

# Candidate nsSNPs That Can Affect the Functions and Interactions of Cell Cycle Proteins

Sevtap Savas,<sup>1,2,3</sup> M. Farhan Ahmad,<sup>1</sup> Mehjabeen Shariff,<sup>1</sup> David Y. Kim,<sup>1</sup> and Hilmi Ozcelik<sup>1,2,3\*</sup>

<sup>1</sup>Fred A. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada

<sup>2</sup>Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, Canada

<sup>3</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

**ABSTRACT** Nonsynonymous single nucleotide polymorphisms (nsSNPs) alter the encoded amino acid sequence, and are thus likely to affect the function of the proteins, and represent potential disease-modifiers. There is an enormous number of nsSNPs in the human population, and the major challenge lies in distinguishing the functionally significant and potentially disease-related ones from the rest. In this study, we analyzed the genetic variations that can alter the functions and the interactions of a group of cell cycle proteins (n = 60) and the proteins interacting with them (n = 26) using computational tools. As a result, we extracted 249 nsSNPs from 77 cell cycle proteins and their interaction partners from public SNP databases. Only 31 (12.4%) of the nsSNPs were validated. The majority (64.5%) of the validated SNPs were rare (minor allele frequencies < 5%). Evolutionary conservation analysis using the SIFT tool suggested that 16.1% of the validated nsSNPs may disrupt the protein function. In addition, 58% of the validated nsSNPs were located in functional protein domains/motifs, which together with the evolutionary conservation analysis enabled us to infer possible biological consequences of the nsSNPs in our set. Our study strongly suggests the presence of naturally occurring genetic variations in the cell cycle proteins that may affect their interactions and functions with possible roles in complex human diseases, such as cancer. *Proteins* 2005;58:697–705.

© 2004 Wiley-Liss, Inc.

**Key words:** single nucleotide polymorphisms; cell cycle; protein function and interaction; protein domains and motifs; computational analysis

## INTRODUCTION

Deregulation of cell cycle and cell proliferation mechanisms has an important role in carcinogenesis. A number of cell cycle genes, such as cyclins, cyclin-dependent kinases (CDKs), and the regulators of the CDKs, are found frequently mutated in many types of cancer.<sup>1,2</sup> In addition, germ-line mutations in several cell cycle control genes such as RB1,<sup>3</sup> BRCA1 and BRCA2,<sup>4</sup> TP53,<sup>5</sup> NF2,<sup>6</sup> and CHEK2<sup>7</sup> have been found to cause strong genetic predisposition to cancer in individuals. However, most of the cancer susceptibility cannot be explained by the presence of

high-penetrant alleles.<sup>8,9</sup> In the remaining cases, genetic variations are hypothesized to contribute to the disease risk.<sup>8–10</sup>

Genetic variations are classified by single nucleotide polymorphisms (SNPs), small insertions/deletions, conversions, and rearrangements.<sup>11–14</sup> SNPs constitute almost 90% of the genetic variations in the human genome. The nonsynonymous SNPs (nsSNPs) that change the amino acids are likely to affect the structure and the function of the proteins, and thus are good candidates for disease-modifying alleles. However, not all amino acid substitutions lead to such an impact; hence it is essential to select the nsSNPs that are either proven to be functional (by means of experimental analyses) or predicted to be functional (using prediction models). Recently many approaches have been developed to predict/assess whether a nsSNP can affect the protein function and structure.<sup>15–19</sup>

Proteins serve specific functions in cells by acting in concert with other proteins and molecules. Both transient or stable binary protein interactions and multisubunit protein complexes are indispensable to promote cellular processes such as signal transduction, DNA repair, and cell cycle.<sup>20,21</sup> For example, formation and the function of the heterodimeric complexes of particular cyclins and CDKs and their interactions with other proteins, such as RB1, are essential for proper initiation and progression of the cell cycle.<sup>2,22</sup> Similarly, the accurate function of the anaphase promoting complex, which contains several subunits, is necessary for appropriate chromosome segregation during mitosis.<sup>23</sup> As anticipated, the abnormalities in such biological interactions/complexes have been implicated in several human diseases including cancer.<sup>2,24,25</sup>

Therefore, in this study, we aimed to characterize not only the genetic variations that can affect the functions and interactions of a set of cell cycle proteins per se, but also the functions and interactions of the proteins interacting with them. To do so, we applied a previously described

Grant sponsor: the Susan Komen Breast Cancer Foundation; Grant number: BCTR0100627.

M. Farhan Ahmad and Mehjabeen Shariff contributed equally to this work.

\*Correspondence to: Hilmi Ozcelik, Ph.D., Mount Sinai Hospital Samuel Lunenfeld Research Institute, 600 University Avenue Room 992A, Toronto, ON, Canada M5G 1X5. E-mail: ozcelik@mshri.on.ca

Received 26 May 2004; Accepted 24 September 2004

Published online 22 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20367

systematic approach to identify the candidate nsSNPs that are likely to affect the protein, and subsequently the cellular functions.<sup>26,27</sup> This strategy has been based on the determination of the evolutionary conservation using the Sorting Intolerant From Tolerant tool (SIFT<sup>16,17</sup>; [http://blocks.fhrc.org/~pauline/SIFT\\_seq\\_submit2.html](http://blocks.fhrc.org/~pauline/SIFT_seq_submit2.html)) as well as the assignment of the functional protein domains and motifs using a variety of web-based tools. We strongly believe that the nsSNPs analyzed and reported here are excellent candidates for functional and cancer-susceptibility studies.

