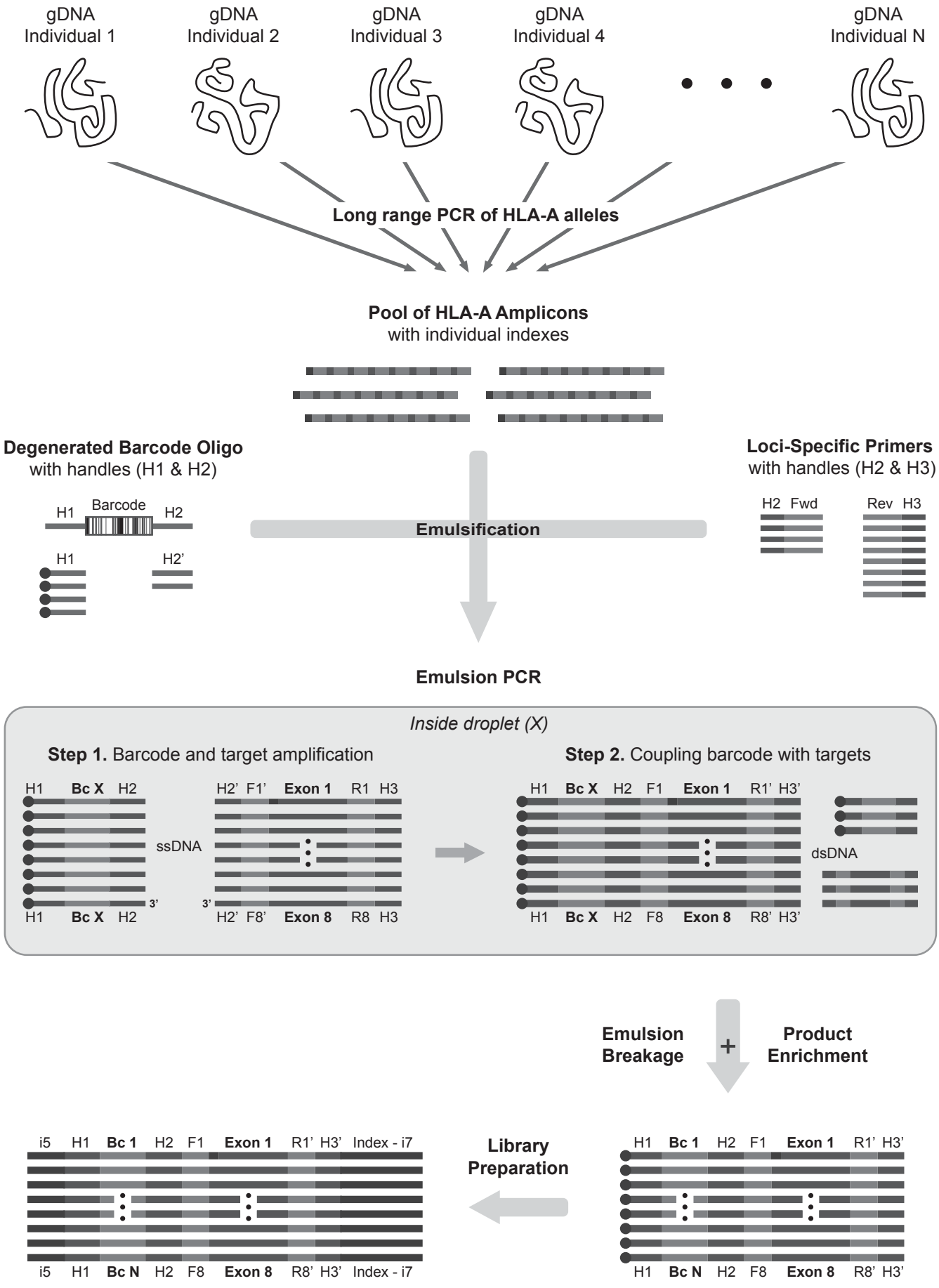
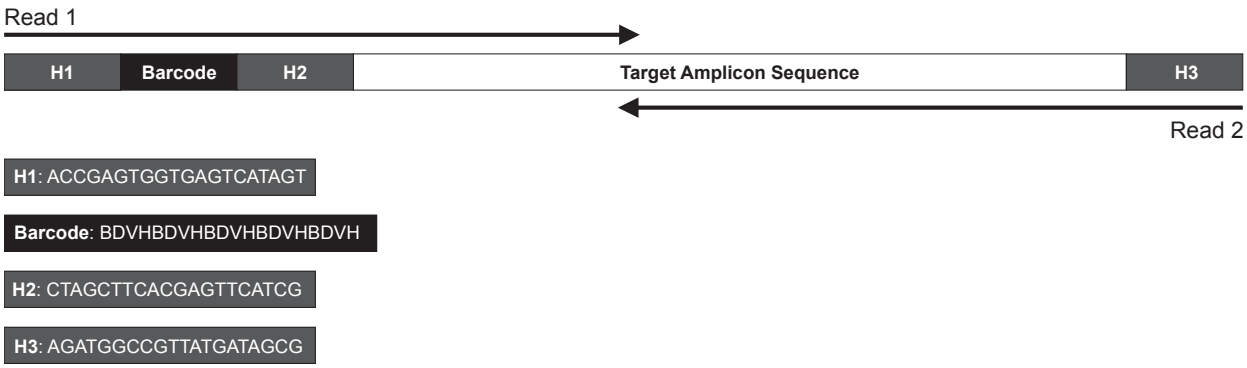
