Impact of missing genotype imputation on the power of Genome Wide Association Studies

Łukasz Król[1], Ghazi Alsbeih[2], Christophe Badie[3], Joanna Polańska[1]

[1] Data Mining Group, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
{lukasz.krol,joanna.polanska}@polsl.pl
[2] Head, Radiation Biology Section, Biomedical Physics Department, King Faisal Specialist Hospital & Research Centre (KFSH&RC), P.O. Box 3354, MBC-03, Riyadh 11211, Kingdom of Saudi Arabia
galsbeih@kfshrc.edu.sa
[3]Public Health England, OX11 ORQ Chilton, Didcot, UK.
christophe.badie@phe.gov.uk

Abstract. Genome Wide Association Studies are often performed on datasets containing a relatively small number of genotypes. In those cases, statistical tests performed on subgroups of those genotypes lack power or may even not fulfill the requirements of minimal number of observations. In this work we present results of running a GWAS analysis including parametric and nonparametric analysis of variance for different stress markers and models, validation by a group of related patients and clustering of results by their chromosomal positions. We show that the selection of imputation method has a significant impact at each phase of the analysis, and that it is worth to use the best method available.

Keywords: Genotype Imputation, Genome Wide Association Studies, Single Nucleotide Polymorphism, SNP Microarrays, PCR, Weighted Nearest Neighbor, fastPHASE, Analysis of Variance, r-scan

1 Introduction

When working with genotype data, one often faces the problem of missing values for certain loci of certain individuals. This can be an issue especially in case of microarray experiments, when we analyze hundreds of thousands of SNPs, but each SNP is observed in a limited number of patients. The limiting factor may be the cost of a microarray chip, or in some cases, like when measuring the effects of irradiation – problems with finding the group of interest. Depending on the type of analysis performed, different steps may be used to handle this issue. The simplest option is to simply skip the missing values. The obvious flaw of this approach is that it limits the power of the experiment – both in terms of its power to find true discoveries and reject false discoveries. Certain types of analysis, like multidimensional analysis where

we are interested in analyzing multiple features at once may be impossible to perform without fixing the missing signals.

There are many approaches to the problem of imputation. In general, one can either employ universal machine learning methods, or deploy algorithms dedicated to genomic data. A good overview of imputation approaches and their performance are presented in [1] and [2]. These domain-specific methods take into consideration the mechanisms behind the process that created the data, so they have a chance of yielding better results. The choice does not depend solely on a method's accuracy. Other factors of importance are time needed to perform the imputation, the availability of tools or the researcher's programming skills. Depending on the technologies used and the format of the dataset, implementing the required solution may be quite time consuming, both in terms of time needed for deployment, and the computation itself. The person analyzing the results may ask herself a question, whether the benefits of imputing the missing data are worth the extra effort.

Our work was aimed at answering this question by performing parallel analysis of genomic dataset imputed using various approaches. Besides capturing their raw imputation accuracy, we examined its impact on the output of the analysis procedure.

2 Materials

In the analysis we made use of two data sources. The primary one were genetic polymorphisms of 130 Caucasian individuals sampled at 565975 loci across 22 pairs of autosomal chromosomes using Affymetrix Axiom GW Hu SNP microarrays. Loci from sex chromosomes, as well as mitochondrial DNA polymorphisms were included in the datasets as well. In case of females, their Y polymorphisms were obviously marked as missing (and not imputed). In case of males, their X and Y mutations were marked as bi-allelic, although only one allele was present. This underlines the fact, that a mutation on the "one and only" allele works similar as a mutation on both alleles.

Table 1. The number of SNPs for each „logical" chromosome is presented along with the percent of missing signals and the average distance between subsequent SNPs. The average is calculated after rejecting the upper and lower half percentile of distances

chromosome:	1	2	3	4	5	6	7	8	9	10	11	12
missing signal ratio:	0,58%	0,60%	0,60%	0,61%	0,60%	0,60%	0,63%	0,61%	0,60%	0,61%	0,60%	0,60%
number of SNPs:	41578	45146	39789	37066	35568	43047	30657	31632	26197	27746	25802	27241
average distance:	5976	5382,2	4970,2	5149,6	5079,3	3966,4	5189,2	4618,9	5377,8	4878,1	5221,5	4904
13	14	15	16	17	18	19	20	21	22	X	Y	MT
0,60%	0,68%	0,59%	0,67%	0,65%	0,59%	0,86%	0,65%	0,61%	0,65%	0,55%	0,34%	0,66%
22343	18290	17016	16364	11636	17731	6890	13285	7894	5515	15291	1995	256
4291,9	4759,8	4832,5	5498,3	6956,7	4392,9	8540,4	4731,2	4716	6205,1	9951,6	28266	62,9

The second data source are gene expression levels for two genes – Ferredoxin Reductase [3] and Cycline G [4]. The levels for each marker were measured before and after exposure to radiation.

