BASIC STATISTICS
and
HUMAN GENETICS

A SWIFT "PERSPECTIVE"

20.400
Perspectives in Biological Engineering
23 & 28 February 2006
W.G.Thilly

20.400
Statistics and Human Genetics
Lecture 1 Basic Statistical Concepts

A. Definitions:
Population (of observations or trials)
Distribution
Mean
Variance
Standard Deviation

B. Operations
Binomial, Poisson and normal distributions and their equivalence.
Variance of a derived variable.
Confidence limits for single and multiple trials. (Bonferroni)

20.400
Statistics and Human Genetics

Lecture 2 Strategy to discover statistical association of common diseases with genes carrying multi-allelic risk.

A. Statistical model and operations
Population of pair-wise trials
Choice of confidence limits
Definition of "power of a test"

B. Scientific model of multi-allelic risk for common diseases (time permitting).

Why master probability and statistics?

1. **Plan experiments with greater chance of success.**
2. **Analyze data without reliance on authors' assertions/conclusions.**

Basically develop your own "bullshit detector".

POPULATION: Set of all possible outcomes of

∞ trials.

Example:

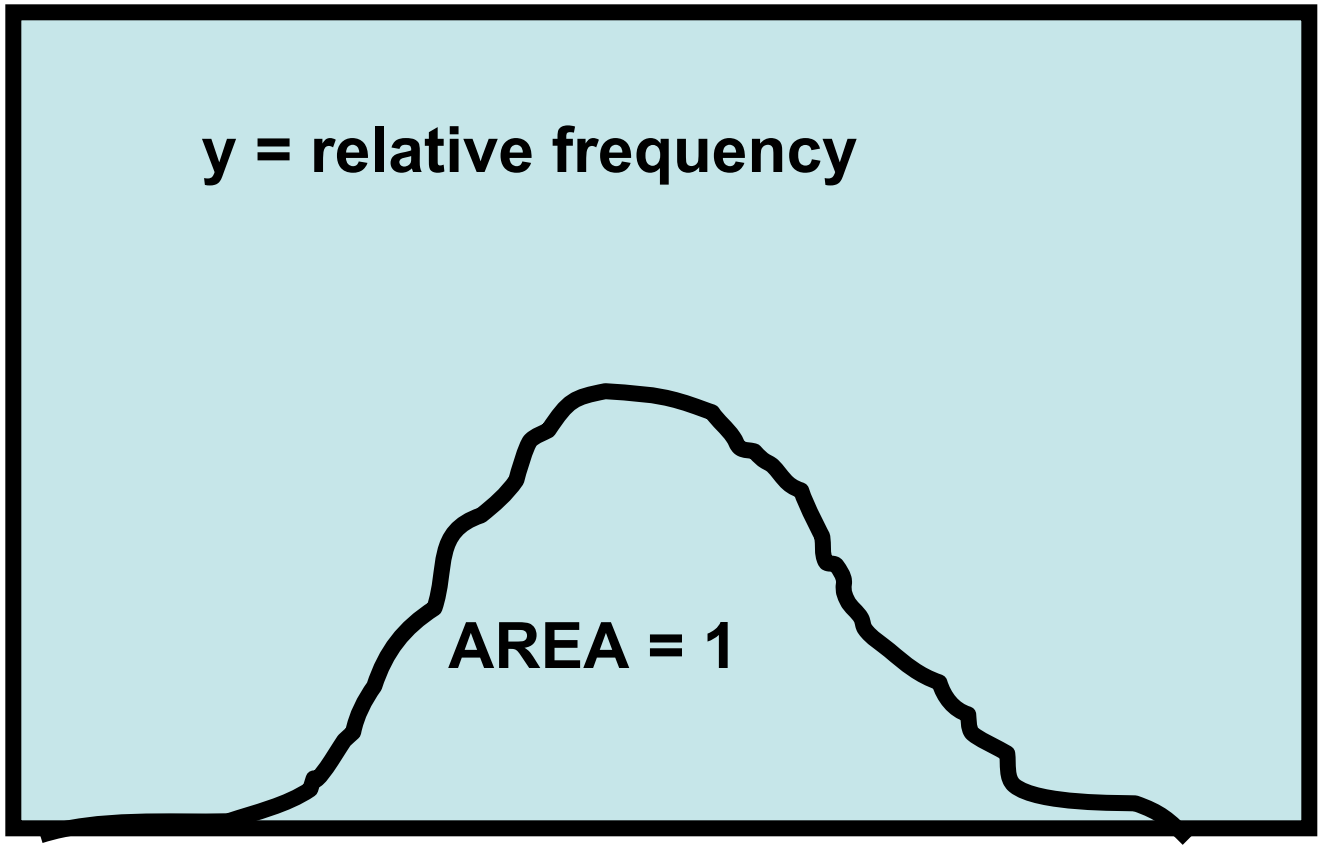
% rock in a 10 million ton coal pile.

D.I.N. Hume, 5.14, 5.194 Analytical Chemistry (1964,66)
G. Wadsworth, 18.10, Applied Statistics (1967)

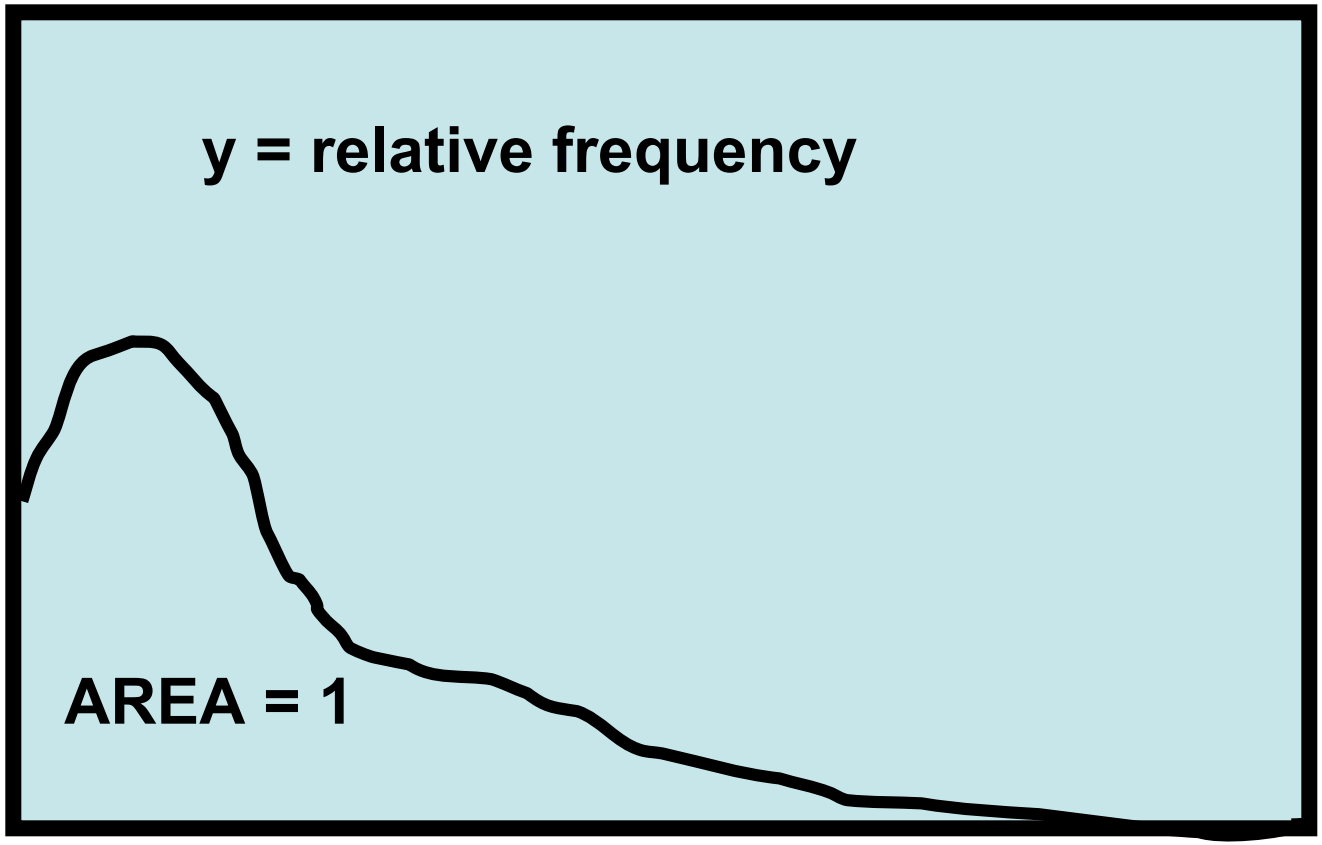
DISTRIBUTION
Set of probabilities of any possible outcome.

z.B.:
Binomial: yes or no
Poisson: any positive integer
Normal: any number

~ Normal Distribution

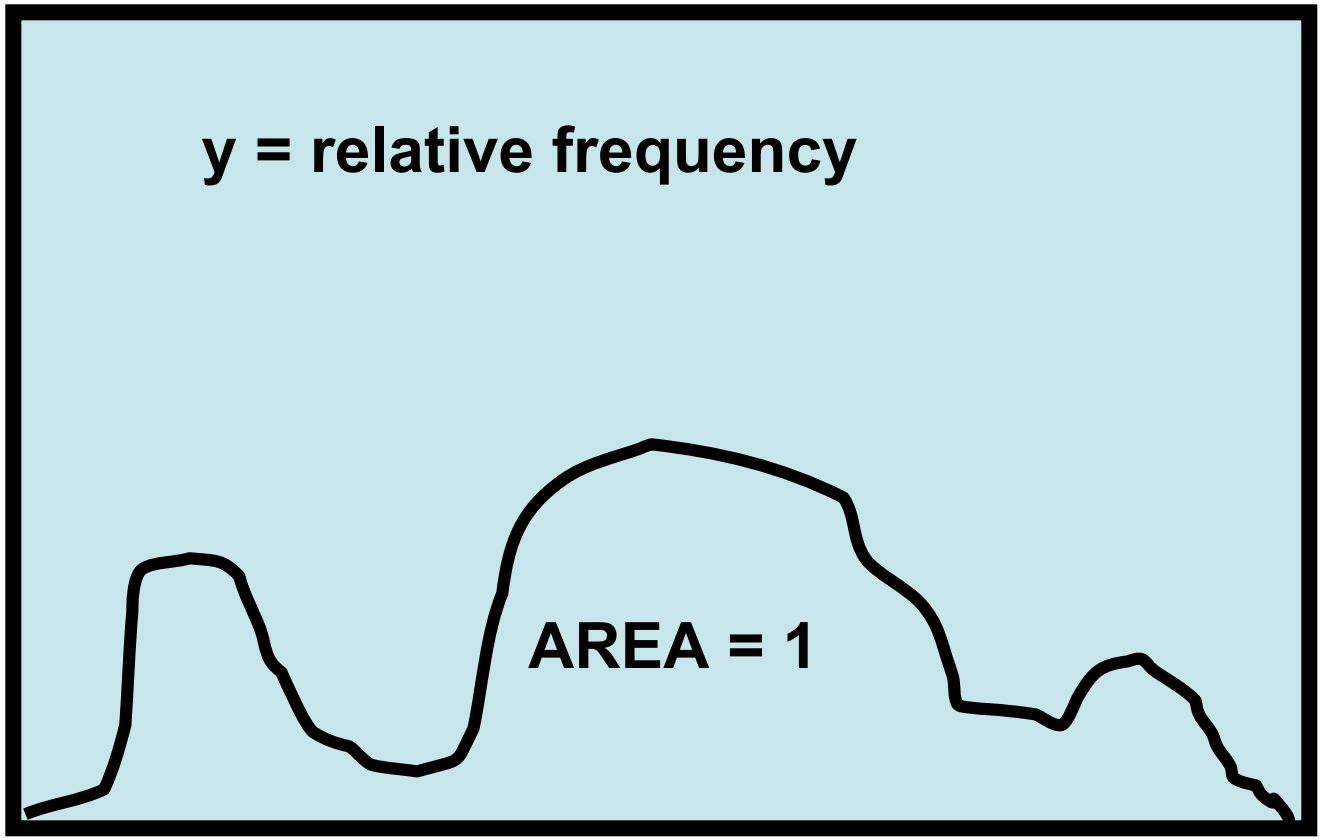


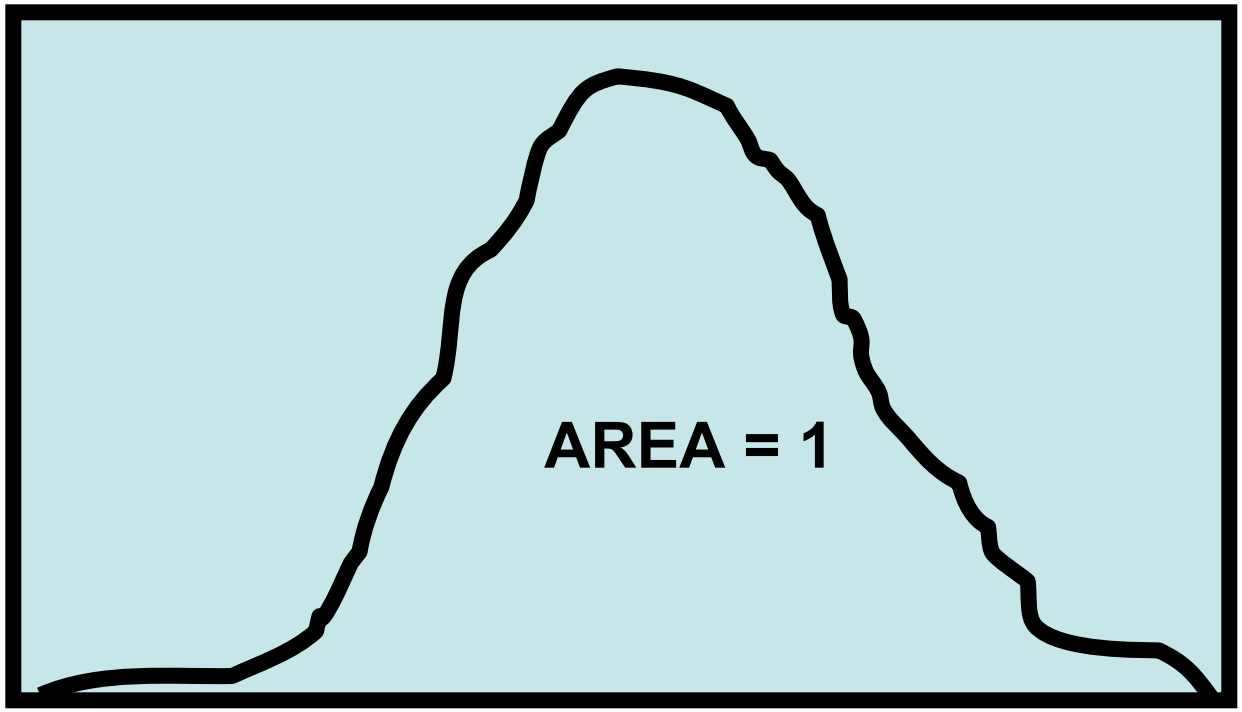
~Poisson Distribution



x = value of observation
x = ln (value of observation)

Typical Ph.D. thesis distribution.





x_i = value of each of n observations

MEAN (x) = $\bar{x} = 1/n \sum(i = 1...n)\ x_i$

VARIANCE(x) = V(x)= $1/n \sum(i = 1...n)\ (x_i - \bar{x})^2$

STANDARD DEVIATION(x) = $V(x)^{1/2}$

Binomial Distribution:

- **n identical independent trials**
- **two possible outcomes**
- **probabilities p and q such that p+q = 1 for any trial**

probability function = $p(x) = n!/(n-x)!x!\ p^x\ q^{(n-x)}$

where n!/(n-x)!x! denotes the number of combinations of n taken x at a time

MEAN = np

VARIANCE = npq

Poisson Distribution

- **n identical independent trials**
- **any positive integer outcomes, 0,1,2,...**
- **MEAN = λ**
- **probability function = $p(x) = \lambda^x e^{-\lambda}/x!$**

VARIANCE =MEAN = λ

$P(x = 0) = e^{-\lambda}$

Very useful characteristics!

