Calling Amplified Haplotypes in Next Generation Tumor Sequence Data

Ninad Dewal[1], Yang Hu[2], Matthew L. Freedman[3,4], Thomas LaFramboise[5], and Itsik Pe'er[2,†]

[1]Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA; [2]Department of Computer Science, Columbia University, New York, NY 10027, USA; [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA; [4]Medical and Population Genetics Program, The Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [5]Department of Genetics, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

SUPPLEMENTARY INFORMATION

These sections supplement the main manuscript with extra detailed information relevant to HATS:

Table of Contents

Table of Figures

Note on data:

The training data was obtained from the 1000 Genomes Project in the form of phased haplotype sequences from HapMap individuals. These sequences can be downloaded from (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/2009_04/). Sites that were not bi-allelic in a respective population were eliminated, as tri-allelic sites are rare within a control population. Several samples (trio children) were also eliminated as mentioned in the main manuscript.

Tumor data (sample TCGA-06-0877) used in evaluations were obtained from The Cancer Genome Atlas (TCGA) (http://tcga-data.nci.nih.gov/tcga/). Specifically, copy number aberration information was available via open-access, while aligned reads were available only via controlled access (via dbGaP). Researchers may apply at dbGaP for access to TCGA data.

Supplementary Results

Real Tumor Data with Stromal Contamination

We consider another region at chr2:30-31Mb within the same sample possessing local $2\overline{\Lambda}$ = 38.7 and local C_a = 2.50, suggesting 50% stromal contamination. Results are displayed in **Supplementary Figure 10**. In this case, the gold standard amplified alleles were instead called by HATS on the sequence data at $2\overline{\Lambda}$, as the naïve method weakens when $C_a < 3$ according to **Figure 2B** and **Supplementary Figure 5**, while HATS remains stronger. The gap between the two methods is wide even at 2Λ = 30 in this region, supporting that a higher 2Λ is required for the naïve method to perform as strongly as HATS for mixtures with $2 < C_a < 3$. Note that the naïve empirical curve deviates from the theoretical curve, possibly due to an imperfect gold standard called by HATS. To test this, we used HATS to call the amplified alleles on simulated data generated with $2\overline{\Lambda}$ = 35 and C_a = 2.50. Using these called alleles as the gold standard, we performed the down-sampling procedure described above on this simulated data and depict the deviation from the theoretical curve in **Supplementary Figure 11**. This deviation, while smaller than that in **Supplementary Figure 10**, supports the culpability of the imperfect gold standard called at $2\overline{\Lambda}$, suggesting a higher $2\overline{\Lambda}$ is needed for a stronger gold standard. An imperfect gold standard for a region may also be caused by lower compatible haplotype representation in the training data, resulting in calls that resemble those of the naïve model; this could pull the naïve curve up from the theoretical curve. To correct for this possibility, we shift both the HATS and naïve curves down in **Supplementary Figure 10** by the mean deviation between the naïve tumor and naïve theoretical curves. This adjustment may better indicate HATS' performance at this copy number. We repeated the analysis on this region, this time calling the gold standard amplified alleles from 43 heterozygous sites (after quality control filtering) from the SNP array data. The results are depicted in **Supplementary Figure 12** and demonstrate again HATS' superior performance over the naïve model despite fewer sites and lesser LD information.

Supplementary Methods

Hidden Markov Models (HMMs)

Copy number detection on more matured platforms, such as arrays, was oftentimes performed using a key computational tool called the Hidden Markov Model (HMM). HMMs have been widely used in technical areas, such as in speech recognition (Baker 1975) as well as in genomics (Durbin 1998; Yoon 2009). An HMM is a probabilistic graphical model that enables the inference of latent causes (modeled as hidden "states") given an observed sequence of events. Each state emits probabilistic observations and transitions probabilistically to other such hidden states next in the sequence. The sequence of observations helps identify the most likely series of hidden states. For example, particular levels of copy number (the states) could be inferred via observed signal intensities (Shah et al. 2006; Wang et al. 2007; Korn et al. 2008; Liu et al. 2010).

Sample-specific Polymorphic Sites

We mentioned in **Methods** that the HMM model includes only those sites polymorphic in the training data. This would exclude sites polymorphic in sample *j* that are monomorphic in *D*, such as sites possessing rare variants or somatic mutations in *j*. To call the amplified allele at each such site *i* while accounting for bias, we first calculate the probability of each allele *x* being amplified, $P(x)$, based on the observed read counts on the Poisson distribution. In the case of a homozygous genotype at *i*, $P(x) = \Pr(r_x; \lambda_x)$, where λ_x is calculated as defined earlier, setting component $G_x = C_a$. In the case of a heterozygote at *i*, $P(x) = \prod_{x=0}^{1} \Pr\left(r_x ; \lambda_x\right)$, in which $G_x = (C_a - 1)$ for a particular *x* while $G_{\bar{x}} = 1$. The function arg-$\max_x\{P(x)\}$ returns the higher probability, thus calling the amplified allele. In the case of a tie, HATS designates ∅ as the amplified allele at *i* to reflect this ambiguity.

One thing to note is that HATS cannot leverage the training data for such sites, as the variant is not present in the training data. However, HATS may still retain a performance gain over the naïve method at such sites due to analysis and correction of allelic bias.

State Pruning and Genotyping Error Correction (GEC)

As described in **Methods**, the set of states S_t for each t in genotype HMM V theoretically contains all states from the Cartesian product of haplotype states $(S_t^A \times S_t^U)$. In practice, not all such states are viable or necessary when analyzing a particular test sample. We can utilize the tumor genotypes provided as part of the input data for sample j to eliminate impossible states.

In particular, if genotype data is completely accurate, then a state $s = (s^A, s^U) \in S_t$ needs to be considered only if the tumor genotype agrees with both of the last symbols in the contained haplotype labels: $\{ h_s^A [l(s^{\boldsymbol{A}})],\ h_s^U [l(s^{\boldsymbol{U}})] \}$. This pruning results in a drastic reduction of states. We enforce this pruning only at higher levels of coverage Λ, in which genotype calling fidelity is strong due to the greater density of reads supporting the call.

However, when coverage levels are low, fewer reads observe each site. This may lead to heterozygous sites being erroneously called as homozygous, as there may be sufficient reads to support one allele but not the other. We therefore cannot rely on homozygous genotype data at low Λ. In this case, a state $s = (s^A, s^U) \in S_t$ is considered only if either $h_s^A [l(s^A)]$ or $h_s^U [l(s^U)]$ agree with the tumor genotype. This results in s representing either a heterozygous or compatible homozygous genotype at i. Including the heterozygous genotype addresses this potential genotyping error and recovers an allele at t for analysis. The cost, however, is a large expansion of states, dramatically increasing execution time and memory usage. This can be assuaged by deactivating GEC at particular sites that possess sufficient reads despite a low Λ.

Estimating Model Bias Parameters from Data

The site-specific bias $b_{i,x}$ represents a potential favoring of read counts towards one of the alleles at site i. During tumor tissue sequencing, overabundance of reads that observe one allele may be attributed to either this bias or different copy number of the two alleles, or both. In matched normal tissues, however, the latter is not an issue in copy neutral regions. This bias may therefore be estimated by analyzing the n

normal samples and is denoted as $\hat{b}_{i,x}$. As *n* increases, $\hat{b}_{i,x}$ approaches the true bias $b_{i,x}$. To model this, we use a *maximum a posteriori* Bayesian approach that identifies $\hat{b}_{i,x}$ as the bias value possessing the maximum posterior probability based on the allele specific read counts.

$\hat{b}_{i,x}$ = arg-max$_b\{P(b \mid \tilde{r}_{i,x})\}$ for $0.01 \leq b \leq 30$, such that:

$$P(b \mid \tilde{r}_{i,x}) = \left[\frac{P(\tilde{r}_{i,x} \mid \tilde{\lambda}_{i,x})}{\int_{b=0.01}^{30} [P(\tilde{r}_{i,x} \mid \tilde{\lambda}_{i,x}) P(b)] db} \right] [P(b)]$$

We cap the range of *b* at 30 as the bias is unlikely to exceed this value. The above formula makes use of a prior distribution on *b* (denoted as *P*(*b*)). We employ a gamma distribution, with shape parameter k = 10.0 and scale parameter θ = 0.1. This results in a desired mean of $k\theta$ = 1.0, which represents the default bias value (translating to no bias). The distribution peak is narrow, symbolizing that most sites will tend not to have bias.