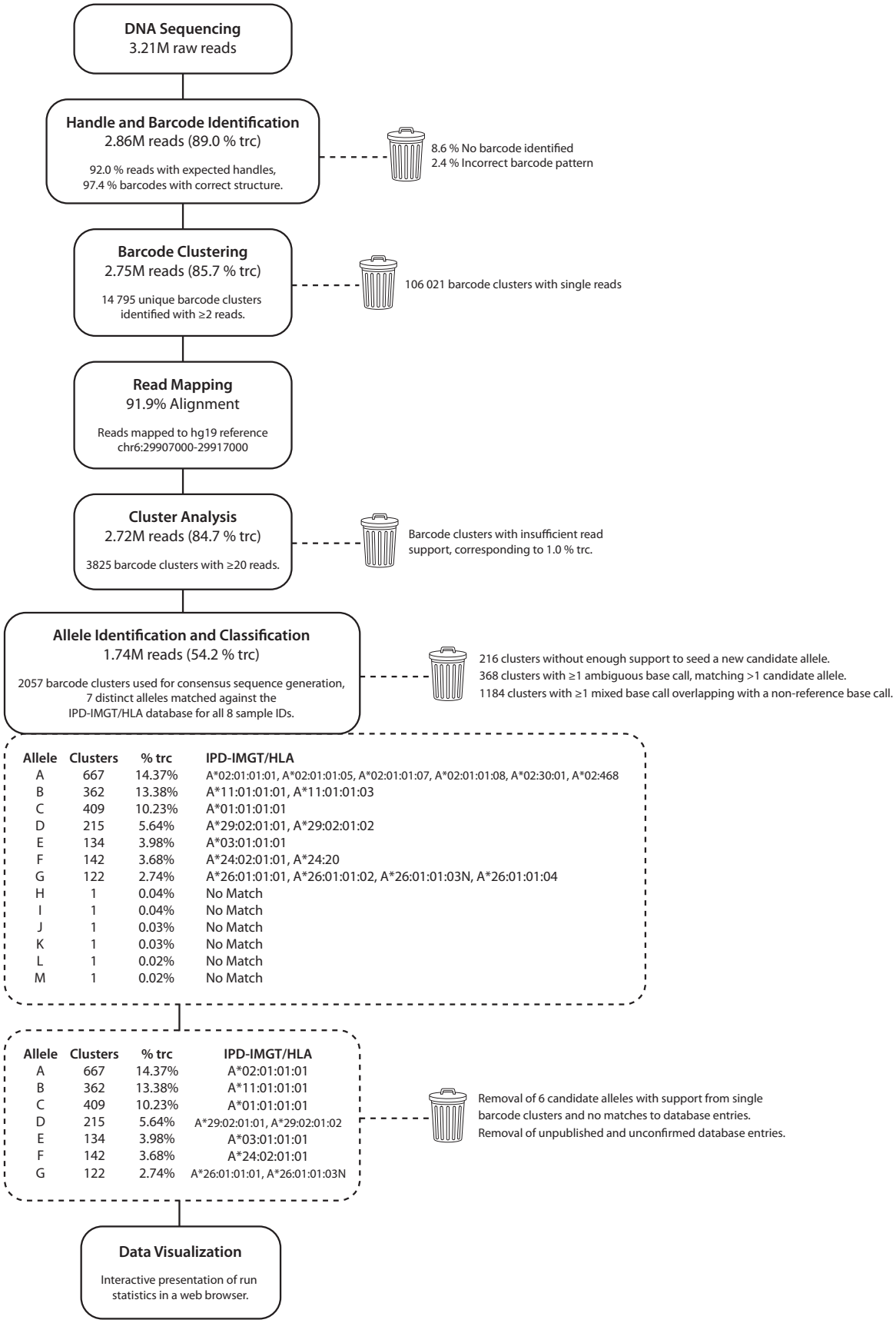

