Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Development of Multi-Null-Hypotheses Method for Detection of Selective Forces at Molecular Level in Evolution of Human Genes involved in DNA-Repair Mechanism Impaired in Cancer Progression

Krzysztof A. Cyran*, Marek Kimmel**

**Institute of Informatics, Silesian University of Technology, Gliwice,*
Poland (Tel: +48-32-2372500; e-mail: krzysztof.cyran@polsl.pl).
***Institute of Automatic Control, Silesian University of Technology, Gliwice,*
Poland (Tel. +48-32-2371051; e-mail: kimmel@rice.edu)

Abstract: We developed a multi-null-hypotheses (MNH) method for testing signatures of natural selection in Kimura's neutral model of human evolution with population growth and limited migration between dems. To evaluate the influence of such time changes and demography in population size, we employed different variants of the null hypothesis, corresponding to constancy, growth, or substructure and growth, respectively. We apply the model for searching in SNP haplotypes from four genes implicated in impairing DNA-repair mechanism in human familial cancers: ataxia telangiectasia (ATM), human helicase RECQL, Bloom's syndrome (BLM) and Werner's syndrome (WRN). The sample is composed of about 600 chromosomes, derived from residents of Houston, TX (USA), representing major ethnic backgrounds: Caucasian, African, Asian and Hispanic. The method is illustrated with Bloom's neutrality tests B and Q. Our study suggests that detected deviations from neutrality may be obscured by presence of recombination, substructure and changes of population size in the case of RECQL haplotypes, while in the ATM haplotypes the signal is rather strong. The BLM and especially WRN haplotypes do not register deviations from neutrality.

Keywords: Bioinformatics, modeling and identification, data mining tools, cancer genes evolution, searching natural selection, DNA-repair mechanism in cancer progression

1. INTRODUCTION

Evolutionary modeling of cancer has resulted recently in numerous discoveries ranging from detection of many new loss-of-function mutations in recessive cancer genes (Bignell et al. 2010) to studies on genome instabilities such as chromosome abnormalities and unstable DNA (Hanahan and Weinberg, 2011), which are mentioned as *Enabling Characteristics* in *New Generation Hallmarks of Cancer*. When the important for precise replication DNA-repair mechanism becomes impaired, the instability of DNA accumulates in cancer cells allowing for unlimited clonal growth of the tumor. The importance of this mechanism implies that genes implicated in a DNA-repair process are targets in a search for signatures of natural selection operating at molecular level.

In a series of our previous papers (Cyran, Polanska, and Kimmel 2004, Cyran 2009, Cyran and Kimmel 2014), we investigated natural selection at SNP haplotypes at four genes implicated in human familial cancers: ataxia telangiectasia (ATM), human helicase RECQL, Bloom's syndrome (BLM) and Werner's syndrome (WRN) (Bonnen et al. 2000, Bonnen et al 2002, Trikka et al. 2002). A number of interesting facts were determined about these genes, but development of multi-null-hypotheses method described in the current paper, allowed for further verification of these results by Wall's (1999, 2001) neutrality tests. ATM gene encodes a large protein involved in response to DNA damage. Abnormalities in this gene with discovered remarkable diversity, include not only point mutations, but also small rearrangements leading to splicing mutations (Teraoka et al. 1999). It has been demonstrated by Li and Swift (2000) that patients who are heterozygous for splice site mutations have significantly longer survival time as compared to those who are homozygous for single truncating mutations. Some of mutations in ATM gene are responsible for ataxia telangiectasia, a serious recessive pleiotropic disorder. By influencing DNA instability, these mutations also increase significantly inherited predisposition to cancer. The age of one ATM mutation, estimated by Broeks *et al.* (2003) using Maximum Likelihood (ML) method to be at least 50,000 years, indicates that the variation present at this locus could be driven by balancing selection.

