Supplementary Information for:

Haplotype-based profiling of subtle allelic imbalance with SNP arrays

Selina Vattathil[1,2*] Paul Scheet[2,1]

[1]Human & Molecular Genetics Program, The University of Texas Graduate School of Biomedical Sciences, Houston, TX 77030

[2]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

*corresponding author: svattathil@utexas.edu

Supplementary Figures

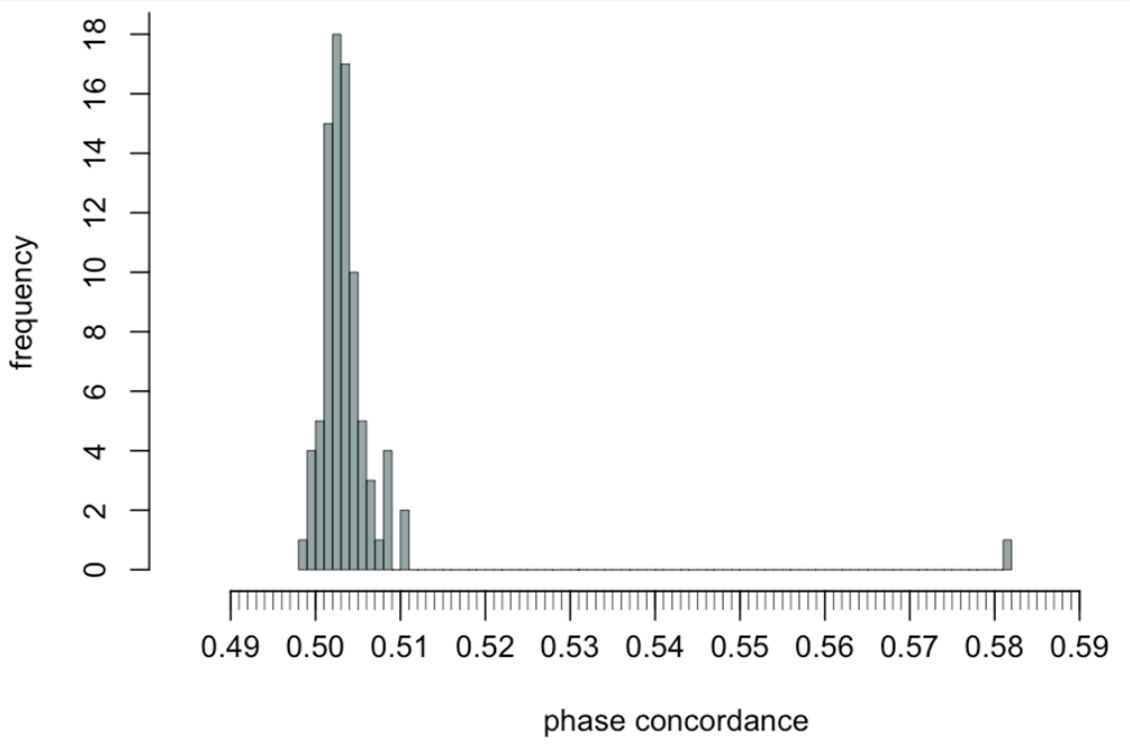

Figure 1: *Distribution of phase concordance for 86 normal samples.* We determined phase concordance rates for 86 paired normal samples from individuals with hepatocellular carcinoma. We downloaded BAFs and genotypes obtained using the Illumina HumanCNV370-Duov1 BeadChip array from GEO (accession GSE32649). We estimated sample haplotypes with fastPHASE using parameters estimated from 120 HapMap CEU haplotypes, as described in the main paper. The BAFs from the sample with the highest phase concordance rate exhibit obvious multi-modality and indicate the sample contains relatively high (in the range of 10%) tumor content.

