



OPEN

# Ultrasensitive amplicon barcoding for next-generation sequencing facilitating sequence error and amplification-bias correction

Ibrahim Ahmed<sup>1,2</sup>, Felicia A. Tucci<sup>3</sup>, Aure Aflalo<sup>1,4</sup>, Kenneth G. C. Smith<sup>1,5</sup> & Rachael J. M. Bashford-Rogers<sup>1,3</sup>✉

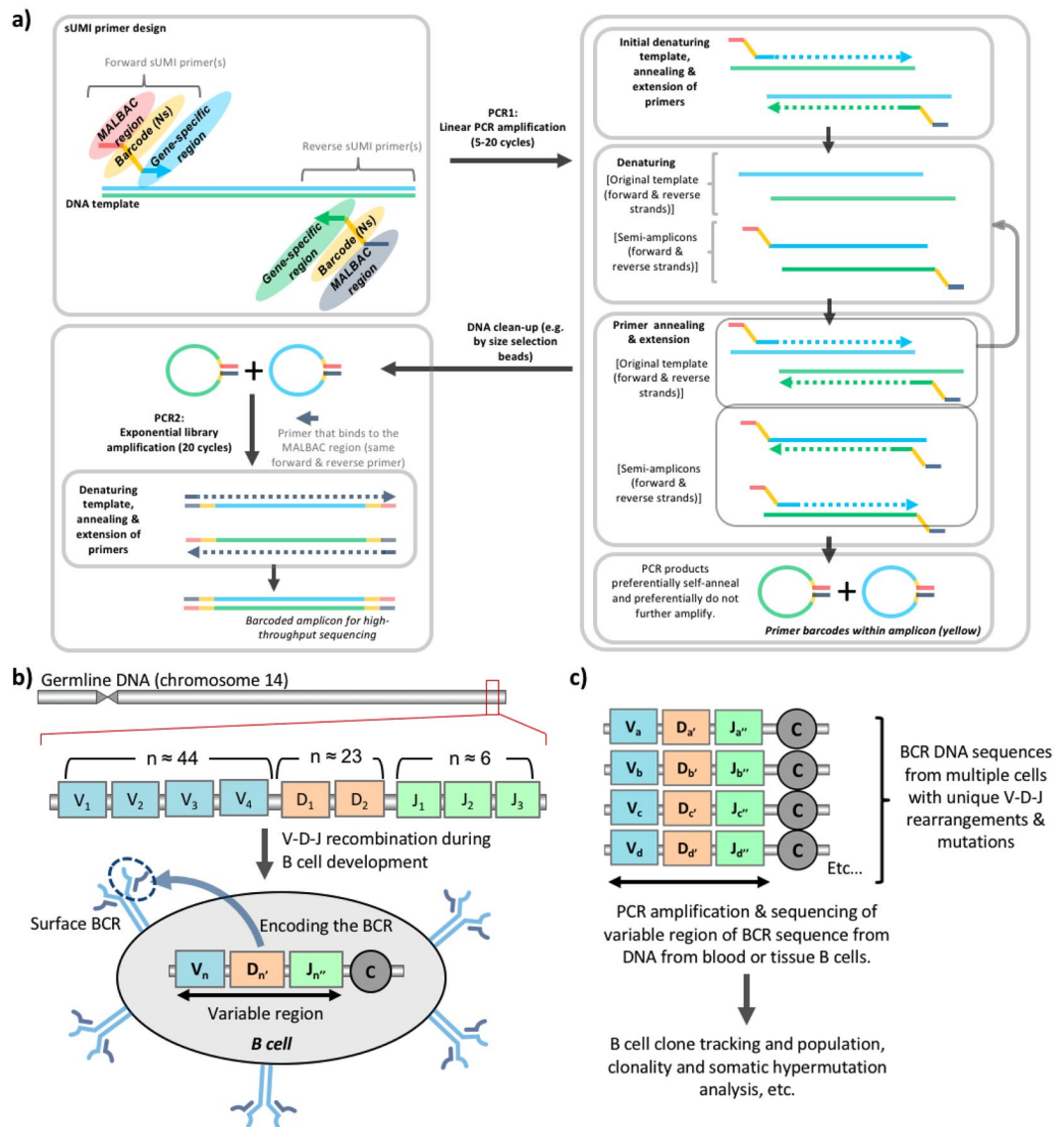
The ability to accurately characterize DNA variant proportions using PCR amplification is key to many genetic studies, including studying tumor heterogeneity, 16S microbiome, viral and immune receptor sequencing. We develop a novel generalizable ultrasensitive amplicon barcoding approach that significantly reduces the inflation/deflation of DNA variant proportions due to PCR amplification biases and sequencing errors. This method was applied to immune receptor sequencing, where it significantly improves the quality and estimation of diversity of the resulting library.

Amplicon sequencing is often the basis for characterizing DNA variant proportions, and is routinely used in many areas including tumor heterogeneity<sup>1</sup>, 16S microbiome<sup>2</sup>, viral<sup>3</sup>, CRISPR/Cas9 library screens<sup>4</sup> and immune receptor sequencing<sup>5</sup>. However, the ability to accurately quantify the proportions of DNA variants is hampered by amplification biases that lead to inflation/deflation of some DNA amplicons, as well as the inability to correct sequencing errors (Fig. S1a). To overcome the amplification biases in DNA-based amplification and sequencing, we developed a novel generalizable ultrasensitive amplicon barcoding approach that significantly reduces the inflation/deflation of DNA variant proportions from PCR amplification biases and sequencing errors.

Amplification biases from RNA starting material have been largely addressed by the introduction of unique molecular identifiers (UMIs) in the reverse transcription primers (barcoded primers), thus subsequent PCR amplification of each cDNA molecule can be quantified and corrected through the capture of the UMI barcode. However, when starting from a DNA template, this approach cannot be used. Previous attempts at generating barcoded PCR amplicons from DNA using barcoded primers via standard exponential PCR amplification leads to the preferential amplification of PCR amplicons rather than template<sup>6</sup> and thus resulting in significant amplification biases (Fig. S1b).

