFWWEFEFDY	0.0	100.0	69.5		0.63
		DTGFHDQDSHYMDY	1.5	98.5	69.5		1.09
Vκ1-39_null_2	VH3-23_null	PAAPFYDEPFDY	0.5	99.5	67.5		1.05
		ATYFWWEFEFDY	0.0	100.0	69.5		0.69
		DTGFHDQDSHYMDY	0.5	99.5	70.0		1.08
Vκ1-39	VH3-23_null	PAAPFYDEPFDY	0.5	99.5	69.0		1.00
		ATYFWWEFEFDY	0.0	100.0	69.0		0.60
		DTGFHDQDSHYMDY	68.7	31.3	70.0		1.08
Vκ1-39_null_1	VH3-23	PAAPFYDEPFDY	1.8	98.2	69.0	75.0	1.06
		ATYFWWEFEFDY	0.0	100.0	69.0	75.0	0.52
		DTGFHDQDSHYMDY	2.0	98.0	69.0		1.10
Vκ1-39_null_2	VH3-23	PAAPFYDEPFDY	1.4	98.6	69.0	71.0	1.08
		ATYFWWEFEFDY	2.3	97.7	69.0	72.0	0.64
		DTGFHDQDSHYMDY	1.8	98.2	69.0		1.10

3.4. Discussion

The grafting of HCDR3 from “parental” frameworks into “null” frameworks was expected not only to change the biophysical characteristics of the de-trained antibody when compared to the parental, but also to change their affinity towards the target. Since null-frameworks reduce potential contacts with the antigen, we predicted that such changes would have an impact on the affinity kinetics towards a given target. This gets even more important if we consider the role of germline sequences within the conformational flexibility hypothesis and polyreactivity of antibodies.[4,5] Antibody diversity is encoded by the rearrangement of variable (V), diversity (D), and joining (J) germline gene segments.[2] The event of somatic recombination prior to antigen exposure encodes a large germline (or natural) repertoire that must be capable of recognizing a large and diverse array of antigens. These natural and unmutated antibodies require a degree of polyreactivity to be able to interact with a number of epitopes that potentially exceeds the combinatorial diversity of the immunoglobulin genes. In fact, polyreactive antibodies constitute up to 20% of antibodies produced by B-lymphocytes in the peripheral blood, and that many have nucleotide and amino acid sequences that closely resemble germline sequences.[4] Such antibodies are capable of binding to a wide range of antigen epitopes but with substantially less affinity than a monoreactive antibody has towards its cognate antigen. Additionally, Rosetta multi-state studies show that germline gene segments are close to ideal for polyspecificity.[5] Thus, the conformational flexibility hypothesis suggests that germline sequences allow for a degree of structural plasticity that facilitates binding to a greater variety of epitopes. Oppositely, highly specific Ag-induced antibodies reveal structural invariance and high rates of somatic mutations.[4–6]

Although the main driver of antibody-antigen specificity is the HCDR3, it is likely that the remaining CDRs play a relevant role in most of the candidates selected from antibody primary libraries such as the ones evaluated in this chapter. It thus comes with no surprise that CDR-null mutations impact binding towards Herceptin for PAAPFYDEPFDY, ATYFWWEFEFDY and DTGFHDQDQSHYMDY, which have all been selected from the FW-κ parental library. On a first instance, the presence of all CDR-null mutations simultaneously was responsible for the complete abrogation of binding to Herceptin in all three case-studies. But, most interestingly, we saw that distinct regions across the Fab domain can have an effect, depending on the HCDR3 used. For PAAPFYDEPFDY, it was the VH3-23, for ATYFWWEFEFDY it was mostly VK1-39, with some degree of cooperation from VH3-23 regarding the dissociation phase, and for DTGFHDQDQSHYMDY it was not only the Vk1-39 but more specifically the LCDR2 region.

The effect of CDR-null mutations on the developability were also evident. Thermal stability tended to decrease when CDR-null mutations were present, as observed previously (see chapter 2). On chapter 2, we observed that CDR-null mutations tended to decrease hydrophobicity for FW-κ. Although true for the WGGDGFYAMDY sequence, it is unlikely that this holds true for all HCDR3 sequences. This was evident by the opposite effects that CDR-null mutations had on the three parental antibodies tested on this chapter. While PAAPFYDEPFDY and DTGFHDQDQSHYMDY had some decrease in hydrophobicity, the nature of their HCDR3 still allowed them to have good hydrophobicity profiles overall. Inversely, CDR-null mutations ameliorated ATYFWWEFEFDY's hydrophobicity while maintaining decent thermal stability values. It has been shown that candidates selected from stable frameworks closely preserved the biophysical features that were characteristic of the parental frameworks, with slight biophysical deviations attributed to the influence of certain HCDR3 sequences.[3] This hints that CDR-null mutations may or may not be detrimental when combined with certain HCDR3 sequences.

Results on this chapter show that modification of germline residues by CDR-null mutations are likely to disrupt the structural plasticity of primary antibodies and impact binding towards their cognate antigen. Whether or not we have a decrease in structural plasticity was not experimentally explored. However, we could confirm that HCDR3 sequences that were previously able to drive affinity towards Herceptin, can no longer do so when grafted into κN1 and κN2 primed frameworks. As such, we predict that CDR-null mutations may be sufficient to change the Vκ1-39/VH3-23 conformational dynamics and favor different HCDR3 during biopanning selection processes, as if it was a complete germline change.

More specifically, we wonder if candidates that would otherwise be selected from FW-κ could also be selected from CDR-null κN1 and κN2 frameworks in some cases, despite the changes. Such will be explored further on chapter 4, by performing a side-by-side comparison of panning results using parental and CDR-null frameworks primary libraries.

3.5. Materials and Methods

Vector cloning

IgG expression plasmids encoding the three anti-Herceptin HCDR3 sequences were kindly provided by Dr. Stefan Ewert (Novartis AG, Basel). The HCDR3 sequences were removed using the BstBI/BlpI restriction enzymes and cloned into κN1 and κN2 heavy chain encoding plasmids.

IgG expression

The expression plasmids were ordered from ThermoFisher's GeneArt platform. The Light-chain (LC) and Heavy-chain (HC) of each IgG were ordered separately and transfected simultaneously (in a 1:1 ratio) with Polyethylenimine (PEI, in a 4:1 ratio with DNA) into 100 x 10^6 human embryonic kidney-293T (HEK- 293T) cells in 18 mL of FreeStyle^{TM} 293 Expression Medium (Life Technologies®). After 4 hours, an additional 20 mL of medium are added to the cells for a final cell concentration of 2.5 x 10^6 cells/mL. Transiently transfected cell cultures were incubated for 4 days in humidified atmosphere of 5% CO_2, 37°C and 140 rpm. After 4 days in culture, transfected cells are centrifuged at 300g for 10 minutes, and their supernatant collected, and vacuum filtered using 0.22 µm pore Steriflips (FisherScientific). The supernatant can be stored at 4°C for a week or at -20°C for extended periods.

IgG purification

IgG purification was performed by Affinity Ligand Chromatography, on Tecan Freedom EVO 200 (equipped with a Liquid Handling arm with 8 stainless steel tips, syringes of 1 mL and TeChrom, to enable fast IgG purification) using MabSelect Sure RoboColumns (Repligen; Ref.: PN 01050408R. Total Column Volumn (CV) = 200 µL). Phosphate Saline Buffer (PBS, pH 7.0) was used as the equilibration buffer. Samples were loaded 1 mL at a time, for a total final load of 35 mL. Retrieval of IgGs was achieved by isocratic elution using 5 CV of 50 mM Citrate-NaCl pH 3.0, for a final eluted volume of 1mL. The pH is neutralized by the addition of 150 µL of

1M Tris-HCL pH 9.0. The sample is then filtered through a 0.22 µm filter pore using a syringe and stored at -20°C. Final volume = 1.15 mL.

IgG quantification

IgGs were quantified via HPLC Affinity Ligand Chromatography (HPLC-ALC), using a POROS™ CaptureSelect™ CH1-XL Affinity HPLC Column 2.1 x 30 mm, coupled to an Agilent 1260 Infinity II (Agilent Technologies). Separation of protein species was achieved using a flow rate of 2 mL/min and detection at 210 nm. Samples are injected directly without any previous dilution (injection volume = 50 µL), and the following method on Table 5 is employed for each individual injection:

Table 5 – ALC-HPLC method. Mobile Phase A: 10 mM NaH_2PO_4, 150 mM NaCl, pH 7.5; Mobile Phase B: 10 mM HCl, 150 mM NaCl, pH 2.0;

Time after injection (in minutes)	Mobile Phase A (in %)	Mobile Phase B (in %)
0	100	0
1.87	100	0
1.88	0	100
4.38	0	100
4.39	100	0

mAb peaks are manually integrated to calculate the Peak Area. Antibody concentration is calculated according to Equation 1.

Equation 1: $$C_A = Peak\ Area_A \times \left(\frac{C_{IS}}{Peak\ Area_{IS}}\right) \times \left(\frac{1}{\frac{RRF_A}{RRF_{IS}}}\right)$$

An internal standard (IS) IgG with known concentration was used to generate an internal response factor (RRF_{IS} = Peak Area $_{IS}$/ Concentration $_{IS}$). Each sample concentration (C_A) was calculated as shown in Rome, K. & McIntyre, A. (2012)[1], by taking into account the concentration of IS (C_{IS}) and by comparing the sample's RRF (RRF_A) with the RRF of IS (RRF_{IS}). (Equation 1)

Size-exclusion chromatography

150 mM Potassium Phosphate pH 6.5 was used to dilute IgG samples to a final concentration of 1 mg/mL. Each candidate was analyzed by size exclusion chromatography on a TSKgel G3000SWXL column (Tosoh Biosciences) using an Agilent 1260 Infinity II HPLC system, equipped with a multi-wavelength detector. A total run time of 35 minutes per sample was employed, after a 10 µg injection of each sample. The mobile phase was 150mM Sodium Phosphate pH 6.0 + 400 mM NaCl. Separation of protein species according to their molecular weight was achieved by applying an isocratic elution using a flow rate of 0.4 mL/min and detection at 210 nm. Peak integration of IgG monomers was done at a retention time around 20 minutes; these are referred to as “main peaks”. Peaks and/or shoulders before the “main peak” are indicative of aggregation and referred to as “high molecular weight species” (HMWs). Peaks and/or shoulders after the main peak are indicative of fragmentation of the IgG monomer and designated “low molecular weight species (LMWs).

Hydrophobic-interaction chromatography

The hydrophobic profile of each candidate was analyzed by hydrophobic-interaction chromatography (HIC) in a TSKgel Butyl-NPR column (4.6 mm ID x 35 mm L) (Tosoh Biosciences). PBS was used to dilute the samples to 1 mg/mL. The mobile phase A was composed by 20 mM His/HCl, pH = 6.0 containing 1.5 M AS. Gradient elution of protein species was achieved by a gradual buffer replacement of mobile phase A with 20 mM His/HCl, pH 6.0 (mobile phase B). The gradient is 20 CV in length and has a slope of – 0.103 M AS per minute. A calibration curve was employed, where the retention time of reference standards was plotted against concentration of AS to calculate the hydrophobicity of the protein molecules.

Differential Scanning Fluorometry

Differential Scanning Fluorometry was performed in BioRad CFX96. Samples were diluted to 0.3 mg/mL (Vf = 50uL) in 43uL of PBS, to which 7 µL of SYPRO orange (previously prepared) was added. Sypro orange preparation was done by diluting the 5000x stock, by pipetting 1.4 µL from the stock solution into 1 mL of H20. The reaction was performed with a temperature increment of 0.5 ºC/min, from 25 °C to 100 °C.

Herceptin Biotinylation

1mg of biotin was dissolved in 166 µL H20 to do a 10 mM solution. Then, 40.5 uL of Sulfo-NHS-SS-Biotin was used to biotinylate 0.5 mL of 10 mg/mL Herceptin. The procedure is done following the directions of EZ-Link Sulfo-NHS-LC-Biotin kit (ThermoScientific A39257 21335). Samples were incubated 1 hour at room temperature. The samples were then passed through a Zeba Spin desalting column, 7K MWCO, 2mL (Thermoscientific, #89891, #QL227761).

Octet affinity measurements of anti-Herceptin antibodies

All kinetic assays were performed on Octet® RED96 (ForteBio), using 96-well plates (Corning), at 30°C and 1000 rpm orbital shake speed. Samples were diluted in freshly prepared kinetic buffer (ForteBio). Biotinylated hHerceptin was loaded onto Streptavidin (SA) Octet biosensor tips, by submerging them for 20 seconds in a 200 µL solution of biotinylated-Herceptin at 0.05 mg/mL. This is followed by a baseline step of 1 minute in kinetic buffer. The Herceptin-loaded tips are then submerged in wells for 900 seconds – the association phase – containing different concentrations of anti-Herceptin antibodies: 75 nM, 37.5 nM, 18.8 nM, 9.4 nM, 4.7 nM, 2.3 nM, 1.2 nM. This step is followed by a 1800 seconds dissociation phase in kinetic buffer. The Herceptin-loaded tips were also dipped in wells that only contained kinetic buffer as a reference.

Estimation of interaction rates and affinity parameters

Binding sensorgrams were first aligned at the beginning of the association phase, and following the single reference subtraction, they were globally fit to a 1:1 binding model, were a single k_{on} and k_{off} is calculated for all binding sensorgrams for every concentration tested.

3.6. Acknowledgements

The author – Jorge Sampaio – was responsible for all the work in this chapter. The author would like to thank Dr. Stefan Ewert and Jonas Schilz for providing the three anti-Herceptin antibodies used in this work. The author also acknowledges Elísio Silva and André F. Faustino for valuable input when evaluating the kinetic measurement data.

3.7. References of Chapter 3

1. Rees, A. R. Understanding the human antibody repertoire. *mAbs* **12**, 1729683 (2020).

2. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).

3. Sela-Culang, I., Kunik, V. & Ofran, Y. The Structural Basis of Antibody-Antigen Recognition. *Front. Immunol.* **4**, (2013).

4. Notkins, A. L. Polyreactivity of antibody molecules. *Trends Immunol.* **25**, 174–179 (2004).

5. Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E. & Meiler, J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. *PLoS Comput. Biol.* **9**, e1003045 (2013).

6. Krishnan, L., Sahni, G., Kaur, K. J. & Salunke, D. M. Role of Antibody Paratope Conformational Flexibility in the Manifestation of Molecular Mimicry. *Biophys. J.* **94**, 1367–1376 (2008).

Chapter 4 – The impact of the CDR-null concept in the antibody discovery process

4.1. Summary

On the previous chapter, we showed how CDR-null mutations are sufficient to disrupt binding of candidates to their target. Here, the CDR-null concept is taken one step forward, with the generation of primed libraries bearing the CDR-null mutations identified previously (see Chapter 2). These newly generated primed libraries were based on FW-κN1 and FW-κN2 and randomized on HCDR3, and were separated on two different pools, according to their HCDR3 sizes: 10-15aa and 16-20aa. The libraries were used in phage-display pannings against Herceptin, and evaluated for their overall quality, reproducibility, and diversity of HCDR3 sequences. All parameters were compared against control libraries derived from FW-κ. In addition, several candidates were selected from all tested conditions and evaluated for their affinity towards Herceptin and overall developability potential. The primed libraries (κN1 and κN2) showed lower diversity when compared with the control (κ), a behavior that can be explained by the lower overall polyreactivity typical of antibodies with less germline residues. Regardless of the lower diversity relative to the control, the κN1 and κN2 primed libraries were able to yield candidates of all HCDR3 sizes, with high affinity towards Herceptin, and equal distribution of affinities when compared with κ library.

4.2. Introduction

Antibody discovery processes at our lab employ phage-display protocols to discover antibodies against targets of interest. For that, we use fully synthetic primary libraries with diversity focused on the HCDR3 loop. This means that while we can select the best HCDR3 sequences after stringent rounds of panning, the remaining of the CDRs are kept unchanged, and thus, do not suffer any selective pressure. Any successful binding will be guided by HCDR3 with the support of germline residues, either by direct binding to the antigen molecule or by stabilizing HCDR3 binding. Primary binders coming from libraries based on FW-κ usually sit on the 10-100 nM scale regarding binding to their target. However, affinity maturation protocols employed on FW-κ primary binders do not always yield optimal results. This is likely happening because of eventual detrimental interactions of the primary binder via their germline CDRs that were not correctly identified, and thus, not optimized. Primed libraries κN1 and κN2 bearing CDR-null mutations aim to reduce the likelihood of germline contacts (specially clashes) with the antigen, and to select candidates that behave better in the affinity maturation steps that follow. Here, we explore how primed libraries κN1 and κN2 behave on a primary panning setting and how they compare with FW-κ.

4.3. Results

Three different frameworks will serve as base to generate six different phage-display primary libraries. These libraries will be challenged against Herceptin in a liquid suspension phage-display protocol. Their outcomes will be compared with the objective of determining wether CDR-null mutations impact the HCDR3 sequences selected after a phage-display protocol, in the following manner: i) Looking for differences in HCDR3 length distribution; ii) Querying the κN1 and κN2 datasets for HCDR3 sequences that were found in past FW-κ pannings against Herceptin; iii) Inspecting the total diversity of HCDR3 sequences by employing a systematic clustering method; iv) Select several κN1 and κN2 candidates and measure their overall affinity and compare it with κ candidates.

4.3.1. Generating test libraries

Six phage-display primary libraries were generated according to their framework and HCDR3 length: framework-κ_10-15aa (κ_10-15); framework-κ_16-20aa (κ_16-20); framework a_null-1 10-15aa (κN1_10-15); framework a_null-1 16-20aa (κN1_16-20); framework a_null-2 10-15aa (κN2_10-15) and framework a_null-2 16-20aa (κN2_16-20) (Table 1, Figure 1).

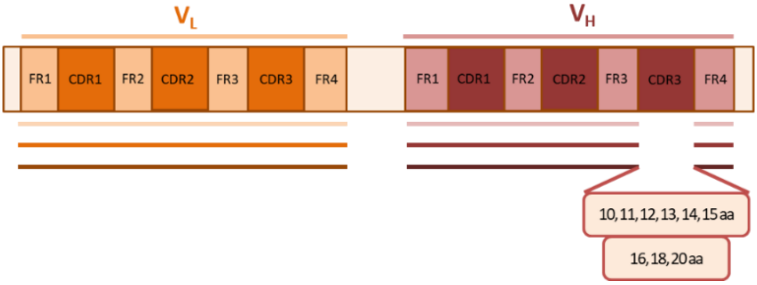


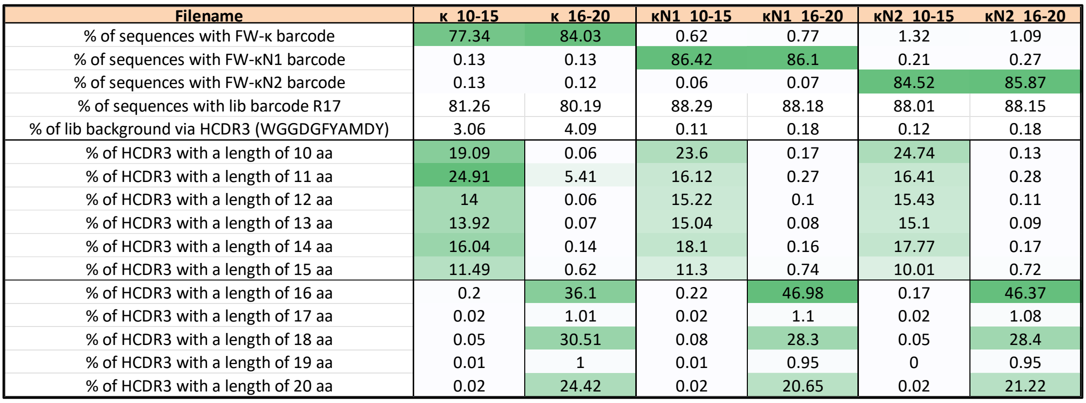
Figure 1 - Schematic representation of the test libraries

The HCDR3 were randomized according to the composition of the CDRs found in naturally occurring human antibody genes (confidential data, not shown). In all lengths and frameworks, most positions have up to 17 different amino acid possibilities, although certain positions in HCDR3 were less randomized or unchanged, due to the strong natural occurrence of certain amino acids for structural or stability reasons. This randomization design is referred to as R17 and has a DNA barcode encoded to it for quality control purposes. Libraries cloned in phagemid vectors were electroporated into TG1 cells and their diversity measured. All libraries had diversities above 10^8 cfu/mL and low vector background (Table 1, see section 4.5. – material and methods).

Table 1 - Test-libraries nomenclature and respective library size after transformation of TG1F+ cells by electroporation. Library size and vector background assessed by counting colony forming units (cfu) after serial dilutions and plating transformed cell after the electroporation protocol (section 4.5.).

ID	Framework	Size Pool	Library Size (cfu/mL)	Vector Background (%)
κ_10-15	FW-κ	10, 11, 12, 13, 14, 15	6.1E+08	1.51%
κ_16-20	FW-κ	16, 18, 20	5.8E+08	1.57%
κN1_10-15	FW-κN1	10, 11, 12, 13, 14, 15	3.3E+08	0.04%
κN1_16-20	FW-κN1	16, 18, 20	2.2E+08	0.06%
κN2_10-15	FW-κN2	10, 11, 12, 13, 14, 15	2.8E+08	0.03%
κN2_16-20	FW-κN2	16, 18, 20	1.3E+08	0.06%

Before moving to the panning protocol, a quality control step to the phage-display libraries was performed (Figure 2). The TG1 cells' DNA was extracted and analysed by Next-generation Sequencing (NGS). All three frameworks have an independent barcode correctly assigned to them (FW-κ, FW-κN1 and FW-κN2), similar percentage of sequences with the randomization barcode R17 (which indicates the correct HCDR3 randomization mentioned in 4.3.1), and low levels of library background (as read by the percentage of the WGGDGFYAMDY sequence).



Filename	κ 10-15	κ 16-20	κN1 10-15	κN1 16-20	κN2 10-15	κN2 16-20
% of sequences with FW-κ barcode	77.34	84.03	0.62	0.77	1.32	1.09
% of sequences with FW-κN1 barcode	0.13	0.13	86.42	86.1	0.21	0.27
% of sequences with FW-κN2 barcode	0.13	0.12	0.06	0.07	84.52	85.87
% of sequences with lib barcode R17	81.26	80.19	88.29	88.18	88.01	88.15
% of lib background via HCDR3 (WGGDGFYAMDY)	3.06	4.09	0.11	0.18	0.12	0.18
% of HCDR3 with a length of 10 aa	19.09	0.06	23.6	0.17	24.74	0.13
% of HCDR3 with a length of 11 aa	24.91	5.41	16.12	0.27	16.41	0.28
% of HCDR3 with a length of 12 aa	14	0.06	15.22	0.1	15.43	0.11
% of HCDR3 with a length of 13 aa	13.92	0.07	15.04	0.08	15.1	0.09
% of HCDR3 with a length of 14 aa	16.04	0.14	18.1	0.16	17.77	0.17
% of HCDR3 with a length of 15 aa	11.49	0.62	11.3	0.74	10.01	0.72
% of HCDR3 with a length of 16 aa	0.2	36.1	0.22	46.98	0.17	46.37
% of HCDR3 with a length of 17 aa	0.02	1.01	0.02	1.1	0.02	1.08
% of HCDR3 with a length of 18 aa	0.05	30.51	0.08	28.3	0.05	28.4
% of HCDR3 with a length of 19 aa	0.01	1	0.01	0.95	0	0.95
% of HCDR3 with a length of 20 aa	0.02	24.42	0.02	20.65	0.02	21.22

Figure 2 - NGS Quality Control of Phage-display Libraries. DNA from electroporated TG1 cells was analyzed on Illumina MiSeq to control for HCDR3 randomization parameters and for correct HCDR3 length distribution.

Regarding the distribution of sizes, a slight bias occurred in the 11aa length for FW-κ, and in 16aa length in FW-κN1 and FW-κN2. Most importantly, the pools have the correct HCDR3 lenghts with approximate relative weights, and display residual distribution of unwanted sizes. This provides confidence that eventual differences in outcome will be explained by the intrinsic differences between frameworks.

4.3.2. Phage-display of test-libraries

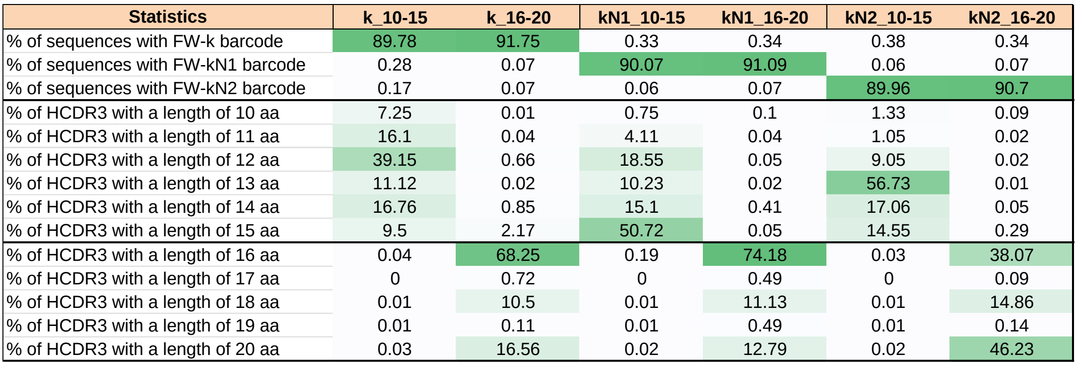
The six libraries were challenged against Herceptin with the goal of sucessfully selecting candidates against this commercial antibody used in breast cancer treatments. Three panning rounds of increase stringency were done following a solution-phase selection strategy[2], due to the constraints associated with traditional solid-phase panning (for more information see chapter 1, section 1.2.4). Solution-phase panning also provides a greater compatibility with automation proceedures when compared with solid-phase panning. More specifically, the KingFisher™ Flex Purification System will be employed, due to its capacity to sequentially operate plates filled with different buffers for washing, incubation and elution purposes, which increases the reproducibility of experiments (see section 4.5 for information on panning conditions). One additional benefit of this system is the background reduction of unspecific binders one compared to microtiter plate selections. This is achieved by the transfer of a minimal volume of magnetic particles from vessel to vessel.[3]. The phage-display was performed as detailed in section 4.5, and the output titers between rounds behaved as expected (Table 2) and were similar across conditions.

Table 17 - Output results from panning campaign. Phage-infected bacteria from each round were serially diluted into plates and counted the next day to calculate the respective round output.

ID	1st Round	2nd Round	3rd Round		
	Herceptin	Herceptin	Herceptin	Control	Mock
κ_10-15	2.25E+07	4.39E+06	2.58E+08	2.24E+04	1.24E+09
κ_16-20	1.48E+07	5.11E+06	2.92E+08	5.46E+04	1.15E+09
κN1_10-15	1.72E+07	3.85E+06	3.04E+08	1.12E+05	1.24E+09
κN1_16-20	1.79E+07	2.69E+06	1.84E+08	3.25E+05	8.72E+08
κN2_10-15	1.77E+07	1.13E+07	6.74E+07	1.25E+05	1.47E+09
κN2_16-20	1.82E+07	8.78E+06	8.03E+08	5.19E+05	1.06E+09

The first round had the lowest wash stringency and aimed to negatively select candidates that do not have affinity towards the target, and to bring the highest amount of positive candidates into the second round. In the second round the level of stringency of selection increased, which not only further removed unwanted sequences but also helped refining the positive binders pool and taking only the best binders into the third round. As a consequence, the second round is normally associated with a decrease in output when compared with the output from the first round (measured as the total number of phage-infected cells, more information on the methods section 4.5.). This was corroborated by the slight decrease from 1st Round to 2nd Round (Table 2). The third and most stringent round is tipically enriched in very good binders that can withstand very harsh selection conditions. Assuming that no overly-stringent selection protocols are employed, such binders cause the third round to be associated with increases in output. This was also observed in our experiment (Table 2).

Although the output results were similar, we observed marked differences between pools and frameworks after inspection of NGS results. Each framework has a distinct HCDR3 length distribution for each HCDR3 length pool. For the 10-15 pool, FW-κ is dominated by 12aa length, followed by 14aa and 11aa. On the other hand, FW-κN1 is dominated by 15aa followed by 12aa and 14aa. Finally, FW-κN2 is dominated by the 13aa length, followed by 14aa and 15aa. For the 16-20 pool, FW-κ and FW-κN1 have a very similar distribution of lengths, with a clear dominance of 16aa over the remaining ones. In contrast, FW-κN2 is dominated by the 20aa length, followed by 16aa and 18aa (Figure 3).

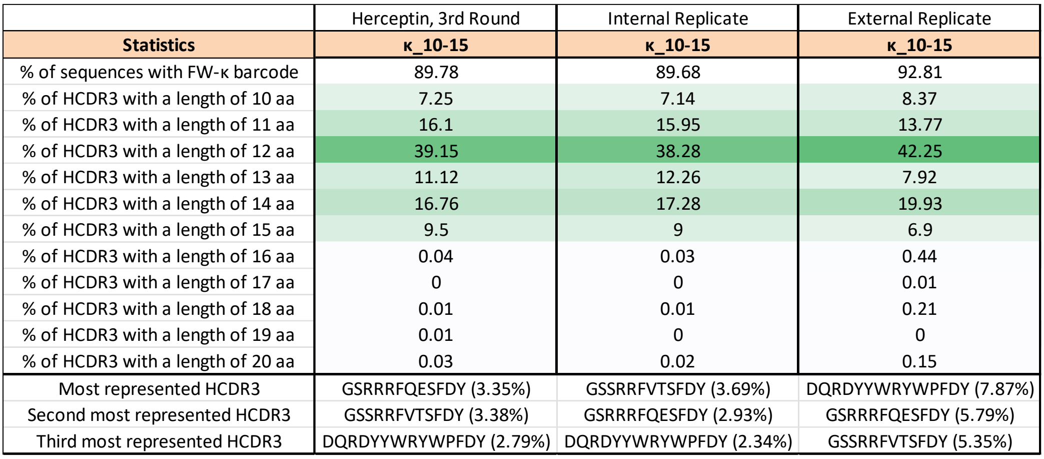


Statistics	k_10-15	k_16-20	kN1_10-15	kN1_16-20	kN2_10-15	kN2_16-20
% of sequences with FW-k barcode	89.78	91.75	0.33	0.34	0.38	0.34
% of sequences with FW-kN1 barcode	0.28	0.07	90.07	91.09	0.06	0.07
% of sequences with FW-kN2 barcode	0.17	0.07	0.06	0.07	89.96	90.7
% of HCDR3 with a length of 10 aa	7.25	0.01	0.75	0.1	1.33	0.09
% of HCDR3 with a length of 11 aa	16.1	0.04	4.11	0.04	1.05	0.02
% of HCDR3 with a length of 12 aa	39.15	0.66	18.55	0.05	9.05	0.02
% of HCDR3 with a length of 13 aa	11.12	0.02	10.23	0.02	56.73	0.01
% of HCDR3 with a length of 14 aa	16.76	0.85	15.1	0.41	17.06	0.05
% of HCDR3 with a length of 15 aa	9.5	2.17	50.72	0.05	14.55	0.29
% of HCDR3 with a length of 16 aa	0.04	68.25	0.19	74.18	0.03	38.07
% of HCDR3 with a length of 17 aa	0	0.72	0	0.49	0	0.09
% of HCDR3 with a length of 18 aa	0.01	10.5	0.01	11.13	0.01	14.86
% of HCDR3 with a length of 19 aa	0.01	0.11	0.01	0.49	0.01	0.14
% of HCDR3 with a length of 20 aa	0.03	16.56	0.02	12.79	0.02	46.23

Figure 3 – Quality Control of anti-Herceptin panning results.

4.3.2.1. Reproducibility

To make sure the differences between conditions was not due to the intrinsic variability of the panning process or related to other uncontrolled phenomena, internal and external replicates for the κ_10-15 library were also added to the experiment. The internal replicate refers to the repetition of the same condition in the same experiment and throughout three rounds of panning. The external replicate refers to the repetition of that same condition in a subsequent experiment that replicated washing conditions, also for three rounds of panning. These were analysed by NGS and compared with the experimental condition (Figure 4).

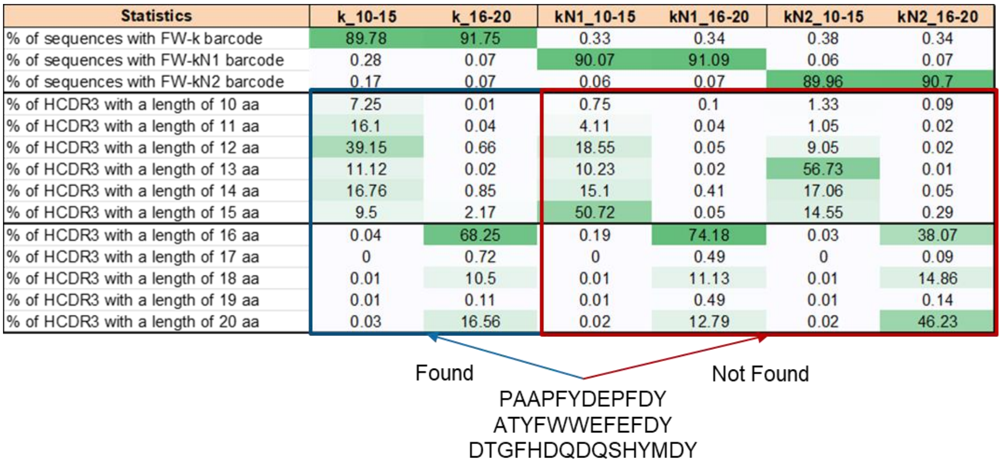


	Herceptin, 3rd Round	Internal Replicate	External Replicate
Statistics	**κ_10-15**	**κ_10-15**	**κ_10-15**
% of sequences with FW-κ barcode	89.78	89.68	92.81
% of HCDR3 with a length of 10 aa	7.25	7.14	8.37
% of HCDR3 with a length of 11 aa	16.1	15.95	13.77
% of HCDR3 with a length of 12 aa	39.15	38.28	42.25
% of HCDR3 with a length of 13 aa	11.12	12.26	7.92
% of HCDR3 with a length of 14 aa	16.76	17.28	19.93
% of HCDR3 with a length of 15 aa	9.5	9	6.9
% of HCDR3 with a length of 16 aa	0.04	0.03	0.44
% of HCDR3 with a length of 17 aa	0	0	0.01
% of HCDR3 with a length of 18 aa	0.01	0.01	0.21
% of HCDR3 with a length of 19 aa	0.01	0	0
% of HCDR3 with a length of 20 aa	0.03	0.02	0.15
Most represented HCDR3	GSRRRFQESFDY (3.35%)	GSSRRFVTSFDY (3.69%)	DQRDYYWRYWPFDY (7.87%)
Second most represented HCDR3	GSSRRFVTSFDY (3.38%)	GSRRRFQESFDY (2.93%)	GSRRRFQESFDY (5.79%)
Third most represented HCDR3	DQRDYYWRYWPFDY (2.79%)	DQRDYYWRYWPFDY (2.34%)	GSSRRFVTSFDY (5.35%)

Figure 4 – Results on the distribution (in %) of each HCDR3 length across replicates.

