MOLECULAR EVOLUTION

● In the context *Phylogenetic methods*: nucleotide substitution models, a practical model choice view of certain aspects of molecular evolution.

● In the course Biometry and bioinformatics III one theme: how to infer (by stat-comp-methods) the effects of natural selection (for example, molecular adaptation) on sequences of nucleotides and proteins. This is another *practical* topic on molecular evolution.

● **Here, at the end of BB_II, we get familiar on certain concepts and characteristics of molecular evolution through a set of scientific papers**. Note that some of these issues are involved in Assignment set 2.

- ● **The following is one part of the exam => a part of the exam will be home-exam.**
- ● You can prepare in advance an answer to one exam question.
- ● The weight of this question is 25% of the exam points:
    - ● Write an one page essay about each paper. In the exam you will be asked to submit one of them, but in advance you don´t know which one => You should read them all and you have, of course, your essays available in your own files when you come to the exam – like all other material, too.

MOLECULAR EVOLUTION , READING SCIENTIFIC PAPERS

● All students, in all disciplines, should read scientific papers!

● It is usually not possible to understand a whole paper and you are supposed to get just some kind of a touch to a given topic in a given paper.

● You can also select a certain theme from a paper, and your essay thus need not cover the whole paper. For example, your essay, after reducing the paper content, can be totally biological or, for example, totally statistical, computational or something between these extremes.

● Your essay cannot contain any kind of copy-pasting from the papers. Only text, produced by you, with your own words, will be acceptable.

Papers in course webpage:

1. The genetic code is one in a million

2. The genetic code constraints yet facilitates Darwinian evolution.

3. Universal trend of amino acid gain and loss in protein evolution

4. Codon usage in eukaryotes.

MUTATIONS IN GENOMES, Background to 9.10 session

● Various kind of mutations, the basic **evolutionary factors**, produce evolutionary raw material. Other evolutionary factors - recombination, natural selection and random drift - dictate the fates of mutations.

● Comparative approaches, involving data from multiple different species, are suitable for detecting past selection. One important tool used to detect selection from genome data is to compare the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site (d_N/d_S) (see last page, the genetic code).

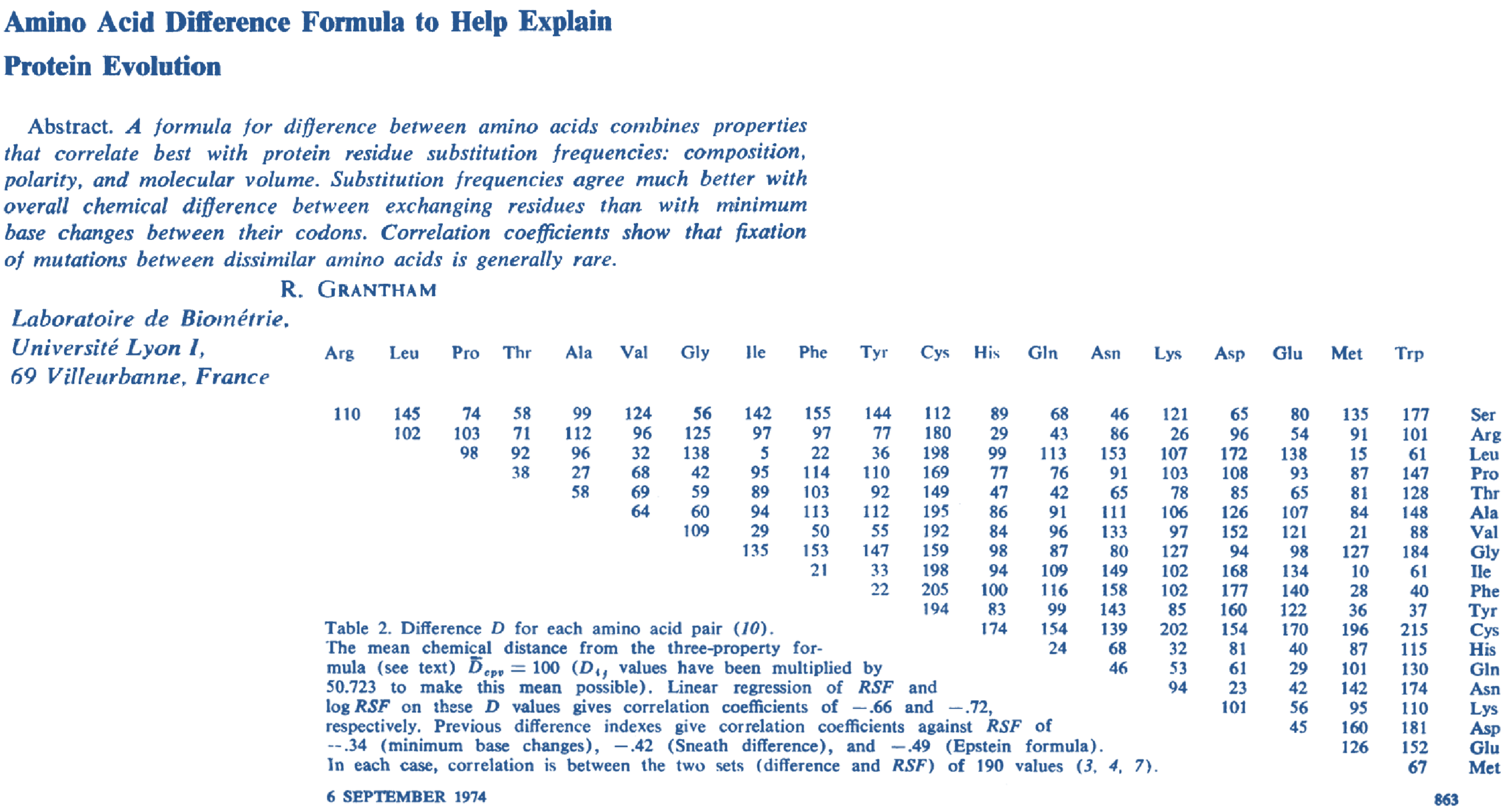
● This set of slides is about current results on profiling, through an evolutionary ”telescope” , amino acid mutations in human genome data.

● Wednesday 9.10 we have a session is for this kind of analysis by widely used softwares SIFT and PolyPhen for certain human genome mutations.

● Which amino acids can be replaced by which - through mutations? An old problem and extremely relavant in many kind of biological and medical questions! Next page shows the historical first step for quantifying the problem.

AMINO ACID DIFFERENCES - GRANTHAM DISTANCE

A classical paper, *Science 185:862-864, 1974*: Grantham distance.



