Published in final edited form as:
Int J Comput Biol Drug Des. 2013 ; 6(4): 279–293. doi:10.1504/IJCBDD.2013.056709.

An Evaluation of Allele Frequency Estimation Accuracy Using Pooled Sequencing Data

Yan Guo[*],
Department of Cancer Biology Vanderbilt University Nashville TN 37232

Amma Bosompem,
Department of Pathology, Microbiology and Immunology Vanderbilt University Nashville TN 37232 Aamma.bosompem@vanderbilt.edu

Xue Zhong,
Department of Cancer Biology Vanderbilt University Nashville TN 37232 xue.zhong@vanderbilt.edu

Travis Clark,
Vanderbilt Technologies for Advanced Genomics Vanderbilt University Nashville TN 37232 Travis.a.clark@vanderbilt.edu

Yu Shyr, and
Department of Biostatistics Vanderbilt University Nashville TN 37232 yu.shyr@vanderbilt.edu

Annette S. Kim[*]
Department of Pathology, Microbiology and Immunology Vanderbilt University Nashville TN 37232

Abstract

MicroRNAseq (miRNAseq) is a form of RNAseq technology that has become an increasingly popular alternative to miRNA expression profiling. Unlike messenger RNA (mRNA), miRNA extraction can be difficult, and sequencing such small RNA can also be problematic. We designed a study to test the reproducibility of miRNAseq technology and the performance of the two popular miRNA isolation methods, mirVana and TRIzol, by sequencing replicated samples using microRNA isolated with each kit. Through careful analysis of our data, we found excellent repeatability of miRNAseq technology. The mirVana method performed better than TRIzol in terms of useful reads sequenced, number of miRNA identified, and reproducibility. Finally, we identified a baseline noise level for miRNAseq technology; this baseline noise level can be used as a filter in future miRNAseq studies.

Keywords

next generation sequencing; microRNA; miRNA; miRVana; TRIzol

[*] Corresponding Author yan.guo@vanderbilt.eduannette.s.kim@vanderbilt.edu.

1 Introduction

In large scale genome-wide association studies (GWAS) over the last decade, researchers have used genotyping array to identify hundreds of loci harboring common variants that are associated with complex traits. In GWAS, an effective strategy, often used to alleviate cost, time and labor, was to genotype pools of DNA from many individuals rather than a single sample. Pooling was first used in 1985 in a case–control association study of HLA class II DR and DQ alleles in type I diabetes mellitus (Arnheim et al., 1985). Since then, the concept of pooling has been extensively applied in other types of genetic studies such as linkage studies in plants (Michelmore et al., 1991), measuring the allele frequencies of microsatellite markers and single nucleotide polymorphisms (SNP) (Pacek et al., 1993, Barcellos et al., 1997, Daniels et al., 1998, Shaw et al., 1998, Krumbiegel et al., 2011, Rivas et al., 2011), homozygosity mapping of recessive diseases in inbred populations (Sheffield et al., 1994, Carmi et al., 1995, Nystuen et al., 1996, Scott et al., 1996), and mutation detection (Amos et al., 2000).

A few previous studies have claimed that data generated from pooling studies are accurate and reliable. Most recently, Huang et.al. claimed that minor allele odds ratio estimated from pooled DNA agree fairly well with the minor allele odds ratio estimated from individual genotyping (Huang et al., 2010). Docherty et al. demonstrate that pooling can be effectively applied to the genome-wide Affymetrix GeneChip Mapping 500 K Array (Docherty et al., 2007). Some studies have even found that pooling designs have an advantage in the detection of rare alleles and mutations. For example, Amos et.al. suggested that mutations in individuals could be more efficiently detected using pools (Amos et al., 2000). However, Gastwirth (Gastwirth, 2000) later pointed out that Amos et al. 's method assumed perfect sensitivity.

The practicality of the pooling strategy applied with high throughput sequencing has been a controversial topic. On the one hand, researchers have claimed significant findings using pooled sequencing. For example, researchers have used high throughput pooled sequencing to identify mutations in NUBPL and FOXRED1 in human complex I deficiency (Calvo et al., 2010), and mutations in CYP7B1 and SPG7 in sporadic spastic paraplegia patients (Schlipf et al., 2011). On the other hand, some researchers have argued that, when compared with individual sequencing, pooled sequencing can generate variant calls with high false-positive rates (Harakalova et al., 2011). Another study claimed that the ability to accurately estimate allele frequency from pooled sequencing is limited (Day-Williams et al., 2011).

There have been very few studies aimed at evaluating the accuracy of allele frequency estimated from pooled sequencing data. In our study, we designed a pooling experiment using 48 subjects to evaluate the accuracy of allele frequency AF_{Pool} estimated from pooled high throughput sequencing data. We measured the accuracy by comparing AF_{Pool} estimated from sequencing data with allele frequency AF_{chip} computed from SNP chips. For simplicity, we assume AF_{chip} as the gold standard. Processing of high throughput sequencing data presents many challenges and any changes in processing or filtering criteria may significantly affect the outcome. Thus we also described a robust protocol for making allele frequency calls from raw data, and the filters applied to achieve optimal results.

