BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 6 2013, pages 799–801
doi:10.1093/bioinformatics/btt041

Sequence analysis

Advance Access publication January 28, 2013

AbCD: arbitrary coverage design for sequencing-based genetic studies

Jian Kang[1,†], Kuan-Chieh Huang[2,†], Zheng Xu[2,3], Yunfei Wang[2,3], Gonçalo R. Abecasis[4] and Yun Li[2,3,5,*]

[1]Faculty of Kinesiology, University of Calgary, Calgary, AB T2N1N4, Canada, [2]Department of Biostatistics, [3]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599-7264, [4]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 and [5]Department of Computer Science, University of North Carolina, Chapel Hill, NC 27599-7264, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Recent advances in sequencing technologies have revolutionized genetic studies. Although high-coverage sequencing can uncover most variants present in the sequenced sample, low-coverage sequencing is appealing for its cost effectiveness. Here, we present AbCD (arbitrary coverage design) to aid the design of sequencing-based studies. AbCD is a user-friendly interface providing pre-estimated effective sample sizes, specific to each minor allele frequency category, for designs with arbitrary coverage (0.5–30×) and sample size (20–10 000), and for four major ethnic groups (Europeans, Africans, Asians and African Americans). In addition, we also present two software tools: ShotGun and DesignPlanner, which were used to generate the estimates behind AbCD. ShotGun is a flexible short-read simulator for arbitrary user-specified read length and average depth, allowing cycle-specific sequencing error rates and realistic read depth distributions. DesignPlanner is a full pipeline that uses ShotGun to generate sequence data and performs initial SNP discovery, uses our previously presented linkage disequilibrium-aware method to call genotypes, and, finally, provides minor allele frequency-specific effective sample sizes. ShotGun plus DesignPlanner can accommodate effective sample size estimate for any combination of high-depth and low-depth data (for example, whole-genome low-depth plus exonic high-depth) or combination of sequence and genotype data [for example, whole-exome sequencing plus genotyping from existing Genomewide Association Study (GWAS)].

Availability and implementation: AbCD, including its downloadable terminal interface and web-based interface, and the associated tools ShotGun and DesignPlanner, including documentation, examples and executables, are available at http://www.unc.edu/~yunmli/AbCD.html.

Contact: yunli@med.unc.edu

Received on November 13, 2012; revised on January 20, 2013; accepted on January 22, 2013

1 INTRODUCTION

Massively parallel sequencing (MPS) technologies have transformed the field of genomic studies (Ansorge, 2009; Mardis, 2008; Metzker, 2010). High-coverage sequencing, which is able to detect essentially every variant present in the sequenced individuals (Coventry *et al.*, 2010; Nelson *et al.*, 2012), has been successful for the identification of genetic variants causing Mendelian diseases (Bamshad *et al.*, 2011; Ionita-Laza *et al.*, 2011; Ng *et al.*, 2010), but it remains cost prohibitive for large samples required for the study of complex traits. Recent studies (Flannick *et al.*, 2012; Li *et al.*, 2011; Pasaniuc *et al.*, 2012) have proposed low-coverage sequencing as a powerful alternative. However, there are no resources or software tools that directly evaluate the statistical power of an arbitrary coverage sequencing-based design to quantitatively guide the design decisions. Here, we present AbCD (arbitrary coverage design), a user-friendly interface with pre-estimated minor allele frequency (MAF) and ethnicity-specific effective sample sizes—which is directly associated with power of subsequent association studies (Pritchard and Przeworski, 2001)—for a wide range of sequencing-based designs with average sequencing depth ranging from 0.5 to 30× and sample size ranging from 20 to 10 000. In addition, we present ShotGun and DesignPlanner, two software tools that generate the effective sample size estimates behind AbCD. We anticipate AbCD and the two associated tools will greatly facilitate the design of sequencing-based studies.

2 IMPLEMENTATION

2.1 AbCD

AbCD has four major features. First, it allows arbitrary coverage sequencing with average coverage/depth (d) ranging from 0.5 to 30×, so that low-, median- and high-coverage sequencing design can be evaluated. Second, AbCD includes a wide range of sequenced sample sizes (n), from 20 to 10 000. We have estimated the MAF-specific effective sample sizes using actual simulations when the number of sequenced individuals (n) takes the following 15 values: 20, 40, 60, 80, 100, 200, 400, 600, 800, 1000, 2000, 4000, 6000, 8000 and 10 000. Simulations involved generation of chromosomes using cosi (Schaffner *et al.*, 2005) that mimics realistic linkage disequilibrium (LD) patterns and demographic histories consistent with real data, generation of short-read sequence data and initial SNP discovery through ShotGun from the simulated chromosomes and final LD-based variant calling and effective sample size estimation using *thunder*

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(Li *et al.*, 2011). Results were averages from 10 regions simulated from the cosi calibrated distributions (see the Simulations Details tab on our website for details). We have previously used the same simulation procedure and found it generates effective sample size estimates close to estimates from real data (Li *et al.*, 2010, 2011). Third, AbCD considers four major ethnic groups—Europeans, Africans, Asians and African Americans—so that multi-ethnic study design can also be accommodated. Finally, AbCD is user-friendly, and it functions through simple mouse clicking. Needless to say, it is fast because no calculation except interpolation for sample sizes not covered (integers between 20 and 10 000, but not among the 15 values aforementioned) is needed.