Amino Acid Difference Formula to Help Explain Protein Evolution

Abstract. *A formula for difference between amino acids combines properties that correlate best with protein residue substitution frequencies: composition, polarity, and molecular volume. Substitution frequencies agree much better with overall chemical difference between exchanging residues than with minimum base changes between their codons. Correlation coefficients show that fixation of mutations between dissimilar amino acids is generally rare.*

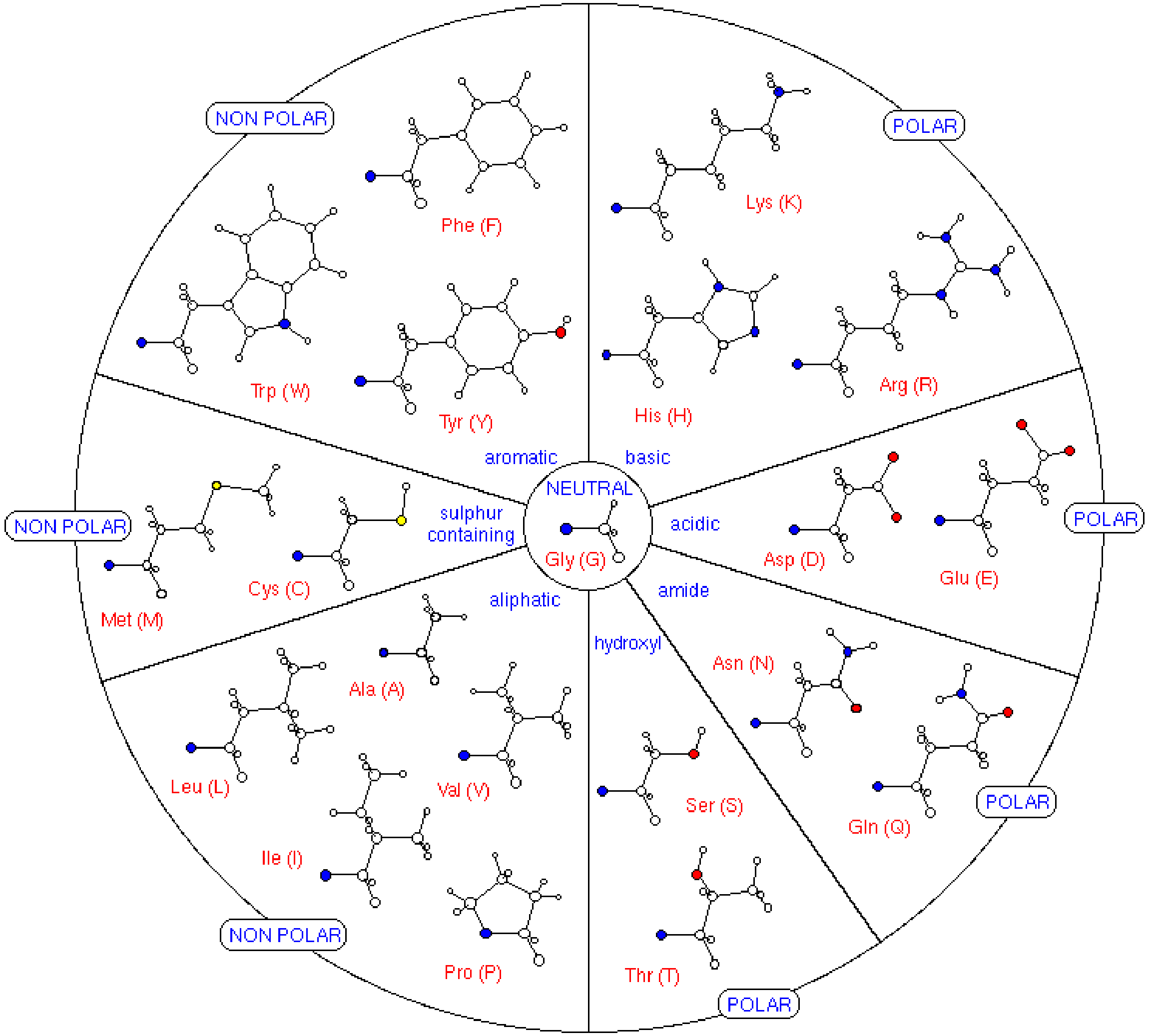
R. GRANTHAM

Laboratoire de Biométrie,
Université Lyon I,
69 Villeurbanne, France

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	100	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

Table 2. Difference D for each amino acid pair (*10*). The mean chemical distance from the three-property formula (see text) $\bar{D}_{cpv} = 100$ (D_{ij} values have been multiplied by 50.723 to make this mean possible). Linear regression of *RSF* and log *RSF* on these D values gives correlation coefficients of −.66 and −.72, respectively. Previous difference indexes give correlation coefficients against *RSF* of --.34 (minimum base changes), −.42 (Sneath difference), and −.49 (Epstein formula). In each case, correlation is between the two sets (difference and *RSF*) of 190 values (*3*, *4*, *7*).

AMINO ACID DIFFERENCES, CHEMICALLY



nSNV´s IN HUMAN GENOMES – GENETIC PROFILING

This chapter is based on Kumar et al. 2011, Trends in Genetics 27: 277-386

● Thousands of individuals in the general public have begun to gain access to their genetic variation profiles by using direct-to-consumer DNA tests available from commercial vendors, which profile hundreds of thousands of genomic markers for low costs.

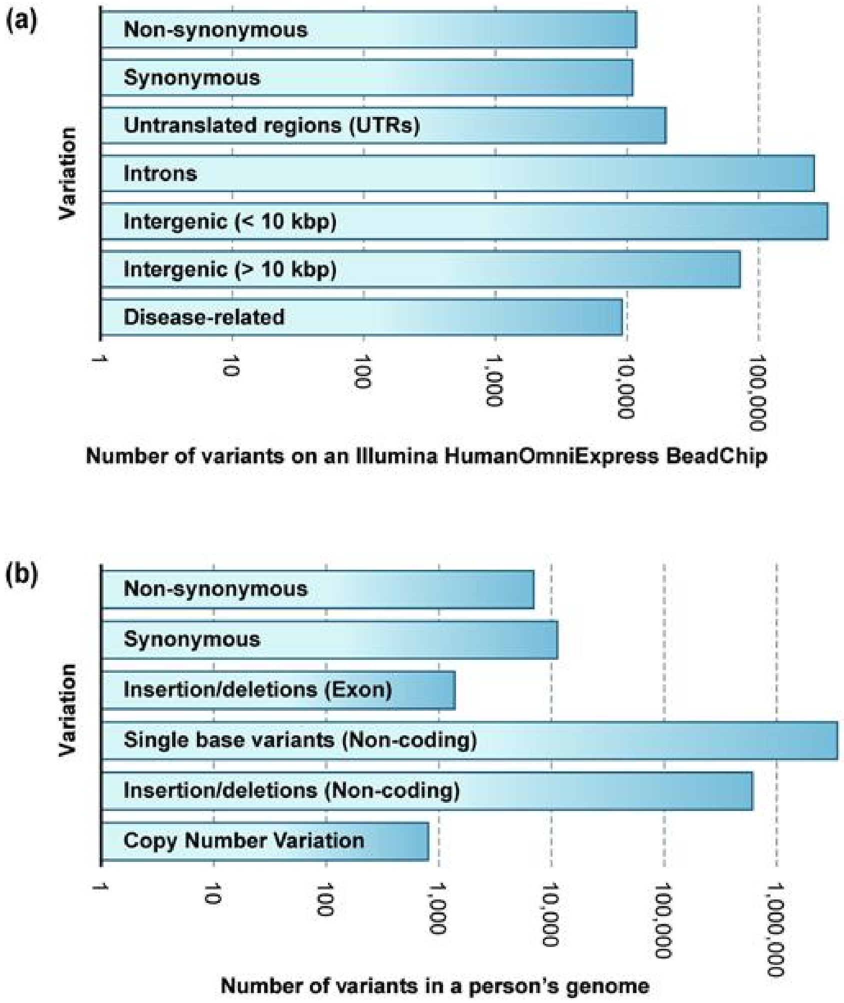
● Through this genetic profiling, individuals hope to learn about not only their ancestry, but also genetic variations underlying their physical characteristics and predispositions to diseases.

