**SNPTools: An integrative variant calling pipeline for accuracy genotype/haplotype inference in low coverage NGS data**

Yi Wang1, James Lu1,3, Jin Yu1, Richard A. Gibbs1,2, Fuli Yu1

1) Human Genome Sequencing Center, Baylor College of Medicine, Houston TX

2) Department of Human and Molecular Genetics, Baylor College of Medicine, Houston TX

3) Department of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston TX

**ABSTRACT [I did not work in this area]**

Next generation sequencing is a powerful approach for discovering genetic variation. Current methods for detecting single nucleotide polymorphisms (SNPs) are primarily designed to detect SNPs from a single individual and not from a population. In this paper, we describe methods for high quality discovery, genotyping and phasing of SNPs from low coverage (~5X) parallel sequencing of populations using a pipeline called SNPTools that contains the following innovations. We introduce, Effective Base Depth (EBD), a non-parametric statistic that summarizes both base and alignment quality information at the nucleotide level and enables statistical modeling on sequencing data. Then, variance ratio based site scoring statistics discovers SNP sites with high sensitivity and specificity. BAM-specific Binomial Mixture Modeling (BBMM) captures the heterogeneity of sequencing data and generates robust raw genotype likelihoods from sequence information. Our novel imputation engine refines raw genotype likelihoods by utilizing LD information and produces high quality genotype and haplotype calls. Our pipeline also allows for integration of multiple existing data sources, such as from genotyping arrays, which improve imputation of genotypes non-trivially. SNPTool’s input/output (I/O) and storage aware design leads to unmatched performance on large sequencing datasets. We apply our pipeline to the 1000 Genomes Project Phase 1 low-coverage dataset and conclude that low coverage sequence genotyping accuracy is comparable with that of microarray with the aforementioned efforts.

**INTRODUCTION [I did not work in this session]**

Next-generation sequencing technologies such as Illumina Hi-Seq and Life Technologies SOLiD are fast becoming a realistic choice for population level genomic studies. Two cost-effective strategies are popular amongst large population sequencing projects are high coverage exon-capture sequencing (>20X) and low coverage (3-5X) whole genome sequencing. The 1000 Genomes Consortium has employed both methods to interrogate population level variation *(1)*. In this report, we focus on low coverage whole genome data although our methodologies can be applied to high coverage exome capture data.

Low coverage population sequencing strategy aims to achieve high sensitivity population level variant discovery and high genotyping accuracy by making use of the n redundancy at specific nucleotide sites across multiple samples. As most genetic variants are inherited rather than generated from *de novo* mutations (e.g. mutation rate of ~1.1 x 10-8)*(2)*, it is possible to discover most genetic variants by combining large sample information with low coverage sequencing. This was initially demonstrated by Liti el al., *(3)* and more recently in the 1000 Genomes low coverage pilot *(1)*. Further, linkage disequilibrium (LD) between neighbouring variants can be used for joint genotype imputation. For example, if two SNPs are tightly linked in the population (r2=1), then their respective read depth can be shared/summed to obtain much more accurate genotype calls at both sites *(4-7)*.

Currently, SNP discovery methods such as Samtools *(8)* and SOAPsnp *(9)* detect SNP sites by evaluating neighborhood sequence information, read alignment quality and other measures in relation to the reference genome for a single sample. For a population based studies, each sample is evaluated separately and then aggregated. For read level information such as mapping and base quality, these methods apply simple heuristics to detect SNP variants; this read level information is not summarized for formal statistical modeling of the data, rather, thresholds are tuned for single or multiple variables *(9, 10)*. Filtering based on cutoffs for these parameters, while straightforward, reduces power to detect variants because variants cannot be jointly analyzed across a population. Further, differences in sequencers, reagents, and mapping make these methods difficult to generalize to new data due to heterogeneity across sequencing platforms *(11, 12)*.

Opportunities exist to take improve SNP detection and genotyping. Proper modeling of mapping and quality information at individual reads when coupled with data across populations improves confidence in and sensitivity of site level calls *(6, 7, 13)*. Population level aggregation of read level information reduces false negatives in low coverage sequencing: for example, a heterozygote site covered by 5x coverage has a (1/2)5=3% probability of missing at least one allele. The error rate drops geometrically with the increase of samples. It is also worth mentioning that the false negative calls in whole exome capture are non- trivial given the unevenness of capture efficiency and random read depth variation in any given capture. In addition, although many SNP microarray genotyping data sets have been accumulated, this information has not been included with NGS data due to the lack of methods for joint analysis. We provide methods for integrating previous genotyping data into our imputation pipeline; combining these data produces more accurate imputed genotypes.

Our pipeline achieves both high sensitivity and specificity SNP discovery and high accuracy genotype calling with computation efficiency. These methods have been utilized to produce one of released call sets for the low-coverage portion of the 1000 Genomes Project Phase 1 (in preparation).

**RESULTS**

***Overview of our SNPTools pipeline***

We develop an integrative pipeline, “SNPTools”, which can achieve high quality (1) SNP discovery, (2) genotype likelihood estimation, and (3) genotype/haplotype inference via imputation in low coverage data collected from heterogeneous NGS platforms. This pipeline takes inputs standard mapping/alignment files in Binary sequence alignment (BAM) format [REF] produced from various widely applied tools (e.g. BWA [REF], BFAST [REF]), and outputs phased genotypes in the Variant Calling Format (VCF) [REF] (Figure 1).