The internal control shows the same % of sequences with the correct barcode, the same distribution of HCDR3 lengths, and the same top three HCDR3 sequences as well as similar relative weight on the total sample. The similarity between the internal replicate and the experimental condition confirms that the variability between samples in the same panning campaign is not due to random variability or any other uncontrolled phenomena. Thus, any differences observed between conditions is considered a direct effect of measurable variables, such as differences in framework, HCDR3 lengths, randomization designs and wash stringency.

Additionally, the external replicate displays a very high level of similarity with the internal replicate and experimental condition for all the aforementioned statistics, which tell us that results of different panning campaigns can be compared without reservations. Considering that distinct panning campaigns against the same target has reproducible results when using the same libraries (Figure 4), we inspected whether we could find the anti-Herceptin HCDR3 sequences reported in Chapter 3. Even though these sequences were captured on a past panning campaign with very different selection protocols, we were able to find all of them in FW-κ samples (Figure 5). In light of the results described in Chapter 3 (anti-Herceptin candidates loss binding/affinity when grafted into CDR-null frameworks), it comes with no surprise that we could not find these same sequences in κN1 and κN2 datasets. Together with the marked difference in length distribution, this result serves as a testament to the impact that CDR-null mutations have on Vκ1-39/VH3-23 conformation dynamics and how it favours different HCDR3 sequences during selection protocols.



Statistics	k_10-15	k_16-20	kN1_10-15	kN1_16-20	kN2_10-15	kN2_16-20
% of sequences with FW-k barcode	89.78	91.75	0.33	0.34	0.38	0.34
% of sequences with FW-kN1 barcode	0.28	0.07	90.07	91.09	0.06	0.07
% of sequences with FW-kN2 barcode	0.17	0.07	0.06	0.07	89.96	90.7
% of HCDR3 with a length of 10 aa	7.25	0.01	0.75	0.1	1.33	0.09
% of HCDR3 with a length of 11 aa	16.1	0.04	4.11	0.04	1.05	0.02
% of HCDR3 with a length of 12 aa	39.15	0.66	18.55	0.05	9.05	0.02
% of HCDR3 with a length of 13 aa	11.12	0.02	10.23	0.02	56.73	0.01
% of HCDR3 with a length of 14 aa	16.76	0.85	15.1	0.41	17.06	0.05
% of HCDR3 with a length of 15 aa	9.5	2.17	50.72	0.05	14.55	0.29
% of HCDR3 with a length of 16 aa	0.04	68.25	0.19	74.18	0.03	38.07
% of HCDR3 with a length of 17 aa	0	0.72	0	0.49	0	0.09
% of HCDR3 with a length of 18 aa	0.01	10.5	0.01	11.13	0.01	14.86
% of HCDR3 with a length of 19 aa	0.01	0.11	0.01	0.49	0.01	0.14
% of HCDR3 with a length of 20 aa	0.03	16.56	0.02	12.79	0.02	46.23

Figure 5 – Anti-Herceptin candidates selected in a previous campaign can be found in FW-κ datasets, but not in FW-κN1 and FW-κN2 datasets.

4.3.2.2. HCDR3 sequence diversity

Using NGS to analyze the output of panning campaigns can be an invaluable tool to select the best candidates out of a very diverse dataset. Classical colony picking only provides a small snapshot of the final output of panning campaigns and can be biased to a handful of more dominant clones. On the other hand, NGS allows not only to inspect the clones that dominated the sample, but also to search for sequence motifs that may be determining antigen binding. Moreover, it allows to get a feel for how diverse our final dataset is, and how effective the panning process was in terms of selection.

The final objective of a panning campaign is undoubtedly to select the candidates that will follow to further characterization, such as K_D measurement and developability analysis. However, if one wants to evaluate the performance of a certain library or process effectiveness, an empirical methodology such as candidate testing constitutes a narrow view of the whole dataset. There is no telling whether the candidates that were not selected for production and testing would perform well or not, and if the user that was selecting those candidates was wrongly biased to certain sequence patterns. Thus, finding ways to retrieve unbiased information about the whole dataset can lay the foundations to compare the behavior of very different libraries across several panning protocols, without having to select candidates.

Here, we were interested on analyzing the diversity of each test library after three rounds of panning (Table 2). Such were measured by counting the number of HCDR3 unique sequences in a dataset, and most importantly, by clustering those sequences based on their similarity. Clustering provides an excellent measurement on the outcome of a panning campaign. The more clusters you get, the more diverse is your dataset in terms of beneficial “motifs” to antigen binding. Likewise, it is expected that increased wash stringency across rounds will reduce the number of clusters in a dataset, narrowing down the dataset to the best group of sequences.

Thus, we analyzed the outputs of the panning campaign with a tool that we recently implemented (see section 4.5 and Annex A) and that uses the density-based spatial clustering of applications with noise (DBSCAN) clustering method. DBSCAN was ran for all datasets and FW-κ libraries showed the greatest variability, with 364 different clusters found across all sizes, while FW-κN1 and FW-κN2 libraries had a total of 129 and 104 clusters, respectively. Outlier HCDR3 sequences represented 24% of the FW-κ dataset, and 11% and 24% of FW-κN1 and FW-κN2 (Figure 6).

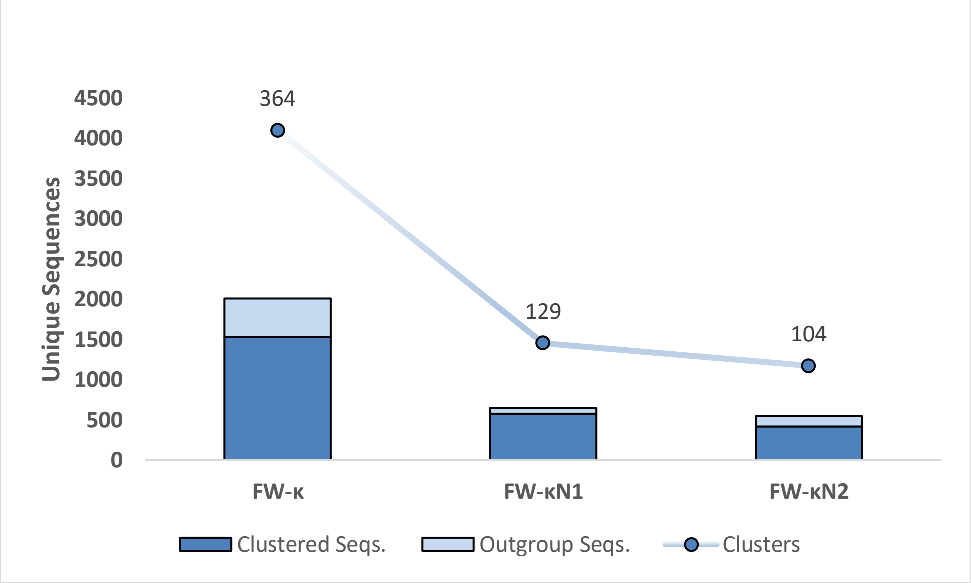


Figure 6 – Unique Sequences and clusters for FW-κ, FW-κN1, FW-κN2.

A size-by-size inspection reveals that the biggest difference in clustering between FW-κ datasets and the others two occurred at length 11aa and 12aa for the small sizes pool (HCDR3 lengths: 10-15aa, Figure 7), and in length 16aa for the bigger sizes pool (HCDR3 lengths 16-20aa, Figure 8). FW-κN1 and FW-κN2 libraries had a similar number of clusters throughout all sizes, with neither being clearly different from each other.

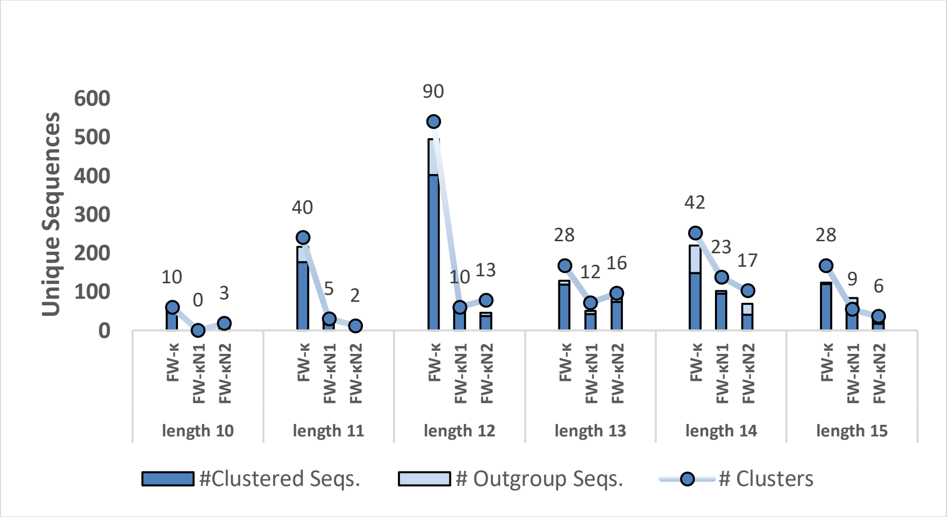


Figure 7 – Comparison of unique sequences and clusters per length, between frameworks, for the 10-15aa pools.

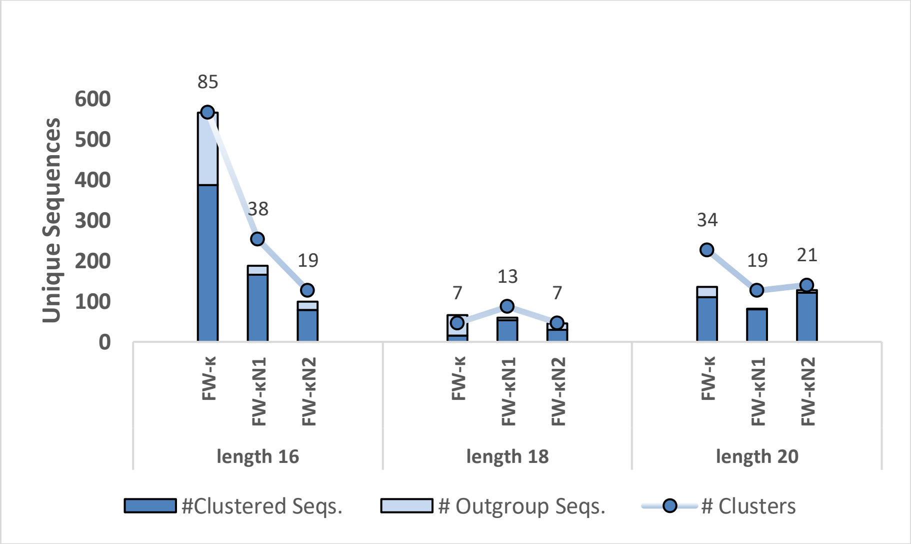


Figure 8 - Comparison of unique sequences and clusters per length, between frameworks, for the 16-20 pools.

To further inspect the behavior of FW-κ, FW-Kn1 and FW-Kn2 libraries under different conditions, a 4th round of panning was employed with different levels of stringency. Two washing conditions were tested with equal protocols but with different buffers. One used the standard PBST buffer, and the other used more stringent buffer, herein referred to as J-buffer (more info in the methods section 4.5). Expectedly, washes from the 4th round decreased the number of unique HCDR3 sequences throughout all datasets, and the number of clusters as well. Interestingly, FW-κN1 and FW-κN2 libraries seemed to resist more the J-buffer washing protocol (Figure 9).

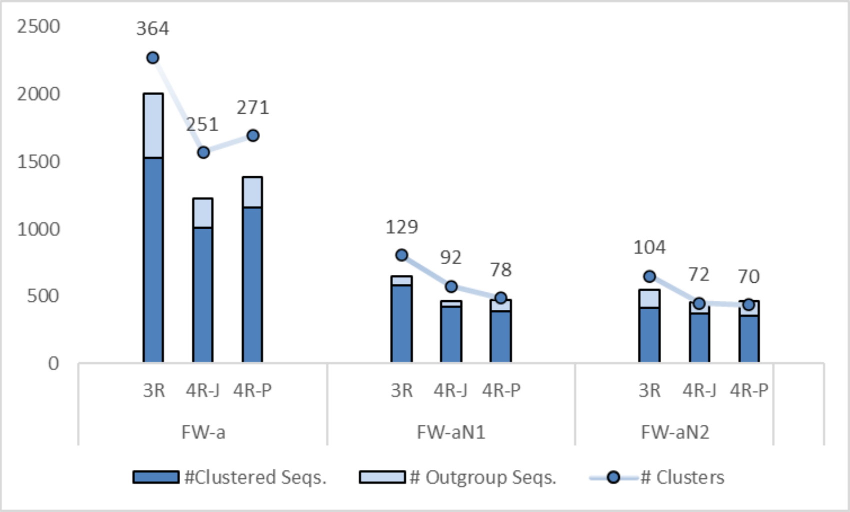


Figure 10 – Effect of 4th Round on unique HCDR3 sequences and clustering.

A size-by-size inspection reveals that FW-κN1 and FW-κN2 libraries seem to suffer more on the small sizes, with next to none amounts of unique HCDR3 sequences, and very few clusters (Figure 10). This goes in line with the results found when the output of 3rd round was analysis (see section 4.3.2., Figure 3). However, they seem to fare better on length 18aa and 20aa, where they slightly outperform FW-κ. The inspection after a 4th round of panning also reveals that even though sometimes different libraries yield the same number of unique sequences, the similarity of those sequences pool can be completely different. A clear example for that are the results of length 15aa where 80 unique sequences are grouped in ~20 different clusters for FW-κ *versus* ~6 different clusters for FW-κN1 (Figure 10).

A major drawback from these datasets is concerning the pooling of HCDR3 lengths. There is no telling what the outcome in terms of sequence diversity would be if each size had the chance to bind to the target antigen without the competition of other sizes. For example, it is likely that FW-κ library did not yield many HCDR3-18aa clusters because of the dominance of HCDR3-16aa sequences (Figure 3).

The more valuable information still comes from the behavior of the frameworks as a whole, as elicited in Figure 6 and Figure 9. In both those cases, FW-κ consistently yielded a more diverse dataset, with more unique HCDR3 sequences and more different clusters, even after stringent washing.

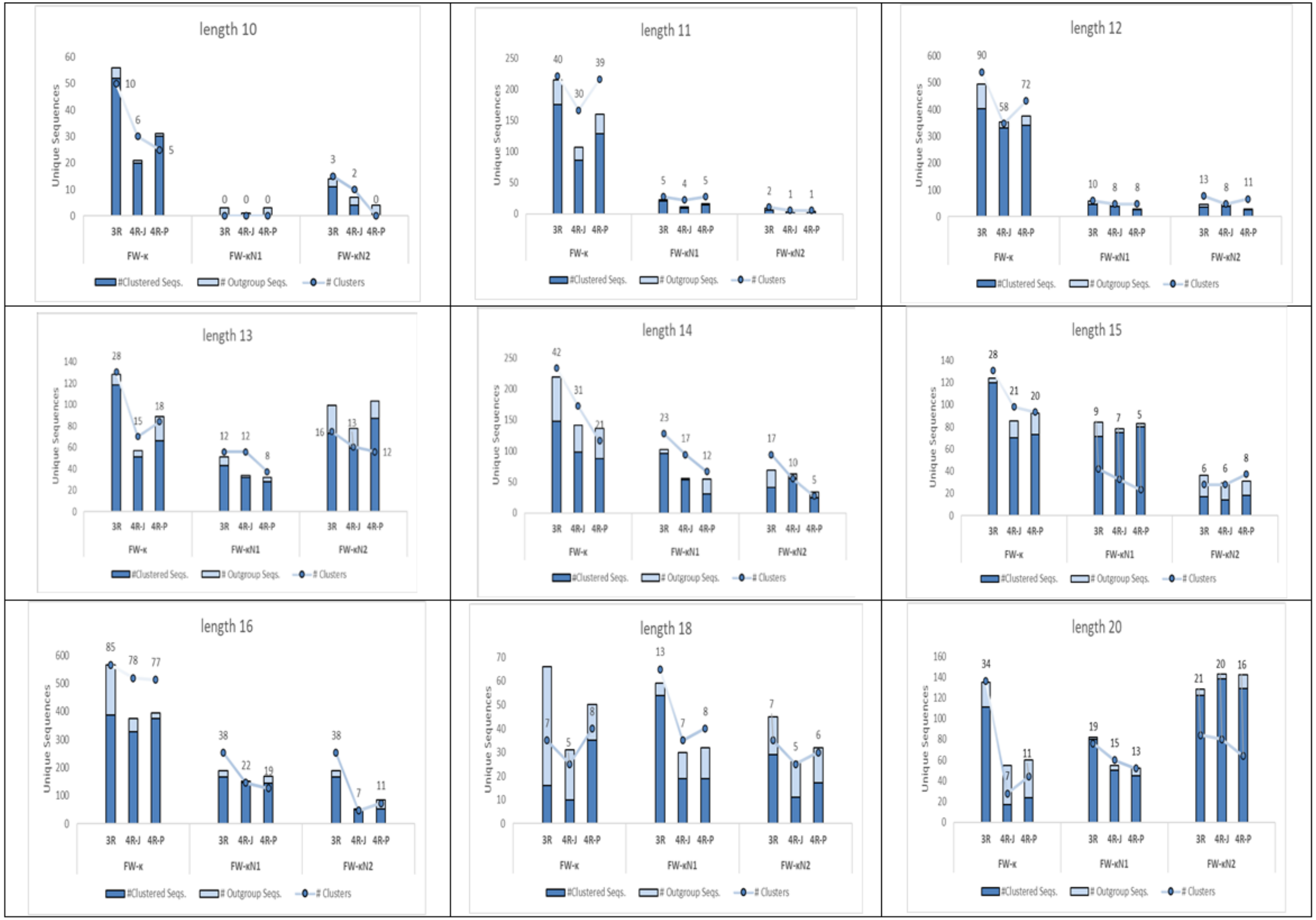


Figure 11 - Effect of 4th round on unique HCDR3 sequences and clusters per length, between frameworks.

4.3.3. Candidate selection and characterization.

Each panning condition was inspected with the aim of selecting the best candidates possible. In principle, binders of higher affinity would have had a bigger chance of infecting bacteria after the selection protocols, systematically increasing their relative weight in the sample after each round. As such, sequences with highest number of occurrences in the antigen and with good enrichments over the mock conditions (which represents the previous round) were prioritized (Table 3). These were expressed as IgGs in HEK293T cells. Upon production, 9 candidates did not express in HEK293T cells. Of those, 8 were within the top 10 clones of their respective condition. This effect was seen across all six libraries.

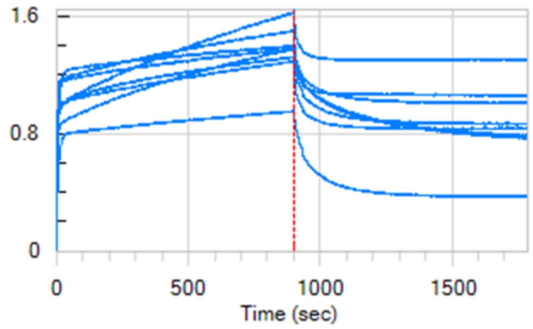
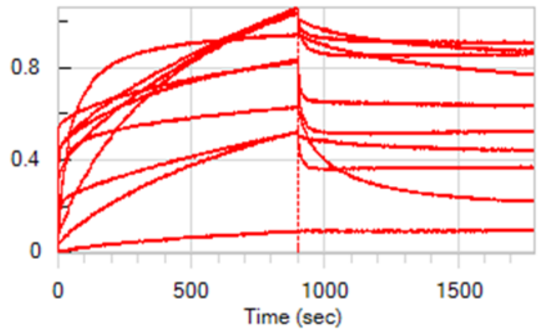
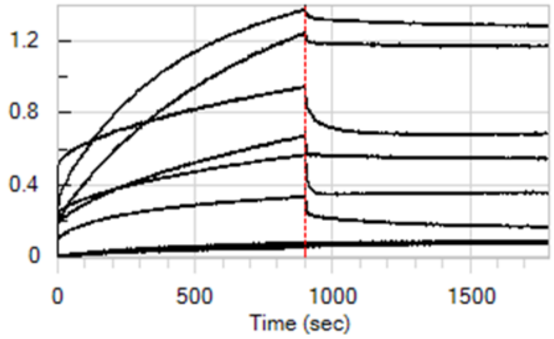
4.3.3.1. Binding Kinetics of Selected Candidates

Successfully produced IgGs were screened by Bio-layer Interferometry on OctetRED96 for their ability to bind to Herceptin®, by a newly devised IgG-IgG interaction protocol devised at the lab (Table 5, for more information see section 4.5)

Table *3* – List of selected candidates for each framework. Thirty-six candidates were chosen, six for each pool, for each framework Sequences from all lengths were chosen whenever possible.

Sequence ID				ALC
FW	length	HCDR3	**Mock Enrichment**	Total mAb (in mg)
FW-κ	10	GQDWEPEFDY	7.18	3.0
	11	QLELFEPELDY	9.37	2.5
	11	SAQYWEPEFDY	9.07	2.8
	12	GSSRRFVTSFDY	15.58	failed
	14	GKFRDWAPEKAFDY	9.81	2.9
	15	AAGWLDTDEGRTMDY	8.81	3.1
	16	DGSGSFLPVEDVSFDY	3.66	3.0
	16	KRGPYYYSFWPYGFDY	7.46	failed
	16	GQWPFAHPEAGLDFDY	3.23	3.0
	18	QQPSSWAGPKYAYHGFDV	2.42	failed
	20	ERPWWGIFSFGYQEEVGMDV	3.76	failed
	20	DRQRVLDLDTYEWAEEYFDV	2.71	3.3
FW-κN1	12	WELRGSPWPFDY	11.89	failed
	12	AQSPFDWADFDY	6.88	2.9
	13	AQGDYLPDDAFDY	5.93	2.4
	13	EGSYKHAEEAFDY	9.5	2.8
	14	DGGPYVQFPEAFDY	10.2	3.0
	15	DDSYQDYYDQGGFDY	16.58	2.5
	16	SPVPWSPYGDDLSFDY	N/A	3.6
	16	HSHLYLEPYWRWRFDY	19.75	failed
	16	DRWGGWDHAAEYLFDY	15.93	4.2
	16	DTDVLTYSFGDYSFDY	8.45	4.3
	18	DKEGDGYDYVTYAGYFDV	10.65	4.3
	20	WADGGAPDYYPQEYELGFDV	6.36	5.0
FW-κN2	12	WEYGPSPYPFDY	25.07	failed
	13	TYGDYYSLESMDY	39.36	3.6
	13	EYGDPYDSYGFDY	29.67	1.3
	13	SQDTYFDDQYFDY	32.67	3.9
	14	GPWHYYPTRGAFDY	62.71	4.8
	15	TTHDYEDWLVSVFDY	25.35	4.0
	16	GYRYARWESSRWRFDY	15.9	3.5
	16	TSSWGHFVDDIEHFDY	43.64	4.4
	18	VAIYAYDHFQDHAAVFDV	12.12	4.1
	20	DSTAWRKGVGGRYYYWAFDV	32.96	failed
	20	THWPHLGGLEYFTYYPYMDV	N/A	4.6
	20	YDSWLGKWRGYYYRYDGFDV	29.64	failed

Table 4 – List of anti-Herceptin candidates, and their overlapped BLI sensograms after affinity screening.

FW	length	HCDR3	bind?	
FW-κ	10	GQDWEPEFDY	Yes	
	11	QLELFEPELDY	Yes	
	11	SAQYWEPEFDY	Yes	
	14	GKFRDWAPEKAFDY	Yes	
	15	AAGWLDTDEGRTMDY	Yes	
	16	DGSGSFLPVEDVSFDY	Yes	
	16	GQWPFAHPEAGLDFDY	Yes	
	20	DRQRVLDLDTYEWAEEYFDV	Yes	
FW-κN1	12	AQSPFDWADFDY	Yes	
	13	AQGDYLPDDAFDY	Yes	
	13	EGSYKHAEEAFDY	Yes	
	14	DGGPYVQFPEAFDY	Yes	
	15	DDSYQDYYDQGGFDY	NO	
	16	SPVPWSPYGDDLSFDY	Yes	
	16	DRWGGWDHAAEYLFDY	Yes	
	16	DTDVLTYSFGDYSFDY	Yes	
	18	DKEGDGYDYVTYAGYFDV	Yes	
	20	WADGGAPDYYPQEYELGFDV	Yes	
FW-κN2	13	TYGDYYSLESMDY	Yes	
	13	EYGDPYDSYGFDY	Yes	
	13	SQDTYFDDQYFDY	NO	
	14	GPWHYYPTRGAFDY	NO	
	15	TTHDYEDWLVSVFDY	NO	
	16	GYRYARWESSRWRFDY	Yes	
	16	TSSWGHFVDDIEHFDY	Yes	
	18	VAIYAYDHFQDHAAVFDV	Yes	
	20	THWPHLGGLEYFTYYPYMDV	Yes	

The affinity screening assays reveal a big variety in the affinity kinetics of each clone. Some display an on-rate that saturates very fast, and after that, a second on-rate slowly builds up to end of the association step. Such is very noticeable in FW-κ binders (Table 4). Others, such as most of the FW-κN2 binders, have slower on-rates overall. Likewise, the off-rates also differ across binders. Some keep bound to the antigen throughout the dissociation phase with very slow off-rate kinetics, which is very notorious in many FW-κN2 binders. Inversely, some binders dissociate very fast upon the start the dissociation phase (Table 5). After the screening step, K_D determination assays were done in the same instrument (see methods in section 4.5.). This assay revealed that FW-κ, FW-κN1 and FW-κN2 libraries originate binders with a roughly similar distribution of binding that ranges, which averages around 68 nM (Figure 11). More specifically, FW-κ and FW-κN1 have an average K_D of 52 nM and 95 nM respectively, while FW-κN2 averages at 49 nM. FW-κN2 binder GYRYARWESSRWRFDY fared worse than the other binders and has a 254 nM affinity towards Herceptin (not shown in Figure 11).

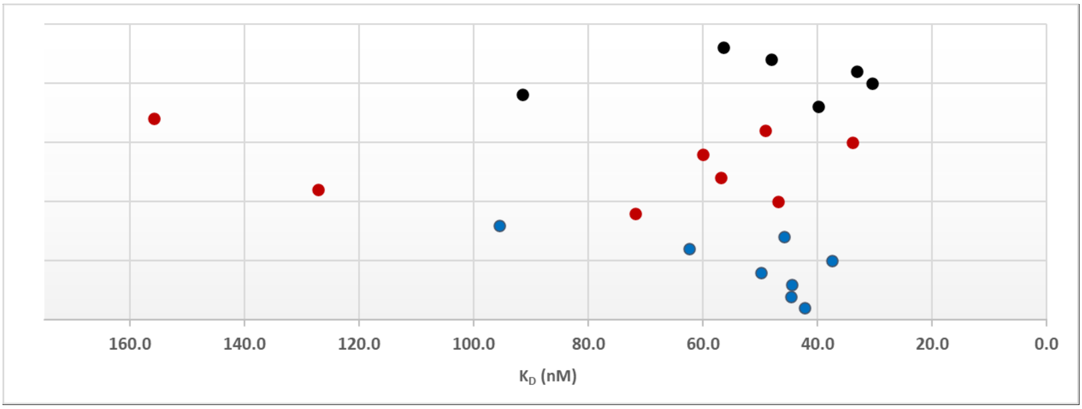


Figure 11 - Affinity values distribution of FW-κ, FW-κN1 and FW-κN2 anti-Herceptin binders: Blue: FW-κ; Red: FW-κN1; Black: FW-κN2.

4.3.3.2. Developability of Selected Candidates

Finally, all candidates were analyzed for their biophysical properties. Except for two FW-κN2 binders, the remaining binders are above the minimum requirements of aggregation and hydrophobicity and no framework-dependent behaviors were observed in that regard (Table 5).

However, the most distinctive trait between frameworks is still the thermal stability of the candidates they originate. As mentioned in Chapter 2, de-stabilization of the Fab domain due to mutations and/or unfavorable HCDR3 sequences can shift the Tm2 values. While FW-κ candidates have consistent Tm2 above 76°C and often 10 degrees above Tm1, FW-κN1 and FW-κN2 candidates frequently display an overlapping of Tm1 with Tm2, which is indicative of instability in the Fab domain.[6–9]

Table *5* - List of produced candidates for each framework and their biophysical characteristics. SEC: size-exclusion chromatography; DSF: Differential Scanning Fluorometry; HIC: Hydrophobic Interaction Chromatography. High-molecular weight species (HMWS) indicative of aggregation above 5% are indicated in yellow. Main Peak (MP) values indicative of monomeric IgG below 95% are indicated in yellow. An Ammonium Sulfate (AS) concentration below 0.8 M is also indicated in yellow.

Sequence ID			Octet	SEC		DSF		HIC
			KD (nM)	HMWS (%)	MP (%)	Tm1 (°C)	Tm2 (°C)	[AS] (M)
FW-κ	10	GQDWEPEFDY	42.3	2.3	97.7	68	76	0.92
	11	QLELFEPELDY	44.7	1.6	98.3	68	76	0.85
	11	SAQYWEPEFDY	44.5	1.5	98.3	68	78	1.02
	14	GKFRDWAPEKAFDY	49.9	2.2	97.8	68	81	1.05
	15	AAGWLDTDEGRTMDY	37.5	2.1	97.9	68	84	1.03
	16	DGSGSFLPVEDVSFDY	62.5	1.6	98.3	68	80	1.03
	16	GQWPFAHPEAGLDFDY	45.8	2.5	97.5	68	79	0.89
	20	DRQRVLDLDTYEWAEEYFDV	95.6	1.8	98.2	68	79	0.97
FW-κN1	12	AQSPFDWADFDY	71.8	3.4	96.6	68		0.87
	13	AQGDYLPDDAFDY	46.9	1.3	98.6	68	76	1.05
	13	EGSYKHAEEAFDY	127.2	2.3	97.6	68	76	1.09
	14	DGGPYVQFPEAFDY	56.9	2.5	97.5	68		0.97
	16	SPVPWSPYGDDLSFDY	254.3	1.8	98.2	68	76	0.95
	16	DRWGGWDHAAEYLFDY	60.0	1.8	98.2	69		0.96
	16	DTDVLTYSFGDYSFDY	34.0	1.1	98.9	69		0.99
	18	DKEGDGYDYVTYAGYFDV	49.2	2.0	98.0	68	80	1.00
	20	WADGGAPDYYPQEYELGFDV	155.9	1.5	98.5	68		0.9
FW-κN2	13	TYGDYYSLESMDY	39.9	1.5	98.5	69		1.03
	13	EYGDPYDSYGFDY	91.5	3.7	96.2	68		0.97
	16	GYRYARWESSRWRFDY	30.5	0.1	99.8	68		1.07
	16	TSSWGHFVDDIEHFDY	33.1	1.7	98.3	69		1.02
	18	VAIYAYDHFQDHAAVFDV	48.0	9.4	90.5	68	78	0.86
	20	THWPHLGGLEYFTYYPYMDV	56.4	4.3	95.7	68	79	0.71

4.4. Discussion

As hinted by chapter 2 and 3 results, the side-by-side panning reveals distinct behaviors between the FW-κ and the CDR null-frameworks libraries (FW-κN1 and FW-κN2). All frameworks showed different bias towards different HCDR3 lengths after 3 rounds of panning against Herceptin, which confirms expectations that CDR-null mutations would affect the HCDR3 sequences selected from panning campaigns.

As stated before, even though HCDR3 tends to contribute the most to antigen binding, all the other five CDR can contribute to antigen binding, depending on the Ab-Ag complex. In 2016 a study on 138 Ab-Ag complexes showed that natural Abs display a higher diversity of paratopes (and in turn allow the libraries to bind to a greater panel of epitopes) when compared with synthetic Abs, which rely more on HCDR3.[10] Moreover, it is also known that germline antibodies retain a degree of structural plasticity in their backbone in order to bind a number of different unrelated antigens, a capacity referred as polyspecificty or polyreactivty. Such conformation flexibility allows the relatively small number of possible germline combinations to adjust to several epitopes.[11,12] It has been reported that polyspecific antibodies often retain a larger proportion of germline gene sequences than more specific antibodies[12,13] . Hence, it comes to no surprise that deviations from the germline ends up restricting the diversity of the dataset after 3 rounds of panning (Figure 6).

On a primary library context, such polyspecificity may be useful to broaden the amount of antigen epitopes a library can bind to. However, it means antibodies coming from such germline-dependent libraries may require several somatic mutations to increase specificity using affinity maturation methods. In line with this conformational flexibility hypothesis, it has been shown that as sequences mature and deviate from germline, the rigidity of their paratopes also increases.[14] Additionally, the HCDR3 length is also found to affect the nature of the binding of other CDRs. On antibodies with longer HCDR3 loops, the HCDR3 is responsible for most of the antibody-antigen interactions, while in antibodies with shorter HCDR3 loops, the remaining CDRs usually assist in antigen binding.[15] This may serve as an explanation to why FW-κN1 and FW-κN2 yield fewer shorter length HCDR3 (Figure 7), and why they seem to match FW-κ diversity in the HCDR3-18aa and HCDR3-20aa lengths (Figures 8 and 10).

