Tracing the genetic etiology of cardiovascular disease using a hierarchy of common genetic variants derived from patient subgroups stratified by differential levels in severity

Sarosh N. Fatakia

Department of Biological Sciences, Tata Institute of Fundamental Research

1 Homi Bhabha Road, Colaba

Mumbai, 400005, INDIA

(Email:) sarosh dot fatakia @ gmail dot com

Abstract

A pedagogical review in informatics based on computational molecular evolution is synopsized to illustrate the role of common genetic variants during adaptation to diverse habitat and physiological functions across disparate species. With that background, an evolutionary perspective is motivated to trace lifestyle-related disease progression in humans. Cardiovascular disease (CVD) is a multi-factorial disease, where maladaptation due to a sedentary lifestyle and faulty diet may influence its prognosis, but a healthy diet and lifestyle may as well restrict it. However, a comprehensive genetic basis for the differential severity in CVD remains unreported. Here, we have computed that CVD severity may be associated with the hierarchical plasticity of genes whose common genetic variants are exclusive among those patients having disparate levels of disease severity. Most importantly, we have used a small and outbred subpopulation to demonstrate that constellation of common variants can be exploited to trace the disease etiology. Taken together, we report a novel perspective to identify a hierarchy of common variants associated with differential CVD severity and subsequently hypothesize how the constellations of common variants may collectively modulate disease severity, which in turn may result in a relatively broad and complex spectrum of severity at the population-level. To corroborate our findings, we report a hierarchical plasticity of *LDLR* gene, which has been previously implicated in the differential response to lipid metabolism, is associated with differential CVD severity in the subpopulation under consideration.

Keywords: Biomedical research, science education, lifestyle-related disease, evolutionary medicine, human genetics, common genetic variants, information theory.

Introduction

Human evolution [1-4] is a continuous process of adaptation [5, 6] during which we have developed immunity against diseases in general [7], as well as local to a region [8], and also kept pace with the dietary changes [9], external conditions of habitat [10-12], and geography [13]. In a comparative genomic study, we may invoke conserved (or invariant) genetic information to trace the shared ancestry [14, 15]. Inspired by a recent report that a healthy lifestyle may reverse some adverse consequences related to cardiovascular disease (CVD) [16], and merging comparative genomics with evolutionary biology, we present insights and new perspectives to rationalize the disease etiology. Moreover, it has been long been suggested that an evolutionary standpoint may improve our understanding of common lifestyle-related diseases such as CVD [17, 18], and recent reports suggest the limitations of converging toward a consensus set of causative variants from diverse studies [19-22].

Cardiovascular disease has been attributed to a plethora of genetic variants, which involve lipid regulation [23-26], inflammation [27-29], as well as other epigenetic agents such as an altered lifestyle [30, 31]. Genome wide association studies involving thousands of individuals, have identified hundreds of causative and associated variants to understand the genetic etiology of CVD [32-35], but it has been estimated that the spectrum of identified variants does not fully represent the inherited risk [22, 32-36]. The genetic etiology of such lifestyle-related diseases can seldom be identical in an outbred population because their ancestries may differ [37]. Hence, not all variants may be identified with high statistical confidence from a specific genome wide association study (GWAS), as suggested schematically in Fig. 1 (adapted from [38]). Therefore, most importantly, results of GWA studies, from distinct subpopulations may differ because the

causative variants may not be identified with similar statistical confidence and the ancestry of individuals with CVD may also differ [36].

Besides a well-documented set of candidate genes associated with cardiac diseases [39], there are other genes that may differentially affect their prognosis, for example DNA damage repair (DDR) genes *TP53* and *MDM2* have been recently implicated in cardiomyocyte function [40]. In this report, we propose that an evolutionary study may be invoked to trace a genetic etiology of CVD in an outbred subpopulation, and overcome the challenges of a small data set. From our proposed formulation, we may be able to corroborate previously implicated CVD genes (and their variants) and also identify new ones. As onset of CVD is largely lifestyle-related in the general population [16], we hypothesize that the disease etiology may be traced by contrasting distinct variants of the same gene. In fact it is well established that specific mutations (generally referred to as variants) of the receptor LDL receptor gene (*LDLR*) cause hypercholesterolemia, and might lead to high risk of myocardial infarction [41], but other mutations of *LDLR* may in fact be protective against increased LDL levels [41].

Using microarray technology, an invariant genotype can be detected from the genome-wide genotype that was probed across various individuals in an outbred subpopulation, at a specifically predetermined single nucleotide polymorphic (SNP) locus. In a comparative genomic study across those subjects, the genotype for a hypothetical common ancestor for the specific cohort may thus be unambiguously determined at that SNP locus. Their hypothetical last common ancestral genome of that outbred subpopulation will also manifest the same invariant genotype at the SNP locus, which has remained invariant subsequent to evolution [42]. Hence, we hypothesize that such SNP loci with invariant genotype (common alleles from that subpopulation), which have

resisted mutations, may be utilized to trace the genetic etiology of CVD cases with respect to age- and gender-matched controls. The occurrence of common genetic variants at any given genomic locus is an inevitable consequence of evolution within an outbred population [43], because those variants may progressively get inherited over generations [44]. Such common variants are likely to be a relatively recent outcome of evolution compared to invariant ones, but presumably older than the relatively rare variants [44]. However, there is a caveat that rare variations are either likely to be new (few generations old) and hence they have not been subject to negative selection for a long time, or are rare because they are being selected against owing to their deleterious genetic nature and are likely have greater penetrance in contrast to that due to common variants [44-46].

By partitioning human noncoding sequence into invariant (conserved) and variant (non-conserved) loci, it has been reported that agents of natural selection have influenced the human genome evolution [47]. Hence, from the perspective of a comparative genomic approach, interactions involving both invariant and variant genomic loci may lead to gain-of-function (adaptation) or loss-of-function [48, 49] or perhaps remain benign without any perceptible change at the genome-level [50, 51]. However, the adverse consequences of these changes may manifest later on during mid-life, much beyond the average reproductive age at the population level. For example, it has been suggested that high levels of cytokine response may confer an innate resistance to infectious diseases at an early age, and that the same response inadvertently becomes adversely associated with CVD at a much later age [52]. An evolutionary adaptation that confers reproducibility success at a relatively early age, but which may adversely affects us later, has also been reported with regard to an ancient variant allele at the growth differentiation factor-5 (*GDF5*) enhancer region [13]. Taken together, the constellation of

genetic variants may have a profound impact at the individual genome level [53-55]. Therefore, by reaching an appreciable frequency in a subpopulation, common variants (evolutionary adaptations) are instrumental for a differential physiological response such that the level of severity may be stratified. Therefore, we propose that the genetic etiology for CVD may be investigated in an outbred subpopulation of patients and age-matched healthy controls using common variants identified from patients with stratified severity level of this disease. We hypothesize that a hierarchical nature of those common genetic variants among patients with disparate severity may help trace the evolution of CVD severity.

Traditional GWA studies for complex multifactorial diseases warrant genome-wide data from thousands of individuals to identify causative variants with high statistical confidence [38]. Therefore, it has been a challenge to characterize the genetic etiology of CVD from India with access to genotype information from nearly one hundred and fifty CVD cases, and an identical number of age- and gender-matched controls. To overcome the limitation of a small data size, we propose insights and perspectives for an inter-disciplinary approach, and hierarchically trace genome-wide common genetic variants among CVD cases with stratified levels of severity that was determined from angiography-based studies [56, 57].

At first, we must present a primer on protein sequence evolution to elucidate its potential using published reports. This synopsis relates to proteins in general, and DNA damage repair proteins in particular, but the insights developed here are for a relatively wider and general application to genomes. Therefore, the insights and perspectives here are addressed to the larger community, inclusive of clinicians and non-clinicians.

Genome sequence evolution using computer-based comparative genomics approaches

The DNA damage repair genes

All eukaryotic living cells experience DNA breaks in their nucleus due to a plethora of endogenous and exogenous agents, and have evolved intrinsic molecular mechanisms to repair those lesions [58]. If left unrepaired, the physiological consequences may be lethal for the cell [59]. Such mechanisms for repair constitute a network of bio-chemical pathways called DNA damage response (DDR) pathways [60], the regulation of which is largely evolutionarily conserved [61]. For example, in the instance of metabolic syndrome, it has been suggested that system-level DDR may impinge on energy metabolism and vascular physiology [60]. Moreover, when the genes that encode human DDR proteins undergo mutations, the inherent genome-wide repair mechanism may breakdown, which may lead to ischemic stress [62], CVD [63, 64], or certain cancers [65-67]. Over a hundred different genes encoding the various human DDR proteins [68, 69], and belonging to disparate DDR pathways [58, 70], are publicly available at: http://sciencepark.mdanderson.org/labs/wood/dna_repair_ genes.html .

Evolutionarily related proteins (or the genes encoding them) are referred to as homologs. The homologs within the same organism, which are a consequence of gene sequences that coalesce [71] at specific within-genome replication event (such as a gene duplication event) are termed paralogs. On the other hand, when gene sequences coalesce [71] at a speciation event they give rise to orthologs. The mathematical formulations of this coalescent theory have been extensively implemented (for example [72]) and covered extensively (for example [73]). Genes encoding orthologs and paralogs have largely evolved by gene duplication events, as well as other evolutionary

processes such as retroposition, intron-mediated recombination (exon shuffling and trans-splicing), horizontal gene transfer and de novo organization (reviewed in [74]). For this comparative genomics approach, an accurate multiple sequence alignment (MSA) is critical for drawing an accurate inference [75]. The MSA is primarily a mathematically derived arrangement of nucleotide or amino acid (AA) sequence (or may merely represent domains / motifs / fragments of the full sequence), of three or more homologs representing their overall ancestry.

Homologous proteins have evolved among diverse species, and the ones that are critical for multiple functions, may largely remain conserved at the amino acid residue (AA) level, and under strong negative selection [76-79]. As most DDR proteins are ubiquitous and play a relatively central role in cellular processes within an organism, it is plausible that they are largely under strong selection pressure and evolutionarily conserved [78, 80]. However, random mutations can give rise to variants at the codon-level, which in may turn out advantageous and may get incorporated or “fixed” in a population, or subpopulation, by natural selection [81]. From among all possible mutations, the ones that are advantageous to the organism’s fitness, and enable it to adapt and thrive in its habitat will be discussed later. These genomic adaptations are not exclusively single nucleotide variants (SNVs), and include instances when molecular adaptions in host organisms were geared to avert insertion of mobile elements, such as transposons, from parasite genomes [82, 83]. However, before a rare mutation gets “fixed” in the population, they first manifest as polymorphisms in subpopulations, and can have interesting physiological consequences [84]. To showcase the importance of genetic invariants and common variants during genome evolution, we synopsize some select reports of molecular adaptations of among various DDR genes, across different species.

To predict the AAs that are important for the protein structure and/or function, computational methods are used to delineate evolutionarily conserved AA positions in a MSA (columns of AAs). DNA repair proteins from *E. coli* and *S. cerevisiae* have been compared to protein sequences encoded in completely sequenced genomes of disparate bacteria, archaea and eukaryotes [85]. In that study, the authors identified conserved domains in DDR proteins across kingdoms and highlighted that horizontal gene transfer and epigenetic regulation influenced their sequence composition [85]. Subsequently, independent studies have identified homologous DDR proteins in disparate species, such as *E. coli*, *S. cervevisiae*, *M. musculus* and *H. sapiens* [61, 86, 87]. Moreover, it is noteworthy that analogous roles of orthologous proteins (such as RecA/*RAD51*) have been traced across eukaryotes: in bacteria, yeast and mammals [61, 87]. As another example, nearly two decades ago, two independent comparative sequence analysis of yeast DDR protein with its orthologs, and *E. coli* DDR protein with its orthologs have led to discoveries of novel orthologous human DDR proteins [88, 89].

An important component of comparative sequence analysis is the *in silico* modeling of the AA sequence evolution. Codon-based models of gene evolution, such as Phylogenetic Analysis by Maximum Likelihood (PAML) [90] and Selecton [91] (see Box 1), use MSAs of closely related nucleotide sequences to compute the evolutionary selection pressure at the codon-level, and differentiate positive selection with respect to negative selection and neutral evolution.

Box 1

Special codon-based models of gene evolution, such as modules/suites available from PAML [90] and Selecton [91], have been developed to simulate single nucleotide substitutions among orthologs and paralogs. Among the various modules implemented, the one most relevant to our report is the one that simulates SNVs to compute the evolutionary selection pressure at the codon-level for a given MSA of homologs. "Silent" mutations, which are synonymous nucleotide substitutions do not change the translated AA sequence so their substitution rate d_s (also called *Ks*) is not subject to a selection pressure on the expressed protein [80, 92]. Such mutations are more prevalent in an organism compared to the nonsynonymous ones. The nonsynonymous substitutions alter the amino acid sequence and their substitution rate d_N (also called K_A) is a function of the selective pressure on the protein [80, 92]. A measure of the selection pressure at the codon level is subsequently obtained from their ratio $\frac{d_N}{d_S}$ (also called $\frac{K_A}{K_S}$), also referred to as ω [80, 92]. To ensure an unbiased estimate of this ratio, an average of one nonsynonymous substitution may be faithfully represented at the codon level (numerator $d_N \sim 1$). If relatively distant sequences are compared, the computation may be biased, as there is ambiguity with regard to whether only a unique single nucleotide substitution, or multiple such substitutions were accrued at the codon level. Hence, if we use distant orthologs, there is a likelihood that $d_N > 1$, and ω may not be faithfully represented [93]. Caveat: an estimate of d_N may be computed at first, before computing ω, using the same suite of software packages [90, 91].