SNPTools is designed and implemented in a modular fashion. Users have the option to apply the pipeline as an integrative NGS variant caller or to apply individual component to generate intermediate results. These intermediate results are compatible with other available tools. For example, one can generate a list of variant sites using Samtools [REF], but then apply SNPTools to generate genotype likelihoods (GL). Genotypes and haplotypes can then be imputed from these genotype likelihoods with SNPTools or with other imputation programs such as Beagle. SNPTools is divided into four modules that each comprise a different type of analysis: (1) Summary of base and mapping quality, (2) Discovery of SNP sites; (3) Estimation of genotype likelihood; and (4) imputation of genotype and haplotypes (Figure 1).

***Effective Base Depth (EBD) summarizes read depth information after recalibration with base quality and mapping quality information***

Despite lower base accuracy in NGS data relative to data from first generation capillary Sanger sequencing [REF: Wheeler, Metzker, Shendure], statistics such as base quality and mapping/alignment quality can be utilized to reduce both Type I ( and Type II errors in variant calling [REF: Heng Li 2011, DePristo 2011]. These quality scores are under continuous improvement and are regularly recalibrated by different tools (REF: DePristo M 2011) to more closely reflect the true underlying sequence. Current methods [REF: Shendure; VarScan 2009] utilize this information by applying a series of filters with empirical cutoffs to reduce errors. For example a user may decide to eliminate all bases with a Phred-like Quality score [ref[ below 20 and reads with Mapping Quality score 0 [ref]. This approach is suitable for high coverage studies [REF: Shendure; VarScan 2009], however it can result in greatly reduced sensitivity in data where average coverage is ~4-6X coverage per individual.

To maximize the usable read depth information, which is necessary to power SNP detection in our model -based framework, we calculate an “effective base depth” (EBD) for each genomic position by aggregating both base and read mapping quality (Equation 1).

[Equation 1]

An EBD value for each nucleotide (A, T, C, G) at a given genomic site is extracted by summing over the base quality and mapping quality for each read that covers the putative site. This value can be considered a weighted read depth in our pipeline and is used in our pipeline, particular for SNP site detection. EBD can be combined across multiple samples in an “mpileup” fashion, so we eliminate the “BAM-mpileup” operation [I don’t know what you mean by this].This file format (.ebd) allows for high performance in I/O when processing a large numbers of samples and sites. The storage footprint of .ebd file is small compare to the corresponding BAMs, with almost 30 fold compression of

***SNP site discovery based on variance ratio statistic***

One major challenge in variant site discovery in low coverage whole genome data is the high level of variance in the distribution of alternate alleles for each sample. The challenge is exacerbated by the relatively high 1-2% error rate from sequencing and mapping which can result in false alternative alleles. Commonly, more than one alternate base is present at a low background level. Further, rare variants with allele frequency < 5% are difficult to identify because they these allele counts are relatively low when compared to the reference allele and may not be discernable from background sequencing noise.

We compensate for these deficiencies by first adopting a bi-allelic SNP assumption. While this limits our ability to detect variant sites with multiple alleles segregating in the population, it nonetheless reduces two thirds of raw sequencing errors and ease downstream analysis. Then, in order to select the correct alternative allele out of three candidate bases, we applied a scoring statistic that derives from a Binomial assumption. We select the alternative base that maximizes the variance scoring statistic (Equation 2).

 [Equation 2]

The scoring statistic is a single-parameter model that takes inspiration from the binomial distribution. In our model *ri* is the reference read count in sample *i*, *ai* is the read depth of the alternative allele in sample *i,* where there are *n* total samples. *N* is the total amount of read depth over n samples, and *p* is the probability of alternative allele that can be attributed to sequencing or mapping/ alignment errors.

In the numerator, we evaluate the binomial variation, which is the difference between the variance seen for a particular alternative allele when compared to the null hypothesis, that alternative alleles are driven by sequencing or mapping error. We then utilize the genotype property of biallelic SNPs; for a SNP, there are only three possible genotypes (Ref/Ref, Ref/Alt, and Alt/Alt) with corresponding binomial parameters (p= 0, 0.5, and 1, respectively). Thus in the demonimator, we calculate the variation when the SNP is a true SNP. This test outperforms Fisher’s index of dispersion test in sites where alternative bases are introduced by high systematic sequencing or mapping error (e.g. 10%).

***Estimation of genotype likelihoods by BAM-specific Binomial Mixture Modeling (BBMM)***

Sequence produced by different sequencing centers can create heterogeneity in BAM files due to differences in aligners, sequencers and operational parameters. Due to this operational heterogeneity, we developed *BAM-specific Binomial Mixture Modeling (BBMM)* to model the heterogeneity within each sample BAM. This allows SNPTools, to maximize the read depth evidence and convert it into accurate genotype likelihoods for each of the three genotypes: Ref/Ref, Ref/Alt, and Alt/Alt.

*BBMM* models each BAM individually. For example, a single individual may have *n* BAMs. Each BAM may be generated using different sequencing platforms, sequenced with capture, or even sequenced on different lanes. We model the *i*th BAM with a mixture of three binomial ratios, with a parameter representing the three possible genotypes (Ref/Ref, Ref/Alt, Alt/Alt) respectively. As these BAMs are drawn from the same individual, we evaluate the joint distribution all BAMs. In the case of single BAM however, the model simplifies as a mixture of three binomial distributions. For a particular site m, the genotype gm is drawn from a genome wide multinomial distribution of three genotype class (Ref/Ref, Ref/Alt, Alt/Alt), denoted as gm ~ multinomial (wr, wh, wa). If we had a microarray dataset of gm, it would straight forward to estimate (pri, phi, pai) and (wr, wh, wa) by simple counting using read depth information.