Even though the total diversity of the dataset drops when using primed libraries κN1 and κN2 in comparation with the control library (FW-κ), it was possible to select anti-Herceptin binders from all CDR-null test libraries, and the average affinity of binders does not differ too much between all the libraries tested. As shown throughout the chapter, different HCDR3 lengths were selected when subjected to the same washing conditions and the amino acid sequences in the datasets were manifestly different. This was easily seen when manually scanning through the data (not shown) and more systematically by looking at the family motifs generated by our recently developed methodology (see Annex A for more detail).

Additionally, with 129 and 104 different clusters to choose from at the end of the 3rd round for FW-κN1 and FW-κN2 libraries, these frameworks are more than able to provide enough options to proceed to candidate production and testing. Producing and testing one representative clone out of each κN1 and κN2 families would surpass the number of clones usually selected for a panning campaign with multiple framework pools and branches, which sits on the mid-dozens. Thus, CDR null-frameworks provide sufficient clonal diversity to follow up for candidate production. Arguably, the decrease in germline residues and the consequent increase in rigidity may help reduce polyspecificity and put CDR-null framework antibodies one step closer to final matured sequences. In addition, the decrease in conformation flexibility may help to narrow down the choice to HCDR3 sequences that have less tendency to bind to other targets, and thus increase the probability of finding an optimal candidate.

Even considering the encouragingly results about the performance of the primed libraries κN1 and κN2, it is important to highlight that CDR-null mutations they bear also clearly destabilize thermal stability, as shown by the greater number of low Tm2 candidates (Table 6, DSF data). As such, we anticipate that primary pannings using primed libraries might need to have some form of thermal challenge to discard unstable binders that may be contaminating the sample. Finally, it is worth noting that CDRs with very similar structures can have very different sequences, and that loops with similar sequences can adopt very different conformations.[16] Therefore, since the current approach only takes in account sequence diversity, it is possible that structurally dissimilar antibodies are being clustered together. Thus, it is not possible to evaluate the structural diversity of the given datasets, and to take definitive conclusions about the potential diversity of paratopes of each library.

4.5. Materials and Methods

Phage-display vectors

To generate phage particles displaying antibody fragments, a phagemid vector was used. The phagemid is based on the M13 filamentous phage, and it encodes a Fab antibody fragment fused via an Amber stop codon (UAG) to a truncated pIII protein (Glycine-rich linker and CT domain, which anchors pIII in the phage coat). The phagemid also carries an ampicillin resistance gene and a M13 origin gene that triggers the packaging signal of the filamentous phage when combined with VSCM13 helper phage within infected bacteria.

Primary Library Preparation

In the primary libraries used in this study, only the HCDR3 is randomized. The other CDRs are from germline origin. FW-κ uses VH3-23 and Vκ1-39 germline sequences. FW-κN1 and FW-κN2 use the same sequences as FW-k but bearing CDR-null mutations. A dummy (or "background") HCDR3 sequence WGGDGFYAMDY is present on all Fab fragments before randomization. HCDR3 randomization is achieved by cloning a mixture of oligonucleotides generated by TRIM (trinucleotide-directed mutagenesis) technology into the HCDR3 region of the Fab fragment. The TRIM technology relies on synthesis of DNA from pre-assembled trinucleotides (trimers). Of the 64 possible combinations of codons, only 20 codons are required to cover the 20 amino acids. Using TRIM, any mixture of amino acids can be adjusted at will at each position of the HCDR3. Frameshifts, stop codons or undesired amino acids can be completely avoided guaranteeing the synthesis of high-quality libraries. The oligos encoding randomized-HCDR3 sequences were ordered from EllaBiotech and mimic the natural amino acid distribution found in human antibodies. Nine different HCDR3 lengths were ordered: 10aa, 11aa, 12aa, 13aa, 14aa, 15aa, 16aa, 18aa, 20aa. These were cloned in equimolar proportions into the HCDR3 region of Fab fragments using unique restriction sites (BssHII and BlpI). In order to achieve a reasonable library size, a total amount of 1 µg of each vector is mixed with at least a 10-fold molar excess of inserts. In this case, it is not a single insert size, but two pools of inserts will be used, one with the sizes of 10-15 aa and another with the sizes of 16-20 aa. The total ligation reaction is prepared in one reaction with a final volume of 160 µL. Ligation ensued overnight, and a de-salting protocol was done to purify the ligated phagemid DNA. The phagemid vectors were transformed via electroporation into electrocompetent *E.coli* TG1 cells (Lucigen). Electroporated cells were recovered in SOC medium for 1 hour before being transferred into 2YT medium with 1% glucose and 100 µg.mL-1 of ampicillin (2YT/A/G) and incubated overnight at 25 °C, 200 rpm on incubator Innova 44. Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol.

Phage production

A sample from each glycerol stock from the libraries was taken to start a 25 mL culture in 2YT/A/G at OD_{600} = 0.1 and grown to mid-log phase (OD600 = 0.5) before being infected with helper phage VCSM13 (Agilent Technologies). Cells were then incubated firstly for 30 min at 37°C in a water bath, and then for 30 min at 37°C shaking at 250 rpm. The infected bacteria were then centrifuged and transferred into a 40 mL culture of 2YT supplemented with 100 µg/mL Ampicillin, 50 µg/mL Kanamycin and 0.25 mM IPTG**.** Phage production ensued overnight at 22°C and 180 rpm. The cultures were centrifuged to remove the cells and the phage-rich supernatant collected into sterile 50 mL Falcon tubes are kept on ice. Phages are precipitated by adding 10 mL of ice cold 20% (w/v) PEG 6K in 2.5 M NaCl into the 40 mL of supernatant and left for 1 hour on ice. After this time, the precipitated solutions were centrifuged at 4000 *g* and 4 ˚C for 30 min (Eppendorf, Ref: 5810 R). The supernatant was discarded, and the precipitated phage pellets were re-suspended with 1 mL of sterile phosphate buffered saline (PBS) and transferred to 1.5 mL Eppendorf tubes. The tubes were then rotated for 30 min on a rotating wheel at 4˚C and then centrifuged at 12 000 *g* and 10 ˚C for 5 min (Eppendorf, Ref: 5810 R) to remove further bacterial debris. Supernatants were filtered into cryovials containing 700 uL of PBS:Glycerol 50:50% (for a final [Glycerol] of 20% v/v).

Herceptin Biotinylation

1mg of biotin was dissolved in 166 µL H20 to do a 10 mM solution. Then, 40.5 uL of Sulfo-NHS-SS-Biotin was used to biotinylate 0.5 mL of 10 mg/mL Herceptin. The procedure is done following the directions of EZ-Link Sulfo-NHS-LC-Biotin kit (ThermoScientific A39257 21335). Samples were incubated 1 hour at room temperature. The samples were then passed through a Zeba Spin desalting column, 7K MWCO, 2mL (Thermoscientific, #89891, #QL227761). The biotinylated Herceptin is kept in PBS pH 7.0, quantified using nanodrop and stored at 4°C.

Phage display panning selections

Phage display protocols were performed using the automated liquid handling functionalities of the KingFisher™ Flex Purification System (ThermoFisher, Catalog number: 5400610). A total of 5 x 10^9 infectious phages corresponding to each primary library were blocked for 1 h in PBS + 0.05% Tween (PBST) supplemented with 0.05% of BSA, in 96 DeepWell plates (Thermo Scientific™ 95040450), followed by in-solution deselection on streptavidin-coated magnetic beads (Dynabeads, Invitrogen, Cat # 112–06) for 30 min, to remove sticky phages that bind to streptavidin beads. Biotinylated Herceptin was added in the corresponding well to each well of sticky-depleted phages and incubated 1h at room temperature (RT) on a micro-plate table. The antigen-antibody complexes were captured from the deep well plates by the streptavidin-coated magnetic beads bound to the KingFisher magnetic rods and transferred to the washing plates sequentially, as shown in Figure 12.

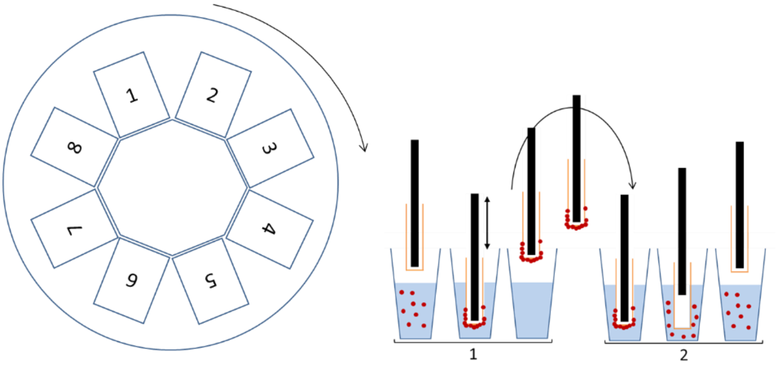


Figure 12 – Schematic representation of kingfisher operation. Streptavidin magnetic-beads bound are put in contact with the phage-antigen-biotin complexes in solution, which are then transferred between wells by plastic-covered rod-shaped magnets. The capture and release movements during transfer and washing protocols are software-driven, and all the parameters such as time, position, and frequency shaking movements can be customized**. Adapted from:** Ch'ng, A.C.W., Ahmad, A., Konthur, Z., and Lim, T.S. (2019). A High-Throughput Magnetic Nanoparticle-Based Semi-Automated Antibody Phage Display Biopanning. In Human Monoclonal Antibodies, M. Steinitz, ed. (New York, NY: Springer New York), pp. 377–400

The washing of bead-antigen-phage complexes was accomplished by washes of increasing shaking vigor, stringency, and duration, on PBST and PBS. Herceptin concentration decreased from round to round to increase selective pressure (Figure 13). The 4th round of panning was separated into two different branches, a more stringent one, and a less stringent one (Figure 14).

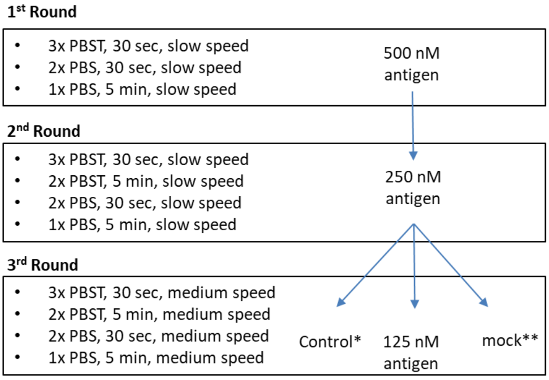

1st Round
3x PBST, 30 sec, slow speed
2x PBS, 30 sec, slow speed
1x PBS, 5 min, slow speed
500 nM antigen
2nd Round
3x PBST, 30 sec, slow speed
2x PBST, 5 min, slow speed
2x PBS, 30 sec, slow speed
1x PBS, 5 min, slow speed
250 nM antigen
3rd Round
3x PBST, 30 sec, medium speed
2x PBST, 5 min, medium speed
2x PBS, 30 sec, medium speed
1x PBS, 5 min, medium speed
Control*
125 nM antigen
mock**

Figure 13 - Three rounds of panning and respective washing procedures. The phages coming from the second round are split into three distinct conditions: The antigen condition, the control condition, and the mock condition. The control condition replicates the current washing protocol but replaces the antigen with any other molecule against which we do not want to retrieve phages. This is essential to discriminate between sticky/unspecific binders and highly specific binders. The mock condition does not go through the washing protocol but rather consists in taking the phages from the second round and using them to directly infect bacteria. Comparing the antigen condition with this one allows for a measure of the enrichment of phages from the 2nd round to the 3rd round.

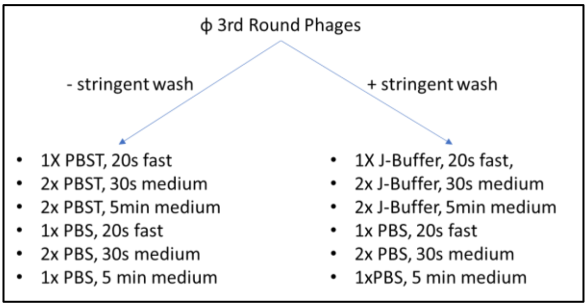


Figure 14 – Washing protocol for the 4th round. Antigen concentration was kept at 125 nM.

Bacterial infection and phage amplification

At the end of each wash protocol of each round, surviving phages were dissociated from the complexes with glycine buffer (10 mM glycine-HCl, pH 2.0) before neutralization with 200 µL Tris-HCl pH 7.5 and infection of a 20 mL mid-log *E.coli* TG1 culture (OD600 = 0.5). The cultures were incubated for 45 min in a water bath at 37ºC before being inoculated into 100 mL 2YT/A/G in 250 ml Erlenmeyer's and let to grow overnight at 25ºC, 150 rpm (Innova 44R, New Brunswick Scientific). Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol. These can be used to produce new phages for subsequent rounds, or to have their DNA extracted for NGS analysis.

DNA preparation and NGS analysis

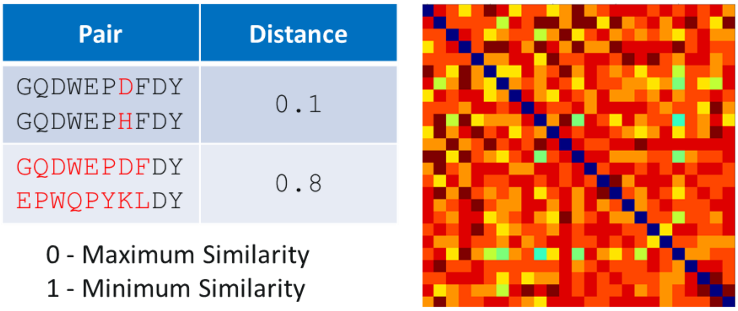
Plasmid DNA was isolated directly from the phage-infected cells from the selection round of interest using the GeneJET Plasmid Miniprep Kit (Thermo Scientific™, K0502). Isolated dsDNA was quantified on the Qubit 3.0 fluorometer using the Qubit® dsDNA HS kit (Invitrogen™ Q32851). The amplicon for HCDR3 sequencing was generated through two PCRs. To amplify the region of interest and to insert the adapter regions for the NGS, the initial PCR utilized a forward primer specific to the vector leader sequence prior to HCDR3 and a reverse primer downstream of HCDR3, near the end of the V_H region where the library barcode and randomization barcode are located. The second PCR inserted the TruSeq universal adapter and the indexes, used to distinguish between different samples (i.e. libraries). Samples were quantified in Qubit 3.0, pooled in equimolar proportions, and ran on an electrophoresis gel. Bands with the appropriate size were excised, purified using the Wizard SV Gel and PCR Clean Up System (Promega, A9281), and quantified on Qubit 3.0. The pool was diluted to a final concentration of 4 nM, spiked with 20% PhiX (Illumina; FC-110-3001), denatured for 5 min in 0.1 N of NaOH (5 µL of DNA+PhiX at 4 nM mixed with 5 µL 0.2 N of NaOH), diluted in HT buffer (provided on the NGS kit; kit details, ahead) to 9 pM and sequenced on the Illumina MiSeq platform using the 150 cycle V3 kit (Illumina; MS-102-2003). The forward read was 75 bp in length while the reverse read was 75 bp. The data analysis of the NGS FastQ output files was performed as described previously.[3] For the panning output of each library, 1 x 10^5 sequences were analyzed using the fixed flanking sequences on the boundary of HCDR3 as template to locate and segment out HCDR3 sequences.

Candidate Selection from NGS datasets

When selecting candidates from NGS datasets, data from three different types of experimental branches will be analyzed: i) Antigen branch: The antigen condition represents the molecule of interest. Candidates that have many occurrences on the antigen dataset are likely good binders. Any experiment may have any given number of antigen branches per library; ii) Control/Counter: In most cases antibodies will not be of interest if they cross-react with certain targets. Sometimes the counter can also be a similar molecule to the antigen, as a way to specifically select antibodies against the domains that differ between the antigen and the counter – e.g. when looking for a candidate against a specific IgG antigen, the counter of choice is usually another unrelated IgG that shares some of the FW regions. Candidates that have high occurrences in the antigen and counter datasets simultaneously are most likely binding to the FC-region or to the shared FW regions and should be discarded. Hence, calculating the ratio of Antigen/Counter provides a critical measurement of a given candidate's cross-reactivity. Any experiment may have any given number of counter branches per library; iii) Mock: The mock represents phages that only went through cycles of production and infection but have not been challenged against the antigen. This provides the baseline values from which candidate enrichment from one round to the other will be calculated from (Antigen/Mock). The mock also provides a good estimation whether some candidates cross-react (Counter/Mock). Only one mock branch per library is used.

Systematic Clustering of HCDR3 sequences using iSERA

iSeRa (interactive sequence ranker) aims to facilitate the selection of candidates from NGS datasets. Firstly, it organizes the dataset in clusters of similar sequences (also referred to as “families”). Secondly, it samples up to two candidates from each family, if they pass a set of selection criteria. Finally, the candidates are ranked according to multiple parameters. In this work, the iSeRa’s systematic clustering method was leveraged to analyze the diversity of NGS datasets in a reproducible and un-biased way. To cluster sequences within a NGS dataset, a similarity score is firstly assigned between each pair of sequences by applying the Smith-Waterman algorithm and a similarity matrix is created. Sequences of different framework and lengths are always clustered separately.



Pair	Distance
GQDWEPDFDY GQDWEPHFDY	0.1
GQDWEPDFDY EPWQPYKLDY	0.8

Figure 15 – Example of sequence pair similarity and final matrix of distances. Smith-Waterman Algorithm is used to compare the distance between each pair of sequences as strings.

A subsequent clustering step is then performed based on the calculated similarity scores, as a way of identifying consensus within the dataset and organizing the candidates. The **D**ensity-**b**ased **s**patial **cl**ustering of **a**pplications with **n**oise (DBSCAN) clustering algorithm is employed at this stage. In contrast with other clustering methods such as k-means[4], DBSCAN does not require the number of clusters to be defined *a priori* to operate. Such allows for an unbiased analysis of broad NGS datasets that come from panning campaigns. Additionally, DBSCAN also allows for outlier sequences (i.e. dissimilar sequences that do not fit into any cluster) into outgroups, which can also serve as a measure of the variability of the dataset.[5] The DBSCAN's clustering method is governed by two variables: min_points, which tells how many datapoints are required to form a cluster, and epsilon (**ε**), which tells how close (i.e. similar) two points must be from each other to belong to the same cluster.

To define these two variables, an iterative approach was employed. The min_points variable was fixed to 2 for all datasets since there is no logical explanation to impede a cluster from being generated if no more than similar two sequences are found. On the other hand, **ε** was iteratively defined by running the algorithm with multiple **ε** values and choosing the one that maximizes the number of clusters (Figure 16).

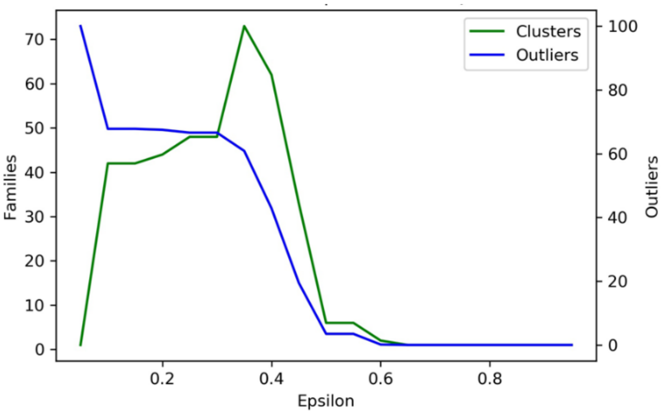
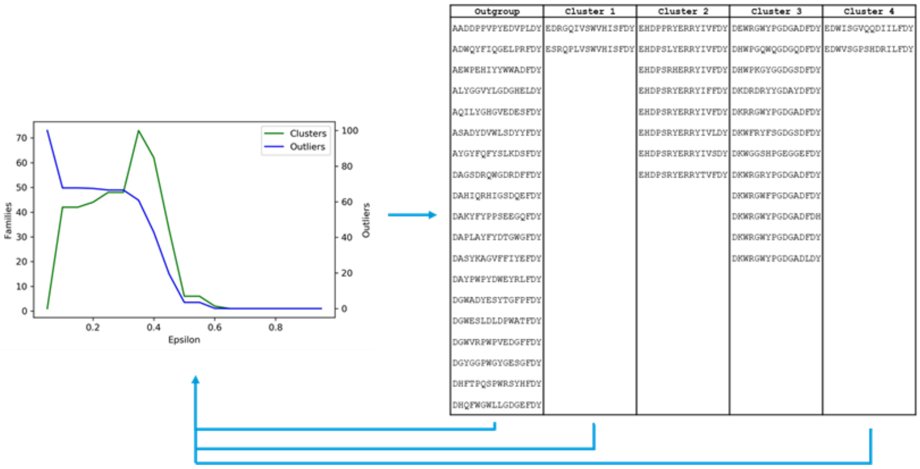


Figure 16 – Cluster-maximizing epsilon screening example.

Sequences that do not fit within any cluster (i.e. are dissimilar to every other sequences) are exported into an outgroup. To circumvent the generation of very big outgroups, the outgroup is grouped with sequences coming from clusters which only contain two sequences, and re-clustered using the same method (Figure 17). The process is stopped once the program can no longer find new clusters with more than 2 unique sequences. This ensures maximal diversity is sampled out of the datasets.



Outgroup	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AADDPPVPYEDVPLDY	EDRGQIVSWVHISFDY	EHDPPRYERRYIVFDY	DEWRGWYPGDGADFDY	EDWISGVQQDIILFDY
ADWQYFIQGELPRFDY	ESRQPLVSWVHISFDY	EHDPSLYERRYIVFDY	DHWPGQWQGDGQDFDY	EDWVSGPSHDRILFDY
AEWPEHIYYWWADFDY		EHDPSRHERRYIVFDY	DHWPKGYGGDGSDFDY	
ALYGGVYLGDGHELDY		EHDPSRYERRYIFFDY	DKDRDRYYGDAYDFDY	
AQILYGHGVEDESFDY		EHDPSRYERRYIVFDY	DKRRGWYPGDGADFDY	
ASADYDVWLSDYYFDY		EHDPSRYERRYIVLDY	DKWFRYFSGDGSDFDY	
AYGYFQFYSLKDSFDY		EHDPSRYERRYIVSDY	DKWGGSHPGEGGEFDY	
DAGSDRQWGDRDFFDY		EHDPSRYERRYTVFDY	DKWRGRYPGDGADFDY	
DAHIQRHIGSDQEFDY			DKWRGWFPGDGADFDY	
DAKYFYPPSEEGQFDY			DKWRGWYPGDGADFDH	
DAPLAYFYDTGWGFDY			DKWRGWYPGDGADFDY	
DASYKAGVFFIYEFDY			DKWRGWYPGDGADLDY	
DAYPWPYDWEYRLFDY				
DGWADYESYTGFPFDY				
DGWESLDLDPWATFDY				
DGWVRPWPVEDGFFDY				
DGYGGPWGYGESGFDY				
DHFTPQSPWRSYHFDY				
DHQFWGWLLGDGEFDY				

Figure 17 – Re-iteration example of the outgroup with clusters with only two sequences

IgG expression

The expression plasmids were ordered from ThermoFisher's GeneArt platform. The Light-chain (LC) and Heavy-chain (HC) of each IgG were ordered separately and transfected simultaneously (in a 1:1 ratio) with Polyethylenimine (PEI, in a 4:1 ratio with DNA) into 100 x 10^6 human embryonic kidney-293T (HEK- 293T) cells in 18 mL of FreeStyle^{TM} 293 Expression Medium (Life Technologies®). After 4 hours, an additional 20 mL of medium are added to the cells for a final cell concentration of 2.5 x 10^6 cells/mL. Transiently transfected cell cultures were incubated for 4 days in humidified atmosphere of 5% CO_2, 37°C and 140 rpm. After 4 days in culture, transfected cells are centrifuged at 300g for 10 minutes, and their supernatant collected, and vacuum filtered using 0.22 µm pore Steriflips (FisherScientific). The supernatant can be stored at 4°C for a week or at -20°C for extended periods.

IgG purification

IgG purification was performed by Affinity Ligand Chromatography, on Tecan Freedom EVO 200 (equipped with a Liquid Handling arm with 8 stainless steel tips, syringes of 1 mL and TeChrom, to enable fast IgG purification) using MabSelect Sure RoboColumns (Repligen; Ref.: PN 01050408R. Total Column Volumn (CV) = 200 µL). Phosphate Saline Buffer (PBS, pH 7.0) was used as the equilibration buffer. Samples were loaded 1 mL at a time, for a total final load of 35 mL. Retrieval of IgGs was achieved by isocratic elution using 5 CV of 50 mM Citrate-NaCl pH 3.0, for a final eluted volume of 1mL. The pH is neutralized by the addition of 150 µL of 1M Tris-HCL pH 9.0. The sample is then filtered trough a 0.22 µm filter pore using a syringe and stored at -20°C. Final volume = 1.15 mL.

IgG quantification

IgGs were quantified via HPLC Affinity Ligand Chromatography (HPLC-ALC), using a POROS™ CaptureSelect™ CH1-XL Affinity HPLC Column 2.1 x 30 mm, coupled to an Agilent 1260 Infinity II (Agilent Technologies). Separation of protein species was achieved using a flow rate of 2 mL/min and detection at 210 nm. Samples are injected directly without any previous dilution (injection volume = 50 µL), and the following method on Table 6 is employed for each individual injection:

Table *6* – ALC-HPLC method. Mobile Phase A: 10 mM NaH_2PO_4, 150 mM NaCl, pH 7.5; Mobile Phase B: 10 mM HCl, 150 mM NaCl, pH 2.0;

Time after injection (in minutes)	Mobile Phase A (in %)	Mobile Phase B (in %)
0	100	0
1.87	100	0
1.88	0	100
4.38	0	100
4.39	100	0

mAb peaks are manually integrated to calculate the Peak Area. Antibody concentration is calculated according to Equation 1.

Equation 1:

$$C_A = Peak\ Area_A \times \left(\frac{C_{IS}}{Peak\ Area_{IS}}\right) \times \left(\frac{1}{\frac{RRF_A}{RRF_{IS}}}\right)$$

An internal standard (IS) IgG with known concentration was used to generate an internal response factor (RRF_{IS} = Peak Area $_{IS}$/ Concentration $_{IS}$). Each sample concentration (C_A) was calculated as shown in Rome, K. & McIntyre, A. (2012)[1], by taking into account the concentration of IS (C_{IS}) and by comparing the sample's RRF (RRF_A) with the RRF of IS (RRF_{IS}). (Equation 1)

Size-exclusion chromatography

50 mM Sodium Phosphate pH 6.5 was used to dilute IgG samples to a final concentration of 1 mg/mL. Each candidate was analyzed by size exclusion chromatography on a SEC BEH 200 column (Waters, 200 Å, 1.7 µm, 4.6 mm x 150mm) using an Agilent 1260 Infinity II HPLC system, equipped with a multi-wavelength detector. A total run time of 35 minutes per sample was employed, after a 2 µg injection of each sample The mobile phase was 50 mM Sodium Phosphate pH 6.0 + 400 mM sodium perchlorate pH 6.0. Separation of protein species according to their molecular weight was achieved by applying an isocratic elution using a flow rate of 0.4 mL/min and detection at 210 nm. Peak integration of IgG monomers was done at a retention time around 20 minutes; these are referred to as "main peaks". Peaks and/or shoulders before the "main peak" are indicative of aggregation and referred to as "high molecular weight species" (HMWs). Peaks and/or shoulders after the main peak are indicative of fragmentation of the IgG monomer and designated "low molecular weight species (LMWs).

Hydrophobic-interaction chromatography

The hydrophobic profile of each candidate was analyzed by hydrophobic-interaction chromatography (HIC) in a TSKgel Butyl-NPR column (4.6 mm ID x 35 mm L) (Tosoh Biosciences). PBS was used to dilute the samples to 1 mg/mL. The mobile phase A was composed by 20 mM His/HCl, pH = 6.0 containing 1.5 M AS. Gradient elution of protein species was achieved by a gradual buffer replacement of mobile phase A with 20 mM His/HCl, pH 6.0 (mobile phase B). The gradient is 20 CV in length and has a slope of – 0.103 M AS per minute. A calibration curve was employed, where the retention time of reference standards was plotted against concentration of AS to calculate the hydrophobicity of the protein molecules.

Differential Scanning Fluorometry

Differential Scanning Fluorometry was performed in BioRad CFX96. Samples were diluted to 0.3 mg/mL (Vf = 50uL) in 43uL of PBS, to which 7 µL of SYPRO orange (previously prepared) was added. Sypro orange preparation was done by diluting the 5000x stock, by pipetting 1.4 µL from the stock solution into 1 mL of H20. The reaction was performed with a temperature increment of 0.5 ºC/min, from 25 °C to 100 °C.

Octet affinity measurements of anti-Herceptin antibodies

All kinetic assays were performed on Octet® RED96 (ForteBio), using 96-well plates (Corning), at 30 °C and 1000 rpm orbital shake speed. Samples were diluted in freshly prepared by diluting 10× Kinetic buffer (PALL) 1:9 in PBS (Gibco). Herceptin, which is a commercial IgG, was loaded either into anti-human Fc (AHC) Octet biosensors tips, by submerging them for 40 seconds in a 200 µL solution of Herceptin at 0.05 mg/mL. This is followed by a baseline step of 1 minute in kinetic buffer. Since mAbs will be assayed for their affinity towards Herceptin, we need to perform a saturation step, to make sure that the AHC biosensor is inaccessible to the mAbs assayed on the following steps. The saturation is achieved by submerging the Herceptin-loaded biosensors in a 200 µL solution of irrelevant-mAb0 at 0.2 mg./mL. This step is also followed by a baseline step of 1 minute in kinetic buffer. The Herceptin-loaded biosensors are then submerged in wells containing different concentrations of mAbs for 900 seconds – the association phase –, followed by a 1800 seconds dissociation phase in kinetic buffer. The mAb-loaded tips were also dipped in wells that only contained kinetic buffer, to serve as the basal reference signal used in the estimation of the affinity parameters step.

Estimation of interaction rates and affinity parameters

Binding sensograms were first aligned at the beginning of the association phase, and following the single reference subtraction, they were globally fit to a 1:1 binding model, were a single k-on and k-off is calculated for all binding sensograms for every concentration tested.

4.6. Acknowledgements

The author – Jorge Sampaio – was responsible for all the work in this chapter. The author acknowledges the valuable contribution of Elísio Silva and André F. Faustino for the development of the IgG-IgG interaction assays on Octet, and for their valuable input when evaluating the kinetic measurement data. The author would also like to thank Miguel A. Antunes and Inês Isidro from the Data Analytics team, for working closely with the author and with other NVS team members to build the iSeRa clustering tool.

4.7. References of Chapter 4

1. Tiller, T. *et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs* **5**, 445–470 (2013).

2. Hawkins, R. E., Russell, S. J. & Winter, G. Selection of phage antibodies by binding affinity. Mimicking affinity maturation. *J. Mol. Biol.* **226**, 889–896 (1992).

3. Ch'ng, A. C. W., Ahmad, A., Konthur, Z. & Lim, T. S. A High-Throughput Magnetic Nanoparticle-Based Semi-Automated Antibody Phage Display Biopanning. in *Human Monoclonal Antibodies* (ed. Steinitz, M.) vol. 1904 377–400 (Springer New York, 2019).

4. Dudik, J. M., Kurosu, A., Coyle, J. L. & Sejdić, E. A Comparative Analysis of DBSCAN, K-Means, and Quadratic Variation Algorithms for Automatic Identification of Swallows from Swallowing Accelerometry Signals. *Comput. Biol. Med.* **59**, 10–18 (2015).

5. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* **42**, 19:1-19:21 (2017).

6. Schaefer, J. V., Sedlák, E., Kast, F., Nemergut, M. & Plückthun, A. Modification of the kinetic stability of immunoglobulin G by solvent additives. *mAbs* **10**, 607–623 (2018).

7. He, F., Hogan, S., Latypov, R. F., Narhi, L. O. & Razinkov, V. I. High throughput thermostability screening of monoclonal antibody formulations. *J. Pharm. Sci.* **99**, 1707–1720 (2010).

8. Vermeer, A. W. P. & Norde, W. The Thermal Stability of Immunoglobulin: Unfolding and Aggregation of a Multi-Domain Protein. *Biophys. J.* **78**, 394–404 (2000).

9. Ionescu, R. M., Vlasak, J., Price, C. & Kirchmeier, M. Contribution of Variable Domains to the Stability of Humanized IgG1 Monoclonal Antibodies. *J. Pharm. Sci.* **97**, 1414–1426 (2008).

10. Burkovitz, A. & Ofran, Y. Understanding differences between synthetic and natural antibodies can help improve antibody engineering. *mAbs* **8**, 278–287 (2016).

11. Romesberg, F. E., Spiller, B., Schultz, P. G. & Stevens, R. C. Immunological origins of binding and catalysis in a Diels-Alderase antibody. *Science* **279**, 1929–1933 (1998).

12. Notkins, A. L. Polyreactivity of antibody molecules. *Trends Immunol.* **25**, 174–179 (2004).

13. Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E. & Meiler, J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. *PLoS Comput. Biol.* **9**, e1003045 (2013).

14. Schmidt, A. G. *et al.* Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc. Natl. Acad. Sci.* **110**, 264–269 (2013).

15. Tsuchiya, Y. & Mizuguchi, K. The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci. Publ. Protein Soc.* **25**, 815–825 (2016).

16. Marks, C. & Deane, C. M. How repertoire data are changing antibody science. *J. Biol. Chem.* **295**, 9823–9837 (2020).

Chapter 5 – CDR-null: Primed libraries for Affinity Maturation

5.1. Summary

On the previous chapter, we showed how primed libraries κN1 and κN2 can generate candidates with similar affinity towards Herceptin, when compared with FW-κ candidates. We also showed that κN1 and κN2 datasets are essentially different from FW-κ, by analyzing HCDR3 sequence diversity, HCDR3 length distribution of outputs, and by visual inspection of datasets. Here, we employ different affinity maturation strategies to FW-κ, FW-κN1 and FW-κN2 candidates to evaluate if using primed libraries κN1 and κN2 leads to manifestly better results than libraries based on FW-κ. The primed framework κN1 was shown to have the biggest rate of success among all conditions (31.3%), specially when combined with the semi-blind affinity maturation method (36.4%). FW-κN1 candidates were also responsible for the biggest fold-increase (FI) in affinity, and often reverted part of their CDR-null residues back to the original germile residues, which conferred them more optimal developability characteristics.