(…Continue main text.)

Differential changes in the AA protein sequence across diverse species.

Computational sequence analysis has made it feasible to identify homologous protein sequence, and subsequently narrow down the evolutionary adaptations (or variations) in those genes. Therefore, one may use a candidate gene-based approach to study select proteins (and their homologs) along with their interacting proteins (dubbed interactors). Here, we have highlighted one such DDR protein, p53, which has been extensively studied, to provide a wider perspective (See Box 2), but in theory any set of proteins may be analogously showcased.

Box 2

P53 as a representative example for comparative study across species

The human tumor suppressor gene (*TP53*) (an oncogene) encodes for the tumor suppressor protein 53, referred to as P53. To showcase its multiple roles in fundamental cellular functions, it is referred to as the "Guardian of the Genome" by David Lane [94], and as the "Cellular Gatekeeper" by Arnold J. Levine [95]. This protein and its homologs have been extensively studied in regard to critical cellular processes such as cell cycle and DDR pathway [96]. Therefore, we synopsized molecular adaptations of P53 homologs and a select subset of its interacting proteins (subsequently referred to as a p53 sub-interactome).

A primer on p53 structural domains

A schematic of the various P53 functional domains is represented as a 1D cartoon in Fig. 2 (Adapted from [97]). The N-terminal contains a short intrinsically disordered region, following that is the transactivation domain (TAD), which is further divided into two subdomains (TAD1 and TAD2) that bind to negative regulators, transcriptional coactivators and components of transcriptional machinery [98-101]. Subsequent to that

is a Proline-rich region (PRR) that favors a polyproline II helix structure and serves as a rigid linker to project the TAD distal to the DNA binding domain (DBD) [102]. In contrast to the TAD, DBD is more conserved across disparate vertebrates. Following the DBD is the C-terminal domain (CTD), comprising the tetramerization domain (TET) and the extreme C-terminus (CT).

The p53 sub-interactome

Coordinated AA interactions are imperative for protein function such as folding, unfolding, and binding to other identical or non-identical proteins to form complexes. For example, it has been demonstrated that the stability of the p53 tetramer is enhanced due to specific AA interactions of the p53 DBD with the p53 N-terminal [97]. The AA critical for such essential function may manifest as an invariant AA across various orthologs from disparate species and resist mutations. Such a conserved position (or column) usually stands out when viewed in a fragmented MSA (or full-length MSA of the protein).

End Box 2.

(…Continue main text.)

Using a candidate gene approach, we considered two molecular interactors of p53: MDM2 and MDMX (aliased MDM4), the genes of which have also been reported as oncogenes, being expressed in some types of cancers [103]. Murine p53 protein interacts with Mdm2 (murine ortholog of MDM2) via its N-terminal as well as core domain [104]. Moreover, it is reported that Mdm2 binds directly to the p53 TAD1 domain to inhibit p53 interactions with its transcriptional coactivators [105] and that their binding is differentially decreased when the murine p53 CTD is deleted, mutated or even acetylated [106]. In addition, the MDMX protein is an important negative regulator of p53 in normal cells. Bista *et al.* [107] have shown that the full-length MDMX contains a sequence of regulatory element (the “WWW element”) that binds to its own N-terminal

domain and therefore blocks the binding of MDMX to p53, and Su *et al.* have shown that there exists a motif 'FXXLW' motif, which is the MDM2 functional motif, in the disordered region of p53 [108]. Such intricate molecular adaptations of these interacting proteins are critical as an inhibition introduces a level of functional regulation making them ideal candidates for study of a p53 sub-interactome with regard to their differential levels in function and protein sequence evolution. On the other hand, the p53 C-terminal binds to RAD51 and RAD54 and additional details of specific interactions are described in Ref [109, 110].

We now synopsize studies that have identified genetic variations as molecular adaptations among orthologs genes of p53 and its sub-interactome. Comparative study of orthologous genes from mammals such as primates, bats and rodents are used here to illustrate how computational biology can help identify codons under positive selection, following which experiments were used to validate those adaptations (A synopsis is presented in Box 3).

Box 3

Synopsis of species-specific adaptations in DDR proteins

We reiterate that the mechanism to repair endogenous and exogenous DNA damages have evolved across disparate species for their survival and proliferation. To maintain a robust DNA-repair pathway, it is plausible that a large fraction of genes are conserved but only a small subset of genes may be under positive selection pressure [78, 80]. Positive selection at two or more unique codons within the same gene and across genes may suggest improvement or modulation of a pre-existing function [111], for example pain-related stimuli in humans [112]. Mutations that tentatively enhance the organism's

fitness are referred to as molecular adaptations and the codons within the same gene, or across different genes, are set to have co-evolved (reviewed in [113]). In this Box, we highlight some preselected reports that have used comparative genomics to identify specific molecular adaptations among DDR proteins across species. We sought out how have the DDR genes had adapted across those diverse species that inhabited harsh and diverse environments, which were not conducive for most mammals. For example, amongst mammals, (i) bats are the only species with flight as their mode of transport, on the other hand; (ii) blind mole rats live in subterranean conditions in extremely cold and hypoxic conditions, where sunlight may not permeate.

1 – Bats

The bat is the only mammal with an innate ability for sustained flight, and therefore a comparative genomic approach with respect to other diverse mammals helps uncover the putative molecular adaptations in select genes. It has been reported that nearly 23% and 5% of bat specific mitochondrial genes and nuclear encoded oxidative phosphorylation (OXPHOS) genes are under significant positive selection respectively [114]. This prompted further investigations to test the hypothesis that positive selection of the genes facilitated in metabolic and energy needs for a sustained flight [114]. Their physical act of flight is perhaps one of the most energy consuming physical activity and therefore it is was suggested that its genes involving energy metabolism and DDR may have evolved unique molecular adaptations, which are absent in all other mammals [115]. Next, to investigate the molecular sequence evolution of the various bat genomes, a comparative genomic approach was used to study genomes from two bat species: *Myotis davidii* (*M. davidii*) and *Pteropus alecto* (*P. alecto*), and compare them with diverse mammals including *Homo sapiens* [115]. In that study, Zhang *et al*. delineated genetic variants from selected mammals for p53 functional domains, and other DDR

genes such as *MDM2*, *ATM*, *RAD50* and *KU80* [115]. Here, in Fig. 3 (adapted from [115]), we showcase a fragment of p53 MSA, from disparate mammals including *M. davidii* and *P. alecto.* Each of the conserved AA at the MSA position suggests zero tolerance to nonsynonymous substitutions, and the rare plasticity of non-conserved positions revealed tolerance to genetic variations, suggesting putative candidates for molecular adaptations. In particular, the authors identified the K321M non-synonymous mutation as the *TP53* molecular adaptation in *M. davidii* and *P. alecto*, but the L323A mutation as an exclusive adaptation in the *M. davidii* sequence [115]. In gene-based studies, they established that the *p53* and *BRCA2* orthologs of were under positive selection pressure in *M. davidii* and that the *LIG4* ortholog was under positive selection pressure in *P. alecto* [115]. Therefore, it is plausible that the molecular adaptations of *M. davidii* and *P. alecto* occurred prior to the divergence from their last common ancestor, within the bat lineage, and the unique species specific adaptations (for example L323A in *TP53* from *M. davidii* and the adaptation in *LIG4* ortholog from *P. alecto*) occurred subsequent to that divergence [115]. In addition, Zhang *et al*. also reported that the human orthologs of *ATM*, *RAD50*, *KU80* and *MDM2* genes are under positive selection pressure, amply suggesting that molecular adaptations within a species usually brings about a concerted genome evolution [115]. We also reiterate that a mutation that impacts fitness (positive or negative) can be fixed by drift as well as selection [22]. A caveat is in order here that independent experimental assays performed in that study, suggest that those putative variants were a result of species adaptation and that the previously mentioned theoretical studies provided insights for an initial working hypothesis [115]. Moreover, experimental assays usually measure fitness consequence in terms of survival probability, and usually not in a complete physical and genomic environment “experienced” by the ancestor. Hence, in most instances, it is generally not possible to experimentally “confirm” a molecular adaptation.

Molecular adaptations among rodent species are synopsized in modules (2) and (3) below, and contrasted with other mammals.

2 – Zokor rats

Zokor rats, blind subterranean rats, root rats and bamboo rats are spalacid rodents that are collectively referred to as mole-rats. These rodents thrive under adverse geographical conditions, for example the Tibetan plateau is a natural habitat for zokor rodents [116]. Recently, two subterranean wild zokor species (highland-dwelling *Myospalax baileyi* and lowland-dwelling *Myospalax cansus*), and one highland-dwelling aboveground species of root vole (*Microtus oeconomus*) were investigated to explore the molecular adaptations in the p53 and contrast them with functional adaptations related to the harsh natural habitat and environmental stress [116].

As previously mentioned, a part of the p53 N-terminal is intrinsically disordered [104], and its MSA with other orthologs represents a high degree of sequence non-conservation [116]. Zhao *et al.* first used cues from protein sequence comparison and computational biology to narrow down the search for putative molecular adaptations at the codon-level in the p53 DBD, which was subsequently cross-checked using site-specific mutagenesis assay [116]. The DBD is relatively more conserved compared to the TAD and CTD and an unambiguous MSA from DBD was first obtained across orthologs, from humans to many different rodents [116]. For the purpose of this synopsis, we highlight that the sequence conservation of domain is disrupted by a solitary non-conserved MSA position, at codon 104 in human p53 DBD. Nonsynonymous variants of p53 at codon 104 were identified in the three rodent species [116]. *P53* orthologs in primates such as *Homo sapiens*, *Macaca fuscata, Macaca*

muleta have a Serine (S) residue at codon 104. However, *Myospalax baileyi* and other mammals such as *Ovis aries* (sheep), *Bos taurus* (cow), *Rattus norvegicus* (rat) and *Mus musculus* (mouse) have Asparagine at that corresponding homologous codon [116]. In addition, the authors reported that for *Myospalax baileyi*, the 104N variant is responsible for the transactivation of apoptotic genes at extremely severe environmental conditions: hypoxia, hypercapnia (acidic stress, high CO_2) and severely low ambient temperatures [116]. Subsequently, they have also reported that the 104E nonsynonymous variant in p53 for *Microtus oeconomus* suppresses apoptotic gene reactivation and cell apoptosis [116].

3 – *Spalax* rodent

Spalax a blind subterranean mole-rat belongs to the Spalacidae family. As shown in Fig. 4, the Arginine residue is conserved across p53 sequences in humans, primates, quadrupeds and numerous rodent species [117]. Comparative sequence analysis of the p53 DBD sequence among various rodent species revealed that the *Spalax*, had acquired the R174K mutation [117]. However, in addition to *Spalax*, the R174K mutation is identified across other species such as *Xenopus laevis*, *Monodelphis domestica*, *Xiphophorus maculatus*, and *Xiphophorus helleri* (Fig. 4, adapted from [117]). The above nonsynonymous substitution is one of the few mutations that was adapted within the *Spalax* genome, and the most interesting feature is that this particular nonsynonymous substitution occurs at an otherwise well conserved position in the MSA. Additional Spalax-specific changes, with respect to human and mice, were also reported using a full-length MSA [117]. All findings motivated the authors to hypothesize that Spalax *p53* gene had adapted, and enabled those rodents to thrive in an underground hypoxic environment. Subsequently, other critical adaptations related to hypoxia were also reported in *Spalax* [118]. (More recently, unique hypoxia-based adaptations have also

been reported in the case of the naked-mole rat, another member of the same Spalacidae family [119]).

We try to present a broader perspective here, that genome-level molecular adaptations have led to the survival and proliferation of a variety of species in harsh, stressful and disparate environments.

(End Box 3)

(…Continue main text.)

Next, to emphasize further that genomes have evolved uniquely across disparate species, we further synopsize some results from DDR-related pathways (See Box 4).

Box 4

Adaptations in DDR pathways

Here, using comparative genomic approaches, we highlight recent reports involving various DDR pathways, and encompassing multiple genes. We highlight recent results from the non-homologous end joining (NHEJ) repair pathway and then draw parallels invoking the homologous recombination (HR) repair pathway in bacteria and mammals. Sequence-based computational biology analysis was used to provide insights and perspectives to predict and rationalize adaptive evolution amongst various interacting proteins. First, we synopsize two independent studies from the NHEJ pathway in this context of an adaptive molecular sequence evolution.

An evolutionary screen of proteins from the yeast genome (*Saccharomyces cerevisiae* and *Saccharomyces paradoxus*) identified various NHEJ proteins to be under adaptive evolution [82]. In that study, the authors identified a total of seventy-two positively

selected genes from the yeast genome and subsequently focused on pathway-centered analysis where they identified statistically significant positive selection from NHEJ repair pathway genes and genes such as *XRS2*, *POL4*, *SAE2* and *NEJ1* [82]. Moreover, they also demonstrated that the identified molecular adaptations are not a result of *Saccharomyces cerevisiae* adapting to laboratory conditions as they obtained similar results by using YJM789, the "wild" *Saccharomyces cerevisiae* isolate [82]. The rapid evolution of those proteins seemed counterintuitive because NHEJ response to DDR is a critical repair response among eukaryotes. In support of their counterintuitive results, the authors hypothesized that those NHEJ genes were under positive selection to counter the integration of Ty LTR-retrotransposons in their genome [82]. In a nutshell, that study primarily identified NHEJ genes manifesting molecular adaptations, in yeast, which were essential for its survival [82].

