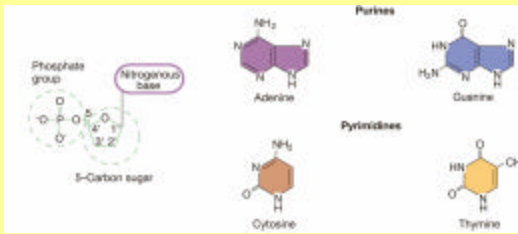


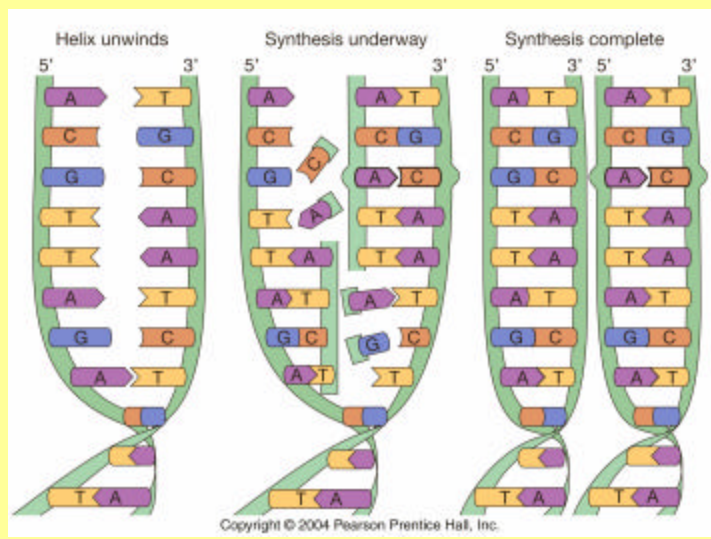
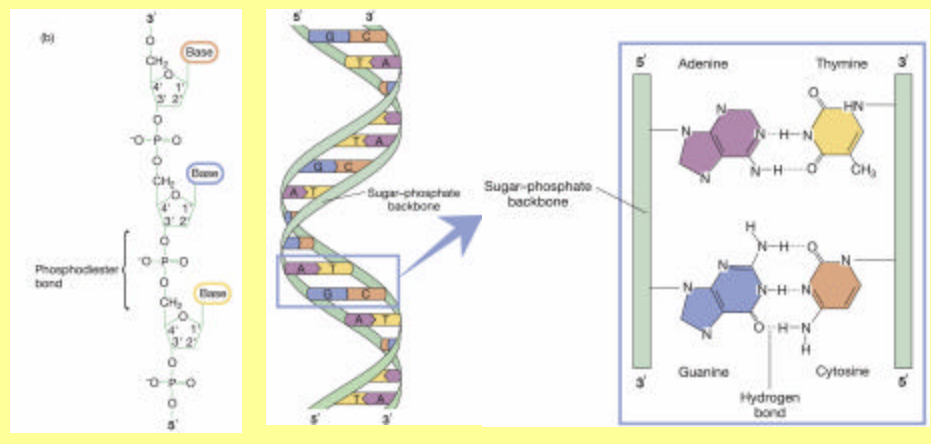
## Research Topic

Computational analyses of human  
disease mutations

Mutations are the ultimate source  
of genetic variation



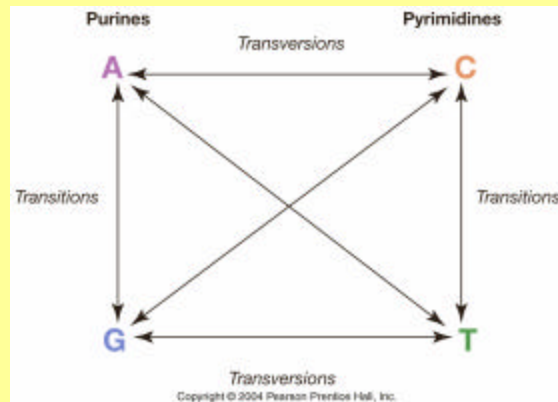
“Opposite” nucleotide types bond to form double-helix



Point mutations are caused by DNA polymerase errors during replication: incorrect nucleotides can sometimes be incorporated into the replicated strand

## Point mutation types

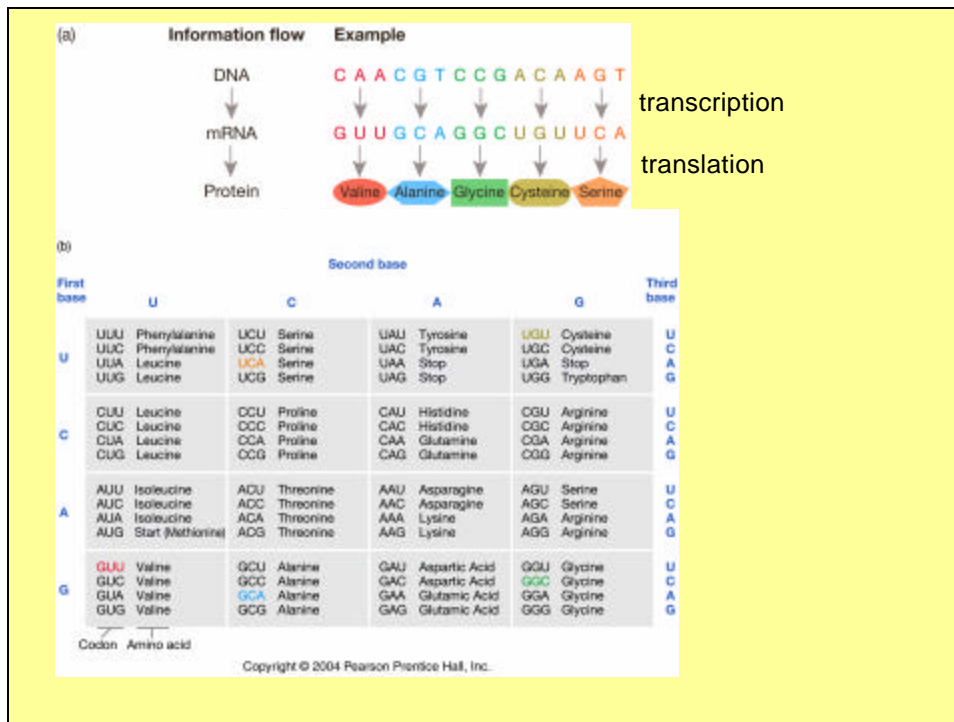
Purine - Purine: Transition  
Pyrimidine - Pyrimidine: Transition  
Purine - Pyrimidine (or vice-versa): Transversion