gDNA Individual 1
gDNA Individual 2
gDNA Individual 3
gDNA Individual 4
gDNA Individual N
Long range PCR of HLA-A alleles
Pool of HLA-A Amplicons
with individual indexes
Degenerated Barcode Oligo
with handles (H1 & H2)
H1 Barcode H2
H1 H2'
Emulsification
Loci-Specific Primers
with handles (H2 & H3)
H2 Fwd
Rev H3
Emulsion PCR
Inside droplet (X)
Step 1. Barcode and target amplification
H1 Bc X H2
ssDNA
H2' F1' Exon 1 R1 H3
3'
H1 Bc X H2
3'
H2' F8' Exon 8 R8 H3
Step 2. Coupling barcode with targets
H1 Bc X H2 F1 Exon 1 R1' H3'
dsDNA
H1 Bc X H2 F8 Exon 8 R8' H3'
Emulsion Breakage
+
Product Enrichment
H1 Bc 1 H2 F1 Exon 1 R1' H3'
H1 Bc N H2 F8 Exon 8 R8' H3'
Library Preparation
i5 H1 Bc 1 H2 F1 Exon 1 R1' H3' Index - i7
i5 H1 Bc N H2 F8 Exon 8 R8' H3' Index - i7

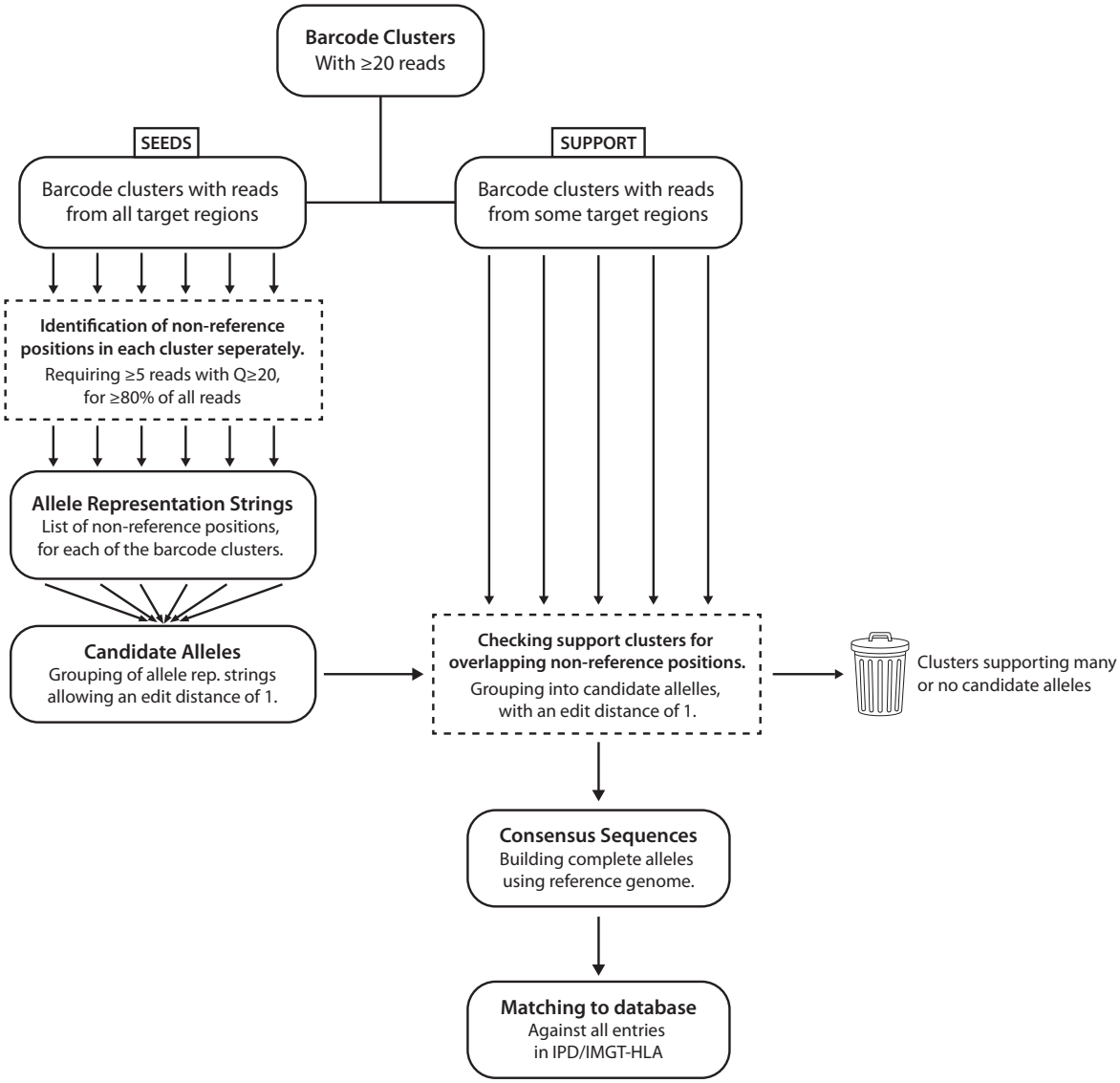
Supplementary Figure 1. Detailed assay overview. First, a long range PCR with sample-specific indexed primers (Supplementary Table 1 and 2) is performed to amplify the target HLA-A gene. An equimolar pool of HLA-A long range amplicons is then prepared and diluted to enable compartmentalization of single molecules by emulsification. Along with a single template molecule, a number of synthetic oligonucleotides (Supplementary Table 3) are incorporated into the emulsion droplets at the same time. Like the pool of template molecules, barcoding oligonucleotides are diluted to be present in single copies in the droplets to yield a unique (albeit unknown) barcode population in each droplet following amplification. Short synthetic oligonucleotides enable amplification of both the barcoding oligonucleotide and seven different target loci of the template molecule. The amplification reaction is asymmetric for both sets of templates (i.e. barcodes and sample templates), resulting in a shift in the equilibrium of the polymerase chain reaction towards single stranded DNA products. This is achieved by having an excess of oligo Bio-H1 in respect to oligo H2' (at a ratio of 8:1), as well as an excess of H3-R1 – H3-R8 oligos in respect to $H2_{10}$-F1 – $H2_{10}$-F8 oligos (at a ratio of 5:1; see Supplementary Table 3). The loci-specific primers were designed to target intronic regions flanking each of the 8 exons of the HLA-A gene. Once a critical concentration of single stranded products has been generated, and the oligonucleotides in limited supply have been depleted, the two PCR products can interact be means of a 10 bp complementary sequence (Emulsion PCR - Step 2). This coupling reaction occurs at a lower annealing temperature to account for the short sequence of complementarity (Supplementary Table 4). When complete, coupling of a clonally amplified barcode to each of the target loci is thereby achieved, yielding double stranded DNA products. Following emulsion breakage, products are then enriched by means of size exclusion to get rid of the excess uncoupled barcoding molecules. The biotinylated H1 oligonucleotide enables a second enrichment step of coupled products to streptavidin-coated beads, to get rid of similarly sized uncoupled loci-specific PCR products. Finally, following enrichment the bead-bound products are released by an indexing PCR to attach sequencing adaptors. The finished library is then sequenced and data is analysed using a custom-built analysis pipeline (see Materials & Methods).



