 doi:10.20944/preprints202108.0084.v1

Understanding the function of a locus using the knowledge available at single-nucleotide polymorphisms

Majid Nikpay[1*], Sepehr Ravati[2], Robert Dent[3], Ruth McPherson[1,4*]

[1]Ruddy Canadian Cardiovascular Genetics Centre, University of Ottawa Heart Institute, Ottawa, Canada

[2]Plastenor Technologies Company, Montreal, Canada

[3]Department of Medicine, Division of Endocrinology, University of Ottawa, the Ottawa Hospital, Ottawa, Canada

[4]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada

[*] **Correspondence:**

Dr. Majid Nikpay, University of Ottawa Heart Institute, 40 Ruskin St – H4208, Ottawa, Canada K1Y 4W7 *mnikpay@ottawaheart.ca*

Dr. Ruth McPherson, University of Ottawa Heart Institute, 40 Ruskin St – H4203, Ottawa, Canada K1Y 4W7 *rmcpherson@ottawaheart.ca*

 doi:10.20944/preprints202108.0084.v1

Abstract:

Understanding the function of a locus is a challenge in molecular biology. Although numerous molecular data have been generated in the last decades, it remains difficult to grasp, how these data are related at a locus? In this study, we describe an analytical workflow that can solve this problem using the knowledge available at single-nucleotide polymorphisms (SNPs) level. The underlying algorithm uses SNPs as connectors to link omics data and identify correlation between them through a joint bioinformatical/statistical approach. We describe its application in finding the mechanism whereby a mutation causes a phenotype and in revealing the path whereby a gene is being regulated and impacts the phenotypes. We translated our workflow into freely available shell scripts that carry out the analyses. Our approach provides a basic framework to solve the information overload problem in biology.

Keywords: Annotation; SNPs; Rare variants; Mendelian randomization; Algorithm

 doi:10.20944/preprints202108.0084.v1

Introduction:

Over the past years, high throughput studies have generated omics data for various biological entities including functional elements and phenotypes. These data are usually generated, analyzed and published separately due to logistical/technical restrains. Genome browsers/databases then add these data to their repertoire and make them available as annotation tracks; however, this has created an information overload problem. For example, a biologist that wants to know the function of a locus starts the task by looking at a stack of annotation tracks provided by a genome browser and investigates the links between them visually. However, this is a tedious task and often not fulfilling. There are also situations where a researcher wants to relate two datasets (e.g. gene expressions with epigenome data) but finds the task cumbersome.

In this study, we describe a workflow that addresses these issues and provides a statistically quantified report for the user. Our approach relies on summary association statistics from GWAS studies of molecular features and phenome. It uses the common identifier (SNPs) between them to combine, identify correlation between their components, and provide a summary of functional elements and their relations at a locus. We demonstrate its application in understanding the mechanism whereby a rare variant causes a disease and in revealing the path whereby a locus is being regulated and impacts the phenotypes. We translated our workflow into shell scripts that are publically available.

Methods:

Advances in sequencing technology allow us to capture rare variants efficiently; however, often it remains elusive among the identified variants which variant (locus) causes the phenotype. Furthermore, it is important to find the path whereby a rare variant impacts a phenotype for therapeutic purposes. From the biological perspective, a rare variant is expected to cause a phenotype by disrupting a functional feature/element, i.e.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

Rare variant → Functional element → Phenotype

Therefore, if we detect the above condition is true for a variant:

i) We identified the variant that causes the phenotype

ii) We identified the underlying mechanism

But how can we relate a rare variant to a functional element and the functional element to the phenotype? The answer is the SNPs. Traditionally SNPs have been used to study both the genetics of phenotypes and the genetics of functional elements. Therefore, if we can find a set of SNPs in a region that meet the following criteria, we have fulfilled our aims (i & ii).