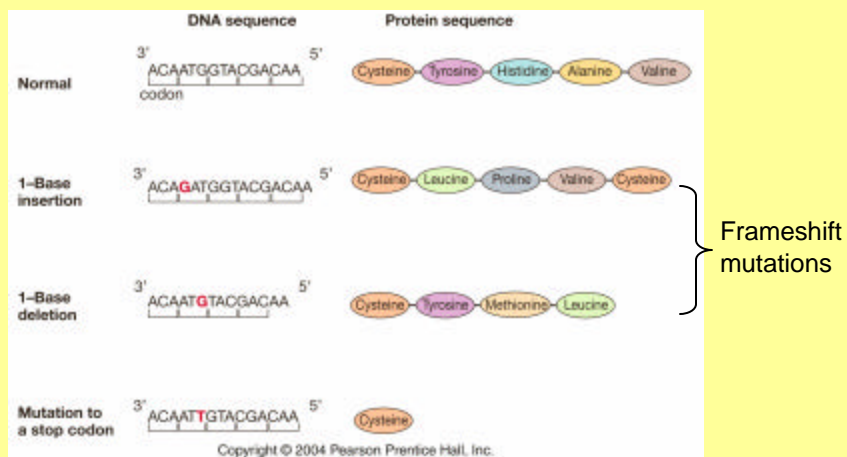
Although the probability of a transversion mutation is initially higher, transition mutations are more commonly observed

## Types of point mutations (cont.)

- Silent mutations: do not cause changes to amino acid sequences
  - Also called synonymous mutations
- Replacement mutations: alter the originally-encoded amino acid
  - Also called non-synonymous mutations



## Other types of mutations: “loss of function” mutations



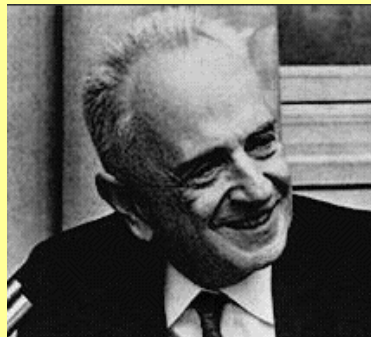
Are there common attributes of deleterious vs. non-deleterious point mutations?

An example using data from human disease mutations and amino acid sequences from different model organisms (i.e., an evolutionary analysis)

A list of different human disease mutation databases:

[http://archive.uwcm.ac.uk/uwcm/mg/docs/oth\\_mut.html](http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html)

**“Nothing in Biology Makes Sense Except in the Light of Evolution”**



Theodosius Dobzhansky

## Statement made by my dentist:

“Someone told me scientists believe that evolution is a *fact*, but gravity is *only a theory*”

## Some definitions...

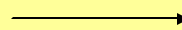
- Hypothesis:
  - A tentative explanation for natural processes that is waiting to be tested. Hypothesis tests can produce *facts*
  - A concept that is not yet verified, but if true would explain certain observations or phenomena
- Theory: Theories are *big deals!*
  - A general principle, based on *facts*, that explains or predicts events or phenomena
  - A theory is generally accepted as valid due to having survived repeated testing
  - Theories provide *mechanisms*, or help us understand *why* things work

Many different scientists



Hypothesis A

Hypothesis B



Theory!!!

## Statement made by my dentist:

“Someone told me that scientists believe that evolution is a *fact*, but gravity is *only a theory*”

- “Evolutionary Theory” is a theory based on accumulated sets of facts!
- Gravity is also a fact! Gravitational theory, however, provides the *mechanistic* explanation for how gravity works!
- Other theories: Theory of Relativity, Quantum Theory, Probability Theory, Big Bang Theory, Game Theory, Chaos Theory

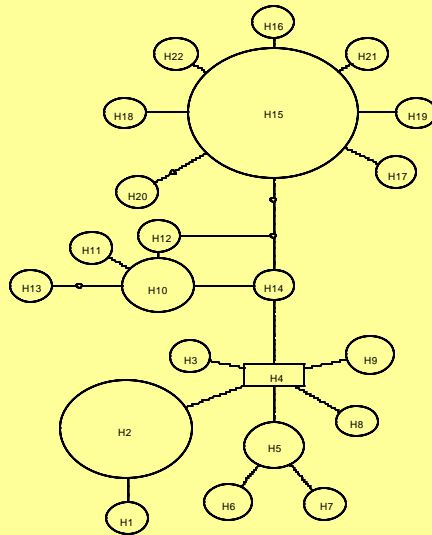
## Something to consider...

- In modern human vernacular, the word “theory” is often used when the word “hypothesis” would be much more appropriate
- How often do you hear people use the word “theory”?
- From now on, ask yourself if their usage is technically correct!!!

