Massive integration of large gene libraries in the chromosome of *Escherichia coli*

Lidia Cerdán[1a], Beatriz Álvarez[1b] and Luis Ángel Fernández[1c]*

1) Department of Microbial Biotechnology, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), Campus UAM Cantoblanco, 28049 Madrid, Spain.

***Running title*:** Massive gene integration in *Escherichia coli*

***Keywords*:** *E. coli* /chromosomal integration/ gene fusions/ gene libraries/ nanobodies/

a- ORCID: 0000-0003-4938-5848

b- ORCID: 0000-0002-9613-5473

c- ORCID: 0000-0001-5920-0638

---

*For correspondence: Dr. Luis Ángel Fernández; E-mail: lafdez@cnb.csic.es;

Phone:+34 91 585 48 54; Fax:+34 91 585 45 06;

Abstract

Large gene libraries are commonly constructed in *Escherichia coli* plasmids, which often show cell toxicity and expression instability due to high-copy gene dosage. These limitations could be minimized by integrating gene libraries in single copy in the bacterial chromosome. In this work, we describe an efficient system for massive integration (MAIN) of large gene libraries in *E. coli* chromosome generating in-frame gene fusions that are stably expressed. MAIN utilizes a thermosensitive integrative plasmid that is linearized *in vivo* to trigger massive integration of the gene library by homologous recombination. Efficient positive and negative selections eliminate bacteria lacking gene integration in the target site. We tested MAIN with a library of 10^7 V_{HH} genes, which encode nanobodies (Nbs). Integration of the V_{HH} genes in a custom target locus of *E. coli* chromosome allowed stable expression and surface display of the Nbs. Next-generation DNA sequencing confirmed that MAIN preserved diversity of the gene library after integration. Lastly, we have screened the integrated library using cell sorting methods to select Nbs binding a specific antigen. These results demonstrate that MAIN allows the massive integration of large gene libraries in *E. coli* chromosome, generating stably expressed in-frame fusions for functional screenings.

Introduction

Cloning of large gene libraries from natural or synthetic repertoires is an essential step of combinatorial biology and directed evolution strategies for the selection of improved enzymes and antibodies (Bornscheuer, et al., 2012, Tiller, et al., 2017, Simon, et al., 2019). *E. coli* is the most common host for cloning of gene libraries due to its high efficiency of transformation and the variety of available expression vectors, mostly based on multicopy plasmids (Rosano and Ceccarelli, 2014). Nonetheless, the use multicopy plasmids induce significant physiological burden for bacteria due to high expression level of the recombinant proteins (Dumon-Seignovert, et al., 2004). Plasmid-based vectors are also inherently unstable and require a selective pressure (e.g., antibiotics) to ensure maintenance. These factors could be limiting for the selection of low-frequency clones from large gene libraries, in which loss of functional expression and overgrowth of non-expressing or plasmid-free clones should be minimized.

The insertion of genes in the bacterial chromosome provides a simple way to overcome many of the drawbacks of plasmid-based systems, stabilizing the exogenous DNA and reducing gene dosage to single-copy (Kuhlman and Cox, 2010, St-Pierre, et al., 2013, Zucca, et al., 2013). Different methodologies have been developed for random or site-specific integration of genes into the bacterial chromosome (Ou, et al., 2018, Li, et al., 2019). Insertion at permissive sites is, in general, a good choice to guarantee similar bacterial growth and fair comparison of gene expression levels among different clones. Site-specific transposases (e.g., Tn7) and integrases from bacteriophages (e.g., λInt) have been used to insert genes at specific attachment (*att*) sites in the chromosome (e.g. *attTn7*, *attB*) (Landy, 2015, Peters, 2019). Engineering these sites at different positions of the chromosome also enables integration at multiple loci (Egger, et al., 2020). Alternatively, transposon-encoded CRISPR-Cas systems with designed crRNA molecules can guide the insertion of the exogenous DNA into a custom target site of the genome (Vo, et al., 2020). However, all these systems require the cargo DNA to be flanked with short DNA sequences (~20-40 bp) recognized by the transposase or integrase, and which are inserted along

with the cargo DNA. The presence of these "scar" sequences makes difficult the generation of in-frame fusions of the exogenous DNA with a target gene in the chromosome.

More versatile scarless integration systems rely on homologous recombination between the chromosome and an exogenous DNA molecule carrying homology regions (HRs) of the target site in the chromosome. Cellular and bacteriophage recombinases (e.g., RecA, λRed) can recombine the chromosome with the HRs of a circular or linear DNA introduced in the bacterium (Datsenko and Wanner, 2000, Kuhlman and Cox, 2010). Since homologous recombination events are rare, double crossovers of the HRs flanking a cargo DNA need to be positively selected, often with an antibiotic resistance (Ab^R) gene marker. These Ab^R markers may be later removed flanking them with short sequences (e.g., *FRT*) recognized by site-specific recombinases (e.g., FLP) (Datsenko and Wanner, 2000).

Strategies for markerless integrations have also been reported (Posfai, et al., 1999, Feher, et al., 2008). Markerless integration depends on the generation of double strand breaks (DSBs) in the chromosome, lethal lesions for bacteria unless repaired by a homologous recombination with the exogenous DNA. DSBs can be induced *in vivo* by expressing a restriction endonuclease such as I-SceI, which recognizes a long target sequence that is not found in bacterial chromosomes but that can be incorporated in a suicide plasmid vector carrying the HRs (Posfai, et al., 1999, Herring, et al., 2003). Thus, after the first homologous recombination that integrates the circular suicide plasmid vector, the chromosome becomes susceptible to cleavage by I-SceI. The DSBs generated by the endonuclease in the chromosome are lethal for bacteria unless they are repaired by a second homologous recombination event that eliminates all vector sequences and leaves a markerless integration of the cargo DNA. Similarly to I-SceI, DSBs can be generated by expression of CRISPR-Cas9 nuclease and crRNAs toward a target sequence of the chromosome (Wang, et al., 2016, Wang, et al., 2022).

However, despite the diversity of available integration systems, there are no reports of the integration of large gene libraries in the chromosome of bacteria (Li, et al., 2019). To the best of our knowledge, the largest integrated gene library in bacteria used random transposition (e.g. mini-Tn5) and contained ca. $1.4x10^5$ independent clones (Scholz, et al., 2019). Other integrated libraries are limited to maximum of $3x10^4$ clones (Cowie, et al., 2006, Elmore, et al., 2017, Biggs, et al., 2020, Saleski, et al., 2021, Parisutham, et al., 2022). The development of efficient systems for integration of large gene libraries would facilitate more effective screening of combinatorial libraries and the implementation of continuous selection processes in synthetic evolution (Simon, et al., 2019).

In this work we have designed an efficient system for the massive integration (MAIN) of large gene libraries into the chromosome of *E. coli*. MAIN uses homologous recombination to generate scarless in-frame fusions with a target gene in the chromosome, producing protein fusions that are stably expressed after integration, allowing the continuous selection of clones with functional activity. MAIN is based on a thermosensitive (*ts*) suicide plasmid with I-*Sce*I sites, that generates *in vivo* a linear double-stranded DNA fragment for homologous recombination, and strong selection and counterselection systems for effective clearance of non-integrants. The linear DNA generated *in vivo* contains HRs flanking the cargo gene (segment) of interest (GOI), which is devoid of a transcriptional promoter to ensure that it is expressed only after its integration in the chromosome. As a proof-of-concept, we have integrated in the chromosome of *E. coli* a library of V_{HH} genes raised against the human epidermal growth factor receptor (EGFR) containing ~$1x10^7$ independent clones (Salema, et al., 2016).

The V_{HH}s are the variable domains of the heavy chain-only antibodies (HCAbs) naturally found in camelids (Hamers-Casterman, et al., 1993, Muyldermans, et al., 2001). The V_{HH} gene segments encode functional single-domain antibody (Ab) fragments, called nanobodies (Nbs), which are attractive molecules for therapeutic and diagnostic applications (Muyldermans, 2013, De Meyer, et al., 2014, Jovčevska and Muyldermans, 2019, Yang and Shah, 2020). We have

used MAIN to integrate this large V_{HH} library in the chromosome of an "acceptor" *E. coli* strain, customized to have in its chromosome a gene segment coding for the outer membrane (OM)-anchoring domain of intimin, called Neae (Bodelón, et al., 2009, Salema, et al., 2013). Gene integration led to in-frame fusions with Neae and the display of the Nbs on the bacterial surface. Massive DNA sequencing of integrated clones showed that MAIN maintained the clonal diversity of original gene library. Lastly, we also demonstrate that EGFR-binding Nbs can be selected by screening the integrated library for antigen-binding clones using magnetic- and fluorescence-activated cell sorting (MACS, FACS) (Salema, et al., 2016, Salema and Fernández, 2017). Taken together, our results demonstrate that MAIN enables massive integration of large gene libraries in the chromosome of *E. coli*, producing in-frame gene fusions that are stably expressed and allow the selection of clones of interest by functional screenings of the integrated library.

Results

Design of the MAIN system.

We designed MAIN combining the approach for scarless gene integration by homologous recombination with selection and counterselection markers for the efficient recovery of integrants and the elimination of non-integrants. The "donor" vector of MAIN is a thermosensitive plasmid in which the library of GOI is cloned flanked by two I-SceI sites and HR sequences of the target gene in the chromosome of an "acceptor" bacterial strain. Replication of the donor plasmid at low temperatures (30 ^{o}C) enables the cloning and propagation of the gene library in *E. coli*. A shift of the incubation temperature (37 ^{o}C) along with the induction of I-SceI endonuclease and λRed proteins (Exo, Bet, Gam) *in vivo* (Herring, et al., 2003) stops plasmid replication and generates a linear double-stranded DNA fragment that can undergo homologous recombination with the target site in the chromosome of the acceptor strain. For the positive selection of integrants, we incorporated the apramycin resistance (Apra^{R}) gene *aac(3)IV* (Magalhaes and Blanchard, 2005) in the donor vector, downstream of the GOI (Figure 1). This marker is not commonly found in *E.*

coli plasmids, allowing the transformation of Apra^R strains with multiple vectors without further manipulation. Nonetheless, FRT-sites flanking the *aac(3)IV* gene were added for its (optional) deletion with FLP recombinase (Datsenko and Wanner, 2000). For effective elimination of non-integrants, we employed the counterselection cassette *tetA-sacB* (Li, et al., 2013). The *tetA-sacB* cassette encodes TetA, a cytoplasmic membrane protein that confers both resistance to tetracycline (Tet^R) and susceptibility to fusaric acid (Fus), and SacB, a levansucrase that converts sucrose (Suc) to levan, which is toxic for the bacteria when accumulated in the periplasm. Thus, *E. coli* bacteria having *tetA-sacB* in the chromosome are killed in rich media containing Fus and Suc. Since both proteins exert their toxic activities independently, the frequency of appearance of spontaneous mutants resistant to the selective Fus+Suc medium ($<10^{-6}$) is orders of magnitude lower than with either of these counterselection genes alone (Li, et al., 2013). This characteristic was suitable for the effective elimination of bacteria lacking the expected homologous recombination, in which *tetA-sacB* is not replaced by the GOI.

Implementing donor vector and acceptor strain for integration of a V_{HH} gene library.