To overcome these issues, we established a sUMI-seq PCR amplification using barcoded primers that generate self-annealing amplicons (Fig. 1a, denoted sUMI-seq primers. Secondary structure-assisted UMI incorporation, amplification and sequencing). These sUMI-seq primers contain three key regions: (1) the target gene-specific region, (2) a UMI primer barcode (8 bp), and (3) a region based on multiple annealing and looping-based amplification cycles (MALBAC) methodology<sup>7</sup>, in which the PCR products are able to self-anneal forming MALBAC amplicon loops. These amplicon loops preferentially do not further amplify due to the thermodynamic and kinetic preference for loop closure compared to further primer annealing to the open and available original DNA template (Fig. S2). This will result in a close-to-linear amplification, rather than standard exponential amplification, of template DNA due to the unavailability of the MALBAC amplicon loops to further amplify. This first PCR (PCR1) is followed by a cleanup step to remove unbound primers and primer dimers. Then a second PCR (PCR2), with primers annealing to the common MALBAC region of PCR1 amplicons, generates linearized amplicons that are amenable for library preparation and high-throughput sequencing. A bioinformatics pipeline was developed to identify the primer barcodes, to correct for amplification frequency and to correct sequencing

<sup>1</sup>Department of Medicine, University of Cambridge, Cambridge, United Kingdom. <sup>2</sup>Present address: Faculty of Biology, Medicine and Health, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK. <sup>3</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom. <sup>4</sup>Department of Pathology, University of Cambridge, Cambridge, United Kingdom. <sup>5</sup>Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, United Kingdom. ✉e-mail: [rbr1@well.ox.ac.uk](mailto:rbr1@well.ox.ac.uk)



**Figure 1.** Schematic of the sUMI-seq PCR with DNA molecular barcoding approach. **(a)** Schematic diagram of sUMI-seq. Quantification of pools of DNA variants can be important across multiple fields in biology, including B cell receptor (BCR) repertoire sequencing. MALBAC-barcoded primer design is shown. Amplicons from PCR1, after a clean-up step, are amplified in PCR2 using forward primers priming gene specific regions (as in PCR1) and reverse universal primers binding to the MALBAC regions **(b)** Schematic diagram of BCR rearrangement: B cells are generated from haematopoietic stem cells. The IGH gene locus on chromosome 14 (in humans) encodes for multiple distinct copies of the variable (V), diversity (D), and joining (J) genes, with functional IGH BCR (one functional allele per cell) generated during B cell differentiation by site-specific V-D-J recombination. Random deletions and insertions of nucleotides during recombination results in sequence diversification at the gene junctional regions. Rearranged BCR genes can be further diversified through somatic hypermutation (SHM) upon B cell activation. **(c)** A BCR repertoire is defined as the BCRs collection from a B cell population, such as from blood or tissues. B cell DNA can be used to construct BCR libraries, using multiplex PCR with forward primers annealing to all the variable VH gene families (IgH V1–7) and reverse primers annealing to the JH gene families (IgH J1–6).

errors through alignment of sequences sharing the same barcode (code made available at [https://github.com/rbr1/sUMI\\_processing\\_pipeline](https://github.com/rbr1/sUMI_processing_pipeline)).

sUMI-seq is made possible by two key innovations: firstly, the PCR amplification step (PCR1) using the sUMI-seq primers allows for preferential amplification of the template DNA and minimal amplification of the MALBAC looped amplicons. Secondly, linearization of the self-annealed PCR amplicons in PCR 2 leads to increased sensitivity in the face of low template DNA input. The use of barcoded sample primers in PCR2 allows for sample pooling, and efficient library preparation and sequencing.

One area in which quantification of DNA variants is important is in B cell receptor sequencing. B cell receptors (BCRs) are membrane-bound immunoglobulins (Igs) which are secreted as antibodies by antibody secreting

cells (plasma cells), which differentiate from naïve or memory B cells upon antigen activation. The huge diversity of the antibody repertoire is due to DNA recombination of variable gene segments (V, (D), J) at the Ig heavy (IgH) and light (IgL) chain loci during B cell ontogeny, and subsequent acquisition of somatic hypermutation (SHM) in activated B cells. Both IgH and IgL variable regions are further subdivided into four framework regions (FWR 1–4), determining the antibody folding, and into three complementarity determining regions (CDR 1–3), involved in antigen binding. BCRs represent unique markers for each B cell clone (Fig. 1b,c). The BCR repertoire analysis by BCR gene deep sequencing allows measurement of the diversity and complexity of B cell response, and to identify clonal related B cells (B cell clones) which correlate with different immunological conditions. BCR repertoire analysis from blood or tissues by high-throughput sequencing has been used to provide powerful insights into B cell biology and tracking B cell clones in the context of health<sup>6,8</sup>, autoimmunity<sup>9</sup>, cancer<sup>10,11</sup>, infection<sup>12</sup>, vaccination<sup>13</sup> and in other diseases. BCR sequencing from RNA has been established using UMIs of each RNA molecule to accurately quantify relative BCR RNA frequencies<sup>6</sup>. Despite the successes of RNA-based BCR repertoire sequencing, RNA can be a sub-optimal substrate for BCR sequencing. The level of Ig transcripts is upregulated through B cell maturation and activation upon antigen encounter, with plasma cells having the highest amount of Ig mRNA per cell. This may lead to inflation or deflation of the detected B cell clonotypes of a repertoire. This was shown clearly in the detection of B cell acute lymphocytic leukaemia<sup>10</sup>, where lower numbers of BCR RNA molecules per leukemic cell compared to non-leukemic B cells lead to a significant underrepresentation of patient tumor proportion. This major limitation can be overcome through sequencing from DNA, as B cells carry one functional BCR allele per cell (one B cell – one antibody). However, no reliable method for molecular barcoding during PCR amplification of DNA has yet been established, thus leading to potential amplification biases in the resulting sequencing data.

Here we develop and validate a novel generalizable ultrasensitive amplicon barcoding approach and apply it to BCR sequencing, where it significantly improves the quality and estimation of diversity of the resulting library.