## Back to the topic at hand...

- Evolutionary (long term fates) of new mutations
  - Deleterious mutations
    - Decrease individual fitness
    - Eliminated or reduced in frequency by natural selection
  - Neutral mutations
    - Have no effect on an individual's fitness
    - May become widespread or may become extinct. The evolutionary process is stochastic for neutral mutations
  - Advantageous mutations
    - Increase individual fitness
    - Favored by natural selection and therefore become common

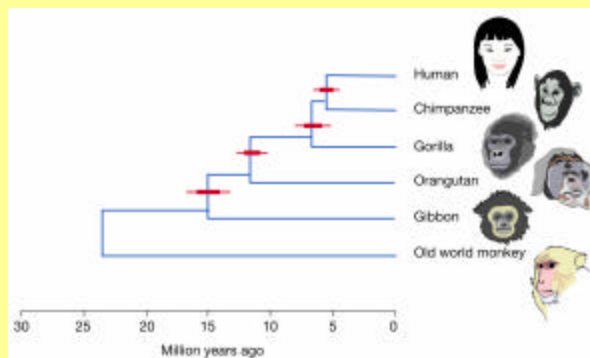
## Mutations are the ultimate source of genetic variation



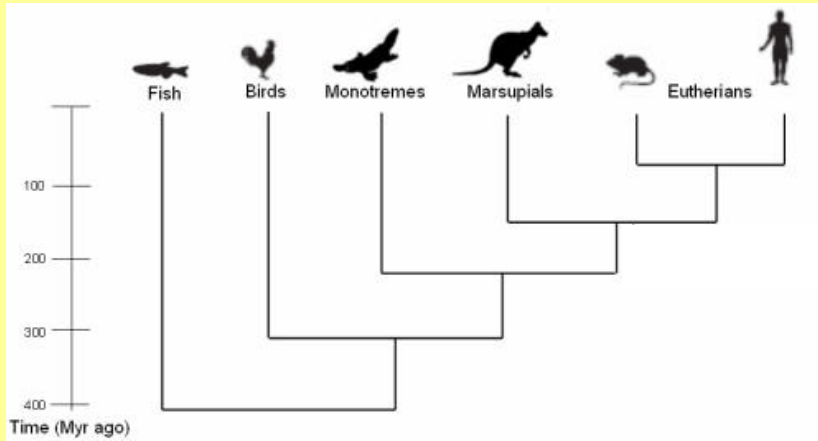
Haplotype network illustrating genealogy of alleles

## Insights from analyses of DNA sequences:

- Split between **orangs** and **chimp-human-gorilla** clade 10 - 13 mya
- Split between **gorillas** and **human-chimp** clade 6 - 9 mya
- Split between **human** and **chimp** from common ancestor 5 - 7 mya

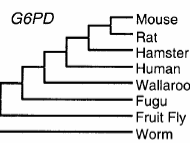
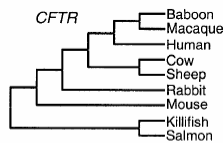


From Stauffer et al. (2001)



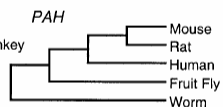
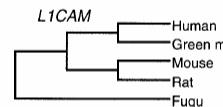
Phylogenies can potentially shed light on the nature of selectively neutral or advantageous mutations. But how can we come to understand the nature of deleterious mutations? What makes a mutation deleterious?

**Cystic fibrosis**



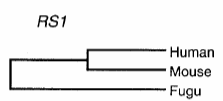
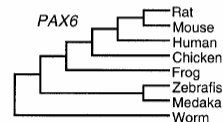
**Severe enzyme deficiencies (Type I) to mild anemias (Types II, III, & IV)**

**MASA syndrome**



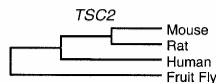
**Phenylketonuria**

**Developmental eye anomalies**



**X-linked retinoschisis**

From Miller and Kumar (2001)



**Tuberous sclerosis**

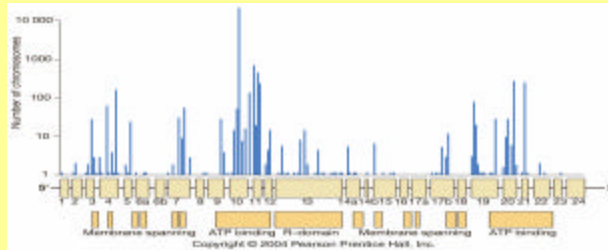
**Table 1.** Disease genes, database web sites and numbers of mutations analyzed from each database

Gene	Web-address for database (reference no.)	Number of mutations analyzed (disease/polymorphic/silent) <sup>a</sup>	****
<i>CFTR</i>	www.genet.sickkids.on.ca/cftr	429/32/61	
<i>G6PD</i>	http://rialto.com/favism/mutat.htm (32)	110/-- <sup>b</sup>	
<i>LI CAM</i>	dnalab-www.uia.ac.be/dnalab/11/ (37)	48/--	
<i>PAH</i>	http://www.mcgill.ca/pahdb/ (38)	270/--	
<i>PAX6</i>	www.hgu.mrc.ac.uk/Softdata/PAX6/ 29/-- (39)		
<i>RS1</i>	www.dmd.nl/rs/rs.html	71/--	
<i>TSC2</i>	expmed.bwh.harvard.edu/ts	47/18/33	

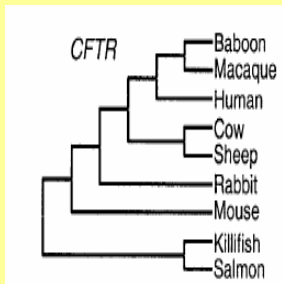
<sup>a</sup>Disease mutations refer to those amino acid changes that produce a disease phenotype. Polymorphic mutations are amino acid changes that are presumably not disease related. Silent mutations are DNA sequence changes that do not alter the encoded amino acid.

<sup>b</sup>The database analyzed contained 48 type I mutations that result in chronic non-spherocytic hemolytic anemic and 62 less severe types II, III or IV mutations.