● Biomedicine scientists have been profiling variations at genomic markers in healthy and diseased individuals at genome scale in a variety of disease contexts and populations:
Discovery of thousands of disease associated genes and DNA variants.

● Any one personal genome contains more than a million variants, the majority of which are **single nucleotide variants, SNVs.**

● Majority of the known disease-associated variants are found within protein-coding genes with genome-wide association studies beginning to reveal also thousands of non-coding variants. Proteins are encoded in genomic DNA by exon regions, which comprise just ~1% of the genomic sequence, **Exome**. This is best understood part: how DNA blueprint sequence relates to function, and is arguably the best chance to connect genetic variations with disease pathophysiology. **A person's exome carries about 6,000 – 10,000 amino-acid-altering nonsynonymous SNVs, nSNVs,** known to be associated with more than a thousand major diseases **.**

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING

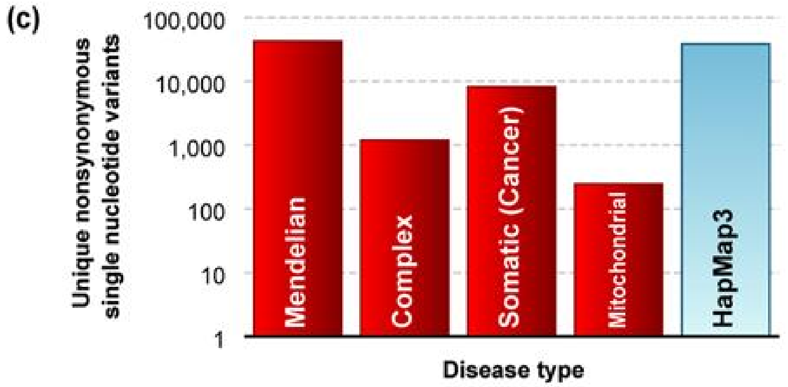


Profiles of personal and population variations.

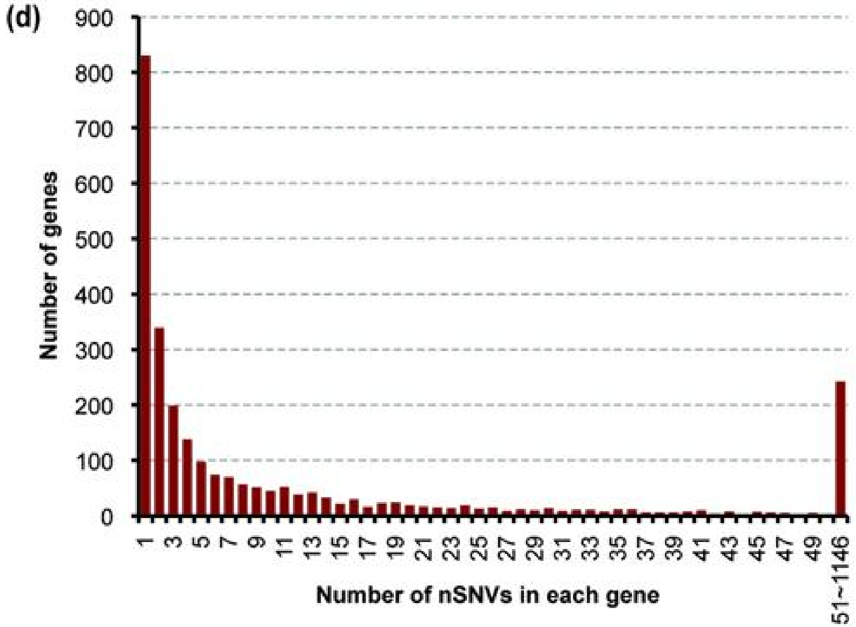
(a) Counts of various types of genetic variants profiled by 23andMe using the Illumina HumanOmniExpress BeadChip. 733,202 SNP identifiers (rsIDs), retrieved from the Illumina website and mapped to the dbSNP database.

(b) The numbers of different types of variants found per human genome.

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING



(c) The numbers of known non-synonymous single nucleotide variants (nSNVs) in the human nuclear and mitochondrial genomes that are associated with Mendelian diseases, complex diseases, and somatic cancers. Compared to complex diseases and somatic cancers, nSNVs related to Mendelian diseases account for the most variants discovered to date.



(**d**) The number of nSNVs in each gene related to Mendelian diseases. The majority of genes have only one or a few mutations, while there are some genes hosting hundreds or even more than 1000 mutations.

The numbers of variants in panels {a–c} (a,b in previous page) are in $\log_{10}$ scale. Information for disease associated variants is shown in red and the personal and population variations are shown in blue.

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING

● Translating a personal variation profile into useful phenotypic information (e.g., relating to predisposition to disease, differential drug response, and other health concerns) is a grand challenge in the field of genomic medicine. Genomic medicine is concerned with enabling healthcare that is tailored to the individual based on genomic information.

● **Phylomedicine:** Through multispecies comparisons of data from various animals in "the tree of life", it is possible to mine this information and evaluate the severity of each variant computationally (*in silico*).

● With the availability of large number genomes from the tree of life, it is becoming clear that evolution can serve as a kind of telescope for exploring the universe of genetic variation. In this evolutionary telescope, the degree of historical conservation of individual position (and regions) and the sets of substitutions permitted among species at individual positions serve as two lenses. This tool has the ability to provide first glimpses into the functional and health consequences of variations that are being discovered by high-throughput sequencing efforts.

● Phylomedicine is an important discipline at the intersection of molecular evolution and genomic medicine with a focus on understanding of human disease and health through the application of long-term molecular evolutionary history. Phylomedicine expands the purview of contemporary evolutionary medicine to use evolutionary patterns beyond the short-term history (e.g., populations) by means of multispecies genomics.

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES

Mendelian (monogenic) diseases

● For centuries it has been known that particular diseases run in families, notably in some royal families where there was a degree of inbreeding. Once Mendel's principles of inheritance became widely known in the early 1900s it became evident from family genealogies that specific heritable diseases fit Mendelian predictions.