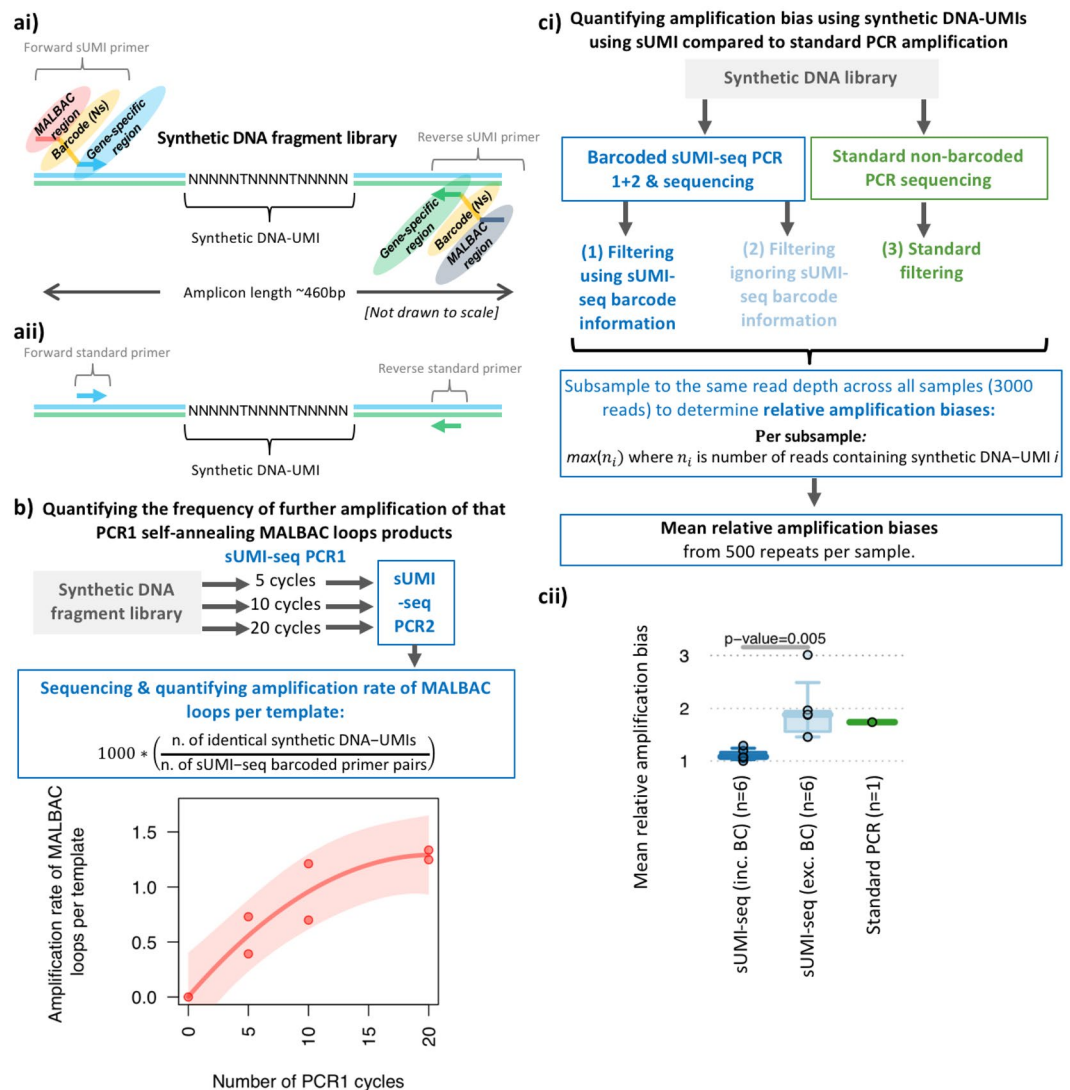
## Results and Discussion

To test the effectiveness of sUMI-seq PCR, a synthetic DNA fragment library was designed containing an internal DNA barcode (referred to as synthetic DNA-UMI) unique to each DNA molecule (Figs. 2a, S3). This synthetic DNA fragment UMI library design was based on a BCR sequence. This means that both the synthetic DNA-UMI and BCR repertoire from clinical samples can be amplified using the same primer sets. Specifically, forward primers anneal to the IgH V genes (FR3) and reverse primers anneal to IgH J genes (Table S1), allowing the amplification of the IgH variable region encompassing the CDR3 which is the major determinant of antibody-binding specificity (Fig. 5ci). Together, this synthetic DNA fragment UMI library design facilitates quantification of the relative amplification of each unique DNA template between methods. The sUMI-seq PCR was applied to the synthetic DNA fragment library using either 5, 10, or 20 PCR cycles in PCR1, to test the effect of different PCR cycles, followed by PCR2 (20 cycles). In addition, a standard non-barcoded PCR using standard non-barcoded primers (i.e. containing the gene-specific annealing region only), using the same synthetic DNA-UMI as template, was amplified with an equivalent approach (see methods). Each reaction condition successfully generated PCR amplicons that were subsequently sequenced by MiSeq (sequencing information in Table S2).

Firstly, we quantified the frequency of further amplification of PCR1 self-annealing MALBAC loops products depending on the number of PCR cycles. This is achieved through the assessment of the frequencies of the synthetic DNA-UMI per primer barcode pair (i.e.  $1000 \times (\text{number of identical synthetic DNA-UMIs} / \text{number of sUMI-seq primer barcode pairs})$ ), (Fig. 2b). As expected, the frequency of duplicated synthetic DNA-UMIs per sUMI-seq primer barcode pair was low (mean rate per sample of 0.392–1.335 per 1000 reads). The rate of duplicated synthetic DNA-UMIs per sUMI-seq primer barcode pair increased with the number of PCR 1 cycles that appeared to asymptote at 1.291 (Fig. 2b, 75% confidence intervals 0.939–2.536, p-value = 0.0168). This demonstrated only a low level of further amplification of the MALBAC loop amplicon in PCR1, and this level can be tailored depending on the number of PCR cycles. Importantly, between 5–10 cycles and between 10–20 cycles, less than a 2-fold increase in MALBAC-loop-specific amplification was observed. This suggests that the PCR1 step MALBAC-loop-specific amplification is non-exponential, and the PCR 1 step predominantly amplifies the original DNA template.

Next, we quantified and compared the relative amplification biases between sUMI-seq and standard non-barcoded PCR through synthetic DNA-UMIs amplification. To determine the effectiveness of the sUMI-seq primers in reducing the amplification biases, we also compared the relative amplification biases after filtering using or ignoring the sUMI-seq barcode information (Fig. 2ci). To account for differences in read depths between the different methods, each filtered dataset was subsampled to the same read depth across all samples (3000 reads), and relative amplification biases were calculated, defined as the maximum number of reads containing the same synthetic DNA-UMI. The mean relative amplification biases was calculated from 500 repeats per sample. (Fig. 2cii). Indeed, the mean relative biases were equivalent between sUMI-seq PCR ignoring the barcode information and the standard non-barcoded PCR. However, the mean relative biases were significantly lower in the sUMI-seq PCR using the barcode information compared to ignoring the barcode information (p-value=0.005). Together, this highlights the need for accounting for amplification biases.

Finally, a quantitative amplicon barcoding method should have a linear correlation between DNA template input and sequence output. To test this, we performed a dilution series of a peripheral blood (PB) DNA sample mixed with the synthetic UMI-DNA library at varying ratios, and sUMI-seq PCR was applied, again using either 5, 10, or 20 PCR1 cycles (Fig. 3a). The PB DNA sample was from a chronic lymphocytic leukaemia (CLL) patient, characterised by a clonal expansion of a single B cell clone, where >50% of all peripheral B cells contain a single IgH VDJ rearrangement (IGHV1–69\*14-IGHJ6\*02), as previously published<sup>5</sup>. Indeed, there was a strong linear relationship between CLL DNA input and proportion of sequencing reads after accounting for barcodes (Figs. 3b,c, S4). This suggests that sUMI-seq primers can be used to accurately correct amplification bias.