In an independent comparative genomic study of proteins from the human NHEJ pathway, the authors have delineated human DDR genes under positive selective pressure [120]. Though the critical functions of DNA repair and checkpoint signaling were largely conserved, they identified signatures of positive selection in five NHEJ genes (*XRCC4, NBS1, Artemis, POLλ and CtIP*), during the evolution from simian primates to humans. Interestingly, only one out of nearly seven hundred and fifty codons (<0.5%) of the *NBS1* gene was under positive selection, yet on the other hand they reported *POLλ* having ~5% codons under positive selection, implying that a complicated systems-level genetic underpinnings has been the basis for such diverse quantum of molecular adaptations. In that study, the authors hypothesized that positive selection at select codons might be responsible in providing an evolutionary advantage to the human

genome (host genome) from parasitic viral genomes, which was critical for human health and survival [120].

Next, a separate study of human genes from Fanconi anemia *BRCA* (FA/BRCA) pathway is synopsized [121]. That theoretical study investigated a select set of genes: *ATM*, *BRCA1*, *BRCA2*, *CHK2*, *NBS1*, *RAD51*, *FANCA*, *FANCB*, *FANCC*, *FANCD2*, *FANCE*, *FANCF*, *FANCG*, *FANCL* and *FANCM*, focusing on a pathway-centric approach to identify positive selection in vertebrates [121], about which we suggest an alternate hierarchical analysis that first considered clades of closely related eukaryotes to ensure that on an average $d_N \sim 1$ at the codon level. Using distantly related orthologs, from humans to fish, the author concluded positive selection at specific codons in most of those fifteen genes but negative selection pressure for *RAD51*, *FANCA* and *FANCG* genes [121].

The divergence of humans from early vertebrates, such as fish, may be traced back to hundreds of million years ago. Ideally, models of codon evolution have been optimized for closely related orthologs, where on the average a single nonsynonymous substitution may be hypothetically anticipated per codon. If judiciously implemented, theoretical approaches from such comparative sequence analysis help identify putative molecular adaptations in a species-specific context, which may subsequently be validated using experimental assays. Using this extensive background, we shall now proceed to appreciate how the genome-level sequence evolution, in the context of human evolution, may help trace the genetic etiology of lifestyle-related complex multifactorial diseases, such as CVD.

End Box 4

(….continue Main text)

Comparative genomic study can determine a unique hierarchy of common genetic variants from patients with stratified levels of disease severity

Modern genotyping platforms, such as microarrays, enable us to systematically probe the inherited genotype at unique and predetermined SNP loci (for example, the Cardio-MetaboChip from Illumina [122]). An angiography-based study of every CVD patient enabled us to designate a SYNTAX-based score for them and helped us stratify their levels of CVD severity [56, 57] (Fig. 5A). For this comparative genomic analysis, we studied the following cohorts of human subjects: (i) all cases and controls (S_{TOT}), (ii) exclusively controls (S_{CTRL}), (iii) exclusively CVD patients (S_{CVD}), (iv) most severe CVD patients (S_{sevCVD_1}), (v) CVD patients with intermediate severity (S_{sevCVD_2}), and (vi) CVD patients with least severity (S_{sevCVD_3}) (Fig. 5A), and probed the genotype of all individuals at SNPs genome-wide. Next, we performed an iterative computational analysis (Fig. 5B), using the genotype from the designated cohorts (Fig. 5A). We first computed the Shannon entropy (S) [123] (See Box 5) at every SNP locus that was probed by our microarray. Furthermore, for quality control purposes, we omitted genotype information recorded by faulty probe(s). After ensuring $S = 0$ for that locus, we also computed the zygosity at that locus (Fig. 5B). Independent instances of autosomal recessive [124] as well as dominant mutations [125, 126] have been traced to familial hypercholesterolemia, which in turn associates with CVD. Hence, we sought to compute both homozygous and heterozygous loci that were exclusively invariant among controls and CVD patients.

Box 5.

Computing the Shannon entropy at the SNP locus

Modern microarray platforms, such as the Cardio-MetaboChip [122], probe the genotypic at predetermined genomic coordinates, which are identical across all subjects – cases and controls. Hence, the genotype data analyzed was analogous to a MSA and ready for a comparative genomic analysis. However, neighboring nucleotides were megabase apart and not adjacent nucleotides from a complete genome. The different genotype probed at a locus, per subject may be annotated in 11 ways. We considered four homozygous calls (AA, TT, GG, CC), six heterozygous calls (AT, AG, AC, TG, TC, GC) and "--" as "bad call" for incorrect / poor quality genomic information that was probed therefore, $N = 11$. Then, for a given study cohort, the Shannon entropy (S) at every SNP locus is defined as [123]:

$$S = \sum_{(genotype)\ i=1}^{N=11} -log_2\ p_i$$

Here, p_i is the probability to find the ith genotype (out of a total of $N = 11$) at a given locus from an MSA column. The conserved or invariant locus has $p_i = 1$ for a specific genotype but zero for all other genotypes, hence $\log_2 p_i = 0$, and therefore Shannon entropy $S = 0$. Once invariant loci are computed, their genotype information is used to identify zygosity at that locus (Fig. 5B).

End Box 5

(…Continue main text.)

CVD cases and corresponding gender- and age-matched controls have a shared ancestry

Although it is common knowledge that the CVD cases and controls from an outbred native subpopulation have a shared ancestry, nevertheless we illustrate this by counting the number of unique invariant genotype for all SNP loci. If individuals have diverged around the same time, from a common ancestor, then the numbers of evolutionarily conserved SNP loci per chromosome may be at par among CVD cases and controls. Such an exercise is motivated because an early onset of CVD is reversible [16] and the human subjects that we studied were all native to India. For a gender- and age-matched case versus control study, the genome-wide identification of statistically consistent number of distinctly invariant loci, on identical chromosomes, may suggest similar evolutionary divergence time from their shared ancestry [36]. We did not do this analysis at the gene-level because most invariant loci were intergenic (unpublished results). Hence, we sought to identify and annotate all invariant loci from the mutually exclusive studies of S_{CTRL} and S_{CVD} respectively. We computed S for all SNP loci genome-wide, from a 3-way comparative study, using those two mutually exclusive genomic data sets: $S_{CTRL} - S_{TOT}$ and $S_{CVD} - S_{TOT}$ (Fig. 6A). Here, the subtracted part S_{TOT} implies that all SNP loci that were computed with invariant genotype from the S_{TOT} study were omitted from those in S_{CTRL} and S_{CVD}.

Here, it is our goal to show that the numbers of exclusively invariant homozygous loci (apportioned per chromosome), and exclusive to either the CVD or control cohorts are at par. Our proposed null hypothesis is that the distribution of all invariant loci from the two mutually exclusive cohorts is statistically consistent. The alternate hypothesis proposed is that the two distributions are statistically inconsistent, suggesting that the two cohorts did not diverge around the same time, from an ancestral population. Using the

nonparametric Kruskal-Wallis test to compute if the likelihood of the two distributions was statistically consistent, we report that the null hypothesis is tenable ($P < 0.23$). In addition, we also computed the ratio of the total number of invariant loci, apportioned per chromosome, from CVD patients with respect to control subjects. We report that the ratio of numbers of invariant loci from the CVD to control cohort is at par (Fig. 6B). On representing the same ratio in a histogram, we report that the histogram mean ≈ 1 (Fig. 6C), which confirms that the number of invariant loci from the cases and controls are at par. On repeating the analysis with invariant loci exclusively from the protein-coding genome, we were able to cross check those results and independently computed another histogram with mean ≈ 1, but with an expectedly greater variance because there were fewer invariant loci (results not shown). Taken together, our results support the fact that the individual CVD cases and controls, have diverged from a common ancestral subpopulation and therefore we may attempt to trace the common evolutionary etiology of CVD by tagging loci that were invariant in a hierarchical manner – for example distinct loci were manifest invariant in the combined cohort, among the controls, and among CVD cases.

3-way comparative genomic study of invariant homozygous loci

Next, using the previously computed loci, we continued to seek additional insights and perspectives from the study of invariant homozygous loci, which were computed in three different ways: from all individuals, only CVD cases and only controls. To identify and annotate all genes with at least one invariant homozygous locus, which was distinct from the CVD versus control cohort, we did a 3-way comparative genomic analysis of S_{TOT}, S_{CTRL} and S_{CVD} cohorts. Using the Venn diagram in Fig. 6A, we represented seven distinct sectors *V1-V7*, each representing the genes or intergenic regions that had such invariant loci from S_{TOT}, S_{CTRL}, and S_{CVD}. We denote those studies as S_{V1}-S_{V7} and will

only focus on the S_{V1} study to identify genes with the three different types of variants that manifest uniquely in these cohorts. More specifically, we sought genes that represented the sector *V1* from the 3-way overlap (Fig. 6A), which may be represented by hypothetical genes *GENE_G1* and *GENE_G3* (Fig. 7). From this analysis we were able to identify *LDLR* along with other genes (unpublished results). As mentioned earlier, distinct *LDLR* variants are implicated in lipid regulation and metabolism [41], which manifested gain-of-function and loss-of-function, and hence our formalism warrants further scrutiny using this evolutionary-based perspective.

Exclusively conserved loci in patients with most severe CVD (*S_{sevCVD_1}*study)

To further the genetic etiology of CVD, we sought out the common variants that were not conserved in CVD cohort, and with a prototypical hierarchy that represented the paradigm exclusively from the hypothetical *GENE_G3* in Fig. 7. As mentioned before, angiography-based studies (SYNTAX-based scoring [56, 57]) were used to stratify the severity in CVD patients. The patients grouped as most severe (S_{sevCVD_1}) had multiple manifestations of severe arterial blockages (unreported results). Therefore, to characterize the genetic etiology of severe CVD cases, using the workflow described earlier (Fig. 5B), we sought to compute those invariant homozygous and heterozygous SNP loci that were exclusive only to that cohort (S_{sevCVD_1}).

This analysis is analogous to seeking invariant genotype in a MSA, where a few aligned sequences from a set of homologs will yield greater conserved positions, and as more and more homologs are added, the number of conserved positions will recede, or at best remain constant. Hence, we anticipate that the study with most severe CVD patients will result is a larger set of invariant loci than the combined study of patients with least severity, intermediate severity and highest severity. As the cohort size for severe CVD is

smaller compared to the full CVD cohort, a genome-wide search for loci with exclusively invariant genotype among them is likely to yield more false positive loci than true positive ones. Therefore, we take a three-tire approach (described next) to sequentially filter out as many false positives as possible, and only compute the homozygous and heterozygous invariant loci that were exclusive to S_{sevCVD_1}.

***An iterative and hierarchical strategy to identify unique common variants from the cohort of most severe CVD cases* (S_{sevCVD_1})**

Using S_{sevCVD_1}, we compute the S for all SNP loci, and delineated the homozygous and heterozygous loci with $S = 0$. In the first tire of analysis, we omitted all invariant loci that were already identified in the S_{TOT} study cohort. Next, in the second tire, using all remaining loci, we subsequently omitted the homozygous / heterozygous invariant loci from the CVD patient study cohort S_{CVD}. Finally, in the final third tire, we combined cohorts of intermediate (S_{sevCVD_2}) and lowest CVD severity (S_{sevCVD_3}) to progressively narrow down our search for invariant SNP loci that were exclusively invariant among the most severe CVD subjects and not otherwise.

In a nutshell, on iterating the analysis outlined here, for different subject cohorts: both case and controls (S_{TOT}), only controls (S_{CTRL}), only CVD patients (S_{CVD}), most severe CVD patients (S_{sevCVD_1}), CVD patients with intermediate severity (S_{sevCVD_2}) and CVD patients with least severity (S_{sevCVD_3}), and filtering out SNV loci obtained from any cohort at higher hierarchy, we will systematically gain access to the genetically invariant loci for cohorts such as S_{sevCVD_1}, S_{sevCVD_2} and S_{sevCVD_3}.Therefore, using a 3-way and three-tire severity-derived CVD data, a robust comparative genomics study is proposed to understand why common genetic variants persist as invariant loci only in very specific

cohorts, such as the most severe CVD cohort, and why the same genes manifest markedly different and hierarchical variants.

Moreover, after classifying CVD patients by their levels of severity, a comparative genomic analysis across the severe category will surely enable us to identify the invariant loci that are unique among identical set of genes or intergenic regions (*GENE_G3* in Fig. 7). We report that such invariant loci are largely inter-genic, with consecutive invariant loci span genomic distances in megabase scale. However, it is beyond our scope to determine if those variants represent a gain-of-function / loss-of-function paradigm [48, 49], or are benign [127].

