J Mol Neurosci (2015) 56:529–530
DOI 10.1007/s12031-015-0557-y

Improving Significance in Association Studies: a New Perspective for Association Studies Submitted to the Journal of Molecular Neuroscience

G. Nicolas[1,2,3] · C. Charbonnier[3] · J. R. M. Oliveira[4]

Received: 23 December 2014 / Accepted: 26 March 2015 / Published online: 14 April 2015

Dear Editor,

The Journal of Molecular Neuroscience has a long tradition of accepting various formats of studies concerning neurosciences, with a great variety of methodological approaches. Amongst them, we find case-control studies to be one of the most frequently occurring formats.

Case-control studies are appropriate to evaluate the influence of genetic variants, often linked to complex disorders. A significant example is the positive association between the ε4 *APOE* allele and Alzheimer's disease, widely replicated since the beginning of the 90s (Ridge et al 2013).

However, we want to keep supporting such studies but with a more stringent selection. This will avoid redundant attempts to analyze complex disorders with unsuitable approaches such as limited number of SNPs and/or small populations.

Basic recommendations include:

> To work as part of a major consortium is a good strategy when limited sample is available and also analyze panels of SNPs including the largest number possible. Studies highlighting few SNPs should not be considered for review, considering the current level of knowledge appointing complex disorders as influenced by hundreds of SNPs, besides environmental variables. In general terms, one should feel that such potential new contributions submitted are actually affecting the field and just adding more ambiguous analysis, without moving the topic forward.

We often receive articles assuming significant associations, after extensive significance tests have been done in which the total patient's group is divided by gender, age of onset, and other features. In this case, the chance of false positives is high and the final results bring a natural feeling of skepticism. We should not proceed with further review in such cases, unless the sub-groups are large enough and include additional data, especially biological markers, which allow the authors to nail actual endophenotypes.

The landscape of case–control association studies, comparing frequencies of SNPs between cases and controls, is evolving from genotyping data to large-scale sequencing. The golden years of genotyping might be behind us, but it is not yet time to turn the page. In particular, genome-wide association studies (GWAS) allow for cartography of frequent SNPs genotyped in an SNP array. These have been widely applied to numerous neuropsychiatric diseases (e.g., Ferreira et al. 2008; Lambert et al. 2013). Issues and clues have been well-established, the most important ones being: (1) power analysis to confirm sample sizes are adequate to detect expected odds ratios, (2) thorough technical quality control, including Hardy–Weinberg equilibrium check, to reduce the risk of genotyping artifacts, (3) adjustment for major epidemiological biases such as population stratification and relatedness, and (4) multiple testing correction of *p*-values.

When targeted genotyping of a limited number of SNPs is used, either in a replication context or in the exploration of candidate genes, strong support for SNP and gene selection must be provided. The same rules apply to

✉ J. R. M. Oliveira
joao.ricardo@ufpe.br

[1] Inserm U1079, IRIB, University of Rouen, Rouen, France

[2] Department of Genetics, Rouen University Hospital, Rouen, France

[3] CNR-MAJ, Rouen, France

[4] Keizo Asami Laboratory and Neuropsychiatry Department, Federal University of Pernambuco, Av. Prof Moraes Rego, 1235, Cidade Universitária, 50670-901 Recife, PE, Brazil

targeted studies as well as to GWAS. However, as population stratification cannot be checked, cases and controls must be clearly matched for ancestry beforehand. The main limitation of GWAS is the restriction to frequent variants, the vast majority of them being intronic or intergenic. Only a few published associated SNPs might have a biological effect by themselves (e.g., the SNPs determining the *APOE* genotypes in Alzheimer's disease). It is largely assumed that most of the associated SNPs are in linkage disequilibrium with functional variants, e.g., ungenotyped non-synonymous exonic.

The advent of massive parallel sequencing (next generation sequencing, NGS) now gives access to nearly all coding and non-coding variants, including rare and private (new) ones, within the whole genome, the whole exome, or a given gene or panel of genes. NGS unveiled the unexpectedly large amount of rare and private variants, which were missed by SNP arrays.

Association studies based on sequencing results are now rising. However, sequencing is not genotyping. There is a crucial need to tackle related new issues: (1) confidence in variant calls, (2) prioritizing strategies for the huge amount of rare variants detected, and (3) statistical significance of rare/private variants. First, stringent quality criteria should be added to the QC check during the bioinformatics pipeline and after variant calling. It is also imperative that all samples should be called simultaneously, so that genotype or missingness information can be gathered across all samples for every identified variant in the whole dataset. The aim of this step is to make sure that information is truly available and of sufficient quality for each considered genomic position, bringing sequencing data closer to genotyping, i.e., is the variant present, absent, or missing (insufficient quality)? Second, filters used to discard the flow of probable neutral variants must be clearly justified and stated. Third, one of the main new issues remains the statistical analysis itself.

Accounting for population stratification is of tremendous importance, although the adequacy of usual methods like principal component analysis (PCA) is still under scrutiny (Liu et al. 2013). Besides, the question of which tests should be applied to rare variants at the variant, gene, and network/pathway levels is still under debate. Several tools have been published and used, such as burden tests (CMC, CAST, WSS, VT, KBAC), variance tests (SKAT, C-alpha) or combined (SKAT-O, MiST) (Lee et al. 2014). As this field is still under development, we invite authors to keep an open mind on new methods.

In parallel, exome arrays have been designed to genotype a defined number of common, low frequency and rare variants, including coding ones. Therefore, this does not allow for the discovery of new variants and gives SNP by SNP data for each individual genotyped, providing a still largely incomplete cartography of the genome.

In conclusion, NGS brings about a double change of paradigm. First, most variants are no longer given but revealed in the sample. The change in technology (genotyping to sequencing) should be accompanied by modified study designs: the discovery of variants should not be limited to cases and then genotyped in controls but also sequenced in controls, comparing the burden of variants in both groups (Li and Leal 2009).

Indeed, if one decides to look at the entire coding region of interest in cases, one should also do this in controls. The holes of the Swiss cheese might indeed contradict the initial plans and re-equilibrate the balance. Second, sequencing points out variants with a putative biological effect. Functional assessment of variants should therefore be encouraged to validate statistical results.

We assume that such recommendations will benefit not only the quality of manuscripts submitted, but also the personal standards of research groups devoted to neuropsychiatric disorders with both genetics and environmental triggers.

References

Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan J, Kirov G, Perlis RH, Green EK, Smoller JW, Grozeva D et al (2008) Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 40(9):1056–1058

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thorton-Wells TA et al (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 45(12):1452–1458

Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95(1): 5–23

Li B, Leal SM (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet 5(5):e1000481

Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, Bodea CA, Muzny D, Reid JG, Banks E, Coon H, Depristo M et al (2013) Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet 9(4): e1003443

Ridge PG, Mukherjee S, Crane PK, Kauwe JS (2013) Alzheimer's disease: analyzing the missing heritability. Alzheimers Dis Genet Consortium 8(11):e79771