Normal Distribution

- **n identical independent trials**
- **any real number outcome**
- **MEAN = μ**
- **VARIANCE = σ^2**
- **probability function =**

$$p(x) = [\exp -(x- \mu)^2/2\sigma^2] / \sigma\sqrt{2\pi}$$

Given μ and σ this is an explicitly integrable function of the form ae^{-bx}.

SEE, FOR INSTANCE,
http://www-stat.stanford.edu/~naras/jsm/FindProbability.html

Calculating the probability that any single observation x_i lies between x_1 and x_2:

$$p(x_1 < x_i < x_2) = \int(x_1, x_2)\, p(x)\, dx$$

Typically one wants to know the probability that a single observation will deviate from the mean by a particular number or *quantile* of standard deviations.

$$p(x > \text{mean}(x) +/- \textit{quantile}\ [V(x)]^{1/2} = \int(x_1, x_2)\, p(x)\, dx$$

where $x_1 = \bar{x} - \textit{quantile}\ [V(x)]^{1/2}$

$x_2 = \bar{x} + \textit{quantile}\ [V(x)]^{1/2}$

One might want to know the range of values of x that have less than some desired probability, z.B. 1%.

A USEFUL FACT
For distributions with large mean values,
z.B. mean >16
the distributions of p(x) are approximately equal
for the binomial (p<<q), Poisson and normal distributions.

STEP 1: DETERMINE IF DISTRIBUTION IS EXPECTED TO BE BINOMIAL, NORMAL OR POISSON.

STEP 2: USE POISSON CHARACTERISTIC
MEAN = VARIANCE = STANDARD DEVIATION^2

STEP 3: USE WEB CALCULATORS FOR AREA UNDER NORMAL DISTRIBUTION FOR *QUANTILES* FOR DESIRED CONFIDENCE LIMITS.

The values of x that define an area under the probability distribution of 0.95 , 0.99, 0.99.9 etc. are called the "confidence limits".

The desired confidence limits may be calculated explicitly from the equation for the normal distribution for any desired area .

IN CLASS DEMO

Calculating *quantiles* for a desired confidence limits.
http://www-stat.stanford.edu/~naras/jsm/FindProbability.html

VARIANCE OF DERIVED VARIABLES

X = X(a,b,c,..) where a, b, c,... are independent variables.

$$V(X) = V(a)\,[\partial X/\partial a]^2 + V(b)\,[\partial X/\partial b]^2 + V(c)\,[\partial X/\partial c]^2 + \ldots$$

Examples:

$$V(a+b) = V(a) + V(b)$$

$$V(a-b) = V(a) + V(b)$$

$$V(ab) = V(a)b^2 + V(b)a^2$$

$$V(a/b) = V(a)/b^2 + V(b)a^2$$

VARIANCE OF DERIVED VARIABLES

Variance of a series of independent processes.

Example:

**Sample n humans from Framingham and
analyze the cholesterol level of each.**

**V(sampling + analysis) =
V(sampling) + V(analysis)**

CONFIDENCE LIMITS FOR MULTIPLE TRIALS

So far we have calculated the probability given p(x) that any single independent trial will have values between x_1 and x_2. By difference, we calculated the probability that any single trial would have a value outside the "confidence limits" x_1 and x_2.

Given the 0.95 confidence limits for p(x) what is the chance that n independent observations all lie within those limits?

n=1? n= 2? n=10? ... n= 10000?..... n=10,000,000?

$$P(n,n) = 0.95^n \quad \text{and} \quad P(0,n) = (1 - 0.95^n)$$

Application of the Bonferroni inequality.....

Application of the "Bonferroni inequality".

Bonferroni discovered that the confidence levels 5%, 1% ... defining the chance none of n independent trials lay outside the confidence interval of the population distribution was simply defined by the single trial intervals intervals defined by 0.05/n, 0.01/n,....

Thus if one wishes to define the interval with a 5% chance that any of n trials lies outside the interval for a known distribution p(x), one simply calculates the
quantile values for 0.05/n.

Example of application of Bonferroni in mRNA or protein array experiments:

Let the set of all pair-wise trials of n macromolecules in samples A and B be $\{A_i, B_i\}$ for i = 1,2,...n and let the distributions be normal.

Let $V(A_i)$ and $V(B_i)$ be defined by repetitive measurements.

$$|A_i - B_i| - 1.96 (V(A_i) + V(B_i)^{1/2} > 0$$

defines the interval in which 95% of all single pair-wise determinations of $|A_i - B_i|$ are expected to fall by chance.

But what about the set of trials of n macromolecules?

Example of application of Bonferroni in mRNA or protein array experiments:

If there are ~25,000macromolecules in the array then using the 95%confidence interval for each comparison of sample A with B would yield:

25,000 x 0.05 = 1250 macromolecules

These 1250 would appear to be "significantly different at the 95% level" BY CHANCE ALONE using a quantile of 1.96.

Such "findings" are called FALSE POSITIVES.

What quantile would required so that there would be less than a 5% chance that any of 25,000 pair-wise comparisons lay outside the confidence interval?

By Bonferronl:

$0.05/25{,}000 = 0.000002 = 2 \times 10^{-6}$

defines the fraction of the normal distribution of a single trial to achieve this degree of certainty.

Using
http://www-stat.stanford.edu/~naras/jsm/FindProbability.html
this works out to *quantile* = 4.76 to achieve an expectation of one FALSE POSITIVE in 25,000 pair-wise trials.

What about FALSE NEGATIVES?

If, for instance, $|A_i - B_i| - 4.76 (V(A_i) + V(B_i)^{1/2} < 0$
but
$|A_i - B_i| - 3.0 (V(A_i) + V(B_i)^{1/2} = 0$ in reality,
then the finding that A_i and B_i
are not significantly different is an example of a FALSE NEGATIVE.

Experimental life is a continuous balancing act between false negative and false positive findings.

Both kinds of false findings can be costly to science and scientists.

LECTURE 1
example from last week's
HUPO Workshop in Dublin.,

Rat brains of inbred strain of ages A and B.

6 major proteomics laboratories prepared one brain of group A and one brain of group B running high resolution 2D gels.

Each laboratory was asked to identify all peptide "spots" that differed between brain A and brain B among thousands of spots.

~360 "spots" identified at least once by 6 labs.
0 found by all 6 labs
1 by 5, 2 by 4,10 by 3, 60 by 2

What was going on here?

20.400
Statistics and Human Genetics

Lecture 2 Strategy to discover statistical association of common diseases with genes carrying multi-allelic risk.

A. Statistical model and operations
Population of pair-wise trials
Choice of confidence limits
Definition of "power of a test"

B. Scientific model of multi-allelic risk for common diseases (time permitting).

A strategy to discover genes that carry multi-allelic and multigenic risk for common diseases: a cohort allelic sums test (CAST).

Stephan Morgenthaler[1] and William G. Thilly[2]
Institute of Mathematics, Ecole polytechnique fédérale de Lausanne, Switzerland[1] and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, U.S.A.[2]

Using high throughput mutational spectrometry (w. Ian Hunter, MIT and Per Ekstrom, Radium Hospital, Oslo) we propose to enumerate the point mutations in case cohorts of ten thousand DNAsamples from each of 100 important common diseases.

The 99 case cohorts of persons without a particular disease would serve as a large control cohort sample.

Given ~ 25,000 genes this strategy involves 25,000 x 100= 2,500,000 pair-wise trials of gene/disease association.

ITEM #2
The observed number and distribution of neutral and disease-causing mutations in human genes.
(Multi-allelic versus mono-allelic risk examples.)

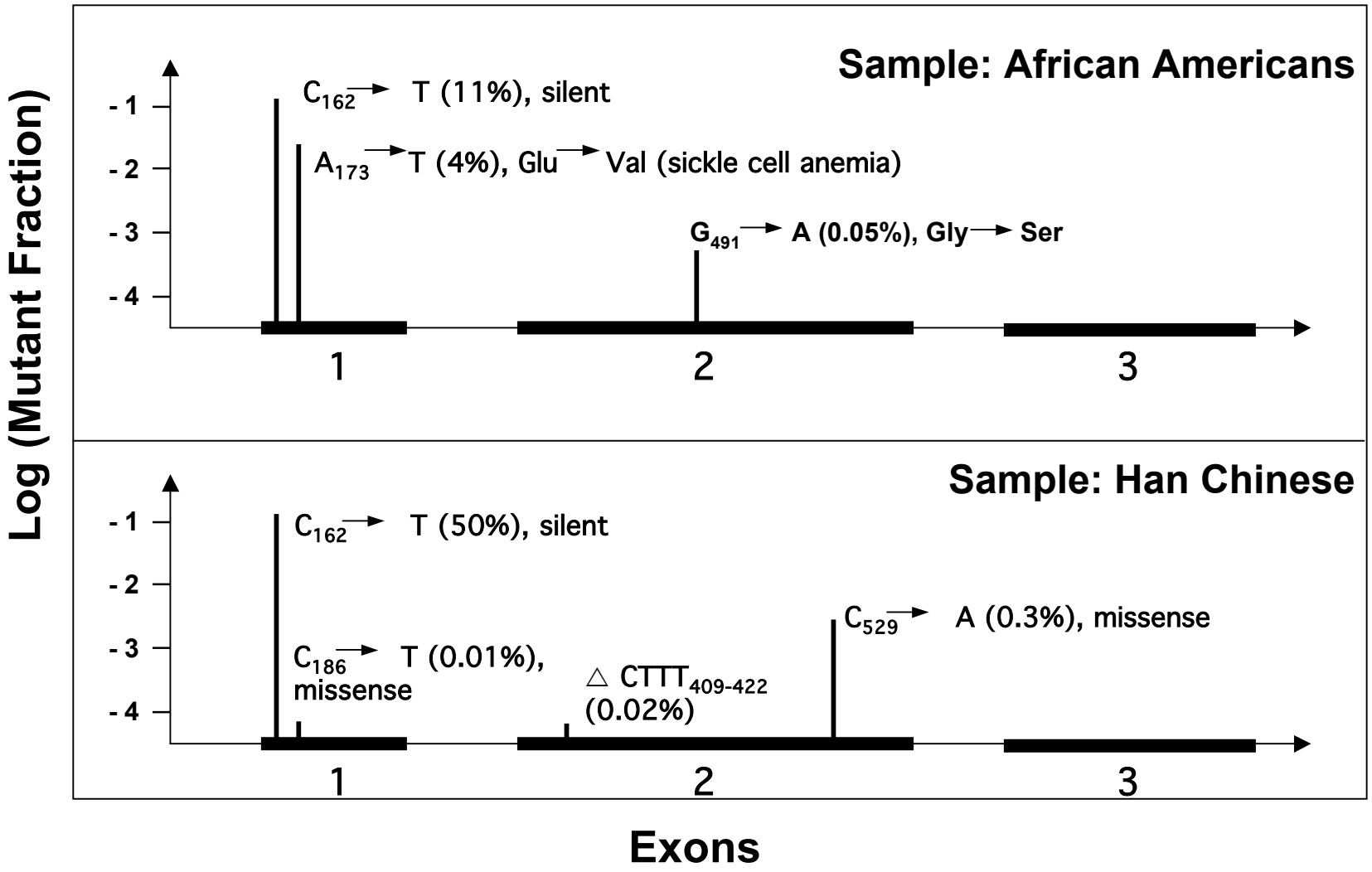
MONOALLELIC (or near monoallelic)DISEASES

1. **sickle cell anemia**
2. **cystic fibrosis in Northern Europeans**
3. **macular degeneration in an American haplotype (subject to validation)**

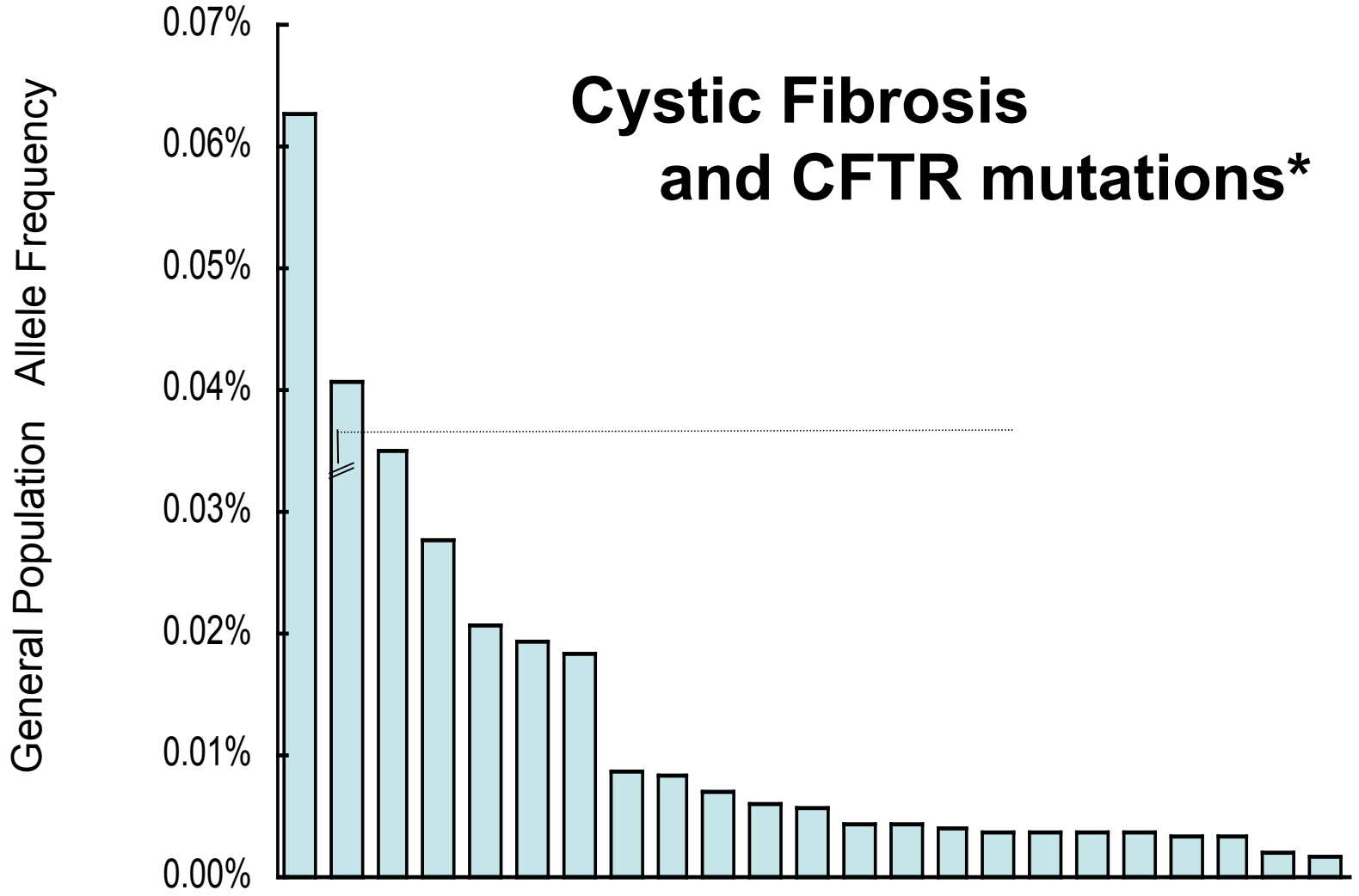
MULTIALLELIC DISEASES

>2000 known gene/disease combinations.
(1750 in HGM Database)

A73T in the betaglobin gene: the original example of a mono-allelic disorder, sickle cell anemia.



del 508 in the CFTR gene: a multi-allelic disorder, cystic fibrosis, confused with mono-allelic disorder.



