Seeking for Genetic Signature of Radiosensitivity- Methods for Data Analysis

Joanna Zyla[1], Paul Finnon[2], Robert Bulman[2], Simon Bouffler[2], Christophe Badie[2] and Joanna Polanska[1]

[1] Silesian University of Technology, Institute of Automatic Control, 44-100 Gliwice, ul. Akademicka 16, 44-100 Gliwice,, Poland
{joanna.zyla, joanna.polanska}@polsl.pl
[2]Health Protection Agency, Centre for Radiation, Chemical & Environmental Hazards, Chilton, Didcot, Oxfordshire, OX11 ORQ, United Kingdom
{paul.finnon, robert.bulman, simon.bouffler, christophe.badie}@hpa.org.uk

Abstract. The aim of the study was to develop a data analysis strategy capable of discovering the genetic background of radiosensitivity. Radiosensitivity is the relative predisposition of cells, tissues, organs or organisms to the harmful effect of radiation. Effects of radiation include the mutation of DNA . Identification of polymorphisms and genes responsible for an organism's radiosensitivity increases the knowledge about the cell cycle and the mechanism of radiosensitivity, possibly providing the researchers with a better understanding of the process of carcinogenesis. To obtain this information, mathematical modelling and data mining methods were used.

1 Introduction

It appears that there are variations in individual responses to radiations, and one of the main issues for future research in radiation protection is the identification of those most at risk in terms of radiation-induced cancer. As with sensitivity to sunlight or to chemotherapeutic drugs, sensitivity to ionising radiation shows variation between individuals. The quantification of the cancer risk associated with ionising radiation requires the mapping and the identification of the genes that affect risk. This will eventually lead to the prediction of individual sensitivity and the evaluation of the risk to individuals. Although a large amount of data has already been obtained, the identification of genes potentially involved in radiosensitivity for the prediction of individual cancer risk is not completed yet and further analysis is required.

2 Materials and methods

2.1 Materials

The data for radiosensitivity evaluation were the results of the G2 chromosomal radiosensitivity assay (G2CR). In the pilot study, the test was performed on 14 inbred mice strains presented in Table 1. From each mouse, splenocytes (irradiated with a dose of 0.5Gy in the G2 cell cycle phase) were isolated. In next step, the measurements (numbers of DNA breaks and gaps per 100 cells) were performed 1, 2, 3, 4 and 5 hours after irradiation.

Table 1. Table of mice strains tested in G2CR assay

No.	Mouse Strain	No.	Mouse Strain
1	A/J	8	C57Bl/6J
2	AKR/J	9	DBA/2J
3	Balb/cAn	10	LP
4	Balb/cByJ	11	NOD/NH
5	C3H/HeHsd	12	NOD/LtJ
6	CBA/Ca	13	NZB/B1NJ
7	CBA/H	14	SJL/J

The second group of data comes from widely available sources of genotyped SNPs for mice. In this study of inter-mouse genetic variation the CGD SNP [1] database was used as the resource. The database contains 7.85 million SNPs genotyped for 74 mice strains. The detailed information on data available for the mouse strains under investigation presents Table 2.

Table 2. Number of SNPs (loci) genotyped for all analysed mouse strains.

Chromosome	No. of SNPs	Chromosome	No. of SNPs
1	694 366	11	258 748
2	520 483	12	395 053
3	507 286	13	397 581
4	476 118	14	345 482
5	494 216	15	337 079
6	508 735	16	304 953
7	405 410	17	265 557
8	444 234	18	289 416
9	361 325	19	221 786
10	398 909	X	222 912

2.2 Mathematical modelling and Gaussian mixture model

The kinetics of the chromosomal aberrations' repair was modelled as an exponential function in time (Eq.1), with two parameters k (gain, responsible for the level of chromosomal aberration) and T (time constant, related to the speed of the DNA repair process) estimated with the use of the least squares method.

$$G2(t) = \begin{cases} k * e^{-\frac{t}{T}} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (1)$$