The term within the first pair of brackets on the right side of the equation above is the normalized likelihood. The numerator of this is P($\tilde{r}_{i,x} \mid \tilde{\lambda}_{i,x}$), which represents the probability of observing $\tilde{r}_{i,x}$ reads for allele *x* at *i* on a Poisson distribution with mean $\tilde{\lambda}_{i,x}$ (tallied across samples). Because the average coverage and read counts of the *n* samples are independent, we use the property of the Poisson distribution in which the sum of the random variables follows a Poisson process with a parameter mean that is equal to the sum of the individual parameter means. As such, $\tilde{R}_{i,x} = \sum_{j=1}^{n} \tilde{R}_{i,j,x}$, $\tilde{r}_{i,x} = \sum_{j=1}^{n} \tilde{r}_{i,j,x}$, and $\tilde{\lambda}_{i,x} = \sum_{j=1}^{n} \tilde{\lambda}_{i,j,x}$. The above posterior formula can then be modified to:

$$P(b \mid \tilde{r}_{i,x}) = \left[\frac{\prod_{j=1}^{n} P(\tilde{r}_{i,j,x} \mid \tilde{\lambda}_{i,j,x})}{\int_{b=0}^{30} \left[\left(\prod_{j=1}^{n} P(\tilde{r}_{i,j,x} \mid \tilde{\lambda}_{i,j,x}) \right) P(b)db \right]} \right] [P(b)]$$

Recall that $\tilde{\lambda}_{i,j,x} = \tilde{\Lambda}_j \cdot \tilde{G}_{i,j,x} \cdot b_{i,x}$; we substitute $b_{i,x}$ by b from the posterior formula. The normalized likelihood term essentially adjusts the prior by adding evidence from sample read counts. The resulting posterior distribution over varying values of b is thus expected to be a narrower, shifted version of the prior distribution. The peak of this distribution represents the b with the maximum probability and is assigned to $\hat{b}_{i,x}$.

Effect of Matched Normal Data Availability on $b_{i,x}$

Oftentimes, matched normal data is available only for a subset of the samples. HATS can still calculate $b_{i,x}$ in this case with the drawback of potentially reduced power. If no matched normal samples are available, HATS may still calculate $b_{i,x}$ with corresponding copy neutral regions in the tumor sample set if available; otherwise, bias is ignored. If $b_{i,x}$ cannot be estimated accurately for such reasons, the genome-wide overdispersion from the Poisson may not be accurately represented either.

Naïve Method: Theoretical Estimation Details

Recall that the naïve model for calling the amplified allele at site i in amplicon a simply picks arg-$\max_x\{r_x\}$, avoiding a call if there is a tie. The naïve call can thus be interpreted as the choosing of the allele for which the maximum likelihood estimate of λ_x (denote as $\hat{\lambda}_x$) is greater. We are therefore able to assess the sensitivity of this model theoretically by determining how often $\hat{\lambda}_x$ resembles $\lambda_{\bar{x}}$ rather than λ_x for a heterozygous genotype at a site i. The result at i can be generalized to any site, as the model considers each site independently of one another.

We compute this sensitivity by summing over all possible pairs of read counts at a site $[(r_x, r_{\bar{x}}) \mid r_{\bar{x}} \geq r_x]$ for a particular $(\lambda_x, \lambda_{\bar{x}})$ pair, assuming *x* is amplified in truth. With *z* → ∞, this can be formalized as:

$$\Pr(\bar{x}\text{ amplified}) = \sum_{r=1}^{z} \sum_{r'=0}^{r-1} \Pr\left(r'; \lambda_x\right) \times \Pr\left(r; \lambda_{\bar{x}}\right) \qquad (10a)$$

$$\Pr(\text{amplification of } x \text{ or } \bar{x} \text{ ambiguous}) = \sum_{r=1}^{z} \Pr\left(r; \lambda_x\right) \times \Pr\left(r; \lambda_{\bar{x}}\right) \qquad (10b)$$

Sensitivity for $(\lambda_x, \lambda_{\bar{x}})$ = 1 – [Pr($\bar{x}$ amplified) + Pr(amplification of *x* or $\bar{x}$ ambiguous)] (10c)

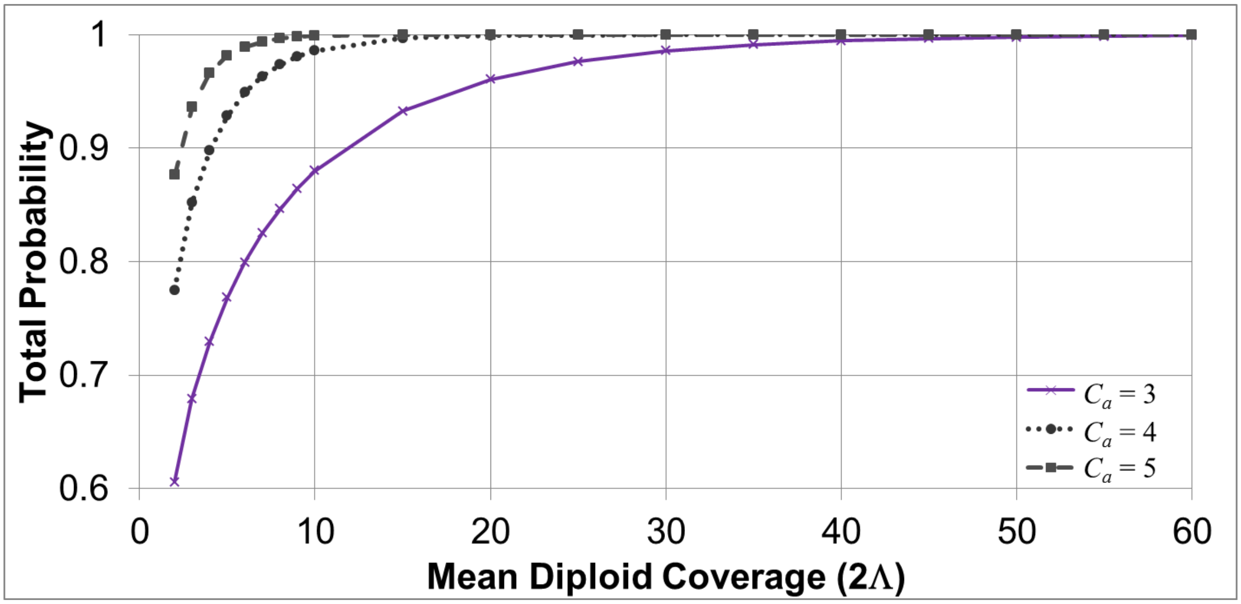
In practice, we set *z* to a large value appropriate for λ_x and $\lambda_{\bar{x}}$, which themselves are calculated for given values of Λ while setting b_x = 1 (no bias), $G_x = C_a - 1$, and $G_{\bar{x}} = 1$. Results over a range of values for diploid coverage 2Λ and C_a are shown in **Supplementary Figure 1**.

Supplementary Tables

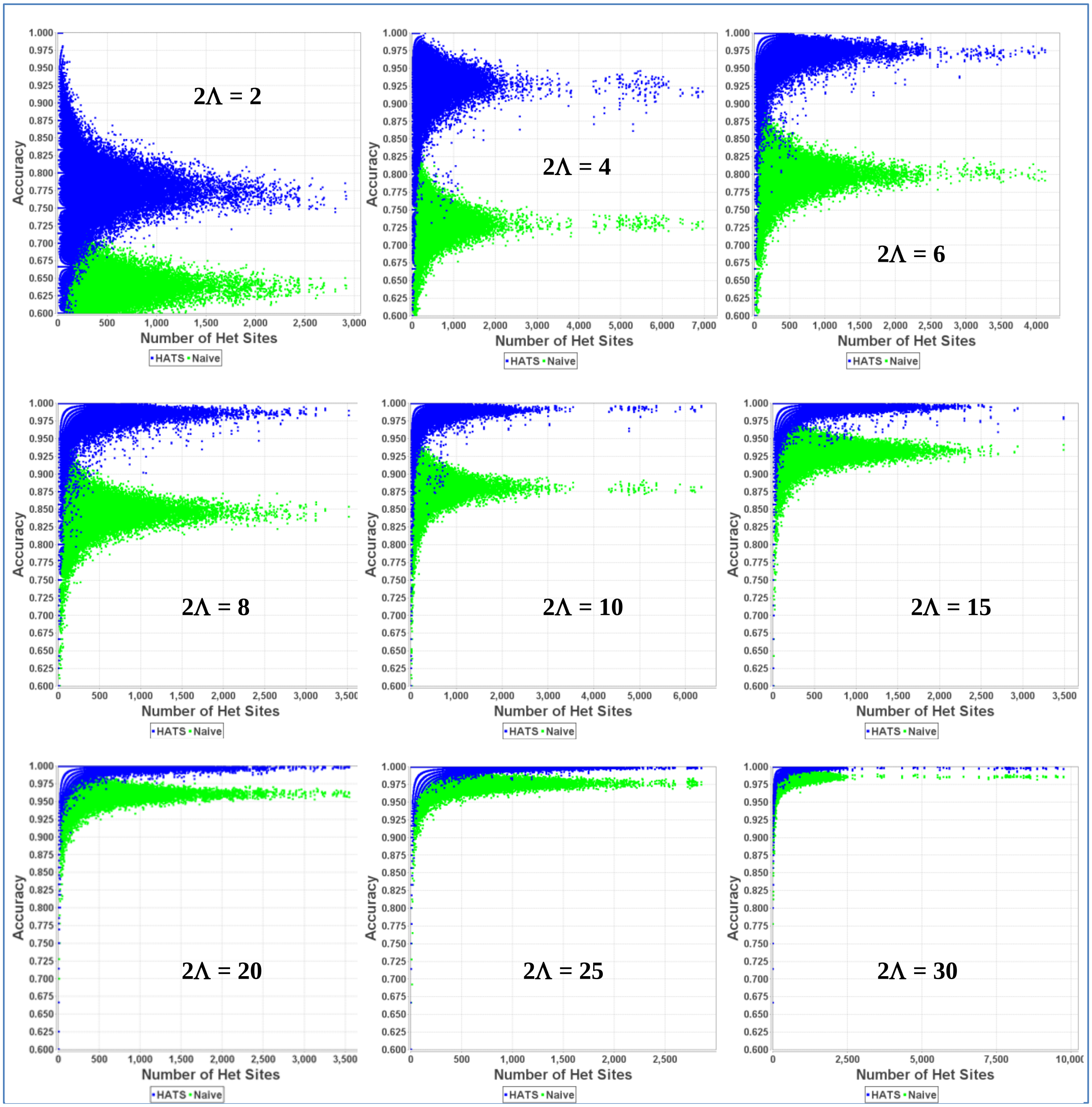
Supplementary Table 1. Accuracies of HATS, Naïve Method, and Germline Phasing, 4 Regions, Glioblastoma Tumor (TCGA-06-877). An alternate method to HATS entails phasing the data first using germline phasing algorithms and then examining the read counts across each reported haplotype to deduce the amplified and non-amplified haplotypes. This alternate approach was compared with HATS and the naïve method on four regions from the indicated glioblastoma tumor sample. Analysis was limited to sites that were both present in the sequence data as well as typed on the SNP array for the same sample, with the gold standard amplified alleles determined by the SNP array calls. In all four regions, HATS performs the same or better than the germline phasing alternative (in this case, Beagle 3.0), thus supporting the relevance for HATS in phasing CNA regions.

