Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing

Anders Ståhlberg, Paul M. Krzyzanowski, Jennifer B. Jackson, Matthew Egyud, Lincoln Stein and Tony E. Godfrey

Supplementary figures

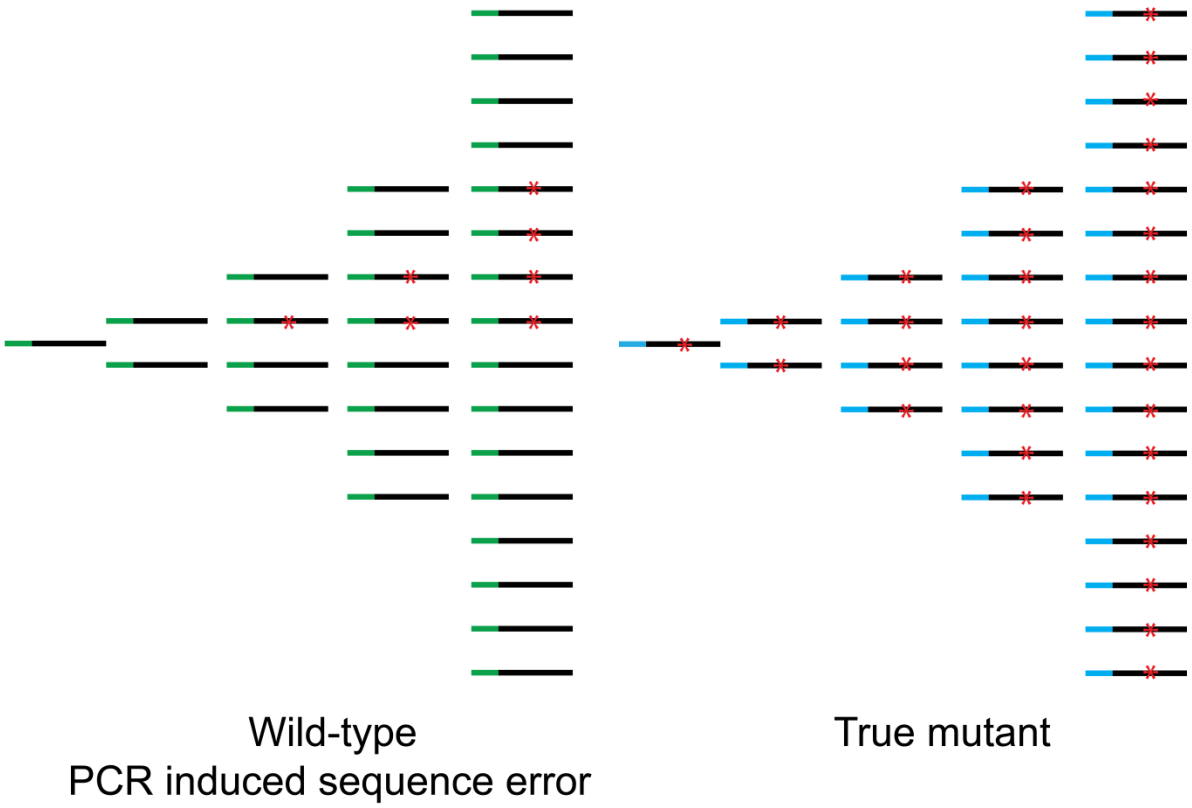

Figure S1. The principle of barcoding. Each target DNA molecule is barcoded with a unique sequence. All PCR amplified molecules that are generated from the same original molecule receive the same barcode. Hence, if a PCR error is introduced in the library construction, only a fraction of all DNA molecules with the same barcode will amplify that specific error (left, green barcode). Conversely, if a mutation is present in the original molecule all downstream generated amplicons with that particular barcode will have the same mutation and can therefore be called a true mutant (right blue barcode).

