Stage_2016 Session 4.2.2016

MODELLING SELECTION, MUTATION, DRIFT

- This lecture is an extension of Hardy-Weinberg, assumptions relaxed => towards more realistic situations.

- The goal is to provide framework for current statistical analysis of real data.

- Included also a set of home-exercises, anwers to be submitted 15.2.2016

- Like in first exercise set, you can work as 2-4 student groups.

- Note two recent review papers (in course webpage) which show how important this kind of (classical) population genetics framework is in understanding, for example, human populations:

  Revising the human mutation rate: implications for understanding human evolution
  Human genomic disease variants: A neutral evolutionary explanation

- We first familiarize with classical statistical population genetics theory and after that have the first glimpse to (human) DNA-sequence polymorphims. Later during the course we work with statistical analyses of polymorphisms.

INTRO

● Population genetics theory gives the basics for understanding how a population evolves under a given set of conditions. Evolution is a forward process: the genetic composition, allele and genotype frequencies change with time.

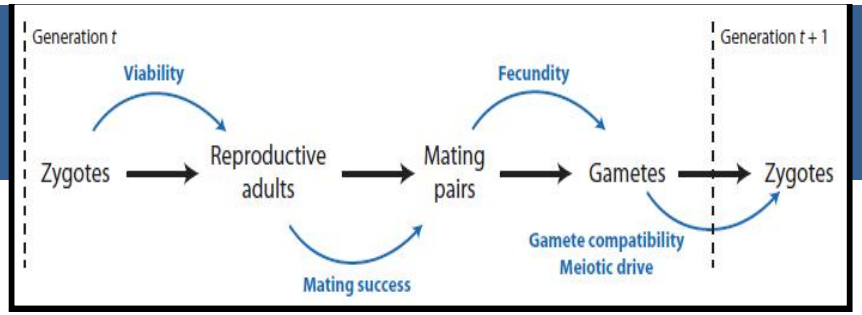
● Hardy-Weinberg-model, the basic null model, states that no change unless evolutionary factors - selection, genetic drift, mutation, gene flow from other populations – are in action.

● *Prospective population genetics theory* dominated for decades, after the seminal work of Sewall Wright, R.A. Fisher, J.B.S. Haldane, and Motoo Kimura. Although all this work is important and provides strong theoretical framework for understanding populations, current data analysis needs another viewpoint, too.

● In practice, in real situations for the researcher, the characteristics of a natural population (or a human population), are examined by taking samples from the population. Interesting biological questions that arise from a sample are mostly *retrospective*, such as the history of the population that gave rise to the sample, or the evolutionary mechanisms responsible for the characteristics observed.

● The accumulation of DNA sequence data since the 1980s has transformed the mainstream of population genetics research from prospective to retrospective, from demonstration of principles to inference of events that happened in the past.

THE BASIC SELECTION MODEL



● One of the messages in HWE is that gene and genotype frequencies do not change from one generation to another – if the assumptions are valid.

Stochastics (genetic drift) has a role (always), maybe also selection, which means that genes and/or genotypes do not perform equally (their fitness-values differ).

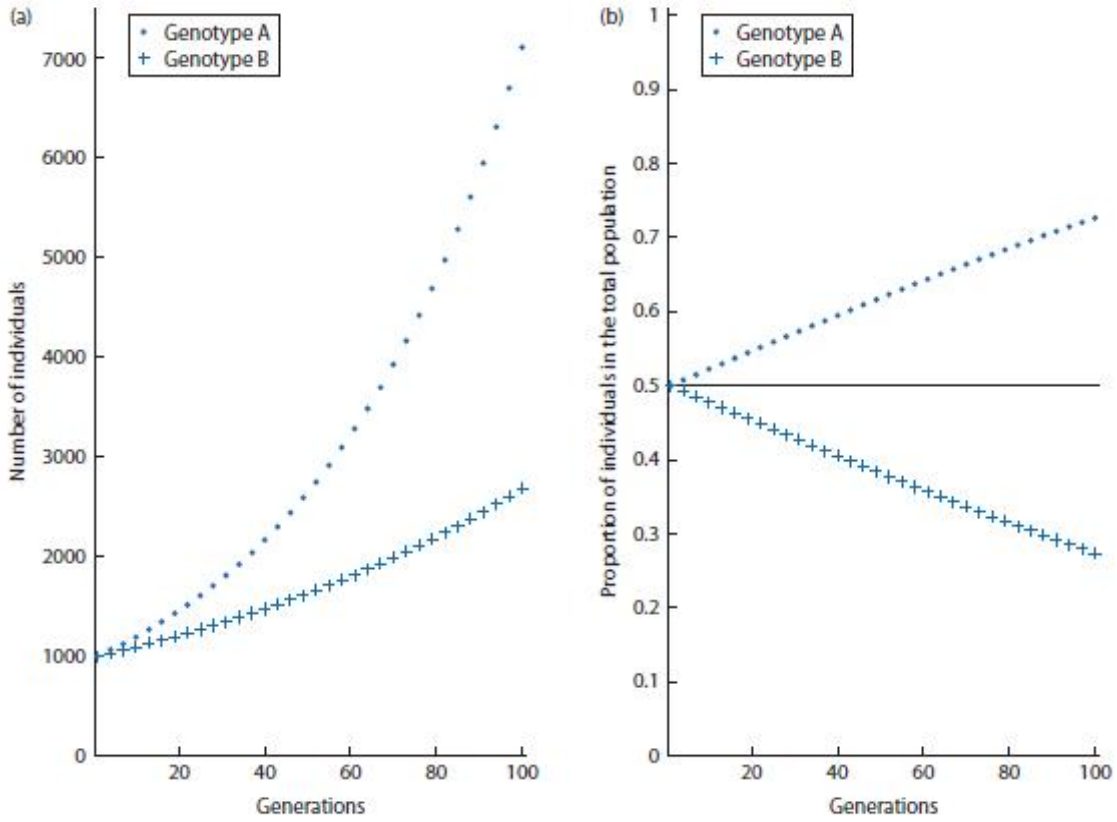
A question of its own is, whether statistical deviation from HWE is a sensitive or practical measure for detecting such evolutionary factors (it seldom is). We don´t consider that here.