Genes computed using this 3-way comparative genomic study represent putative candidates for multifactorial CVD severity. Therefore, we hypothesize here that genes that manifest evolutionary plasticity from: (i) the 3-way analysis study of S_{V1}, (ii) with exclusively invariant homozygous and/or (iii) heterozygous loci in the severe CVD cohorts, may be used to trace the hierarchical etiology of this complex multifactorial disease, either directly modulating physiological function or via its interactors, so as to reverse the disease prognosis [16]. We also emphasize that our results are along the lines of the seminal discovery involving evolutionary plasticity of the *LDLR* gene, wherein it was shown that unique variants of the gene gave rise to both loss-of-function and gain-of-function [41]. Hence, more importantly a candidate-gene based search strategy may putatively enable us to identify common genetic variants having unique evolutionary plasticity that is manifest in diseases with differential severity.

Conclusion

In this perspective, we have presented alternate perspectives and insights to trace the etiology of CVD from a very small study set (50% CVD cases and 50% controls) representing an outbred subpopulation from India. We show that it is possible to identify genetic variants from CVD patients stratified by disease severity. The hierarchical plasticity of those variants led us to identify a constellation of genetic variants unique to CVD cases that were stratified by differential levels of disease severity. Consequently, interactions among various proteins that originate from variant genes may give rise to a differential response. Taken together, this work opens up avenues for inter-disciplinary analyses that may complement genome wide association studies.

As a primer to computational molecular evolution, we have synopsized that genomes have evolved uniquely in among disparate species, from independently published reports. Accordingly, we highlighted instances of diverse molecular adaptations among interacting proteins across species. Novel insights and perspectives developed here are for motivating a paradigm shift and for wider education that may lead to a better understanding of disease paradigm using an evolution-based approach.

Outlook

The perspective presented is not restricted to CVD or lifestyle-related diseases and may be explored in the context of diverse diseases, for example autism where in a rational basis for stratification of disease severity exists.

Acknowledgements

SNF would like to acknowledge Tata Institute of Fundamental Research (TIFR)-DAE, Government of India and Professor BJ Rao of TIFR, Mumbai for financial support.

Doctor TF Ashavaid, P.D. Hinduja Hospital and Medical Resarch Centre, Mumbai for sharing a microarray-based genomic data from 149 CVD cases and 149 controls.

Table

Table 1. The paralogs of *RAD51* human DNA repair genes and their orthologs from bacteria (*recA* genes) and yeast are illustrated. (Adapted from Reference [86]).

Homolog category	**Bacteria**[@]	**Budding yeast** [$]	**Fission yeast** [π]	**Mammals** [§]
Orthologs	*recA*	*RAD51*	*rhp51*$^{+}$	*RAD51*
Paralogs		*RAD55*	*rhp55*$^{+}$	*RAD51L1/B/REC2/R51H2*
		RAD57	*rhp57*$^{+}$	*RAD51L2/C/R51H4*
		DMC1	*rlp1*$^{+}$	*RAD51L3/D/R51H3/TRAD*
			dmc1^{+}	*XRCC2*
				XRCC3
				DMC1/LIM15

[@] *Escherichia coli*; [$] *Saccharomyces cerevisiae*; [π] *Schizosaccharomyces pombe*; [§] *Homo sapiens* and *Mus musculus*.

Figure Legends

Figure 1. A schema reveals the trend for volume of data needed to characterize the genetic etiology of a multigenic disease. Shaded hypothetical clouds illustrate a continuum of values of allele frequency, which may range from the low end of the spectrum (lighter shade of triangle) to the high end of the spectrum (darker shade of triangle). The region of interest probed from low volume data studies may only pertain to identification of common variants, which are considered causative for relatively low penetrance (bottom right corner, or middle right region of graph). GWAS = genome wide association study.

Figure 2. The functional domains from the p53 protein sequence are described. The human p53 molecule comprises of 393 amino acid residues, whose positions ID are

sequentially illustrated on the top and below the schematic. The N-terminal comprises of a transactivation domain (TAD) and proline-rich region (PRR). It is followed by the DNA binding domain (DBD), and subsequently followed by the C-terminal. (Adapted from Reference [97]). **(Inset for Box 1.)**

Figure 3. **Representative MSA of p53 and MDM2 orthologs showing the different AA variants unique to bats.** These AA variants were detected in the functionally relevant regions of the p53 nuclear localization signal and MDM2 nuclear export signal (shaded). The p53 MSA (shown on left) is representative MSA from nuclear localization signal. MDM2 MSA (shown on right) is representative for nuclear export signal. The dot connotes AA identical to the one in the first row of the fragmented MSA (Adapted from [115]). **(Inset for Box 2)**

Figure 4. Comparing the p53 DNA binding domain across species. A fragment MSA describing the DNA binding domain, across forty four disparate p53 orthologs, is represented to illustrate the conserved (invariant) amino acid residues and the variants. (Adapted from Reference [117]). **(Inset for Box 2)**

Figure 5. The outline of our bioinformatics analysis used to identify the hierarchy of genetic variants in subjects with disparate levels of disease severity. A. The hierarchical organization of data from human subjects used to demonstrate our working hypothesis. **B.** The flowchart describing the iterative comparative genomic analysis. The workflow to describe the computation and quality control steps for identification of invariant genotype across the human genome, for over one million SNP loci. At different stages of analysis, the input dataset for this procedure is different, for example genotypic dataset from (i) case and controls (S_{TOT}), (ii) only cases (S_{CVD}) (iii) only controls (S_{CTRL}),

(iv) only severe CVD cases (S_{sevCAD_1}), (v) only CVD cases with intermediate severity (S_{sevCAD_2}), (vi) only CVD cases with low severity (S_{sevCAD_3}).

Figure 6. Comparative genomic study demonstrates similar evolutionary divergence times of CVD cases versus age- and gender-matched controls. (A) A Venn diagram that represents the study plan for a 3-way comparative genomics approach. **(B)** A bar graph describes the ratio of total number of exclusively invariant loci, apportioned per chromosome, obtained from the CAD cohort with respect to the controls, is illustrated here. All loci that were identical to both cohorts were subtracted first. **(C)** The same ratio from (B) is now described using a histogram, which has been fitted to a Gaussian function (normal distribution).

Figure 7. Demonstrating a rational for a hierarchical comparative genomic analysis using hypothetical DNA sequences obtained from cases and controls. Hypothetical DNA sequences obtained from cases and controls for a specific genomic region of interest are illustrated as points. The dotted lines indicate the presence of a string of bases representing the genotype (of a person) at various SNP loci (for example, genome-wide loci probed by a microarray chip). The black rectangles below the sequences reflect hypothetical functional elements (UTRs, exons interspersed by introns) at the positions indicated in the sequence above. A vertical line (grey lines, black arrows with single and double arrowhead) represents the invariant bases (completely conserved genotype) at the given genomic locus. The grey vertical line represents a locus that has invariant genotype in all cases and controls, and with a single arrowhead (double arrowhead) represents exclusively conserved genotype from instances of cardiovascular disease CVD cases (age- and gender-matched controls). A black line with two dots represents systematic and complete genotype conservation at that locus,

which exclusively only to severe CVD cases, but not in the remaining cases or controls. The CVD severity among cases is stratified in three tires: (1) highest severity, (2) intermediate severity, and (3) lowest severity, as shown on the right margin.

References

1. Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, et al. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature. 2016;530(7591):429-33. Epub 2016/02/18. doi: 10.1038/nature16544. PubMed PMID: 26886800; PubMed Central PMCID: PMCPMC4933530.
2. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338(6104):222-6. Epub 2012/09/01. doi: 10.1126/science.1224344. PubMed PMID: 22936568; PubMed Central PMCID: PMCPMC3617501.
3. Pennisi E. Human evolution. More genomes from Denisova Cave show mixing of early human groups. Science. 2013;340(6134):799. Epub 2013/05/21. doi: 10.1126/science.340.6134.799. PubMed PMID: 23687020.
4. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43-9. Epub 2013/12/20. doi: 10.1038/nature12886. PubMed PMID: 24352235; PubMed Central PMCID: PMCPMC4031459.
5. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nat Rev Genet. 2007;8(11):857-68. doi: 10.1038/nrg2187. PubMed PMID: 17943193; PubMed Central PMCID: PMCPMC2933187.
6. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-8. Epub 2007/10/19. doi: 10.1038/nature06250. PubMed PMID: 17943131; PubMed Central PMCID: PMCPMC2687721.
7. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science. 2011;334(6052):89-94. Epub 2011/08/27. doi: 10.1126/science.1209202. PubMed PMID: 21868630; PubMed Central PMCID: PMCPMC3677943.
8. Karlsson EK, Harris JB, Tabrizi S, Rahman A, Shlyakhter I, Patterson N, et al. Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. Sci Transl Med. 2013;5(192):192ra86. Epub 2013/07/05. doi: 10.1126/scitranslmed.3006338. PubMed PMID: 23825302; PubMed Central PMCID: PMCPMC4367964.
9. Perry GH, Kistler L, Kelaita MA, Sams AJ. Insights into hominin phenotypic and dietary evolution from ancient DNA sequence data. Journal of human evolution. 2015;79:55-63. Epub 2015/01/08. doi: 10.1016/j.jhevol.2014.10.018. PubMed PMID: 25563409.
10. Yang J, Jin ZB, Chen J, Huang XF, Li XM, Liang YB, et al. Genetic signatures of high-altitude adaptation in Tibetans. Proc Natl Acad Sci U S A. 2017;114(16):4189-94.

Epub 2017/04/05. doi: 10.1073/pnas.1617042114. PubMed PMID: 28373541; PubMed Central PMCID: PMCPMC5402460.
11. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-8. doi: 10.1126/science.1190371. PubMed PMID: 20595611; PubMed Central PMCID: PMCPMC3711608.
12. Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-7. Epub 2014/07/22. doi: 10.1038/nature13408. PubMed PMID: 25043035; PubMed Central PMCID: PMCPMC4134395.
13. Capellini TD, Chen H, Cao J, Doxey AC, Kiapour AM, Schoor M, et al. Ancient selection for derived alleles at a GDF5 enhancer influencing human growth and osteoarthritis risk. Nat Genet. 2017. Epub 2017/07/04. doi: 10.1038/ng.3911. PubMed PMID: 28671685.
14. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science. 2003;299(5611):1391-4. Epub 2003/03/01. doi: 10.1126/science.1081331. PubMed PMID: 12610304.
15. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science. 2003;301(5629):71-6. Epub 2003/05/31. doi: 10.1126/science.1084337. PubMed PMID: 12775844.
16. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. The New England journal of medicine. 2016;375(24):2349-58. doi: 10.1056/NEJMoa1605086. PubMed PMID: 27959714; PubMed Central PMCID: PMCPMC5338864.
17. Nesse RM, Stearns SC, Omenn GS. Medicine needs evolution. Science. 2006;311(5764):1071. Epub 2006/02/25. doi: 10.1126/science.1125956. PubMed PMID: 16497889.
18. Culotta E, Pennisi E. Breakthrough of the year: evolution in action. Science. 2005;310(5756):1878-9. Epub 2005/12/24. doi: 10.1126/science.310.5756.1878. PubMed PMID: 16373538.
19. Ding K, Kullo IJ. Evolutionary genetics of coronary heart disease. Circulation. 2009;119(3):459-67. Epub 2009/01/28. doi: 10.1161/circulationaha.108.809970. PubMed PMID: 19171868.
20. Kullo IJ, Fan X, Ding K. Genetic Risk, Lifestyle, and Coronary Artery Disease. The New England journal of medicine. 2017;376(12):1192-3. Epub 2017/03/24. doi: 10.1056/NEJMc1700362. PubMed PMID: 28332385.
21. Manolio TA. Genomewide association studies and assessment of the risk of disease. The New England journal of medicine. 2010;363(2):166-76. doi: 10.1056/NEJMra0905980. PubMed PMID: 20647212.
22. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356-69. doi: 10.1038/nrg2344. PubMed PMID: 18398418.
23. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet. 2008;40(2):189-97. doi: 10.1038/ng.75. PubMed PMID: 18193044; PubMed Central PMCID: PMCPMC2682493.
24. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet.