[ need to decipher the code before I can write this section]]]

However in our situation, gm is not observed, but it can be treated as latent variable that can be solved using the Expectation-Maximization (EM) algorithm [dempster].

 [is this the prior over the read depth of reference allele for individual I]

 [ this formula is not correct, you can’t set the conditional on its self… also, his parameters keep changing… in this writeup ]

***Raw genotype likelihood*** can be obtained directly once we obtain the estimation of the genotype ratios: p(rm,am|gm=g)=B(am|pg,rm+am). We use these likelihoods as HMM emission probabilities in our imputation engine detailed later.

where  is the number of reference reads and  is the number of alternative reads.

*BAM-specific Binomial Mixture Modeling (BBMM)* was developed to model the heterogeneity within each BAM that may arise due because of different aligners, sequencers and parameters and then convert corresponding read depth information into the genotype likelihood of one of three genotypes: Ref/Ref, Ref/Alt, and Alt/Alt. To evaluate the quality of each BAM, we generate

the emission probability of a read being Ref given a Ref/Ref genotype against

the probability of a read being Ref given a Alt/Alt genotype in Figure 3b. We found these two values to be negatively correlated with an r2=0.72. However this plot also revealed samples located on the up-left part of Figure 3b that exhibit a strong deviation from the clusters for Illumina, SOLiD and Solexa 454 samples. Upon examination, these sequencing runs were considered of low quality due to operational issues such as lane swap, or poor coverage statistics.

*The BBMM algorithm generates raw genotype likelihoods in a* three dimensional vector (2 degree of freedoms) representing the Ref/Ref, Ref/Alt and Alt/Alt genotype. In order to visualize the genome wide distribution of alleles, we plotted the histogram of reference count for a single representative sample, HG00176, in Figure 3c. We found that the three genotypes segregate and cluster independently. Note that for sites that were not covered by reads, we could not generate genotype likelihoods and so all genotypes are equally likely. The expected genotypes will have to be imputed from neighboring sites depending on LD.

***Imputation provides accurate genotypes***

Using the genotype likelihoods generated in the previous step, SNPTools can impute haplotype and genotype calls with confidence scores stored in VCF format. (See Methods).

*Identification of parameters for SNPTools through testing on Chr20*

Two key parameters affect the accuracy and speed of the imputation engine, "chunk size" or the two of sites to be imputed, and MCMC (markov chain monte carlo) sampling cycles. We generated several genotype imputation call sets of chromosome 20 using different parameter settings and evaluated the results by comparing the genotype concordance against known genotypes from HapMap3, OMNI and Affymetrix Axiom data sets. We measure the error by the discordance rate (%) for the genotype classes, Alt/Alt, Ref/Alt and Ref/Ref and also evaluate an overall discordance rate and a non-ref/ref discordance rate.

With an MCMC =200, and chunk size =200, the best overall discordance rate was 0.61%, 0.67% and 0.62% when compared to HapMap3, OMNI and Axiom respectively. This is similar to single site high coverage capture discordance rate of (what is the rate) [ref] and is close to microarray errors (0.2~0.3%) [ref]. We also found that the alt/alt and ref/alt genotype error rate was 1.21% and 0.89%, respectively. These error rates are higher than the ref/ref error rate when evaluated using HapMap3 genotypes (Figure 5a). While imputation accuracy can be improved with increased expenditure of computation resources, as Figure 5b shows, the amount of improvement after utilizing 1000 CPU equivalent months is marginal. A CPU month is a unit of computation defined as 1 CPU core working for 1 month.

*SNPTools generates accurate haplotype phases*

We next assess the accuracy of haplotype phasing by comparing our phasing results for our imputed genotypes to both HapMap3 CEU and YRI trios from the 1000 Genomes Pilot [ref]. We mirror our error rate using Marchini’s definition of Incorrect Genotype Proportion [ref]*(17)*, except that we define the error as the percentage of genotypes incorrectly phased. In Figure 5c, we found that the phasing error race was less than 1% with stretches of nucleotides less than 50kbp. We also found that the error rate grows approximately linearly with the length of alleles being phased. When compared to Beagle [ref], SNPTools also produces marginally higher quality phasing. **[do we want this instead... i actually don't have the real results nor had i planned on generating them.... this is hyun's result]**

***Integration of previous genotype datasets can improve genotype concordance***

To date, numerous types of genetic variants have been discovered by various technologies. For example, for samples in the 1000 Genomes project, SNP genotypes exist from the HapMap project. Further genotypes were interrogated using Illumina OMNI and Affymetrix Axiom arrays. We expect that there will continue to be heterogeneity in data sets as whole exome catpure, whole genotype seqeunce and array based technologies develop to interrogate INDELs, structural variants and copy number variants.

With our pipeline we are able to integrate known and future array data sets from different platforms to improve genotype accuracy. We incorporate these genotypes as high confidence priors, which can be used to update the corresponding raw genotype likelihoods. Incorporation of strong priors improves imputation and reduces genotype errors both for on-array sites, and for off array sites that are imputed using LD (Figure 6a).

We completed three imputations on chromosome 20 by integrating HapMap3, OMNI and Axiom data sets independently and then evaluated their discordance rate again the other two array genotypes. Results are shown in Figure 6b. We found improvements both for genotypes at array locations and off-array. For on-array sites, genotype concordance against other data sets was improved; for example when we integrate Axiom’s genotypes, the discordance rate when compared to HapMap3 and OMNI dropped from 0.34% to 0.14% and from 0.30% to 0.09%, respectively. Off-array genotypes are also improved slightly, genotype dis-concordance dropped ~0.01%- 0.04% when compared to non integrated imputed genotypes.