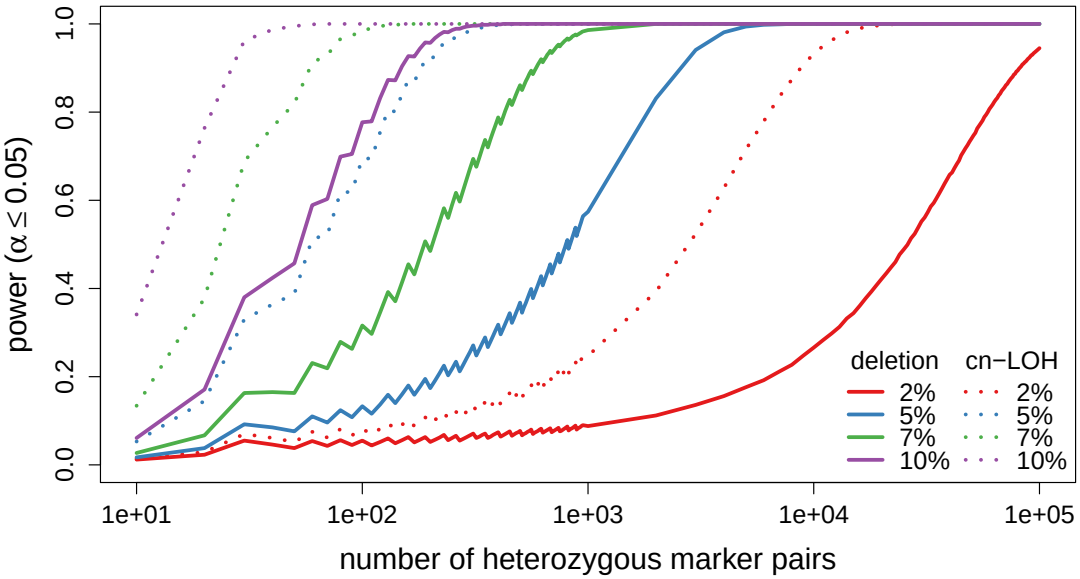

Figure 2: *Power to detect allelic imbalance.* We show power as a function of the number of informative sites in the tested region, assuming the entire tested region was affected by AI, for hemizygous deletion and cn-LOH at four tumor proportions. Power will be lower if imbalance only affects a portion of the tested region. The results were calculated using the phase concordance rate per event type and tumor proportion observed from the simulated CRL-2324 data described in the main text. The maximum false positive rate allowed was 5%; the sawtooth appearance of the curve at the smaller sample sizes is due to fluctuation in the actual false positive rate due to the discreteness of the binomial probability distribution.