As a proof of concept of the MAIN system, we chose the integration of a V_{HH} gene library of ca. $1x10^7$ independent clones (Salema, et al., 2016) in the chromosome of an acceptor *E. coli* strain (EcM1-NL; Supporting Table S1) modified with a genetic construct encoding the intimin Neae fragment (Salema, et al., 2013, Salema and Fernández, 2017). This genetic construct was integrated in the *flu* gene of *E. coli* K-12 chromosome (van der Woude and Henderson, 2008) under the control of the Ptac promoter (de Boer, et al., 1983) and a strong ribosome binding sequence from T7 bacteriophage (RBS_{T7}) (Figure 1). This construct resulted in the constitutive leaky expression of Neae-V_{HH} fusions in absence of the inducer IPTG (Supporting Figure S1). We have previously showed the constitutive expression of Neae-V_{HH} gene fusions integrated in the *flu* site are stable and express the fusion proteins in the OM, displaying functional Nbs on the bacterial surface (Piñero-Lambea, et al., 2015, Al-ramahi, et al., 2021). Lastly, the acceptor EcM1-NL strain was engineered to contain the *tetA-sacB* cassette downstream of the Neae gene

segment (Figure 1) for counterselection of bacteria in which a V_{HH} gene was not integrated replacing *tetA-sacB* cassette.

The donor vector pRecomb-TS (Figure 1, Supporting Table S2) was based on the backbone of the thermosensitive plasmid pGETS (Ruano-Gallego, et al., 2015), which contains a kanamycin resistance marker (Km^R), Ori101, RepA101(ts), and a multiple cloning site (MCS) flanked by two I-SceI sites. Ori101 confers a low-copy plasmid phenotype (ca. 5 copies per bacterium at 30ºC) (Hashimoto-Gotoh and Sekiguchi, 1977) and exhibits class A theta replication (Lilly and Camps, 2015). In pRecomb-TS, the MCS was modified with a DNA insert comprising: 1) the first homology region (HR1) of ~500 bp corresponding to the 3'-end of the Neae gene segment; 2) the GOI (i.e., a stuffer DNA or a V_{HH} in frame with the upstream Neae) between unique *Sfi*I a *Not*I sites; 3) a short DNA segment encoding an in frame C-terminal myc-tag, a translation stop codon and a downstream transcriptional terminator (T0); 4) the $Apra^R$ gene marker with flanking FRT-sequences; and 5) the second homology region (HR2) of ~500 bp corresponding to the 3' end of the *flu* locus of *E. coli* K-12 chromosome. The whole DNA insert can be released *in vivo* by I-SceI cleavage, generating a linear DNA fragment carrying the HRs, the V_{HH} and $Apra^R$ gene and which can be integrated through a double recombination event in the target site of the chromosome of EcM1-NL (i.e., *Neae-tetA-sacB*). This would result in the assembly of full-length in-frame Neae-V_{HH} fusions that are expressed from the chromosomal Ptac promoter (Figure 1). The donor plasmid does not have any promoter driving transcription of the cloned V_{HH} to avoid any potential toxicity due to Nb expression from this high-copy plasmid. Further, in case of spurious transcription from the plasmid, the polypeptide encoding the Nb does not contain a RBS for translation and cannot be displayed on the bacterial surface as is fused only to a truncated extracellular region of Neae (residues 493 to 654 of intimin), which lacks both the N-terminal signal peptide and the β-barrel domain required for OM localization (Bodelón, et al., 2009, Fairman, et al., 2012).

In order to minimize the size of the donor vector, the genes encoding the I-*Sce*I enzyme and λ-Red products were not included in pRecomb-TS but co-expressed from the chloramphenicol-resistant (Cm^R) helper plasmid pACBSR (Figure 1, Supporting Table S2) upon induction with L-arabinose (L-Ara) (Herring, et al., 2003). After L-Ara induction at the restrictive temperature (37 ^oC), bacteria with the correct double-recombination event were expected to have in frame Neae-Nb fusions and the $Apra^R$ marker (Figure 1). These recombinants would have lost the *tetA-sacB* cassette, and therefore any remaining non-integrant bacteria in the $Apra^R$ population could be cleared in the Fus + Suc medium.

Validation of the MAIN system using an immune V_{HH} library against human EGFR.

For validation we chose an immune V_{HH} library against human EGFR that contains ~1.3 x 10^7 clones (Salema, et al., 2016). These V_{HH} sequences were excised from the original replicative plasmid library in pNeae2 (Salema, et al., 2016) and ligated between the *Sfi*I and *Not*I sites of pRecomb-TS. A library of ~1.5 x 10^7 clones was obtained in the highly competent *E. coli* DH10BT1R strain (Supporting Table S1) at 30 ºC, a permissive temperature for pRecomb-TS replication. The DNA plasmid library, named pRecomb-TS-V_{HH} EGFR (Supporting Table S2), was purified from the bacterial pool and electroporated into the acceptor strain EcM1-NL carrying the pACBSR helper plasmid (Figure 2). We obtained ~8x10^7 independent transformants at 30 ºC in LB-agar plates containing Cm and Km. To integrate the V_{HH} library, these bacteria were harvested and grown to exponential phase at 30 ºC in an LB liquid culture supplemented with Cm and Apra. At this step, Apra is used instead of Km for selection of the donor plasmid since only the $Apra^R$ marker is maintained after integration. Then, expression of I-*Sce*I and λ-Red from pACBSR was induced with L-Ara and temperature was shifted to 37 ºC (Figure 2) to trigger the *in vivo* generation of the linear DNA containing HR1-V_{HH}-$Apra^R$-HR2 for integration into the bacterial chromosome.

To select the integrant clones we performed two successive selection steps. In the first step bacteria were plated on large (150 mm agar plates containing LB+Apra+Cm and grown at 37 ^oC

for positive selection of bacteria with the Apra^R marker integrated in the chromosome (Figure 2, step 5). A total of ~2x10^7 Cm^R and Apra^R CFU grew on plates according to parallel bacterial counting of serial dilutions, a number above the size of the original V_{HH} library. After growth, bacteria were collected from these plates and directly plated on Fus + Suc agar medium for the second selection step, which depleted non-integrant bacteria still having the *tetA-sacB* cassette (Figure 2, step 6). Bacterial counting revealed a total of ~9x10^8 CFU growing on plates with Fus+Suc, a number exceeding the size of the original library.

Since Nb surface display depends on the correct in-frame fusion of the V_{HH} to the chromosomal Neae gene segment, we analyzed the percentage of bacteria displaying Nbs in the different bacterial populations to estimate the percentage of correctly integrated bacteria after each selection step. The Nb display was analyzed by flow cytometry by staining the myc-tag located at the C-terminus of the Nb (Figure 3A). Results revealed that ~33% of bacteria obtained after the first selection step (Apra+Cm) displayed Nbs on their surface (Figure 3B), while this number increased to ~88% after the counterselection step (Fus + Suc medium) (Figure 3B). In this final population we found ~12 % of bacteria negative for Nb display. These percentage is similar to the number of non-expressing clones (~10%) found in the original library in pNeae2 (Figure 3B). Interestingly, similar fluorescence signals of Nb-display were observed from the replicative high-copy plasmid pNeae2 (Salema, et al., 2016) and from single-copy integration in the bacterial chromosome (Figure 3B), which indicated a similar surface display level of Nbs in both expression systems. Hence, these results indicate that the integration process of the V_{HH} genes had occurred as expected, generating in-frame fusions that are displayed on the bacterial surface.

Diversity analysis of the integrated library.

To evaluate whether the integration process affected the library diversity, we performed high-throughput DNA sequencing analysis and compared the sequence diversity of the V_{HH} library before and after integration. To this end, we obtained by PCR two ~400 bp amplicons comprising

the V_{HH}s from the library before (EcM1-NL with pRecomb-TS-V_{HH} EGFR library) and after integration (EcM1-NL-V_{HH} EGFR library after the second selection step). These DNA amplicons were sequenced in an Illumina Miseq platform with paired end (length ~2 x 300 bp) to acquire ~500,000 reads per sample (ca. 5% of the estimated library size of ~$1x10^7$ clones). As the most variable region of the V_{HH}s is the CDR3, which is located close to the 3'-end of the V_{HH}, only reverse sequence reads were compared in this study. High-quality reads (ca. 75% of the raw reads; Experimental procedures) were organized in clusters of sequences with ≥ 98% of sequence identity (ID). A diversity value (DV) was established as the ratio between the number of identified clusters and the total number of analyzed sequences expressed in percentage. Thus, a DV of 100% would mean that all the sequences in the collection are different while decreased DVs would indicate the presence of repeated sequences. With this approach two sequences with an ID ≥ 98% are considered the same. Data analysis revealed that the non-integrated plasmid library had a DV of 9.80% while a DV of 7.15% was estimated for the integrated library (Table 1). Therefore, only a small reduction in the DV (%) is found after integration, indicating that the overall diversity of the V_{HH} library is maintained after its chromosomal integration.

Another variability parameter analyzed was the size of the clusters. Clusters with a large number of members would indicate the presence of repeated V_{HH} sequences in the library, so reduced variability. We compared the pattern of size distribution of the clusters before and after chromosomal integration. We found that cluster distribution was similar in the plasmid and in the integrated library (Figure 4). We also observed that most clusters from the plasmid (ca. 92%) and the integrated (ca. 86%) libraries had less than 20 members indicating a high sequence diversity in both cases. In accordance to the DV values, gene cluster distribution also indicates that the diversity of the V_{HH} library was well-maintained after the massive integration process.

Selection of Nbs binding to EGFR antigen using the integrated library

Next, we evaluated whether Nbs binding EGFR could be selected from the integrated V_{HH} library. For this purpose, bacteria from the integrated EcM1-NL-V_{HH} EGFR library were incubated with the ectodomain of human EGFR (eEGFR) labelled with biotin. Antigen-binding bacteria were enriched using MACS and FACS (Experimental procedures). Approximately $2x10^8$ bacteria were subjected to two consecutive rounds of MACS using biotinylated eEGFR-Fc as a bait. Enriched bacteria after MACS cycles were subjected to a single round of FACS by incubation with biotinylated eEGFR-Fc and c-myc mAb, to select the double positive population in antigen-binding and Nb display (Figure 5A).

Nb display levels and binding to biotinylated eEGFR-Fc of the bacterial populations after the MACS and FACS cycles were analyzed by flow cytometry (Supporting Figure S3). The results showed a consistent gradual enrichment in positive eEGFR-Fc binders, from the baseline of ca. 0.7% of the integrated library to 8.7% after second round of MACS. Remarkably, ca. 98.6% of bacteria were positive in antigen-binding after FACS (Supporting Figure S3). Nb display levels remained constant in the library along MACS selections with a significant reduction in the number of non-expressing bacteria after the FACS (Supporting Figure S3).

For the isolation of specific EGFR binders, 195 individual colonies from the FACS selection were analyzed by flow cytometry for eEGFR-Fc binding and V_{HH} sequence determination (see Experimental procedures). Eighty percent of the positive binders corresponded to a highly frequent clone also in the original pNeae2 library, known as VEGFR1 (Salema, et al., 2016), named Nb1-EGFR in this study. We identified six additional different Nbs, named sequentially from Nb2- to Nb7-EGFR depending on their frequency of isolation. The frequency of these Nbs and the amino acid sequence in their CDR3 are summarized in Table 2. From them, only Nb3-EGFR was previously isolated from the original plasmid library (Salema, et al., 2016). As determined by flow cytometry (Figure 5B), all the Nbs identified in the screening of the integrated library showed specific binding to eEGFR-Fc and not to human Fc, as nonspecific control antigen. Interestingly, Nbs with both intermediate and high antigen-binding signals were identified (Figure

5B), suggesting that Nbs of different affinities were selected in this screening. Taken together, these results demonstrate that MAIN allowed the expression of the gene fusions generated after the massive V_{HH} gene integration, enabling the screening of the integrated *E. coli* library for the selection of Nbs with specific antigen-binding capabilities.

