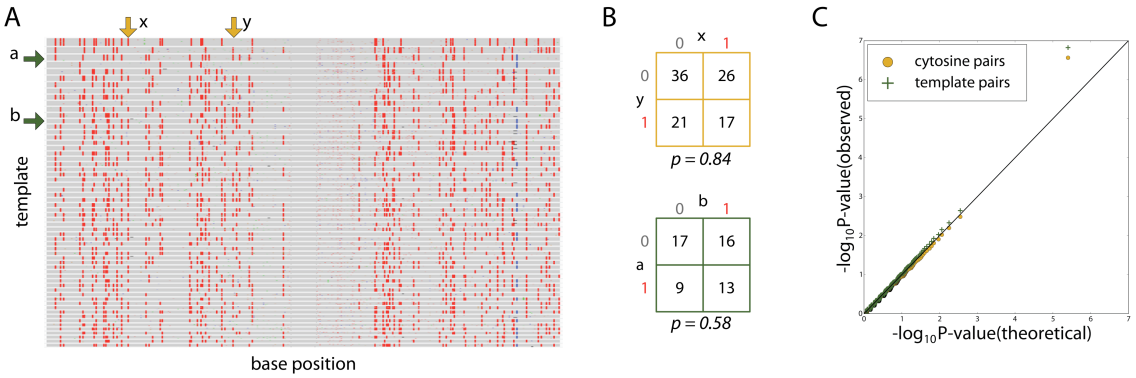
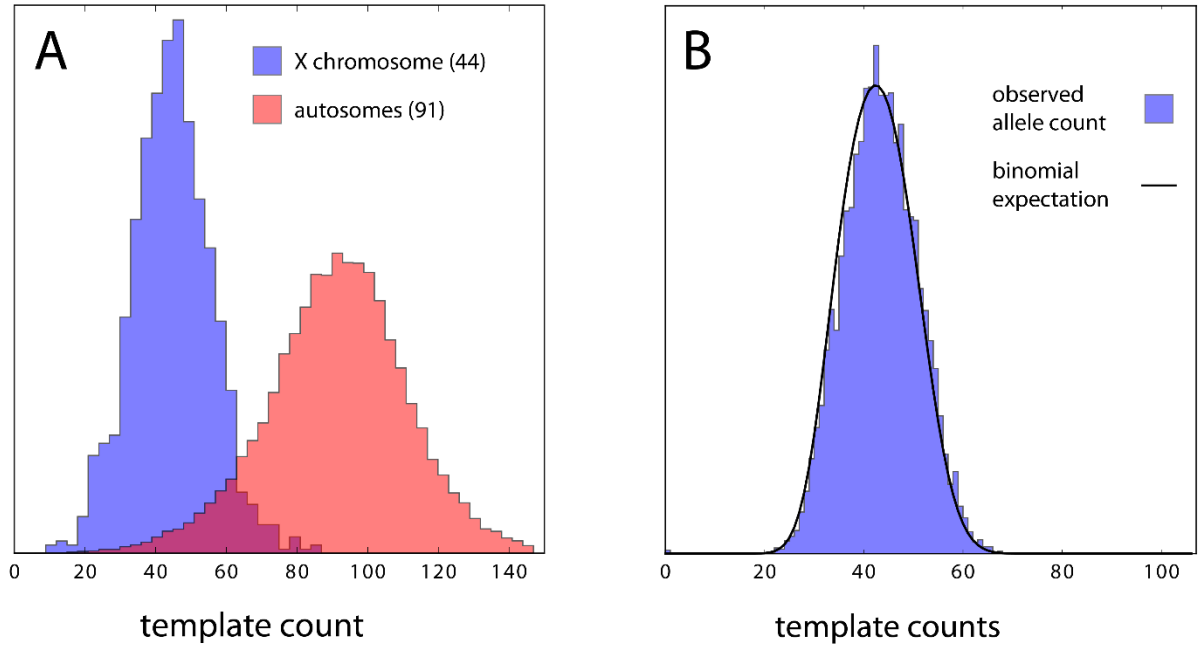