2 Materials and Methods

For a successful DNA pooling study, an essential requirement is that the pool must contain equal amounts of DNA from each sample so that a robust PCR and library can be obtained. Pools in this study were constructed following previously suggested protocols (Gaukrodger et al., 2005, Sham et al., 2002, Lavebratt and Sengul, 2006, Nejentsev et al., 2009) with modifications to ensure equal amounts were pooled from each subject. DNA concentration was measured twice using the Hoechst Dye method and the PicoGreen method. The DNA concentration for each individual sample was then averaged from the 4 measurements and normalized to 100 ng/μl. The DNA pool construction protocol is shown in Table S1. Briefly, we randomly selected 48 subjects who had SNP Chip data from the Shanghai Women's Health Study and designed a pooling experiment with 8 overlapping pools (pool A-H). Pools A, B, C, and D each contained one DNA sample. Pool E contained 12 samples including samples in Pools A to D. Pool F contained 24 samples including all samples included in Pool E. Pool G contained 36 samples including all samples included in Pool F. Pool H contained 48 samples including all samples included in Pool G. Equal amounts of DNA from each individual DNA sample constituting a pool was added to one tube by a PerkinElmer JANUS liquid handling system. All 48 subjects were genotyped using the Affymetrix SNP 6.0 chip; detailed genotyping methods and stringent quality control criteria are described in Zheng et al. (Zheng et al., 2009).

All data in this study were generated from targeted sequencing on kinome regions (Manning et al., 2002) on 2 lanes of a Illumina HighSeq 2000 sequencer at the Illumina service center. The kinome target regions contain 11,229 intervals, total of 3,212,495 bp, 704 genes, and a median length of 241 bp (range: 116 - 17160). We aligned the FASTQ (Cock et al., 2010) files to National Center for Biotechnology Information (NCBI) human reference genome version 37 (HG19) using the program Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). We then marked duplicates with Picard and carried out regional realignment and quality score recalibration using Genome Analysis Toolkit (GATK) (McKenna et al., 2010).

Since Pool A to Pool D contained only a single subject, we performed SNP calls on Pool A to Pool D using GATK's Unified Genotyper (McKenna et al., 2010) to evaluate the overall batch quality. SNP consistency rate with genotyping chip for those 4 pools were computed as a quality control measure to ensure our data batch has the most optimal quality for pooling analysis.

Using the bam files (Li et al., 2009) from recalibration step, we produced pileup files using Samtools' mpileup command (Li et al., 2009). After applying a variety of combinations of base alignment quality Phred scores (BAQ from samtools mpileup) (Cock et al., 2010), mapping quality Phred score (MAPQ) and depth as filters, we calculated an allele count for each of the four nucleotides at each aligned position. Based on the allele counts we computed allele frequency for the SNPs that overlap with SNP chip data. Since the SNP chip data were from Affymetrix SNP 6.0 chip, which uses HG18 annotation, we had to convert all HG18 locations on the chip to HG19 using the Liftover tool developed by UCSC.

Two different statistics were used to measure the accuracy of allele frequency estimated from pooled sequencing data: Pearson's correlation coefficient and error rate. Pearson's correlation were commonly used as the primary measurement for allele frequency accuracy (Day-Williams et al., 2011, Huang et al., 2010); it was computed between the allele frequency estimated from pooled sequencing data and SNP chip data. However, we will show that high correlation does not necessarily mean allele frequency estimated from pooled sequence data is not significantly different from allele frequency estimated from SNP chip. Thus error rate was calculated as a secondary measurement. Error rate was computed for each SNP and is defined as $|AF_{chip} - AF_{Pool}| / ((AF_{chip} + AF_{Pool})/2$ where AF_{Pool} the allele equency estimated from sequencing data, and AF_{chip} is the allele frequency estimated from SNP chip data. We choose the denominator as $((AF_{chip} + AF_{Pool})/2)$ instead of to avoid dividing by zero.

Previously, we have shown that sequencing data generated outside capture regions can still produce reliable SNP calls (Yan Guo, 2012). In current study, because a significant portion of the reads aligned outside kinome as expected (Table 1), we separately examined SNPs inside kinome and outside kinome. To further evaluate pooling sequencing data quality, we examined SNPs in different minor allele frequency intervals (0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, and 0.4-0.5) and we separately examined allele frequency for each of the four nucleotides.

3 Results

3.1 GWAS Data Quality

The 48 subjects were genotyped using the Affymetrix SNP 6.0 array in a GWAS (Zheng et al., 2009). The original study included three quality control samples in each 96-well plate, and the SNP calls showed a very high concordance rate (mean 99.9%; median 100%) for the quality control samples. In addition, 742 SNPs were genotyped using alternative genotyping platforms for a subset of subjects; they also had a high concordance rate with genotypes obtained from the SNP chip (mean 99.1%; median 99.8%). The SNP chip call rate for the 48 samples used in this study ranged from 97.82% to 97.84%.

3.2 Sequencing Data Quality

We achieved very high quality sequencing data for all pools. From the two lanes on Illumina HighSeq 2000, we obtained a total of 707 million reads. On average, each pool had 88.4 million reads (range: 79.2 - 96.8), about 98.7% (range: 98.6% -98.8%) of the reads were properly aligned; 69% (range 64.7%- 71.7%) of the reads were aligned inside the kinome. The median depth inside the kinome region was 1064 (range: 971 - 1188); the median base alignment quality score of the kinome region was 37.1 (range: 36.1 - 37.5) and the median mapping quality score for the kinome region was 60 (range: 60 - 60). Among the 3,212,495 bp in the kinome target regions, 99.4% (range: 99.3% - 99.4%) of those base pairs had a depth ≥ 10; 98.4% of those base pairs have a depth ≥ 100. Detailed information regarding alignment quality can be found in Table 1.