## METHODS

### SNP Resource

The list of group A cell cycle proteins was manually selected from the CGAP-GAI database (<http://lpgws.nci.nih.gov/><sup>28</sup>), LocusLink resource of NCBI (<http://www.ncbi.nlm.nih.gov/LocusLink/><sup>29</sup>), and from relevant literature reports. The nsSNPs were extracted from the dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/><sup>30</sup>), HGVBbase (<http://hgvbbase.cgb.ki.se/><sup>31</sup>), CGAP-GAI (<http://lpgws.nci.nih.gov/><sup>28</sup>), and SNP500 (<http://bumper.nci.nih.gov/home.cfm/><sup>32</sup>) databases using the gene symbol as the search key. The 100–120 base pair long SNP-flanking sequences were blasted against the transcribed sequences deposited in the GenBank<sup>33</sup> using the BLAST against gene transcripts tool (<http://lpgws.nci.nih.gov:80/perl/blast2/><sup>34</sup>) to identify the nsSNP location on a particular transcript<sup>26</sup>; 1) only the SNP-flanking sequence/transcript sequence alignments with no gaps or mismatches were accepted, 2) partial sequence alignments were accepted, and 3) the sequences that match with only the transcript of interest were evaluated as gene specific, and included in this study. The SNP flanking-sequences were also blasted against the human genome using the BLAST tool of NCBI (<http://www.ncbi.nlm.nih.gov/BLAST/><sup>29</sup>) to eliminate the SNPs derived from more than one region of the genome<sup>35</sup> as explained in Savas et al.<sup>26</sup> In this manuscript, we annotated a nsSNP as a “validated nsSNP” if and only if it was found in  $\geq$  two chromosomes (to eliminate possible genotyping errors) in a sample panel with a size of  $\geq$  48 chromosomes.<sup>36</sup> In the case of BRCA1, the SNP entries with the frequency determination in cancer-affected individuals/families were excluded from the data.

No validated nsSNPs were found in the following genes: Group A: ANAPC11, ANAPC2, ANAPC4, ANAPC5, ANAPC7, APC10, CCNB1, CCNB2, CCNB3, CCNC, CCND1, CCNF, CCNG1, CCNL1, CDC14A, CDC16, CDC2, CDC20, CDC23, CDC25B, CDC2L5, CDC37, CDC40, CDC42, CDC45L, CDC5L, CDC7, CDK2, CDK3, CDK4, CDK5, CDK6, CDK7, CDK9, CDKN1C, CDKN3, E2F4, E2F6, JUN, MDM2, NFKBIA, RBBP8. Group B: ABL2, BCL3, BCR, CRK, CSNK2A1, CSNK2B, CTNNA1, DNMT1, ESR1, HDAC1, HSPA8, HSPCA, LMNA, PSMA3, STAT3, TFDP2.

### Evolutionary Conservation Analysis

Evolutionary conservation analysis based on homologous protein alignments was performed using the SIFT

tool ([http://blocks.fhrc.org/~pauline/SIFT\\_seq\\_submit2.html/](http://blocks.fhrc.org/~pauline/SIFT_seq_submit2.html/)<sup>16,17</sup>). A SIFT result with a normalized probability of  $\geq$  0.05 predicts the substitution tolerated whereas  $<$  0.05 predicts the substitution affecting the function of the protein.<sup>16,17</sup> In cases when the median sequence conservation score was  $\geq$  3.25, we interpreted the SIFT predictions as possibly tolerated or possibly affecting.<sup>26</sup> Whenever the protein size was too large to be analyzed by SIFT, we used a smaller portion of the protein for SIFT analysis. The SIFT program was run for the amino acid sequence under the accession number obtained by the BLAST against gene transcripts tool.<sup>34</sup>

The mouse orthologs of the human proteins were extracted from the Swiss-Prot database.<sup>37</sup> The human protein sequence (under the accession numbers given in Tables IIA and IIB) was aligned with that of the mouse using the ClustalW<sup>38</sup> program to identify the mouse amino acid at the SNP location. The Swiss-Prot IDs for the mouse orthologs are as follows: Group A: abl1: P00520, apc: Q61315, ccna2: P51943, ccnd3: P30282, ccng2: O08918, ccnh: Q61458, ccni: Q9Z2V9, cdc27: Q8R568, cdc6: O89033, cdkn1a: P39689, cdkn1b: P46414, cdkn2a: P51480, chek1: O35280, e2f1: Q61501, e2f2: P56931, e2f3: O35261, rb1: P13405, sertad1: Q9JL10. Group B: brca1: Q9Z1D2, nfkb1: P25799.

### Protein Domain/Motif Analysis

Protein domain/motif information was extracted from the Swiss-Prot feature table (<http://us.expasy.org/sprot/><sup>37</sup>), InterPro (<http://www.ebi.ac.uk/InterPro/scan.html/><sup>39</sup>), Human Protein Reference Database (HPRD; <http://www.hprd.org/><sup>40</sup>), and Molecular Interactions Databases (MINT; <http://160.80.34.4/mint/><sup>41</sup>) or using the InterProScan tool (<http://www.ebi.ac.uk/InterPro/scan.html/><sup>42</sup>) as described in Savas et al.<sup>27</sup> Protein domain and interaction information was not available for all proteins in the databases described above (data not shown, available upon request).

The proteins that are annotated as interacting with the group A cell cycle proteins in HPRD and MINT databases (group B proteins) were analyzed for the presence of nsSNPs as described above. In group B, excluding the HIV TAT protein, no nsSNP was found in the SNP databases that are searched for the human F-actin, MYOD1, and RYBP proteins. The gene symbols approved by the HUGO Human Genome Nomenclature (<http://www.gene.ucl.ac.uk/nomenclature/><sup>43</sup>) were included into this manuscript. Information curated from different databases and reported here is as of January 2004.

## RESULTS

We analyzed the nsSNPs in 60 cell cycle genes (group A) and 26 genes that were reported to interact with them (group B). Group B consisted of proteins that function in diverse cellular processes such as cell cycle, DNA repair, and apoptosis. Four proteins (RB1, MDM2, CDC16, and ABL1) are present in both of the groups, resulting in a nonredundant set of a total of 81 proteins. For simplicity, we discussed the RB1, MDM2, CDC16, and ABL1 proteins in the group A category only.