Each individual was accompanied by information about his phenotype, and his level of relatedness with other individuals. In general, three groups of relatedness exist in the data. 44 unrelated individuals serve as a basic investigation group. The dizygotic group of 56 siblings was as a validation group. These individuals are genetically different, but were assumedly raised in the same conditions.

The third group – homozygotic twins – was left for further studies.

3 Methods

3.1 Data preparation

Before the imputation and analysis steps could be implemented, some additional steps like data integration and cleaning had to be performed in order to create a consistent dataset to be used through this and following analyzes. The original data contained chromosomal positions of the polymorphism that were consistent with the positions from the *hg18* [5] assembly. We updated those positions with those from the *hg19* [5] assembly in hope that the more precise positions would slightly enhance the imputation accuracy.

3.2 Imputation techniques

Population mode.

This is a simple and straightforward method. Each polymorphism is imputed individually by selecting the value which is most common amongst the population. The fact that a big part of population usually shares the same haplotype block [6] makes its accuracy higher than in case of random signal distribution.

Although primitive, this method may be useful for creating a reference baseline for measuring the performance of other, more advanced methods.

Weighted Nearest Neighbor.

The Weighted Nearest Neighbor algorithm [2], inspired by the universal k-nearest neighbor algorithm may be viewed as a local and direct alternative to more advanced and time consuming but precise fastPHASE [7]. It is local in the sense that it includes only a limited neighborhood of the loci being imputed, and direct in the sense that it imputes the genetic signals without inferring the underlying haplotype phases.

The SNPs are imputed using orthogonal coding, that is no-mutation, one-mutation and two-mutation homozygotes are encoded as (001), (010) and (100) respectively. The missing signals are encodes as (000). As in case of any “neighborlike” classifier, a proximity measure between observations must be defined. The observation itself is a vector of 3 x w binary numbers, where w is the width of the neighborhood included in the experiment. With adopting the Hamming Distance [8] – the number of differences between binary vectors - in the proximity measure, we obtain the same distances for all three possible signal values.

The most interesting feature of this algorithm is that the neighboring loci's influence on the proximity measure declines with its distance from the loci being imputed. This reflects the fact that SNPs that are close to each other are likely to be inherited together and come from the same haplotype block, so the vector around the SNP of interest may form a pattern that will be found in other individuals.

In [2], special steps are performed in case of equal values of the proximity measure for two or more neighbors. These include recruiting additional observations to the voting pool, and ultimately extending the window. We use a simplified approach, and draw a sample of one neighbor from the nearest ones.

fastPHASE

fastPHASE is the imputation method introduced in [7]. It models the patients haplotypes as a mosaic of haplotype blocks common through the population. Furthermore, block membership of the alleles is assumed to change following a Hidden Markov Model [9]. The parameters of the whole model are fitted using the EM algorithm [10].

3.3 Measuring imputation performance

In order to measure the overall imputation performance, we created a test dataset from the main analysis group by removing any loci containing missing values. Then, for a given imputation algorithm and missing signal ratio, we marked random loci of random individuals as missing. Comparing the original values with those imputed allowed us to observe, besides the overall crude accuracy, individual accuracies for three possible values of input signals, with respect to the kind of error made by the imputation algorithm.

3.4 Analysis

In our analysis we were interested in picking up polymorphisms responsible for variations in levels of expression of the FDXR and CYCG1 stress markers. This scenario suits the ANOVA statistical procedure, with the values of polymorphisms being our independent variable, and the gene expression levels being our dependent variable. The statistical procedure was performed twice – once for the pre-irradiation expression levels, and once for the irradiated samples. We were interested in polymorphisms, for which significant differences in expression levels are observed for groups of genomic signals after irradiation, while the null hypothesis remains not rejected for original expression levels. This accounts for situation, where the level of organism's reaction to stress marker is a direct result of its genotype.

