

## **SNPs - An Exploratory Study**

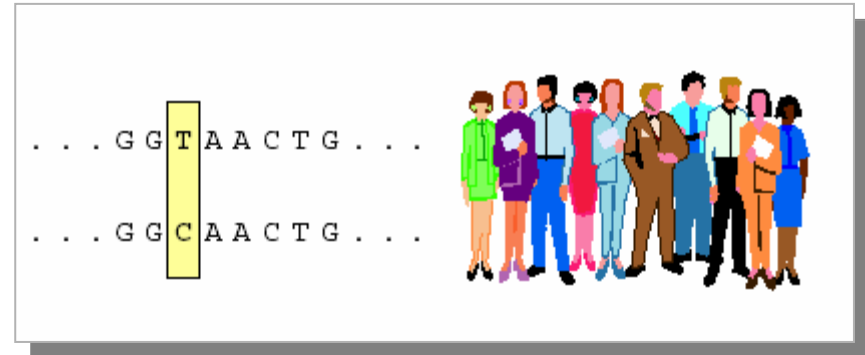
**Prediction of functionally important SNPs using computational tools**

By: M. Farhan Ahmad  
Summer 2002

## 1. Introduction

### What are SNPs?

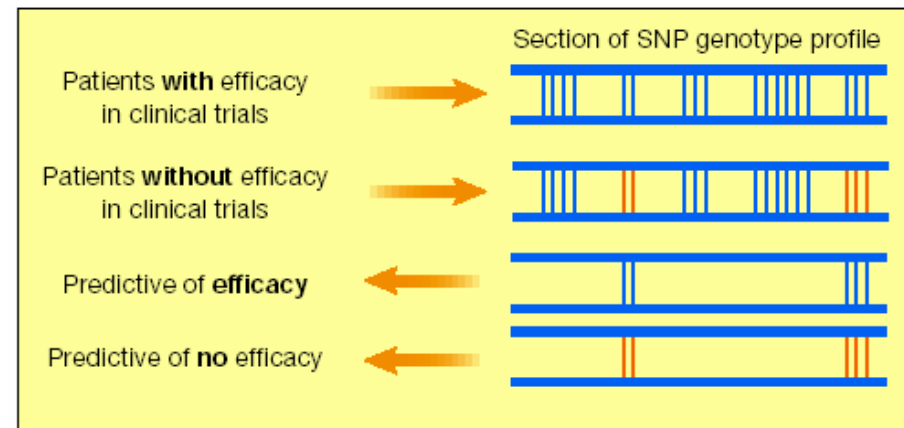
SNPs stands for **Single Nucleotide Polymorphism's**. They are single base pair changes or variations within a genome. For example, a C changing to A/G/T.



- Most common type of variation - account for 90% of sequence differences (Collin et al., 1998).
- Found in the human genome at an average frequency of 1 per 1000 base pairs (Brookes, A.J., 1999)
- Most are found in non-coding regions because coding regions are under selective pressure to be conserved.

### Why are they important?

- Diagnostics: Compare SNPs (in or close to genes related to a disease) in effected vs. normal individuals.
- Drug Design: Differences in drug response w.r.t. SNP patterns.



## 2. Importance to Cancer and Past Success

### Common Variations

Diseases such as *cancer* caused by common defects.

Therefore, find many SNPs and compare SNP profiles of diseased and normal individuals in terms of frequencies and modification of SNP expressions etc.

Most SNPs predictive of abnormal phenotypic expressions are found around or within genes involved in mechanisms related to diseases.

One strategy is to look at SNPs within coding regions.

### Past Success

French Group - 1996

By comparing “risk genes” b/w normal and those affected by Crohn’s disease, they found *13 candidate SNPs*.

Crohn’s patients carried at least one of three candidate SNPs more frequently than normal individuals. All three fell within a specific gene’s (NOD2) coding region.