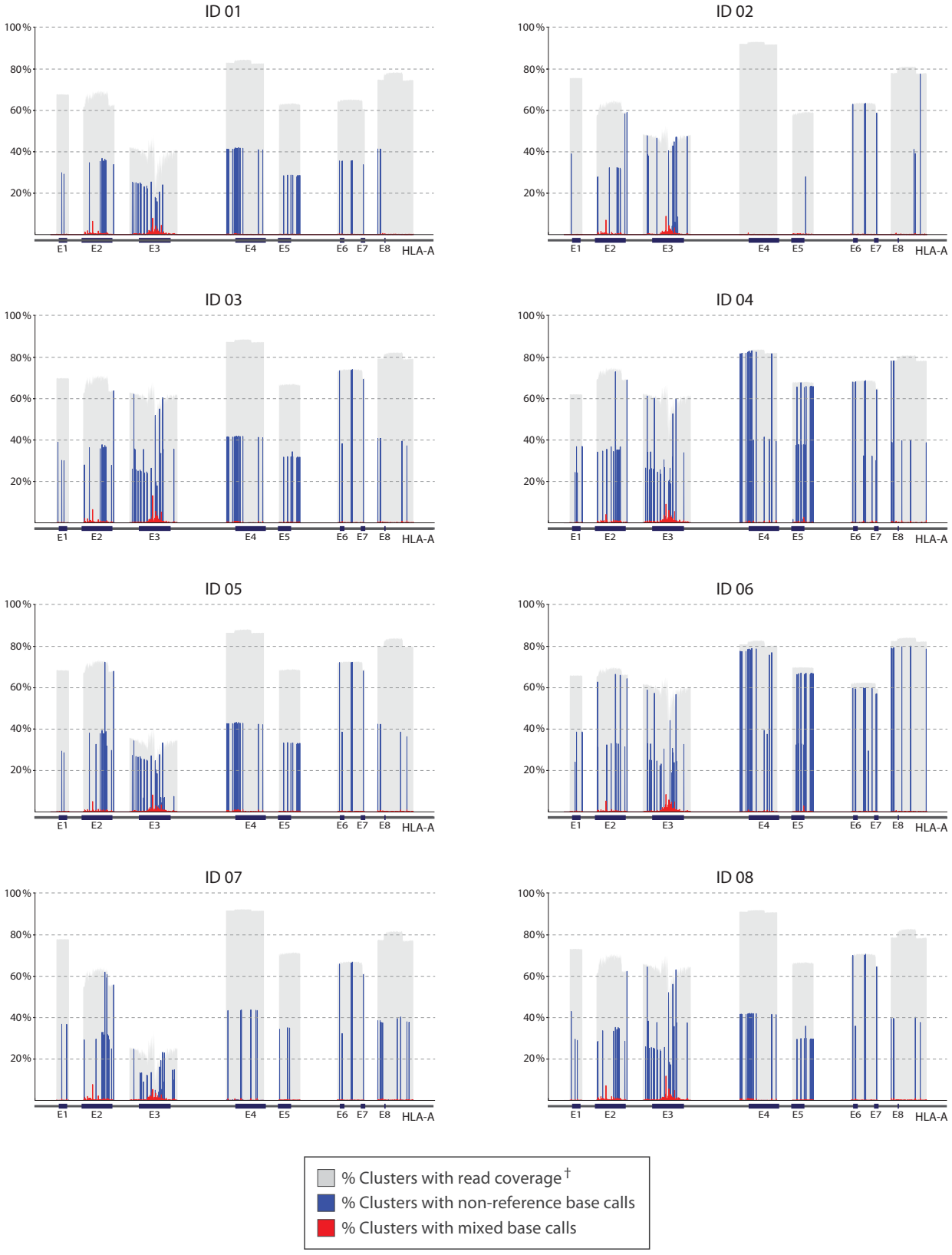
Supplementary Figure 2. Structure of sequencing reads and handles.