The remaining three analyzed loci encode for human DNA helicases. Their function is related to DNA metabolism, including transcription, accurate chromosomal segregation, recombination, and repair. Helicase-dependent DNA repair include such molecular mechanisms as mismatch repair, nucleotide excision repair, and direct repair. Bloom and Werner syndromes, rare autosomal recessive disorders, are both characterized by a high predisposition to malignancies (Siitonen et al. 2003). Cells in Bloom syndrome exhibit hypermutability including hyperrecombinality between sister chromatids and homologous chromosomes (Yusa et al.

2004). Karow et al. (2000) and Wu and Hickson (2003) emphasize the role of BLM as an antirecombinase for suppresion of tumorigenesis. Adams et al. (2003) concluded that BLM product maintains genomic stability.

Since, the genes considered are implicated (by proper or impaired DNA repair) in accumulation of autosomal mutations in tumors in human familial cancers, they could be potentially subject to natural selection. With that motivation, in the present paper, we apply to our data Wall's neutrality tests, which are sensitive to different deviations from neutrality, in an attempt to determine whether signatures of selection are present at these gene loci. By now, there are a number of interesting examples of natural selection at molecular level (Gilad et al. 2002, Toomajian and Kreitman 2002), with perhaps the most spectacular one being the ASPM locus, a major contributor to brain size regulation in primates (Zhang 2003). There exist two general types of tests of natural selection at molecular level: First, whenever the data consists of entire or partial coding sequences of a gene, comparison of frequencies of silent substitutions at the third codon position to the frequencies of substitutions on the first and second position, allows to measure selective pressure.

Unfortunately, in many cases, we have to do with a second type of data, which consists of sequences that are not only non-coding, but also composed of nucleotides located at considerable distance from each other. Single Nucleotide Polymorphisms (SNP) in intronic regions of a target gene provide one example of data of the second type. They form haplotypes, which serve as tools to investigate the genetic diversity and disease association of the target gene.

In such cases, a model for neutral evolution of the sequence has to be determined and then its predictions compared to data. Usually, this model is a form of the Wright-Fisher model of genetic drift with mutation (Jobling et. al. 2004). Significant departure from predictions under neutrality (which serves as the null hypothesis) may provide evidence for selection (the desirable alternative hypothesis). However, there exist a number of undesirable alternatives, which may cause departures from the null, and be confused with selection. The most common of these are population substructure and past change in population size (Nielsen 2001). Sometimes, these influences cannot be disentangled from selection. One way to deal with this problem is to apply a number of tests, each one sensitive to different combination of factors, and compare the results. The substructure problem can be approached by considering data from different subpopulations separately.

2. MATERIALS AND METHODS

We analyzed a total of 45 Single Nucleotide Polymorphisms (SNPs) located on intronic and other non-coding sequences in the four helicases: ATM, BLM, WRN, and RECQL. Detailed information on names, positions and variations of the analyzed SNPs, primer sequences, PCR conditions and product size for each of the polymorphic sites, as well as the ASO sequences and wash conditions for each SNP variant are presented in Bonnen et al. (2000) for the ATM gene and in Trikka et al. (2002) for the BLM, WRN and RECQL.

Blood samples were collected from the individuals, residents of Houston, TX, from four major ethnic groups: Caucasians, African-Americans, Hispanics, and Asians (Table 1). The group called "Global" was composed of the above-mentioned ethnic groups and 40 individuals, the members of the CEPH pedigrees. Haplotypes were inferred and their frequencies were estimated by using the Expectation-Maximization (EM) algorithm (Dempster et al. 1977, Excoffier and Slatkin 1995, Polanska 2003). The most frequent haplotypes (with the estimated frequency greater than 5%) and their frequencies among each ethnic group are presented in Tables 1 – 4.

Table 1. The most frequent ATM haplotypes and their frequencies among the analyzed ethnic groups.