The distributions of individual values of each of these two parameters, together with AUC (area under the curve) were subjected to Gaussian Mixture Model (GMM) decomposition [2]. It allows for identification of mice subpopulations characterized by different kinetics of DNA repair. Another option could be any standard clustering technique, like k-means method, allowing for parameter grouping. However, in contrast to k-means, GMM has the ability to create soft boundaries between clusters - a point in space can belong to any class with a certain probability. Mixture model is used when the data follow a distribution being a mixture of basic distributions (for example Gaussian distributions) - which makes the probability density function a convex combination of the various probability functions. The optimal numbers of Gaussian components were chosen with the use of Bayesian Information Criterion (BIC).

2.3 SNP selection

In the presented study, we are looking for genetic signature differentiating the radiation response in subpopulations of mice (detected by the methodology presented in section 2.2). This statement led us to the following selection process of SNPs: if the genotyping of a given SNP for all mice strains assigned to the radiosensitivity subpopulation is be the same and simultaneously is different but identical among strains classified as normal, the SNP is deemed relevant. To better understand this process, Table 3 presents examples of relevant SNPs.

Table 3. Examples of relevant SNPs selected for the further analysis, where: "R.S mice" represent radiosensitive strains;"A", "G", "T" represent genotype of SNP.

SNP ID	R.S. mice	Normal mice	Normal mice	Normal mice	Normal mice	R.S. mice	Normal mice	R.S. mice
SNP1	A	G	G	G	G	A	G	A
SNP2	G	T	T	T	T	G	T	G

2.4 The distribution of relevant SNPs along the chromosomes

To verify the hypothesis on differences in frequency of polymorphic loci between mouse strains showing high and low induction of chromosome aberrations after irradiation, Fisher's exact test was performed per each loci along every chromosome.

Next step involves r-scan test, which allows for testing the null hypothesis that the locations of chosen loci are iid *(independent and identically distributed)* uniformly distributed random variables with range [0, L]. The alternative hypothesis states that points occur in an overly dispersed fashion [3, 4].

2.5 Analysis of nonsynonymous SNPs

The selection process led to the identification of relevant SNPs, the most interesting of those being nonsynonymous SNPs (nsSNP). Polymorphisms of this type result in an amino acid change. If the nucleobase change does not lead to a change of the amino acid in the protein sequence, the polymorphism is called synonymous.
To assess the impact of nsSNP to the organism, widely available algorithms were used: PHANTER [5], PhD-SNP [6], SNAP [7], SIFT [8] and PolyPhen-2 [9]. Each of them predict, with some probability, if the amino acid change could cause a deleterious effect. Most of the algorithms use the information about protein sequence conservation. Some of them (e.g. PolyPhen-2) are using additional information about annotation and protein structure. Additionally, when nsSNPs were substitution of amino acids involved in the process of phosphorylation (changing Serine, Threonine or Tyrosine), it is possible to assess the group of protein kinases (PK) that could be blocked in this position. For this problem the algorithm GSP 2.1[10] was used.

2.6 *In silico* prediction for other radiosensitive strains of mice.

The data used in the performed analysis were based on chosen mouse strains with measured radiosensitivity. However, the CGD database contains information for 74 strains. For the remaining strains of mice, the estimation of other radiosensitive strains of mice was performed. To carry out calculations, only the relevant SNPs were taken. A reference group of mice was taken from the radiosensitive group. Then the similarity of genotypes between the group of radiosensitive mice and each of the remaining mouse strain were evaluated. Similarity was understood as the percentage of identically genotyped relevant SNPs along the genome. The most similar to the radiosensitive mice mouse strains were defined as the mild outliers in the similarity measure distribution, to detect them the method proposed by Hubert and Vandervieren [11] was used.