**Figure 2.** sUMI-seq more accurately quantified B cell receptor sequence repertoires. **(a)** Schematic diagram of the synthetic barcoded DNA fragments (representing a BCR sequence) used to quantify the DNA capture and amplification biases of DNA sequencing methods. A library of synthetic DNA fragments was designed based on a BCR sequence and containing a region of random nucleotides such that each DNA fragment is unique (termed synthetic DNA-UMIs). These were amplified using the sUMI-seq primers (**a i**, **a ii**) and standard PCR primers with gene specific sequences. **(b)** Quantifying the frequency of further amplification of that PCR1 self-annealing MALBAC loops products under 5, 10 and 20 PCR 1 cycles performed in duplicate. **(c)** Quantification of the amplification bias using synthetic DNA-UMIs and sUMI-seq compared to standard PCR amplification. **(i)** Schematic diagram of the experimental design and **(ii)** a boxplot of the mean relative amplification biases between sUMI-seq (filtering using barcode information), sUMI-seq (filtering ignoring barcode information), and standard PCR amplification without barcodes. The number of samples per box is provided in brackets. Wilcoxon tests were performed in R.

We next applied sUMI-seq PCR to BCR sequencing of clinical samples, namely on a well-characterized cohort of peripheral blood mononuclear cell (PBMC) samples from 11 healthy individuals and 4 chronic lymphocytic leukemia (CLL) patients who have previously been sequenced using the conventional BCR non-barcoded amplification method<sup>14</sup>. Subsequent linearization of the amplicons in PCR2 with the inclusion of a sample-specific barcode in the linearization primers was used to facilitate efficient sample pooling before library preparation and high-throughput sequencing (Fig. 4a). This yielded between 4228–29372 unique sUMI-barcodes per sample after filtering for BCRs, comprising between 1055–4688 unique IgH V-D-J rearrangement per sample (Table S1). As previously observed, the healthy individual samples yielded diverse BCR repertoires (Figs. 4b and S6), whereas the CLL samples were characterized by the clonal expansion of a single malignant B cell clone (Fig. 4b), demonstrated by the increased maximum clone size and clonal diversification indices. The BCR sequences of dominant malignant BCR clones identified by sUMI-seq were identical to that of conventional BCR non-barcoded amplification methods and BCR amplification by RNA as previously published<sup>5</sup> (Fig. S7). Furthermore, the frequency

of each B cell clone, as defined by the CDR3 of the BCR sequence, was highly correlated with that of the conventional DNA amplification method (Fig. 4c). Together, this demonstrated that sUMI-seq PCR could be used to efficiently capture BCR repertoire data from DNA sources.

We next determined whether the capture of BCR repertoires were significantly improved using sUMI-seq. Given that sUMI-seq benefits from both error-correction and amplification bias-correction (see methods), we hypothesized that the estimation of (1) clonal diversity, (2) the level of somatic hypermutation and (3) mean amplicon length would be improved compared to standard non-barcoded BCR PCR amplification methods.

Firstly, the relative clonal diversity of all clones representing >1% of the total repertoire in each sample was compared between filtering using sUMI-seq barcode information to correct for amplification biases and filtering ignoring the sUMI-seq barcode information, whilst accounting for read depth (Fig. 5a,bi). Indeed, the use of the sUMI-seq barcode information resulted in a significant reduction in estimated clonal diversity in all clones tested in both healthy (diverse) and CLL (clonal) BCR repertoires ( $p$ -values <  $1e-10$ , Fig. 5bii). This suggests that standard PCR amplification methods overestimate the diversity of DNA pools due to the introduction of PCR amplification and sequencing errors, which can be corrected through the use of sUMI-seq primer barcoding. Secondly, the estimation of the level of somatic hypermutation (SHM) was significantly reduced when the sUMI-seq barcode information was used for filtering. This was demonstrated by a significantly higher proportion of unmutated BCR sequences (i.e. the IGHV region within 1 bp difference from the closest germline reference gene) when using the sUMI-seq primer barcoding (Fig. 5cii). The nature of the mutations, often reported in BCR sequencing studies<sup>15</sup>, was also significantly different when using error and amplification bias correction. This was shown both in terms of the lower silent-to-non-silent mutation ratio ( $p$ -value = 0.033, Fig. 5biii) and the locations of the mutations: namely a higher proportion of mutations occurring in the CDRs compared to the FWRs ( $p$ -value = 0.00059, Fig. 5civ). The latter is in agreement with previous studies where mutations are known to preferentially occur in the CDRs compared to the FWRs<sup>16,17</sup>.

Furthermore, the PCR amplification is known to preferentially amplify shorter amplicons. The CDR3 is the most variable region of the BCR sequence, driven in part by the combinations of different IGHV-D-J regions that are recombined during B cell maturation (Fig. 5ci). Indeed, longer CDR3 lengths (longer than ~20 amino acids) are associated with both auto- and poly-reactivity and are often interrogated in BCR repertoire studies<sup>18</sup>. The mean CDR3 region length can be determined from the BCR sequencing data, and, indeed, significantly increased mean CDR3 lengths were observed using amplification-bias-correction via sUMI-seq compared to when error correction was not used (Fig. 5cv).

Together, this data suggests that the sUMI-seq barcoded approach represents a closer representation of the “ground truth” of the BCR repertoire compared to the non-barcoded repertoires. This demonstrates that the estimation of diversity, mutation and amplicon lengths of a mixed DNA pool are all significantly improved by sUMI-seq compared to conventional non-barcoded methods.

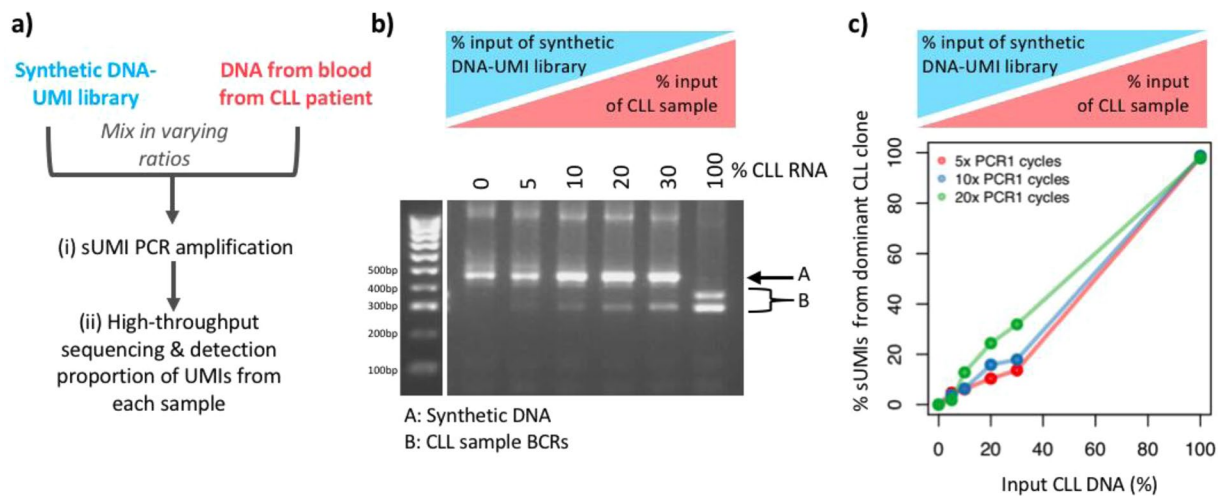
In summary, the sUMI-seq strategy allows for ultrasensitive barcoding of PCR amplicons from DNA for high-throughput sequencing, benefiting from significantly reduced PCR and sequencing errors and amplification biases leading to more accurate characterization of mixed DNA samples. We applied this method to immune receptor repertoire (BCR) profiling, where sUMI-seq captured both highly diverse and highly clonal B cell repertoires from healthy and CLL patients, respectively. sUMI-seq allowed for more accurate estimation of diversity, mutation and amplicon lengths, which are key analyses in many studies of mixed DNA variant pools. sUMI-seq can be easily applied to any PCR amplicons, and benefits from simplicity of primer design, straightforwardness of the amplification protocol with few steps, and streamlined method for incorporating sample barcodes in the second PCR. We have demonstrated the utility and power of this method in the characterization of complex immune receptor repertoire profiles, and this may be applied to a wide range of other applications in which characterizing DNA variants may be obscured by amplification bias or sequence error.