ATM Haplotype	African-Americans	Caucasians	Hispanics	Asians	Global
ACTCTACTTCTTTC	0.1899	0.2922	0.3149	0.5000	0.3044
ACTCTCCTTCTTTC	-	0.0649	-	-	-
ATTCTACTTCTTTC	-	-	-	0.0513	-
ACTTTACTTCCCTC	0.2183	-	-	-	0.0983
ACTTTACTTCTTTC	0.0775	-	-	-	-
TTCTCACTCTCCTA	-	0.1753	-	-	-
TTTTCACCCTCCTC	0.1408	0.0974	0.1096	0.0680	0.1175
TTTTCACTCTCCTC	0.1617	-	-	-	0.0735
TTTTCATCCTCCCC	0.1127	0.3506	0.3630	0.2910	0.2485

Table 2. The most frequent RECQL haplotypes and their frequencies among the analyzed ethnic groups.

RECQL Haplotype	African-Americans	Caucasians	Hispanics	Asians	Global
ACCTAGTAGCT	0.1770	0.0900	0.0615	0.0541	0.1035
ACCTAGTAGTT	0.1540	0.1819	0.1928	0.2136	0.1806
ATCTAGTAGCT	0.0748	-	-	-	-
ATCTAGTAGTT	-	0.0599	0.0644	0.1508	0.0762
ATGGGACGATT	0.1061	0.3825	0.3948	0.4966	0.3314
ATGGGGTGACT	0.0711	-	-	-	-
ATGGGGTGGCT	0.0957	-	-	-	-
ATGGGGTGGTT	0.0519	-	-	-	-

Table 3. The most frequent WRN haplotypes and their frequencies among the analyzed ethnic groups.

WRN Haplotype	African-Americans	Caucasians	Hispanics	Asians	Global
ATCAGGTACGGG	-	0.1594	0.1261	0.0757	0.1000
ATCTGGTACGGG	0.0501	-	-	-	-
ATCTGGTATGGG	0.0521	-	-	-	-
ATTTAATCCGGG	-	-	-	0.0542	-
ATTTAATCTGGG	0.0827	0.1705	0.2188	0.1766	0.1602
ATTTGATCTGGG	0.0783	-	-	-	-
ATTTGGGACGGG	0.2450	0.2165	0.2509	0.3892	0.2538
ATTTGGGATGGG	-	0.1087	0.0825	0.0595	0.0776
ATTTGGTACGGG	0.1343	-	-	-	-
ATTTGGTATGGG	0.1113	-	-	-	0.0550
GTTTGGGACGGG	-	-	0.1115	0.1026	-

Table 4. The most frequent BLM haplotypes and their frequencies among the analyzed ethnic groups.

BLM Haplotype	African-Americans	Caucasians	Hispanics	Asians	Global
CATCGCAG	-	-	0.0915	0.1428	0.0745
CATCTGCG	-	-	-	0.0528	-
CGCCGCAG	0.0715	-	-	-	-
CGTCGCAG	-	-	0.0542	0.1226	-
CGTCTGCG	-	0.0509	-	-	-
TGCCGCAG	0.2260	0.0785	0.0971	-	0.1116
TGCGGCAA	0.0715	-	0.0796	0.0809	0.0849
TGTCGCAG	0.0649	0.2564	0.3050	0.2987	0.2146
TGTCGGCG	0.0833	-	-	-	-
TGTCTGCG	0.1311	0.2152	0.1830	0.0698	0.1684

Mathematics of natural selection – contrary to that of genetic drift and mutation – results in significantly increasing complexity of equations even with the small increase in a number of loci and alleles considered. For the simplest case of single locus and two-allele model, the change $\Delta_s p$ of the frequency p of allele A_1 can be explicitly derived as a function of current frequencies p and q of allele A_1 and A_2 respectively, as

$$\Delta_s p = \frac{pq\left(p\left(1-\frac{w_{12}}{w_{11}}\right)+q\left(\frac{w_{12}}{w_{11}}-\frac{w_{22}}{w_{11}}\right)\right)}{p^2+2pq\frac{w_{12}}{w_{11}}+q^2\frac{w_{22}}{w_{11}}}, \qquad (1)$$

where w_{11}, w_{12}, and w_{22}, represent viabilities (probabilities to survive to give birth to progeny) of genotypes A_{11}, A_{12}, and A_{22}, respectively. To identify particular phenomena in more complex, and therefore interesting from practical perspective, models, i.e. models with more loci and more alleles, the dynamics of selection process needs to be studied by other, computational methods. One of such methods is testing against numerous null hypotheses using statistics with unknown critical values, and hence demanding large scale coalescent-based simulations for appropriate modeling of signatures of natural selection.