**DISCUSSION**

Accurate detection and genotyping of SNPs is important for the detection of common and rare sequence variants that contribute to population variation and medical disease. Low – coverage whole genome information poses computational challenges for the accurate detection and genotyping of population level SNP variation. Methods have been developed to aggregate population level genetic information [ref]; however we provide a new pipeline, SNPTool that can detect, impute and phase SNPs with high sensitivity and specificity from low coverage whole genome sequencing for population studies. We apply this pipeline to the production of 1000 Genomes Phase 1 low coverage genotypes, but expect that this pipeline will be particularly valuable to new genome wide association studies.

We propose several novel methods to the problem of SNP detection and genotyping in low coverage whole genome sequencing. We first aggregate mapping and base quality errors for each read into a single parameter, EBD, which weighs the evidence from each read in determining each potential SNP. We then use the EBD, and leverage site level genomic information across multiple individuals to provide high sensitivity SNP site detection. Given a site list, genotype likelihoods for each SNP in each individual are calculated using BBMM. BBMM models the heterogeneity of each BAM separately thus preventing the variation between sequencers and sequencing centers from overwhelming the evidence for each genotype. In addition, BBMM does not require the site list to emerge from the SNPTools pipeline, rather BBMM can intake any site list and evaluate the likelihood of potential SNPs for each sample. A comprehensive SNP list from a continually updated database such as dbSNP, ensures that every known SNP is examined in future genome wide studies.

After generating genotype likelihoods, we are able to impute genotypes from these genotype likelihoods using site- level evidence and LD. At this step we further provide a mechanism to incorporate SNP array genotypes as strong priors. Given the large amount of knowledge that already and will continue to originate from array based technology, our pipeline provides opportunities to integrate those results with next generation sequencing with trivial effort. With integration, SNPTools provides genotype accuracy, even with low coverage whole genome sequences, similar to array based technologies.

Lastly our imputation platform also provides opportunities to incorporate genotype likelihoods for bi-allelic structural variants, copy number variants and INDELs produced by other variant callers [ref- dindel, szyergy, gatk, samtools, …etc….]. Our imputation pipeline has shown superior imputation concordance with short INDELs in 1000G Pilot 1 data relative to IMPUTE2 [under review]. We look forward to forward to improving this feature which has the potential to greatly increase the types of variation available to researchers in future studies.

**MATERIALS AND METHODS**

**Data Set Description**

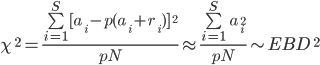
We apply our pipeline on 1000 Genomes Project Phase I low coverage BAMs (20110213 BAM index file). The data set contains 1094 individuals (1103 BAMs) from 14 populations (ASW, CEU, CHB, CHS, CLM, FIN, GBR, IBS, JPT, LWK, MXL, PUR, TSI, YRI). These BAMs were sequenced by three different platforms: Illumina (946 BAMs), SOLiD (142 BAMs) and 454 (15 BAMs) with an average coverage 5X. These BAMs passed a series of consensus preprocessing procedures described in the Pilot paper [ref].

**Effective Base Depth**

Base error and mapping error are calculated from base quality and mapping quality using the Phred score definition in SAM format specification [ref]. EBD can be considered as a weighted read depth in the whole SNPTools pipeline. For reads with zero base quality or mapping quality, that information is not included. EBD increases with mapping and base quality, so if both values are high, then EBD is 1. For multiple reads that cover a particular site, we sum the EBD of the same bases to get the site level EBDs of the four bases respectively. We use the term ‘read depth’ as site level EBD later.

***Variant allele selection using a goodness-of-fit***

We used a goodness-of-fit test based on binomial assumption to select the correct alternate allele (i.e. variant allele). We assume that the site is non-polymorphic under null hypothesis and there is a small chance (p~1%) that we observe alternative base due to sequencing or mapping error. Then the pooled chi-squared statistics can be written as



where ai, ri is the alternative read depth and reference read depth of the the *i*th individal respectively, N is the total read depth of the entire population. The approximation is reasonably good when p(ai+ri) is close to zero which is realistic in low-pass case.

**SNP Site Discovery**

*Site filtering*

Filtering of the site list is an optional procedure that increases the specificity of SNP discovery, however with some reduction in sensitivity. Our pipeline employs four widely used criteria: (1) maximum population read depth; (2) minimum population read depth; (3) strand bias as tested by 2x2 contingency table (rows are reference and alternative bases and columns are positive and negative strand), and (4) position bias. The mean of the reference base and the alternative base for each of the above criteria is evaluated by t-test.

**Genotype Likelihood Generation**

***BAM-specific Binomial Mixture Modeling (BBMM)*** was developed to model the heterogeneity of BAMs and convert read depth information into the likelihood of one of three genotypes, Ref/Ref, Ref/Alt and Alt/Alt. Given n BAMs of one individual, the BAMs could be sequenced by different platforms (e.g. with/without capture) or on different lanes. We model the *i*th BAM with a mixture of three binomial ratios, each with a parameter p (pri, phi, pai) for the three genotypes (Ref/Ref, Ref/Alt, Alt/Alt) respectively. Since these n BAMs are from the same individual, their information can be evaluated by joint estimation. For a particular site m, the genotype gm is drawn from a genome wide multinomial distribution of three genotype class (Ref/Ref, Ref/Alt, Alt/Alt), denoted as gm ~ multinomial (wr, wh, wa). If we had a microarray dataset of gm, it would straight forward to estimate (pri, phi, pai) and (wr, wh, wa) by simple counting using read depth information.