Human mutations along the CFTR gene



### Amino acid sequence



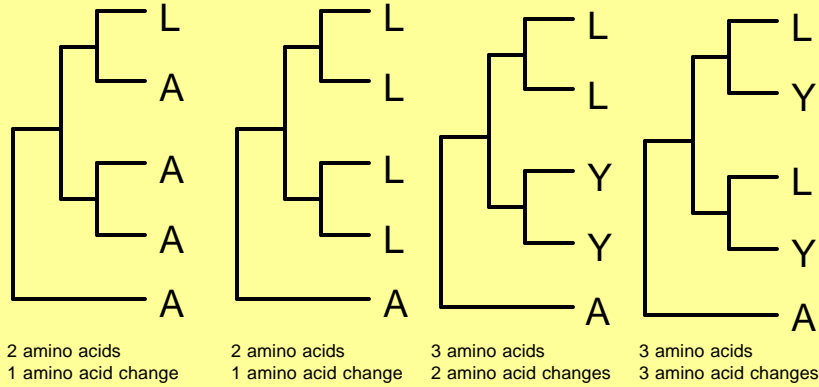
MQRSPLEKASVVS~~K~~LF~~F~~SWTRPILR~~R~~KGYRQRL~~E~~LS~~D~~IYQI~~P~~SAD~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKASVVS~~K~~LF~~F~~SWTRPILR~~R~~KGYRQRL~~E~~LS~~D~~IYQI~~P~~SAD~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKASVVS~~K~~LF~~F~~SWTRPILR~~R~~KGYRQRL~~E~~LS~~D~~IYQI~~P~~S~~V~~D~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKASVVS~~K~~V~~F~~FSWTRPIL~~R~~KGYRQRL~~E~~LS~~D~~IYH~~I~~SS~~S~~D~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKASVVS~~K~~LF~~F~~SWTRPIL~~R~~KGYRQRL~~E~~LS~~D~~IYH~~I~~SS~~S~~D~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKAGVLS~~K~~LF~~F~~SWTRPILR~~R~~KGYRQRL~~E~~LS~~D~~IYQI~~P~~SAD~~S~~ADNLS~~E~~KLEREWDR  
 MQRSPLEKAS~~F~~IS~~K~~LF~~F~~SWTTPILR~~R~~KGYRHH~~L~~ELSDIYQAP~~S~~AD~~S~~ADHL~~S~~E~~K~~LEREWDR  
 MQRSPVEDANFLS~~R~~FVFWITPLLR~~R~~KGF~~T~~KKLELTDVYKAP~~S~~FDIAD~~T~~L~~S~~E~~K~~LEREWDR  
 MQRSPVEDANFLS~~K~~Y~~F~~FWIT~~S~~PLLR~~R~~KGF~~R~~KKLELTDVYKAP~~S~~FDIAD~~N~~L~~S~~E~~K~~LEREWDR

# of amino acid changes among species over evolutionary history

001001010212013101002000100112201010102100201002000100000000

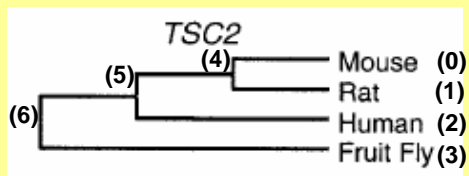
### Site variability

## Mapping amino acid changes on a phylogeny



Main point...the phylogeny of organisms is important, not just the number of different amino acid residues present in the alignment!

## Representing a phylogenetic tree in a 2-dimensional array (Treeconfig)



```

Treeconfig[0,0]:= -1;
Treeconfig[0,1]:= -1;
Treeconfig[1,0]:= -1;
Treeconfig[1,1]:= -1;
Treeconfig[2,0]:= -1;
Treeconfig[2,1]:= -1;
Treeconfig[3,0]:= -1;
Treeconfig[3,1]:= -1;
Treeconfig[4,0]:= 0;
Treeconfig[4,1]:= 1;
Treeconfig[5,0]:= 4;
Treeconfig[5,1]:= 2;
Treeconfig[6,0]:= 5;
Treeconfig[6,1]:= 3;
    
```

## Analysis of the CFTR gene: STATISTICS!

Site Variability	# of sites in gene	Proportion	# of observed mutations at each type of site
0	706	0.4770	286
1	430	0.2905	96
2	217	0.1466	40
3	105	0.0709	11
4	20	0.0135	3
5	2	0.0014	0
Total	1480	1	436

How many disease mutations should we EXPECT to observe at each type of site?

## Analysis of the CFTR gene

Site Variability	# of sites in gene	Proportion	# of observed mutations at each type of site	Expected # of mutations	Calculation for expected value
0	706	0.4770	286	207.98	$0.4770 * 436$
1	430	0.2905	96	126.68	$0.2905 * 436$
2	217	0.1466	40	63.93	$0.1466 * 436$
3	105	0.0709	11	30.93	$0.0709 * 436$
4	20	0.0135	3	5.89	$0.0135 * 436$
5	2	0.0014	0	0.59	$0.0014 * 436$
Total	1480	1	436	436	