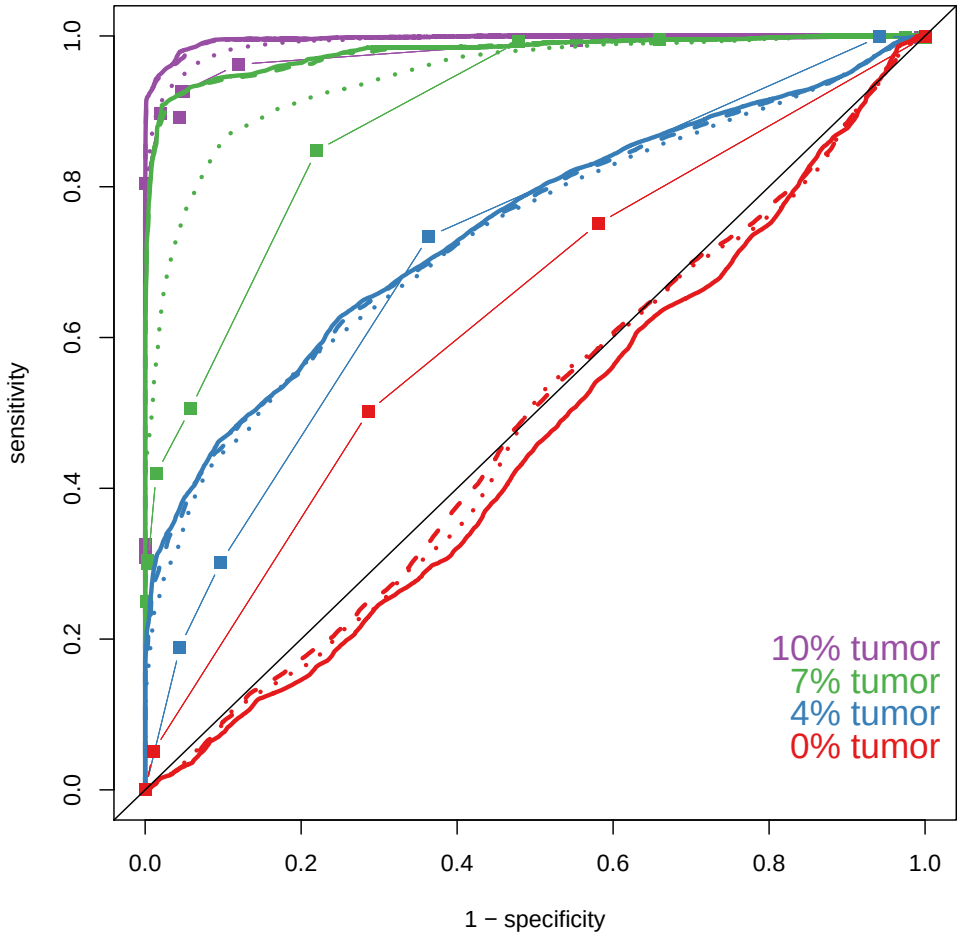

Figure 3: *ROC plot.* *BAFsegmentation* results are plotted as squares (with a thin line), and *hapLOH* results are shown as thick lines: solid for runs using an estimated transition probability matrix (TPM) with pseudocounts corresponding to mean event length of 20 Mb (as described in Methods), dashed for runs using an estimated TPM with pseudocounts corresponding to mean event length of 5 Mb, and dotted for a run using a fixed TPM assuming a mean event length of 5 Mb. To calculate sensitivity, we classify a call as correct if it identifies AI, regardless of whether it correctly identifies the event type. *BAFsegmentation* performance was assesed by running the method at a series of mBAF thresholds (a parameter of their algorithm; the default is 0.56). One curious observation is that the *BAFsegmentation* curve for the 0% tumor sample is elevated from the diagonal. This inflation is not replicated by our method. We note that all of the events identified by *BAFsegmentation* for this control sample cover nearly the entire chromosome on which they are located, and they occur disproportionately on chromosomes with high event proportion in the curated tumor sample. We postulate that the results reflect an artifact of one of the normalization procedures applied to the data. If this is the case, this apparent inflation of sensitivity may be propagated in results for the lower tumor proportions as well. The true mean event length in this dataset is 24 Mb, so the prior of 5 Mb is considerably smaller than the truth. The results for the 7% and 10% tumor samples indicate that if the prior on the event size is poorly specified, and the data are sufficiently informative, allowing the method to obtain data-driven estimates of the transition probabilities should yield an improvement in results. At 4% tumor, however, where the data are more noisy, estimation does not improve results even when the mean event size distribution is rather poorly specified. We also ran the method using fixed transition probabilities corresponding to 20 Mb mean event length. The ROC curve for those results looks like the curves shown here for the runs using the estimated parameters.