We have modeled the effects of differences at the levels of single alleles in three ways. In the dominant model, we assumed that the organism's reaction is triggered by mutation of one allele, so the heterozygotes and dominant homozygotes were grouped together. In the recessive model, when two mutations are required for the effect to take place, the heterozygotes were combined with recessive homozygotes. In the third

approach, we tried to model a situation where the effect is visible with one mutation, but is still enhanced by the second mutation, so we analyzed the groups separately. Each SNP was – if possible – analyzed with all three models. If multiple models passed validation by the related individuals group, that with the lowest p-value was chosen as that best describing the underlying mechanism.

The ANOVA procedure requires fulfillment of requirements of normality and equality of variances of the populations from which the groups were drawn. This was assessed by testing the groups for differences from normal distribution using the Shapiro-Wilk tests. The main reason for choosing it over more robust Anderson-Darling test was its low minimal sample size requirement. With a small population of 44 individuals, and given that most of them usually share the same haplotype block, testing for the normality of the smallest group would be impossible for most of the polymorphisms if using tests with larger minimal sample size requirements. The validation phase of the analysis should reject most of the false discoveries that arise from this simplification of the assumptions verification. The other assumption – the assumption of group variance equality – was verified using Bartlett test [11].

The method of handling a SNP for certain model (dominant, recessive, cumulative), marker (FDXR, CYCG1) and time (pre and post irradiation) depended on the level of fulfillment of ANOVA assumptions. If any of the tests could not be performed, the SNP was dropped from analysis. If all the assumptions were met, standard ANOVA was performed. In case of any problems with normality of the groups, Kruskal-Wallis rank test [12] was being performed instead of ANOVA. In case of issues with equal variances of the groups, Welsch t-test [13] was performed on 2-group models, and Kruskal-Wallis test in case of the cumulative model. If the level of fulfillment was different for the pre and post irradiation expression levels, then the most robust method was chosen to analyze both sets.

A unique feature of our analysis pipeline is the validation step where we used the group of related individuals to reject potential false discoveries. We deployed two validation conditions – one aimed at testing for the presence of acquired, not genetically-dependent reactions to irradiation, and one condition aimed at detecting the effects of experiment conditions. The first condition was implemented by arranging the pairs of related validation individuals so that they would represent different groups of the model currently being validated. If the model is valid, and the previously observed values of gene expressions were purely result of genetic differences, then we should observe significant differences of gene expressions as well. However, if the previously observed differences were results of acquired features, then the expression levels should be pulled together as the siblings have assumedly acquired the same features. In the second validation condition, we tested for effects of experiment conditions. In this case, we were operating on individuals sharing the same model state. If the observed levels of expressions in the groups were different, then the SNP was rejected.

3.5 Filtering out the most interesting results

All the SNPs picked up at the preceding phase are considered valid results that can be further investigated manually. However, it is worth to check those that are most promising first. A cluster of validated loci with low chromosomal distance may be the location worth investigating.

Our method of clustering is based on the r-scan statistics introduced by Karlin and Macken in [14]. Those statistics provide the information how unlikely would it be for a k-th smallest (or largest) distance to exist, if a larger distance was divided into n sub-distances at random (the null hypothesis). Using these statistics, a top-down or bottom-up clustering algorithm can be implemented. In the top-down approach, the set of SNPs is recursively divided into clusters, as long as the largest distance is too large to appear by chance alone. In the bottom-up approach, loci are grouped together as long as their distance – the outer linkage distance between the closest loci of two clusters – is too small to appear by chance. We decided to use the second approach, as we are interested in detecting clumping rather than overdispersion. The algorithm stops when the distances become likely to occur by chance, so no further analysis of the dendrogram is needed.

3.6 Capturing imputation's impact

In order to capture imputation's impact on the obtained results, the number of SNPs that passed each step of the data analysis pipeline for each gene-model-imputation algorithm is examined. Besides capturing the difference in total number of loci, two set differences between the unimputed and imputed are calculated in order to capture shifts in SNPs between analysis methods used.

To verify imputation's effect on p-values obtained for specific loci, paired Wilcoxon ranked tests [15] were performed for the p-values of loci validated for both methods to verify if the latter are lower.