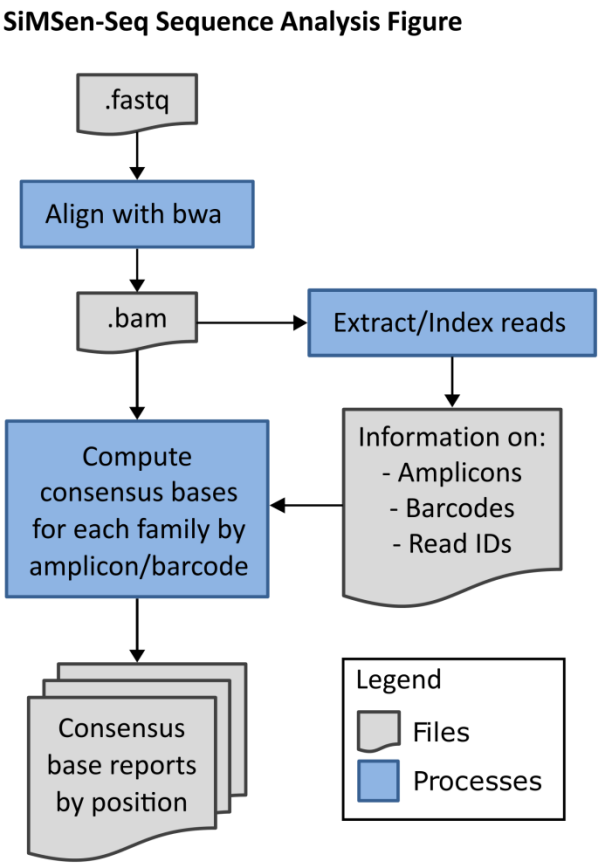

Figure S2. Schematic of bioinformatics workflow used to process SiMSenSeq data. All reads are first aligned to the appropriate genome, and the corresponding bam file is used to identify locations of valid reads containing an adaptor and barcode sequence. Consensus bases are computed by collapsing reads that have the same barcode and amplicon combination according to rules described in Material and Methods.

