Supplementary information of

"Accurity: Accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants"

Table of Contents

1. Accurity on simulation data with fewer heterozygous SNPs and more subclones

We used 1.5million heterozygous SNPs from the 1000Genome project in simulations. Here we explored if reducing the number of heterozygous SNPs can have an adverse effect on the performance of Accurity. Figure S1 and S2 showed Accurity performance did not deteriorate in low-coverage (5X) simulations with 150K SNPs (10% of usual setting) or 15K SNPs (1% of usual setting).

We further explored if increasing the number of subclones can have an adverse effect on the performance of Accurity. Figure S3 (two-subclone low-coverage) and S4 (three-subclone low-coverage) showed that Accurity performed similarly as it did in the two-subclone high-coverage setting of Figure 3C.

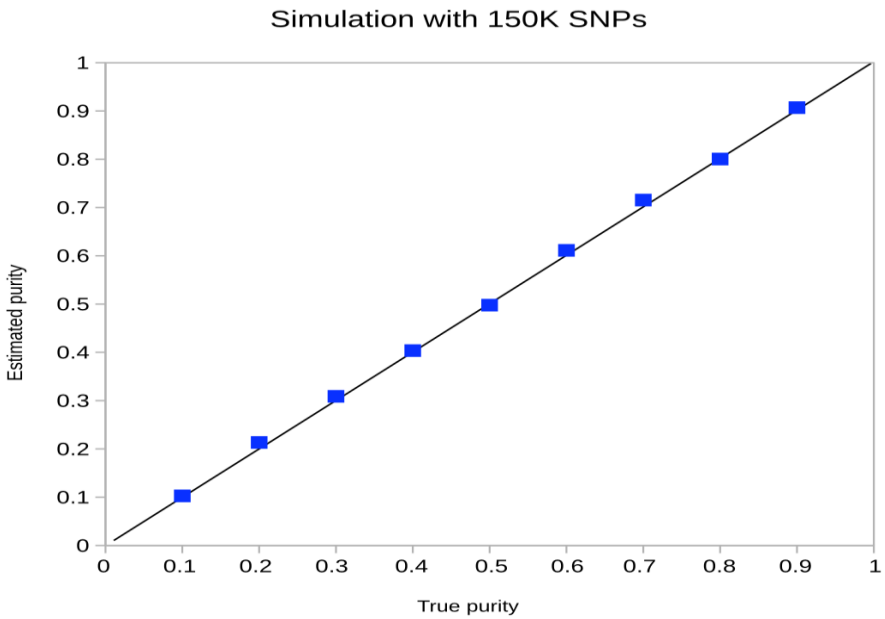

Figure S1. Accurity results on 5X simulation data that contains about 150,000 SNPs, which is about 10% of what we normally use in simulation. Other settings are the same as the ones in Figure 3B.