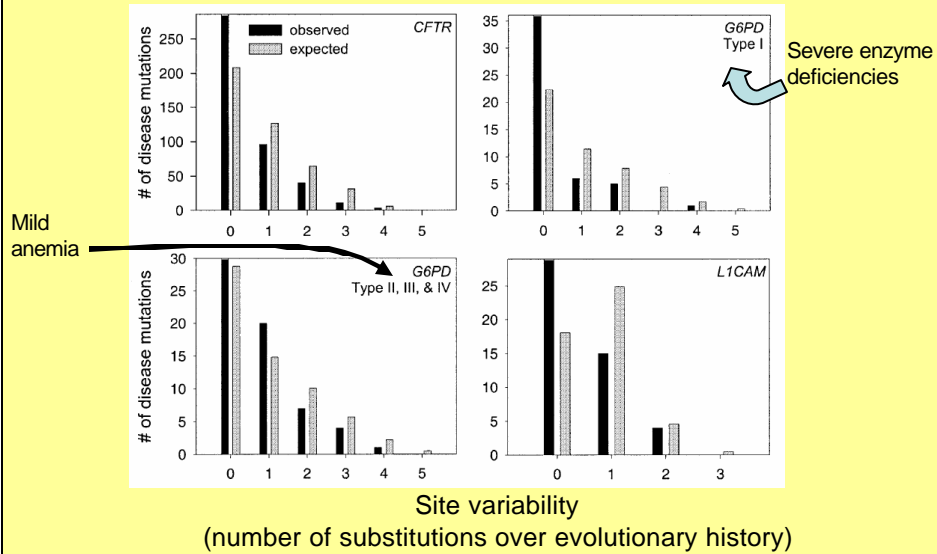
## Analysis of the CFTR gene

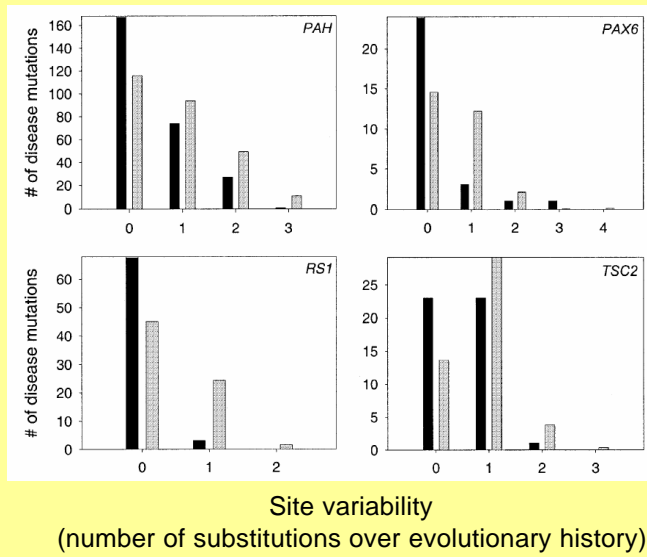
Site Variability	# of sites in gene	Proportion	# of observed mutations at each type of site	Expected # of mutations	Calculation for expected value
0	706	0.4770	286	207.98	0.4770 * 436
1	430	0.2905	96	126.68	0.2905 * 436
2	217	0.1466	40	63.93	0.1466 * 436
3	105	0.0709	11	30.93	0.0709 * 436
4	20	0.0135	3	5.89	0.0135 * 436
5	2	0.0014	0	0.59	0.0014 * 436
Total	1480	1	436	436	

Statistics: Can use a  $\chi^2$  goodness of fit test with N-1 degrees of freedom (5).

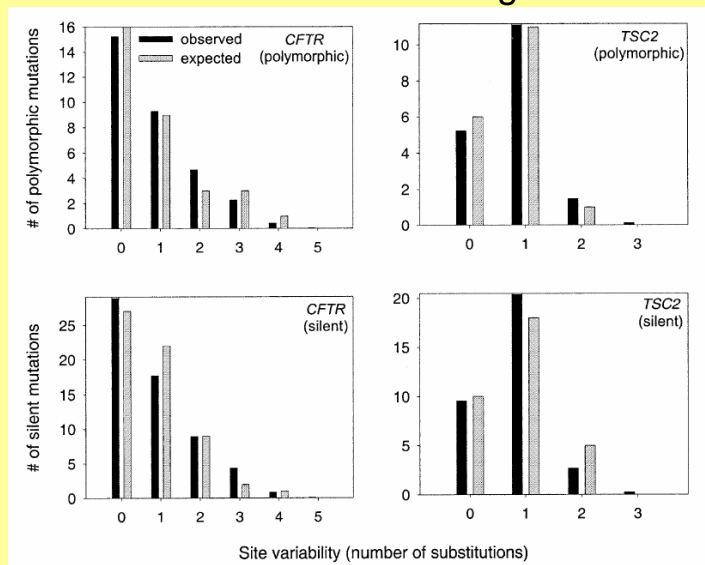
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad \chi^2 = 60.51$$

## How are disease associated point mutations distributed within a gene?





### How are non-disease associated point mutations distributed within a gene?



## Important inference:

- Disease mutations are overabundant at conserved amino acid sites!
- Suggests that evolutionary conserved amino acid residues are *critical* for proper protein function.
- Suggests that these specific amino acid residues have remained constant over evolutionary history due to strong selective pressure.

## What else can be learned through evolutionary analyses and sequence comparisons?

- Examining the frequency with which a given amino acid changes to another amino acid over time
- Are there differences between amino acid changes that have been “accepted” by natural selection and those that produce heritable diseases?

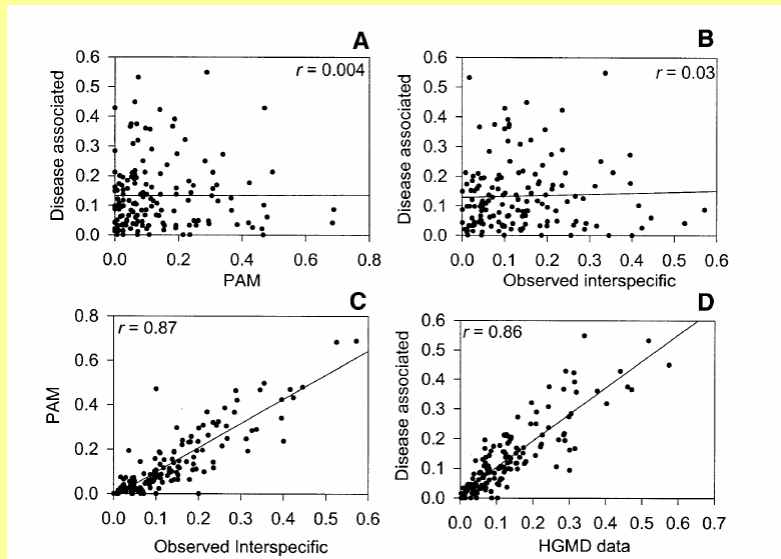
# The PAM matrix (Point Accepted Mutations)

