Stat Biosci (2013) 5:3–25
DOI 10.1007/s12561-012-9067-4

Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data

Yun Li · Wei Chen · Eric Yi Liu · Yi-Hui Zhou

Received: 30 November 2011 / Accepted: 10 May 2012 / Published online: 29 June 2012

Abstract Massively parallel sequencing (MPS), since its debut in 2005, has transformed the field of genomic studies. These new sequencing technologies have resulted in the successful identification of causal variants for several rare Mendelian disorders. They have also begun to deliver on their promise to explain some of the missing heritability from genome-wide association studies (GWAS) of complex traits. We anticipate a rapidly growing number of MPS-based studies for a diverse range of applications in the near future. One crucial and nearly inevitable step is to detect SNPs and call genotypes at the detected polymorphic sites from the sequencing data. Here, we review statistical methods that have been proposed in the past five years for this purpose. In addition, we discuss emerging issues and future directions related to SNP detection and genotype calling from MPS data.

Keywords Massively parallel sequencing · Next-generation sequencing · SNP detection · Genotype calling · Linkage disequilibrium (LD)

Y. Li (✉)
Department of Genetics, University of North Carolina, Chapel Hill, NC 27599-7264, USA
e-mail: yunli@med.unc.edu

Y. Li · Y.-H. Zhou
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7264, USA

W. Chen
Division of Pediatric Pulmonary Medicine, Allergy and Immunology, Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh School of Medicine, Pittsburgh, PA 15224, USA

Y. Li · E.Y. Liu
Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599-7264, USA

Table 1 Abbreviations

Abbreviation	Description (section)
BAM	Binary SAM, Sequence Alignment/Map format (Sect. 2.1)
GLF	Genotype Likelihood Format (Sects. 2, 3.2.1)
GWAS	Genome-wide Association Studies (abstract, Sect. 1)
HTS	High Throughput Sequencing (Sect. 1)
Kb	Kilobase (Sect. 1)
LD	Linkage Disequilibrium (Sects. 1, 3.3.2)
MPS	Massively Parallel Sequencing
MS-LD*	Multi-Sample Linkage Disequilibrium genotype calling method (Sect. 3.3.2)
MS-SS*	Multi-Sample Single-Site genotype calling method (Sect. 3.3.1)
NGS	Next Generation Sequencing (Sect. 1)
SAM	Sequence Alignment/Map format (Sect. 2.1)
SNP	Single Nucleotide Polymorphism
SS*	Single-Sample genotype calling method (Sect. 3.2)
VCF	Variant Call Format (Sect. 2)

1 Introduction

Since 1977, Sanger capillary sequencing [1] had predominated the field of DNA sequence generation. It was essentially the single viable DNA sequencing technology for almost three decades. After more than two decades of gradual improvement, the costs of Sanger sequencing in the early 2000s were on the order of US $0.5 per kilobase (Kb) [2], taking ∼100 minutes [3] to sequence a Kb. This cost and throughput prohibited its application to large-scale sequencing-based studies. Massively parallel sequencing (MPS; see Table 1 for a list of abbreviation), also known as next-generation sequencing (NGS), and high-throughput sequencing (HTS), marked this debut in 2005 [4]. These new sequencing technologies are able to generate 1 Kb sequence data at the cost of US $0.00005 in ∼0.002 minute. The growth pattern has been more remarkable than that in Moore's Law [5].

Besides ultra-low costs and ultra-high throughput as compared to Sanger sequencing technology, these new technologies have two other hallmarks highly pertinent to our topic of SNP detection and genotype calling: first, relatively short read and second, high per-base sequencing error rate. Compared to Sanger sequencing, which can generate reads up to ∼1 Kb with a per-base error rate <0.001 % [2], MPS technologies generate short reads (typically 30–400 base pairs [bp] in length) with much higher error rate (0.5–1.0 % error per raw base is typical) [4, 6]. Such high error rates entail redundant sequencing of each base to distinguish sequencing errors from true polymorphisms when SNP detection and genotype calling are performed at the level of a single individual.

Commonly used MPS technologies in the market include the Illumina Solexa sequencing-by-synthesis [6], Roche 454 pyrosequencing [4], Applied Biosystem SOLiD [7], Helicos Biosciences [8], Pacific Biosciences [9]. Excellent review papers [2, 3, 10–14] exist covering various aspects of these new sequencing technolo-

gies and it is clear that MPS technologies have transformed the field of genomic studies [10–12]. In particular, in the field of gene mapping for human disease or traits, these technologies have led to successful identification of causal variants for several rare Mendelian disorders [15–21]. They also begin to explain some of the missing heritability from Genome-wide association studies (GWAS) [22–24]. For example, whole genome sequencing of ∼1000 individuals from an isolated population has allowed the rediscovery of a coding variant which is known to affect plasma low-density lipoprotein levels through direct sequencing but was missed using standard GWAS and imputation [25]. We anticipate MPS to play an increasingly important role in genomic studies.

One crucial step for the successful application of MPS is variant detection and genotype calling at detected variant loci. In this review, we will focus on SNP detection and genotype calling at detected SNPs. The remainder of the review will be organized as follows: we will first introduce a typical workflow of SNP detection and genotype calling from sequence data. We will then provide a detailed discussion of methods to detect SNPs and/or perform genotype calling at detected SNPs. In particular, we categorize the methods into two general groups: those that detect SNPs or estimate allele frequencies without individual-level genotype calls, and those that generate individual-level genotype calls. Our focus will be on the latter group, which we further break down into three types: single-sample (SS), multi-sample single-site (MS-SS), and multi-sample linkage disequilibrium (LD) based (MS-LD). We will present representative methods from each category and demonstrate their relative performance using real data from the 1000 Genomes Project. We will then discuss the implication of the newly developed SNP detection and genotype calling methods for the design of sequencing-based association studies, particularly for the study of complex traits. Finally, we will discuss emerging issues and future directions.

2 A Typical Workflow

Figure 1 depicts a typical pipeline for SNP detection and genotype calling from MPS data. In this diagram, we start with sequence read data in fastq format files (details below in Sect. 2.1). The fastq files are generated by base-calling methods from a series of images directly from sequencing machines. An excellent review paper [26] and several methods papers [27–30] have been published on base-calling from image data. Our starting data, the fastq format files, contain the sequence of nucleotides and their corresponding per-base quality scores, which are typically not very well calibrated (see Sect. 2.3 for more). At this stage, we have millions or even billions of short reads from unknown genomic positions. We determine their genomic positions through read alignment (or, interchangeably, called read mapping) where we map the short read against the entire reference genome or reference transcriptome (depending on the application) to find the genomic location of each short read. Short reads are mapped to their most likely genomic positions with varying levels of uncertainty by alignment algorithms (details in Sect. 2.2). These algorithms provide the most likely mapped position along with a mapping quality score for each mapped read, which are together stored in Sequence Alignment/Map (SAM) or BAM (binary SAM) format

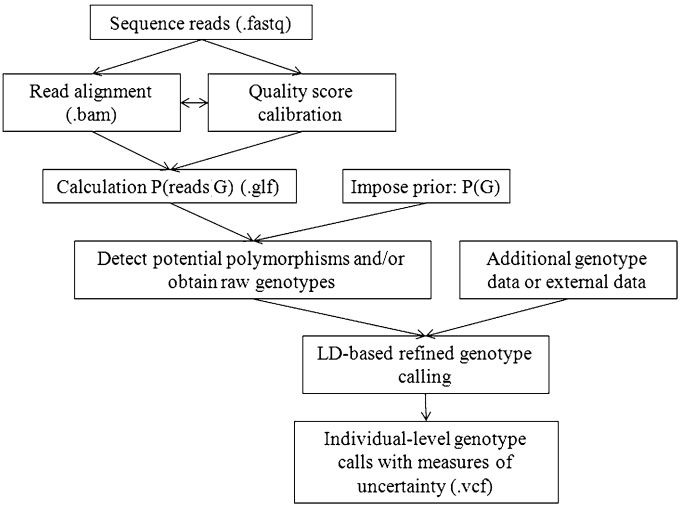

Fig. 1 A typical workflow for SNP detection and genotype calling. We show a typical workflow for SNP detection and genotype calling from massively parallel sequencing data, starting from unmapped reads (in fastq format)

files [31]. Along with or after read alignment, per-base quality score recalibration is performed (details in Sect. 2.3).