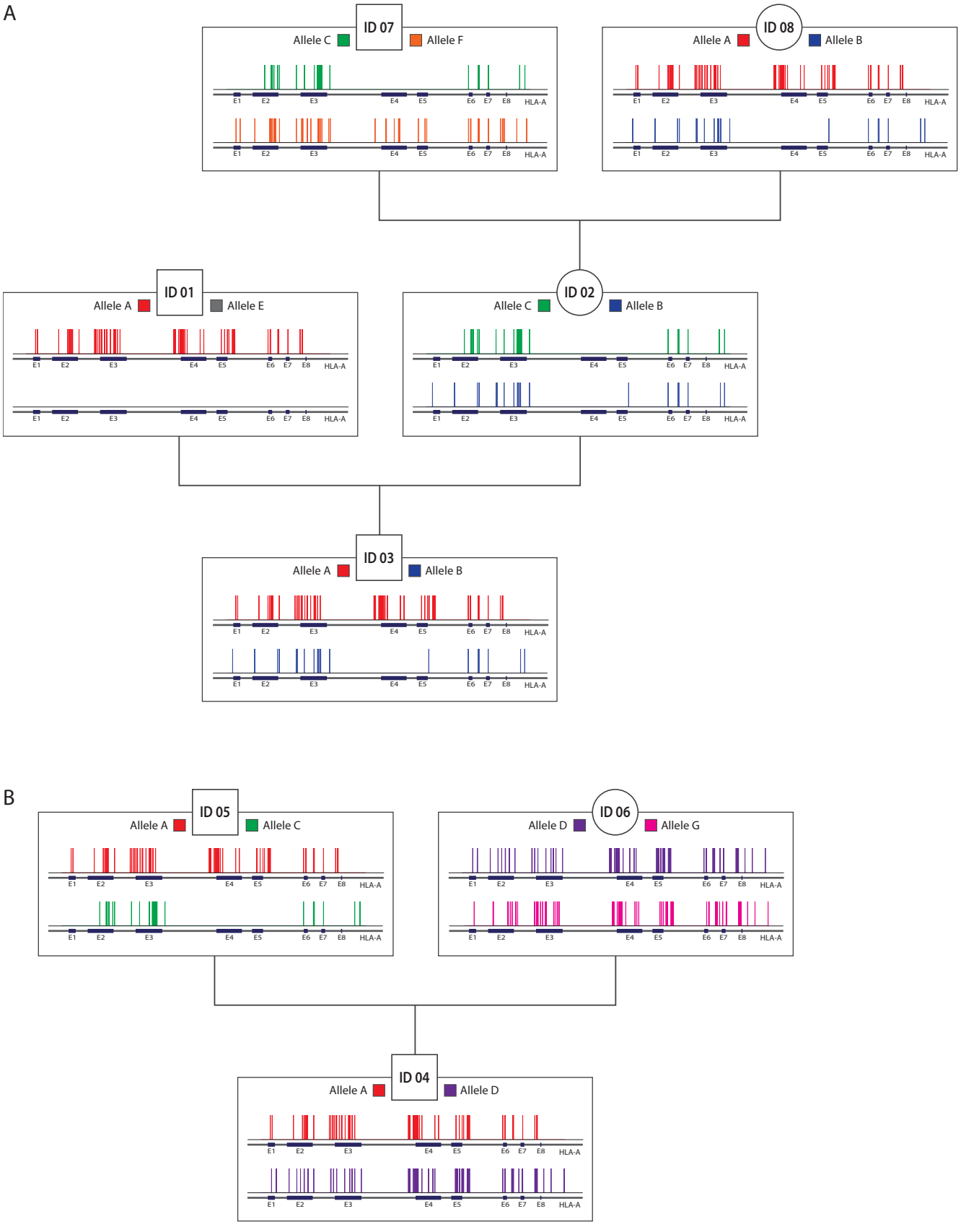
Supplementary Figure 3. Flowchart of data analysis steps, detailing read counts and barcode cluster counts for the 8-plex reaction. % trc corresponds to percentage of the total read count (i.e. raw reads).



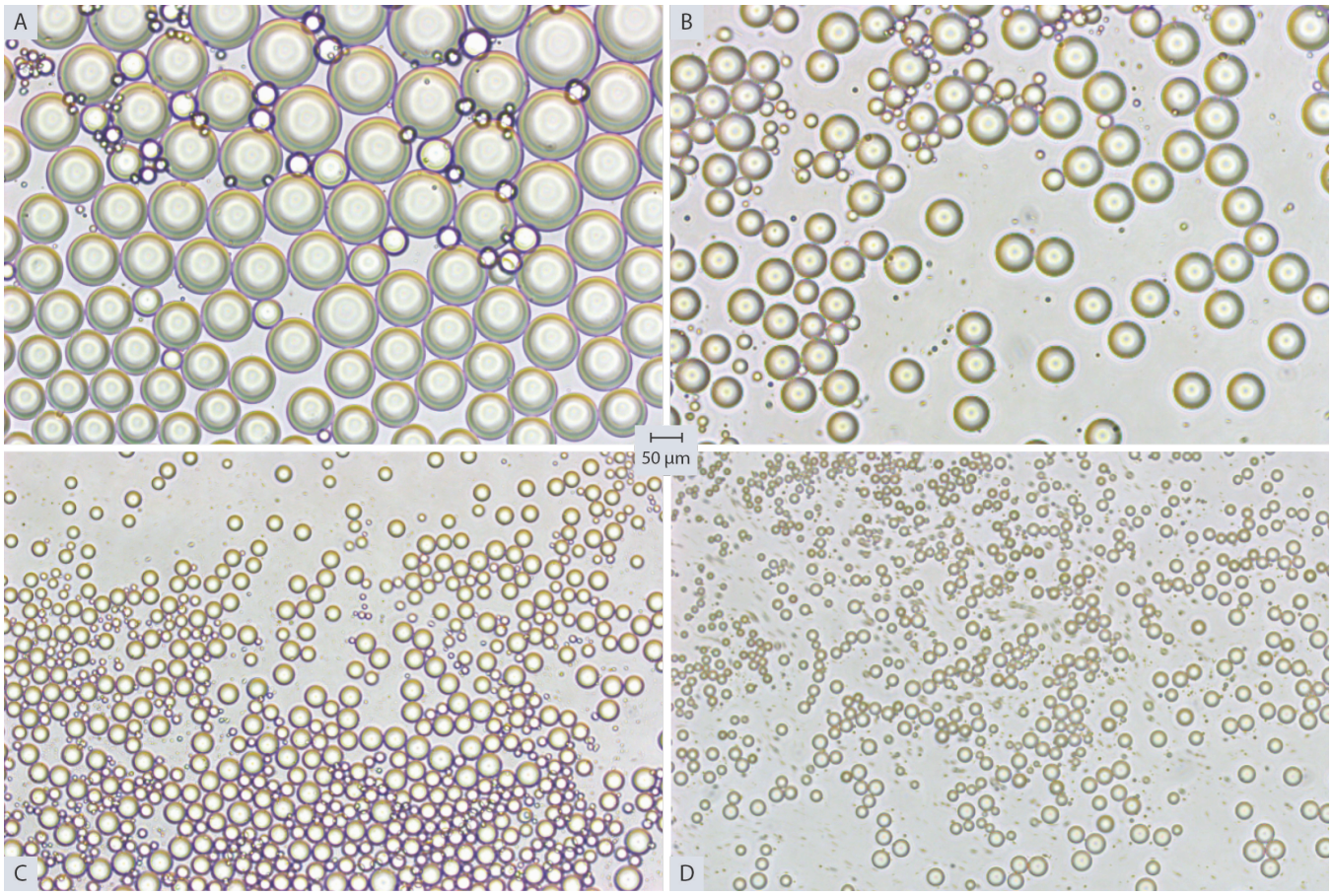
Supplementary Figure 4. Flowchart detailing the procedure for building and identifying alleles from barcode clusters with at least 20 reads. It is worth noting that this bioinformatic solution does not distinguish between supporting barcode clusters that are truly non-clonal clusters (i.e. those that stem from multiple input molecules), and those with an error profile that exceeds our cutoff of allowing (at the most) an edit distance of 1. This filtering of supporting barcode clusters could most likely be made less stringent, but considering the purpose of this study is to perfectly call all variant alleles we did not deem it necessary.



