PNAS NEXUS

PNAS Nexus, 2025, 4, pgaf006
https://doi.org/10.1093/pnasnexus/pgaf006
Advance access publication 9 January 2025
Research Report

Targeted barcoding of variable antibody domains and individual transcriptomes of the human B-cell repertoire using Link-Seq

Hongxing Hu[a,b,*,1], Fan Zhou[c,1], Xiaoli Ma[a], Karl Albert Brokstad[c,d], Leonie Kolmar[a], Charles Girardot[b], Vladimir Benes[b], Rebecca J. Cox[c,e,*] and Christoph A. Merten[a,*]

[a]Institute of Bioengineering, School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[b]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, 69117 Germany
[c]Department of Clinical Sciences, Influenza Centre, University of Bergen, Bergen, N5021, Norway
[d]Department of Safety, Chemistry and Biomedical Laboratory Sciences, Western Norway University of Applied Sciences (HVL), Bergen, N5020, Norway
[e]Department of Microbiology, Haukeland University Hospital, Bergen, N5021, Norway
*To whom correspondence should be addressed: Email: hongxing.hu@epfl.ch (H.H.); Email: rebecca.cox@uib.no (R.J.C.); Email: christoph.merten@epfl.ch (C.A.M.)
[1]H.H. and F.Z. contributed equally to this work.
Edited By Hidde Ploegh

Abstract

Here, we present Link-Seq, a highly efficient droplet microfluidic method for combined sequencing of antibody-encoding genes and the transcriptome of individual B cells at large scale. The method is based on 3′ barcoding of the transcriptome and subsequent single-molecule PCR in droplets, which freely shift the barcode along specific gene regions, such as the antibody heavy- and light-chain genes. Using the immune repertoire of COVID-19 patients and healthy donors as a model system, we obtain up to 91.7% correctly paired immunoglobulin heavy and light chains. Furthermore, we map the V(D)J usage and obtain sensitivities comparable with the current gold-standard 10× Genomics commercial systems while offering full flexibility in experimental setup and significant cost savings. A further unique feature of Link-Seq is the possibility of barcoding multiple target genes in a site-specific manner. Based on the open character of the platform and its conceptual advantages, we expect Link-Seq to become a versatile tool for single-cell analysis, especially for applications requiring additional processing steps that cannot be implemented on commercially available platforms.

Keywords: immune repertoire sequencing, single-cell analysis, droplet microfluidics, SARS-CoV-2 infection, RNA-seq

Significance Statement

As SARS-CoV-2 continues to circulate, the stochastic emergence of novel variants will pose a threat for the foreseeable future. Minimizing their impact requires identifying and containing each variant at source to halt onward circulation. Our results highlight that, due to the lag between exposure and death, identifying a novel variant using mortality data is unlikely before substantial spatial spread. On the other hand, our results suggest that by the start of 2022, multiple countries were sequencing sufficient infections to make successfully identifying and containing variants via genomic surveillance realistic, contingent on fast sequence processing, dissemination, and analysis times. As a consequence, while genomic surveillance makes early variant identification possible, it is not feasible to precisely quantify the threat posed before a proportionate containment response is needed.

Introduction

High-throughput single-cell RNA-sequence (scRNA-seq) technologies, such as Drop-seq, inDrop, Seq-Well, and SPLiT-seq, have revolutionized the way we study biology (1–4). The COVID-19 pandemic reinforced the importance of combining single B-cell immune repertoire sequencing with single-cell transcriptome sequencing to study adaptive immunity after SARS-CoV-2 infection (5–11). However, based on the limited read length of common sequencing methods such as the Illumina platform, barcoding the 3′ end of the transcript is not compatible with sequencing the V(D)J gene at the 5′ end and hence limits immune repertoire analyses.

Previously, two B-cell receptor (BCR) gene-specific sequencing approaches for single B-cell immune repertoires have been established: DART-Seq (12) and RAGE-Seq (13), which, however, both suffer from low pairing efficiency, ranging from 15 to 42.6%. In

Competing Interest: The authors declare no competing interests.
Received: June 18, 2024. **Accepted:** December 18, 2024

addition, DART-Seq can only provide sequences of the CDR3 region. In contrast, RAGE-Seq makes use of Nanopore full-length BCR gene sequencing, but it shows a lower accuracy compared with Illumina sequencing platforms. A variation using 5′ end barcoding developed by 10× Genomics enables the V(D)J sequences and single-cell transcriptome to be measured simultaneously. However, this workflow is fully locked and cannot be modified for the integration of further steps, such as the labeling of newly synthesized mRNA, which has previously been shown to significantly improve single immune cell clustering approaches (14). Hence, a flexible, open-access platform for combined immune repertoire and single-cell transcriptome sequencing with high sensitivity and accuracy is still missing.

Here, we developed Link-Seq, a single-cell immune repertoire and mRNA sequencing method based on 3′-end barcoding technology and subsequent droplet-based fusion PCR. The latter step allows the barcode to move next to any sequence of interest, which we used to generate short, barcoded V(D)J amplicons for Illumina sequencing. By performing the fusion PCR on individually encapsulated DNA template molecules, we drastically reduce amplification bias and sequence contamination. We further developed two A-tailing strategies that can prevent sequencing contaminations based on intermolecular hybridization and overlapping extension after breaking the emulsion. Therefore, Link-Seq can detect extremely rare V and J genes and improve the efficiency of retrieving paired BCR genes. Of note, Link-Seq can also be applied directly to already existing cDNA that has been prepared with any other 3′-end barcoding methods for retrieving valuable BCR repertoire information after designing specific primers for the in-droplet fusion PCR (for details see Table S1).

Results

Link-Seq workflow

Link-Seq is based on 3′ barcoding of the transcriptome and subsequent single-molecule PCR in droplets, enabling the barcode to freely shift along specific gene regions, such as the BCR V(D)J regions (Fig. 1). In a first step, a modified Drop-seq workflow was used to barcode single-cell mRNA. After cDNA amplification, part of the cDNA was used for short-read Illumina sequencing to generate single-cell gene expression profiles. In parallel, an aliquot of the barcoded cDNA was encapsulated into droplets at single-molecular concentration, together with PCR mix, using a microfluidic chip (Fig. 1a). Droplets with a diameter between 40 and 55 μm were generated at a rate of 35 million per hour (Fig. S1a). Within the droplets, the 5′-end BCR V(D)J sequence and the 3′-end-cell-barcode-unique molecular identifier (CB-UMI) sequence from the same cDNA template were amplified and fused using overlapping primers to form a shortened recombinant fragment with a length of about 600 base pairs (bp) for light chains and a length of about 670 bp for heavy chains (Fig. S1c and d). A major challenge in linking the 3′-end CB-UMI sequence to its own 5′-end BCR V(D)J sequence was to avoid intermolecular hybridization during overlapping extension and downstream nested PCR amplifications.

To avoid intermolecular hybridization during the overlapping extension step, we encapsulated the BCR transcripts at a lambda (average number of transcripts per droplet volume) of less than one on average to ensure only single BCR transcripts per droplet. To prevent remaining nonfused fragments from hybridizing with those from other droplets after breaking the emulsion, we further developed two A-tailing strategies: Moloney Murine Leukemia Virus (MMLV) Reverse Transcriptase-based A-tailing and lambda exonuclease/terminal deoxynucleotidyl transferase (L/T)-based A-tailing. MMLV Reverse Transcriptase can add poly-A to blunt-ended DNA fragments (15). As a result, the treated, nonfused fragments can no longer anneal to any other DNA molecule in the pooled sample, hence robustly preventing intermolecular hybridization of fragments that initially were not in the same droplet. A similar result can be achieved using terminal deoxynucleotidyl transferase (TDT), which can add poly-A to the 3′ end of single-stranded DNA, while being considerably less efficient for double-stranded DNA (16). Therefore, we introduced lambda exonuclease treatment prior to TDT treatment for converting double-stranded DNA into single-stranded DNA. Lambda exonuclease cleavage requires 5′ phosphorylation, which was introduced in the primers used for amplification of the individual fragments. However, after overlap extension PCR, this modification is no longer present in the fused fragments, for which reason only remaining nonfused fragments were affected by the L/T treatment (Fig. 1b).