Given data consisting of mapped reads, per-base quality scores, and read-level mapping quality scores, we can calculate the probabilities of the data conditional on any of the possible true genotypes for each diploid individual at each base. These probabilities are stored in Genotype Likelihood Format (GLF) files.[1] Together with a prior on the distribution for the possible true genotypes, one can obtain genotype calls by applying Bayes' rule, which forms the basis of most non-LD-based genotyping methods. LD-based methods take an additional step to refine genotype calls by borrowing information from other individuals who carry similar haplotypes, where a haplotype is a specific allele combination across SNPs. The final calls typically consist of the most likely genotype call for each individual at each polymorphic locus, along with measures of uncertainty, typically stored in Variant Call Format (VCF) files (for format details, refer to http://www.1000genomes.org/node/101).

We will now provide detailed explanations for every aforementioned step in the text to follow.

2.1 Sequence Data: fastq Format Files

Before introducing methods to analyze sequence data, we view it helpful to show what MPS data look like. As mentioned above, the raw sequence data are actually a series of images, from which base-calling methods infer the sequence of nucleotides and their corresponding per-base quality scores for each short read. The sequence

[1] For format details, refer to ftp://share.sph.umich.edu/1000genomes/pilot1/GLF1.pdf, an excerpt from an early version of http://samtools.sourceforge.net/SAM1.pdf.

```
@ERR009169.17725968 IL18_2954:8:100:1790:1881/2
CTAAAAATACAAAAAAAAAAAAAGAAAAAAAATGCTGAGCATCGTGGCGGATGGCTGTAAACCCAGCTACTCGGGA
+
@BBBBBBBBABCBCBBBBBCBB@=@BBBBBB@<@7?=?;15)9=/@@6AB6*6%(%5&2=%*,'2))-2.12?A:(
@ERR009169.17725969 IL18_2954:8:100:1790:1768/2
ACTACCTATGAAGTGGGAACATTTTAAAGGCAAGAAATCAGAGCTCAGAAGTCAAGTAACTTACTCAAGATCACAC
+
BBBBABCBBCBBBBBBCBBBBBBBBBBBBB@BBBBBBBABBAC@B>BCB@B>>AABAAB@BAB@B>B@B@@=BABA
@ERR009169.17725970 IL18_2954:8:100:1790:480/2
ACAAAATACAGCCAATTCTTGCTATTTGCAGTAGTGAGGTTCTAGAAAGTCACCGTGAACGCTGAGCTGCCACTCC
+
??@@?=@@???>????@???@?@????<;>?=???9?>>?8=????>>??=>><=?7?:?=?====>=4==6=<===
@ERR009169.17725972 IL18_2954:8:100:1790:1563/2
CGGTAACTGCTATGTGTAAAGGCTTAGGGCACTTTACACCTGTCAGACTGACAAATCAGACAGTGGAATCATGCAA
+
=>>??@?@@??@??>@????@@>??????>?<????>?>???>???>=???7=>???>?===>7?>==?4>=?=6&
@ERR009169.17725973 IL18_2954:8:100:1790:1246/2
TTCCTTTGAGTAAGATATGGGATGTTATTAATTGATTAATCTCCCTCCCTATTCTTAAAAATGATTTAAGGAGGGT
+
BA@ABB8%4A=A?A<@?BAA?=A@;A=BA:3@A??AA=;?=>?>6==6?=9;1'@=2+282=>3?8997=44=778
```

Fig. 2 Example fastq file. We show records from a standard format for unmapped reads: fastq format file

of nucleotides and per-base quality scores are typically stored in fastq files (for format details, see http://en.wikipedia.org/wiki/FASTQ_format). Figure 2 shows a few records from a fastq file for a CEU (Utah residents [CEPH] with Northern and Western European ancestry) individual (ID for this individual is NA12878) sequenced by the 1000 Genomes Project [32]. The fastq files are available at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NA12878/sequence_read/ERR009169.filt.fastq.gz.

Information in Fig. 2 is for five short reads, with four lines constituting one read. We will take the first read as an example. The first line contains read identifier information. The unique ID for this particular read is ERR009169.17725968. The ID line always starts with the @ character and may contain additional information. The next line contains the actual sequence of nucleotides called and is a string made up of four possible characters, A, C, G, and T, for the four possible nucleotides, Adenine, Cytosine, Guanine, and Thymine, respectively. Sometimes, a fifth character, N, is also included for "no call." We can also see from this nucleotide line that this particular short read is of length 76 base pairs. The next line has a single character "+" and sometimes copies the read ID after the "+" character. The last line contains the per-base quality score in ASCII characters. From these ASCII characters, we can obtain per-base phred quality scores [33, 34], denoted by Q:

$$Q = -10 \times \log_{10}(e), \quad \text{where } e \text{ is the per-base sequencing error.} \tag{1}$$

Given the above definition, a phred score of 10 corresponds to one error every 10 bases (or sequencing error rate of 0.1); 20 every 100 bases (or sequencing error rate of 0.01); and 30 every 1000 bases (or sequencing error rate of 0.001). In the example fastq, the formula to calculate phred score from the ASCII characters is:

$$Q = \text{ASCII} - 33. \tag{2}$$

Note: The conversion formula may vary with the source of the fastq file. For example, some newer versions use $Q = \text{ASCII} - 64$.

Here, the first base in the first read has an ASCII character '@' corresponding to a numeric value of 64. Using the above formula, we get the phred score of 31, which

indicates an estimated sequencing error of 0.00079. Similarly we can calculate the phred scores for the remaining 75 bases in the read.

2.2 Read Alignment/Mapping

The next crucial step in the analysis of MPS data is read alignment. A large number of methods have been developed in the past five years for efficiently mapping short reads to a reference sequence. An incomplete list of commonly used methods includes MAQ [35], BWA [36, 37], stampy [38], SOAP2 [39], novalign (www.novocraft.com), BFAST [40], SSAHA [41] most commonly used for DNA sequencing data; and BOWTIE [42], TOPHAT [43], MapSplice [44], GSNAP [45], and RUM [46] most commonly used for RNA/transcriptome sequencing data. For a more complete list of methods and software available, see earlier review articles [47–49] and the following wiki page: http://en.wikipedia.org/wiki/List_of_sequence_alignment_software.

2.3 Quality Score Recalibration

As previously mentioned, the per-base quality scores estimated by base-calling methods are typically not well calibrated. For example, when the called nucleotides are compared with experimental genotypes with comparison restricted at homozygous genotypes (so that any nucleotide other than the allele underlying the homozygous genotype can be viewed as a sequencing error), the discordance/error rates typically do not agree with what is implicated by the per-base quality scores. Since these per-base quality scores play an important role in SNP detection and genotype calling (see, for example, Sect. 3.2.1), it is essential to perform quality score recalibration analysis. One typical procedure as implemented in GATK [50] flows as follows: first we bin the data according to factors that affect calibration precision. The factors include read cycle (or position along the read), raw per-base quality score, genomic context (nucleotides before and after the investigated base). Other factors, particularly those that are specific to a certain MPS technology, have been reported previously [51–53] and can also be useful for quality score recalibration [54]. After binning, we calculate the mismatch rate within each bin, at homozygous genotypes when external genotypes are available (for example, all individuals sequenced by the 1000 Genomes Pilot Project [32] had been genotyped previously by the International HapMap Projects [55, 56]), or at non-dbSNP [57] sites under the rationale that almost all individuals are homozygous for the reference allele at these sites. Finally, we reset the per-base quality scores accordingly to Eqs. (1) and (2) in Sect. 2.1, where e in Eq. (1) is set to be the mismatch rate calculated. The three above steps are iterated until the final per-base quality scores stabilize.

