DNA Repair 16 (2014) 97–103

Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems

Manika Sehgal, Tiratha Raj Singh*

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Waknaghat, Solan, HP 173234, India

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form
13 December 2013
Accepted 22 January 2014
Available online 16 February 2014

Keywords:
GAS
Haplotype
nsSNP
LD
Phosphorylation

ABSTRACT

We present DR-GAS[1], a unique, consolidated and comprehensive DNA repair genetic association studies database of human DNA repair system. It presents information on repair genes, assorted mechanisms of DNA repair, linkage disequilibrium, haplotype blocks, nsSNPs, phosphorylation sites, associated diseases, and pathways involved in repair systems. DNA repair is an intricate process which plays an essential role in maintaining the integrity of the genome by eradicating the damaging effect of internal and external changes in the genome. Hence, it is crucial to extensively understand the intact process of DNA repair, genes involved, non-synonymous SNPs which perhaps affect the function, phosphorylated residues and other related genetic parameters. All the corresponding entries for DNA repair genes, such as proteins, OMIM IDs, literature references and pathways are cross-referenced to their respective primary databases. DNA repair genes and their associated parameters are either represented in tabular or in graphical form through images elucidated by computational and statistical analyses. It is believed that the database will assist molecular biologists, biotechnologists, therapeutic developers and other scientific community to encounter biologically meaningful information, and meticulous contribution of genetic level information towards treacherous diseases in human DNA repair systems. DR-GAS is freely available for academic and research purposes at: http://www.bioinfoindia.org/drgas.

1. Introduction

DNA repair is a very complex and vital process through which a cell recognizes damage to the DNA caused by endogenous or environmental insults, as well as genetic defects that result in incomplete repair. The cell tries to repair these damages to retain the integrity of their genome. DNA repair is present in both prokaryotes and eukaryotes, whereas in the later the genome and the repair mechanisms are much more complex [1]. Various factors involved for incorporating changes or aberrations in DNA molecule such as reactive oxygen species [2], replication errors, ultraviolet radiations, X-rays, gamma rays, thermal disruption and viruses can all result in the DNA damage [3]. There is a high rate of recurrence for endogenous DNA damage as compared to exogenous damage and the type of damages produced due to both factors is roughly indistinguishable [4]. The damage to the DNA is caused by multiple factors such as oxidation of bases, generation of DNA strand interruptions, alkylation of bases [5], bulky adduct formation, mismatches and pyrimidine dimers that often trigger viral interactions [6].

The elimination of damaged DNA from the genome is a complicated process involving a number of repair proteins like *DDB2*, *MLH1*, *XPA* and different associated mechanisms for diverse type of lesions. The numerous known mechanisms by which the damaged DNA is repaired includes BER, NER, MMR, HRR, NHEJ, DDS and TLS which have different set of genes, enzymes and pathways for repairing the DNA. These mechanisms not only maintain the genetic stability but also prevent the genome from carcinogenesis, pre-mature aging, cockayne syndrome, xeroderma pigmentosum, progeria and several other disorders [7–20] which we are intended to analyze in this study for better understanding of the entire process.

Innumerable genetic sequence patterns comprising of common haplotype blocks, essential genetic markers and LD plots are associated with DNA repair related disorders, where some of the DNA repair genes have already been analyzed for its strong involvement with the genetic diseases. The *XRCC1* DNA repair gene is said to be involved in pancreatic cancer and its haplotype analysis

Abbreviations: BER, base excision repair; NER, nucleotide excision repair; MMR, mismatch repair; HRR, homologous recombination repair; NHEJ, non-homologous end joining; TLS, trans lesion synthesis; DDS, DNA damage signaling; HNPCC, hereditary non-polyposis colorectal cancer; LD, linkage disequilibrium; nsSNPs, non-synonymous SNPs; NCBI, national center for biotechnology information; OMIM, online mendelian inheritance in man; dbSNP, single nucleotide polymorphism database; GAD, genetic association database; KEGG, Kyoto encyclopedia for genes and genomes; SIFT, sorting intolerant from tolerant.

* Corresponding author. Tel.: +91 01792 239385.
E-mail addresses: manika.sehgal22@gmail.com (M. Sehgal), tiratharaj@gmail.com (T.R. Singh).

[1] DNA Repair-Genetic Association Studies.

1568-7864/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dnarep.2014.01.004

showed a strong statistical association with the disorder [21]. Saadat et al. [22] demonstrated that preeclampsia disorder is linked with higher frequency of "194R-399Q" haplotype in *XRCC1* gene with a confidence of 95% as compared to the control. Moreover, variations in *ERCC5* repair gene have been reported in gastric cancer which serves as an important marker for the disease [23]. Similar studies on *ERCC1* and *ERCC2* DNA repair genes were performed to prominently demonstrate the association of the two genes with lung adenocarcinoma [24]. Above mentioned studies suggest the involvement and association of LD and common haplotype patterns with countless DNA repair disorders [25]. Therefore, there is a need to analyze these indispensible genetic parameters in an efficient way to understand the mechanisms of DNA repair related disorders.

The intrinsic properties of many DNA repair proteins are found to be affected by the altered phosphorylation sites, since its state may govern the risk of developing cancers [26–28]. The phosphorylation takes place at serine (S), threonine (T) or tyrosine (Y) residues [29] in the proteins which not only influences the structure but also affects the function, stability, sub-cellular localizations and interaction with other proteins [30,31]. Few cases of key DNA repair proteins, where nsSNPs such as S31R (CDKN1A), S326C (OGG1) and T241M (XRCC3) have already been associated with risks for endometrial [32], esophageal [33,34], lung [35] and breast cancers [36] due to change in their phosphorylation states. The change may be from phosphorylated to dephosphorylated residue or vice-versa affecting the activity of the repair proteins. The change due to nsSNPs in DNA repair proteins is also found to be defensive against certain disorders, for example T241M mutation in XRCC3 protein is protective against bladder cancer in heavy smokers [37]. Since, the phosphorylation states play an imperative role in the regulation of multitude of cellular processes, gene expression, signal transduction, apoptosis, homeostasis and DNA damage recognition and its repair [38], it is important to thoroughly analyze the phosphorylation states of diverse DNA repair proteins.