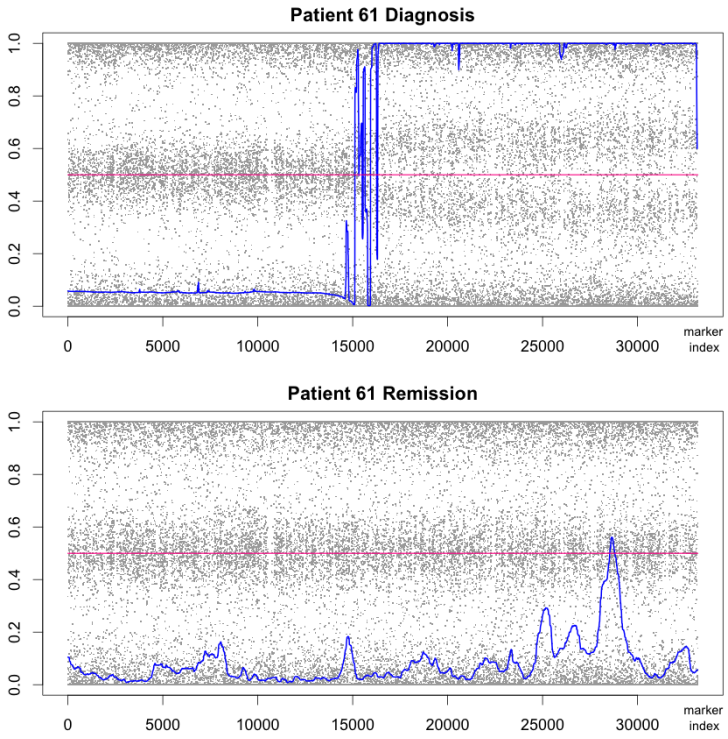

Figure 4: *Large partial cn-LOH event.* BAFs and posterior probability of AI for chromosome 11 in Patient 61. The pink line at 0.5 indicates the threshold posterior probability value used to call events.

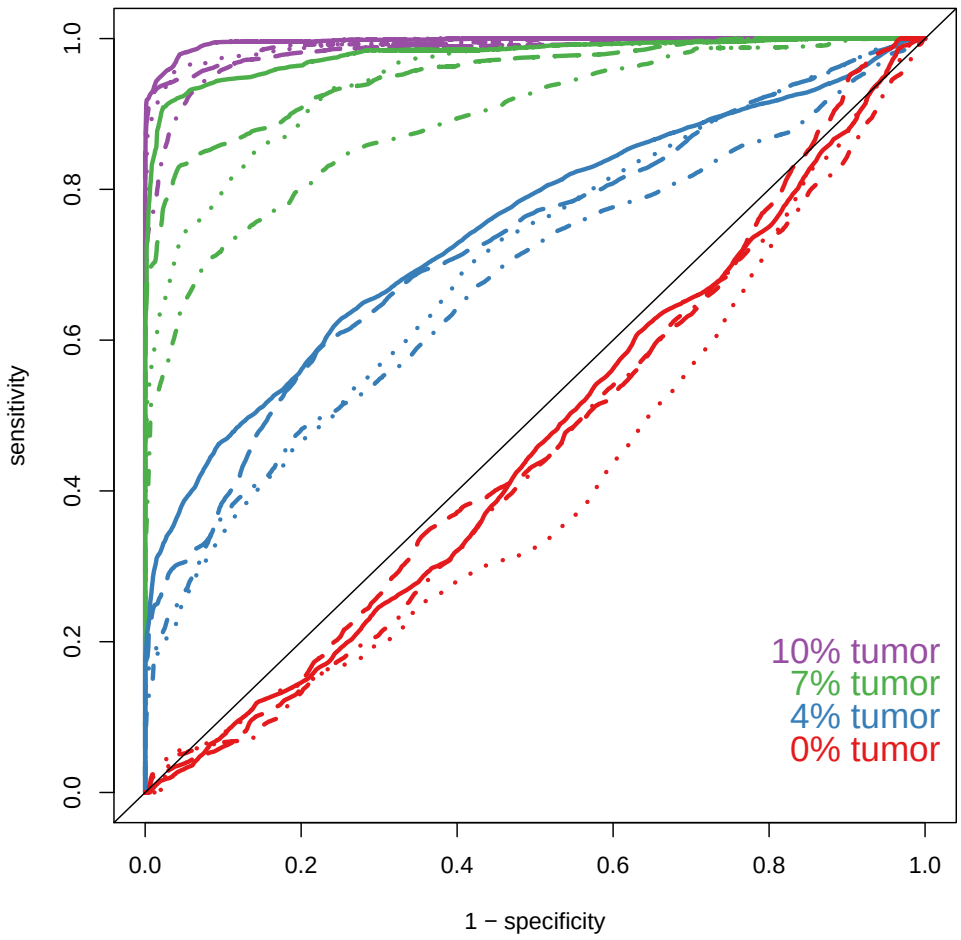

Figure 5: *Robustness to germline phasing errors.* The *fastPHASE* germline haplotype estimates used for the analyses throughout this paper have switch accuracies of about 93% (estimated by *fastPHASE*). To investigate the robustness of the method to errors in the statistical germline phasing, we introduced additional switch errors into the *fastPHASE* haplotypes and calculated ROC curves for hapLOH localization results as in Supplemental Figure 3. The HMM was fit as described in Methods. Results are shown for runs using the original data (solid lines), and for runs using the fastPHASE estimates with switches introduced independently at each marker interval at rates of .05 (dashed lines), .10 (dotted lines), and .15 (dashed-dotted lines). Phasing errors decrease sensitivy slightly, especially in the lower tumor proportions. We note that state-of-the-art phasing packages generally have switch accuracies of 90% or higher.