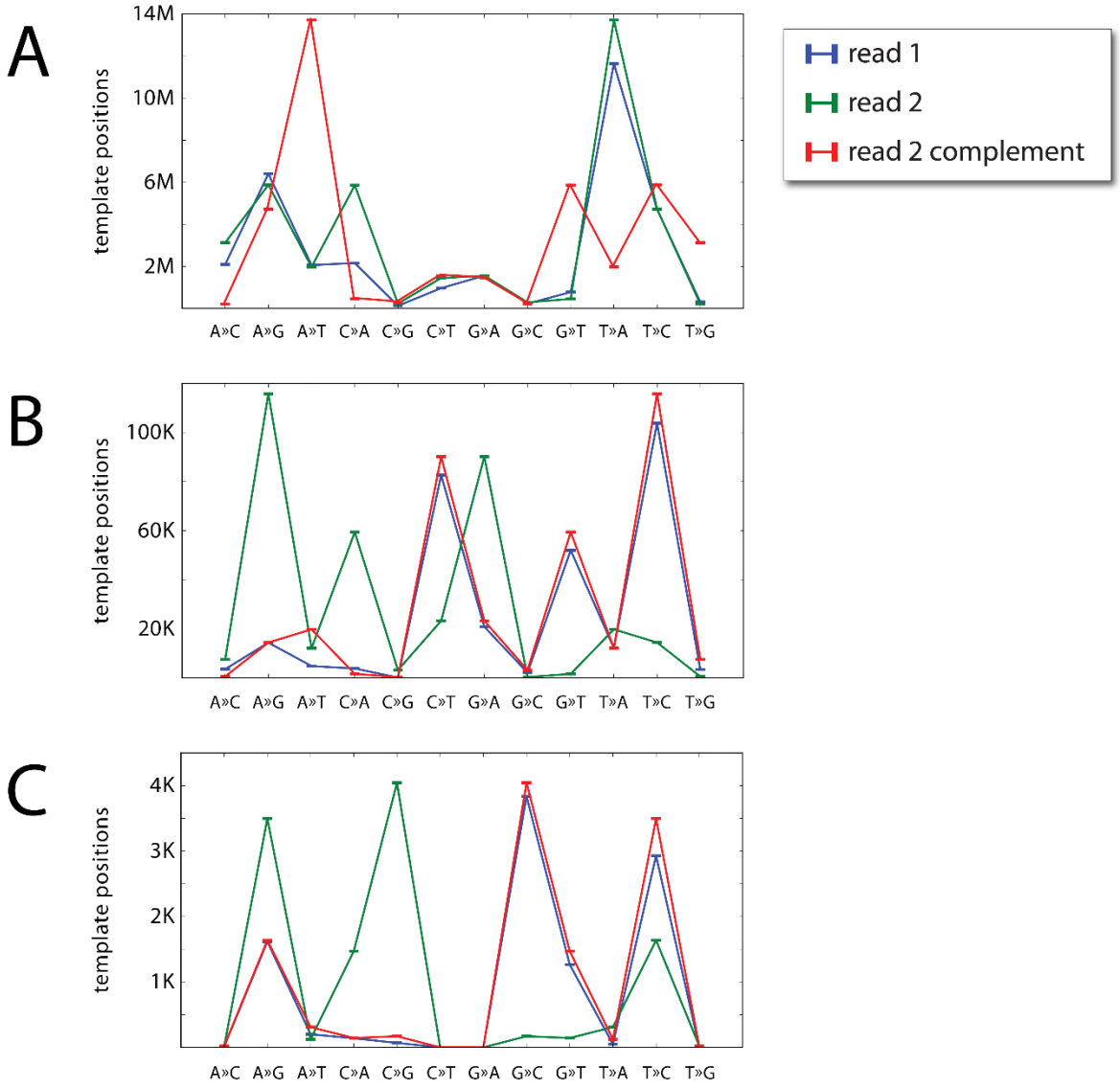
Supplementary Data