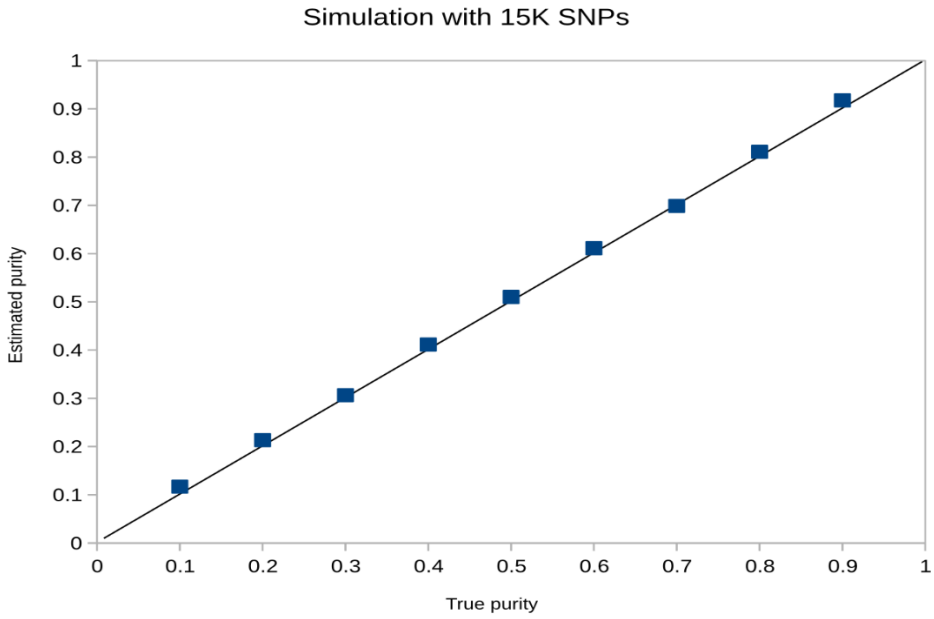

Figure S2. Accurity results on 5X simulation data that contains about 15,000 SNPs, which is about 1% of what we normally use in simulation. Other settings are the same as the ones in Figure 3B.

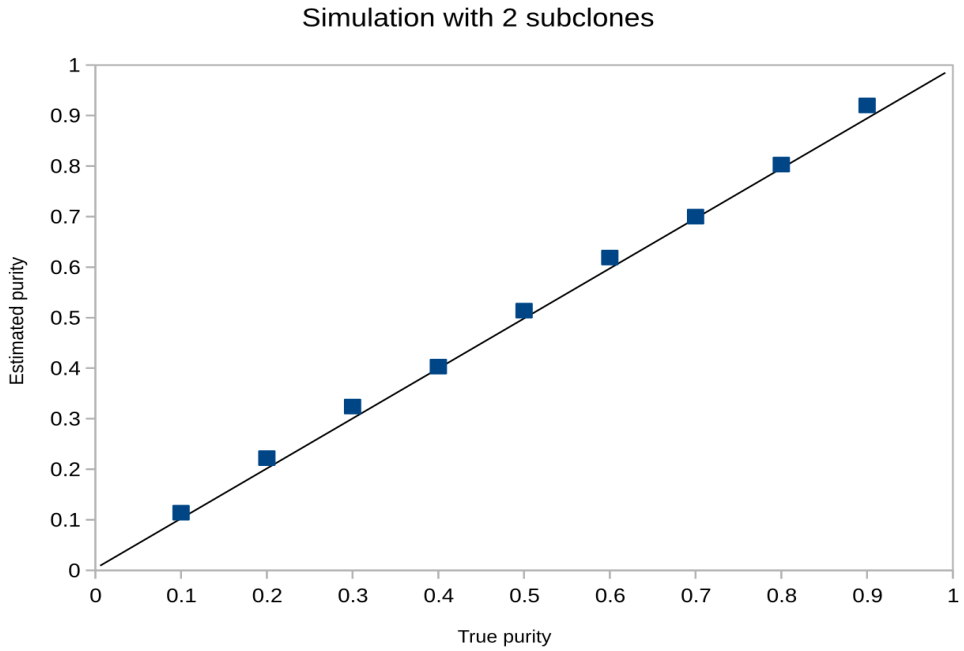

Figure S3. Accurity results on 5X simulation data that contains two subclones. Other settings are the same as the ones in Figure 3B.