4 Results

4.1 Simulated imputation accuracies

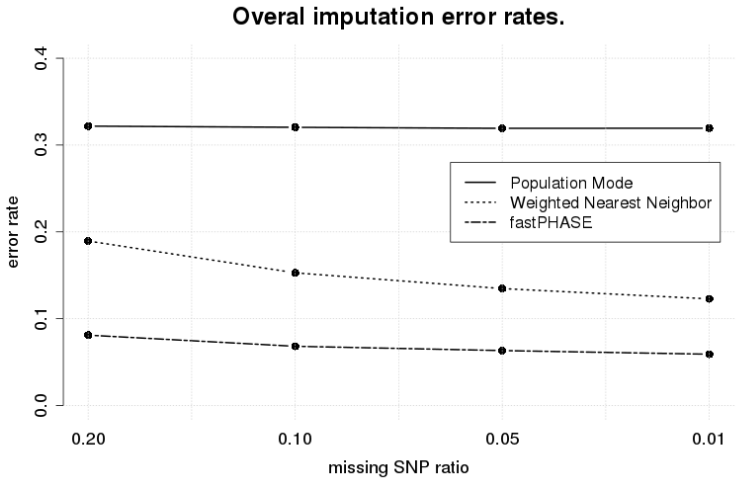

Fig. 1. The total imputation accuracies (the fraction of simulated missing signals imputed correctly) for the methods employed.

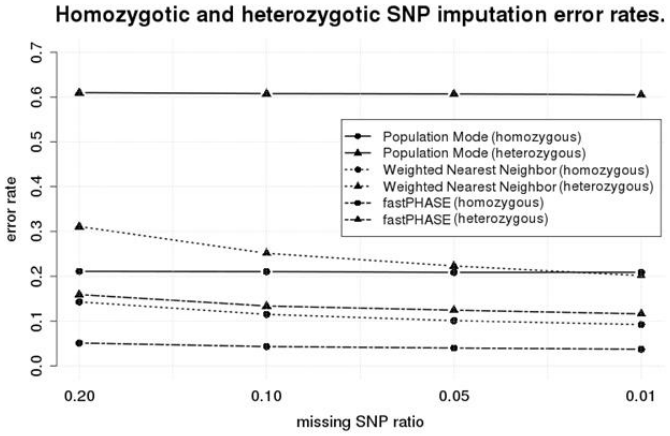

Fig. 2. Separate error rates for heterozygous and homozygous signals.

The results are consistent with those presented in [2]. The primitive method of replacing missing values with population mode has achieved worst results, although better than could be expected if the underlying genotypes were evenly distributed in the population. It's performance is not influenced by the number of missing signals – this

is because sampling does not change the underlying distributions. fastPHASE has obviously achieved highest results, while Weighted Nearest Neighbor is much faster at the price of imputation accuracy. Also, its performance seems to be the one most influenced by missing signal ratio.

4.2 Impact of imputation on analysis results

Table 2. Differences in the number of SNPs at different stages of the analysis expressed in terms of set differences. Values for the unimputed dataset are absolute, while others represent difference in regards to them.

gene	model	imp. method	applicable analysis					
			ANOVA		robust		invalid	
CYCG1	dominant	none	249247		90976		225752	
		mode	+1035	-978	+979	-1035	0	-1
		wnn	+2080	-1563	+1751	-1515	0	-753
		fph	+2067	-1557	+1746	-1499	0	-757
	recessive	none	276473		101250		188252	
		mode	+1329	-1208	+1208	-1328	0	-1
		wnn	+2336	-1978	+2158	-1815	0	-701
		fph	+2336	-1961	+2134	-1834	0	-675
	additive	none	132222		91296		342457	
		mode	+827	-988	+989	-827	0	-1
		wnn	+1792	-1440	+1882	-1137	0	-1097
		fph	+1805	-1452	+1935	-1140	0	-1148
FDXR	dominant	none	277509		62678		225788	
		mode	+527	-470	+470	-526	0	-1
		wnn	+1690	-1080	+1262	-1113	0	-759
		fph	+1718	-1076	+1252	-1132	0	-762
	recessive	none	308831		68852		188292	
		mode	+660	-540	+540	-659	0	-1
		wnn	+1921	-1197	+1375	-1399	0	-700
		fph	+1920	-1225	+1387	-1410	0	-672
	additive	none	159298		64211		342466	
		mode	+529	-498	+498	-528	0	-1
		wnn	+1817	-1054	+1418	-1082	0	-1099
		fph	+1897	-1060	+1431	-1118	0	-1150