Theoretically, the recalibration procedure should be iterated with read alignment because per-base quality scores and aligned positions affect each other. For example, if several bases in a read have much lower recalibrated per-base quality scores, the read may match better to other genomic positions. Conversely, when reads are mapped to different places in the genome, the configuration of each bin changes accordingly, which in turn leads to differently calibrated per-base quality scores. In practice, read alignment is typically not repeated. This is partly because reads most

susceptible to changes in per-base quality scores tend to be poorly mapped in the first place, thus the information from these reads will be downweighted in subsequent analysis. The time and resources required for read alignment also pose a challenge to iteration of recalibration and alignment.

3 Methods for SNP Detection and Genotype Calling

We use "SNP detection" to refer to the inference regarding which base has a variant allele, that is, an allele other than the reference. We use "genotype calling" to refer to the estimation of genotypes for each individual at detected SNP loci. In this section, we will first briefly discuss selected methods that detect SNPs or estimate allele frequencies but do not estimate individual genotypes (Sect. 3.1). We will then focus on methods that detect SNPs as well as generate individual-level genotype calls, breaking the methods into three types: single-sample genotype calling (Sect. 3.2), multi-sample single-site genotype calling (Sect. 3.3.1) and multi-sample LD-based genotype calling (Sect. 3.3.2). Note that we use sample to refer to a diploid individual throughout the review. Hereafter, we will use sample, individual, diploid individual interchangeably. This review, ignoring the literature for SNP detection from Sanger capillary sequencing data, for example, methods underlying PolyBayes [58], PolyPhred [59, 60], and PolyScan [61], focuses on methods developed for MPS data.

3.1 SNP Detection or Allele Frequency Estimation Methods

Brockman et al. [51] and VarScan [62] detect SNPs using largely heuristic approaches. VarScan, for example, takes specific features of different sequencing platforms (Roche 454 and Illumina Solexa considered) and different read alignment methods (compatible with five methods: BLAT [63], Newbler (Roche), cross_match, BOWTIE [42], and novalign) into consideration. SNP detection is achieved by applying a series of filters according to thresholds on total read depth (total number of reads covering the base investigated), strand-specific depths (number of reads in forward and reverse strand separately), per-base quality scores, and number of reads carrying the minor allele.

ProbHD proposed by Hoberman and colleagues [64] used a machine learning approach that considers multiple features to generate a heterozygosity score for each base. Their method, designed specifically for Roche 454 data, considers a large number of features including total read depth, strand-specific depths, read cycle (within-read relative position), per-base quality scores, read alignment quality, and homopolymer length. They used the random forest method [65] which builds multiple decision trees using these features to classify whether a base is heterozygous. The proportion of trees that classify a base as heterozygous is used to construct a heterozygosity score for each diploid individual. Evidence can be accumulated across individuals to improve detection sensitivity at controlled false call rate.

Atlas-SNP2 [54] detects SNPs in two steps. In the first step, it recalibrates per-base quality scores for every base carrying the non-reference allele using a logistic regression on a training data set. In its real data example, the training data set is an independent pre-existing data set generated by the same Roche 454 Titanium technology

and by the same base-calling method as in the data set under study. Predictors considered in the logic regression include raw per-base quality score, neighboring quality standard [66], GC content, read cycle, genomic context (flanking nucleotides), and features specific to the Roche 454 platform (e.g., homopolymer length mentioned above). In the second step, Atlas-SNP2 accumulates information across all reads carrying the non-reference allele using the recalibrated per-base quality scores, and adopts a Bayesian approach to include read depth and prior knowledge of the overall sequencing error rates into the modeling framework.

Allele frequency estimation has many important applications for disease mapping [67, 68] and in the field of population genetics [69, 70]. Although the SNP detection methods discussed above can either estimate allele frequencies or can easily extend to do so, there are methods that were developed more specifically to fulfill this important task [71–76]. For example, Kim et al. used likelihood-based methods to estimate allele frequencies under three different scenarios: when genotypes are already called from MPS data; when genotypes are not called but the minor allele is obvious; and when genotypes are not called and the minor allele is not obvious. The distinction between the second and third scenarios lies mostly between common and rare variants. For common SNPs, the minor allele frequency (MAF) is high enough, such that the second most frequently occurring allele can be easily identified from the three non-reference alleles. However, for rare SNPs, all three non-reference alleles may appear similar number of times due to the confounding from sequencing errors.

3.2 Single-Sample (SS) Genotype Calling

3.2.1 Genotype Likelihood Calculation

As introduced in Sect. 2, the typical step after read alignment and quality score recalibration is to calculate likelihood of the observed sequence data given possible true genotypes at each base and for every diploid individual. Although, one could determine *the* alternative allele (assuming SNPs are bi-allelic and thus there is only one alternative allele) first and calculate likelihood given three possible true genotypes, namely homozygote for the reference allele; heterozygote, or homozygote for the alternative allele. Most methods calculate all ten possible true genotypes at every base pair, as implemented in SAMtools [31]. The calculation involves three pieces of information: the called nucleotides at each base for each read, per-base quality scores (better if calibrated), and read-level mapping quality scores.

We will start with a simple scenario where we observe only two alleles at a particular base from the sequencing data of a particular diploid individual. Denote the two alleles by A and B where each takes one of the four possible values {A, C, G, T} corresponding to the four possible nucleotides. The three possible true genotypes therefore are A/A, A/B, and B/B. Further denote N_A the random variable for the number of reads carrying allele A, and n_A and n_B the observed number of reads carrying alleles A and B, respectively. If we assume a uniform per-base sequencing

error rate of δ and further assume that the probability of misreading allele A as allele B is the same as the probability of misreading allele B as allele A, we have:

$$N_A \sim \begin{cases} \text{Binomial}(n_A + n_B, 1-\delta) & G = A/A \\ \text{Binomial}(n_A + n_B, 0.5) & G = A/B \\ \text{Binomial}(n_A + n_B, \delta) & G = B/B \end{cases} \tag{3}$$

In practice, such simple binomial distribution approximations do not perform well for several reasons. First, the per-base sequencing error rates are not uniform (we have per-base quality scores which are estimates for the base-specific sequencing error rate). Second, they do not take into account mapping quality information at the read level. A base called with high confidence still should not be trusted if the read it belongs to is mapped to the current position with low confidence. Third, sequencing errors tend to be correlated instead of independent. To solve the second issue, Li and Durbin [35] proposed capping the per-base quality scores by the mapping score of their residing read. To model the base-specific error rates and dependency among sequencing errors, Li et al. borrow ideas from Huang and Madan [77]. In particular, the overall error probability of observing n_A reads carrying allele A and n_B reads carrying allele B given the true genotype being B/B, denoted by $\text{ERROR}_{n_A,(n_A+n_B)}$ will change from Eq. (4) [according to Eq. (3) above] to Eq. (5):

$$\text{ERROR}_{n_A,(n_A+n_B)} = \binom{n_A + n_B}{n_B} \times \delta^{n_A} \times (1-\delta)^{n_B} \tag{4}$$

$$\text{ERROR}_{n_A,(n_A+n_B)} = C_{n_A,(n_A+n_B)} \prod_{i=0}^{n_A-1} \delta_{(i+1)}^{\theta^i} \tag{5}$$

where $\delta_{(i+1)}$ indicates the $(i+1)$th lowest base error rate and $C_{n_A,(n_A+n_B)}$ is a function of the per-base error rate estimates δ_i's but varies little with these δ_i's (details in Li and Durbin [35] Supplementaries 3.1 and 3.2). θ by default is set at 0.85, which the authors found a reasonable value for Illumina Solexa data. The particular form in Eq. (5) effectively downweights information from bases with lower quality scores in a gradually more aggressive fashion. Suppose $n_A = 3$ and that the sorted corresponding per-base error rates are 0.0001, 0.001, and 0.01, respectively. In particular, the product term in (5) will be $(0.0001)^{\theta^0} \times (0.001)^{\theta^1} \times (0.01)^{\theta^2}$. With the default value $\theta = 0.85$, it becomes $(0.0001)^1 \times (0.001)^{0.85} \times (0.01)^{(0.85)^2} = (0.0001) \times (0.002818) \times (0.03589)$.