We observed many high quality SNPs in the four pools that contained only individual sample. Using GQ ≥ 10 and DP ≥ 10 as quality control filters, on average, we observed 2759

(range: 2666 −2911) SNPs per sample in all regions that overlap with SNP chip. The average consistency rate with genotyping chip is 99.8% (range: 99.6% - 99.8%) (Table S2). The high quality SNPs observed in these four pools provided us extra confidence in our data quality.

3.3 Allele Frequency Estimation Accuracy

We computed allele frequency for all genomic positions inside and outside kinome that passed the filtering criteria. The filtering criteria consisted of combinations of three measurements: base alignment quality (BAQ), mapping quality (MAPQ), and depth of coverage. Pearson's correlation coefficient was computed between allele frequency estimated from sequencing data for each pool and allele frequency estimated from their matched SNP chip data.

We found that there was no significant correlation difference between SNPs inside capture regions (Figure 1) and SNPs outside capture regions (Figures S1). For example, in Pool E, the correlation was 0.993 for SNPs inside capture regions, and the correlation was 0.988 for SNPs outside capture regions. Outside capture regions contained a slightly larger number of SNPs. The number of SNPs outside capture regions is highly dependent on capture efficiency of the capture kit and the size of capture regions. The error rate for SNPs outside capture regions and inside capture regions were also very similar. Furthermore, from pooled sequencing data we separately computed allele frequency for the four nucleotides, and calculated their Pearson's correlation coefficients with SNP chip data. We observed high correlations for all four nucleotides in Pool E to Pool H regardless of capture region (Figure S2).

To ensure the robustness of this study, we also computed allele frequency correlation and error rate at different minor allele frequency intervals for inside capture regions (Figure 2) and outside capture regions (Figure S3). SNPs at lower minor allele frequency (MAF) had higher allele frequency correlation and higher error rate, SNPs at higher MAF had lower allele frequency correlation and lower error rate. The correlation at each minor allele frequency interval was lower than the overall correlation. This was due to an artifact from MAF estimation from SNP chip data, which was not truly continuous. For example, in a pool of 12 samples, only 12 possible MAF values are possible. Thus, when dividing into small MAF intervals, the correlation will decrease. Such effect can also be visualized in Figure 1.

We have found that, in general, MAPQ is negatively correlated with error rate and have little effect on total number of SNPs and correlation (Table 2). For example, in Pool E and in capture regions, at BAQ ≥ 0, MAPQ ≥ 20 and Depth ≥ 100, there were 739 SNPs that overlap with Affymetrix 6.0 SNP chip. By fixing BAQ and Depth and increasing MAPQ ≥ 40, the total number of SNPs stayed at 739, correlation only dropped 0.1%, and the mean error rate decreased from 56.62% to 49.80%. Also in Pool E and in capture regions, by increasing depth filter from 50 to 500 had little effect on correlation. The total number of SNPs, however, dropped from 739 to 665, roughly a 10% decrease; and the error rate increased slightly from 56.62% to 58.58%. Surprisingly, increasing BAQ filter did not produce any positive result. For example, in Pool E and in capture regions, at depth ≥ 100,

MAPQ ≥ 20 and BAQ ≥ 0, there were 739 SNPs. By fixing depth, MAPQ and increasing BAQ to ≥ 40, the number of SNPs decreased to 469, correlation decreased from 0.993 to 0.179 and the mean error rate increased from 56.62% to 180.46%. Similar patterns were observed in all four pools (E-H).

Based on the results observed (Table 2), we do not recommend using BAQ computed by Samtools as a filtering criterion when using pooled sequencing to estimate allele frequency. Increasing BAQ filter will significantly decrease the number of SNPs identifiable, and significantly lower the correlation and increase the error rate. On the other hand, using MAPQ and depth as filters can produce very high correlation between allele frequencies estimated from pooled sequencing data and SNP chip data without losing too many SNPs. The ideal MAPQ filter is around 20, and the ideal depth filter varies, depending on the sequencing depth. In our study, depth ≥ 100 produced good results. The details of associations between different filter criteria and results were documented in Table S3-S10 for all four pools (E-H) for both inside capture regions and outside capture regions, respectively.

Moreover, we computed the accuracy for each pool using error rate ≤ 1%, 2%, 3%, 4%, 5% as acceptable thresholds. Accuracy is computed as number of SNPs with error rate greater than the threshold divided by the total number of SNPs. Pool E to Pool H showed similar poor accuracy for all error rate thresholds. For example, using error rate ≤ 5% as the acceptable threshold, the accuracy is only around 25% for Pool E to Pool H. The graphical representation can be visualized in Figure 3, S4 for inside and outside capture regions respectively.

4 Conclusion and Discussion

Our unique study design provided us with an opportunity to evaluate the practicality of allele frequency estimation using pooled sequencing data. Our analysis results suggest that pool size does not make a significant difference on allele frequency estimation accuracy. SNPs outside capture regions and SNPs inside capture regions have comparable allele frequency estimation accuracy and error rate. However, the quality and number of SNPs outside capture regions may vary depending on the capture efficiency of the capture kit and capture region length. We observed no significant difference among the four nucleotides in allele frequency accuracy.