## 3. Our Approach

- A) Search for SNPs in Cancer related genes
- B) Combining information from different SNP databases
- C) Searching for Protein Domains
- D) Visualizing information - localization of SNPs in protein domains
- E) How tolerated a SNP is w.r.t. how conserved the original nucleotide is

## 3. Our Approach

### A) Search for SNPs in Cancer related genes

**We used 4 databases to search for SNPs:**

**i) HGVbase**

<http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+query+-l+hgbase>

**ii) dbSNP**

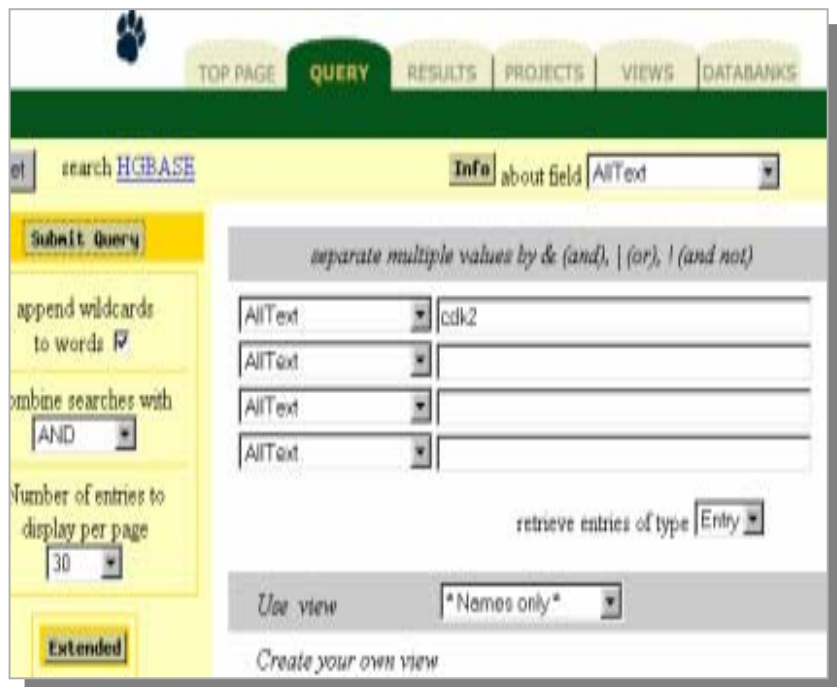
<http://www.ncbi.nlm.nih.gov/SNP/snpLocus.html>

**iii) CGAP-GAI**

<http://lpgws.nci.nih.gov:82/perl/snpbr>

**iv) GeneSNP**

<http://www.genome.utah.edu/genesnps/>



[TOP PAGE](#)
[QUERY](#)
[RESULTS](#)
[PROJECTS](#)
[VIEWS](#)
[DATABANKS](#)

search [HGVBASE](#)
 Info about field AllText

append wildcards to words

combine searches with

Number of entries to display per page

separate multiple values by & (and), | (or), ! (and not)

AllText

AllText

AllText

AllText

retrieve entries of type

Use view

[Create your own view](#)

This entry is from: [HGVBASE SNP001026531](#)

[HGVBASE](#)

ID	SNP001026531
VarType	SNP
VarStatus	Proven
CurationStatus	---
DBRefMolecule	EMBL::AC025162.30:99770 (FlankSeq: <a href="#">100</a> <a href="#">200</a> <a href="#">300</a> <a href="#">Choose</a> )
bpStreamSeq	TTGGAAAAAGCATCGAGAGGGACAGT
Allele	A
Allele	C
DnStreamSeq	CGGAGTTGTGTACAAAGCCAGAAAC
XX	
GeneID	GID000001217
GeneSymbol	<a href="#">CDK2</a>
GeneName	Cyclin-dependent kinase 2
GeneType	Functional Gene
GeneRegion	5-EX1CD8+
ProtSeqChanged	Yes
Feature	DNA
Feature	/seq:0
Feature	DNA
Feature	/seq:c
Feature	RNA
Feature	/seq:a
Feature	/codon:acc
Feature	RNA
Feature	/seq:c
Feature	/codon:acc
Feature	AA
Feature	/seq:Y
Feature	AA
Feature	/seq:0
XX	
DbXRef	EMBL::AC025162.30
DbXRef	<a href="#">RefSeq</a> ::NM_001798.1
DbXRef	<a href="#">EGP Database</a>
XX	
PopulationID	POP000001262 (from SRC000000888)
Population	NIH Coriel panel (346 individuals)
Frequency	99.40
Frequency	0.60
XX	
SourceID	SRC000000888
Citation	EGP Database
Submitter	Gene K. S. Wong ( <a href="#">SUB000000053</a> )
SourceComment	cdk2-e1+c73

**HGVbase**

Display:  ?