Although the presentation above assumed only two alleles, the formulae directly apply to all four nucleotides because the formulae only depend on the count and quality scores (again, including per-base and the mapping quality scores) of "error" bases. Once conditional on the true genotype, it is obvious which bases are sequencing errors. For example, if the true genotype is A/C at a locus for a diploid individual, any read carrying G or T allele at that locus manifests a sequencing error.

3.2.2 Genotype Prior and Calling via Bayes' Rule

Once the ten genotype likelihoods are calculated, that is, once we have $\Pr(\text{Data} \mid G)$ where G is the true genotype, inferring genotype becomes trivial. We will only need

a prior on the true genotypes $\Pr(G)$. With these two, namely likelihood $\Pr(\text{Data} \mid G)$ and prior $\Pr(G)$, we can easily call genotypes via Bayes' rule. In particular, the posterior probability of the true genotype $\Pr(G \mid \text{Data})$ can be expressed as follows:

$$\Pr(G \mid \text{Data}) \propto \Pr(G) \times \Pr(\text{Data} \mid G)$$

The genotype with the highest posterior probability is then the most likely genotype call and measures of calling uncertainty can be easily derived. Such a Bayesian framework underlies common single-sample genotyping methods though many of the different methods use different priors. For example, MAQ [35] uses priors in which the two possible homozygous genotypes (with only the two alleles with largest number of read support retained) have equal prior probability and the heterozygote has a prior probability r. The MAQ authors set $r = 0.001$ to discover new SNPs, and $r = 0.2$ for known SNPs. At known SNP loci, more informative priors leveraging the allele frequency information can help genotype calling when coverage is low ($<$5X) or medium ($<$10–15X). Priors can also be made more informative by distinguishing homozygous genotype for the reference allele from homozygous genotype for the alternative, and distinguishing transition (A $\leftrightarrow$ G, C $\leftrightarrow$ T) mutations from transversion (A/G $\leftrightarrow$ C/T) mutations as considered in SOAP-SNP [78].

3.3 Multi-Sample (MS) Genotype Calling

In the previous section, we have laid out the common statistical framework for inferring genotypes for one diploid individual: calculate genotype likelihood, impose a prior on true genotypes, and then estimate posterior probabilities via Bayes' rule. These single-sample methods rely on redundant sequencing of each base to distinguish sequencing errors from true polymorphisms [6, 79]. For example, 30X read depth (where each base is covered by an average of 30 reads) typically results in $>$99 % genotyping accuracy [6]. These methods perform well with high depth data but have unacceptable performance when applied to single individuals with low depth data. For example, Li et al. [78] reported a per-base false positive rate (FPR) of 0.04 % for a single individual sequenced at 4X, implying a cumulative per-base FPR of $1 - (1 - 0.04\,\%)^{100} = 4\,\%$ when applied to 100 independent individuals. This corresponds to one false positive per 25 bases, and implies that $\sim$90 % of the SNPs called are false positives assuming one true SNP per 200–300 bases. In addition, at depth 4X, the probability that both alleles at a locus are covered at least once is $\sim$75 % (assuming the number of times each allele is covered follows a Poisson distribution with mean 2), implying that $>$25 % of heterozygotes cannot possibly be inferred properly.

In an attempt to mitigate these issues and to improve the per-depth information obtainable, a number of multi-sample methods have been proposed in the last two years that generate high quality genotypes for medium coverage data (10–20X per individual), and even for low coverage data (down to 2–4X per individual). We classify these methods into two categories: multi-sample single-site where information is integrated across individuals but at each site separately; and multi-sample LD-based where information is borrowed both across individuals and from flanking sites.

3.3.1 Single-Site Inference

There are at least two places where information across individuals can facilitate inference. First, at the per-base quality score recalibration step, information from other individuals, particularly when sequenced together, can be used to form the bins introduced previously in Sect. 2.3. This leads to bins with a larger number of observations, thus better at avoiding sparse bins and eventually leading to more reliable recalibration. SNIP-Seq [80], for example, using information across individuals, was able to partition their sequencing reads into as many as $36 \times 2 \times 3$ bins, according to read cycle (their read length is 36 bases), strand (forward and reverse), and raw per-base quality (0–9, 10–19, and 20–30). These very fine bins allowed more accurate calibration of the per-base quality scores, which improves SNPs detection and genotype calling accuracy.

Secondly, information across individuals can be used to form more informative prior on true genotypes. We mentioned in Sect. 3.2.2 that allele frequency of known SNPs can be used to form informative prior. The allele frequency can either come from previous data, or be estimated using multiple samples sequenced under the current study. These allele frequency estimates together with Hardy–Weinberg equilibrium [81, 82] can be used to specify a prior on the probabilities of the true genotypes, as in the SeqEM [83] framework. SeqEM adopts an empirical Bayesian approach. It uses sequence data consisting of multiple samples to estimate prior parameters including sequencing error rate and allele frequency.

3.3.2 LD-Based Inference

Integrating information across individuals at the single site level can improve SNP detection and genotype calling accuracy such that inference on medium coverage (10–20X) data is possible [80]. To further improve the per-coverage information gains, multi-sample LD-based methods have been proposed. There are currently three published methods that fall into this category: MARGARITA [84] + QCALL [85], GATK [50] + BEAGLE [86], and glfMultiples + *thunder* [87].

The MARGARITA + QCALL method was developed by Le and Durbin at the Wellcome Trust Sanger Institute. The method first performs non-LD-based analysis to detect potential polymorphisms. The non-LD-based analysis integrates information across individuals to estimate the probability of being polymorphic at each base. Bases that are inferred with an SNP probability exceeding a prespecified threshold (in this case, 90 %) are carried on to their LD-based analysis. In the LD-based analysis, genealogy inference is first performed using existing genotype data for the individuals currently sequenced. The inferred genealogy is in the form of ancestry recombination graph, which is a coalescent tree describing how chromosomes or haplotypes from a population-based sample are related to each other, through recombination, mutation, and coalescence, back to a common ancestor. These coalescent trees, defining haplotype sharing patterns among individuals sequenced, can be used to make accurate genotype calls as long as the alleles defining a local haplotype or a section of a local haplotype can be determined by one of the individuals carrying it. The accuracy of the genealogy inference thus directly affects the final genotype calling accuracy. The

authors recommend using phased haplotypes for more accurate genealogy inference with MARGARITA.

The GATK + BEAGLE pipeline also starts with potential polymorphism generation. The candidate SNP generation is fulfilled using an E-M algorithm [50, 88] where allele frequency at each base is estimated based on information across all sequenced individuals. Again, bases with high probability of being polymorphic are carried on to LD-based analysis, using an imputation method implemented in software BEAGLE. BEAGLE [86, 89] uses a variable length Markov model to describe local LD structure and is able to generate genotype calls even at bases with low coverage for a particularly individual, by borrowing information from other individuals carrying similar haplotypes in local regions but having reasonable coverage at the investigated bases.

