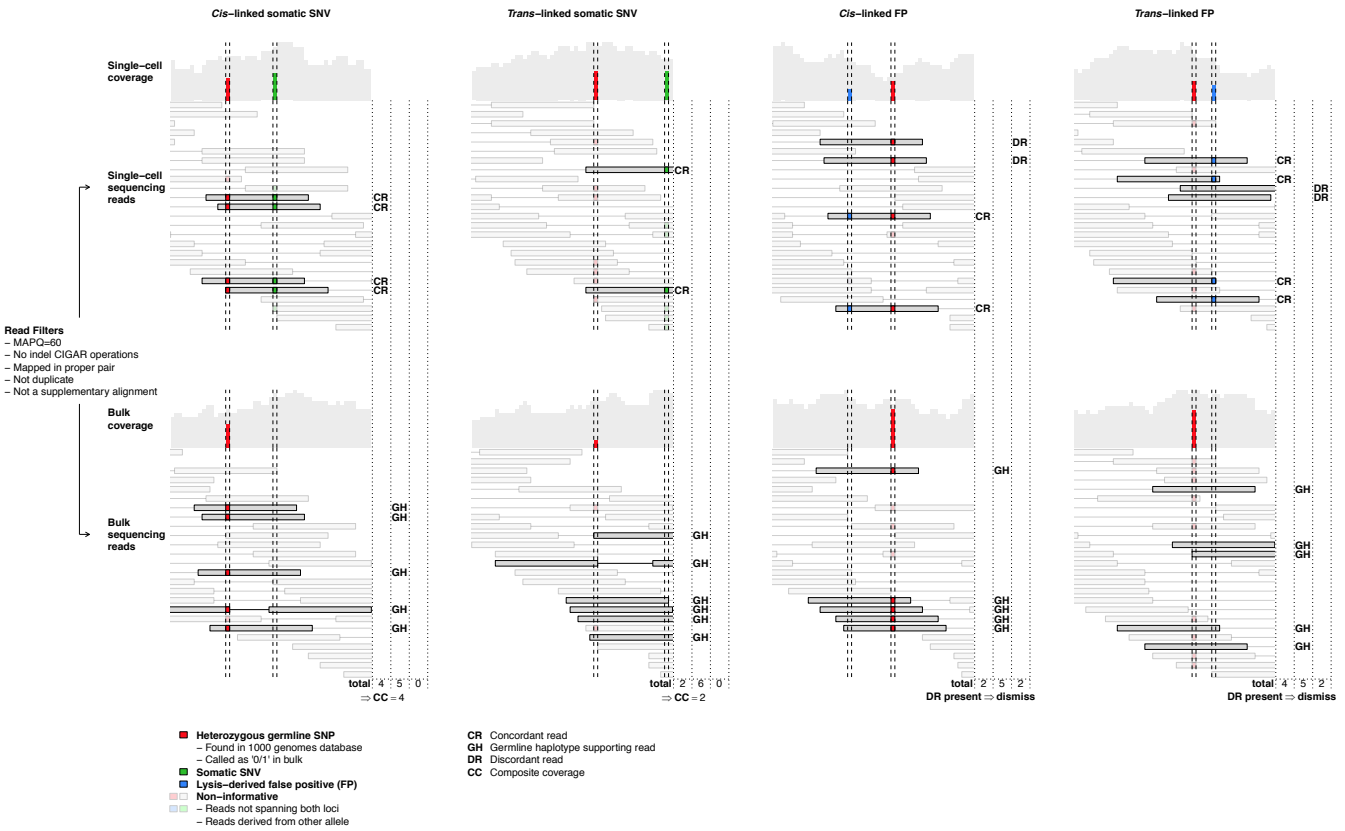


Figure S1: Linkage cases in LiRA

LiRA uses population polymorphic germline heterozygous SNPs (gHets) found in bulk sequencing data to evaluate nearby somatic SNV (sSNV) candidates. Given a set of gHets and candidate sSNVs, LiRA considers a filtered subset of reads and mate pairs covering the position of the candidate sSNV and linked gHet allele in single-cell and bulk sequencing data (spanning reads). Reads covering only the sSNV, and not the gHet, or reads derived from the unlinked allele are non-informative. For somatic SNVs, which are derived from fixed mutations, all spanning reads in single-cell data support the somatic SNV call, while the those in bulk data do not. We call these two types of reads 'concordant reads' (CRs) and 'germline haplotype supporting reads' (GHs). In contrast, a subset of spanning reads for a lysis-derived FP have a reference call at the somatic SNV position ('discordant reads'; DR). sSNV candidates with DRs present are filtered. For DR-free sSNVs, composite coverage (CC) is computed as the minimum of the GH depth in bulk and CR depth in single-cell sequencing data.

