SCIENTIFIC REPORTS
nature research

OPEN

Quantifying Hematopoietic Stem Cell Clonal Diversity by Selecting Informative Amplicon Barcodes

Emily M. Teets[1,3], Charles Gregory[1,3], Jami Shaffer[1], James S. Blachly[1,2,4] & Bradley W. Blaser[1,4*]

Hematopoietic stem cells (HSCs) are functionally and genetically diverse and this diversity decreases with age and disease. Numerous systems have been developed to quantify HSC diversity by genetic barcoding, but no framework has been established to empirically validate barcode sequences. Here we have developed an analytical framework, Selection of informative Amplicon Barcodes from Experimental Replicates (SABER), that identifies barcodes that are unique among a large set of experimental replicates. Amplicon barcodes were sequenced from the blood of 56 adult zebrafish divided into training and validation sets. Informative barcodes were identified and samples with a high fraction of informative barcodes were chosen by bootstrapping. There were 4.2 ± 1.8 barcoded HSC clones per sample in the training set and 3.5 ± 2.1 in the validation set (p = 0.3). SABER reproducibly quantifies functional HSCs and can accommodate a wide range of experimental group sizes. Future large-scale studies aiming to understand the mechanisms of HSC clonal evolution will benefit from this new approach to identifying informative amplicon barcodes.

Hematopoietic stem cells (HSCs) are increasingly recognized to be functionally and genetically heterogeneous[1]. HSC clonal diversity has important implications for the study of hematopoiesis and blood disorders[2–6]. Methods for estimating HSC clonal diversity include counting viral integration sites in transduced bone marrow[7], single-cell or limiting dilution transplantation[8–11], sequencing transposon insertion sites[12], SNP analysis in genomic[13,14] or mtDNA[15], and genetic barcoding using CRISPR/Cas9 or Cre-Lox-based recombination[16–24]. An ideal method would sample a large fraction of the HSCs in an organism, be able to discriminate many clones with little or no ambiguity, and would mark HSC clones without causing any alteration in cell function. Further, the mark should be induced prior to any experimental intervention, be inherited by all progeny of the HSC, and be detectable using reproducible and cost-efficient means with a minimum investment of labor and computational time. Optimizing these parameters would allow researchers to perform more powerful experiments to address mechanisms of hematopoiesis and leukemogenesis.

The transgenic zebrafish system, Genome Editing of Synthetic Target Arrays for Lineage Tracing (GESTALT), has emerged as a powerful tool for studying cellular phylogeny[16,25]. GESTALT zebrafish carry a single germline copy of a synthetic array consisting of 10 tandem CRISPR/Cas9 target sites. By microinjecting synthetic guide RNAs (sgRNAs) targeting this array with either Cas9 mRNA or recombinant Cas9 protein into the single-cell zebrafish embryo, double strand breaks are induced within the array and then repaired by non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ). The combinatorial effect of editing the 10 sites of the array induces thousands of unique genetic barcodes[16]. Barcoding ceases when all available sites have been edited or when cell division or degradation has reduced the effective concentration of editing reagents below a critical threshold. For experiments using recombinant Cas9 protein, this is estimated to be between 4–5 hours post-fertilization (hpf)[16]. An inducible genetic system has been developed which extends the editing window beyond this time[25,26]. The genetic barcoding can be used to trace cell phylogeny and in theory could be combined with conditional transgenics, mutants, or other genetic or chemical modifications to understand how these experimental conditions affect clonal diversity of the blood system.

[1]The Ohio State University College of Medicine, Department of Medicine, Division of Hematology, The Ohio State University Comprehensive Cancer Center, Ohio, USA. [2]The Ohio State University College of Medicine, Department of Biomedical Informatics, Ohio, USA. [3]These authors contributed equally: Emily M. Teets and Charles Gregory. [4]These authors jointly supervised this work: James S. Blachly and Bradley W. Blaser. *email: bradley.blaser@osumc.edu

There are a number of limitations to the GESTALT method. The fraction of barcoded cells in the GESTALT zebrafish depends on the integrity and quantity of the reagents used and the efficiency with which the injection solution is delivered to the embryo. The NHEJ and MMEJ repair mechanisms can produce stereotypical repair patterns, reducing the actual diversity of barcodes observed. The published bioinformatic analyses do not provide a means to identify these uninformative variants or a systematic approach to exclude samples with a low fraction of informative barcodes. Unique molecular identifiers (UMIs) have been used for sequencing error, PCR error, and PCR bias correction[27], however the UMI PCR protocol more than doubles the sample preparation time and reagent cost, and the extent to which UMIs improve accuracy over standard PCR has not been demonstrated in this setting.

Here we have developed a sequencing and informatics pipeline optimized for the quantification of amplicon barcodes from genomic DNA which we entitle Selection of Amplicon Barcodes from Experimental Replicates (SABER). By analyzing a large number of zebrafish blood samples, we were able to define a threshold for discriminating informative GESTALT barcodes from uninformative variants and then optimize this threshold through iteration and modeling. Using bootstrap analysis, we provide a method to rationally exclude samples with a low fraction of informative barcodes. We also show that the results of a standard PCR protocol are nearly indistinguishable from a UMI-based PCR protocol. We believe that this experimental method, conceptual framework and reference dataset will be useful to laboratories studying cellular phylogeny, clonal diversity and clonal evolution using amplicon barcodes.

Results

Analytical pipeline, sample generation and sequencing. SABER is divided into three components: (1) core functions for processing sequence data, aligning to reference sequence and calling variants, (2) an optional module to handle UMI-based PCR amplicons and (3) functions to identify informative barcodes, select samples and perform statistical analysis (Fig. 1a, Supplementary Fig. S1).

To generate samples for this analysis, hemizygous GESTALT transgenic embryos were injected at the single cell stage with an injection mix containing pooled GESTALT sgRNAs and EnGen Spy Cas9 NLS enzyme (NEB) (Fig. 1b). Injected zebrafish were grown to 3 months post-fertilization (mpf) and peripheral blood was collected via retro-orbital (RO) venipuncture. GESTALT barcodes were amplified from genomic DNA via standard or UMI-based PCR. A representative gel with 5 such samples is shown in Fig. 1c. Unfragmented amplicons were sequenced using Illumina-based technology (2×250 bp reads, Amplicon EZ, Genewiz) and data was delivered as separate FASTQ files for reads 1 and 2. In the entire dataset, an average of $64{,}022 \pm 13{,}499$ (mean $\pm$ S.D.) paired-end reads were obtained per sample and $51{,}502 \pm 14{,}587$ were aligned to the reference sequence after filtering. Supplementary Fig. S2 summarizes the sequencing metrics for the entire dataset.

For this analysis, 56 peripheral blood samples from 3 independent experiments were evenly divided into training and validation sets of 28 samples each. The founder of this line was previously shown to harbor a single insertion of the GESTALT transgene[16]. Each clutch of fish was derived from an outcross of a single GESTALT homozygote and a *casper* zebrafish.

Standard PCR and UMI-based PCR produce similar GESTALT variant allele frequencies. UMI-based PCR techniques reduce sequencing error by generating consensus alleles from reads with identical UMIs and mitigate PCR bias by reducing read groups with the same UMI to a single deduplicated read[27]. This protocol incorporates a series of annealing/extension steps to tag single genomic DNA template molecules with the UMI barcode. Genomic DNA from a subset of 5 samples was processed using the UMI-based PCR protocol. In parallel, a standard PCR (i.e. without the preceding annealing/extension steps), was performed on the same samples using the same UMI-tagged forward primer and reverse primer. All 10 samples (5 UMI PCR and 5 standard PCR) were processed using the core pipeline with UMI deduplication in the UMI PCR samples. Variant allele frequencies for the uncut GESTALT sequence plus the top 20 variants are shown for a representative sample (Fig. 2a). Linear regression analysis showed a high degree of correlation between VAFs derived from standard and UMI-based PCR techniques (average adjusted R^2: 0.998, $N = 5$ samples, representative sample with $R^2 = 0.9998$ shown in Fig. 2b). In the context of this system, we conclude that UMI-based PCR techniques do not significantly improve the accuracy of variant quantification and so have used the standard PCR protocol and analysis for the subsequent analysis.