3 Results

According to the methodology presented in section 2.2, the set of individual kinetics models was obtained, and the distributions of gain parameter (k), time constant (T) and area-under-curve (AUC) values were decomposed into GMMs. Final model consisted of two components for k and AUC, and one component for T parameter. The decomposition of the k and AUC parameters allowed for identification of threshold value and the detection of two subpopulations of mice (Fig. 1) - one of them repre-

sents increased radiosensitivity, the second one represents normal response. Figure 2 shows the kinetic of DNA repair on mice with GMM-distinguished subpopulations.

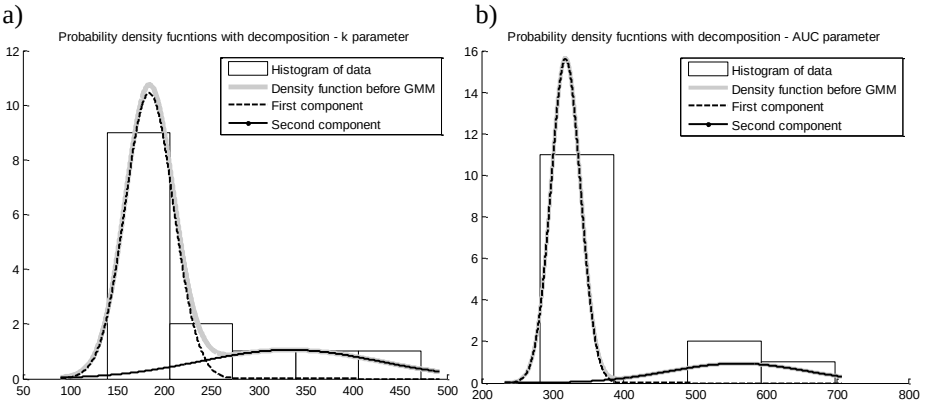

Fig. 1. The GMM decomposition of kinetic model parameters. a) The decomposition of the k parameter. b) the decomposition of the AUC parameter.

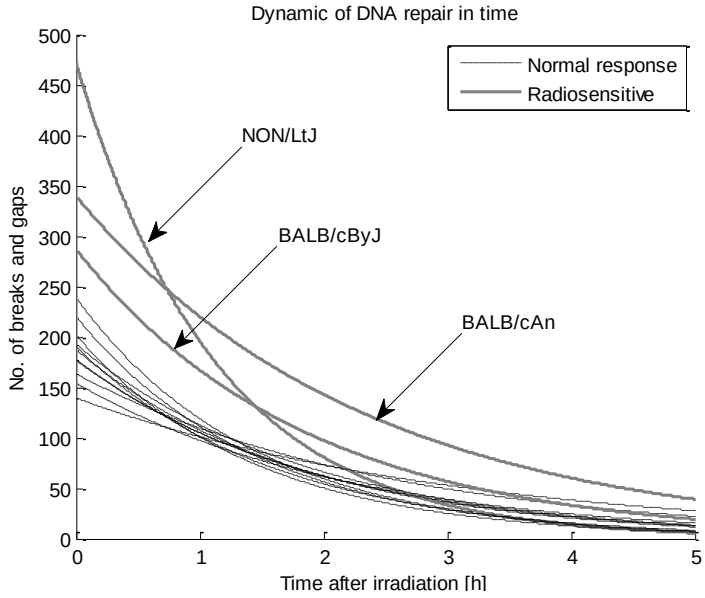

Fig. 2. Kinetics of DNA repair for mice tested in G2CR Grey lines represent strains of mice detected as radiosensitive. Black, dash lines represent strains of mice with normal DNA repair.

The following mice strains were classified into the group of radiosensitive: BALB/cAn, BALB/cByJ and NON/LtJ. Similar classification results were obtained by applying the outlier detection technique proposed by Hubert and Vandervieren [11] directly to the distribution of the analyzed parameters. With knowledge about

mice subpopulations, the SNP selection process was performed following algorithm described in section 2.3. Since the average overall rare allele frequency is equal to 9.62% (Table 4), the probability of observing such structure for single SNP is equal to: $p = 0.0962^3 (1-0.0962)^{11} = 0.0029263$.