Figure 1 shows the interface of our downloadable AbCD (our web-based interface is similar), displaying results for the design with $n = 85$ (note that it is not one of the 15 values directly evaluated through actual simulations), $d = 2$ and ethnicity = EU (Europeans). We can see that rarer variants are more challenging for detection. For example, this particular design can only detect 17.68% rare variants with MAF 0.5–1%, but it can discover essentially all (99.35%) common variants with MAF of >5%. Note that the detection power is upper bounded by the percentage of variants in the MAF category that segregate (i.e. polymorphic) among the sequenced samples, which are 83.94% (392/467) and 100%, respectively. In addition, information content as measured by r^2 and effective sample size (nr^2) also tends to decrease with decreasing MAF. For example, although 85 individuals are sequenced in this design, the effective sample size is only 46 for the 17.68% discovered rare variants with MAF of 0.5–1%, but it is much higher (75) close to the sequenced sample size 85 for common variants with MAF of >5%.

2.2 ShotGun and DesignPlanner

ShotGun is a short-read simulator that can (i) *allow cycle-specific sequencing error rates*. It is known that MPS data quality deteriorates with the biochemical reactions, such that nucleotides called from cycles towards the end are usually of inferior quality (Bansal *et al.*, 2010). Therefore, it is important to allow cycle-specific error rates to reflect this feature; (ii) handle o*ver dispersed read depth distribution*. Ideally, the read depths follow a Poisson distribution when the starting positions of sequence reads follow a discrete uniform distribution in the targeted region. However, stochastic and experimental limitations result in over dispersed read depths across bases. Following the study conducted by Sampson *et al.* (2011), we use a Gamma mixture of Poisson's to account for the overdispersion. The level of overdispersion is controlled by a shape parameter that can be specified by the user; (iii) *control false-positive rate*. When simulating a large number of individuals in a long region, it is desirable to filter out base pairs that are unlikely to be polymorphic. We apply a multi-sample single-site statistic $\sum_i[MI_i(MI_i+1)]$, where the sum is over all individuals, and MI_i is the minor allele count of the ith individual (Li *et al.*, 2010). We use simulations to determine the empirical threshold for this statistic at any user-specified false-positive rate and filter out bases unlikely to be polymorphic accordingly.

DesignPlanner is a full pipeline from sequencing data simulation using ShotGun, to LD-based genotype calling, to effective sample size estimation in different MAF categories in DesignPlanner and to facilitate the design of MPS-based studies not covered in AbCD.

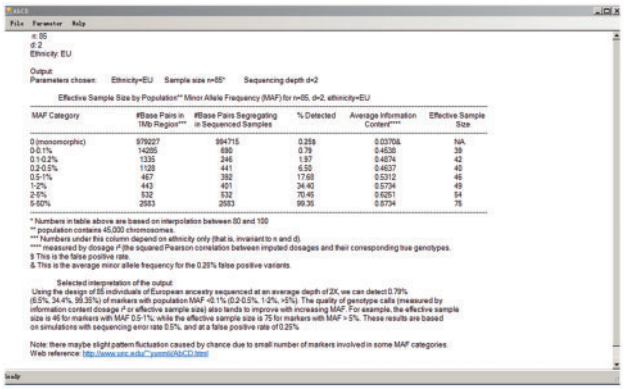

Fig. 1. AbCD downloadable terminal interface: screenshot for the designs with $n = 85$, $d = 2$ and ethnicity = EU (Europeans)

3 CONCLUSIONS

We present AbCD and two associated software tools (ShotGun and DesignPlanner) to facilitate the design of arbitrary coverage sequencing-based studies. Our software, tutorial and simulated datasets are available at http://www.unc.edu/~yunmli/AbCD.html.

ACKNOWLEDGEMENT

The authors thank Dr Mingyao Li and Dr Wei Chen for reviewing early versions of this manuscript.

Funding: NIH (R01-HG006292 and R01-HG006703 to Y.L., HG007022 and HG005581 to G.R.A.).

Conflict of Interest: none declared.

REFERENCES

Ansorge,W.J. (2009) Next-generation DNA sequencing techniques. *N. Biotechnol.*, **25**, 195–203.

Bamshad,M.J. *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.

Bansal,V. *et al.* (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.*, **20**, 537–545.

Coventry,A. *et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.*, **1**, 131.

Flannick,J. *et al.* (2012) Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS Comput. Biol.*, **8**, e1002604.

Ionita-Laza,I. *et al.* (2011) Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am. J. Hum. Genet.*, **89**, 701–712.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Li,Y. *et al.* (2011) Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, **21**, 940–951.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Nelson,M.R. *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**, 100–104.
Ng,S.B. *et al.* (2010) Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.*, **42**, 30–35.
Pasaniuc,B. *et al.* (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.*, **44**, 631–635.
Pritchard,J.K. and Przeworski,M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
Sampson,J. *et al.* (2011) Efficient study design for next generation sequencing. *Genet. Epidemiol.*, **35**, 269–277.
Schaffner,S.F. *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–1583.