Optimization of A-tailing and template concentration during BCR library preparation

To test these two A-tailing methods, we amplified and purified the light-chain VJ fragments and CB-UMI fragments and applied them to overlap extension PCR. The MMLV-based A-tailing using the supplier's commercial reaction buffer greatly reduced the overlapping extension, while using a homemade buffer (15) was less efficient. The L/T-based A-tailing eliminates the overlapping extension to an extent that no corresponding band was visible on the gel anymore (Fig. S1b). For a more systematic assessment of the random pairing rate, we also generated two synthetic fragments, each harboring a different barcode and an IgG domain of known sequence, facilitating molecular assays for chimera generation. qPCR analysis revealed a roughly 60% reduced occurrence of chimeras (Fig. S2).

We further compared MMLV and L/T treatments in eliminating intermolecular hybridization in BCR repertoire sequencing. We performed single-cell barcoding of B cells isolated from a healthy donor (HD). cDNA was amplified from an estimated 800–1,000 single-cell transcriptomes attached to microparticles (STAMPs). Overlapping extension PCR was performed by encapsulating cDNA in droplets at a concentration of 10 pg/mL ($\lambda = 0.5$) or 1 pg/mL ($\lambda = 0.05$), respectively. Emulsion PCR products were subsequently processed with MMLV-based A-tailing or L/T-based A-tailing (Fig. 2a). After two rounds of nested PCR, the barcoded BCR libraries were combined and sequenced on the MiSeq platform. A-tailing treatments introduced after the emulsion PCR efficiently prevented intermolecular overlap extension, thereby reducing artifactual inflation of the BCR diversity. For the L/T-treated samples, the top 10 most abundant clones cumulatively represented a larger fraction of the entire immune repertoire when compared with the MMLV RT-treated samples and the control groups, indicating that L/T-based A-tailing prevented intermolecular hybridization most efficiently (Fig. 2b). Due to the prevention of intermolecular hybridization using MMLV-based A-tailing or L/T-based A-tailing methods, the two BCR immune repertoires in 10 pg/mL groups are more similar to each other (0.912) than to the untreated control BCR immune repertoire (0.902/0.903). This pattern was, however, only observed for the 10 pg/mL ($\lambda = 0.5$) group, while not being visible in the 1 pg/mL group. The 10-fold reduced amount of cDNA probably resulted in insufficient sampling of the BCR diversity, so that sampling variation overwhelmed the differences based on the A-tailing treatments (Fig. 2c). A-tailing treatments and less input cDNA (1 pg/mL) reduced the

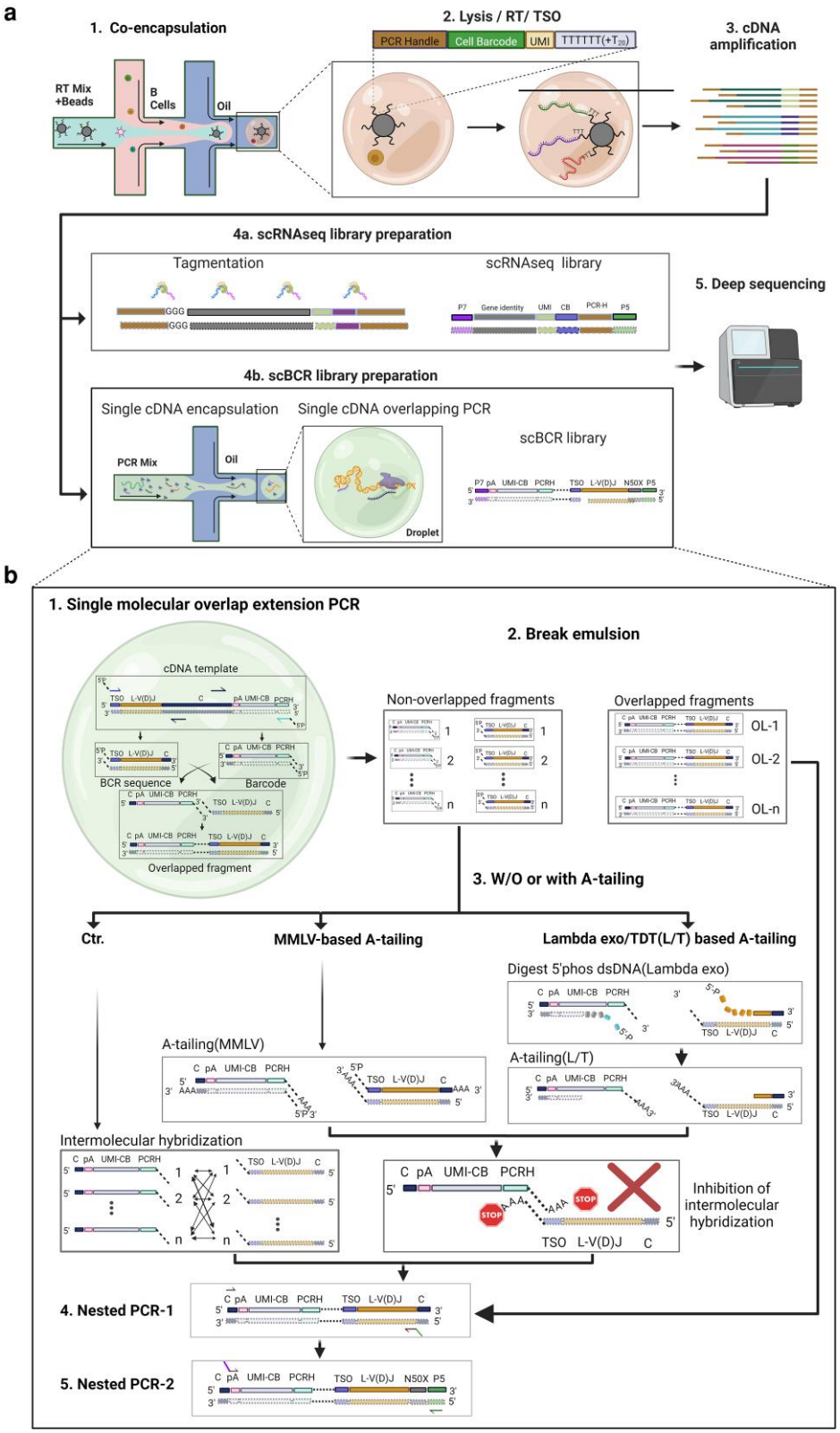

Fig. 1. Link-Seq workflow. a) (1) Single B cells were co-encapsulated with barcoded beads and RT-PCR mix. (2) Within the droplets, cell lysis, reverse transcription, and template switching were performed to generate barcoded cDNA. (3) Subsequently, cDNA was amplified, split, and simultaneously subjected to (4a) tagmentation for scRNA-seq library preparation and (4b) single B-cell BCR (scBCR) library preparation. The scRNA-seq library was sequenced using the Illumina NextSeq platform, while the scBCR library was sequenced using the Illumina MiSeq platform. b) scBCR library preparation workflow and prevention of undesired intermolecular hybridization. (1) Single-molecule overlap extension PCR was performed in droplets. (2) After breaking the emulsion, (3) Lambda exonuclease/terminal deoxynucleotidyl transferase (L/T)-based A-tailing or MMLV-based A-tailing was performed to prevent intermolecular overlap extension. (4–5) Subsequently, two rounds of nested PCR were performed to further enrich the BCR genes and to introduce the Illumina sequencing adapters. pA, poly-A tail; C, constant region; TSO, template-switching oligo; UMI-CB, unique molecular identifier-cell-barcode; PCRH, PCR handle.