5.2. Introduction

Regardless of the origin and platform used, antibodies generated against a given target can have their affinity improved, besides other advantageous characteristics. In vitro affinity maturation panning constitutes a high-throughput alternative to the classical approach based on the analysis of X-ray crystallography data. While *in vitro* affinity maturation panning provides throughput and generalization potential, it is not as case-specific as direct structural inspections and may lead to inconsistent results. *In vitro* affinity maturation libraries work as traditional antibody libraries but use a single parental antibody as their starting point. In most cases, the HCDR3 and antibody framework are maintained, while the other CDRs are diversified. Typical approaches involve the diversification of each CDR individually in separate cassettes. These can be used to select several beneficial mutations within each CDR, in parallel. Mutations selected in the first rounds of selection can be combined to search for synergistic effects. However, there is no telling if synergistic mutations were selected out during the first steps of selection, when cassettes were separated from each other and when single-point mutations were being sampled. Such may indicate why sometimes there are no noticeable gains in affinity after such strategies, a phenomenon that we also saw in *in-house* results (not shown). Ideally, innovative affinity maturation methods should be generalizable to provide high-throughput results while maintaining a certain degree of specificity towards the antibody structure being considered. As such they require attention to be paid to specific regions, such as the ones likely to be in contact with the antigen or regions that influence the antibodies' structural integrity and overall developability. A "semi-blind" affinity maturation process can be proposed as a compromise between generalization and specificity. In this chapter, we propose a "semi-blind" affinity maturation process were structural hotspots on primed libraries kN1 and kN2 are targeted for maturation with the objective of maximizing the likelihood of finding beneficial mutations that improve affinity.

5.3. Results

Several different candidates where identified after three rounds of panning with different antibody libraries (FW-κ, FW-κN1 and FW-κN2). Of these, 23 had affinity towards herceptin and were deemed suitable to proceed to affinity maturation. They were subjected to two different *in vitro* affinity maturation randomizations as to improve their affinity towards Herceptin®.

5.3.1. Two Randomization designs: L3/H2 Cassette versus TWIST

5.3.1.1. Generation of blind affinity maturation libraries using L3/H2 cassettes

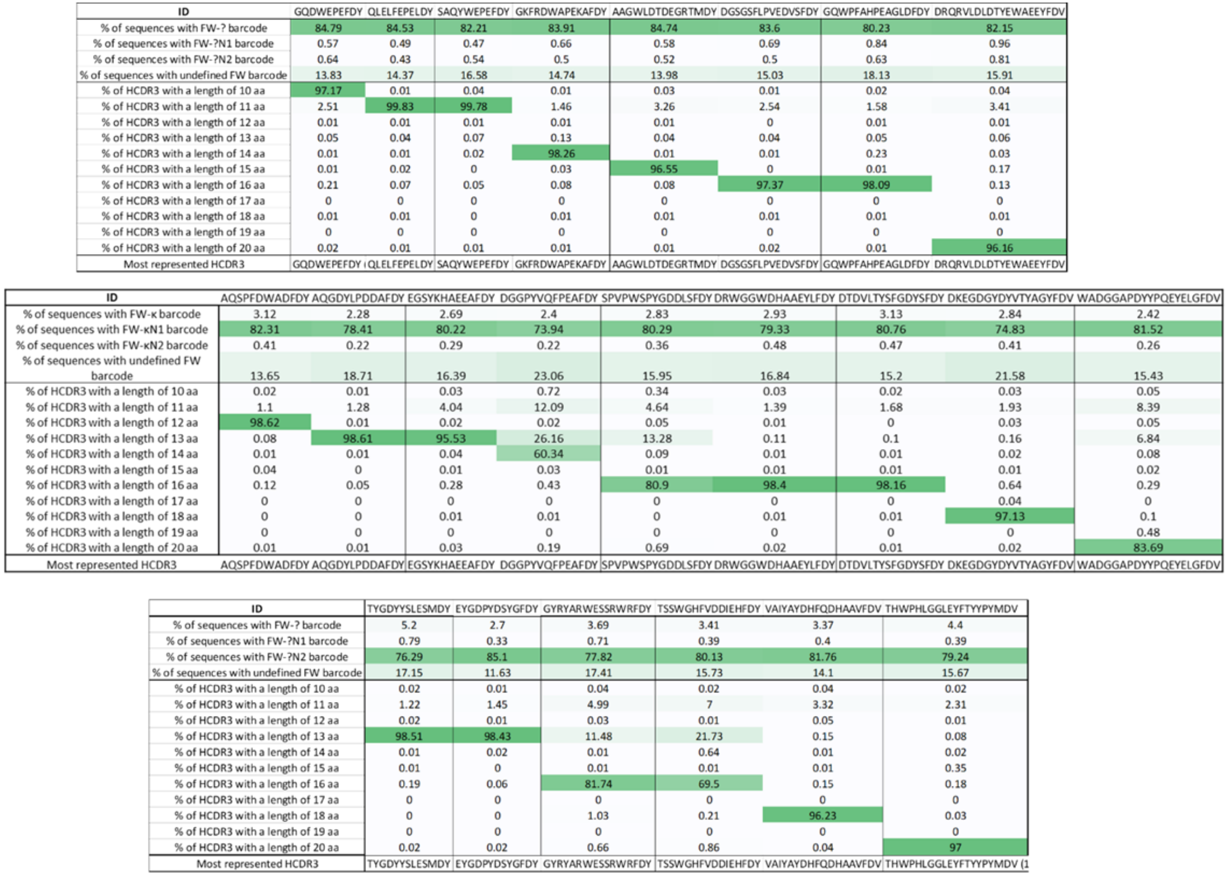
The LCDR3 and HCDR2 are major contributors for antigen binding (see section 1.1.3.). Hence, the first affinity maturation method involves cloning the parental HCDR3 sequences into frameworks randomized in the LCDR3 and HCDR2 loops (L3/H2 cassettes). This is the standard affinity maturation method used at the lab (along with LTM libraries, not explored on this thesis). The randomization design mimics the composition of the CDRs found in naturally occurring human rearranged antibody genes. The remaining CDR and FR sequences remain unchanged and are the same as their respective parental framework.

The LCDR3 and HCDR2 sequences were randomized via PCR, by amplification of those regions with randomized TRIMoligos (for more information see section 5.5.). The HCDR3 parental sequences were cloned into the appropriate L3/H2-randomized phage-display vectors and transformed into TGF1+ cells by electroporation. All libraries had diversities above 10^8 cfu/mL and low vector background. (Table 1)

The libraries on table 1 were inspected by NGS, as to confirm that the HCDR3 was correctly clones into L3/H2-randomized frameworks (Figure 2), and that the LCDR3 and HCDR2 randomization designs were according to plan (data not shown).

Table 1 – L3/H2 cassette affinity maturation libraries' size after transformation of TG1F+ cells by electroporation. Library size assessed by counting colony forming units (cfu) after serial dilutions and plating transformed cell after the electroporation protocol (for more details see section 5.5).

Parental HCDR3 sequences	Framework	Library Size (cfu/mL)	Estimated Vector Background
GQDWEPEFDY	FW-κ	2.43E+08	3.38%
QLELFEPELDY	FW-κ	1.06E+09	0.80%
SAQYWEPEFDY	FW-κ	3.62E+08	2.30%
GKFRDWAPEKAFDY	FW-κ	1.93E+09	0.44%
AAGWLDTDEGRTMDY	FW-κ	2.83E+08	2.92%
DGSGSFLPVEDVSFDY	FW-κ	3.14E+08	2.65%
GQWPFAHPEAGLDFDY	FW-κ	1.89E+09	0.45%
DRQRVLDLDTYEWAEEYFDV	FW-κ	9.28E+08	0.91%
AQSPFDWADFDY	FW-κN1	1.26E+09	0.67%
AQGDYLPDDAFDY	FW-κN1	5.78E+08	1.45%
EGSYKHAEEAFDY	FW-κN1	7.16E+08	1.18%
DGGPYVQFPEAFDY	FW-κN1	2.27E+09	0.37%
SPVPWSPYGDDLSFDY	FW-κN1	1.73E+09	0.49%
DRWGGWDHAAEYLFDY	FW-κN1	8.76E+08	0.96%
DTDVLTYSFGDYSFDY	FW-κN1	2.05E+09	0.41%
DKEGDGYDYVTYAGYFDV	FW-κN1	1.65E+09	0.51%
WADGGAPDYYPQEYELGFDV	FW-κN1	1.82E+09	0.47%
TYGDYYSLESMDY	FW-κN2	1.95E+09	0.44%
EYGDPYDSYGFDY	FW-κN2	1.56E+09	0.54%
GYRYARWESSRWRFDY	FW-κN2	6.59E+08	1.28%
TSSWGHFVDDIEHFDY	FW-κN2	3.80E+08	2.19%
VAIYAYDHFQDHAAVFDV	FW-κN2	1.49E+09	0.57%
THWPHLGGLEYFTYYPYMDV	FW-κN2	1.20E+09	0.71%



ID	GQDWEPEFDY	QLELFEPELDY	SAQYWEPEFDY	GKFRDWAPEKAFDY	AAGWLDTDEGRTMDY	DGSGSFLPVEDVSFDY	GQWPFAHPEAGLDFDY	DRQRVLDLDTYEWAEEYFDV
% of sequences with FW-? barcode	84.79	84.53	82.21	83.91	84.74	83.6	80.23	82.15
% of sequences with FW-?N1 barcode	0.57	0.49	0.47	0.66	0.58	0.69	0.84	0.96
% of sequences with FW-?N2 barcode	0.64	0.43	0.54	0.5	0.52	0.5	0.63	0.81
% of sequences with undefined FW barcode	13.83	14.37	16.58	14.74	13.98	15.03	18.13	15.91
% of HCDR3 with a length of 10 aa	97.17	0.01	0.04	0.01	0.03	0.01	0.02	0.04
% of HCDR3 with a length of 11 aa	2.51	99.83	99.78	1.46	3.26	2.54	1.58	3.41
% of HCDR3 with a length of 12 aa	0.01	0.01	0.01	0.01	0.01	0	0.01	0.01
% of HCDR3 with a length of 13 aa	0.05	0.04	0.07	0.13	0.04	0.04	0.05	0.06
% of HCDR3 with a length of 14 aa	0.01	0.01	0.02	98.26	0.01	0.01	0.23	0.03
% of HCDR3 with a length of 15 aa	0.01	0.02	0	0.03	96.55	0	0.01	0.17
% of HCDR3 with a length of 16 aa	0.21	0.07	0.05	0.08	0.08	97.37	98.09	0.13
% of HCDR3 with a length of 17 aa	0	0	0	0	0	0	0	0
% of HCDR3 with a length of 18 aa	0.01	0.01	0	0.01	0.01	0.01	0.01	0.01
% of HCDR3 with a length of 19 aa	0	0	0	0	0	0	0	0
% of HCDR3 with a length of 20 aa	0.02	0.01	0.01	0.01	0.01	0.02	0.01	96.16
Most represented HCDR3	GQDWEPEFDY (	QLELFEPELDY	SAQYWEPEFDY	GKFRDWAPEKAFDY	AAGWLDTDEGRTMDY	DGSGSFLPVEDVSFDY	GQWPFAHPEAGLDFDY	DRQRVLDLDTYEWAEEYFDV

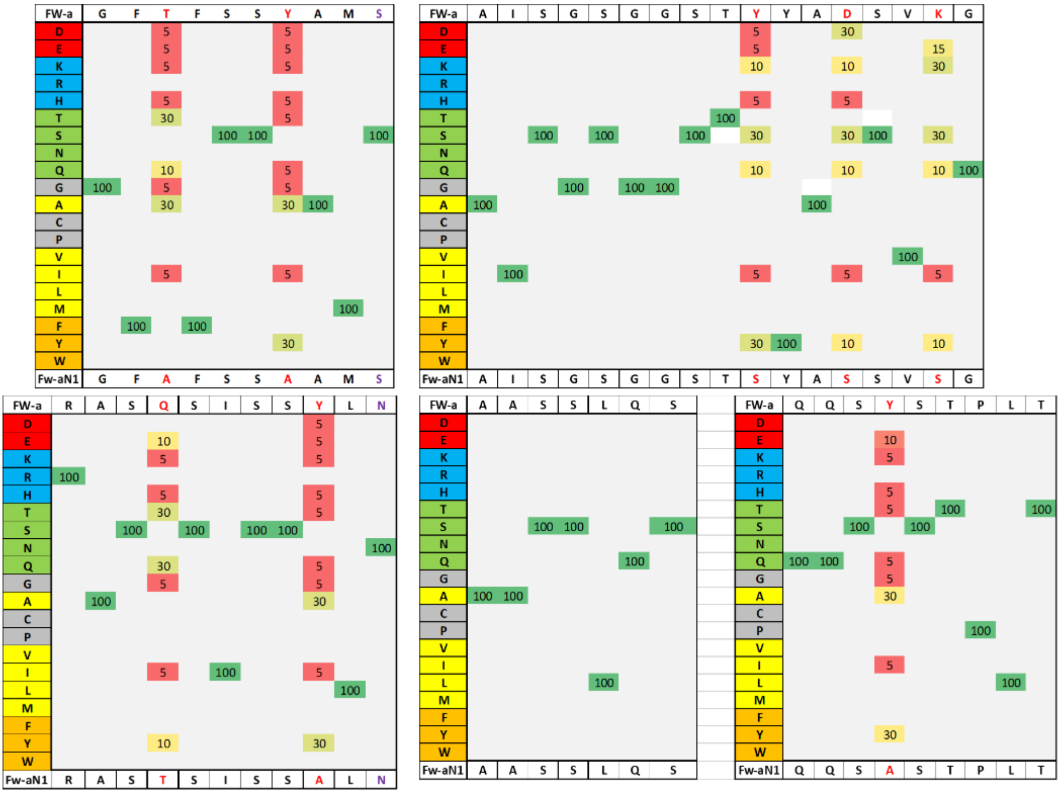
ID	AQSPFDWADFDY	AQGDYLPDDAFDY	EGSYKHAEEAFDY	DGGPYVQFPEAFDY	SPVPWSPYGDDLSFDY	DRWGGWDHAAEYLFDY	DTDVLTYSFGDYSFDY	DKEGDGYDYVTYAGYFDV	WADGGAPDYYPQEYELGFDV
% of sequences with FW-κ barcode	3.12	2.28	2.69	2.4	2.83	2.93	3.13	2.84	2.42
% of sequences with FW-κN1 barcode	82.31	78.41	80.22	73.94	80.29	79.33	80.76	74.83	81.52
% of sequences with FW-κN2 barcode	0.41	0.22	0.29	0.22	0.36	0.48	0.47	0.41	0.26
% of sequences with undefined FW barcode	13.65	18.71	16.39	23.06	15.95	16.84	15.2	21.58	15.43
% of HCDR3 with a length of 10 aa	0.02	0.01	0.03	0.72	0.34	0.03	0.02	0.03	0.05
% of HCDR3 with a length of 11 aa	1.1	1.28	4.04	12.09	4.64	1.39	1.68	1.93	8.39
% of HCDR3 with a length of 12 aa	98.62	0.01	0.02	0.02	0.05	0.01	0	0.03	0.05
% of HCDR3 with a length of 13 aa	0.08	98.61	95.53	26.16	13.28	0.11	0.1	0.16	6.84
% of HCDR3 with a length of 14 aa	0.01	0.01	0.04	60.34	0.09	0.01	0.01	0.02	0.08
% of HCDR3 with a length of 15 aa	0.04	0	0.01	0.03	0.01	0.01	0.01	0.01	0.02
% of HCDR3 with a length of 16 aa	0.12	0.05	0.28	0.43	80.9	98.4	98.16	0.64	0.29
% of HCDR3 with a length of 17 aa	0	0	0	0	0	0	0	0.04	0
% of HCDR3 with a length of 18 aa	0	0	0.01	0.01	0	0.01	0.01	97.13	0.1
% of HCDR3 with a length of 19 aa	0	0	0	0	0	0	0	0	0.48
% of HCDR3 with a length of 20 aa	0.01	0.01	0.03	0.19	0.69	0.02	0.01	0.02	83.69
Most represented HCDR3	AQSPFDWADFDY	AQGDYLPDDAFDY	EGSYKHAEEAFDY	DGGPYVQFPEAFDY	SPVPWSPYGDDLSFDY	DRWGGWDHAAEYLFDY	DTDVLTYSFGDYSFDY	DKEGDGYDYVTYAGYFDV	WADGGAPDYYPQEYELGFDV

ID	TYGDYYSLESMDY	EYGDPYDSYGFDY	GYRYARWESSRWRFDY	TSSWGHFVDDIEHFDY	VAIYAYDHFQDHAAVFDV	THWPHLGGLEYFTYYPYMDV
% of sequences with FW-? barcode	5.2	2.7	3.69	3.41	3.37	4.4
% of sequences with FW-?N1 barcode	0.79	0.33	0.71	0.39	0.4	0.39
% of sequences with FW-?N2 barcode	76.29	85.1	77.82	80.13	81.76	79.24
% of sequences with undefined FW barcode	17.15	11.63	17.41	15.73	14.1	15.67
% of HCDR3 with a length of 10 aa	0.02	0.01	0.04	0.02	0.04	0.02
% of HCDR3 with a length of 11 aa	1.22	1.45	4.99	7	3.32	2.31
% of HCDR3 with a length of 12 aa	0.02	0.01	0.03	0.01	0.05	0.01
% of HCDR3 with a length of 13 aa	98.51	98.43	11.48	21.73	0.15	0.08
% of HCDR3 with a length of 14 aa	0.01	0.02	0.01	0.64	0.01	0.02
% of HCDR3 with a length of 15 aa	0.01	0	0.01	0.01	0.01	0.35
% of HCDR3 with a length of 16 aa	0.19	0.06	81.74	69.5	0.15	0.18
% of HCDR3 with a length of 17 aa	0	0	0	0	0	0
% of HCDR3 with a length of 18 aa	0	0	1.03	0.21	96.23	0.03
% of HCDR3 with a length of 19 aa	0	0	0	0	0	0
% of HCDR3 with a length of 20 aa	0.02	0.02	0.66	0.86	0.04	97
Most represented HCDR3	TYGDYYSLESMDY	EYGDPYDSYGFDY	GYRYARWESSRWRFDY	TSSWGHFVDDIEHFDY	VAIYAYDHFQDHAAVFDV	THWPHLGGLEYFTYYPYMDV (1

Figure 2 – Quality control of Affinity Maturation Libraries L3/H2. The CDR cassettes allow up to 11 amino acids at each position. The LCDR3 has 1.6 x 10^5 possible combinations and the HCDR2 3.0 x 10^8 possible combinations. The Theoretical diversity of LCDR3 + HCDR2 cassettes = 4.8 x 10^{13}

5.3.1.2. Generation of semi-blind affinity maturation libraries using TWIST

The second affinity maturation method also involves cloning the HCDR3 sequences into an affinity maturation framework. In this case, instead of randomizing several aminoacids across LCDR3 and HCDR2, only the CDR-null positions were targeted for diversification (Figure 3). This was accomplished using Twist Bioscience's proprietary silicon-based DNA synthesis platform technology. These will be hereinafter reffered to as Twist Libraries for simplicity.



FW-a	G	F	T	F	S	S	Y	A	M	S
D			5				5			
E			5				5			
K			5				5			
R										
H			5				5			
T			30				5			
S					100	100				100
N										
Q			10				5			
G	100		5				5			
A			30				30	100		
C										
P										
V										
I			5				5			
L										
M									100	
F		100		100						
Y							30			
W										
Fw-aN1	G	F	A	F	S	S	A	A	M	S

FW-a	A	I	S	G	S	G	G	S	T	Y	Y	A	D	S	V	K	G
D										5			30				
E										5						15	
K										10			10			30	
R																	
H										5			5				
T									100								
S			100		100			100		30			30	100		30	
N																	
Q										10			10			10	100
G				100		100	100										
A	100											100					
C																	
P																	
V															100		
I		100								5			5			5	
L																	
M																	
F																	
Y										30	100		10			10	
W																	
Fw-aN1	A	I	S	G	S	G	G	S	T	S	Y	A	S	S	V	S	G

FW-a	R	A	S	Q	S	I	S	S	Y	L	N
D									5		
E				10					5		
K				5					5		
R	100										
H				5					5		
T				30					5		
S			100		100		100	100			
N											100
Q				30					5		
G				5					5		
A		100							30		
C											
P											
V											
I				5		100			5		
L										100	
M											
F											
Y				10					30		
W											
Fw-aN1	R	A	S	T	S	I	S	S	A	L	N

FW-a	A	A	S	S	L	Q	S
D							
E							
K							
R							
H							
T							
S			100	100			100
N							
Q						100	
G							
A	100	100					
C							
P							
V							
I							
L					100		
M							
F							
Y							
W							
Fw-aN1	A	A	S	S	L	Q	S

FW-a	Q	Q	S	Y	S	T	P	L	T
D									
E				10					
K				5					
R									
H				5					
T				5		100			100
S			100		100				
N									
Q	100	100		5					
G				5					
A				30					
C									
P							100		
V									
I				5					
L								100	
M									
F									
Y				30					
W									
Fw-aN1	Q	Q	S	A	S	T	P	L	T

Figure 3 - TWIST libraries randomization design. For every diversified position, a 30% probability was given the original FW-κ amino acid and another 30% to the CDR-null mutation. The remaining 40% were split among four groups: Positive, Negative, Polar, Hidrophobic and Aromatic aminoacids. Positive and negative aminoacids where given the same weight whenever possible.

Diversification of CDR-null hotspots provides two advantadges. Firstly, it consitutes a generalized (or blind) method that can be applied to any primary candidate coming from a primary panning. Secondly, and as mentioned throughout this thesis, such positions were identified as likely to target the antigen. This adds focus to the randomization protocol and increases the likelihood of finding beneficial mutations, hence the name semi-blind. Specifically, the 60% allocation of FW-κ and FW-κN1/2 residues to each position means that most combinations will tend to respect the original conformation of the parental antibody and that deviations from the original format will need to be very beneficial to dominate the sample and be selected. The frameworks were cloned into the appropriate phage-display vectors and transformed into TGF1+ cells by electroporation. All libraries had diversities above 10^8 cfu/mL and low vector background. (Table 2, for more details see section 5.5.)

Table *2* - TWIST Affinity maturation libraries' size after transformation of TG1F+ cells by electroporation.

ID	Framework	Library Size (cfu/mL)	Estimated Vector Background (%)
GQDWEPEFDY	FW-κ	5.39E+08	1.56%
QLELFEPELDY	FW-κ	6.97E+08	1.21%
SAQYWEPEFDY	FW-κ	3.57E+08	2.33%
GKFRDWAPEKAFDY	FW-κ	4.24E+08	1.97%
AAGWLDTDEGRTMDY	FW-κ	7.65E+08	1.10%
DGSGSFLPVEDVSFDY	FW-κ	1.24E+09	0.69%
GQWPFAHPEAGLDFDY	FW-κ	3.43E+08	2.42%
DRQRVLDLDTYEWAEEYFDV	FW-κ	3.43E+08	2.42%
AQSPFDWADFDY	FW-κN1	3.22E+08	2.58%
AQGDYLPDDAFDY	FW-κN1	3.74E+08	2.23%
EGSYKHAEEAFDY	FW-κN1	1.08E+09	0.78%
DGGPYVQFPEAFDY	FW-κN1	8.03E+08	1.05%
SPVPWSPYGDDLSFDY	FW-κN1	6.77E+08	1.24%
DRWGGWDHAAEYLFDY	FW-κN1	3.73E+08	2.24%
DTDVLTYSFGDYSFDY	FW-κN1	3.26E+08	2.55%
DKEGDGYDYVTYAGYFDV	FW-κN1	2.86E+08	2.89%
WADGGAPDYYPQEYELGFDV	FW-κN1	5.69E+08	1.48%
TYGDYYSLESMDY	FW-κN2	1.71E+09	0.50%
EYGDPYDSYGFDY	FW-κN2	1.79E+09	0.47%
GYRYARWESSRWRFDY	FW-κN2	5.62E+08	1.49%
TSSWGHFVDDIEHFDY	FW-κN2	5.24E+08	1.60%
VAIYAYDHFQDHAAVFDV	FW-κN2	1.12E+09	0.75%
THWPHLGGLEYFTYYPYMDV	FW-κN2	1.01E+09	0.84%

5.3.2. Selection of candidates from affinity maturation pannings

In total, 46 affinity maturation libraries (2 designs x 23 parentals) were challenged against Herceptin® with the goal of maturing the selected candidates. Two panning rounds of increased stringency will be done following a solution-phase selection strategy coupled with KingFisher™ Flex Purification System as described in section 4.3.2. (chapter 4). The first round had the same wash stringency as the primary panning's third round (see section 4.3.) but decreased the antigen concentration. On top of that, an off-rate selection protocol was implemented to positively select binders with longer off-rates, based on a paper from Zahnd *et al.*[1] The second round further decreased the antigen concentration to 1 nM and did not employ an off-rate seletion. For complete information see the methods section 5.5. The affinity maturation output data can be found in the table below.

Table 3 - Output results from affinity maturation panning. Phage-infected bacteria from each round were serially diluted into plates and counted the next day to calculate the respective round output. The “mock” outputs were 3 x 10^7 on average (not shown).

ID	1st Round		2nd Round	
	L3/H2	TWIST	L3/H2	TWIST
GQDWEPEFDY	4.70E+05	6.57E+05	1.81E+06	3.47E+05
QLELFEPELDY	1.30E+05	7.21E+05	2.84E+05	3.02E+06
SAQYWEPEFDY	1.61E+06	2.04E+06	2.18E+06	1.41E+07
GKFRDWAPEKAFDY	1.78E+05	6.44E+05	8.74E+05	2.02E+06
AAGWLDTDEGRTMDY	1.11E+05	1.50E+06	5.46E+04	7.32E+06
DGSGSFLPVEDVSFDY	1.12E+05	9.03E+05	1.34E+06	1.70E+07
GQWPFAHPEAGLDFDY	1.12E+05	9.05E+04	2.91E+05	2.97E+05
DRQRVLDLDTYEWAEEYFDV	8.82E+04	1.85E+05	9.66E+04	4.82E+05
AQSPFDWADFDY	4.30E+06	3.92E+04	3.08E+04	8.42E+06
AQGDYLPDDAFDY	2.87E+06	1.76E+05	3.14E+05	8.78E+06
EGSYKHAEEAFDY	1.28E+06	1.54E+07	1.08E+06	8.78E+06
DGGPYVQFPEAFDY	2.74E+05	2.44E+06	8.40E+04	6.00E+06
SPVPWSPYGDDLSFDY	4.03E+06	1.95E+06	6.22E+05	5.82E+06
DRWGGWDHAAEYLFDY	3.19E+07	1.49E+08	2.60E+06	1.51E+08
DTDVLTYSFGDYSFDY	2.69E+05	6.46E+05	5.60E+04	2.17E+06
DKEGDGYDYVTYAGYFDV	1.23E+06	7.78E+05	2.91E+05	6.99E+06
WADGGAPDYYPQEYELGFDV	1.14E+06	3.78E+04	3.64E+05	5.05E+07
TYGDYYSLESMDY	1.36E+06	6.11E+05	2.24E+04	2.15E+06
EYGDPYDSYGFDY	1.85E+06	3.69E+05	9.24E+04	6.05E+05
GYRYARWESSRWRFDY	3.14E+06	1.51E+05	9.24E+04	3.36E+05
TSSWGHFVDDIEHFDY	4.03E+06	7.73E+05	1.46E+05	4.81E+06
VAIYAYDHFQDHAAVFDV	6.63E+06	2.40E+07	5.88E+04	6.54E+06
THWPHLGGLEYFTYYPYMDV	5.02E+06	9.63E+05	1.20E+05	2.84E+06

After two panning rounds, the samples were analysed by NGS using a coupled V_L:V_H analysis. This newly developed technique allows to correlate mutations on V_L with mutations on V_H to find rare clones with increased affinity (Manuscript under submission, for more information see section Annex B).

From inspecting the NGS datasets, 81 mature candidates were selected from 23 different parental antibodies (46 datasets). The candidates' sequences can be found on Tables 4-6.

Table 4 - Mutations of matured FW-κ antibodies. Successfully matured candidates highlighted in bold. Mutations highlighted in red.

Parental HCDR3	Parental K_D (nM)	Maturation Method	Final K_D (nM)	LCDR1	LCDR3	HCDR1	HCDR2
				ASQSISSYLN	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG
GQDWEPEFDY	34.5	L3/H2	n/a	....	...A.....	....	.........S..S..S.
		L3/H2	38	....	.......Y.	....	....
		L3/H2	40	....	.......Y.	....	..T....YH........
		TWIST	33	....	...S.....	....	....
QLELFEPELDY	38.1	L3/H2	18	....	....	....	..YS.A.HR........
		L3/H2	17	....	....	....	E..S.....R.......
		L3/H2	57	....	.......W.	....	....
		L3/H2	19	....	.......W.	....	E..S.....R.......
		TWIST	24	..T.......	...Q.....	.A.......	.........Q..S..S.
SAQYWEPEFDY	38	L3/H2	56	....	.......Y.	....	....
		L3/H2	61	....	.......W.	....	....
		L3/H2	78	....	.....V.W.	....	....
		L3/H2	45	....	.....V.Y.	....	....
		TWIST	39	..T.......	....	.A...A...	....
GKFRDWAPEKAFDY	45	L3/H2	52	....	.....V...	....	....
		L3/H2	37	....	....EV...	....	....
		TWIST	60	....	...S.....	....	....
AAGWLDTDEGRTMDY	32	L3/H2	185	....	....	....	E.A.....Y........
		L3/H2	n/a	....	....	....	Y.T..GRYH........
		L3/H2	n/a	....	.......W.	....	E.A.....Y........
		L3/H2	2044	....	.......W.	....	Y.T..GRYH........
		TWIST	15	....	....	.A...I...	.........K.......
DGSGSFLPVEDVSFDY	60	L3/H2	187	....	....	....	G......YYH.......
		TWIST	33	..T.......	....	.A...A...	............S....
GQWPFAHPEAGLDFDY	41	L3/H2	78	....	.......Y.	....	....
		L3/H2	10	....	.......Y.	....	..TAHGYY.........
		TWIST	535	..T.......	....	.A.......	.........K.......
		TWIST	634	.......T..	...I.....	.....I...	.........Q..H..S.
		TWIST	1734	.......G..	...K.....	....	.........K..I....
DRQRVLDLDTYEWAEEYFDV	84	L3/H2	58	....	....	....	G................
		L3/H2	119	....	....	....	G.....D..........
		TWIST	71	....	....	.K.......	....
		TWIST	n/a	.......Q..	....	.D...A...	.........K..K..S.

Table *5* - Mutations of of matured FW-κN1 antibodies. Successfully matured candidates highlighted in bold. Mutations highlighted in red.

Parental HCDR3	Parental K_D (nM)	Maturation Method	Final K_D (nM)	LCDR1	LCDR3	HCDR1	HCDR2
				ASTSISSALN	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG
AQSPFDWADFDY	38	L3/H2	36	....	....	....	..........D......
		L3/H2	28	....	...Y...Y.	....	....
		L3/H2	49	....	...Y...W.	....	....
		TWIST	38	....	....	.H...E...	....
AQGDYLPDDAFDY	47	L3/H2	48	....	....	....	.............R...
		L3/H2	94	....		....	.T...............
		L3/H2	228	....	...Y...Y.	....	....
		L3/H2	381	....	...Y...W.	....	....
		TWIST	105	..Q.......	....	.E...H...	............D....
		TWIST	160	....	...E.....	.T...Y...	.........Y..Y..K.
EGSYKHAEEAFDY	127	L3/H2	24	....	..VY.....	....	....
		L3/H2	11	....	....	....	.........Y..D..K.
		TWIST	27	..Q.......	...Y.....	.T.......	.........Y..D..K.
		TWIST	26	..Q....E..	...I.....	....	.........Y..D..K.
DGGPYVQFPEAFDY	57	L3/H2	394	....	..VY.....	....	....
		TWIST	26	..Q....E..	...Y.....	.T.......	............D..K.
SPVPWSPYGDDLSFDY	254	L3/H2	18	....	...YSY...	....	....
		L3/H2	18	....	...YSY...	....	..........S......
		TWIST	n/a	....	....	....	.........K.....K.
DRWGGWDHAAEYLFDY	60	L3/H2	116	....	....	....	.M...............
		L3/H2	42	....		....	.......R.........
		L3/H2	72	....	...Y...Y.	....	....
		L3/H2	114	....	...Y...W.	....	....
		TWIST	0.05	..Q.......	...Y.....	.....Y...	............D..K.
DTDVLTYSFGDYSFDY	34	L3/H2	34	....	...YEV...	....	....
		TWIST	1182	.......G..	....	.....Y...	.........Y..Y..E.
DKEGDGYDYVTYAGYFDV	49	L3/H2	589	....	...Y...Y.	....	.L...............
		L3/H2	318	....	...Y...W.	....	....
		TWIST	41	....	...Y.....	.....Y...	...............K.
WADGGAPDYYPQEYELGFDV	156	L3/H2	452	....	...Y...Y.	....	....
		L3/H2	326	....	...Y...W.	....	....
		TWIST	489	....	...Y.....	.Q...Y...	.........K.....Y.

Table 6 - Mutations of of matured FW-κN2 antibodies. Successfully matured candidates highlighted in bold. Mutations highlighted in red.