## Materials and Methods

**Samples.** Peripheral blood mononuclear cells (PBMCs) were isolated from 10 mL of whole blood from healthy volunteers and CLL patients using Ficoll gradients (GE Healthcare). Total RNA was isolated using TRIzol and purified using RNeasy Mini Kit (Qiagen), including on-column DNase digestion according to the manufacturer’s instructions. Ethical approval for this study was obtained from the Eastern NHS Multi Research Ethics Committee (07/MRE05/44). Informed consent was obtained from all subjects enrolled and all experiments were performed in accordance with relevant guidelines and regulations.

**Design and amplification with barcoded primers.** Gene specific sUMI-seq primers were designed according to Fig. S8.

*sUMI-seq PCR 1 amplification with barcoded MALBAC primers.* PCR1 was performed using 15  $\mu$ L KAPA buffer (2 $\times$ ) (KAPA HIFI Hotstart PCR kit, Kapa Biosystems), 1  $\mu$ L MALBAC IgH V (FR3) forward primer mix (10  $\mu$ M) (containing 7 family specific primers designed to target the FR3 regions of VH1 through VH7 variable gene families) and 1  $\mu$ L MALBAC reverse primer JH\_ (10  $\mu$ M) (consensus sequence), 8  $\mu$ L nuclease-free water, 5  $\mu$ L DNA template (20 ng/ $\mu$ L), in a final volume of 30  $\mu$ L. sUMI-seq primer sequences that amplify the BCR repertoire are provided in Table S1. The synthetic DNA library (UMI\_DNA) was designed to be amplified with the same primer sets. The thermal cycling conditions for sUMI-seq PCR 1 were as follows: 1 cycle (95  $^{\circ}$ C–5 min); 5 cycles (98  $^{\circ}$ C–5 sec; 72  $^{\circ}$ C–2 min); 5 cycles (65  $^{\circ}$ C–10 sec, 72  $^{\circ}$ C–2 min); 5, 10, 20, or 30 cycles (98  $^{\circ}$ C–20 sec, 60  $^{\circ}$ C–1 min, 72  $^{\circ}$ C–2 min); 1 step (72  $^{\circ}$ C–10 min). PCR1 amplicons were then cleaned-up using 0.8x Agencourt AMPure XP beads (Beckman Coulter).



**Figure 3.** Titration of DNA pools using sUMI-seq. PCR (a) Schematic of titration experiment: To test the quantitative nature of sUMI-seq, a dilution series was performed of a peripheral blood (PB) DNA sample, from a chronic lymphocytic leukaemia (CLL) patient, mixed with the synthetic barcoded BCR DNA fragment library at varying ratios. sUMI-seq PCR was applied, again using either 5, 10, or 20 cycles in PCR1. (b) DNA agarose gel showing the BCR amplification of a titration of a CLL patient PB DNA sample into synthetic barcoded BCR fragments using 5 cycles in PCR1 and (c) the corresponding linear correlation between CLL DNA patient input and proportion of CLL BCR sequencing reads after accounting for primer barcodes based on PCR1 cycles.

*sUMI-seq PCR 2 amplification (without sample IDs).* The PCR2 reaction was performed using 17.5  $\mu$ L of KAPA buffer (2 $\times$ ) (KAPA HIFI Hotstart PCR kit, Kapa Biosystems), 1  $\mu$ L of 10  $\mu$ M IgH V (FR3) forward primer mix and 1  $\mu$ L of 10  $\mu$ M MALBAC\_UNI primers, 5.5  $\mu$ L of nuclease-free water, 10  $\mu$ L of DNA template (from PCR1), in a final volume of 35  $\mu$ L. The thermal cycling conditions were as follows: 1 cycle (95  $^{\circ}$ C–5 min); 5 cycles (98  $^{\circ}$ C–5 sec; 72  $^{\circ}$ C–2 min); 5 cycles (65  $^{\circ}$ C–10 sec, 72  $^{\circ}$ C–2 min); 20 cycles (98  $^{\circ}$ C–20 sec, 60  $^{\circ}$ C–1 min, 72  $^{\circ}$ C–2 min); 1 step (72  $^{\circ}$ C–10 min).