Identification of informative barcodes from GESTALT variants. In cellular phylogenetic terms, a GESTALT sequence variant can be considered an informative genetic barcode if it is unique to the clade of cells descended from the ancestral cell in which the barcoding was performed. The theoretical number of unique barcodes in the GESTALT system is much larger than the number of cells in the embryo at 4–5 hpf, when barcoding likely is complete[16]. However in practice the diversity of barcodes is limited by stereotypical repair patterns arising from the NHEJ and MMEJ double strand break-repair mechanisms[28]. InDelphi was used to predict GESTALT variants with 1% or greater likelihood of occurrence after Cas9 cutting and repair at each target site in isolation[29]. Between 11 and 22 GESTALT variants were predicted across the 10 target sites (Supplementary Fig. S3). However, some variants were predicted to be strongly favored over others, with the most likely predicted variant at each locus having a frequency of between 15.4% and 40.0%. The Shannon entropy calculation for the effective number of barcodes at each locus ranged from 2.5 to 3.3, corresponding to 72,304 potential GESTALT barcodes[30]. This is far lower than the 5.0×10^{12} barcodes predicted had each individual variant been equally likely. Barcode diversity is limited further by GESTALT sequences that are never edited and large deletions removing one or more adjacent targets. For two cells from the same organism that carry the same GESTALT sequence, it is impossible to know if that sequence was generated in their last common ancestor cell (meaning that the cells were correctly assigned to the same clade) or if the GESTALT sequence was generated independently in separate clades. The latter GESTALT

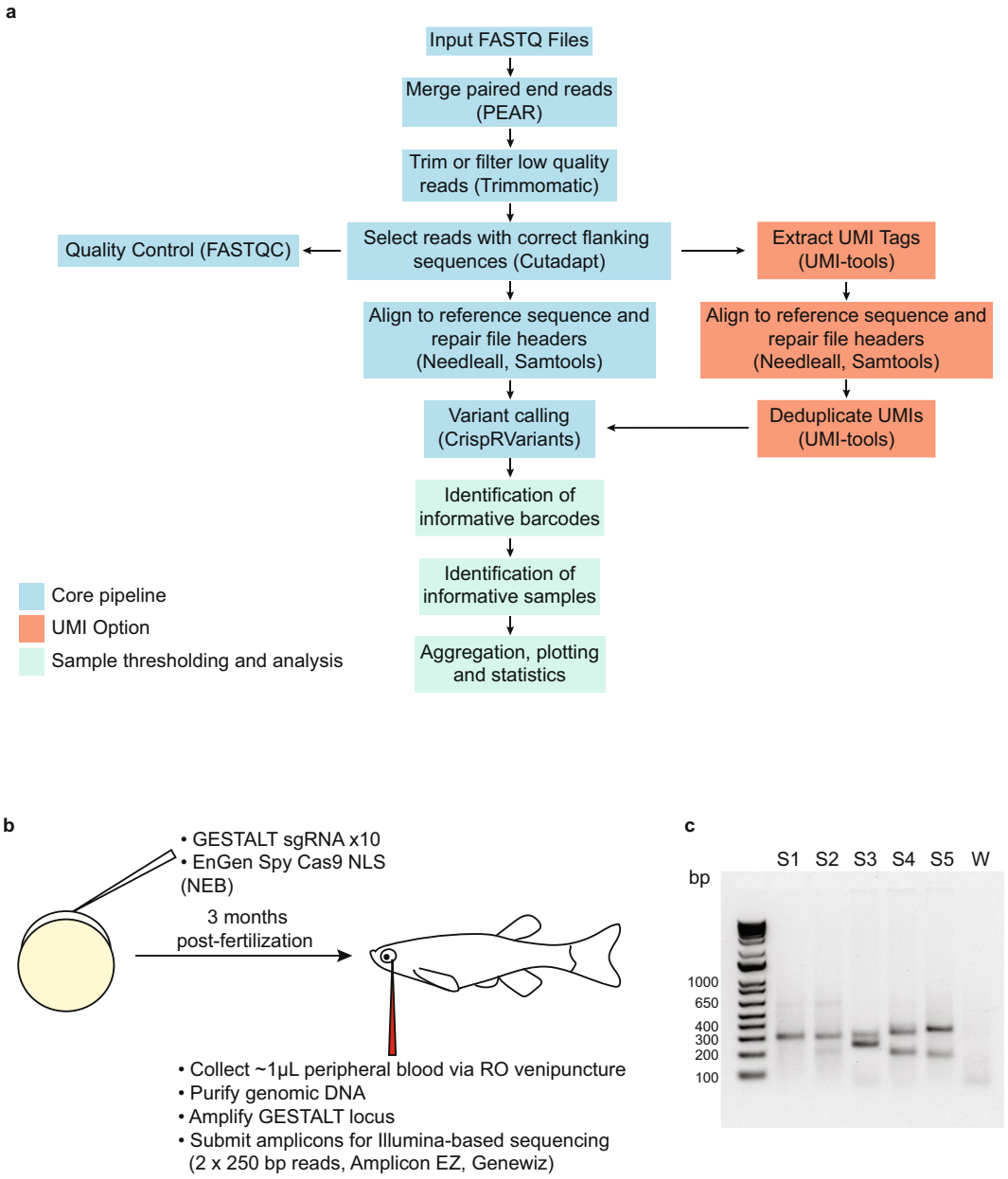

Figure 1. SABER analysis pipeline and experimental outline. (**a**) Major components of the SABER analytical pipeline pictured in blue, red and green boxes. (**b**) Experimental outline for GESTALT barcoding. (**c**) Representative agarose gel electrophoresis of PCR products generated from peripheral blood genomic DNA in GESTALT barcoding experiments. S1-5: 5 samples from the training cohort. W: water control.

sequences would be expected to appear repeatedly in a set of samples that can have no cellular phylogenetic relationship (i.e. from different animals), so we developed a method to search for variants shared between samples within the training set and classify them as uninformative GESTALT sequences.

The *m*-by-*n* matrix, A, of 28 samples (*m*) and 26,877 unique GESTALT variants (*n*) is highly sparse, suggesting that there are few shared alleles (Supplementary Fig. S4). Unsupervised hierarchical clustering of A showed a group of samples with a large number of reads in a common variant, in this case the unedited GESTALT sequence (asterisk, Fig. 3a). Other than the unedited sequence, one GESTALT variant was shared at relatively high frequency between two samples (variant 125:1I with 20,225 reads and VAF = 0.29 in sample AB042 and 4073 reads and VAF = 0.08 in sample AB053, arrow, Fig. 3a). The 20 most common GESTALT variants detected in the training set are shown in Supplementary Fig. S5. To quantify the degree to which any two samples in the training set shared GESTALT alleles, we developed the Sharing Factor (see Methods/Data Analysis) and calculated this for each pair of samples in the training set. The Sharing Factor can be expressed as the proportion of reads attributed to variants shared between two samples. This is plotted for all pairwise combinations of samples, including all GESTALT variants (edited plus unedited) or only edited GESTALT variants, in Fig. 3b. The mean Sharing Factor

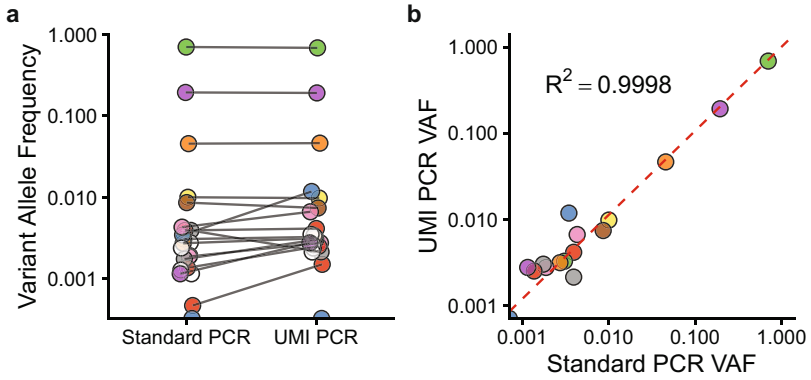

Figure 2. Comparison of Standard and UMI-based PCR protocols. (**a**,**b**) A representative GESTALT blood sample was amplified with UMI-tagged primers using either a standard or UMI-based PCR protocol. (**a**) Variant allele frequencies for the top 20 most common variants in each sample plus the unaltered GESTALT allele. Identical variants are joined; variants missing in either sample are colored white. (**b**) Scatter plot of variant allele frequencies from the same sample. Linear regression model plotted in red (slope = 0.97, intercept = 0.001, $R^2 = 0.9998$).