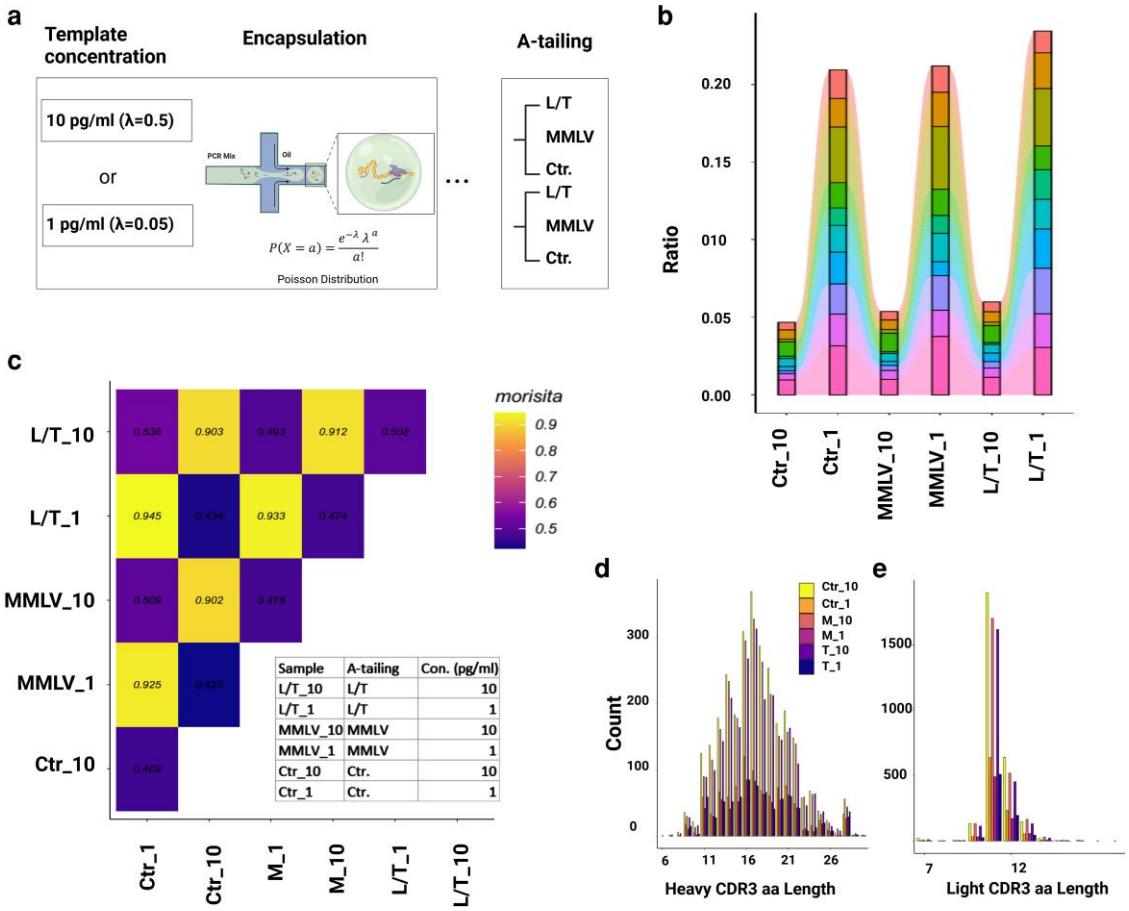

Fig. 2. Optimization of conditions for the preparation of BCR libraries. a) Overview of the experimental workflow for comparing two A-tailing methods. cDNA from 800 to 1,000 STAMPs was encapsulated at 10 or 1 pg/mL for overlap extension PCR in droplets. After breaking the emulsion, PCR products were treated with L/T or MMLV prior to nested PCR. An untreated group was used as a control for each template concentration. The final BCR libraries were sequenced, and BCR repertoires were analyzed and compared. b) Frequency of the 10 most abundant clones in each repertoire. c) Morista similarity between different BCR repertoires. d) Density graph showing the heavy-chain CDR3 amino acid length distribution and e) the light-chain CDR3 amino acid length distribution. L/T_10 (or 1): cDNA was encapsulated at 10 (or 1) pg/mL for overlap extension PCR in droplets, and the emulsion PCR products were treated with L/T before nested PCR. MMLV_10 (or 1): cDNA was encapsulated at 10 (or 1) pg/mL for overlap extension PCR in droplets, and the emulsion PCR products were treated with MMLV before nested PCR. Ctr_10 (or 1): cDNA was encapsulated at 10 (or 1) pg/mL for overlap extension PCR in droplets, and the emulsion PCR products were not treated before nested PCR.

number of clones detected at a given CDR3 length but barely changed the frequency of CDR3 amino acid length distribution in the repertoire (Fig. 2d and e). Taken together, the combination of encapsulation of cDNA at λ of 0.5 and L/T-based A-tailing provided the best performance in terms of good coverage and prevention of intermolecular hybridization. Therefore, this combination was used for the final Link-Seq protocol.

It should be noted that in theory other methods for preventing intermolecular hybridization could be used as well. For example, suppression PCR has been used very successfully for the prevention of chimera formation of T-cell receptors and could be applicable here as well (17). Furthermore, gel purification of the fused band prior to nested PCR (after breaking the emulsion) could be performed. However, this method is not suitable for samples with very limited quantities due to reduced recovery rates, as demonstrated here (Fig. S3).

Studying the B-cell immune repertoire of COVID-19 patients and HDs with Link-Seq

After establishing Link-Seq, we set out to validate our method in a real-world case study. To do so, we analyzed the B-cell immune repertoires of four hospitalized, COVID-19 patients and two HDs. In particular, we were wondering (i) whether the transcriptomic data (scRNA-seq library) would allow reliable identification of different B-cell subtypes, (ii) whether the scBCR library would allow recovery of full-length antibody sequences, and (iii) whether the method would allow tracking of clonal expansion over time using longitudinal samples. Peripheral blood mononuclear cells (PBMCs) were isolated from four hospitalized moderately ill COVID-19 patients and HDs and cryopreserved at −150 °C until use (Fig. 3a and Table S5). Using a negative selection method, B cells were isolated from cryopreserved PBMCs (Fig. 3b). Among the four COVID-19 patients, patient 36 was particularly interesting as PBMCs were collected at three different time points, including 4 days postsymptom onset (DPSO), 45 DPSO, and 169 DPSO. We therefore selected this patient for generating longitudinal data on BCR repertoires and single-cell transcriptomes using Link-Seq.

Identification of B-cell subtypes and changes in their composition during acute SARS-CoV-2 infection

The resulting scRNA-seq data were applied to SingleR package for predicting the exact B-cell types based on gene expression profiles (18) (Fig. 4a). Detailed analysis of the relative cell-type abundance

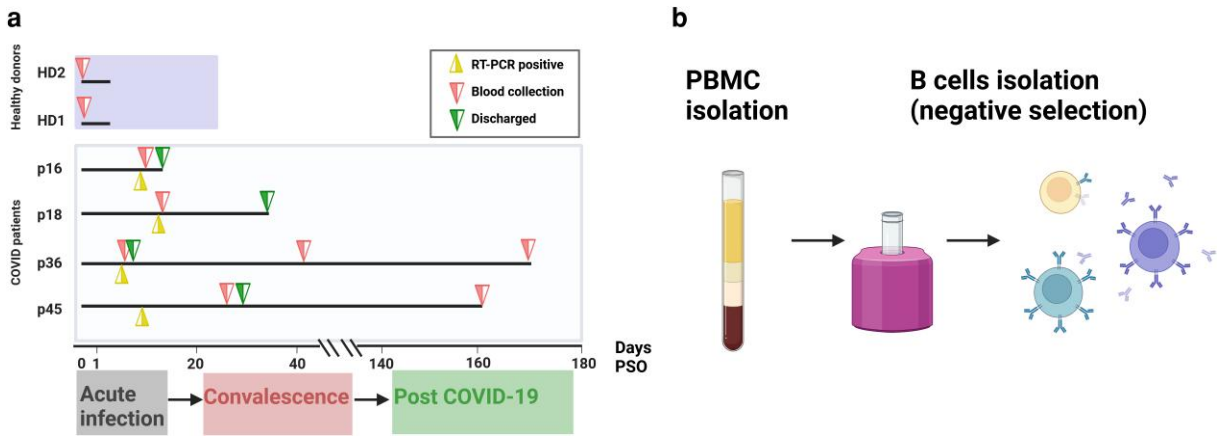

Fig. 3. Processing of samples from COVID-19 patients and HDs. a) Top: Timeline illustrating sample collection from HDs and hospitalized moderately ill patients (p). For all patients, symptom onset was defined as day 0. Acute infection was defined as day 1 to day 23 postsymptom onset (PSO); convalescence was defined as day 24 to day 45 PSO; and post-COVID-19 was defined as day 46 to day 169 PSO. b) PBMCs were isolated using Cell Preparation Tubes (BD) and cryopreserved at −150 °C until use. B cells from individual patients were isolated from PBMCs using negative immunomagnetic enrichment.