2009;41(1):56-65. doi: 10.1038/ng.291. PubMed PMID: 19060906; PubMed Central PMCID: PMCPMC2881676.
25. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707-13. Epub 2010/08/06. doi: 10.1038/nature09270. PubMed PMID: 20686565; PubMed Central PMCID: PMCPMC3039276.
26. White J, Swerdlow DI, Preiss D, Fairhurst-Hunter Z, Keating BJ, Asselbergs FW, et al. Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes. JAMA cardiology. 2016;1(6):692-9. Epub 2016/08/04. doi: 10.1001/jamacardio.2016.1884. PubMed PMID: 27487401.
27. Hansson GK. Inflammation, atherosclerosis, and coronary artery disease. The New England journal of medicine. 2005;352(16):1685-95. doi: 10.1056/NEJMra043430. PubMed PMID: 15843671.
28. Hansson GK, Libby P. The immune response in atherosclerosis: a double-edged sword. Nat Rev Immunol. 2006;6(7):508-19. doi: 10.1038/nri1882. PubMed PMID: 16778830.
29. Libby P, Ridker PM, Maseri A. Inflammation and atherosclerosis. Circulation. 2002;105(9):1135-43. Epub 2002/03/06. PubMed PMID: 11877368.
30. Fernando E, Razak F, Lear SA, Anand SS. Cardiovascular Disease in South Asian Migrants. The Canadian journal of cardiology. 2015;31(9):1139-50. Epub 2015/09/01. doi: 10.1016/j.cjca.2015.06.008. PubMed PMID: 26321436.
31. Patel JV, Vyas A, Cruickshank JK, Prabhakaran D, Hughes E, Reddy KS, et al. Impact of migration on coronary heart disease risk factors: comparison of Gujaratis in Britain and their contemporaries in villages of origin in India. Atherosclerosis. 2006;185(2):297-306. doi: 10.1016/j.atherosclerosis.2005.06.005. PubMed PMID: 16005463.
32. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell. 2012;148(6):1242-57. Epub 2012/03/20. doi: 10.1016/j.cell.2012.03.001. PubMed PMID: 22424232; PubMed Central PMCID: PMCPMC3319439.
33. Kullo IJ, Leeper NJ. The genetic basis of peripheral arterial disease: current knowledge, challenges, and future directions. Circulation research. 2015;116(9):1551-60. Epub 2015/04/25. doi: 10.1161/circresaha.116.303518. PubMed PMID: 25908728; PubMed Central PMCID: PMCPMC4410432.
34. O'Donnell CJ, Nabel EG. Genomics of cardiovascular disease. The New England journal of medicine. 2011;365(22):2098-109. doi: 10.1056/NEJMra1105239. PubMed PMID: 22129254.
35. Ozaki K, Tanaka T. Molecular genetics of coronary artery disease. Journal of human genetics. 2016;61(1):71-7. Epub 2015/07/03. doi: 10.1038/jhg.2015.70. PubMed PMID: 26134515.
36. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17(R2):R156-65. doi: 10.1093/hmg/ddn289. PubMed PMID: 18852205; PubMed Central PMCID: PMCPMC2782356.
37. Ioannidis JP, Ntzani EE, Trikalinos TA. 'Racial' differences in genetic effects for complex diseases. Nat Genet. 2004;36(12):1312-8. doi: 10.1038/ng1474. PubMed PMID: 15543147.
38. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. doi: 10.1038/nature08494. PubMed PMID: 19812666; PubMed Central PMCID: PMCPMC2831613.

39. Liu H, Liu W, Liao Y, Cheng L, Liu Q, Ren X, et al. CADgene: a comprehensive database for coronary artery disease genes. Nucleic Acids Res. 2011;39(Database issue):D991-6. doi: 10.1093/nar/gkq1106. PubMed PMID: 21045063; PubMed Central PMCID: PMCPMC3013698.
40. Stanley-Hasnain S, Hauck L, Grothe D, Aschar-Sobbi R, Beca S, Butany J, et al. p53 and Mdm2 act synergistically to maintain cardiac homeostasis and mediate cardiomyocyte cell cycle arrest through a network of microRNAs. Cell cycle (Georgetown, Tex). 2017:1-16. Epub 2017/07/27. doi: 10.1080/15384101.2017.1346758. PubMed PMID: 28745540.
41. Goldstein JL, Brown MS. A century of cholesterol and coronaries: from plaques to genes to statins. Cell. 2015;161(1):161-72. Epub 2015/03/31. doi: 10.1016/j.cell.2015.01.036. PubMed PMID: 25815993; PubMed Central PMCID: PMCPMC4525717.
42. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science. 2004;304(5675):1321-5. Epub 2004/05/08. doi: 10.1126/science.1098119. PubMed PMID: 15131266.
43. Zanetti D, Carreras-Torres R, Esteban E, Via M, Moral P. Potential Signals of Natural Selection in the Top Risk Loci for Coronary Artery Disease: 9p21 and 10q11. PLoS One. 2015;10(8):e0134840. Epub 2015/08/08. doi: 10.1371/journal.pone.0134840. PubMed PMID: 26252781; PubMed Central PMCID: PMCPMC4529309.
44. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135-45. doi: 10.1038/nrg3118. PubMed PMID: 22251874; PubMed Central PMCID: PMCPMC4408201.
45. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. Clinical genetics. 2011;79(3):199-206. doi: 10.1111/j.1399-0004.2010.01535.x. PubMed PMID: 20831747; PubMed Central PMCID: PMCPMC3652532.
46. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. American journal of human genetics. 2008;82(1):100-12. doi: 10.1016/j.ajhg.2007.09.006. PubMed PMID: 18179889; PubMed Central PMCID: PMCPMC2253956.
47. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A. 2007;104(30):12410-5. doi: 10.1073/pnas.0705140104. PubMed PMID: 17640883; PubMed Central PMCID: PMCPMC1941483.
48. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823-8. Epub 2012/02/22. doi: 10.1126/science.1215040. PubMed PMID: 22344438; PubMed Central PMCID: PMCPMC3299548.
49. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506-11. Epub 2013/09/17. doi: 10.1038/nature12531. PubMed PMID: 24037378; PubMed Central PMCID: PMCPMC3918453.
50. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-7. doi: 10.1038/nature04240. PubMed PMID: 16237444.
51. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive evolution in the human genome. PLoS Genet. 2007;3(6):e90. doi: 10.1371/journal.pgen.0030090. PubMed PMID: 17542651; PubMed Central PMCID: PMCPMC1885279.

52. Van Den Biggelaar AH, De Craen AJ, Gussekloo J, Huizinga TW, Heijmans BT, Frolich M, et al. Inflammation underlying cardiovascular mortality is a late consequence of evolutionary programming. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2004;18(9):1022-4. Epub 2004/04/16. doi: 10.1096/fj.03-1162fje. PubMed PMID: 15084512.
53. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017;544(7649):235-9. doi: 10.1038/nature22034. PubMed PMID: 28406212.
54. Stitziel NO, Khera AV, Wang X, Bierhals AJ, Vourakis AC, Sperry AE, et al. ANGPTL3 Deficiency and Protection Against Coronary Artery Disease. Journal of the American College of Cardiology. 2017;69(16):2054-63. Epub 2017/04/03. doi: 10.1016/j.jacc.2017.02.030. PubMed PMID: 28385496; PubMed Central PMCID: PMCPMC5404817.
55. Timpson NJ, Walter K, Min JL, Tachmazidou I, Malerba G, Shin SY, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. Nature communications. 2014;5:4871. Epub 2014/09/17. doi: 10.1038/ncomms5871. PubMed PMID: 25225788; PubMed Central PMCID: PMCPMC4167609.
56. Palmerini T, Caixeta A, Genereux P, Cristea E, Lansky A, Mehran R, et al. Comparison of clinical and angiographic prognostic risk scores in patients with acute coronary syndromes: Analysis from the Acute Catheterization and Urgent Intervention Triage StrategY (ACUITY) trial. American heart journal. 2012;163(3):383-91, 91 e1-5. Epub 2012/03/20. doi: 10.1016/j.ahj.2011.11.010. PubMed PMID: 22424008.
57. Palmerini T, Genereux P, Caixeta A, Cristea E, Lansky A, Mehran R, et al. A new score for risk stratification of patients with acute coronary syndromes undergoing percutaneous coronary intervention: the ACUITY-PCI (Acute Catheterization and Urgent Intervention Triage Strategy-Percutaneous Coronary Intervention) risk score. JACC Cardiovasc Interv. 2012;5(11):1108-16. Epub 2012/11/24. doi: 10.1016/j.jcin.2012.07.011. PubMed PMID: 23174634.
58. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, et al. A systems approach to mapping DNA damage response pathways. Science. 2006;312(5776):1054-9. Epub 2006/05/20. doi: 10.1126/science.1122088. PubMed PMID: 16709784; PubMed Central PMCID: PMCPMC2811083.
59. Aten JA, Stap J, Krawczyk PM, van Oven CH, Hoebe RA, Essers J, et al. Dynamics of DNA double-strand breaks revealed by clustering of damaged chromosome domains. Science. 2004;303(5654):92-5. Epub 2004/01/06. doi: 10.1126/science.1088845. PubMed PMID: 14704429.
60. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature. 2009;461(7267):1071-8. Epub 2009/10/23. doi: 10.1038/nature08467. PubMed PMID: 19847258; PubMed Central PMCID: PMCPmc2906700.
61. Polo SE, Jackson SP. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. Genes & development. 2011;25(5):409-33. Epub 2011/03/03. doi: 10.1101/gad.2021311. PubMed PMID: 21363960; PubMed Central PMCID: PMCPMC3049283.
62. Mocanu MM, Yellon DM. p53 down-regulation: a new molecular mechanism involved in ischaemic preconditioning. FEBS Lett. 2003;555(2):302-6. Epub 2003/12/04. PubMed PMID: 14644432.
63. Schneider JG, Finck BN, Ren J, Standley KN, Takagi M, Maclean KH, et al. ATM-dependent suppression of stress signaling reduces vascular disease in metabolic syndrome. Cell Metab. 2006;4(5):377-89. Epub 2006/11/07. doi: 10.1016/j.cmet.2006.10.002. PubMed PMID: 17084711.

64. Vousden KH, Lane DP. p53 in health and disease. Nat Rev Mol Cell Biol. 2007;8(4):275-83. Epub 2007/03/24. doi: 10.1038/nrm2147. PubMed PMID: 17380161.
65. Diderich K, Alanazi M, Hoeijmakers JH. Premature aging and cancer in nucleotide excision repair-disorders. DNA repair. 2011;10(7):772-80. Epub 2011/06/18. doi: 10.1016/j.dnarep.2011.04.025. PubMed PMID: 21680258; PubMed Central PMCID: PMCPmc4128095.
66. Garinis GA, van der Horst GT, Vijg J, Hoeijmakers JH. DNA damage and ageing: new-age ideas for an age-old problem. Nat Cell Biol. 2008;10(11):1241-7. Epub 2008/11/04. doi: 10.1038/ncb1108-1241. PubMed PMID: 18978832; PubMed Central PMCID: PMCPmc4351702.
67. Hoeijmakers JH. DNA damage, aging, and cancer. The New England journal of medicine. 2009;361(15):1475-85. Epub 2009/10/09. doi: 10.1056/NEJMra0804615. PubMed PMID: 19812404.
68. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutation research. 2005;577(1-2):275-83. Epub 2005/06/01. doi: 10.1016/j.mrfmmm.2005.03.007. PubMed PMID: 15922366.
69. Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science. 2001;291(5507):1284-9. Epub 2001/02/22. doi: 10.1126/science.1056154. PubMed PMID: 11181991.
70. Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, 3rd, Hurov KE, Luo J, et al. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science. 2007;316(5828):1160-6. Epub 2007/05/26. doi: 10.1126/science.1140321. PubMed PMID: 17525332.
71. Kingman JFC. On the genealogy of large populations. Journal of Applied Probability. 1982;19A:27-43.
72. Beerli P, Felsenstein J. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A. 2001;98(8):4563-8. Epub 2001/04/05. doi: 10.1073/pnas.081068098. PubMed PMID: 11287657; PubMed Central PMCID: PMCPMC31874.
73. Felsenstein J. Inferring Phylogenies: Macmillian Education; 2004.
74. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313-26. Epub 2010/07/24. doi: 10.1101/gr.101386.109. PubMed PMID: 20651121; PubMed Central PMCID: PMCPmc2945180.
75. Valencia A. Multiple sequence alignments as tools for protein structure and function prediction. Comparative and functional genomics. 2003;4(4):424-7. Epub 2008/07/17. doi: 10.1002/cfg.313. PubMed PMID: 18629077; PubMed Central PMCID: PMCPmc2447362.
76. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12(6):962-8. Epub 2002/06/05. doi: 10.1101/gr.87702. Article published online before print in May 2002. PubMed PMID: 12045149; PubMed Central PMCID: PMCPmc1383730.
77. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics. 2001;158(2):927-31. Epub 2001/06/30. PubMed PMID: 11430355; PubMed Central PMCID: PMC1461684.
78. Pal C, Papp B, Hurst LD. Genomic function: Rate of evolution and gene dispensability. Nature. 2003;421(6922):496-7; discussion 7-8. Epub 2003/01/31. doi: 10.1038/421496b
421496b [pii]. PubMed PMID: 12556881.