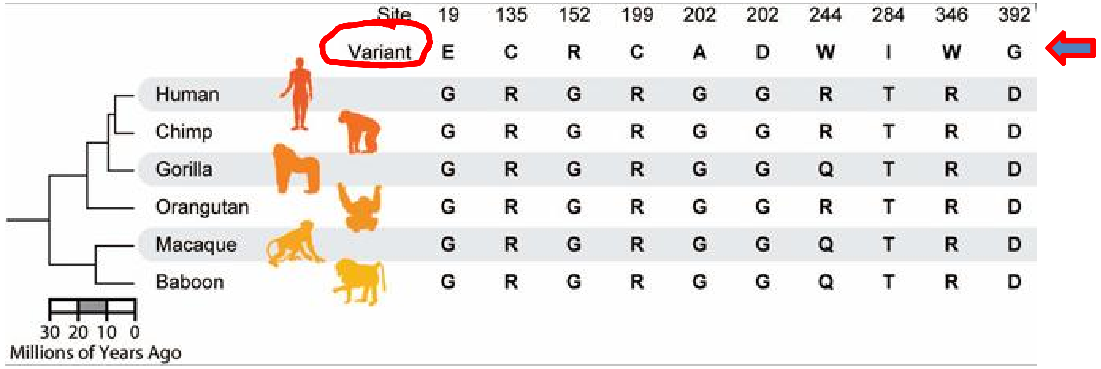
● Over the last three decades, mutations in single (candidate) genes in many families have been linked to individual Mendelian diseases. Sometimes more than a hundred SNVs in the same gene have been implicated in a particular disease. For example, by the turn of this century, individual patient and family studies revealed over 500 nSNVs in the Cystic fibrosis transmembrane conductance regulator (*CFTR*) gene for cystic fibrosis (CF). *This enabled first efforts to examine evolutionary properties of the positions harboring CFTR nSNVs.*

- ● *The disease-associated nSNVs were found to be overabundant at positions that had permitted only a very small amount of change over evolutionary time.*
- ● This trend was confirmed at the proteome scale in analyses of thousands of nSNVs from hundreds of genes.
- ● These patterns were in sharp contrast to the variations seen in non-patients, which are enriched in the fast evolving positions. In population polymorphism data, faster evolving positions also show higher minor allele frequencies than those at slow evolving positions, which translates into *an enrichment of rare alleles in slow-evolving and functionally important genomic positions.*

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES

An example

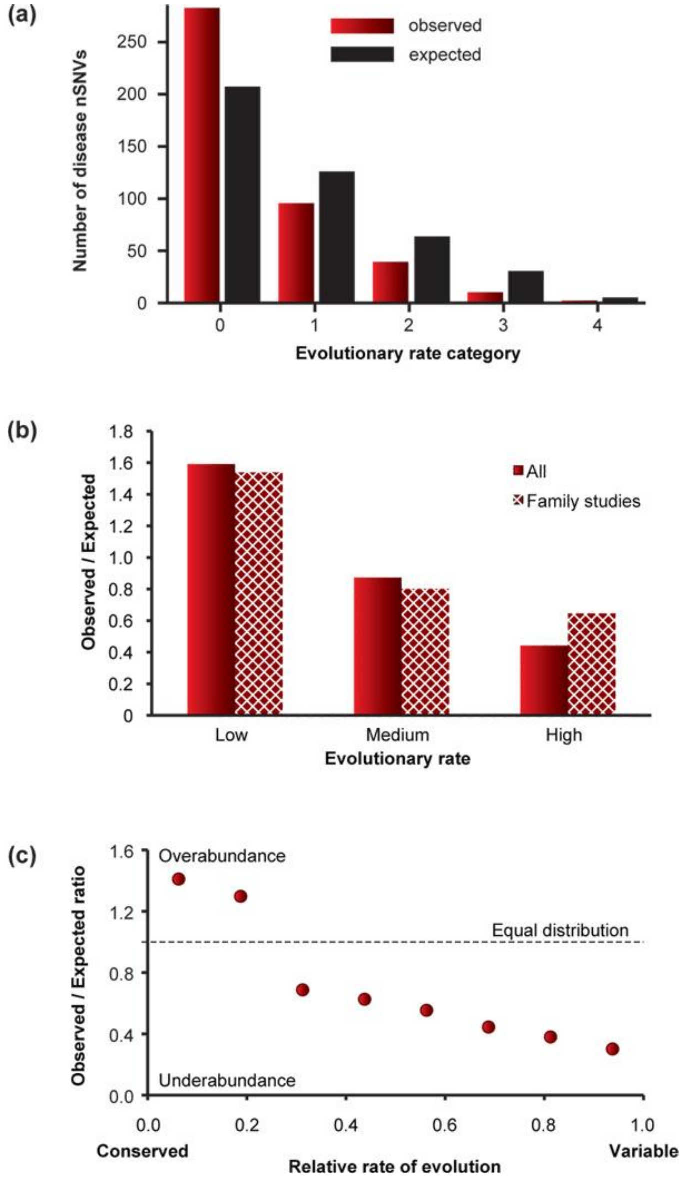
Miller syndrome is a rare genetic disorder characterized by distinctive craniofacial malformations that occur in association with limb abnormalities. It is a typical Mendelian disease that is inherited as an autosomal recessive genetic trait. By sequencing the exomes of four affected individuals in three independent kindreds, ten mutations in a single candidate gene, *DHODH*, were found to be associated with this disease. They are in slow-evolving sites that are highly conserved not only in primates, but also among distantly related vertebrates. Specifically, 50% of these mutations are found at completely conserved positions among 46 vertebrates, including human. The average evolutionary rate for sites containing these disease-related mutations is 0.50 substitutions per billion year, which is ~40% slower than those sites hosting four non-disease-related population polymorphisms of DHODH available in the public databases.



Site	19	135	152	199	202	202	244	284	346	392
Variant	E	C	R	C	A	D	W	I	W	G
Human	G	R	G	R	G	G	R	T	R	D
Chimp	G	R	G	R	G	G	R	T	R	D
Gorilla	G	R	G	R	G	G	Q	T	R	D
Orangutan	G	R	G	R	G	G	R	T	R	D
Macaque	G	R	G	R	G	G	Q	T	R	D
Baboon	G	R	G	R	G	G	Q	T	R	D

Ten amino acid altering mutations at sites 19, 135, etc. referring to the protein sequence positions

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



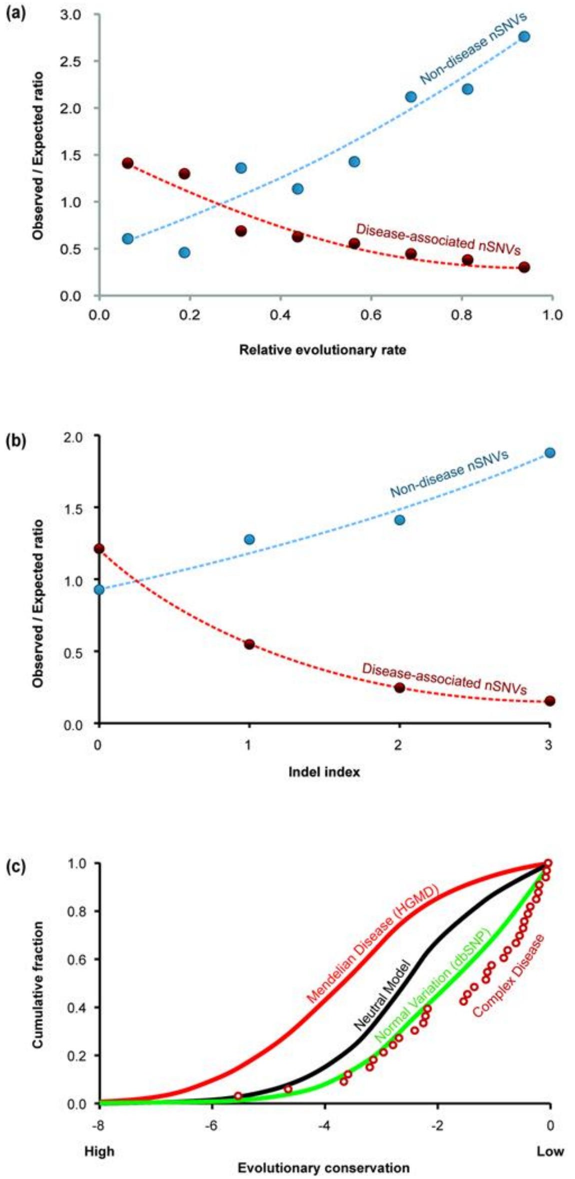
Evolutionary properties of positions afflicted with disease-associated nonsynonymous single nucleotide variants (nSNVs).