Supplementary Figure 1. Randomness of sodium bisulfite conversion.
(A) An IGV screenshot of muSeq reads from the same 320-bp restriction fragment as in Figure 2C. For each pair of bit positons (yellow arrows) and templates (green arrows), we determine a contingency table, and **(B)** apply the Fisher exact test to measure the probability of the observation under the assumption of independence (italics). Examining 1000 randomly selected fragments generates 2.9 million pairwise-bit comparisons and 5.1 million pairwise-template comparisons. **(C)** A Q-Q plot is shown with the sorted observed Fisher exact p-values against the sorted theoretical p-values, assuming independence.

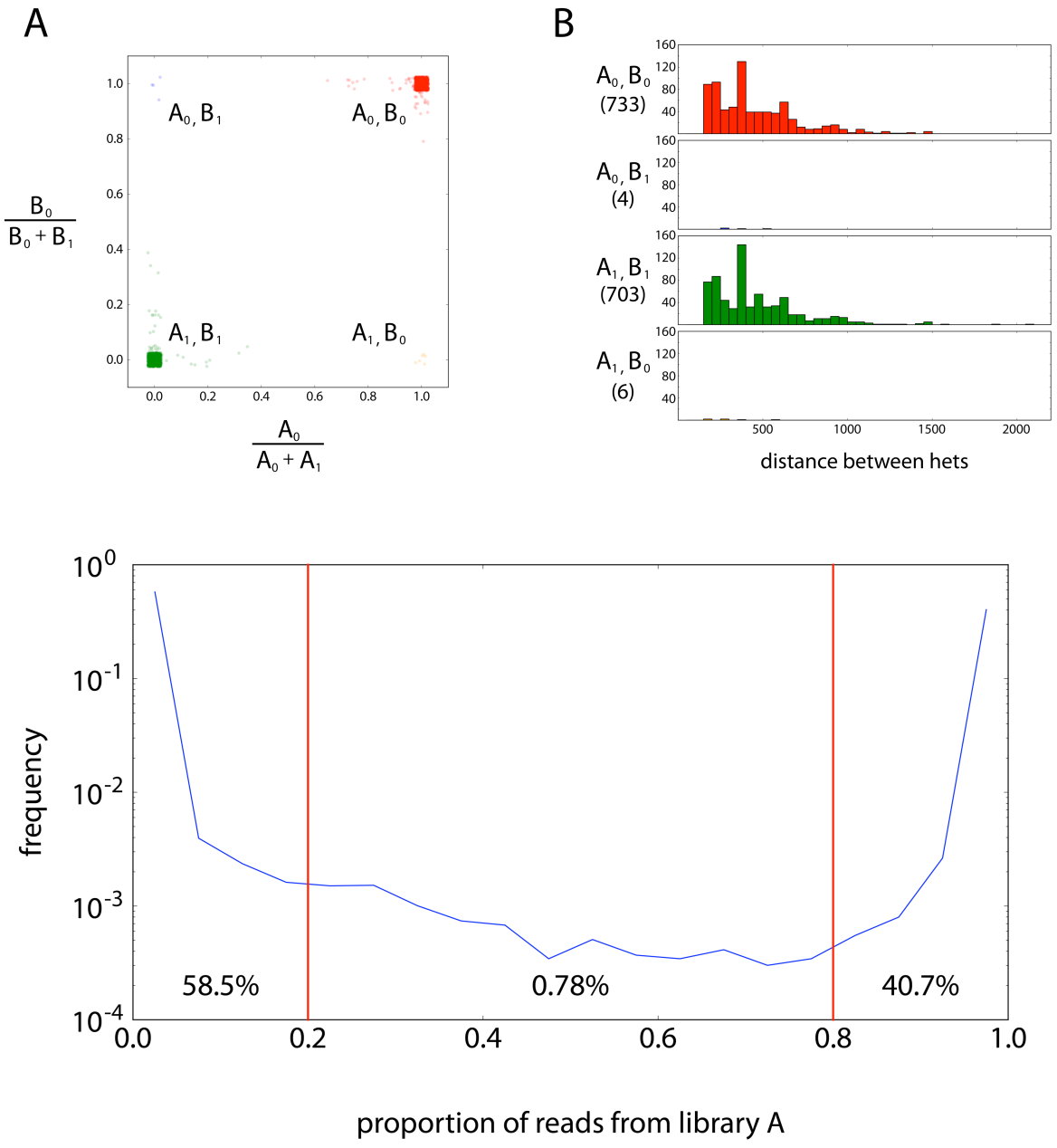


Supplementary Figure 2. Template count distributions.
The 6 minute muSeq library was sampled at a concentration of about 91 haploid genomes. This has two measurable consequences: **(A)** first, fragments from the autosomes and pseudo-autosomal regions have counts of 91 on average (red histogram), whereas fragments derived from the proper X chromosome have half as many copies (blue histogram); and **(B)** fragments that contain heterozygous loci should show the alternative allele half the time. The blue histogram shows the distribution of