Chr	Position Range	# Het Sites	% Amplified Alleles at Het Sites Called Correctly by:		
	Start - End		Germline Phasing	HATS	Naive
2	20Kb - 990Kb	29	96.6%	100.0%	89.5%
2	9.9Mb - 10.4Mb	49	87.8%	100.0%	94.0%
2	30Mb - 31Mb	45	100.0%	100.0%	96.0%
19	2.18Mb - 2.54Mb	3	66.6%	100.0%	100.0%

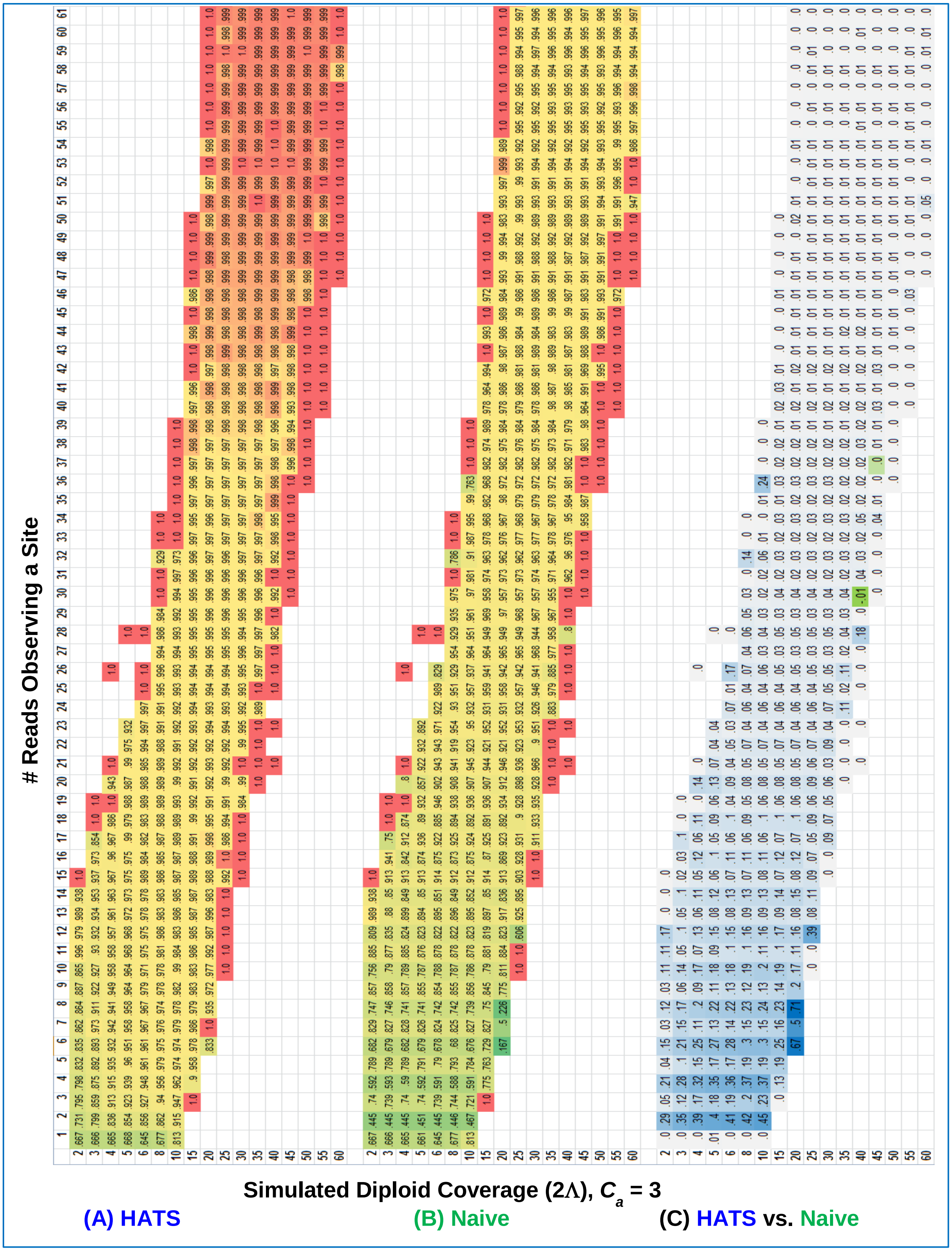
Supplementary Figures



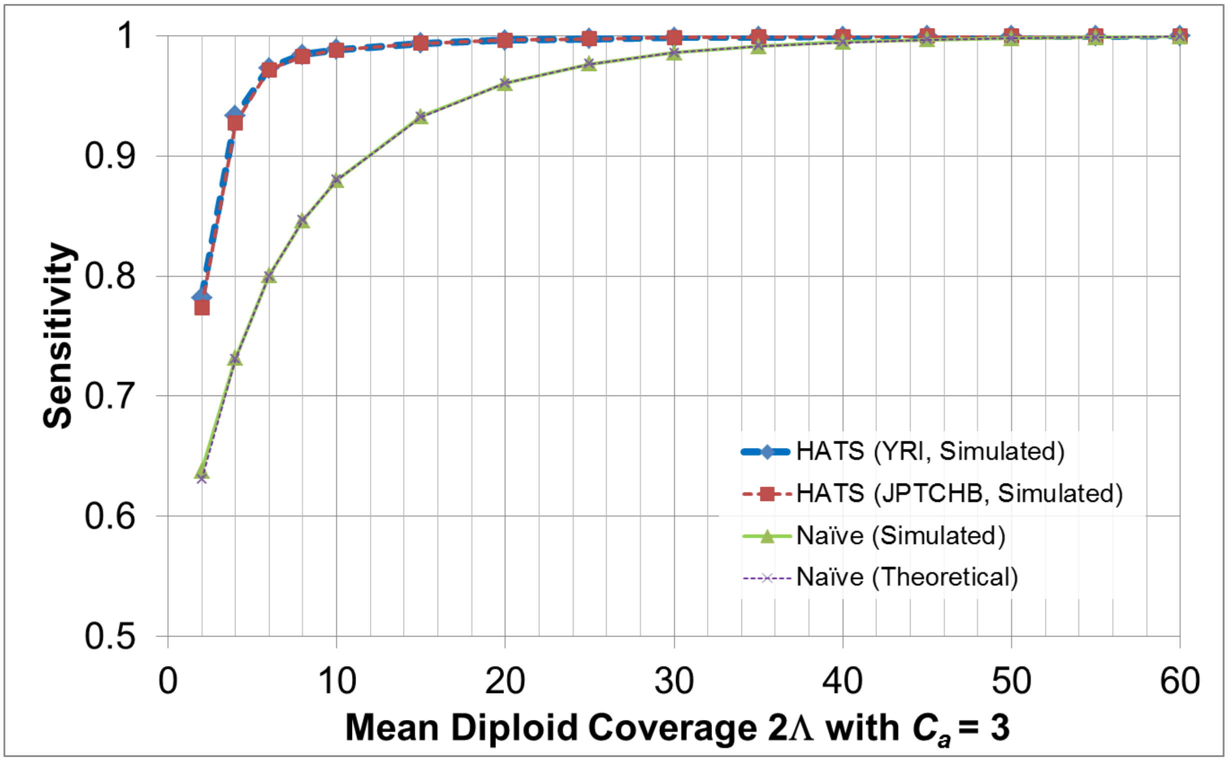
Supplementary Figure 1. Theoretical accuracy of the naïve model across varying genome-wide coverage levels and amplicon-specific copy number levels. The figure above displays the sensitivities for varying copy number levels for an amplicon (3 – 5) over varying diploid coverage levels. The naïve method performs quite well theoretically at high copy number levels, except for low coverage levels. This is especially visible with the default value of copy number 3.