(a) The observed and expected numbers of disease associated nSNVs in positions that have evolved with different evolutionary rates in the *CFTR* protein (cystic fibrosis). The disease associated nSNVs are enriched in positions evolving with the lowest rates, which belong to the rate category 0.

(b) The ratio of observed to expected numbers of nSNVs in different rate categories for all *CFTR* variants (solid pattern; 431 variants) and those reported in publications profiling one or more families (hatched pattern; 59 variants).

(c) The proteome-scale relationship of the observed/expected ratios of Mendelian disease-associated nSNVs in positions that have evolved with different evolutionary rates. The results are from an analysis of disease associated nSNVs from 2,717 genes (public release of HGMD). Just as for individual diseases, nSNVs are enriched in positions evolving with the lowest rates.

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



The enrichment of disease-associated nSNVs (red) and the deficit of population polymorphisms (blue) in human amino acid positions

(a) evolving with different rates and

(b) with differ degrees of insertion-deletions. In both cases, smaller numbers on the x axis correspond to more conserved positions. There is an enrichment of disease associated nSNVs and a deficit of population nSNPs in conserved positions. This trend is reversed for the fastest evolving positions.

(c) The cumulative distributions of the evolutionary conservation scores for nSNVs associated with Mendelian diseases (solid red line), complex diseases (open red circles), and population polymorphisms (green line). The shift towards the left in Mendelian nSNVs indicates higher position specific evolutionary conservation. Conversely, a shift towards the right in complex disease nSNVs indicates lower evolutionary conservation, which overlaps with normal variations observed in the population. Data for the neutral model (black line) is from a simulation.

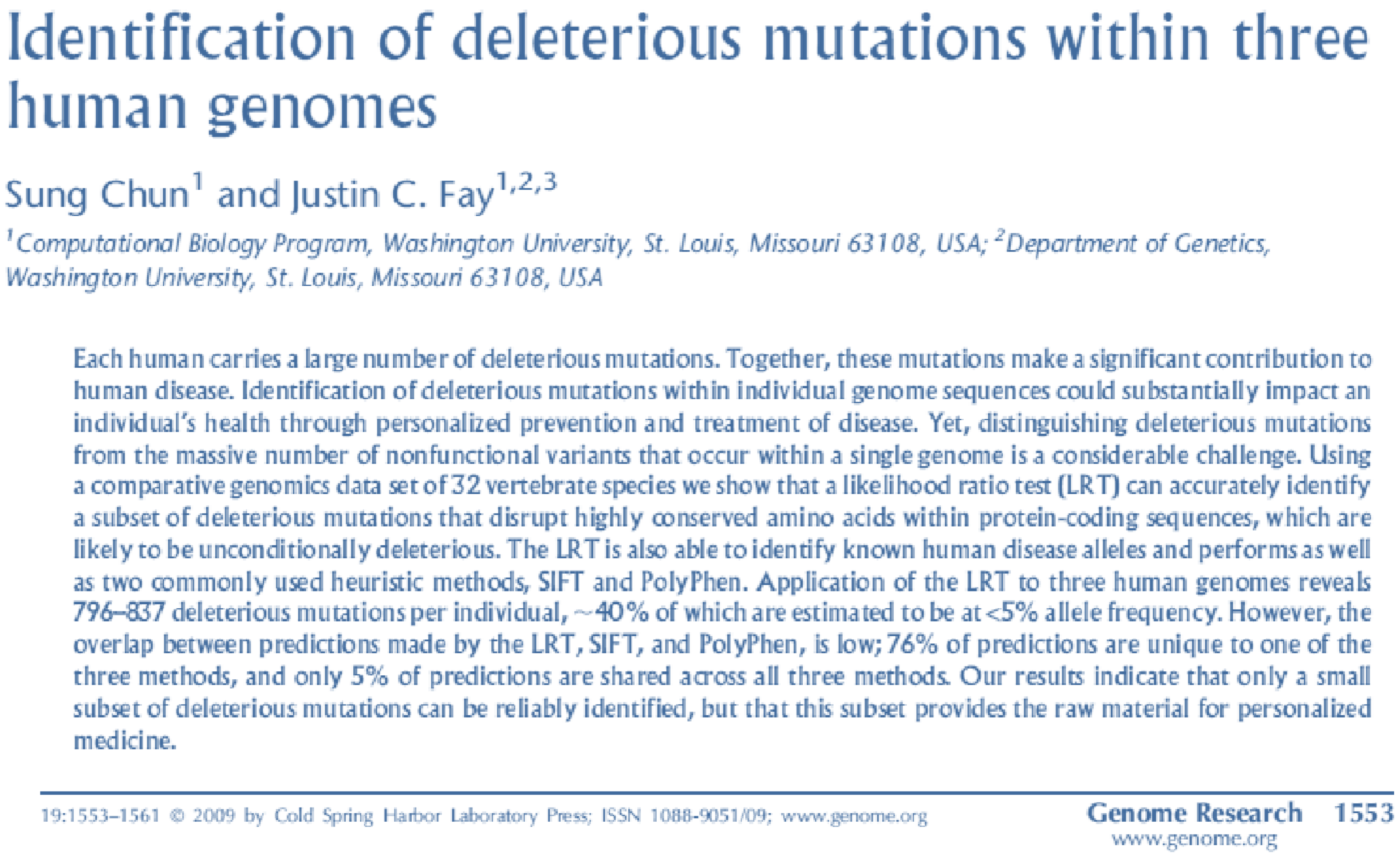
nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES

● Patterns of evolutionary retention at positions, another type of evolutionary conservation, a similar pattern is noticed: positions preferentially retained over the history of vertebrates were more likely to be involved in Mendelian diseases as compared to the patterns of natural variation. Somatic mutations in a variety of cancers have also been found to occur disproportionately at conserved positions. A similar pattern has emerged for mitochondrial disease-associated nSNVs.

● The relationship between evolutionary conservation and disease association has been explained by the effect of natural selection:

- ● There is a high degree of purifying selection on variation at highly conserved positions because of their potential effect on inclusive fitness (fecundity, reproductive success) due to the functional importance of the position.
- ● At the faster-evolving positions, many substitutions have been tolerated over evolutionary time in different species.
- ● This points to the “neutrality” of some mutations that spread through the population primarily by the process of random genetic drift and appear as fixed differences between species.
- ● Therefore, fewer mutations are culled at fast-evolving positions, producing a relative under-abundance of disease mutations at such positions. Of course, the above arguments hold true only when the functional importance of a position has remained unchanged over evolutionary time, an assumption that is expected to be fulfilled for a large fraction of positions in orthologous proteins.