template counts for the non-reference allele over 6310 fragments with heterozygous loci. The black curve reveals the aggregate distribution of expected counts assuming binomial distributions (p=0.5) over all loci, each conditioned on the total coverage.

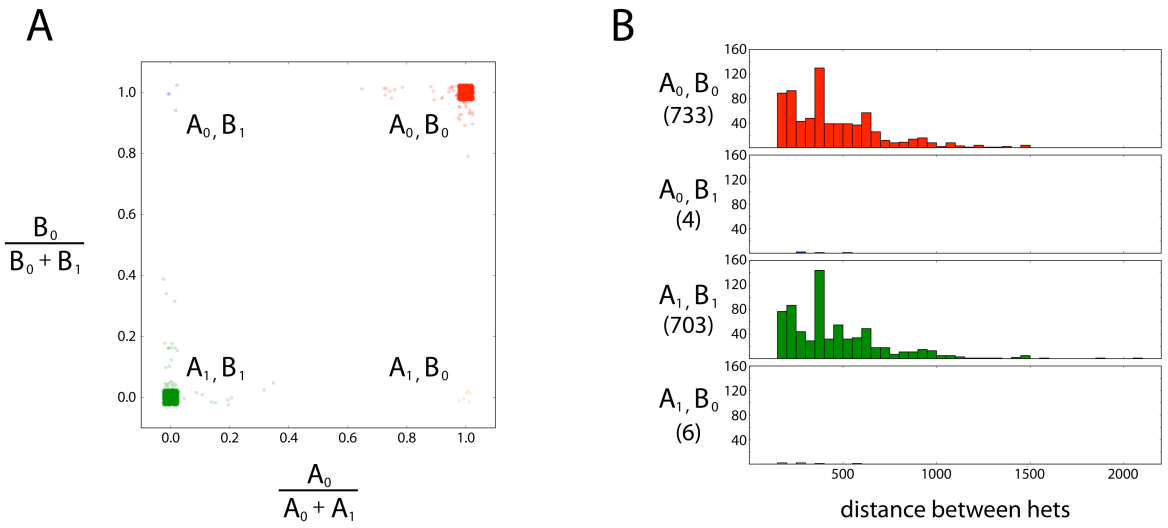


Supplementary Figure 3. Base conversions by error type.
In Figure 3, we divide the space of mutational errors into three types: (1) type A errors which are almost entirely the reference base; (2) type B errors which have non-trivial signal for both reference and non-reference bases; and (3) type C errors that are almost entirely non-reference. Here we show each type of error and examine the number of template positions (y-axis) for which we observe a particular conversion (reference >> alternate, x-axis). Conversions may be considered two ways: (1) with respect to the sequence of the machine generated reads, read 1 (blue) and read 2 (green); or (2) with respect to the sequence of the template molecules, read 1 as before but now paired with the complement of read 2 (red). **(A)** Type A errors have strong agreement between read 1 and read 2, suggesting that these conversions are largely the result of the sequencing instrument. **(B)** Type B errors have strong