*Twenty-three most common mutations of unselected CFTR mutant alleles. Delta 508, accounting for 80-95% of point mutations in Caucasian samples, is excluded to demonstrate absolute frequencies of unselected alleles in this recessive deleterious condition

---

As of 15/11/2004, HGMD contains 49335 mutations in 1954 genes and provides 1745 reference cDNA sequences

---

Mutation type	No. of entries
Micro-lesions -	
Missense/nonsense	28309
Splicing	4668
Regulatory	599
Small deletions	8181
Small insertions	3268
Small indels	470
Gross lesions -	
Repeat variations	121
Gross insertions & duplications	437
Complex rearrangements	582
Gross deletions	2700
Total	*49335*

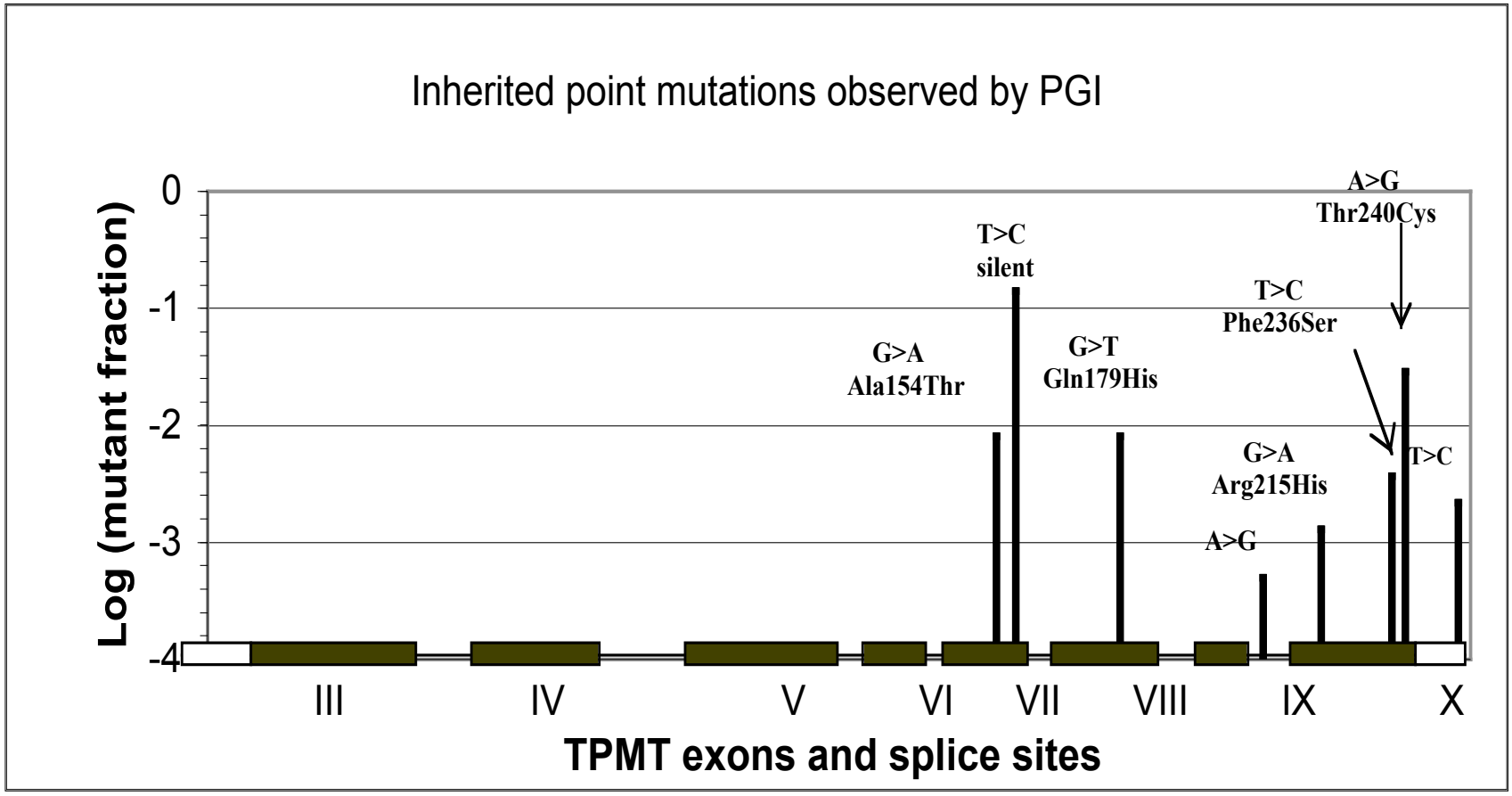
1. **49335 mutations coding for inherited disease in 1954 genes.**

2. **44,896/49335 = 91% of disease-causing mutations are point mutations within the exons and splice sites.**

3. **23 disease-causing point mutations/gene.**

THESE DATA SUPPORT THE GENERAL MODEL OF MULTI-ALLELIC RISK FOR INHERITED DISEASES FOR BOTH DOMINANT AND RECESSIVE DELETERIOUS CONDITIONS OF RISK.

A monogenic, multi-allelic, "non-deleterious" disorder, thiopurine sensitivity caused by thiopurine methyltransferase (TPMT) mutations with q~ 0.06.



dbSNP Build 120, Human Genome Build 34 v3 released Feb 10, 2004 5/20/04

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17*
Chr #	Approx. Size (Mbp)	Approx. # of Genes	Total # of reference SNPs	Total # of Synon + Nonsynon Coding SNPs	Total # of Nonsynon Coding SNPs	Nonsynon Coding SNPs as a % of Total Coding SNPs	Nonsynon Coding SNPs as a % of Total SNPs	Total # of Nonsynon TERM SNPs in Coding Regions	Nonsynon TERM Coding SNPs as a % of Total Coding SNPs	Nonsynon TERM Coding SNPs as a % of Total SNPs	Total # of INDEL SNPs	Total # of Synon + NonSynon INDEL TERM SNPs in Coding Regions	Coding INDEL TERM SNPs as a % of Total SNPs	Total # of Splice Site SNPs	Splice Site SNPs as a % of Total SNPs	Total # of Obligatory Knockout SNPs
1	245.2	2,396	641,605	7,538	4,306	57.12%	0.67%	104	1.38%	0.016%	45,656	40	0.006%	49	0.008%	193
2	243.3	2,377	512,110	4,405	2,456	55.75%	0.48%	76	1.73%	0.015%	49,925	13	0.003%	25	0.005%	114
3	199.4	1,948	443,294	3,366	1,831	54.40%	0.41%	45	1.34%	0.010%	56,304	60	0.014%	33	0.007%	138
4	191.6	1,872	426,913	2,362	1,305	55.25%	0.31%	24	1.02%	0.006%	39,800	16	0.004%	17	0.004%	57
5	181.0	1,768	415,247	3,036	1,617	53.26%	0.39%	43	1.42%	0.010%	37,202	18	0.004%	11	0.003%	72
6	170.7	1,668	505,521	4,151	2,381	57.36%	0.47%	50	1.20%	0.010%	40,717	23	0.005%	38	0.008%	111
7	158.4	1,548	397,625	3,608	2,017	55.90%	0.51%	53	1.47%	0.013%	34,770	28	0.007%	24	0.006%	105
8	145.9	1,426	340,052	2,278	1,256	55.14%	0.37%	30	1.32%	0.009%	29,423	17	0.005%	17	0.005%	64
9	134.5	1,314	378,876	3,137	1,809	57.67%	0.48%	42	1.34%	0.011%	22,236	10	0.003%	30	0.008%	82
10	135.5	1,324	402,871	3,172	1,779	56.08%	0.44%	35	1.10%	0.009%	32,157	14	0.003%	24	0.006%	73
11	135.0	1,319	397,693	4,906	2,787	56.81%	0.70%	50	1.02%	0.013%	26,116	32	0.008%	31	0.008%	113
12	133.5	1,304	368,109	3,671	1,954	53.23%	0.53%	50	1.36%	0.014%	28,182	15	0.004%	39	0.011%	104
13	114.2	1,115	270,435	1,347	731	54.27%	0.27%	18	1.34%	0.007%	35,433	4	0.001%	7	0.003%	29
14	105.3	1,029	201,653	2,023	1,069	52.84%	0.53%	24	1.19%	0.012%	19,691	7	0.003%	13	0.006%	44
15	100.1	978	191,300	2,393	1,370	57.25%	0.72%	39	1.63%	0.020%	18,048	19	0.010%	19	0.010%	77
16	90.0	879	220,997	3,279	1,770	53.98%	0.80%	46	1.40%	0.021%	16,289	16	0.007%	31	0.014%	93
17	81.7	798	191,180	4,015	2,175	54.17%	1.14%	39	0.97%	0.020%	18,340	26	0.014%	31	0.016%	96
18	77.8	760	189,265	1,051	599	56.99%	0.32%	21	2.00%	0.011%	16,781	9	0.005%	9	0.005%	39
19	63.8	623	152,707	5,288	2,855	53.99%	1.87%	54	1.02%	0.035%	11,437	23	0.015%	31	0.020%	108
20	63.6	622	230,682	2,609	1,373	52.63%	0.60%	22	0.84%	0.010%	28,408	37	0.016%	67	0.029%	126
21	47.0	459	110,903	944	519	54.98%	0.47%	20	2.12%	0.018%	9,317	3	0.003%	9	0.008%	32
22	49.5	483	148,211	2,410	1,385	57.47%	0.93%	37	1.54%	0.025%	10,265	19	0.013%	30	0.020%	86
X	152.6	1,491	271,679	2,098	1,183	56.39%	0.44%	31	1.48%	0.011%	17,091	12	0.004%	14	0.005%	57
Y	51.0	498	37,134	269	170	63.20%	0.46%	1	0.37%	0.003%	1,649	2	0.005%	1	0.003%	4
Total	**3,071**	**30,000**	**7,446,062**	**73,356**	**40,697**			**954**			**645,237**	**463**		**600**		**2,017**
		Avg or %	**2,425**	**0.99%**	**0.55%**	**55.7%**	**0.60%**	**0.013%**	**1.3%**	**0.014%**	**8.67%**	**0.006%**	**0.007%**	**0.01%**	**0.009%**	**0.03%**
			SNPs/Mb	**of total SNPs**				**of total SNPs**			**of total SNPs**					**of total SNPs**
																2.75%

***Column 17 = Column 9 + Column 13 + Column 15**

of total Coding SNPs

dbSNP SUMMARY and COMMENTS

7,444,062 separate mutations recorded.
73,356 mutations recorded in exons and splice sites.

40,697 nonsynonymous mutations in exons.

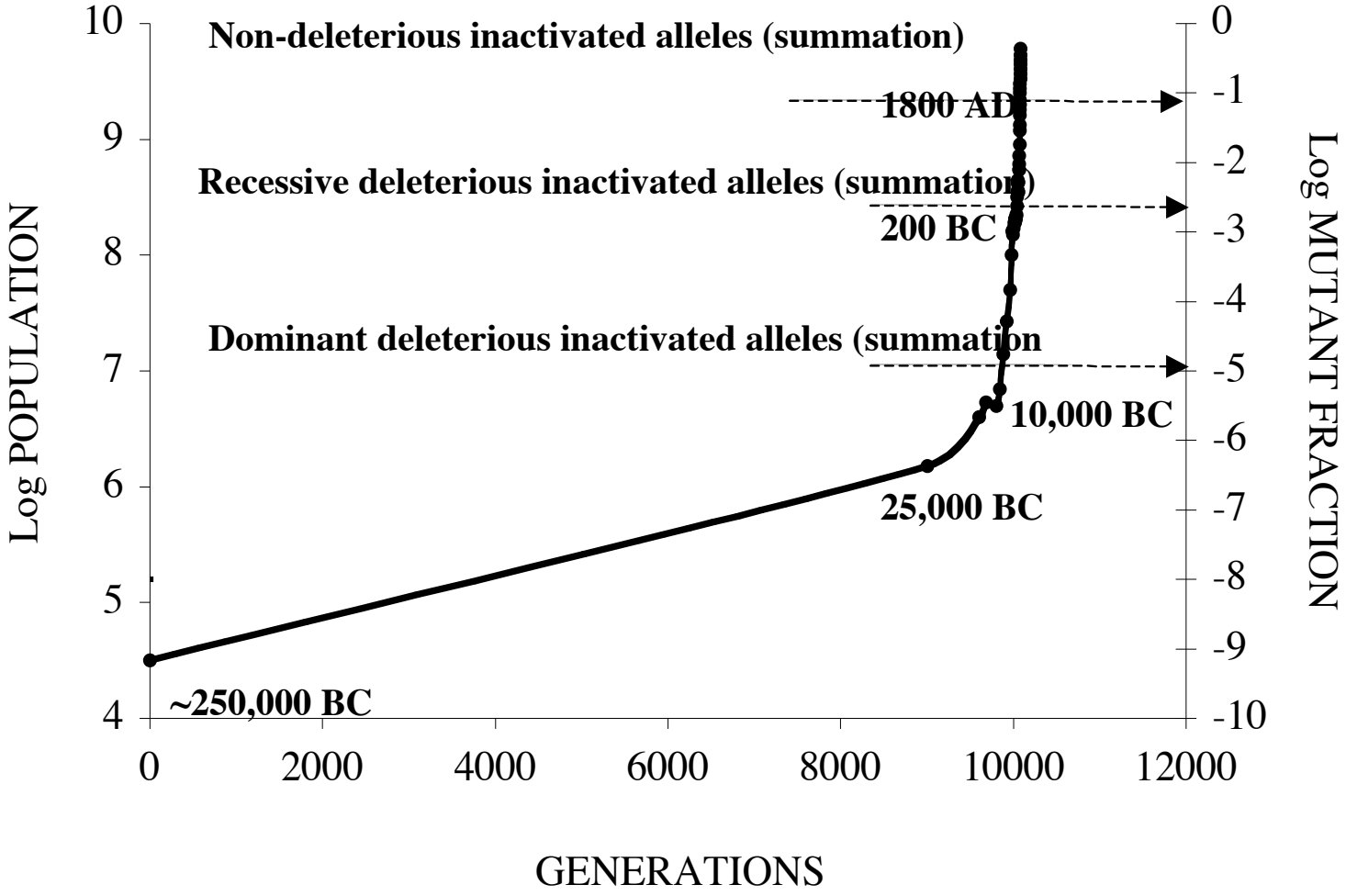
2017 obligatory knockout mutations (OKOs) recorded in exons and splice sites of about 1900 known genes.

OKOs can be used to identify genes carrying non-deleterious inactivating mutations.

We observe that ~1/3 of disease causing mutations are obligatory knockouts.

0.00027 of all recorded mutations
0.027 of all mutations in exons and splice sites.

History of Human Genetic Variation



BASIC STATISTICAL ARGUMENT

Pair wise trial of a single gene and a single disease.
S(controls) = sum of mutant "sequences" for N(controls)
S(cases) = sum for the N(cases)

If the gene is not associated with the disease, we expect
$$|S(cases)/2N(cases) - S(controls)/2N(controls)| = 0 \text{ or}$$
$$|S(cases) - S(controls)N(cases)/N(controls)| = 0.$$

Significant values of this difference would indicate an association of the mutant alleles of the gene with risk of the disease.

Letting S*(controls) = S(controls)N(cases)/N(controls) reduces this expression of the null hypothesis to
$$|S(cases) - S^*(controls)| = 0$$

The interval of a normal distribution spanned by mean ±1.96(*quantile*) variance $^{1/2}$ comprises 95% of the area. The resulting statistical test has a 5% "per comparison error" rate (PCE) of a false positive discovery.

This would be much too liberal in our case in which the differences in total mutant allele numbers are in turn tested for each of some 25,000 separate genes and 100 diseases.

In such a pangenomic scan for a single disease and a quantile of 1.96, some 5% or 1250 genes would be expected to lie outside this interval by chance alone. With a scan through 100 common diseases this would result in the paradoxical outcome of 125,000 genes exceeding this "95%" *quantile* or five diseases associated with each gene by chance.