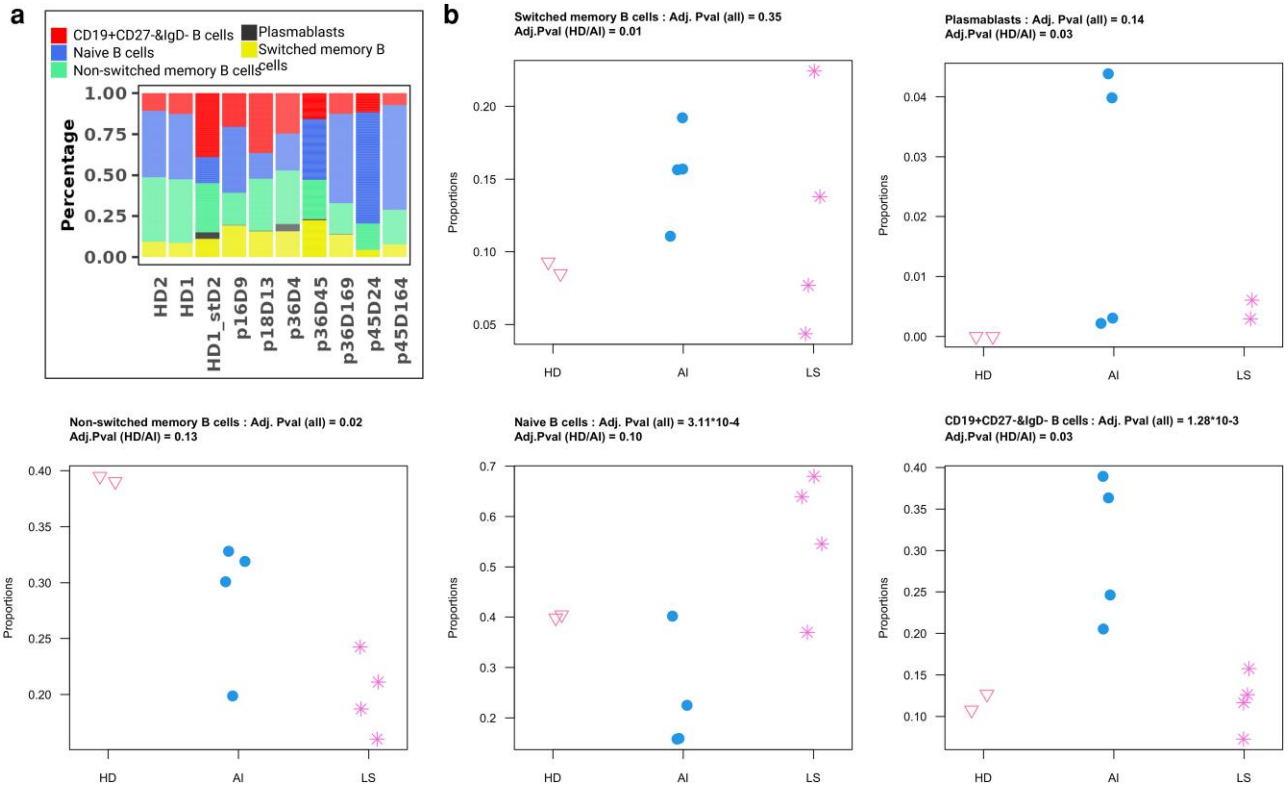

Fig. 4. B-cell composition changes during and after SARS-CoV-2 infection. a) Results for all samples processed in this study. HD_stD2 corresponds to HD-1 B cells, stimulated with CD40L and CpG for 2 days in vitro. b) Statistical analysis of the cell-type distributions across all samples shown in (a). Samples were grouped into HDs ($n = 2$), acute infection samples (AI; including the single in vitro–stimulated sample from HD; $n = 4$) and later time point samples (LS; referring to all samples that were taken >24 DPSO; $n = 4$). Adjusted *P*-values were determined based on an *f* test (comparison across all three sample categories, "(all)" values shown above each plot), or t test (comparison between categories 1 and 2, only, "(HD/AI)" values shown inside the plot.

over the course of the disease indicated significant changes in the composition of the B-cell immune repertoire (Fig. 4b). The proportions of B-cell subsets were very similar in the two HDs, indicating typical homeostasis of B-cell subsets circulating in the peripheral blood. Consistent with other reports (19–21), a surge of plasmablasts up to 4% of total B cells (p36D4) was observed in some of the acute infection samples but was undetectable in HDs. Similarly, the percentage of exhausted B cells in acute infection samples increased dramatically, from on average 12 to 27%, compared with HD B cells. In contrast (19), the percentage of naive B cells was lower in acute infection samples as compared with B cells from HDs (on average 26% in acute infection vs. 40% in HDs). Interestingly, a pattern similar to the SARS-CoV-2 acute infection could be observed for HD B cells, stimulated with CD40L and CpG for 2 days in vitro, e.g. an increase in plasmablasts from undetectable levels before stimulation to 4% after

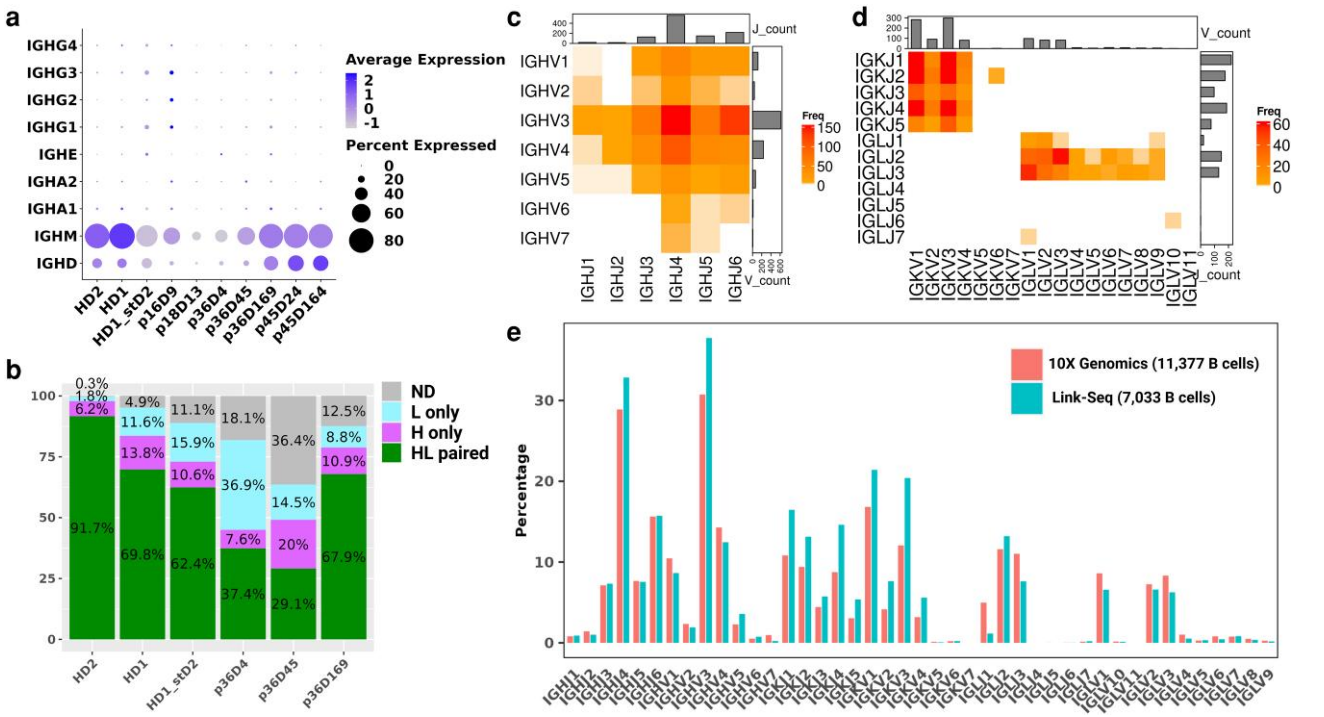

Fig. 5. Pairing efficiency and sensitivity of Link-Seq. a) BCR isotype gene expression levels over time in different patient samples. b) Percentage of B cells for which only the heavy chain (H only), only the light chain (L only), paired heavy and light chains (HL paired) or no IgG sequences could be obtained (ND). c) Detection of V and J genes for p36D169 (with 1,376 B cells) heavy chains and d) for light chains. e) V and J gene percentage for heavy, kappa and lambda chains as detected by Link-Seq (7,033 B cells) or 10× Genomics (11,377 B cells). Data for Link-Seq were obtained using a mix of HD and COVID-19 patient B cells, while 10× Genomics data were reprocessed based on an earlier publication (19), also from a mixture of HD and COVID-19 patient B cells.

stimulation. Taken together, these results nicely illustrate how Link-Seq can be used to monitor changes in the immune cell composition over the course of a disease.