**TABLE I. Categorization of the Validated nsSNPs Based on Their Minor Allele Frequencies and SIFT Prediction**

	Group A	Group B	Total
All nsSNPs (n)	23	8	31
nsSNPs < 5%	14/23 (60.9%)	6/8 (75.0%)	20/31 (64.5%)
nsSNPs ≥ 5%	9/23 (39.1%)	2/8 (25.0%)	11/31 (35.5%)
Functional by SIFT <sup>a</sup> (n)	4	1	5
nsSNPs < 5%	3/4 (75.0%)	1/1 (100.0%)	4/5 (80.0%)
nsSNPs ≥ 5%	1/4 (25.0%)	0 (0%)	1/5 (20.0%)
Tolerated by SIFT <sup>b</sup> (n)	15	7	22
nsSNPs < 5%	10/15 (66.6%)	5/7 (71.4%)	15/22 (68.2%)
nsSNPs ≥ 5%	5/15 (33.3%)	2/7 (28.6%)	7/22 (31.8%)
No prediction by SIFT <sup>c</sup> (n)	4	0	4
nsSNPs < 5%	1/4 (25.0%)	0 (0%)	1/4 (25.0%)
nsSNPs ≥ 5%	3/4 (75.0%)	0 (0%)	3/4 (75.0%)

<sup>a</sup>Affecting and possibly affecting upon SIFT analysis.

<sup>b</sup>Tolerated and possibly tolerated upon SIFT analysis. The BRCA1-Q356R (with both less and higher than 5% minor allele frequencies) is included into the nsSNP group with > = 5% minor allele frequency.

<sup>c</sup>No SIFT prediction could be made due to the presence of less than six proteins in the alignment.

We have extracted and analyzed a total of 249 nsSNPs (155 group A and 94 group B nsSNPs) from 59 group A and 18 group B proteins from SNP databases, respectively (see [www.ozceliklab.com/savas\\_cell\\_cycleA](http://www.ozceliklab.com/savas_cell_cycleA) and [www.ozceliklab.com/savas\\_cell\\_cycleB](http://www.ozceliklab.com/savas_cell_cycleB)). Altogether, 87.6% of the nsSNPs were not validated whereas 12.4% were validated/proven (found in at least two chromosomes). Validated nsSNPs are presented in Tables I, IIA, and IIB, and constitute the main focus of this manuscript. A total of 20 (64.5%) validated nsSNPs were reported with minor allele frequencies < 5%, whereas 11 (35.5%) nsSNPs had minor allele frequencies ≥ 5% in the samples analyzed, (upon analysis of different samples, BRCA1-Q356R was reported both in ≥ 5% and < 5% of the samples and for simplicity, it is categorized as a nsSNP with allele frequency ≥ 5% throughout the manuscript) (Table I).

Comparison of allelic frequencies of nsSNPs among different ethnic groups demonstrated a total of 10 nsSNPs with some degree of difference (Table III, see below). For example, CCNA2-V163I, CHEK1-V471I, BRCA1-S1040N, and BRCA1-M1652I were not found in African population, but found in either East Asian or Caucasian populations. Similarly, the BRCA1-Q356R and BRCA1-D693N were found only in Caucasian samples. The CDKN1A-F63L was found in African but not in East Asian or Caucasian samples. The minor allele frequencies for the CDKN1A-S31R were comparatively different among the African (25%), East Asian (~46%), and Caucasian (~5%) populations. Interestingly, the allele frequency of the BRCA1-P871L was, to a certain extent, different between the Africans in which the common allele was L871 (~85%), and Caucasians in which the common allele was P871 (~68%).

The evolutionary conservation analysis using the SIFT tool<sup>16,17</sup> (Table I) was inconclusive in four group A cell cycle nsSNPs. However, reliable SIFT results were available in 82.6% of the validated nsSNPs from groups A and B genes. Among this group, one nsSNP was interpreted as

affecting (with a median sequence conservation score of < 3.25). In eight of the nsSNPs, the SIFT median sequence conservation score was ≥ 3.25, indicating that the proteins in the alignment are highly similar to each other. Thus we interpreted such nsSNPs as possibly affecting (see Savas et al.<sup>26</sup>).

Protein domain and motif search results using the Swiss-Prot,<sup>37</sup> InterPro,<sup>39</sup> InterProScan,<sup>42</sup> MINT,<sup>41</sup> and HPRD<sup>40</sup> algorithms and databases for validated nsSNPs from groups A and B proteins are shown in Tables IIA and IIB, respectively: 11/23 group A nsSNPs and 7/8 group B validated nsSNPs were found in a functional domain/motif. Among these, CDC6-D295N, CDKN1A-F63L, and CDKN2A-A148T in group A and the BRCA1-D693N in group B were predicted to have functional consequences upon SIFT analysis (Tables I, IIA, and IIB); all of these nsSNPs had minor allele frequencies < 5%.

Based on the evolutionary conservation status as well as the location in a functional protein domain/motif, we were able to infer possible biological functions for our set of validated nsSNPs analyzed in this study. When the nsSNPs resulting in evolutionarily intolerated substitutions (affecting and possibly affecting nsSNPs) are considered, it can be speculated that: the CDC6-D295N can affect the potential ATP-dependent protein clamp activity of CDC6; CDKN1A-F63L can affect the CDKN1A interaction with CSNK2B and thus the down-regulation of CSNK2B activity;<sup>44</sup> and BRCA1-D693N can affect BRCA1 binding to, 1) ZNF350, and thus the transcriptional corepressor function of BRCA1,<sup>45</sup> 2) STAT1, and thus STAT1-mediated transcriptional responses and the IFN-gamma-dependent tumor surveillance function of BRCA1,<sup>46</sup> 3) RAD50 and thus BRCA1's response to DNA damage, 4) AR, and thus enhancement of AR-dependent transactivation.<sup>47</sup>

Finally, the comparison between the human and mouse orthologs revealed that in case of 1) 16 nsSNPs, mouse had the human common allele at the nsSNP position; 2) five nsSNPs, the mouse ortholog had the variant amino acid

residue at the nsSNP position; 3) seven nsSNPs, the mouse ortholog had neither the human wild type nor the human variant amino acid residue at the nsSNP position (Tables IIA and IIB). For three nsSNPs (CDC27-Y496H, CDKN2A-A148T, and BRCA1-S1140G), this comparison could not be done.