If we set the rigorous criterion of less than a 5% chance that <u>any</u> difference S(cases)–S*(controls) of the 25,000 x 100 gene/disease comparisons erroneously leads to rejection of the null hypothesis, we can apply the reasoning of Bonferroni and determine the *quantile* values of the normal distribution such that they include all but 0.05/100 x 25,000 = $2x10^{-8}$ of the area under the normal distribution. The corresponding *quantile* value to achieve this degree of certainty is 5.61.

This approach controls the "family-wise error" rate (FWE) of a false discovery within the whole family of comparisons we are making.

Choices between a stringent "family wise error" rate and a loose "per comparison error" rate are possible and important.

Suppose we were to test a total of h =100 x 25,000 = 2.5 million null hypotheses.
Suppose further that the number of rejected hypotheses (significant results) is U.
Of these discoveries some are false positives (F) and some are true positives (T), that is U = T+F.

The ratio F/(T+F) is the proportion of false discoveries or "false discovery rate", FDR.

The "per comparison error rate", PCE, is the average value of F/h, i.e. the proportion of false discoveries among all tests.
The "family wise error "rate, FWE, is the probability of F > 0, i.e. the probability of at least one false discovery among all tests.

It is critical to recognize that "true" in this context means that repetition of the test with independent samples will yield another statistically significant result and that "false" means that such repetitions will not.

Absent biases, most geneticists would be pleased to have an FDR (false discovery rate) of 50% or even higher as the perceived value of a single accurate finding would be many times the scientific/economic value of the cost of retesting and rejecting a false positive result. However, The "great SNP search" has had an FDR >99.9% and 0 or 1 valid gene/disease associations have been made.

"Beware the Jabberwock, my son."

"True" in this statistical sense does not mean that the test has necessarily revealed an actual gene-disease relationship. Each experimental strategy may carry unperceived biases.

In practice, one would like to detect as many genes as possible that carry risk-conferring or risk-deferring alleles and would not want to reject such genes on the basis of an overly rigorous statistical criterion.

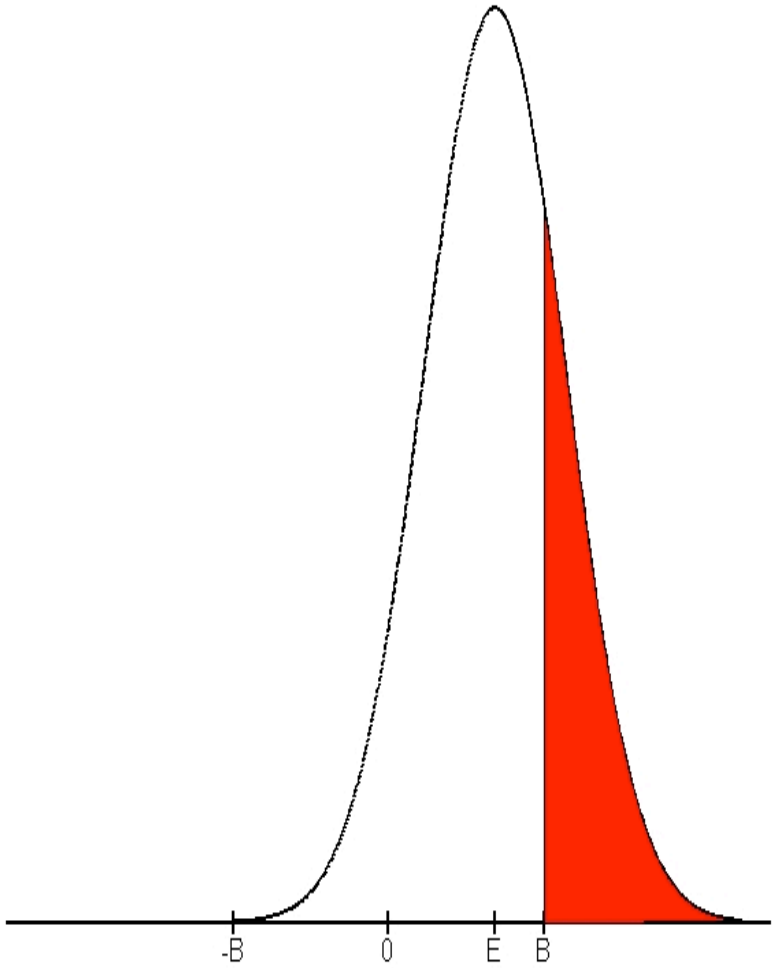
As additional biological tests would have to be performed to confirm or reject any statistical association discovered in a pangenomic scan, one should be willing to countenance a certain number of false positive results in order to capture as many true positives as possible.

If one accepted on average one false positive result per disease or 1 per 25,000 pair-wise trials, then the *quantile* encompassing all but 4×10^{-5} of the normal distribution would be chosen, which is 4.11. This is one possible compromise between stringency and flexibility.

N.B. The decision of which confidence limits to choose is the intellectual responsibility of the researcher.

It depends critically on accurately accounting the costs of both false positives which decrease and false negatives which increase with statistical stringency.

Quantile assignments are dependent on a touching faith in a normally distributed and well-mixed universe of samples seldom encountered in reality. Somewhat more than one false negative per disease must thus be expected applying the value of 4.11 as a _quantile_ value. A pangenomic experiment, of course, would provide 25,000 distributions of mutant frequencies per gene over 100 samples and allow a better informed estimates of the expected dispersions in estimates of $|S(cases)–S^*(controls)|$ and would then be employed in preference to the initial assumption of a normal distribution.

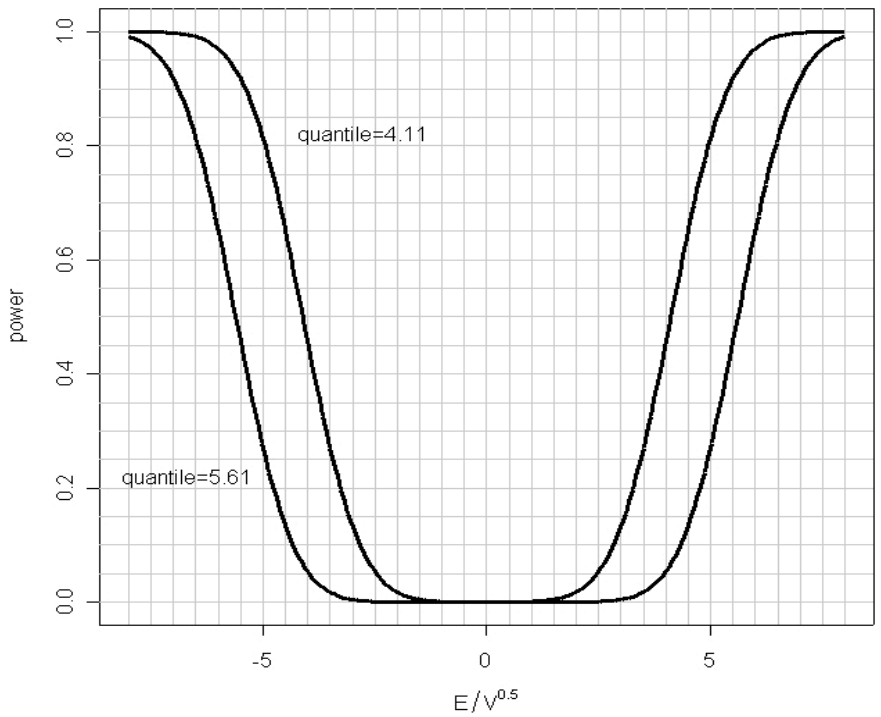


from Morgenthaler & Thilly, 2006

Figure 1. Logical basis for the calculation of the power of the test.

For a specific real
gene-disease relationship
S(cases)-S*(controls) = E
For the null hypothesis
S-S* = 0 and its
confidence limits are +/-B/.

The power of the test given E and B is just the fraction of probability distribution with mean E that lies outside the null hypothesis confidence limits 0 +/- B.



U-shaped curves show the power of the test as a function of the quotient $E/V^{0.5}$. The inner curve has been calculated for *quantile* = 4.11 and the outer for *quantile*= 5.61.

Larger values for *quantile* lead to less powerful tests as they increase the probability that a true positive result will be rejected. Increasing sample sizes increases $E/V^{0.5}$ and therefore increases the power of the test. For simplicity one notes that for results for which S(cases)–S*(controls) – *quantile* $V^{0.5} = 0$ the power of the test is approximately 0.5 and rises rapidly toward 1.0 with increasing values of S(cases)–S*(controls) – *quantile* $V^{0.5}$.

N(controls)	N(cases)	S(cases)– S*(controls)	VARIANCE$^{0.5}$	TEST STATISTIC - 4.11 VARIANCE$^{0.5}$	POWER
1,000,000	10,000	6120	163	5450*	~1
100,000	10,000	6120	168	5428*	~1
50,000	10,000	6120	174	5404*	~1
10,000	10,000	6120	215	5237*	~1
10,000	5000	3060	134	2508*	~1
10,000	2000	1224	77	906*	~1
10,000	1000	612	53	395*	~1
10,000	500	306	37	155*	~1
10,000	250	153	26	47*	~0.95
10,000	125	76	18	2*	~0.55
1000	1000	612	68	334*	~1
500	**500**	**306**	**48**	**109***	**~0.97**
250	250	153	34	14*	~0.63
125	125	76	24	(–22)	~0.15

MONOGENIC RISK FOR COMMON DISEASE

N(controls)	N(cases)	M	S(cases)-S*(controls)	VARIANCE$^{0.5}$	TEST STATISTIC -4.11 VARIANCE$^{0.5}$	POWER
1,000,000	10,000	1	7191	162	6524*	~1
1,000,000	10,000	1/2	3381	150	2764*	~1
1,000,000	10,000	1/3	2108	146	1510*	~1
1,000,000	10,000	1/4	1475	143	886*	~1
1,000,000	10,000	1/5	1094	142	511*	~1
1,000,000	**10,000**	**1/6**	**835**	**141**	**256***	**~0.97**
1,000,000	10,000	1/7	659	140	82*	~0.75
1,000,000	10,000	1/8	523	140	-52	~0.35
1.000,000	10,000	1/9	416	139	-157	~0.17
1,000,000	10,000	1/10	332	139	-240	~0.02
10,000	10,000	1	7191	211	6322*	~1
10,000	10,000	1/2	3381	202	2550*	~1
10,000	10,000	1/3	2108	199	1291*	~1
10,000	**10,000**	**1/4**	**1475**	**197**	**665***	**~1**
10,000	10,000	1/5	1094	196	288*	~0.92
10,000	10,000	1/6	835	196	32*	~0.55

MULTI-GENIC RISK FOR COMMON DISEASE

Random additional notes not used in 20.400 StatGen lectures but of possible interest to students.

Hardy-Weinberg Law: $(p + q)^2 = 1 = \mathbf{p^2 + 2pq + q^2}$

Hardy-Weinberg Equilibrium for Populations

Number of new inactivated gene copies created by mutation/generation
=
Number of inactivated gene copies lost by related disease/generation

Using the Equilibrium Equation permits us to set expectations for the population fraction of inactivated gene copies,**q**, given the rate of mutation/generation, **R**, and specifying the kind of genetic conditions that are deleterious.

ITEM #1

Dominant, recessive and non-deleterious disease-causing mutations. (Expectations based on the average mutation rate per generation in humans.)

For the 22 autosomal chromosomal pairs a person may have either zero, one or two inactivated (or functionally altered) gene copies.

By convention:

p = fraction of active, unaltered(wild-type) gene copies in a population or population sample.

q = fraction of inactivated or functionally altered copies.

By definition: $\mathbf{p + q = 1}$

Hardy-Weinberg Law: $(p + q)^2 = 1 = \mathbf{p^2 + 2pq + q^2}$

Dominant Deleterious Gene Inactivating Mutations
(One inactivated allele/person reduces fecundity)

Hardy-Weinberg Law: $(p + q)^2 = 1 = \mathbf{p^2 + 2pq + q^2}$

Number lost per generation = $(2pq)\ N_{generation}$

Number gained per generation = $2R\ N_{generation}$

$$\mathbf{(2pq)\ N_{generation} = 2R\ N_{generation},\quad p \sim 1}$$

$$\mathbf{q \approx R}$$

DOMINANT DELETERIOUS MUTATIONS

$$q \approx R$$

Estimates of R derived from hundreds of different genes average close to 3×10^{-6} inactivating mutations/ gene copy x generation.

Thus for an average gene in the set of genes that experience dominant deleterious mutations about 6 conceptions per 1,000,000 would be affected.

If there were ~8000 genes of this sort about 5% of all conceptions would eventually be lost due to dominant deleterious mutations.

Recesssive Deleterious Inactivating Mutations
(Two inactivated allele/person reduces fecundity)

Hardy-Weinberg Law: $(p + q)^2 = 1 = \mathbf{p^2 + 2pq + q^2}$

Number lost per generation = $2(q^2)\ N_{generation}$

Number gained per generation = $2R\ N_{generation}$

$$\mathbf{2(q^2)\ N_{generation} = 2R\ N_{generation},}$$

$$\mathbf{q^2 \approx R}$$

RECESSIVE DELETERIOUS MUTATIONS

$$q^2 \approx R \approx 3 \times 10^{-6}$$

Estimate of R is 10^{-5} inactivating mutations/generation.

$$q \approx (3 \times 10^{-6})^{1/2} \approx 1.7 \times 10^{-3}$$

Homozygotes' (q^2) two mutant gene copies are lost to future generations. For an average gene in the set of genes that experience recessive deleterious mutations about 3 conceptions per 1,000,000 would be affected.

If there were ~8,000 genes of this sort about 2.5% of all conceptions would eventually be lost to recessive deleterious mutations.

At population equilibrium, 2pq are heterozygotes:

$$2pq \approx 2 [(1 - 1.7 \times 10^{-3}) (1.7 \times 10^{-3})] \approx 0.0034$$

RECESSIVE DELETERIOUS MUTATIONS

$$q^2 \approx R \approx 3 \times 10^{-6},$$
$$q \approx 1.7 \times 10^{-3}$$

At equilibrium, however, 2pq of the population are unaffected heterozygotes.

$$2pq \approx 0.0033$$

If there were ~8,000 genes of this sort each person would on average be heterozygous for ~ 26 of them.

Non-deleterious Gene Inactivating Mutations
(Inactivated alleles do not reduce fecundity)

Hardy-Weinberg Law: $(p + q)^2 = 1 = \mathbf{p^2 + 2pq + q^2}$

Number lost per generation = 0

Number gained per generation = $2R\ N_{generation}$

$R \sim 3 \times 10^{-6}$ mutations per gene copy per generation.

Number of human generations, g_{now}, from human speciation some 250,000 years ago ~ 10,000.

Expected value of $q = R\ g_{now} \sim 3 \times 10^{-6} \times 10{,}000 \sim 0.03$

$$\mathbf{q_{nondeleterious} \sim 0.03}$$

Non-deleterious Gene Inactivating Mutations

(Inactivated alleles do not reduce fecundity)

$$q_{nondeleterious} \sim 0.03$$

Hardy-Weinberg Law: $(p + q)^2 = 1 = p^2 + 2pq + q^2$

$$p^2 = (1- 0.03)^2 = 0.94$$

$$2pq = 2(1- 0.03)(0.03) = 0.058$$

$$q^2 = (0.03)^2 = 0.0009$$

N.B. There may be ~16,000+ genes of this sort so any population is a complex mixture to say the least. Each person would then carry ~930 conditions of heterozygosity and ~15 conditions of "nullizygosity".

EXPECTATIONS for $R = 3 \times 10^{-6}$ and $g_{now} = 10{,}000$.

	DOMINANT	RECESSIVE	NON-DELETERIOUS
p^2	~1	~0.937	~0.94
$2pq$	$\sim 6 \times 10^{-6}$	~0.0063	~0.058
q^2	~0	~0.00001	~0.0009
gene #	<9000	<9000	>16,000

N.B. These expectations are for the mean of many genes. The distribution around the mean is as yet undefined but appears to be <u>much</u> broader than expected for a simple Poisson or Normal distribution.