agreement across template symmetries, consistent with early PCR synthesis error. **(C)** Type C errors omit bit conversions (C >> T and G >> A); this too shows template symmetry, but one that is not identical to that seen with type B errors. Type C errors are not symmetric within the template and therefore unlikely to be double-stranded somatic mutations, leaving initial template damage as the likely source for this type of conversion.



Supplementary Figure 4. Library mixing.
We prepared two sequencing libraries (A and B) from two separate aliquots drawn from the same converted cDNA. The reads in each library have unique barcodes typically used to multiplex libraries in a single sequencing lane. To measure the success of conversion patterns in correctly clustering reads from the same template, we processed the two libraries as one. Ideally, each cluster would then have reads from only one library. We restricted to 116,000 clusters with ≥20 reads and recorded the proportion of reads from library A. More than 99.2% of clusters were almost entirely from a single library. The plot shows a log histogram of the observed ratios.



Supplementary Figure 5: SNP phasing across assembled transcripts.
From previous experiments (single cell sperm sequencing and family information), we had prior knowledge of haplotype phasing for many heterozygous loci ("hets") in our donor sample. For two such loci A and B, we denote the heterozygous bases for A as A_0 and A_1 and for B as B_0 and B_1. The letter refers to the locus, and the number indicates the haplotype such that A_0 and B_0 occur on the same molecule (e.g. A_1 and B_1). For each transcript cluster passing quality controls for mixing between samples (<0.05) and tags (<0.1), we identify regions of contiguous sequence ("contig") in which every base is either observed in a read or as a spliced intron. For every pair of hets in a contig, we computed the distance in the transcript and recorded the base counts for A_0, A_1, B_0 and B_1. We considered only those pairs in which each locus was covered by ≥2 reads and the loci are at least 150 bp apart, so as to eliminate het pairs in the same read. **(A)** We show a scatter plot of the ratios $A_r = A_0 / (A_0 + A_1)$ on the x-axis compared to $B_r = B_0 / (B_0 + B_1)$ on the y-axis, with some jitter added to distribute points at the corners. The four quadrants in the plot indicate the most likely configuration given the observed ratios. The vast majority of points are consistent with known phase information, occurring in either in the (A_0,B_0) or (A_1,B_1) quadrants (733 and 703 pairs, respectively). In contrast, the quadrants that disagree with the known phase, (A_0,B_1) and (A_1,B_0), have 4 and 6 pairs, respectively. **(B)** We show the distribution of distances between hets for each of the four quadrants. The quadrants match the color in the scatter plot and are labeled with the most likely het pair called. Beneath the label, we show the number of pairs for each quadrant in parentheses. Note that we obtained correct phasing for hets that are widely separated (over 2000 bp) in the template.

Supplementary File. Transcriptional variants (transcriptional variants.pdf).
Similar in structure to Figure 5, we included auto-generated illustrations for genes with ≥1 muSeq cluster >500 bp in length. There are several differences with respect to Figure 5. Here we show all proper and complete assemblies. Gaps within an assembly that are due to splice junctions are shown as green lines. Gaps that are due to unobserved sequence between paired-end reads are shown as a dashed orange lines. All observed muSeq transcripts are shown in black starting at 0, whereas the annotated GENCODE transcripts fall below the x-axis and are shown in color.

Supplementary Manuscript. Kumar and Levy, 2015 (Supplementary Manuscript.pdf)
Vijay Kumar and Dan Levy, Clustering by transitive propagation (2015) **arXiv:1506.03072**.