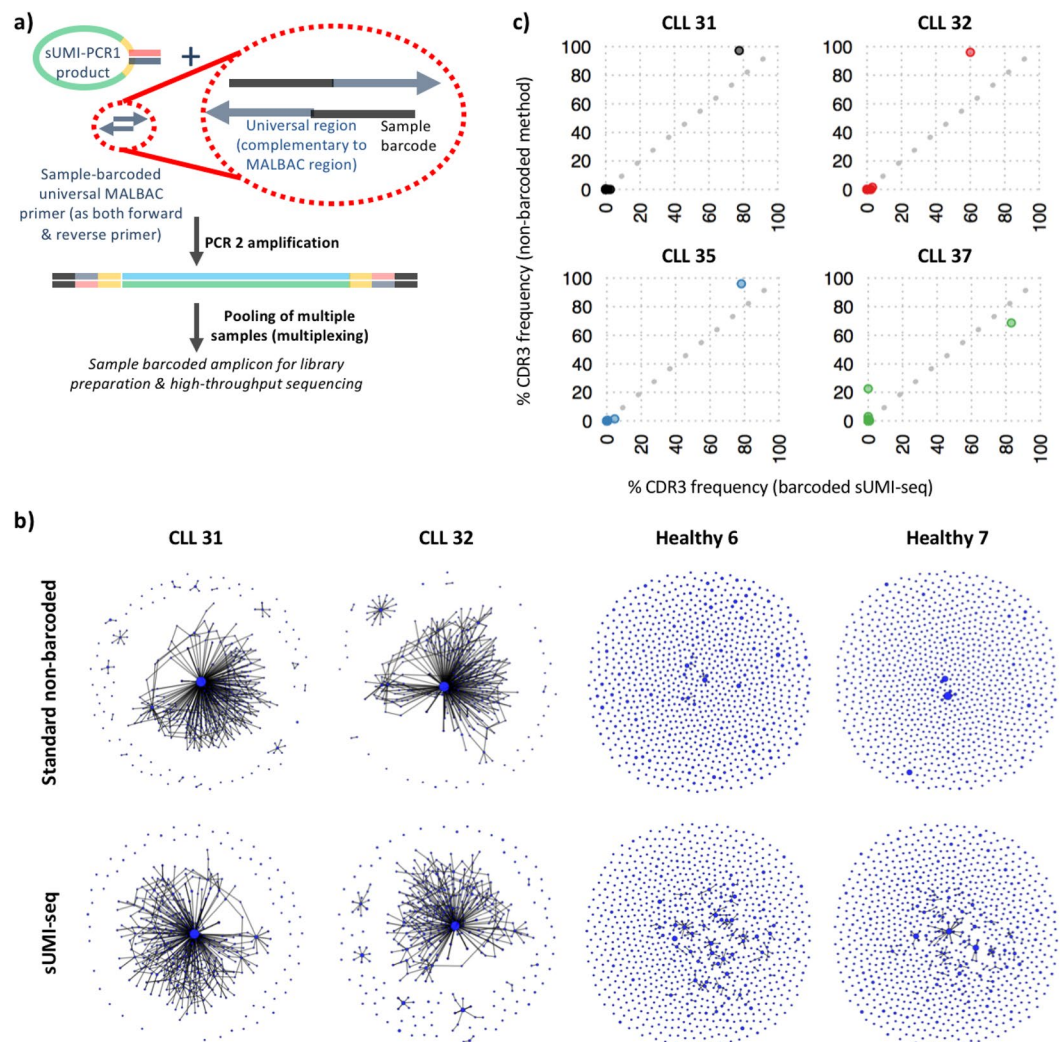
*sUMI-seq PCR 2 amplification (with sample barcode IDs).* The PCR2 reaction was performed using 17.5  $\mu$ L KAPA buffer (2 $\times$ ) (KAPA HIFI Hotstart PCR kit, Kapa Biosystems), 1  $\mu$ L MALBAC IgH V (FR3) forward primer mix (10  $\mu$ M), and 1  $\mu$ L MALBAC\_UNI\_Ind primer (10  $\mu$ M) (choice of 1–12 barcodes) (Table S1), 5.5  $\mu$ L nuclease-free water, 10  $\mu$ L DNA template (from PCR1), for a total volume of 35  $\mu$ L. The thermal cycling conditions were as follows: 1 cycle (95  $^{\circ}$ C–5 min); 5 cycles (98  $^{\circ}$ C–5 sec; 72  $^{\circ}$ C–2 min); 5 cycles (65  $^{\circ}$ C–10 sec, 72  $^{\circ}$ C–2 min); 20 cycles (98  $^{\circ}$ C–20 sec, 60  $^{\circ}$ C–1 min, 72  $^{\circ}$ C–2 min); 1 step (72  $^{\circ}$ C–10 min).

**Standard non-barcoded PCR amplification.** This was performed using 15  $\mu$ L KAPA buffer (2 $\times$ ) (KAPA HIFI Hotstart PCR kit, Kapa Biosystems), 1  $\mu$ L IgH V (FR3) forward primer mix (10  $\mu$ M) (standard non-barcoded primers), and 1  $\mu$ L reverse IgH-J (10  $\mu$ M) (standard primers), 8  $\mu$ L nuclease-free water, 5  $\mu$ L DNA template (20 ng/ $\mu$ L), for a total volume of 30  $\mu$ L. The thermal cycling conditions were as follows: 1 cycle (95  $^{\circ}$ C–5 min); 5 cycles (98  $^{\circ}$ C–5 sec; 72  $^{\circ}$ C–2 min); 5 cycles (65  $^{\circ}$ C–10 sec, 72  $^{\circ}$ C–2 min); 5, 10, or 20 cycles (98  $^{\circ}$ C–20 sec, 60  $^{\circ}$ C–1 min, 72  $^{\circ}$ C–2 min); 1 step (72  $^{\circ}$ C–10 min).

**High-throughput sequencing and QC.** PCR2 DNA amplicons were cleaned-up using 0.8x Agencourt AMPure XP beads (bead-based size selection) and checked using electrophoresis on a 2% agarose gel. MiSeq libraries were prepared using KAPA protocols (KK8722 and KK8504) and sequenced using 300 bp pair-end MiSeq (Illumina). Raw MiSeq reads were filtered for base quality (median Phred score >32) using QUASR (<http://sourceforge.net/projects/quasr/>)<sup>3</sup>.

MiSeq forward and reverse reads were merged together if they contained an identical overlapping region of >50 bp, or otherwise discarded.

**For the sUMI-seq filtering pipeline.** Universal barcoded regions were identified in reads and orientated to read from forward (IgH V)-primer to reverse (IgH-J) region primer. The barcoded region within each primer was identified and checked for conserved bases. **Error-correction and amplification bias correction:** Groups of sequencing reads containing the same sUMI-seq primer UMIs originate from the same DNA template, and therefore a consensus sequence was generated from these groups. This reduces amplification biases (i.e. the effect of differential amplification of DNA templates), as well as correcting potential PCR/sequencing errors. Consensus sequences were retained only if there was a per-base agreement of 80% between all sequencing reads containing the same UMI. For groups of 4 or fewer sequencing reads containing the same UMI, there needed to be complete agreement between sequences after alignment, otherwise were discarded. This is summarised in Fig. S5.