Discussion

The MAIN system is intended to facilitate that large gene libraries could be efficiently integrated in a target gene of the bacterial chromosome, leading to large collections of in-frame fusions expressed in single copy. The design of the MAIN system needs two customized elements: 1) a thermosensitive donor plasmid for cloning the library of GOI with HRs of the target gene in the chromosome; 2) an acceptor *E. coli* strain carrying the *tetA-sacB* cassette integrated in the target gene of the chromosome. Using these elements we have demonstrated that the MAIN system enables the massive integration of large gene libraries of $>10^7$ clones in a custom target gene of the *E. coli* chromosome. The acceptor *E. coli* strain used in our study is RecA+, a cellular recombinase that promotes homologous recombination (del Val, et al., 2019). Nonetheless, since the Bet recombinase from λRed is coexpressed with I-SceI from the helper plasmid, homologous recombination between HRs of plasmid and chromosome could also occur in *E. coli* hosts with *recA* mutation (Murphy Kenan, 2016).

MAIN takes advantage of the *in vivo* expression of I-*Sce*I endonuclease, which linearizes the thermosensitive plasmid having I-SceI sites but not the bacterial chromosome in which I-SceI sites are absent (Herring, et al., 2003). In addition, co-expression of λRed recombination system (*exo gam* and *bet* genes) from the helper plasmid protects the linear double-stranded DNA from host exonucleases and facilitate homologous recombination (Pines, et al., 2015). We have found no defect in the growth of bacteria carrying the helper plasmid pACBSR in the absence of L-Ara inducer. Therefore, curing of this helper plasmid is not needed after the integration process. *In*

vivo linearization of plasmids combines the advantages of using plasmids for propagation of the GOI and linear DNA for integration. This strategy was originally reported for site-specific integration of a heterologous gene into the bacterial chromosome (Herring, et al., 2003). Since the double recombination event occurs between two different homology regions (HR1 and HR2) that flank the heterologous GOI, the integration process is stable, mono-directional, and irreversible.

MAIN system includes positive selection of integrant bacteria with the $Apra^R$ marker (Magalhaes and Blanchard, 2005), located in the donor plasmid downstream of GOI, and the counterselection cassette *tetA-sacB* (Li, et al., 2013) in the target locus of the acceptor strain for removal of non-integrant clones. The combination of positive selection using the antibiotic Apra combined with counterselection of non-integrant bacteria on Fus + Suc medium led to a highly efficient selection process resulting in a bacterial population fully composed by integrant clones. Our data indicate that a single selection step with antibiotic (Apra) is not sufficient to obtain a large gene library fully composed by integrant bacteria. In our case, only 33% of the bacteria harvested from Apra-containing plates were integrants correctly displaying the Nb. The high percentage of non-integrant bacteria after growth in Apra-plates could be attributed to an insufficient selection pressure of the antibiotic on the plates due to the high bacterial density plated when working with large libraries. However, we cannot rule out the presence of undigested plasmids in some of these bacteria or off-target integrations that could confer resistance to the antibiotic. We found that counterselection in the Fus + Suc medium was crucial to fully eliminate the non-integrant bacteria carrying *tetA-sacB* in the chromosome. The effectiveness of the *tetA-sacB* cassette in Fus+Suc medium is likely due to the low frequency of resistant bacteria ($\leq 6x10^{-7}$) using this double counterselection system (Li, et al., 2013).

We have demonstrated the potential of the MAIN system by the successful integration of an immune V_{HH} library with ~$1x10^7$ clones (Salema, et al., 2016), which represents the largest integrated library in a bacterial chromosome. Gene libraries with size $\geq 10^7$ clones are essential

in combinatorial biology and synthetic evolution strategies for the successful selection of novel peptides, enzymes and antibodies from libraries (Dufner, et al., 2006, Löfblom, 2011, Simon, et al., 2019, Rees, 2020). One important aspect of MAIN is its scalability by increasing volume of *E. coli* cultures and the plating surface for $Apra^R$ selection and Fus^R+Suc^R counterselection. In our experiments we have used agar plates of 150 mm diameter for selection and counterselections, but for integrating higher size libraries (e.g., ~10^9 clones) the plating surface should be increased to ensure that integrant clones are properly selected by these media. In theory, the maximum size of libraries that could be integrated following MAIN is only limited by the cloning efficiency in *E. coli* (~10^{10} clones for the largest reported libraries) provided that sufficient plating surface is used for selections.

In our proof of concept, the V_{HH} genes of the immune library were integrated in frame with the gene segment Neae for display of the Nbs on the bacterial surface (Salema and Fernández, 2017). The Neae fragment corresponds to the N-terminal OM-anchoring domain (residues 1-654) of EHEC intimin (Bodelón, et al., 2009, Fairman, et al., 2012, Salema, et al., 2013). According to flow cytometry analysis, ~12% of the integrants did not display correctly the Nbs with C-terminal myc tag. This was likely due to the presence of truncated V_{HH} genes (i.e., having premature stop codons and out of frame nucleotide insertions) in the original pNeae2 library (Salema, et al., 2016) and not to the integration process itself. A similar number of non-expressing clones was detected by flow cytometry in the original pNeae2 library. Since the display of Nb on the bacterial surface relies on the correct in-frame fusion of the V_{HH} with the chromosomal Neae gene segment, any duplication or gene rearrangement during integration would result in a substantial reduction in Nb display levels in the integrated library. The comparable display levels observed in both libraries, the replicative pNeae2 and the integrated library, suggest that duplications or gene rearrangements post-integration are either absent or extremely rare. This demonstrates that the MAIN system is a robust method to create in frame fusions in the chromosome by homologous recombination, which is very useful for the integration

of gene libraries encoding specific protein domains of large domain proteins, keeping the rest of the target gene unaffected.

Importantly, we have also demonstrated by high-throughput DNA sequencing that the MAIN system does not compromise the library diversity. In the case of V_{HH} and Ab libraries, a large diversity is essential for the identification of high affinity binders during the selection process, existing a correlation between the library size and the probability to found high affinity clones among the population (Bradbury and Marks, 2004). In this work, the large size of the integrated V_{HH} library allowed the identification of seven different Nbs binding EGFR specifically. Remarkably, only two of these Nbs were previously isolated from this immune V_{HH} library in the replicative plasmid pNeae2 (Salema, et al., 2016). The identification of novel Nbs binders in the integrated library suggest that low frequent V_{HH} sequences could be selected more effectively in the integrated library.

The expression of gene libraries from the bacterial chromosome leads to a more stable gene expression as compared to those constructed in high-copy plasmids. In our work, flow cytometry peaks associated to the Nb surface display levels were more heterogeneous in bacteria carrying plasmid pNeae2 than in the integrated library, in which more uniform and narrower peak was reproducibly observed. In fact plasmid expression from the Plac promoter required IPTG for induction, whereas in the case of the integrated library, leaky expression from the Ptac promoter (Wilson, et al., 2007) was sufficient to achieve good levels of display in the absence of IPTG. Integration in the chromosome of V_{HH} libraries displayed on *E. coli* could also facilitate downstream affinity maturation processes of selected nanobodies by ssDNA recombineering (Al-ramahi, et al., 2021) or the use of base deaminase-T7 RNA polymerase fusions (Álvarez, et al., 2020), as pRecomb-TS contains a reverse T7 promoter downstream of the V_{HH} (Supporting Table S2).

Another important aspect of the MAIN system is its ability to be customized for different applications beyond V_{HH} libraries. For this, the HRs of pRecomb-TS donor plasmid strain should be simply changed for those of the target integration site in the chromosome. The acceptor *E. coli* strain should also be modified to contain the *tetA-sacB* cassette in the new integration site, as shown in this work for the synthetic *flu::Neae* site. MAIN could be applied for the chromosomal insertion of large gene libraries encoding enzymes of industrial interest (Intasian, et al., 2021), optimizing screenings from metagenomic studies (Ngara and Zhang, 2018) and from directed evolution approaches after *in vitro* or *in vivo* mutagenesis (Zeymer and Hilvert, 2018). Metabolic enzymes are frequently found in chromosomal operons and their optimization for metabolic engineering requires their correct expression within the operon (Fisher, et al., 2014). MAIN allows that gene libraries of enzymes could be inserted in their natural chromosomal context, ensuring a balanced expression with other enzymes of the metabolic pathway. Another potential applications of MAIN are the integration of synthetic DNA sequences for information storage in bacterial populations (Hao, et al., 2020, Bencurova, et al., 2023) or the tagging of synthetic bacteria with DNA barcodes for tracking of individual strains (Tellechea-Luzardo, et al., 2022).

Conclusions

MAIN represents a powerful, scalable, and customizable strategy for massive integration of large gene libraries in the chromosome of *E. coli* enabling the generation of precise gene fusions for screening and selection of protein variants of interest. Adaptation of MAIN to other bacterial species beyond *E. coli* (and closely related enterobacteria) will require appropriated selection and counterselection markers for the different bacteria, but the basic principles of I-SceI cleavage for linearization of a conditional replicative plasmid for homologous recombination can be applied in many bacterial host that are amenable to genetic manipulation, such as *Pseudomonas* (Martínez-García and de Lorenzo, 2011) *Streptomyces* (Fernández-Martínez and Bibb, 2014), *Bacillus* (Wang, et al., 2018), *Clostridium* (Zhang, et al., 2015), *Lactobacillus* (Van Zyl, et al., 2019) and *Mycoplasma* (Piñero-Lambea, et al., 2022).

Experimental procedures

Bacterial strains, media, and growth conditions.

The *E. coli* strains used in this work are described in Supporting Table S1. The strain DH10BT1R (Durfee, et al., 2008) was used for propagation of plasmids and cloning. Bacterial strains were grown in liquid or solid agar lysogeny broth (LB) (Miller, 1992) at 30°C or 37 °C, as indicated. Bacteria for electrocompetent cell preparation were grown at 37 °C with shaking (250 rpm). Bacteria carrying cloned V_{HH} genes in plasmids pNeae2 or pRecomb-TS (Supporting Table S2) were grown at 30 °C with shaking (170 rpm), unless indicated otherwise. For preparation of LB solid medium, 1.5 % (w/v) agar (Gibco, Thermo Fisher Scientific) was added. Bacteria grown on solid agar media were spread to obtain individual colonies in conventional Petri dishes (∅90 mm). In the case of libraries, bacteria were grown as lawns on large plates (∅150 mm, p150) and serial dilutions were also plated on conventional Petri dishes for CFU counting. Starter liquid cultures of 10 mL were inoculated with individual colonies when working with single clones or, in the case of libraries, from a mixture of bacteria freshly harvested from plates (initial OD_{600} of the culture ~0.05). Starter cultures were grown overnight (O/N) at 30°C under static conditions, unless otherwise indicated. For depletion of non-integrants, bacteria were grown for 48 h at 37 ºC on agar Tet/SacB counter-selection medium, prepared as described previously (Li, et al., 2013). This solid medium contains per liter: 15 g of agar, 4 g of tryptone, 4 g of yeast extract, 8 g of NaCl, 8 g of NaH_2PO_4 H_2O, 0.11 g $ZnCl_2$, 24 mg fusaric acid (Fus) and 60 g sucrose (Suc), referred in this work as Fus+Suc medium. When required, antibiotics and inducers were added to the media at the following concentrations: ampicillin (Amp) at 150 μg/ml, chloramphenicol (Cm) at 30 μg/ml, kanamycin (Km) at 50 μg/ml, apramycin (Apra) at 50 μg/ml and tetracycline (Tet) at 15 μg/ml, isopropylthio-β-D-galactoside (IPTG) at 0.1 mM, L-arabinose (L-Ara) at 0.2% (w/v)

unless otherwise indicated. Antibiotics were obtained from Duchefa-Biochemie. Chemical reagents were obtained from Merck-Sigma.

Plates and starter liquid cultures of bacteria carrying derivatives of pNeae2 (Supporting Table S2) contained 2% (w/v) glucose for repression of the *lac* promoter before induction. To induce a culture with IPTG, bacteria corresponding to an OD_{600} of 0.5 were harvested by centrifugation (4000 x g, 5 min) from a starter culture, washed twice with 1 volume of liquid LB, and resuspended in 10 ml volume of LB with 0.1 mM IPTG. Induced bacteria were incubated at 30 ºC for 3 h with shaking (170 rpm).