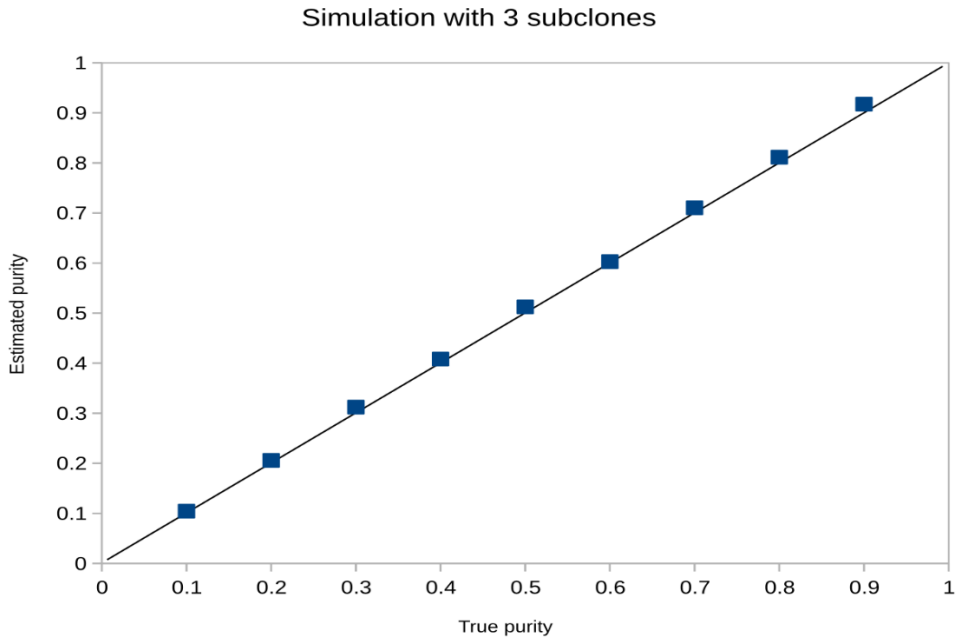

Figure S4. Accurity results on 5X simulation data that contains three subclones. Other settings are the same as the ones in Figure 3B.

2. Performance of MixClone on simulation data with 2 subclones

In addition to CNAnorm(Gusnanto, et al., 2012) and Absolute(Carter, et al., 2012) that we have compared Accurity with in the main text, we also included MixClone(Li and Xie, 2015) in one of our benchmarks. MixClone is a recent method that also considers intra-tumor heterogeneity in estimating tumor purity. The input information it extracts from the tumor-normal pair of sequencing data include the coverage data of SCNAs and allelic coverage at heterozygous germline SNPs, similar to Accurity. However, that is where similarity between the two software ends. MixClone estimates LOH (loss of heterozygosity) at each segment and uses that estimate to group segments into baseline (non-LOH) segments (which means copy-number=2 normal segments based on its publication) and LOH segments. One potential issue using non-LOH segments as normal is that it might include copy-number>2 segments which contain both alleles but are not of allelic balance. In contrast, Accurity identifies candidate peaks of normal segments from the histogram of TRE using auto-correlation analysis and then selects the final peak corresponding to normal segments through model selection on a joint likelihood of segmental and allelic coverage data. The second notable difference is that MixClone assumes one genomic region belongs to only one subclone. In bulk sequencing of an intra-heterogeneous tumor, DNA segments mapped to the same genomic region could come from different cancer subclones. We are not clear how the approximation will affect its actual performance. The third difference is technical. MixClone uses EM to estimate parameters and a heuristic approach for model selection. Accurity uses MLE for parameter estimation and BIC for model selection.

In our testing on a series of low-coverage (5X) two-subclone sequencing data, Figure S5, MixClone estimated correctly for tumor purity=0.3-0.5 and purity=0.9 but erred on the high side (~0.1 higher than truth) for purity=0.2 and 0.6-0.8. It failed to complete estimation for purity=0.1 after 10 days, which looked like being trapped in an infinite loop based on its real-time output. Accurity estimated correctly (within 0.015 of truth) for all nine samples, Figure S3. Implemented in Python, MixClone is slower than Accurity. The average running time of MixClone on nine finished samples is 110 hours (4.6 days), vs 40 minutes of Accurity. Due to its lengthy running time, we balked at running MixClone on additional simulation data. Based on our benchmark on nine pairs of tumor-normal pair sequencing data whose true purity ranges from 0.1 to 0.9, we reach the conclusion that Accurity performs better than MixClone for both accuracy and speed.

