

MEETING ABSTRACT

Open Access

Power and sample size of two-stage extreme phenotype sequencing design for next generation sequencing studies

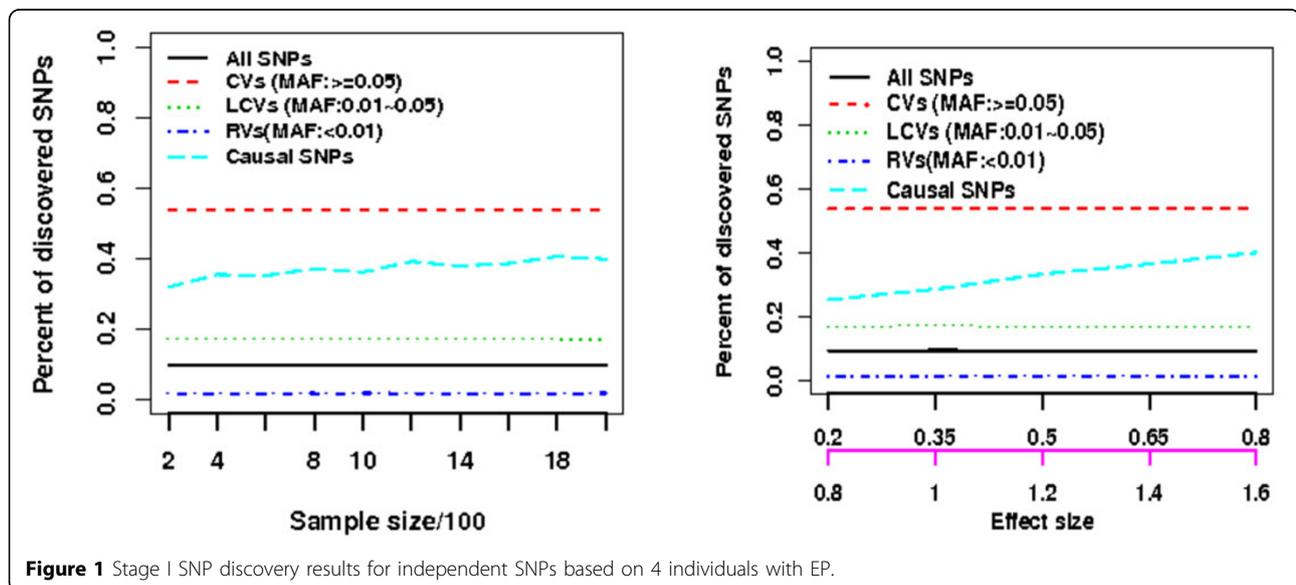
Guolian Kang

From 12th Annual UT-ORNL-KBRIN Bioinformatics Summit 2013
Buchanan, TN, USA. 22-24 March 2013

Background

Next-generation sequencing technology is changing the genomic research due to its huge sequencing capacity which is used to identify rare susceptibility variants that affect complex diseases. Because of its high cost, a two-stage extreme phenotype sequencing (TS-EPS) design is an alternative approach [1] in which the genetic variants are discovered by whole-genome or whole-exome sequencing individuals with extreme phenotypes in stage I and

the association variants are detected by sequencing a large number of individuals on the discovered variants in stage II. TS-EPS can efficiently discover more than half of the causal variants using about 0.2% of all individuals [2] and can therefore have higher power than random sampling given the sample size and effect sizes of the causal variants [2,3]. Using simulated data for unrelated individuals, we further explore the efficiency of TS-ESP in term of different sample sizes and varying effect sizes.



Correspondence: Guolian.Kang@stjude.org
Department of Biostatistics, St. Jude Children's Research Hospital, Memphis,
TN 38105, USA