Parental HCDR3	Parental K_D (nM)	Maturation Method	Final K_D (nM)	LCDR1	LCDR3	HCDR1	HCDR2
				ASTSISSALN	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG
TYGDYYSLESMDY	40	L3/H2	34	....	....	....	...VH....A..D..K.
		TWIST	18	..E....D..	...G.....	.....Y...	............K..Y.
EYGDPYDSYGFDY	49	TWIST	65	..I....K..	....	.....I...	............K....
GYRYARWESSRWRFDY	30.5	L3/H2	49	....	....	....	..H......Y..D..K.
		L3/H2	77	....		....	S.H....DAY..D..K.
		L3/H2	97	....	...YSV...	....	....
		TWIST	71	..Q.......	...E.....	....	....
		TWIST	55	..Y....G..	...E.....	....	.........D.....Y.
TSSWGHFVDDIEHFDY	33	L3/H2	32	....	....	....	.M...............
		TWIST	31	....	....	....	.........I.......
VAIYAYDHFQDHAAVFDV	48	L3/H2	52	....	....	....	........P.S......
		TWIST	212	..Q....Y..	....	.T...Y...	.........Y..D..K.
THWPHLGGLEYFTYYPYMDV	56	L3/H2	144	....	....	....	..........D......
		TWIST	467	..E.......	....	.....Y...	.........D..Y..E.
		TWIST	34	..Q.......	...E.....	.G.......	............D..K.
		TWIST	53	..Q.......	....	.H.......	.........K..Q..Q.

5.3.3. Kinetic constants determination of affinity maturation candidates

The 81 candidates were firstly analyzed for their ability to bind to Herceptin by BLI on Octet RED96. The IgG-IgG interaction method explored on the previous chapter was used (section 4.5.) and their on-rates and off-rates plotted in Figures 4-6. Parental sequences and their respective outcomes will be hereinafter named after their parental HCDR3 sequence.

5.3.3.1. Overall affinity maturation success rates

The affinity maturation sucess rate as whole was of 24.7%, with 20 antibodies out of 81 significantly improving their affinities in comparison with their respective parental. More specifically, 24.2% of candidates tested for FW-κ (8 out of 33) were able to improve affinity, 31.3% of candidates tested for FW-κN1 (10 out of 32) were able to improve affinity and 12.5% of candidates tested for FW-κN2 (2 out of 16) were able to improve affinity. Regarding the methodology used, the L3/H2-cassette was sucessful in 22% (11 out of 50) of cases and TWIST in 29% (9 out of 31) of cases. (Table 7, Figures 4-6)

Table 7 – Affinity Maturation success rates of frameworks and methods. Based on the improvement of affinity kinetic constants calculated by BLI shown on Figures 4-6.

	L3/H2	Twist	Overall
FW-K	22.7%	27.3%	24.2%
FW-κN1	28.6%	36.4%	31.3%
FW-κN2	0%	22.2%	12.5%
Overall	22%	29%	24.7%

5.3.3.1. Kinetic constants of matured FW-κ candidates

For FW-κ, 22.7% (5 out of 22) of antibodies were able to significantly improve affinity towards Herceptin using the L3/H2 cassette approach. For the TWIST approach, 27.3% (3 out of 11) were able to significantly improve affinity towards Herceptin.

QLELFEPELDY had visible improvements on its off-rate through the cassette approach, and a mild improvement to both rates using the TWIST method (Figure 4). GKFRDWAPEKAFDY and GQWPFAHPEAGLDFDY improved their off-rate after mutations on LCDR3 and HCDR2 respectively, using the cassette method. AAGWLDTDEGRTMDY and DGSGSFLPVEDVSFDY improved both rates using the TWIST method. The latter contained 4 mutations that are the same as the CDR-null mutations imposed to FW-kN1 and FW-kN2. DRQRVLDLDTYEWAEEYFDV improved its on-rate via the cassette method. GQDWEPEFDY and SAQYWEPEFDY had no visible improvement. (Figure 4, Table 4).

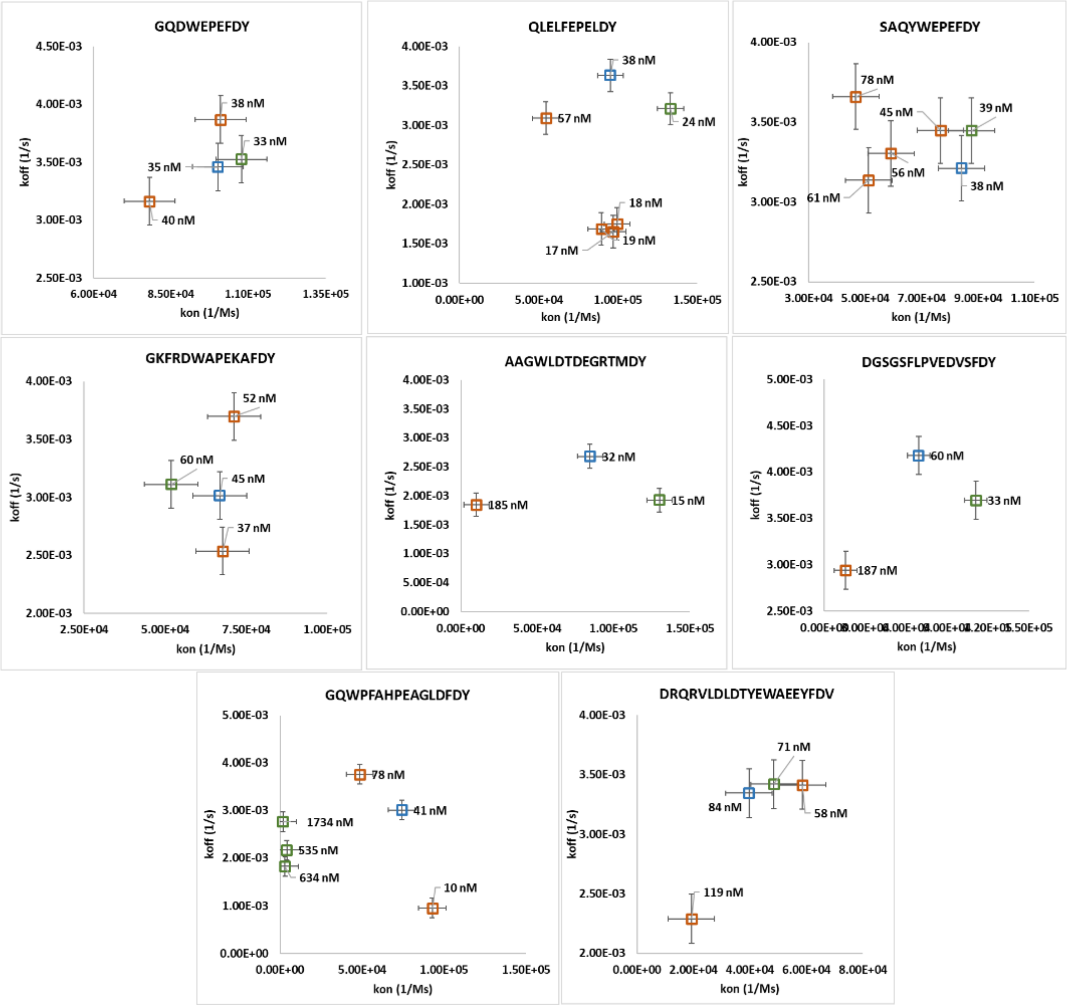


Figure 4 – Association and Dissociation constants of matured FW-κ antibodies. In blue: parental antibody. In Orange: L3/H2 cassette-derived antibodies. In Green: TWIST-derived antibodies. Candidates without any binding are not displayed (4 out of 22). k-off error = ±0.0002 (1/s); k-on error = ± 8226.1 (1/M.s)

5.3.3.2. Kinetic constants of matured FW-κN1 candidates

For FW-κN1, 28.6% (6 out of 21) of antibodies were able to significantly improve affinity towards Herceptin using the L3/H2 cassette method. For the TWIST method, 36.4% (4 out of 11) were able to significantly improve affinity towards Herceptin (Figure 5, Table 7). For AQSPFDWADFDY, the L117W mutation using the cassette approach greatly improved the on-rate at the expense of a faster off-rate. EGSYKHAEEAFDY saw an improvement in both constants, regardless of the method. DGGPYVQFPEAFDY saw both rates improve using the TWIST method. SPVPWSPYGDDLSFDY improved its on-rate dramatically via a S115Y mutation discovered through the cassette approach. DRWGGWDHAAEYLFDY saw the biggest increase in affinity due to a big improvement to its off-rate, using the TWIST method. Interestingly, the 5 residues mutated are equal to the original residues of FW-κ sequence before the CDR-null mutations – i.e. equal to germline residues. DKEGDGYDYVTYAGYFDV saw slight improvements to its on-rate at the expense of a faster off-rate using the TWIST method. AQGDYLPDDAFDY, DTDVLTYSFGDYSFDY and WADGGAPDYYPQEYELGFDV had no visible improvements (Figure 5, Table 5).

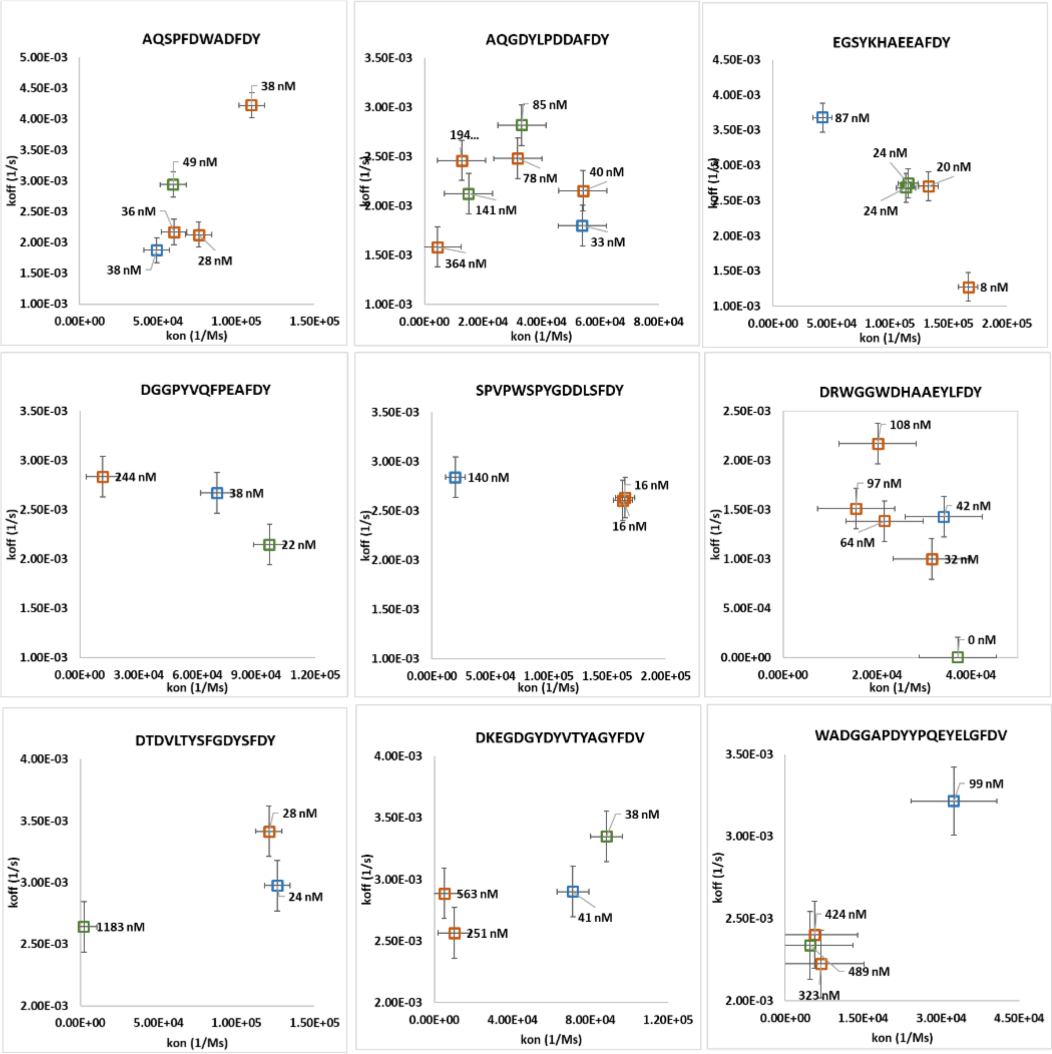


Figure 5 - Association and Dissociation constants of matured FW-κN1 candidates. In blue: parental antibody. In Orange: L3/H2 cassette-derived antibodies. In Green: TWIST-derived antibodies. Candidates without any binding are not displayed (1 in 21). k-off error = ±0.0002 (1/s); k-on error = ± 8226.1 (1/M.s)

5.3.3.3. Kinetic constants of matured FW-κN2 candidates

For FW-κN2, no antibodies were able to significantly improve affinity towards Herceptin using the L3/H2 cassette approach. For the TWIST approach, 22.2% (2 out of 9) were able to significantly improve affinity towards Herceptin. (Figure 6, Table 7). For TYGDYYSLESMDY, several HCDR2 mutations using the cassette approach greatly improved the on-rate at the expense of a faster off-rate, while the TWIST method was able to improve both constants. For THWPHLGGLEYFTYYPYMDV, the major improvement was at the off-rate level. The remaining candidates did not improve. (Figure 6, Table 6).

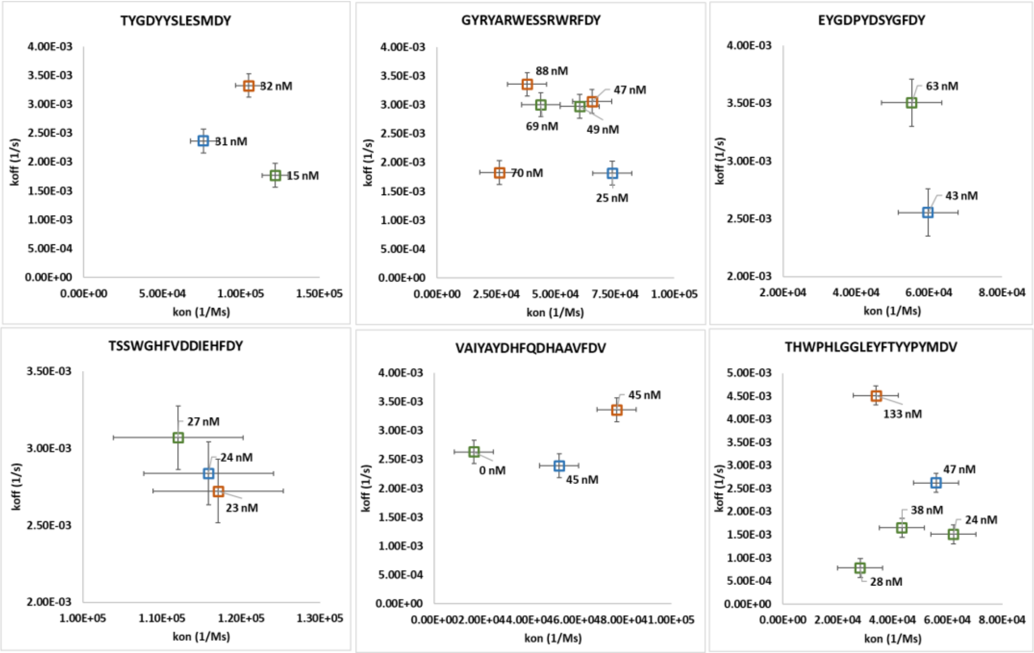


Figure 6 - Association and Dissociation constants of matured FW-κN2 candidates. In blue: parental antibody. In Orange: L3/H2 cassette-derived antibodies. In Green: TWIST-derived antibodies. k-off error = ±0.0002 (1/s); k-on error = ± 8226.1 (1/M.s)

5.3.4. Developability of matured selected candidates

The sucesfully matured candidates were also analysed by DSF, SEC and HIC to uncover their developability characteristics (Table 8-9).

5.3.4.1. Developability of successfully matured FW-κ candidates

Modifications to the HCDR2 sequence using the cassette approach in QLELFEPELDY antibodies led to losses in thermal stability of the Fab region, as shwon by the decrease in Tm2 values. On the other hand the L118W mutation led to a big increase in stability and hidrophobicity decrease. It's presence was able to compensate for the loss of stability driven by mutations on HCDR2 and led to a clone with low hidrophobicity and Tm2 = 77.5°C. The TWIST approach only slightly reduced thermal stability by 1°C, but its tendency to aggregate (HWMS = 8.5%) is a cause for concern. GKFRDWAPEKAFDY and GQWPFAHPEAGLDFDY both had mild improvements to their stability, ending with Tm2 = 83°C and Tm2 = 81.5°C. However, while GKFRDWAPEKAFDY improved its hidrophobicity profile ([AS] = 1.11 M), GQWPFAHPEAGLDFDY saw a big increase in hidrophobicity ([AS] = 0.77 M), an expected behaviour due to the introduction of four hidrophobic residues in "TAHGYY" in the place of the more neutral SGSGGS sequence. DRQRVLDLDTYEWAEEYFDV, which was also improved via the cassette method, had a improvement in it's stability and ended up with a Tm2 = 83.5°C. AAGWLDTDEGRTMDY and DGSGSFLPVEDVSFDY, which were both improved using the TWIST method, saw their Fab thermal stability drop to 75°C, but without compromising hidrophobicity and aggregation profiles.

Table 8 – Developability of matured FW-κ, FW-κN1 candidates. ΔTm2 and Δ represent the difference to the parental. Whenever Tm1 and Tm2 are not discernible – i.e. they are overlapped –, then Tm1 = Tm2**.** Hidrophobicity profiles and percentage of high molecular weight species (HMWS) below the acceptable threshold ([AS] < 0.8 M and HMWS > 5%, respectively) are indicated in yellow.

	Parental HCDR3	Maturation Method	LCDR1	LCDR3	HCDR1	HCDR2	Octet	DSF			HIC		SEC	
			ASQSISSYLN	**QQSYSTPLT**	**FTFSSYAMS**	**AISGSGGSTYYADSVKG**	KD (nM)	Tm1 (°C)	Tm2 (°C)	Δ Tm2	[AS] (M)	Δ	HMWS (%)	MP (%)
FW-κ	QLELFEPELDY	L3/H2	ASQSISSYLN	QQSYSTPLT	FTFSSYAMS	AIYSSAGHRYYADSVKG	18	68.0	71.00	-5.00	0.99	0.14	0.8	99.2
	QLELFEPELDY	L3/H2	ASQSISSYLN		FTFSSYAMS	EISSSGGSTRYADSVKG	17	69.0	73.50	-2.50	0.87	0.02	4.7	95.3
	QLELFEPELDY	L3/H2	ASQSISSYLN	QQSYSTPWT	FTFSSYAMS	EISSSGGSTRYADSVKG	19	68.0	77.50	1.50	1.01	0.16	4.0	96.0
	QLELFEPELDY	**TWIST**	ASTSISSYLN	QQSQSTPLT	FAFSSYAMS	AISGSGGSTQYASSVSG	24	69.0	75.00	-1.00	0.89	0.04	8.5	91.5
	AAGWLDTDEGRTMDY	**TWIST**	ASQSISSDLN	QQSYSTPLT	FAFSSIAMS	AISGSGGSTKYADSVKG	15	68.5	75.00	-9.00	1.06	0.03	0.8	99.2
	DGSGSFLPVEDVSFDY	**TWIST**	ASTSISSYLN	QQSYSTPLT	FAFSSAAMS	AISGSGGSTYYASSVKG	33	69.0	75.00	-5.00	1.08	0.05	2.9	97.1
	GQWPFAHPEAGLDFDY	L3/H2	ASQSISSYLN	QQSYSTPYT	FTFSSYAMS	AITAHGYYTYYADSVKG	10	66.5	81.50	2.50	0.77	-0.12	1.8	98.2
	DRQRVLDLDTYEWAEEYFDV	L3/H2	ASQSISSYLN	QQSYSTPLT	FTFSSYAMS	GISGSGGSTYYADSVKG	58	67.0	82.50	3.50	1.03	0.06	3.2	96.8

	Parental HCDR3	Maturation Method	LCDR1	LCDR3	HCDR1	HCDR2	Octet	DSF			HIC		SEC	
			ASTSISSALN	**QQSASTPLT**	**FAFSSAAMS**	**AISGSGGSTSYASSVSG**	KD (nM)	Tm1 (°C)	Tm2 (°C)	Δ Tm2	[AS] (M)	Δ	HMWS (%)	MP (%)
FW-κN1	AQSPFDWADFDY	L3/H2	ASTSISSALN	QQSYSTPYT	FAFSSAAMS	AISGSGGSTSYASSVSG	28	67.5	78.5	10.50	1.05	0.18	1.8	98.2
	EGSYKHAEEAFDY	L3/H2	ASTSISSALN	QQVYSTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	20	67.0	76.5	0.50	1.15	0.06	1.8	98.2
	EGSYKHAEEAFDY	L3/H2	ASTSISSALN	QQSASTPLT	FAFSSAAMS	AISGSGGSTYYADSVKG	8	66.0	81.0	5.00	1.15	0.06	2.3	97.7
	EGSYKHAEEAFDY	**TWIST**	ASQSISSALN	QQSYSTPLT	FTFSSAAMS	AISGSGGSTYYADSVKG	24	69.0	81.0	5.00	1.09	0.00	2.7	97.3
	EGSYKHAEEAFDY	**TWIST**	ASQSISSELN	QQSISTPLT	FAFSSAAMS	AISGSGGSTYYADSVKG	24	69.0	78.5	2.50	1.09	0.00	1.8	98.2
	DGGPYVQFPEAFDY	**TWIST**	ASQSISSELN	QQSYSTPLT	FTFSSAAMS	AISGSGGSTSYADSVKG	22	69.5	69.5	1.50	0.96	-0.01	0.0	100.0
	SPVPWSPYGDDLSFDY	L3/H2	ASTSISSALN	QQSYSYPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	16	67.0	79.0	3.00	1.03	0.08	2.6	97.4
	SPVPWSPYGDDLSFDY	L3/H2	ASTSISSALN		FAFSSAAMS	AISGSGGSTSSASSVSG	16	66.5	75.0	-1.00	1.04	0.09	1.4	98.6
	DRWGGWDHAAEYLFDY	**TWIST**	ASQSISSALN	QQSYSTPLT	FAFSSYAMS	AISGSGGSTSYADSVKG	0.05	69.0	79.5	10.50	0.98	0.02	0.7	99.3
	DKEGDGYDYVTYAGYFDV	**TWIST**	ASTSISSALN	QQSYSTPLT	FAFSSYAMS	AISGSGGSTSYASSVKG	38	69.0	84.0	4.00	0.99	-0.01	2.1	97.9

Table 9 – Developability of matured FW-κN2 candidates. ΔTm2 and Δ represent the difference to the parental. Whenever Tm1 and Tm2 are not discernible – i.e. they are overlapped –, then Tm1 = Tm2**.** Hidrophobicity profiles and percentage of high molecular weight species (HMWS) below the acceptable threshold ([AS] < 0.8 M and HMWS > 5%, respectively) are indicated in yellow.

	Parental HCDR3	Maturation Method	LCDR1	LCDR3	HCDR1	HCDR2	Octet	DSF			HIC		SEC	
			ASTSISSALN	QQSASTPLT	FAFSSAAMS	AISGSGGSTSYASSVSG	KD (nM)	Tm1 (°C)	Tm2 (°C)	Δ Tm2	[AS] (M)	Δ	HMWS (%)	MP (%)
FW-κN2	TYGDYYSLESMDY	**TWIST**	AS**E**SISS**D**LN	QQS**G**STPLT	FAFSS**Y**AMS	AISGSGGSTSYA**K**SV**Y**G	15	69.5		0.50	0.78	-0.25	3.7	96.3
	THWPHLGGLEYFTYYPYMDV	**TWIST**	AS**E**SISSALN	QQSASTPLT	FAFSS**Y**AMS	AISGSGGST**D**YA**Y**SV**E**G	28	68.5	83.0	4.00	0.78	0.07	6.0	94.0

5.3.4.2. Developability of successfully matured FW-κN1 candidates

For AQSPFDWADFDY, the L117W and L117Y mutation using the cassette approach greatly improved stability. For EGSYKHAEEAFDY, since both the cassette and the TWIST method lead to the rescue of germline residues in the HCDR2, all candidates that did so were able to improve its thermal stability. DGGPYVQFPEAFDY candidates' characteristics remained mostly similar to it's parental. Tm1 and Tm2 were overlapped and peaked at 69.5°C, while the parental had a clear peak at 68°C, so we assumed a slight increase in thermal stability of the Fab domain. SPVPWSPYGDDLSFDY improved its on-rate dramatically via the cassette approach saw its thermal stabiltiy increase to 79°C in the best clone. DRWGGWDHAAEYLFDY, which saw the biggest increase in affinity of the project using the TWIST method, also increased its stability by 10.5 degrees, ending up with a Tm2 = 79.5°C. This is undoubtly related with the stability of germline residues. DKEGDGYDYVTYAGYFDV, which also incorporated germline residues after the TWIST method, also increased its stability to 84°C.

5.3.4.3. Developability of successfully matured FW-κN2 candidates

Finally, TYGDYYSLESMDY candidates' thermal stability did not change much in relation with the parental antibody. However, in the case of the TWIST method, the increased affinity came at the cost of increased hidrophobicity ([AS] = 0.78) in comparison with the cassette method ([AS] = 1.05).

5.4. Discussion

On this chapter, 23 different candidates, from 3 different frameworks (FW-κ, FW-κN1 and FW-κN2) were subjected to 2 distinct generalizable affinity maturation methods. A fully blind method aims to randomize the LCDR3 and the HCDR2 sequences, which have been shown to be very important for antigen binding.[2–4] The other method tries a semi-blind approach, where hotspots sitting in a structurally advantageous position for antigen binding are diversified with the hope of increasing contacts with the antigen.

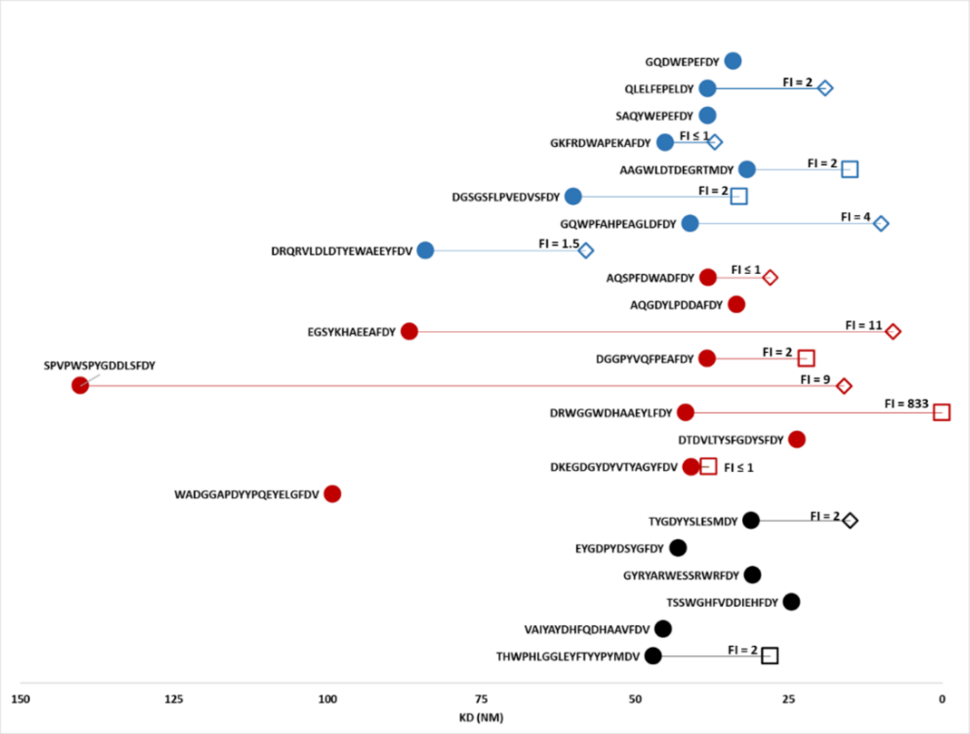


Figure 7 - Lead candidates for each HCDR3. In blue: FW-κ; In Red: FW-κN1; In black: FW-κN2 binders; Filled circles represent the parental sequences. Squares represent the lead candidate using the TWIST method. Diamonds represent the lead candidate using L3/H2 cassette method. The fold-increase (FI) in affinity for each mature candidate in comparison with the parental is also shown.

FW-κN1 and FW-κN2 were structurally primed to avoid clashes with the antigen molecule and, most importantly, to provide hotspots that can be acted upon to yield better affinity maturation results. The best mature candidate of each sucessfully mature parental are shown in Figure 7.

Of 14 lead candidates obtained on this work, only one achieved sub-nanomolar affinitites towards herceptin. (DRWGGWDHAAEYLFDY: K_D = 0.05 nM; FI = 833, FW-κN1, Table 8, Figure 7). As such, we wonder wether higher levels of stringency should have been used, either by increasing wash duration, prolonging off-rate selections, or by decreasing the concentration of biotinylated antigen (see section 5.5). Nonetheless, the use of bio-layer interferometry (BLI) to measure the kinetic constants of candidates may confound this analysis. Octet RED96 (which uses BLI technology) is a good solution to screen candidates and rank them acording to their affinity towards a given target, in a high-throughput manner. It also allows the determination of kinetic constants, up to a certain point. As it reaches the 1 nM mark, BLI starts losing sensibility, and other methods such as Surface Plasmon Resonance (SPR) are more suitable. More specically, accuracy tends to falter when determining the on-rates, more than when determining off-rates (Figures 4-6, k-on error = ± 8226.1 (1/M.s), k-off error = ±0.0002 (1/s);). This is also consistent with the literature and is often attributed to sensor-related artifacts and mass transport phenomena that are not observed in other techniques.[5] In that case, we can be more confident affinity gains that arise from improvements in the off-rate (e.g. Figure 5, DRWGGWDHAAEYLFDY and EGSYKHAEEAFDY). Finally, because our IgG-IgG interaction assay relies on the high-density immobilization of an irrelevant antibody on the sensor, it can further decrease sensibility and lead to underestimations of affinity (section 5.5.). As many candidates sit on the borderline of the instrument's sensibility, it is relevant to analyze the experiment from additional points of view.

The comparison of the different affinity maturation panning outcomes can be achieved by various ways: i) The rate of success for each condition, i.e. how many affinity maturation candidates have higher affinity towards the cognate antigen per condition; ii) Fold-increase in affinity towards the cognate antigen for each condition; iii) Quality of “lead” candidates generated per condition, i.e. do mature candidates display a very high affinity coupled with good developability characteristics, which would deem them ideal to move to the later stages of the antibody discovery pipeline?

In this chapter, the primed framework kN1 was shown to have a bigger rate of sucess (31.3%) than the control framework (FW-k, 24.2%). This was true regardless of the affinity maturation method used. Additionally, the highest affinity maturation rate of sucess was achieved when combining the primed library kN1 with the semi-blind affinity maturation method – 36.4%. (see overall success rates on section 5.3.3.1). This phenomenon is a direct consequence of the overall strategy of this work, that aimed to improve affinity maturation outcomes by structurally priming frameworks on the primary panning, and then directing diversity towards the primed hotspots that would likely be in contact with the antigen. FW-κN1 was also responsible for the biggest fold-increase (FI) in affinities in this body of work (EGSYKHAEEAFDY: K_D = 10 nM, FI = 11; SPVPWSPYGDDLSFDY: K_D = 16 nM, FI = 9; DRWGGWDHAAEYLFDY: K_D = 0.05 nM; FI = 833. Figure 7). This highlights the power of the strategy behind this work, where a small compromise in the affinity of primary binders is compensated by a bigger gain in affinity after affinity maturation (see aims and goals, section 1.4.). Other reasonable affinities were obtained from other candidates, but without a high fold-increase in affinity (GQWPFAHPEAGLDFDY: K_D = 8 nM, FI = 4; AAGWLDTDEGRTMDY: 15 nM, FI = 2; TYGDYYSLESMDY: 15 nM, FI = 2).

A curious trait about FW-kN1 binders is regarding their acquired mutations when coupled with the semi-blind method. These binders tended to acquire changes to CDR-null positions that reverted those positions back to the original germline residue that was mutated in the first place. For example, out of 8 possible positions, the lead candidate from EGSYKHAEEAFDY accumulated 6 mutations that are equal to the germline residues, while the remaining two positions were kept unchanged. This lead to a 11-fold increase in affinity (from 87 nM to 8 nM, Table 5 and Figure 7) and to noticeable improvements in the thermal stability of the Fab domain (Tm2 = 81°C, ΔTm2 = + 5°C, Table 8, FW-kN1). The biggest increase in affinity of this work came from DRWGGWDHAAEYLFDY, which also went through a partial germline reversion, by accumulating 5 mutations equal to germline residues. This lead to a 833-fold increase in affinity (from 60 nM to 0.05 nM, Table 5 and Figure 8), and to big improvements in thermal stability of the Fab domain (Tm2 = 79.5°C; ΔTm2 = + 10.5°C, Table 8, FW-kN1). A partial germline reversion was also seen on the lead candidates of DGGPYVQFPEAFDY and DKEGDGYDYVTYAGYFDV but without the same affinity gains (5 positions yielding a FI = 2, and 3 positions yielding a FI ≤ 1, respectively. Table 5 and Figure 7).

The big gains in thermal stability of EGSYKHAEEAFDY and DRWGGWDHAAEYLFDY do not come as a surprise, in light of the results we have shown in chapter 2, where we show that germline sequences are naturally optimized towards stability. But why did FW-κN1 often choose a germline residue over any other amino acid from the remaining chemical groups, when given the chance? Most likely because the eight germline residues chosen (Q48, Y53, Y113A, T283, Y287, Y315, D318, K321) are tipically involved in polar contacts and salt bridges. This enables FW-κN1 to increase its affinity towards the cognate antigen, whille simultaneously gaining a boost in thermal staility in the process. However, that same strategy does not seem to work with the κN2 primed library. This probably stems from two distinct reasons related with the experimental setup. The first reason is related with the off-rate selection condition on the first round of the affinity maturation panning. Altough primary binders from κN2 libraries had roughly the same affinity of other candidates (see section 4.3.3.1, Figure 11), their kinetics were substantially different. FW-κN2 binders did not have fast on-rates, but dissociated very slowly from the antigen, which conferred them similar K_D values to FW-κ and FW-κN1 binders, which had a mix of fast on-rates with fast off-rates (see section 4.3.3.1, Table 5). This means that FW-k and FW-κN1 binders are in a better position to further increase their affinity after the off-rate selection employed on this set of results (see results on section 5.3.2, and methods on section 5.5 for more information on off-rate experiments). Affinity maturation libraries derived from FW-κN2 parentals would most likely have benefited from having a more stringent on-rate selection – e.g. by decreasing the antigen concentration even further –, rather than a further optimization of their off-rates.