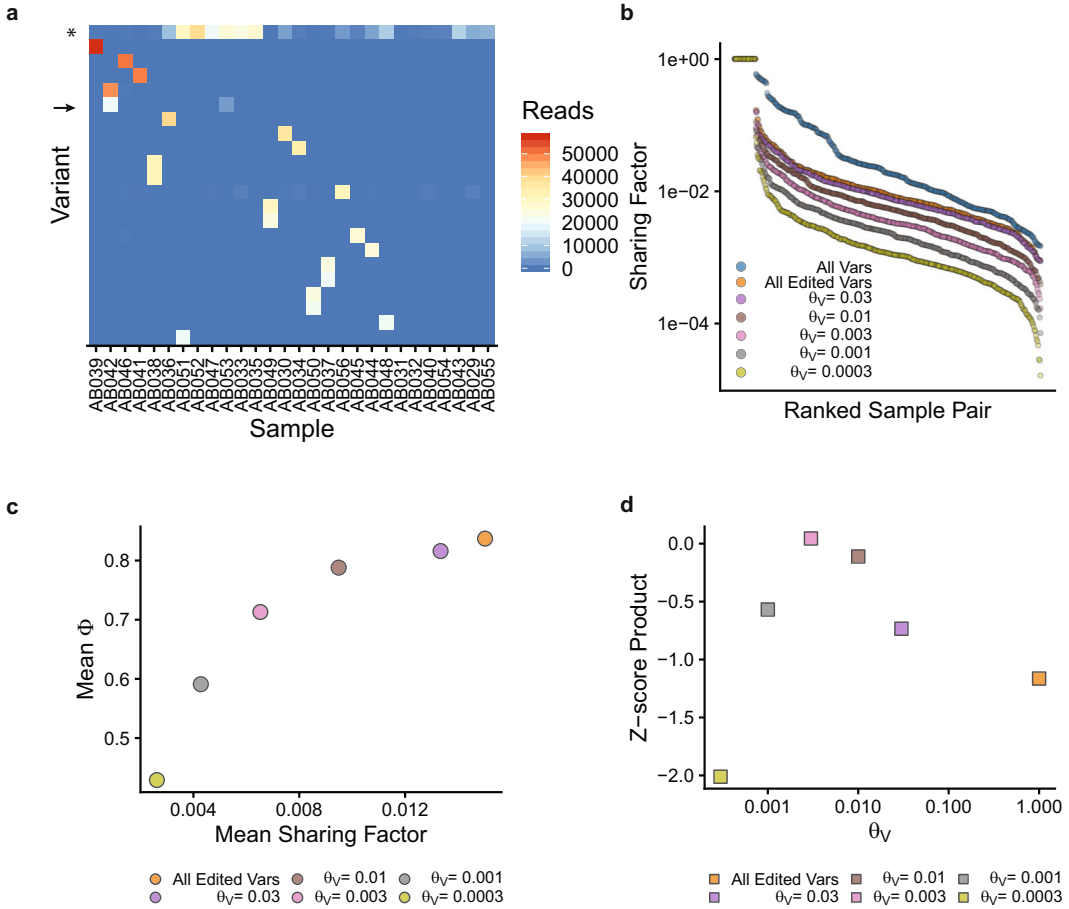

Figure 3. Identification of informative GESTALT barcodes. (**a**) Plot of GESTALT variants by sample. Only variants with read count > 20,000 in any sample are shown. Color indicates read count for each variant. Asterisk indicates the unedited GESTALT allele. Arrow indicates a high-frequency shared allele. (**b**) Sharing Factor curves for all pairwise combinations of samples, grouped by threshold, θ_V. All Vars: All detected GESTALT variants; All Edited Vars: All edited GESTALT variants with $\theta_V = 1.0$. (**c**) The mean Sharing Factor and mean fraction of informative reads remaining after removing common variants and unedited GESTALT alleles (mean Φ) are plotted. (**d**) The Z-score of mean Φ times the Z-score of 1-mean Sharing factor (Z-score Product) is plotted at each value of θ_V.

in the training set (excluding identical comparisons) was 0.06 for all GESTALT alleles. After removing unedited alleles, the mean Sharing Factor was 0.015.

We sought to cap the maximum frequency permitted for a variant shared between any two samples and to further reduce the Sharing Factor across the dataset. A variant allele fraction threshold (θ_V) was introduced into SABER with the following heuristic: if a variant is detected above θ_V in more than one sample, it is labeled as a "common variant" and excluded from further analysis (Supplementary Fig. S1). A very stringent (low) value for θ_V would be expected to identify and exclude a large number of common variants, to minimize the Sharing Factor, and to put a stringent cap on the VAF permitted for any variant shared between two samples; this would also, however, limit the number of informative reads remaining for each sample. To generate a model for minimizing inter-sample variant sharing and maximizing the fraction of informative reads (Φ, defined as the number of reads assigned to informative barcodes divided by the total number of aligned reads) remaining in each sample, the analysis was iterated with θ_V = 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, and 0.3. Plotting these in descending rank order generated the family of curves shown in Fig. 3b (the curves for $\theta_V = 0.1$ and 0.3 are identical to the curve for all edited variants and are omitted for clarity). Mean Sharing Factor and mean Φ values are plotted for the 28 training samples analyzed under each condition (Fig. 3c). To select an optimal value for θ_V, a standardized measure was calculated by taking the Altman Z-score of the mean Φ at each value of θ_V and multiplying this by the Z-score of 1 minus the mean Sharing Factor at each value of θ_V (Fig. 3d). The value of θ_V corresponding to the maximum of this Z-score product was 0.003. At this threshold setting, variants accounting for more than 0.3% of reads in more than one sample are identified as common variants and excluded. The mean Sharing Factor at this threshold was 0.0065 and the mean Φ was 0.71. All variants not identified as common variants or unedited GESTALT alleles are considered valid GESTALT barcodes.

Selection of samples based on the fraction of informative barcodes. We next recognized that samples with a low Φ might be less representative of the true number of HSC clones compared to samples with a high Φ. Excluding low Φ samples in a systematic, objective way could improve the accuracy of enumerating HSC clones. For each sample in the training set, Φ and the number of unique valid barcodes with VAF > 0.02 ($B_{0.02}$) was calculated. Bootstrap analysis (N = 1000 repetitions) was performed and the standard deviation of $B_{0.02}$ and mean Φ across all replicates was plotted ($r = -0.34$, Fig. 4a). The inverse relationship between these values supports the notion that samples with a high Φ are less variable and more precise predictors of the number of HSC clones than samples with a low Φ. To objectively identify a cutoff point between high and low Φ samples, the bootstrap estimate and 95% confidence intervals were calculated (mean Φ bootstrap estimate = 0.71, 95% CI = [0.59,0.81], Fig. 4b). The lower bound of the 95% confidence interval, plotted as a horizontal dashed line, classified 18 samples as high Φ and 10 as low Φ (Fig. 4c). The distribution of $B_{0.02}$ for high and low Φ samples was significantly different (Kolmogorov-Smirnov test $p = 1.5 \times 10^{-7}$, Fig. 4d). The mean Φ in the low and high Φ samples was 0.34 and 0.92, respectively. The mean Sharing Factor for the 18 high Φ samples was 0.0024, a 25-fold reduction from the original training set (Fig. 4e).

Enumerating barcoded HSC clones from GESTALT variants. On average, the number of detected GESTALT barcodes in the training set was large (1254 ± 563 barcodes per sample, range 590–2663). The distribution was heavily right-skewed with 98% of reads assigned to 9 or fewer variants per sample (representative sample with the top 25 of 1046 variants shown, Fig. 5a). This presents a challenge when seeking to quantify HSC clones from variant data. Applying a flat cutoff to variant data (e.g., VAF > 0.02 defines an HSC clone) is currently the standard for quantifying HSC clones in the clinical and research setting[3,4,31]. Weighted averages derived from ecology and population science such as Shannon entropy[30,32] and the inverse Simpson diversity index[33] can account for all variants in the sample. However, the weight given to common versus uncommon variants in these schemes is arbitrary. We reasoned that the number of HSC clones detectable in a population of animals should follow a normal distribution and asked whether these calculations (counting clones with VAF > 0.02, Shannon entropy, or Simpson diversity) provided normally distributed data in the training set of samples. For all three calculations, the distribution was not significantly different from normal (Shapiro-Wilk test: p = 0.15, p = 0.79 and p = 0.068 for VAF > 0.02, Shannon entropy and Simpson diversity, respectively, Fig. 5b–d). The Shannon entropy and Simpson diversity calculations were no more resistant to outliers than using the flat cutoff and the distribution for the Simpson diversity index was particularly right-skewed. These data support using the VAF > 0.02 cutoff or the Shannon entropy calculation in this experimental context.