In our detailed analysis, we applied Wall's (1999) tests *B* and *Q*. Statistic *B* is the normalized number *B*' of pairs of adjacent congruent (inducing identical partitions) segregating sites. To be normalized, *B*' is divided by the total number (K – 1) of pairs of adjacent segregating sites

$$B = \frac{B'}{K-1}. \qquad (2)$$

If we indicate by *A* the set of all distinct partitions induced by pairs of adjacent congruent segregating sites, then the statistic *Q* is defined as

$$Q = \frac{B + card(A)}{K}, \qquad (3)$$

where is the number of elements in *A* and the power of a test becomes less sensitive to the recombination, because of compensation of the decrease of *B* by an increase of card (*A*).

Family of null hypotheses employed

In order to fine-tune the outcome of testing, we employed, for Wall's (1999) tests, three different null hypotheses:

- H_{00}, panmictic population, with size constant in time,
- H_{01}, panmictic population, with size increasing 10 times over the period of 5000 human generations, according to an exponential function,
- H_{02}, population with internal structure, composed of 4 demes with between-demes migration rate mN=10, with total size increasing 10 times over the period of 5000 human generations, in a growth modeled by an exponential function.

Each of the above null hypotheses assumes selective neutrality.

Hypotheses H_{01} and H_{02} are less conservative than H_{00}, although they are still conservative, considering the feasible human population history scenarios. Wall's (1999) tests and coalescent simulations for critical values were conducted with the use of Rozas's DNASP and Filatov's ProSeq software.

3. RESULTS

Application of Wall's neutrality tests to data

Our analysis in Wright-Fisher neutral model of four loci, ATM, RecQL, WRN and BLM, reveals that they fall into two categories A (ATM, RecQL) and B (WRN, BLM). This partition is clearly observed according to Wall's B and Q tests, (Tables 5 – 8) though almost no outcome is significant (except for Asian population for RecQL locus) in these tests against null hypothesis H_{00} (constant size, panmictic and selectively neutral population). Tables 5 – 8 present for each test the value of the corresponding statistic and the critical values at 0.05 significance level under null hypotheses H_{00} H_{01} and H_{02}.

Table 5 Results of the Wall's B and Q tests for ATM locus

ATM	*B*				*Q*			
	Value	*H₀₀*	*H₀₁*	*H₀₂*	Value	*H₀₀*	*H₀₁*	*H₀₂*
AfAm	0	0.33	0.08	0.17	0	0.46	0.14	0.29
Caucasian	0.23	0.39	**0.08**	**0.15**	0.36	0.43	**0.14**	**0.25**
Asian	0.15	0.39	**0.08**	**0.15**	0.29	0.50	**0.14**	**0.29**
Hispanic	0.08	0.39	**0.08**	0.15	0.14	0.43	**0.14**	0.21

Results of these tests against the slightly less conservative null hypotheses H_{01} and H_{02} are significant for genes ATM and RecQL for almost all considered subpopulations, but still not significant for any population at BLM and WRN loci. The expected frequencies of types of segregating sites under null hypotheses H_{00}, H_{01} and H_{02} at the ATM locus in African American population are depicted in Figure 1.