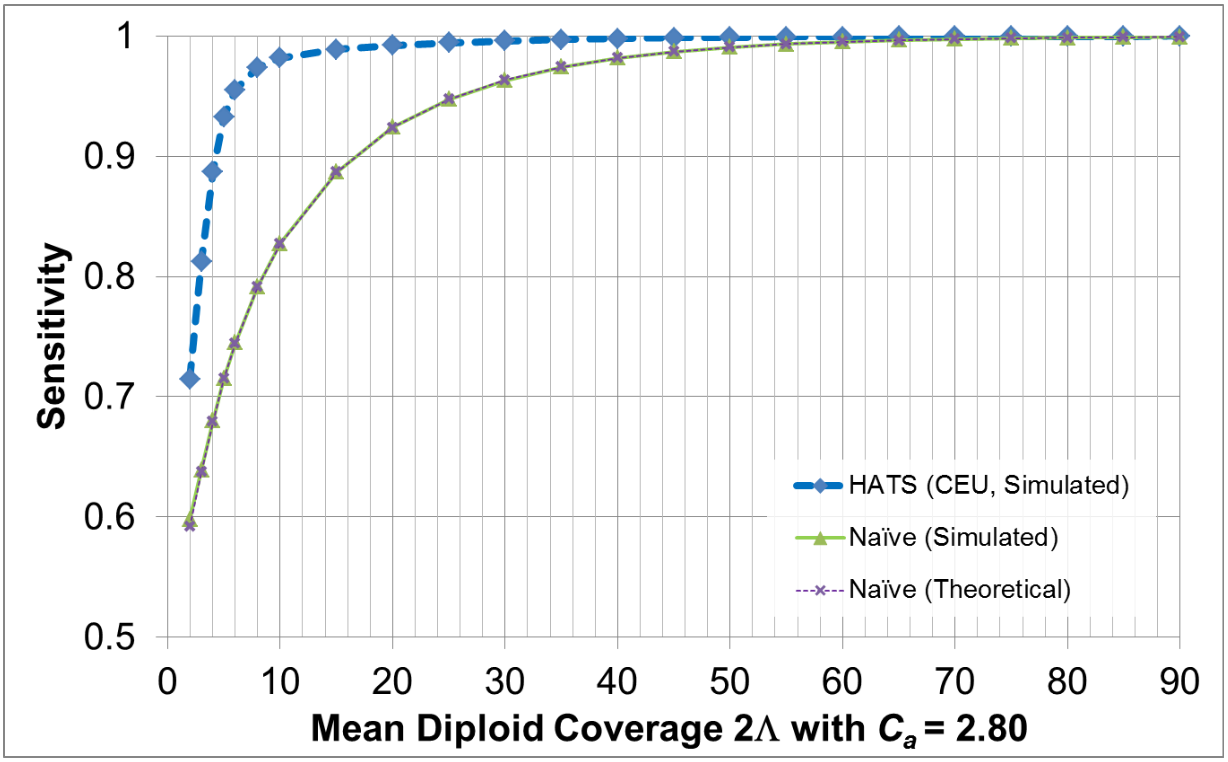
Supplementary Figure 2: Accuracy Plots over Varying Coverage Levels for HATS and the Naïve Method from Simulations, European (CEU) Training Dataset. The above nine accuracy plots depict the accuracies for both methods over hundreds of trial runs over varying levels of diploid coverage (2Λ). Each point in an accuracy dot plot represents the accuracy for a method in a particular amplicon *a* in sample *j* per trial. As the number of heterozygous sites in *a* increases, the accuracies converge to a peak for each method. Even when the number of heterozygous sites is low within an amplicon *a*, HATS consistently outperforms the naïve method in that amplicon. The training dataset was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).


# Reads Observing a Site
Simulated Diploid Coverage (2Λ), C_a = 3
(A) HATS
(B) Naive
(C) HATS vs. Naive

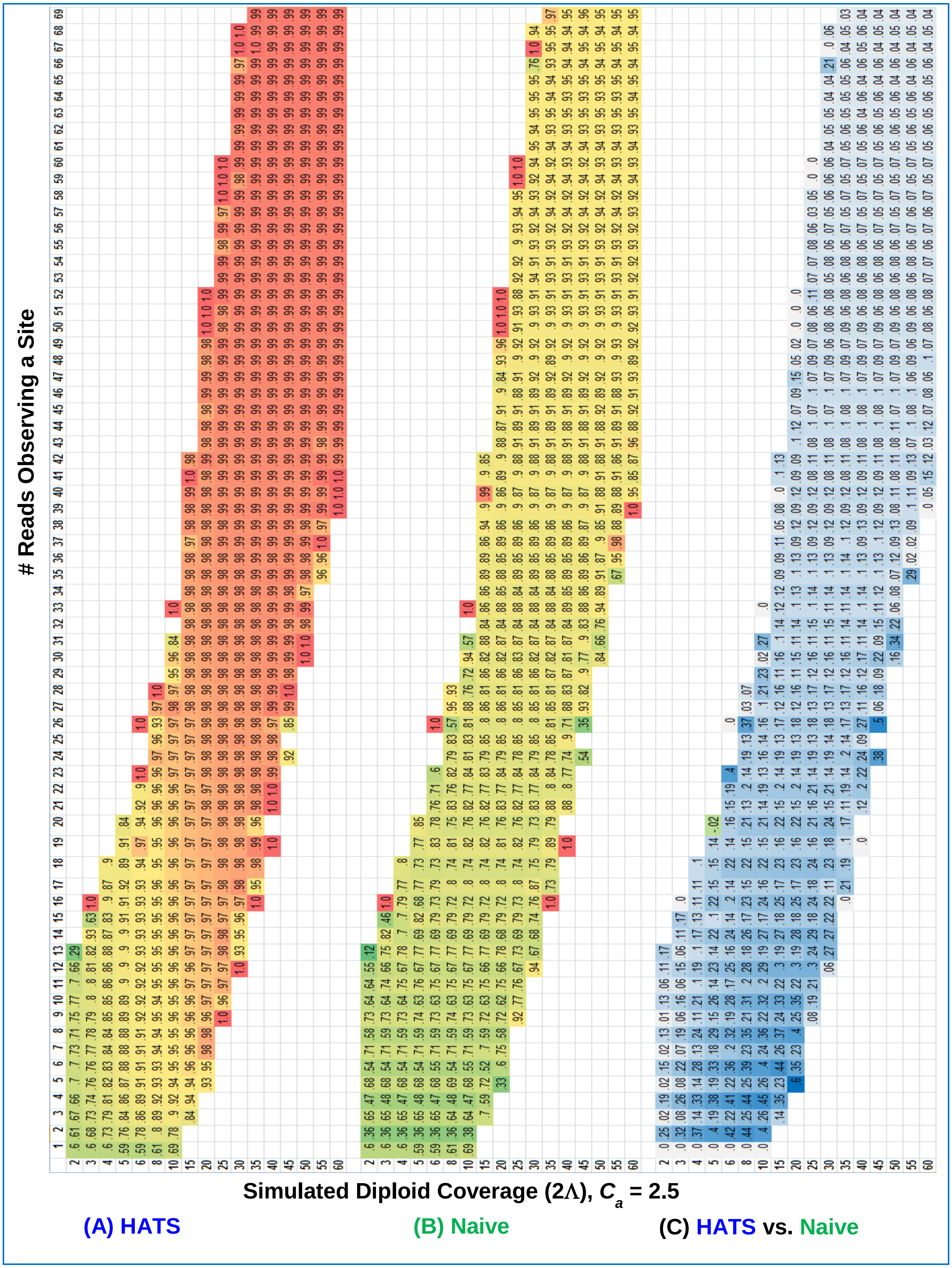
Supplementary Figure 3: Sensitivity of HATS and the Naïve Model for Breakdown of Reads Observing a Site vs. Diploid Coverage Simulated, with Copy Number of 3. The first two plots **(A)** and **(B)** depict the sensitivity for each y-axis bin, where each bin represents a particular number of total reads observing a site. The reads in a bin are generated by a Poisson distribution with diploid mean 2Λ. A range of values for 2Λ are displayed on the x-axis. This figure is generated from simulated data with amplicon copy number of 3. Blank bins represent read counts not generated from a particular 2Λ. Sensitivity spans from low (green) to high (red) and is scaled for both methods together. Note that the naïve method possesses periodicity – namely, even numbered bins perform poorer than adjacent odd numbered bins. The reason is that even numbered bins can result in ties in allelic counts, which disable the naïve method from making a call. Odd numbered bins cannot result in ties by definition. **(C)** This plot represents the differences in sensitivities between HATS and the naïve method. The cells are shaded in a gradient corresponding to the sensitivity differences: green (negative difference, naïve is better) to off-white (no difference, labeled as ".0") to blue (positive difference, HATS is better). Note that the periodicity of the naïve model is reflected in the sensitivity differences. This figure depicts HATS generally outperforming the naïve especially when 2Λ is low or the bin numbers are low. The training dataset was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).