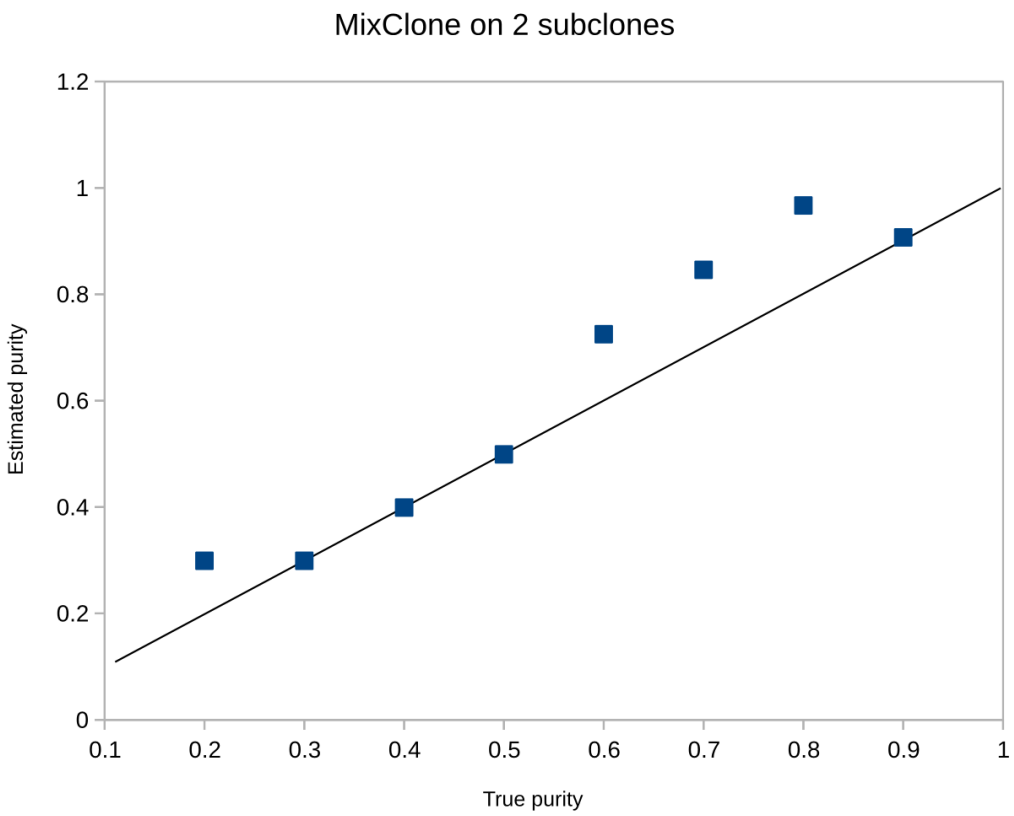

Figure S5. MixClone results on 5X simulation data that contains two subclones, which is the same dataset used in Figure S3. It fails to complete for the sample of purity=0.1 after 9 days. The average running time for MixClone is 110 hours (4.6 days).

3. Performance of Patchwork on one simulated and one TCGA sample

We ran Patchwork(Mayrhofer, et al., 2013), a related software, on one 5X coverage simulation data and one real-sequencing data, TCGA-AA-3553. For both datasets, Patchwork obtained results comparable to those of Accurity. For the 5X simulation data (true tumor-purity=0.6), Patchwork estimated its purity to be 0.596, vs 0.598 by Accurity. For TCGA-AA-3553, Patchwork obtained a tumor purity estimate of 0.412, vs 0.384 by Accurity. The SCNA profiles in both cases are also identical to that of Accurity.

However, halfway during its pipeline, it requires the user to manually inspect the allelic-imbalance-vs-normalized-coverage plot of all SNVs and determine 1) the cluster of SNVs that correspond to copy number two, 2) the allelic copy numbers to which sub-clusters of SNVs correspond. This manual inspection requires substantial user expertise of how patchwork works. For real-sequencing data, this manual inspection is non-trivial, as can be seen in the left panel of Figure S6. In rare cases where the copy number two cluster does not exist, it would fail to make correct estimates. It also hinders fast processing of large amount of cancer genomic data.