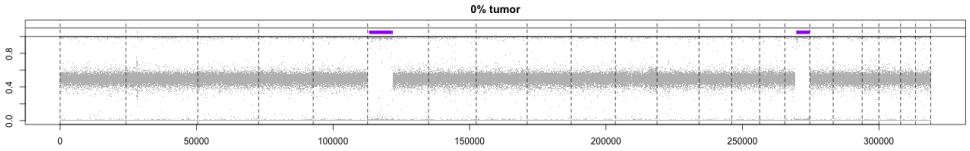

Figure 6: *Uncurated normal sample.* *BAFsegmentation* was run using default mBAF threshold 0.56 on the normal sample.

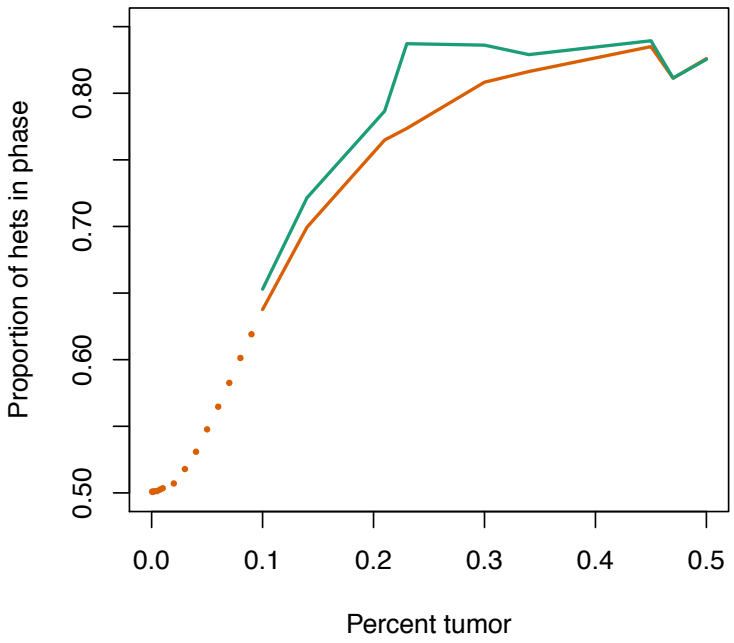

Figure 7: *Calibration of computational simulations of BAFs.* The total proportion of heterozygotes over which haplotypes from BAFs and statistical estimation are "in phase" is plotted by percent tumor. The real data (_) and computational dilutions (_) show similar patterns. The computational dilutions are slightly conservative, indicating that perhaps more of the events in the real data are due to cn-LOH than what we assumed for these dilutions.