Organism:  ?

Locus associated with:  ?

Query:   ?

**Query examples:**

Gene Symbol	LPL
Gene Product	lipoprotein lipase
Accession Number	NP_000228
Gene Ontology (GO)	fatty acid metabolism

Gene Model (contig mRNA transcript) information from genome sequence for [XCM\\_049150](#)

Contig accession	Contig position	Protein accession	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
<a href="#">NT_009458</a>	884126	<a href="#">XP_049150</a>	contig reference	C	Thr [T]	2	200
			non-synonymous change	G	Ser [S]	2	200

**LocusLink:** no link established by BLAST analysis of mRNA sequences

### Integrated Maps: ↑

**NCBI Map Viewer:** [rs2069413](#) maps exactly once on NCBI human [chromosome](#)

Chromosome	Contig accession	Contig Position	Chromosome Position	Hit orientation
12	<a href="#">NT_009458.10</a>	884126	57300525	minus strand

**Project Ensembl:** Query [rs2069413](#) in Ensembl.

**UC Santa Cruz Genome Assembly:** Query [rs2069413](#) on the Santa Cruz Assembly.

### Variation Summary: ↑

Assay sample size (number of chromosomes): 222  
 Population data sample size (number of chromosomes): 708  
 Total number of populations with frequency data: 1  
 Total number of individuals with genotype data: 90  
 Average estimated heterozygosity: 0.012

Average Allele Frequency:

C	0.989
G	0.011

Average Genotype Frequency:

CC	0.989
CG	0.011

**dbSNP**

## CGAP SNP index

This is an index of [UniGene](#) assemblies and our most current build of predicted SNP locations, searchable by gene, description, or nucleotide sequence. See also our [translations](#) of SNPs onto reference sequences.

Search for:  Organism/dataset:

Keyword, gene symbol ([Uni](#)), accession, or location:

ID #:

NOTE: wildcards (\*) may be used, for example, \*GST\*\* would match GSTA1, GSTA2, GSTM1, etc. If multiple keywords are specified they must appear sequentially in the description to match.

Search by:

[BLAST](#) nucleotide search:

Database:

Expect value:

## dbSNP <=> refseq

- [Get results in tab-delimited format](#)

Search results for [NM\\_001798](#) (Homo sapiens cyclin-dependent kinase 2 (CDK2), transcript variant 1, mRNA):

Accession	Version	Snp type	Snp id	Base number	CDS SNP	Overlap bases	Identity percent	Validated
NM_001798	2	GAI	<a href="#">888199</a>	818	<a href="#">Chr19:53u</a>	611	100	-
NM_001798	2	GAI	<a href="#">888200</a>	1881	-	999	99	-

# CGAP-GAI



**GeneSNPs** My Genes ▾ Gene Lists ▾ Resources ▾ Help ▾ Search

GeneSNPs list of genes for search:

[Search HUGO with this query.](#) Click for [Help](#)

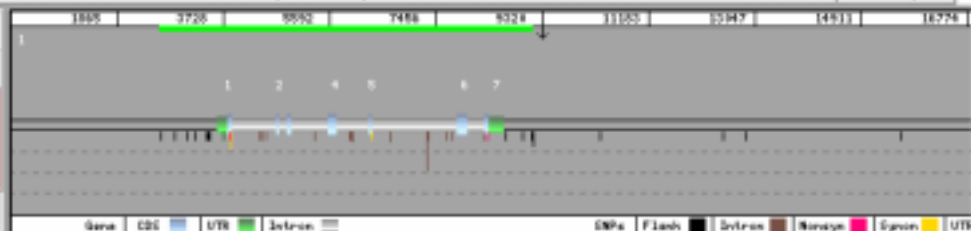
SNPCard	Length	SNPs	Status	Class	Name
<a href="#">CDK2</a>	6514	38	OPSCU	Cell Cycle	cyclin-dependent kinase 2