79. Hirsh AE, Fraser HB. Protein dispensability and rate of evolution. Nature. 2001;411(6841):1046-9. Epub 2001/06/29. doi: 10.1038/35082561. PubMed PMID: 11429604.
80. Hurst LD. Evolutionary genomics: A positive becomes a negative. Nature. 2009;457(7229):543-4. Epub 2009/01/30. doi: 457543a [pii] 10.1038/457543a [doi]. PubMed PMID: 19177117.
81. Orr HA. The genetic theory of adaptation: a brief history. Nat Rev Genet. 2005;6(2):119-27. Epub 2005/02/18. doi: nrg1523 [pii] 10.1038/nrg1523. PubMed PMID: 15716908.
82. Sawyer SL, Malik HS. Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis. Proc Natl Acad Sci U S A. 2006;103(47):17614-9. Epub 2006/11/15. doi: 10.1073/pnas.0605468103. PubMed PMID: 17101967; PubMed Central PMCID: PMCPmc1693795.
83. Demogines A, Abraham J, Choe H, Farzan M, Sawyer SL. Dual host-virus arms races shape an essential housekeeping protein. PLoS biology. 2013;11(5):e1001571. Epub 2013/06/01. doi: 10.1371/journal.pbio.1001571. PubMed PMID: 23723737; PubMed Central PMCID: PMCPmc3665890.
84. Quintela-Fandino M, Hitt R, Medina PP, Gamarra S, Manso L, Cortes-Funes H, et al. DNA-repair gene polymorphisms predict favorable clinical outcome among patients with advanced squamous cell carcinoma of the head and neck treated with cisplatin-based induction chemotherapy. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24(26):4333-9. Epub 2006/08/10. doi: 10.1200/jco.2006.05.8768. PubMed PMID: 16896002.
85. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 1999;27(5):1223-42. Epub 1999/02/12. PubMed PMID: 9973609; PubMed Central PMCID: PMCPmc148307.
86. Kawabata M, Kawabata T, Nishibori M. Role of recA/RAD51 family proteins in mammals. Acta medica Okayama. 2005;59(1):1-9. Epub 2005/05/21. PubMed PMID: 15902993.
87. Su TT. Cellular responses to DNA damage: one signal, multiple choices. Annual review of genetics. 2006;40:187-208. Epub 2006/06/30. doi: 10.1146/annurev.genet.40.110405.090428. PubMed PMID: 16805666.
88. Hofmann RM, Pickart CM. Noncanonical MMS2-encoded ubiquitin-conjugating enzyme functions in assembly of novel polyubiquitin chains for DNA repair. Cell. 1999;96(5):645-53. Epub 1999/03/25. PubMed PMID: 10089880.
89. Gibbs PE, McGregor WG, Maher VM, Nisson P, Lawrence CW. A human homolog of the Saccharomyces cerevisiae REV3 gene, which encodes the catalytic subunit of DNA polymerase zeta. Proc Natl Acad Sci U S A. 1998;95(12):6876-80. Epub 1998/06/17. PubMed PMID: 9618506; PubMed Central PMCID: PMCPmc22668.
90. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91. Epub 2007/05/08. doi: msm088 [pii] 10.1093/molbev/msm088 [doi]. PubMed PMID: 17483113.
91. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 2007;35(Web Server issue):W506-11. Epub 2007/06/26. doi: gkm382 [pii] 10.1093/nar/gkm382 [doi]. PubMed PMID: 17586822; PubMed Central PMCID: PMC1933148.
92. Hurst LD. Fundamental concepts in genetics: genetics and the understanding of selection. Nat Rev Genet. 2009;10(2):83-93. Epub 2009/01/03. doi: nrg2506 [pii] 10.1038/nrg2506 [doi]. PubMed PMID: 19119264.

93. Bielawski JP. Detecting the signatures of adaptive evolution in protein-coding genes. Curr Protoc Mol Biol. 2013;Chapter 19:Unit 19 1. Epub 2013/01/05. doi: 10.1002/0471142727.mb1901s101. PubMed PMID: 23288462.
94. Lane DP. Cancer. p53, guardian of the genome. Nature. 1992;358(6381):15-6. Epub 1992/07/02. doi: 10.1038/358015a0. PubMed PMID: 1614522.
95. Levine AJ. p53, the cellular gatekeeper for growth and division. Cell. 1997;88(3):323-31. Epub 1997/02/07. PubMed PMID: 9039259.
96. Hainaut P, Wiman KG. 30 years and a long way into p53 research. The Lancet Oncology. 2009;10(9):913-9. Epub 2009/09/01. doi: 10.1016/s1470-2045(09)70198-6. PubMed PMID: 19717093.
97. Natan E, Baloglu C, Pagel K, Freund SM, Morgner N, Robinson CV, et al. Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. J Mol Biol. 2011;409(3):358-68. Epub 2011/04/05. doi: 10.1016/j.jmb.2011.03.047. PubMed PMID: 21457718; PubMed Central PMCID: PMCPmc3176915.
98. Di Lello P, Jenkins LM, Jones TN, Nguyen BD, Hara T, Yamaguchi H, et al. Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIH and the activation domain of p53. Molecular cell. 2006;22(6):731-40. Epub 2006/06/24. doi: 10.1016/j.molcel.2006.05.007. PubMed PMID: 16793543.
99. Lill NL, Grossman SR, Ginsberg D, DeCaprio J, Livingston DM. Binding and modulation of p53 by p300/CBP coactivators. Nature. 1997;387(6635):823-7. Epub 1997/06/19. doi: 10.1038/42981. PubMed PMID: 9194565.
100. Teufel DP, Bycroft M, Fersht AR. Regulation by phosphorylation of the relative affinities of the N-terminal transactivation domains of p53 for p300 domains and Mdm2. Oncogene. 2009;28(20):2112-8. Epub 2009/04/14. doi: 10.1038/onc.2009.71. PubMed PMID: 19363523; PubMed Central PMCID: PMCPmc2685776.
101. Thut CJ, Chen JL, Klemm R, Tjian R. p53 transcriptional activation mediated by coactivators TAFII40 and TAFII60. Science. 1995;267(5194):100-4. Epub 1995/01/06. PubMed PMID: 7809597.
102. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, et al. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci U S A. 2008;105(15):5762-7. Epub 2008/04/09. doi: 10.1073/pnas.0801353105. PubMed PMID: 18391200; PubMed Central PMCID: PMCPmc2311362.
103. Wade M, Li YC, Wahl GM. MDM2, MDMX and p53 in oncogenesis and cancer therapy. Nature reviews Cancer. 2013;13(2):83-96. doi: 10.1038/nrc3430. PubMed PMID: 23303139; PubMed Central PMCID: PMCPMC4161369.
104. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science. 1994;265(5170):346-55. Epub 1994/07/15. PubMed PMID: 8023157.
105. Lin J, Chen J, Elenbaas B, Levine AJ. Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1B 55-kD protein. Genes & development. 1994;8(10):1235-46. Epub 1994/05/15. PubMed PMID: 7926727.
106. Poyurovsky MV, Katz C, Laptenko O, Beckerman R, Lokshin M, Ahn J, et al. The C terminus of p53 binds the N-terminal domain of MDM2. Nat Struct Mol Biol. 2010;17(8):982-9. Epub 2010/07/20. doi: 10.1038/nsmb.1872. PubMed PMID: 20639885; PubMed Central PMCID: PMCPmc2922928.
107. Bista M, Petrovich M, Fersht AR. MDMX contains an autoinhibitory sequence element. Proc Natl Acad Sci U S A. 2013;110(44):17814-9. Epub 2013/10/16. doi:

10.1073/pnas.1317398110. PubMed PMID: 24127580; PubMed Central PMCID: PMCPmc3816421.
108. Su CT, Chen CY, Hsu CM. iPDA: integrated protein disorder analyzer. Nucleic Acids Res. 2007;35(Web Server issue):W465-72. Epub 2007/06/08. doi: 10.1093/nar/gkm353. PubMed PMID: 17553839; PubMed Central PMCID: PMCPmc1933224.
109. Buchhop S, Gibson MK, Wang XW, Wagner P, Sturzbecher HW, Harris CC. Interaction of p53 with the human Rad51 protein. Nucleic Acids Res. 1997;25(19):3868-74. Epub 1997/10/10. PubMed PMID: 9380510; PubMed Central PMCID: PMCPmc146972.
110. Linke SP, Sengupta S, Khabie N, Jeffries BA, Buchhop S, Miska S, et al. p53 interacts with hRAD51 and hRAD54, and directly modulates homologous recombination. Cancer research. 2003;63(10):2596-605. Epub 2003/05/17. PubMed PMID: 12750285.
111. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A. 2001;98(5):2509-14. Epub 2001/02/28. doi: 10.1073/pnas.051605998 051605998 [pii]. PubMed PMID: 11226269; PubMed Central PMCID: PMC30168.
112. Fatakia SN, Costanzi S, Chow CC. Molecular evolution of the transmembrane domains of G protein-coupled receptors. PLoS One. 2011;6(11):e27813. Epub 2011/12/02. doi: 10.1371/journal.pone.0027813. PubMed PMID: 22132149; PubMed Central PMCID: PMCPMC3221663.
113. de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet. 2013;14(4):249-61. Epub 2013/03/06. doi: 10.1038/nrg3414. PubMed PMID: 23458856.
114. Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP. Adaptive evolution of energy metabolism genes and the origin of flight in bats. Proc Natl Acad Sci U S A. 2010;107(19):8666-71. Epub 2010/04/28. doi: 10.1073/pnas.0912613107. PubMed PMID: 20421465; PubMed Central PMCID: PMCPmc2889356.
115. Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, et al. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science. 2013;339(6118):456-60. Epub 2012/12/22. doi: 10.1126/science.1230835. PubMed PMID: 23258410.
116. Zhao Y, Ren JL, Wang MY, Zhang ST, Liu Y, Li M, et al. Codon 104 variation of p53 gene provides adaptive apoptotic responses to extreme environments in mammals of the Tibet plateau. Proc Natl Acad Sci U S A. 2013;110(51):20639-44. Epub 2013/12/04. doi: 10.1073/pnas.1320369110. PubMed PMID: 24297887.
117. Ashur-Fabian O, Avivi A, Trakhtenbrot L, Adamsky K, Cohen M, Kajakaro G, et al. Evolution of p53 in hypoxia-stressed Spalax mimics human tumor mutation. Proc Natl Acad Sci U S A. 2004;101(33):12236-41. Epub 2004/08/11. doi: 10.1073/pnas.0404998101. PubMed PMID: 15302922; PubMed Central PMCID: PMCPmc514462.
118. Avivi A, Ashur-Fabian O, Joel A, Trakhtenbrot L, Adamsky K, Goldstein I, et al. P53 in blind subterranean mole rats--loss-of-function versus gain-of-function activities on newly cloned Spalax target genes. Oncogene. 2007;26(17):2507-12. Epub 2006/10/18. doi: 10.1038/sj.onc.1210045. PubMed PMID: 17043642.
119. Park TJ, Reznick J, Peterson BL, Blass G, Omerbasic D, Bennett NC, et al. Fructose-driven glycolysis supports anoxia resistance in the naked mole-rat. Science. 2017;356(6335):307-11. Epub 2017/04/22. doi: 10.1126/science.aab3896. PubMed PMID: 28428423.
120. Demogines A, East AM, Lee JH, Grossman SR, Sabeti PC, Paull TT, et al. Ancient and recent adaptive evolution of primate non-homologous end joining genes.

PLoS Genet. 2010;6(10):e1001169. Epub 2010/10/27. doi: 10.1371/journal.pgen.1001169. PubMed PMID: 20975951; PubMed Central PMCID: PMCPmc2958818.
121. O'Connell MJ. Selection and the cell cycle: positive Darwinian selection in a well-known DNA damage response pathway. J Mol Evol. 2010;71(5-6):444-57. Epub 2010/11/09. doi: 10.1007/s00239-010-9399-y. PubMed PMID: 21057781.
122. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012;8(8):e1002793. Epub 2012/08/10. doi: 10.1371/journal.pgen.1002793. PubMed PMID: 22876189; PubMed Central PMCID: PMCPMC3410907.
123. Cover TM, Thomas JA. Elements of Information Theory: Wiley; 1991.
124. Garcia CK, Wilund K, Arca M, Zuliani G, Fellin R, Maioli M, et al. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. Science. 2001;292(5520):1394-8. Epub 2001/04/28. doi: 10.1126/science.1060458. PubMed PMID: 11326085.
125. Leren TP. Mutations in the PCSK9 gene in Norwegian subjects with autosomal dominant hypercholesterolemia. Clinical genetics. 2004;65(5):419-22. Epub 2004/04/22. doi: 10.1111/j.0009-9163.2004.0238.x. PubMed PMID: 15099351.
126. Timms KM, Wagner S, Samuels ME, Forbey K, Goldfine H, Jammulapati S, et al. A mutation in PCSK9 causing autosomal-dominant hypercholesterolemia in a Utah pedigree. Human genetics. 2004;114(4):349-53. Epub 2004/01/17. doi: 10.1007/s00439-003-1071-9. PubMed PMID: 14727179.
127. Cooper GM, Brown CD. Qualifying the relationship between sequence conservation and molecular function. Genome Res. 2008;18(2):201-5. Epub 2008/02/05. doi: 10.1101/gr.7205808. PubMed PMID: 18245453.