Supplementary Tables

sample	startsnp	endsnp	chr	startbp	endbp	event size (Mb)	informative marker count	observed concordance rate	-log10(p-value)
Patient_9_Dx	rs17025785	rs9813622	3	30,642,429	30,723,944	0.08	29	0.93	6.1
Patient_9_Dx	rs7630256	rs12486635	3	67,867,664	79,280,400	11.41	1001	0.89	153
Patient_9_R	rs17649641	rs8080254	17	41,353,200	42,702,891	1.35	120	0.75	7.7
Patient_9_R	rs2825523	rs2826530	21	19,627,751	21,108,259	1.48	111	0.67	3.5
Patient_21_R	rs7528118	rs1022636	1	56,741,343	58,133,153	1.39	208	0.57	1.7
Patient_21_R	rs7642123	rs6797289	3	153,610,044	157,080,481	3.47	249	0.60	3.1
Patient_21_R	rs7692447	rs13151254	4	135,190,819	139,767,513	4.58	298	0.59	2.8
Patient_21_R	rs1461349	rs11821682	11	37,157,502	41,821,239	4.66	296	0.62	4.3
Patient_21_R	rs7488279	rs17728942	12	5,438,481	8,179,670	2.74	193	0.61	2.7
Patient_21_R	rs1342606	rs418853	13	82,759,884	87,877,460	5.12	328	0.58	2.5
Patient_21_R	rs9559492	rs1948851	13	87,991,344	89,058,641	1.07	14	0.79	1.5
Patient_21_R	rs1409911	rs4773607	13	89,165,586	90,412,505	1.25	129	0.59	1.4
Patient_21_R	rs1906160	rs2291620	15	37,813,417	38,116,265	0.30	16	0.75	1.4
Patient_21_R	rs3107997	rs8093901	18	26,174,242	27,288,097	1.11	146	0.60	1.7
Patient_21_R	rs6025034	rs13038808	20	54,655,431	54,878,739	0.22	41	0.61	1
Patient_23_Dx	rs478410	rs7325798	13	29,554,081	30,495,592	0.94	136	0.63	2.9
Patient_23_Dx	rs2911851	rs951443	15	29,118,254	31,705,123	2.59	229	0.62	3.6
Patient_23_R	rs13191136	rs9283859	6	167,606,330	169,721,000	2.11	190	0.62	3.1
Patient_25_R	rs494723	rs4574630	6	10,854,213	11,726,173	0.87	90	0.63	1.9
Patient_25_R	rs11794457	rs12235266	9	7,358,850	8,292,479	0.93	161	0.62	2.7
Patient_25_R	rs2507903	rs592525	11	114,504,525	118,889,063	4.38	355	0.59	3.7
Patient_25_R	rs3741808	rs4019400	12	67,011,218	74,096,025	7.08	625	0.60	6.7
Patient_25_R	rs785450	rs2596210	15	27,928,636	31,579,227	3.65	316	0.67	9.2
Patient_25_R	rs874187	rs12460033	19	49,509,008	51,098,538	1.59	76	0.78	5.6
Patient_29_Dx	rs774381	rs10878795	12	65,726,355	66,911,125	1.18	119	0.62	2.1
Patient_55_Dx	rs882311	rs622898	10	1,219,548	3,512,302	2.29	330	0.66	8.9
Patient_55_R	rs11166104	rs945748	1	98,989,371	103,256,102	4.27	374	0.60	4.1
Patient_55_R	rs6880142	rs153478	5	149,105,642	150,619,632	1.51	156	0.63	3.3
Patient_55_R	rs7758899	rs4946399	6	116,374,038	119,470,457	3.10	185	0.63	3.5
Patient_55_R	rs7077992	rs7904349	10	934,496	3,411,055	2.48	335	0.69	11.1
Patient_55_R	rs2356535	rs2069002	14	50,445,000	51,320,440	0.88	87	0.62	1.8
Patient_61_Dx	rs6442749	rs4493418	3	2,730,983	2,811,418	0.08	17	1.00	5.1
Patient_61_Dx	rs9422881	rs7075452	10	127,110,624	127,340,239	0.23	33	0.88	5.3
Patient_61_Dx	rs4930561	rs7481750	11	67,688,337	68,904,167	1.22	82	0.78	6.8
Patient_61_Dx	rs6606662	rs9630218	11	68,915,925	69,588,488	0.67	38	0.79	3.1
Patient_61_Dx	rs674374	rs3017478	11	69,817,313	70,375,682	0.56	44	0.86	6.3
Patient_61_Dx	rs1028050	rs6592575	11	71,929,039	74,000,654	2.07	88	0.82	8.6
Patient_61_Dx	rs605954	rs1965277	11	74,258,861	134,355,825	60.10	5091	0.86	653.0
Patient_61_Dx	rs1556999	rs11148069	13	46,870,340	47,117,281	0.25	15	0.93	2.4
Patient_61_R	rs2714298	rs6717502	2	14,769,944	16,858,388	2.09	198	0.62	3
Patient_61_R	rs354707	rs1580063	2	143,602,907	148,794,127	5.19	287	0.60	3.3
Patient_61_R	rs17341291	rs9843725	3	133,383,722	137,185,555	3.80	403	0.60	4.5
Patient_61_R	rs6879951	rs17517907	5	123,873,768	124,554,764	0.68	83	0.71	4.1
Patient_61_R	rs513870	rs12211264	6	11,616,840	16,021,001	4.40	431	0.59	3.8
Patient_61_R	rs12701976	rs12375125	7	42,474,899	47,349,043	4.87	352	0.57	2.2
Patient_61_R	rs11218071	rs1238553	11	120,322,070	121,207,450	0.89	121	0.62	2.3