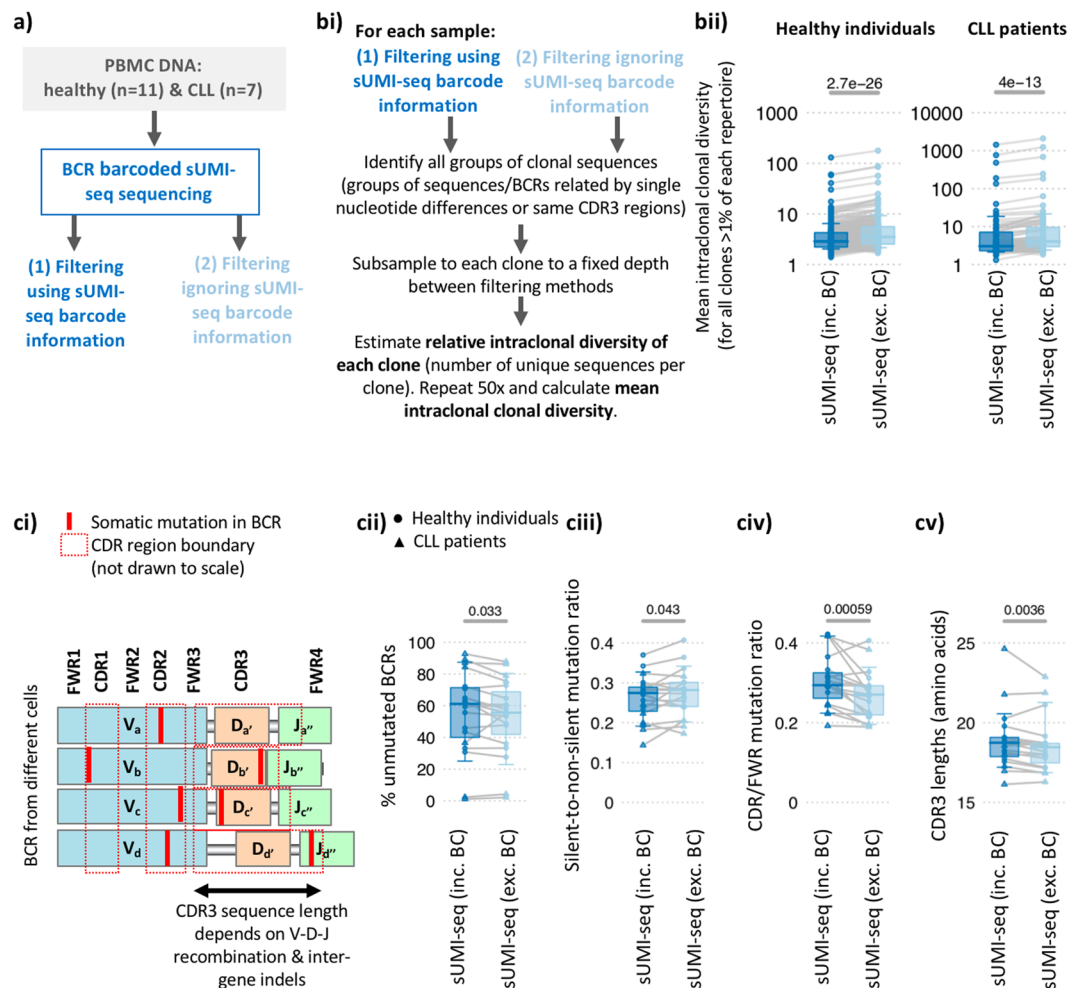


**Figure 4.** BCR repertoire sequencing by sUMI-seq. PCR (a) Schematic diagram of multiplex (sample-barcode) sUMI-seq. used in PCR2 (b) Representative network plots of 2 chronic lymphocytic leukaemia (CLL) and 2 healthy individual BCR repertoires derived from PBMC DNA amplified by (top) the standard non-barcode BCR amplification approach and (bottom) by sUMI-seq. Each vertex represents a unique BCR sequence (B cell clone), where relative vertex size is proportional to the number of identical BCR reads. Edges join vertices that differ by single nucleotide non-indel differences and clusters are collections of related, connected vertices. Networks are comprised of a subsample of 500 BCRs per sample. (c) Correlation of the frequencies of each B cell clone (defined by the CDR3 sequence) within the CLL patient samples derived from the sUMI-seq method versus the standard PCR method. The grey dotted line corresponds to  $y = x$ , and each point corresponds to a different CDR3 (B cell clone) sequence frequency.

**For the standard filtering pipeline.** The primer regions within the sequencing reads were determined. All sequences without identifiable primer annealing regions were discarded.

**Quantifying the frequency of further amplification of that PCR1 self-annealing MALBAC loops products.** For each sequence within the synthetic UMI-DNA datasets, the synthetic DNA-UMIs and primer barcode pairs were identified. From this, the proportion of sequences which contained DNA-UMIs associated with more than one primer barcode pairs was determined, and normalised to the total number of reads (provided as a rate per 1000 reads):  $1000 * \left( \frac{\text{number of identical synthetic DNA - UMIs} * \text{number of sUMI - seq primer barcode pairs}}{\text{number of BC - MALBAC primer pairs} / \text{number of reads}} \right)$

**Quantifying amplification bias using the synthetic DNA-UMIs library and comparing sUMI PCR to standard PCR amplification.** The relative amplification biases were compared between (1) the sUMI-seq method using sUMI-seq barcode information (using the sUMI-seq filtering pipeline), (2) the sUMI-seq method ignoring the sUMI-seq barcode information (using the standard filtering pipeline), and (3) using the standard non-barcode PCR amplification method and the standard filtering pipeline. To account for differences in read depths between the different methods, each filtered dataset was subsampled to the same read



**Figure 5.** Comparison of BCR repertoire features between sUMI-seq PCR, filtering with and without barcode information **(a)** Schematic diagram of the comparison between sUMI-seq (filtering using barcode information) and sUMI-seq (filtering ignoring barcode information). **(b)** Quantifying the mean intraclonal diversity: the relative intraclonal diversity of all clones >1% of the total repertoire in each sample was compared between the two methods. To account for the differences in read depth, the BCRs from each clone were subsampled to a fixed depth between filtering methods (0.75x the minimum number of sequences per clone across methods). The relative intraclonal diversity of each clone was defined as number of unique BCR sequences per clone after subsampling, and the mean intraclonal relative intraclonal diversity was determined through calculating the mean of 50 repeats. **(ii)** Boxplots of the mean intraclonal diversity between methods for all healthy (left) and CLL (right) patient samples, plotted on a logarithmic scale. Grey lines connect the mean relative clonal diversity measurements between methods. **(c)** The schematic demonstrates the relative locations of the CDRs and FWRs within the BCR sequences (IgH VDJ). Boxplots quantify the differences in commonly used BCR repertoire features between methods including **(cii)** the proportion of unmutated BCRs (i.e. BCRs with no somatic hypermutations), **(ciii)** the ratio of silent-to-non-silent mutations, **(civ)** the CDR/FWR mutation ratio and **(cv)** CDR3 lengths (amino acids). The y axis provides the mean value per sample for each BCR repertoire feature, and grey lines the measurements between methods. Wilcoxon tests were performed in *R*.

depth across all samples (3000 reads), defined as the maximum number of reads containing the same synthetic DNA-UMI:

$$\text{amplification bias (per subsample)} = \max(n_i)$$

where  $n_i$  is number of reads containing synthetic DNA – UMI  $i$ . The mean relative amplification biases was calculated from 500 repeats per sample. Wilcoxon tests were performed in *R*.

**BCR sequence filtering.** Sequences without complete reading frames and non-immunoglobulin sequences were removed and only reads with significant similarity to reference IgH variable genes (V-D-J) from the IMGT database were retained using BLAST<sup>19</sup>. Sequence annotation, including somatic hypermutation, CDR3 regions and IGHV gene usages, were defined via IMGT V-QUEST, where repertoire differences were performed by custom scripts in Python, and statistics were performed in *R* using Wilcoxon tests for significance.