Analysis of the validation set. Sequence data from the 28 validation samples were analyzed without UMI deduplication. Between 25,111 and 78,609 reads were obtained per sample (Fig. 6a, Supplementary Fig. S6). Iterative analysis and algorithmic selection of θ_V identified an optimal value of 0.003 corresponding to a mean Sharing Factor of 0.035 and a mean Φ of 0.44 (Fig. 6b–d). Bootstrapping identified 14 high Φ samples (mean Φ = 0.75, Fig. 6e) with a mean Sharing Factor of 0.008 (Fig. 6f,g). The number of HSC clones with VAF > 0.02 in the training and validation sets was similar (4.2 ± 1.8 vs 3.5 ± 2.1, respectively, p = 0.31, Fig. 6h).

Two additional independent datasets were analyzed to demonstrate the utility of the SABER experimental and computational framework in identification of inadequately barcoded samples and uninterpretable experiments in general (Supplementary Figs. S7–10). Supplementary Figs. 7 and 8 show data from samples that were generated in a fashion similar to the training and validation sets. Barcoding efficiency was poor in this experiment (95–98% of alleles were unedited, Supplementary Figs. S7a and 8). The algorithmically-selected value for θ_V in this experiment was 0.001 (Supplementary Fig. S7b–d). After selecting the most informative samples, the mean Sharing Factor was 0.51 with a mean Φ of 0.01 (Supplementary Fig. S7e). With over half of barcodes shared between any two samples despite removing 99% of the most redundant sequences, SABER clearly identifies this as an uninterpretable experiment. Both conditions elicit warning messages from SABER.

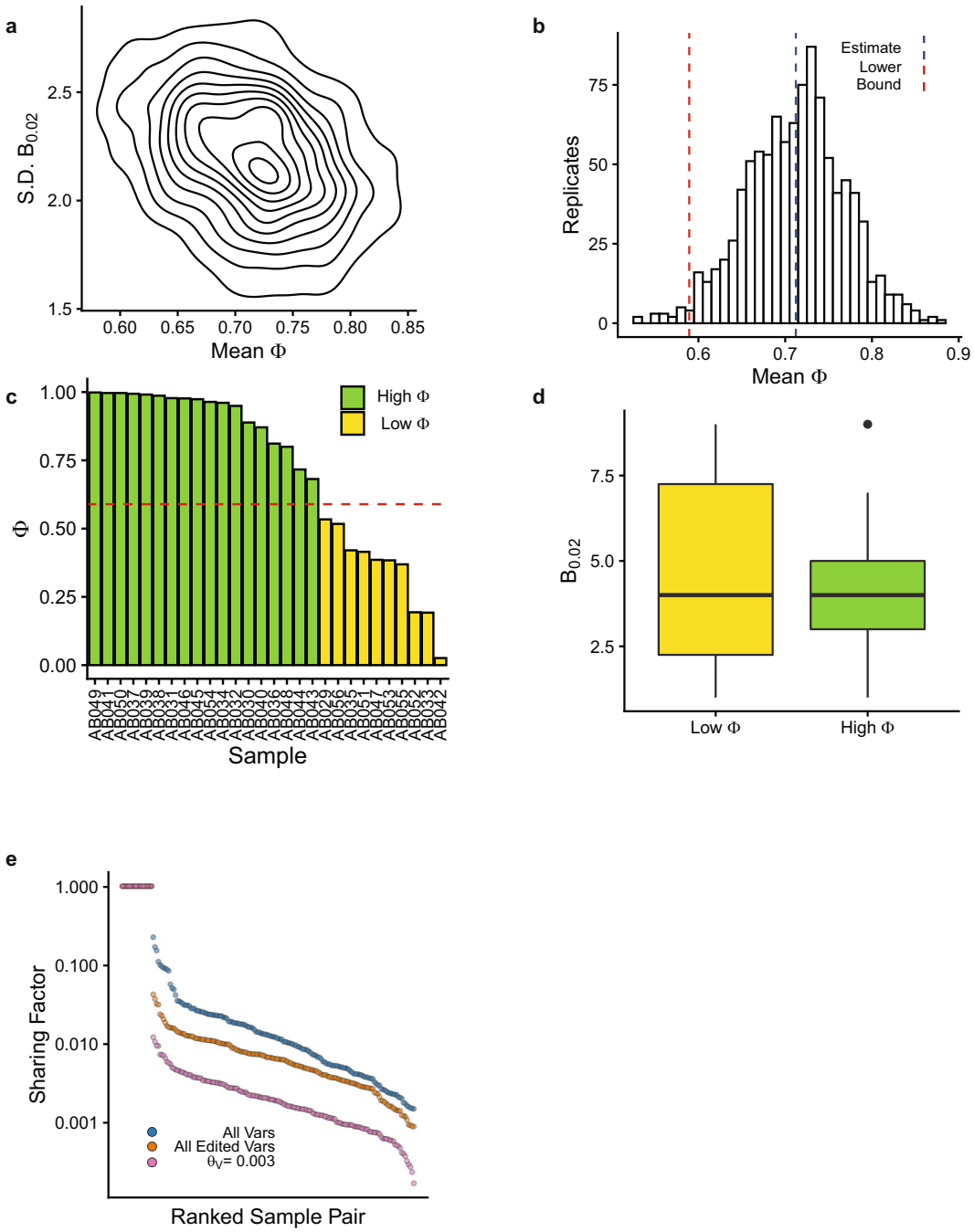

Figure 4. Systematic selection of samples with a high fraction of informative reads. (**a,b**) The fraction of informative reads and number of barcodes with VAF > 0.02 was calculated for each sample in the training set. (**a**) Contour plot of N = 1000 bootstrap replicates for the standard deviation of the number of barcodes with VAF > 0.02 and the mean Φ, r = −0.34. (**b**) Histogram showing the bootstrap estimate and lower bound of the 95% confidence interval using the bias-corrected, accelerated method. (**c**) Fraction of informative reads in each training sample. Red line indicates the confidence interval from b. (**d**) Number of barcodes with VAF > 0.02 in the training samples with low and high fraction of informative reads. Box and whisker plots show median, 1st and 3rd quartiles, and 1.5 x the interquartile range. (**e**) Sharing Factor curves for the samples with a high Φ including all GESTALT variants (blue), all edited GESTALT variants (orange), and all remaining variants at the algorithmically selected final value of θ_V (pink).

A second dataset was generated using an inducible rather than microinjected CRISPR/Cas9 system with heat-shock induction of barcoding at 26 hpf (Supplementary Figs. S9 and S10)[25]. In this experiment, the labeling efficiency was high, but one editing pattern involving 5 GESTALT sites was identified at very high frequency in 14/16 samples (Supplementary Fig. S9a, arrow indicates common variant, asterisk indicates unedited GESTALT allele; Supplementary Fig S10, variant 122:108D). The algorithmically selected value of θ_V was 0.3 (Supplementary

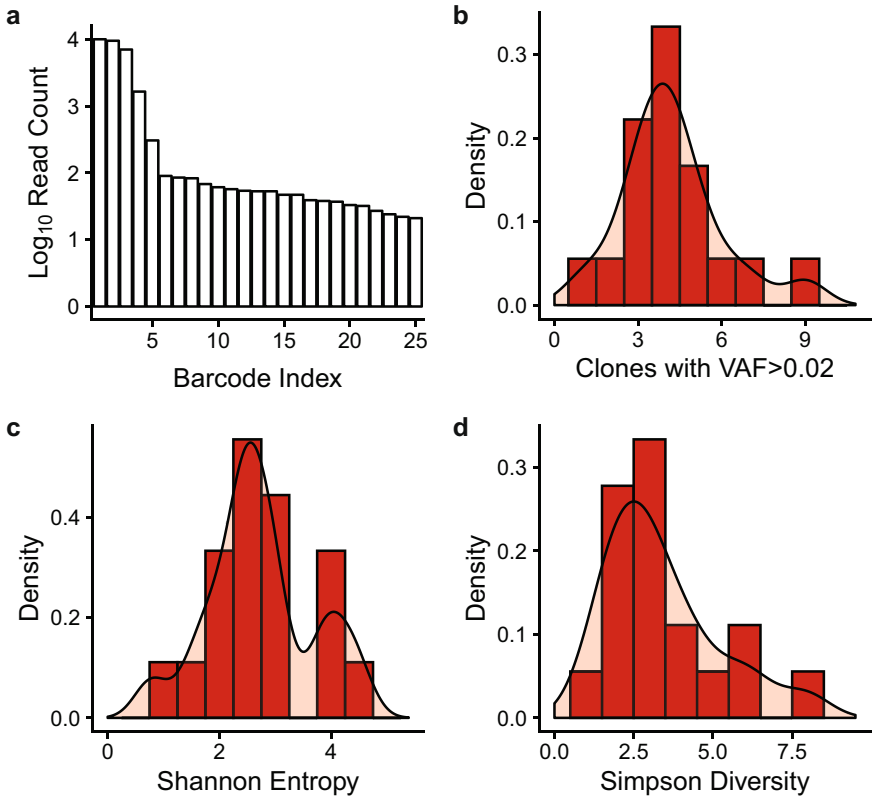

Figure 5. Enumeration of HSC clones from barcoding data. (**a**) Read counts for the top 25 most abundant GESTALT barcodes from a representative sample. (**b**–**d**) Histogram and cumulative density plots for the number of GESTALT HSC clones as determined by (**b**) a flat cutoff of VAF > 0.02, (**c**) Shannon Entropy, and (**d**) the inverse Simpson diversity index.