Similarly, glfMultiples + *thunder* [87] first promote a set of candidate polymorphisms using Bayesian framework. Starting with genotype likelihood $\Pr(\text{Data} \mid G)$ where G is again the true genotype and taking ten possible values, glfMultiples infers the following posterior probability of being polymorphic at each base:

$$\Pr(M = 1_{\{A,B\}} \mid \text{Data}) \propto \Pr(M = 1_{\{A,B\}}) \times \Pr(\text{Data} \mid M = 1_{\{A,B\}})$$

where $M = 1_{\{A,B\}}$ if the base is polymorphic for alleles A and B. The posterior probability is proportional to the product of the likelihood $\Pr(\text{Data} \mid M = 1_{\{A,B\}})$ and the prior on M. The following prior on M is used, according to population genetics principles [90] and knowledge on mutation type relative to the reference allele, specifically that transitions are twice as likely as transversions [91, 92]:

$$\Pr(M = 1) = \theta \times \sum_{i=1}^{2n} 1/i$$

$$\Pr(M = 1_{\{A,B\}}) = c \times \Pr(M = 1) \times \mu$$

where θ is the pairwise nucleotide difference, estimated to be 0.001 [93, 94], n is the number of diploid individuals sequenced, c is a normalizing constant chosen such that probabilities for all the configurations sum to one, and μ is a constant set to be $2/3$ if A is the reference allele and B is the transition mutation; to be $1/6$ if A is the reference allele and B is the transition mutation; and to be $1/1000$ otherwise.

To infer the desired posterior probability $M = 1_{\{A,B\}}$, glfMultiples first maximizes the following likelihood as a function of p_A, the frequency for allele A:

$$\begin{aligned} L(P_A) &= \prod_{i=1}^{n} \Pr(\text{Data}_i \mid M = 1_{\{A,B\}}) \\ &= \prod_{i=1}^{n} \left\{ \sum_{g} \left[\Pr(G_i = g \mid M = 1_{\{A,B\}}) \times \Pr(\text{Data}_i \mid G_i = g) \right] \right\} \end{aligned}$$

where $\Pr(g \mid M = 1_{\{A,B\}}) = (p_A)^2$ if $g = A/A$; $(1 - p_A)^2$ if $g = B/B$; $2p_A(1 - p_A)$ if $g = A/B$; and 0 otherwise.

Again, bases with posterior polymorphic probability exceeding a prespecified threshold are carried into a hidden Markov model-based method that takes LD into

account [95]. The LD-based calling method in both BEAGLE and *thunder* adopts essentially the same statistical framework as used for genotype imputation, which makes inference on missing genotypes by borrowing information from other individuals carrying similar haplotypes. To read more about genotype imputation, see review articles [96, 97].

All the three MS-LD methods discussed above share the same two major components: candidate SNP generation using information across individuals, at each base separately; and LD-based genotype calling at candidate sites. All were used to generate genotype calls with similar accuracy for the 1000 Genomes Pilot Project, where individuals were sequenced at an average coverage of ∼4X. Combining results of the three methods into a consensus call sets improved calling accuracy. For example, average genotype concordance, when compared with experimental genotypes from the International HapMap Projects, improved to 98.69 % from 97.56–98.01 % by a single method [32]. This observation suggests that each individual method can be further improved. For example, the analysis of sequencing data generated by the 1000 Genomes Main Project has suggested the merit of using BEAGLE-inferred haplotypes as starting point for the hidden Markov model in *thunder* (personal communications with Drs. Gonçalo Abecasis and Hyunmin Kang). Please see the following wiki page for more information: http://genome.sph.umich.edu/wiki/UMAKE. For another example, Yu and colleagues at the Baylor College of Medicine developed methods that also show promising results in the analysis of data generated by the 1000 Genomes Project. Their methods are implemented in SNPTools, which is available at http://www.hgsc.bcm.tmc.edu/cascade-tech-software_snp_tools-ti.hgsc.

A more complete list of available software is summarized in Table 2.

3.4 High Level Comparison of the Genotype Calling Methods

We compared the relative performance of the three classes of genotype calling methods on CEU individuals sequenced by the 1000 Genomes Project. There are two individuals, NA12891 and NA12892, who were sequenced at high coverage (average depth ∼43X), and 70 other individuals sequenced at low depth (average depth ∼4X). We used samtools mpileup, GATK UnifiedGenotyper, and glfMultiples on each of the 72 individuals separately (therefore SS methods), samtools mpileup, GATK UnifiedGenotyper, and glfMultiples on all 72 individuals together (therefore MS-SS methods), and glfMultiples + thunder on all 72 individuals together (therefore a MS-LD method). For the two high coverage individuals, we performed calling on randomly selected $10, 20, \ldots, 90$ % of the sequencing reads to compare relative performance of the three methods at different read depths. For each of the 72 individuals, we counted the number of true heterozygote sites (defined based on HapMap experimental genotypes) called by each of the seven methods and compared with the corresponding HapMap experimental genotypes. We applied all seven methods to chromosome4: 57–62 Mb, a region with moderate level of LD, as measured by physical distance of half-life r^2 [98, 99]. In this region, there are 690 and 772 true heterozygous sites for NA12891, NA12892, and on average 647 (standard deviation: 88, range: 376–841) per person for the 70 low coverage individuals.

Figure 3 shows the results for the two high coverage individuals. All methods achieve very low genotype discordance rate when the coverage is high. For example,

Table 2 Available software for SNP detection or genotype calling from MPS data

Method	Website	Method category	Reference
VarScan	http://varscan.sourceforge.net/	SNP detection	[62]
ProbHD	http://seqanswers.com/wiki/ProbHD	SNP detection	[64]
Atlas-SNP2	http://www.hgsc.bcm.tmc.edu/cascade-tech-software_atlas_snp-ti.hgsc	SNP detection	[54]
NA	ftp://ftp.sanger.ac.uk/pub/pathogens/pools/	allele frequency estimation	[72]
NA	http://www.genetics.org/content/suppl/2009/03/16/genetics.109.100479.DC1	allele frequency estimation	[73]
MapNext	http://evolution.sysu.edu.cn/english/software/mapnext.htm	allele frequency estimation	[74]
testassoc	http://www.biomedcentral.com/1471-2105/12/231/additional	allele frequency estimation	[75]
SNVer	http://snver.sourceforge.net/	SNP detection and allele frequency estimation	[76]
samtools	http://samtools.sourceforge.net/	genotype calling (both SS & MS-SS)	[31]
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	genotype calling (both SS & MS-SS)	[50]
SOAP-SNP	http://soap.genomics.org.cn/soapsnp.html	genotype calling (both SS & MS-SS)	[78]
glfMultiples	http://genome.sph.umich.edu/wiki/GlfMultiples	genotype calling (both SS & MS-SS)	[87]
SNIP-Seq	http://polymorphism.scripps.edu/~vbansal/software/SNIP-Seq/	MS-SS genotype calling	[80]
SeqEM	http://hihg.med.miami.edu/software-download/seqem-version-1.0	MS-SS genotype calling	[83]
MARGARITA + QCALL	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	MS-LD genotype calling	[84, 85]
GATK + BEAGLE	http://faculty.washington.edu/browning/beagle/beagle.html	MS-LD genotype calling	[50, 86]
thunder	http://genome.sph.umich.edu/wiki/Thunder	MS-LD genotype calling	[87]
SNPTools	http://www.hgsc.bcm.tmc.edu/cascade-tech-software_snp_tools-ti.hgsc	MS-LD genotype calling	NA