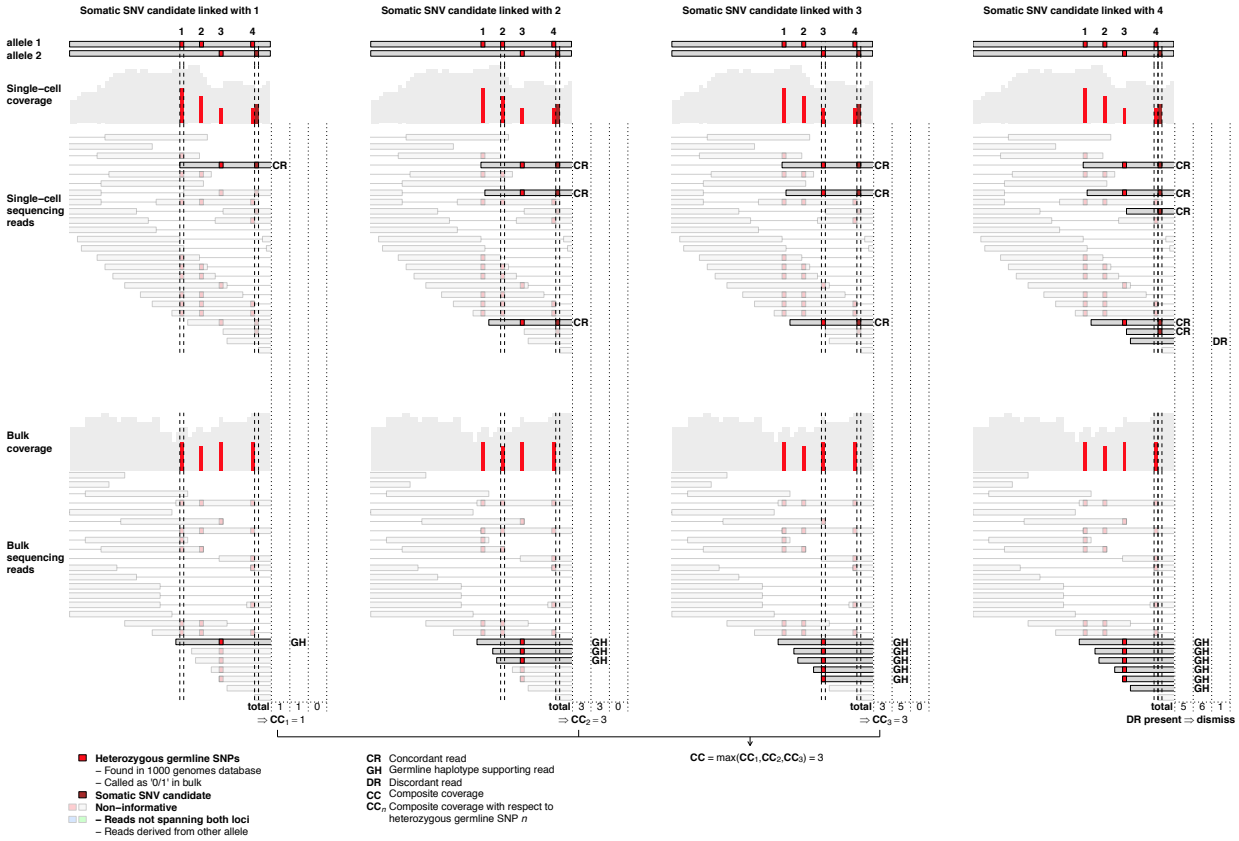


Figure S2: Resolution of linkage of an sSNV to multiple gHets
When an sSNV candidate is linked with reads to multiple gHets, LiRA considers linkage with each gHet separately.The overall composite coverage (CC) for an sSNV candidate is taken as the maximum CC measured. If there is at least one sSNV candidate-gHet pair without discordant reads, the sSNV candidate is not filtered and composite coverage is measured.