Taking under consideration multiple testing performed for all available loci, the expected number of false discoveries equals to 2297. While applying the proposed methodology to our data we get 1856 relevant SNPs. Table 5 presents the distribution of the relevant SNPs across the genome. Detailed inspection of distribution of the relevant SNPs along the chromosomes shows that there are some chromosomes with significantly higher number of relevant SNPs and other chromosomes with significantly lower number of that type of loci. It suggests that there are chromosomes with clumped distribution of relevant polymorphic loci. To check on this, the r-scan test was applied to verify the hypothesis on uniformity of location [3]. Applying r-Scan U_{max} test (r=1) to relevant SNPs distribution along chromosomes gives p-values much less than 1e-12 and allows for rejection of null hypotheses that relevant SNPs are not clumped. Figure 3 present graphical illustration of Relevant SNPs distribution. The same analysis might be done with the use of ChromoScan software [4].

Table 4. Estimation of variant (rare) allele frequency, pSNPs stands for polymorphic loci.

Chr	No of SNPs	pSNPs		Overall variant freq	Chr	No of SNPs	pSNPs		Overall variant freq
		N	variant freq				N	variant freq	
1	694366	341063	22.62%	11.11%	11	258748	146431	21.39%	12.11%
2	520483	262264	19.99%	10.07%	12	395053	198106	20.62%	10.34%
3	507286	229264	23.42%	10.59%	13	397581	181873	20.01%	9.16%
4	476118	219418	21.97%	10.13%	14	345482	213003	20.98%	12.94%
5	494216	211104	21.05%	8.99%	15	337079	150984	20.91%	9.37%
6	508735	232322	20.74%	9.47%	16	304953	114988	19.99%	7.54%
7	405410	207630	19.50%	9.99%	17	265557	135794	21.48%	10.98%
8	444234	218490	18.36%	9.03%	18	289416	119201	21.14%	8.71%
9	361325	183990	17.64%	8.98%	19	221786	94203	20.53%	8.72%
10	398909	125856	20.10%	6.34%	X	222912	44286	24.91%	4.95%

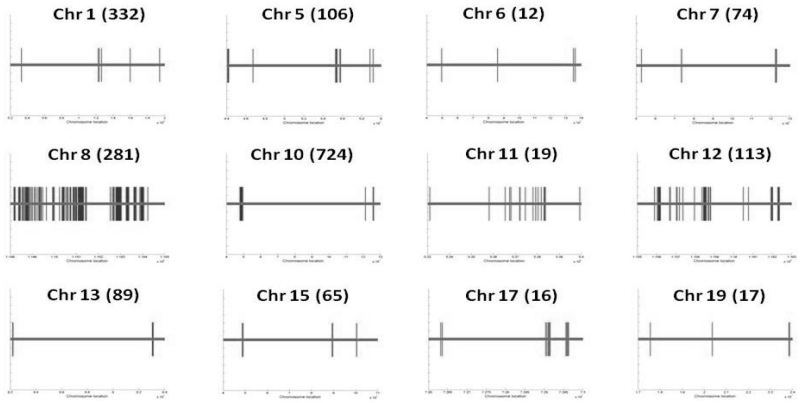

Fig. 3. The clumping of relevant SNPs along the chromosomes.

Table 5. Distribution of relevant SNPs along the chromosomes. #exp – expected number of relevant SNPs; #obs – number of observed relevant SNPs; p-value – Fisher exact test for overrepresentation of the relevant SNPs (- light greay) or underrepresentation of the relevant SNPs (-dark greay).

