European Journal of Human Genetics (2009) 17, 383–386
© 2009 Macmillan Publishers Limited All rights reserved 1018-4813/09 $32.00
www.nature.com/ejhg

SHORT REPORT

SNP frequency estimation using massively parallel sequencing of pooled DNA

Max Ingman*,[1] and Ulf Gyllensten[1]

[1] *Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden*

Resequencing of genomic regions that have been implicated by linkage and/or association studies to harbor genetic susceptibility loci represents a necessary step to identify causal variants. Massively parallel sequencing (MPS) offers the possibility of SNP discovery and frequency determination among pooled DNA samples. The strategies of pooling DNA samples and pooling PCR amplicons generated from individual DNA samples were evaluated, and both were found to return accurate estimates of SNP frequencies across varying levels of sequence coverage.

European Journal of Human Genetics (2009) **17,** 383–386; doi:10.1038/ejhg.2008.182; published online 15 October 2008

Keywords: SNP; sequencing

Introduction

Massively parallel sequencing (MPS) represents a major advance in sequencing technology[1] as it has enabled the determination of the nucleotide sequence of a targeted genomic region or a complete genome at an unprecedented rate. A key application for MPS is to determine the sequence of genomic regions that harbor genetic susceptibility factors for common diseases in affected individuals. The very high coverage possible with MPS also provides an opportunity for SNP discovery in pooled samples. Pooling of DNA from many individuals could be used to determine the SNP spectrum or for combined SNP discovery and association analyses. However, the performance of MPS for SNP discovery from pooled DNAs needs to be evaluated and optimal conditions must be established before the potential of this method can be determined. Here, we have studied the effect of DNA pooling strategy and depth of sequence coverage on SNP discovery using sequence data generated by the Roche 454 GS-FLX high-throughput sequencer.

Materials and methods

To compare SNP frequencies estimated by MPS of pooled individuals with those calculated from individual genotypes, we selected a genomic region of $\sim$17 kb in length (chr9:114 677 513–114 694 357) that contains one gene. This gene, *TSCOT*, has previously shown linkage to cervical cancer *in situ* in affected sib-pair (ASP) analyses and has been proposed to affect susceptibility to persistent infection by human papillomavirus.[2] DNA from 96 cervical cancer patients, in which 16 SNPs in or around this gene had previously been studied by TaqMan assays, were chosen for these analyses. High molecular weight genomic DNA was extracted from PBLs using an organic extraction technique (phenol–chloroform), the DNA precipitated with ethanol and dissolved in water. The DNA concentration of both the DNA stock and diluted DNA was measured using nano-drop.

The 17 kb region was amplified in four long-range PCR amplicons (primers described in Supplementary Table 1). DNA amplification was made using Phusion High Fidelity DNA Polymerase (Finnzymes, Finland) in HF buffer (fidelity $= 4.4 \times 10^{-7}$). Reagent amounts and cycling conditions were as per the manufacturer's instructions. Two pooling strategies were used: (1) the four PCR amplicons were generated separately from each of the 96 DNA samples and normalized prior to pooling (as amplicon yield varied between individual PCRs

*Correspondence: Dr M Ingman, Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala 751 85, Sweden.
Tel: +46 0 18 471 4969; Fax: +46 0 18 471 4931;
E-mail: max.ingman@genpat.uu.se

Received 9 April 2008; revised 5 August 2008; accepted 4 September 2008; published online 15 October 2008

(Supplementary Table 1)), and (2) equal quantities of genomic DNA from each of the 96 individuals were pooled and the four long-range PCRs were run separately on the DNA pool and the amplicons were combined in equal quantities (normalized for sequence length) for sequencing. The amplicons from both these pools were sequenced on a single slide (physically separated) using 454 GS-FLX MPS technology. We obtained an average read length of approximately 200 bp and an average of 80 000 reads for each pool. These reads were aligned to the 17 kb reference sequence using BLAST and the output parsed for the location of the alignment and the position and type of mismatched bases. Alignments were required to be at least 90% the length of the read with a maximum of three mismatches. With these criteria, 45 000 reads were uniquely mappable to the reference translating to just over 500 times coverage for each of the two pools. Mismatches were considered to be valid SNPs if they had a quality score ≥20, were found in more than one read and these reads had more than one start point. We also excluded repeated regions of low (<15%) divergence to prevent calling SNPs based on mismapped reads.

To assess the effect of coverage depth on the ability to discover SNPs of different frequency and the accuracy of the frequency estimation, we randomly removed reads from each dataset to create four subsets of data from each pool with a mean coverage of 500 ×, 250 ×, 100 × and 50 ×. We compared the frequencies of the 16 SNPs previously determined by TaqMan assays with those estimated in the four subsets of our two types of pools. These 16 SNPs have minor allele frequencies ranging from 3.7% to over 40%, and therefore should provide a good indication of the performance of the two pooling strategies.