In human molecular systems, DNA repair is a very crucial process for which the main challenge lies in the development of a platform where one could easily access and retrieve the integrated information for several genetic parameters involved in DNA repair. Currently, not many resources are available which provide information on DNA repair. Few such resources are REPAIRtoire [39] and Repair funmap [40], while to the best of our knowledge there is no database till date which provides all the information associated with genetic parameters of human DNA repair system in a comprehensive way. Additionally, none of the available resources provide information concerning the functional association of nsSNPs and their phosphorylation states for human DNA repair system.

Keeping in view all the above mentioned requirements and to fill up this research gap for DNA repair systems, first a widespread computational analysis on the genotype data was performed and then a database named, DR-GAS (DNA Repair Genetic Association Studies) was compiled for various genetic features, for instance haplotype blocks, LD plots, essential genetic markers and their respective statistical parameters. This database also includes nsSNPs and their putative functional effect on the genome through their phosphorylation states amongst all DNA repair mechanisms. DR-GAS database is a unique and most comprehensive database regarding DNA repair genes which include their involvement in various repair mechanisms, associated pathways and diseases which could be of utmost use to the researchers involved in the study of DNA repair.

2. Materials and methods

In this study, we applied an integrated approach which is a combination of *in silico* and quantitative genetic studies, being performed on 215 DNA repair genes, their proteins, associated pathways, diseases, etc., obtained from NCBI and other published studies. On the basis of literature and information collected from numerous relevant resources, we categorized all DNA repair genes into major 16 classes as shown in Fig. 1.

2.1. Database design and content

To build DR-GAS database, we designed a comprehensive workflow (refer graphical abstract) which explains the overall process of data collection, computational analysis, database design and implementation. It consists mainly of 4 parts i.e. (i) Collection of genotype data and quantitative genetic studies, (ii) Identification of nsSNPs and their functional effect on human repair systems, (iii) Detection of putative phosphorylation sites, (iv) Collection and computational verification of essential diseases and pathways associated with human DNA repair systems.

2.1.1. Collection of genotype data and quantitative genetic studies

The genotype data has been pulled out from The International HapMap Project [41] and was analyzed for several quantitative genetic parameters. The parameters such as haplotype which is the combination of alleles at neighboring loci on the chromosome being transmitted together, LD i.e., the involvement of alleles in a non-random mode in the population, and additional important genetic markers, for which analysis was performed using Haploview [42]. The main parameters for the genetic association provided by haploview were D' and r^2. D' is the measure of LD between the two blocks which is calculated from the equation:

$$D' = \frac{D}{D_{\max}}$$

$$\text{where,}\quad D = [(F11)(F12) - (F12)(F21)]$$

and "$D_{\max}$" depends upon the sign of D. If D is positive, then

$$D_{\max} = \min[(m1m2)\quad \text{or}\quad (m2m1)]$$

While if, D is negative, then

$$D_{\max} = \min[(m1m1)\quad \text{or}\quad (m2m2)]$$

Value of D in the vicinity of zero provides greater amount of historical recombination between the two blocks. Here, $m1$ and $m2$ are the frequencies of alleles at SNP1, $n1$ and $n2$ are the frequencies of alleles at SNP2 and $F11$, $F12$, $F21$, $F22$ are the possible haplotype frequencies.

Another crucial parameter used was r^2 i.e. the correlation coefficient which is calculated by:

$$r = \frac{D}{(m1m2m1m2)^{1/2}}$$

The squared coefficient of correlation (r^2) is frequently used to eliminate the arbitrary sign thus introduced in the correlation value.

2.1.2. Identification of nsSNPs and their functional effect on human repair systems

The nsSNPs information was gathered by analyzing the DNA repair genes in dbSNP [43] and other SNP based databases [44]. Investigating the effect of SNPs in coding sequences is very complex and expensive through experimental methods, consequently the genetic mutations in the genes have been linked to deviations in the phenotypes using a bioinformatics algorithm i.e. SIFT [45]. This algorithm is purposeful to study the genetic variants that may affect the phenotypic characteristics. SIFT prediction tool has been used for the protein conservation analysis which is based on the

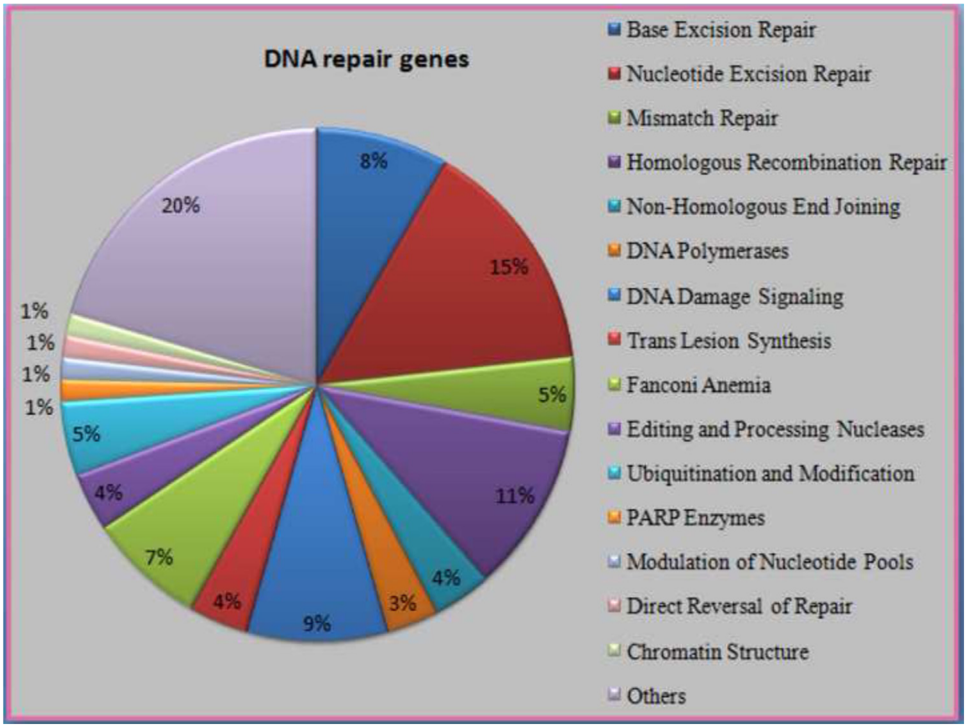


Fig. 1. DNA repair genes and associated pathways.
215 DNA repair genes are represented in DR-GAS. The percentage (%) shows the number of genes present in each pathway.