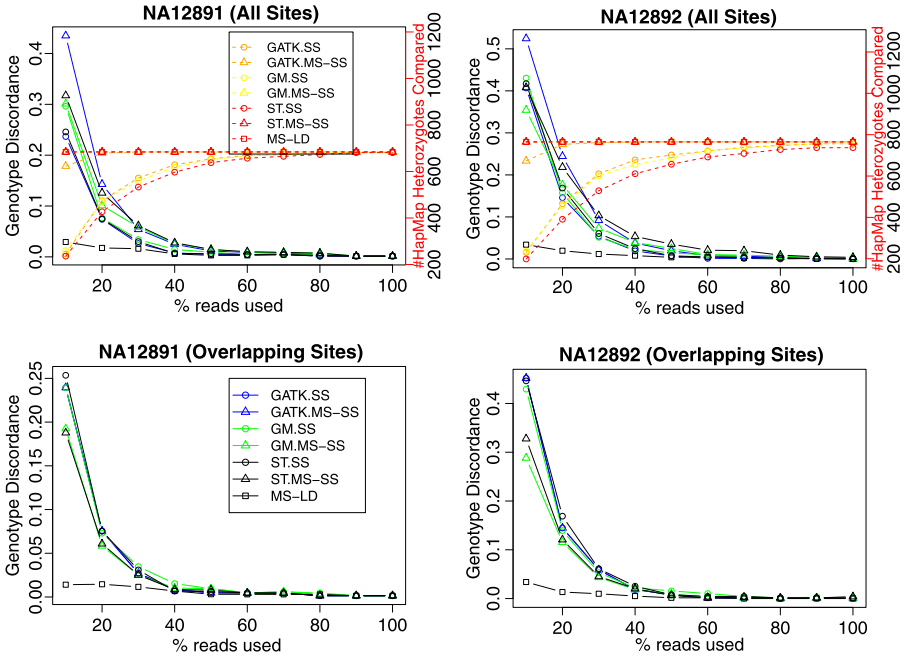

Fig. 3 Comparison of methods on high coverage data from the 1000 genomes pilot project. Three classes of methods, namely SS, MS-SS, and MS-LD, are compared in terms of both number of heterozygotes detected and genotype concordance with experimental genotypes (from the International HapMap project) at detected sites, for NA12891 and NA12892 who were sequenced to a high coverage (average depth ∼40X) in the 1000 Genomes Pilot Project. The right Y-axis shows the number of sites where the method generates a genotype call and where the experimental genotype is heterozygous. Warm color (*red*, *yellow* and *orange*) *dotted lines* and *points* use this axis. The left Y-axis shows the genotype discordance rate at the compared heterozygotes. Cool color (*blue*, *green* and *black*) *solid lines* and *points* use this axis. For both SS and MS-SS, three methods are used: GATK (GATK UnifiedGenotyper), GM (glfMultiples), and ST (samtools). For clarity, the right-Y axis legend is only shown in NA12891 (all sites) and the left-Y axis legend only shown in NA12891 (overlapping sites). (Color figure online)

genotype discordance rate is <0.5 % for all seven methods attempted when 90 % of the sequencing data are used for genotype calling. MS-LD method manifests its advantages when average read depth is moderate or low (<20X coverage when <50 % of reads are used). For example, when 10 % of the reads for NA12892 are used for calling, the discordance rate is 40.7–43.0 %, 35.4–52.5 % and 3.4 % respectively for SS, MS-SS, and MS-LD methods (Fig. 3 top panel lines and dots in cool colors: black, blue, and green). In general, MS-LD method generates higher quality calls than SS and MS-SS. Both MS-LD and MS-SS are able to produce genotype calls at more heterozygous sites than SS. Although MS-SS sometimes generate calls of lower quality than SS (Fig. 3 top panel), it is largely because of the extra sites detected that are generally harder to call. For example, when restricting concordance analysis to sites that are detected by all seven methods, MS-SS always outperform SS (Fig. 3 bottom panel).

Overall, within each category, methods perform very similarly. For the three SS methods, glfMultiples (GM, yellow) and GATK (orange) tend to call at slightly more

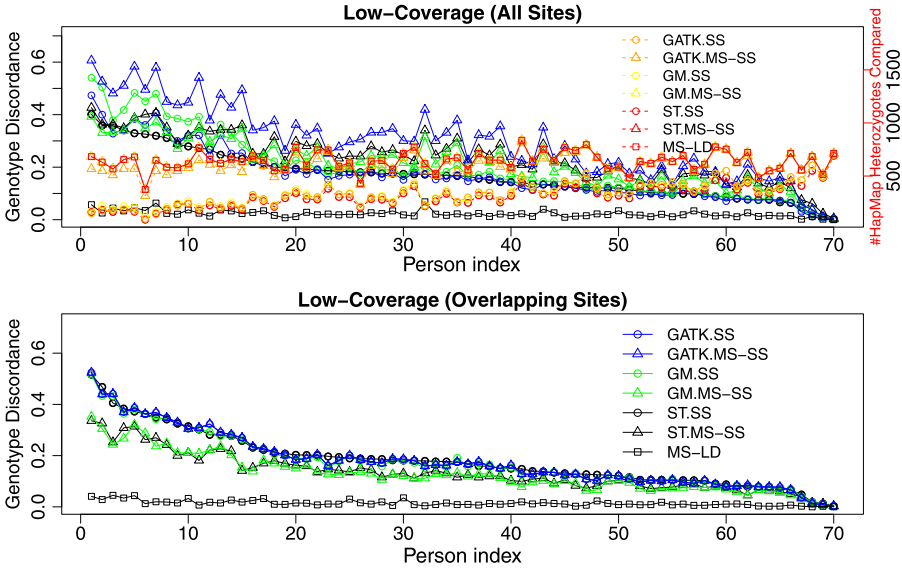

Fig. 4 Comparison of methods on low coverage data from the 1000 genomes pilot project. Three classes of methods, namely SS, MS-SS, and MS-LD, are compared in terms of both number of heterozygotes detected and genotype concordance with experimental genotypes (from the International HapMap project) at detected sites, for 70 individuals sequenced at low coverage (average depth ∼4X) in the 1000 Genomes Pilot Project. The right Y-axis shows the number of sites where the method generates a genotype call and where the experimental genotype is heterozygous. Warm color (*red*, *yellow* and *orange*) *dotted lines* and *points* use this axis. The left Y-axis shows the genotype discordance rate at the compared heterozygotes. Cool color (*blue*, *green* and *black*) *solid lines* and *points* use this axis. For both SS and MS-SS, three methods are used: GATK (GATK UnifiedGenotyper), GM (glfMultiples), and ST (samtools). For clarity, the right-Y axis legend is only shown in the *top panel* (low-coverage all sites) and the left-Y axis legend only shown in the *bottom panel* (low-coverage overlapping sites). (Color figure online)

heterozygous sites than samtools (ST, red). GATK (blue) and samtools (ST, black) generate slightly more accurate calls at heterozygous sites than glfMultiples (GM, green) (Figs. 3 and 4 top panel, circle points). At the overlapping sites, glfMultiples generates slightly more accurate calls than samtools and GATK (Figs. 3 and 4 bottom panel, circle points). For the three MS-SS methods, GATK (orange) tends to call at slightly fewer heterozygous sites than glfMultiples (GM, yellow) and samtools (ST, red) and calls genotypes with slightly less accurate quality at both all-heterozygous and overlapping sites compared (Figs. 3 and 4, triangle points). We only included one MS-LD method since systematic comparisons have been reported elsewhere [32, 87] as discussed at the end of Sect. 3.3.2.

Figure 4 shows the results for the 70 low coverage individuals. Consistent with observations from the two high coverage individuals when a small percentage of reads are used for calling, MS-LD is the only viable method when dealing with low coverage MPS data. The multi-sample methods (MS-LD and MS-SS) have higher power to detect SNPs. For example, the average discordance rate for the 70 individuals is 16.90–19.70 %, 21.26–29.44 %, and 2.26 % (Fig. 4 top panel lines and points in cool colors) at an average of 314–343, 604–641, and 641 detected heterozygote sites (Fig. 4 top panel lines and dots in red) using SS, MS-SS, and MS-LD, respectively.