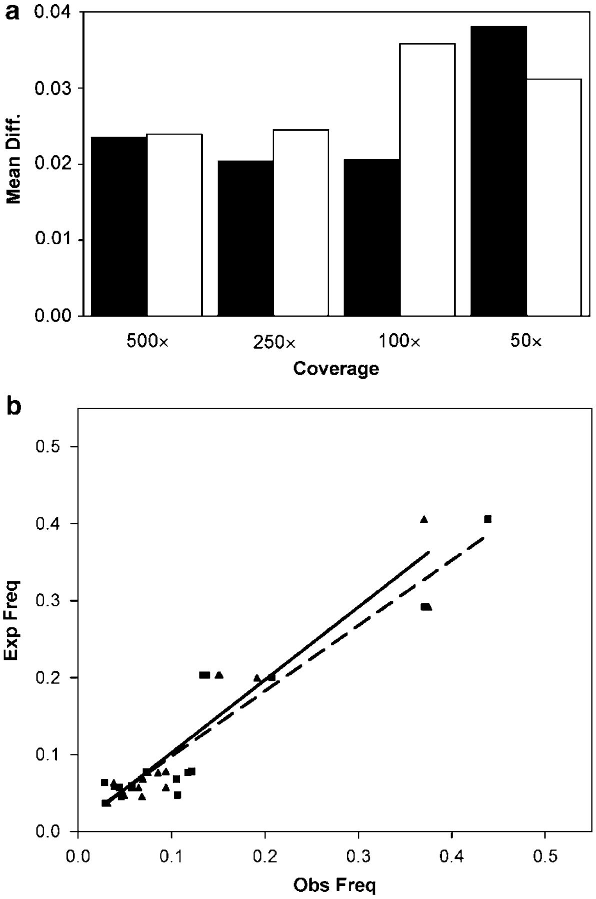

Figure 1 (**a**) Mean difference between the frequencies of 16 SNPs derived from TaqMan analysis and their frequencies estimated from sequencing pooled PCR products (solid bars) and pooled DNA samples (open bars) with varying degrees of coverage. (**b**) Relationship between SNP frequencies derived from TaqMan analysis (Exp Freq) and those derived by sequencing pooled PCR products (Obs Freq) with 500 × (triangle, solid line; $r^2=0.904$, $P<0.0001$) and 50 × (square, dashed line; $r^2=0.878$, $P<0.0001$) coverage.

Results

Comparisons with data from dbSNP build 129[3] showed that although many previously identified SNPs were detected by high coverage sequencing, several new SNPs were also found (Supplementary Table 2). We detected 59 SNPs that were present in both the pooled PCR product as well as the pooled DNA datasets, of which 26 are present in dbSNP and 33 are not. The majority of the novel SNPs (29/33) are at less than 5% frequency but the fact that they are found in both datasets indicates their validity. We also find 83 SNPs that are detected in one dataset but not the other. All but three of these are below 1% frequency and none are found in dbSNP.

The mean difference across the 16 SNPs between the TaqMan-derived SNP frequencies and those estimated through sequencing was calculated in each subset of data (Figure 1a). The 500 × and 250 × coverage provide somewhat closer estimates to the TaqMan frequencies than 100 × and 50 ×, but the precision of estimates is not consistently higher across the different levels of coverage for the pooled PCR products as compared with the pooled DNA samples.

We plotted the individual frequencies of the 16 SNPs at 500 × and 50 × coverage from the pooled amplicon dataset against the data for individual genotypes (TaqMan) (Figure 1b). In both instances the r^2 is high (500 ×, $r^2=0.904$, $P<0.0001$; 50 ×, $r^2=0.878$, $P<0.0001$, respectively), indicating that the frequency estimates are consistently accurate, even at relatively low coverage. In fact, none of the 16 SNPs from the four subsets of pooled PCR amplicons deviated from the expected (TaqMan) frequencies (Fisher's Exact test), whereas one SNP showed deviations in both the 500 × and 100 × coverage subsets of the pooled genomic DNA dataset. Therefore, the two pooling strategies appear to have similar performance in terms of their ability to estimate the frequency of SNPs across a wide frequency range. Although both pooling strategies can be used to estimate SNP frequencies, the variance of the SNP frequency increases with lower coverage (Supplementary Table 3).

A comparison of the SNPs identified in the 500× coverage dataset for the pooled PCR products with those found in the other coverage/pooling groups, showed that no SNPs with a frequency of greater than 10% in the 500× coverage pooled PCR dataset were missed in any of the other seven datasets (Supplementary Table 4). However, with decreasing coverage, increasing numbers of low frequency (<5%) SNPs were not identified.