However in our situation, gm is not observed, but it can be treated as latent variable that can be solved using the Expectation-Maximization (EM) algorithm [dempster]. Note that in the case of single BAM, the model can be simplified as a mixture of three binomial distributions, which is called BBMM by us originally.

 [is this the prior over the read depth of reference allele for individual I]

 [ this formula is not correct, you can’t set the conditional on its self… also, his parameters keep changing… in this writeup ]

***Raw genotype likelihood*** can be obtained directly once we obtain the estimation of the genotype ratios: p(rm,am|gm=g)=B(am|pg,rm+am). We use these likelihoods as HMM emission probabilities in our imputation engine detailed later.

where  is the number of reference reads and  is the number of alternative reads.

**Genotype and Haplotype Imputation**

***Existing methods*** for haplotype inference and genotype imputation methods are based on the following iterative framework: initialize all haplotypes by random guesses, predict and update each individual’s haplotypes in turn by assuming other haplotypes in the population are known. Different methods employ different models and strategies in the prediction step. For example, EM algorithm models a haplotype as a random drawn from a population haplotype pool without mutation and recombination. While the popular FDSL distribution models a haplotype as a mutated mosaic combination of genomic segments from several haplotypes in the population. The FDSL distribution and corresponding estimation techniques have been applied on real datasets successfully in the past few years.

***Challenges*** for the FDSL distribution comes from two issues. The major and practical issue is that the estimation of such multi-haplotype mosaic model by HMM is expensive due to the large number of hidden status (O (N2) states and O (N4) transitions). This is problem becomes more serious in low coverage sequencing case as large sample size are desired and more iterations are required due to genotype uncertainty. The minor and theoretical issue is that the mosaic combination involving more than two haplotypes in the population is not biologically meaningful. For example, under FDSL distribution an A-B-C hybrid haplotype may be as probable as an A-B-A hybrid haplotype, but the latter is more possible due to the biological nature of diploid recombination.

***Constrained FDSL distribution*** models a haplotype as a mutated recombinant of two parental haplotypes in the population (Figure 4b). This model is more biologically meaningful and eases the corresponding estimation. The joint distribution of an individual’s genotype probabilities vector G, his haplotypes h, the 4-parental haplotypes set H from the population, a constant mutation rate M and a recombination rate vector R can be written as:

P(G|h) is the genotype likelihood and P(h | H, M, R) is modeled by a 4-state-HMM which is detailed in the supplement. It is also possible for us to obtain the marginal likelihood of the 4 parental haplotypes by integrating the hidden child haplotype h.

Due to the small number of HMM status and efficient implementation, both (1) and (2) can be evaluated or estimated efficiently in O (L) time (~10-5s ), where L is the number of SNP loci.

Formula (2) enables us to search a good parental haplotype set H for an individual carrying G. Although the candidate space of H is still huge (O (N4)), we are able to sample H in constant time (~10-2s) by a Metropolis-Hastings (MH) Sampler. The MH sampler proposes an alternation of one haplotypes in H at a time, evaluate the proposal by (2) and determine whether accept such change or not. Once we find a good set of parental haplotypes, it is straight forward for us to sample the hidden haplotype h using H according to (1).

***The algorithm*** is summarized as follow (Figure 4a):

1. Guess all haplotypes randomly
2. For each individual
   1. Sample(search) his 4 parental haplotypes H by a MH sampler in light of (2)
   2. Sample(infer) his haplotypes using parental templates H according to (1)
3. Repeat 2 for a given number of times

**Discordance Rates**

Total/pooled discordance rate and “non-ref” discordance rate (defined as <http://www.broadinstitute.org/gsa/wiki/index.php/File:GenotypeConcordanceGenotypeErrorRate.png> ) are also reported in Figure 5a. [define in methods]

**REFERENCE**

FISIEER, R. A. (1954). Statistical Methods for Research Workers, sect. 19, 12th ed. Edinburgh: Oliver and Boyd.

1. R. M. Durbin et al., A map of human genome variation from population-scale sequencing, *Nature* **467**, 1061-1073 (2010).

2. J. C. Roach et al., Analysis of genetic inheritance in a family quartet by whole-genome sequencing., *Science (New York, N.Y.)* **328**, 636-9 (2010).

3. G. Liti et al., Population genomics of domestic and wild yeasts., *Nature* **458**, 337-41 (2009).

4. T. International, H. Consortium, A haplotype map of the human genome., *Nature* **437**, 1299-320 (2005).

5. C. S. Carlson et al., Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium., *American journal of human genetics* **74**, 106-20 (2004).

6. J. Duitama et al., Linkage disequilibrium based genotype calling from low-coverage shotgun sequencing reads., *BMC bioinformatics* **12 Suppl 1**, S53 (2011).

7. S. Q. Le, R. Durbin, SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples., *Genome research* **21**, 952-60 (2011).

8. H. Li et al., The Sequence Alignment/Map format and SAMtools., *Bioinformatics (Oxford, England)* **25**, 2078-9 (2009).

9. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores., *Genome research* **18**, 1851-8 (2008).

10. R. Li et al., SNP detection for massively parallel whole-genome resequencing., *Genome research* **19**, 1124-32 (2009).

11. S. Suzuki, N. Ono, C. Furusawa, B.-W. Ying, T. Yomo, Comparison of sequence reads obtained from three next-generation sequencing platforms., *PloS one* **6**, e19534 (2011).

12. O. Harismendy et al., Evaluation of next generation sequencing platforms for population targeted sequencing studies., *Genome biology* **10**, R32 (2009).