When restricting analysis to the overlapping sites (average 296 because all sites detected by SS are detected by MS-SS and MS-LD), the average discordance rate is 17.77–18.24 %, 12.88–18.05 %, and 1.41 %, respectively (Fig. 4 bottom panel).

4 Implications for the Design of MPS-Based Genetic Association Studies

The availability of statistical methods to generate accurate genotype calls for low to medium coverage data has important implications for the design of sequencing-based studies. Le and Durbin [85] evaluated, in terms of SNP detection power, five different designs with the same total sequencing investment of 1600X. The five designs—50 individuals at 32X, 100 individuals at 16X, 200 individuals at 8X, 266 individuals at 6X, and 400 individuals at 4X—allowed evaluations of the trade-off between sample size and sequencing depth. While reducing the per-sample sequencing depth reduces power to detect variants in the sample, increasing sample size is likely to include more copies of the rare alleles in the sample. For example, Le and Durbin reported on one hand a loss of 187 SNPs when the depth dropped from 32X to 16X among the 50 sequenced individuals, while on the other hand a gain of 3628 detected SNPs because of the extra 50 individuals sequenced. In general, their results showed that sequencing a large number of individuals with low depth (4–6X) is more powerful for rare SNP discovery than sequencing a small number of individuals at high depth.

Li et al. [87] also investigated the optimal design problem from an imputation perspective. In particular, they quantified the trade-off between number of SNPs detected and the quality of imputation for these detected SNPs when imputed into an external sample without sequencing data. In particular, they compared two designs: 60 individuals sequenced at 16X and 400 individuals sequenced at 2X. Both were used for imputation into an independent sample of 500 individuals with GWAS level (in this case, roughly 300–600 K SNPs genome-wide) data. They found that the low coverage design is advantageous in terms of both SNP detection power and imputation quality in the external sample for SNPs with MAF >0.5 %. For example, the low coverage design resulted in $\sim$14 % more imputable SNPs and $\sim$7 % increase in average information content, for SNPs with MAF 1–2 %.

The simulations discussed above by Le and Durbin, and Li et al., underpinning the initial design of the low-coverage 1000 Genomes Project, focus mainly on the design of MPS-based reference panels that can be utilized by multiple disease/trait-oriented studies. There are also studies that gauge different design options more explicitly according to statistical power to detect association with phenotypic trait(s).

For instance, Li et al. [87] evaluated 24 different designs for detecting a single disease causing variant. The 24 designs investigated included genotyping tagging SNPs only as in a typical GWAS study in a sample of 3000 individuals, sequencing a subset of individuals of different sizes (400, 1000, 2000, and 3000 individuals) at different depths (2X, 4X, 6X, 12X, and 30X), and imputation into individuals not sequenced. They found the low coverage design (2–4X) a powerful alternative for studying complex traits where a large sample is typically needed, particularly for the detection of uncommon disease causing variants.

Sampson et al. [100] proposed likelihood ratio test statistics on sequencing data to find efficient MPS-based study designs for association analysis with human disease,

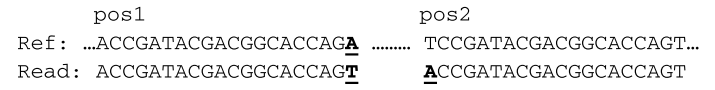

Fig. 5 Alignment and SNP detection are rivals. We illustrate using a toy example that alignment and SNP detection are two competing goals in the sense that standard alignment methods favor mapping reads to genomic positions that would lead to under-calling of SNPs

with a particular focus on discovering rare polymorphisms among the sequenced individuals in the first place and ultimately on detecting rare disease susceptible variants. Their simulations have lead to similar conclusions. Specifically, they found that the optimal depth per sample is 2–8X for detecting rare polymorphisms; and that sequencing as many individuals as possible at depths as shallow as 1X is preferable for association analysis. Among studies considering the design of MPS-based studies [101–109], Kim et al. [104], Wang et al. [106], Ionita-Laza and Laird [109], and Lee et al. [103] also evaluated the impact of sequencing depth. In particular, Kim et al. explicitly assessed the trade-off between sequencing depth and sample size, also finding that sequencing a larger number of individuals at shallower depth is more powerful than sequencing a smaller number of individuals at higher depth.

5 Remaining Issues and Future Directions

Despite the numerous methods developed recently for SNP detection and genotype calling from MPS data, there are still many remaining issues that can benefit from more powerful statistical methods or more efficient computational algorithms.

One issue concerns read alignment. SNP calling methods discussed in this review all assume that the short reads are correctly aligned. Some only collect count information while the most sophisticated methods developed so far take mapping quality into account. Theoretically, the presence of SNPs can affect read alignment. In particular, reads carrying the non-reference allele (i.e., reads that support the presence of SNPs) tend to be biased against during read alignment. For example, Degner et al. [110] reported a significant bias towards higher mapping rate of the reference allele. Indeed, read alignment and SNP detection can be viewed as rivals as illustrated by the toy example in Fig. 5.

The read in Fig. 5 can be mapped to two places in the genome, pos1 and pos2, each with one mismatch. Specifically the read can map to pos1 with a mismatch at the last base, or to pos2 with a mismatch at the first base. Further suppose that the phred score at the first base (A) of the read is 50 and at the last base (T) is 10. Read alignment would favor aligning the read to pos1 because the probability that the last base T is a sequencing error is 10,000 times that of the first base A. But for the same reason that the mismatched base has lower quality, SNP detection at this locus would be favored against. Although being conservative is preferable to having outrageous false positive rates (FPR), SNP detection power can likely be enhanced at a controlled FPR using either SNP-tolerant alignment methods [45] or SNP detection methods that take into account alternative mapping positions.

Further method development is also desired in LD-based genotype calling. First, all the LD-based methods developed are computationally intensive. For example,

genome-wide application of the three LD-based methods to 60 CEU individuals group sequenced by the 1000 Genomes Pilot Project took one to two weeks. Computational burden, increasing at the maximum cubically with sample size, can become prohibitive when sample size exceeds 1000. One potential solution is through cloud computing, as adopted by Myrna for RNA-sequencing differential expression analysis [111]. In addition, existing methods were developed largely for a sample of unrelated individuals; extending these methods to allow family data [112, 113] would be valuable and could be advantageous for rare variant discovery and subsequent association mapping. For example, Chen et al. proposed a method to consider both LD patterns and the constraints imposed by family structure when assigning individual genotypes and haplotypes. Their method implemented in TrioCaller demonstrates that trios provide both higher genotype calling and phasing accuracy across frequency spectrum, both overall and at hard-to-call heterozygous sites.

Finally, the ultimate goal of genomic studies is almost never detecting SNPs or obtaining SNP genotypes but rather to detect SNPs or genes that are associated with phenotypic trait(s) of interest. Therefore, it is desirable to have statistical methods that can incorporate uncertainty in genotype calls, for subsequent imputation and eventually for association mapping [114–116]. There is a rich recent literature for testing rare variants detected in sequencing-based studies. See review articles [117, 118] but there is no consensus on the most powerful method(s). In addition, population stratification, a potential confounder for association analysis, warrants further research in the new sequencing context [119]. It is unclear whether common genetic variants alone suffice for population substructure inference, or whether rare variants detected through sequencing can improve the precision of ancestry inference, which would eventually lead to enhanced power in association analysis. All the aforementioned tasks are directly pertinent to association mapping and can be greatly affected by SNP detection and genotype calling. Although some genotype-free methods [120] have been proposed for various association and population genetics analyses, the vast majority of analyses rely heavily on accurate SNP detection and genotype calling methods. We anticipate more research, both in statistical methodology and computational algorithms, in this important arena.