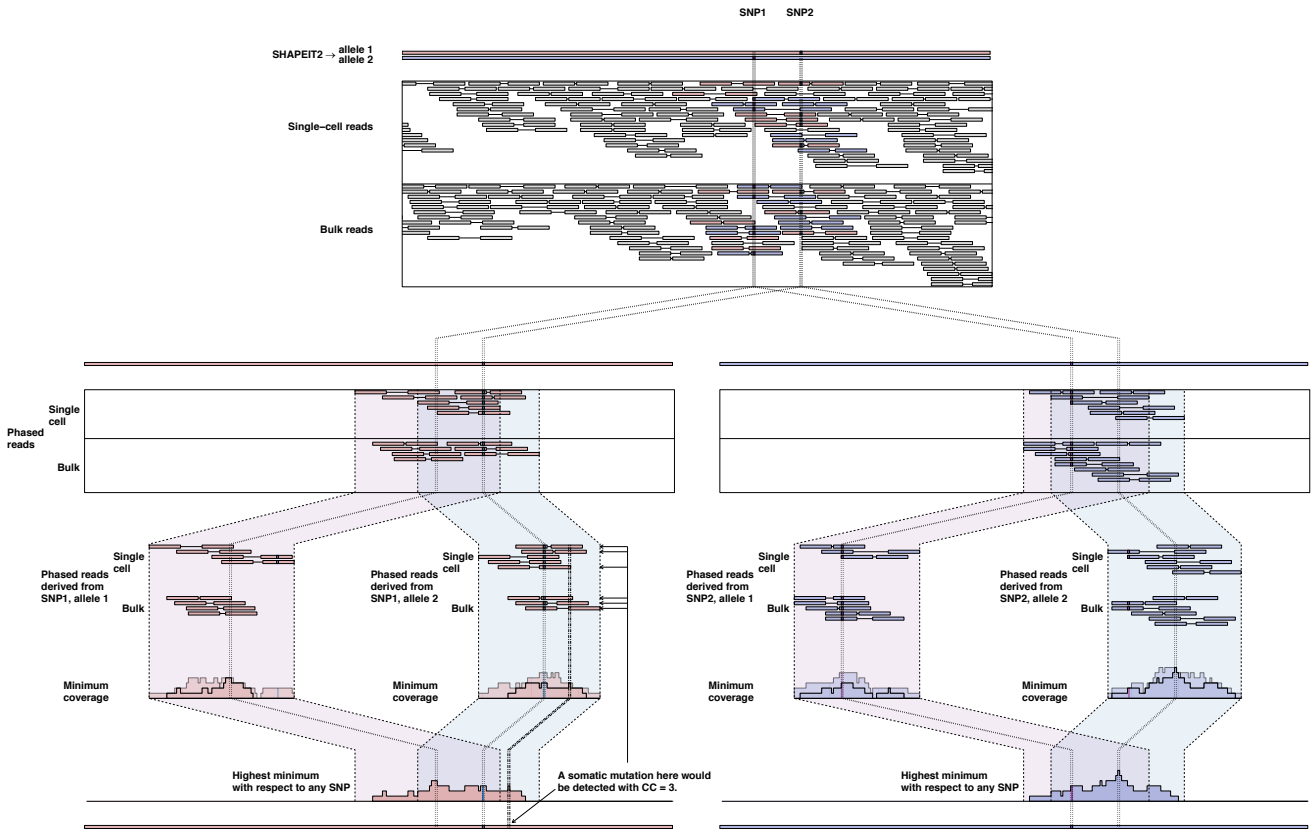
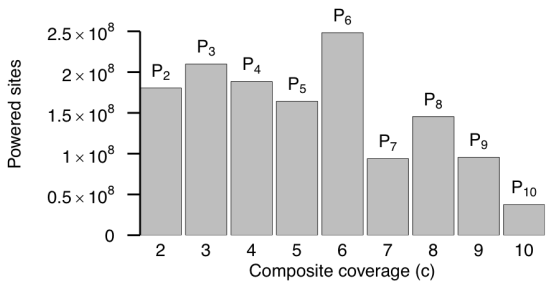
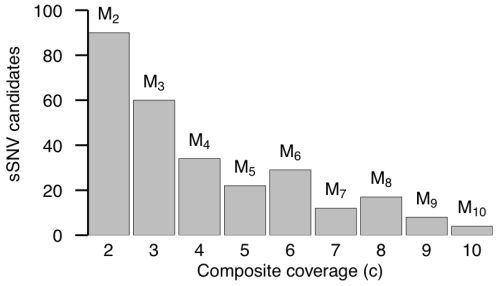


Figure S3: Measurement of power to detect somatic sSNVs
To accurately measure the sSNV rate as a function of composite coverage (CC), LiRA measures the power it had to detect an sSNV at each genomic position on both alleles in the diploid genome. To do this, it first uses SHAPEIT2 to infer the haplotype of origin for each gHet allele (ref/alt). Here the ref/alt alleles of SNP1 are phased to allele 1 / allele 2, and the ref/alt alleles of SNP2 are phased to allele 2/allele 1. Next, LiRA isolates the sets of reads from single-cell and bulk sequencing data that are derived from each allele (phased reads; pink/blue). Many reads do not align to a gHet position (gray) and as such their allele of origin cannot be inferred. Among each set of phased reads, LiRA considers the set of single-cell and bulk reads supporting each gHet (SNP1; purple, and SNP2; blue) separately, and measures the minimum depth of single-cell and bulk coverage at each position covered. Had an sSNV occurred at one of these positions, on the allele under consideration, it would have been linked with to the gHet under consideration with a CC value equal to this minimum. Finally, to account for positions that are linked with multiple gHets, LiRA takes the maximum of these minimum coverage profiles for each allele (bottom). This trace represents the overall CC value with which an sSNV at each position would have been detected, had one occurred.





