Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article DOI - 10.26479/2017.0302.06

SELECTING SINGLE-NUCLEOTIDE POLYMORPHISMS (SNPS) MARKERS FOR PHARMACOGENOMICS USE

N. Zaïd [1, 2, 3], Y. Souilmi[1, 4], N. El Bilali[2], N. El Kadmiri[5*], S. Amzazi[1, 3]

1 Biology of Human Pathologies Laboratory, Genomics of Human Pathologies Center, Faculty of Science, Mohammed V University, Rabat, Morocco.

2 Microbiology Immunology and Infectious Diseases Department, Montreal University, Montreal, Canada.

3 Biochemistry and Immunology Laboratory, Faculty of Science, Mohamed V University, Rabat, Morocco.

4 Australian Center for Ancient DNA, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia.

5 LBVE, Sciences and Techniques Department, Polydisciplinary Faculty of Taroudant , IBN ZOHR University, B.P: 271, 83 000, Taroudant, Morocco.

ABSTRACT: Genome-wide genetic association studies (GWAS) represent the analyses of several genetic variations in many individuals to study in order their correlations with phenotypic traits. These studies aim to identify genetic susceptibility factors of multifactorial diseases. The very large number of statistical tests performed requires considerable computing power and corrections on multiple tests, which reveal more the number of SNPs used in the study is higher; more the p-value should be smaller to be significant. According to the literature, it is necessary to have a high number of samples (up to 35,000) to perform a GWAS. If the number of samples in the study is reduced, there is less risk of finding a SNP with a significant p value. In this current study, we aim to make a selection of SNP markers from our cohort based on criteria described in the literature in order to retain only the SNPs likely to be used in the pharmacogenomics use. A comparative study between genotyping or enrichment platforms was subsequently done. The major goal of this study is to check whether our selection is covered by current commercially available chips. To validate this protocol, a GWAS study, with only 2084 samples, was made on all the SNPs retained. The result was compared to other GWAS study related to complete set of SNPs before the selection. We proof that we have made the right selection by finding a SNPs among the selected ones with a significant p value.

KEYWORDS: Pharmacogenomics; GWAS; SNPs; database.

***Corresponding Author: Prof. Nadia El Kadmiri** Ph.D.
LBVE, Sciences and Techniques Department, Polydisciplinary Faculty of Taroudant, IBN ZOHR
University, B.P: 271, 83 000, Taroudant, Morocco.
Email Address: elkadmiri1979@gmail.com

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are simple nucleotide variations occurring when a reference nucleotide is substituted with an alternative base at the same position without insertion. SNPs are very abundant in the human genome, with over 154 million SNPs reported on dbSNP (https://www.ncbi.nlm.nih.gov/dbvar/content/org_summary/), over 95,000 SNPs with well-established and validated medical effects (SNPedia), and 5,000 SNPs with pharmacogenomics associations on PharmGKB. They occur throughout the genome generating different alleles[1] . The sequence of all SNPs a given region of the genome, forms a haplotype. The total number of SNPs estimated at the beginning was around 300,000. This number reached 1.4 million in 2001[2] then 3.1 million in 2007[3] to reach over 6.4 Million SNPs stored on the HapMap project database currently. SNPs represent more than 90% of human genetic polymorphism[4,5] and we found one SNP every 500 to 1000 bp. To make a Genome Wide Association Study (GWAS), a large representative number of individuals is needed for both cases and controls. If necessary, the study should be carried out on a reduced number of SNPs in order to have a significant results. The aim of this work is to explore the possibility to reduce the number of SNPs markers according to criteria described in the literature to retain only a small number allowing to make an association study on a cohort of 2084 samples. For this, we performed an automated selection within the 518,066 SNPs of the database while passing through standardization steps by organizing the data in a standard format and unifying the complementation build by looking for the positions of the SNPs in the chromosomes as well as the flanking regions of each SNP, and eliminating duplicates for purification to avoid any kind of redundancy. The operation resulted in a significant reduction in SNPs from 518,066 to 37,167 SNPs which allowed us to do a GWAS study of the 2084 samples.