Acknowledgements The authors thank Mingyao Li and Andrea Byrnes for critical reading of earlier versions of the manuscript. We are also grateful to an anonymous reviewer, whose comments have resulted in an improved manuscript. This research was supported by the National Institute of Health Grants R01 HG006292-01 and HG006703-01 (to Y.L.).

References

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12):5463–5467
2. Shendure J, Ji HL (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145
3. Shendure J et al (2004) Advanced sequencing technologies: methods and goals. Nat Rev Genet 5(5):335–344
4. Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
5. Moore GE (1998) Cramming more components onto integrated circuits. Proc IEEE 86(1):82–85. (Reprinted from Electronics, pp. 114–117, April 19, 1965)

6. Bentley DR et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59
7. Valouev A et al (2008) A High-resolution, nucleosome position map of C. Elegans reveals a lack of universal Sequence-dictated positioning. Genome Res 18(7):1051–1063
8. Ozsolak F et al (2009) Direct RNA sequencing. Nature 461(7265):814–818
9. Eid J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138
10. Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25(4):195–203
11. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
12. Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46
13. Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15(12):1767–1776
14. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16(6):545–552
15. Ng SB et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461(7261):272–U153
16. Ng SB et al (2010) Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42(1):30–35
17. Ng SB et al (2010) Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome. Nat Genet 42(9):790–793
18. Ng SB et al (2010) Massively parallel sequencing and rare disease. Hum Mol Genet
19. Nikopoulos K et al (2010) Next-generation sequencing of a 40 MB linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. Am J Hum Genet 86(2):240–247
20. Roach JC et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328(5978):636–639
21. Lupski JR et al (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362(13):1181–1191
22. Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456(7218):18–21
23. Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747–753
24. Eichler EE et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6):446–450
25. Sidore C et al (2011) Whole genome sequencing of 1000 individuals in an isolated population (Platform 188). Presented at the 12th international congress of human Genetics/61st annual meeting of the American Society of Human Genetics, Montreal, Canada
26. Nielsen R et al (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12(6):443–451
27. Quinlan AR et al (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods 5(2):179–181
28. Erlich Y et al (2008) Alta-Cyclic: a selfoptimizing base caller for next-generation sequencing. Nat Methods 5(8):679–682
29. Kao WC, Stevens K, Song YS (2009) BayesCall: a model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res 19(10):1884–1895
30. Kao WC, Song YS (2011) naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. J Comput Biol 18(3):365–377
31. Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079
32. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
33. Ewing B et al (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185
34. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186–194
35. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18(11):1851–1858
36. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760
37. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595

38. Lunter G, Goodson M (2010) Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res
39. Li R et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25(15):1966–1967
40. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS ONE 4(11):e7767.
41. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11(10):1725–1729
42. Langmead B et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
43. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111
44. Wang K et al (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res
45. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26(7):873–881
46. Grant GR et al (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27(18):2518–2528
47. Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6(11 Suppl):S6–S12
48. Li H, Homer N (2010) A survey of sequence alignment algorithms for Next-generation sequencing. Brief Bioinform 11(5):473–483
49. Trapnell C, Salzberg SL (2009) How to map billions of short reads onto genomes. Nat Biotechnol 27(5):455–457
50. McKenna A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303
51. Brockman W et al (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18(5):763–770
52. Dohm JC et al (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16):10
53. Ossowski S et al (2008) Sequencing of natural strains of arabidopsis thaliana with short reads. Genome Res 18(12):2024–2033
54. Shen Y et al (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20(2):273–280
55. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
56. The International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58
57. Sherry ST et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311
58. Marth GT et al (1999) A general approach to single-nucleotide polymorphism discovery. Nat Genet 23(4):452–456
59. Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25(14):2745–2751
60. Stephens M et al (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38(3):375–381
61. Chen K et al (2007) PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. Genome Res 17:659–666
62. Koboldt DC et al (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25(17):2283–2285
63. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12(4):656–664
64. Hoberman R et al (2009) A probabilistic approach for SNP discovery in High-throughput human resequencing data. Genome Res 19(9):1542–1552
65. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
66. Altshuler D et al (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407(6803):513–516
67. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

68. Frazer KA et al (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10(4):241–251
69. Nielsen R et al (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8(11):857–868
70. Keinan A et al (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39(10):1251–1255
71. Van Tassell CP et al (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5(3):247–252
72. Holt KE et al (2009) Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics 25(16):2074–2075
73. Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics 182(1):295–301
74. Bao H et al (2009) MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads. BMC Genomics 10(Suppl 3):S13
75. Kim SY et al (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinform 12:231
76. Wei Z et al (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res 39(19):e132
77. Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9(9):868–877
78. Li RQ et al (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19(6):1124–1132
79. Ley TJ et al (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456(7218):66–72
80. Bansal V et al (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res 20(4):537–545
81. Hardy HG (1908) Mendelian proportions in a mixed population. Science 28:49–50
82. Weinberg W (1908) On the demonstration of heredity in man. In: Papers on human genetics. Prentice Hall, Englewood Cliffs (1963, translation by S. H. Boyer)
83. Martin ER et al (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26(22):2803–2810
84. Minichiello MJ, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. Am J Hum Genet 79(5):910–922
85. Le SQ, Durbin R (2010) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res
86. Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85(6):847–861
87. Li Y et al (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res
88. DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
89. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84(2):210–223
90. Hudson RR (1991) Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J (eds) Oxford surveys in evolutionary biology. Oxford University Press, New York, pp 1–44
91. Zhao Z, Boerwinkle E (2002) Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. Genome Res 12(11):1679–1686
92. Zhang ZL, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res 31(18):5338–5348
93. Collins FS et al (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931–945
94. Sachidanandam R et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928–933
95. Li Y et al (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816–834
96. Li Y et al (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406
97. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11(7):499–511

98. Smith AV et al (2005) Sequence features in regions of weak and strong linkage disequilibrium. Genome Res 15:1519–1534
99. Liu EY et al (2011) MaCH-Admix: genotype imputation for admixed populations (submitted)
100. Sampson J et al (2011) Efficient study design for next generation sequencing. Genet Epidemiol
101. Liu DJ, Leal SM (2010) Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 87(6):790–801
102. Schaid DJ, Sinnwell JP (2010) Two-stage Case-control designs for rare genetic variants. Hum Genet 127(6):659–668
103. Lee JS et al (2011) On optimal pooling designs to identify rare variants through massive resequencing. Genet Epidemiol
104. Kim SY et al (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. Genet Epidemiol 34(5):479–491
105. Yang F, Thomas DC (2011) Two-stage design of sequencing studies for testing association with rare variants. Hum Hered 71(4):209–220
106. Wang T et al (2010) Resequencing of pooled DNA for detecting disease associations with rare variants. Genet Epidemiol 34(5):492–501
107. Feng B-J et al (2011) Design considerations for massively parallel sequencing studies of complex human disease. PLoS ONE 6(8):e23221
108. Edwards TL, Song Z, Li C (2011) Enriching targeted sequencing experiments for rare disease alleles. Bioinformatics 27(15):2112–2118
109. Ionita-Laza I, Laird NM (2010) On the optimal design of genetic variant discovery studies. Stat Appl Genet Mol Biol 9(1):Article33
110. Degner JF et al (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics
111. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with myrna. Genome Biol 11(8):R83
112. Chen W et al (2010) An efficient LD based variant calling and phasing method for next generation sequencing in trios. ASHG Program # 134
113. Li B, Chen W, Abecasis G (2010) Variant calling from low-pass next generation sequence data in families. ASHG Program # 2993
114. Li Y, Byrnes AE, Li M (2010) To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet 87(5):728–735
115. Wu MC et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93
116. Zawistowski M et al (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87(5):604–617
117. Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308
118. Bansal V et al (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11(11):773–785
119. Price AL et al (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459–463
120. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993