Chr	No of SNPs	Variant freq [%]	# exp	# obs	p-value	Chr	No of SNPs	Variant freq [%]	# exp	# obs	p-value
1	694366	11.11	260	332	0.0017	11	258748	12.11%	111	19	<1e-6
2	520483	10.07	165	0	<1e-6	12	395053	10.34%	131	113	ns
3	507286	10.59	175	6	<1e-6	13	397581	9.16%	106	89	ns
4	476118	10.13	152	0	<1e-6	14	345482	12.94%	163	0	<1e-6
5	494216	8.99	127	106	ns	15	337079	9.37%	93	65	0.0157
6	508735	9.47	144	12	<1e-6	16	304953	7.54%	55	2	<1e-6
7	405410	9.99	126	74	0.0001	17	265557	10.98%	98	16	<1e-6
8	444234	9.03	115	281	<1e-6	18	289416	8.71%	70	0	<1e-6
9	361325	8.98	92	0	<1e-6	19	221786	8.72%	54	17	6.3e-6
10	398909	6.34	49	724	<1e-6	X	222912	4.95%	15	0	3.1e-5

In total 1856 relevant SNPs were detected, nonuniformly distributed along the chromosomes. The detailed analysis of these SNPs revealed that 47 of 1856 are located in exons, 882 in introns, 13 in UTR regions, and 914 in intergenic regions. Eight SNPs appeared to be nonsynonymous (nsSNP). It was shown that relevant SNPs concentrate in 29 clusters located in 28 genes. Eight nsSNPs occurred only in 4 genes. Using widely available algorithms to predict an effect of nsSNP on protein function, it was possible to check the influence of the obtained nsSNPs. Additionally nsSNPs with substitution of amino acids involved in the process of phosphorylation were checked with the GPS 2.1 algorithm in order to predict their effect on protein kinases (PK). Two nsSNPs present increased probability of having a deleterious effect and one of

them could disorder phosphorylation with 14 PKs. Genes with large numbers of SNPs or nsSNPs had their gene ontology and signalling pathway participation analyzed.

4 Conclusion

The proposed strategy for data analysis, which is a combination of mathematical modelling and data mining techniques, allowed for the discovery of the candidate genetic signature of radiosensitivity. From the group of differentiating genes, two of them are, according to the literature study, highly significant for the analyzed phenomena of radiosensitivity. One might be responsible for the process of DNA damage repair. The second is indirectly responsible for cell adhesion and it was observed to be up regulated in breast cancer patients that are one of the groups more frequently exposed for the radiation does. The biological and functional validation of the obtained relevant SNPs is necessary and will be performed very soon.

5 References

1. Szatkiewicz, J.P, at.el.: An imputed genotype resource for the laboratory mouse. Mamm. Genome (2008) 19(3):199-208
2. McLachlan, G.J.; Peel, D.: Finite Mixture Models. Wiley (2000)
3. Ewens, W., Grant, G.: Statistical methods in bioinformatics. An introduction. Springer (2001)
4. Karlin, S., Macken, C.A.: Assesment of inhomogeneities in E.Coli physical map. Nucleic Acids Research 1991, 19:4241-4246
5. Thomas, P.D., Kejariwal, A.: Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences molecular effects. PNAS. (2004) 101(43):15398-15403
6. Capriotti, E., Calabrese, R., Casadio, R.: Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. (2006) 22:2729-2734
7. Bromberg, Y., Rost, B.: SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Aicid Res. (2007) 34(11):3823-3835
8. Kumur, P., Henikoff, S., Ng, P.C.: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. (2009) 4(7):1073-1081
9. Adzhubei, I.A., at.el.: A method and server for predicting damaging missense mutations. Nat. Methods. (2010) 7(4):248-249
10. Xue, J., at. el.: GPS 2.0,a Tool to Predict Kinase-specific Phosporylation Sites in Hierarchy. Mol. Cell. Proteomics. (2008) 7:1598-1608
11. Hubert, M., Vandervieren, E.: An adjusted boxplot for skewed distributions. Computational Statistics & Data Analysis. (2008) 52(12):1933-1940

6 Acknowledgement

The work was partially financially supported by MNiSW grant NN519579938 and SUT grant BK-2013.