Table 6 Results of the Wall's B and Q tests for RECQL

RECQL	*B*				*Q*			
	Value	H_{00}	H_{01}	H_{02}	Value	H_{00}	H_{01}	H_{02}
AfAm	0.20	0.40	**0.10**	**0.10**	0.36	0.46	**0.18**	**0.18**
Caucasian	0.20	0.40	**0.10**	**0.10**	0.36	0.56	**0.18**	**0.18**
Asian	0.44	**0.44**	**0.11**	**0.22**	0.60	**0.50**	**0.20**	**0.30**
Hispanic	0	0.40	0.10	0.10	0	0.46	0.18	0.27

Table 7 Results of the Wall's B and Q tests for WRN

WRN	*B*				*Q*			
	Value	H_{00}	H_{01}	H_{02}	Value	H_{00}	H_{01}	H_{02}
AfAm	0	0.36	0.09	0.22	0	0.50	0.17	0.30
Caucasian	0	0.36	0.09	0.18	0	0.50	0.17	0.25
Asian	0	0.38	0.13	0.25	0	0.56	0.22	0.33
Hispanic	0	0.36	0.09	0.18	0	0.50	0.17	0.25

Table 8 Results of the Wall's B and Q tests for BLM

BLM	*B*				*Q*			
	Value	H_{00}	H_{01}	H_{02}	Value	H_{00}	H_{01}	H_{02}
AfAm	0	0.43	0.14	0.14	0	0.50	0.25	0.25
Caucasian	0	0.44	0.14	0.14	0	0.50	0.25	0.25
Asian	0	0.43	0.14	0.14	0	0.63	0.25	0.25
Hispanic	0	0.43	0.14	0.14	0	0.50	0.25	0.25

Results for the global samples are presented in the Table 9, below

Table 9 Wall's B and Q tests for all loci and global samples

Global	*B*				*Q*			
	Value	H_{00}	H_{01}	H_{02}	Value	H_{00}	H_{01}	H_{02}
ATM	0	0.31	0.08	0.08	0	0.43	0.14	0.14
RecQL	0	0.30	0.10	0.10	0	0.46	0.18	0.17
WRN	0	0.33	0.09	0.09	0	0.42	0.17	0.17
BLM	0	0.27	0.14	0.14	0	0.38	0.25	0.25

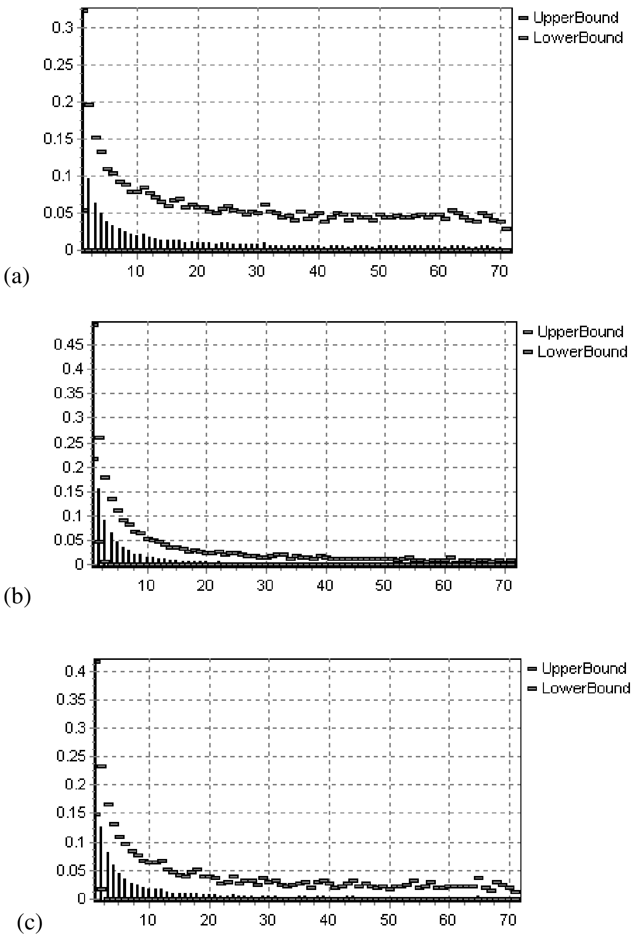

Fig. 1. Frequencies of segregating site types under null hypotheses (a) H_{00}, (b)H_{01} and (c)H_{02}.