Plasmids, DNA cloning and oligonucleotides.

Plasmids used in this work are detailed in Supporting Table S2. Cloning procedures were performed using standard techniques of DNA manipulation, ligation and transformation (Ausubel, et al., 2002). Details of DNA constructs are described in Supporting Experimental procedures. All DNA constructs were sequenced using the chain-termination method (Macrogen). The thermosensitive plasmid pRecomb-TS (Km^R) is a derivative of pGETS (Km^R, pSC101-ts origin of replication) (Ruano-Gallego, et al., 2015) used for cloning the V_{HH} gene library and chromosomal integration. The insert of pRecomb-TS comprises: a ~500 bp HR1 of the 3'-end of intimin N-terminal domain ('Neae) corresponding to residues 493 to 654 from enterohemorrhagic *E. coli* (EHEC) intimin (Salema, et al., 2013), a ~1 kb stuffer DNA (*xylE*) between unique *Sfi*I and *Not*I sites, a c-myc-epitope tag, a stop codon, a reverse T7 promoter (taatacgactcactataggg), a transcriptional terminator (T0) (http://parts.igem.org/Part:BBa_B0010), the Apramycin resistance ($Apra^R$) marker (Magalhaes and Blanchard, 2005) flanked by FRT sites, and ~500 bp HR2 of the 3'-end of *E. coli* K-12 *flu* gene (van der Woude and Henderson, 2008). Oligonucleotides used for DNA amplification and sequencing were synthetized by Sigma and are listed in Supporting Table S3.

Genome modifications of *E. coli* strains EcM1-Ptac-Vgfp and EcM1-NL.

The *E. coli* strain EcM1-Ptac-NVgfp (Supporting Table S1) was generated from an EcM1Δ*lacI* strain (Supporting Experimental procedures) by integration in the chromosomal *flu* locus of a gene cassette having the *lacI*^{q}*-Ptac* region (Amann, et al., 1988) controlling expression of an intimin Neae-Nb fusion binding GFP (NVgfp) (Salema, et al., 2013) followed by the T0 terminator and the Apra^{R} marker. This integration was performed using the thermosensitive plasmid pGETS*flu*NVgfp-Apra^{R} (Supporting Table S2). The acceptor strain EcM1-NL was originated from EcM1-Ptac-NVgfp strain by replacing the DNA region comprising Vgfp and Apra^{R} by the *tetA-sacB* counterselection cassette (Li, et al., 2013) using pRecomb-TS-tetAsacB (Supporting Table S2). Details of construction of EcM1-Ptac-NVgfp and EcM1-NL strains are described in the Supporting Experimental procedures.

Cloning of the V_{HH} gene library

The DNA sequences of V_{HH}s from the immune library against human EGFR were excised from the plasmid pool of pNeae2-V_{HH} EGFR library (Salema, et al., 2016) by *Sfi*I and *Not*I digestion and cloned into the same sites of pRecomb-TS (Km^{R}), replacing the stuffer DNA *xylE*. Digested V_{HH} DNA fragments were run into an agarose gel (1% w/v), stained with SYBR safe and visualized under a blue light transilluminator (ThermoFisher Scientific). The band of ~400 bp corresponding to the V_{HH} fragments was cut and purified using PureLink™ Quick Gel Extraction Kit (ThermoFisher Scientific). Five hundred ng of purified V_{HH} fragments (in ddH_2O) were ligated with the backbone of pRecomb-TS vector (Supporting Table S2) previously digested with *Sfi*I (ThermoFisher Scientific) and *Not*I (New England Biolabs) and gel-purified as above. The ligation reaction was prepared in a final volume of 0.2 ml at a 3:1 insert:vector molar ratio using a final vector concentration of ~2 ng/μl and 5U of T4 DNA ligase (Merck-Roche). After O/N incubation at 16 °C, the ligation products were ethanol-precipitated and resuspended in ddH_2O in a final DNA concentration of ~50 ng/μl. Ligation products were electroporated into *E. coli* DH10B-T1^{R} cells (~500 ng of DNA per aliquot of 100 μl of competent cells). Six 100 μl aliquots of electrocompetent cells were used to reach the library size ~1.3 x 10^{7} clones. Plasmids from the

library were purified (Midi-prep kit, Qiagen) from the pool of transformed *E. coli* DH10B-T1^R (grown as a lawn on LB) and transferred by electroporation into *E. coli* EcM1-NL electrocompetent cells containing the pACBSR (Cm^R) (Herring, et al., 2003) to obtain >1.3 x 10^7 clones. Library size was determined by plating serial dilutions on LB-Km or LB-Km-Cm plates and CFU counting.

Integration of the V_{HH} gene library in the EcM1-NL chromosome.

EcM1-NL/pACBSR transformants carrying pRecomb-TS-V_{HH} EGFR library (Supporting Table S2) were harvested from p150 LB-Km-Cm plates and used to prepare a single 5 ml starter LB liquid culture supplemented with Apra and Cm at an initial OD_{600} of 0.05. This culture was incubated at 30 °C with shaking (250 rpm) until reaching an OD_{600} of 0.5. Next, culture was induced for integration by adding L-arabinose (L-Ara) at 0.2% (w/v) and the temperature was increased up to 37 °C. After 3.5 h of induction, 2 ml of bacterial culture were plated as a lawn (0.5 ml/plate) on LB-Apra-Cm p150 plates (First selection step). In parallel serial dilutions were plated on conventional Petri dishes with the same medium for CFU counting. Once a library size of >1.3 x 10^7 clones was obtained, bacteria were harvested from these plates using 4 ml of liquid LB medium per plate and a cell spreader (Digralsky spatula). Finally, 1.5 ml of harvested bacteria were re-plated as a lawn (0.5 ml/plate) on Fus+Suc medium p150 plates for depletion of non-integrant bacteria by counterselection (Second selection step). In parallel, serial dilutions were plated on conventional Petri dishes with the same selective medium. After 48 h of incubation at 37 ºC, library size was assessed by CFU counting (>1.3 x 10^7) and bacterial lawn was harvested.

High-throughput DNA sequencing of V_{HH} sequences for diversity analysis.

To analyze the diversity of the V_{HH} library, high-throughput DNA sequencing of the V_{HH} sequences was performed. To do that, the V_{HH} sequences (~400 bp) were amplified by PCR using primers CS1-E-tag and CS2-c-myc-tag (Supporting Table S3) and as DNA template the plasmid pRecomb-TS-V_{HH} EGFR library (before integration) or purified chromosomal DNA from

the integrated EcM1-NL-V_{HH} EGFR library (Supporting Table S1). Plasmid Midi-prep kit (Qiagen) and GENOME® DNA isolation kit (MP Biomedicals) were used for the purification of plasmid and genomic DNA, respectively. Each PCR reaction contained: 1.5 μl of CS1-E-tag oligo at a concentration of 10 μM, 1.5 μl of reverse CS2-c-myc-tag oligo at a concentration of 10 μM, 5 μl of GoTaq® G2 Flexi Reaction Buffer (10X) (Promega), 4.5 μl of $MgCl_2$ (Promega), 1 μl of dNTP mix (dA;dC;dG;dT) at a concentration of 2.5 mM, 0.5 μl of GoTaq® G2 Flexi DNA Polymerase (Promega) and ddH_2O up to a final volume of 50 μl. In these reactions 18.9 pg of the V_{HH} library in pRecomb-TS plasmid or 31.2 ng of genomic DNA from the V_{HH} library integrated in EcM1-NL were used as templates in the PCR. These amounts correspond to ~$3x10^5$ and $6x10^5$ molecules, of pRecomb-TS-NL (5500 bp) and *E. coli* chromosomal DNA ($4.6x10^6$ bp), respectively. PCR program included: 1 cycle of 2 min at 94 °C, 30 cycles of 1 min at 94 °C, and 2 min at 72 °C (amplification) and a final cycle of 10 min at 72 ºC. Obtained amplicons of ca. 500 bp were run into agarose gels (0.8% w/v), the corresponding band was excised and purified with PureLink™ Quick Gel Extraction Kit (ThermoFisher Scientific) and finally sequenced using Illumina Miseq platform with paired-end (length > 2 x 300 bp) to acquire ~500,000 reads per sample in the Genomic Unit service of the Technologic Park of Madrid.

For this study, ~500,000 reverse reads containing the CDR3 regions of the V_{HH} sequences were analyzed per pool (pool 1= Plasmid library, pool 2= Integrated library). The reads were filtered and adaptor trimmed for improving their quality with the AlienTrimmer 0.4.0 software (Criscuolo and Brisse, 2013) with the following parameters: minimum quality (q)=28, conservativity (k)=10, maximum mismatch (m)=10 and minimum length (l) =175. The quality of the reads was determined by FastQC software (Patel and Jain, 2012, Leggett, et al., 2013) before and after trimming. Improved sequences were then clustered attending to their sequence identity (ID) using CD-HIT-EST online server (Huang, et al., 2010) and setting an ID threshold of 98%. Obtained data were subsequently used to estimate the diversity of the library attending to two different parameters: i) the diversity value (DV); [DV= (total No. of clusters/total analyzed reads)

x 100] and ii) the percentage (%) of low membership clusters. Arbitrary, clusters with 20 or less members were considered as low membership clusters.

E. coli magnetic cell sorting and fluorescence-activated cell sorting.

For selection of bacteria binding eEGFR-Fc (the ectodomain of human EGFR fused to the Fc domain of human IgG1), magnetic cell sorting (MACS) and fluorescence activated cell sorting (FACS) were used. First, MACS was performed to enrich binders from the library. Bacteria of the integrated EcM1-NL-V_{HH} EGFR library were harvested from lawns grown on Fus+Suc p150 plates and used as inoculum of a 10 ml LB-Apra liquid culture at an initial OD_{600} of 0.5 (>100 times more CFU than the clonal size of the library). After 3 h of growth with shaking (170 rpm) at 30 ºC (final OD_{600} ~1.5), ~$5x10^8$ bacteria were harvested by centrifugation (4000 x g, 3 min, RT) and washed twice with 2 ml of PBS (1X). Then, bacteria were resuspended in 100 µl of PBS-BSA [PBS (1X) supplemented with 0.5% (w/v) of BSA] and mixed with 100 µl of biotinylated eEGFR-Fc in the same buffer at a final concentration of 100 nM. Biotinylation of eEGFR-Fc protein (R&D Systems) was performed as described previously (Salema, et al., 2016). After 1 h of incubation at RT, bacteria were harvested by centrifugation, washed twice with 2 ml of PBS-BSA and resuspended in 100 µl of the same buffer with 20 µl of anti-biotin paramagnetic beads (Miltenyi Biotec). After 20 min of incubation at 4 °C, bacteria were harvested by centrifugation, washed, resuspended in 500 µl of PBS-BSA and loaded onto a MACS MS column (Miltenyi Biotec) which had been previously equilibrated with 500 µl of the PBS-BSA and placed on an OctoMACS Separator (Miltenyi Biotec). Unbound bacteria were collected, and the MACS column was washed 3 times with 500 µl of PBS-BSA. Unbound bacteria and washed volumes were combined as “unbound fraction”. Next, MACS column was removed from the separator and placed onto a collector tube for elution of the column-bound bacteria with 2 ml of LB. This “bound fraction” was plated as a lawn on LB-Apra p150 plates (0.5 ml/plate) for later bacterial harvesting. In parallel, serial dilutions of bound and unbound fractions were plated on the same medium for CFU counting. For a subsequent MACS round, bacterial lawn grown from the previous MACS

were harvested and used as inoculum for a new starter culture, which was grown and processed as before.