Fig. S9b–d). Selection of such a high value by SABER suggests that the majority of barcode sharing is driven by one or a few common alleles, a condition that elicits a warning. After selection of informative samples, the mean Sharing Factor was 0.012 and the mean Φ was 0.55, both of which elicit warnings. Because of the suboptimal labeling, SABER identifies this as an uninterpretable dataset. These parameters (θ_V, Φ and mean Sharing Factor) should be reported as quality control measures in any experiment analyzed using SABER.

Longitudinal analysis of clonal dynamics. To demonstrate the ability to reproducibly detect HSC clones and track the dynamic changes in clonal output over time, a subset of animals in the training cohort were bled again at 12 mpf. GESTALT barcodes were amplified, sequenced and processed using SABER. 3-mpf and 12-mpf samples were matched to unique fish identifiers by manual inspection of the top 5 most frequent barcodes in each sample (Fig. 7a). Barcodes were filtered at each time point to include only those present at a VAF > 0.02 and these were normalized to a cumulative frequency of 1.0. HSC clonal dynamics from 3 to 12 mpf are shown for these 4 unique fish in Fig. 7b.

Discussion

HSC heterogeneity has been quantified in many terms using *in vitro* assays, single-cell RNA sequencing, transplantation studies and barcoding under conditions of native hematopoiesis[1]. Although conceptually more abstract than transplantation-based assays, studying barcoded native hematopoiesis has the advantage of minimizing potential artifact and bias induced by the requisite stress imposed on the HSC compartment by transplantation. The challenges encountered by various barcoding techniques lie in barcode induction, barcode validity and barcode interpretation. Here we have used the transgenic zebrafish line, GESTALT, and developed SABER as an analytical framework to attempt to overcome these challenges.

Methods to induce barcode labeling in experimental models include CRISPR/Cas9 editing of non-critical regions of DNA[16,17,23–25,34], random Cre-Lox recombination[20,22,35–37], random oligonucleotide sequences embedded in transgenes[38], and transposon mobilization[12]. The percent of hematopoietic cells barcoded using these methods ranges from 30 to over 90. Here we provide evidence that animals with a low fraction of informative barcodes have a higher variance between samples compared to animals with a high fraction of informative barcodes. Using the Sharing Factor statistic and bootstrap analysis, we have developed a systematic, objective method for identifying these less-accurate samples. After removing these from the analysis, the remaining embryos in the training and validation cohorts displayed successful editing of between 75% and 92% of the detected GESTALT sequences.

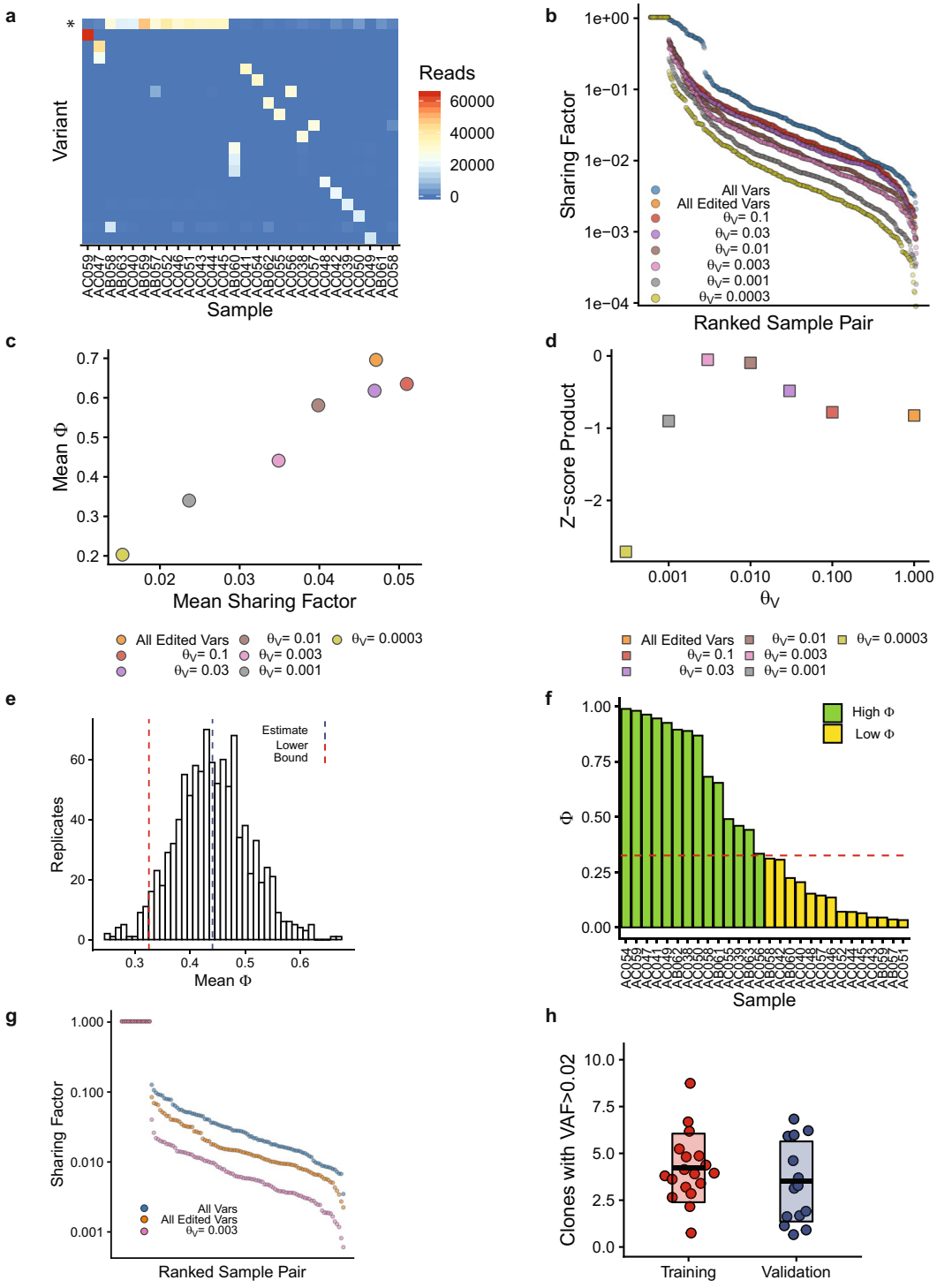

Figure 6. Comparison of training and validation sets. (**a**) Heatmap of GESTALT variant read counts. Asterisk indicates the unedited GESTALT allele. (**b**) Sharing factor curves for all GESTALT variants, all edited GESTALT variants and at indicated values of θ_V. (**c**) Mean Φ plotted versus mean Sharing Factor at indicated values of θ_V. (**d**) Z-score product plotted at indicated values of θ_V. (**e**) Bootstrap analysis of the validation set with estimate and lower bound of the 95% confidence interval. (**f**) Histogram showing samples with a high and low fraction of informative reads. (**g**) Sharing Factor curves at the algorithmically selected final value of θ_V. (**h**) Number of GESTALT HSC clones with VAF > 0.02 in the training and validation sets. Boxes show mean and standard deviation.