Supplementary Figure 5. Haplotyping results visualized for all sample ID singleplex reactions, with data from all clusters with ≥20 reads (i.e. before filtering of barcode clusters not contributing to the formation of consensus sequences, as depicted in Supplementary Figure 4). † Base calling of each position supported by ≥5 reads with a quality score ≥20. Dips in the sequencing coverage for the middle of the target covering exon 3 is due to this region not being covered by read 1, in combination with a high sequencing error profile at the end of read 2.



Supplementary Figure 6. Pedigree visualization of allele heredity for (A) extended family of Trio 1 and 2, and (B) Trio 3; showing positions of all non-reference base calls across the HLA-A gene, for each allele identified within the 8 samples.



Supplementary Figure 7. Emulsion droplets generated by controlled oscillation at set frequencies; (A) 11 Hz, (B) 13 Hz, (C) 15 Hz, and (D) 17 Hz. For reasons detailed in Supplementary Note 1, emulsions generated at 15 Hz were deemed the best our assay reactions.

Supplementary Table 1. HLA-A Long range PCR indexing of Coriell gDNA samples

Sample ID	ID Index Sequence	Long range PCR Forward primer	Coriell gDNA Individual	Coriell Trio (Family relation)	Sex
01	ACAGTC	LR.F01	NA12877	Trio 2: Parent	Male
02	TGATGC	LR.F02	NA12878	Trio 1: Child Trio 2: Parent	Female
03	GTCGAT	LR.F03	NA12882	Trio 2: Child	Male
04	TGAGTG	LR.F04	NA10860	Trio 3: Child	Male
05	GTACTG	LR.F05	NA11992	Trio 3: Parent	Male
06	TAGCTG	LR.F06	NA11993	Trio 3: Parent	Female
07	AGTCTG	LR.F07	NA12891	Trio 1: Parent	Male
08	TCACGT	LR.F08	NA12892	Trio 1: Parent	Female

Supplementary Table 2. Synthetic oligonucleotides for long range PCR and library preparation.
Note: Index sequence for i7'-H3 oligo were variable TruSeq LT indexes for Illumina's sequencing platform.

Name	Sequence (5' - 3')
LR.F01	TTAGTCTCCGACGGCAGGCTTCAATACAGTCACGCACCCACCGGGACTCAG
LR.F02	TTAGTCTCCGACGGCAGGCTTCAATTGATGCACGCACCCACCGGGACTCAG
LR.F03	TTAGTCTCCGACGGCAGGCTTCAATGTCGATACGCACCCACCGGGACTCAG
LR.F04	TTAGTCTCCGACGGCAGGCTTCAATTGAGTGACGCACCCACCGGGACTCAG
LR.F05	TTAGTCTCCGACGGCAGGCTTCAATGTACTGACGCACCCACCGGGACTCAG
LR.F06	TTAGTCTCCGACGGCAGGCTTCAATTAGCTGACGCACCCACCGGGACTCAG
LR.F07	TTAGTCTCCGACGGCAGGCTTCAATAGTCTGACGCACCCACCGGGACTCAG
LR.F08	TTAGTCTCCGACGGCAGGCTTCAATTCACGTACGCACCCACCGGGACTCAG
LR.R	ATATAACCATCATCGTGTCCCAAGGTTC
H1	ACCGAGTGGTGAGTCATAGT
H3	AGATGGCCGTTATGATAGCG
i5-H1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCTACCGAGTGGTGAGTCATAGT
i7'-H3	CAAGCAGAAGACGGCATACGAGAT-INDEX-ACGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTAGATGGCCGTTATGATAGCG

Supplementary Table 3. Sequence specification and composition of oligonucleotides in emulsion reactions.