One phenomenon we observed was that even when the correlation is very high (>0.99), the error rate can still remain at an alarming level, which is not acceptable in many situations where high accuracy is required. Correlation has been commonly used to assess allele frequency accuracy (Day-Williams et al., 2011, Huang et al., 2010). However, correlation only measures the linear relationship between two variables. With allele frequencies accuracy measurement, we are more interested in the actual difference between allele frequencies estimated from pooled sequencing data and SNP chip data rather than their linear relationship. Error rate on the other hand, is a more appropriate and informative representation of the actual difference between the two allele frequencies. Thus, higher

priority should be given to error rate instead of correlation when conducting such an evaluation.

Several sources may contribute to the high error rate we have observed. Compared to genotyping of individual samples, pooled DNA comparisons may involve additional sources of variation. The sources include unequal quantity of DNA contributions to the pool by each sample, differential allelic amplification and measurement errors. Such problems have been documented for SNP chip-based studies (Barratt et al., 2002, Downes et al., 2004, Le Hellard et al., 2002, Hoogendoorn et al., 2000). For high throughput sequencing-based pooled DNA studies, additional source of variation could potentially also include sample quality and sequencing depth, library preparation, fragmentation, among others. In practice, even if the pool was constructed perfectly and the quantity of each subject's DNA in the pool was equal, the sequencing data generated from the pool may still not represent each sample equally. This may be due to the nature of high throughput sequencing technology where variation among samples could be introduced as statistical variation in the sequencing process. Under the same conditions, it is known that higher quality samples tend to yield more reads on Illumina sequencer.

Pooling strategy combined with high throughput sequencing technology may permit researchers to conduct their research at a lower cost. However, due to the high level of variation associated with the pooling sequencing data, we do not recommend using pooled sequencing as a substitute for individual sequencing in studies where the primary goal requires estimating allele frequency.

Acknowledgments

This work was partly supported by NIH grants R01HG004517 (JH, CL) and R01CA124558 and R01CA62477 (JL, JH, WZ, CL). Patient recruitment and sample collection was supported by R01CA64277. Sample preparation was conducted at the Survey and Biospecimen Core, which is supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA68485). We also would like to thank Peggy Schuyler for her editorial support.

Biography

Biographical Notes: Yan Guo in a research assistant professor of Department of Cancer Biology at Vanderbilt University. He received his PhD in computer science from University of South Carolina.

Amma Bosompem is a research assistant at Department of Pathology, Microbiology and Immunology.

Xue Zhong is a post doc fellow at Department of Cancer Biology.

Travis Clark is the Technical Director of Sequencing at Vanderbilt Technologies for Advanced Genomics.

Yu Shyr is a professor of Department of Biostatistics and he is the director for Center for Quantitative Sciences.

Annette S. Kim is an assistant professor at Department of Pathology, Microbiology and Immunology.

Reference

AMOS CI, FRAZIER ML, WANG W. DNA pooling in mutation detection with reference to sequence analysis. American journal of human genetics. 2000; 66:1689–92. [PubMed: 10733464]

ARNHEIM N, STRANGE C, ERLICH H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. Proceedings of the National Academy of Sciences of the United States of America. 1985; 82:6970–4. [PubMed: 2995996]

BARCELLOS LF, KLITZ W, FIELD LL, TOBIAS R, BOWCOCK AM, WILSON R, NELSON MP, NAGATOMI J, THOMSON G. Association mapping of disease loci, by use of a pooled DNA genomic screen. American journal of human genetics. 1997; 61:734–47. [PubMed: 9326338]

BARRATT BJ, PAYNE F, RANCE HE, NUTLAND S, TODD JA, CLAYTON DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Annals of human genetics. 2002; 66:393–405. [PubMed: 12485472]

CALVO SE, TUCKER EJ, COMPTON AG, KIRBY DM, CRAWFORD G, BURTT NP, RIVAS M, GUIDUCCI C, BRUNO DL, GOLDBERGER OA, REDMAN MC, WILTSHIRE E, WILSON CJ, ALTSHULER D, GABRIEL SB, DALY MJ, THORBURN DR, MOOTHA VK. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nature genetics. 2010; 42:851–8. [PubMed: 20818383]

CARMI R, ROKHLINA T, KWITEK-BLACK AE, ELBEDOUR K, NISHIMURA D, STONE EM, SHEFFIELD VC. Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. Human molecular genetics. 1995; 4:9–13. [PubMed: 7711739]

COCK PJ, FIELDS CJ, GOTO N, HEUER ML, RICE PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010; 38:1767–71. [PubMed: 20015970]

DANIELS J, HOLMANS P, WILLIAMS N, TURIC D, MCGUFFIN P, PLOMIN R, OWEN MJ. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. American journal of human genetics. 1998; 62:1189–97. [PubMed: 9545387]

DAY-WILLIAMS AG, MCLAY K, DRURY E, EDKINS S, COFFEY AJ, PALOTIE A, ZEGGINI E. An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies. PloS one. 2011; 6:e26279. [PubMed: 22069447]

DOCHERTY SJ, BUTCHER LM, SCHALKWYK LC, PLOMIN R. Applicability of DNA pools on 500 K SNP microarrays for cost-effective initial screens in genomewide association studies. BMC genomics. 2007; 8:214. [PubMed: 17610740]