Barcode validity is determined by the dynamic range of the barcoding system and the resolution of the method employed to "read" the barcodes. Barcodes with a genome-wide distribution such as random transgene insertions[38], transposon integration sites[12], CRISPR or Cre-Lox recombination targets present as high-multicopy

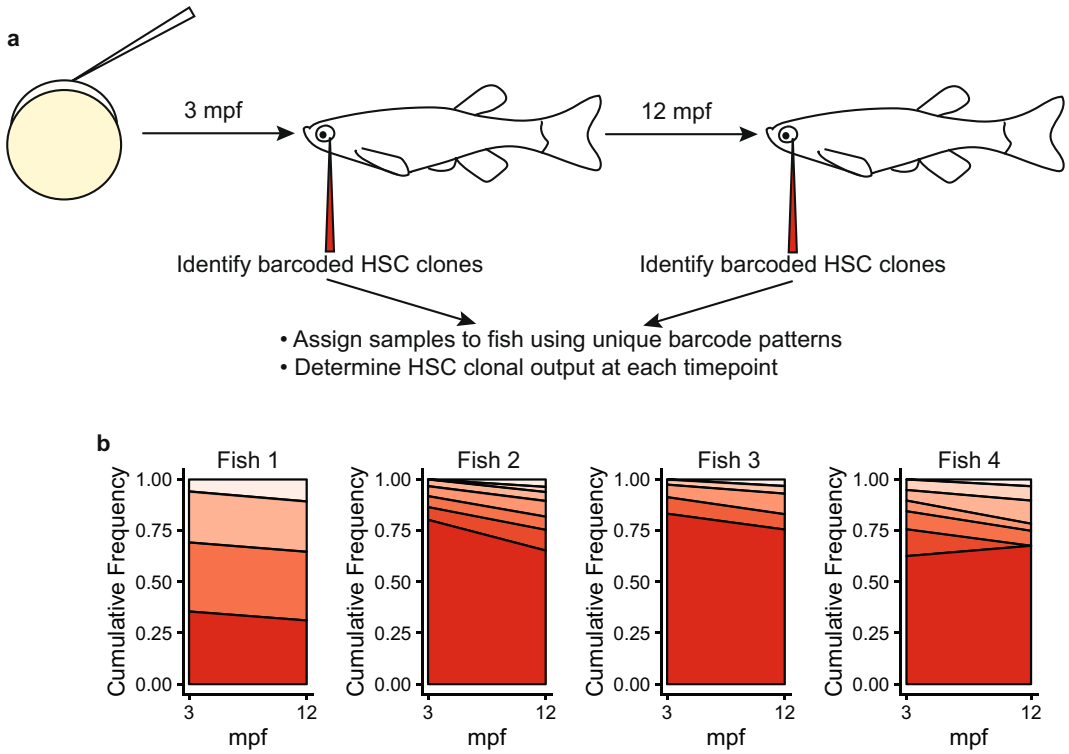

Figure 7. Longitudinal analysis of clonal dynamics. (**a**) Experimental design. (**b**) Stacked area plots showing HSC clones with VAF > 0.02 at 3 and 12 mpf. Each polygon represents a unique HSC clone. The height of each polygon at the 3 and 12 mpf marks corresponds to the frequency of that clone at that time point normalized to a cumulative frequency of 1.0.

transgenes[17,20], or naturally arising SNPs[13] have a very high dynamic range. The resolution of genetic barcodes is limited only by sequencing error (on the order of 10^{-6} for indels using Illumina-based platforms[39]) whereas the resolution of optical barcodes is limited by the instrumentation and clustering algorithm used. Our data show that the 10 tandem GESTALT CRISPR targets are subject to stereotypical repair patterns which limits the dynamic range of this system. Here we have taken an empiric approach to identifying uninformative GESTALT variants by creating a threshold to define such variants shared across samples and remove them from the analysis. Modeling the fraction of informative reads remaining and the degree of variant sharing by calculating the Sharing Factor at multiple threshold settings allowed us to algorithmically select optimal parameters for the analysis.

The design of genetic barcodes largely dictates the sequencing approach required to read and interpret them. Single-locus barcodes over 300–400 bp total are not compatible with Illumina-based methods and require long-read sequencing technology. Distributed multicopy barcodes require single cell sequencing techniques, which at present severely limit the number of samples that can be assayed[17,23,38]. Because of the single PCR-amplicon structure of the barcode, GESTALT is ideally suited for high-throughput analysis of clonal distribution in many samples. This in turn permits large, well-controlled experiments addressing chemical, genetic or other factors that may affect the clonal distribution but with small effect sizes. Based on our sequencing metrics, throughput and cost per sample could be further improved with 2x or 4x sample multiplexing.

Complex transgenic barcodes like GESTALT have been introduced with the aim of recording the lineage history of each cell in a complex, vertebrate organism[23–25,40]. This requires the genetic recorder to operate continuously over the developmental period of interest and for the recording media to be large and complex enough to encode this history. The molecular and computational requirements for such a system are demanding[41]. Further, if one were to introduce an experimental variable, it may be challenging to disentangle effects of that variable on the biologic system versus effects on the recorder itself. Instead, we have used a pulse-labeling approach in which thousands of barcodes are introduced within the first 4–5 hours after fertilization, prior to gastrulation and well before HSC specification. Any experimental interventions taking effect thereafter (e.g. conditional or tissue-specific transgenes) will not alter barcoding per se and can be interpreted in terms of the effect on the cell lineage being tested. The barcodes detected in the blood samples in this study can be traced to ancestral cells from 4–5 hpf that gave rise to GESTALT-barcoded HSCs. The number of GESTALT-barcoded HSCs detected using this labeling strategy is necessarily lower than strategies where barcoding occurs later in development and the number of cells available for barcoding is larger[20,25]. However, our results are similar to those seen in the original report of the GESTALT line[16], and extend these findings to a large number of samples with confirmation in an independent cohort of animals.

In this study, we have provided a conceptual framework and analytical approach for quantifying functional GESTALT-barcoded HSC clones from whole blood. It is important to consider that zebrafish red blood cells are

nucleated and so this system is insensitive to the hematopoietic lineage bias (e.g. myeloid versus lymphoid) that might occur with the introduction of oncogenes. Sorting cell populations prior to barcode analysis or performing barcode analysis using a single cell sequencing platform[25,26,42–44] would allow simultaneous recovery of hematopoietic and clonal lineage data. The SABER approach to identifying valid barcodes and selecting samples could also be applied to these methodologies. Indeed, SABER could readily be used to study the clonal diversity of other organ systems or tumor tissues with any single-amplicon transgenic barcode sequence by isolating the cell type of interest through bulk sorting, selection or single cell isolation techniques followed by sequencing of the amplicon barcode. Alternatively, the barcode could be expressed in a tissue-specific manner and the barcode read from cDNA prepared from the organ of interest, analogous to what has been done using the conditional *hsp70l* promoter[25]. As we have shown, SABER can be used to follow longitudinal clonal dynamics and so could be used to study the initiation and progression of hematologic neoplasms or solid tumors and the regeneration of tissues that can be repeatedly sampled such as blood, epidermis and caudal fin structures.

The study of HSC phylogeny and clonal evolution is critically important for our understanding of normal and malignant hematopoiesis[1,2,45,46]. SABER provides a framework to address some of the most fundamental challenges inherent to using amplicon barcodes to study cellular phylogeny: barcode uniqueness, sample selection and experimental quality control. Quantification of non-unique barcodes with the Sharing Factor and rational selection of informative samples are important concepts that are generally applicable to future studies in this field. We anticipate that SABER will be a useful experimental tool to study novel factors that control long-term cellular fate.

Methods

Zebrafish. The GESTALT v7 and heat-shock Cas9 (*Tg(hsp70l:zCas9-2A-EGFP,5* × (*U6:sgRNA*))) lines were a kind gift from A. Schier[16]. *Casper* zebrafish were a kind gift from L. Zon[47]. Zebrafish were housed at the Ohio State University Comprehensive Cancer Center. Animals were maintained and experiments were approved and performed under Ohio State University Institutional Laboratory Animal Care and Use Committee (IACUC) protocol 2018A00000012. All work was performed in accordance with American Association for Accreditation of Laboratory Animal Care (AAALAC) guidelines[48].

CRISPR/Cas9-based GESTALT barcoding. GESTALT sgRNAs were generated by *in vitro* transcription using established methods[16,49]. A 10 μL injection mix containing approximately 20 ng of each sgRNA (200 ng total), 20 pg control plasmid DNA, 200 ng Tol2 mRNA, 60 pmol EnGen Spy Cas9 NLS recombinant protein (NEB) and phenol red was generated and 1 nL was injected into hemizygous GESTALT embryos at the single cell stage. For experiments using the heat-shock Cas9 line, GESTALT barcoding was performed by incubating embryos at 40 °C for 30 min in embryo medium beginning at 26 hpf.