FAMILIAL MONOGENIC RISK EXPECTATIONS

GENERAL POPULATION	FIRST DEGREE RELATIVES	COMPUTED for q = 0.1
q^2	$q / q^2 = 1/q$	10.00
$2pq$	$0.5 / 2pq$	2.78
p^2	$p / p^2 = 1/p$	1.11

N.B. The literature for late onset colon cancers was interpreted to suggest that familial risk was about 1.5 to 1.8. In the summer of 2004 in Lausanne we re-examined the underlying premises and found an egregious error. With this corrected, we find familial risk for late onset colon cancer to be about 2.5 to 2.8.

CFTR as an example of a gene carrying recessive deleterious mutations.

(~85% of northern European CFTR mutations causing cystic fibrosis are the delta 508 allele that apparently protects against typhus and typhoid and other g.i. infections. The numbers here are for the set of other CF causing mutations.)

Summation of mutant fractions for 23 most common CFTR mutations causing CF:

~ $3.2 \times 10^{-3} = q_{CFTR}$

Expected sum of inactivating alleles for an average gene carrying recessive inactivating mutations:

~$1.7 \times 10^{-3} = q$

EVIDENCE FOR MULTI-ALLELIC RISK AS A GENERAL MODEL

www.hgmd.org

Welcome to the Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff.

(In association with CELERA)

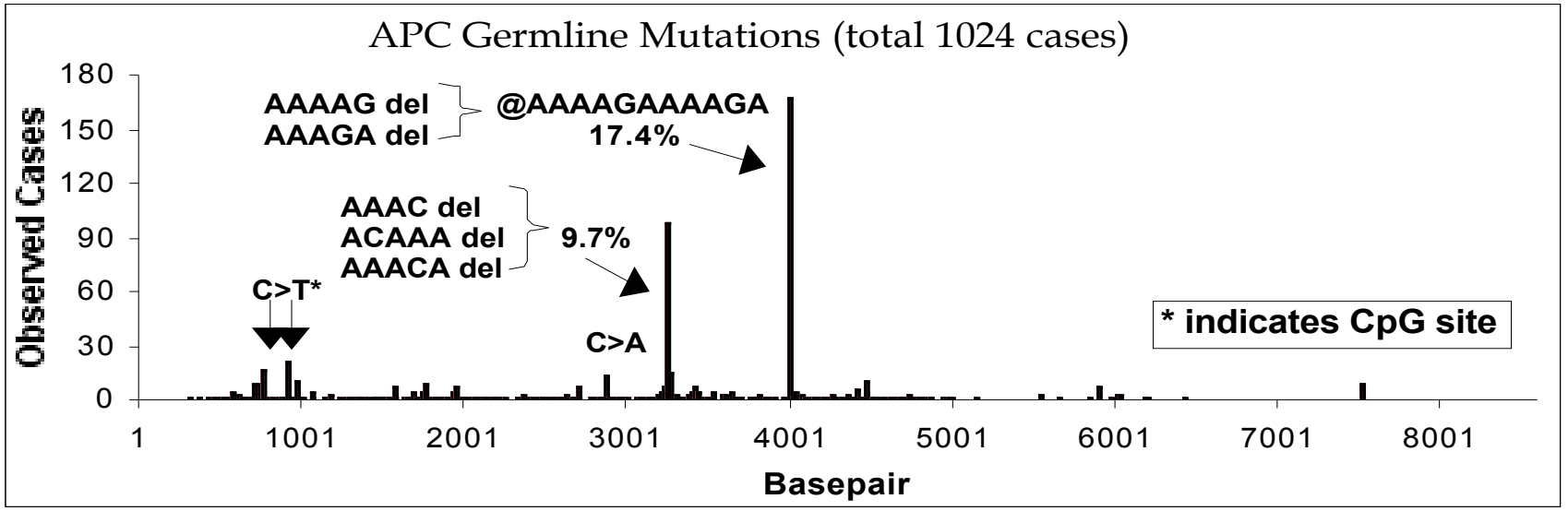
HGMD Search Statistics What's new HGMD Background

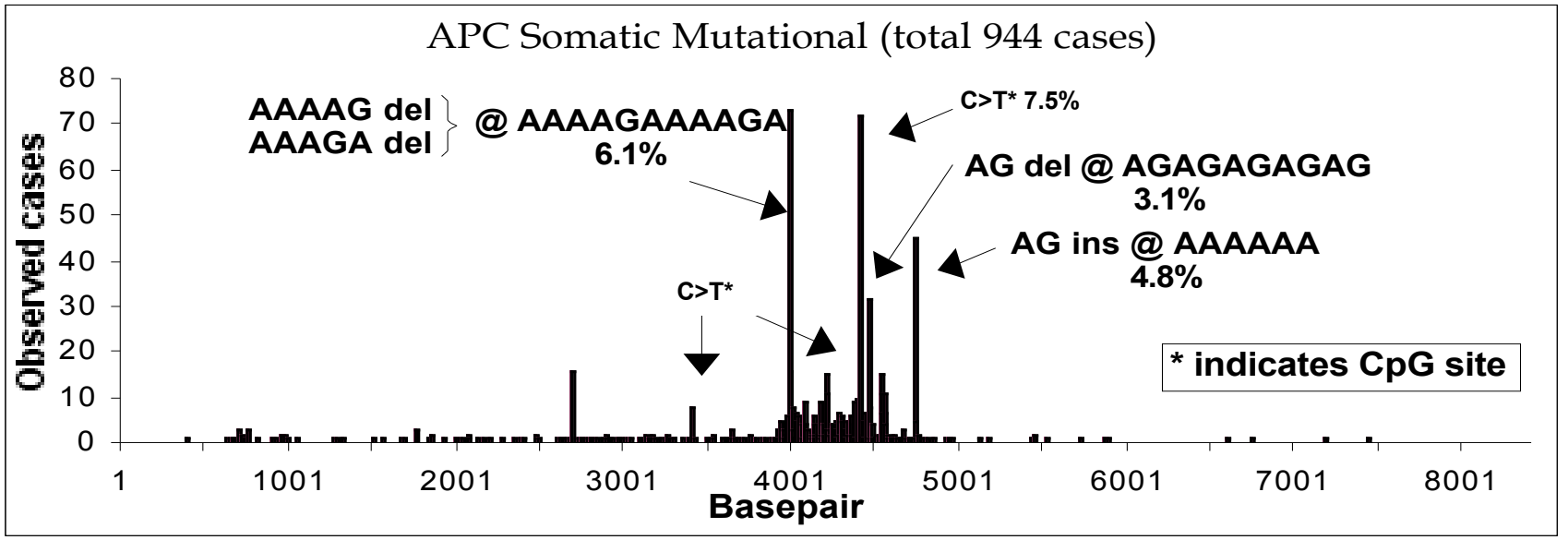
---

* HGMD publications
* Newly added genes
* Locus-specific databases
* Electronic mutation submission
* Comment form

* Supplementary information
* Other useful links
* Meetings & news
* Mutation nomenclature
* Nomenclature publications

A monogenic, multi-allelic, deleterious disorder, FAMILIAL ADENOMATOUS POLYPOSIS COLI





THE NATIONAL SNP DATABASE
(prepared by Brian Glassner w. W.Thilly)

A rough approximation of point mutations found by "random" sequencing of whole genome for ~10 persons and nonrandom reports from genome regions sequenced from an unknown number of individuals.

Contaminated with random reports from diseased cohorts, e.g. Van Hippel Landau syndrome.

We scanned these data to get an estimate of what fraction were in exons and splice sites and, of these, what fraction might represent gene inactivating mutations.

dbSNP SUMMARY and COMMENTS

7,444,062 separate mutations recorded.
73,356 mutations recorded in exons and splice sites.
40,697 nonsynonymous mutations in exons.

32,659 synonymous mutations in exons, each with mutant fractions < 0.5.

N.B. These data permit us to estimate that the number of neutral mutant alleles per gene is, on average, less than 1.0

HUMAN GENETIC HISTORY

DOMINANT DELETERIOUS MUTATIONS:
Mostly arise in parental generations and are extinct in 1-10 generations. Distribution over gene reflects mutational spectrum of gametogenesis. **$q \sim 6 \times 10^{-6}$**

RECESSIVE DELETERIOUS MUTATIONS
Arose in relatively recent genetic history and become extinct in several hundred generations by give and take of equilibrium. Distribution over gene approximates gametic mutational spectrum for large populations absent selection of heterozygotes. **$q \sim 0.0017$**

NON-DELETERIOUS MUTATIONS
Accumulate over all time. A mixture of the most frequently occurring gametic mutations plus rare mutations surviving by chance. **$q \sim 0.03$ with wide variation, up to 0.8 known**.

NON-DELETERIOUS MUTATIONS (comment)

Accumulate over all time. A mixture of
the most frequently occurring gametic
mutations plus rare mutations surviving by chance.

q~ 0.03 with wide variation,.

The mutational spectrum of inactivating mutations for
a gene carrying non-deleterious alleles is expected to arise

~ 50% from a small number of 10-20 highly mutable positions
known as “hotspots” and

~50% from a very large number of positions with average or lower
than average mutation rates.

Thus some “SNPs” arise from multiple and some from single ancestors. Local groups of SNPs form haplotypes that collectively mark a common ancestor.

Item #3
MC1R, the first example of a gene coding for risk for a common disease, skin cancer.

The genetics of sun sensitivity in humans.

Rees JL.
Systems Group, Dermatology, University of Edinburgh, Edinburgh, United Kingdom.

Humans vary >100-fold in their sensitivity to the harmful effects of ultraviolet radiation. The main determinants of sensitivity are melanin pigmentation and less-well-characterized differences in skin inflammation and repair processes. Pigmentation has a high heritability, but susceptibility to cancers of the skin, a key marker of sun sensitivity, is less heritable. Despite a large number of murine coat-color mutations, only one gene in humans, the melano-cortin 1receptor (MC1R), is known to account for substantial variation in skin and hair color and in skin cancer incidence. MC1R encodes a 317-amino acid G-coupled receptor that controls the relative amounts of the two major melanin classes, eumelanin and pheomelanin. Most persons with red hair are homozygous for alleles of the MC1R gene that show varying degrees of diminished function. **More than 65 human MC1R alleles with nonsynonymous changes have been identified,** and current evidence suggests that many of them vary in their physiological activity, such that a graded series of responses can be achieved on the basis of (i) dosage effects (of one or two alleles) and (ii) individual differences in the pharmacological profile in response to ligand. Thus, a single locus, identified within a Mendelian framework, can contribute significantly to human pigmentary variation.

Am J Hum Genet. 2004 Nov;75(5):739-51.

Defining the quantitative contribution of the melanocortin 1 receptor (MC1R) to variation in pigmentary phenotype.

Ha T, Naysmith L, Waterston K, Oh C, Weller R, Rees JL.
Dermatology, University of Edinburgh, Edinburgh EH3 9YW, United Kingdom.

The melanocortin 1 receptor (MC1R) is a key determinant of pigmentary phenotype. Several sequence variants of the MC1R have been described, many of which are associated with red hair and cutaneous sensitivity to ultraviolet radiation even in the absence of red hair. Red hair approximates to an autosomal recessive trait, and most people with red hair are compound heterozygote or homozygous for limited numbers of mutations that show impaired function in in vitro assays. There is a **clear heterozygote effect on sun sensitivity (even in those without red hair) and with susceptibility to the most common forms of skin cancer.**

Ann N Y Acad Sci. 2003 Jun;994:339-47

The melanocortin 1 receptor (MC1R): more than just red hair.

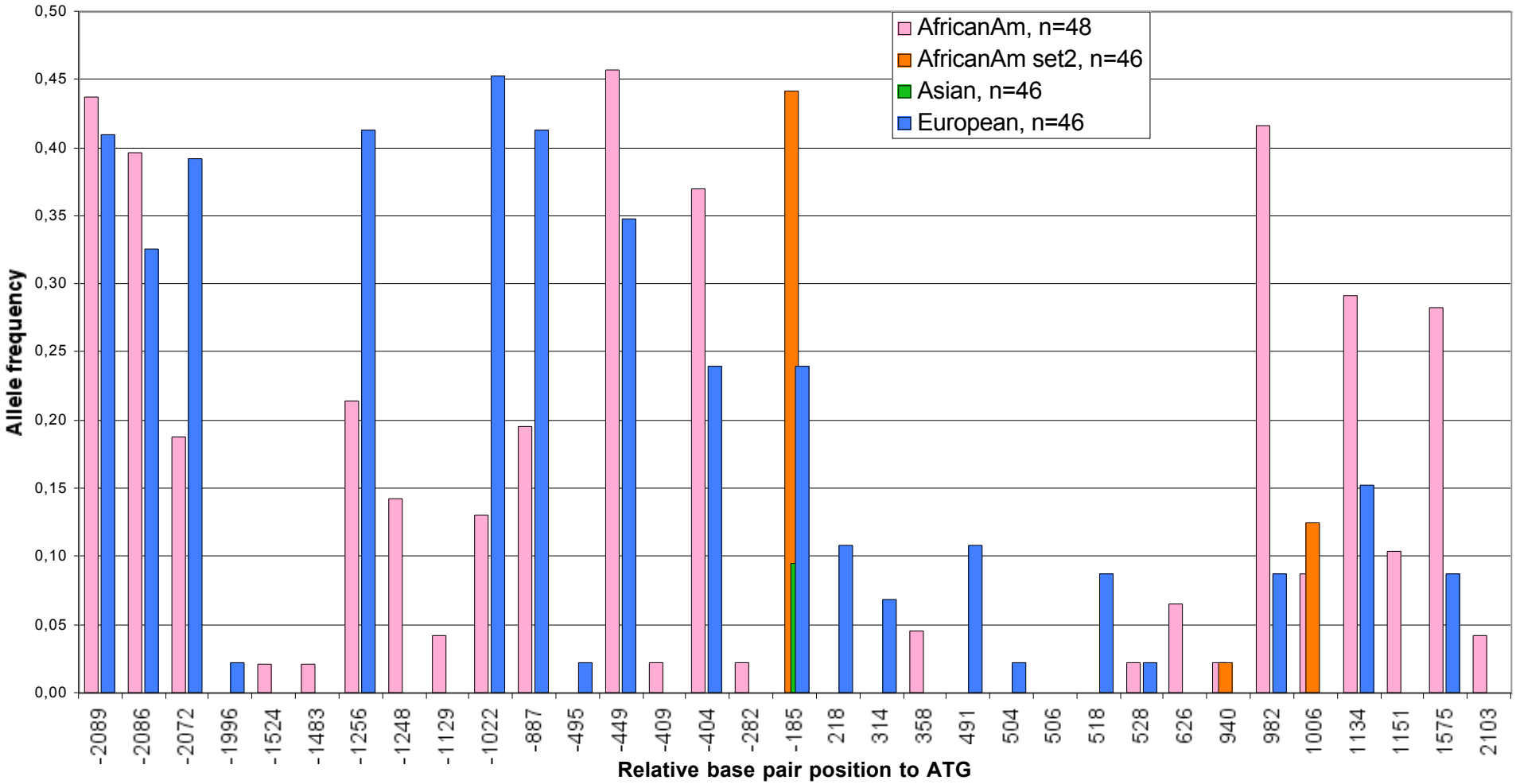
Rees JL.
University of Edinburgh, Royal Infirmary, United Kingdom.

The melanocortin 1 receptor, a seven pass transmembrane G protein coupled receptor, is a key control point in melanogenesis. Loss-of-function mutations at the MC1R are associated with a switch from eumelanin to phaeomelanin production, resulting in a red or yellow coat colour. Activating mutations, in animals at least, lead to enhanced eumelanin synthesis. In man, a number of loss-of-function mutations in the MC1R have been described. The majority of red-heads (red-haired persons) are compound heterozygotes or homozygotes for up to five frequent loss-of-function mutations. A minority of redheads are, however, only heterozygote. The MC1R is, therefore, a major determinant of sun sensitivity and a genetic risk factor for melanoma and non-melanoma skin cancer. **Recent work suggests that the MC1R also shows a clear heterozygote effect on skin type, with up to 30% of the population harbouring loss-of-function mutations.** Activating mutations of the MC1R in man have not been described. The MC1R is particularly informative and a tractable gene for studies of human evolution and migration. In particular, study of the MC1R may provide insights into the lightening of skin colour observed in most European populations. The world wide pattern of MC1R diversity is compatible with functional constraint operating in Africa, whereas **the greater allelic diversity seen in non-African populations is consistent with neutral predictions rather than selection.** Whether this conclusion is as a result of weakness in the statistical testing procedures applied, or whether it will be seen in other pigment genes will be of great interest for studies of human skin colour evolution.
Pigment Cell Res. 2000 Jun;13(3):135-40.