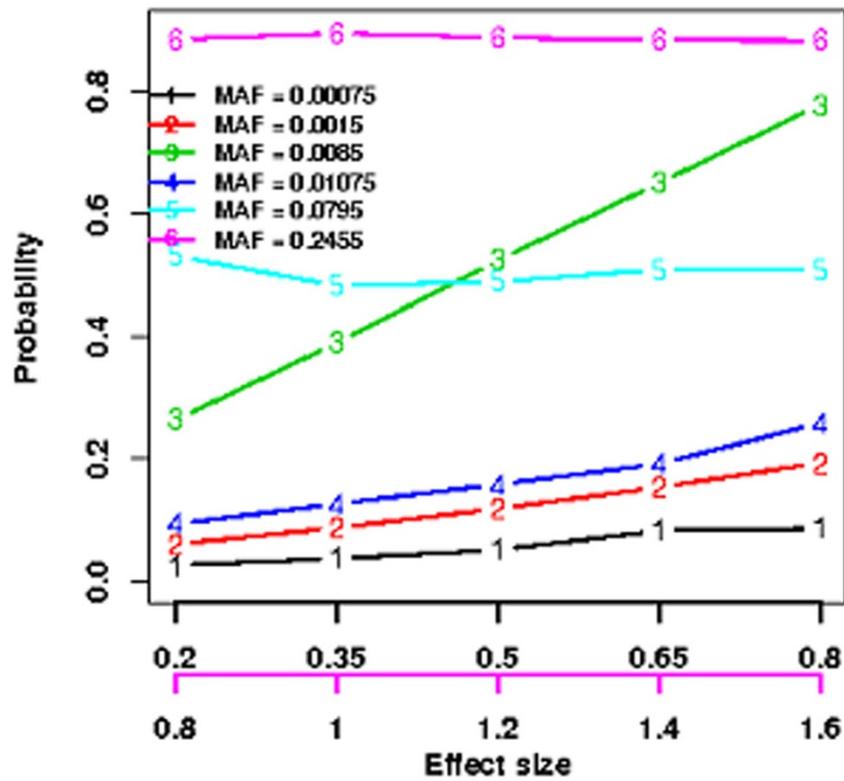


Figure 2 Enrichment of causal SNPs in 4 individuals with EP.

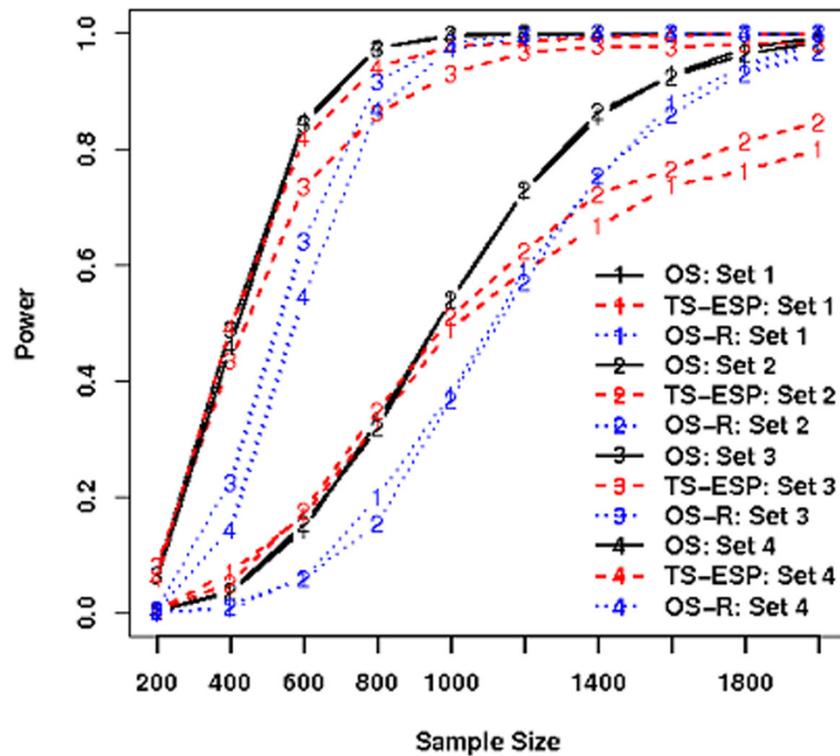


Figure 3 Power of TS-EPs for independent SNPs based on 4 individuals at a nominal level of 5×10^{-5} . OS: one-stage design; TS-ESP: two-stage extreme phenotype sequencing design; OS-R: one-stage design exclude 4 individuals.

Results

When four individuals with the first four most extreme trait values are sequenced in stage I, we found that 1) TS-EPS can discover the constant numbers of CVs, LCVs, and RVs regardless of the sample size and effect size, however the increasing numbers of causal variants with increasing sample size and effect size (Figure 1); 2) the probability of discovering a causal CV is constant regardless of its effect sizes but the quantity depends on its minor allele frequency; however, the probability of discovering RVs is a complex function of their effect sizes and minor allele frequencies given disease model and sample size (Figure 2). 3) Therefore, using an optimal unified association test for gene-based association analyses, the power of TS-ESP is comparable to one-stage (OS) design in which all individuals are sequenced for association testing if the rare causal variants have large effect sizes (Figure 3).

Published: 22 October 2013

References

1. Emond MJ, Louie T, Emberson J, Zhao W, Mathias RA, Knowles MR, Wright FA, Rieder MJ, Tabor HK, Nickerson DA, Barnes KC, National Heart, Lung, and Blood Institute (NHLBI) Go Exome Sequencing Project, Lung GO, Gibson RL, Bamshad MJ: **Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis.** *Nature Genet* 2012, **44**(8):886-889.
2. Kang G, Lin D, Hakonarson H, Chen J: **Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power.** *Hum Hered* 2012, **73**(3):139-147.
3. Barbett JJ, Lee S, Lin S: **Detecting rare variant effects using extreme phenotype sampling in sequencing association studies.** *Genet Epidemiol* 2013, **37**(2):142-151.

doi:10.1186/1471-2105-14-S17-A16

Cite this article as: Kang: Power and sample size of two-stage extreme phenotype sequencing design for next generation sequencing studies. *BMC Bioinformatics* 2013 **14**(Suppl 17):A16.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