$$R_c \sim B(M_c + 0.5,\ P_c - M_c + 0.5)$$

Solve $\underset{K_1 \geq 0,\ K_2 \geq 0}{\operatorname{argmin}} \sum_c \mathrm{Var}(R_c)^{-1}(K_1 E(c) + K_2 T(c) - \bar{R}_c)^2$

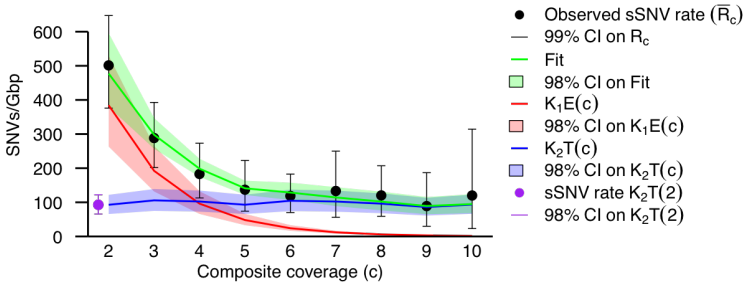
$\mathrm{Fit} = K_1 E(c) + K_2 T(c)$

for $s \in \{1, 2, \ldots, 100\}$

Sample R_c to obtain $R_c^{(s)}$

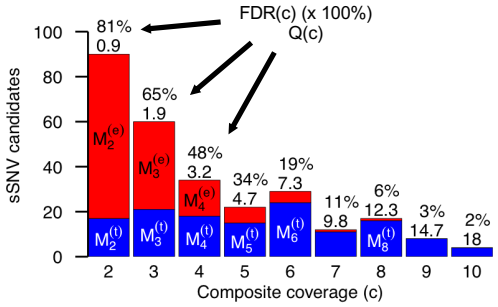
Solve $\underset{K_1^{(s)} \geq 0,\ K_2^{(s)} \geq 0}{\operatorname{argmin}} \sum_c \mathrm{Var}(R_c)^{-1}(K_1^{(s)} E(c) + K_2^{(s)} T(c) - R_c^{(s)})^2$

$\mathrm{Fit}^{(s)} = K_1^{(s)} E(c) + K_2^{(s)} T(c)$



$$\mathrm{FDR}(c) = \frac{K_1 E(c)}{K_1 E(c) + K_2 T(c)}$$

$$Q(c) = -10 \log_{10} \mathrm{FDR}(c)$$



$$M_c^{(e)} = [\mathrm{FDR}(c) M_c]$$

$$M_c^{(t)} = [(1 - \mathrm{FDR}(c)) M_c]$$

$$\mathrm{FDR}_{agg}(c_m) = \frac{\sum_{c \geq c_m} M_c^{(e)}}{\sum_{c \geq c_m} (M_c^{(e)} + M_c^{(t)})}$$

$$\Rightarrow c^* = 7$$

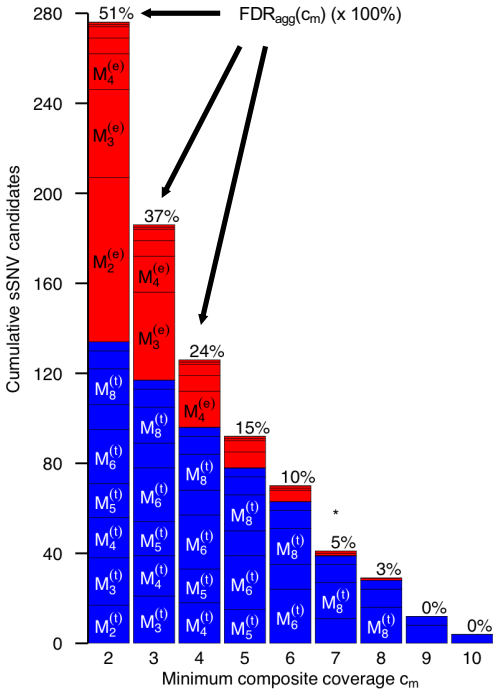


Figure S4: Two-component model workflow (simulated data)
After measuring the number of sites available for sSNV calling (P_c, upper left barplot) and the number of sSNVs observed (M_c, upper right barplot) at each level of composite coverage c, LiRA computes the sSNV rate as a function of composite coverage (black points). LiRA models this observed rate as the linear combination of two components (Fit, green): an error component ($E(c) = K_1(1/2)^{c-2}$, red line) and a 'true mutation' component (T(c), blue line), the expected relationship between the observed rate and c if the observed sSNV set were composed entirely of fixed heterozygous mutations. T(c) is generated by sampling linked heterozygous germline mutations (online methods). To assess robustness, LiRA then applies this same procedure to samples from the beta distributions of the sSNV rate at each level of composite coverage. The ranges of values obtained over 100 samples for the error component (min/max of $K_1^{(s)}E(c)$), the 'true mutation component' (min/max of $K_2^{(s)}T(c)$), and the overall fit (min/max of $Fit^{(s)}$) are shown in corresponding transparent colors and constitute 98% confidence intervals. Then, using the relative values of E(c) and T(c), LiRA estimates the false discovery rate at each level of composite coverage FDR(c), and uses this to choose a threshold c value such that the FDR across all accepted mutations is less than 10%.