4. DISCUSSION

Nielsen (2001) suggests to be conservative in concluding about detection of selection based on tests using only haplotype spectrum data, because other alternative hypotheses lead to similar results. The main alternative is that of population growth, which can be easily mistaken for a selection sweep. Further analysis of obtained results shows, however, that these concerns, which are especially important in the case of selective sweeps leading to an excess of young mutations (Fu 1997), are not applicable directly to this study, with samples displaying excess of old mutations.

The interpretation of the results from the *B* and *Q* tests is not difficult in the light of known demography of human population. Although for all genes Wall's *B* and *Q* tests did not indicate any statistical significance against constant population size, totally panmictic and selectively neutral model under null hypothesis H_{00}, the partition of genes into categories A and B (Tables 1-4) is evident based on these tests. The explanation why the results are not significant for ATM and RECQL when tested against classical null hypothesis, lays in the nature of the tests which paradoxically may lose power for relatively large samples. The superiority of *B* and especially of *Q* over other tests was confirmed in Wall's (1999, Fig. 7) simulations, in terms of power to detect departures from selective neutrality under recombination. Perhaps Wall's (1999) conclusions should be restricted to relatively small sample sizes, if the value of Watterson's estimate $\hat{\theta}_W$ is smaller than value used in Wall's simulations ($\hat{\theta}_W = 7.5$).

In our studies we obtained, for different genes and populations, $\hat{\theta}_W$ ranging from 1.1 to 2.8 for much larger sample sizes than that analyzed by Wall, the power of Wall's tests for our data was less than powers of tests previously analyzed. Therefore, we were able to increase the power of the tests by changing the null hypothesis H_{00} to H_{01} which incorporates population growth. In doing so, we tried to be as conservative as possible: we incorporated population growth in a very conservative range, just 10-fold over 5,000 generations.

We expected the tests to be less conservative against null hypothesis H_{01} because we already had strong indications about dealing with balancing selection (which, contrary to selective sweeps, causes genetic variability very dissimilar to that caused by population growth). The obtained results (Tables 5 – 8) confirm that ATM and RecQL display statistically significant departures from neutrality if H_{01} is used as the null hypothesis. Therefore, we tested neutrality against null hypothesis H_{02}, the scenario with 10-fold population growth and population substructure with 4 demes and normalized migration rate $Nm = 10$. This time it was difficult to predict à priori whether testing using H_{02} will be more or less conservative than testing using H_{00}, because of mutually canceling effects of population growth and population stratification, whereas it was obvious that testing using H_{02} was more conservative than using H_{01}. The results indicated that powers P_i of testing against H_{0i} ($i = 0, 1, 2$) satisfied the relationship: $P_0 < P_2 < P_1$. Similarly as against null hypothesis H_{01}, testing against H_{02} gave in majority of cases significant departures from neutrality for ATM and RECQL but not for WRN and BLM.

The observation that Wall's *B* and *Q* statistics gave for all genes values equal zero when applied to global samples, at first looked contrary to expectations, because these tests were explicitly designed to detect departures similar to those produced by population stratification. However, we observe that with the increase of the sample size *n* the ratio of the length of any particular branch of the coalescent tree to the length of the whole sample genealogy becomes smaller. Therefore, for given number of segregating sites, the probability of finding *any* adjacent congruent segregating sites decreases and tends to zero as $n \to \infty$. On the other hand, the total number of segregating sites should increase with the increase of sample size, which theoretically could compensate the previous effect. However since the rate of growth of the number of segregating sites with the growth of *n* also decreases with the increase of *n*, for really large sample sizes, the compensation is perhaps not strong enough to avoid the disappearance of congruent segregating sites. Detailed additional studies are required to determine the exact relationships between sample size and the possible reduction of the number of congruent sites to zero, but looking at the definition of *B* and *Q* tests explains why in the absence of congruent segregating sites these tests may yield null values.