principle that protein evolution has a strong correlation with protein function. The highly conserved positions suffer from fewer substitutions whereas the ones weakly conserved, can tolerate more substitutions. If the SIFT score is less than the threshold value, the substitution is said to have an effect on the protein otherwise no major changes in the protein are confirmed.

2.1.3. Detection of putative phosphorylation sites

The key amino acids within the repair proteins could be analyzed by identifying their phosphorylation sites and such predictions could provide insights into the biochemical actions of the analyzed proteins. For the prediction of potential phosphorylation sites at S, T, and Y residues in the repair protein sequences, NetPhos [46] algorithm was used. It is based on artificial neural networks which are extensively trained from various samples. For the training of NetPhos program, PhosphoBase [47], a database of phosphorylated proteins has been used for pattern recognition. The analysis of phosphorylation state of repair proteins is a significant phase of this study as many cellular processes and DNA damage recognition and its repair mechanism are affected by it.

2.1.4. Collection and computational verification of essential diseases and pathways associated with human DNA repair systems

The information for a variety of diseases related to DNA repair genes like multiple kind of cancers, xeroderma pigmentosum, cockayne syndrome and fanconi anemia were collected from the literature and databases such as OMIM [48], GAD [49] and Gene Cards [50]. The information thus obtained was integrated and represented in Table 1 along with the number of predicted nsSNPs in each mechanism. The pathways wherein there is association of several DNA repair genes were collected and analyzed from GenBank annotations and variety of additional resources like KEGG [51], Gene Cards, REACTOME [52], etc. to verify their respective entities. This process applied manual curation of data to remove any likelihood of redundancy in the ultimate collected information.

3. Results and discussions

In this study, 215 DNA repair genes are categorized on the basis of the mechanisms in which they are involved. DR-GAS database provides an easy and effective way for the search and retrieval of genetic essential markers and associated information for repair genes as shown in Fig. 2. We have collected the LD, haplotype, markers, nsSNPs, pathways and disease related information for 215 repair genes which have been classified into main pathways such as BER, NER, MMR, HRR, NHEJ, DDS, TLS, DNA Polymerases,

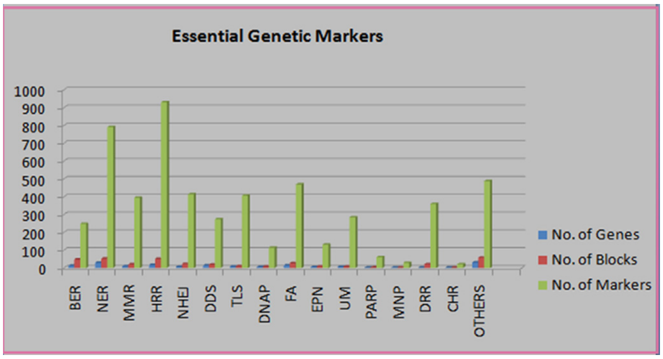
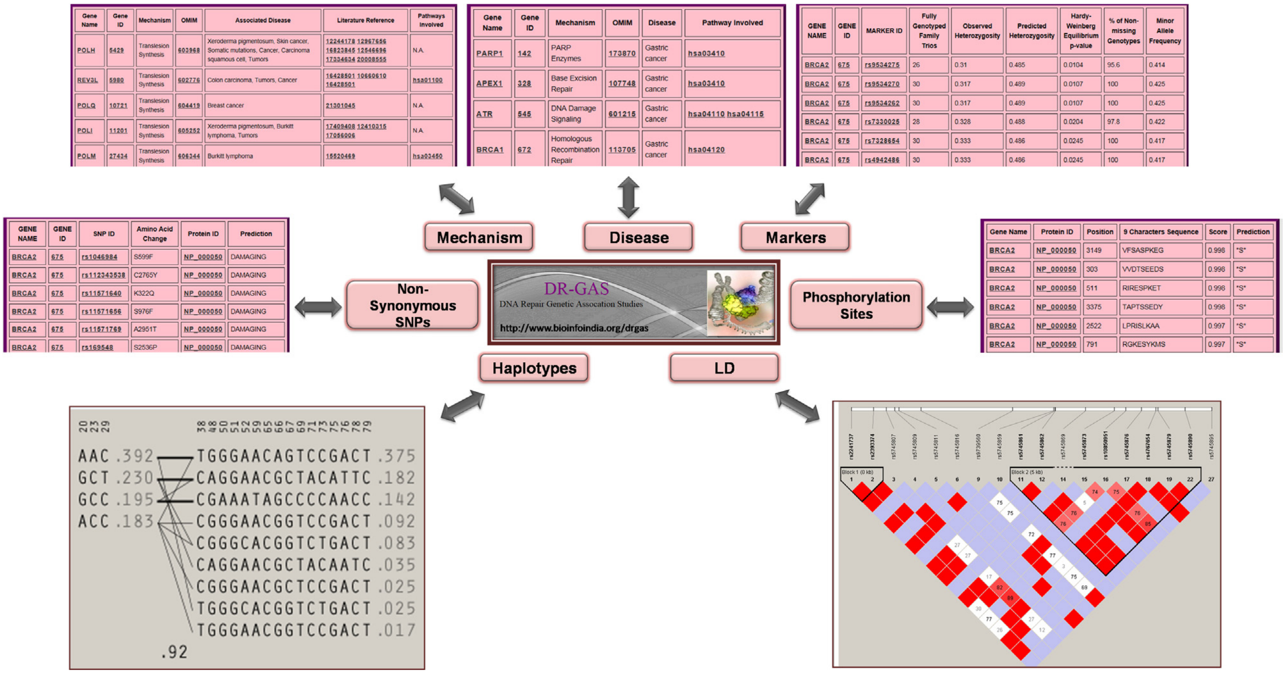