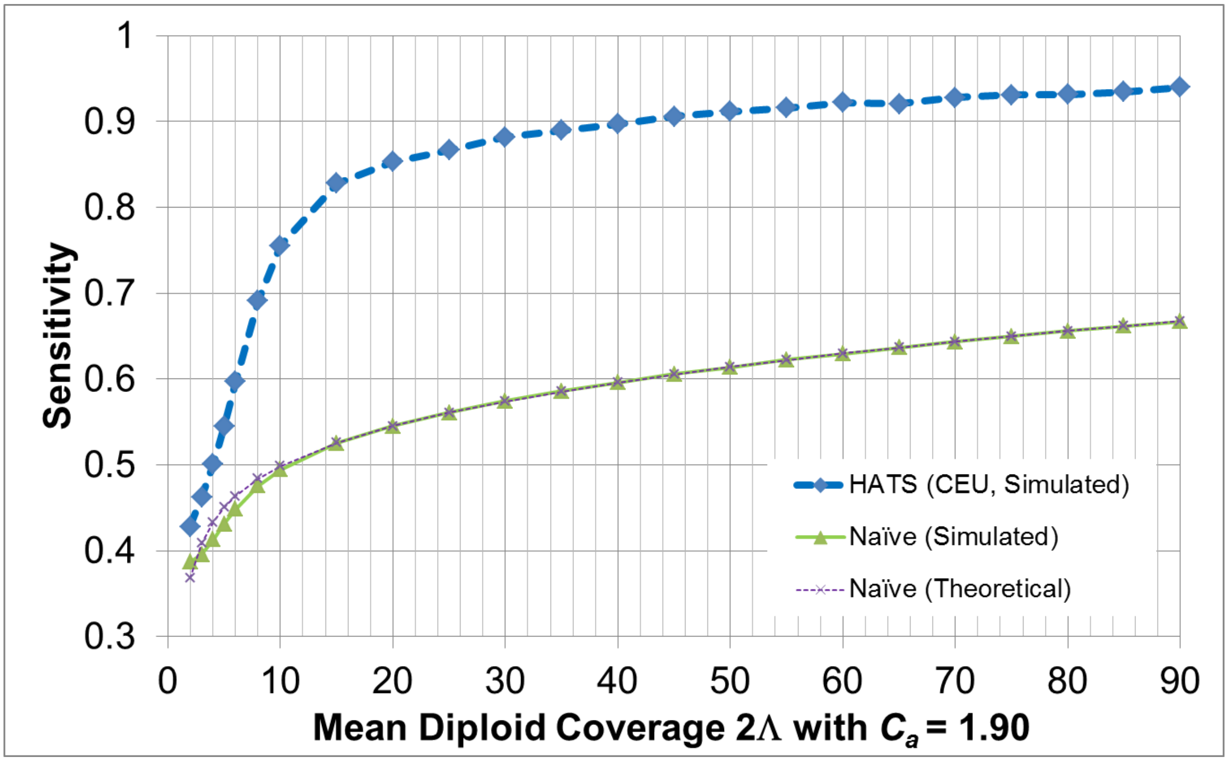
Supplementary Figure 4. Sensitivity of HATS and the Naïve model from simulations, Yoruban (YRI) and Japanese and Chinese (JPTCHB) Training Datasets. This figure displays the sensitivity results for HATS (without Genotyping Error Correction) and the naïve model from the simulation runs. Since the naïve curves for both populations are congruent with each other and the theoretical curve, the same naïve simulation curve is displayed here. The HATS curves for the two populations are also nearly congruent and outperform the naïve model until the latter catches up past a diploid coverage of 40. The training datasets were obtained from the 1000 Genomes Project (http://www.1000genomes.org/).



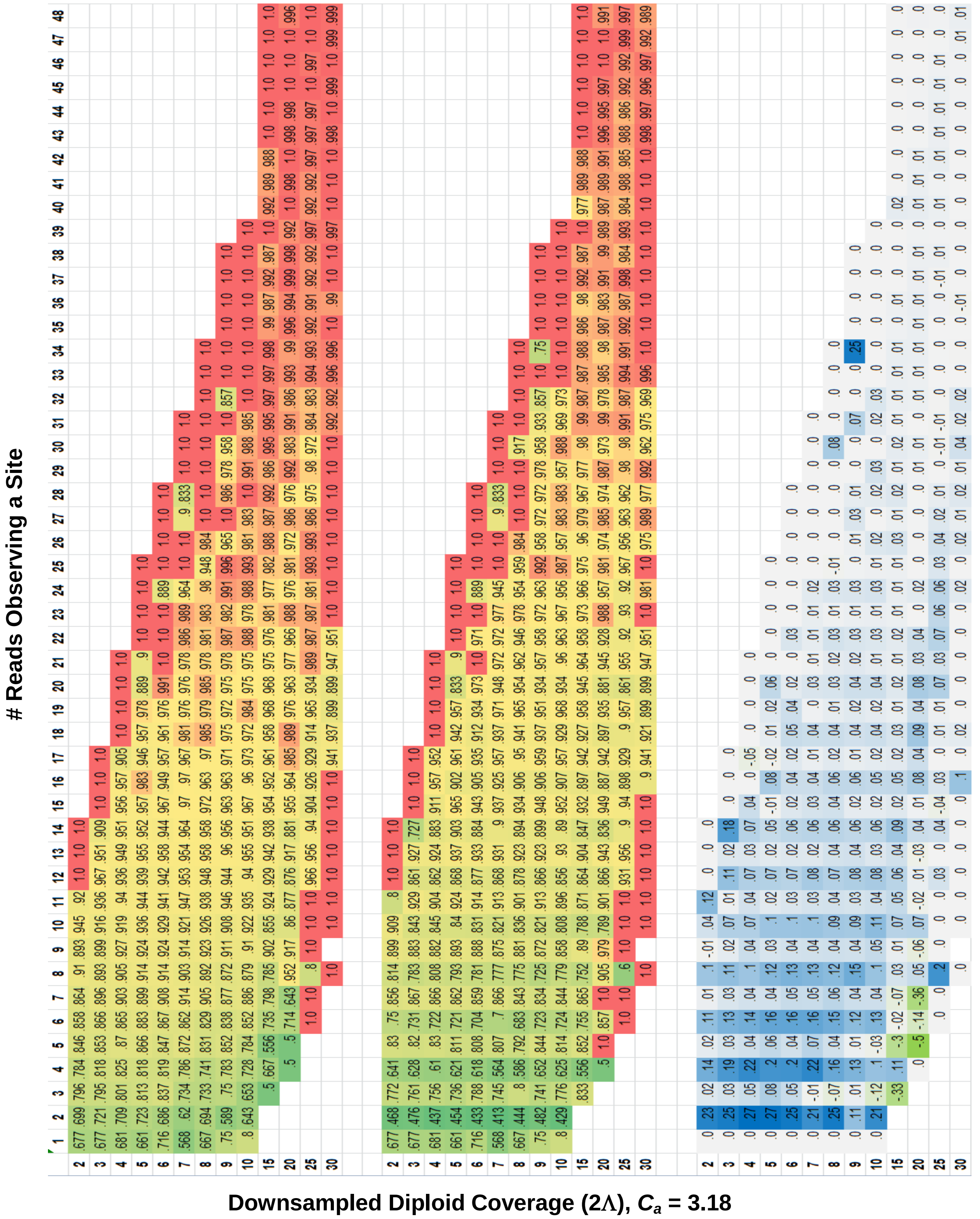
Supplementary Figure 5. Sensitivity of HATS and the Naïve model from simulations with stromal contamination, European (CEU) Training Dataset, Copy Number 2.8. This figure displays the simulation sensitivity results for HATS as well as for the naïve model given a copy number of 2.8, which represents a tumor of copy number 3 with 20% stromal contamination of copy neutral healthy cells. Note that the performance gap between the two methods remains wide, and the naïve method does not catch up to HATS even at a high diploid coverage of 60. The training dataset was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).