gene	model	imp. method	null rejected				validated			
			ANOVA		robust		ANOVA		robust	
CYCG1	dominant	none	11588		5211		125		69	
		mode	+247	-235	+71	-91	+4	-2	+2	-1
		wnn	+360	-327	+142	-150	+5	-4	+3	-2
		fph	+352	-320	+142	-153	+5	-4	+3	-2
	recessive	none	12789		5732		161		75	
		mode	+279	-253	+112	-130	+2	-5	+2	-6
		wnn	+401	-372	+190	-182	+4	-5	+2	-5
		fph	+390	-366	+180	-181	+4	-5	+2	-6
	additive	none	6192		3462		2		4	
		mode	+149	-154	+79	-83	+1	0	0	-1
		wnn	+226	-223	+130	-107	+1	0	0	-1
		fph	+237	-205	+127	-104	+2	0	0	-1
FDXR	dominant	none	14455		2143		418		48	
		mode	+278	-282	+45	-44	+10	-10	0	0
		wnn	+439	-404	+93	-90	+11	-11	+2	-1
		fph	+421	-404	+92	-93	+12	-11	+2	-2
	recessive	none	16194		2240		550		66	
		mode	+330	-341	+63	-39	+14	-10	+1	-3
		wnn	+518	-446	+86	-82	+24	-24	+2	-5
		fph	+507	-469	+88	-84	+19	-20	+2	-5
	additive	none	8484		1105		3		2	
		mode	+194	-198	+33	-32	0	0	0	0
		wnn	+298	-255	+59	-51	0	0	0	0
		fph	+294	-255	+60	-53	0	0	0	0

Table 3. illustrates an obvious flaw of the primitive method of subsetting the missing signal with population mode – it never yields the least represented signal, so it doesn't increase the amount of loci that can be analyzed for a specific model. However, it affects the fulfillment of ANOVA assumptions as there are shifts between parametric and nonparametric branches of the analysis pipeline. Weighted Nearest Neighbor and fastPHASE have – on the contrary – allowed several hundred more SNPs to be analyzed.