Fig. 2. The indispensable genetic markers for various DNA repair mechanisms.
Statistical distribution of number of genes in different mechanisms, important haplotype blocks and genetic markers associated with these genes. BER stands for Base Excision Repair, NER is Nucleotide Excision Repair, MMR is Mismatch Repair, HRR is Homologous Recombination Repair, NHEJ is Non-Homologous End Joining, DDS is DNA Damage Signaling, TLS is Translesion Synthesis, DNAP is DNA Polymerases, FA is Fanconi Anemia, EPN is Editing and Processing Nucleases, UM is Ubiquitination and Modification, PARP is PARP Enzymes, MNP is Modulation of Nucleotide Pools, DRR is Direct Reversal of Repair, CHR is Chromatin Structure.

Table 1
No. of predicted nsSNPs that alter amino acid sequence and associated diseases with various DNA repair mechanisms.

Mechanisms	No. of genes	Gene names	No. of predicted damaging nsSNPs	Associated major disease
Base Excision Repair	18	MBD4, MPG, MUTYH, NEIL1, NEIL2, NEIL3, NTHL1, OGG1, APEX1, APEX2, LIG3, PNKP, XRCC1, SMUG1, TDG, UNG, APLF, HUS1	40	Rett Syndrome, Angelman Syndrome, Nsclc, Lynch syndrome, Xeroderma pigmentosum, Parkinson disease, Alzheimers disease, Diphtheria, Bloom syndrome, Ataxia telangiectasia and various kinds of cancers.
Nucleotide Excision Repair	32	CCNH, CDK7, CETN2, DDB1, DDB2, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ERCC6, ERCC8, GTF2H1, GTF2H2, GTF2H3, GTF2H4, GTF2H5, LIG1, MMS19L, MNAT1, RAD23A, RAD23B, RPA1, RPA2, RPA3, TFIIH, XAB2, XPA, XPC, UVSSA, RFC1, CUL4A	122	Retinoblastoma, Leukemia, Lymphoma, Xeroderma pigmentosum, Alzheimers disease, Nsclc, Cockayne syndrome, Trichothiodystrophy, Glioma, Necrosis, Fanconianemia, Ataxia telangiectasia and various kinds of cancers.
Mismatch Repair	10	MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1, PMS2, PMS2P3	246	Lynch syndrome, HNPCC, Muir-torre syndrome, Turcot syndrome, Werner syndrome, Hnscc, Nsclc, Peutz-jeghers syndrome and various kinds of cancers.
Non-Homologous End Joining	8	DCLRE1C, LIG4, NHEJ1, PRKDC, XRCC4, XRCC5, XRCC6, WRN	48	Omenn syndrome, Ataxia telangiectasia, Nijmegen breakage syndrome, Lig4 syndrome, Leukemia, Bloom syndrome and various kinds of cancers.
Homologous Recombination Repair	23	BRCA1, DMC1, EME1, EME2, GIYD1, GIYD2, MRE11A, MUS81, NBN, RAD50, RAD51, RAD51L1, RAD51L3, RAD52, RAD54B, RAD54L, RBBP8, SHFM1, XRCC2, XRCC3, BLM, RAD51B, RAD51C	74	Fanconi anemia, Ataxia telangiectasia, Cowden disease, Nijmegen breakage syndrome, Bloom syndrome, HNPCC, Werner syndrome, Canavan disease, Xeroderma pigmentosum, Lynch syndrome, Sickle cell disease, Hodgkin disease, Leukemia, Anemia, Hnscc, Nsclc and various kinds of cancers.
DNA Damage Signaling	19	ATM, ATR, ATRIP, CDKN1A, CHEK1, CHEK2, DCLRE1A, DCLRE1B, GPS1, MDC1, RAD1, RAD9A, RAD17, RFC2, RFC3, RFC4, RFC5, TOPBP1, TP53	92	Ataxia telangiectasia, Osteoporosis, Atherosclerosis, Necrosis, Nsclc, Hnscc, Xeroderma pigmentosum, Cardiovascular diseases, Cockayne syndrome, Alzheimers disease, Malaria, Pituitary diseases, Burkitt lymphoma, Mental retardation, Thalassemia, Epilepsy, Gaucher disease, Aids and Liver cirrhosis.
Trans Lesion Synthesis	8	POLM, POLN, POLQ, REV1L, POLH, POLI, POLK, REV3L	76	Mitochondrial diseases, Werner syndrome, Glioma, Alpers syndrome, Epilepsy, Neurodegenerative diseases, Liver diseases, Parkinson disease, Xeroderma pigmentosum and various kinds of cancers.
Fanconi Anemia	16	BRCA2, BRIP1, BTBD12, FAAP24, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCL, FANCM, FANCI, PALB2, C1ORF86	762	Fanconi anemia, Ataxia telangiectasia, Bloom syndrome, Lynch syndrome, HNPCC, Tay-sachs disease, Necrosis, Hodgkin disease, Hnscc, Nijmegen breakage syndrome, Alzheimers disease and various kinds of cancers.
Editing and Processing Nucleases	8	APTX, EXO1, FEN1, MTMR15, SPO11, TREX1, TREX2, TTRAP	14	Ataxia telangiectasia, HNPCC, Neurodegenerative diseases, Werner syndrome, Parkinson’s disease and various kinds of cancers.
Ubiquitination and Modification	10	HLTF, RAD18, RNF4, RNF8, RNF168, SHPRH, UBE2A, UBE2B, UBE2N, UBE2V2	9	Colorectal cancer, Colon cancer, Gastric cancer, Adenoma, Papilloma.
PARP Enzymes	3	PARP1, PARP2, PARP3	14	Ataxia telangiectasia, Werner syndrome, Fanconi anemia, Cockayne syndrome, Parkinson disease, Diabetes mellitus, Asthma and various kinds of cancers.
Modulation of Nucleotide Pools	3	DUT, NUDT1, RRM2B	–	Herpes simplex, Hiv infections, Parkinson disease, Dysplasia and various kinds of cancers.
Direct Reversal of Damage	3	ALKBH2, ALKBH3, MGMT	10	Glioblastoma, Nsclc, Stable disease, Xeroderma pigmentosum, Viral infection and various kinds of cancers.
Chromatin Structure	3	CHAF1A, H2AFX, SETMAR	–	Ataxia telangiectasia, Nijmegen breakage syndrome, Bloom syndrome, Fanconi anemia, Necrosis, Leukemia and various kinds of cancers.
DNA Polymerases	7	PCNA, POLB, POLD1, POLE, POLG, MAD2L2, POLL	24	Werner syndrome, Aids, Glioma, Alpers syndrome, Mitochondrial diseases, Epilepsy, Neurodegenerative diseases, Liver diseases, Parkinson disease, Xeroderma pigmentosum, Burkitt lymphoma and various kinds of cancers.
Others[a]	44	TDP1, ABL1, ADA, ATRX, BARD1, TTDN1, CIB1, CLK2, COPS2, CRY1, GADD45A, GADD45G, HEL308, LHX3, OBFC2B, PER1, POLD3, POLDIP2, POLDIP3, POLE2, POLE3, POLE4, POLR2E, POLR2F, POLR2H, POLR2K, POLR2L, PRPF19, RAD21, RDM1, RECQL, RECQL4, RECQL5, RPA4, SIRT1, SMC1A, SUMO1, TP53BP1, UNG2, GEN1, PMS6, XSE6, YKU80P, ZFP276	62	Fanconi anemia, Werner syndrome, Bloom syndrome, Progeria, Nijmegen breakage syndrome, Cataract, Ataxia telangiectasia, Osteoporosis, Atherosclerosis, Necrosis, Nsclc, Hnscc, Xeroderma pigmentosum, Cardiovascular diseases, Cockayne syndrome, Alzheimers disease, Malaria, Kallmann syndrome, Pituitary diseases, Burkitt lymphoma, Mental retardation, Thalassemia, Epilepsy, Gaucher disease, Aids, Liver cirrhosis, Brain tumors, Rheumatoid Thyroid cancer, Hematologic disorders and various kinds of cancers.