DOWNES K, BARRATT BJ, AKAN P, BUMPSTEAD SJ, TAYLOR SD, CLAYTON DG, DELOUKAS P. SNP allele frequency estimation in DNA pools and variance components analysis. BioTechniques. 2004; 36:840–5. [PubMed: 15152604]

GASTWIRTH JL. The efficiency of pooling in the detection of rare mutations. American journal of human genetics. 2000; 67:1036–9. [PubMed: 10986050]

GAUKRODGER N, MAYOSI BM, IMRIE H, AVERY P, BAKER M, CONNELL JM, WATKINS H, FARRALL M, KEAVNEY B. A rare variant of the leptin gene has large effects on blood pressure and carotid intima-medial thickness: a study of 1428 individuals in 248 families. Journal of medical genetics. 2005; 42:474–8. [PubMed: 15937081]

HARAKALOVA M, NIJMAN IJ, MEDIC J, MOKRY M, RENKENS I, BLANKENSTEIJN JD, KLOOSTERMAN W, BAAS AF, CUPPEN E. Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm regions of interest-challenges and limitations. Journal of cardiovascular translational research. 2011; 4:271–80. [PubMed: 21360310]

HOOGENDOORN B, NORTON N, KIROV G, WILLIAMS N, HAMSHERE ML, SPURLOCK G, AUSTIN J, STEPHENS MK, BUCKLAND PR, OWEN MJ, O'DONOVAN MC. Cheap, accurate

and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Human genetics. 2000; 107:488–93. [PubMed: 11140947]

HUANG Y, HINDS DA, QI L, PRENTICE RL. Pooled versus individual genotyping in a breast cancer genome-wide association study. Genetic epidemiology. 2010; 34:603–12. [PubMed: 20718042]

KRUMBIEGEL M, PASUTTO F, SCHLOTZER-SCHREHARDT U, UEBE S, ZENKEL M, MARDIN CY, WEISSCHUH N, PAOLI D, GRAMER E, BECKER C, EKICI AB, WEBER BH, NURNBERG P, KRUSE FE, REIS A. Genome-wide association study with DNA pooling identifies variants at CNTNAP2 associated with pseudoexfoliation syndrome. European journal of human genetics : EJHG. 2011; 19:186–93. [PubMed: 20808326]

LAVEBRATT C, SENGUL S. Single nucleotide polymorphism (SNP) allele frequency estimation in DNA pools using Pyrosequencing. Nature protocols. 2006; 1:2573–82.

LE HELLARD S, BALLEREAU SJ, VISSCHER PM, TORRANCE HS, PINSON J, MORRIS SW, THOMSON ML, SEMPLE CA, MUIR WJ, BLACKWOOD DH, PORTEOUS DJ, EVANS KL. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. Nucleic acids research. 2002; 30:e74. [PubMed: 12140336]

LI H, DURBIN R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]

LI H, HANDSAKER B, WYSOKER A, FENNELL T, RUAN J, HOMER N, MARTH G, ABECASIS G, DURBIN R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]

MANNING G, WHYTE DB, MARTINEZ R, HUNTER T, SUDARSANAM S. The protein kinase complement of the human genome. Science. 2002; 298:1912–34. [PubMed: 12471243]

MCKENNA A, HANNA M, BANKS E, SIVACHENKO A, CIBULSKIS K, KERNYTSKY A, GARIMELLA K, ALTSHULER D, GABRIEL S, DALY M, DEPRISTO MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303. [PubMed: 20644199]

MICHELMORE RW, PARAN I, KESSELI RV. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proceedings of the National Academy of Sciences of the United States of America. 1991; 88:9828–32. [PubMed: 1682921]

NEJENTSEV S, WALKER N, RICHES D, EGHOLM M, TODD JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324:387–9. [PubMed: 19264985]

NYSTUEN A, BENKE PJ, MERREN J, STONE EM, SHEFFIELD VC. A cerebellar ataxia locus identified by DNA pooling to search for linkage disequilibrium in an isolated population from the Cayman Islands. Human molecular genetics. 1996; 5:525–31. [PubMed: 8845847]

PACEK P, SAJANTILA A, SYVANEN AC. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR methods and applications. 1993; 2:313–7. [PubMed: 8324505]

RIVAS MA, BEAUDOIN M, GARDET A, STEVENS C, SHARMA Y, ZHANG CK, BOUCHER G, RIPKE S, ELLINGHAUS D, BURTT N, FENNELL T, KIRBY A, LATIANO A, GOYETTE P, GREEN T, HALFVARSON J, HARITUNIANS T, KORN JM, KURUVILLA F, LAGACE C, NEALE B, LO KS, SCHUMM P, TORKVIST L, DUBINSKY MC, BRANT SR, SILVERBERG MS, DUERR RH, ALTSHULER D, GABRIEL S, LETTRE G, FRANKE A, D'AMATO M, MCGOVERN DP, CHO JH, RIOUX JD, XAVIER RJ, DALY MJ. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics. 2011; 43:1066–73. [PubMed: 21983784]

SCHLIPF NA, SCHULE R, KLIMPE S, KARLE KN, SYNOFZIK M, SCHICKS J, RIESS O, SCHOLS L, BAUER P. Amplicon-based high-throughput pooled sequencing identifies mutations in CYP7B1 and SPG7 in sporadic spastic paraplegia patients. Clinical genetics. 2011; 80:148–60. [PubMed: 21623769]

SCOTT DA, CARMI R, ELBEDOUR K, YOSEFSBERG S, STONE EM, SHEFFIELD VC. An autosomal recessive nonsyndromic-hearing-loss locus identified by DNA pooling using two inbred Bedouin kindreds. American journal of human genetics. 1996; 59:385–91. [PubMed: 8755925]

SHAM P, BADER JS, CRAIG I, O'DONOVAN M, OWEN M. DNA Pooling: a tool for large-scale association studies. Nature reviews. Genetics. 2002; 3:862–71.