**BCR repertoire generation and network analysis.** The network generation algorithm and network properties were calculated as in Bashford-Rogers *et al.*<sup>5</sup>: each vertex represents a unique sequence, where relative vertex size is proportional to the number of identical reads. Edges join vertices that differ by single nucleotide non-indel differences and clusters are collections of related, connected vertices.

A **clone (cluster)** refers to clonally-related B cells, containing BCRs with identical CDR3 regions and IgH gene usage, or differing by single point mutations, such as through somatic hypermutation.

**Clonality diversity** refers to the relative number of clonally-related, but distinct, B cells within a clone. In the context of BCR sequencing, this is a measure of the number of unique clonally-related BCRs (clone members). Sequence repertoire parameters that were dependent on sequencing depth were generated by subsampling each sequencing sample to a specified clone depth. This includes the Clonal Diversification index, was measured by cluster Renyi Index as defined in Bashford-Rogers *et al.*<sup>5</sup>. This is calculated from the distribution of the number of unique VDJ region sequences per clone within subsampled BCR repertoires at specified depth of 1000 clones. The mean of 100 repeats of resulting Clonal Diversification indices was determined. Clone size distributions were also calculated from the same subsamples and a mean of 100 repeats was determined.

**BCR network sampling to preserve the overall clonal structure of visual representation.** To obtain representative subgraph of a network that preserves the overall relative clonal architecture whilst providing visual representations that distinguish between samples of different clonalities, clone subsampling was used as described in<sup>15</sup>. One thousand clones are subsampled and a network generated from all BCRs from these clones. Subsampling was performed 100 times, and the sample that contained a maximum clone size closest to the median of all subsamples greater than the unsampled maximum clone size was chosen.

**Ethics approval and consent to participate.** Ethical approval for this study was obtained from the Eastern NHS Multi Research Ethics Committee (07/MRE05/44). Informed consent was obtained from all subjects enrolled.

### Data availability

Code is made available at [https://github.com/rbr1/sUMI\\_processing\\_pipeline](https://github.com/rbr1/sUMI_processing_pipeline). The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. All sequencing data will be uploaded to the EGA.

Received: 8 November 2019; Accepted: 1 June 2020;

Published online: 29 June 2020

### References

- McKerrell, T. *et al.* Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep* **10**, 1239–1245, <https://doi.org/10.1016/j.celrep.2015.02.005> (2015).
- Human Microbiome Project, C. A framework for human microbiome research. *Nature* **486**, 215–221, <https://doi.org/10.1038/nature11209> (2012).
- Watson, S. J. *et al.* Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120205, <https://doi.org/10.1098/rstb.2012.0205> (2013).
- Wei, L. *et al.* Genome-wide CRISPR/Cas9 library screening identified PHGDH as a critical driver for Sorafenib resistance in HCC. *Nat Commun* **10**, 4681, <https://doi.org/10.1038/s41467-019-12606-7> (2019).
- Bashford-Rogers, R. J. M. *et al.* Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* **23**, 1874–1884, <https://doi.org/10.1101/gr.154815.113> (2013).
- Petrova, V. N. *et al.* Combined Influence of B-Cell Receptor Rearrangement and Somatic Hypermutation on B-Cell Class-Switch Fate in Health and in Chronic Lymphocytic Leukemia. *Frontiers in Immunology* **9**, <https://doi.org/10.3389/fimmu.2018.01784> (2018).
- Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626, <https://doi.org/10.1126/science.1229164> (2012).
- Galson, J. D. *et al.* In-Depth Assessment of Within-Individual and Inter-Individual Variation in the B Cell Receptor Repertoire. *Front Immunol* **6**, 531, <https://doi.org/10.3389/fimmu.2015.00531> (2015).
- Bashford-Rogers, R. J. M., Smith, K. G. C. & Thomas, D. C. Antibody repertoire analysis in polygenic autoimmune diseases. *Immunology*, <https://doi.org/10.1111/imm.12927> (2018).
- Bashford-Rogers, R. J. M. *et al.* Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia*, <https://doi.org/10.1038/leu.2016.142> (2016).
- Bashford-Rogers, R. J. M. *et al.* Dynamic variation of CD5 surface expression levels within individual chronic lymphocytic leukaemia clones. *Exp Hematol*, <https://doi.org/10.1016/j.exphem.2016.09.010> (2016).
- Galson, J. D. *et al.* Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine* **2**, 2070–2079, <https://doi.org/10.1016/j.ebiom.2015.11.034> (2015).
- Ma, L. *et al.* Characteristics Peripheral Blood IgG and IgM Heavy Chain Complementarity Determining Region 3 Repertoire before and after Immunization with Recombinant HBV Vaccine. *PLoS One* **12**, e0170479, <https://doi.org/10.1371/journal.pone.0170479> (2017).
- Bashford-Rogers, R. J. *et al.* Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol* **15**, 29, <https://doi.org/10.1186/s12865-014-0029-0> (2014).
- Bashford-Rogers, R. J. M. *et al.* Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*, <https://doi.org/10.1038/s41586-019-1595-3> (2019).
- Lin, M. M., Zhu, M. & Scharff, M. D. Sequence dependent hypermutation of the immunoglobulin heavy chain in cultured B cells. *Proc Natl Acad Sci USA* **94**, 5284–5289 (1997).
- Yaari, G. *et al.* Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* **4**, 358, <https://doi.org/10.3389/fimmu.2013.00358> (2013).
- Meffre, E. *et al.* Immunoglobulin heavy chain expression shapes the B cell receptor repertoire in human B cell development. *J Clin Invest* **108**, 879–886, <https://doi.org/10.1172/JCI13051> (2001).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410, [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (1990).

## Acknowledgements

This work was supported by the Wellcome Trust. We would like to thank the Cambridge Cancer Trials Centre and nurse specialists Gwyn Stafford, Rosie Tween, Lisa Walbridge and Joanna Baxter, and the patients and staff of Addenbrooke's Haematology Translational Research Laboratory. This work was supported by the Wellcome Trust (grant WT106068AIA) and the Amgen Foundation.

## Author contributions

R.J.M.B.-R. and K.G.C.S. planned the study. R.J.M.B.-R., I. A., A. A. and F.A.T. performed BCR amplification and R.J.M.B.-R. analysed sequencing data. All authors provided intellectual contributions to experiments and/or analyses. R.J.M.B.-R. wrote the manuscript. All authors edited the manuscript.

## Competing interests

R.J.M.B.-R. is a co-founder and consultant for Alchemab Therapeutics Ltd and a consultant for Imperial College London and VHSquared. F.A.T. is a consultant for Alchemab Therapeutics Ltd. K.G.C.S. is a co-founder of Rheos Medicines and PredictImmune. I. A. and A. A. declare no conflicts of interest.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-67290-1>.

**Correspondence** and requests for materials should be addressed to R.J.M.B.-R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020