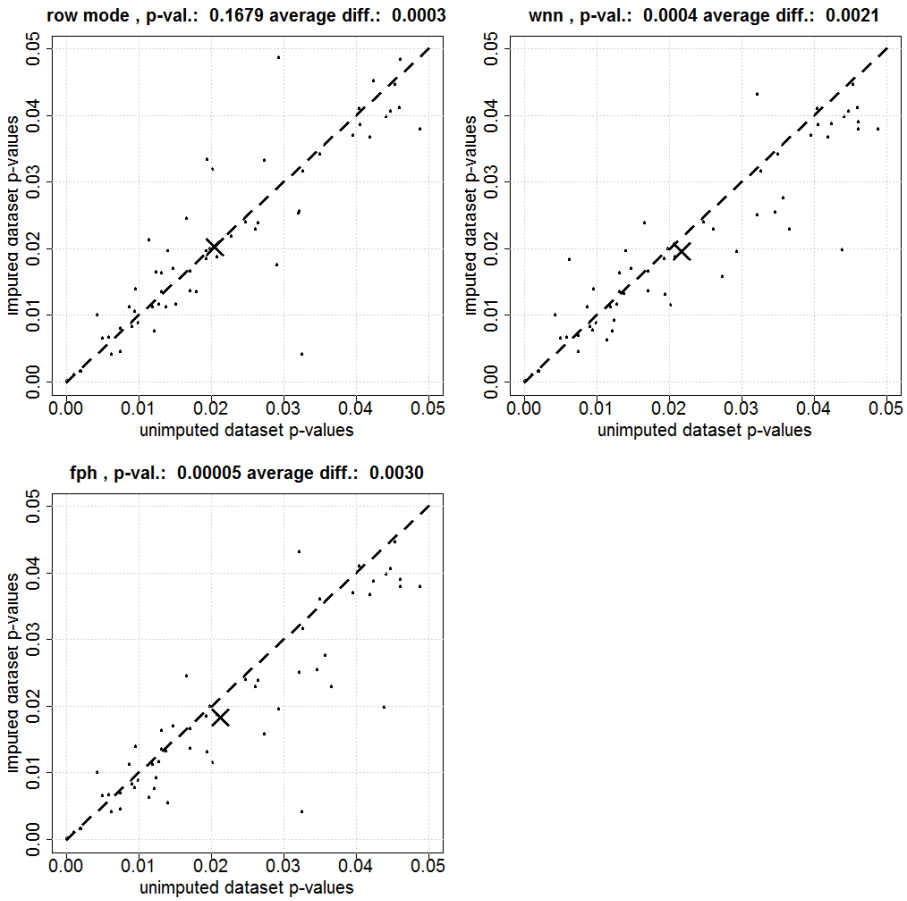

Fig. 3. A comparison of Cycline G1 p-values obtained for analogous loci of the original dataset and a dataset imputed by the primitive method of population mode, the Weighted Nearest Neighbor and fastPHASE. Above each scenario, the paired Wilcoxon test probability for the second group of p-values being lower under the null hypothesis is presented, along with the average difference in p-values. The X mark represents geometric center of the pairs of p-values.

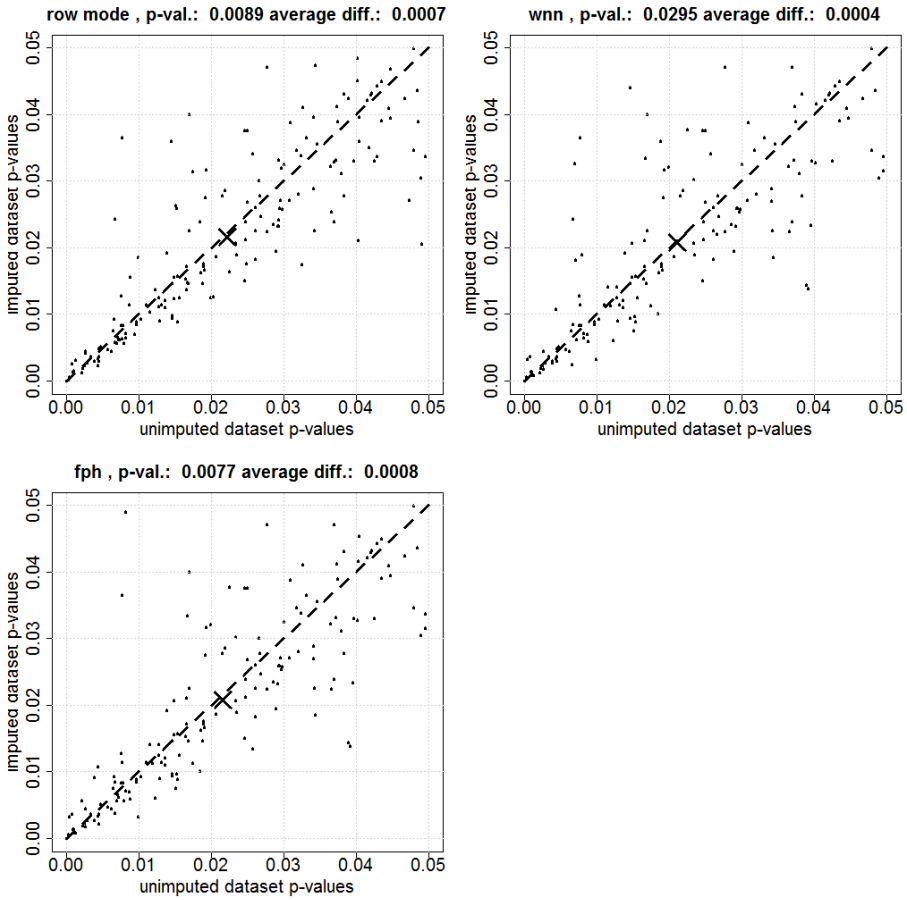

Fig. 4. An analogous to *Fig. 4* plot for Ferredoxin Reductase.

For all genes and imputation algorithms, the loci p-values observed for the imputed dataset are lower than those for the unimputed one. In case of Cycline G1 the differences for fastPHASE are the biggest and most significant, with the method of subsetting with row mode being the opposite. In case of Ferredoxin Reductase the differences are not as clear however.