2. MATERIALS AND METHODS

Genome-wide genetic association studies (GWAS) represent the analyzes of several genetic variations in many individuals in order to study their correlations with phenotypic traits. These studies aim to identify genetic susceptibility factors of multifactorial diseases. The very large number of statistical tests performed requires considerable computing power and corrections on multiple tests which mean more the number of SNPs used in the study is higher, more the p-value should be smaller to be significant.

$$P \quad value < \frac{0.05}{number \quad of \quad SNPs}$$

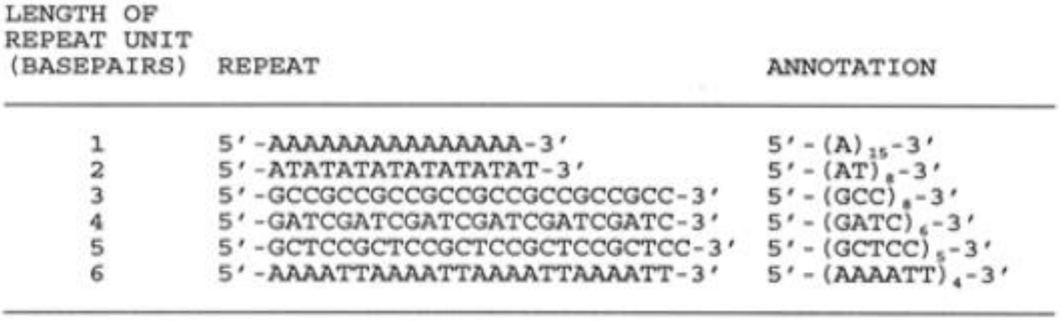
According to the literature, it is necessary to have a high number of samples (up to 35,000) to make a GWAS. If the number of samples in the study is reduced, there is less risk to find a SNP with a significant p value. Some questions are necessary to ask: Is there any way to carry out an association study with a small number of samples? Is it really necessary to recruit more patients to increase the number of samples, or finding an alternative solution by reducing the number of SNPs? And according to which criteria will we make our selection? In this work, we will study the possibility to perform a selection of SNP markers from our cohort based on criteria described in the literature in order to retain only the SNPs likely to be used in the pharmacogenomics use. To validate this protocol, a GWAS study, including 2084 samples, was made on all the SNPs retained. The result was compared to other GWAS study on SNPs complete set before the selection. We proof that we have realized the right selection by finding a SNPs among the selected ones with a significant p value. A comparative study between genotyping or enrichment platforms was subsequently done. The goal of this study is to check whether our selection is covered by current commercially available chips.

2.1. SNPs selection criteria

The choice of SNPs used in pharmacogenomics is based on a number of criteria [6], summarized as below:

- **Validity status:** to verify whether the SNP being processed is real and unique in the genome. The Validate SNP () boolean function ensures the validity of the pointed SNP by returning the value 'true' if it's the case. This information is available for each SNP element being processed under the 'snp_class' code. In the worst case, the program will stop the processing and switch to the next element;
- **Biallelic character:** to ensure the absence of mutations as well as a reliable allelic frequency. The Biallelique () function checks the number of alleles between the code 'variation' and it returns the Boolean value to 'True' if the SNP is biallelic and 'False' in the opposite case;
- **Primer test:** a good ratio of GC in the vicinity of the SNP studied; The percentage of **GC** represent the proportion of cytosine and guanine bases [7]. Primers should generally have a GC percentage between 45% and 60% to get a more specific PCR [8]. This is due to the fact that the G-C bond is constructed by three hydrogen bonds that requires a high temperature for denaturation, whereas the A-T bond is constructed by two hydrogen bonds.
- **Absence of polybases and Short Sequence Repeats (SSRs)**

Flanking sequences should not contain polybases sequences (sequences composed of identical consecutive bases) or microsatellites, also called "Simple Sequences" or "Short Sequence Repeats" (SSRs) composed of a k pattern repeated n times (figure1). The k pattern length varies from 1 to 6 bases repeated in tandem, and the different categories are designated according to the number of contiguous bases constituting k [9]. A proper detection of SNPs should be performed on the molecular level.



LENGTH OF REPEAT UNIT (BASEPAIRS)	REPEAT	ANNOTATION
1	5'-AAAAAAAAAAAAAAA-3'	5'-$(A)_{15}$-3'
2	5'-ATATATATATATATAT-3'	5'-$(AT)_{8}$-3'
3	5'-GCCGCCGCCGCCGCCGCCGCCGCC-3'	5'-$(GCC)_{8}$-3'
4	5'-GATCGATCGATCGATCGATCGATC-3'	5'-$(GATC)_{6}$-3'
5	5'-GCTCCGCTCCGCTCCGCTCCGCTCC-3'	5'-$(GCTCC)_{5}$-3'
6	5'-AAAATTAAAATTAAAATTAAAATT-3'	5'-$(AAAATT)_{4}$-3'

Figure 1: Diagram of microsatellites. Examples of homogeneous simple sequence motifs consisting of repeating units varying from 1 to 6 nucleotides in length[10].