After MACS, the final selection of EGFR-binding bacteria was performed by FACS. Bacteria captured by MACS were used to inoculate a starter LB-liquid culture (OD_{600}~0.05), which was incubated O/N at 30 °C under static conditions followed by 2 h at the same temperature and 170 rpm (final OD_{600}~1.5-2). Bacteria were harvested by centrifugation, washed, and resuspended like in MACS, followed by a double staining for surface display expression and antigen binding as detailed below for flow cytometry analysis. Finally, samples were resuspended in 1 ml of PBS and sorted in a FACS vintage SE sorter cytometer (Becton Dickinson). The bacterial population positive to both fluorophores was collected in a sterile tube containing 2 ml of liquid LB-Apra medium and grown as a lawn (0.5 ml/plate) on LB-Apra p150 plates. A total of ~$1x10^7$ bacteria were processed by FACS.

Flow cytometry analysis of bacteria for Nb surface display and antigen binding.

The Nb display levels and antigen binding capacity of selected bacteria (individual clones or libraries) were analyzed by flow cytometry. For these assays, ~10^9 *E. coli* cells from LB liquid cultures inoculated with a single colony (individual clones) or with lawn harvested bacteria at an OD_{600} of ~0.5 (libraries) and grown for 3h at 30 °C with shaking (170 rpm) were harvested by centrifugation (4000 x g, 3 min, RT), washed 3 times with 500 μl of PBS (1X) and resuspended in 400 μl of the same buffer. Then, 90 μl of cell suspension (~$2x10^8$ bacteria) was taken and incubated for 1h at RT with 10 μl of primary antibody (for Nb display analysis) and/or labeled proteins diluted in PBS (1X) (for binding analysis). Mouse anti-c-myc monoclonal Ab (1:500; 9B11 clone; Cell Signaling, Ref: 2276) was used for staining of bacteria displaying Nbs. Biotinylated eEGFR-Fc and human Fc (prepared as described in Supporting Experimental procedures) were used at 50 or 100 nM for antigen-binding analysis. Then, bacteria were washed 3 times with 500 μl of PBS (1X) and incubated for 1h at 4 °C in the darkness with 100 μl of PBS (1X) containing

the corresponding secondary reagent. Goat anti-mouse IgG-Alexa 488 conjugated polyclonal Ab (1:500; ThermoFisher Scientific, Ref: A11029) was used for the detection anti-myc mAb (Nb surface display) on bacteria while Streptavidin-APC (1:100; Beckman Coulter, Ref: 733001) was used for detection of biotinylated antigen on bacteria (antigen binding). Finally, samples were washed 3 times with PBS (1X) and resuspended in 500 μl of the same buffer for analysis in a Gallios cytometer (Beckman Coulter). Around $5x10^4$ bacteria were analyzed per sample.

Identification and DNA sequencing of specific V_{HH} clones.

The V_{HH} sequences of the specific anti-EGFR clones integrated in EcM1-NL chromosome were amplified by colony PCR with oligonucleotides eae5 and HindIII-Ter T0 (Supporting Table S3) using the GoTaq® G2 Flexi DNA Polymerase (Promega). Obtained PCR products of ca. 600 bp were run into an agarose gel (0.8% w/v) and purified from agarose bands. The PCR amplicons were sequenced by the chain-termination method (Macrogen) using the primer eae5. To identify low-frequent clones different from the most frequent clone of the library (Nb1-EGFR), individual colonies were screened by PCR to detect colonies carrying this Nb with the reverse oligo VEGFR1-CDR3 and the universal forward oligo VHH-Sfi2 (Supporting Table S3). Colony PCR reactions were carried out using the NZYTaq II 2x Green Master Mix (NZYtech, Ref: MB358) following manufacturer's instructions. Colonies that did not shown amplification bands (~400 bp) were analyzed by flow cytometry for antigen binding capacity and the cloned V_{HH} sequenced as described above.

Data Availability

Sequencing data from highthrough-put DNA sequencing experiments are deposited in Sequencing Read Archive (SRA) in the Bioproject ID PRJNA1000930 and are freely available from the following URL https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1000930.

AlienTrimmer 0.4.0 software (Criscuolo and Brisse, 2013) was used to trim adapter sequences and low-quality bases from raw Illumina reads. The software is freely available for download from the following URL https://bioweb.pasteur.fr/packages/pack@AlienTrimmer@0.4.0 The source code is also available upon request.

FastQC software (Andrews, 2010) was used to evaluate the quality of the trimmed reads. The software is freely available for download from the following URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. The source code is also available upon request.

CD-HIT-EST online server (Li and Godzik, 2006) was used to cluster the assembled transcripts. The server was freely available for use at the following URL https://sites.google.com/view/cd-hit/web-server. The on-line version is not available but the program can be run as a command line tool or a local CD-HIT server can also be downloaded from this URL: https://github.com/weizhongli/cdhit-web-server.

Genebank accession numbers for plasmids: pGETSfluNVgfp (OR359883), pRecomb-TS-tetAsacB (OR359884), pRecomb-TS (OR359885), pRecomb-TS-Vgfp (OR359886).

Acknowledgements

We thank the excellent technical support of CNB-CSIC core scientific facility "Flow Cytometry" and of the Flow Cytometry service at SidI-UAM. We also thank to the Genomic Unit of "*Parque Científico de Madrid*" for their technical work in massive DNA sequencing.

Funding

This work was supported by research grants to L.A.F.: MCIN/AEI BIO2017-89081-R, MCIN/AEI and NextGeneration EU/ PRTR (PLEC2021-007739), and the European Union's Horizon 2020

Future and Emerging Technologies research and innovation program (FET Open 965018-BIOCELLPHE).

Author contributions

Lidia Cerdán: Conceptualization (equal); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); visualization (lead); validation (supporting); writing – original draft (equal); writing – review and editing (supporting).

Beatriz Álvarez: Conceptualization (equal); formal analysis (supporting); investigation (supporting); methodology (supporting); supervision (supporting); validation (supporting); writing – original draft (equal); visualization (supporting); writing – review and editing (supporting).

Luis Ángel Fernández: Conceptualization (equal); funding acquisition (lead); formal analysis (supporting); methodology (supporting); resources (lead); supervision (lead); validation (lead); visualization (supporting); writing – review and editing (lead).

Competing interests

The authors declare that they have no competing interests.

References

Al-ramahi, Y., Nyerges, A., Margolles, Y., Cerdán, L., Ferenc, G., Pál, C., et al. (2021) ssDNA recombineering boosts *in vivo* evolution of nanobodies displayed on bacterial surfaces, *Communications Biology* **4**: 1169.

Álvarez, B., Mencía, M., de Lorenzo, V., and Fernández, L.Á. (2020) *In vivo* diversification of target genomic sites using processive base deaminase fusions blocked by dCas9, *Nature Communications* **11**: 6436.

Amann, E., Ochs, B., and Abel, K.J. (1988) Tightly regulated *tac* promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli*, *Gene* **69**: 301-315.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (2002) *Short Protocols in Molecular Biology*. New York: John Wiley & Sons, Inc.

Bencurova, E., Akash, A., Dobson, R.C.J., and Dandekar, T. (2023) DNA storage—from natural biology to synthetic biology, *Computational and Structural Biotechnology Journal* **21**: 1227-1235.

Biggs, B.W., Bedore, S.R., Arvay, E., Huang, S., Subramanian, H., McIntyre, E.A., et al. (2020) Development of a genetic toolset for the highly engineerable and metabolically versatile Acinetobacter baylyi ADP1, *Nucleic Acids Research* **48**: 5169-5182.

Bodelón, G., Marín, E., and Fernández, L.Á. (2009) Role of periplasmic chaperones and BamA (YaeT/Omp85) in folding and secretion of intimin from enteropathogenic *Escherichia coli* strains, *J Bacteriol* **191**: 5169-5179.

Bornscheuer, U.T., Huisman, G.W., Kazlauskas, R.J., Lutz, S., Moore, J.C., and Robins, K. (2012) Engineering the third wave of biocatalysis, *Nature* **485**: 185-194.
Bradbury, A.R., and Marks, J.D. (2004) Antibodies from phage antibody libraries, *J Immunol Methods* **290**: 29-49.
Cowie, A., Cheng, J., Sibley, C.D., Fong, Y., Zaheer, R., Patten, C.L., et al. (2006) An integrated approach to functional genomics: construction of a novel reporter gene fusion library for Sinorhizobium meliloti, *Appl Environ Microbiol* **72**: 7156-7167.
Criscuolo, A., and Brisse, S. (2013) AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads, *Genomics* **102**: 500-506.
Datsenko, K.A., and Wanner, B.L. (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products, *Proc Natl Acad Sci U S A* **97**: 6640-6645.
de Boer, H.A., Comstock, L.J., and Vasser, M. (1983) The tac promoter: a functional hybrid derived from the trp and lac promoters, *Proc Natl Acad Sci U S A* **80**: 21-25.
De Meyer, T., Muyldermans, S., and Depicker, A. (2014) Nanobody-based products as research and diagnostic tools, *Trends Biotechnol* **32**: 263-270.
del Val, E., Nasser, W., Abaibou, H., and Reverchon, S. (2019) RecA and DNA recombination: a review of molecular mechanisms, *Biochemical Society Transactions* **47**: 1511-1531.
Dufner, P., Jermutus, L., and Minter, R.R. (2006) Harnessing phage and ribosome display for antibody optimisation, *Trends Biotechnol* **24**: 523-529.
Dumon-Seignovert, L., Cariot, G., and Vuillard, L. (2004) The toxicity of recombinant proteins in *Escherichia coli*: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3), *Protein Expr Purif* **37**: 203-206.
Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., 3rd, Burland, V., Mau, B., et al. (2008) The complete genome sequence of *Escherichia coli* DH10B: insights into the biology of a laboratory workhorse, *J Bacteriol* **190**: 2597-2606.
Egger, E., Tauer, C., Cserjan-Puschmann, M., Grabherr, R., and Striedner, G. (2020) Fast and antibiotic free genome integration into *Escherichia coli* chromosome, *Scientific reports* **10**: 16510.
Elmore, J.R., Furches, A., Wolff, G.N., Gorday, K., and Guss, A.M. (2017) Development of a high efficiency integration system and promoter library for rapid modification of Pseudomonas putida KT2440, *Metabolic Engineering Communications* **5**: 1-8.
Fairman, J.W., Dautin, N., Wojtowicz, D., Liu, W., Noinaj, N., Barnard, T.J., et al. (2012) Crystal structures of the outer membrane domain of intimin and invasin from enterohemorrhagic *E. coli* and enteropathogenic Y. pseudotuberculosis, *Structure* **20**: 1233-1243.
Feher, T., Karcagi, I., Gyorfy, Z., Umenhoffer, K., Csorgo, B., and Posfai, G. (2008) Scarless engineering of the *Escherichia coli* genome, *Methods Mol Biol* **416**: 251-259.
Fernández-Martínez, L.T., and Bibb, M.J. (2014) Use of the Meganuclease I-SceI of *Saccharomyces cerevisiae* to select for gene deletions in actinomycetes, *Scientific reports* **4**: 7100.
Fisher, A.K., Freedman, B.G., Bevan, D.R., and Senger, R.S. (2014) A review of metabolic and enzymatic engineering strategies for designing and optimizing performance of microbial cell factories, *Computational and Structural Biotechnology Journal* **11**: 91-99.
Hamers-Casterman, C., Atarhouch, T., Muyldermans, S., Robinson, G., Hamers, C., Songa, E.B., et al. (1993) Naturally occurring antibodies devoid of light chains, *Nature* **363**: 446-448.
Hao, M., Qiao, H., Gao, Y., Wang, Z., Qiao, X., Chen, X., and Qi, H. (2020) A mixed culture of bacterial cells enables an economic DNA storage on a large scale, *Commun Biol* **3**: 416.
Hashimoto-Gotoh, T., and Sekiguchi, M. (1977) Mutations of temperature sensitivity in R plasmid pSC101, *J Bacteriol* **131**: 405-412.
Herring, C.D., Glasner, J.D., and Blattner, F.R. (2003) Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli*, *Gene* **311**: 153-163.
Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics* **26**: 680-682.
Intasian, P., Prakinee, K., Phintha, A., Trisrivirat, D., Weeranoppanant, N., Wongnate, T., and Chaiyen, P. (2021) Enzymes, *In vivo* Biocatalysis, and Metabolic Engineering for Enabling a Circular Economy and Sustainability, *Chemical Reviews* **121**: 10367-10451.