A. The location of rare variant is within the coordinates of the SNP set

B. The SNP set is associated with the functional element and the phenotype

C. The functional element causes the phenotype

Based on these criteria, we devised an analytical pipeline (**Figure 1**). Below, we describe each step and the reasoning behind it:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

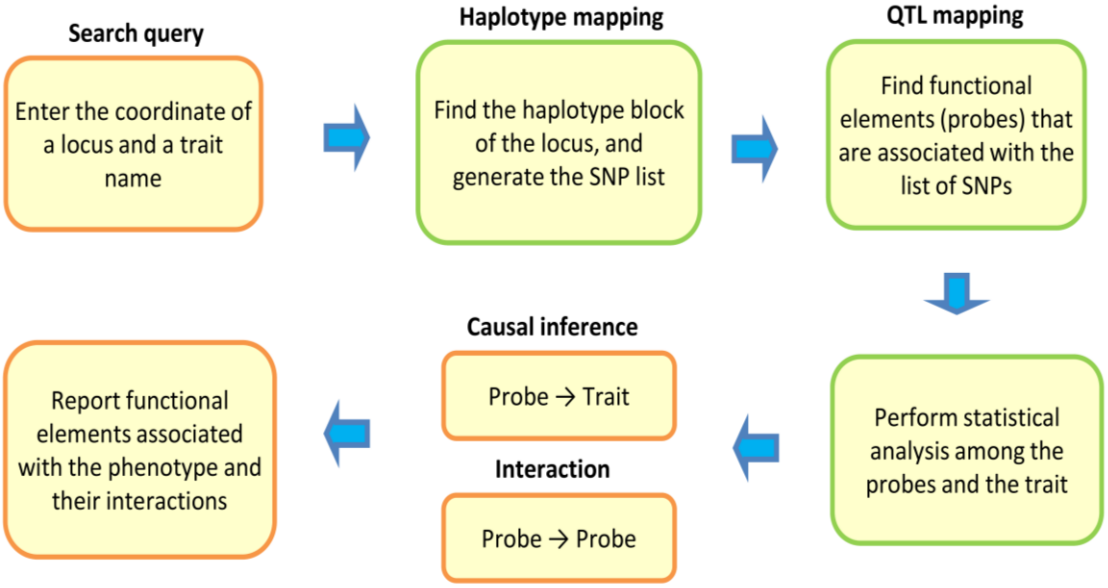

Figure 1. The design of our workflow to understand the function of a locus and its impact on a phenotype. The search starts by finding the haplotype block that the queried locus resides within it. Our algorithm then uses SNPs in the haplotype block to find functional elements (probes) associated with them (P<5e-8). The next step is to test the association between the probes and the trait using Mendelian randomization (MR). Our algorithm also uses MR to test whether the identified probes act together (interaction) or independently. The final step is to generate a report that summarizes the nature and the magnitude of association between each probe and the trait as well as between the probes. In Figure 2 and 3, we provided examples and interpreted the meaning of statistics. Detailed description of each step is provided in the Methods section.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

Haplotype mapping: We start by finding the haplotype block that the queried mutation resides within it. After finding the haplotype block, our algorithm retrieves the list of SNPs within the block for QTL mapping. A haplotype block is a genomic region whereby SNPs within it are in linkage disequilibrium (LD). Haplotype mapping has been traditionally used to identify genes for Mendelian diseases. A mutation that arises in a block is expected to cause the disease by impacting the function of the block. Therefore, haplotype mapping is a viable step towards narrowing our search space for functional elements.

We estimated the boundaries of human haplotypes using PLINK (v 1.9)[1], based on definition of blocks suggested by Gabriel et al.[2] and using European sample (n = 503) of the 1000 genomes project. The reason was because as compare to African and Asian populations, Europeans have larger haplotype blocks; furthermore, majority of existing GWAS data come from studies conducted in European populations and choosing a different population could cause downstream issues such as population stratification. Nonetheless, our pipeline is flexible and if a researcher wants to use a different definition of a block or uses data another population it can be done by simply replacing the input file for haplotype blocks by another file.