Discussion

All the 16 SNPs, including those at low frequency (five are under 7%), were identified at all coverage levels in both the pooled DNA samples and pooled PCR product datasets. However, the amount of coverage provides a limitation on the frequencies that are possible to discern. As we require an SNP to be supported by at least two sequence reads, the lowest identifiable SNP frequency would be 0.4% for 500× coverage and 4% for 50× coverage. As a singleton will have a frequency of approximately 0.5% in a pool of 96 samples, 500× coverage should be sufficient to resolve the majority of all SNPs. This assumption is affected by the error rate of PCR and the MPS technique, as well as the sampling distribution. The mismatch error rate of the 454 GS20 MPS technology has been estimated to be less than 0.1%.[4] Assuming an error rate of 10^{-3} per position, among the 9×10^6 mapped nucleotides (45 000 reads of 200 bp), we expect 9×10^3 errors. Most of these will be represented by variants in a single read. The probability of two reads having the same SNP is on the order of 10^{-6}, or less than 10 SNPs in each dataset. However, as we filtered SNPs in individual reads based on quality score, our mismatch error rate is likely to be lower than 0.1% as evidenced by the fact that we find approximately 10 times less SNPs covered by single reads than the expected 9×10^3 errors (Supplementary Table 4). Therefore, most of the low frequency SNPs found in at least two reads are likely to be true. The error rate reduces further if three reads are required to verify the presence of an SNP (error probability $= 10^{-9}$). This more stringent criterium reduces the number of SNPs found in the pooled DNA dataset but not the pooled PCR product dataset or vice versa from 83 to 20 and all these SNPs are likely to be real. The detection rate of rare alleles will also be affected by the number of individuals included in the DNA pool and the sequence coverage of each position. For instance, when pooling 100, 200 and 500 chromosomes, the probability of sampling an SNP with a population frequency of 1%, is 63, 87 and 99%, respectively. Also, the probability that an SNP with a frequency of 1% in a DNA pool will be represented by two or more sequence reads with 100×, 250× and 500×, coverage is 26, 71 and 96%, respectively. Therefore, although we have detected a substantial number of SNPs with a frequency <5% in our sample of 96 individuals, projects focusing on detection of low frequency variants should aim for a larger DNA pool size and high coverage.

Our results show that MPS using either pooled amplicons from PCRs on individual DNA samples or PCR on pooled DNA are viable strategies for SNP discovery and allele frequency estimation. Amplifying each sample DNA is considerably more labor intensive and costly than pooling the DNAs prior to PCR. Although the former method ensures the template pool for sequencing has equal representations of each individual sample, in our study it was not statistically superior. Interestingly, very high coverage (500× coverage) did not perform significantly better than the low 50× coverage for the 16 SNPs studied. This implies that for a pool of 96 samples, coverage levels of 50× to 100× should be sufficient to detect common SNPs, enabling longer genomic regions to be studied in the same MPS run. The main advantage of using 500× coverage comes when the target SNPs are expected to have very low frequencies in the material to be examined, such as somatic mutations in a tumor or low frequency private congenic mutations in a pool of DNA from many patients. The limitation to detection of low frequency variants in such a case may be the inherent error rate of the PCR or the sequencing technology used.

A limitation to the present strategy is the need to establish long-range PCR for large genomic regions. Long-range PCR is viable, but will be costly and time-consuming for megabase regions. In addition, even when DNA samples are extracted using the same method, it is unlikely that each sample will amplify with the same efficiency, introducing bias when DNA samples are pooled prior to PCR. Also, coverage tends to be higher both at the ends of amplicons (due to bias in DNA shearing) and where amplicons overlap, potentially introducing bias making rare variants more likely to be identified in such regions. Technologies that provide targeted enrichment of specific genomic regions (eg.[5–7]) hold great promise to enrich for genomic regions to be resequenced in single individuals, or by tagging in smaller pools. Whether these technologies are also suitable when used on a large number of pooled DNA samples as a tool for SNP discovery remains to be determined.

References

1 Margulies M, Egholm M, Altman WE *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**: 376–380.

2 Engelmark MT, Ivansson EL, Magnusson JJ *et al*: Identification of susceptibility loci for cervical carcinoma by genome scan of affected sib-pairs. *Hum Mol Genet* 2006; **15**: 3351–3360.

3 Sherry ST, Ward MH, Kholodov M *et al*: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308–311.

4 Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**: R143.

5 Albert TJ, Molla MN, Muzny DM *et al*: Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007; **4**: 903–905.

6 Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M: Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* 2005; **33**: e71.

7 Hodges E, Xuan Z, Balija V *et al*: Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* 2008; **39**: 1522–1527.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)