Coming back to FW-κN1, it is important to note that EGSYKHAEEAFDY, DRWGGWDHAAEYLFDY, and other FW-κN1 candidates were not found in FW-κ datasets. This means that while some germline residues were advantageous for the HCDR3 sequences in question, the simultaneous presence of all of the eight germline residues deemed these HCDR3 candidates unviable/weak. This opens an opportunity to increase the functional diversity of the original FW-κ libraries – i.e. selecting more diverse candidates without increasing theoretical randomization (for more information on the diversity conundrum see section 1.3.1, or refer to the following chapter 6, where the topic is approached with more depth). As stated in chapter 4, the HCDR3 randomization is equal for FW-κ and FW-κN1 (see section 4.3.1), but very different sequences will be sampled from those two very similar frameworks that differ in only 8 residues. Using primed libraries allows to sample extra candidates from the same diversity pool, and then approximate them to the germline by coupling those candidates with a semi-blind affinity maturation method, to achieve gains in affinity and stability.

5.5. Methods

Phage-display vectors

To generate phage particles displaying antibody fragments, a phagemid vector is used (see section 1.2.3. on chapter 1). The phagemid is based on the M13 filametous phage, and it encodes a Fab antibody fragment fused via an Amber stop codon (UAG) to a truncated pIII protein (Glycine-rich linker and CT domain, which anchors pIII in the phage coat). The phagemid also carries an ampicillin resistance gene and a M13 origin gene that triggers the packaging signal of the filamentous phage when combined with VSCM13 helper phage within infected bacteria.

Generating LCDR3 and HCDR2 randomized cassettes

The LCDR3 and HCDR2 sequences were randomized via PCR, by amplification of those regions with randomized TRIMoligos manufactured at EllaBiotech. Two different TRIMoligos will be used as primers for the PCR reaction. The forward primer has a constant sequence that is homologous to the LFR3 in the forward strand (see section 1.1.1. Figure 1) and to the first and second codon of LCDR3. The randomization proportion arises by controlling the stoichiometric ratio of codons during the TRIM gene synthesis.[6] The reverse primer has a constant sequence that is homologous to HFR3 in the reverse strand (see section 1.1.1. Figure 1). After the PCR reaction, the samples (100 µl) were prepared for electrophoresis gel and ran for 60 min at 100V in a 1% gel. The bands were excised and purified by gel extraction using the manufacturer's instructions (Wizard SV Gel and PCR Clean Up System, Promega). The purified DNA was digested using BamHI and BstBI and ligated with a phagemid vector digested with the same enzymes. The phagemid vector carries a dummy HCDR3 sequence. The phagemid vectors were transformed via electroporation into electrocompetent *E.coli* TG1 cells (Lucigen). Electroporated cells were recovered in SOC medium for 1 hour before being transferred into 2YT medium with 1% glucose and 100 µg.mL-1 of ampicillin (2YT/A/G) and incubated overnight at 25 °C, 200 rpm on incubator Innova 44. Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol. The HCDR3 parental sequences were cloned into the appropriate L3/H2-randomized phage-display vectors and transformed into TGF1+ cells by electroporation.

Affinity maturation libraries based on L3/H2-cassettes

The DNA of TG1+ cells containing L3/H2-randomized phagemids was extracted and digested using BlpI and BstBI. A total of 23 parental antibodies were digested using the same enzymes to clone the parental HCDR3 sequences into the phagemid vectors encoding an affinity maturation framework randomized on LCDR3 and HCDR2.

Affinity maturation libraries based on TWIST

A total of 23 HCDR3 sequences from parental antibodies were cloned into in-house phagemid vectors encoding an affinity maturation framework diversified on LCDR1, LCDR3, HCDR1 and HCDR2. These were designed *in-house* and manufactured by Twist Bioscience and follow the design shown on Figure 4. The phagemid vectors were transformed via electroporation into electrocompetent *E.coli* TG1 cells (Lucigen). Electroporated cells were recovered in SOC medium for 1 hour before being transferred into 2YT medium with 1% glucose and 100 µg.mL-1 of ampicillin (2YT/A/G) and incubated overnight at 25 °C, 200 rpm on incubator Innova 44. Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol.

Phage production

A sample from each glycerol stock was taken to start a 25 mL culture in 2YT/A/G at OD_{600} = 0.1 and grown to mid-log phase (OD600 = 0.5) before being infected with helper phage VCSM13 (Agilent Technologies). Cells were then incubated firstly for 30 min at 37°C in a water bath, and then for 30 min at 37°C shaking at 250 rpm. The infected bacteria were then centrifuged and transferred into a 40 mL culture of 2YT supplemented with 100 µg/mL Ampicillin, 50 µg/mL Kanamycin and 0.25 mM IPTG**.** Phage production ensued overnight at 22°C and 180 rpm. The cultures were centrifuged to remove the cells and the phage-rich supernatant collected into sterile 50 mL Falcon tubes and kept on ice. Phages are precipitated by adding 10 mL of ice cold 20% (w/v) PEG 6K in 2.5 M NaCl into the 40 mL of supernatant, 1 hour on ice. After this time, the precipitated solutions were centrifuged at 4000 *g* and 4 ˚C for 30 min (Eppendorf, Ref: 5810 R). The supernatant was discarded and the precipitated phage pellets were re-suspended with 1 mL of sterile phosphate buffered saline (PBS) and transferred to 1.5 mL Eppendorf tubes. The tubes were then rotated for 30 min on a rotating wheel at 4˚C and then centrifuged at 12 000 *g* and 10 ˚C for 5 min (Eppendorf, Ref: 5810 R) to remove further bacterial debris. Supernatants were filtered into cryovials containing 700 uL of PBS:Glycerol 50:50% (for a final [Glycerol] of 20% v/v).

Herceptin Biotinylation

1mg of Biotin was dissolved in 166 µL H20 to do a 10 mM solution. Then, 40.5 uL of Sulfo-NHS-SS-Biotin was used to biotinylate 0.5 mL of 10 mg/mL Herceptin. The procedure is done following the directions of EZ-Link Sulfo-NHS-LC-Biotin kit (ThermoScientific A39257 21335). Samples were incubated 1 hour at room temperature. The samples were then passed through a Zeba Spin desalting column, 7K MWCO, 2mL (Thermoscientific, #89891, #QL227761).

Phage display panning selections

Phage display protocols were performed using the automated liquid handling functionalities of the KingFisher™ Flex Purification System (ThermoFisher, Catalog number: 5400610, Figure 8)

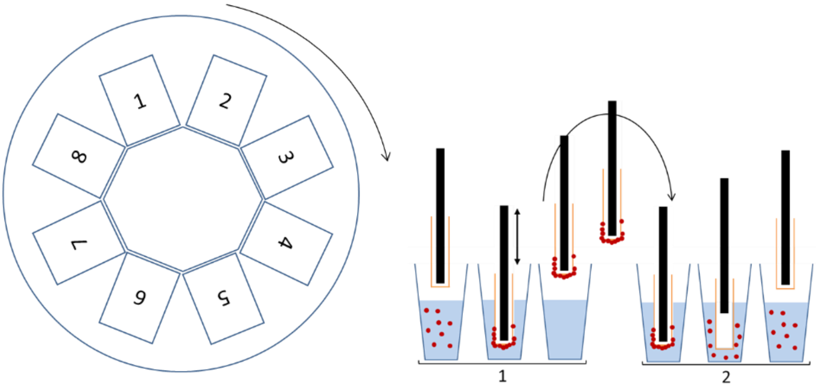


Figure 8 – Schematic representation of kingfisher operation. Streptavidin magnetic-beads bound are put in contact with the phage-antigen-biotin complexes in solution, which are then transferred between wells by plastic-covered rod-shaped magnets. The capture and release movements during transfer and washing protocols are software-driven, and all the parameters such as time, position, and frequency shaking movements can be customized**. Adapted from:** Ch'ng, A.C.W., Ahmad, A., Konthur, Z., and Lim, T.S. (2019). A High-Throughput Magnetic Nanoparticle-Based Semi-Automated Antibody Phage Display Biopanning. In Human Monoclonal Antibodies, M. Steinitz, ed. (New York, NY: Springer New York), pp. 377–400

A total of 5 x 10^9 infectious phages corresponding to each primary library were blocked for 1 h in PBS + 0.05% Tween (PBST) supplemented with 0.05% of BSA, in 96 DeepWell plates (Thermo Scientific™ 95040450), followed by in-solution deselection on streptavidin-coated magnetic beads (Dynabeads, Invitrogen, Cat # 112–06) for 30 min, to remove sticky phages that bind to streptavidin beads.

Biotinylated Herceptin was added in the corresponding well to each well of sticky-depleted phages and incubated 1h at room temperature (RT) on a micro-plate table. The antigen-antibody complexes were captured from the deep well plates by the streptavidin-coated magnetic beads bound to the KingFisher magnetic rods and transferred to the washing plates sequentially, as shown in Figure 8. The washing of bead-antigen-phage complexes was accomplished by washes of increasing shaking vigor, stringency, and duration, on PBST and PBS.

Affinity maturation pannings will have two different rounds. The first round aims to improve the *k-off* of binders using a off-rate selection protocol coupled with a conservative concentration of Herceptin (50 nM, equal to the average affinity of binders in the third round. section 4.3.). From the pool of good off-rate binders, the second round will try to enrich the sample in binders with better on-rates using a very low Herceptin concentration. The washes in both rounds will be the same as those used in the third round of the primary panning. (Figure 9)

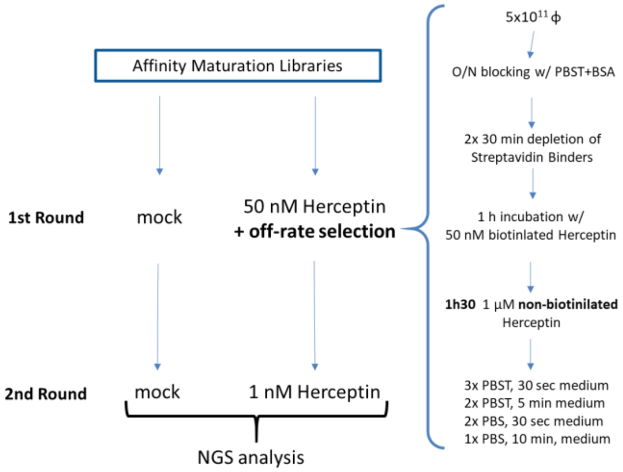


Figure 9 – Schematic representation of affinity maturation rounds

The off-rate selection used in the first round takes advantadge of the liquid selection panning system, that uses a biotinilated antigen.

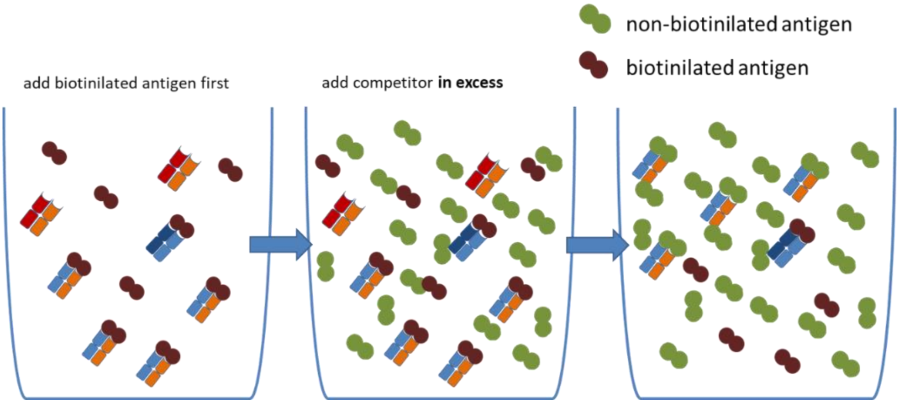


Figure 10 – Schematic representation of off-rate selection protocol

First we add the biotinilated antigen, and then we add a competitor in excess afterwards. Fast off-rate binders will unbind from the biotinilated antigen and likely bind to the excess competitor. Slow off-rate binders will remain bound to the biotinilated antigen overtime. Hence, we can selectively recover slow off-rate binders with streptavidin/neutravidin beads while the fast-off rate binders will remain bound to the un-captured antigen (Figure 10).

To effectively determine which concentrations of competitor antigen we should use, and for how long, we simulated an off-rate experiment using equation 1, from Zahnd *et al.*[1]

Equation *1*. θ_i is the fraction of a given library member (ligand i) that is still bound to biotinylated antigen, τi is dimensional time (t) multiplied by the dissociation rate constant (k_{off} i), βi is the concentration of biotinylated antigen (B) divided by the equilibrium dissociation constant of this complex (K_Di) and µi is the concentration of unbiotinylated competitor (U) divided by K_Di.

$$\mathrm{R} = \frac{\theta_2}{\theta_1} = \frac{e^{-\tau_2} + \frac{\beta_2}{\beta_2 + \mu_2 + 1} \times [1 - e^{-\tau_2} + (\beta_2 + \mu_2)^{-1} \left(e^{-(\beta_2+\mu_2+1)\tau_2} - e^{-\tau_2}\right)]}{e^{-\tau_1} + \frac{\beta_1}{\beta_1 + \mu_1 + 1} \times [1 - e^{-\tau_1} + (\beta_1 + \mu_1)^{-1} \left(e^{-(\beta_1+\mu_1+1)\tau_1} - e^{-\tau_1}\right)]}$$

$$\beta_i = \frac{B}{K_{Di}}; \qquad \mu_i = \frac{U}{K_{Di}}; \qquad \tau_i = t \times koffi$$

Equation 1 represents enrichment ratio (R) of a binary 'library', comprised of just two unique members, θ_2 and θ_1 . The model assumes that θ_2 and θ_1 share the same on-rate, but have unequal dissociation constants, where $k_{off}\,\theta_2 < k_{off}\,\theta_1$, and thus, $K_D\,\theta_2 < K_D\,\theta_1$. For simplicity, we considered θ_1 to be the estimated average affinity of our phage pools towards Herceptin and defined θ_2 as our target affinity. We performed three estimations with different θ_1 values (50 nM, 25 nM and 15 nM) and determined that a 90 minute off-rate selection procol with a competitor concentration (U = 1 µM) and a biotinilated concentration (B = 10 nM) for about 90 minutes was enough to reach between 9-fold to 17-fold enrichment in θ_2 binders in our sample. (Figure 11)

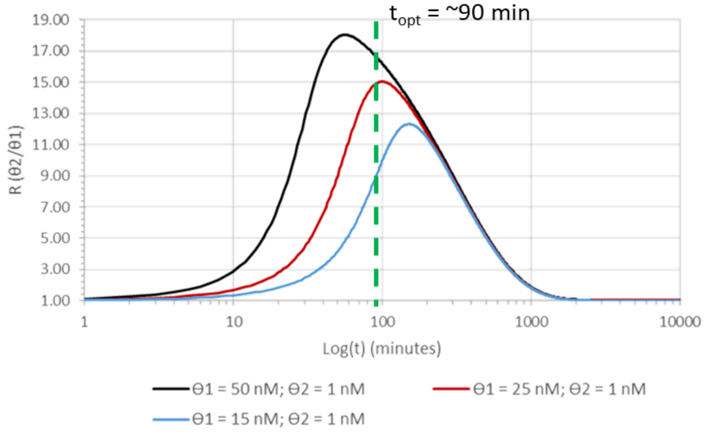


Figure 11 - off-rate simulation of affinity maturation phage pools

Bacterial infection and phage amplification

At the end of each wash protocol of each round, surviving phages were dissociated from the complexes with glycine buffer (10 mM glycine-HCl, pH 2.0) before neutralization with 200 µL Tris-HCl pH 7.5 and infection of a 20 mL mid-log *E.coli* TG1 culture (OD600 = 0.5). The cultures were incubated for 45 min in a water bath at 37ºC before being inoculated into 100 mL 2YT/A/G in 250 ml Erlenmeyer's and let to grow overnight at 25ºC, 150 rpm (Innova 44R, New Brunswick Scientific). Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol. These can be used to produce new phages for subsequent rounds, or to have their DNA extracted for NGS analysis.

DNA preparation and NGS analysis

Plasmid DNA was isolated directly from the phage-infected cells from the selection round of interest using the GeneJET Plasmid Miniprep Kit (Thermo Scientific™, K0502). Isolated dsDNA was quantified on the Qubit 3.0 fluorometer using the Qubit® dsDNA HS kit (Invitrogen™ Q32851). The generation of V_L:V_H amplicon for sequencing was generated through two PCRs. To amplify the region of interest and to insert the adapter regions for the NGS, the initial PCR utilized a forward primer specific to the vector leader sequence prior to LCDR1 and, since we did not need HCDR3 information, a reverse primer downstream of HCDR2. The second PCR inserted the TruSeq universal adapter and the indexes, used to distinguish between different samples (i.e. libraries). Samples were quantified in Qubit 3.0, pooled in equimolar proportions, and ran on an electrophoresis gel. Bands with the appropriate size were excised, purified using the Wizard SV Gel and PCR Clean Up System (Promega, A9281), and quantified on Qubit 3.0. The pool was diluted to a final concentration of 4 nM, spiked with 20% PhiX (Illumina; FC-110-3001), denatured for 5 min in 0.1 N of NaOH (5 µL of DNA+PhiX at 4 nM mixed with 5 µL 0.2 N of NaOH), diluted in HT buffer (provided on the NGS kit; kit details, ahead) to 7.2 pM and sequenced on the Illumina MiSeq platform using the 500 cycle V2 kit (Illumina; MS-102-2003). The forward read was 230 bp in length while the reverse read was 270 bp. R1 retrieves information on LCDR1, LCDR2 (non-diversified), and LCDR3. R2 retrieves information on HCDR1 and HCDR2. R1 and R2 are matched using their specific cluster coordinates, as explained on Annex A. This generates concatenated R1+R2 reads that allow for the correlation of V_L:V_H information. The data analysis of the NGS FastQ output files was performed as described previously.[7] For the panning output of each library, 1 x 10^5 sequences were analyzed using the fixed-by-design flanking sequences on the boundary of diversified positions as template to locate and segment out mutations. Full CDR sequences were reconstructed by coupling the regions fixed-by-design with the information on the diversified regions.

<u>IgG expression</u>

The expression plasmids were ordered from ThermoFisher's GeneArt platform. The Light-chain (LC) and Heavy-chain (HC) of each IgG were ordered separately and transfected simultaneously (in a 1:1 ratio) with Polyethylenimine (PEI, in a 4:1 ratio with DNA) into 100 x 10^6 human embryonic kidney-293T (HEK- 293T) cells in 18 mL of FreeStyle^{TM} 293 Expression Medium (Life Technologies®). After 4 hours, an additional 20 mL of medium are added to the cells for a final cell concentration of 2.5 x 10^6 cells/mL. Transiently transfected cell cultures were incubated for 4 days in humidified atmosphere of 5% CO_2, 37°C and 140 rpm. After 4 days in culture, transfected cells are centrifuged at 300g for 10 minutes, and their supernatant collected, and vacuum filtered using 0.22 µm pore Steriflips (FisherScientific). The supernatant can be stored at 4°C for a week or at -20°C for extended periods.

<u>IgG purification</u>

IgG purification was performed by Affinity Ligand Chromatography, on Tecan Freedom EVO 200 (equipped with a Liquid Handling arm with 8 stainless steel tips, syringes of 1 mL and TeChrom, to enable fast IgG purification) using MabSelect Sure RoboColumns (Repligen; Ref.: PN 01050408R. Total Column Volumn (CV) = 200 µL). Phosphate Saline Buffer (PBS, pH 7.0) was used as the equilibration buffer. Samples were loaded 1 mL at a time, for a total final load of 35 mL. Retrieval of IgGs was achieved by isocratic elution using 5 CV of 50 mM Citrate-NaCl pH 3.0, for a final eluted volume of 1mL. The pH is neutralized by the addition of 150 µL of 1M Tris-HCL pH 9.0. The sample is then filtered trough a 0.22 µm filter pore using a syringe and stored at -20°C. Final volume = 1.15 mL.

IgG quantification

IgGs were quantified via HPLC Affinity Ligand Chromatography (HPLC-ALC), using a POROS™ CaptureSelect™ CH1-XL Affinity HPLC Column 2.1 x 30 mm, coupled to an Agilent 1260 Infinity II (Agilent Technologies). Separation of protein species was achieved using a flow rate of 2 mL/min and detection at 210 nm. Samples are injected directly without any previous dilution (injection volume = 50 µL), and the following method on Table 10 is employed for each individual injection:

Table *10* – ALC-HPLC method. Mobile Phase A: 10 mM NaH_2PO_4, 150 mM NaCl, pH 7.5; Mobile Phase B: 10 mM HCl, 150 mM NaCl, pH 2.0;

Time after injection (in minutes)	Mobile Phase A (in %)	Mobile Phase B (in %)
0	100	0
1.87	100	0
1.88	0	100
4.38	0	100
4.39	100	0

mAb peaks are manually integrated to calculate the Peak Area. Antibody concentration is calculated according to Equation 1.

Equation 1: $$C_A = Peak\ Area_A \times \left(\frac{C_{IS}}{Peak\ Area_{IS}}\right) \times \left(\frac{1}{\frac{RRF_A}{RRF_{IS}}}\right)$$

An internal standard (IS) IgG with known concentration was used to generate an internal response factor (RRF_{IS} = Peak Area $_{IS}$/ Concentration $_{IS}$). Each sample concentration (C_A) was calculated as shown in Rome, K. & McIntyre, A. (2012)[1], by taking into account the concentration of IS (C_{IS}) and by comparing the sample's RRF (RRF_A) with the RRF of IS (RRF_{IS}). (Equation 1)

Size-exclusion chromatography

50 mM Sodium Phosphate pH 6.5 was used to dilute IgG samples to a final concentration of 1 mg/mL. Each candidate was analyzed by size exclusion chromatography on a SEC BEH 200 column (Waters, 200 Å, 1.7 µm, 4.6 mm x 150mm) using an Agilent 1260 Infinity II HPLC system, equipped with a multi-wavelength detector. A total run time of 35 minutes per sample was employed, after a 2 µg injection of each sample The mobile phase was 50 mM Sodium Phosphate pH 6.0 + 400 mM sodium perchlorate pH 6.0. Separation of protein species according to their molecular weight was achieved by applying an isocratic elution using a flow rate of 0.4 mL/min and detection at 210 nm. Peak integration of IgG monomers was done at a retention time around 20 minutes; these are referred to as “main peaks”. Peaks and/or shoulders before the “main peak” are indicative of aggregation and referred to as “high molecular weight species” (HMWs). Peaks and/or shoulders after the main peak are indicative of fragmentation of the IgG monomer and designated “low molecular weight species (LMWs)

Hydrophobic-interaction chromatography

The hydrophobic profile of each candidate was analyzed by hydrophobic-interaction chromatography (HIC) in a TSKgel Butyl-NPR column (4.6 mm ID x 35 mm L) (Tosoh Biosciences). PBS was used to dilute the samples to 1 mg/mL. The mobile phase A was composed by 20 mM His/HCl, pH = 6.0 containing 1.5 M AS. Gradient elution of protein species was achieved by a gradual buffer replacement of mobile phase A with 20 mM His/HCl, pH 6.0 (mobile phase B). The gradient is 20 CV in length and has a slope of – 0.103 M AS per minute. A calibration curve was employed, where the retention time of reference standards was plotted against concentration of AS to calculate the hydrophobicity of the protein molecules.

Differential Scanning Fluorometry

Differential Scanning Fluorometry was performed in BioRad CFX96. Samples were diluted to 0.3 mg/mL (Vf = 50uL) in 43uL of PBS, to which 7 µL of SYPRO orange (previously prepared) was added. Sypro orange preparation was done by diluting the 5000x stock, by pipetting 1.4 µL from the stock solution into 1 mL of H20. The reaction was performed with a temperature increment of 0.5 ºC/min, from 25 °C to 100 °C.

Octet affinity measurements of anti-Herceptin antibodies

All kinetic assays were performed on Octet® RED96 (ForteBio), using 96-well plates (Corning), at 30 °C and 1000 rpm orbital shake speed. Samples were diluted in freshly prepared by diluting 10× Kinetic buffer (PALL) 1:9 in PBS (Gibco). Herceptin, which is a commercial IgG, was loaded either into anti-human Fc (AHC) Octet biosensors tips, by submerging them for 40 seconds in a 200 µL solution of Herceptin at 0.05 mg/mL. This is followed by a baseline step of 1 minute in kinetic buffer. Since mAbs will be assayed for their affinity towards Herceptin, we need to perform a saturation step, to make sure that the AHC biosensor is inaccessible to the mAbs assayed on the following steps. The saturation is achieved by submerging the Herceptin-loaded biosensors in a 200 µL solution of irrelevant-mAb0 at 0.2 mg./mL. This step is also followed by a baseline step of 1 minute in kinetic buffer. The Herceptin-loaded biosensors are then submerged in wells containing different concentrations of mAbs for 900 seconds – the association phase –, followed by a 1800 seconds dissociation phase in kinetic buffer. The mAb-loaded tips were also dipped in wells that only contained kinetic buffer, to serve as the basal reference signal used in the estimation of the affinity parameters step.

Estimation of interaction rates and affinity parameters

Binding sensograms were first aligned at the beginning of the association phase, and following the single reference subtraction, they were globally fit to a 1:1 binding model, were a single k-on and k-off is calculated for all binding sensograms for every concentration tested.

5.6. Acknowledgments

The author – Jorge Sampaio – was responsible for all the work in this chapter. The author acknowledges the valuable support of André F. Faustino in the troubleshooting of affinity maturation libraries generation, as well as Elísio Silva for valuable inputs on the bio-layer interferometry assays for kinetic constants determination.

5.7. References of Chapter 5

1. Zahnd, C., Sarkar, C. A. & Plückthun, A. Computational analysis of off-rate selection experiments to optimize affinity maturation by directed evolution. *Protein Eng Des Sel* **23**, 175–184 (2010).
2. Tiller, T. *et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs* **5**, 445–470 (2013).
3. Kunik, V. & Ofran, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering, Design and Selection* **26**, 599–609 (2013).
4. Persson, H. *et al.* CDR-H3 Diversity Is Not Required for Antigen Recognition by Synthetic Antibodies. *Journal of Molecular Biology* **425**, 803–811 (2013).
5. Estep, P. *et al.* High throughput solution-based measurement of antibody-antigen affinity and epitope binning. *MAbs* **5**, 270–278 (2013).

6. Shim, H. Synthetic approach to the generation of antibody diversity. *BMB Reports* **48**, 489–494 (2015).
7. Liu, G. *et al.* Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2020).

Chapter 6 – Final Remarks

Therapeutic monoclonal antibodies (mAbs) are one of the main drivers of revenue of the pharmaceutical market. The global mAbs market is valued at around 115 billion US dollars and is expected to grow to about 300 billion US dollars until 2025.[1] As of November 2020, 88 mAb products were under late-stage clinical investigation (6 for COVID-19).[2] In May 2021, it was announced that the FDA granted marketing approval for its 100th mAb product.[3] Monoclonal antibody discovery production was firstly achieved by hybridoma technology, on the seminal paper by Kohler and Milstein.[4] Iterations to the mAb discovery process, such as chimeric recombinant antibodies[5] and humanized antibodies[6] were developed to minimize the immunogenicity that arises from using animal systems to discover mAbs. Ultimately, *in vitro* discovery platforms such as phage-display were developed to answered to the limitations of animal immunization approaches, from operational and ethical standpoints. Phage-display provides additional experimental control (e.g. epitope-specific selections), parallelization, automation, and miniaturization. To date, a total of 14 antibodies derived from phage display were approved for use in the clinic, and more than 70 have undergone or are undergoing clinical evaluation (Table 1).[7,8]

Regardless of the origin and platform used, antibodies generated against a given target may have room for improvement. Using *in vitro* affinity maturation libraries aims to surpass the throughput limitations of classical X-ray crystallography affinity maturation approaches, by providing a generalizable approach (or blind) that can be applied to many candidates. Cassette randomization[9] and Look-through mutagenesis (LTM)[10] aim to capture beneficial mutations in a generalizable way, but they do not always assure that synergistic mutations are found and may not respect the structural constraints of the IgG molecule and lead to inconsistent results. Ideally, innovative affinity maturation methods should be generalizable to provide high-throughput results while maintaining a certain degree of specificity towards the antibody structure being considered.

As such they require attention to be paid to specific regions, such as the ones likely to be in contact with the antigen, or regions that influence the antibodies' structural integrity and overall developability.

In this thesis, we developed a semi-blind affinity maturation approach using structurally primed primary libraries, to reach a compromise between generalization and precision. The primed libraries showed similar performance in primary pannings when compared to the control library and outperformed the control library in an affinity maturation setting, specially when combined with our newly developed affinity maturation design (TWIST approach). This and other relevant advances related to this work will be further explored in the following paragraphs.

6.1. Germline Sequences are Optimized Towards Stability

Germline sequences encompass all the possible sequences that arise from V(D)J recombination processes, that have yet to be presented to an antigen and, hence, have not suffered selective pressure nor acquired somatic mutations upon antigen contact. Intuitively, having a large repertoire of germline antibodies increases the chance of finding paratopes that bind to the antigen. Some estimates of naïve human repertoires go from around 10^{11}-10^{12} up to 10^{15} - 10^{18} sequences.[11–14] However, these number seem unlikely to occur in a single individual, since these numbers surpass by far the total number of cells in the human body (10^{13}), let alone the number of circulating peripheral naïve mature B-cells (CD27−/IgD+) at any point in time (10^{9}).[15,16] The overestimation likely occurs because many individuals are used in the naïve datasets, which increase the probability of finding unique sequences. Rather than each individual having a $>10^{15}$ repertoire, this value is the representation of the sum of the overlapping repertoires within the total human population.[13,17] The actual diversity of a single healthy human individual is estimated to be around $>10^{7}$ unique sequences.[17,18] However, since sequences found in circulation are clearly biased to certain subsets of V_H families, and κ and λ families, the functional diversity is expected to be a fraction of this value.[19–21] This means that the immune system probably has far less antibody sequences than the number of epitopes on foreign antigens to which one could be exposed.

To combat this, germline sequences have adapted to the “shape space” of antigen epitopes and evolved to be polyreactive.[22,23] In this model, each individual antibody structure is able to bind to a given structural shape. This means that a single antibody structure may recognize several unrelated epitopes, provided that they present similar shapes. This structural redundancy is most commonly referred to as polyreactivity or polyspecificity, and has been vastly associated with antibodies triggered early in the response (e.g. IgMs) and germline sequences.[17,24–26] This mechanism has been recently termed “conformation flexibility hypothesis”. It suggests that germline gene-coded antibodies retain a degree of structural plasticity in their backbone in order to maximize the number of different unrelated antigens that they can recognize. A study of 137 therapeutic mAbs showed that the absence of somatic mutations in germline sequences is a good predictor of polyreactivity.[24] Older studies also report that poly-specific antibodies retain a larger amount of germline sequences than more specific antibodies.[27,28]

Hence, germline sequences provide poly-reactive surfaces that can bind to a wide range of structural antigen epitopes with sufficient affinity to initiate an immune response. This allows for a limited diversity repertoire to screen a panel of epitopes that is potentially bigger than its sequence-encoded diversity, in a resource efficient manner. Our results show that besides polyreactivity, germline sequences may also be optimized towards stability. When generating primed frameoworks derived from FW-κ and FW-λ, we saw that all the mutations that deviated from the germline led to a decrease in thermal stability, regardless of framework and/or combination of mutations. A fraction of those mutations showed no effect on thermal stability unless combined with other mutations. Decreases in thermal stability were confined to the

Fab domain (Tm2, section 2.3. Table 10-11), and were sometimes accompanied by improvements in hydrophobicity (for FW-κ, Table 10) and aggregation (for FW-λ, Table 11). The severity of thermal de-stabilization varies greatly and tended to increase with the number of mutations (Table 10-11). Our data also confirms literature reports that the stability of antigen-binding Fab domains plays a crucial role in the overall stability of the IgG.[29,30].

High thermal stability of a mAb indicates a well-packed structure that requires more energy to unfold, and thus serves as a good predictor of robustness to destabilizing factors such as temperature, pH and pressure. Indeed, it was shown that stably folded antibodies have a lower tendency to aggregate.[29,31–34] Moreover, mAbs with worse thermal stabilities were reported to be poorly expressed.[30,34].Taking into account that around 20% of B-lymphocytes in the peripheral blood make polyreactive antibodies,[25] – and that these constitute the first line of adaptive immunities defense –, it makes sense that these antibodies are also highly stable. Unstable primary antibodies would signify a loss in functional diversity and inefficient use of cellular resources.

These results also raised important questions about the design of antibody libraries and their expected quality and effective/functional clonal diversity. Extensive developability tests are done to antibody frameworks to ensure they can accommodate variations to their CDRs and still provide viable candidates.[35–37] Many state-of-the-art libraries rely on the extensive randomization of all CDRs simultaneously to discover binders against an antigen of interest. We postulate that extensive simultaneous randomization of several CDR-loops may generate very unstable clones that will not be selected during phage-display or other *in vitro* mAb-discovery procedures. This is critical as it may play a big role in the difference between the theoretical versus the actual diversity of the library. This is consistent with a study that a HCDR3+LCDR3-randomized library had lower fitness than a

HCDR3-only randomized library, a trait attributed to the higher likelihood of generating inter-CDR structural clashes and dysfunctional clones.[11]

Our lab uses libraries with diversity focused on the HCDR3, based on frameworks with proven developability advantages. As such, an eventual de-stabilization caused by a particularly unstable HCDR3 is expected to be accommodated by the framework. Additionally, its germline-rich sequence can also be advantageous for the initial screen of primary binders before affinity maturation. However, a study from 2016 showed that synthetic antibodies overly rely on HCDR3-mediated interactions with the antigen, at the expense of HCDR2 and HCDR1, when compared to natural Abs.[38] This can reduce the amount of possible epitopes accessible to synthetic libraries. Looking at a HCDR3-only randomized, this problem is exacerbated since candidates selected after a primary panning are most likely the ones with a very dominant HCDR3 that can sustain the Ag-Ab interaction despite the typically weak support of the germline residues.