Our global samples were much greater than subpopulation samples and one order of magnitude larger than those analyzed by Wall (1999, Fig 7). Furthermore, our estimates of $\hat{\theta}_W$ were 3 to 7 times smaller than those used by Wall in his simulations. As a result, at least for our data, we have observed the effect of disappearance of adjacent congruent sites, obtaining global samples with no such sites. For this reason, the *B* and *Q* statistics were equal to zero for all loci under consideration. Furthermore, we confirmed this effect by testing with twice smaller samples obtained by random drawing of the half of chromosomes from original samples. With smaller sample sizes the global samples still yielded zeros but some of the tests originally yielding zeros, e.g., for African-American population at ATM locus and for Hispanic population at RECQL locus, presented positive results statistically significant against H_{01} and H_{02}.

With above discussion it is plausible that the non-significant results at ATM and RECQL do not contradict the possibility of balancing selection operating at regions linked to ATM and RECQL introns, as determined by all previous results.

In conclusion, the only type of selection at a linked locus resulting in positive values of tests obtained for ATM and RECQL is balancing selection (Fu 1996). It is so, because advantageous selection sweeps always result in negative values of *F*'(*r*, *r*') tests (Fu 1997, Kreitman 2000). This conclusion is also confirmed by the results of Wall's *B* and *Q* tests, which seemed to be more conservative against null hypothesis H_{00} (without growth) than against H_{01} (with growth). The deleterious mutation model cannot be excluded, but it demands special mutation rates from advantageous to inferior allele and different rates in opposite direction. Therefore, for patterns observed at ATM and RECQL the most feasible conclusion seems to be the over-dominant selection, always resulting in selection-mutation balance, which in turn is responsible for the observed excess of old mutations in ATM and RECQL (but not in WRN) haplotype data. The conclusion about BLM is not univocal because majority of tests of class *F*'(*r*, *r*') give results on the boundary of significance (also in the direction of the excess of old mutations), and even if *B* and *Q* do not confirm this departure it could be simply because of relatively strong recombination operating in this locus as detected by Fu (1997) test F_s.

ACKNOWLEDGEMNTS

The research leading to these results has been performed at William Marsh Rice University in Houston, TX and at Silesian University of Technology in Gliwice, Poland. It has received funding from the PEOPLE Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 298995 to support research work by K.A. Cyran during his

International Outgoing Fellowship to Rice University. Research work of M. Kimmel was supported by NCN MAESTRO grant no. DEC-2012/04/A/ST7/00353.

REFERENCES

Adams, M. D., McVey, M., and Sekelsky, J. J. (2003) Drosophila BLM in double-strand break repair by synthesis-dependent strand annealing. *Science*, 299, 265-267.

Bignell, G. R., Greenman, Ch. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., Widaa, S., Hinton, J., Fahey, C., Fu, B., Swamy, S., Dalgliesh, G. L., Teh, B. T., Deloukas, P., Yang, F., and Campbell, P. J. (2010). Signatures of mutation and selection in the cancer genome. *Nature*, 463, 893-900.

Bonnen, P. E., Story, M. D., Ashorn, C. L., Buchholz, T. A, Weil, M. M., and Nelson, D. L. (2000) Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am J Hum Genet*, 67, 1437-1451.

Bonnen, P. E., Wang, P. J., Kimmel, M., Chakraborty, R., and Nelson D. L. (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res*, 12, 1846-1853.

Broeks, A., Urbanus, J. H. M., de Knijff, P., Devilee, P., Nicke, M., Klopper, K., Dork, T., Floore, A. A. N., and van't Veer, L. J. (2003) IVS10-6T-G, an ancient ATM germline mutation linked with breast cancer. *Hum. Mutat.*, 21, 521-528.

Cyran, K. A., Polanska, J., Kimmel, M. (2004) Testing for signatures of natural selection at molecular level. *Journal of Medical Informatics & Technologies*. 8, 31-39.

Cyran, K. A. (2009) Quasi Dominance Rough Set Approach in Testing for Traces of Natural Selection at Molecular Level. In K.A. Cyran et al. (ed.)*, Advances in Intelligent and Soft Computing 59*, 163-172. Springer, Berlin Heidelberg.