## DISCUSSION

Among the 31 validated nsSNPs, more than half (64.5%) had minor allele frequencies less than 5%. This data indicates that there is an excess of rare nsSNPs in our data set, which is in good agreement with the data obtained from larger genetic variation sets.<sup>48–50</sup> In evolutionary terms, the low frequency of an allele can be explained by either being i) new and not fixed in the population yet, ii) deleterious/slightly deleterious and subject to purifying selection.<sup>50–55</sup> In order to evaluate the functional consequences of the amino acid substitutions, we utilized the SIFT tool.<sup>16,17</sup> SIFT makes evolutionary conservation predictions based on the alignment of the similar proteins in the protein databases and automatically calculates scores for the substitutions. We had previously decided to utilize a modified interpretation of the SIFT predictions based on the number of proteins in the alignment as well as the median sequence conservation score.<sup>26</sup> When applied to the validated nsSNPs set discussed here, the SIFT results demonstrated that four out of 20 rare nsSNPs were predicted either affecting (CDC6-D295N and BRCA1-D693N), or possibly affecting (CDKN1A-F63L and CDKN2A-A148T), suggesting that 20% of the rare nsSNPs are likely to be under negative selection. We believe that once the rare SNPs are further validated in larger number of samples to better determine their allelic frequencies, the nsSNPs discussed above will be excellent candidates for further analyses.

The higher allelic frequency, on the other hand, can mostly predict the nature of the variant as either neutral, slightly deleterious, or beneficial (positively selected). However, the exceptions apply if the variant of interest provides a selective advantage to the heterozygotes (when the allele frequency of the deleterious allele becomes presumably higher<sup>52,53</sup>). Another possible scenario is based on the possibility of the existence of hot-spot mutations that have detrimental effects: in this case, the minor allele frequency and the deleterious effect will not inversely correlate with each other, either. We evaluated the eight common nsSNPs with minor allele frequencies of  $\geq 5\%$  under the light of these possibilities, and found that there was only one nsSNP found possibly affecting by SIFT analysis (CDC6-V441I). The CDC6-V441I was not found exclusively conserved among the homologous proteins though (evaluated as possibly affecting by SIFT), it still can be suggested that this variant may affect the function. Therefore, CDC6-V441I could represent a beneficial variant, or a deleterious variant (with heterozygote advantage), or a hot-spot mutation with a deleterious consequence. In addition, due to both its possible functional character as well as its relatively higher allelic frequency (20–44% in North American

and Asian populations; dbSNP: rs13706), CDC6-V441I can be a good candidate for cancer/disease-association studies.

In order to evaluate whether the common alleles were fixed or under recent selection, we compared the human proteins at the SNP location with those of the mouse (assuming that the protein sequences deposited in the GenBank<sup>33</sup> for human and the Swiss-Prot<sup>37</sup> for mouse represent the common alleles) (Tables IIA and IIB). This comparison demonstrated that: i) in 51.6% of the cases, the mouse and the human common alleles had the same amino acids, suggesting that the major alleles have been inherited from the human-mouse ancestor and have been fixed in these organisms; ii) in 16.1% of the cases, mouse had the minor allele, suggesting that either the ancestral allele is being rapidly replaced during human evolution (recent selection), or there might be an epistatic interaction (see below); and, iii) in 22.6% of the cases, mouse had an amino acid different from both the common and the rare allele in human, suggesting that these amino acids (or perhaps, the proteins) have evolved/diverged between the two organisms.

We also compared the minor allele frequencies among African, East Asian, and Caucasian samples (Table III) for the nsSNPs with available data in at least two ethnic populations. The African population, being considered the oldest population on earth, is associated with higher molecular diversity.<sup>56</sup> Existence of some nsSNPs in Caucasian and/or East Asian populations but not in African population may indicate that these variations are relatively new (Table III). Similarly, different population-specific selection pressures can explain the variation in frequencies among ethnic populations.

Results of protein domain and motif analyses using Swiss-Prot,<sup>37</sup> InterPro,<sup>39</sup> InterProScan,<sup>42</sup> MINT,<sup>41</sup> and HPRD<sup>40</sup> databases and algorithms indicated that a considerable number of nsSNPs ( $> 60\%$ ) lied in a functional protein domain/motif (Tables IIA, IIB). These domains/motifs were implicated in diverse functions ranging from molecular interaction domains to protein activity. A total of three validated functional nsSNPs (predicted as affecting and/or possibly affecting upon SIFT analysis) were also found to be located in a protein domain/motif in the cell cycle proteins and their interaction partners: none of these three nsSNPs had minor allele frequencies  $\geq 5\%$ . This data supports the previous scientific findings in the literature that indicated the presence of a strong purifying pressure against the amino acid alterations in protein domains.<sup>57,58</sup> Besides, other evolutionary forces, such as bottleneck and drift<sup>54,55</sup> can also explain these differences. However, due to the fact that the herein discussed data is based on relatively small number of samples, we strongly believe that data from larger samples sizes are required to validate these suggestions.