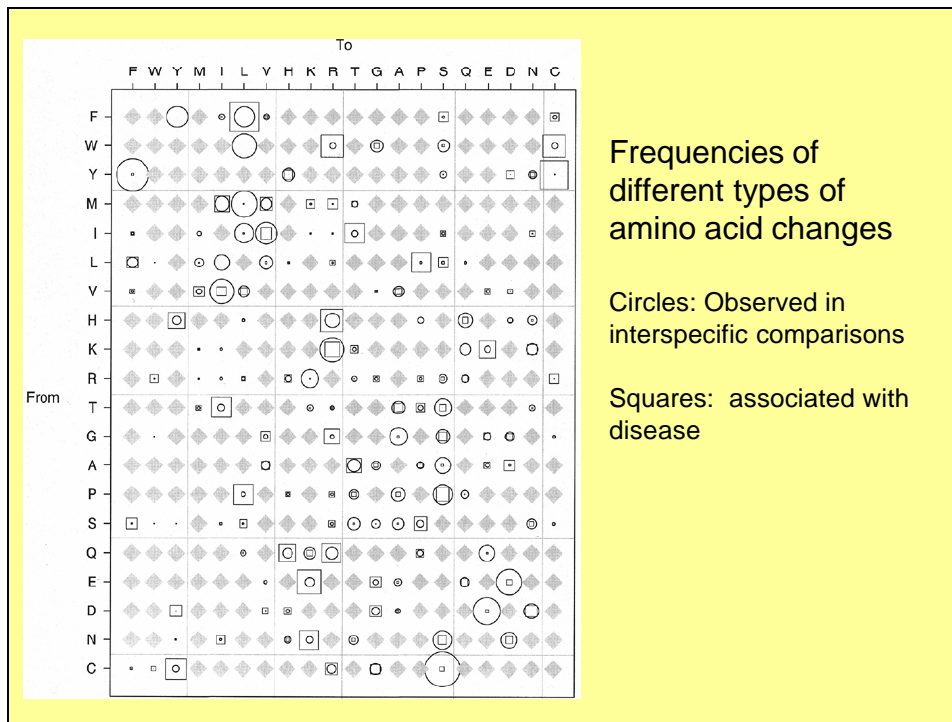
**PAM 250  
matrix**

```

C 12
S 0 2
T -2 1 3
P -3 1 0 6
A -2 1 1 1 2
G -3 1 0 -1 1 5
W -4 1 0 -1 0 0 2
D -5 0 0 -1 0 1 2 4
E -5 0 0 -1 0 0 1 3 4
Q -5 -1 -1 0 0 -1 1 2 2 4
H -3 -1 -1 0 -1 -2 2 1 1 3 6
R -4 0 -1 0 -2 -3 0 -1 -1 1 2 6
K -5 0 0 -1 -1 -2 1 0 0 1 0 3 5
M -5 -2 -1 -2 -1 -3 -2 -3 -2 -1 -2 0 0 6
I -2 -1 0 -2 -1 -3 -2 -2 -2 -2 -2 -2 2 5
L -6 -3 -2 -3 -2 -4 -3 -4 -3 -2 -2 -3 -3 4 2 6
V -2 -1 0 -1 0 -1 -2 -2 -2 -2 -2 -2 2 4 2 4
F -4 -3 -3 -5 -4 -5 -4 -6 -5 -5 -2 -4 -5 0 1 2 -1 9
W 0 -3 -3 -5 -3 -5 -2 -4 -4 -4 0 -4 -4 -2 -1 -1 -2 7 10
Y -8 -2 -5 -6 -6 -7 -4 -7 -7 -5 -3 2 -3 -4 -5 -2 -6 0 0 17
C S T P A G N D E Q H R K M I L V F W Y
    
```

Developed by Dayhoff et al. (1978)





## Important conclusion:

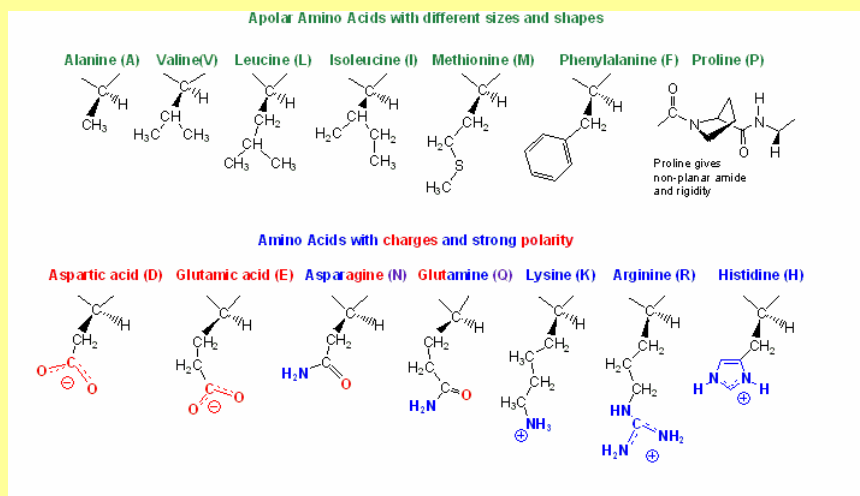
- The types of amino acid changes associated with disease are generally not the same as those types of amino acid changes that are accepted over the long course of evolutionary history

Why???