## GeneSNPs

**GeneSNPs** My Genes ▾ Gene Lists ▾ Resources ▾ Help ▾ Search

**CDK2** cyclin-dependent kinase 2 [Cell Cycle](#)

Menu  
 GenoTables  
 NIEHS\_RESEQ  
 Submitted  
 UUGC



Controls Image: Full View ◀ 1000 ▶ R SNP: C All SNPs High Freq All Subs

**All Coding SNPs from 0.0kb to 18.6kb**

Implication	Position	Allele	Freq	Heter.	#Chrom.	PopID	Submitter	SS#
nonsyn Tyr 15 Ser	3836	A/C	0.01		83	<a href="#">NIHPDR</a>	UWGC-NIEHS	<a href="#">4318859</a>
synon Glu 26 Glu	3876	G/A	0.10		162	<a href="#">NIH90</a>	UUGC-NIEHS	<a href="#">2981455</a>
synon Glu 26 Glu	3876	A/G	0.07		81	<a href="#">NIHPDR</a>	UWGC-NIEHS	<a href="#">4318858</a>
synon Glu 195 Glu. 3'splice(-4)	6357	A/G	0.02		89	<a href="#">NIHPDR</a>	UWGC-NIEHS	<a href="#">4318861</a>
synon Glu 195 Glu. 3'splice(-4)	6357	G/A	0.03		180	<a href="#">NIH90</a>	UUGC-NIEHS	<a href="#">2981463</a>
nonsyn Thr 290 Ser	8381	C/G	0.01		176	<a href="#">NIH90</a>	UUGC-NIEHS	<a href="#">2981470</a>
nonsyn Thr 290 Ser	8381	C/G			42		YUSUKE	<a href="#">3205744</a>

## B) Combining information from different databases

Each database provides SNPs based on different mRNA sequences. However, all provide short sequences upstream and downstream of the SNPs.

This information can be used to localize all these different SNPs on to a common reference strand using SNP Blast - <http://lpgfs.nci.nih.gov:82/perl/blast2>

### BLAST against gene transcripts

This form uses the [BLAST](#) program to translate a specified SNP onto a GenBank sequence and then do a substitution. For example, the sequence TGAAAGATAAGAAAGAACTAGAAGGT[G/T]CTGGGAAGGTGAGTCAAATAAAT. Click [here](#) to see a demo.

This is an experimental service; please report any problems to [edmonson@msk.gov](mailto:edmonson@msk.gov) with as much detail as possible.

SNP sequence:

[Instructions...](#)

Select:

Search: -or-

Single sequence:

```

      8
      agaaagaactagaaggtkctgggag
      |||
3081 agaaagaactagaaggtgctgctactgaagcaatagaaaactccgaaacggttgtttctt
      K K E L E G A G K I A T E A I E N F R T V V S L
      887 889 891 893 895 897 899 901 903 905 907 90
      86 888 890 892 894 896 898 900 902 904 906 908

3151 gectcaggagcagaagtttgaacatatgtatgctcagagtttgcaggtaccatcacgaaactctttgagg
      T Q E Q K F E H E Y A Q S L Q V P Y R N S L R
      9 911 913 915 917 919 921 923 925 927 929 931
      910 912 914 916 918 920 922 924 926 928 930 932
    
```

id	description	score	overlap bases	overlap identity %
<a href="#">NM_000927</a>	Homo sapiens ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), mRNA	66.8	35	100.0

- Query sequence: TGAAAGATAAGAAAGAACTAGAAGGT[G/T]CTGGGAAGGTGAGTCAAATAAAT. The SNP site is represented in the BL search by an [ambiguity symbol](#)
- Template sequence: [NM\\_000927](#) (Homo sapiens ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), mRNA ), version 2
- Protein change: yes (Ala892Ser)
- Click [here](#) to search for other known SNPs in [NM\\_000927](#).

## C) Protein Domain Searches

The last step is to find the domains in the gene product and see if the SNPs found are in the domains or not. SNPs in domains are more likely to cause functional changes than SNPs in other parts of a protein.