It is also important to note that not all human genome is within haplotype blocks; therefore, in situations where a variant cannot be assigned to a haplotype block, our algorithm searches for nearby SNPs that are within 5 Kbp of the mutation and if the search result is null it increments its search window by 10 Kbp and stops until it reaches the search window of 45 Kbp (the largest size of an average haplotype block across human populations).[2]

QTL mapping: List of SNPs obtained from the previous step then is used to identify functional elements (probes) associated (P<5e-8) with them. We use the SMR software (version 1.03)[3] for QTL mapping. It reduces the file size by keeping QTL data in a binary format; moreover, it provides

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

flexible options to query the data. QTL data can be obtained from previous studies; we also provided a collection of QTL data (see the Data/Code availability section).

Statistical analysis: The aim of this stage is to test if any of functional elements obtained from the previous step is associated with the studied phenotype. For this purpose, we used Mendelian randomization that can infer causality between a functional element and a trait. QTLs included in the instrument for the MR must be non-pleiotropic ($P<0.01$), be associated with the functional element ($P<5e-8$) and be in linkage equilibrium ($r^2<0.05$). MR analysis was done using the GSMR algorithm implemented in GCTA software (version 1.92)[4]. In this stage, if we find a significant association between a functional element and a trait, we can conclude our queried variant impacts the trait through the functional element.

Test of interaction: The purpose of this step is to obtain additional functional insight by testing whether the identified probes contribute to the trait independently or not. The underlying script uses MR to carry out pairwise interaction analysis between the probes; however, in addition to causality (Probe A → Probe B) it also examines the presence of pleiotropy (Probe A ← SNPs → Probe B) between the probes by keeping the pleiotropic SNPs in the instrument. This helps to identify probes that are under the regulatory impact of the same set of SNPs.

In the following section, we further describe our workflow through examples of well-studied loci.

Results

PCSK1

PCSK1 was one of the first genes linked to monogenic early-onset obesity. It encodes the enzyme proprotein convertase 1. Substrates of PCSK1 enzyme such as POMC, insulin, NPY, ghrelin and GLP-1 are involved in the regulation of energy homeostasis and food behaviour. Patients with mutations in *PCSK1* develop a profound appetite that results in significant weight gain and eventually obesity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

early in life.[5] *PCSK1* is located in chromosome 5q15 and rare variants within this region are reported for obesity in ClinVar db (**Table S1**). In this study, we tested whether our algorithm can link the rare variants to this gene and this gene to obesity. We passed the location of a rare variant within this locus and the name of relevant phenotype (BMI) to the wrapper script and executed it as:

bash wrapper.sh chr5:95734724 BMI_PMID30239722

The first argument represents the coordinate of the variant in human reference genome (build GRCh37). The second argument indicates the phenotype name and its study identifier (represented by PMID). Our algorithm performed haplotype and QTL mapping and retrieved the list of biomarkers (functional elements) tagged by the variant. It then tested the association between the biomarkers and BMI using MR and reported the result. The output indicated PCSK1.13388.57.3 which is a biomarker that measures the level of PCSK1 in the blood is associated with obesity (B=-0.02, P=4e-19). This finding indicates the queried rare variant is among pQTLs for PCSK1 (**Figure 2a**) and as such, contributes to obesity by lowering the level of this protein in the blood (**Figure 2b**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

