BCR network sampling to preserve the overall clonal structure

We aimed to obtain a graphical representation of an individual's BCR repertoire (or network) that preserves the overall relative clonal architecture of the samples while allowing visualization of differences in clonality between samples. We assessed three methods of network sampling to generate representative visualisations of the overall BCR repertoire, namely (a) Read sampling, (b) Cluster-enforced linkage sampling (CC), and (c) clone subsampling.

Let $G = (V, E)$ represent the graph, where V is the set of nodes and E is the set of edges in the graph. Each edge $e \in E$ can be described as a tuple of the form (vi,vj) where $vi,vj \in V$. Given a sampling fraction φ, the goal is to create a sample graph $Gs = (Vs, Es)$ such that $|Vs|/|V| = \varphi$, that preserves the structure of the original network. Here we quantify the overall structure of the original network as the maximum cluster size (the number of vertices in the largest cluster divided by the total number of vertices), which provides representative visualisation of clonal connectivity and diversification of the overall network.

Sampling methods:

a) Read sampling

A fixed number of unique BCRs was subsampled and a network generated from these BCRs. Subsampling was performed 20 times, and from these, the sample that contained a maximum clone size closest to the true maximum clone size was chosen.

Pros:

- Fast and computationally efficient
- Truly random sampling

Cons

- Loss of connectivity between nodes within the same original cluster leads to significant differences in network structure compared to the original unsampled network, particularly when the sample size is much lower than the original network size (i.e. when $|Vs|/|V|$ is small).

b) Cluster-enforced linkage sampling (CC)

Most graph sampling algorithms have two basic components: (1) vertex selection, and (2) induced graph formation. The CC algorithm employs a third step to account for loss of connectivity between vertices in clusters due to lack of sampling, as well as loss of highly connected vertices that provide information on the manner of B-cell clonal diversification (Extended Data Figure 5a):

(1) Vertex selection

Vertices are randomly sampled until the number of desired clusters in the original network G are represented.

(2) Cluster-vertex migration

For each cluster in the original network which contains more than one vertex that was sampled, vertices were re-selected such that the cluster connectivity was retained in the sampled network. This was performed by:

- Let *NVsc* denote the number of sampled vertices corresponding to original network cluster *c*, and *Vc* denote all the vertices in the original network cluster *c*. Select the vertex in *Vc* with the highest connectivity (degree) as the first re-selected vertex. This vertex will have the highest importance to the network structure.
- Then iteratively randomly select vertices that are within 1 edge of this that are already selected. Repeat this process until *NVsc* vertices are selected.

Replace the originally sampled vertices from each cluster with these re-selected vertices.

(3) Induced graph formation

Graph induction selects the set of edges (Es) to be included in the sampled graph. Total graph induction is used in CC, selecting all edges incident on the sampled vertices are included in the sampled graph.

This process was repeated 20 times, and the subsample that most closely represented the true (unsampled) maximum cluster size was retained and plotted.

Pros:

- Given that this algorithm enforces connectivity between all sampled vertices within the same original cluster, then the expectation value of sampled cluster size for cluster *i*, *E(Csi),* will equal the proportion of vertices in the original cluster, , thus faithfully representing the original network (i.e. sampled cluster sizes approximate true cluster sizes even at low (sample depths i.e. when $|Vs|/|V|$ is small).

Cons

- The nodes of greatest connectivity have a higher probability of sampling, thus biasing the plotted network.
- When $|Vs|/|V|$ is small, the cluster sizes are determined by the number of clusters that can be plotted within a single visual representation. Between 2000-5000 clones has been found to be optimal for visual representation BCR networks allowing the identification of isotype/node colour and observing network edges. However, this method does not easily visually distinguish between networks of low clonality (maximum clone sizes <5%).

c) Clone sampling

A fixed number of clones were 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.

Pros:

- This algorithm captures the full diversity of sampled clones in the visual representation including SHM and CSR.
- Maximum clone sizes are highly correlated with true maximum clone sizes. In addition, for the samples with low maximum clone sizes (<5%), differences between samples can be more easily visually distinguished due to accentuation of larger clones.

Cons

- More subsamples are required to identify a network that best represents the original sample.

How each sampling method performs

All samples were subjected to all three sampling methods to a depth of 2000 clusters, and the subsampled maximum cluster sizes were determined (Extended Data Figure 5b). The *read sampling method* significantly underestimated the large clone sizes due to loss of connectivity through network node reduction, the *clones sampling method* strongly correlates with true maximum clone size with over-representation of the larger clones, whereas the *CC sampling method* faithfully represents the maximum clone sizes. Given that visualisation of more than 2000 clones leads to overcrowding of individual nodes, the CC sampling method would not clearly distinguish between samples (Extended Data Figure 5c), whereas the *clones sampling method* best visually discriminated most clearly between the clonalities of the non-CLL samples.

Therefore, we visualised BCR repertoire networks using the *clones sampling method* in the manuscript, as this visually discriminated most clearly between the clonalities of the non-CLL samples, whilst faithfully correlating with the overall clonality of the total repertoire.