We use 3 databases to search for protein domains:

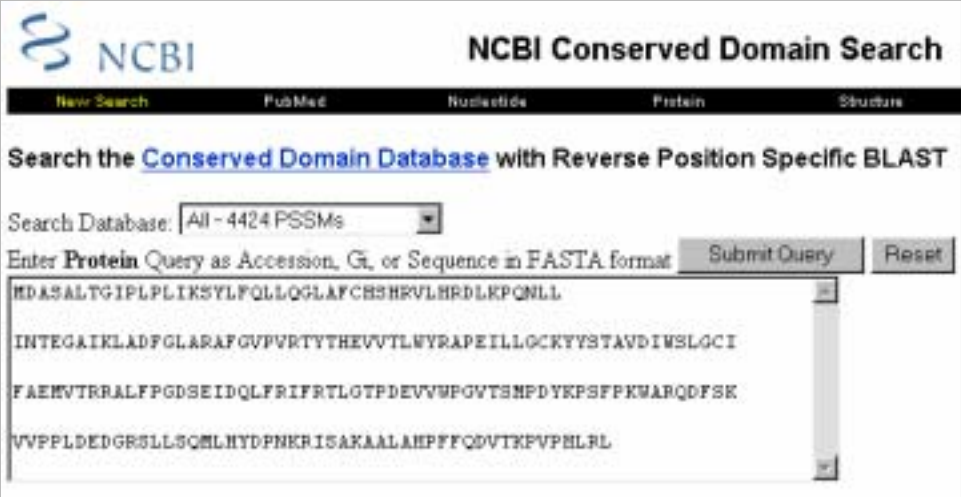
.**NCBI CD-Search** - <http://www.ncbi.nlm.nih.gov/SNP/>

.**Pfam** - <http://pfam.wustl.edu/hmmsearch.shtml>

.**SMART** - <http://smart.embl-heidelberg.de>

In each case you simply copy and paste the amino sequence of each gene on to which all SNPs have been standardized. Record the amino acid positions for the start and end of all domains.

Note, the amino acid sequence is found by going to <http://www.ncbi.nlm.nih.gov/> and entering the nucleotide accession number.



**NCBI Conserved Domain Search**

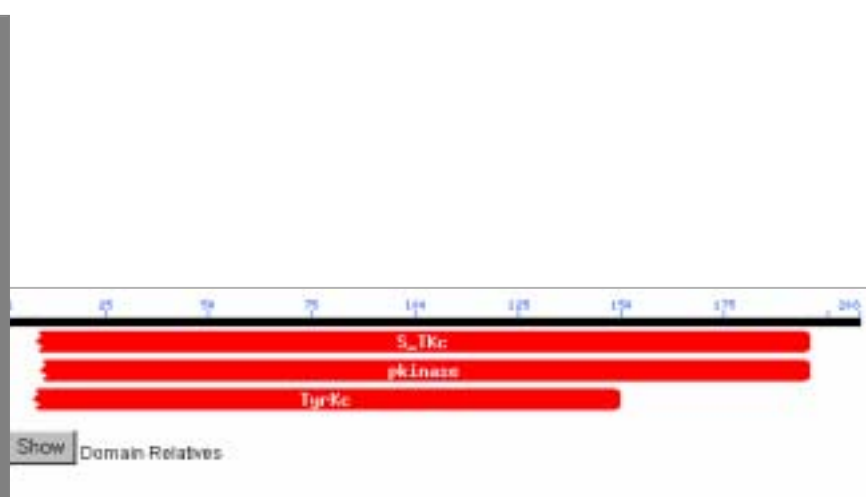
Search Database: All - 4424 PSSMs

Enter **Protein** Query as Accession, GI, or Sequence in FASTA format

Submit Query    Reset

```

MDASALTGIPLPKSYLQQLQGLAFCHSHRVLHRDLKPNLL
INTEGAIKLADFGLARAFGVVVRTYTHEVVTLEWYRAPEILLOCKYVSTAVDIWSLGC
FAEMVTRRALFPGDSEIDQLFRIFRITLQTPDEVVMPGVTSNPDYKPSFKWARQDFSK
VVPPLDEDGRSLLSQMLHYDPNKRISAKAALAHPPFFQDVTKVPVHLRL
    
```



Sequence alignment visualization showing domains: S\_TKc, pkinase, TyrKc.