The functional protein region information not only enabled us to get insight into the distribution of (functional) nsSNPs along the protein domains/motifs but also to infer a biological role for the nsSNPs in the cell cycle proteins and their interaction partners. The CDC6 protein had two validated and presumably functional nsSNPs, which are

TABLE IIA. Validated nsSNPs of the Cell Cycle Proteins<sup>†</sup>

Gene	SNP ID <sup>a</sup>	msSNP	Residue in mouse <sup>b</sup>	Frequency <sup>c</sup>	SIFT <sup>d</sup>	Protein domain/motif <sup>e</sup>	Function <sup>f</sup>
ABL1	rs1064152	P140L	L	1	P. tolerated	SH2 domain/SH2 domain (IPR000980)	For regulation of the intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides
APC	rs459552	D182V	A	2	NP	—	—
CCNA2	rs769242	V163I	L	1	Tolerated	—	—
CCND3	rs3218089	P134S	P	1	Tolerated	[Cyclin domain (IPR006670)]; [Cyclin, N-terminal (IPR006671)]	[ <sup>64,65</sup> For protein-protein interaction?; <sup>66</sup> For binding to CDKs?]; [ <sup>64,65</sup> For protein-protein interaction?]
CCNG2	rs1051130	S259A	A	2	Tolerated	Cyclin C terminal domain (IPR004367)	?
CCNH	rs4150050	L4V	L	1	NP	—	—
	SNP000064444/ rs2266691	K138R	R	1	Tolerated	[Cyclin domain (IPR006670)]; [Cyclin N terminal domain (IPR006671)]	[ <sup>64,65</sup> For protein-protein interaction?; <sup>66</sup> For binding to CDKs?]; [ <sup>64,65</sup> For protein-protein interaction?]
	SNP000064445/ rs2266690	V270A	V	2	Tolerated	—	—
CCNI	rs4252903	V207I	V	1	Tolerated	—	—
CDC27	SNP000019837/ rs13666	Y496H	Na	2	NP	Protein-protein interaction	Binds to CDC16
CDC6	rs4135012	D295N	D	1	Affect	AAA ATPase domain (IPR003593)/AAA ATPase, central region domain (IPR003959)	Act as ATP-dependent protein clamps
	SNP000009441/ rs13706/ GAI1516788	V441I	V	2	P. affect	—	—
CDKN1A	SNP00003435/ rs1801270/ GAI1503061	S31R	R	2	P. tolerated	C4-type zinc finger domain (potential)	<sup>67</sup> Mediates DNA binding, protein-protein, and protein-lipid interactions
CDKN1B	rs4986867	F63L	F	1	P. affect	Protein-protein interaction	Binds to CSNK2B
CDKN2A	rs2066827	V109G	V	2	Tolerated	—	—
CHEK1	rs3731249	A148T	Na	1	P. affect	—	—
E2F1	rs506504	V471I	V	1	P. tolerated	—	—
	rs3213173	V276M	V	1	Tolerated	Dimerization domain (potential)	<sup>68</sup> Protein dimerization/heterodimerization
	rs3213176	G393S	G	1	P. tolerated	Transactivation domain	<sup>68</sup> Transcriptional activation
E2F2	rs2075995	Q226H	Q	2	Tolerated	Dimerization domain (potential)	<sup>68</sup> Protein dimerization/heterodimerization
E2F3	rs4134982	D389N	D	1	P. tolerated	—	—
RB1	rs4151539	A525G	A	1	Tolerated	[Pocket domain (binds T and E1A) / Domain A]; [Retinoblastoma-associated protein, A-box (IPR002720)]; [protein-protein interaction]	[ <sup>69</sup> For modulation of RB1 function upon binding to viral and cellular proteins]; [appears to be required for the stable folding of the B box; contain the cyclin-fold structural motif]; [binds to HDAC1, DNMT1, ABL1, MDM2]
SERTAD1	rs268687	A31T	P	2	NP	—	—

TABLE IIB. Validated nsSNPs of the Interaction Partners of the Cell Cycle Proteins<sup>†</sup>

Gene	SNP ID <sup>a</sup>	nsSNP	Residue in mouse <sup>b</sup>	Frequency <sup>c</sup>	SIFT <sup>d</sup>	Protein domain/motif <sup>e</sup>	Function <sup>f</sup>
ABLI BRAC1	See Table IIA						
	SNP00002456/ rs1799950	Q356R	P	2	Tolerated	Protein–protein interaction	Binds to RB1, ZNF350, VCP, NMI, RAD50, AR
	rs4986850	D693N	A	1	Affect	Protein–protein interaction	Binds to ZNF350, STAT1, RAD50, AR
	SNP000007492/ rs799917	P871L	L	2	Tolerated	Protein–protein interaction	Binds to RAD51L1, RAD51, MSH6, MSH2, AR
rs4986852	S1040N	G	1	P. tolerated	Protein–protein interaction	Binds to RAD51L1, RAD51, MSH2, MSH6, AR, MLH1	
rs2227945	S1140G	Na	1	P. tolerated	Protein–protein interaction	Binds to MSH6, MSH2, MLH1, AR	
SNP00002485/ rs1799967	M1652I	M	1	P. tolerated	[ <i>BRCT1 domain/BRCT domain</i> (IPR001357)]; [protein–protein interaction]	[ <sup>61</sup> Protein–protein interaction]; [binds to DDX39, POLR2A, BRCA2, CTBP2, CTBP1, RBBP4, EP300, RBBP7, CREBBP, RBBP8, BACH1, MSH6, MSH2, MLH1, DHX9, SP1, LMO4, AR, RNA Polymerase II complex, POLR2K, HDAC2, HDAC1]	
NFKB1	rs4648072	M507V	L	1	Tolerated	—	—
RSI	rs4648099	H712Q	H	1	Tolerated	[ANK 5 repeat/ANK repeat (IPR002110)]	Protein–protein interaction

<sup>†</sup>The nsSNPs with minor allele frequencies in group A and group B cell cycle proteins are listed in Table IIA and Table IIB, respectively.

<sup>a</sup>rs, SNP, and GAI prefixes represent the SNP IDs in dbSNP,<sup>30</sup> HGvbase,<sup>31</sup> and CGAP-GAI<sup>26</sup> SNP databases, respectively.

<sup>b</sup>In this analysis, although we cannot totally exclude the possibility that both the human proteins (the protein sequences under the accession numbers) and the mouse proteins (the protein sequences under the Swiss-Prot IDs) could represent the variant sequences, we assumed that the common alleles had a higher chance of being captured and analyzed, and thus being deposited into the GenBank<sup>33</sup> and the Swiss-Prot<sup>37</sup> databases that are utilized in this study. “Na” stands for not available; we could not determine the mouse residue at the SNP location for three nsSNPs: CDC27-Y496H (mouse orthologue was truncated), CDKN2A-A148T (this region had diverged between mouse and human), and the BRCA1-S1140G (mouse had a deletion of the corresponding amino acid).