Retrieving paired BCR for individual B cells and benchmarking the performance of Link-Seq for detection of rare V/J genes

After identifying the cell types, paired BCR genes were retrieved from the BCR repertoire sequencing data by matching cell barcodes to those in scRNA-seq data. During BCR library preparation, we noticed that the BCR library yields for B cells from acute infection and convalescent samples were much lower than those of post-COVID-19 samples from the same patients (Fig. S4), probably indicating cellular stress and mirroring the high level of exhausted B cells observed in these samples (Fig. 4a). We also observed a significant down-regulation of IGHM gene expression levels in B cells isolated from all three acute SARS-CoV-2 infection samples (mean percentage expression of 40% for acute infection samples compared with 82% for HDs; Fig. 5a). A similar down-regulation of the BCR gene was also observed in HD B cells stimulated in vitro with CD40L and CpG for 2 days, with the proportion of B cells expressing the IGHM gene decreasing from 84% in HD1 to 70% in HD1_stD2 (Fig. 5a). These results nicely resemble what had been observed in the acute SARS-CoV-2 infection group and potentially suggest a common mechanism for down-regulation of BCR gene expression levels by B cells in response to SARS-CoV-2 infection or combined CpG and CD40L stimulation. As a direct consequence of the BCR gene down-regulation during acute SARS-CoV-2 infection, the proportion of single B cells with matched heavy and light BCR genes was also reduced. For example, in p36D4 B cells, paired BCRs could only be retrieved from 37% of B cells (Fig. 5b), a proportion consistent with the percentage of B cells expressing heavy-chain genes (Fig. 5a). In contrast, 68–92% of paired BCRs were retrieved from B cells from HDs or late convalescent COVID-19 patients. For in vitro stimulated B cells, 62% paired BCRs could be obtained, which is slightly less than before stimulation, most likely due to the fact that we did not include IGHE primers in our protocol, despite the abundance of IGHE-expressing B cells in both CD40L/CpG-stimulated B cells and SARS-CoV-2 acute infection samples (Fig. S5a).

It is very important to cover rare V and J genes in an unbiased manner when sequencing the single B-cell immune repertoire. Potential bias is mainly based on (i) incomplete coverage of the diversity of the sequences flanking the V(D)J region, despite the use of a large set of primers; (ii) differences in annealing temperature, secondary structure, mispriming, and the ratio of individual primers can further cause large systematic biases; and (iii) abundance of different BCR genes in the gene pool, with highly abundant sequences dominating amplification. To overcome these limitations, we designed forward primers hybridizing to the template-switching oligo sequence, rather than to the leader sequences or V-gene framework-one sequences. By doing so, we reduced the number of forward primers from 21 (22) to just 1. We further reduced potential amplification bias by compartmentalizing individual BCR templates in droplets, thus preventing competition for PCR reagents and excluding chimeric hybridization. A direct indication for unbiased amplification of BCR genes is the detection of rare heavy- and light-chain V/J genes in the BCR sequencing repertoire. Previous studies have shown that several V gene families (e.g. IGHV7; IGKV5, 6, and 7; IGLV4, 10, and 11) are expressed at very low frequencies in the human immune repertoire (23). Consistently, a VH-VL library generated from 61,000 fresh IgG + B cells failed to detect the rare IGHV7 and IGKV7 (19).

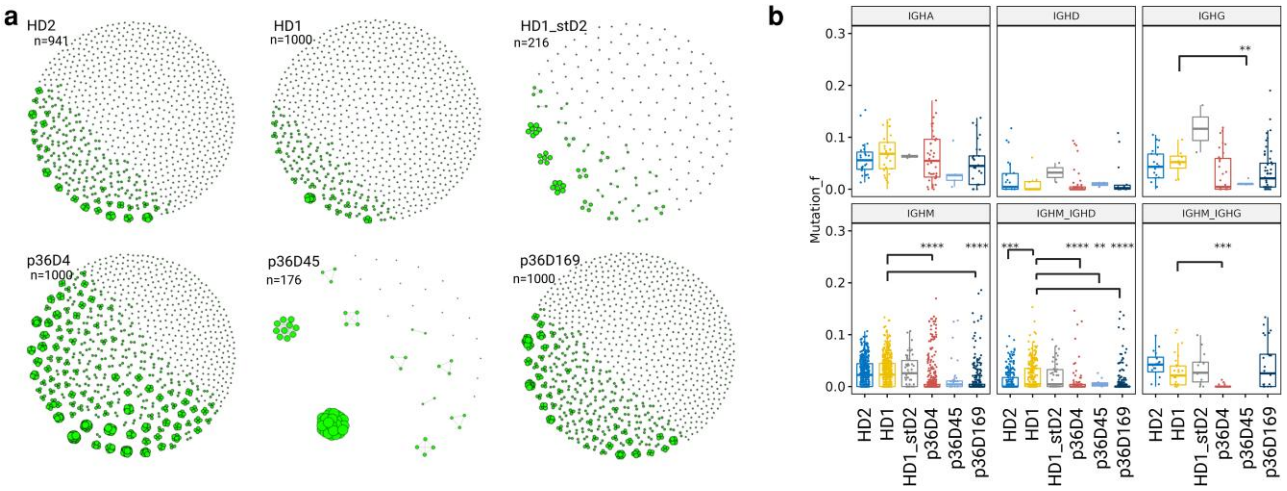

Fig. 6. Characterization of the BCR repertoires from healthy and diseased donors. a) Network graphs illustrating the clonal expansion of BCR repertoires for B cells from HDs (HD1/HD2), in vitro–stimulated HD B cells (HD1_stD2) or patient-36 B cells collected at multiple time points. Each dot represents a single B cell within a sample of size = *n*. Dots were connected if their heavy-chain CDR3 amino acid sequences had the same length and showed >80% similarity. b) Box plots show the SHM detected in COVID-19 patient samples and HDs used as controls. Asterisks indicate a significant difference in SHM levels between groups (*0.05; unpaired t test).

Applying Link-Seq, to a post-COVID-19 BCR library with only 1,376 B cells (p36D169), we could detect all IGHV genes, including the rare IGHV7, IGKV6, IGLV4, and IGLV10 variants (Fig. 5c and d). Notably, the extremely rare IGLJ6 gene, estimated at a frequency of 0.018% in an 11,377 B-cell library (19) prepared using the gold-standard 10× Genomics immune repertoire analysis method, was also detected in our Link-Seq sample (Fig. 5d). Based on the previously observed frequency, this is most likely based on the amplification of a single copy within our library. A more comprehensive comparison was performed between the same 10× Genomics library and a 7,033 B-cell library prepared with Link-Seq. The extremely rare IGLJ4, IGKV7, and IGLV11 not detected in our Link-Seq library were neither observed in the 1.62-fold larger 10× Genomics B-cell library. The only gene detected in the 10× Genomics library, but missing in our Link-Seq library is IGLJ5, which was only previously detected in 1 out of 11,377 B cells (Fig. 5e). Taken together, optimized primer design and single-molecular amplification of the BCR genes in droplets reduced the PCR bias and facilitated the detection of extremely rare BCR genes using Link-Seq.