SHAW SH, CARRASQUILLO MM, KASHUK C, PUFFENBERGER EG, CHAKRAVARTI A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome research. 1998; 8:111–23. [PubMed: 9477339]

SHEFFIELD VC, CARMI R, KWITEK-BLACK A, ROKHLINA T, NISHIMURA D, DUYK GM, ELBEDOUR K, SUNDEN SL, STONE EM. Identification of a Bardet-Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. Human molecular genetics. 1994; 3:1331–5. [PubMed: 7987310]

YAN GUO JL, HE JING, LI CHUNG-I, CAI QIUYIN, SHU XIAO-OU, ZHENG WEI, LI CHUN. Exome Sequencing Generates High Quality Data in Non-Target Regions. BMC genomics. 2012

ZHENG W, LONG J, GAO YT, LI C, ZHENG Y, XIANG YB, WEN W, LEVY S, DEMING SL, HAINES JL, GU K, FAIR AM, CAI Q, LU W, SHU XO. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature genetics. 2009; 41:324–8. [PubMed: 19219042]

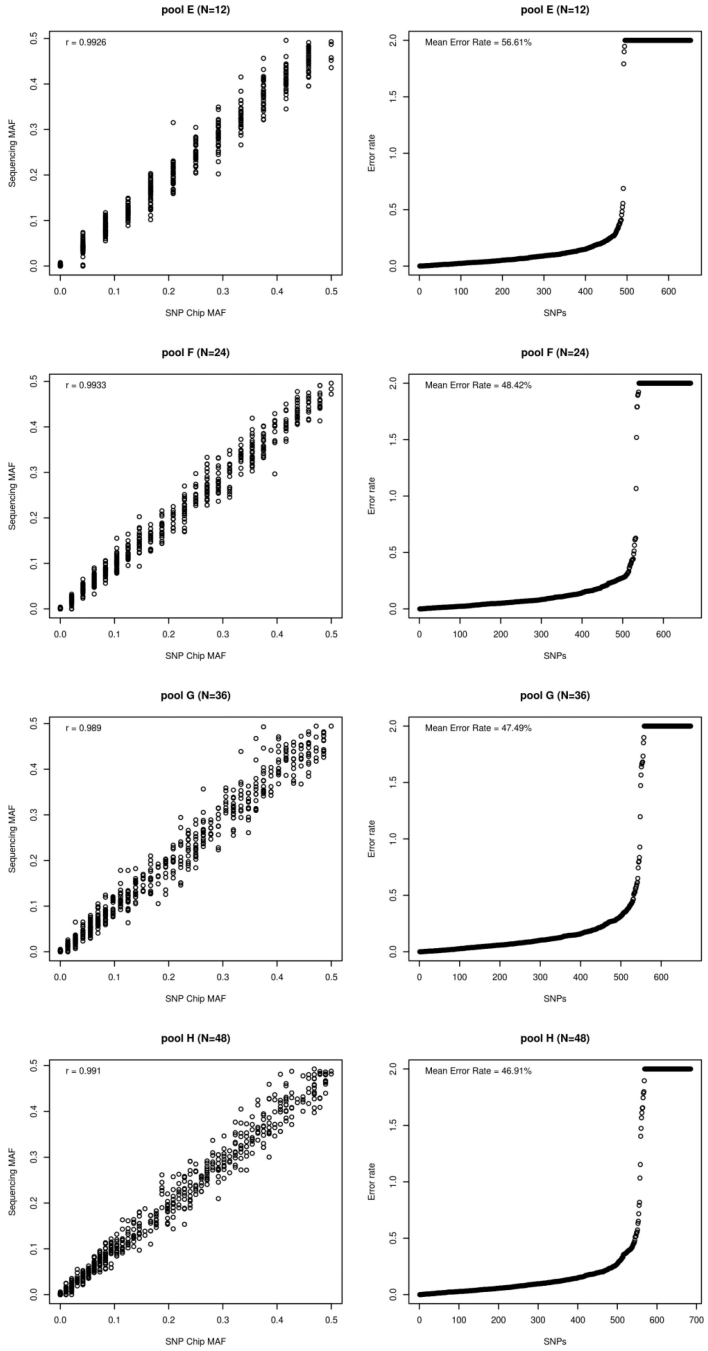

Figure 1.

Correlation of allele frequencies and error rate between sequencing and SNP chip inside capture regions.