## What about properties of individual amino acids?

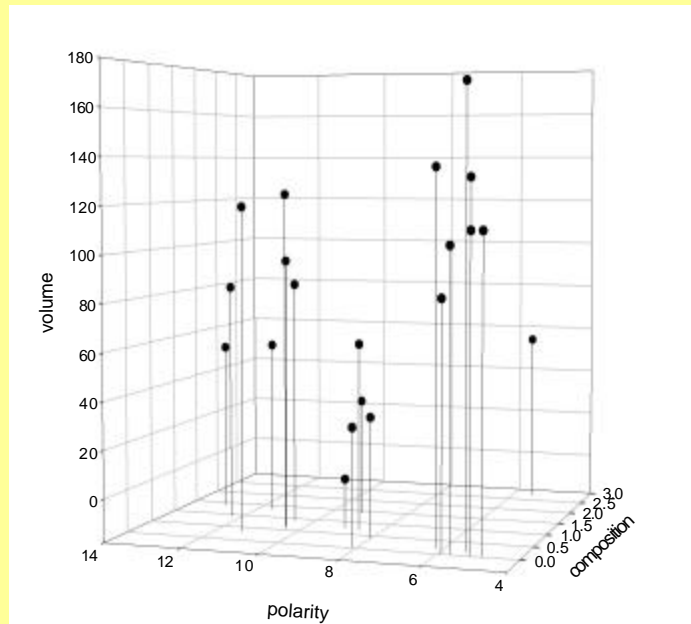
- There are 20 different amino acids
- Each has a different polarity, volume, and side chain composition
- Grantham (1974) developed a “chemical difference” formula to quantify the relative differences between each amino acid

## Examples of variation among amino acids



## Grantham's quantitative assessment of individual amino acid compositions, polarities, and volumes

Amino acid	Property		
	<i>c</i>	<i>p</i>	<i>v</i>
Ser	1.42	9.2	32
Arg	0.65	10.5	124
Leu	0	4.9	111
Pro	0.39	8.0	32.5
Thr	0.71	8.6	61
Ala	0	8.1	31
Val	0	5.9	84
Gly	0.74	9.0	3
Ile	0	5.2	111
Phe	0	5.2	132
Tyr	0.20	6.2	136
Cys	2.75	5.5	55
His	0.58	10.4	96
Gln	0.89	10.5	85
Asn	1.33	11.6	56
Lys	0.33	11.3	119
Asp	1.38	13.0	54
Glu	0.92	12.3	83
Met	0	5.7	105
Trp	0.13	5.4	170
$\bar{D}$	0.739	3.134	50.06



## Grantham's (1974) chemical difference matrix

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asp	Pro	Glu	Arg	Ser	Thr	Val	Try	Tyr
Ala	--	195	126	107	113	60	86	94	106	96	84	111	27	91	112	99	58	64	148	112
Cys	195	--	154	170	205	159	174	198	202	198	196	139	169	154	180	112	149	192	215	194
Asp	126	154	--	45	177	94	81	168	101	172	160	23	108	61	96	65	85	152	181	160
Glu	107	170	45	--	140	98	40	134	56	138	126	42	93	29	54	80	65	121	152	122
Phe	113	205	177	140	--	153	110	21	102	22	28	158	114	116	97	155	103	50	40	22
Gly	60	159	94	98	153	--	98	135	127	138	127	80	42	87	125	56	59	109	184	147
His	86	174	81	40	110	98	--	94	32	99	87	68	77	24	29	89	47	84	115	83
Ile	94	198	168	134	21	135	94	--	102	5	10	149	95	109	97	142	89	29	61	33
Lys	106	202	101	56	102	127	32	102	--	107	95	94	103	53	26	121	78	97	110	85
Leu	96	198	172	138	22	138	99	5	107	--	15	153	98	113	102	145	92	32	61	36
Met	84	196	160	126	28	127	87	10	95	15	--	142	87	101	91	135	81	21	67	36
Asp	111	139	23	42	158	80	68	149	94	153	142	--	91	46	86	46	65	133	174	143
Pro	27	169	108	93	114	42	77	95	103	98	87	91	--	76	103	74	38	68	147	110
Glu	91	154	61	29	116	87	24	109	53	113	101	46	76	--	43	68	42	96	130	99
Arg	112	180	96	54	97	125	29	97	26	102	91	86	103	43	--	110	71	96	101	77
Ser	99	112	65	80	155	56	89	142	121	145	135	46	74	68	110	--	58	124	177	144
Thr	58	149	85	65	103	59	47	89	78	92	81	65	38	42	71	58	--	69	128	92
Val	64	192	152	121	50	109	84	29	97	32	21	133	68	96	96	124	69	--	88	55
Try	148	215	181	152	40	184	115	61	110	61	67	174	147	130	101	177	128	88	--	37
Tyr	112	194	160	122	22	147	83	33	85	36	36	143	110	99	77	144	92	55	37	--

**Table 2.** Average chemical differences of amino acid changes in disease-associated mutations, polymorphic replacement mutations and mutations observed in interspecific comparisons of humans and other metazoan species

Gene	Average chemical difference (SE)		
	Disease <sup>a</sup>	Polymorphic	Interspecific
<i>CFTR</i> <sup>b</sup>	88.53 (2.26)	68.84 (9.96)	55.82 (1.17)
<i>TSC2</i> <sup>b</sup>	91.32 (7.66)	56.22 (8.26)	60.53 (1.30)
<i>G6PD</i> <sup>c</sup>	95.13 <sup>d</sup> (8.54)	–	54.58 (1.89)
	78.23 <sup>e</sup> (6.29)		
<i>LICAM</i>	111.40 (7.49)	–	61.08 (1.45)
<i>PAH</i>	86.33 (3.00)	–	59.01 (2.28)
<i>RSI</i>	109.27 (7.05)	–	51.87 (4.00)
<i>PAX6</i>	84.55 (9.12)	–	56.84 (2.20)

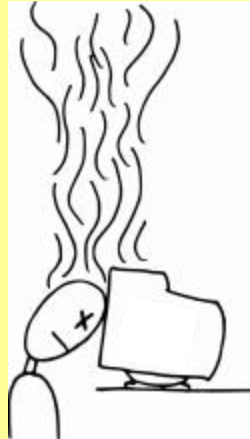
## Another important inference

- Chemical difference scores for amino acid changes associated with disease-causing replacement mutations are higher than those for non-disease replacement mutations
- Chemical difference scores for amino acids associated with disease-causing replacement mutations are greater than those for changes observed among species (i.e., over evolutionary history)
- Natural selection has selected against extreme amino acid changes over the course of evolutionary history

## How were all of these analyses performed?

- No existing software
- No one had ever thought to perform these types of analyses
- All we had was access to some databases:
  - [http://archive.uwcm.ac.uk/uwcm/mg/docs/oth\\_mut.html](http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html)

Solution: spend lots of time  
compiling raw data and writing  
code to implement analyses



A sample data file

Questions?  
Comments?  
Hypotheses?