Table 1: *AML events.* Events called using a threshold of 0.5 for the marginal conditional probability of AI.

chr	startbp	endbp	SNP count	size (Mb)
2	31,982,105	33,266,534	131	1.28
2	38,755,294	38,887,094	16	0.13
6	99,536	68,754,442	9,968	68.65
11	71,640,522	134,435,899	7,795	62.80
12	28,466,092	28,491,511	8	25.42
12	31,157,554	31,298,174	20	0.14
16	30,423,993	88,690,776	5,604	58.27

Table 2: *Excluded regions.* Upon visual examination of the tQN-normalized array data from the normal cell line sample, we noticed several obvious structural variants. In order to assess the sensitivity and specificity rates for our methods and the other methods, we decided to mask these regions of the genome to prevent them from being called as false positives. We ran *BAFsegmentation* on the normal sample using default parameters. The method identified 8 segments on 4 chromosomes. The larger of these are identifiable in Supplementary Figure 6. One segment corresponded to a visible deletion on chromosome 16, and two contiguous segments corresponded to a visible deletion on chromosome 6. In these cases the segments did not cover the entire region that could be visually identified by looking carefully at the BAFs and LRRs. For these events, we excluded a region that included both the *BAFsegmentation* segments and any additional contiguous loci that appeared (by inspection) to be part of the same event. Another region on chromosome 6 was identified (11 SNPs only), approximately 700 SNPs downstream from the other segments. The exclusion region on chromosome 6 was extended to include this event. A segment on chromosome 2 corresponds to an obvious increase in copy number. Another small segment (16 SNPs) was identified about 800 SNPs downstream. These two regions were excluded according to the *BAFsegmentation* segment coordinates. A segment on chromosome 12 corresponds to a small duplication event. *BAFsegmentation* also identified another small event about 400 SNPs upstream. Both of these regions were excluded according to the *BAFsegmentation* segment coordinates.An apparently heterogeneous region on chromosome 11 was also excluded.

chr	copy number	SNP count	size (Mb)
2	1	5,951	60.56
3	2	3,351	32.54
4	1	1,450	11.17
4	1	1,251	10.16
4	1	5,536	49.40
5	2	3,917	30.78
7	2	1,250	9.88
8	1	6,200	47.22
9	1	3,351	22.47
9	1	2,951	20.87
10	2	1,910	12.76
12	1	10,725	95.04
13	2	301	2.39
13	2	6,830	53.46
14	1	951	9.32
14	2	351	2.88
14	2	1,401	11.65
15	1	3,550	34.92
18	2	1,700	11.53
18	1	3,101	18.96
19	1	1,450	11.37
19	1	2,316	19.92
21	1	251	1.52
22	1	800	5.86
22	2	2,141	13.14

Table 3: *Simulated events.*

Supplementary Note

AML diagnosis and remission samples on Affymetrix 6.0

We obtained Affy 6.0 CEL files for 19 normal-karyotype acute myeloid leukemia (NK-AML) samples (>90% blasts), and matched remission samples (<5% blasts) for 11 of them (GEO accession GSE21780). The samples are described in detail in Barresi et al. [2010b]. BAFs and logRRs were extracted from the CEL files using PennCNV (with 77 additional CEL files from the HapMap2 CEU data to help train the algorithm), and genotypes were called using Birdseed v2 [Korn et al. 2008].