Genomic DNA isolation, PCR and amplicon sequencing. Peripheral blood was collected from anesthetized adult (3 mpf) zebrafish by retro-orbital venipuncture and diluted into 50 μL PBS containing 2% FBS and heparin[20]. Genomic DNA was isolated using the Quick-DNA Miniprep Kit (Zymo Research). 5 μL (approximately 250 ng) was used for the standard PCR protocol and 23 μL (approximately 1.15 μg) was used for the UMI-based PCR protocol. Primer sequences are listed in Supplementary Table S1. The standard PCR protocol (using primers v6_7_F_illum and v6_7_R_illum at 0.4 μM each or v6_7_UMI_F and v6_7_R at 0.5 μM each) was initial denaturation at 98 °C × 5 min, 45 cycles of denaturation at 98 °C × 30 sec, annealing at 56 °C × 30 sec and extension at 72 °C × 1 min, with a final extension at 72 °C × 7 min. The UMI-based PCR protocol (starting with v6_7_UMI_F at 4 μM) was denaturation at 98 °C × 5 min followed by 10 cycles of annealing at 56 °C × 20 sec and extension at 72 °C × 1 min. This was followed by 2 rounds of bead purification (Ampure, Beckman-Coulter) to remove excess primer. The entire eluate was then used as a template for a standard PCR reaction with the GC_tag and v6_7_R primers.

GESTALT amplicons were bead purified (Ampure XP, Beckman Coulter) and resuspended at 20 ng/μL. 25 μL (500 ng) samples were submitted for Amplicon-EZ sequencing, an Illumina-based sequencing service compatible with amplicons 150–500 bp in length that does not include a fragmentation step in library preparation (Genewiz). Amplicons were sequenced as 2 × 250 bp paired reads and demultiplexed prior to delivery. Agarose gel electrophoresis images were acquired on an Aplegen gel documentation system with automatic settings and inverted in Photoshop.

Data analysis. A reproducible analysis pipeline with sample data is freely available at https://github.com/blachlylab/SABER. The sample data are from barcoded GESTALT blood samples similar to those analyzed in Results. SABER is Linux/Mac compatible and requires only installation of Snakemake (via conda) which deploys all other software dependencies. The original R script for the main analysis program, and scripts for comparing clone numbers between experiments and for longitudinal analysis are available at the same site.

Paired-end reads were merged with PEAR, trimmed, and filtered for quality using Trimmomatic (SLIDINGWINDOW:4:15 MINLEN:100). Reads containing incorrect, absent, or multiple flanking primer sequences were filtered out using Cutadapt. Merged reads retained after filtering were aligned to the GESTALT reference sequence using Needleall, an implementation of the Needleman-Wunsch algorithm (gap open penalty = 10, gap extension penalty = 0.25). This aligner is superior to the Burrows-Wheeler aligner for mapping highly divergent sequences such as edited GESTALT variants to a small reference sequence[50]. GESTALT variants were called using the CrispRVariants package using the option split.snv = FALSE to collapse all non-indel mutations into the "no variant" allele[51].

We sought to quantify the degree to which all pairwise combinations of GESTALT samples shared variants in common. Any variant detected in more than one sample above a certain frequency threshold, θ_V, might then be excluded from the analysis as non-informative. In this way we could expect to generate a list of unique GESTALT barcodes with a known maximum amount of sharing across the dataset together with the proportion of reads accounting for these unique barcodes. CrispRVariants was used to generate matrices *C* and *P* in which each

column refers to a unique sample and each row refers to either the read counts *(C)* or proportion *(P)* for a specific variant in each sample, 1…N, in the dataset (schematic in Supplementary Fig. S1, see Supplementary Figs. S5, S6, S8 and S10 for examples of CrispRVariants output). Rows are identified either as "no variant" or by a systematically generated name derived from the CIGAR string produced by the aligner. Only indel length and position are considered in order to reduce the number of false barcodes arising from base miscalls. SABER then identifies all variants in P where the proportional abundance exceeds θ_V in more than 1 sample and adds these variant names to a list of common variants. Matrix C is split into individual read count tables, $C_1 \ldots C_N$ for each sample. Common variants previously identified in P are marked on $C_1 \ldots C_N$ and their read counts summed for each sample. This sum is added as an additional row labeled "common variant sum" on each table and the individual common variant rows are dropped. The resulting data tables, $C'_1 \ldots C'_N$ contain read counts for unedited GESTALT alleles, aggregated common variants and all unique variants as defined by θ_V.

To optimize the value of θ_V through iteration and modeling, we developed a statistic to quantify the magnitude of GESTALT variant sharing across the dataset. For a given variant v, sample pair (p_1, p_2), and variant read counts v_1 and v_2, respectively, we define the sharing coefficient for this variant for the pair (p_1, p_2) as:

$$s = \frac{2 \times \min(v_1, v_2)}{v_1 + v_2} \tag{1}$$

This coefficient has the following properties:

(a) $0 \leq s \ll 1$ when $\frac{v_1}{v_2}$ or $\frac{v_2}{v_1}$ approaches 0.
(b) $0 \ll s \leq 1$ when $\frac{v_1}{v_2}$ or $\frac{v_2}{v_1}$ approaches 1.
(c) $s = 0$ when the variant is unique to one sample in the pair.
(d) $s = 1$ when $v_1 = v_2$.

Considering the matrix,

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ \vdots & \vdots \\ b_{m1} & b_{m2} \end{bmatrix} \tag{2}$$

where m is the number of distinct GESTALT variants detected in (p_1, p_2) and b_{ij} is the number of reads of the i^{th} variant detected in sample j, we define the Sharing Factor, S, for (p_1, p_2) as

$$S = \frac{2 \times \sum_{i=1}^{m} pmin_i}{\sum_{i=1}^{m} b_{i1} + b_{i2}} \tag{3}$$

Where:

$$pmin_i = \min_{j=[1\ 2]} b_{i,j}, \ \ i = 1, \ldots, m \tag{4}$$

The denominator of Eq. (3) is the sum of all reads in sample pair (p_1, p_2). S and s are analogous in that $S = 0$ for a pair of samples with no shared variants, $S = 1$ when comparing a pair of identical samples, S is close to 0 when shared variants are rare in the pair and S is close to 1 when one or more shared variants are common between the pair.

The Sharing Factor, S, can be expressed as the proportion of all variants shared between a given pair of samples. SABER calculates S and the number of reads assigned to informative barcodes divided by the total number of aligned reads (fraction of informative reads, Φ) for all pairwise combinations of samples in a given dataset with $\theta_V = 0.0003$, 0.001,0.003,0.01,0.03,0.1,0.3 and 1. The mean S for the dataset (excluding identical comparisons) versus the mean Φ is then plotted for each value of θ_V (e.g. Fig. 3c). To select an optimal value for θ_V, SABER calculates the Altman Z-score for the mean Φ and 1- mean S at each value of θ_V and stores these as numeric vectors[52]. The element-wise product of these two vectors is calculated and the maximum of the resulting vector is taken to correspond to the optimal value of θ_V (e.g. Fig. 3d).

Bootstrapping was performed using the "boot" package in R by first calculating the number of valid GESTALT barcodes with VAF > 0.02 in each sample and then taking the mean and standard deviation of this value over 1000 replicates. Bootstrap confidence intervals were calculated using the bias-corrected, accelerated method.

Shannon entropy was calculated as

$$H' = -\sum_{i=1}^{R} p_i \ln p_i \tag{5}$$

where R is equal to richness, or the total number of observed barcodes and p_i is equal to the proportion of all barcodes represented by the i^{th} barcode.

The inverse Simpson Diversity index was calculated as

$${}^2D = \frac{1}{\sum_{i=1}^{R} p_i^2} \tag{6}$$

Statistical methods. Statistical analysis was performed using R, version 3.6.1. All statistical methods are explicitly detailed in Results and in the available software. Box and whisker plots are in the style of Tukey and show median, 1st and 3rd quartiles and 1.5 x the interquartile range. HSC clone data in the training and validation groups were first shown to be normal using the Shapiro-Wilk test and then two-tailed p values were calculated using Student's t-test. All numeric data are expressed as mean ± standard deviation unless otherwise indicated.

Data availability

The full datasets generated and analyzed during the current study are available in the Sequence Read Archive (SRA) repository, with the primary accession code PRJNA563355 (http://www.ncbi.nlm.nih.gov/bioproject/563355).