DELETERIOUS MUTATIONS IN HUMANS



Identification of deleterious mutations within three human genomes

Sung Chun[1] and Justin C. Fay[1,2,3]

[1]Computational Biology Program, Washington University, St. Louis, Missouri 63108, USA; [2]Department of Genetics, Washington University, St. Louis, Missouri 63108, USA

Each human carries a large number of deleterious mutations. Together, these mutations make a significant contribution to human disease. Identification of deleterious mutations within individual genome sequences could substantially impact an individual's health through personalized prevention and treatment of disease. Yet, distinguishing deleterious mutations from the massive number of nonfunctional variants that occur within a single genome is a considerable challenge. Using a comparative genomics data set of 32 vertebrate species we show that a likelihood ratio test (LRT) can accurately identify a subset of deleterious mutations that disrupt highly conserved amino acids within protein-coding sequences, which are likely to be unconditionally deleterious. The LRT is also able to identify known human disease alleles and performs as well as two commonly used heuristic methods, SIFT and PolyPhen. Application of the LRT to three human genomes reveals 796–837 deleterious mutations per individual, ~40% of which are estimated to be at <5% allele frequency. However, the overlap between predictions made by the LRT, SIFT, and PolyPhen, is low; 76% of predictions are unique to one of the three methods, and only 5% of predictions are shared across all three methods. Our results indicate that only a small subset of deleterious mutations can be reliably identified, but that this subset provides the raw material for personalized medicine.

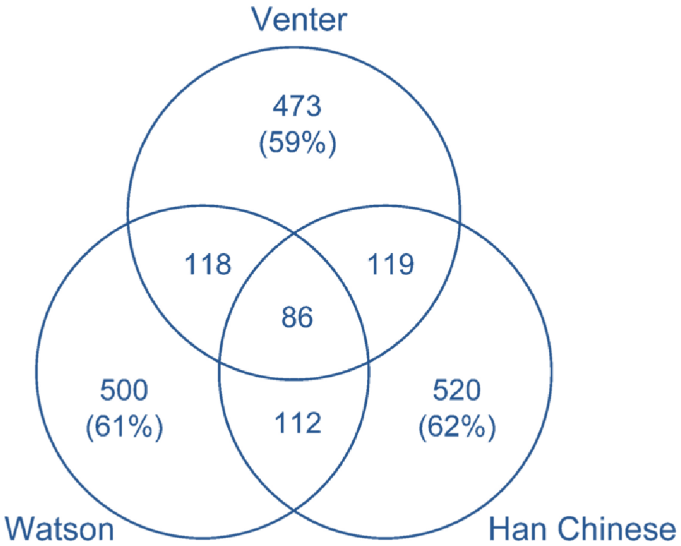
19:1553–1561 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Genome Research 1553
www.genome.org

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES

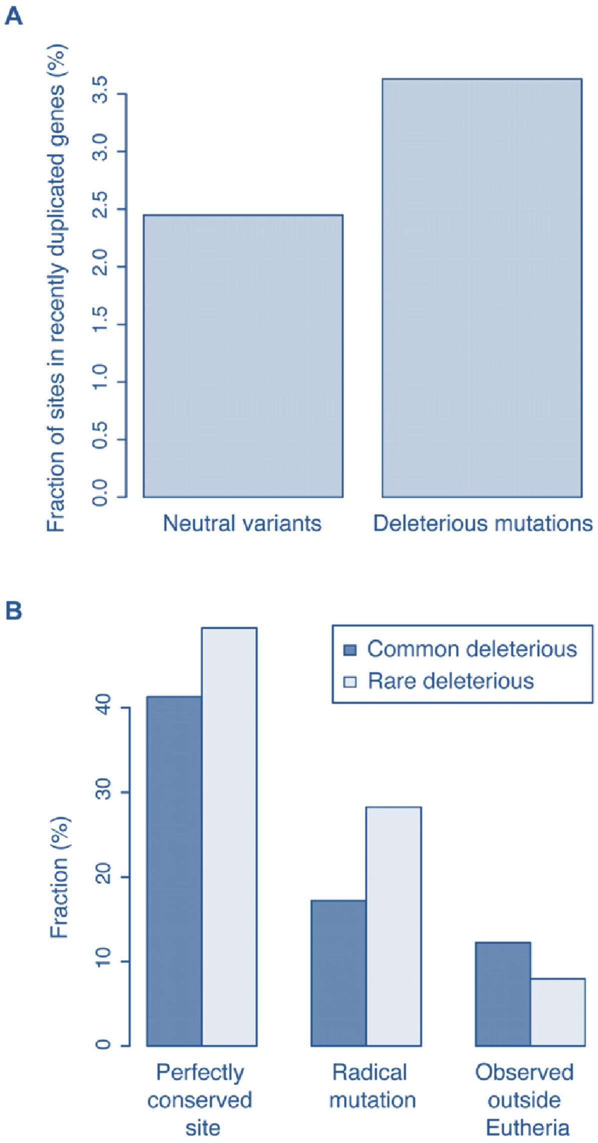
● The paper has data from a complete catalog of SNPs from J. Craig Venter (Google who he is if you don´t know), from a Han Chinese male from their respective websites (http://www.jcvi.org/cms/research/projects/huref/ and http://yh.genomics.org.cn), and for James D. Watson (from "Watson – Crick")

● Nonsynonymous and synonymous SNPs were identified using known genes in Ensembl release 49. Coding SNPs in ambiguous reading frames, due to overlap of adjacent genes or frame shifts between known splice variants, or in known pseudogenes, were excluded.



● The percentage of individual-specific deleterious mutations found in each genome is shown in parentheses.