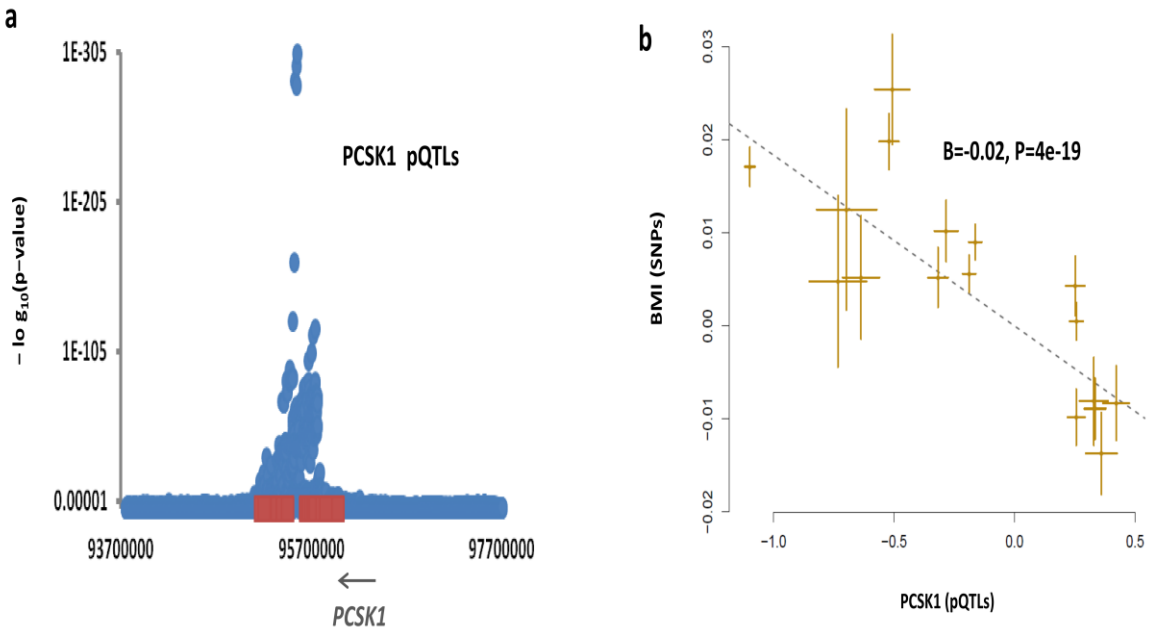

Figure 2. Impact of rare variants within 5q15 on obesity

Rare variants within 5q15 chromosome band are reported for obesity in ClinVar database. In this region PCSK1 deficiency is known to cause obesity. We investigate whether our workflow can link these variants to *PCSK1* and PCSK1 protein level to obesity. The haplotype mapping indicates the reported variants are within coordinate of pQTLs of PCSK1 (panel **a**). MR analysis indicates subjects that are genetically susceptible to have lower level of PCSK1 in blood tend to have higher risk of obesity (panel **b**). MR was done using non-pleiotropic SNPs (P>0.01) that are independently (r^2<0.05) and significantly associated (P<$5e^{-8}$) with PCSK1 protein level. Each point on the MR plot represents a SNP, the x-value of a SNP is its beta effect size on PCSK1 protein level and the horizontal error bar, represents the standard error around the beta. The y-value of the SNP is its beta effect size on BMI and the vertical error bar represents the standard error around its beta. The dashed line represents the line of best fit (a line with the intercept of 0 and the slope of β from the MR test). pQTLs summary statistics (Beta and SE) were obtained from PMID: 29875488. For BMI, we obtained these data from PMID: 30239722.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

APOE

APOE is involved in the metabolism of lipids and mutations in this gene cause abnormality of lipids. We tested whether our approach can link this gene to lipid phenotypes? For this purpose, we passed the genomic coordinates of the locus (chr19:45409039-45412650) and the name of the lipid phenotype, to the wrapper script to initiate the search. The result (**Figure 3**) indicates higher protein level of APOE in the blood is associated with higher LDL (B=0.8, P=$1.2e^{-83}$), total cholesterol (B=0.5, P=$1.3e^{-56}$) and lower HDL (B=-0.14, P=$9e^{-23}$). The report generated by the interaction test indicates APOE protein level is under the impact of cg13375295 site. Higher methylation at this site was associated (B=-3.5, P=$4.2e^{-9}$) with lower level of APOE in the blood (**Figure 3**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