Cyran, K. A., and Kimmel, M. (2014) Comparison of Connectionist and Rough Set Based Knowledge Discovery Methods in Search for Selection in Genes Implicated in Human Familial Cancer. A. Gruca, T. Czachórski, and S. Kozielski (ed.), *Advances in Intelligent Systems and Computing 242*, 163-171. Springer, Heidelberg New York Dordrecht London.

Dempster, A. P., Laird, N. M., and Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. With discussion. *J Roy Statist Soc Ser B,* 39, 1–38.

Excoffier, L, Slatkin, M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12, 921-927.

Fu, Y. X. (1996) New Statistical Tests of Neutrality for DNA Samples From a Population. *Genetics*, 143, 557-570.

Fu, Y. X. (1997) Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *Genetics*, 147, 915-925.

Gilad, Y., Rosenberg, S., Przeworski, M., Lancet, D., and Skorecki, K. (2002), Evidence for positive selection and population structure at the human MAO-A gene. *Proc. Natl. Acad. Sci.*, 99, 862-867.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144 (5), 646–674.

Jobling, M.A., Hurles, M. E., and Tyler-Smith, C. (2004) *Human Evolutionary Genetics: origins, peoples & disease*, Garland Science, New Delhi, India.

Karow, J. K., Constantinou, A., Li, J.-L., West, S. C. , and Hickson, I. D. (2000) The Bloom's syndrome gene product promotes branch migration of Holliday junctions. *Proc. Nat. Acad. Sci.*, 97, 6504-6508.

Kreitman, M. (2000) Methods to Detect Selection in Populations with Application to the Human. *Annu. Rev. Genomics Hum. Genet.*, 01, 539-559.

Li, A., and Swift M. (2000) Mutations at the ataxia-telangiectasia locus and clinical phenotypes of A-T patients. *Am. J. Med. Genet.*, 92, 170-177.

Nielsen, R. (2001), Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86, 641-647.

Polanska, J. (2003) The EM algorithm and its implementation for the estimation of the frequencies of SNP-haplotypes. *Int. J. Appl. Math. Comput. Sci.*, 13, 419-429.

Siitonen, H. A., Kopra, O., Haravuori, H., Winter, R. M., Saamanen, A. M., Peltonen, L., and Kestila, M. (2003) Molecular defect of RAPADILINO syndrom expands the phenotype spectrum of RECQL diseases. *Hum. Mol. Genet.,* 12(21), 2837-2844.

Teraoka, S. N., Telatar, M., Becker-Catania, S., Liang, T., Onengut, S., Tolun, A., Chessa, L., Sanal, O., Bernatowska, E., Gatti, R. A., and Concannon, P. (1999) Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. *Am. J. Hum. Genet.*, 64, 1617-1631.

Toomajian, C., and Kreitman, M. (2002) Sequence Variation and Haplotype Structure at the Human HFE Locus. *Genetics,* 161, 1609-1623.

Trikka, D., Fang, Z., Renwick, A., Jones, S. H. , Chakraborty, R., Kimmel, M., and Nelson D. L. (2002) Complex SNP-based haplotypes in three human helicases: implications for cancer association studies. *Genome Res,* 12, 627-639.

Wall, J. D. (1999) Recombination and the power of statistical tests of neutrality. *Genet. Res.*, 74, 65-79.

Wall, J. D. (2001) Coalescent Simulations and Statistical tests of Neutrality. *Mol. Biol. Evol.*, 18, 1134-1135.

Wu, L., and Hickson, I. D. (2003) The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature*, 426, 870-874.

Yusa, K., Horie, K., Kondoh, K. G. , Kouno, M., Maeda, Y., Kinoshita, T., and Takeda, J. (2004) Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. *Nature*, 429, 896-899.

Zhang, J. (2003) Evolution of the Human ASPM Gene, a Major Determinant of Brain Size. *Genetics*, 165, 2063-2070.