BCR clonal expansion and progressive low somatic hypermutation in different B-cell subsets during SARS-CoV-2 infection

Following antibody repertoires over time in infected patients also allowed us to analyze clonal expansion of the B-cell repertoire. B cells with CDRH3 amino acid sequences of the same length and >80% similarity were considered to be clonally related and were connected in the clone network plot (24) (Fig. 6a). The clonal expansion of patient 36 started as early as day 4 PSO and diminished on day 169, still being higher than that of the two HDs. The B cells from patient 36 on day 45 also showed strong clonal expansion; however, due to poor cell viability, much fewer B cells were recovered, making a direct comparison with the two other time points very difficult.

Next, we analyzed the dynamics of somatic hypermutations (SHMs) of the BCR genes during SARS-CoV-2 infection. It is well known that SARS-CoV-2-reactive antibodies generally show very low levels of SHM. However, the dynamics of the low SHM and its characterization in different B-cell subsets have rarely been described. For this reason, we analyzed the characteristic low SHM across different time points in different B-cell subsets during SARS-CoV-2 infection. We observed a significantly lower level of BCR SHM in IGHM-positive B-cell subsets, as well as in IGHM/IGHD double-positive B-cell subsets. This was the case for all time points from day 4 to day 169 PSO, except for IGHM from p36D45, where the SHM level was lower when compared with HD1, but not statistically significant (Fig. 6b). These B-cell subsets are generally thought to consist mainly of naïve B cells and non-switched memory B cells.

However, the IGHM/IGHG double-positive B-cell subsets, which are in a transitional phase of class switching, showed a significantly lower level of SHM only during acute infection (p36D4), but no longer on day 169 (Fig. 6b and Table S8). For p36D45, only 176 B cells were recovered; among them, no IGHM/IGHG double-positive B cells were detected. In contrast, low SHM levels in the fully switched IGHA-positive B-cell subsets and IGHG-positive B-cell subsets were not observed in the acute infection sample (p36D4) or the post-COVID-19 stage sample (p36D169), but only in the convalescent phase (p36D45). Taken together, the characteristic low SHM levels observed in fully switched IGHA- or IGHG-expressing B cells lagged behind that of IGHM, IGHM/IGHD (naïve and nonswitched memory B cells), or IGHM/IGHG (isotype switching) B cells (Fig. 6b).

While much larger patient cohorts would be needed to dig deeper into the general mechanisms on SHM during SARS-COVID-19 infection, our case study nicely illustrates the technical possibilities offered by Link-Seq and the parameters that can be measured. The fact that our observations and sensitivities are in line with previous literature indicate no obvious biases. Based on many optimization steps, Link-Seq represents a mature method that can be easily applied to real-world samples of human disease survivors, enabling detailed insights into their immune repertoire.

Materials and methods

Patient samples and isolation of pan B cells

Patients were prospectively recruited during the first pandemic wave (March to June 2020) after the diagnosis of SARS-CoV-2

infection and hospitalization in Haukeland University Hospital, Bergen, Norway. Informed consent was obtained prior to recruitment. The study was approved by the Regional Committee for Medical and Health Research Ethics in Western Norway (#118664). PBMCs were isolated using Cell Preparation Tubes (BD) and cryopreserved at −150 °C until use. Human B cells were isolated from thawed PBMCs using EasySep Human B-cell isolation kit (Stemcell, catalog #19674), following the manufacturer's instructions.

B-cell stimulation

B cells were cultured in RPMI1640 complete medium at 37 °C in a 5% CO_2 atmosphere. Medium was supplemented with 10% FBS, 1 μg/mL CD40L (InvivoGen, catalog # rcyec-hcd40l) and 0.35 μM CpG2006 (25) (synthesized by Sigma) for 2 days.

Device fabrication and chip preparation

Microfluidic molds were fabricated using soft lithography. For the co-encapsulation of cells with barcode beads, the design described in Drop-seq (1) was used to manufacture a master mold with 87 μm height channels on a 3-inch silicon wafer. For the encapsulation of single cDNA molecules, a modified Drop-seq chip with a nozzle width and height of ∼40×50 μm was used. For chip production, polydimethylsiloxane (PDMS) base was mixed with cross-linker (Dow Corning) at a 10:1 ratio, degassed and poured into a Petri dish containing a mold. PDMS chips were then degassed again and incubated at 65 °C overnight to cross-link the PDMS base. After that, PDMS chips were cut out from the molds; inlets and outlets were punched with 0.75 mm biopsy punches (Harris Uni-Core) and treated with oxygen plasma before being bonded to 76×26 mm glass slides (Thermo Scientific). To render the PDMS channels hydrophobic, Aquapel (PGW Auto Glass, LLC) was injected into PDMS chips and incubated for a few minutes. Subsequently, the chips were flushed with air and baked at 80 °C for 10 min to evaporate any residual Aquapel. Finally, chips were flushed with Novec-7500 oil.

scRNA-seq library preparation

scRNA-seq libraries were generated using the Drop-seq method with some modifications. Briefly, isolated B cells were filtered through a 40-μm filter and resuspended in PBS (0.04% BSA) at a density of ∼100 cells/μL, before being co-encapsulated with barcode beads resuspended in cell lysis and RT buffer (2). For encapsulation, cells were injected using a syringe pump operating at a flow rate of 2,500 μL/h. Novec-7500 oil supplemented with 0.75% EA surfactant (RainDance) was injected at a flow rate of 5,000 μL/h. Subsequent to encapsulation, the emulsion was incubated at room temperature for 30 min, before being transferred to a preheated incubator at 42 °C for 90 min. Finally, we followed the original Drop-seq protocol to complete the recovery of barcode beads.

For cDNA amplification, barcoded beads were resuspended in 1× Kappa HiFi PCR mix at a concentration of 120 beads/μL and 0.2 μM cDNA multiplex primers (Table S1) were added. The PCR mixture with barcode beads was encapsulated using the original Drop-seq chip using Bio-Rad oil (QX200 Droplet Generation Oil for EvaGreen; Bio-Rad). Emulsions were aliquoted into PCR tubes for cDNA amplification over 16–18 cycles. After the PCR, 0.6 volumes of perfluorooctanol (PFO) were added to break the emulsion. The tubes were inverted several times to mix the emulsion with PFO and spun in a benchtop centrifuge at maximum speed for 12 min. The supernatant containing the amplified cDNA was transferred to a new tube, purified with 0.6 volumes of AMPure XP beads and resuspended in 15 μL H_2O. Amplified cDNA was measured with Qubit and analyzed with a fragment analyzer using the high-sensitivity DNA kit. Tagmentation was performed using homemade Tn5 produced by the EMBL Protein Core Facility and the EPFL Protein Core Facility, following the previously described Tn5 tagmentation protocol (26), but replacing the use of 25% DMF with 4% PEG8000 (27). The labeled libraries were analyzed using Qubit and Fragment Analyzer devices. The final libraries were sequenced on the Illumina NextSeq platform (75 bp, single end) using Drop-seq sequencing primers (1) to obtain single-cell transcriptome data.

Preparation of single B-cell BCR libraries

Encapsulation of single cDNA molecules

An average molecular size of 1,500 bp was assumed to calculate the number of cDNA molecules in each sample, which seems to be coherent with the obtained Fragment Analyzer profiles. For encapsulation, droplets with a diameter of about 40–55 μm were generated (Fig. S1a).

The number of cDNA molecules encapsulated per droplet is dictated by Poisson statistics. An adjusted λ of 0.5 was used for the encapsulation of cDNA molecules. According to Poisson statistics, 60% of these droplets are empty ($x = 0$), 30% have only one cDNA molecule ($x = 1$), and the probability (P) of having more than one cDNA molecule in a droplet is 9% ($x > 1$). However, according to our HD scRNA-seq data, heavy-chain BCR genes account for only 1.06% and light-chain BCR genes for only 1.5% of the total transcripts. Thus, the effective λ for heavy-chain genes is 0.0057, resulting in a probability of having more than one heavy-chain BCR transcript in a droplet of just 0.002%. Similarly, the probability of having more than one light-chain transcript in a droplet is 0.003% under these conditions.

$$P(x) = \frac{e^{-\lambda} \times \lambda^{x}}{x!}$$

For human B cells isolated from cryopreserved PBMCs, we detected an average of 300–500 UMIs per B cell. Based on our HD scRNA-seq data, heavy-chain BCR genes account for only 1.06% of the total transcripts, implying that each cell shows an average of 3–5 UMIs for heavy-chain transcripts. Considering a redundancy ratio (R = reads/UMIs) between 2 and 4 and using droplets with a diameter (d) of 40 μm, an emulsion with a total volume (V) of 2 mL PCR reaction and a cDNA concentration of 25 pg/mL ($\lambda = 0.5$) is sufficient to cover about, 8,000 to 27,000 single B cells.

$$\text{Number} = P(x) \times \frac{V}{(4/3) \times \pi \times (d/2)^3 \times R}$$

Single-molecule overlapping PCR in droplets

PCR mix for heavy- and light-chain libraries was prepared separately (Tables S2 and S3). The flow rate was adjusted to 1,200 μL/h for the PCR mix and 2,000 μL/h for oil (QX200 Droplet Generation Oil for EvaGreen; Bio-Rad).