13. V. Bansal et al., Accurate detection and genotyping of SNPs utilizing population sequencing data., *Genome research* **20**, 537-45 (2010).

14. J. Shendure, H. Ji, Next-generation DNA sequencing., *Nature biotechnology* **26**, 1135-45 (2008).

15. I. Comino-Méndez et al., Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma., *Nature genetics* **43**, 663-667 (2011).

16. Y. Shi et al., J. Hoh, Ed. Exome Sequencing Identifies ZNF644 Mutations in High Myopia, *PLoS Genetics* **7**, e1002084 (2011).

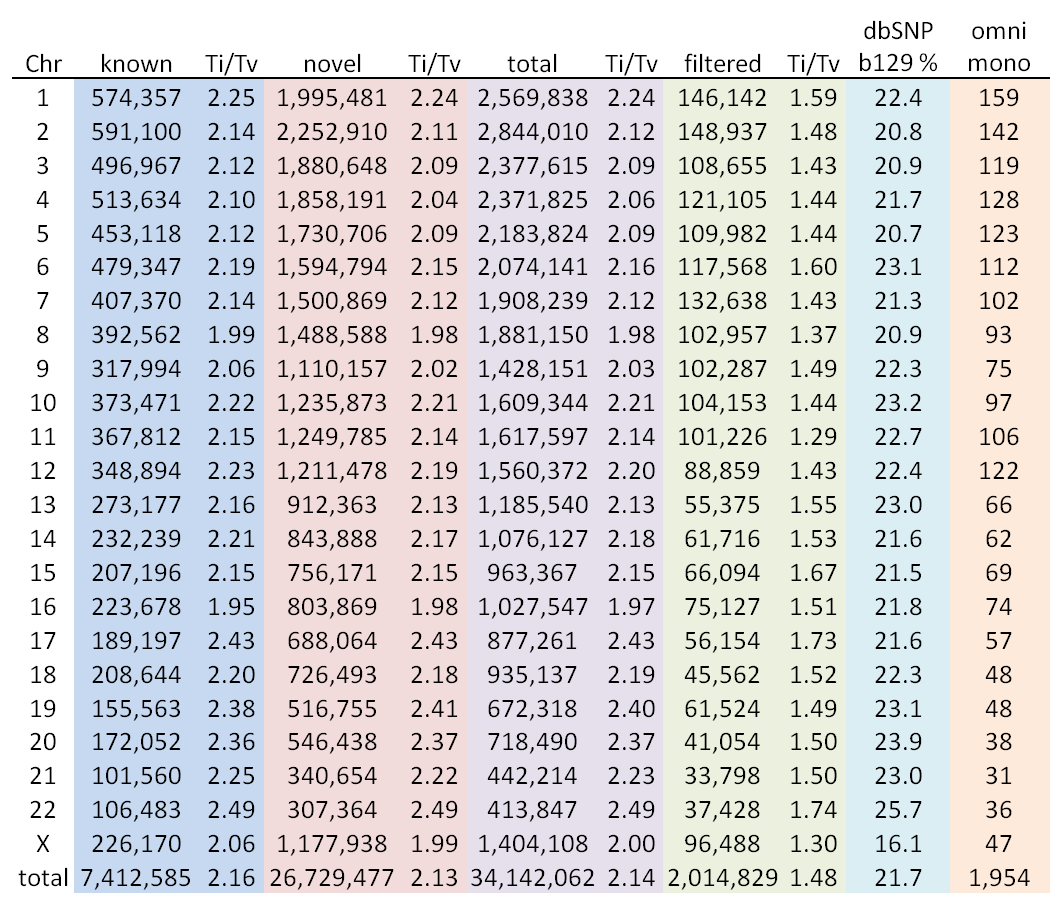
17. J. Marchini et al., A comparison of phasing algorithms for trios and unrelated individuals., *American journal of human genetics* **78**, 437-50 (2006).

**TABLES**

**FIGURE LEGEND**

**Figure 1: SNPTools Pipeline:** The SNPTools pipeline utilizes binary sequence map (BAM) files and then processes them through 5 modular steps: Calculation of effective base depth (EBD), SNP site discovery, BAM-specific binomial mixture modeling to calculate genotype likelihoods, and genotype imputation.