Melanocortin 1 receptor variants in an Irish population.

Smith R, Healy E, Siddiqui S, Flanagan N, Steijlen PM, Rosdahl I, Jacques JP, Rogers S, Turner R, Jackson IJ, Birch-Machin MA, Rees JL.
Department of Dermatology, University of Newcastle upon Tyne, UK.

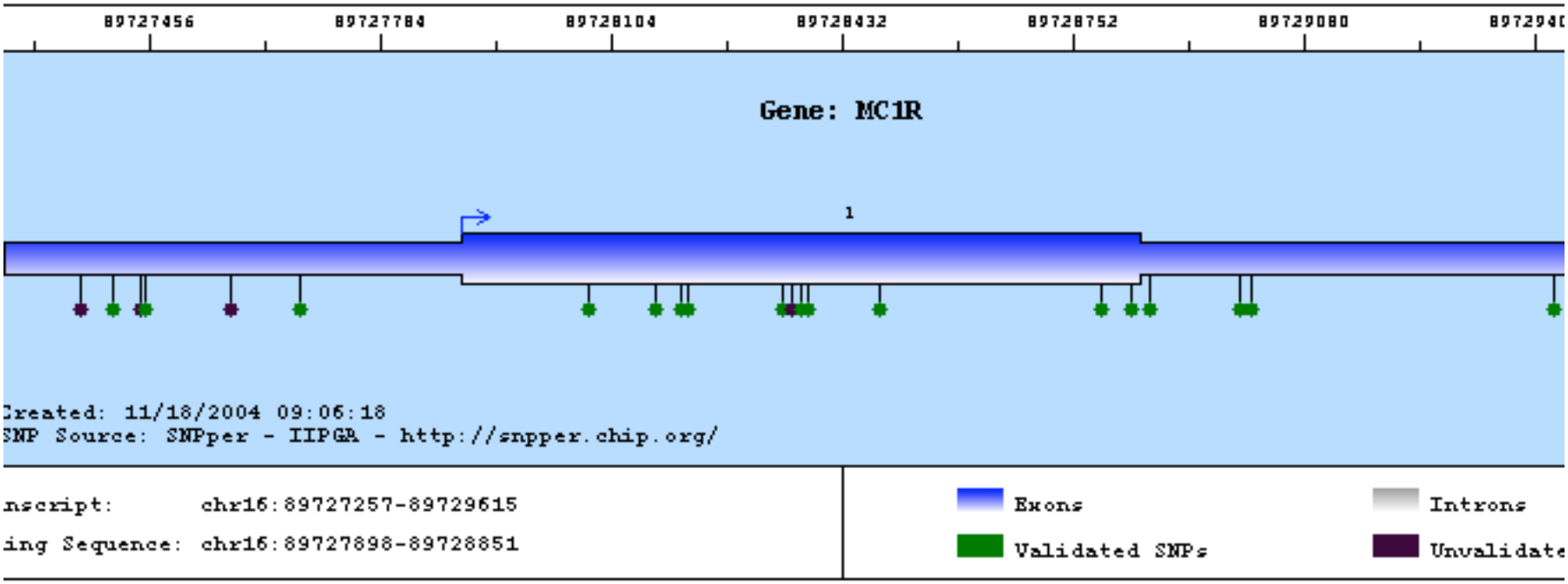
The identification of an association between variants in the human melanocortin 1 receptor (MC1R) gene and red hair and fair skin, as well as the relation between variants of this gene and coat color in animals, suggests that the MC1R is an integral control point in the normal pigmentation phenotype. In order to further define the contribution of MC1R variants to pigmentation in a normal population, we have looked for alterations in this gene in series of individuals from a general Irish population, in whom there is a preponderance of individuals with fair skin type. **Seventy-five per cent contained a variant in the MC1R gene, with 30% containing two variants.** The Arg151Cys, Arg160Trp, and Asp294His variants were significantly associated with red hair ($p = 0.0015$, $p < 0.001$, and $p < 0.005$, respectively). Importantly, no individuals harboring two of these three variants did not have red hair, although some red-haired individuals only showed one alteration. The same three variants were also o ver-represented in individuals with light skin type as assessed using a modified Fitzpatrick scale. Despite these associations many subjects with dark hair/darker skin type harbored MC1R variants, but there was no evidence of any particular association of variants with the darker phenotype. The Asp294His variant was similarly associated with red hair in a Dutch population, but was infrequent in red-headed subjects from Sweden. The Asp294His variant was also significantly associated with nonmelanoma skin cancer in a U.K. population. The results show that the Arg151Cys, Arg160Trp, and Asp294His variants are of key significance in determining the pigmentary phenotype and response to ultraviolet radiation, and suggest that in many cases the red-haired component and in some cases fair skin type are inherited as a Mendelian recessive. J Invest Dermatol. 1998 Jul;111(1):119-22.



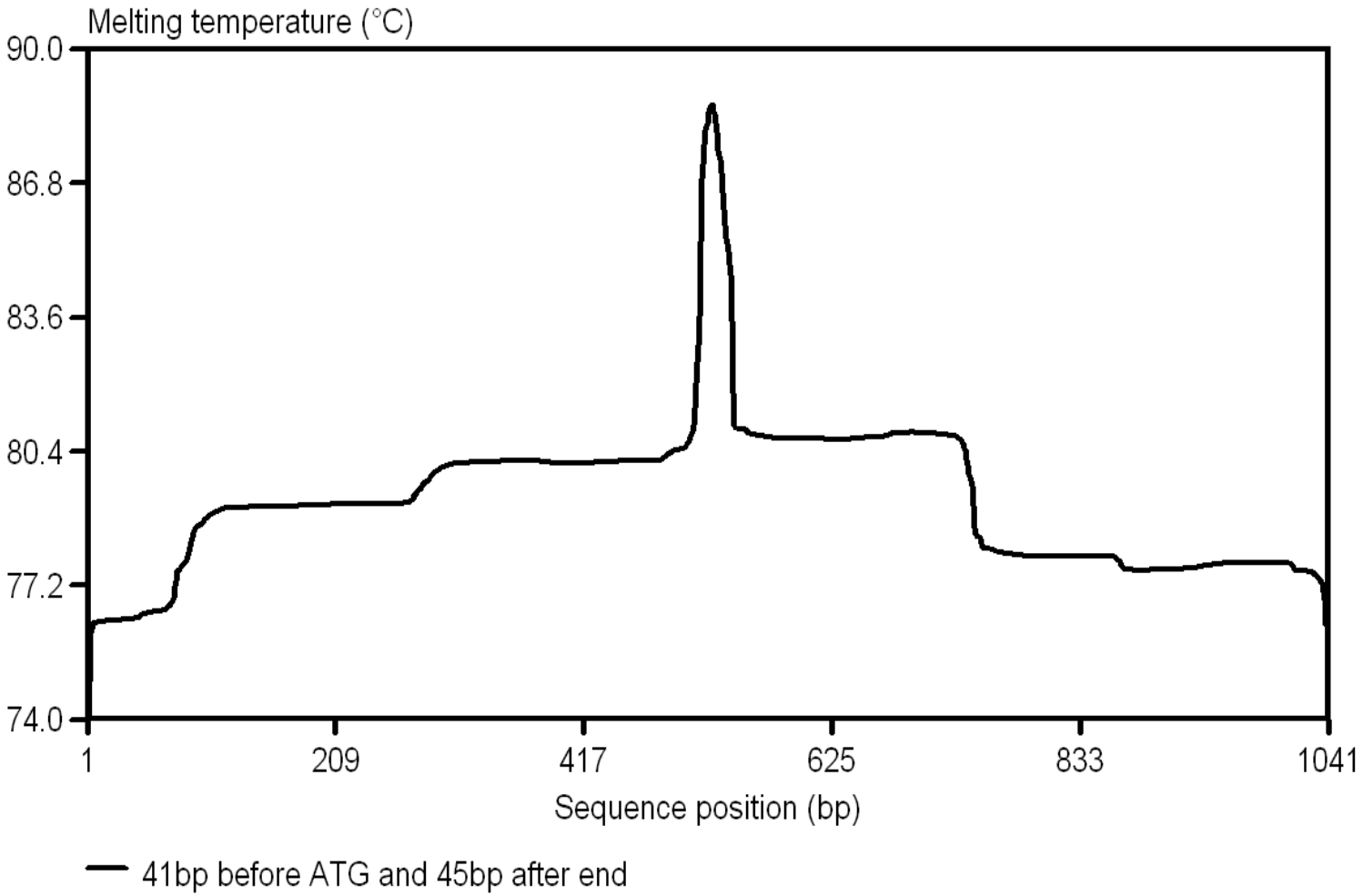
Highest frequency alleles compiled by Per Olaf Ekstrom from NCBI SNP database.
No data yet at HGMD for known (Rees et al.) low frequency alleles.

MC1R (melanocortin 1 receptor)

Chr16 (89,727,257-89,729,615) (exon #1 = 2358 base pair, coding region = 954 base pair)
OMIM info about MC1R
http://www.ncbi.nih.gov/entrez/dispomim.cgi?id=155555



MC1R MELTING MAP (Per Olaf Ekstrom)



MC1R appears to be the first real example of a single gene carrying multiple non-deleterious alleles that govern risk for a common disease.

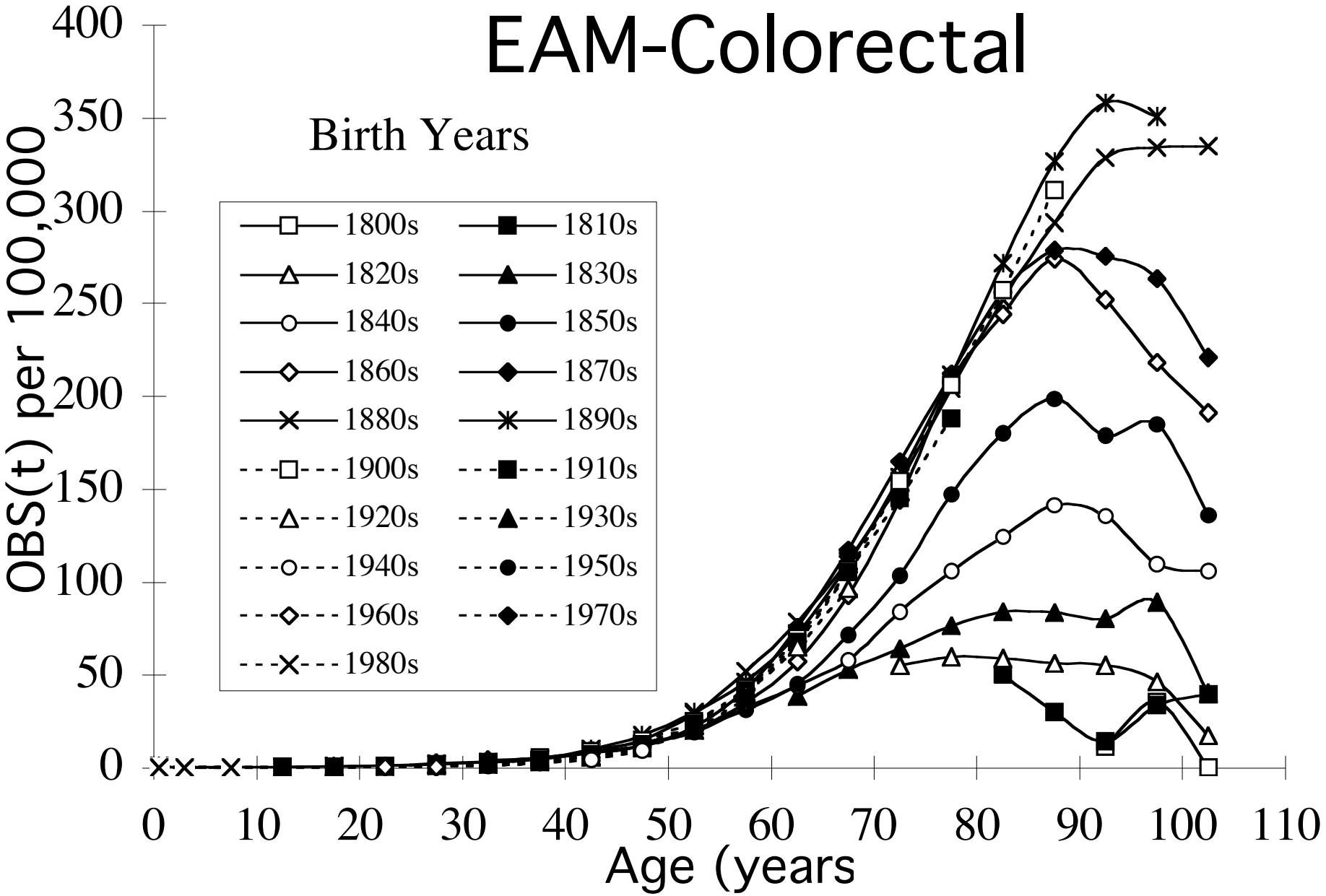
Risk for melanoma and the much more common basal and squamous cell carcinomas of the skin appear to be encoded in Europeans' and Asians' MC1R gene copies.

The MC1R mutant copies are generally not inactivating but of altered function in the physiologic response to the most ubiquitous environmental carcinogen, sunlight.

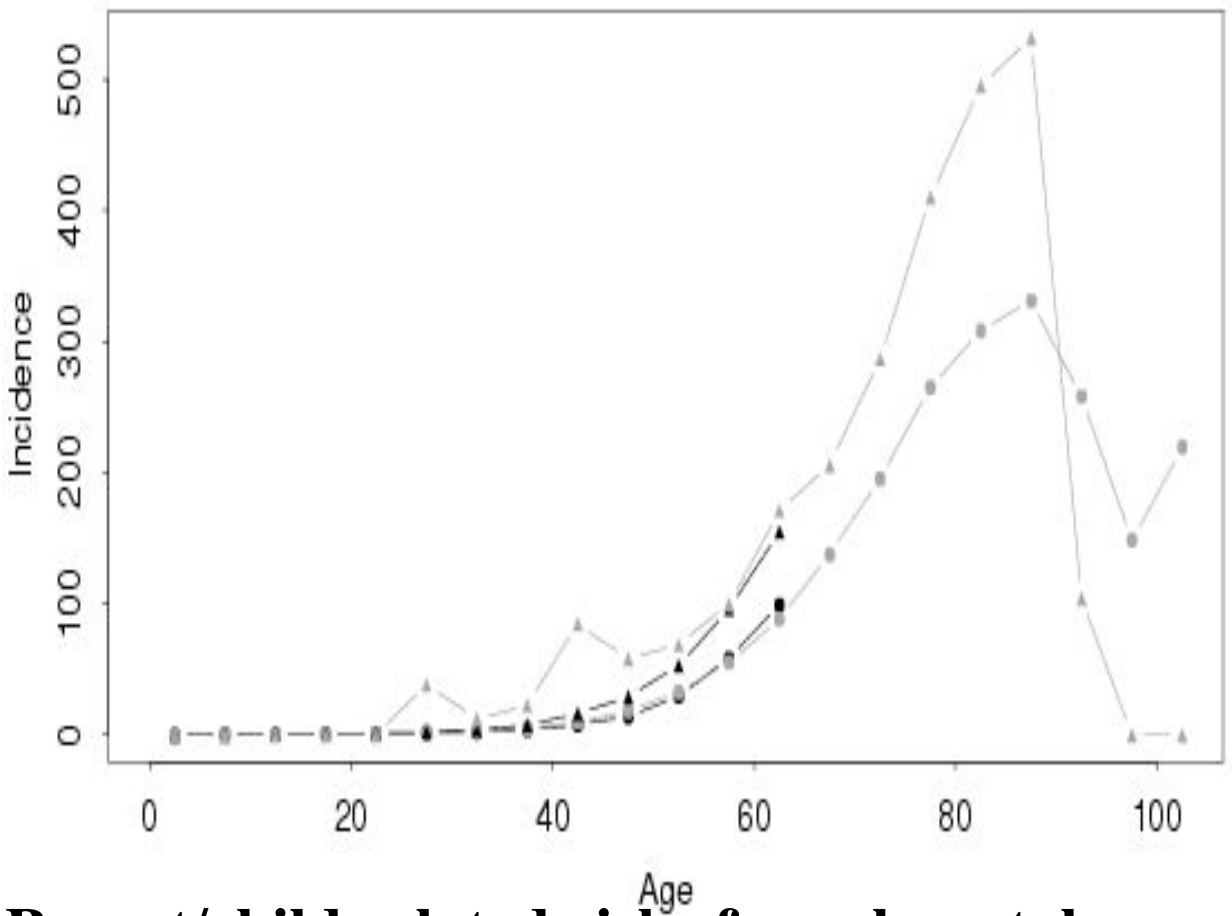
Insofar as unknown environmental changes drove up the risk of many common cancer types in the 19th century, could the underlying genetic risk be physiologically analogous to MC1R and sunlight?