Jovčevska, I., and Muyldermans, S. (2019) The Therapeutic Potential of Nanobodies, *BioDrugs* **34**: 11-26.
Kuhlman, T.E., and Cox, E.C. (2010) Site-specific chromosomal integration of large synthetic constructs, *Nucleic Acids Res* **38**: e92.
Landy, A. (2015) The λ Integrase Site-specific Recombination Pathway, *Microbiology spectrum* **3**: Mdna3-0051-2014.
Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D., and Davey, R.P. (2013) Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics, *Front Genet* **4**: 288.
Li, L., Liu, X., Wei, K., Lu, Y., and Jiang, W. (2019) Synthetic biology approaches for chromosomal integration of genes and pathways in industrial microbial systems, *Biotechnol Adv* **37**: 730-745.
Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* **22**: 1658-1659.
Li, X.T., Thomason, L.C., Sawitzke, J.A., Costantino, N., and Court, D.L. (2013) Positive and negative selection using the tetA-sacB cassette: recombineering and P1 transduction in *Escherichia coli*, *Nucleic Acids Res* **41**: e204.
Lilly, J., and Camps, M. (2015) Mechanisms of Theta Plasmid Replication, *Microbiology spectrum* **3**: Plas-0029-2014.
Löfblom, J. (2011) Bacterial display in combinatorial protein engineering, *Biotechnol J* **6**: 1115-1129.
Magalhaes, M.L., and Blanchard, J.S. (2005) The kinetic mechanism of AAC3-IV aminoglycoside acetyltransferase from *Escherichia coli*, *Biochemistry* **44**: 16275-16283.
Martínez-García, E., and de Lorenzo, V. (2011) Engineering multiple genomic deletions in Gram-negative bacteria: analysis of the multi-resistant antibiotic profile of Pseudomonas putida KT2440, *Environ Microbiol* **13**: 2702-2716.
Miller, J.H. (1992) *A short course in bacterial genetics: a laboratory manual and handbook for Escherichia coli and related bacteria*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.
Murphy Kenan, C. (2016) λ Recombination and Recombineering, *EcoSal Plus* **7**: 10.1128/ecosalplus.ESP-0011-2015.
Muyldermans, S. (2013) Nanobodies: natural single-domain antibodies, *Annu Rev Biochem* **82**: 775-797.
Muyldermans, S., Cambillau, C., and Wyns, L. (2001) Recognition of antigens by single-domain antibody fragments: the superfluous luxury of paired domains, *Trends Biochem Sci* **26**: 230-235.
Ngara, T.R., and Zhang, H. (2018) Recent Advances in Function-based Metagenomic Screening, *Genomics, Proteomics & Bioinformatics* **16**: 405-415.
Ou, B., Garcia, C., Wang, Y., Zhang, W., and Zhu, G. (2018) Techniques for chromosomal integration and expression optimization in *Escherichia coli*, *Biotechnol Bioeng* **115**: 2467-2478.
Parisutham, V., Chhabra, S., Ali, M.Z., and Brewster, R.C. (2022) Tunable transcription factor library for robust quantification of regulatory properties in *Escherichia coli*, *Molecular Systems Biology* **18**: e10843.
Patel, R.K., and Jain, M. (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data, *PLOS ONE* **7**: e30619.
Peters, J.E. (2019) Targeted transposition with Tn7 elements: safe sites, mobile plasmids, CRISPR/Cas and beyond, *Mol Microbiol* **112**: 1635-1644.
Pines, G., Freed, E.F., Winkler, J.D., and Gill, R.T. (2015) Bacterial Recombineering: Genome Engineering via Phage-Based Homologous Recombination, *ACS synthetic biology* **4**: 1176-1185.
Piñero-Lambea, C., Bodelón, G., Fernández-Periañez, R., Cuesta, A.M., Álvarez-Vallina, L., and Fernández, L.Á. (2015) Programming controlled adhesion of *E. coli* to target surfaces, cells, and tumors with synthetic adhesins, *ACS synthetic biology* **4**: 463-473.
Piñero-Lambea, C., Garcia-Ramallo, E., Miravet-Verde, S., Burgos, R., Scarpa, M., Serrano, L., and Lluch-Senar, M. (2022) SURE editing: combining oligo-recombineering and programmable insertion/deletion of selection markers to efficiently edit the Mycoplasma pneumoniae genome, *Nucleic Acids Res* **50**: e127.

Posfai, G., Kolisnychenko, V., Bereczki, Z., and Blattner, F.R. (1999) Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome, *Nucleic Acids Res* **27**: 4409-4415.
Rees, A.R. (2020) Understanding the human antibody repertoire, *mAbs* **12**: 1729683-1729683.
Rosano, G.L., and Ceccarelli, E.A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges, *Front Microbiol* **5**: 172.
Ruano-Gallego, D., Álvarez, B., and Fernández, L.A. (2015) Engineering the Controlled Assembly of Filamentous Injectisomes in *E. coli* K-12 for Protein Translocation into Mammalian Cells, *ACS synthetic biology* **4**: 1030-1041.
Salema, V., and Fernández, L.Á. (2017) *Escherichia coli* surface display for the selection of nanobodies, *Microb Biotechnol* **10**: 1468-1484.
Salema, V., Mañas, C., Cerdán, L., Piñero-Lambea, C., Marín, E., Roovers, R.C., et al. (2016) High affinity nanobodies against human epidermal growth factor receptor selected on cells by *E. coli* display, *MAbs* **8**: 1286–1301.
Salema, V., Marín, E., Martínez-Arteaga, R., Ruano-Gallego, D., Fraile, S., Margolles, Y., et al. (2013) Selection of single domain antibodies from immune libraries displayed on the surface of *E. coli* cells with two β-domains of opposite topologies, *PLoS ONE* **8**: e75126.
Saleski, T.E., Chung, M.T., Carruthers, D.N., Khasbaatar, A., Kurabayashi, K., and Lin, X.N. (2021) Optimized gene expression from bacterial chromosome by high-throughput integration and screening, *Sci Adv* **7**.
Scholz, S.A., Diao, R., Wolfe, M.B., Fivenson, E.M., Lin, X.N., and Freddolino, P.L. (2019) High-Resolution Mapping of the *Escherichia coli* Chromosome Reveals Positions of High and Low Transcription, *Cell Syst* **8**: 212-225 e219.
Simon, A.J., d'Oelsnitz, S., and Ellington, A.D. (2019) Synthetic evolution, *Nat Biotechnol* **37**: 730-743.
St-Pierre, F., Cui, L., Priest, D.G., Endy, D., Dodd, I.B., and Shearwin, K.E. (2013) One-Step Cloning and Chromosomal Integration of DNA, *ACS synthetic biology* **2**: 537-541.
Tellechea-Luzardo, J., Hobbs, L., Velázquez, E., Pelechova, L., Woods, S., de Lorenzo, V., and Krasnogor, N. (2022) Versioning biological cells for trustworthy cell engineering, *Nature Communications* **13**: 765.
Tiller, K.E., Chowdhury, R., Li, T., Ludwig, S.D., Sen, S., Maranas, C.D., and Tessier, P.M. (2017) Facile Affinity Maturation of Antibody Variable Domains Using Natural Diversity Mutagenesis, *Front Immunol* **8**: 986.
van der Woude, M.W., and Henderson, I.R. (2008) Regulation and function of Ag43 (flu), *Annu Rev Microbiol* **62**: 153-169.
Van Zyl, W.F., Dicks, L.M.T., and Deane, S.M. (2019) Development of a novel selection/counter-selection system for chromosomal gene integrations and deletions in lactic acid bacteria, *BMC Mol Biol* **20**: 10.
Vo, P.L.H., Ronda, C., Klompe, S.E., Chen, E.E., Acree, C., Wang, H.H., and Sternberg, S.H. (2020) CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering, *Nat Biotechnol* **39**: 480-489.
Wang, H., La Russa, M., and Qi, L.S. (2016) CRISPR/Cas9 in Genome Editing and Beyond, *Annual Review of Biochemistry* **85**: 227-264.
Wang, J.Y., Pausch, P., and Doudna, J.A. (2022) Structural biology of CRISPR–Cas immunity and genome editing enzymes, *Nature Reviews Microbiology* **20**: 641-656.
Wang, T., Wang, D., Lyu, Y., Feng, E., Zhu, L., Liu, C., et al. (2018) Construction of a high-efficiency cloning system using the Golden Gate method and I-SceI endonuclease for targeted gene replacement in Bacillus anthracis, *Journal of Biotechnology* **271**: 8-16.
Wilson, C.J., Zhan, H., Swint-Kruse, L., and Matthews, K.S. (2007) The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding, *Cell Mol Life Sci* **64**: 3-16.
Yang, E.Y., and Shah, K. (2020) Nanobodies: Next Generation of Cancer Diagnostics and Therapeutics, *Frontiers in oncology* **10**: 1182.
Zeymer, C., and Hilvert, D. (2018) Directed Evolution of Protein Catalysts, *Annu Rev Biochem* **87**: 131-157.

Zhang, N., Shao, L., Jiang, Y., Gu, Y., Li, Q., Liu, J., et al. (2015) I-SceI-mediated scarless gene modification via allelic exchange in Clostridium, *Journal of Microbiological Methods* **108**: 49-60.
Zucca, S., Pasotti, L., Politi, N., Cusella De Angelis, M.G., and Magni, P. (2013) A standard vector for the chromosomal integration and characterization of BioBrick™ parts in *Escherichia coli*, *Journal of biological engineering* **7**: 12.