Show Domain Relatives

3D structure. To display structure, download [Cn3D!](#)

PSSMs producing significant alignments:

Accession	Description	Score	E value
<a href="#">gn CDD 5785</a>	smart00220, S_TKc, Serine/Threonine protein kinases, catalytic...	195	8e-52
<a href="#">gn CDD 5852</a>	pfam00069, pkinase, Protein kinase domain	171	2e-44
<a href="#">gn CDD 3848</a>	smart00219, TyrKc, Tyrosine kinase, catalytic domain; Phosphot...	79.1	1e-16

---

[gn|CDD|5785](#), smart00220, S\_TKc, Serine/Threonine protein kinases, catalytic domain; Phosphotransferases. Serine or threonin

CD-Length = 256 residues, only 63.7% aligned  
Score = 195 bits (498), Expect = 8e-52

```

Query:  8  GIPLPKSYLQQLQGLAFCHSHRVLHRDLKPNLLINTEGAIKLADFGLARAFGVVVR  67
Sbjct: 94  RLSEDEARFYARQILSALEYLHENGIIHRDLKPFENILLDSDGHVKLADFGLAKQLDGGGT 153

Query:  68  TYTHEVVTLEWYRAPEILLOCKYVSTAVDIWSLGCIFAEEMVTRRALFPGDSEIDQLFRIFR 127
Sbjct: 154 LLTTFVGTPEYMAPEVLLG-EGYGRKAVDIWSLGVILYELLTGRKPPFGDDQLDALFKKIG 212

Query: 128  TLQTPDEVVMPGVTSNPDYKPSFKWARQDFSKVVPPLDEDGRSLLSQMLHYDPNKRISA 187
Sbjct: 213  K-----P-----KISPEAKDLIKLLVKDPEKSLTA 247

Query: 188  KAALAHPPFF 196
Sbjct: 248  EEALEHPPFF 256
    
```

---

[gn|CDD|5852](#), pfam00069, pkinase, Protein kinase domain.

## NCBI CD-Search

**Washington University in St. Louis** **Pfam**  
Analyze a query sequence

[Pfam \(St. Louis\)](#) | [Pfam \(Cambridge\)](#) | [Pfam \(Stockholm\)](#) | [Pfam \(France\)](#) | [HMMER](#) | [TIGR](#)  
[Home](#) | [Protein search](#) | [DNA search](#) | [Browse Pfam](#) | [Keyword search](#) | [Send Pfam](#) | [Contact](#)

### Analyze a query sequence by searching Pfam HMMs

**Important Note:** Internet Explorer on the Macintosh has a bug which will often cause searches to return a message saying that you already have a search in the queue. We're working on this, meanwhile, try using Netscape or another browser.

**Protein sequence query.** Cut and paste your sequence here. FASTA format or raw sequence are acceptable.

```
MFASALTSIPSLLEKDTLFGQLLQGLAFCHSEKFLKDFGLFQKLL  
INTEGALKLADFGKARAFQVFRITSTHEVVTLSRAPEILGGCKYVSTAVDINSKGC  
FAKVFTRALFFGQSEIDQLYRIFRILGTPQEVVVFQVTRFDYEPFFQVARDQFSE  
VVVFLDEKQKLLGQKLYDPMKRIKAAALAHFFQGVTRFVFLSL
```

Or: Select the query sequence file you wish to use

**PFAM**

## Pfam HMM search results, glocal+local alignments merged (Pfam\_ls+Pfam\_fs)

[\[Go here for an explanation of the format of the results\]](#)

Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!! <a href="#">pkinase</a>	1	196	1	294	104.3	<b>1.5e-28</b>		glocal Protein kinase domain



**pkinase 1-196**



## SMART

Domains within the query sequence of 208 residues



Mouse over domain / undefined region to see the limits; click on it to go to further annotation; right-click to see Transmembrane segments as predicted by the [TMHMM2](#) program (■), coiled coil regions determined by the [Coiled](#) determined by the [SEC](#) program (—) signal peptides determined by the [SignalP](#) program (—), GPI anchors (—) for hits in the schnipsel database and (■) for hits against PDB. Regions containing repeats detected by [ ]

### Architecture analysis

[Display](#) all proteins with similar domain [organisation](#).  
[Display](#) all proteins with similar domain [composition](#).