Item #4

Colon cancer, the evidence that lifetime risk may be encoded by a single unknown gene.



ESTIMATED FRACTION AT RISK: 0.18-0.20 (EAM, 1890s)



Parent/child related risks for colorectal cancer in Sweden. Uncorrected original data (K.Hemminki) suggested late-onset familial risk of 1.5-1.6.

FAMILIAL MONOGENIC RISK EXPECTATIONS

GENERAL POPULATION	FIRST DEGREE RELATIVES	COMPUTED for q = 0.1
q^2	$q / q^2 = 1/q$	10.00
$2pq$	$0.5 / 2pq$	2.78
p^2	$p / p^2 = 1/p$	1.11

N.B..
Re-examining the underlying premises and found an egregious error. With this corrected, we find familial risk for late onset colon cancer to be about 2.5 to 2.8. (Morgenthaler, Hemminki, Thilly)

COLORECTAL CANCER FACTS

Integration under curve of incidence rate up to age 125 indicates at least 0.2 of American males are at risk. This agrees with St.Lukes Hospital (London) studies over 30 years finding ~ 20% of males show at least one colorectal polyp. (W.Atkin)

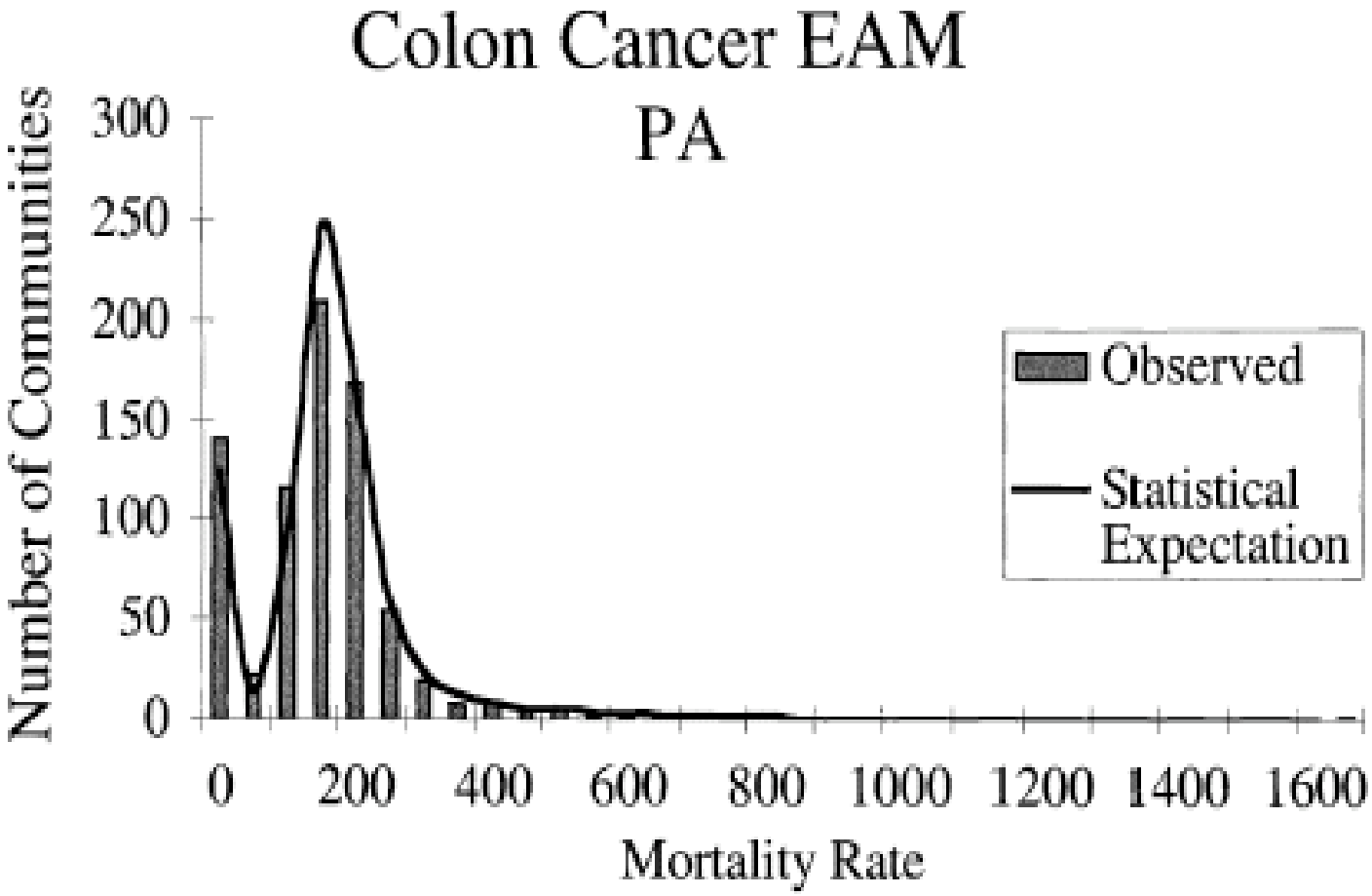
Hemminki data shows a relative familial risk of 2.5 -2.8 for *late-onset* CRC.

Hemminki data shows ZERO spousal risk for CRC for cohabitants of >30 years. Concludes that in present day Sweden environmental risk for CRC is ubiquitous. i.e. equals close to 1.0.

If environmental risk is close to 1.0 then genetic risk is close to 0.2.

We performed an independent trial.

MODERN CANCER RATES AMONG U.S. COMMUNITIES DISTRIBUTE ACCORDING TO THE NULL HYPOTHESIS



520 Pennsylvania communities 1958-1995 (Dr. Janice Vatland)

COLORECTAL CANCER (late onset)

CALCULATIONS FOR CASE OF MONOGENIC RISK AND UBIQUITOUS ENVIRONMENTAL RISK

CASE I.Risk derived from heterozygosity:
LIFETIME GENETIC RISK ~ 0.18-0.20
2pq ~ 0.18-0.2

$$q \sim 0.100 - 0.113$$

LATE ONSET FAMILIAL RISK ~ 2.5 -2.8
0.5/2pq ~ 2.5-2.8
2pq~ 0.2 - 0.18

$$q \sim 0.113 - 0.100$$

Such agreements are encouraging. These data/calculations are the first indicating that simple monogenic risk may account for major forms of cancer.

COLORECTAL CANCER (late onset)

CALCULATIONS FOR CASE OF MONOGENIC RISK AND UBIQUITOUS ENVIRONMENTAL RISK

CASE II.Risk derived from homozygosity

LIFETIME GENETIC RISK ~ 0.18-0.20
q^2 ~ 0.18-0.2
q ~ 0.424 - 0.447

LATE ONSET FAMILIAL RISK ~ 2.5 -2.8
1/q ~ 2.5-2.8
q ~0.40- 0.36

These results do not seem to support the case for risk conferred by homozygosity. Hemminki, Morgenthaler and MIT group are continuing data collection and analyses.

COLORECTAL CANCER (late onset)

CALCULATIONS FOR CASE OF MONOGENIC RISK AND UBIQUITOUS ENVIRONMENTAL RISK

CASE III.Risk from multigenic (n=32) heterozygosity.

LIFETIME GENETIC RISK ~ 0.18- 0.20

$[1 - e^{-32\,(2pq)}]$ **~ 0.18-0.2**

2pq ~ 0.0062- 0.0070

q ~ 0.0031-0.0035

LATE ONSET FAMILIAL RISK ~ 2.5

$[(0.5 + 0.2)/ (1 - e^{-32(2pq)}]$ **~ 2.5 -2.8**

2pq ~0.010-0.009

q ~ 0.005-0.0045

These results do not support the case for risk by multigenic heterozygosity.

IMPLICATIONS FOR REQUIRED COHORT SIZES

Assuming as many as 1 neutral mutant per gene copy, conditions of monogenic heterozygote risk with q<0.25 are expected to be detected with cohort sizes of 1000.

As q -> 0.4 larger cohort sizes approaching 10,000 persons are required. (q=0.5 cannot be detected.)

In general, homozygote risk, monogenic or higher, requires smaller cohort sizes than heterozygote risk.

The average for q is expected to be around 0.03 for the genes that carry nondeleterious inactivating mutations.

But genes with q = 0.3-0.8 are already known. (CYP2D6, GSMT).

SCIENTIFIC VIEWS & COHORT SIZES

Wandering the globe, I find a growing recognition that linkage studies using high frequency neutral alleles has not worked. A research director at a regional Pennsylvania hospital last year referred to the "snipe hunts' used as an initiation for naive summer campers in his youth. (See Mickey Mouse Club, circa 1955.)

However, this recognition tends to reach toward larger cohorts while continuing to use linkage to high frequency alleles as a potential answer. Population geneticists are stil saying that monogenic risk would have already been discovered by linkage studies without considering that multi-allelic risk defeats this strategy even if cohorts were as large as might be imagined.

PANGENOMIC SCANNING

DCE & competing technologies:

Cost & feasibility for scanning exons and splice sites for 25,000 genes for 100 common disease cohorts of 10,000 persons each.

Assume ~250,000 exons, ~500,000 isomelting domains for DCE preps.

Approximate scanning dimensions:

$2 \times 25{,}000 \times 100 \times 10{,}000 \sim 5 \times 10^{10}$ gene copy scans

or

$\sim 5 \times 10^{11}$ exon copy scans

or

$\sim 10^{14}$ base pair scans

or

10^{12} DCE isomelting domain scans

DCE & competing technologies.

Direct Sequencing

Sample costs +

@$0.01/base pair x 10^{14} = 10^{12}

@$0.001/base pair x 10^{14} = 10^{11}

DCE

Sample costs +

@$800*/domain x 5 x 10^5 domains = $4 x 10^8

assuming 100 persons/DCE run using oscillating DCE and multiple runs per capillary a' la Ekstrom.

*** X.-C. Li-Sucholeiki and W.Thilly appear to have devised a less expensive approach that also permits larger pools. (November 2004)**

DCE & competing technologies

DCE acceptance is unlikely for most medical geneticists using high frequency allele linkage analysis to scan the genome independent of cohort size.
(This now appears to include DeCode.)

The exceptions will be geneticists with physiology-driven hypotheses who think they have excellent gene candidates and perceive the need for cohort sizes of 1000 to 10,000.

Because DCE comprehensively detects nearly all point mutations it gives the greatest chance of detecting a real multi-allelic risk.

A SMALL DIGRESSION FROM POPULATION GENETICS

STRUCTURE FUNCTION RELATIONSHIPS OF ENZYMES AND OTHER PROTEINS

a.) With selection: set of mutations inactivating or functionally altering protein revealed.

b.) Without selection: set of all mutations revealed.

c.) Original goal of mutational spectrometry at MIT in HPRT gene. Re-introduced as “TILLING”.

SOURCES OF POINT MUTATIONS

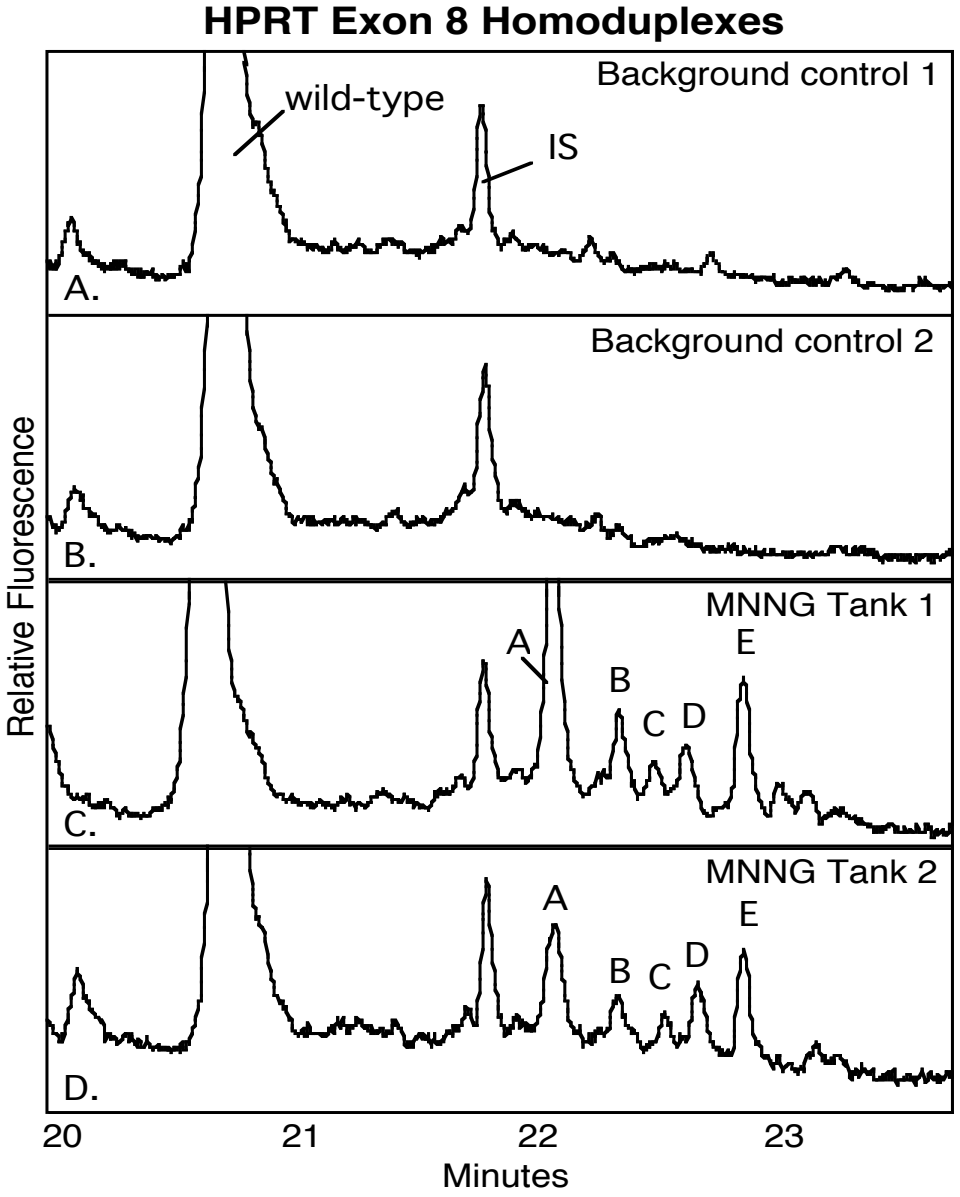
a.) errors of DNA polymerases copying undamaged DNA possibly during “DNA turnover”.