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES

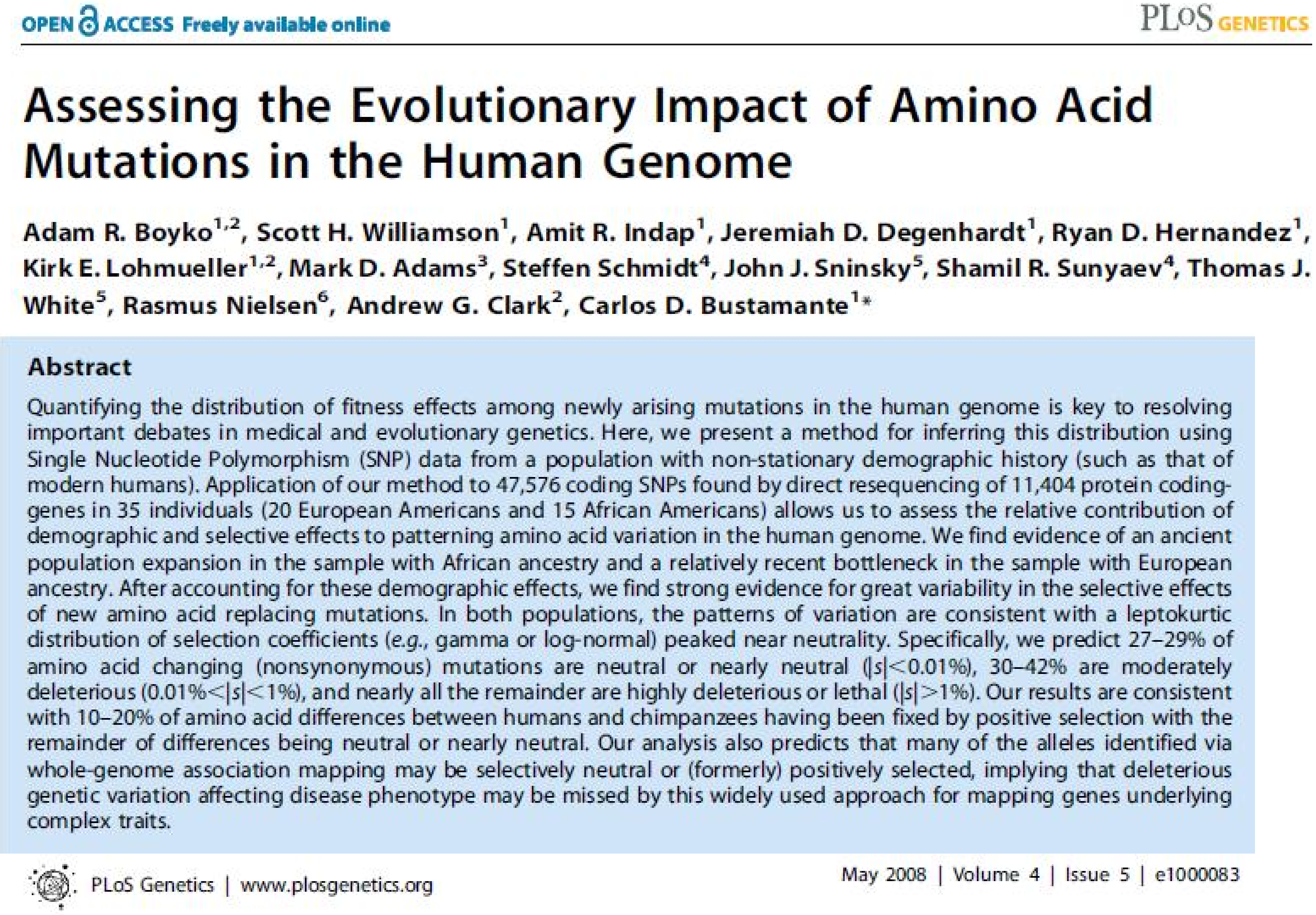


Characteristics of deleterious mutations.

(*A*) Deleterious mutations (*n* = 1928) are more likely to occur in recently duplicated genes relative to neutral variants (*n* = 8287).

(*B*) Mutations at perfectly conserved sites, mutations that cause radical amino acid changes, defined by BLOSUM62 ≤ −2, and mutations to amino acids that are not observed outside of eutherian mammals are more frequent among rare (*n* = 807) compared with common deleterious mutations (*n* = 1121).

EXAMPLE 1



OPEN ACCESS Freely available online

PLoS GENETICS

Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome

Adam R. Boyko[1,2], Scott H. Williamson[1], Amit R. Indap[1], Jeremiah D. Degenhardt[1], Ryan D. Hernandez[1], Kirk E. Lohmueller[1,2], Mark D. Adams[3], Steffen Schmidt[4], John J. Sninsky[5], Shamil R. Sunyaev[4], Thomas J. White[5], Rasmus Nielsen[6], Andrew G. Clark[2], Carlos D. Bustamante[1]*