<sup>c</sup>1: minor allele frequency < 5%; 2: minor allele frequency > = 5%; 1 and 2: minor allele frequencies reported as both < 5% and > = 5% in different submissions/samples.

<sup>d</sup>These predictions were done running the SIFT program;<sup>16,17</sup> “p.” stands for possibly (see Ref. <sup>26</sup>).

<sup>e</sup>The protein domain information was extracted from the Swiss-Prot,<sup>37</sup> InterPro,<sup>39</sup> MINT,<sup>41</sup> and HPRD<sup>40</sup> protein databases as well as using the InterProScan program;<sup>42</sup> Swiss-Prot entries are shown in italics; InterPro domain IDs are in parentheses; the protein regions extracted from MINT and HPRD databases were simply annotated as “protein–protein interaction domain.”

<sup>f</sup>This information was extracted from InterPro database or relevant publications. AMK: ankryin.

ATP: Adenosine triphosphate; CDK: cyclin-dependent kinase, SH2: Src-homology domain 2.

The GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) accession numbers for these genes are as follows: **Table IIA:** ABL1 (NM\_005157.2), APC (NM\_000038.2), CCNA2 (NM\_001237.2), CCND3 (NM\_001760.2), CCNG2 (NM\_004354.1), CCNH (NM\_001239.2), CCNI (NM\_006835.2), CDC27 (NM\_001256.2), CDC6 (NM\_001254.2), CDKN1A (NM\_000389.2), CDKN1B (NM\_004064.2), CDKN2A (NM\_000077.2), CHEK1 (NM\_001274.2), E2F1 (NM\_005225.1), E2F2 (NM\_004091.2), E2F3 (NM\_001949.2), RB1 (NM\_000321.1), SERTAD1 (NM\_013376.1). **Table IIB:** BRCA1 (NM\_007294.1), NFKB1 (NM\_003998.2).

**TABLE III. nsSNPs That Have Different Minor Allele Frequencies in Different Ethnic Groups<sup>†</sup>**

Gene	nsSNP	African	East Asian	Caucasian
CCNA2	V163I	48 Chr. G = 1.000 A = 0.000	48 Chr. G = 0.958 A = 0.042	—
CDKN1A	S31R	48 Chr. C = 0.750 A = 0.250	48 Chr. C = 0.542 A = 0.458	62 Chr. C = 0.952 A = 0.048
CDKN1A	F63L	48 Chr. G = 0.958 T = 0.042	48 Chr. G = 1.000 T = 0.000	62 Chr. G = 1.000 T = 0.000
CDKN2A	A148T	—	48 Chr. G = 1.000 A = 0.000	62 Chr. G = 0.919 A = 0.081
CHEK1	V471I	48 Chr. G = 1.000 A = 0.000	—	62 Chr. G = 0.935 A = 0.065
BRCA1	Q356R	48 Chr. A = 1.000 G = 0.000	48 Chr. A = 1.000 G = 0.000	58 Chr. A = 0.948 G = 0.052
BRCA1	D693N	48 Chr. G = 1.000 A = 0.000	48 Chr. G = 1.000 A = 0.000	58 Chr. G = 0.948 A = 0.052
BRCA1	P871L	48 Chr. T = 0.854 C = 0.146	—	62 Chr. C = 0.677 T = 0.323
BRCA1	S1040N	48 Chr. G = 1.000 A = 0.000	—	62 Chr. G = 0.952 A = 0.048
BRCA1	M1652I	48 Chr. G = 1.000 A = 0.000	—	62 Chr. G = 0.968 A = 0.032

<sup>†</sup>Only those nsSNPs that were found in at least two chromosomes in a sample panel of at least 48 chromosomes were included into that table. The frequency information belonging to the Multinational or North American populations are not included here. Under the population columns are the number of chromosomes analyzed and the allelic frequencies. Chr stands for chromosomes.

likely to affect the function of the protein. One of them, CDC6-D295N, was located in the ATPase domain of the CDC6 protein. This domain is essential for CDC6 binding to the double-stranded DNA for DNA replication in another organism.<sup>59,60</sup> Thus, it can be speculated that this nsSNPs can alter the DNA replication function of the CDC6 protein.

In addition, a total of 18 nsSNPs from the cell cycle and interacting proteins were found to be located in domains/motifs designated for protein–protein/DNA/microtubule/lipid interactions. Apparently, the vast majority of the domains were the ones involved in protein–protein interactions, such as the SH2/SH3 domains and ankyrin-repeats,<sup>21</sup> BRCT1 domains,<sup>61</sup> and the protein regions retrieved from the MINT<sup>41</sup> and the HPRD<sup>40</sup> databases that were reported as interacting with other proteins. One of the nsSNPs located in protein–protein interaction domains and motifs in group B was predicted as functional by the SIFT analysis: BRCA1-D693N. BRCA1 is a tumor suppressor protein, inactivation of which accounts for 5–10% of the breast cancers.<sup>25</sup> The presumably functional BRCA1 nsSNP (BRCA1-D693N) was located in a protein region that have been shown to interact with a total of four different proteins (Table IIB). Our analysis suggests that this BRCA1 nsSNP is likely to affect the interactions of the BRCA1 protein, and influence both the BRCA1-dependent transcriptional regulations through its interactions with the AR, ZNF350, and STAT1 proteins and the DNA repair and genome integrity functions of BRCA1 through its interaction with RAD50 DNA repair protein.<sup>44–47</sup>