Figure S5: Two-component model fits for each single cell analyzed

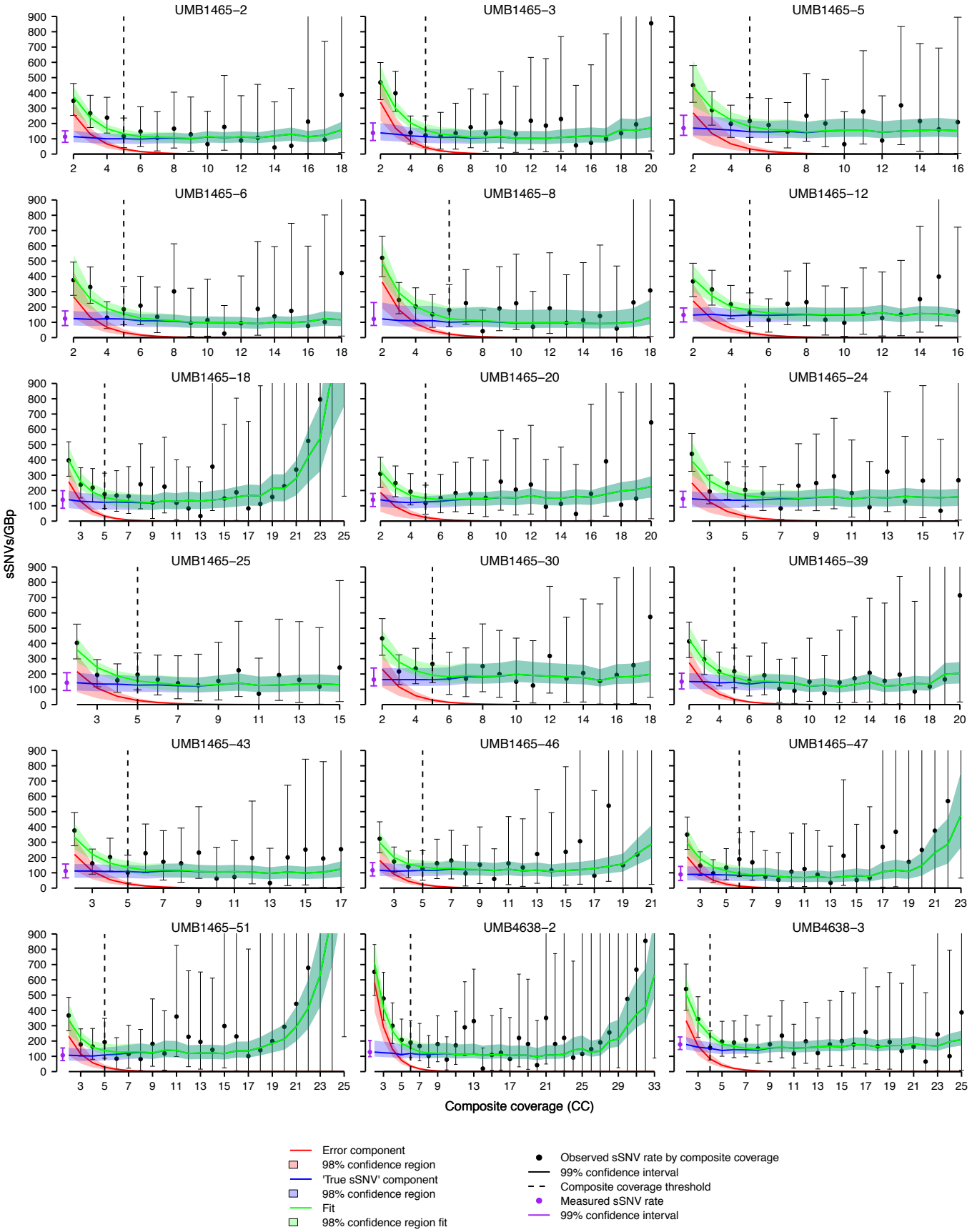
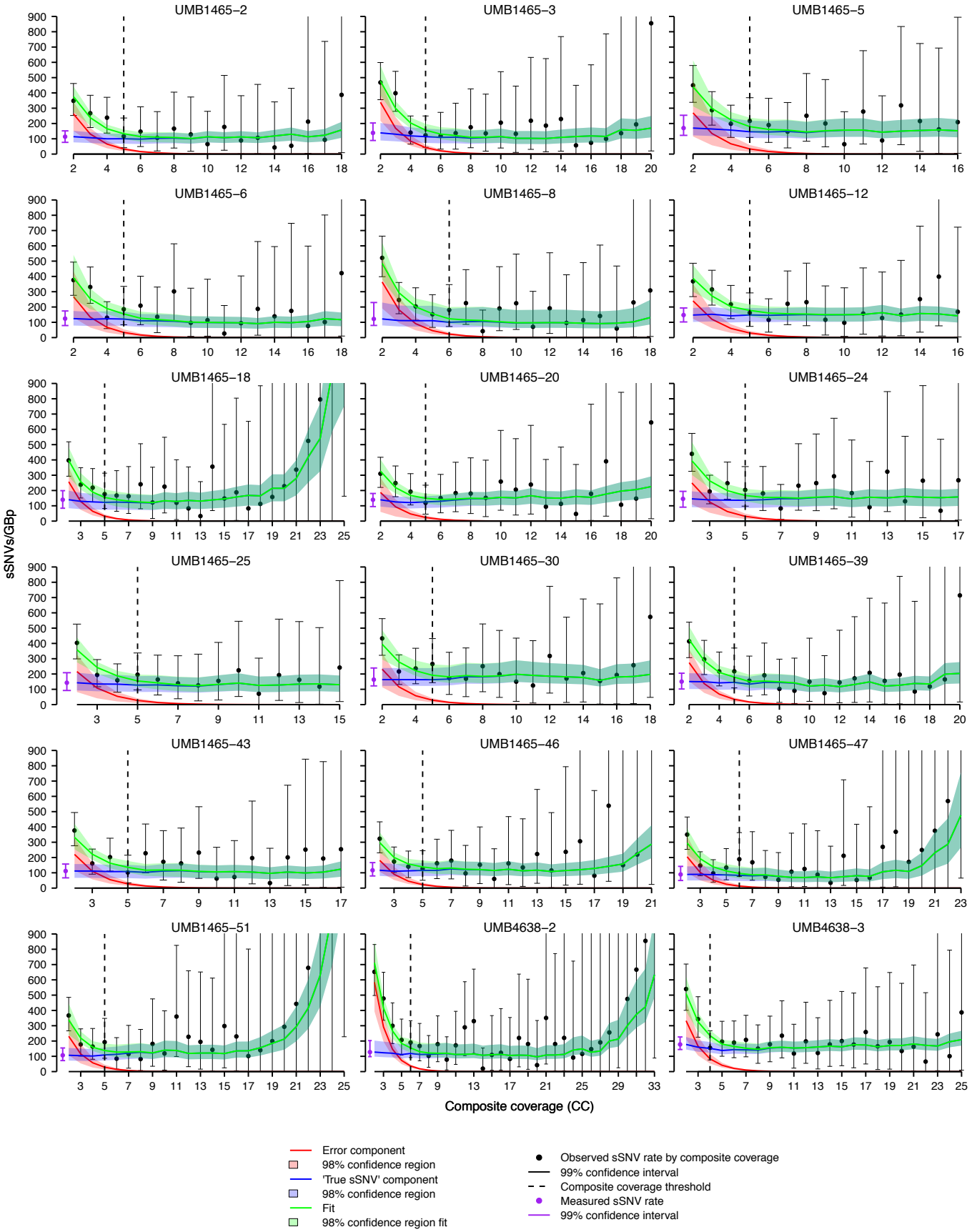