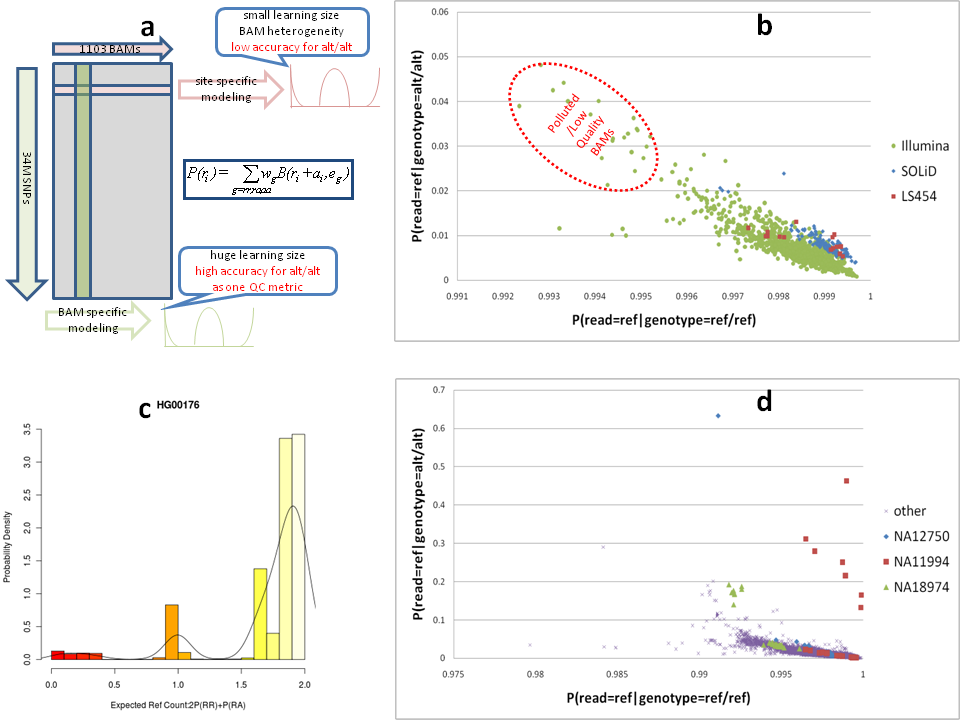
**FIGURES**



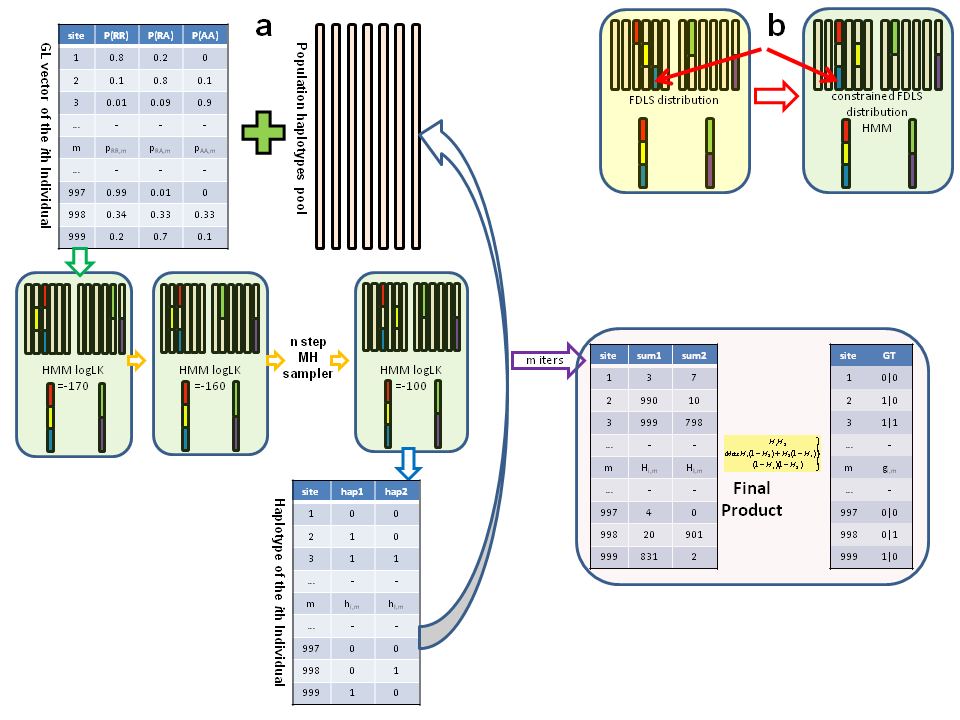
**Figure 3: Bam-Specific Binomial Mixture Modeling:** (a)BAM specific binomial mixture modeling

****

Do we have the Figure (b) and (c) separate?