Considering the above, the author postulates whether a progressive randomization could provide a compromise between the polyreactivity and stability of germline sequences, and the affinity and epitope discovery capability of somatic mutations. Starting from a HCDR3-only randomized library, one or two rounds of panning under slightly stringent conditions would provide a way of removing negative binders and recovering a wide spectrum of positive binders. Afterwards, a step of polyclonal affinity maturation could be employed by a process of vector reformatting (i.e. cloning all HCDR3 sequences in bulk to a new phagemid vector which is randomized in the remaining CDR loops), which would go through further rounds of panning to select beneficial germline deviations on the remaining CDR loops.

The advantages would be four-fold: i) Polyclonal affinity maturation saves time by skipping candidate production and affinity screening; ii) Polyclonal affinity maturation provides an un-biased way of selecting HCDR3 sequences and effectively expands the universe of tested candidates, by testing all of them; iii) Separation of randomization designs across different vectors mitigates the limitations imposed by the practical diversity of libraries, by filtering out unwanted HCDR3 sequences previously – i.e. a bigger diversification of non-HCDR3 residues can be imposed compared to the ones employed on six-CDR-randomized libraries; iv) Mitigation of the number of dysfunctional clones with inter-CDR clashes between HCDR3 and other CDR loops, due to substantially less HCDR3 sequences on the polyclonal affinity maturation stage.

6.2. Developing a tool for library comparison and automated candidate selection

Candidate selection from phage-display campaigns was traditionally done by manual colony picking and Sanger Sequencing[39]. Such allowed for the selection of dominant clones but provided a small snapshot of the total diversity of candidates. Even though automated colony picking strategies have been implemented to achieve the inspection of up to >10^3 clones per experiment, the usage of next-generation sequencing (NGS) approaches has allowed a much deeper inspection of candidate pools after the final round of panning.[40–42] Classical colony picking only provides a small snapshot of the final output of panning campaigns and will be biased to a handful of more dominant clones. On the other hand, NGS allows not only to inspect the clones that dominated the sample, but also to search for rare clones that enriched throughout the process, and to search for sequence motifs that may be determining antigen binding. In this thesis, we took advantage of the Illumina NGS platform to select candidates and to compare the performance of our antibody libraries. Candidates were selected considering the parameters detailed in section 4.5.

Candidates have been used as read-outs of library performance, since they provide meaningful parameters that can be compared across conditions (e.g. Positive Clone %, Hit Rate %, Average K_D).[11,43] Comparing libraries using candidates is a major endeavor that requires several targets to be tested in different conditions (usually 6-10 targets in both solid-phase and liquid-phase selections), and a big number of assayed clones (from 100 to 10000). The possible presence of selection biases is the major caveat of this methodology, since there is no telling whether the candidates that were not selected for production and testing would perform well or not, and if the user selecting those candidates was wrongly biased to certain sequence patterns. This is especially critical when comparing libraries from distinct companies and/or laboratories that will have distinct SOPs of their own.

On this thesis, we tried to compare the performance of three distinct libraries (FW-κ, FW-κN1, and FW-κN2), but the volume of work required to achieve the aforementioned benchmarks would turn this task unfeasible. Instead of using candidates as read-outs of our libraries' performance, we decided to inspect the NGS datasets as whole to get some answers. We accomplished that by counting the number of HCDR3 unique sequences in a dataset, and most importantly, by clustering those sequences based on their similarity. Generally speaking, the more clusters you get, the more diverse is your dataset in terms of paratopes (some rare exceptions include similar sequences having different structures[45] or dissimilar sequences sharing the same epitope shape space[46]). Hence, clustering provides an excellent measurement on the outcome of a panning campaign, regarding the amount of paratopes it produces. It also provides a good comparator for wash stringency, since harsher washes reduce the number of clusters in a dataset (see section 4.3.), narrowing down the dataset to the best group of paratopes. This provided a rapid, systematic, and unbiased way of evaluating the NGS datasets and comparing library performance, as well as wash efficacy. We were, to our best knowledge, the first group to use sequence clustering for that specific purpose.

Systematic sequence clustering also opened the window for better candidate selection. The final objective of a panning campaign is undoubtedly to select the candidates that will follow to further characterization, such as K_D measurement and developability analysis. To maximize positive outcomes, expert users select a diverse dataset of candidates to test, since different sequences tend to generally bind to different epitopes. Having a clustering method allows users to sample candidates representative of the different clusters, and thus maximize cohort diversity. To further improve this methodology, an automatic sequence ranking tool was implemented following the clustering. For each cluster, in descending order from the top of the antigen list, the tool selects the first two candidates that pass through two of selection filters – Antigen/Mock > 1, Counter/Mock ≤ 1.5. These are put into a pre-selection list, and the candidates ranked according to their values on the following three ratios: Antigen/Mock, Antigen/Counter and Counter/Mock. The latter contributes subtracts to the score, while the other two add to the score. Higher scores are given to clones on the top of the list (i.e., higher ratios. For more information on these ratios see section 4.5.). The clones are then ranked according to their cumulative score for the user to choose from.

Data science has taken big steps in becoming a major tool for drug discovery.[46] This tool represents an institutional effort (see section 4.6. and Annex A) to accelerate the choice of candidates from a project, by mimicking the expert-users algorithm, while removing user biases and maximizing diversity of selected cohorts. Unfortunately, the tool in question was not available upon the selection of candidates from the primary pannings (section 4.3) for the work of this Thesis, and was only used for assessing paratope diversity later on. We expect nonetheless that this tool has a meaningful impact on the Novartis' antibody discovery pipeline. Additionally, this tool provides a solid scaffold for more complex methods to be built on top of. (e.g. add aminoacid similarity scores[47] to the distance score of sequences, upgrading the clustering method to HDBSCAN[48], motif finding across different sequence lengths, etc.).

6.3. Capturing Diversity in six CDR loops simultaneously

A major challenge in using NGS for the antibody discovery is related with the length of genes encoded in antibody libraries and the total read length of different NGS platforms. Typically, phage-display libraries present antibody formats such as Fab and single chain variable-fragment (scFv), rather than the full-length antibodies, as they are easier to express in bacterial systems. Additionally, both formats contain V_L and V_H sequences and can be readily re-formatted to IgG if needed, as IgG is the most accepted format in therapy nowadays. A scFv library is 700-800 bp in length, while a Fab library can reach up to 1500 bp due to the additional C_L and C_{H1} domains. NGS techniques with bigger depths (i.e. highest number of reads) are the most suitable to analyse the high diversity of phage-display outputs. On the other hand, the NGS techniques with the longest read lengths only do so at the expanse of throughput (e.g. PacBio), which turns them unsuitable for most library-based methods.[49] Currently, the longest read length available with reasonable throughput is provided by Illumina, with 600 bp reads yielding around 10^7 sequences. Still, 600bp are not enough to recover information on the six CDRs involved in antigen-binding. In Fab libraries, this also means that NGS will only yield information on either V_L or V_H, but never on V_L:V_H pairs. The lack of V_L and V_H paired data means that researchers need to opt between randomizing less CDRs or to perform analysis of V_L and V_H separately. This is particularly critical due to the relevance of LCDR3 and HCDR3 for antigen binding.

When confronted with the necessity of inspecting the diversity of LCDR1, LCDR3, HCDR1 and HCDR2 simultaneously on our affinity maturation libraries, we developed a NGS method on the widely used MiSeq (Illumina) to provide accurate information about all the six CDRs simultaneously for Fab phage display systems (for more info see section 5.5 and Annex B).

Our methodology yields trustworthy sequences that would otherwise be lost in V_L and V_H independent analysis and it also provides more reads than single-cell alternatives (~ 5 x 10^4 reads)[50] and SMRT sequencing based on PACbio (~ 8 x 10^4 reads).[51] Other alternatives, such as Kunkel mutagenesis, aim to physically concatenate CDRs *in vitro* by employing extensive PCR-based manipulations to remove unwanted segments between CDRs.[52] While effective, this approach is more error-prone and leads to high experimental burden often incompatible with library-based methods.[53,54] Inversely, our approach leverages existing run data to match forward and reverse reads to concatenate CDRs *in silico*, and as such requires minimal adaptation of lab protocols.

The most straightforward application of NGS is to try to sample high frequency clones near the top of the dataset, with the hope that such dominance translates into better affinity towards the intended target. Other NGS applications require a deeper inspection of datasets, such as when looking for motifs and clusters within the dataset[55], when looking for mutations of interest that were predicted by *in silico* tools.[56], or when looking for rare clones that were enriched throughout the selection rounds. We have shown that, without our analysis, some top clones can be eliminated from the dataset and hurt even the most straightforward analysis Critically, we have also shown that there is only 5% convergence between the real V_L:V_H pairing dataset and the dataset inferred from the separate V_L and V_H analysis. This highlights the relevance of our work, and we expect it to potentiate all the aforementioned NGS applications.

The use of NGS also goes beyond the scope of analyzing phage-display panning outputs and initial libraries and can be used to assess diversity of naive and immune libraries. One of the major hurdles of assessing diversity in immune and naive libraries is that combinatorial assembly of antibody heavy and light chains may generate V_L:V_H pairings that were not part of the donor's original repertoire and, hence, provide inaccurate estimations of diversity and potentially non-functional V_L:V_H pairs.[17,57] To mitigate the possibility of inaccurate V_L:V_H pairings, single-cell sequencing is employed, at the expense of depth. Our approach surpasses these problems while also allowing DNA from cells to be sequenced in bulk without having to go through single-cell isolation procedures. This provides higher throughput to *in vivo*-based antibody discovery systems, while also increasing sequencing depth.

In summary, we believe this methodology expands the capabilities of both *in vivo* and *in vitro* antibody discovery methods, while simultaneously tackling several challenges of previous V_L:V_H pairing approaches. It allows for the V_L:V_H pairing to be done in the most widely used NGS platform, without loss of throughput and high read fidelity, and without increasing the experimental burden. Most importantly, it provides a suitable V_L:V_H methodology for sequencing repertoires from *in vivo* samples and from extensively diversified *in vitro* Fab and scFv libraries.

6.4. Primed Libraries expand functional diversity in antibody libraries

The aim of this thesis was to build libraries with improved affinity maturation outcomes. For that, we structurally primed antibody frameworks by mutating hotspots that had a high likelihood of binding to antigen molecules. By inspecting two crystal structures from distinct frameworks (FW-κ and FW-λ), polar and charged residues pointing outwards towards the solvent were replaced by serines and alanines, and the effect of such mutations on the antibodies' developability was evaluated in terms of aggregation, hydrophobicity, and thermal stability. After careful analysis of the biophysical parameters achieved by those mutations, two final primed frameworks were selected (κN1 and κN2, bearing 8 and 11 mutations to CDR positions, respectively). These frameworks served as the basis for generating primed libraries.

Despite being very precise, the analysis of crystal structures is also very rigid and does not allow for the accurate prediction of outcomes upon mutation, or on how a given paratope will behave when meeting any given epitope. Future works on this topic should include molecular simulations, and preferably use several HCDR3 sequences representative of each length (e.g. from 10aa to 20aa) rather than just the sequence used on this study (WGGDGFYAMDY, see section 2.3). The 3D modelling of HCDR3 sequences is often very difficult to reliably accomplish (resolution < 1.0 Å), but that is not the case for HCDR1/2 and LCDR1/2/3, which are not expected to deviate much from a well detailed set of canonical structures.[58–62] As such, molecular simulations could provide a basis to interrogate which residues are consistently pointing outwards, and/or if these depend on the length of HCDR3 sequence in question. Ultimately, this could mean that we need to design a handful of "primed designs" that are dependent on the HCDR3 length and on its amino acid content.

Despite the known limitations of rigid structure analysis, our chosen mutations showed great effect in abrogating binding against Herceptin, on three very different and well-known anti-Herceptin FW-κ antibodies (see section 3.3.). This meant that the CDR-null mutations were sufficient to change the framework's conformational dynamics and favor different HCDR3 sequences during the panning selection processes. This was confirmed by the very different datasets arising from primary pannings, between all three frameworks (κ , κN1 and κN2). Besides yielding very different HCDR3 sequences, the primed libraries (κN1 and κN2) also showed lower diversity when compared with the control (κ), a behavior that can be explained by the lower overall polyreactivity typical of antibodies with less germline residues.[25,26] However, they were able to yield candidates of all HCDR3 sizes, with high affinity towards Herceptin, and equal distribution of affinities when compared with κ library. After the primary panning rounds and candidate selection, these CDR-null positions were targeted during a “semi-blind” affinity maturation protocol to maximize the likelihood of finding beneficial mutations that improve affinity. The primary binders were also subjected to the standard affinity maturation method used at the lab, based on the randomization of LCDR3 and HCDR2. Consequently, the primed framework κN1 was shown to have a bigger rate of affinity maturation success (31.3%) than the control framework (FW-κ, 24.2%), regardless of the affinity maturation method used. This value was increased when combining the primed library κN1 with the semi-blind affinity maturation method – 36.4%. (see overall success rates on section 5.3.3.1). FW-κN1 was also responsible for the biggest fold-increase (FI) in affinities in this body of work. This highlights the power of the strategy behind this work, where a small compromise in the affinity of primary binders is compensated by a bigger gain in affinity after affinity maturation (see aims and goals, section 1.4.). Finally, the partial germline reversion of some FW-κN1 candidates led them to simultaneously acquire affinity (due to the polar nature of those specific germline residues) and thermal stability, a major predictor of antibody developability success.[29–34]

Some caveats can be nonetheless identified. Although successful, this work requires further testing with other frameworks (such as FW-λ, which was dropped after Chapter 2) and targets (IL-1β and HSA were tested along the results shown in Chapter 4 but were discarded due to experimental burden). It is also likely that different types of targets will require different structural priming. Lessons from Janssen Bio’s pIX V3.0 library[37] can be used in this occasion, since their diversification design focused on positions frequently found in contact with protein and peptide targets.[63,64] More specifically, their diversification was based on a study[64] that said anti-protein antibodies have a large amount of antigen binding residues located at the edge of the protein surface while anti-hapten antibodies tend to have hotspots of contacts buried deep in the V_L:V_H interface. Another caveat from this thesis concerns the selection of primary binders that will follow through to affinity maturation. If full library comparison is to be attained, then a polyclonal affinity maturation process is likely the best choice. This eliminates eventual selection biases in the primary panning while also reduces the experimental steps of the study, a very important requirement in such a big endeavor.

All in all, this body of work opens a new paradigm in antibody discovery, by employing a somewhat counter-intuitive approach in the primary panning to achieve better outcomes in the later stages of antibody discovery. Additionally, this method was able to extract more HCDR3 sequences out of the randomization design (for more information on the diversity conundrum see section 1.3.1), effectively expanding the functional diversity of the original libraries against the same target.

6.5. References of Chapter 6

1. Lu, R.-M. et al. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci 27, 1 (2020).

2. Kaplon, H. & Reichert, J. M. Antibodies to watch in 2021. mAbs 13, 1860476 (2021).

3. Mullard, A. FDA approves 100th monoclonal antibody product. Nature Reviews Drug Discovery (2021) doi:10.1038/d41573-021-00079-7.

4. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature 256, 495–497 (1975).

5. Morrison, S. L., Johnson, M. J., Herzenberg, L. A. & Oi, V. T. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. Proc. Natl. Acad. Sci. U.S.A. 81, 6851–6855 (1984).

6. Jones, P. T., Dear, P. H., Foote, J., Neuberger, M. S. & Winter, G. Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature 321, 522–525 (1986).

7. Alfaleh, M. A. et al. Phage Display Derived Monoclonal Antibodies: From Bench to Bedside. Front. Immunol. 11, 1986 (2020).

8. Nagano, K. & Tsutsumi, Y. Phage Display Technology as a Powerful Platform for Antibody Drug Discovery. Viruses 13, 178 (2021).

9. Chan, D. T. Y. & Groves, M. A. T. Affinity maturation: highlights in the application of in vitro strategies for the directed evolution of antibodies. Emerging Topics in Life Sciences ETLS20200331 (2021) doi:10.1042/ETLS20200331.

10. Rajpal, A. et al. A general method for greatly improving the affinity of antibodies by using combinatorial libraries. PNAS 102, 8466–8471 (2005).

11. Mahon, C. M. et al. Comprehensive Interrogation of a Minimalist Synthetic CDR-H3 Library and Its Ability to Generate Antibodies with Therapeutic Potential. Journal of Molecular Biology 425, 1712–1730 (2013).

12. Alberts, B. et al. The Generation of Antibody Diversity. Molecular Biology of the Cell. 4th edition (2002).

13. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566, 393–397 (2019).

14. Schroeder, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Developmental & Comparative Immunology 30, 119–135 (2006).

15. Boyd, S. D. & Joshi, S. A. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiol Spectr 2, (2014).

16. Morbach, H., Eichhorn, E. M., Liese, J. G. & Girschick, H. J. Reference values for B cell subpopulations from infancy to adulthood. Clin Exp Immunol 162, 271–279 (2010).

17. Rees, A. R. Understanding the human antibody repertoire. mAbs 12, 1729683 (2020).

18. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. Journal of Allergy and Clinical Immunology 125, S41–S52 (2010).

19. Glanville, J. et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proceedings of the National Academy of Sciences 106, 20216–20221 (2009).

20. Hong, B. et al. In-Depth Analysis of Human Neonatal and Adult IgM Antibody Repertoires. Front. Immunol. 9, 128 (2018).

21. Soto, C. et al. High frequency of shared clonotypes in human B cell receptor repertoires. Nature 566, 398–402 (2019).

22. Perelson, A. S. & Weisbuch, G. Immunology for physicists. Rev. Mod. Phys. 69, 1219–1268 (1997).

23. Perelson, A. S. & Oster, G. F. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. Journal of Theoretical Biology 81, 645–670 (1979).

24. Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. Molecular Immunology 112, 338–346 (2019).

25. Notkins, A. L. Polyreactivity of antibody molecules. Trends in Immunology 25, 174–179 (2004).

26. Willis, J. R., Briney, B. S., DeLuca, S. L., Crowe, J. E. & Meiler, J. Human Germline Antibody Gene Segments Encode Polyspecific Antibodies. PLoS Comput Biol 9, e1003045 (2013).

27. Crouzier, R., Martin, T. & Pasquali, J. L. Heavy chain variable region, light chain variable region, and heavy chain CDR3 influences on the mono- and polyreactivity and on the affinity of human monoclonal rheumatoid factors. J Immunol 154, 4526–4535 (1995).

28. Harindranath, N., Ikematsu, H., Notkins, A. L. & Casali, P. Structure of the VH and VL segments of polyreactive and monoreactive human natural antibodies to HIV-1 and Escherichia coli beta-galactosidase. Int Immunol 5, 1523–1533 (1993).

29. Brader, M. L. et al. Examination of thermal unfolding and aggregation profiles of a series of developable therapeutic monoclonal antibodies. Mol Pharm 12, 1005–1017 (2015).

30. Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. PNAS 114, 944–949 (2017).

31. Barthelemy, P. A. et al. Comprehensive Analysis of the Factors Contributing to the Stability and Solubility of Autonomous Human V H Domains. Journal of Biological Chemistry 283, 3639–3654 (2008).

32. Chennamsetty, N., Voynov, V., Kayser, V., Helk, B. & Trout, B. L. Design of therapeutic proteins with enhanced stability. Proceedings of the National Academy of Sciences 106, 11937–11942 (2009).

33. He, F., Hogan, S., Latypov, R. F., Narhi, L. O. & Razinkov, V. I. High throughput thermostability screening of monoclonal antibody formulations. Journal of Pharmaceutical Sciences 99, 1707–1720 (2010).

34. Seeliger, D. et al. Boosting antibody developability through rational sequence optimization. mAbs 7, 505–515 (2015).

35. Tiller, T. et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. mAbs 5, 445–470 (2013).

36. Knappik, A. et al. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides 1 1Edited by I. A. Wilson. Journal of Molecular Biology 296, 57–86 (2000).

37. Shi, L. et al. De Novo Selection of High-Affinity Antibodies from Synthetic Fab Libraries Displayed on Phage as pIX Fusion Proteins. Journal of Molecular Biology 397, 385–396 (2010).

38. Burkovitz, A. & Ofran, Y. Understanding differences between synthetic and natural antibodies can help improve antibody engineering. MAbs 8, 278–287 (2016).

39. Selection of human antibody fragments by phage display | Nature Protocols. https://www.nature.com/articles/nprot.2007.448.

40. 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–e308 (2017).

41. Rouet, R., Jackson, K. J. L., Langley, D. B. & Christ, D. Next-Generation Sequencing of Antibody Display Repertoires. Front. Immunol. 9, 118 (2018).

42. Noh, J. et al. High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library. mAbs 11, 532–545 (2019).

43. Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. Antibodies 8, 44 (2019).

44. Kovaltsuk, A. et al. How B-Cell Receptor Repertoire Sequencing Can Be Enriched with Structural Antibody Data. Front Immunol 8, 1753 (2017).

45. Krawczyk, K. et al. Structurally Mapping Antibody Repertoires. Front Immunol 9, 1698 (2018).

46. Ferrero, E. et al. Ten simple rules to power drug discovery with data science. PLOS Computational Biology 16, e1008126 (2020).

47. Kim, Y., Sidney, J., Pinilla, C., Sette, A. & Peters, B. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior. BMC Bioinformatics 10, 394 (2009).

48. Castro Gertrudes, J., Zimek, A., Sander, J. & Campello, R. J. G. B. A unified view of density-based methods for semi-supervised clustering and classification. Data Min Knowl Discov 33, 1894–1952 (2019).

49. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).

50. DeKosky, B. J. et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 31, 166–169 (2013).

51. Nannini, F. et al. Combining phage display with SMRTbell next-generation sequencing for the rapid discovery of functional scFv fragments. mAbs 13, 1864084 (2021).

52. Barreto, K. et al. Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. Nucleic Acids Research 47, e50–e50 (2019).

53. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering 96, 317–323 (2003).

54. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl 1, (2014).

55. Norman, R. A. et al. Computational approaches to therapeutic antibody design: established methods and emerging trends. Briefings in Bioinformatics 21, 1549–1567 (2020).

56. Campbell, S. M. et al. Combining random mutagenesis, structure-guided design and next-generation sequencing to mitigate polyreactivity of an anti-IL-21R antibody. mAbs 13, 1883239 (2021).

57. Harel Inbar, N. & Benhar, I. Selection of antibodies from synthetic antibody libraries. Archives of Biochemistry and Biophysics 526, 87–98 (2012).

58. Kuroda, D., Shirai, H., Jacobson, M. P. & Nakamura, H. Computer-aided antibody design. Protein Engineering Design and Selection 25, 507–522 (2012).

59. Almagro, J. C. et al. Antibody modeling assessment: Antibody Modeling. Proteins 79, 3050–3066 (2011).

60. Almagro, J. C. et al. Second antibody modeling assessment (AMA-II): 3D Antibody Modeling. Proteins 82, 1553–1562 (2014).

61. Chothia, C. & Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196, 901–917 (1987).

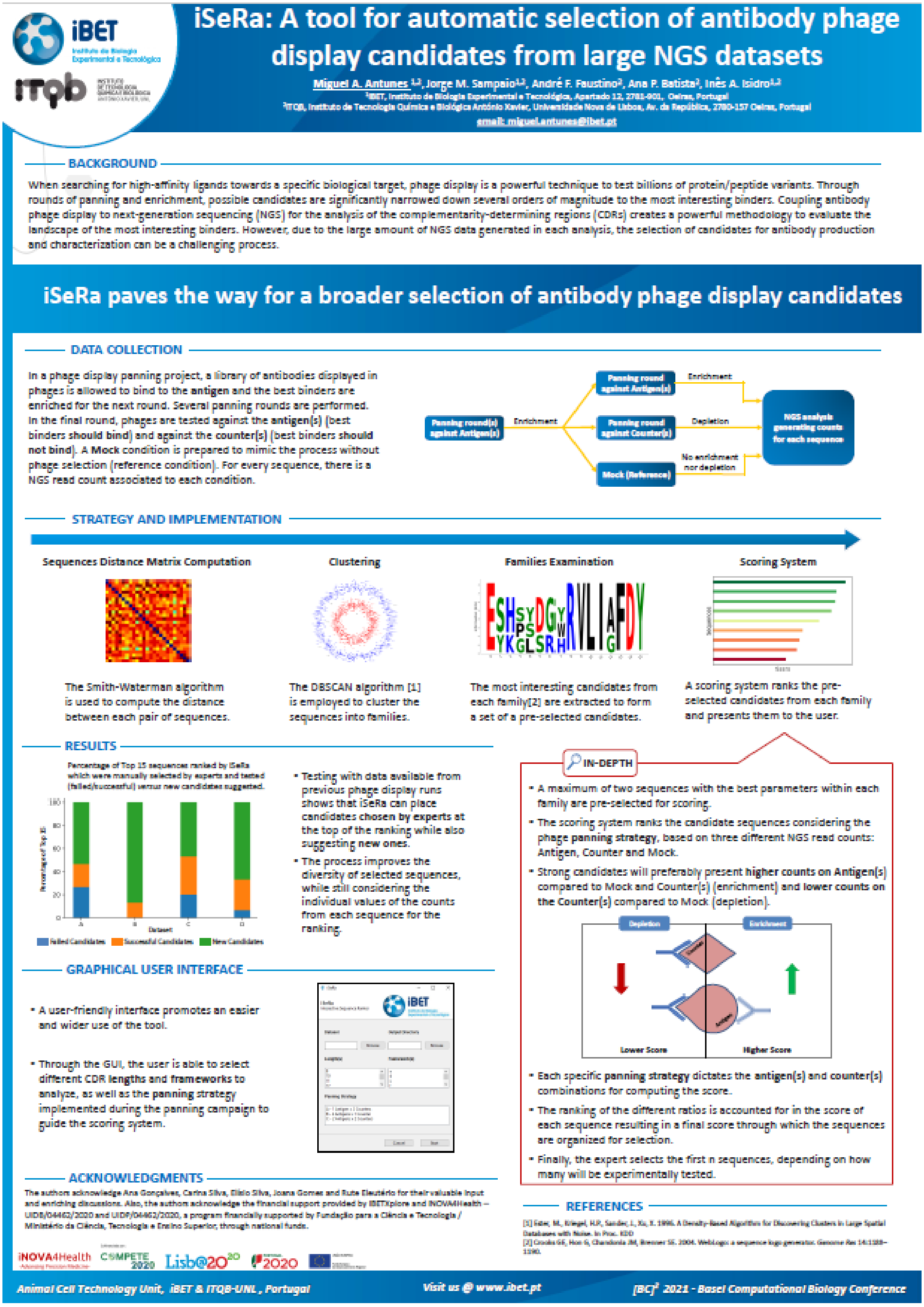
62. North, B., Lehmann, A. & Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. Journal of Molecular Biology 406, 228–256 (2011).

63. Raghunathan, G., Smart, J., Williams, J. & Almagro, J. C. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. Journal of Molecular Recognition 25, 103–113 (2012).

64. Almagro, J. C. Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. Journal of Molecular Recognition 17, 132–143 (2004).

Appendix

Annex A: Automatic Selection of Antibody Phage Display Candidates



iSeRa: A tool for automatic selection of antibody phage display candidates from large NGS datasets

Miguel A. Antunes [1,2], Jorge M. Sampaio[1,2], André F. Faustino[2], Ana P. Batista[2], Inês A. Isidro[1,2]
[1] iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2781-901, Oeiras, Portugal
[2] ITQB, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal
email: miguel.antunes@ibet.pt

BACKGROUND

When searching for high-affinity ligands towards a specific biological target, phage display is a powerful technique to test billions of protein/peptide variants. Through rounds of panning and enrichment, possible candidates are significantly narrowed down several orders of magnitude to the most interesting binders. Coupling antibody phage display to next-generation sequencing (NGS) for the analysis of the complementarity-determining regions (CDRs) creates a powerful methodology to evaluate the landscape of the most interesting binders. However, due to the large amount of NGS data generated in each analysis, the selection of candidates for antibody production and characterization can be a challenging process.

iSeRa paves the way for a broader selection of antibody phage display candidates

DATA COLLECTION

In a phage display panning project, a library of antibodies displayed in phages is allowed to bind to the **antigen** and the best binders are enriched for the next round. Several panning rounds are performed. In the final round, phages are tested against the **antigen(s)** (**best binders should bind**) and against the **counter(s)** (**best binders should not bind**). A **Mock** condition is prepared to mimic the process without phage selection (reference condition). For every sequence, there is a NGS read count associated to each condition.

Panning round(s) against Antigen(s)
Enrichment
Panning round against Antigen(s)
Enrichment
Panning round against Counter(s)
Depletion
Mock (Reference)
No enrichment nor depletion
NGS analysis generating counts for each sequence

STRATEGY AND IMPLEMENTATION

Sequences Distance Matrix Computation

The Smith-Waterman algorithm is used to compute the distance between each pair of sequences.

Clustering

The DBSCAN algorithm [1] is employed to cluster the sequences into families.

Families Examination

The most interesting candidates from each family[2] are extracted to form a set of a pre-selected candidates.

Scoring System

A scoring system ranks the pre-selected candidates from each family and presents them to the user.

RESULTS

Percentage of Top 15 sequences ranked by iSeRa which were manually selected by experts and tested (failed/successful) versus new candidates suggested.

- Testing with data available from previous phage display runs shows that iSeRa can place candidates chosen by experts at the top of the ranking while also suggesting new ones.
- The process improves the diversity of selected sequences, while still considering the individual values of the counts from each sequence for the ranking.

IN-DEPTH

- A maximum of two sequences with the best parameters within each family are pre-selected for scoring.
- The scoring system ranks the candidate sequences considering the **phage panning strategy**, based on three different NGS read counts: Antigen, Counter and Mock.
- Strong candidates will preferably present **higher counts on Antigen(s)** compared to Mock and Counter(s) (enrichment) and **lower counts on** the Counter(s) compared to Mock (depletion).

- Each specific **panning strategy** dictates the **antigen(s)** and **counter(s)** combinations for computing the score.
- The ranking of the different ratios is accounted for in the score of each sequence resulting in a final score through which the sequences are organized for selection.
- Finally, the expert selects the first n sequences, depending on how many will be experimentally tested.

GRAPHICAL USER INTERFACE

- A user-friendly interface promotes an easier and wider use of the tool.
- Through the GUI, the user is able to select different **CDR lengths** and **frameworks** to analyze, as well as the **panning strategy** implemented during the panning campaign to guide the scoring system.

ACKNOWLEDGMENTS

The authors acknowledge Ana Gonçalves, Carina Silva, Elísio Silva, Joana Gomes and Rute Eleutério for their valuable input and enriching discussions. Also, the authors acknowledge the financial support provided by iBETXplore and iNOVA4Health – UIDB/04462/2020 and UIDP/04462/2020, a program financially supported by Fundação para a Ciência e Tecnologia / Ministério da Ciência, Tecnologia e Ensino Superior, through national funds.

REFERENCES

[1] Ester, M., Kriegel, H.P., Sander, J., Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proc. KDD
[2] Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188–1190.

Animal Cell Technology Unit, iBET & ITQB-UNL, Portugal *Visit us @ www.ibet.pt* *[BC]^2 2021 - Basel Computational Biology Conference*

Annex B: V_L:V_H pairing and six CDR assessment

Antibody discovery powered by NGS: Retrieving Fab sequences with coupled V_L:V_H information and six CDR assessment by cluster coordinate matching.

Jorge Moura-Sampaio[1,2], André F. Faustino[1], Remi Boeuf[3], Miguel A. Antunes[1], Stefan Ewert[3], Ana P. Batista[1,*]

1 – iBET, Instituto de Biologia Experimental e Tecnológica, Apartado 12, 2781-901 Oeiras, Portugal

2 – Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal;

3 – Novartis Institutes for BioMedical Research, Basel, Switzerland.

* e-mail: abatista@ibet.pt; Phone number: +351 214 469 741

Keywords:

Next-generation Sequencing, Phage-display, Synthetic Libraries, Randomization, Diversity

ABSTRACT

Next-generation sequencing (NGS) has become an indispensable tool in antibody discovery projects. Its superior depth has allowed to select more diverse candidates when compared with traditional selection methods based on colony picking and Sanger sequencing. However, the limits on NGS read length make it difficult to reconstruct full antibody sequences from the sequencing runs, especially if the six CDRs are randomized, or when dealing with the longer Fab sequences. To overcome that, we devised a simple method that retrieves paired sequencing data of V_L and V_H regions in Fab sequences, while maintaining the high amount of reads necessary for antibody discovery campaigns, with a high degree of fidelity. We rely on *in silico* cluster coordinate information, and not on extensive *in vitro* manipulation, making the protocol easily deployable and less prone to PCR-derived errors. This sequencing approach potentiates not only phage-display and synthetic library-based discovery methods, but also the NGS-driven analysis of naïve and immune libraries.