Figure S5: Two-component model fits for each single cell analyzed



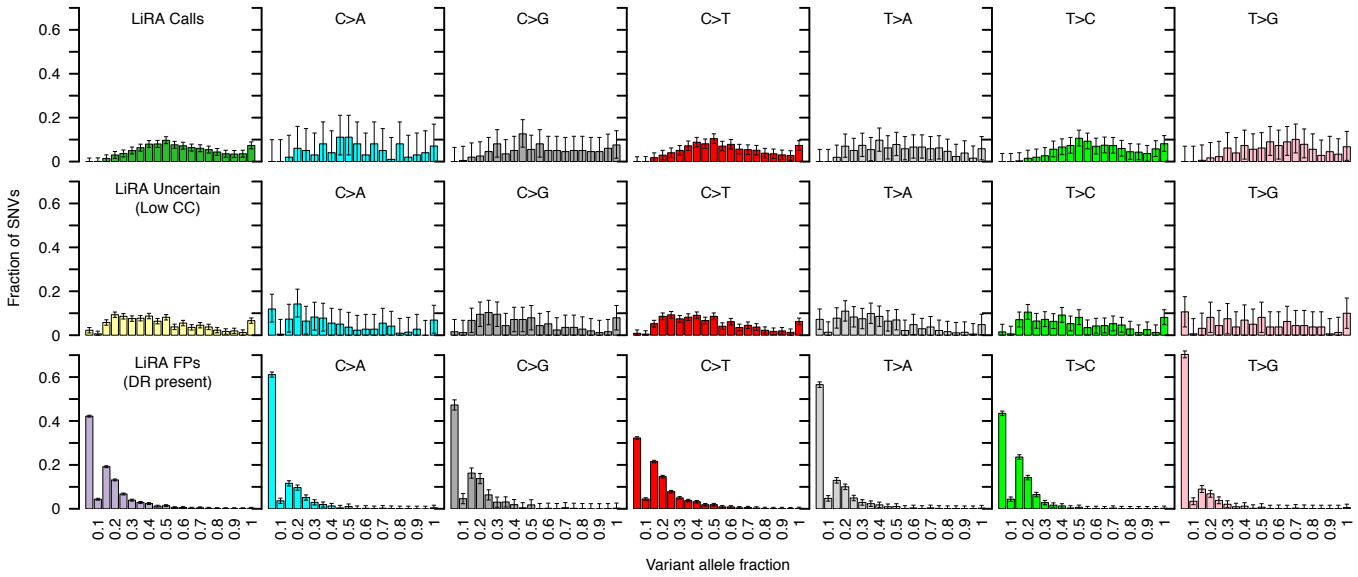


Figure S6: Variant allele fraction distributions for LiRA calls, LiRA uncertain calls, and LiRA FPs by SNV type (99% CI on SNV frequencies) Over individual mutation types, LiRA calls remain centered about ~0.5 variant allele fraction, consistent with gHet calls. Uncertain calls and to a greater extent FPs are shifted towards lower variant allele fraction values.

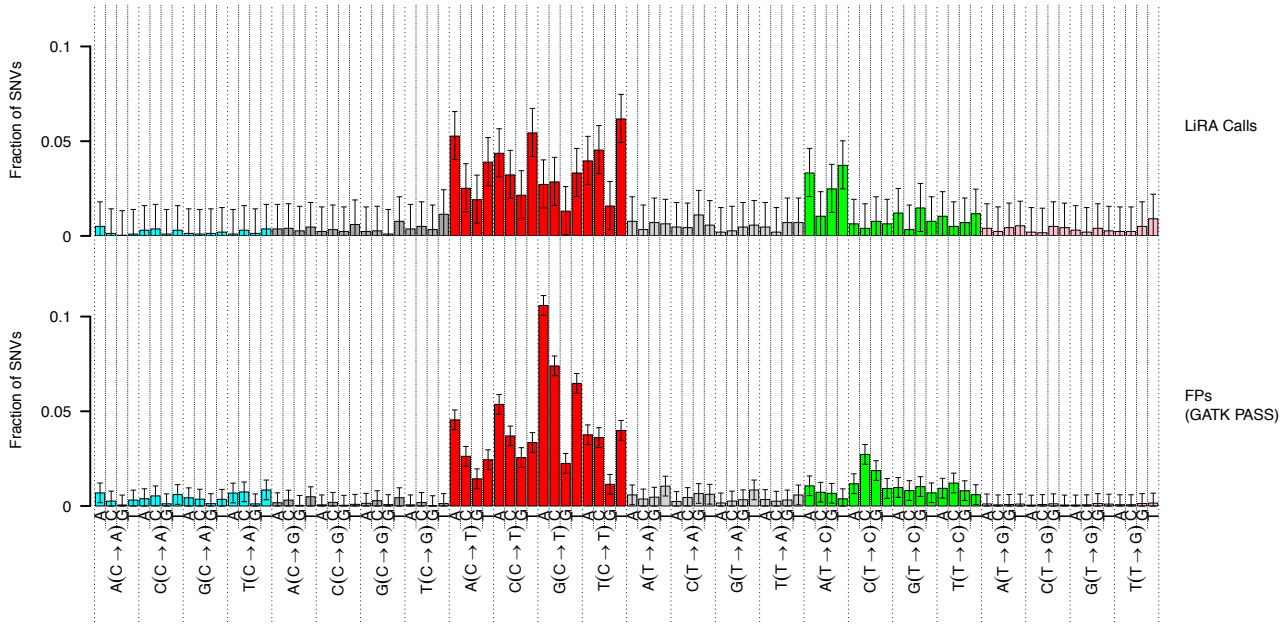


Figure S7: Trinucleotide context of LiRA calls and LiRA FPs (99% CI on SNV frequencies) LiRA calls and high-quality (filtered as GATK PASS) LiRA FPs have different mutational frequencies when viewed by trinucleotide context.