Name	Sequence (5' - 3')	Reaction Concentration
Bio-H1	Biotin-ACCGAGTGGTGAGTCATAGT	400 nM
Barcoding Oligo	ACCGAGTGGTGAGTCATAGTBDVHBDVHBDVHBDVHBDVH CTAGCTTCACGAGTTCATCG	100 fM
H2'	CGATGAACTCGTGAAGCTAG	50 nM
$H2_{10}$-F1	GAGTTCATCGTTAGTCTCCGACGGCAGGCTTCAAT	40 nM
$H2_{10}$-F2	GAGTTCATCGCCCAGGCTCYCACTCCATGA	40 nM
$H2_{10}$-F3	GAGTTCATCGGCGCCTTWACCCGGTTTCATT	50 nM
$H2_{10}$-F4	GAGTTCATCGCCCGGGTGTCCTGTCCATTC	30 nM
$H2_{10}$-F5	GAGTTCATCGCCCAGAGCYRTCTTCCCAGC	20 nM
$H2_{10}$-F6:7	GAGTTCATCGGCCCTGYCTCCTTTCTGGTATC	40 nM
$H2_{10}$-F8	GAGTTCATCGTGAGCATGCGATGGGCTGTT	40 nM
H3-R1	AGATGGCCGTTATGATAGCGTCCCCGCAGAGGCGGTTTCC	200 nM
H3-R2	AGATGGCCGTTATGATAGCGGGGATGAGRGGTCSTGACCT	200 nM
H3-R3	AGATGGCCGTTATGATAGCGTGTTGGTCCCAATTGTCTCCC	250 nM
H3-R4	AGATGGCCGTTATGATAGCGTACCCCATCTCAGGGTGAGG	150 nM
H3-R5	AGATGGCCGTTATGATAGCGTTGTGGAHGGTGCTTCCCAG	100 nM
H3-R6:7	AGATGGCCGTTATGATAGCGGGYCCTSCARGCTCTCACCTT	200 nM
H3-R8	AGATGGCCGTTATGATAGCGCAGCTCAGTGCACCATGAAGT	200 nM

Supplementary Table 4. Emulsion PCR protocol.

Step	Cycles	Temperature	Duration	Ramp Speed
1	1	94°C	3 min	40%
2	5	94°C	1 min	40%
2	5	45°C	2 min	40%
2	5	72°C	2 min	20%
3	21	94°C	1 min	40%
3	21	55°C	2 min	40%
3	21	72°C	2 min	20%
4	12	94°C	1 min	40%
4	12	25°C	1 min	20%
4	12	72°C	2 min	10%
5	1	72°C	5 min	100%
6	1	20°C	1 min	10%
7	1	10°C	HOLD	10%

Supplementary Table 5. Data quantified in terms of generated barcode clusters for each assay reaction.

Sample Reaction	Read pairs with barcode	Total Barcode Clusters	Barcode Clusters (1 read)	Barcode Clusters (≥2 reads)	Barcode Clusters (≥20 reads)	Filtered Barcode Clusters [% tot reads]
01	1 487 060	74 709	64 975	9 734	2 069	1 231 [53.1%]
02	1 320 082	56 289	48 935	7 354	1 488	981 [60.9%]
03	1 359 780	61 071	53 515	7 556	2 097	1 504 [72.9%]
04	1 306 578	57 455	49 738	7 717	1 247	788 [57.3%]
05	1 564 702	77 826	67 713	10 113	1 411	955 [63.5%]
06	1 513 428	88 414	76 525	11 889	1 335	917 [64.1%]
07	1 262 517	57 155	47 904	9 251	2 054	1 491 [63.6%]
08	1 372 462	60 055	52 580	7 475	2 377	1 544 [69.0%]
8-Plex	2 859 364	120 816	106 021	14 795	3825	2057 [54.2%]

Supplementary Note 1. Generating emulsions.

Generating droplets using shaking of two immiscible phases (aqueous and oil) leads to a broad range of droplet in different sizes. But by controlling the conditions (oscillation frequency, time, volumes and reagents) we can tune the size range and obtain these results in a reproducible fashion. Using a Tissuelyser (Qiagen) instrument (typically used to generate cell suspensions from tissues), the oscillation frequency can be set from 3.0 Hz to 30.0 Hz (+/- 0.1 Hz), and for emulsions this offers control of the range of droplet sizes. We have found that the tube volume, aqueous phase volume, and oil phase volume all influence the size distribution due to their effect on forces generated during shaking. To simplify the optimization process we decided to keep the variables of shaking time and volumes constant to study the effect of oscillation frequency on droplet sizes. Supplementary Figure 7 shows a representative view of emulsions generated with four different oscillation frequencies (11 Hz, 13 Hz, 15 Hz and 17 Hz). In such emulsions we can observe a varying range of droplet sizes, as well as the spread of this range. When the oscillation frequency is low (11 Hz) we see droplets ranging from 5 μm (65 fl) to 200 μm (4.2 nl), while when the oscillation frequency is high (17 Hz) we observe a range from 1 μm (0.5 fl) to 30 μm (14.1 pl). We came to the conclusion that emulsions generated with a frequency of 15 Hz were the best as they yielded the highest proportion of droplets within the range of 10-30 μm (motivated by discussion below). For emulsions generated at an oscillation frequency of 15 Hz we estimated the average droplet size to be 15-20 μm, and based assay parameters on this assumption (Supplementary Note 2).