---

The SMART diagram above represents a summary of the results shown below. Domains with scores less than 0.01 are also not shown when two or more occupy the same piece of sequence; the priority for display is given by SM Transmembrane > Coiled coil > Low complexity. In either case, features not shown in the above diagram are not

Confidently predicted domains, repeats, motifs

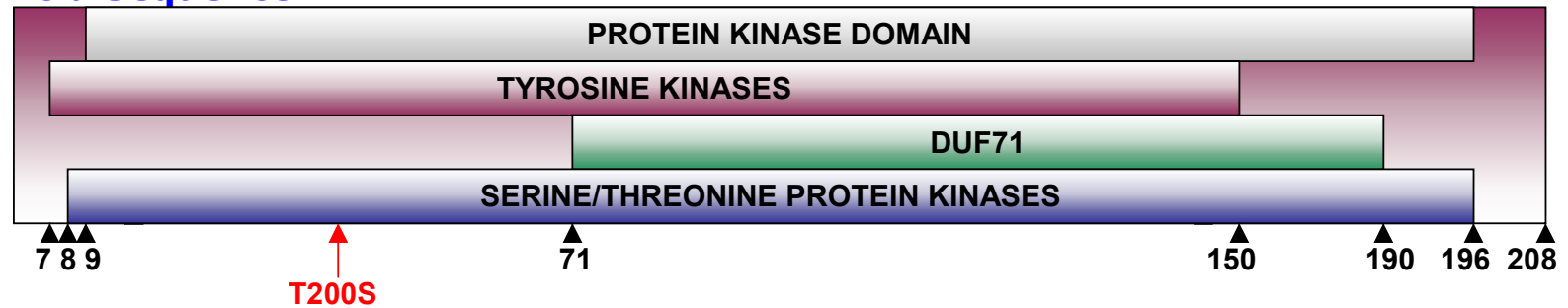
name	begin	end	E-value
<a href="#">STYKc</a>	4	196	2.07e-02

## D) Visualizing Information

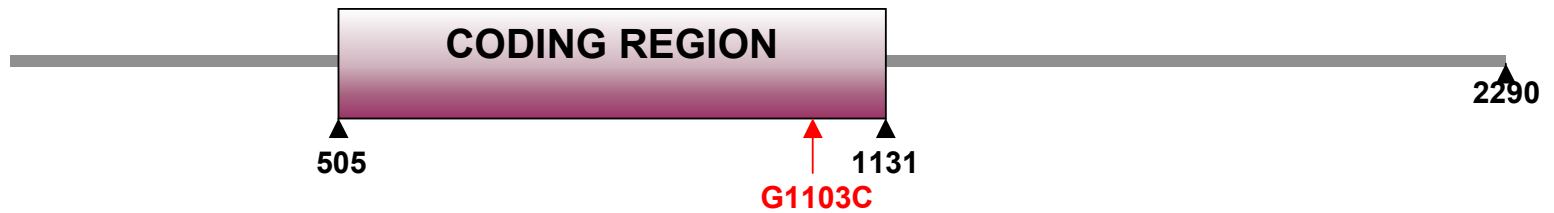
### SNP Locations

#### CDK2

#### Amino Acid Sequence




#### Nucleotide Sequence



## E) SNPs - Amino Acid change tolerated or not

We want to check whether the amino acid change is tolerated or not. This is done by comparing the SNP change across several species.

The program we use is [SIFT v.2](http://blocks.fhcrc.org/~pauline/SIFT_seq_submit2.html) - [http://blocks.fhcrc.org/~pauline/SIFT\\_seq\\_submit2.html](http://blocks.fhcrc.org/~pauline/SIFT_seq_submit2.html)



**SIFT v.2**

SIFT version 2. Sequences are chosen by a conservation cutoff [more details]. Sequences are chosen. The procedure is described in *Genome Research* 10:117-125 (2000).

This will take 10-15 min. because we must search your protein sequence against a database to pick out [sequence and related sequences](#) or [aligned sequences](#) if you already have them.