The existence of microsatellites and polybases neighbouring a SNP creates interferences making its detection on the molecular level more difficult, hence the uselessness of studying polymorphism. Mono-nucleotide (N) type microsatellites have at least 10 repeat units, 6 units for di-nucleotides (NN), tri- (NNN), tetra- (NNNN) and hexa-nucleotides (NNNNN)[11].

- **Absence of secondary structures**

Flanking sequences should not have a secondary structures such as hairpin structures formed by GGCCs or GCGCs [12].

- **Absence de clustering**

Flanking sequence should not contain additional SNPs or substitution sites neighbouring the SNP. Only SNPs that are spaced with than 1000 Kb (1 Kb) are kept. Cluster elimination involves two steps. (i) Creating an index that contains the LSID (SNP identification number) of all the SNPs and their positions during the parsing of the document. (ii) Evaluating the distance between each SNP and its neighbours by insertion of all the SNPs which were validated by tests mentioned above in the database.

- **Polymorphic content:** a good discrimination index (Dp) as well as a low linkage imbalance (LD);

$$\boldsymbol{D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{s} x_j(x_j - 1)}$$

The polymorphic content is measured by the discrimination index. 'N' is the total number of strains in the population (sample), 's' is the total number of types described, and 'x_j' is the number of strains belonging to the 'j^th^ ' type [13].

These different filters have been implemented in the form of seven functions executed sequentially, ie. If the SNP fails in the test, it will be rejected and the program will be switched directly to the next element.

3. RESULTS AND DISCUSSION

Initially, JAVA program imports XML files, processes them by browsing SNPs one by one, by applying the filters, and then inserts only those retained in an Oracle database. This first part of the work is crowned by the creation of a WEB application accessible via a website allowing the exploitation of this data. The diagram below summarizes the logic followed (figure 2). The second part consists on conducting association studies with all the SNPs selected with a reduced number of samples (2,000). This will allow to check if there are SNPs markers whose p value is significant that could not be detected by studying all the starting SNPs.

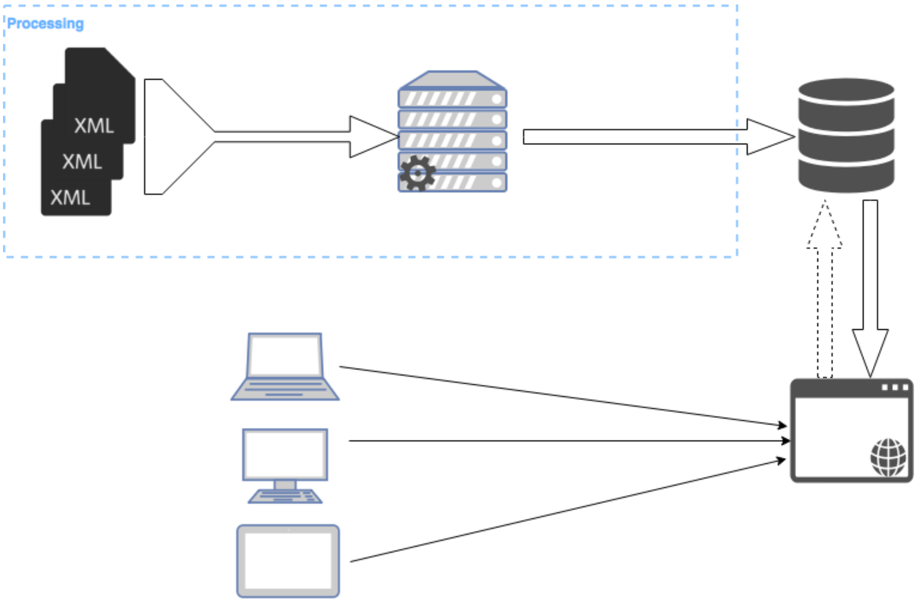


Figure 2: Diagram of the proceeding approach

A comparison between the genotyping and enrichment platforms will be made. The purpose of this step is to know which between current marketed chips provide a good coverage for our batch of SNPs after selection. We started with a database of 518.066 SNPs markers that have been selected. The result highlights a significant reduction of SNPs, which drop to 37.167 only.

New p value:

$$p \;\; value < \frac{0.05}{37294} = 1.34e^{-6}$$

Instead of:

$$p \;\; value < \frac{0.05}{518066} = 9.65e^{-8}$$

In order to simplify the study, only SNPs with a significant GWAS *p-value* were selected. We performed an intelligence quotient (IQ) study using the 2084-cohort, and only 9 SNPs conceded the stringent p-value filtration (Table 1).