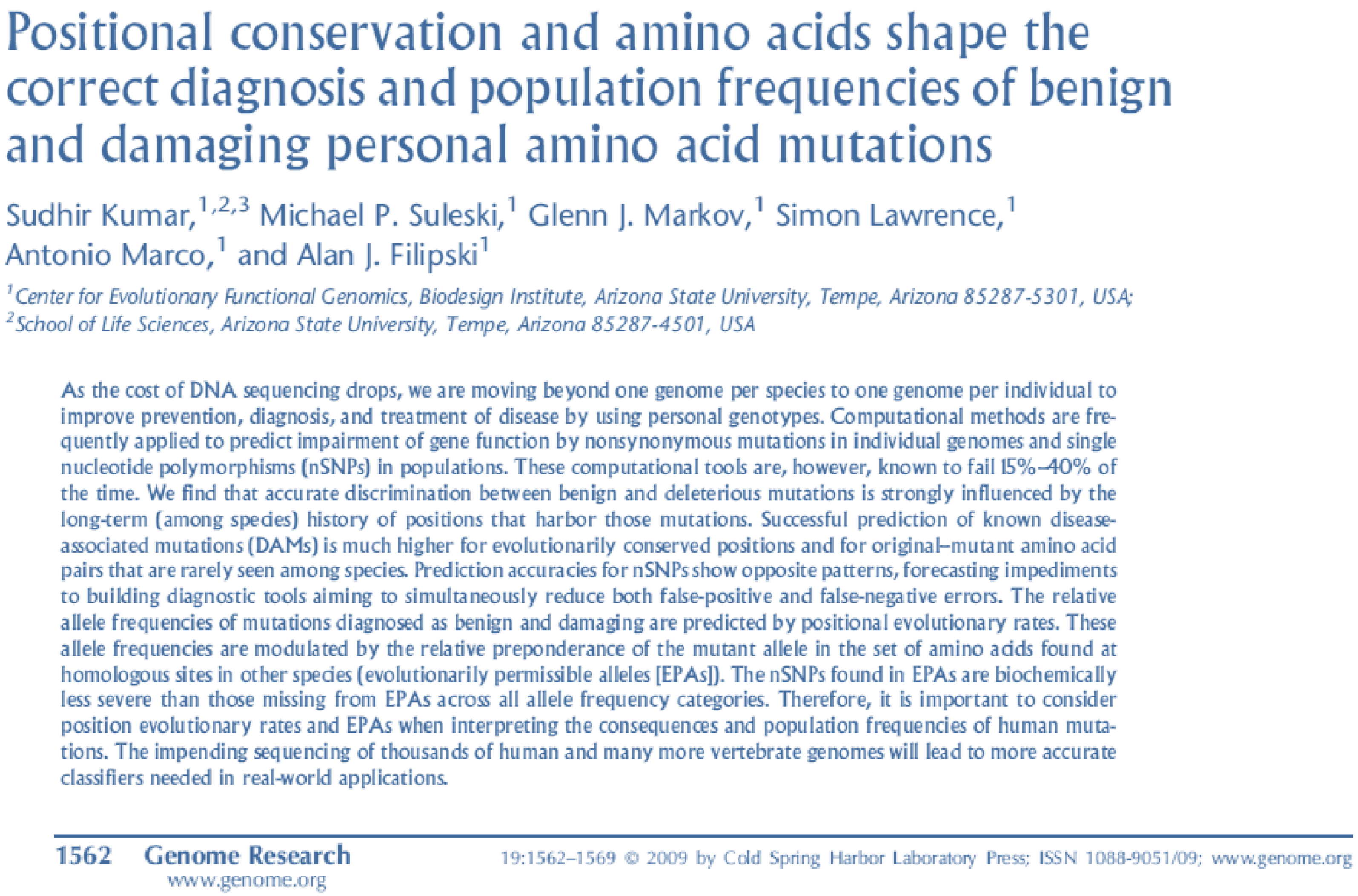
Abstract

Quantifying the distribution of fitness effects among newly arising mutations in the human genome is key to resolving important debates in medical and evolutionary genetics. Here, we present a method for inferring this distribution using Single Nucleotide Polymorphism (SNP) data from a population with non-stationary demographic history (such as that of modern humans). Application of our method to 47,576 coding SNPs found by direct resequencing of 11,404 protein coding-genes in 35 individuals (20 European Americans and 15 African Americans) allows us to assess the relative contribution of demographic and selective effects to patterning amino acid variation in the human genome. We find evidence of an ancient population expansion in the sample with African ancestry and a relatively recent bottleneck in the sample with European ancestry. After accounting for these demographic effects, we find strong evidence for great variability in the selective effects of new amino acid replacing mutations. In both populations, the patterns of variation are consistent with a leptokurtic distribution of selection coefficients (*e.g.*, gamma or log-normal) peaked near neutrality. Specifically, we predict 27–29% of amino acid changing (nonsynonymous) mutations are neutral or nearly neutral ($|s|<0.01\%$), 30–42% are moderately deleterious ($0.01\%<|s|<1\%$), and nearly all the remainder are highly deleterious or lethal ($|s|>1\%$). Our results are consistent with 10–20% of amino acid differences between humans and chimpanzees having been fixed by positive selection with the remainder of differences being neutral or nearly neutral. Our analysis also predicts that many of the alleles identified via whole-genome association mapping may be selectively neutral or (formerly) positively selected, implying that deleterious genetic variation affecting disease phenotype may be missed by this widely used approach for mapping genes underlying complex traits.

PLoS Genetics | www.plosgenetics.org

May 2008 | Volume 4 | Issue 5 | e1000083

EXAMPLE 2



Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations

Sudhir Kumar,[1,2,3] Michael P. Suleski,[1] Glenn J. Markov,[1] Simon Lawrence,[1] Antonio Marco,[1] and Alan J. Filipski[1]

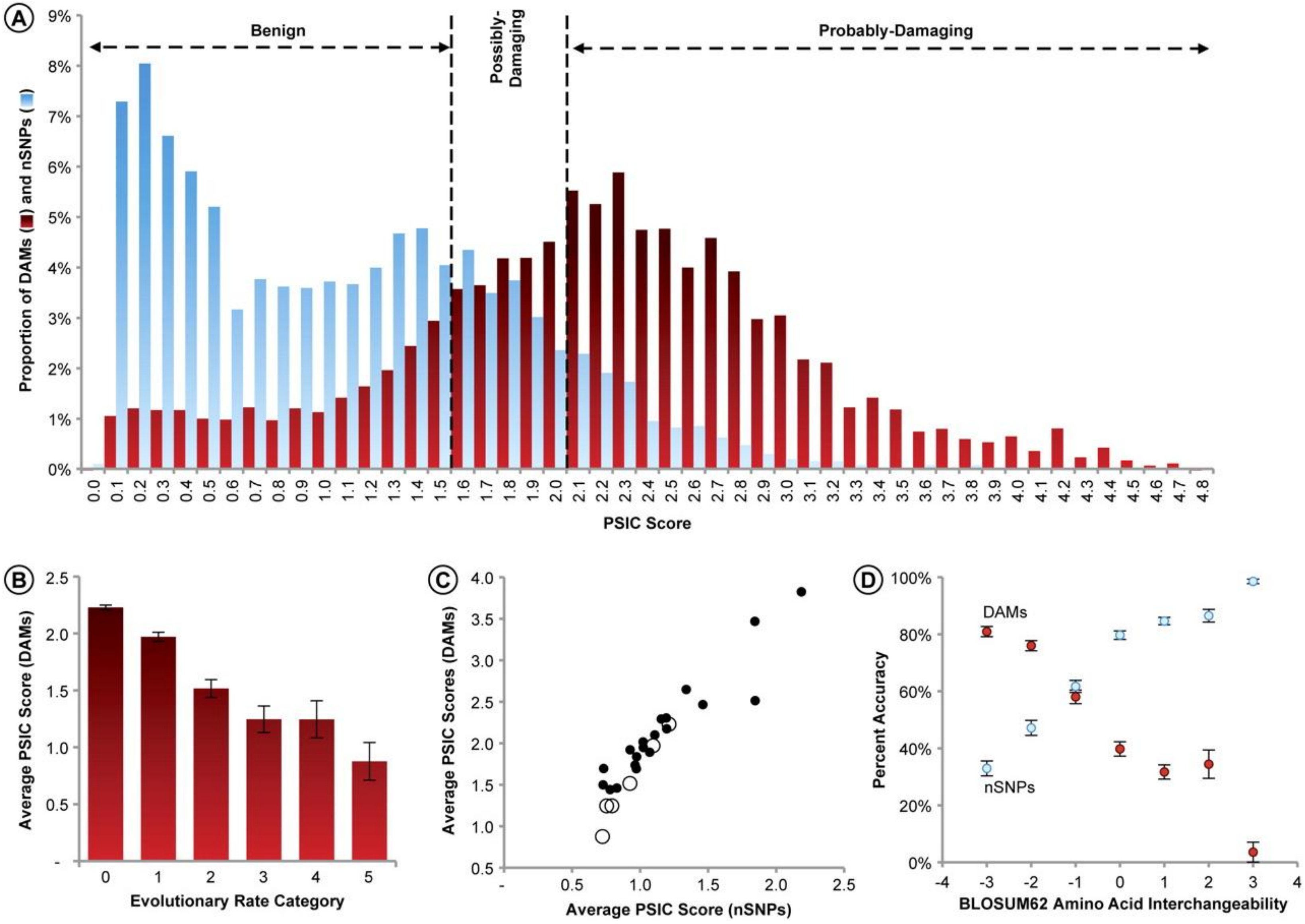
[1] Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301, USA;
[2] School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501, USA

As the cost of DNA sequencing drops, we are moving beyond one genome per species to one genome per individual to improve prevention, diagnosis, and treatment of disease by using personal genotypes. Computational methods are frequently applied to predict impairment of gene function by nonsynonymous mutations in individual genomes and single nucleotide polymorphisms (nSNPs) in populations. These computational tools are, however, known to fail 15%–40% of the time. We find that accurate discrimination between benign and deleterious mutations is strongly influenced by the long-term (among species) history of positions that harbor those mutations. Successful prediction of known disease-associated mutations (DAMs) is much higher for evolutionarily conserved positions and for original–mutant amino acid pairs that are rarely seen among species. Prediction accuracies for nSNPs show opposite patterns, forecasting impediments to building diagnostic tools aiming to simultaneously reduce both false-positive and false-negative errors. The relative allele frequencies of mutations diagnosed as benign and damaging are predicted by positional evolutionary rates. These allele frequencies are modulated by the relative preponderance of the mutant allele in the set of amino acids found at homologous sites in other species (evolutionarily permissible alleles [EPAs]). The nSNPs found in EPAs are biochemically less severe than those missing from EPAs across all allele frequency categories. Therefore, it is important to consider position evolutionary rates and EPAs when interpreting the consequences and population frequencies of human mutations. The impending sequencing of thousands of human and many more vertebrate genomes will lead to more accurate classifiers needed in real-world applications.

1562 Genome Research
www.genome.org

19:1562–1569 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

UNDERSTANDING EVOLUTIONARY PATTERNS OF MUTATIONS


A
Benign
Possibly-Damaging
Probably-Damaging
Proportion of DAMs (■) and nSNPs (■)
PSIC Score
B
Average PSIC Score (DAMs)
Evolutionary Rate Category
C
Average PSIC Scores (DAMs)
Average PSIC Score (nSNPs)
D
DAMs
nSNPs
Percent Accuracy
BLOSUM62 Amino Acid Interchangeability