[[SIFT home page](#)] [[SIFT help](#)] [[Preventing connection failures](#)]

Protein sequence

Name of file containing query sequence ([fasta format](#))

-or-

Paste in your query sequence ([fasta format](#))

```
MDASALTGIFLPLIKSYLFQLLQGLAFCHSERVLRDLKPQMLL
INTEGAIKELADFGLARAFGVPVRYTYTHEVVTLWYRAPEILLGCKYYSTAVIDNSLGGI
FAENVTRRALFPGDSEIDQLFRIFRTLGTPEVWVWPGVTSNPDYKPSFPRMARQDSEK
VVPPLDEDGRSLLSQNLHYDPNKRISAKAALAHPI
```

Predict Not Tolerated	Position	Seq	Rep	Predict Tolerated
dh g n e c s w k r y p q t a f v i	1M	0.80	L M	
w c f y m h i v p g a k L R S	2E	0.88	T N Q D E	
	3D	0.96	M i p v g L S E D T A N Q K R	
k q h n r d g e p c t s a m v i w l	4Y	0.97	F Y	
	5T	0.98	C p H g L r I D N V T A S K Q E	
c w f m d i v y p h l a t e S N G	6K	0.98	O R K	



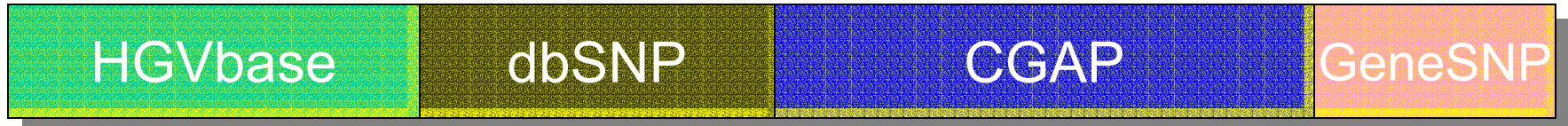
## 4. RESULTS

### A) Gene Families Studied - 71 Cell Cycle Genes

- Cell Division Cycle (CDC) - 22
- Cyclin (CCN\_) - 14
- Cyclin-dependent Kinase (CDK) - 10
- Cyclin-dependent Kinase Inhibitor (CDKN) - 7
- Growth Arrest and DNA-damage-inducible (GADD) - 3
- Other genes - 15

## B) Statistical Results

Total: 337 SNPs



Unique a.a.: 164 changes

After SNP Blast  
Unique a.a.: 131 changes

30 Unique a.a. changes  
Not Tolerated

38 Unique a.a. changes  
in Functional Domains

## Unique a.a. changes  
Not Tolerated

## Unique a.a. changes  
Not Tolerated



# Single Nucleotide Polymorphisms (SNPs)

June 19, 2002



SNPs Present in Functional Domains and Not Tolerated						
Gene	A.A. Change	Domains Affected		Validation Status		
1	CDK5	Asp38Tyr	S_TKc	P_Kinase	Tyr K	U
2	CDK9	Val45Leu	S_TKc	P_Kinase	Tyr K	U
3	CDC10	Ser45Leu	GTP_CDC	Feo_B	RAB	U
4	CDC14A	Trp182Cys	PTPc_DSPc	Y-phosphatase		S
5	CDC2	Arg59Cys	S_TKc	P_Kinase	Tyr K	U
6		Arg59Gly	S_TKc	P_Kinase	Tyr K	U
7	CDC20	Leu467Pro	WD40			
8	CDC25B	Pro492Ser	Rhodanese	RHOD		
9	CDC27	Gly574Arg	TPR			
10		Asn575Asp	TPR			
11		Tyr635Asn	TPR			
12	DDC	P210L	pyridoxal_deC			

SNPs Not Present in Functional Domains and Not Tolerated				Validation
1	CCND3	Ala15Val		
2	CDC20	Ala117Val		
3		Ile405Thr		
4		Leu493Phe		
5	CDC23	Pro3Leu		
6	CDC25B	Ser590Thr		
7	CDC27	Phe26Ser		
8		Leu27Pro		
9		Tyr48Ser		
10		Tyr73Cys		
11		Ser230Tyr		
12		Glu534Lys		
13		Ser566Leu		
14	CDC2L5	Thr671Ala		
15	CDC37	Gly360Glu		
16	CDKN1A	Asp149Gly		
17	CDKN1B	Arg15Trp		
18	DDC	M17V		
19	PKMYT1	Val445Ala		

## ACKNOWLEDGEMENTS

**Sevtap Savas**

Hilmi Ozcelik

Hamdi Jarjanazi

Priscilla Chan

Jane Figueiredo

Susan Lau

Hong Li

Venus Onay

Sean Wells

Denise Yee