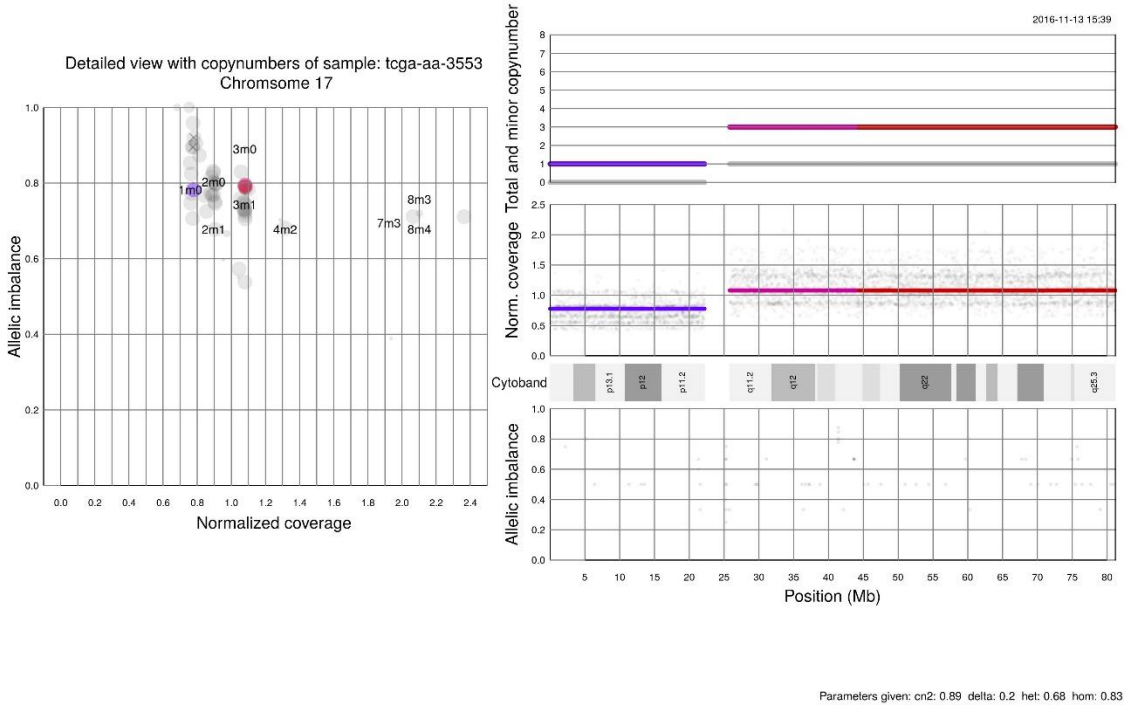

Figure S6. Result of patchwork on TCGA-AA-3553, the same sample from Figure 6. Only chromosome 17 is shown. Patchwork requires manual inspection of the left panel by a user to identify: 1) the cluster of SNVs that correspond to copy number two; 2) the allelic copy numbers to which sub-clusters of SNVs correspond. In this non-trivial real sequencing data example, our manual inspection was successful and enabled patchwork to make correct inferences.

4. Performance of Accurity on mixed HCC1187 cell line data

To test Accurity on a more realistic setting than our EAGLE-generated pure simulation data, we tested Accurity on nine pairs of mixed HCC1187 cell line data. The HCC1187 and its matched normal WGS data was generated by Illumina Inc. on a HiSeq 2000 machine. These data retains the characteristics (GC-bias, etc.) of real sequencing data. Details of mixing are in Methods section 2.9. Accurity performed well on these data, Figure S7, which suggests Accurity can handle real sequencing data.