[a] This category includes all the DNA repair genes which have not been classified on the basis of mechanism in which they are involved.



Gene Name	Gene ID	Mechanism	OMIM	Associated Disease	Literature Reference	Pathways Involved
POLH	5429	Translesion Synthesis	603968	Xeroderma pigmentosum, Skin cancer, Somatic mutations, Cancer, Carcinoma squamous cell, Tumors	12244178 12967656 16823845 12546696 17334634 20008555	N.A.
REV3L	5980	Translesion Synthesis	602776	Colon carcinoma, Tumors, Cancer	16428501 10660610 16428501	hsa01100
POLQ	10721	Translesion Synthesis	604419	Breast cancer	21301045	N.A.
POLI	11201	Translesion Synthesis	605252	Xeroderma pigmentosum, Burkitt lymphoma, Tumors	17409408 12410315 17056006	N.A.
POLM	27434	Translesion Synthesis	606344	Burkitt lymphoma	15520469	hsa03450

Gene Name	Gene ID	Mechanism	OMIM	Disease	Pathway Involved
PARP1	142	PARP Enzymes	173870	Gastric cancer	hsa03410
APEX1	328	Base Excision Repair	107748	Gastric cancer	hsa03410
ATR	545	DNA Damage Signaling	601215	Gastric cancer	hsa04110 hsa04115
BRCA1	672	Homologous Recombination Repair	113705	Gastric cancer	hsa04120

GENE NAME	GENE ID	MARKER ID	Fully Genotyped Family Trios	Observed Heterozygosity	Predicted Heterozygosity	Hardy-Weinberg Equilibrium p-value	% of Non-missing Genotypes	Minor Allele Frequency
BRCA2	675	rs9534275	26	0.31	0.485	0.0104	95.6	0.414
BRCA2	675	rs9534270	30	0.317	0.489	0.0107	100	0.425
BRCA2	675	rs9534262	30	0.317	0.489	0.0107	100	0.425
BRCA2	675	rs7330025	28	0.328	0.488	0.0204	97.8	0.422
BRCA2	675	rs7328654	30	0.333	0.486	0.0245	100	0.417
BRCA2	675	rs4942486	30	0.333	0.486	0.0245	100	0.417

GENE NAME	GENE ID	SNP ID	Amino Acid Change	Protein ID	Prediction
BRCA2	675	rs1046984	S599F	NP_000050	DAMAGING
BRCA2	675	rs112343538	C2765Y	NP_000050	DAMAGING
BRCA2	675	rs11571640	K322Q	NP_000050	DAMAGING
BRCA2	675	rs11571656	S976F	NP_000050	DAMAGING
BRCA2	675	rs11571769	A2951T	NP_000050	DAMAGING
BRCA2	675	rs169548	S2536P	NP_000050	DAMAGING

Gene Name	Protein ID	Position	9 Characters Sequence	Score	Prediction
BRCA2	NP_000050	3149	VFSASPKEG	0.998	"S"
BRCA2	NP_000050	303	VVDTSEEDS	0.998	"S"
BRCA2	NP_000050	511	RIRESPKET	0.998	"S"
BRCA2	NP_000050	3375	TAPTSSEDY	0.998	"S"
BRCA2	NP_000050	2522	LPRISLKAA	0.997	"S"
BRCA2	NP_000050	791	RGKESYKMS	0.997	"S"

Fig. 3. Demonstration and implementation of DR-GAS.
Web interface of DR-GAS: A database of DNA Repair genes and the Genetic Association Studies with various available search and advanced options. The illustrated output from the repository is represented in a combined image for all the results generated. The haplotype patterns and LD plot is also revealed at the bottom of the image.