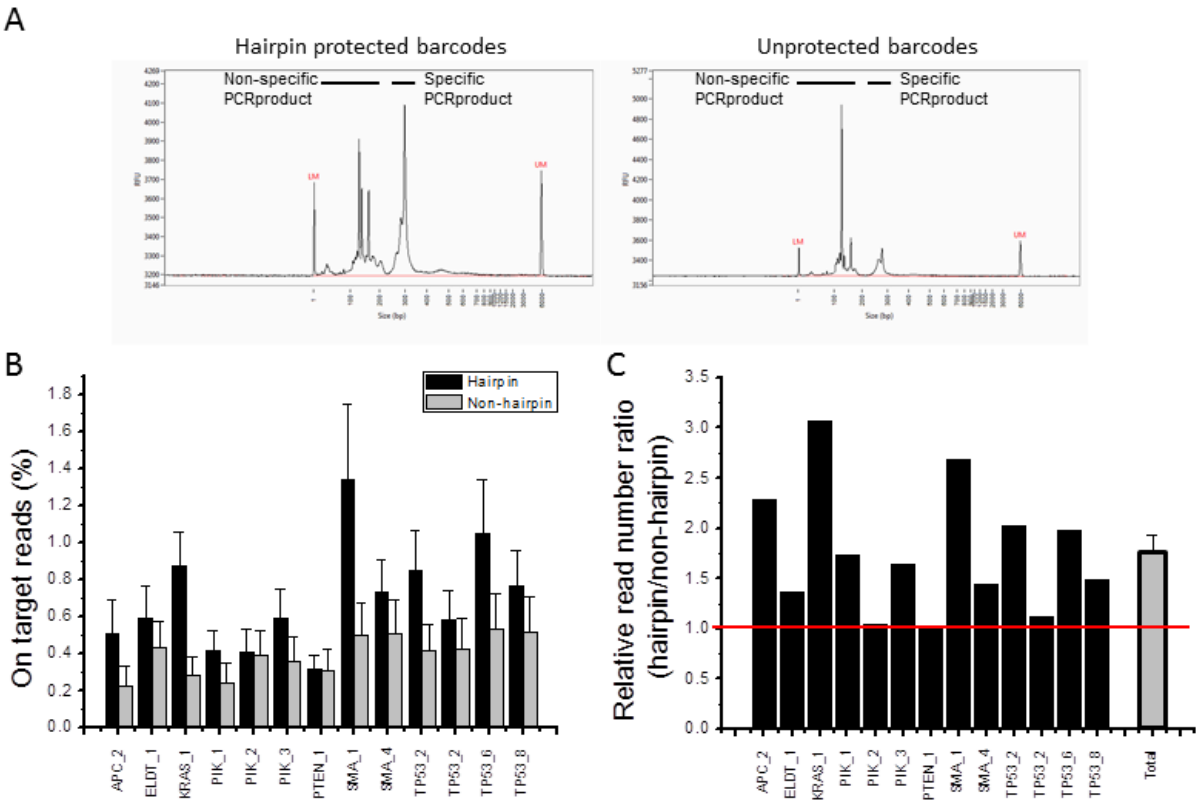

Figure S3. Comparison between hairpin and non-hairpin protected barcodes using SiMSen-Seq. (A) Electropherograms of unpurified libraries targeting one DNA sequence (TP53_6) with and without hairpin protected barcodes using the Fragment Analyzer are shown. Hairpin protected barcodes generated significantly more specific PCR product than unprotected barcodes (average = 1.73; $p < 0.01$; n = 37). Thirteen individual assays were analyzed in replicates (Table S1). (B) On target reads as a percentage of total reads for each of the 13 assays in (A) generated as a 13-plex library. Unpurified libraries were analyzed in triplicate. Mean ± SD is shown. The coefficient of variation between hairpin protected barcodes and unprotected barcodes was equal. (C) All 13 assays with hairpin protected barcodes generated more reads than unprotected barcodes. On average 1.76 times more reads were observed for hairpin protected barcodes ($p < 0.01$). Mean ± SEM is shown.

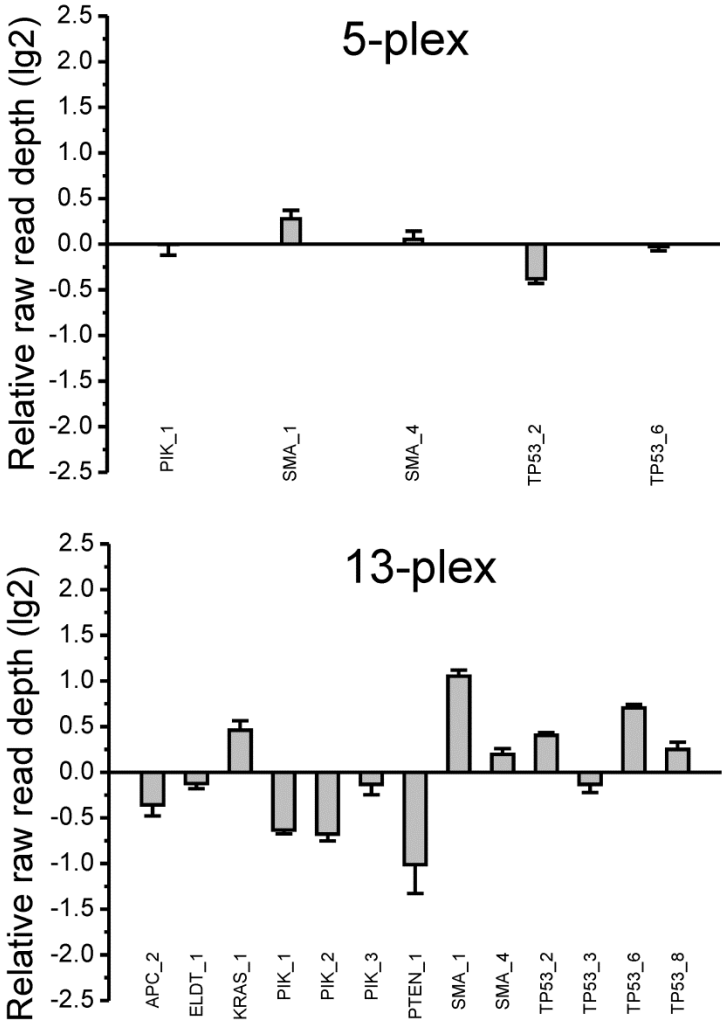

Figure S4. **Relative raw reads uniformity between individual amplicons using SiMSen-Seq.** Relative raw read depth of 5 and 13 multiplexed amplicons were analyzed. The average raw read depth was $2.3*10^6$ and $1.1*10^4$ per amplicon for the 5- and 13-plex libraries, respectively. DNA from tumor cell line CP-A was used for all experiments. Mean ± SD is shown ($n_{5\text{-plex}} = 12$, $n_{13\text{-plex}} = 3$).