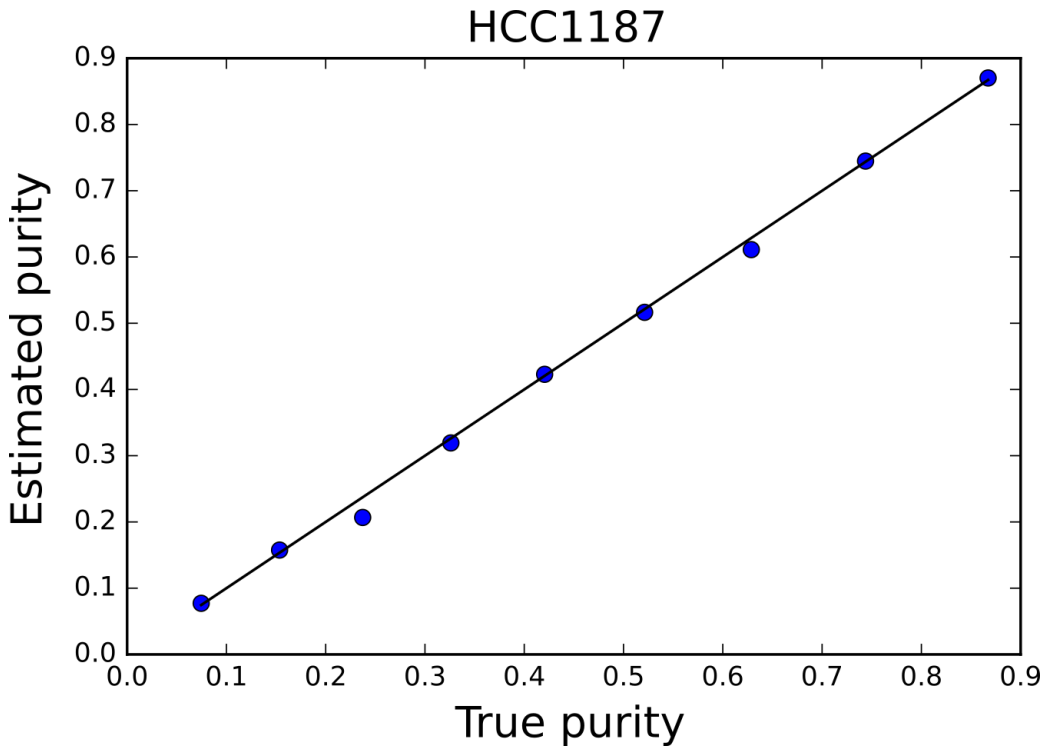

Figure S7 Purity estimates by Accurity on nine pairs of mixed HCC1187 cell line data. X-axis is the true purity from mixture. Y-axis is the estimated purity by Accurity.

5. Comparison of Accurity with other methods on 172 pairs TCGA samples

Limited by computing storage resources and internet speed between China and US, we managed to download 172 pairs of TCGA samples, ran Accurity on all of them, and compared Accurity purity estimates with those of ABSOLUTE, ESTIMATE, and LUMP, as reported in Aran et al. 2015 (Aran, et al., 2015), Table S1. Accurity successfully produced purity estimates for 111 pairs and data of the rest were too noisy for Accurity to produce a robust estimate. The IHC purity measurement, as estimated by image analysis of haematoxylin and eosin stain slides produced by the Nationwide Children's Hospital Biospecimen Core Resource, is regarded as the gold standard.

Software	Coverage <5	Coverage 5~10	Coverage >10
Accurity	0.291	0.187	0.456
ABSOLUTE	0.074	0.460	0.444
ESTIMATE	0.027	0.355	0.092
LUMP	0.172	0.352	0.316

Table S2. Performance of different methods over varying coverage on TCGA samples. Each cell is the Spearman rank correlation between estimates of that method (row name) and IHC.