Ubiquitination and Modification, Fanconi anemia, Editing and Processing Nucleases. A category referred as ‘Others’ has been created in the database for the genes which are reported as DNA repair genes but are not being classified in any of the main pathways.

3.1. Web interface

The DR-GAS offers the facility to the user to browse the repository using seven different types of search options. One can search for mechanism, nsSNPs, haplotypes, LD, genetic markers, diseases and phosphorylation sites. It also provides an option for the advanced search for the user’s convenience where the hybrid data is provided for few categories. The web interface with required details and some indicative results using DR-GAS is shown in Fig. 3. In the mechanism menu, user can search for any DNA repair mechanism and retrieve information regarding all the genes involved in various mechanisms, their Gene ID’s, OMIM ID’s, pathways involved, associated diseases and the appropriate literature references corresponding to the disorders which are linked to NCBI, OMIM, KEGG and PUBMED databases, respectively.

The nsSNPs in DNA repair genes are easily accessible through nsSNPs search option in DR-GAS. The nsSNP when introduced in the gene sequence can result in a change in amino acid sequence that could, in theory, disrupt function to promote inefficient DNA repair. By clicking on Get nsSNPs button, various nsSNPs for the selected gene (Gene ID) are displayed. SNP ID’s are linked to dbSNP, amino acid change column gives the position of the change in the sequence and the amino acid being replaced, while the prediction column shows whether the change is damaging or not.

In the haplotypes search option, users can acquire exhaustive information regarding the essential haplotypes which is combination of alleles at adjacent locations (loci) on the chromosome that are transmitted together. One can also explore parameters like block which gives the current number of blocks in a particular query gene, number of markers column gives the amount of markers present in a block. On clicking the block option, an easy and inferable view of the haplotypes in the specified gene is generated where the marker numbers are depicted on the top and the tag SNPs (if any) are highlighted with a triangular pointer. Here, haplotypes are the blocks of associated SNPs which are conserved throughout the genome in the form of patterns called “haplotype blocks”. These blocks correspond to the set of consecutive sites which either has small or no indication of historical recombination. Population frequencies of each haplotype are shown and common crossings from one block to the next are represented by lines, where thicker lines portray more common crossings than the thinner ones. The blocks present the correlation of various residue states among the polymorphic sites across the genome. The multilocus D prime value (D') is also being specified at the bottom of the image.

The LD information of various repair genes is analyzed and congregated in the LD search option. Here, we have considered only those loci’s whose r^2 value i.e. the correlation coefficient between the two loci is $\geq$0.6 as these are the most substantial ones. It includes loci 1 and loci 2, which are the two loci under study, D' value between the two loci, LOD which is the log of the likelihood odds ratio i.e. a measure of confidence in the value of D', correlation coefficient value between loci 1 and loci 2, CIlow and CIhi column represents 95% confidence lower bound and upper bound on D', distance (in bases) between the loci, LD image column gives an interactive image of the LD plot thus generated.

In the markers search option, significant markers identified in the study have been compiled and incorporated. It provides a facility to the user to access the marker specific parameters like Marker ID, fully genotyped family trios for the marker (0 for datasets with unrelated individuals), the marker’s observed heterozygosity, predicted heterozygosity of the marker calculated from:

$$[2 \times \mathrm{MAF} \times (1 - \mathrm{MAF})]$$

where MAF is Minor Allele Frequency, Hardy-Weinberg (H-W) equilibrium p value, i.e. the probability that its deviation from H-W equilibrium could be explained by chance, the percentage of non-missing genotypes for the marker and MAF for the given marker. Additionally, moving further to the database options, there is a disease menu, which comprises of numerous diseases that have been

Table 2
Identification and verification of phosphorylation sites in DNA repair proteins.

Mechanisms	No. of proteins	Predicted phosphorylated residues			Experimental validations (PubMed IDs)[a]
		S	T	Y	
Base Excision Repair	18	295	98	57	18971944, 15073047, 18669648
Nucleotide Excision Repair	32	729	285	165	12140753, 17081983, 18669648, 16964243
Mismatch Repair	10	382	132	70	17525332, 18669648, 16964243
Non-Homologous End Joining	8	298	99	74	14599745, 18669648, 16097034
Homologous Recombination Repair	23	639	189	72	14701743, 22084686, 14749735
DNA Damage Signaling	19	705	269	156	22084686, 18971944, 8084608
Trans Lesion Synthesis	8	591	143	84	18669648
Fanconi Anemia	16	905	273	132	12815053, 18669648, 11855836, 17525332
Editing and Processing Nucleases	8	157	50	23	–
Ubiquitination and Modification	10	227	70	33	21150323, 21098111
PARP Enzymes	3	63	26	21	–
Modulation of Nucleotide Pools	3	28	7	7	8389461, 18669648
Direct Reversal of Damage	3	22	8	7	–
Chromatin Structure	3	58	28	16	15302935, 17525332
DNA Polymerases	7	158	72	65	18669648
Others[b]	44	1062	333	188	17081983, 18669648, 15298678, 17525332

[a] The PubMed IDs of some of the research articles showing the experimental validations for the occurrence of important phosphorylation sites in DNA repair proteins have been stated in this column.

[b] The category includes all the DNA repair genes which have not been classified on the basis of mechanism in which they are involved.

reported because of the mutations or any additional aberrations in DNA repair genes. The disease information includes the details of the mechanism, OMIM ID's and the pathways involved.