Individual amino acid substitutions either on the same or functionally/physically interacting proteins can compensate for each other's deleterious effects or present an additive effect within certain environmental condition (coevolution, epistasis<sup>52,54,62,63</sup>). For example, previously the comparison of human disease-associated protein sequences with mouse orthologs showed that in 160 cases (2.2%), the mouse protein contained the disease-associated residues of human proteins, suggesting the presence of such dependences.<sup>57</sup> In our data set, we have shown the existence of multiple nsSNPs along the coding region of some proteins (Tables IIA and IIB). The herein presented strategy cannot evaluate the epistatic effect of multiple

amino acid substitutions on the same or interacting proteins, however we feel that such proteins represent excellent candidates for elaborate experimental designs to reveal such relations. Additionally, we also propose that the interacting proteins, each carrying functional nsSNPs, are also good candidates for evaluation of interprotein epistasis<sup>62</sup> (Tables IIA, IIB). Identification of such protein couples or complexes would also provide an invaluable resource of nsSNP combinations that can actually compensate for or aggravate each other's effects. No matter what the functional consequences of an individual nsSNP are, these candidates could promote the elucidation of complex genetic variation-disease susceptibility relationship.

## CONCLUSION

Dynamic/transient protein–protein interactions (such as the ones that recruit the effector proteins to the receptors in signal transduction), or more organized/stable ones that form the multisubunit protein complexes (such as the transcriptional machinery) are essential for cells and organisms. Accordingly, in this study, we analyzed the genetic variations that can alter the functions and the interactions of a group of cell cycle protein and proteins interacting with them. We strongly believe that the nsSNPs that are predicted to be functional upon evolutionary conservation analysis and are located in functional protein domains and motifs constitute an excellent resource for specific molecular studies to elucidate the direct consequences of these variations on protein function and interaction, which could also help the molecular epidemiology and genetic studies aiming to reveal the genetic variation–disease risk association.

## ACKNOWLEDGMENTS

This work was supported by a grant (BCTR0100627) from the Susan Komen Breast Cancer Foundation, USA. S. Savas is supported, in part, by a "CIHR Strategic Training Program Grant—The Samuel Lunenfeld Research Institute Training Program: Applying Genomics to Human Health" fellowship. The authors thank Hamdi Jarjanazi for his help with the URL tables.

### ELECTRONIC-DATABASE INFORMATION

Electronic URL addresses for the databases and algorithms used in this article are as follows:

BLAST: <http://www.ncbi.nlm.nih.gov/BLAST/>  
 BLAST against gene transcripts: <http://lpgws.nci.nih.gov:80/perl/blast2/>  
 CGAP-GAI: <http://lpgws.nci.nih.gov/>  
 ClustalW: <http://www.ebi.ac.uk/clustalw/>  
 dbSNP: <http://www.ncbi.nlm.nih.gov/SNP/>  
 GenBank: <http://www.ncbi.nih.gov/Genbank/>  
 HGVbase: <http://hgvbase.cgb.ki.se/>  
 HPRD: <http://www.hprd.org/>  
 HUGO Human Genome Nomenclature: <http://www.gene.ucl.ac.uk/nomenclature/>  
 LocusLink: <http://www.ncbi.nlm.nih.gov/LocusLink/>  
 InterProScan: <http://www.ebi.ac.uk/InterPro/scan.html/>  
 SIFT: [http://blocks.fhcr.org/~pauline/SIFT\\_seq\\_submit2.html/](http://blocks.fhcr.org/~pauline/SIFT_seq_submit2.html/)  
 SNP\_500: <http://bump.nci.nih.gov/home.cfm/>  
 Swiss-Prot: <http://us.expasy.org/sprot/>  
 MINT: <http://160.80.34.4/mint/>  
 The entire group A nsSNPs data: [http://www.ozceliklab.com/savas\\_cell\\_cycleA](http://www.ozceliklab.com/savas_cell_cycleA)  
 The entire group B nsSNPs data: [http://www.ozceliklab.com/savas\\_cell\\_cycleB](http://www.ozceliklab.com/savas_cell_cycleB)

### REFERENCES

- Evan GI, Vousden KH. Proliferation, cell cycle and apoptosis in cancer. *Nature* 2001;411:342–348.
- Malumbres M, Barbacid M. To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer* 2001;1:222–231.
- Knudson AG. Cancer genetics. *Am J Med Genet* 2002;111:96–102.
- Venkitaraman AR. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 2002;108:171–182.
- Robles AI, Harris CC. p53-mediated apoptosis and genomic instability diseases. *Acta Oncol* 2001;40:696–701.
- Reed N, Gutmann DH. Tumorigenesis in neurofibromatosis: new insights and potential therapies. *Trends Mol Med* 2001;7:157–162.
- Bell DW, Varley JM, Szydlow TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF, Birch JM, Li FP, Garber JE, Haber DA. Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. *Science* 1999;286:2528–2531.
- Ponder BAJ. Cancer genetics. *Nature* 2001;411:336–341.
- Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. *Nat Genet* 2003;Suppl:238–244.
- Chakravarti A. Population genetics—making sense out of sequence. *Nat Genet* 1999;21(1 Suppl):56–60.
- Miller RD, Kwok P-Y. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum Mol Genet* 2001;10:2195–2198.
- Taylor JG, Choi EH, Foster CB, Chanock SJ. Using genetic variation to study human disease. *Trends Mol Med* 2001;7:507–512.
- Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 2000;9:2403–2408.
- Shastri BK. SNP alleles in human disease and evolution. *J Hum Genet* 2002;47:561–566.
- Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol* 2001;307:683–706.
- Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863–874.
- Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 2002;12:436–446.
- Wang Z, Moutl J. SNPs, protein structure, and disease. *Hum Mutat* 2001;17:263–270.
- Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* 2002;30:3894–3900.
- Nourry C, Grant SG, Borg JP. PDZ domain proteins: plug and play! *Sci STKE* 2003;179:RE7.
- Pawson T, Nash P. Assembly of cell regulatory systems through protein interaction domains. *Science* 2003;300:445–452.
- Murray AW. Recycling the cell cycle: cyclins revisited. *Cell* 2004;116:221–234.
- Chan GK, Yen TJ. The mitotic checkpoint: a signaling pathway that allows a single unattached kinetochore to inhibit mitotic exit. *Prog Cell Cycle Res* 2003;5:431–439.
- Wijnhoven BP, Dinjens WN, Pignatelli M. E-cadherin-catenin cell-cell adhesion complex and human cancer. *Br J Surg* 2000;87:992–1005.
- Jhanwar-Uniyal M. BRCA1 in cancer, cell cycle and genomic stability. *Front Biosci* 2003;8:s1107–1117.
- Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. *Cancer Epidemiol Biomarkers Prev* 2004;13:801–807.
- Savas S, Kim DY, Ahmad MF, Shariff M, Ozcelik H. Systematic evaluation of functional consequences of nsSNPs in DNA repair pathway using computational tools. Submitted for publication.
- Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH. Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. *Genome Res* 2001;10:1259–1265.
- Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 2004;32:D35–40.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–311.
- Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res* 2002;30:387–391.
- Packer BR, Yeager M, Staats B, Welch R, Crenshaw A, Kiley M, Eckert A, Beerman M, Miller E, Bergen A, Rothman N, Strausberg R, Chanock SJ. SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res* 2004;32:D528–532.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. *Nucleic Acids Res* 2004;32:D23–26.
- Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* 2004;20:1006–1014.
- Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet* 2002;11:1987–1995.
- Kruglyak L, Nickerson DA. Variation is the spice of life. *Nat Genet* 2001;27:234–236.
- O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: Swiss-Prot and TrEMBL. *Brief Bioinform* 2002;3:275–284.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.
- Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* 2003;31:315–318.
- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, et al. Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;13:2363–2371.
- Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INteraction database. *FEBS Lett* 2002;513:135–140.
- Zdobnov EM, Apweiler R. InterProScan—an integration platform