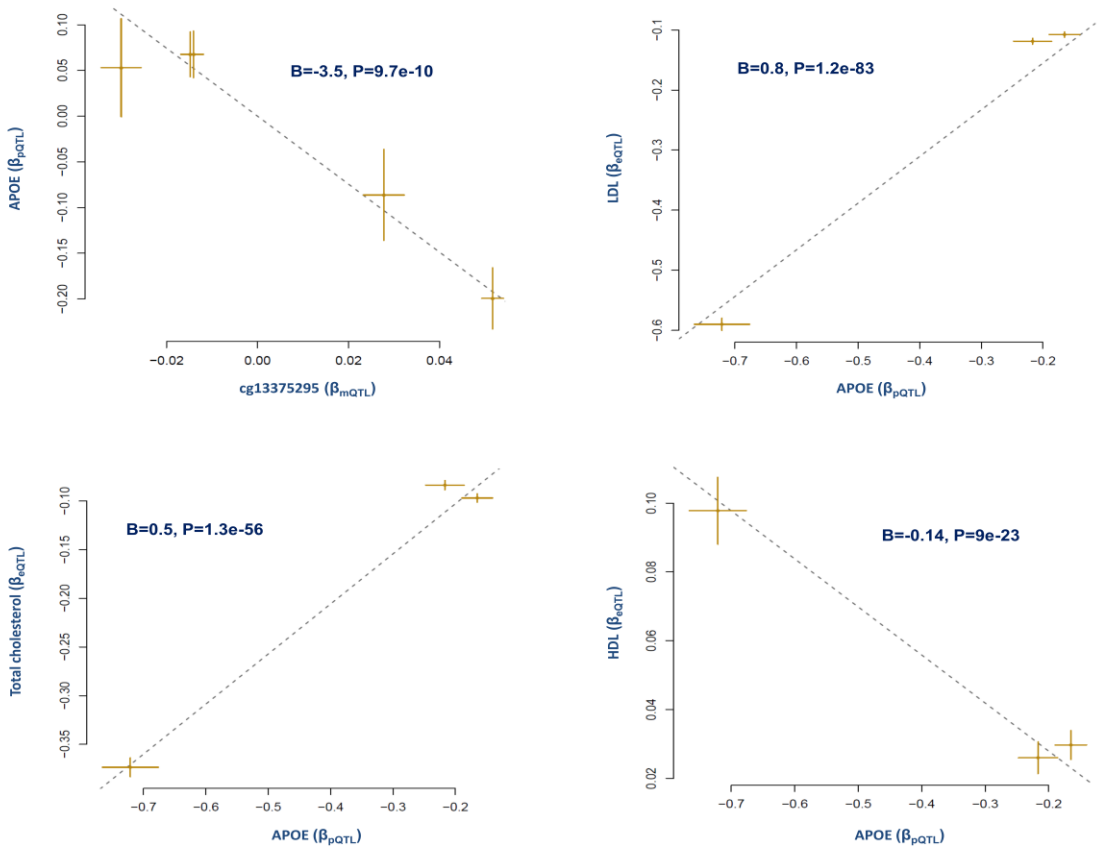

Figure 3. Impact of APOE blood level on lipids phenotype.

In this example, we tested whether our workflow can detect the association of APOE protein level with lipid phenotypes. The results indicate APOE level in blood is under the regulatory impact of cg13375295 site upstream of this gene. Furthermore, we detected higher level of APOE is associated with higher LDL, higher total cholesterol but lower HDL. Summary statistics provided in parentheses are from the Mendelian Randomization analysis and using non-pleiotropic SNPs (P>0.01) that are independently (r^2<0.05) and significantly associated (P<$5e^{-8}$) with the exposure. Each point on the plots represents a SNP, the x-value of a SNP is its beta effect size on the exposure (cg13375295) and the horizontal error bar, represents the standard error around the beta. The y-value of the SNP is its beta effect size on the outcome and the vertical error bar represents the standard error around its beta. The dashed line represents the line of best fit (a line with the intercept of 0 and the slope of β from the MR test). pQTLs summary statistics (Beta and SE) are from PMID: 29875488, mQTLs summary statistics are from PMID: 30401456 and BMI summary statistics are from PMID: 30239722.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 August 2021 doi:10.20944/preprints202108.0084.v1

Discussion

In this study we devised a workflow to understand the function of a locus using knowledge available at SNP level and we demonstrate its application through examples for PCSK1 and APOE locus. The underlying algorithm that carries out the task is written in shell scripting language. This allows the use of parallel computing and therefore the possibility to conduct screening at phenome/genome-wide scales. Considering the volume of existing and upcoming functional data, parallel computing will become a necessity to integrate/relate various layers of omics data in a time efficient manner.