In the phosphorylation search option, user can acquire information for the predicted phosphorylation sites in the DNA repair protein sequences. The major information includes the protein ID of the repair protein sequence, position of the phosphorylation site in the sequence, 9 character sequence representing the phosphorylated residue at the exact center of the sequence, prediction scores above 0.5 has been chosen as a criteria for the selection of potential phosphorylation sites and the prediction column gives the putative phosphorylated residues (S or T or Y). The experimental validation for many phosphorylation sites in the DNA repair proteins has been made from diverse research articles and intense literature survey. Total number of phosphorylated residues (S, T, Y) for each mechanism, and PubMed IDs for few of these verified research articles have been mentioned in Table 2.

For effortless understanding of the results and their interpretations, following examples have been taken from the repository and illustrated in Fig. 3. Here, “Translesion Synthesis” has been chosen from the mechanism menu, ‘675′ as Gene ID of *BRCA2* “GenBank ID: 675” for nsSNPs and markers search boxes, “*BRCA2*” as gene name for phosphorylation site prediction, “Gastric cancer” as disease name from disease menu, ‘7515′ as Gene ID of *XRCC1* “GenBank ID: 7515” for haplotype search option and ‘5985′ as Gene ID of *RFC5* “GenBank ID: 5985” for LD block generations. DR-GAS is first of its kind model where the users could easily retrieve and explore the quantitative genetic parameters and the phosphorylation states of DNA repair genes altogether.

4. Conclusion

DR-GAS is a compendium and comprehensive resource of DNA repair genes, their association studies with disease, other genetic parameters and phosphorylation states. There is no such catalog for DNA repair genes available which provides all these essential quantitative genetic details including LD, haplotype, SNPs, disease related information, and their phosphorylation states on a common platform. This database will help the researchers or scientists to study the repair genes in depth and will provide useful insight for future analysis and studies. This repository will also help for easy understanding and investigation of many DNA repair related disorders and moreover will provide useful genes and proteins related information. This database will be of utmost use to the researchers who are focused in developing therapeutic targets for precarious diseases like multiple forms of cancers, skin diseases and neurodegenerative disorders through the genetic factor's information, which is the basic foundation for the analysis and treatment of diseases. It is anticipated that this web based comprehensive resource would serve as a useful accompaniment for analyzing DNA repair systems for human and will also contribute scientific knowledge towards better understanding of other mammalian repair systems.

Conflict of interest statement

The Authors declare that there are no conflicts of interest.

Acknowledgements

We would like to express our gratitude to the editor and two anonymous reviewers whose constructive and insightful comments were very helpful in strengthening this paper. We thank Priya P. Panigrahi for critically reading the manuscript. This work was funded by Department of Science and Technology (DST), India through FASTTRACK Scheme grant (SR/FT/LS-026/2009) to TRS.

References

[1] G.A. Cromie, J.C. Connelly, D.R. Leach, Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans, Mol. Cell. 8 (2001) 1163–1174.

[2] G. Slupphaug, B. Kavli, H.E. Krokan, The interacting pathways for prevention and repair of oxidative DNA damage, Mutat. Res. 531 (2003) 231–251.

[3] A. Roulston, R.C. Marcellus, P.E. Branton, Viruses and apoptosis, Annu. Rev. Microbiol. 53 (1999) 577–628.

[4] A.L. Jackson, L.A. Loeb, The contribution of endogenous sources of DNA damage to the multiple mutations in cancer, Mutat. Res. 477 (2001) 7–21.

[5] T. Lindahl, DNA repair enzymes, Annu. Rev. Biochem. 51 (1982) 61–89.

[6] C.E. Lilley, R.A. Schwartz, M.D. Weitzman, Using or abusing: viruses and the cellular DNA damage response, Trends Microbiol. 15 (2007) 119–126.

[7] M.C. Moraes, J.B. Neto, C.F. Menck, DNA repair mechanisms protect our genome from carcinogenesis, Front. Biosci. 17 (2012) 1362–1388.

[8] J. Knoch, Y. Kamenisch, C. Kubisch, M. Berneburg, Rare hereditary diseases with defects in DNA-repair, Eur. J. Dermatol. 22 (2012) 443–455.

[9] Best P. Benjamin, Nuclear DNA damage as a direct cause of aging, Rejuven. Res. 12 (2009) 199–208.

[10] T.L. Timme, R.E. Moses, Diseases with DNA damage-processing defects, Am. J. Med. Sci. 295 (1988) 40–48.

[11] J.H. Hoeijmakers, DNA damage, aging, and cancer, N. Engl. J. Med. 361 (2009) 1475–1485.

[12] S. Hassen, N. Ali, P. Chowdhury, Molecular signaling mechanisms of apoptosis in hereditary non-polyposis colorectal cancer, World J. Gastrointest. Pathophysiol. 3 (2012) 71–79.

[13] M. Sehgal, T.R. Singh, Identification and analysis of biomarkers for mismatch repair proteins: a bioinformatic approach, J. Nat. Sc. Biol. Med. 3 (2012) 139–146.

[14] D. Tamura, S.G. Khan, M. Merideth, J.J. Digiovanna, M.A. Tucker, et al., Effect of mutations in XPD (ERCC2) on pregnancy and prenatal development in mothers of patients with trichothiodystrophy or xeroderma pigmentosum, Eur. J. Hum. Genet. 20 (2012) 1308–1310.

[15] J. Oshima, G.M. Martin, F.M. Hisama, Werner syndrome, in: R.A. Pagon, T.D. Bird, C.R. Dolan, et al. (Eds.), GeneReviews^{TM}, University of Washington, Seattle, 2002 http://www.ncbi.nlm.nih.gov/books/NBK1514/

[16] A.N. Suhasini, R.M. Brosh Jr., Fanconi anemia and Bloom's syndrome crosstalk through FANCJ-BLM helicase interaction, Trends Genet. 28 (2012) 7–13.

[17] K. Savitsky, A. Bar-Shira, S. Gilad, G. Rotman, Y. Ziv, et al., A single ataxia telangiectasia gene with a product similar to PI-3 kinase, Science 268 (1995) 1749–1753.