On to some more recent analyses...what other type of work has been performed?

**Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods**

Ewy Mathe<sup>1,2</sup>, Magali Olivier<sup>1</sup>, Shunsuke Kato<sup>3</sup>, Chikashi Ishioka<sup>3</sup>, Pierre Hainaut<sup>1</sup> and Sean V. Tavtigian<sup>1,4</sup>

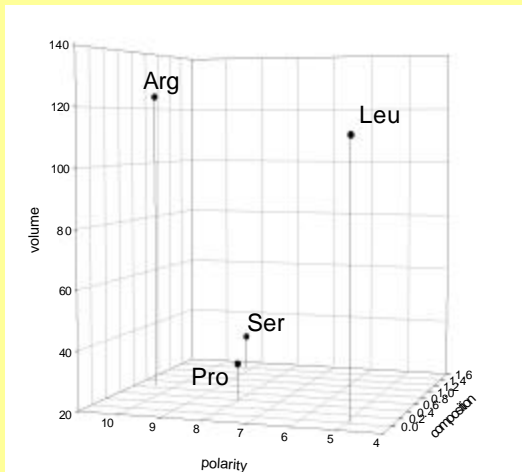
<sup>1</sup>International Agency for Research on Cancer, Lyon, France, <sup>2</sup>Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA, USA and <sup>3</sup>Department of Clinical Oncology, Institute of Development Aging and Cancer, Tohoku University, Sendai 980-8575, Japan

## The p53 tumor suppressor gene

- Mutations in this gene are associated with greater likelihood of different types of cancers
- Mathe et al. created a multiple sequence alignment (MSA: human, monkey, sheep, dog, mouse, rat, chicken, frog, and fish)
- Also assembled a large database of 1514 replacement point mutations from humans

## Some measurements for each amino acid site in the protein

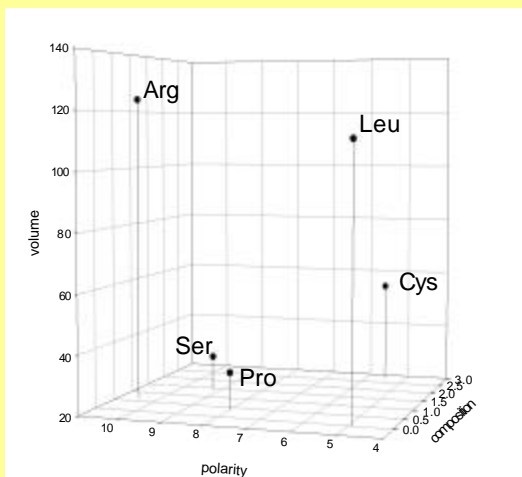
- Grantham Variation (GV)



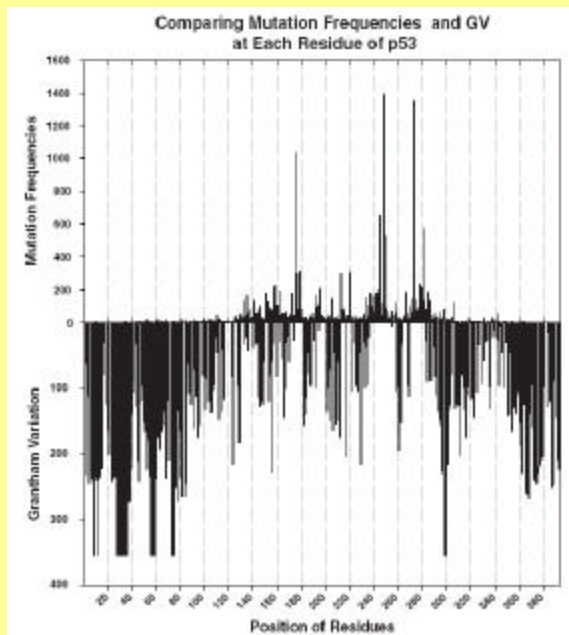
- 1) Plot amino acid residues from MSA in multidimensional space
- 2) Identify smallest "box" that encloses all points
- 3) Calculate GV as the Euclidean length of the longest diagonal of the box
- 4) GV is supposed to quantify biochemical variation at an amino acid site

## A measurement for each disease-associated mutation

- Grantham Deviation (GD)



- 1) Plot mutation on graph
- 2) GD is calculated as Euclidean distance between mutation and the closest point on the box defined by MSA
- 3) GD=0 if mutation lies within box
- 4) GD measures biochemical difference between mutation and observed interspecific variation at site



Some results...

## Some other insights...

- If  $GD=0$ , then mutation is likely neutral
- If  $(GV > 61.3)$  and  $(0 < GD \leq 61.3)$ , then the mutation is likely neutral
- If  $(GV=0)$  and  $(GD > 0)$  then the mutation is likely deleterious
- If  $(0 < GV \leq 61.3)$  and  $(GD > 0)$  then the mutation is likely deleterious

**Table 1.** Distribution of classifications made by Align-GVGD

	Eutherian <sup>a</sup>	Chicken <sup>b</sup>	Frog <sup>c</sup>	Fish <sup>d</sup>
Neutral	765	766	794	800
Deleterious	637	637	608	607
Unclassified	112	111	112	107

<sup>a</sup>MSA used includes all sequences but chicken, frog and fish.

<sup>b</sup>MSA used includes all sequences but chicken and frog.

<sup>c</sup>MSA used includes all sequences but frog.

<sup>d</sup>MSA used includes all sequences.

## Some limitations of study

- Does not take phylogeny into consideration!
- Why use a box to quantify Grantham Variation? Biologically relevant?
- Had no prior knowledge of whether or not a mutation was actually deleterious!