b.) miscopying deaminated methyl cytosines and cytosines

c.) copying over cell-generated methylated bases

d.) copying over other DNA damage sites

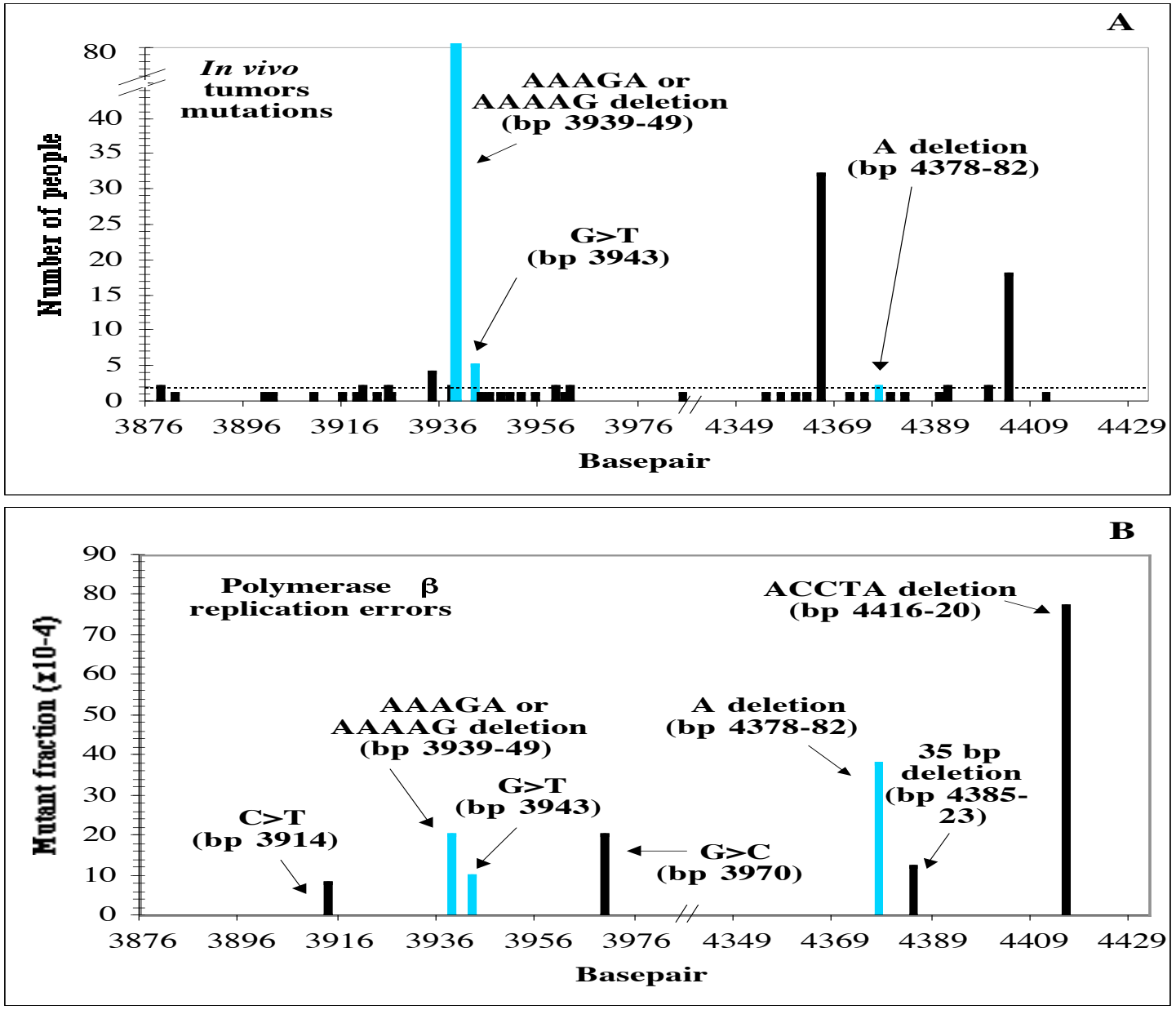
CDCE/hifiPCR mutational spectrometry has a sensitivity of < 10^{-5} without selection < 10^{-8} with selection.



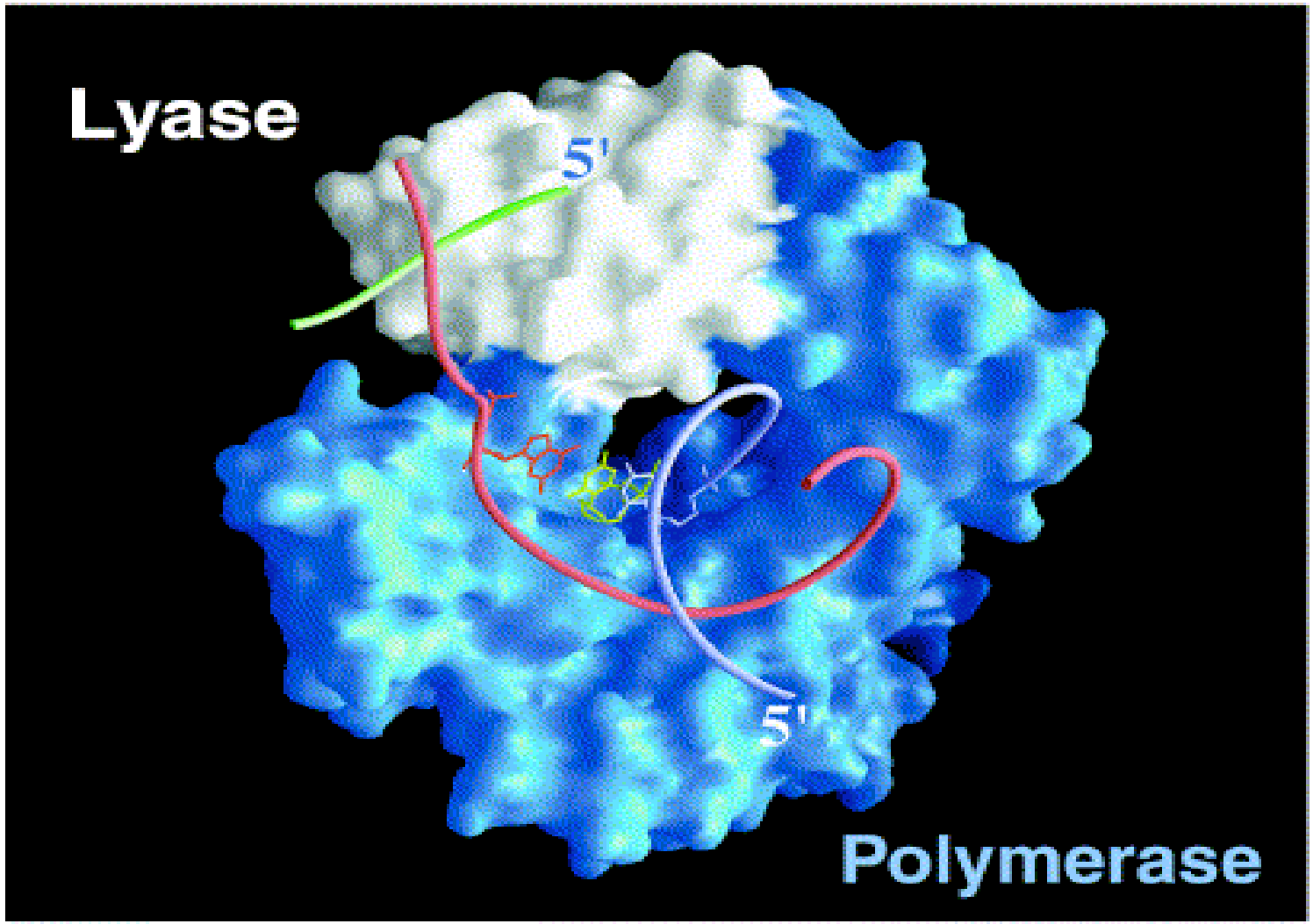
Example of HPRT 6-TG resistant mutants (A,B,C,D,E) each present at ~10^{-8} in human cell culture.

(Aoy Tomita-Mitchell)

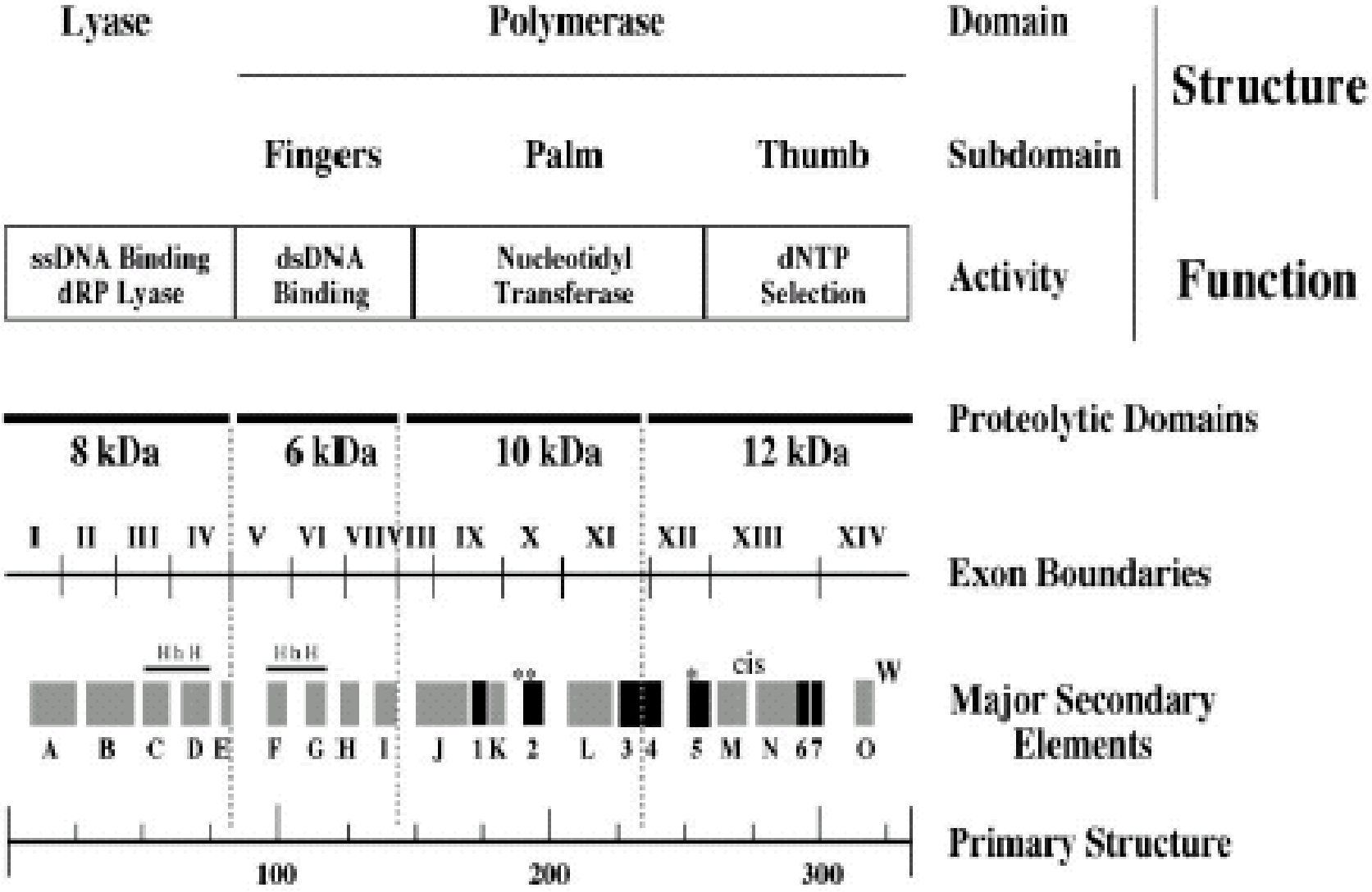
POL B induced hotspots match ~55% of APC exon 15 hotspots in CRC.



DNA Polymerase Beta

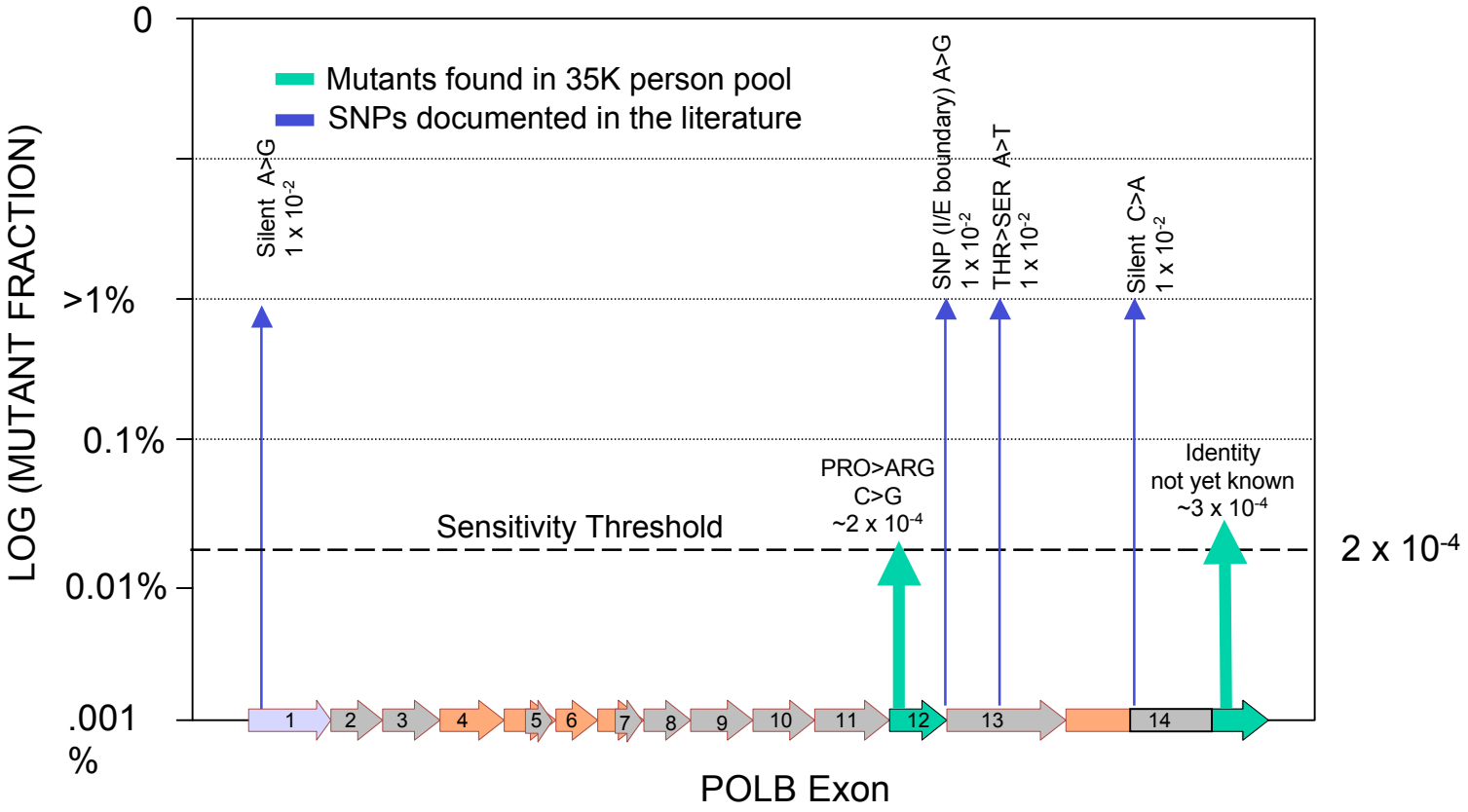


Pol Beta Structural Summary



Pol Beta: DNA polymerase beta

➢Novel mutants were discovered in 2 of 14 exons; the remaining 10 were vacant at the sensitivity of the screen.



- Gene: CTLA4 “nominated” as carrying mutations causing juvenile diabetes.
- Population Size- Young adults: **78,200 individuals**
- Population Size- Juvenile diabetics: **3800 individuals**