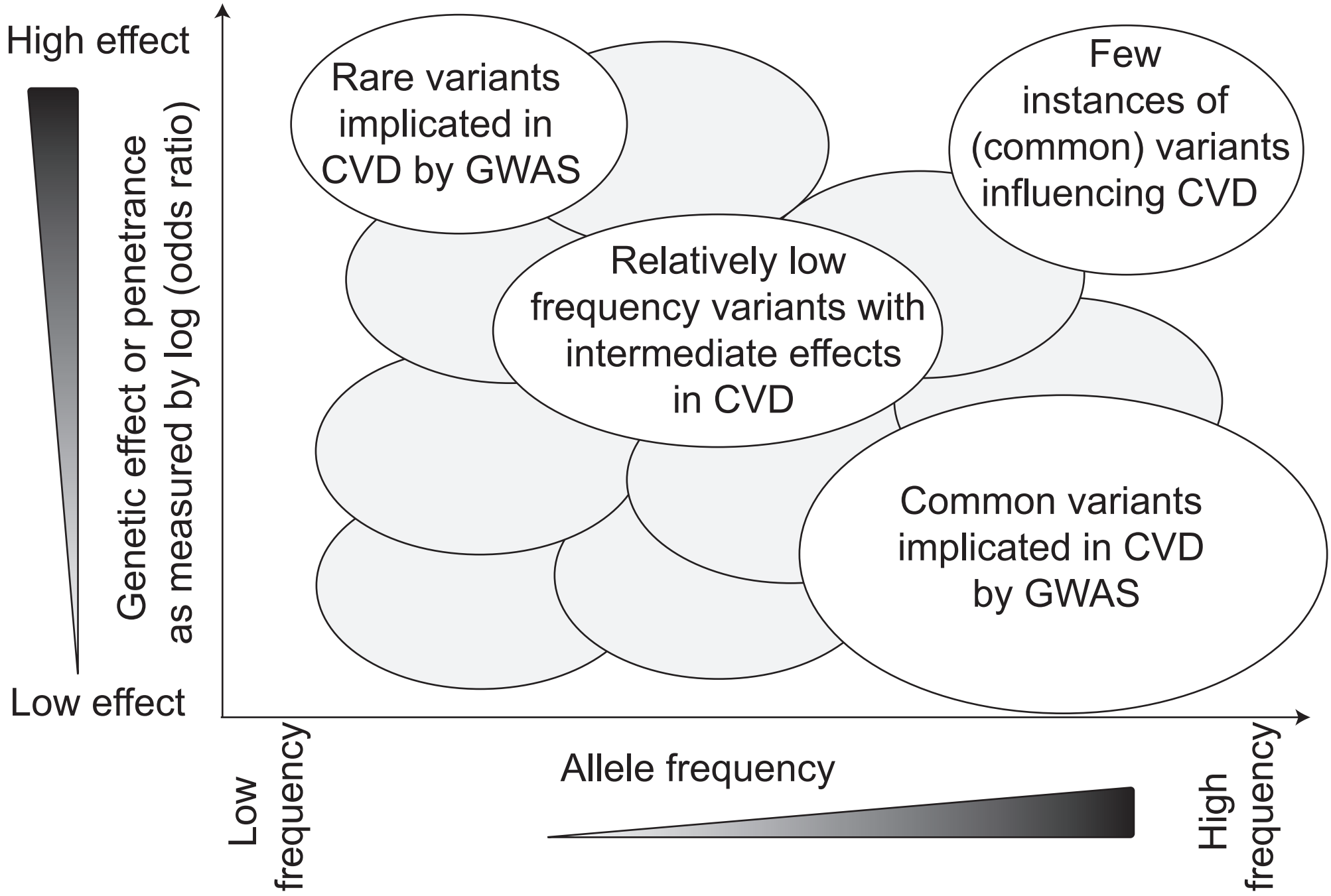

High effect
Low effect
Genetic effect or penetrance as measured by log (odds ratio)
Rare variants implicated in CVD by GWAS
Few instances of (common) variants influencing CVD
Relatively low frequency variants with intermediate effects in CVD
Common variants implicated in CVD by GWAS
Low frequency
Allele frequency
High frequency

Figure 1

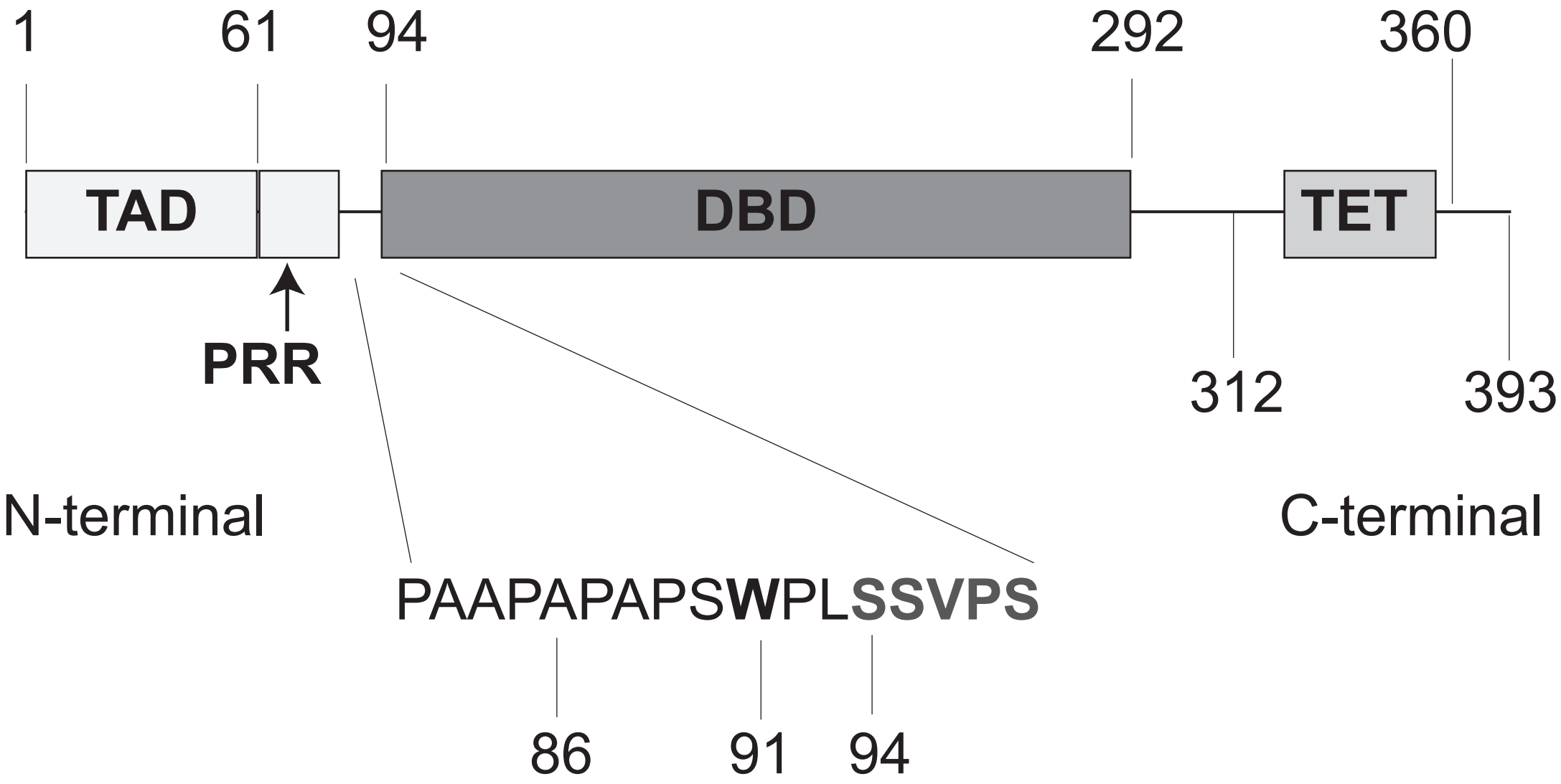


Figure 2 (inset for Box 2)

```
H. sapiens        STKRALPNNTSSSPQPKKKPL        SLSFDESLALC
M. musculus       .A...........PQ.....         .....P..G..
B. taurus         ...................          ...........
A. melanoleuca    ..............PQ.....        ...........
F. catus          ..............PQ.....        ...........
C. familiaris     ..............PQ.....        ...........
E. caballus       ....V.........PQ......        ......G....
P. alecto         ..............PSP.M..        .F.........
M. davidii        ..............P-P...A        .F.........
```

Figure 3 (inset for Box 3)

		174		209
Spalax	KKSQHMTEVV	**K**RCPHHERCS	GNLRAEYLDD	**K**HTFRHSVVV
Oryctolagus cuniculus	KKSQHMTEVV	RRCPHHERCS	GNLRAEYLDD	RNTFRHSVVV
Homo sapiens	KQSQHMTEVV	RRCPHHERCS	GNLRVEYLDD	RNTFRHSVVV
Cercopithecus aethiops	KQSQHMTEVV	RRCPHHERCS	GNLRVEYSDD	RNTFRHSVVV
Macaca muleta	KQSQHMTEVV	RRCPHHERCS	GNLRVEYSDD	RNTFRHSVVV
Macaca fuscata	KQSQHMTEVV	RRCPHHERCS	GNLRVEYSDD	RNTFRHSVVV
Macaca fascicularis	KQSQHMTEVV	RRCPHHERCS	GNLRVEYSDD	RNTFRHSVVV
Tupaia belangeri	KQSQYVTEVV	RRCPHHERCS	GNLHAEYSDD	RNTFRHSVVV
Cavia porcellus	KKSQHMTEVV	RRCPHHERCS	GNLHAEYVDD	RTTFRHSVVV
Mesocricetus auratus	KKLQYMTEVV	RRCPHHERSS	GNMHAEYLDD	**K**QTFRHSVVV
Cricetulus griseus	KKLQYMTEVV	RRCPHHERSS	GNLHAEYLDD	**K**QTFRHSVVV
Rattus norvegicus	KKSQHMTEVV	RRCPHHERCS	GNPYAEYLDD	RQTFRHSVVV
Mus musculus	KKSQHMTEVV	RRCPHHERCS	GNLYPEYLED	RQTFRHSVVV
Felis catus	KKSEFMTEVV	RRCPHHERCP	GNLHAKYLDD	RNTFRHSVVV
Ovis aries	KKLEHMTEVV	RRSPHHERSS	GNLRAEYFDD	RNTFRHSVVV
Bos taurus	KKLEHMTEVV	RRCPHHERSS	GNLRAEYLDD	RNTFRHSVVV
Bos indicus	KKLEHMTEVV	RRCPHHERSS	GNLRAEYLDD	RNTFRHSVVV
Bos primigenius	KKLEHMTEVV	RRCPHHERSS	GNLRAEYLDD	RNTFRHSVVV
Meriones unguiculatus	KNSQHMTEVV	RRCPHHERCS	GNLHAEYVDD	RQTFRHSVLV
Spermophilus beecheyi	KKSQHMTEVV	RRCPHHERCS	GNLHAEYVDD	RQTFRHSVLV
Mastomys natalensis	KKSQHMTEVV	RRCPHHERCT	GNLNAEYLDD	**K**QTFRHSVVV
Gallus gallus	KKSEHVAEVV	RRCPHHERCG	GNPQARYHDD	ETTKRHSVVV
Microtus arvalis	GGMNRMTEVV	RRCPHHERSS	GNLRAEYLDD	RNTFRHSVVV
Microtus rossiaeme	KKSQHMTEVV	RRCPHHERCS	GNLRAEYLDD	RQTFRHSVVV
Xenopus laevis	KKSEHVAEVV	**K**RCPHHERSV	GNLQAYYMED	VNSGRHSVCV
S.irideus	KKLSDVADVV	RRCPHHQSTS	GNQRSEYMED	GNTLRHSVLV
Danio rerio	KKSEHVAEVV	RRCPHHERTP	GNQRANYRED	NITLRHSVFV
Barbus barbus	KKSEHVAEVV	RRCPHHERTP	GNSRALYRED	DVNSRHSVVV
Ictalurus punctatus	KRSEHVAEVV	RRCPHHERSN	GNSRAVYQED	GNTQAHSVVV
Canis familiaris	KKSEFVTEVV	RRCPHHERCS	GNLRAKYLDD	RNTFRHSVVV
Equus caballus	KKSEFMTEVV	RRCPHHERCS	GNLRAEYLDD	RNTFRHSVVV
Sus scrofa	KKSEYMTEVV	RRCPHHERSS	GNLRAEYLDD	RNTFRHSVVV
Monodelphis domestica	KKSEHMTEVV	**K**RCPHHEQCT	GNLQAEYLID	ATTKRQSVSV
Oncorhynchus keta	KKLSDVADVV	RRCPHHQSTS	GNQRSEYMED	RNTLRQSVLV
Platichthys flesus	KKTEHVADVV	RRCPHHQTED	GSQRALYFED	PHTKRQSVTV
Tetraodon miurus	KKTEHVAEVV	RRCPHHQNED	GSERAQYFEH	PHTKRQSVTV
Oryzias latipes	KKTEHVADVV	RRCPHHQNED	GSQLAQYFED	PYTKRQSVTV
Xiphophorus maculatus	KKTEHVGEVV	**K**RCPHHQSED	GSQLAQYFED	PNTRRHSVTV
Xiphophorus helleri	KKTEHVGEVV	**K**RCPHHQSED	GSQLAQYFED	PNTRRHSVTV
Loligo forbesi	MKPEHVQEVV	**K**RCPNHATAK	HKL-AKYHED	**K**YSGRQSVLI
Equus asinus	KKSEFMTEVV	RRCPHHERCS	GNLRAEYLDD	RNTLRHSVVV
Delphinapterus leucas	KKSEYMTEVV	RRCPHHERCS	GNLRAEYLDD	RNTFRHSVVV
Sigmodon hispidus	KKSQHMTEVV	RRCPHHERCS	XNLRAEYLDD	**K**QTFRHKCGG
Mya arenaria	MKPEHVQEAV	**K**RCPNHATSK	HKV-SKYVED	PYTNRQSVLI

Figure 4 (inset for Box 3)