- for the signature-recognition methods in InterPro. *Bioinformatics* 2001;17:847–848.
43. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet* 2001;109:678–680.
  44. Gotz C, Kartarius S, Scholtes P, Montenarh M. Binding domain for p21(WAF1) on the polypeptide chain of the protein kinase CK2 beta-subunit. *Biochem Biophys Res Commun* 2000;268:882–885.
  45. Zheng L, Pan H, Li S, Flesken-Nikitin A, Chen PL, Boyer TG, Lee WH. Sequence-specific transcriptional corepressor function for BRCA1 through a novel zinc finger protein, ZBRK1. *Mol Cell* 2000;6:757–768.
  46. Ouchi T, Lee SW, Ouchi M, Aaronson SA, Horvath CM. Collaboration of signal transducer and activator of transcription 1 (STAT1) and BRCA1 in differential regulation of IFN-gamma target genes. *Proc Natl Acad Sci USA* 2000;97:5208–5213.
  47. Park JJ, Irvine RA, Buchanan G, Koh SS, Park JM, Tilley WD, Stallcup MR, Press MF, Coetzee GA. Breast cancer susceptibility gene 1 (BRCA1) is a coactivator of the androgen receptor. *Cancer Res* 2000;60:5946–5949.
  48. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999;22:231–238.
  49. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet* 1999;22:239–247.
  50. Sunyaev SR, Lathe WC 3rd, Ramensky VE, Bork P. SNP frequencies in human genes: an excess of rare alleles and differing modes of selection. *Trends Genet* 2000;16:335–337.
  51. Graur D, Li W-H. Dynamics of genes in populations. In: *Fundamentals of molecular evolution*, 2<sup>nd</sup> edition. Sunderland, MA: Sinaur Associates, Inc.; 2000. p 41.
  52. Fay JC, Wu CI. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* 2003;4:213–235.
  53. Dean M, Carrington M, O'Brien SJ. Balanced polymorphism selected by genetic versus infectious human disease. *Annu Rev Genomics Hum Genet* 2002;3:263–292.
  54. Ohta T. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci USA* 2002;99:16134–16137.
  55. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 2003;33 Suppl:266–275.
  56. Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet* 2002;3:611–621.
  57. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, et al. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;420:520–562.
  58. Miller MP, Parker JD, Rissing SW, Kumar S. Quantifying the intragenic distribution of human disease mutations. *Ann Hum Genet* 2003;67:567–579.
  59. Grabowski B, Kelman Z. Autophosphorylation of archaeal Cdc6 homologues is regulated by DNA. *J Bacteriol* 2001;183:5459–5464.
  60. Kelman LM, Kelman Z. Archaea: an archetype for replication initiation studies?. *Mol Microbiol* 2003;48:605–615.
  61. Yu X, Chini CC, He M, Mer G, Chen J. The BRCT domain is a phospho-protein binding domain. *Science* 2003;302:639–642.
  62. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect in humans. *Hum Mol Gen* 2002;11:2463–2468.
  63. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. *Science* 2002;296:750–752.
  64. Noble ME, Endicott JA, Brown NR, Johnson LN. The cyclin box fold: protein recognition in cell-cycle and transcription control. *Trends Biochem Sci* 1997;22:482–487.
  65. Branden C, Tooze J. Folding and flexibility. In: *Introduction to protein structure*, 2<sup>nd</sup> edition. New York: Garland; 1999. p 107.
  66. Kobayashi H, Stewart E, Poon R, Adamczewski JP, Gannon J, Hunt T. Identification of the domains in cyclin A required for binding to, and activation of, p34cdc2 and p32cdk2 protein kinase subunits. *Mol Biol Cell* 1992;3:1279–1294.
  67. Matthews J.M, Sunde M. Zinc fingers—folds for many occasions. *IUBMB Life* 2002;54:351–355.
  68. de Bruin A, Maiti B, Jakoi L, Timmers C, Buerki R, Leone G. Identification and characterization of E2F7, a novel mammalian E2F family member capable of blocking cellular proliferation. *J Biol Chem* 2003;278:42041–42049.
  69. Goodrich DW. How the other half lives, the amino-terminal domain of the retinoblastoma tumor suppressor protein. *J Cell Physiol* 2003;197:169–180.