After single-molecule encapsulation, the emulsion was aliquoted into PCR tubes and overlapping PCR was performed in droplets to link the V(D)J fragment with the cell-barcode-UMI fragment. The PCR program used for heavy-chain emulsion PCR was as follows: 3 min at 95 °C, 12 cycles of 10 s at 98 °C, 30 s at 66.4 °C, and 30 s at 72 °C, followed by 20 cycles of 10 s at 98 °C,

30 s at 57.5 °C, and 45 s at 72 °C. Ultimately, final strand extension was performed for 5 min at 72 °C.

The following PCR program was used for light-chain emulsion PCR: 3 min at 95 °C, 13 cycles of 10 s at 98 °C, 30 s at 66.4 °C, and 30 s at 72 °C, followed by 17 cycles of 10 s at 98 °C, 30 s at 57.5 °C, and 45 s at 72 °C, completed by final strand extension for 5 min at 72 °C. After the overlapping PCR, 0.6× volumes of PFO were added to the emulsion. The tubes were inverted vigorously several times, and spun at maximum speed on a benchtop centrifuge for 10 min to break the emulsion. The supernatants were first purified with a PCR product recovery column (QIAquick PCR Purification Kit). Subsequently, two further purification cycles using 0.5× and 0.8× AMPure XP beads were performed. Finally, the beads were resuspended in 46 µL ddH_2O.

L/T-based A-tailing

The purified emulsion PCR products were treated with lambda exonuclease to degrade the 5′ phosphorylated linear dsDNA from 5′ to 3′ direction. To do so, the following components were mixed: 1.2 µL of lambda exonuclease (5 U/µL), 6 µL of 10× buffer, and 40 µL purified emulsion PCR product in a total volume of 60 µL. The reactions were carried out at 37 °C for 30 min. After that, 1.8 µL TDT (20 U/µL), 6 µL dATP (1 mM), and 18 µL TDT buffer (5×) were added directly to the 60 µL reaction. The resulting 90 µL sample was incubated at 37 °C for 30 min. After the L/T treatment, the products were purified with 0.8× AMPure XP beads and resuspended in 45 µL H_2O.

Nested PCR

The first round of nested PCR was performed using a total reaction volume of 50 µL consisting of the following reagents: 21 µL of overlapping PCR products, 2 µL sense 10 µM primer mix (IgG:IgM:IgA:IgD = 4:3:2:1), 2 µL 10 µM reverse primers, and 25 µL 2× Kapa Hifi mix (Table S1). The PCR program for heavy-chain library NEST1 was as follows: 3 min at 95 °C, nine cycles of 20 s at 98 °C, 30 s at 62 °C, and 45 s at 72 °C, completed by final strand extension for 5 min at 72 °C. The PCR program for light-chain library NEST1 was the same, except with two cycles less (seven cycles in total). The PCR products were purified with 0.5× SPRI beads and resuspended in 22 µL H_2O.

The second round of nested PCR was performed using a total reaction volume of 50 µL consisting of the following reagents: 10 µL of first round nested PCR products, 2 µL sense primer (10 µM) and 2 µL reverse primers (10 µM; Table S1), and 25 µL 2× Kapa Hifi mix. The second round nested PCR program for heavy- and light-chain library was as follows: 3 min at 95 °C, and 12 cycles of 20 s at 98 °C, 30 s at 55 °C, and 25 s at 72 °C, completed by final strand extension for 5 min at 72 °C. The PCR products were recovered with 0.6×SPRI beads and resuspended in 15 µL H_2O. The final BCR library was measured with Qubit and Fragment Analyzer devices. The final heavy-chain BCR library consists of fragments with a size around 670 bp, while the light BCR library shows a peak at about 600 bp. Equimolar heavy- and light-chain libraries were mixed and sequenced using the Illumina MiSeq platform (V2, 500 bp). Custom sequencing primers (Table S4) were used for read-1, index read-1, and read-2. For the index read-1, the primers consist of a mixture of heavy- and light-chain sequencing primers at equal molar concentrations for each allele, while kappa primers were mixed with lambda primers at a 4:1 ratio. Overall, heavy- and light-chain primer mixtures were used at equimolar concentrations. Twenty-one base pairs were assigned for read 1, 250 bp for index read 1, 8 bp for index read 2 (when sequencing multiple samples), and the remaining read length for read 2. All primer-binding sites are illustrated in Fig. S6.

Bioinformatic methods

All bioinformatic methods and tools are comprehensively described in the Supplementary Material. Data on sequencing quality and B-cell clustering approaches are summarized in Tables S6–S8 and Figs. S7–S9.

Discussion

Here, we present Link-Seq, a generalizable experimental workflow to integrate short-read single-cell gene expression profiles with long-read sequences of interest, such as V(D)J sequences of BCR transcripts, located away from the 3′-end-cell barcodes. Based on single-molecule amplification of the BCR gene in droplets, Link-Seq enables us to detect extremely rare V and J genes as benchmarked using the gold-standard commercial 10× Genomics immune repertoire sequencing method. As a result, up to 91.7% of paired heavy and light chains, genes can be retrieved for B cells isolated from HDs, which is comparable with 10× Genomics technology (19) and is more than double of other open platforms, such as DART-Seq (12) and RAGE-Seq (13). In addition to the high efficiency for getting paired BCR genes, Link-Seq also allows us to retrieve the full-length BCR genes via direct assembly of paired MiSeq reads. Missing FR1 and CDR1 regions of long heavy chains can be reconstructed from germline sequences, which should have only small impact on antigen-binding properties. As an alternative, the enriched BCR libraries could also be sequenced on PacBio sequencing platform to retrieve full-length BCR variable sequences.

In parallel, Link-Seq allowed us to detect different B-cell subtypes and their changes in composition, over the course of a disease. Taken together, Link-Seq complements the existing toolbox for immune repertoire sequencing by enabling monitoring and characterization of the expansion of specific BCR clones, while getting transcriptomic data on the entire pan-B-cell repertoire. Comparable commercial transcriptomic platforms, such as the 10× Genomics immune repertoire sequencing method, are very powerful, but they are very costly and represent closed systems that cannot be modified for specific applications. In contrast, Link-Seq's ability to freely fuse any sequence of interest at any position within a gene with the cell barcode and UMI, opens further interesting possibilities. Given that the resulting constructs are short enough to be compatible with the read length of, for example, the Illumina MiSeq platform, it also overcomes the need for long reads in many applications.

Furthermore, Link-Seq offers the unique feature of barcoding multiple "targeted" genes within a given cell, paving the way for a variety of correlative approaches. We show here how two independently expressed genes (heavy- and light-chain antibody genes) can be barcoded at specific sites. This should be considered as a general proof of concept for barcoding specific parts of multiple long gene sequences, for subsequent correlation. An immediate expansion of the applications shown here is the paired sequencing of TCRs, which can be done with very little modifications of the current workflow. Furthermore, we believe Link-Seq could also be useful for monitoring co-evolution of proteins in cancer signaling pathways (e.g. monitoring mutations in different proteins of the same pathway) (28–30) or viral mutations in the context of host cell adaptation and development of resistances (e.g. changes in the envelope protein that correlate with

mutations of the core proteins) (31, 32). Sequencing exclusively the barcoded genes of interest, rather than the entire transcriptome, can increase throughput by orders of magnitude while reducing overall cost.