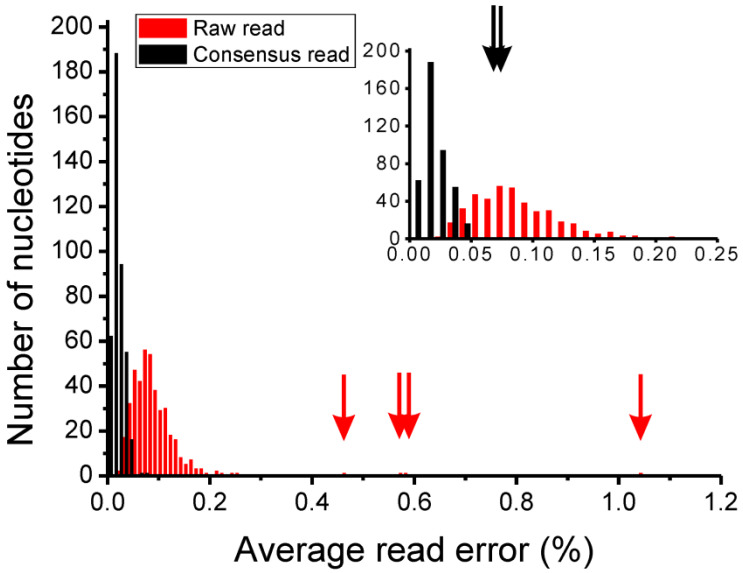

Figure S5. Read error parameters. Distribution of average read errors for total raw and consensus reads. Arrows indicate single nucleotides.

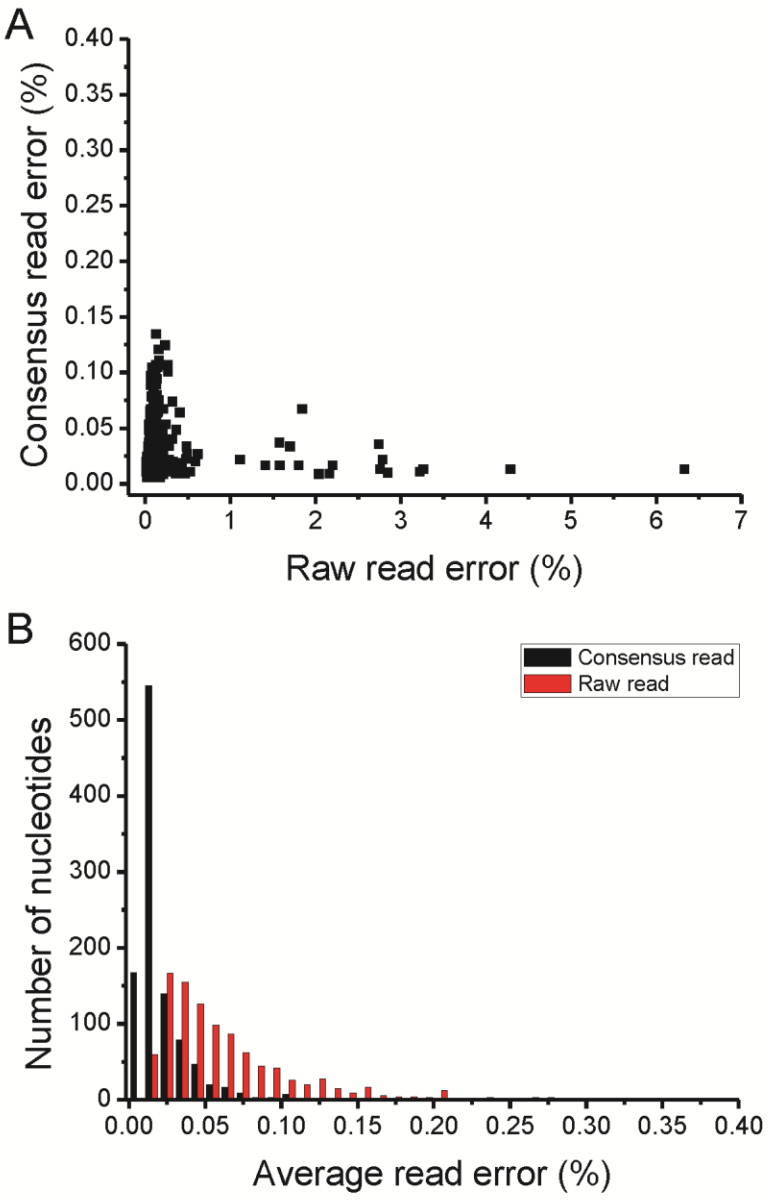

Figure S6. **Read error parameters.** (A) Total consensus versus raw read error for 1042 nucleotides in 13 amplicons. (B) Distribution of average read errors for total raw and consensus reads. Raw read errors above 1% are not shown.