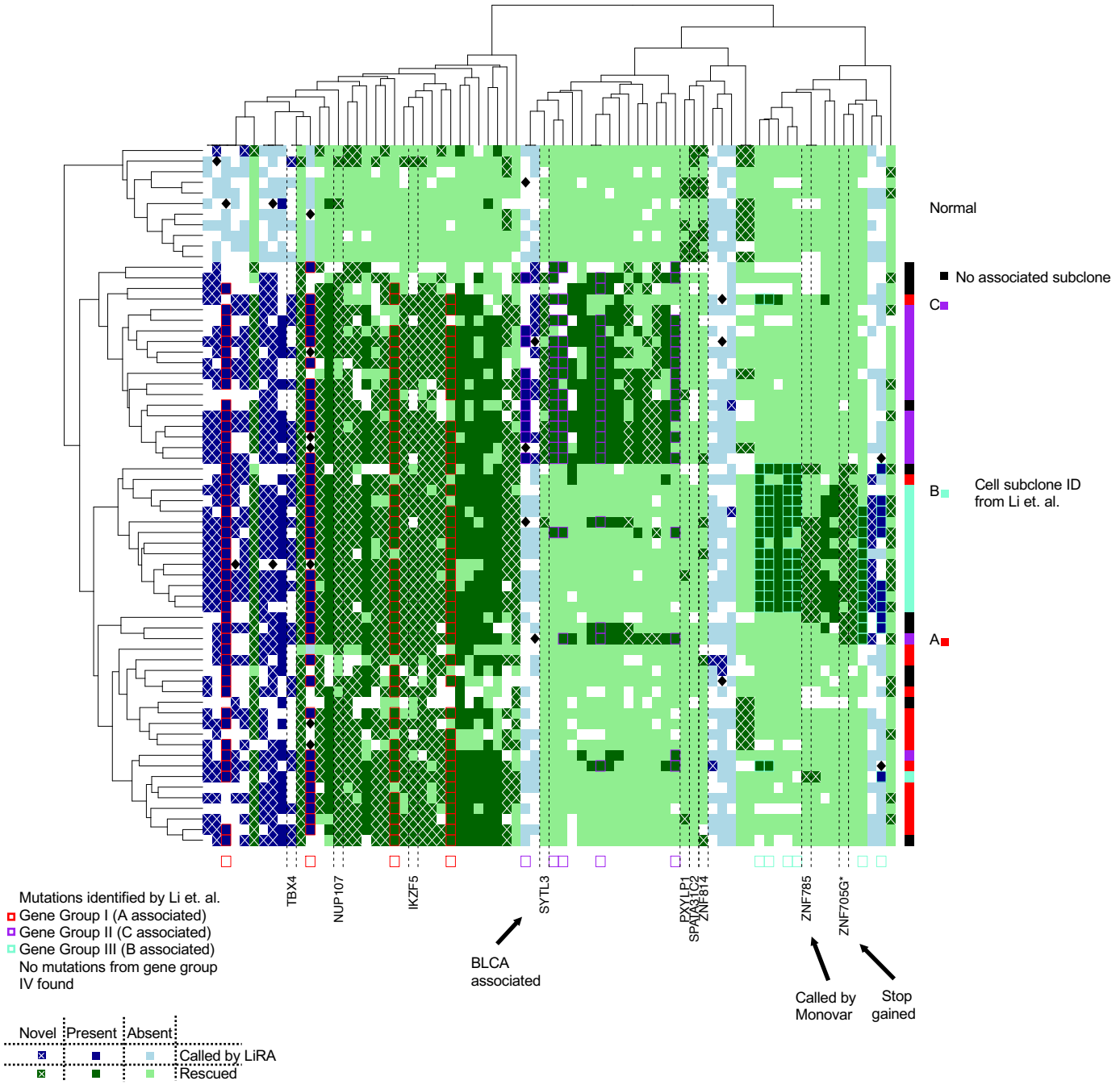


Figure S8: Analysis of Li et. al.[1] bladder cancer exome dataset using LiRA
Hierarchical clustering of bladder cancer and normal cells using LiRA calls (blue) and highly associated, 'rescued' unlinked sSNV candidates. Calls described by Li. et. al. are boxed in the color of their corresponding gene group, while novel calls found here are marked with a white X. The subclone assignment given in Li et. al. for each cancer cell is shown on the right. Coding mutations identified among LiRA calls or rescued sSNV candidates are labeled with the gene in which they occur. Calls in exons not made by Li. et. al. are labeled with gene names. With the exception of the call in ZNF785 labeled above, none of these were called by Monovar.[2] One gene in which a novel call was found (SYTL3) has been previously investigated in bladder cancer.[3]

1. Li, Y. *et al.* Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaSci* **1,** 69 (2012).

2. Zafar, H., Wang, Y., Nakhleh, L., Navin, N. & Chen, K. Monovar: single-nucleotide variant detection in single cells. *Nat Meth* **13,** 505–507 (2016).

3. Ho, J. R. *et al.* Deregulation of Rab and Rab Effector Genes in Bladder Cancer. *PLoS ONE* **7,** e39469–16 (2012).