Already for the "standard" application of monitoring immune cell repertoires, Link-Seq is significantly more cost-efficient when compared with 10× Genomics, while offering a similar performance and sensitivity. In terms of cost, our approach sums up to about 8.6 cents per cell, which is roughly five times cheaper than 10× Genomics.

Another advantage of Link-Seq is its open and modular character, which is generally also compatible with more complex and customized single-cell sequencing workflows, including the selective labeling of new mRNA to improve clustering resolution (14, 33). The labeling relies on immobilized mRNA that can be exposed to chemical treatments and washing steps, which can be implemented in a 10× Genomics workflow only when using fixed cells exposed to a specific treatment condition in bulk (34). Upon encapsulation into droplets, the mRNA hybridizes to soluble barcodes and is directly converted into cDNA, preventing any further labeling steps. In contrast, Link-Seq, derived from Drop-seq, retains the option to perform the cDNA synthesis step in bulk after breaking the droplets, using mRNA that is still bound to barcoded beads. Therefore, the chemical conversion can be carried out at this stage, enabling to still "record" any transcriptional changes inside the droplets. Given the many possibilities of generating biologically and chemically different droplets (35–37), this opens the way for interesting multiplexing approaches, in which the effect of many different stimuli on a cell's transcriptome can be analyzed.

While Link-Seq is fully open, flexible, and more cost-efficient than commercial platforms, its establishment in biology labs requires at least some basic microfluidics equipment and expertise. It should be noted, however, that methods of similar technological complexity, such as Drop-seq, have been picked up by hundreds of nonspecialized biology labs around the world. On the hardware side, one needs much less than what is required for phenotypic droplet analyses platforms, which have been shared in great detail with the scientific community recently (38). All required modules for the generation of droplets can also be obtained from commercial suppliers, such as Dolomite Microfluidics, microfluidic ChipShop, and other companies.

Acknowledgments

The authors thank the COVID-19 patients for altruistically contributing to this research. They thank Dr Kristin Mohn, Professors Bjørn Blomberg, Camilla Tøndel, and Nina Langeland for patient recruitment and members of Bergen COVID-19 research group for processing blood samples. Figures 1, 2a, 3b, and S6 were created with BioRender.com

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

C.A.M. acknowledges generous funding through an SNSF Bridge grant (40B2-0_218721). The Influenza Centre in Bergen, Norway, is supported by the Ministry of Health and Care Services, Norway; Helse Vest (F-11628), the Trond Mohn Stiftelse (TMS; TMS2020TMT05), the European Union (EU IMI 2 101007799 Inno4Vac, H2020 874866 INCENTIVE, H2020 101037867 Vaccelerate), and the Norwegian Research Council Globvac program (284930).

Author Contributions

Conceptualization: H.H., F.Z., K.A.B., R.J.C., and C.A.M. Experimental work: H.H., F.Z., and L.K. Sequencing and data analysis: X.M., C.G., and V.B. Supervision: R.J.C. and C.A.M. Manuscript writing: H.H., F.Z., X.M., K.A.B., L.K., R.J.C., and C.A.M.

Data Availability

All raw data underlying this article are publicly available on the Zenodo repository (https://doi.org/10.5281/zenodo.11079548 and https://doi.org/10.5281/zenodo.11086086).

References

1 Macosko EZ, *et al.* 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 161:1202–1214.
2 Klein AM, *et al.* 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 161:1187–1201.
3 Gierahn TM, *et al.* 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods*. 14: 395–398.
4 Rosenberg AB, *et al.* 2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. 360:176–182.
5 Cao Y, *et al.* 2020. Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. *Cell*. 182:73–84.e16.
6 Dugan HL, *et al.* 2021. Profiling B cell immunodominance after SARS-CoV-2 infection reveals antibody evolution to non-neutralizing viral targets. *Immunity*. 54:1290–1303.e7.
7 Ehling RA, *et al.* 2022. SARS-CoV-2 reactive and neutralizing antibodies discovered by single-cell sequencing of plasma cells and mammalian display. *Cell Rep*. 38:110242.
8 Ju B, *et al.* 2020. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature*. 584:115–119.
9 Scheid JF, *et al.* 2021. B cell genomics behind cross-neutralization of SARS-CoV-2 variants and SARS-CoV. *Cell*. 184:3205–3221.e24.
10 Wang S, *et al.* 2021. A single-cell transcriptomic landscape of the lungs of patients with COVID-19. *Nat Cell Biol*. 23:1314–1328.
11 Woodruff MC, *et al.* 2020. Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nat Immunol*. 21:1506–1516.
12 Saikia M, *et al.* 2019. Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nat Methods*. 16:59–62.
13 Singh M, *et al.* 2019. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun*. 10:3120.
14 Qiu Q, *et al.* 2020. Massively parallel and time-resolved RNA sequencing in single cells with scNT-Seq. *Nat Methods*. 17:991–1001.
15 Ohtsubo Y, Nagata Y, Tsuda M. 2017. Efficient N-tailing of blunt DNA ends by Moloney murine leukemia virus reverse transcriptase. *Sci Rep*. 7:41769.
16 Motea EA, Berdis AJ. 2010. Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim Biophys Acta*. 1804:1151–1166.
17 Turchaninova MA. 2013. Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol*. 43:2507–2515.

18 Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. 2013. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*. 14:632.
19 Zhang J-Y, *et al.* 2020. Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol*. 21:1107–1118.
20 Wilk AJ, *et al.* 2020. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med*. 26:1070–1076.
21 Kaneko N, *et al.* 2020. Loss of Bcl-6-expressing T follicular helper cells and germinal centers in COVID-19. *Cell*. 183:143–157.e13.
22 Ippolito GC, *et al.* 2012. Antibody repertoires in humanized NOD-scid-IL2Rγ(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS One*. 7:e35497.
23 DeKosky BJ, *et al.* 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol*. 31:166–169.
24 Paschold L, *et al.* 2022. Rapid hypermutation B cell trajectory recruits previously primed B cells upon third SARS-Cov-2 mRNA vaccination. *Front Immunol*. 13:876306.
25 Hartmann G, Krieg AM. 2000. Mechanism and function of a newly identified CpG DNA motif in human primary B cells. *J Immunol*. 164:944–953.
26 Hennig BP, *et al.* 2018. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3 (Bethesda)*. 8:79–89.
27 Picelli S, *et al.* 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res*. 24: 2033–2040.
28 Pietilä EA, *et al.* 2021. Co-evolution of matrisome and adaptive adhesion dynamics drives ovarian cancer chemoresistance. *Nat Commun*. 12:3904.
29 Ciriello G, *et al.* 2024. Cancer evolution: a multifaceted affair. *Cancer Discov*. 14:36–48.
30 Zhang M, *et al.* 2013. Using the theory of coevolution to predict protein-protein interactions in non-small cell lung cancer. *Chin J Cancer*. 32:91–98.
31 Quadeer AA, Morales-Jimenez D, McKay MR. 2018. Co-evolution networks of HIV/HCV are modular with direct association to structure and function. *PLoS Comput Biol*. 14:e1006409.
32 Champeimont R, Laine E, Hu S-W, Penin F, Carbone A. 2016. Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. *Sci Rep*. 6:26401.
33 Hendriks G-J, *et al.* 2019. NASC-seq monitors RNA synthesis in single cells. *Nat Commun*. 10:3138.
34 Neuschulz A, *et al.* 2023. A single-cell RNA labeling strategy for measuring stress response upon tissue dissociation. *Mol Syst Biol*. 19:e11147.
35 Miller OJ, *et al.* 2012. High-resolution dose-response screening using droplet-based microfluidics. *Proc Natl Acad Sci U S A*. 109: 378–383.
36 Kulesa A, Kehe J, Hurtado JE, Tawde P, Blainey PC. 2018. Combinatorial drug discovery in nanoliter droplets. *Proc Natl Acad Sci U S A*. 115:6685–6690.
37 Mathur L, *et al.* 2022. Combi-seq for multiplexed transcriptome-based profiling of drug combinations using deterministic barcoding in single-cell droplets. *Nat Commun*. 13:4450.
38 Panwar J, Autour A, Merten CA. 2023. Design and construction of a microfluidics workstation for high-throughput multi-wavelength fluorescence and transmittance activated droplet analysis and sorting. *Nat Protoc*. 18:1090–1136.