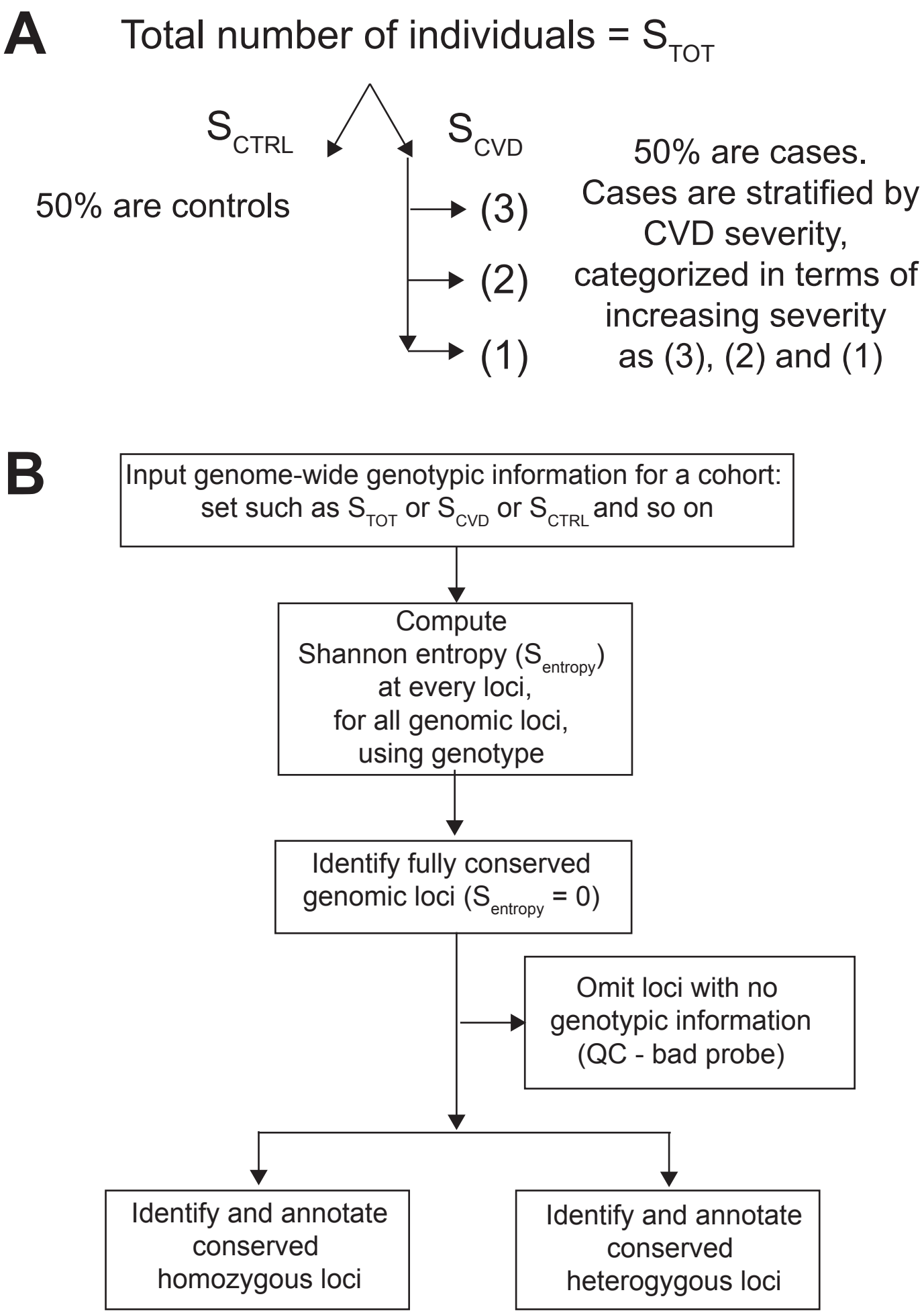
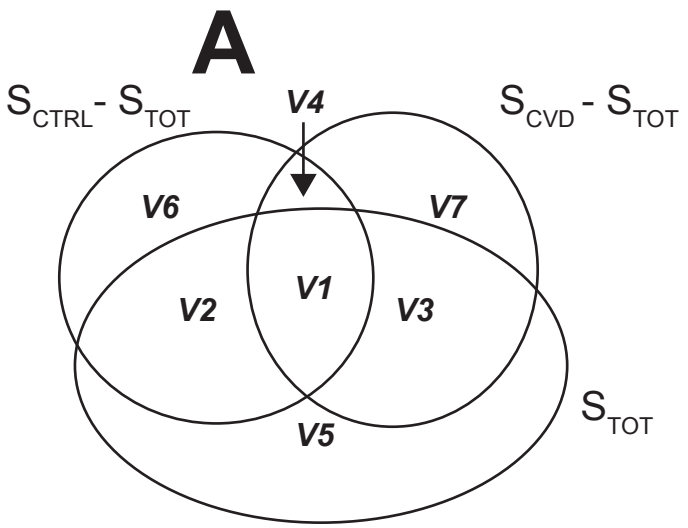
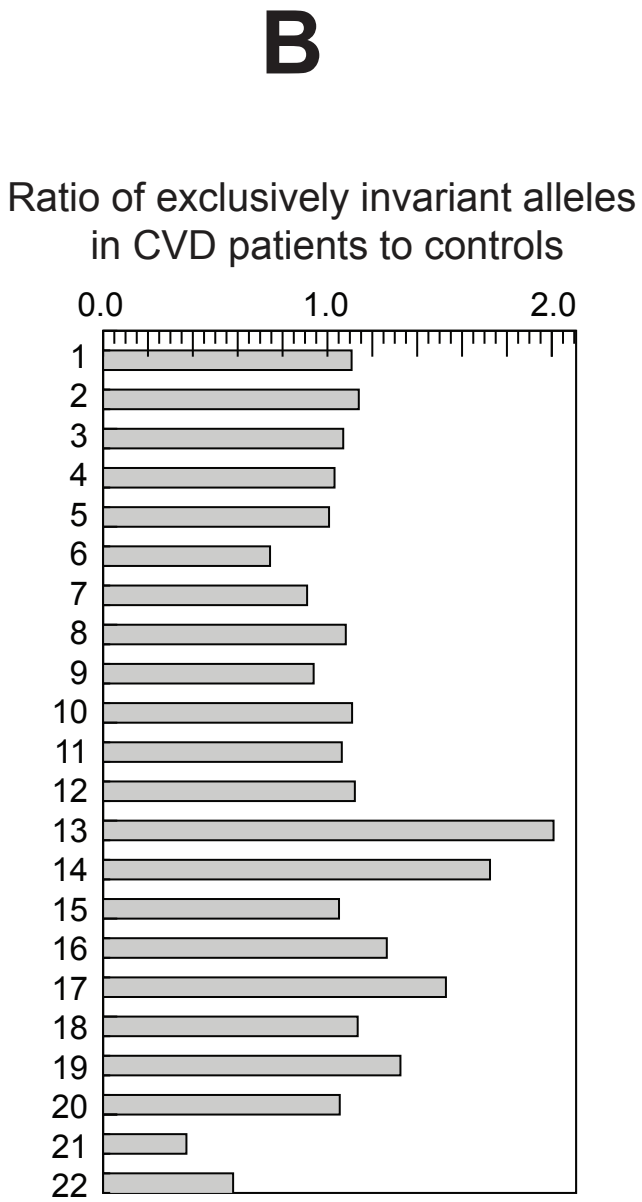

A
Total number of individuals = S_{TOT}
S_{CTRL}
S_{CVD}
50% are controls
(3)
(2)
(1)
50% are cases.
Cases are stratified by
CVD severity,
categorized in terms of
increasing severity
as (3), (2) and (1)
B
Input genome-wide genotypic information for a cohort:
set such as S_{TOT} or S_{CVD} or S_{CTRL} and so on
Compute
Shannon entropy ($S_{entropy}$)
at every loci,
for all genomic loci,
using genotype
Identify fully conserved
genomic loci ($S_{entropy}$ = 0)
Omit loci with no
genotypic information
(QC - bad probe)
Identify and annotate
conserved
homozygous loci
Identify and annotate
conserved
heterogygous loci

Figure 5


A
S_{CTRL} - S_{TOT}
V4
S_{CVD} - S_{TOT}
V6
V7
V1
V2
V3
V5
S_{TOT}


B
Ratio of exclusively invariant alleles in CVD patients to controls
0.0
1.0
2.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

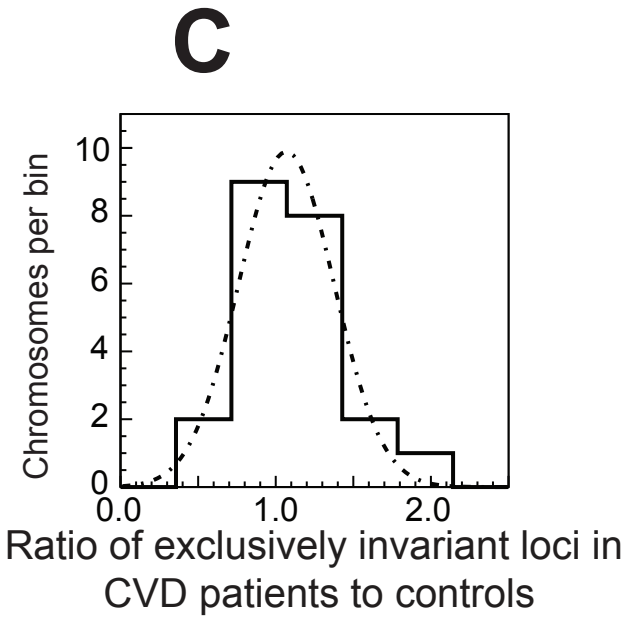
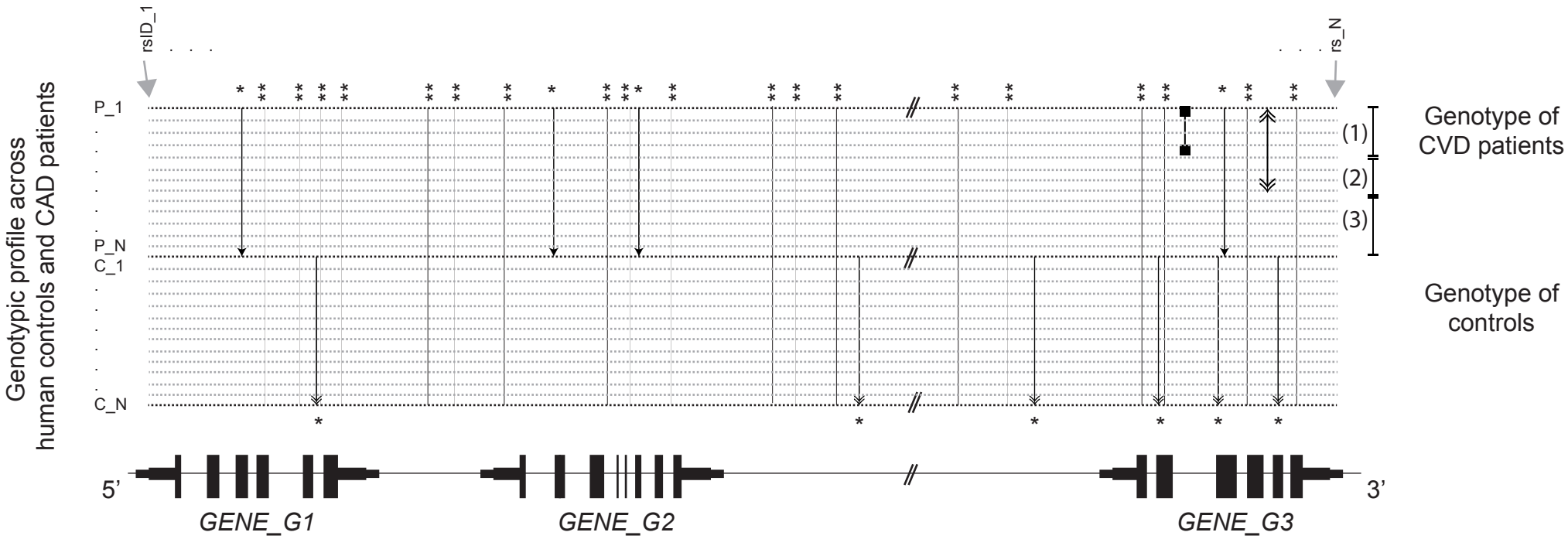

C
Chromosomes per bin
10
8
6
4
2
0
0.0
1.0
2.0
Ratio of exclusively invariant loci in CVD patients to controls

Figure 6



Conserved genotype across all individuals studied: CAD patients and age-, gender-matched controls (** shown on top margin)

Conserved genotype in all CAD cases only but not fully conserved in control cohort (* shown on top margin)

Conserved genotype in from all controls only but not fully conserved in CAD cohort (* shown on bottom margin)

Conserved genotype across all severe CAD cases (1), but not invariant across all CAD cases or in controls

Conserved genotype across all severe and moderately severe CAD patients ((1) and (2)), but not invariant across all patients or in controls

// A discontinuity in genomic coordinates

All CVD cases are categorized as per their SYNTAX scores: an angiographic tool stratifying the complexity of CVD with highest severity (1), intermediate severity (2), and least severity (3) - shown on right margin.

Schematic representation of genomic elements is not to scale,
CVD patients are represented as P_1 to P_N and controls C_1 to C_N on the left margin

Figure 7