4.3 r-scan clustering

Table 3. The differences between the number of clusters of specific sizes obtained for different imputation techniques using r-scan clustering. The values obtained for the unimputed dataset are expressed as absolute values, while the others are relative to them. The cluster size of 1 represents unclustered SNPs.

gene		CYCG1				FDXR			
cluster size		1	2	3	4+	1	2	3	4+
imp. method	none	252	43	16	9	528	86	45	40
	mode	-6	1	-1	1	10	-4	3	-1
	wnn	-1	-2	0	1	9	-5	-2	2
	fph	-10	1	1	1	-11	1	0	2

Table 4. A comparison of the number of clustered SNPs in the unimputed and imputed datasets. The values being presented are the numbers of actual SNPs being removed or introduced.

gene		CYCG1		FDXR	
imp. method	none	180		555	
	mode	+4	-2	+13	-21
	wnn	+11	-12	+18	-29
	fph	+12	-4	+30	-22

Table 4. shows that imputing less than a percent of missing values has increased the clumping of results – the number of results participating in a cluster – by a magnitude of several percent.

5 Conclusion

Even with a ratio of missing values below one percent, introduction of imputation causes shifts in the data analysis pipeline – it increases the number of SNPs valid for analysis, influences results of statistical procedures and may change patterns of chromosomal positions of the result locations. Multidimensional analysis of polymorphisms could even be impossible to perform without it. For these reasons, when it is employed, it should be done the best way possible. The easiest method of replacing a missing value with population mode for a specific SNP has many flaws, the most significant one being discrimination of the least represented signal, which is a key to increasing power of the experiment. An obvious solution is a mature imputation algorithm like fastPHASE.

There are two apparent advantages of the substitution by population mode – facility of deployment and low processing time. The latter is rivalled by Weighted Nearest Neighbor, while the first one could be overcome with deployment of on-line applications that would allow researchers to benefit from its high speed and accuracy without carrying the burden of implementing it by themselves. This, combined with its development potential like employment of dynamic weighting functions reflecting local characteristics of the chromosome makes it a method worth of further investigation.

6 Acknowledgements

We would like to thank Dr. S. Majid, Ms. N. Al-Harbi, Ms. S. Al-Qahtani for running the Axiom Affymetrix platform and Sylwia Kabacik and Paul Finnon for run the H2AX test. The work was financially supported by NCN grant HARMONIA UMO-2013/08/M/ST6/00924 (JP,LK), the National Institute for Health Research Centre for Research in Public Health Protection at Public Health England (CB) and the National Science, Technology & Innovation Plan (NSTIP) Project 11-BIO1429-20 (KFSHRC RAC# 2120 003). Project was partially financed by NCBiR project no POIG.02.03.01-24-099/13 "Upper Silesian Center for Computational Science and Engineering".

7 References

1. Sharon R. Browning: Missing data imputation and haplotype phase inference for genome-wide association studies, Human Genetics December 2008, Volume 124, Issue 5, pp 439-450
2. Yining Wang, Zhipeng Cai, Paul Stothard, Steve Moore, Randy Goebel, Lusheng Wang, Guohui Lin: Fast accurate missing SNP genotype local imputation, BMC Research Notes, August 2012, 5:404
3. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/gene/2232
4. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/gene/900
5. University of California Santa Cruz, https://genome-euro.ucsc.edu/cgi-bin/hgGateway
6. International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
7. Scheet P, Stephens M.: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006 Apr;78(4):629-44
8. Hamming, Richard W.: Error detecting and error correcting codes, Bell System Technical Journal, Volume 29, Issue 2, pages 147–160, April 1950
9. Rabiner, L.:A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE (Volume:77 , Issue: 2), 257-286
10. A. P. Dempster; N. M. Laird; D. B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38.
11. M. S. Bartlett: Properties of Sufficiency and Statistical Tests, Proc. R. Soc. Lond. A May 18, 1937 160 901 268-282
12. William H. Kruskal, W. Allen Wallis: Use of Ranks in One-Criterion Variance Analysis, Journal of the American Statistical Association, Volume 47, Issue 260, 1952, pp. 583-621
13. Welch, B. L.: The generalization of "Student's" problem when several different population variances are involved, Biometrika. 1947;34(1-2):28-35.
14. Karlin S, Macken C.: Assessment of inhomogeneities in an E. coli physical map, Nucleic Acids Res. 1991 Aug 11;19(15):4241-6.
15. Wilcoxon, Frank.: Individual comparisons by ranking methods, Biometrics Bulletin 1 (6): 80–83