[18] C.A. Strathdee, M. Buchwald, Molecular and cellular biology of Fanconi anemia, Am. J. Pediatr. Hematol. Oncol. 14 (1992) 177–185.

[19] I. Kamileri, I. Karakasilioti, A. Sideri, T. Kosteas, A. Tatarakis, et al., Defective transcription initiation causes postnatal growth failure in a mouse model of nucleotide excision repair (NER) progeria, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 2995–3000.

[20] H. Vogel, D.S. Lim, G. Karsenty, M. Finegold, P. Hasty, Deletion of Ku80 causes early onset of senescence in mice, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 10770–10775.

[21] M. Nakao, S. Hosono, H. Ito, M. Watanabe, N. Mizuno, et al., Selected polymorphisms of base excision repair genes and pancreatic cancer risk in Japanese, J. Epidemiol. 22 (2012) 477–483.

[22] I. Saadat, Z. Beyzaei, F. Aghaei, S. Kamrani, M. Saadat, Association between polymorphisms in DNA repair genes (XRCC1 and XRCC7) and risk of preeclampsia, Arch. Gynecol. Obstet. 286 (2012) 1459–1462.

[23] Z. Duan, C. He, Y. Gong, P. Li, Q. Xu, et al., Promoter polymorphisms in DNA repair gene ERCC5 and susceptibility to gastric cancer in Chinese, Gene 511 (2012) 274–279.

[24] J. Yin, U. Vogel, Y. Ma, R. Qi, H. Wang, et al., HapMap-based study of a region encompassing ERCC1 and ERCC2 related to lung cancer susceptibility in a Chinese population, Mutat. Res. 713 (2011) 1–7.

[25] P. Van Eerdewegh, R.D. Little, J. Dupuis, R.G. Del Mastro, K. Falls, et al., Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness, Nature 418 (2002) 426–430.

[26] A. Sarasin, An overview of the mechanisms of mutagenesis and carcinogenesis, Mutat. Res. 544 (2003) 99–106.

[27] J. Thacker, M.Z. Zdzienicka, The mammalian XRCC genes: their roles in DNA repair and genetic stability, DNA Rep. (Amst) 2 (2003) 655–672.

[28] N. Motoyama, K. Naka, DNA damage tumor suppressor genes and genomic instability, Curr. Opin. Genet. Dev. 14 (2004) 11–16.

[29] L.N. Johnson, M. O'Reilly, Control by phosphorylation, Curr. Opin. Struct. Biol. 6 (1996) 762–769.

[30] P. Cohen, The regulation of protein function by multisite phosphorylation–a 25 year update, Trends Biochem. Sci. 25 (2000) 596–601.

[31] T. Pawson, Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems, Cell 116 (2004) 191–203.

[32] J.W. Roh, J.W. Kim, N.H. Park, Y.S. Song, I.A. Park, et al., p53 and p21 genetic polymorphisms and susceptibility to endometrial cancer, Gynecol. Oncol. 93 (2004) 499–505.

[33] M.T. Wu, D.C. Wu, H.K. Hsu, E.L. Kao, C.H. Yang, et al., Association between p21 codon 31 polymorphism and esophageal cancer risk in a Taiwanese population, Cancer Lett. 201 (2003) 175–180.

[34] D.Y. Xing, W. Tan, N. Song, D.X. Lin, Ser326Cys polymorphism in hOGG1 gene and risk of esophageal cancer in a Chinese population, Int. J. Cancer 95 (2001) 140–143.

[35] H. Sugimura, T. Kohno, K. Wakai, K. Nagura, K. Genka, et al., hOGG1 Ser326Cys polymorphism and lung cancer susceptibility, Cancer Epidemiol. Biomarkers Prev. 8 (1999) 669–674.

[36] J.C. Figueiredo, J.A. Knight, L. Briollais, I.L. Andrulis, H. Ozcelik, Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the Breast Cancer Family Registry, Cancer Epidemiol. Biomarkers Prev. 13 (2004) 583–591.

[37] M. Shen, R.J. Hung, P. Brennan, C. Malaveille, F. Donato, et al., Polymorphisms of the DNA repair genes XRCC1, XRCC3, XPD, interaction with environmental exposures, and bladder cancer risk in a case-control study in northern Italy, Cancer Epidemiol. Biomarkers Prev. 12 (2003) 1234–1240.

[38] J.D. Graves, E.G. Krebs, Protein phosphorylation and signal transduction, Pharmacol. Ther. 82 (1999) 111–121.

[39] K. Milanowska, J. Krwawicz, G. Papaj, J. Kosinski, K. Poleszak, et al., REPAIRtoire–a database of DNA repair pathways, Nucleic Acids Res. 39 (2011) D788–D792.

[40] Wen. Liting, Jin-An Feng, Repair-FunMap: a functional database of proteins of the DNA repair system, Bioinformatics 20 (2004) 2135–2137.

[41] G.A. Thorisson, A.V. Smith, L. Krishnan, L.D. Stein, The International HapMap Project Web site, Genome Res. 15 (2005) 1592–1593.

[42] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263–265.

[43] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, et al., dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.

[44] A.J. Brookes, H. Lehvaslaiho, M. Siegfried, J.G. Boehm, Y.P. Yuan, et al., HGBASE: a database of SNPs and other variations in and around human genes, Nucleic Acids Res. 28 (2000) 356–360.

[45] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.

[46] N. Blom, S. Gammeltoft, S. Brunak, Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351–1362.

[47] A. Kreegipuu, N. Blom, S. Brunak, PhosphoBase, a database of phosphorylation sites: release 2.0, Nucleic Acids Res. 27 (1999) 237–239.

[48] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (2005) D514–D517.

[49] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, The genetic association database, Nat. Genet. 36 (2004) 431–432.

[50] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, D. Lancet, GeneCards: integrating information about genes, proteins and diseases, Trends Genet. 13 (1997) 163.

[51] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular datasets, Nucleic Acids Res. 40 (2012) D109–D114.

[52] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, et al., Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res. 39 (2010) D691–D697.