# Reads Observing a Site
Simulated Diploid Coverage (2Λ), C_a = 2.5
(A) HATS
(B) Naive
(C) HATS vs. Naive

Supplementary Figure 6: Sensitivity of HATS and the Naïve Model for Breakdown of Reads Observing a Site vs. Diploid Coverage Simulated, with Copy Number of 2.5. The first two plots **(A)** and **(B)** depict the sensitivity for each y-axis bin, where each bin represents a particular number of total reads observing a site. The reads in a bin are generated by a Poisson distribution with diploid mean 2Λ. A range of values for 2Λ are displayed on the x-axis. This figure is generated from simulated data with amplicon copy number of 2.5. Blank bins represent read counts not generated from a particular 2Λ. Sensitivity spans from low (green) to high (red) and is scaled for both methods together. Note that the naïve method possesses periodicity – namely, even numbered bins perform poorer than adjacent odd numbered bins. The reason is that even numbered bins can result in ties in allelic counts, which disable the naïve method from making a call. Odd numbered bins cannot result in ties by definition. **(C)** This plot represents the differences in sensitivities between HATS and the naïve method. The cells are shaded in a gradient corresponding to the sensitivity differences: green (negative difference, naïve is better) to off-white (no difference, labeled as ".0") to blue (positive difference, HATS is better). Note that the periodicity of the naïve model is reflected in the sensitivity differences. This figure depicts HATS generally outperforming the naïve method when stromal contamination is present, even at high 2Λ or when the bin numbers are high. The training dataset was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).

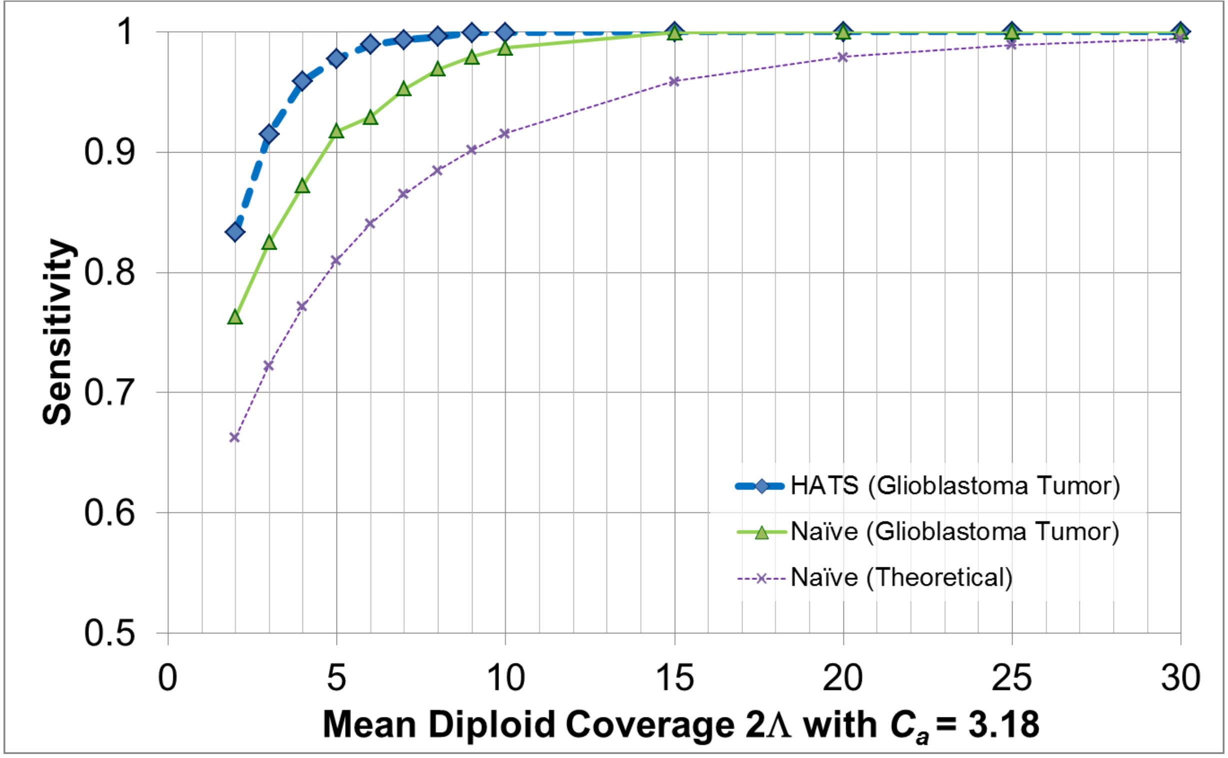


Supplementary Figure 7. Sensitivity of HATS and the Naïve model from simulations, CEU Training Dataset, Copy number 1.9. This figure displays the simulation sensitivity results for HATS and the naïve model with copy number of 1.9, representing a heterozygous deletion mixture. While performance between the two methods remains similar at very low coverage, increasing coverage slightly results in a tremendous performance difference between the two. The training dataset was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).