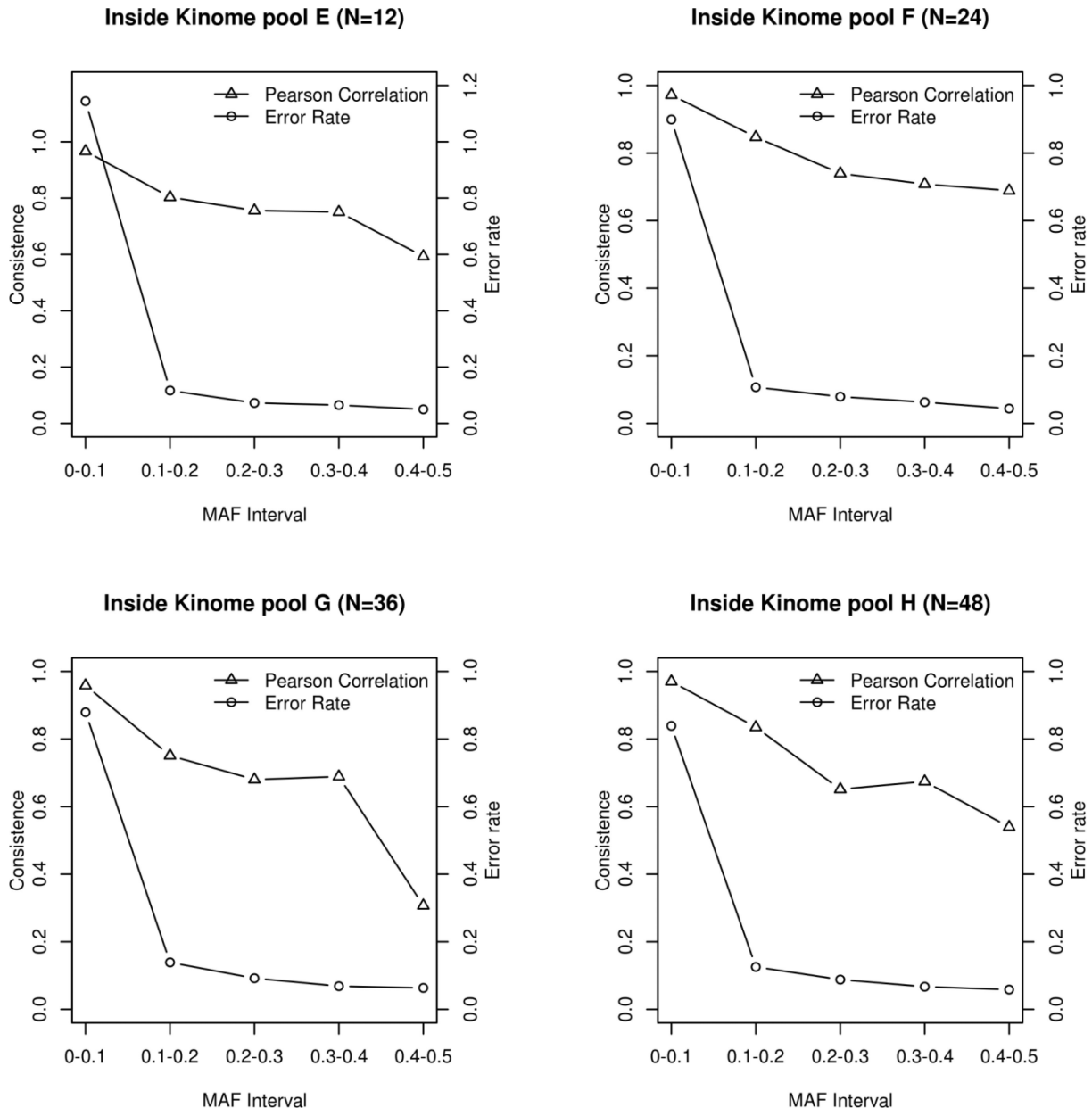

Figure 2.

Correlation of allele frequencies and error rate between sequencing and GWAS by MAF intervals inside capture regions