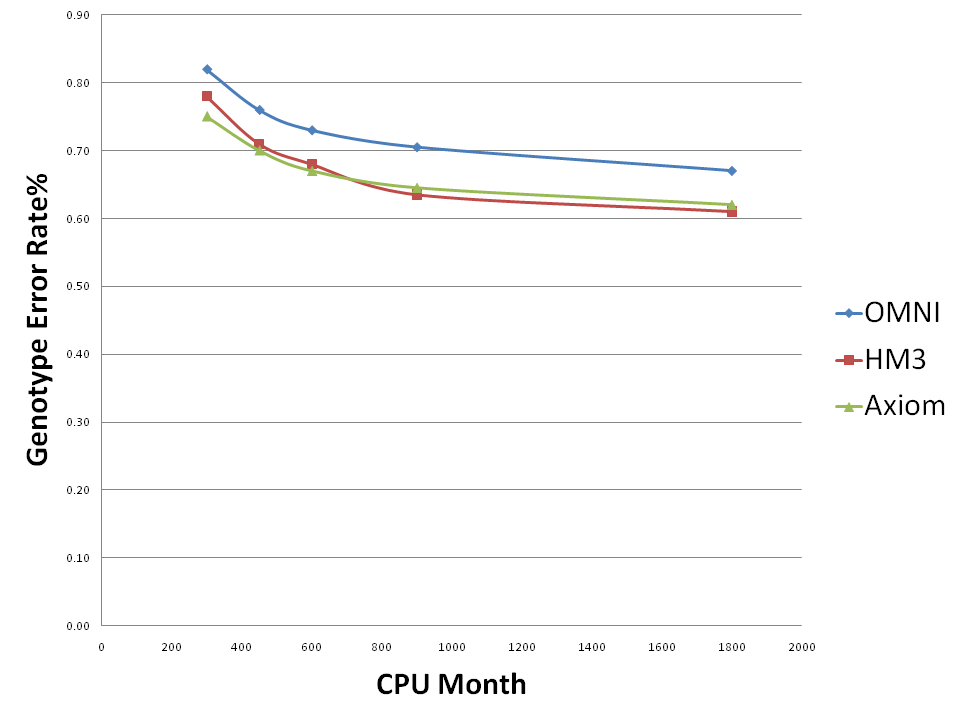


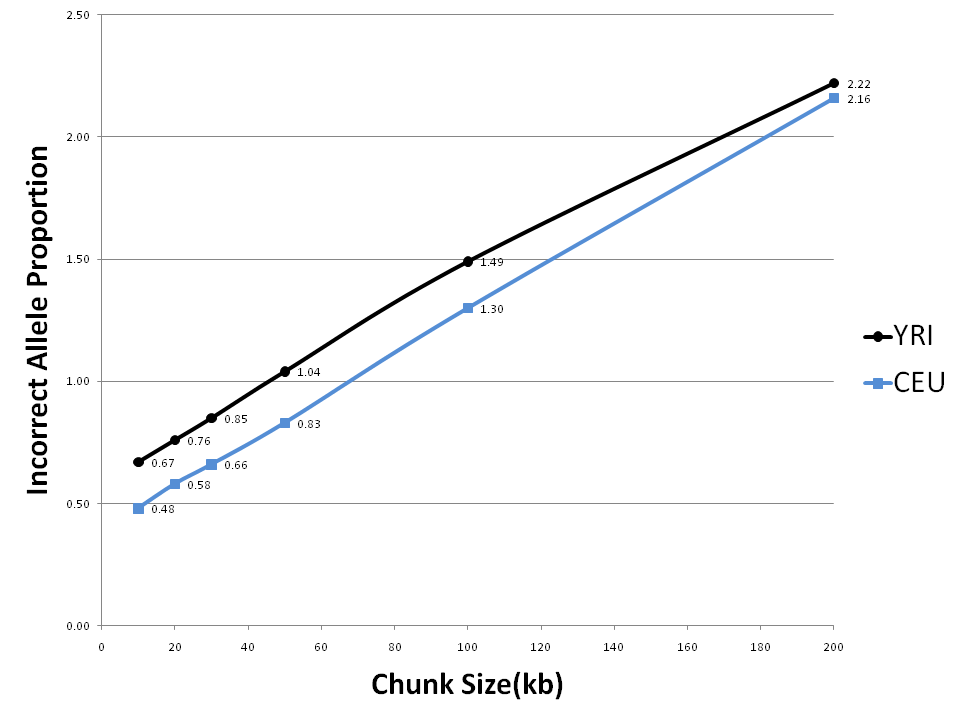
**Figure 4 Genotype Imputation:** Imputation of



**Figure 5 Imputed Genotypes for chr20:** (a) Discordance rates for imputed genotypes on chr20, created using different Chunk Sizes and MCMC, compared to genotypes from three array datasets, HapMap3, Illumina OMNI and Affymetrix Axiom. MCMC of 200 produced the lowest error rate. (b) Computational burden of increased MCMC, showed asymptotic reductions in performance after 1000 CPU Months. (c) Haplotype Phasing, when compared to high coverage CEU and YRI trios showed ~ 1% error for lengths less than 50kb. Error rates increased linearly with haplotype length.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Reference** | **MCMC** | **Chunk Size** | **Ref/Ref** | **Ref/Alt** | **Alt/Alt** | **Total** | **Non-Ref** |
| HapMap3 | 30 | 1024 | 0.36% | 1.14% | 1.44% | 0.78% | 1.67% |
| 50 | 1024 | 0.31% | 1.04% | 1.37% | 0.71% | 1.52% |
| 65 | 1024 | 0.29% | 0.99% | 1.32% | 0.68% | 1.45% |
| 100 | 1024 | 0.26% | 0.94% | 1.26% | 0.64% | 1.36% |
| 100 | 512 | 0.27% | 0.92% | 1.25% | 0.63% | 1.36% |
| 200 | 1024 | 0.25% | 0.89% | 1.21% | 0.61% | 1.30% |
| OMNI | 30 | 1024 | 0.43% | 1.67% | 1.59% | 0.82% | 2.53% |
| 50 | 1024 | 0.40% | 1.52% | 1.53% | 0.76% | 2.34% |
| 65 | 1024 | 0.39% | 1.44% | 1.46% | 0.73% | 2.24% |
| 100 | 1024 | 0.37% | 1.45% | 1.43% | 0.72% | 2.21% |
| 100 | 512 | 0.37% | 1.35% | 1.40% | 0.69% | 2.14% |
| 200 | 1024 | 0.35% | 1.34% | 1.35% | 0.67% | 2.07% |
| Axiom | 30 | 1024 | 0.27% | 1.82% | 1.55% | 0.75% | 2.25% |
| 50 | 1024 | 0.24% | 1.69% | 1.50% | 0.70% | 2.10% |
| 65 | 1024 | 0.23% | 1.62% | 1.45% | 0.67% | 2.01% |
| 100 | 1024 | 0.21% | 1.58% | 1.41% | 0.65% | 1.93% |
| 100 | 512 | 0.22% | 1.53% | 1.40% | 0.64% | 1.91% |
| 200 | 1024 | 0.21% | 1.52% | 1.37% | 0.62% | 1.86% |





**Figure 6: Integration of SNP array genotypes:** (a) Array genotypes are integrated into genotype likelihoods as a high confidence prior for one of the genotypes. This improves both on -aray and off-array genotype accuracy. (b) We imputed genotypes for chromosome 20 by integrating HapMap3, OMNI and Axiom genotypes independently , and then evaluated discordance against the other two arrays. We found that integration non-trivially improves genotype quality.