The idea of using emulsions generated by shaking to enable high throughput assays is not novel; it for instance constitutes a fundamental step in the process of preparing sequencing libraries for the 454 and Ion Torrent sequencing platforms. However, most of the recently published and/or commercialized emulsion-based technologies rely on droplets generated using microfluidic devices to obtain a more monodisperse population of reaction compartments. While this reduces the need for optimization of reagent concentrations, there are important drawbacks that limit the spread of this technology. Setting up microfluidics-based droplet generation requires at its minimum a set of very precise pumps and microfluidic chips for consumption. Such items are available from commercial vendors (e.g. Dolomite Microfluidics) but the expense of single-use microfluidic chips means it is not a feasible solution for most labs. Instead, the alternative to produce chips in-house is what most labs working with droplet technologies have opted for, but the manufacture of high quality devices requires expertise not readily available.

Regardless if droplets are generated on-chip or by shaking, the most important aspect of emulsion technologies is the stability of droplets, in particular when exposed to thermal cycling (e.g. PCR). Droplet stability is typically improved by adding an amphiphilic chemical surfactant that sits in the interface between water and oil, stabilizing the droplets to prevent coalescence (droplet merging). Very few surfactants are good enough to allow emulsion reactions to undergo long PCR reactions; we have found the commercially available 008-FluoroSurfactant (RAN Biotechnologies) to be the best. Other factors that can influence droplet stability are the carrier oil, the composition of the aqueous phase, and the size the generated droplets. Large droplets are typically less stable than smaller droplets since they have a higher surface area (with stabilizing surfactant molecules) in relation to their volume. Our observations show that droplets up to ~50 μm can be PCR cycled up to 35 cycles without significant merging. Although stable through PCR, small droplets (<10 μm) will statistically be less likely to contain all the necessary reagents for the assay, so in order to ensure assay efficiency the size distribution of droplets has been optimized.

Supplementary Note 2. Evaluating monoclonality based on emulsion parameters.

Unique droplet barcoding in this assay is based on the dilution of barcoding oligonucleotides and template molecules to obtain a single copy of each molecule in emulsion droplets. Random distribution of molecules is modeled by the Poisson distribution where an occupancy of, for example 0.1 molecules per droplet (m.p.d.), will result in ~90% empty droplets and ~95% of the occupied droplets will be monoclonal (i.e. only contain one molecule). Having two reagents that each ought to be present in single copies effectively doubles the effect of the Poisson distribution. Two sets of independent molecules, each loaded with 0.1 m.p.d., would thus yield roughly 0.095*0.095 ≈ 0.9% of droplets containing one of each molecule.

Determining the number of active droplets in emulsions generated by shaking is no trivial matter, but our observations of emulsions generated at 15 Hz suggest that we can estimate a conservative average of droplet size of 15 – 20 um in diameter (Supplementary Note 1). While this seems like a small range, there is roughly a two-fold difference in the average volume (1.8 – 4.2 pl droplets in a 50 ul reaction); yielding an estimated range that spans from 12 million to 28 million droplets total in our assay. Given that we have approximately 3 million barcoding oligonucleotides and 2 million HLA template molecules (in the 8-plex reaction), we ought to expect the following theoretical yields of droplet barcodes:

$$Barcode\ m.p.d.\ \times\ Template\ m.p.d.\ \ = \%\ Droplets\ with\ both\ barcode\ and\ template$$

$$[15\ \mu m\ average] \leftrightarrow \sim 28 \cdot 10^6\ droplets$$

$$\frac{3 \cdot 10^6\ barcodes}{28 \cdot 10^6\ droplets} \times \frac{2 \cdot 10^6\ templates}{28 \cdot 10^6\ droplets} = 0.76\%\ Occupancy \rightarrow 215\,000\ droplet\ barcodes$$

$$[20\ \mu m\ average] \leftrightarrow \sim 12 \cdot 10^6\ droplets$$

$$\frac{3 \cdot 10^6\ barcodes}{12 \cdot 10^6\ droplets} \times \frac{2 \cdot 10^6\ templates}{12 \cdot 10^6\ droplets} = 4.17\%\ Occupancy \rightarrow 500\,000\ droplet\ barcodes$$

Our results show that we have identified 120 816 barcode clusters, corresponding to an estimated efficiency of 24% – 56%. It is worth noting that the sequencing depth heavily influences this measure of efficiency, and the limited 3.2M reads generated for the 8-plex sample is likely to be detrimental to this estimate. A large proportion of barcode clusters identified only had one read pair, which was then excluded from downstream data analysis steps (Supplementary Figure 3).

Given an estimated number of droplets we can calculate theoretical rate of monoclonality for the template molecules in our assay, according to the Poisson distribution we can expect between 92% and 96% monoclonal droplets (11). Counting the number of barcode clusters in which more than one different HLA template is present can experimentally validate this figure. We found that 112 out of 3825 clusters with ≥20 reads contained reads from multiple sample ID-tags, corresponding to a doublet rate of 2.93% of clusters and 2.51% of the total read population. It should be acknowledged however, that this measure is based on the assumption that the targets covering exon 1 has enough read coverage for both template molecules, which may not always be the case. Unfortunately there is no way of identifying the frequency of droplets with more than one barcoding molecule, but given the number of input molecules the rate of monoclonality should be within the same range as that of template molecules.