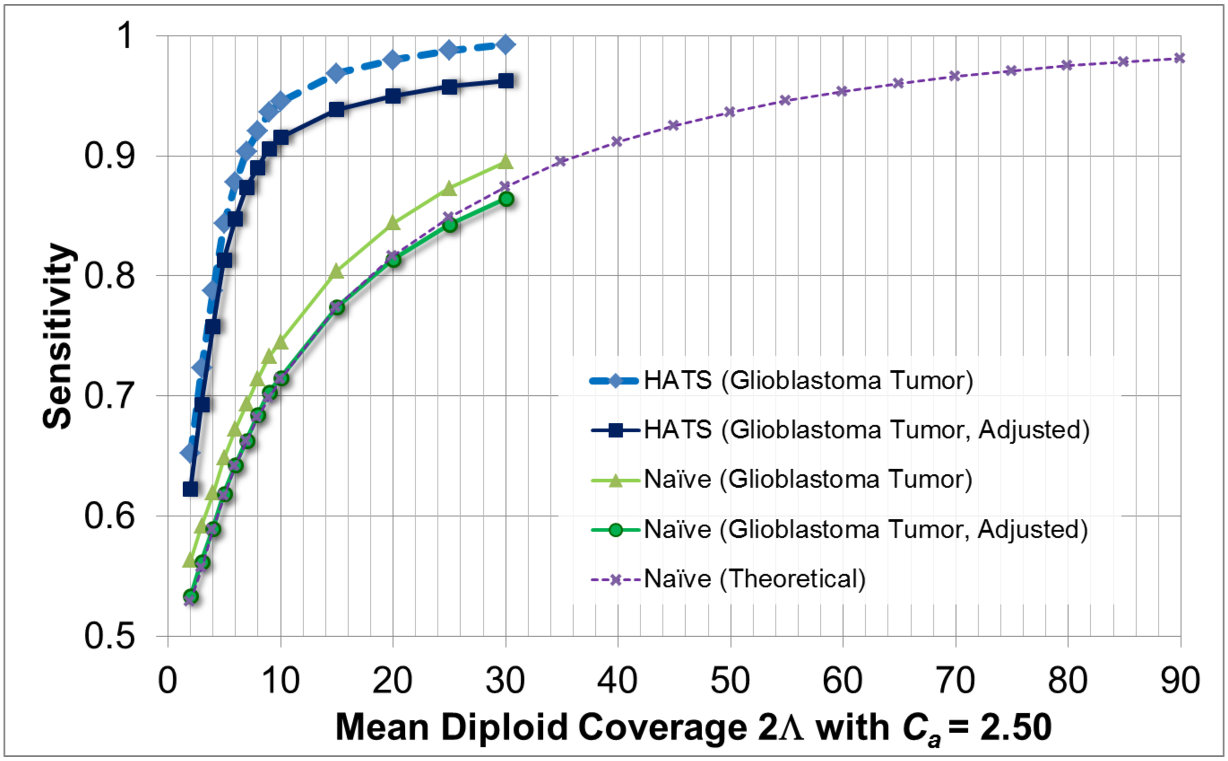


(A) HATS (B) Naive (C) HATS vs. Naive

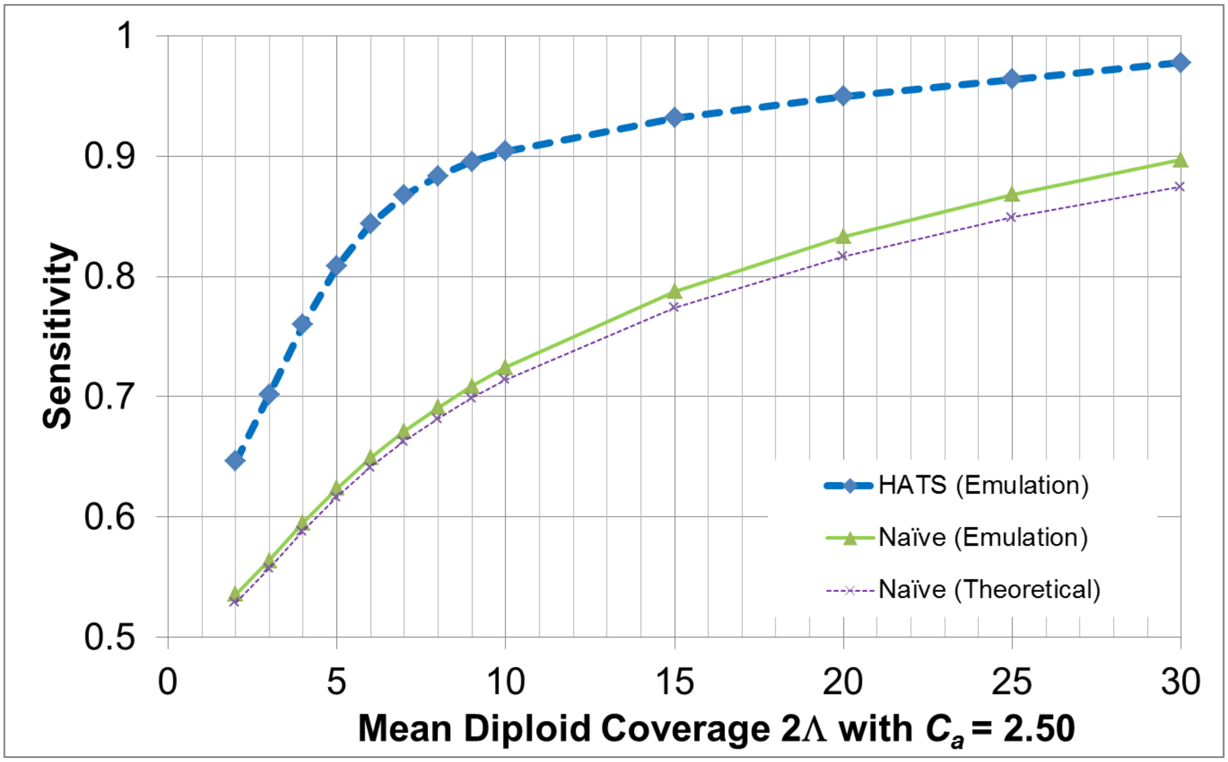
Supplementary Figure 8: Sensitivity of HATS and the Naïve Model for Breakdown of Reads Observing a Site vs. Downsampled Diploid Coverage on a Real Tumor Amplicon, with Copy Number of 3.18. This figure is generated from real data with amplicon copy number of 3.18 undergoing the downsampling evaluation procedure. The first two plots **(A)** and **(B)** depict the sensitivity for each y-axis bin, where each bin represents a particular number of total reads observing a site. The reads in a bin result from a Poisson distribution with diploid mean 2Λ. A range of values for 2Λ are displayed on the x-axis. Blank bins represent read counts not resulting from a particular 2Λ. Sensitivity spans from low (green) to high (red) and is scaled for both methods together. Note that the naïve method possesses periodicity – namely, even numbered bins perform poorer than adjacent odd numbered bins. The reason is that even numbered bins can result in ties in allelic counts, which disable the naïve method from making a call. Odd numbered bins cannot result in ties by definition. **(C)** This plot represents the differences in sensitivities between HATS and the naïve method. The cells are shaded in a gradient corresponding to the sensitivity differences: green (negative difference, naïve is better) to off-white (no difference, labeled as ".0") to blue (positive difference, HATS is better). Note that the periodicity of the naïve model is reflected in the sensitivity differences. This figure depicts HATS generally outperforming the naïve especially when 2Λ is low or the bin numbers are low. However, there are a handful of cases in which the naïve method performs better. Also note that the HATS sensitivities for a bin are not consistent across coverage levels; this is due to the effect of the training data being incorporated. The TCGA data for this patient was obtained from http://tcga-data.nci.nih.gov/tcga/ (with tumor and alignment files obtained from dbGaP with accession number: phs000178.v4.p4).



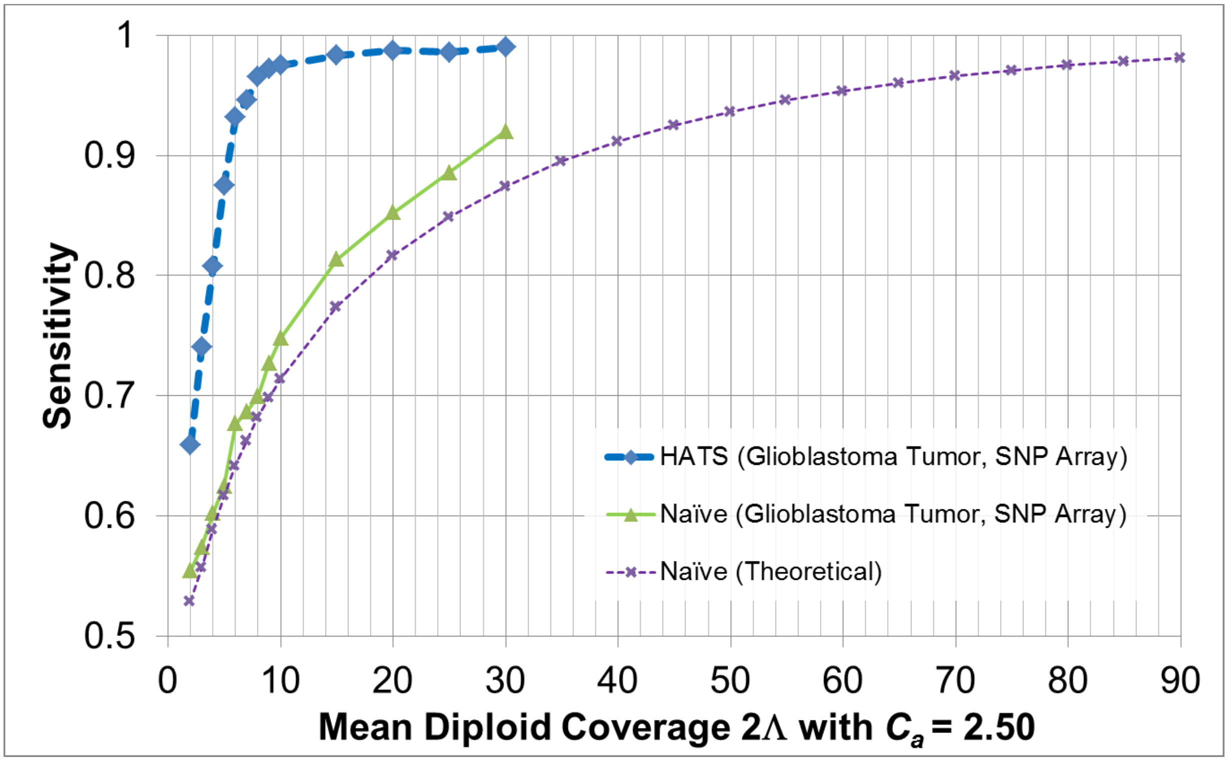
Supplementary Figure 9: Empirical Sensitivity of HATS and the Naïve Model, TCGA Glioblastoma sample (TCGA-06-0877), Chr 19, SNP Array Gold Standard. This figure displays the sensitivity results for HATS and the naïve method on an amplified region (Chromosome 19: 2,181,615 – 2,541,253) in a glioblastoma patient (TCGA-06-0877) obtained from TCGA with local copy number of 3.18. The naïve theoretical curve is included for comparison purposes. The gold standard amplified alleles were obtained via the amplified allelic calls made on SNP array data for the same region from the same sample. Only three heterozygous gold standard sites remained after quality control filtering. The sequencing data read counts were then randomly down-sampled to result in varying coverage levels as displayed on the x-axis (with 400 trials of down-sampling performed per coverage level). The down-sampled read counts were passed to both HATS and the naïve method. The reported amplified alleles were compared with the gold standard to indicate sensitivity. Note that for higher coverages, the performance of both HATS and the naïve method is strong, which is expected as this was observed in the simulations. As coverage decreases, HATS maintains a marked performance improvement over the naïve method. What is further remarkable is that HATS maintains this advantage over the naïve model even with such few heterozygous sites (three). Another thing to note is that the naïve method performs better on sites with an odd number of reads versus an even number, as there is no chance of ties in the former. This explains the slight step effect in the naïve model's sensitivity at $2\Lambda = 5$. The TCGA data for this patient was obtained from http://tcga-data.nci.nih.gov/tcga/ (with tumor and alignment files obtained from dbGaP with accession number: phs000178.v4.p4).