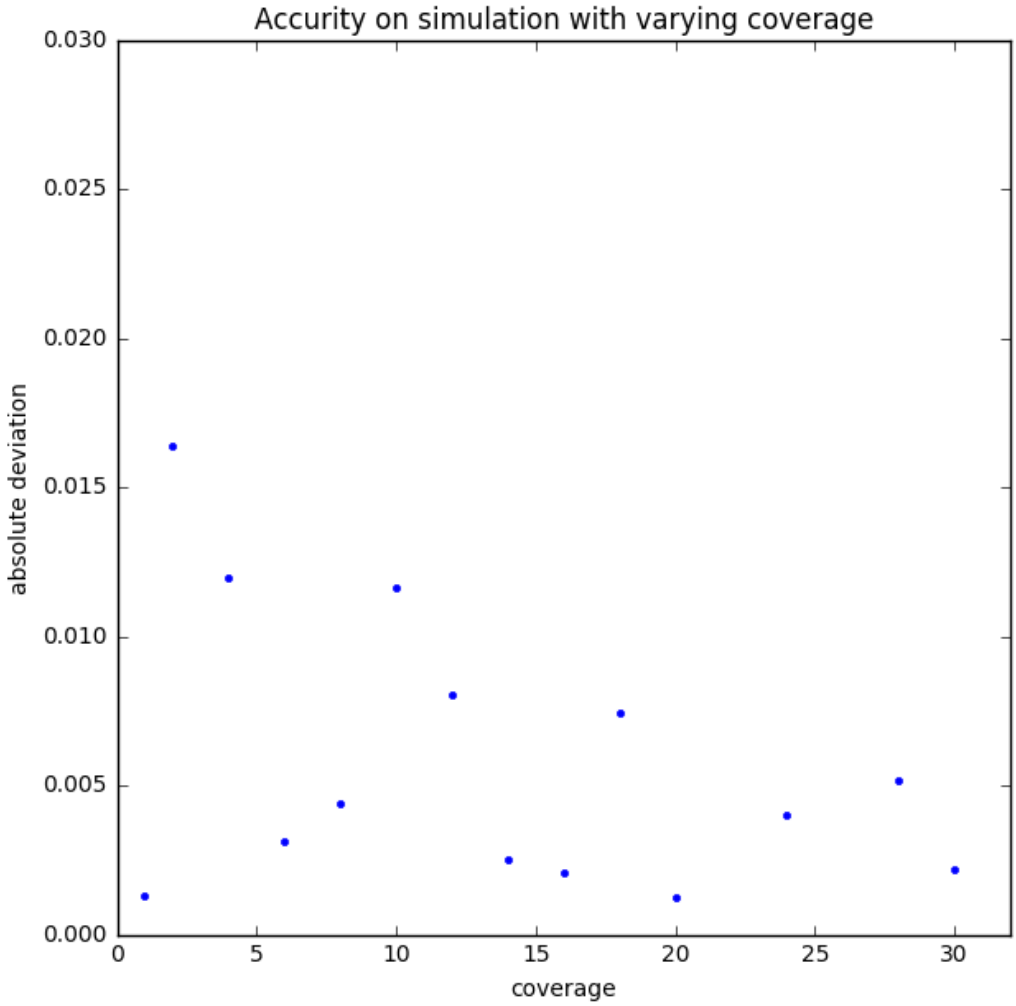

Figure S8. Accurity performance, measured as the absolute deviation from the true purity level on the Y-axis, over the coverage, X-axis.

References

Aran, D., Sirota, M. and Butte, A.J. (2015) Systematic pan-cancer analysis of tumour purity, *Nature communications*, **6**, 8971.

Carter, S.L.*, et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer, *Nature biotechnology*, **30**, 413-421.

Gusnanto, A.*, et al.* (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data, *Bioinformatics*, **28**, 40-47.

Li, Y. and Xie, X. (2015) MixClone: a mixture model for inferring tumor subclonal populations, *BMC genomics*, **16 Suppl 2**, S1.

Mayrhofer, M., DiLorenzo, S. and Isaksson, A. (2013) Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue, *Genome biology*, **14**, R24.