INTRODUCTION

Antibody discovery has been potentiated by the use of *in vitro* display methods and, more recently, by using Next-generation Sequencing (NGS) to analyze immunoglobulin variable regions.[1–4] *In vitro* display platforms answer to limitations of animal experimentation from both operational and ethical standpoints, also providing greater experimental control over antigen presentation. The most popular *in vitro* methods are based on microbial systems – phage and yeast display – and thus have very high potential regarding parallelization, automation, and miniaturization. *In vitro* protocols require multiple rounds of selection to discover the lead candidates against a given target. In most cases, the dominant clones will have superior affinity towards the antigen in comparison with the rest of the sample. The identification of dominant clones was traditionally done by manual colony picking and Sanger Sequencing. Such allowed for the selection of dominant clones but provided a small snapshot of the total diversity of candidates. While automated colony picking strategies have been implemented to achieve the inspection of >10^3 clones per experiment, this only accounts for about 0.1% of the total yield of a phage-display protocol (10^5 - 10^8 cfu). On the other hand, NGS has allowed a much deeper inspection of candidate pools after the final round of panning, by retrieving up to 10^7 sequences.[1,5,6] The bigger depth of NGS analysis allows for the selection of more diverse candidate dataset by providing alternatives to the dominant “top clones”. These alternatives arise from careful inspection of patterns across conditions and controls. For instance, enrichment ratios can be calculated between the condition of interest and a “mock” condition (a condition with successive cycles of infection and phage production, but not challenged against an antigen) to discard clones that have been displayed better but that did not present superior binding. Likewise, enrichment ratios between two similar targets can be calculated to select clones directed against a specific epitope. Finally, the sheer number of unique sequences achieved by NGS allows for the discovery of beneficial motifs and clustering of sequences against a given target. Besides panning analysis and candidate selection, NGS has also proven to be a useful tool for the quality control of antibody libraries, of both naive and synthetic origin, in terms of their CDR length distribution, germline frequencies and clone redundancy. [7–9]

A major challenge in using NGS for the aforementioned purposes is related with the length of genes encoded in antibody libraries and the total read length of different NGS platforms. Typically, phage-display libraries present antibody formats such as Fab and single chain variable-fragment (scFv), rather than the full-length antibodies, as they are easier to express in bacterial systems. Additionally, both formats contain V_L and V_H sequences and can be readily re-formatted to IgG if needed, as IgG is the most accepted format in therapy nowadays. A scFv library is 700-800 bp in length, while a Fab library can reach up to 1500 bp due to the additional C_L and C_{H1} domains. NGS techniques with bigger depths (i.e. highest number of reads) are the most suitable to analyse the high diversity of phage-display outputs. On the other hand, the NGS techniques with the longest read lengths only do so at the expanse of throughput (e.g. PacBio), which turns them unsuitable for most library-based methods.[10] Currently, the longest read length available with reasonable throughput is provided by Illumina, with 600 bp reads yielding around 10^7 sequences. Still, 600bp are not enough to recover information on the six CDRs involved in antigen-binding. In Fab libraries, this also means that NGS will only yield information on either V_L or V_H, but never on V_L:V_H pairs. The lack of V_L and V_H paired data means that researchers need to opt between randomizing less CDRs or to perform analysis of V_L and V_H separately. This is particularly critical due to the relevance of LCDR3 and HCDR3 for antigen binding. Some efforts have successfully sequenced V_L:V_H at the expense of some drawbacks. DeKosky *et al.* were able to retrieve information on V_L:V_H pairs after single-cell sequencing of 5×10^4 B-cells,[11] which is not compatible with all antibody discovery platforms. Nanini *et al.* employed coupled single-molecule real time (SMRT) sequencing to retrieve information on V_L:V_H pairs, but only did so for scFv sequences, while using the lower throughput PacBio equipment.[12] Finally, Barreto et al. used Kunkel mutagenesis during the amplification step to remove intervening framework regions between four randomized CDRs, using the Ion Torrent platform.[13] The latter strategy is able to sample diversity from distant CDRs in Fab sequences, but at the expense of greater experimental burden and higher error likelihood, a typical drawback of PCR-based techniques.[1]

The objective of this work was to develop a NGS method on the widely used MiSeq (Illumina) to provide accurate information about all the six CDRs simultaneously for Fab phage display systems. Here, paired-end sequencing of V_L (forward read) and V_H (reverse read) is performed after amplifying the whole region of interest. Then, information on cluster coordinates that arise from the Bridge-PCR of Illumina sequencing is used to match the forward and reverse reads. This method has the advantage of being processed on the most widely used NGS platform (Illumina) with minimal adjustment of protocols and with higher number of reads than single-cell and SMRT sequencing approaches. Additionally, since it relies on cluster information, it can be applied to regions of interest that are far apart, making it compatible with Fab sequences, which is not the case of long-read applications such as PACbio that only achieved read fidelity on scFv sequences. Moreover, the CDRs are concatenated *in silico*, rather than *in vitro* (as the Kunkel mutagenesis approaches do), making our strategy less prone to errors and less extensive from the operational standpoint.

RESULTS

Cluster coordinates allow the interrogation of six CDRs simultaneously

Currently, the longest reads produced on an Illumina platform is accomplished by paired-end sequencing on the MiSeq Reagent Kit v3 (600-cycle), which produces a forward read (R1) and a reverse read (R2) whose lengths sum up to a total of 600 bp. 300 bp is enough to cover the distance between the first residue of CDR1 and the last residue of CDR3, for both V_L and V_H, even when considering an extra 12-15 bp in both 5' and 3' ends (used for primer annealing during the amplification step and for query purposes during the data analysis step). Thus, if R1 and R2 are correctly assigned to each other, a full Fab fragment can be reconstructed with the information on all six CDRs. Illumina systems make use of Bridge-PCR to replicate the desired amplicons into a cluster of identical amplicons, as a way to increase the signal.[14] Sequencing ensues on sequencing-by-synthesis manner, where a mix of reversible dye-terminated nucleotides are added to mixture and imaged in a cyclical manner. If guidelines are thoroughly followed, each cluster should be perfectly distinguished from each other as to provide a confident signal to the sequencing equipment. On each image taken, all clusters will have a specific ID that indicate their lane, tile and XY coordinates inside the composite image generated by the imaging software. (Figure 1)

As long as cluster information is retained, sequence information can be paired confidently. We have leveraged sequence coordinate information to retrieve information on regions of interest that are far apart in big amplicons, such as the CDRs from Fab and scFv molecules. As such, we have developed a simple methodology which concatenates R1 and R2 reads that share the same cluster ID, by appending dummy n nucleotides along with the reverse complement of R2 read to the end of R1 (Figure 2a,b). This information is appended to all sequences by the instrument[15] and R1 and R2 can be easily matched using simple scripts on any programming language commonly used in biosciences or bioinformatics. While it is obvious that Fab sequences benefit from this application, scFv sequences may also be sequenced by longer 600 bp reads to get correlated information on up to 5 CDRs. However, long reads come at a cost of lower read quality. Paired-end sequencing allows better read quality than a single-read across the same amplicon, and, hence, makes this approach also useful for scFv sequencing.

One common concern in Bridge-PCR protocols is that the quality of the experiment tends to decrease with increasing amplicon lengths, since longer amplicons lead to clusters with larger diameters with a higher probability to overlap. Adjusting the concentration of DNA loaded into the instrument is a suitable way of controlling cluster overlapping. We have tested that a loading concentration of 7.2 pM is sufficient to yield good overall cluster quality without compromising the total number of reads and their fidelity. A total of 1.68 x 10^7 reads were obtained, with 81.7% having a quality score above 30 (Table S1). This method allowed us to extract information on diversified positions (in this case, 8 different positions across 4 CDRs), with a Q-score > 30 for each nucleotide inspected (24 nt.), for at least 100.000 reads for each library.

Coordinate matching reveals hidden V_L:V_H pairs and avoids mispairings that arise from inference based on independent analysis.

We decided to apply the aforementioned analysis to 23 different affinity maturation projects, which had diversity in LCDR1, LCDR3, HCDR1 and HCDR2. Firstly, we challenged these affinity maturation libraries against an industry-standard antigen. Secondly, we analyzed these results using two different methods. One method tries to infer V_L:V_H pairs from independent V_L and V_H sequencing runs (R1 and R2, respectively), by matching the top clones from each of the independently analyzed datasets. The other method uses cluster coordinates to match V_L and V_H reads (R1 and R2, respectively) to produce sequences with the full information on V_L and V_H diversity. The poor matching between the top clones generated from inference compared to the real clones generated by the cluster coordinate method highlights the relevance of the suggested new approach (Figure 3, Table S1).

Around 5% of the Top 100 inferred pairs corresponded to real sequences, with others being artificial sequences that did not appear in the real dataset (at least with meaningful representativeness). This effect improved only slightly when the Top50 and Top25 were compared. Most importantly, even when sampling the Top10 sequences, only 17±6.2% of sequences converged between correlated and independent datasets (Figure 3, Table S2). The effects of analyzing V_L and V_H independently can also be seen when looking exclusively at the top clone of each dataset. We searched the top hits of each dataset within the inferred dataset and found that in 13 projects, the real Top1 candidates could have not been found if sequence coordinate matching had not been performed. In the remaining projects, there was a match between the top clones of the datasets in 9 of them (Top 1), and in another project, there was a match between Top2 of the correlated and inferred dataset (Table 1).

DISCUSSION

The implementation of NGS technologies in phage-display applications has provided insights on all stages of antibody discovery: library generation and diversity assessment, quality control, and candidate selection after panning campaigns. Although HCDR3 is widely recognized as the most important CDR for antigen binding the interplay of all six CDRs is necessary for the antibody-antigen interaction.[16–18] While information on all CDRs can be accomplished by traditional sequencing of single-colonies, the read length limitations of NGS means that V_L and V_H information comes separately when high-throughput approaches are necessary. Primary library strategies have increasingly relied on the randomization of other CDRs to discover antibodies against antigens. Ylanthia (randomized in LCDR3 and HCDR3)[9] and HuCAL

PLATINUM® (all six CDRs)[19] constitute clear examples of state-of-the-art synthetic libraries that employ such strategies.

In this work we devised a simple and effective protocol to match V_L and V_H sequences for NGS applications, using the widespread MiSeq platform and, hence, recover information on all CDRs in a high-throughput manner. Our methodology yields trustworthy sequences that would otherwise be lost in V_L and V_H independent analysis and it also provides more reads than single-cell alternatives (~ 5 x 10^4 reads)[11] and SMRT sequencing based on PACbio (~ 8 x 10^4 reads).[12] Other alternatives, such as Kunkel mutagenesis, aim to physically concatenate CDRs *in vitro* by employing extensive PCR-based manipulations to remove unwanted segments between CDRs.[13] While effective, this approach increases the accumulation of errors that stem from PCR-based techniques, and increases the experimental load as higher amounts of DNA are required to fulfill the PCR, electrophoresis and purification steps. Both these drawbacks are exacerbated when working in antibody library and phage-display settings, where diversity, throughput and quality-control are adamant.[20,21] Inversely, our approach leverages existing run data to match forward and reverse reads to concatenate CDRs *in silico*, and as such requires minimal adaptation of lab protocols. The amplicons generated by our approach are bigger than the ones usually used in MiSeq protocols.

We used this method to investigate the outcome of 23 different affinity maturation projects, with libraries that had diversity in LCDR1, LCDR3, HCDR1, and HCDR2, and compared it with the previous strategies, that relied on the independent analysis of V_L and V_H. In this work we have also pointed out how the independent analysis of V_L and V_H generates incorrect V_L:V_H combinations that do not correspond to the real dataset. This effect was seen in both high frequency clones and low frequency clones. This highlights the importance of correctly matching V_L and V_H pairs, and the impact that independent V_L and V_H analysis have throughout all types of NGS-based applications. The most straightforward application of NGS is to try to sample high frequency clones near the top of the dataset, with the hope that such dominance translates into better affinity towards the intended target. We have shown that, without our analysis, some top clones can be eliminated from the dataset and hurt even the most straightforward analysis. Other NGS applications require a deeper inspection of datasets, such as when looking for motifs and clusters within the dataset[22], when looking for mutations of interest that were predicted by *in silico* tools.[2], or when looking for rare clones that were enriched throughout the selection rounds. Critically, we have also shown that there is only 5% convergence between the real V_L:V_H pairing and the ones inferred from the separate analysis when going as deep as 100 clones. This highlights the relevance of our work and we expect it to potentiate all the aforementioned NGS applications.

The use of NGS also goes beyond the scope of analyzing phage-display panning outputs and initial libraries, and can be used to assess diversity of naive and immune libraries. One of the major hurdles of assessing diversity in immune and naive libraries is that combinatorial assembly of antibody heavy and light chains may generate V_L:V_H pairings that were not part of the donor’s original repertoire and, hence, provide inaccurate estimations of diversity and potentially non-functional V_L:V_H pairs.[23,24] To mitigate the possibility of inaccurate V_L:V_H pairings, single-cell sequencing is employed, at the expense of depth. Our approach surpasses these problems while also allowing DNA from cells to be sequenced in bulk without having to go through single-cell isolation procedures. This provides higher throughput to *in vivo*-based antibody discovery systems, while also increasing sequencing depth. Moreover, since the proposed method is purely based on DNA sequencing, it can be applied to any system in which the regions of interest to be sequenced are far apart (up to a reasonable amplicon length; 1250 bp in our case), having a known or non-interesting region in between.

In summary, we believe this work expands the capabilities of both *in vivo* and *in vitro* antibody discovery methods, while simultaneously tackling several challenges of previous V_L:V_H pairing approaches. It allows for the V_L:V_H pairing to be done in the most widely used NGS platform, without loss of throughput and high read fidelity, and without increasing the experimental burden. Most importantly, it provides a suitable V_L:V_H methodology for sequencing repertoires from *in vivo* samples and from extensively diversified *in vitro* Fab and scFv libraries.

METHODS

Fab affinity maturation libraries generation: A total of 23 HCDR3 sequences from parental antibodies were cloned into in-house phagemid vectors encoding an affinity maturation framework diversified on 8 different amino acid positions, across LCDR1, LCDR3, HCDR1 and HCDR2. These were designed *in-house* and manufactured by Twist Bioscience. The phagemid vectors were transformed via electroporation into electrocompetent *E.coli* TG1 cells (Lucigen). Electroporated cells were recovered in SOC medium for 1 hour before being transferred into 2YT medium with 1% glucose and 100 µg.mL-1 of ampicillin (2YT/A/G) and incubated overnight at 25 °C, 200 rpm on incubator Innova 44. Glycerol stocks were established the next day by storing the cells in 2YT/A/G/ supplemented with 10% glycerol.

Phage production: A sample from each glycerol stock was taken to start a 25 mL culture in 2YT/A/G at OD_{600} = 0.1 and grown to mid-log phase (OD600 = 0.5) before being infected with helper phage VCSM13 (Agilent Technologies). Cells were then incubated firstly for 30 min at 37°C in a water bath, and then for 30 min at 37°C shaking at 250 rpm. The infected bacteria were then centrifuged and transferred into a 40 mL culture of 2YT supplemented with 100 µg/mL Ampicillin, 50 µg/mL Kanamycin and 0.25 mM IPTG**.** Phage production ensued overnight at 22°C and 180 rpm. The cultures were centrifuged to remove the cells and the phage-rich supernatant collected into sterile 50 mL Falcon tubes and kept on ice. Phages are precipitated by adding 10 mL of ice cold 20% (w/v) PEG 6K in 2.5 M NaCl into the 40 mL of supernatant, 1 hour on ice. After this time, the precipitated solutions were centrifuged at 4000 *g* and 4 ˚C for 30 min (Eppendorf, Ref: 5810 R). The supernatant was discarded, and the precipitated phage pellets were re-suspended with 1 mL of sterile phosphate buffered saline (PBS) and transferred to 1.5 mL eppendorf tubes. The tubes were then rotated for 30 min on a rotating wheel at 4˚C and then centrifuged at 12 000 *g* and 10 ˚C for 5 min (Eppendorf, Ref: 5810 R) to remove further bacterial debris. Supernatants were filtered into cryovials containing 700 uL of PBS:Glycerol 50:50% (for a final [Glycerol] of 20% v/v).

Phage display panning selections: Phage display protocols were performed using the automated liquid handling functionalities of the KingFisher™ Flex Purification System (ThermoFisher, Catalog number: 5400610). Phages corresponding to each affinity maturation library were blocked for 1 h in PBS + 0.05% Tween (PBST) supplemented with 0.05% of BSA, in 96 DeepWell plates (Thermo Scientific™ 95040450), followed by in-solution deselection on streptavidin-coated magnetic beads (Dynabeads, Invitrogen, Cat # 112–06) for 30 min. To each well of sticky-depleted phages, 50 nM of biotinylated antigen was added in the corresponding well and incubated 1h at room temperature (RT) on a micro-plate table. The antigen-antibody complexes were captured from the deep well plates by the streptavidin-coated magnetic beads bound to the KingFisher magnetic rods and transferred to the washing plates sequentially. Washing of bead-antigen-phage complexes was accomplished by 10 washes of increasing vigor, stringency, and duration, on PBST and PBS. At the end of the wash protocols, phages were dissociated from the complexes with glycine buffer (10 mM glycine-HCl, pH 2.0) before neutralization with 200 µL Tris-HCl pH

7.5 and infection of a 20 mL mid-log *E.coli* TG1 culture (OD600 = 0.5). The cultures were incubated for 45 min in a water bath at 37ºC before being inoculated into 100 mL 2YT/A/G in 250 ml Erlenmeyer's and let to grow overnight at 25ºC, 150 rpm (Innova 44R, New Brunswick Scientific).

DNA preparation and NGS analysis: Plasmid DNA was isolated directly from the phage-infected cells from the selection round of interest using the GeneJET Plasmid Miniprep Kit (Thermo Scientific™, K0502). Isolated dsDNA was quantified on the Qubit 3.0 fluorometer using the Qubit® dsDNA HS kit (Invitrogen™ Q32851). The generation of V_L:V_H amplicon for sequencing was generated through two PCRs. To amplify the region of interest and to insert the adapter regions for the NGS, the initial PCR utilized a forward primer specific to the vector leader sequence prior to LCDR1 and, since we did not need HCDR3 information, a reverse primer downstream of HCDR2. The second PCR inserted the TruSeq universal adapter and the indexes, used to distinguish between different samples (i.e. libraries). Samples were quantified in Qubit 3.0, pooled in equimolar proportions, and ran on an electrophoresis gel. Bands with the appropriate size were excised, purified using the Wizard SV Gel and PCR Clean Up System (Promega, A9281), and quantified on Qubit 3.0. The pool was diluted to a final concentration of 4 nM, spiked with 20% PhiX (Illumina; FC-110-3001), denatured for 5 min in 0.1 N of NaOH (5 µL of DNA+PhiX at 4 nM mixed with 5 µL 0.2 N of NaOH), diluted in HT buffer (provided on the NGS kit; kit details, ahead) to 7.2 pM and sequenced on the Illumina MiSeq platform using the 500 cycle V2 kit (Illumina; MS-102-2003). The forward read was 230 bp in length while the reverse read was 270 bp. R1 retrieves information on

LCDR1, LCDR2 (non-diversified), and LCDR3. R2 retrieves information on HCDR1 and HCDR2. Note: we performed a sequencing between LCDR1 and HCDR2 with a 500 cycle V2 kit, but the procedure is directly transferrable to a sequencing between LCDR1 and HCDR3 with a 600 cycle v3 kit (Illumina, MS-102-3003) by sequencing 230+370 bp (as explained on Figure 2). The data analysis of the NGS FastQ output files was performed as described previously.[3] For the panning output of each library, 1 x 10^5 sequences were analyzed using the fixed-by-design flanking sequences on the boundary of diversified positions as template to locate and segment out mutations. Full CDR sequences were reconstructed by coupling the regions fixed-by-design with the information on the diversified regions. This analysis was performed on the two distinct datasets (as explained on Figure 3). The first dataset contains correlated V_L:V_H information after concatenation of R1 and R2 using their sequence's coordinates, as described in the result section. The second dataset is generated only after the independent analysis of R1 and R2 outputs, and tries to infer to infer V_L:V_H pairs from sequence frequency and relative postion within the dataset (i.e. the most frequent clone of R1 is matched with the most frequent clone of R2 and so forth). Sequences in both datasets that only had one occurrence were removed from the analysis.

AUTHOR CONTRIBUTIONS

JMS and AFF conceived and designed the experiments. JMS performed the experiments, analyzed the results, and draft the manuscript. AFF critically discussed the data and helped to draft the manuscript. RB devised the in-house NGS analysis tool used for CDR determination. MAA programmed the sequence matching protocol. APB and SE coordinated the study, critically discussed the data, and helped to draft the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We acknowledge Inês A. Isidro for all the insightful comments and discussions. This work was supported by iNOVA4Health – UIDB/04462/2020 and UIDP/04462/2020, a program financially supported by Fundação para a Ciência e Tecnologia / Ministério da Ciência, Tecnologia e Ensino Superior, through Portuguese national funds. JMS personally acknowledges FCT-MEC fellowship PD/BD/128321/2017.

REFERENCES

1. Rouet, R., Jackson, K. J. L., Langley, D. B. & Christ, D. Next-Generation Sequencing of Antibody Display Repertoires. Front. Immunol. 9, 118 (2018).

2. Campbell, S. M. et al. Combining random mutagenesis, structure-guided design and next-generation sequencing to mitigate polyreactivity of an anti-IL-21R antibody. mAbs 13, 1883239 (2021).

3. Liu, G. et al. Antibody complementarity determining region design using high-capacity machine learning. Bioinformatics 36, 2126–2133 (2020).

4. Senatore, A. et al. Protective anti-prion antibodies in human immunoglobulin repertoires. EMBO Mol Med 12, e12739 (2020).

5. 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–e308 (2017).

6. Noh, J. et al. High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library. mAbs 11, 532–545 (2019).

7. Zhai, W. et al. Synthetic Antibodies Designed on Natural Sequence Landscapes. Journal of Molecular Biology 412, 55–71 (2011).

8. Ravn, U. et al. Deep sequencing of phage display libraries to support antibody discovery. Methods 60, 99–110 (2013).

9. Tiller, T. et al. A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. mAbs 5, 445–470 (2013).

10. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).

11. DeKosky, B. J. et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 31, 166–169 (2013).

12. Nannini, F. et al. Combining phage display with SMRTbell next-generation sequencing for the rapid discovery of functional scFv fragments. mAbs 13, 1864084 (2021).

13. Barreto, K. et al. Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. Nucleic Acids Research 47, e50–e50 (2019).

14. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next Generation Sequencing Technologies. Curr Protoc Mol Biol 122, e59 (2018).

15. File Format. https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/FileFormat_FASTQ-files_swBS.htm.

16. Schroeder, H. W. & Cavacini, L. Structure and function of immunoglobulins. Journal of Allergy and Clinical Immunology 125, S41–S52 (2010).

17. North, B., Lehmann, A. & Dunbrack, R. L. A New Clustering of Antibody CDR Loop Conformations. Journal of Molecular Biology 406, 228–256 (2011).

18. Sela-Culang, I., Kunik, V. & Ofran, Y. The Structural Basis of Antibody-Antigen Recognition. Frontiers in Immunology 4, (2013).

19. Prassler, J. et al. HuCAL PLATINUM, a Synthetic Fab Library Optimized for Sequence Diversity and Superior Performance in Mammalian Expression Systems. Journal of Molecular Biology 413, 261–278 (2011).

20. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering 96, 317–323 (2003).

21. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl 1, (2014).

22. Norman, R. A. et al. Computational approaches to therapeutic antibody design: established methods and emerging trends. Briefings in Bioinformatics 21, 1549–1567 (2020).

23. Rees, A. R. Understanding the human antibody repertoire. mAbs 12, 1729683 (2020).

24. Harel Inbar, N. & Benhar, I. Selection of antibodies from synthetic antibody libraries. Archives of Biochemistry and Biophysics 526, 87–98 (2012).

FIGURES

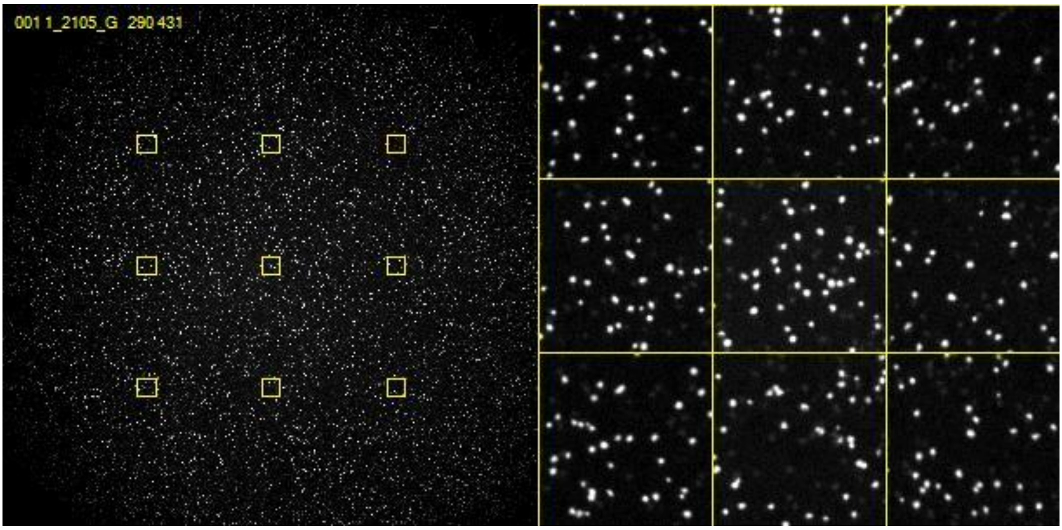
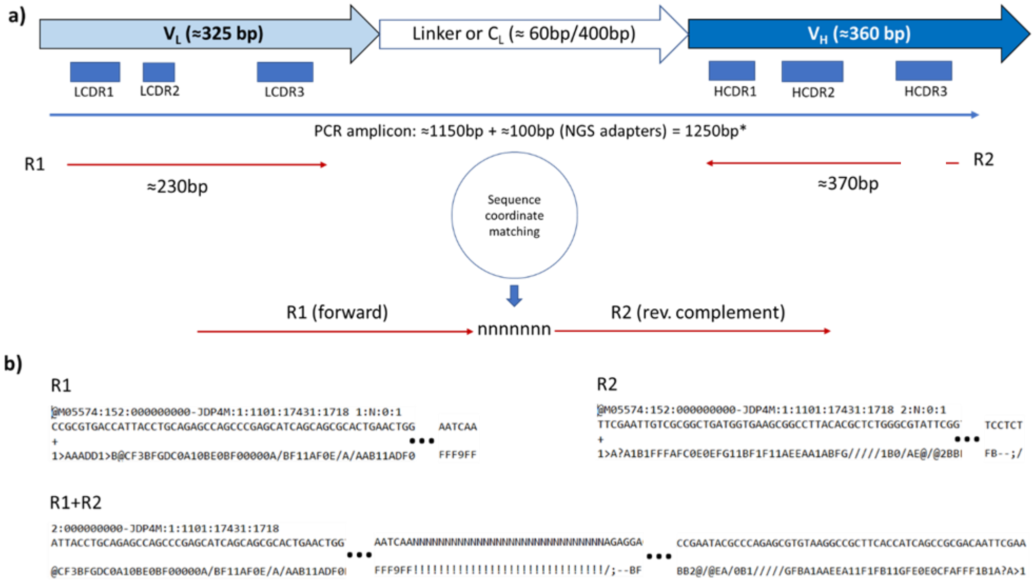


Figure 1 – NGS cycle snapshot example. The MiSeq flow cell is imaged cyclically, with each cycle corresponding to a different dye-terminated nucleotide mix (in this case, a mix of G nucleotides) that will have had hybridized to available complementary nucleotides in each cluster. All clusters are identifiable by their coordinates within the composite image (on the right). The information on <lane>:<tile>:<x-pos>:<y-pos> is then exported to the resulting sequence on the raw data FASTQ file.



R1

```
@M05574:152:000000000-JDP4M:1:1101:17431:1718 1:N:0:1
CCGCGTGACCATTACCTGCAGAGCCAGCCCGAGCATCAGCAGCGCACTGAACTGG ••• AATCAA
+
1>AAADD1>B@CF3BFGDC0A10BE0BF00000A/BF11AF0E/A/AAB11ADF0     FFF9FF
```

R2

```
@M05574:152:000000000-JDP4M:1:1101:17431:1718 2:N:0:1
TTCGAATTGTCGCGGCTGATGGTGAAGCGGCCTTACACGCTCTGGGCGTATTCGG ••• TCCTCT
+
1>A?A1B1FFFAFC0E0EFG11BF1F11AEEAA1ABFG/////1B0/AE@/@2BB     FB--;/
```

R1+R2

```
2:000000000-JDP4M:1:1101:17431:1718
ATTACCTGCAGAGCCAGCCCGAGCATCAGCAGCGCACTGAACTGG ••• AATCAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGAGGA ••• CCGAATACGCCCAGAGCGTGTAAGGCCGCTTCACCATCAGCCGCGACAATTCGAA

@CF3BFGDC0A10BE0BF00000A/BF11AF0E/A/AAB11ADF0     FFF9FF!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!/;--BF     BB2@/@EA/0B1/////GFBA1AAEEA11F1FB11GFE0E0CFAFFF1B1A?A>1
```

Figure 2 – Sequence coordinate matching. a) Schematic representation of paired-end reading and sequence coordinate matching to retrieve correlated V_L:V_H on pairs. Big Fab amplicons comprise both V_L and V_H, with C_L in between. In the case of scFV sequences, the total length of the amplicon decreases to about 900 bp. The R1 and R2 reads add up to a maximum of 600 bp (using Illumina's MiSeq system), with R1 shorter than R2 due to the bigger HCDR2 and HCDR3 loops. **b)** Sequence coordinates are the first row on FASTQ raw data files identifying each cluster. The first row contains the following elements: @<instrument>:<run number>:<flow cell ID>:<lane>:<tile>:<x-pos>:<y-pos>:<UMI> <read>:<is filtered>:<control number>:<index>. The second line identifies the nucleotide sequence while the third line (after "+") indicates the quality score of each sequenced nucleotide. R1 + R2 is composed by R1 and the reverse complement of R2, united by a string of N nucleotides.

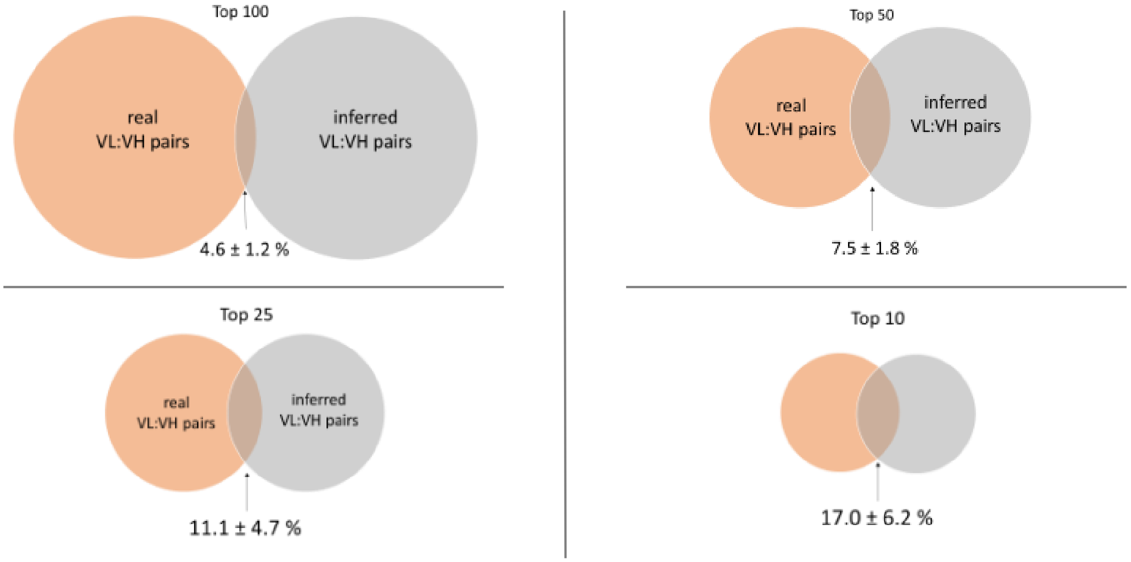


Figure 3 – Intersecting sequences between correlated and independent datasets. Each of the 23 affinity maturation projects was analyzed by the sequence coordinate matching method to get correlated information on V_L:V_H pairs (real dataset – orange circles). Those same affinity maturation projects were also analyzed using the current alternative, which analyses V_L and V_H separately. Independently analyzed V_L and V_H were then combined to generate inferred V_L:V_H pairs, based on the frequency of each clone (inferred dataset – grey circles). The real and inferred datasets were ordered by the occurrences of each sequence and, then, compared to discover identical sequences on the Top 10, 25, 50 and 100, for each affinity maturation project. (see also Table S2). The majority of the top sequences are not found in both datasets simultaneously, strengthening the correlated approach (coordinate matched dataset).

TABLES

Table 1 - Top clone comparison between the real dataset after V_L:V_H coordinate matching *versus* the inferred dataset using data from analyzing VL and VH independently. The parental CDR sequences are highlighted on the third row. Amino acids shown in black indicate deviations from the parental sequence that are equal in both datasets. Amino acids shown in red indicate deviations from the parental sequences that are different between datasets. That same top clone was searched on the inferred dataset and its ranking across the whole dataset annotated. The LCDR2 was not randomized, and thus, omitted from the table. Sequences indicated with a N/A were not found within the inferred dataset.

Real Top Clone (VL:VH coordinate matching)				**Ranking within independent dataset**
LCDR1	LCDR3	HCDR1	HCDR2	
ASTSISSYLN	QQSYSTPLT	FTFSSYAMS	AISGSGGSTYYADSVKG	
..Q.......	...**Y**.....	....	....	#N/A
....	...**Y**.....	.A.......	.........Q..S..S.	#N/A
....	....	.**A**...A...	.........S.......	#N/A
..Q.......	....	....	....	1
..Q....D..	...**D**.....	.A...**I**...	.........K.......	#N/A
....	....	.A...A...	...........S.....	1
....	....	.A.......	.........K.......	1
....	...**Y**.....	.**K**.......	...............S.	#N/A
.......A..	...A.....	.H...E...	.........S..S..S.	1
..Q....A..	...A.....	....	...........Y.....	1
..Q....A..	...A.....	.....A...	....	1
..Q....E..	...E.....	.....A...	.........S.......	1
.......A..	...A.....	.A...A...	.........K.......	#N/A
..**T**....A..	...A.....	.A.......	.........S..S....	#N/A
..**Y**....A..	...A.....	.**E**...**K**...	.........**S**..**D**..**E**.	#N/A
.......A..	...A.....	.Q.......	...........S.....	1
.......A..	...A.....	.Q.......	.........K..S..Y.	1
.......D..	...D.....	.A.......	.........S..S..S.	1
..I....K..	...K.....	.A...I...	.........S..K..S.	1
..Q....A..	...A.....	.A...A...	.........S..S..S.	1
..**T**....A..	...A.....	.A...A...	.........I..S..S.	#N/A
..**T**.......	....	....	....	#N/A
..E....A..	...A.....	.Q.......	.........S.....Q.	5

ITQB-UNL | Av. da República, 2780-157 Oeiras, Portugal
Tel (+351) 214 469 100 | Fax (+351) 214 411 277

www.itqb.unl.pt