Received: 19 September 2019; Accepted: 24 January 2020;
Published online: 07 February 2020

References

1. Haas, S., Trumpp, A. & Milsom, M. D. Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. *Cell Stem Cell* **22**, 627–638, https://doi.org/10.1016/j.stem.2018.04.003 (2018).
2. Mossner, M. *et al*. Mutational hierarchies in myelodysplastic syndromes dynamically adapt and evolve upon therapy response and failure. *Blood* **128**, 1246–1259, https://doi.org/10.1182/blood-2015-11-679167 (2016).
3. Jaiswal, S. *et al*. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498, https://doi.org/10.1056/NEJMoa1408617 (2014).
4. Genovese, G. *et al*. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487, https://doi.org/10.1056/NEJMoa1409405 (2014).
5. Xie, M. *et al*. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478, https://doi.org/10.1038/nm.3733 (2014).
6. Ptashkin, R. N. *et al*. Prevalence of Clonal Hematopoiesis Mutations in Tumor-Only Clinical Genomic Profiling of Solid Tumors. *JAMA oncology*, https://doi.org/10.1001/jamaoncol.2018.2297 (2018).
7. Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927, https://doi.org/10.1016/0092-8674(86)90566-0 (1986).
8. Yamamoto, R. *et al*. Large-Scale Clonal Analysis Resolves Aging of the Mouse Hematopoietic Stem Cell Compartment. *Cell Stem Cell* **22**, 600–607 e604, https://doi.org/10.1016/j.stem.2018.03.013 (2018).
9. Yamamoto, R. *et al*. Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154**, 1112–1126, https://doi.org/10.1016/j.cell.2013.08.007 (2013).
10. Dykstra, B. *et al*. Long-Term Propagation of Distinct Hematopoietic Differentiation Programs *In Vivo*. *Cell Stem Cell* **1**, 218–229, https://doi.org/10.1016/j.stem.2007.05.015 (2007).
11. Morita, Y., Ema, H. & Nakauchi, H. Heterogeneity and hierarchy within the most primitive hematopoietic stem cell compartment. *J. Exp. Med.* **207**, 1173, https://doi.org/10.1084/jem.20091318 (2010).
12. Sun, J. *et al*. Clonal dynamics of native haematopoiesis. *Nat.* **514**, 322–327, https://doi.org/10.1038/nature13824 (2014).
13. Osorio, F. G. *et al*. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308–2316 e2304, https://doi.org/10.1016/j.celrep.2018.11.014 (2018).
14. Lee-Six, H. *et al*. Population dynamics of normal human blood inferred from somatic mutations. *Nat.* **561**, 473–478, https://doi.org/10.1038/s41586-018-0497-0 (2018).
15. Ludwig, L. S. *et al*. Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325–1339 e1322, https://doi.org/10.1016/j.cell.2019.01.022 (2019).
16. McKenna, A. *et al*. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Sci.* **353**, aaf7907, https://doi.org/10.1126/science.aaf7907 (2016).
17. Spanjaard, B. *et al*. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473, https://doi.org/10.1038/nbt.4124 (2018).
18. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature*, https://doi.org/10.1038/nature25969 (2018).
19. Ganuza, M. *et al*. The global clonal complexity of the murine blood system declines throughout life and after serial transplantation. *Blood* **133**, 1927–1942, https://doi.org/10.1182/blood-2018-09-873059 (2019).
20. Henninger, J. *et al*. Clonal fate mapping quantifies the number of haematopoietic stem cells that arise during development. *Nat. Cell Biol.* **19**, 17–27, https://doi.org/10.1038/ncb3444 (2017).
21. Coombs, C. C. *et al*. Identification of clonal hematopoiesis mutations in solid tumor patients undergoing unpaired next-generation sequencing assays. *Clin Cancer Res*, https://doi.org/10.1158/1078-0432.CCR-18-1201 (2018).
22. Pei, W. *et al*. Polylox barcoding reveals haematopoietic stem cell fates realized *in vivo*. *Nat.* **548**, 456–460, https://doi.org/10.1038/nature23653 (2017).
23. Chan, M. M. *et al*. Molecular recording of mammalian embryogenesis. *Nat.* **570**, 77–82, https://doi.org/10.1038/s41586-019-1184-5 (2019).
24. Kalhor, R. *et al*. Developmental barcoding of whole mouse via homing CRISPR. *Science*, https://doi.org/10.1126/science.aat9804 (2018).
25. Raj, B. *et al*. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol*, https://doi.org/10.1038/nbt.4103 (2018).
26. Raj, B., Gagnon, J. A. & Schier, A. F. Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR-Cas9 barcodes by scGESTALT. *Nat. Protoc.* **13**, 2685–2713, https://doi.org/10.1038/s41596-018-0058-x (2018).
27. Clement, K., Farouni, R., Bauer, D. E. & Pinello, L. AmpUMI: design and analysis of unique molecular identifiers for deep amplicon sequencing. *Bioinforma.* **34**, i202–i210, https://doi.org/10.1093/bioinformatics/bty264 (2018).
28. Chen, W. *et al*. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Res*, https://doi.org/10.1093/nar/gkz487 (2019).
29. Shen, M. W. *et al*. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nat.* **563**, 646–651, https://doi.org/10.1038/s41586-018-0686-x (2018).
30. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (1948).
31. Steensma, D. P. Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood Adv.* **2**, 3404–3410, https://doi.org/10.1182/bloodadvances.2018020222 (2018).
32. Makishima, H. *et al*. Dynamics of clonal evolution in myelodysplastic syndromes. *Nat. Genet.* **49**, 204–212, https://doi.org/10.1038/ng.3742 (2017).
33. Simpson, E. H. Measurement of Diversity. *Nat.* **163**, 688–688, https://doi.org/10.1038/163688a0 (1949).

34. Frieda, K. L. *et al*. Synthetic recording and *in situ* readout of lineage information in single cells. *Nat.* **541**, 107–111, https://doi.org/10.1038/nature20777 (2017).
35. Pan, Y. A. *et al*. Zebrabow: multispectral cell labeling for cell tracing and lineage analysis in zebrafish. *Dev.* **140**, 2835–2846, https://doi.org/10.1242/dev.094631 (2013).
36. Livet, J. *et al*. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nat.* **450**, 56–62, https://doi.org/10.1038/nature06293 (2007).
37. Snippert, H. J. *et al*. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144, https://doi.org/10.1016/j.cell.2010.09.016 (2010).
38. Wagner, D. E. *et al*. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, https://doi.org/10.1126/science.aar4362 (2018).
39. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinforma.* **17**, 125, https://doi.org/10.1186/s12859-016-0976-y (2016).
40. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, https://doi.org/10.1242/dev.169730 (2019).
41. Salvador-Martinez, I., Grillo, M., Averof, M. & Telford, M. J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *Elife* **8**, https://doi.org/10.7554/eLife.40292 (2019).
42. Paul, F. *et al*. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677, https://doi.org/10.1016/j.cell.2015.11.013 (2015).
43. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289, https://doi.org/10.1038/nbt.3129 (2015).
44. Macaulay, I. C. *et al*. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protoc.* **11**, 2081–2103, https://doi.org/10.1038/nprot.2016.138 (2016).
45. Scala, S. & Aiuti, A. *In vivo* dynamics of human hematopoietic stem cells: novel concepts and future directions. *Blood Adv.* **3**, 1916–1924, https://doi.org/10.1182/bloodadvances.2019000039 (2019).
46. Nangalia, J., Mitchell, E. & Green, A. R. Clonal approaches to understanding the impact of mutations on hematologic disease development. *Blood*, https://doi.org/10.1182/blood-2018-11-835405 (2019).
47. White, R. M. *et al*. Transparent adult zebrafish as a tool for *in vivo* transplantation analysis. *Cell Stem Cell* **2**, 183–189, https://doi.org/10.1016/j.stem.2007.11.002 (2008).
48. Westerfield, M. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio Rerio)*. (M. Westerfield, 2007).
49. Hwang, W. Y. *et al*. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229, https://doi.org/10.1038/nbt.2501 (2013).
50. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453, https://doi.org/10.1016/0022-2836(70)90057-4 (1970).
51. Lindsay, H. *et al*. CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat. Biotechnol.* **34**, 701–702, https://doi.org/10.1038/nbt.3628 (2016).
52. Altman, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Financ.* **23**, 589–609, https://doi.org/10.1111/j.1540-6261.1968.tb00843.x (1968).

Acknowledgements

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number K08DK111920.

Author contributions

E.M.T. performed experiments, acquired data and edited the manuscript; C.T.G. analyzed data, revised software and edited the manuscript; J.S. performed experiments, acquired data and edited the manuscript; J.S.B. revised software, edited the manuscript, and supervised the work; B.W.B. conceived and designed the work, wrote the initial version of the software, analyzed the data, wrote the manuscript, and supervised the work. All authors have approved the submitted version of the manuscript and are personally accountable for their own contributions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59119-8.

Correspondence and requests for materials should be addressed to B.W.B.

Reprints and permissions information is available at www.nature.com/reprints.

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