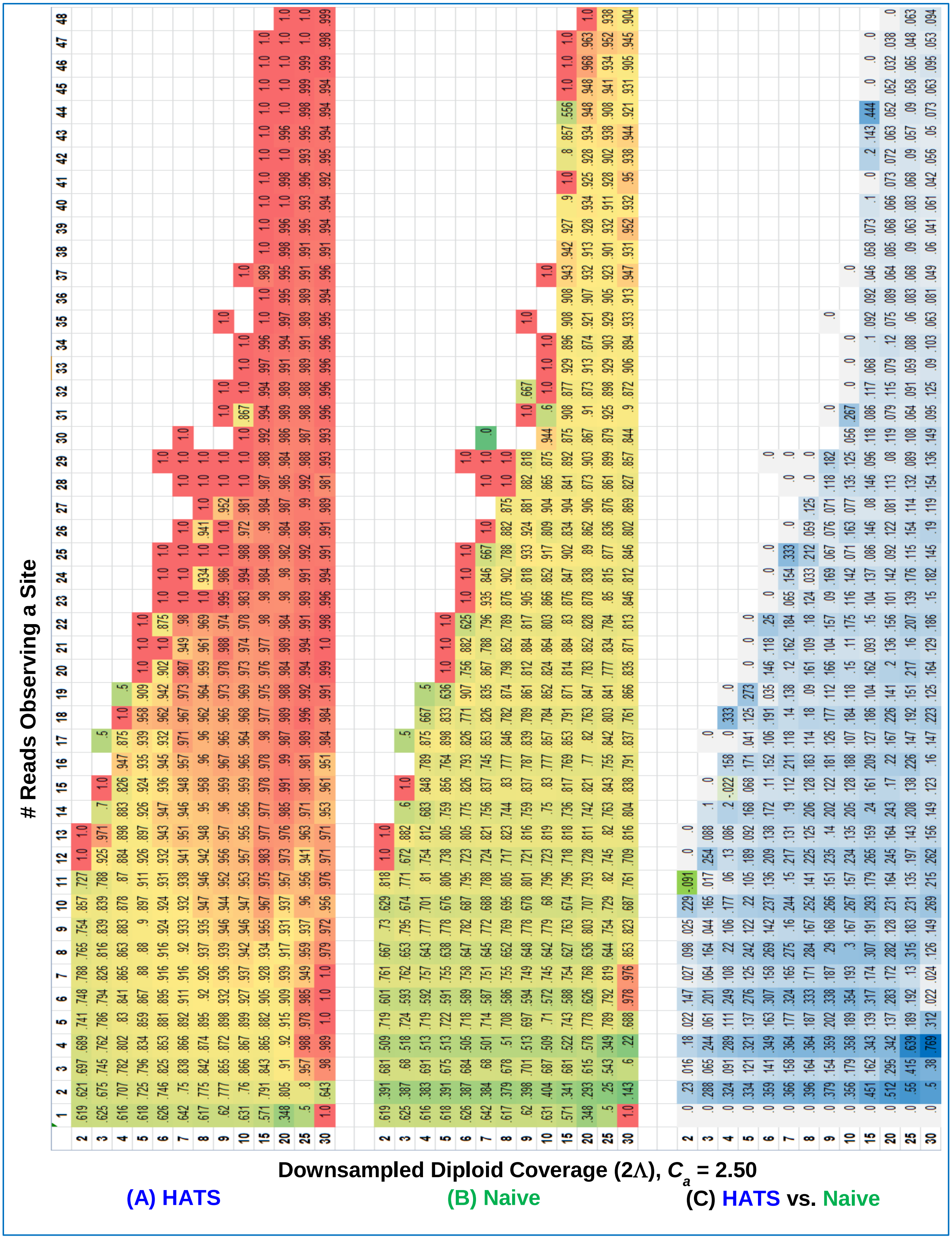
Supplementary Figure 10. Empirical sensitivity of HATS and the Naïve model, TCGA Glioblastoma sample (TCGA-06-0877), Chr 2. This figure displays the sensitivity results for HATS and the naïve method on an amplified sub-region (Chromosome 2: 30,117,665 – 31,000,001) in a glioblastoma patient obtained from TCGA with copy number of 2.50. The empirical curves only extend to a diploid coverage of 30, as the original diploid coverage in the data was ~38.7x for this region. The theoretical curve extends past this to display that very high coverage is necessary in order for the naïve method to theoretically match the performance of HATS. Note that the naïve curve deviates from the theoretical curve, likely due to an imperfect gold standard called by HATS at 38.7x. To correct for this, the average deviation is calculated between the naïve curve and naïve theoretical curve. Both the naïve and HATS curves are then shifted down by this amount to result in the respective adjusted curves, which more accurately indicate HATS' performance at this copy number. The TCGA data for this patient was obtained from http://tcga-data.nci.nih.gov/tcga/ (with tumor and alignment files obtained from dbGaP with accession number: phs000178.v4.p4).



Supplementary Figure 11. Sensitivity of HATS and the naïve model, simulated data with gold standard called on high coverage (emulating Real Data tumor evaluation). HATS called the amplified alleles (assigned as the gold standard) on simulated data generated at 35x diploid coverage with copy number of 2.5. This simulated data is then analyzed as real data would be evaluated, via a down-sampling procedure that uses the called gold standard alleles. The figure displays the results, indicating the deviation of the naïve emulated curve from the theoretical curve. This suggests that the imperfect gold standard used is a cause of this deviation. The training dataset (CEU) was obtained from the 1000 Genomes Project (http://www.1000genomes.org/).



Supplementary Figure 12: Empirical sensitivity of HATS and the Naïve model, TCGA Glioblastoma sample (TCGA-06-0877), Chr 2, SNP Array Gold Standard. This figure displays the sensitivity results for HATS and the naïve method on an amplified sub-region (Chromosome 2: 30,117,665 – 31,000,001) in a glioblastoma patient obtained from TCGA with copy number of 2.50. The empirical curves only extend to a diploid coverage of 30, as the original diploid coverage in the data was ~38.7x for this region. The theoretical curve extends past this to display that very high coverage is necessary in order for the naïve method to theoretically match the performance of HATS. The gold standard amplified alleles were obtained via the amplified allelic calls made on SNP array data for the same region from the same sample. Only 43 heterozygous gold standard sites remained after quality control filtering. The sequencing data read counts were then randomly down-sampled to result in varying coverage levels as displayed on the x-axis (with 400 trials of down-sampling performed per coverage level). With this stromal contamination present, HATS outperforms the naïve method at all diploid coverage levels. The TCGA data for this patient was obtained from http://tcga-data.nci.nih.gov/tcga/ (with tumor and alignment files obtained from dbGaP with accession number: phs000178.v4.p4).



# Reads Observing a Site
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
2 3 4 5 6 7 8 9 10 15 20 25 30
Downsampled Diploid Coverage (2Λ), C_a = 2.50
(A) HATS
(B) Naive
(C) HATS vs. Naive

Supplementary Figure 13: Sensitivity of HATS and the Naïve Model for Breakdown of Reads Observing a Site vs. Downsampled Diploid Coverage on a Real Tumor Amplicon, with Copy Number of 2.50. This figure is generated from real data with amplicon copy number of 2.50 undergoing the downsampling evaluation procedure. The first two plots **(A)** and **(B)** depict the sensitivity for each y-axis bin, where each bin represents a particular number of total reads observing a site. The reads in a bin result from a Poisson distribution with diploid mean 2Λ. A range of values for 2Λ are displayed on the x-axis. Blank bins represent read counts not resulting from a particular 2Λ. Sensitivity spans from low (green) to high (red) and is scaled for both methods together. Note that the naïve method possesses periodicity – namely, even numbered bins perform poorer than adjacent odd numbered bins. The reason is that even numbered bins can result in ties in allelic counts, which disable the naïve method from making a call. Odd numbered bins cannot result in ties by definition. **(C)** This plot represents the differences in sensitivities between HATS and the naïve method. The cells are shaded in a gradient corresponding to the sensitivity differences: green (negative difference, naïve is better) to off-white (no difference, labeled as ".0") to blue (positive difference, HATS is better). Note that the periodicity of the naïve model is reflected in the sensitivity differences. This figure depicts HATS generally outperforming the naïve method when stromal contamination is present, even at high 2Λ or large bin number. Also note that the HATS sensitivities for a bin are not consistent across coverage levels; this is due to the effect of the training data being incorporated. The TCGA data for this patient was obtained from http://tcga-data.nci.nih.gov/tcga/ (with tumor and alignment files obtained from dbGaP with accession number: phs000178.v4.p4).

Software

HATS was implemented in the Java programming language. The source code is available at: https://sourceforge.net/projects/tumorhats/. We recommend using JRE 1.6 or higher.

While a non-trivial amount of time might be required of the user to organize data into HATS' input file format (described at the linked site), processing the data itself is very fast. On short regions, HATS takes several seconds to process the regions (including time to load the training data files), as was measured on a Core 2 Duo 2.4 GHz machine with 6 GB RAM. Wider regions will take proportionally longer (when ignoring the overhead of loading the training data). HATS is memory intensive, and as such, a 64-bit machine with 4+ GB of RAM is recommended, depending on the width of regions being analyzed and whether the especially memory intensive Genotype Error Correction feature is activated.

References

Baker J. 1975. The DRAGON system--An overview. *Acoustics, Speech and Signal Processing, IEEE Transactions on* **23**(1): 24-29.

Durbin R. 1998. *Biological sequence analysis : probabalistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK New York.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**(10): 1253-1260.

Liu Z, Li A, Schulz V, Chen M, Tuck D. 2010. MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PLoS One* **5**(6): e10909.

Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**(14): e431-439.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**(11): 1665-1674.

Yoon BJ. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics* **10**(6): 402-415.