Tables

Table 1. Diversity of the plasmid and integrated libraries

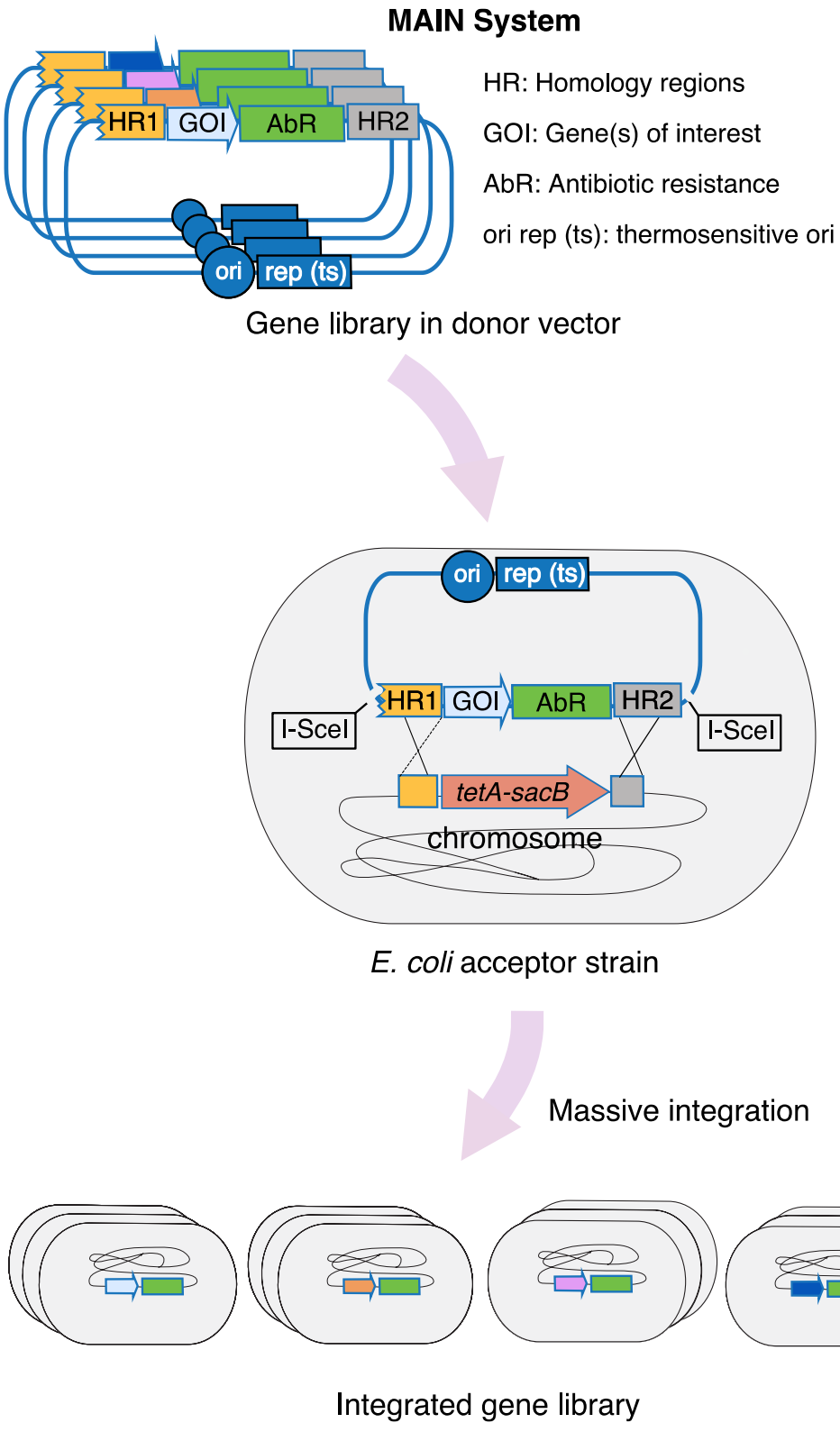
	Total reads	HQ reads	No. of clusters*	Diversity value**(%)
pRecomb-TS-V_{HH} library	697458	518361	50813	9.80%
Integrated library	495202	377690	26989	7.15%

* A cluster is defined as a group of sequences having ≥ 98% of identity.
**Diversity value (%) = (No. of clusters / High-quality reads) x 100.

Table 2. Nanobodies binding EGFR from the integrated *E. coli* library

Nb	Frequency	CDR3
Nb1-EGFR	167/195	DKWSSSRRSVDYD
Nb2-EGFR	10/195	STTWGRPSYVYR
Nb3-EGFR	6/195	STYSRDTIFTKWANYN
Nb4-EGFR	3/195	DKWASSTRSIDYD
Nb5-EGFR	2/195	SRIIYSYVNYVNPGEYD
Nb6-EGFR	1/195	STYSRDTIFTNRANYN
Nb7-EGFR	1/195	DRRSTTDLKTLRAD

Graphical Abstract



Figures Legends

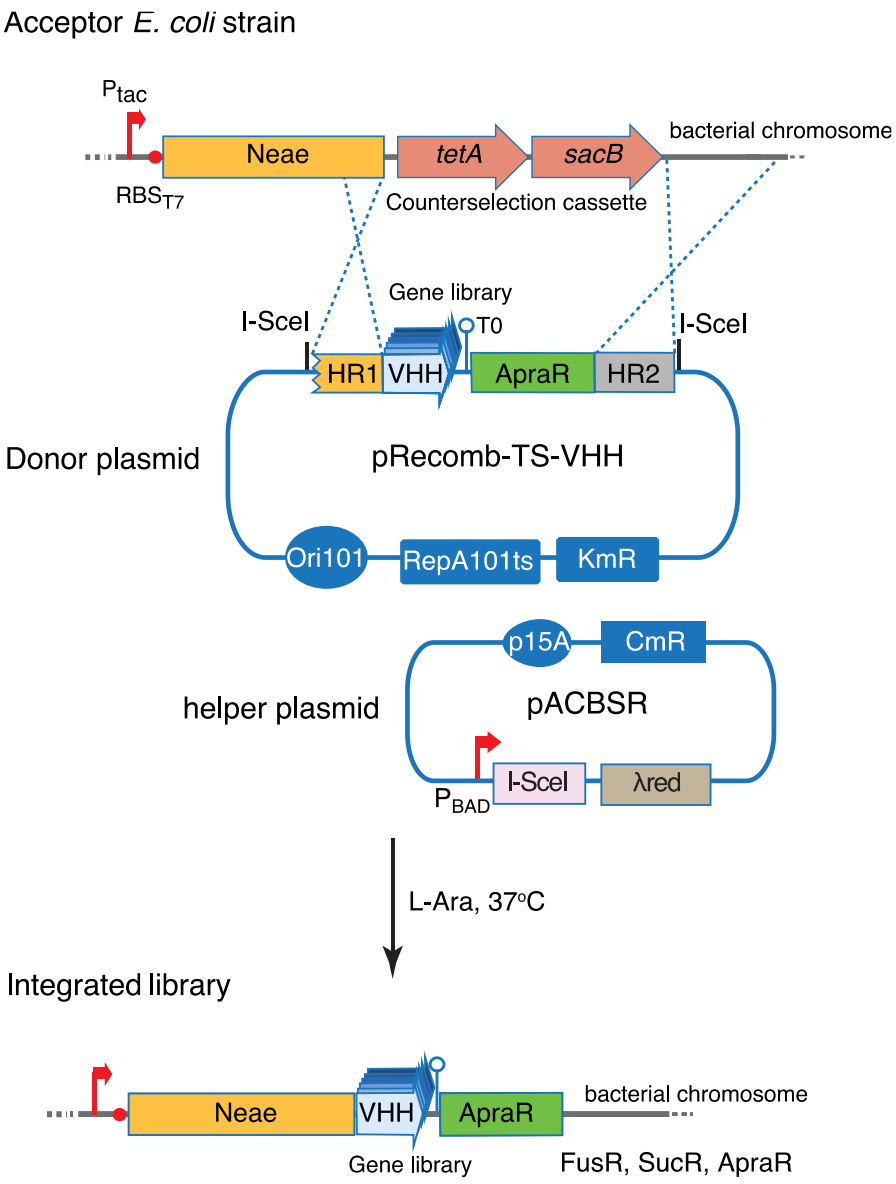


Figure 1. Schematic representation of the genetic elements comprising the MAIN system for integration of V_{HH} gene libraries. The acceptor *E. coli* strain (EcM1-NL) contains the Neae gene construct and the counterselection cassette *tetA-sacB* inserted in the chromosomal *flu* gene. The thermosensitive donor vector (pRecomb-TS) carries the cloned V_{HH} gene library and a downstream apramycin resistance marker ($Apra^R$) flanked by homology regions (HR1 and HR2) and I-SceI restriction sites. The inducible expression of I-SceI enzyme and the λ-Red products with L-arabinose (L-Ara) from a helper plasmid (pACBSR) mediates the *in vivo* digestion of the donor plasmid releasing a linear DNA fragment (HR1-V_{HH}-$Apra^R$-HR2) that can be integrated in the chromosome of the acceptor strain by a double homologous recombination event. A temperature shift from 30 to 37 ºC hinders further replication of undigested donor plasmids. The resulting bacteria carry the Neae-V_{HH} in-frame fusion integrated in single-copy at the *flu* site of *E. coli* chromosome. The following regulatory elements are indicated: *tac* promoter (P_{tac}) and T7 ribosome binding site (RBS_{T7}) in the acceptor strain, the transcriptional terminator T0 (T0) and the thermosensitive replication (ori101, *repA101ts*) in the donor vector and the arabinose promoter (P_{BAD}) in the helper plasmid. The resistance (R) of the bacterial strains and plasmids to fusaric acid (Fus), sucrose (Suc), kanamycin (Km), apramycin (Apra), and chloramphenicol (Cm) are indicated.

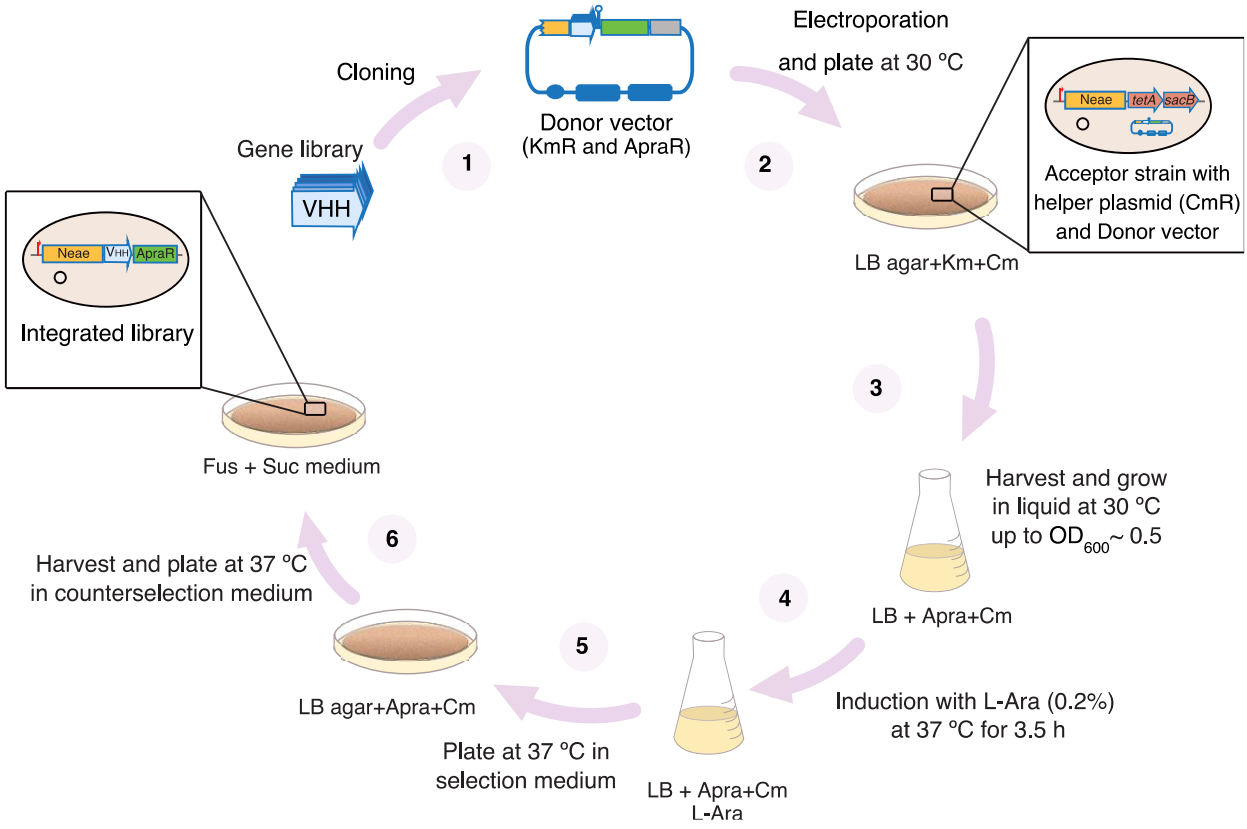


Figure 2. Scheme of the integration process for the V_{HH} gene library using MAIN system. A V_{HH} gene library was built in pRecomb-TS using *E. coli* cloning strain DH10BT1R at 30 ºC (step 1). The pRecomb-TS-V_{HH} plasmid library was electroporated in the acceptor strain EcM1-NL carrying the pACBSR helper plasmid (step 2). Transformants with pRecomb-TS-V_{HH} library and pACBSR plasmid were harvested from plates and grown in liquid LB medium supplemented with Apra and Cm at 30 ºC up to exponential phase (step 3). The I-SceI meganuclease and λRed recombination system were induced from pACBSR with L-Ara 0.2% (w/v) and the temperature was increased to 37ºC to hinder pRecomb-TS replication (step 4). After 3.5 h of induction bacteria were plated as a lawn in LB agar supplemented with Apra and Cm for a first selection step of bacteria containing the GOI-Apra^R cassette integrated into the chromosome (step 5). Bacteria from the first selection step were harvested and directly plated in Fus+Suc counterselection medium to remove non-integrant bacteria carrying *the tetA-sacB* cassette (step 6).