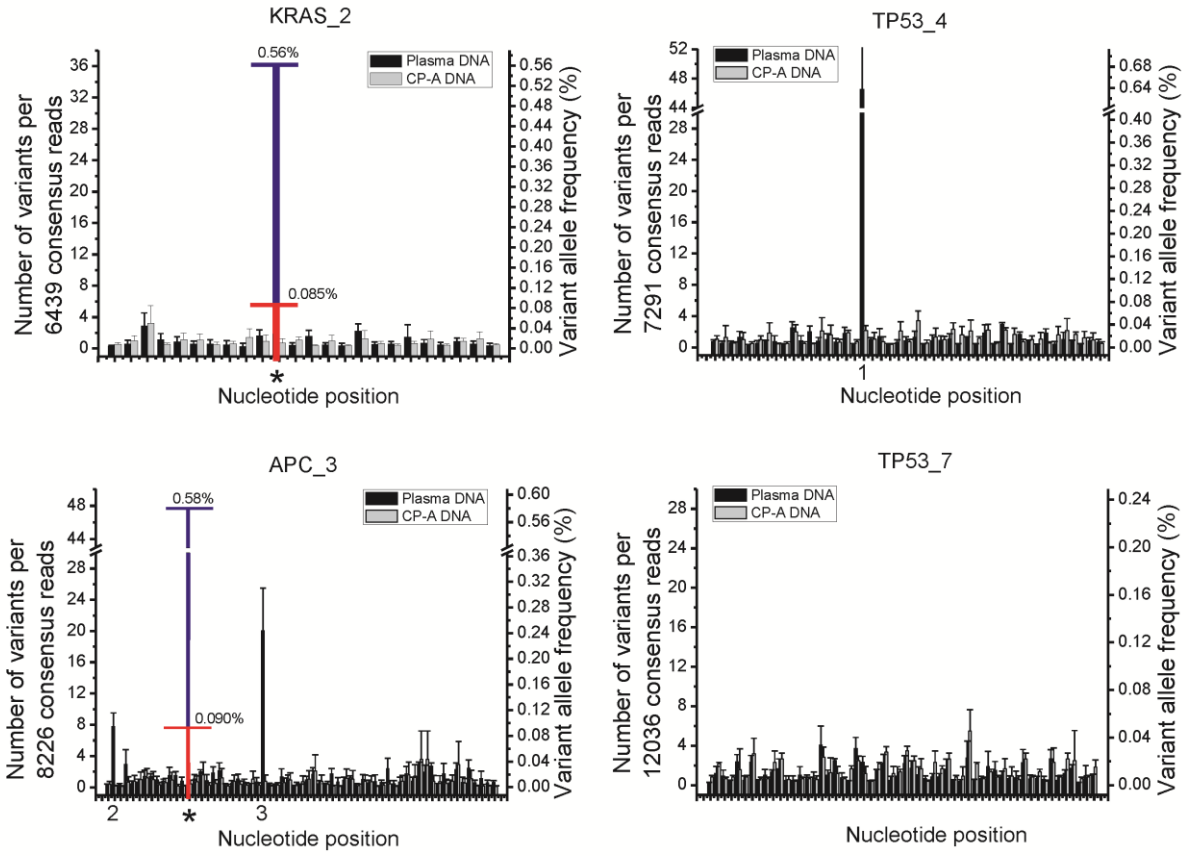

Figure S7. Rare mutation detection in *APC*, *KRAS* and *TP53*. Number of variants per nucleotide is shown with corresponding variant allele frequency on the right side y-axis. Pooled plasma DNA from more than 10 individuals and DNA from a clonal derived cell line (CP-A) were analyzed with SiMSen-Seq (n = 3 - 4). Primary tumor DNA with known mutations (marked *) were spiked into the plasma DNA using different mutated DNA concentrations using 10-fold dilution (blue and red marked bar). Additional variants are indicated by number. These variants most likely originated from the plasma DNA and not the spiked in tumor DNA, since their frequencies remained almost constant regardless the amount of spiked in primary tumor DNA. Detailed variant analysis is shown in Table S2. For most experiments, we used clonally derived cell line DNA in order to minimize the amount of true, low-level mutations. Interestingly, when we changed to plasma DNA for spike-in experiments, we identified several base positions with consistent variant allele frequencies above background (0.10-0.64%). Plasma used for this experiment was purchased from a commercial provider and is allegedly pooled from blood of healthy individuals. Our data suggests that there may be biological background (true low-level variations) in plasma DNA that occur at variable allele frequency. If true, understanding this background among individuals will be important for applications such as early cancer detection.