In this study, we used Mendelian randomization to quantify association between two entities. MR allows incorporating summary association statistics from large GWAS consortia and therefore provides higher statistical power as compared to traditional association studies conducted in a sample of individuals. MR design also provides a shield against the confounding effect of environmental factors because it uses a set of independent SNPs (an instrument) to gauge the relationship between two entities and alleles of independent SNPs are allocated to offspring randomly. Therefore an instrument of SNPs is inherently immune to the confounding effect of known/unknown factors. Of note, the use of SNPs as an instrument also brings the caveat of weak instrument bias when testing the association between two entities that are highly polygenic (e.g. complex traits); however, this is a lesser issue for a functional element that is under the regulatory impact of fewer SNPs.

Our workflow relies on the power of big data to provide molecular insight; however, downloading and reformatting these data could be cumbersome for a user. This is a limitation of our approach. One solution would be to provide an online platform where data are collected and centralized and a user can obtain the results by simply entering a search term. Unfortunately, we did not receive support to set up such a platform; however, this is a direction of work for future studies that are

 doi:10.20944/preprints202108.0084.v1

interested in this subject. Another limitation of our approach is that QTL data are mainly available at transcriptome, proteome and methylome levels and to a lesser extant at other levels of functional annotations. Nonetheless, our approach is easily extendable to all types of functional data, if the practice of reporting the results with regard to their associations with SNPs becomes more frequent.

In summary, this study provides a solution to navigate through various layers of functional annotations in order to understand the function of a locus. We show its application in finding the mechanism whereby a mutation causes a disease and in revealing the path whereby a gene is being regulated and impacts the phenotypes. We provided freely available shell scripts that can carry out the tasks.

Author Contributions

Conceptualization, M.N., R.M.; formal analysis, M.N.; investigation, S.R.; resources, R.M. and R.D.; draft preparation, M.N.; editing, R.M. and S.R. All authors have read and agreed to the published version of the manuscript.

Funding

Supported by the Canadian Institutes of Health Research # FDN-154308 (RM).

Data/Code Availability

Data, instructions and shell scripts to carry out the analyses are available from:
https://github.com/mnikpay/locus-annotator.git

Acknowledgments

This research was enabled in part by computational resources and support provided by the Compute Ontario and the Compute Canada.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015, *4*, doi:10.1186/s13742-015-0047-8.

2. Gabriel, S.; Schaffner, S.; Nguyen, H.; Moore, J.; Roy, J.; Blumenstiel, B.; Higgins, J.; DeFelice, M.; Lochner, A.; Faggart, M.; et al. The Structure of Haplotype Blocks in the Human Genome. *Science (New York, N.Y.)* 2002, *296*, 2225-2229, doi:10.1126/science.1069424.

3. Zhu, Z.; Zhang, F.; Hu, H.; Bakshi, A.; Robinson, M.R.; Powell, J.E.; Montgomery, G.W.; Goddard, M.E.; Wray, N.R.; Visscher, P.M.; et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016, *48*, 481-487, doi:10.1038/ng.3538.

4. Zhu, Z.; Zheng, Z.; Zhang, F.; Wu, Y.; Trzaskowski, M.; Maier, R.; Robinson, M.R.; McGrath, J.J.; Visscher, P.M.; Wray, N.R.; et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature communications* 2018, *9*, 224, doi:10.1038/s41467-017-02317-2.

5. Ramos-Molina, B.; Martin, M.G.; Lindberg, I. PCSK1 Variants and Human Obesity. *Prog Mol Biol Transl Sci* 2016, *140*, 47-74, doi:10.1016/bs.pmbts.2015.12.001.