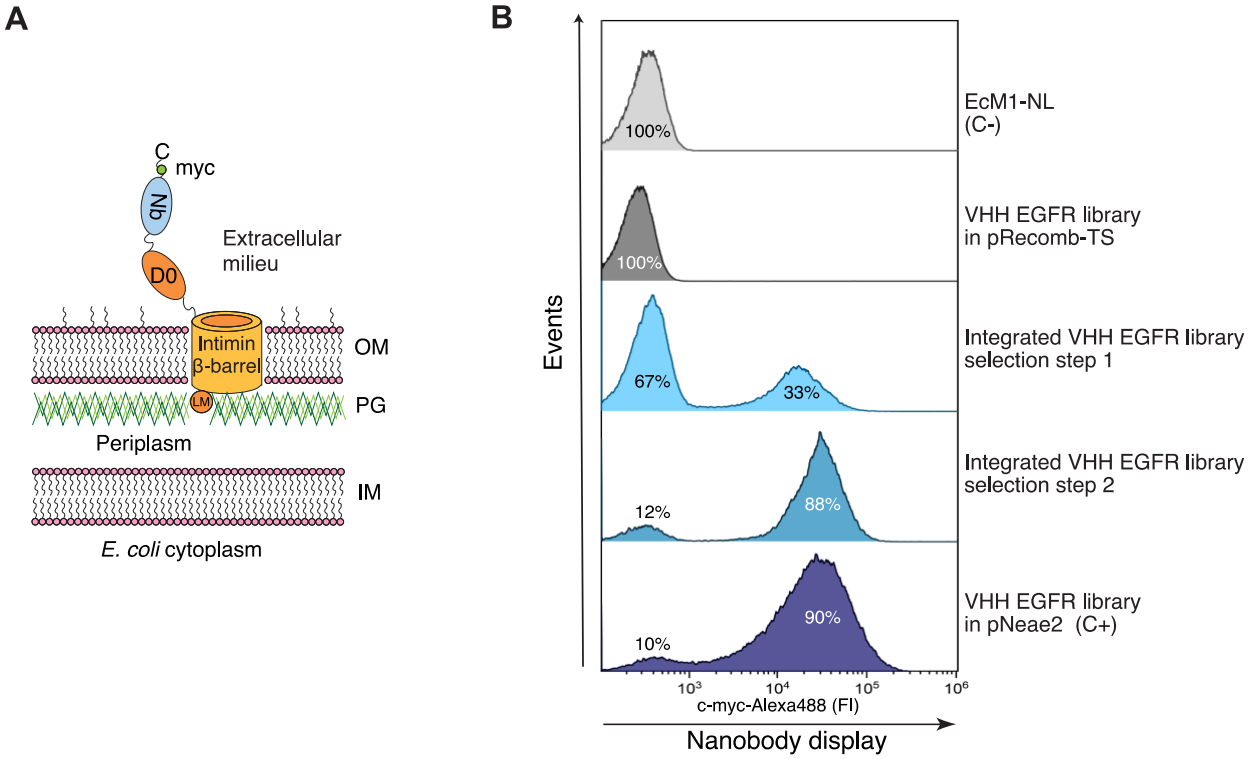


Figure 3. Nb display levels of the integrated library. (A) Scheme representing the display of Nbs in the outer membrane (OM), indicating the periplasmic LysM domain (LM) binding the peptidoglycan (PG), the β-barrel domain anchored in the OM, and the extracellular Ig-like (D0) and Nb domains with C-terminal myc-tag (myc). **(B)** Flow cytometry analysis of Nb display levels in the indicated bacterial populations stained with anti-myc mAb and secondary anti-mouse IgG-Alexa488 conjugate. Bacteria analyzed from top to bottom: EcM1-NL strain as a negative control (C-); EcM1-NL with pRecomb-TS-V_{HH} EGFR library grown at 30ºC; integrated EcM1-NL-V_{HH} EGFR library grown at 37 ºC after the first selection step in Apra+Cm LB agar, and after the second selection step in Fus + Suc medium; EcM1 with pNeae2-V_{HH} EGFR library grown at 30 ºC and induced with IPTG, as positive control (Salema, et al., 2016).

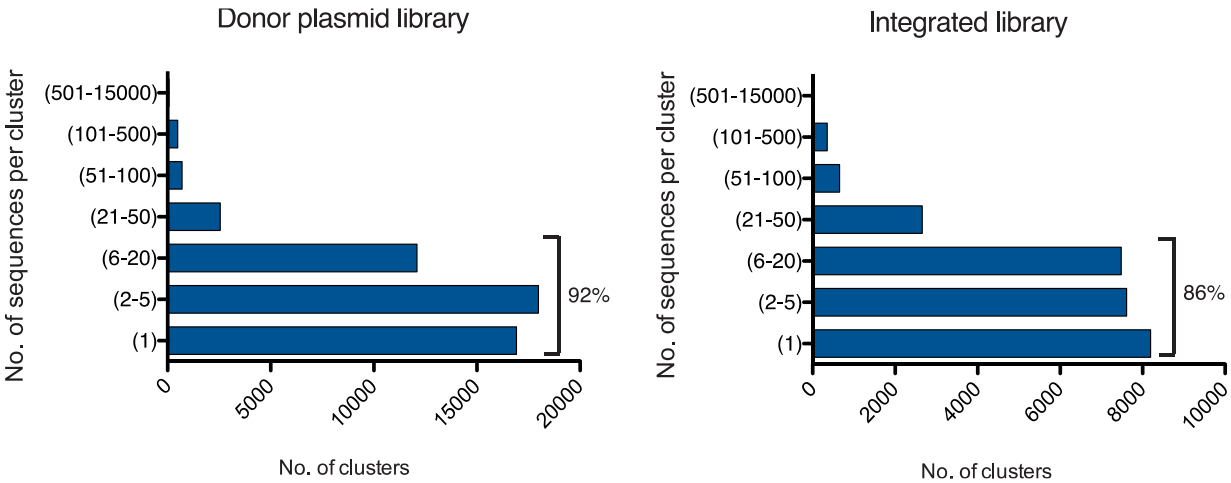


Figure 4. Cluster distribution of V_{HH} sequences in the library before and after integration. Graphs represent the cluster size distribution. Data of the V_{HH} library in plasmid (left) and integrated (right) are shown. The percentage of clusters with ≤ 20 members is indicated in each graph.

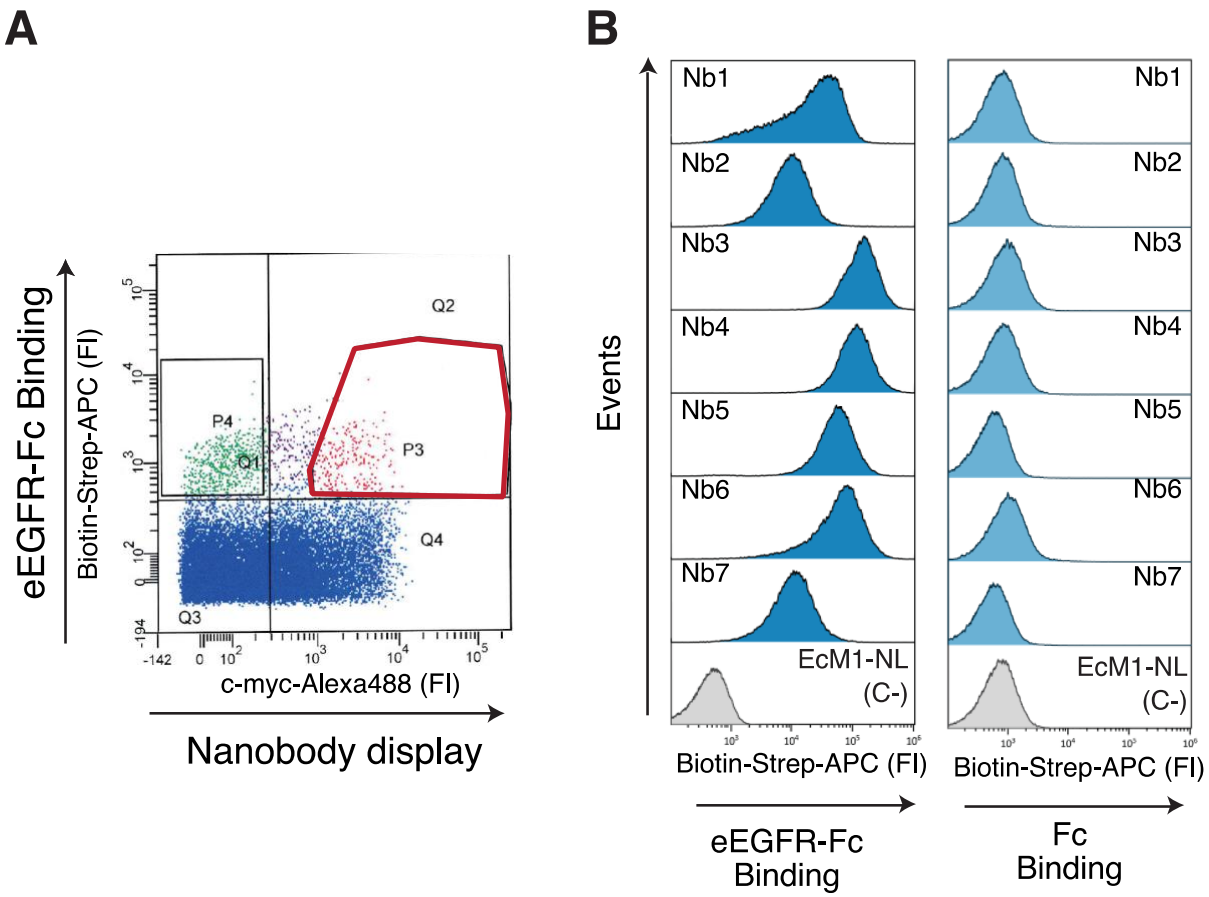


Figure 5. Selection of Nbs binding EGFR from the integrated V_{HH} library. (A) Fluorescence activated cell sorting (FACS) for selection of bacteria displaying Nbs binding eEGFR-Fc antigen. Histogram shows fluorescence intensity (FI) signals of bacteria from the integrated V_{HH} EGFR library enriched by MACS and incubated with biotinylated eEGFR-Fc (100 nM) and anti-c-myc mAb, followed by Streptavidin-APC and anti-mouse IgG-Alexa 488 as secondary reagents. The double-labeled sorted population is indicated as P3. **(B)** Flow cytometry analysis of *E. coli* clones displaying anti-EGFR Nbs (Nb1 to Nb7) identified from the integrated library. Binding analyses of bacterial clones against the target antigen (eEGFR-Fc) and a negative control antigen (human Fc) were performed to assess the specificity. The parental strain EcM1-NL was used as a negative control (C-). Nb display and antigen binding were stained as in A but concentration of biotinylated antigen (either eEGFR-Fc or Fc) was reduced to 50 nM.