Table 1: GWAS IQ Result.

CHR	SNP	UNADJ	GC	BONF
10	**rs2671709**	1.229e-08	2.613e-06	**0.000434**
21	**rs11910387**	4.435e-07	3.119e-05	**0.01566**
6	**rs9373437**	4.907e-07	3.345e-05	**0.01733**
3	**rs1512044**	1.064e-06	5.714e-05	**0.03756**
10	**rs10999632**	1.186e-06	6.161e-05	**0.04187**
14	**rs4982394**	1.281e-06	6.499e-05	**0.04523**
2	**rs7426114**	2.15e-06	9.306e-05	0.07594
18	**rs2938033**	2.921e-06	0.0001151	0.1032
20	**rs6048743**	3.999e-06	0.0001431	0.1412

Among the six SNPs identified as associated with the IQ, only one SNP (rs2671709) have been captured by studding 518.066 SNPs. Subsequently, we compared the two enrichment chips, HaloPlex and SureSelect provided from the Agilent platform. The percentage of coverage via these chips is theoretical which means that it is based on baits only. The covers of our selection of SNPs for these two chips were 92% and 65% for Haloplex and SureSelect respectively. This suggest that Haloplex can cover the SNPs selected significantly. However, this coverage is theoretical and it would be desired to recalculate the real data that will allow us to have a practice coverage depending on the plate depth. In this study, we aimed to find a new method for association studies with a reduced number of samples. Indeed, we developed an algorithm capable of screening large number of SNPs in order to retain only small number that could be used in pharmacogenomics. The selection of the SNPs was conducted on the basis of previously detailed criteria (criteria qualified as filters). Our data highlights a significant reduction of SNPs number from 518,066 to 37,167, with a filtration ratio of 8%. The protocol was validated by the results of the GWAS study which confirmed that it was impossible to detect the marker signal without reducing their number, as well as a comparative study between the current genotyping and enrichment chips.

We conclude that the SNPs selected according to the six criteria (filters) selected are needed to be used in pharmacogenomics research with significant and reliable manner, which have never been done before. However, it would be preferred to add a seventh, namely "the discrimination index" whose implementation will certainly bring more acutely to our results. The comparison between the two enrichment chips, HaloPlex and SureSelect, from the Agilent platform allowed us to have a coverage of 92% of our selected SNPs. This coverage is ideal, but it must be compared with other platforms to assess the perfection, including genotyping platforms. This coverage remains theoretical (based on baits only) however a comparative study between these platforms will be performed on samples. This allows us to have practical coverage of our selection of SNPs. Such perspectives will have the advantage of fine-tuning the work and endow it with more specificity with regard to the pharmacogenomics.

CONFLICT OF INTEREST

The authors declare that no competing interests exist.

ACKNOWLEDGEMENT

We would like to thank the staff of Biochemistry and Immunology Laboratory and the staff of Montreal University for their technical assistance.

REFERENCES

[1] Brookes AJ. The essence of SNPs. Gene 1999; 234:177–86.

[2] Thorisson GA, Stein LD. The SNP Consortium website: past, present and future. Nucleic Acids Res 2003;31:124–7.

[3] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449: 851–61. doi:10.1038/nature06258.

[4] Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8:1229–31.

[5] McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, et al. A physical map of the human genome. Nature 2001;409:934–41. doi:10.1038/35057157.

[6] Phillips C. Selecting single nucleotide polymorphisms for forensic applications. Int Congr Ser 2004;1261:18–20. doi:10.1016/j.ics.2003.12.001.

[7] Prescott LM, Harley JP, Klein DA, Calberg CMB, Dusart J. Microbiologie. Paris: Deboeck; 2003.

[8] Scott T, Mercer EI, editors. Concise encyclopedia biochemistry and molecular biology. English language ed., 3rd ed. Berlin ; New York: Walter de Gruyter; 1997.

[9] Balaresque P. Les microsatellites des génomes eucaryotes: De leur cycle de vie et de leur neutralité. Médecine/Sciences 2007;23:729–34. doi:10.1051/medsci/20072389729.

[10] van Belkum A, Scherer S, van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev MMBR 1998;62:275–93.

[11] Varshney R, Horres R, Molina C, Nayak S, Jungmann R, Swamy P, et al. Extending the repertoire of microsatellite markers for genetic linkage mapping and germplasm screening in chickpea. Journal of SAT Agriculture 2007;5.

[12] Phillips C. Using online databases for developing SNP markers of forensic interest. Methods Mol Biol Clifton NJ 2005;297:83–106.

[13] Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 1988;26:2465–6.