We applied *hapLOH* to each sample, setting transition parameters for a mean event length of 2.5 Mb and a 10% genome-wide prevalence of AI. Using a threshold posterior probability of 0.5 and combining the probability of AI states we identified 46 AI regions, with a median event size of 1.55 Mb. Supplementary Table 1 summarizes the called events.

The focus of the original study of these samples was to identify tumor-associated copy-number aberrations or cn-LOH that could be used to classify these normal-karyotpe patients into risk categories. In our analysis, 4 out of 11 paired sets harbored AI events in only the diagnosis sample, including two events larger than 10 Mb in two distinct samples. One of these covered 11 Mb on chromosome 3 in the diagnosis sample of Patient 9. The logRR indicates that it is a hemizygous deletion. This loss was also identified by Barresi *et al.*

The other large (>10 Mb) event covered 60 Mb in the terminal region of 11q in the diagnosis sample of Patient 61 (Supplementary Fig. 4). Inspection of the logRR in this region suggests cn-LOH. This event includes 5,091 heterozygous genotypes (16,828 total markers) and had a phase concordance of 0.86. This event was not reported in the original publication. Interestingly, a small (88 Kb) event in the remission sample of the same patient overlaps the large event in the diagnosis sample, but the observed BAF and phase concordance indicate a much less prevalent event.

In several blood cancers, cn-LOH at 11q is associated with mutations in the gene Casitas B-lineage lymphoma (CBL) [Dunbar et al. 2008]. Another study [Barresi et al. 2010a] that used Affymetrix SNP 6.0 genotyping at multiple timepoints during the progression of a patient through early disease (refractory anemia with excess blasts [RAEB-2]), remission, and late disease (acute myeloid leukemia) was able to document the change in the proportion of cells harboring cn-LOH at 11q. They observed a partial event soon after RAEB-2 diagnosis, apparent dissapearance of the aberration during remission, and complete or near-complete cn-LOH during late disease. These observations provide evidence for the expansion of the cn-LOH-harboring clone. Our observation of a low-prevalence 11q cn-LOH clone in the diagnosis sample of Patient 61 may therefore provide insight into a possible route of progression of the

patient's disease.

In their investigation, Barresi et al. [2010b] found an average of 50 cn-LOH events larger than 1 Mb per sample (including 3 events larger than 10 Mb that are clearly visible in the BAF plots), and on the order of 10 deletion events and 10 gain events larger than 50 Kb and 80 Kb, respectively. We did not call these 3 large cn-LOH events and called many fewer smaller events. The discrepancy in the number of called events could be due to false positives (all of the cn-LOH events in the 1-10 Mb range occured in both the diagnosis and remission sample when paired samples were available, and some may be due to chance runs of homozygosity), or more likely because the events occurred in a high proportion of the sample and therefore included few markers that were informative for our method. It is important to note that our method is intended to be complementary to methods that identify events occurring in high proportion.

Bibliography

Barresi V, Palumbo G, Musso N, Consoli C, Capizzi C, Meli C, Romano A, Di Raimondo F, and Condorelli D. 2010a. Clonal selection of 11q CN-LOH and CBL gene mutation in a serially studied patient during MDS progression to AML. *Leukemia Research*, 34(11):1539–1542.

Barresi V, Romano A, Musso N, Capizzi C, Consoli C, Martelli M, Palumbo G, Di Raimondo F, and Condorelli D. 2010b. Broad copy neutral-loss of heterozygosity regions and rare recurring copy number abnormalities in normal karyotype-acute myeloid leukemia genomes. *Genes, Chromosomes and Cancer*, 49(11):1014–1023.

Dunbar A, Gondek L, O'Keefe C, Makishima H, Rataul M, Szpurka H, Sekeres M, Wang X, McDevitt M, and Maciejewski J. 2008. 250k single nucleotide polymorphism array karyotyping identifies acquired uniparental disomy and homozygous mutations, including novel missense substitutions of c-cbl, in myeloid malignancies. *Cancer Research*, 68(24):10349.

Korn J, Kuruvilla F, McCarroll S, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins P, Darvishi K, et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics*, 40(10):1253–1260.