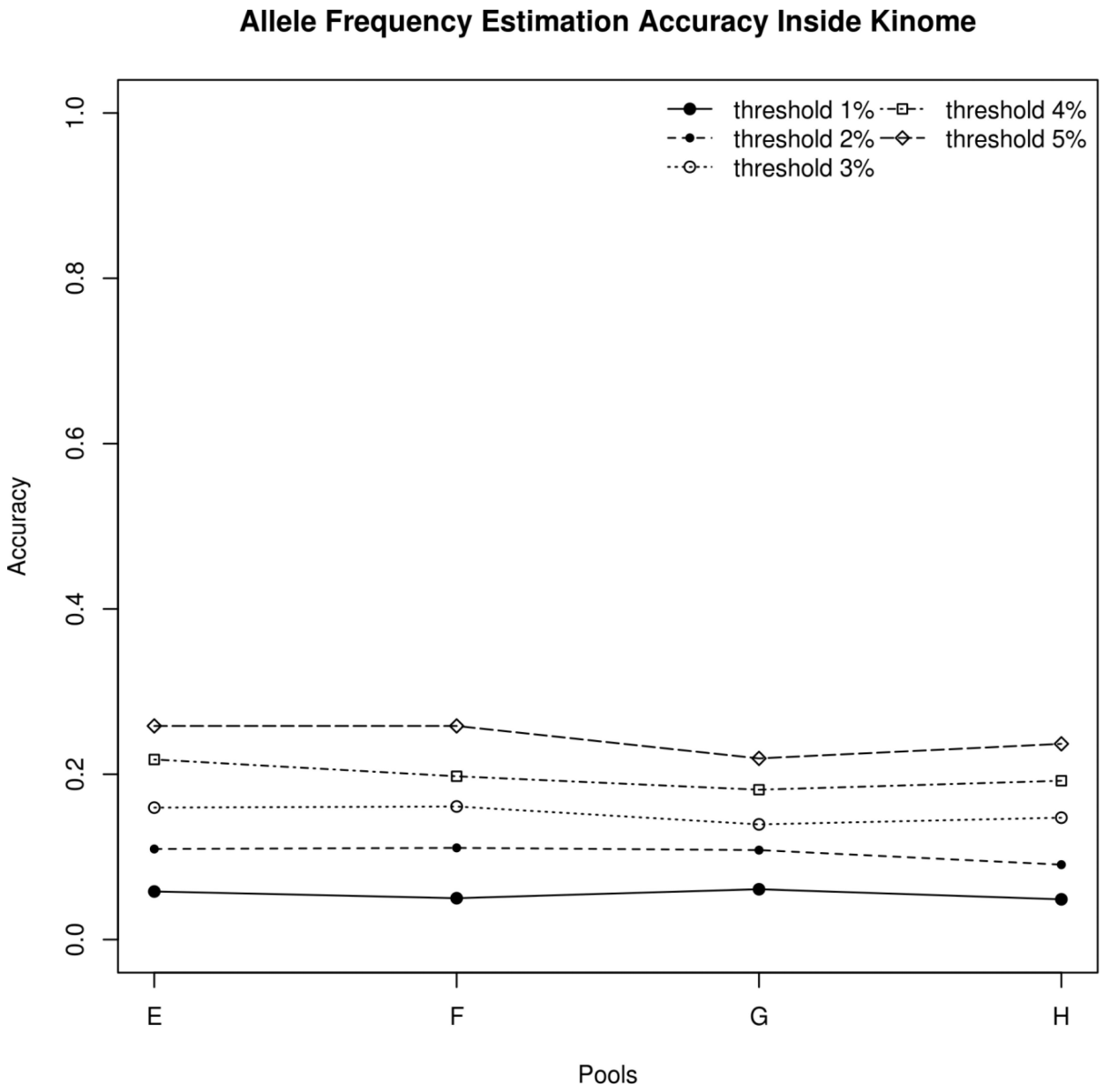

Figure 3.
Overall accuracy acceptable error rate at 1-5%.

Table 1

Alignment Quality Summary

Pool	Total (m[1])	Aligned (m)	Properly Paired (m)	Reads Aligned In Kinome (m)	Median Depth In Kinome	Median BAQ in Kinome	Median MAPQ in Kinome	Kinome % with depth ≥ 1	Kinome % with depth ≥ 10	Kinome % with depth ≥ 100	Kinome % with depth ≥ 200
A (n=1)	81.7	80.8	80.2	57.1	998	37	60	100	99	98	96
B (n=1)	96.8	95.4	94.6	62.6	1091	36	60	100	99	98	96
C (n=1)	90.0	88.9	88.3	64.3	1113	37	60	100	99	98	96
D (n=1)	95.3	94.2	93.4	68.3	1188	37	60	100	99	98	97
E(n=12)	90.5	89.2	88.4	60.3	1054	37	60	100	99	98	96
F (n=24)	89.2	88.1	87.3	61.5	1073	37	60	100	99	98	96
G (n=36)	79.2	78.2	77.6	55.0	971	37	60	100	99	98	96
H(n=48)	84.7	83.5	82.8	58.8	1023	37	60	100	99	98	96
Mean	**88.4**	**87.3**	**86.6**	**61.0**	**1063.9**	**37.1**	**60.0**	**99.5**	**99.4**	**98.4**	**96.3**
SD	**6.2**	**6.1**	**6.0**	**4.2**	**69.0**	**0.4**	**0.0**	**0.0**	**0.0**	**0.1**	**0.2**

1. Read count in millions

Table 2

Filtering criteria's effect on SNP number, correlation and error rate

	Inside Capture Regions						Outside capture Regions					
	BAQ	**MAPQ**	**DP**	**SNPs**	**Correlation**	**Mean Error**	**BAQ**	**MAPQ**	**DP**	**SNPs**	**Correlation**	**Mean Error**
Fixing Depth and MAPQ	0	20	100	739	0.993	56.62%	0	20	100	875	0.987	45.76%
	10	20	100	739	0.99	48.56%	10	20	100	869	0.986	38.84%
	20	20	100	736	0.928	45.38%	20	20	100	845	0.935	38.12%
	30	20	100	703	0.507	115.50%	30	20	100	730	0.596	109.76%
	40	20	100	469	0.179	180.46%	40	20	100	301	0.151	191.40%
Fixing Depth and BAQ	0	20	100	739	0.993	56.62%	0	20	100	875	0.987	45.76%
	0	30	100	739	0.992	49.98%	0	30	100	866	0.983	40.20%
	0	40	100	739	0.991	49.80%	0	40	100	864	0.982	40.40%
Fixing BAQ and MAPQ	0	20	50	739	0.993	56.62%	0	20	50	1107	0.978	44.86%
	0	20	100	739	0.993	56.62%	0	20	100	875	0.987	45.76%
	0	20	200	736	0.993	56.72%	0	20	200	653	0.99	49.28%
	0	20	300	722	0.993	57.32%	0	20	300	494	0.991	51.02%
	0	20	400	688	0.993	57.88%	0	20	400	374	0.991	49.94%
	0	20	500	665	0.993	58.58%	0	20	500	277	0.99	48.22%