● The point is that HWE is a basic model, and based on this model, classical population genetics includes some other basic models.

● In practice (data analysis) the basic selection models have very little value. However, it is good to know about them. Of special importance is the concept balancing selection which maintains polymorphism. The amounts of polymorphisms are enormous and their maintenance is enigmatic. The bulk of variation is (or might be) neutral, but how much, and which parts of them are maintained by selection?

THE BASIC SELECTION MODEL

- Population growth, two haploid genotypes (like bacteria, or gametes of a diploid individual), *N* individuals, *λ* is the finite rate of increase



Genotype A grows 3% per generation (λ = 1.03) and genotype B grows 1% per generation (λ = 1.01).

(a) Individuals of both genotypes increase in number over time.

(b) Because the genotypes grow at different rates, their relative proportions in the total population change over time. The solid line shows the initial equal proportions. Eventually, genotype A will approach 100% and genotype B 0%:

Genotype A is fixed and B is lost and A-B polymorphism is lost.

	Genotype	
	A	B
Generation *t*		
Initial frequency	p_t	q_t
Genotype-specific growth rate (absolute fitness)	λ_A	λ_B
Relative fitness	$w_A = \frac{\lambda_A}{\lambda_A}$	$w_B = \frac{\lambda_B}{\lambda_A}$
Frequency after natural selection	$p_t w_A$	$q_t w_B$
Generation *t* + 1		
Initial frequency p_{t+1}	$\frac{p_t w_A}{p_t w_A + q_t w_B}$	$\frac{q_t w_B}{p_t w_A + q_t w_B}$
Change in genotype frequency	$\Delta p = p_{t+1} - p_t$	$\Delta q = q_{t+1} - q_t$
Generation *t*		
Initial frequency	$p_t = 0.5$	$q_t = 0.5$
Genotype-specific growth rate (absolute fitness)	$\lambda_A = 1.03$	$\lambda_B = 1.01$
Relative fitness	$w_A = \frac{\lambda_A}{\lambda_A} = \frac{1.03}{1.03} = 1.0$	$w_B = \frac{\lambda_B}{\lambda_A} = \frac{1.01}{1.03} = 0.981$
Frequency after natural selection	$p_t w_A = (0.5)(1.0) = 0.5$	$q_t w_B = (0.5)(0.981) = 0.4905$
Generation *t* + 1		
Initial frequency p_{t+1}	$\frac{0.5}{0.5 + 0.4905} = 0.5048$	$\frac{0.4905}{0.5 + 0.4905} = 0.4952$
Change in genotype frequency	$0.5048 - 0.5 = 0.0048$	$0.4952 - 0.05 = -0.0048$

Haploid selection.

The top section of the table gives expressions for the general case. The bottom part of the table uses absolute and relative fitness values to show the change in genotype proportions for the first generation of natural selection. The absolute fitness of the A genotype is highest and is therefore used as the standard of comparison for relative fitness.

- The relative fitness can be used to determine the change in frequency of a genotype over time $\Delta p = p_{t+1} - p_t$. In the haploid example, with relative fitnesses w_A and w_B,

$$\Delta p = p_t w_A / (p_t w_A + p_t w_B) - p_t$$

THE BASIC SELECTION MODEL

- Assumptions of a general model

  - Diploid individuals
  - One locus with two alleles
  - Obligate sexual reproduction
  - Generations do not overlap
  - Mating is random
  - Mechanism of natural selection is genotype-specific differences in (fitness) that lead to variable genotype-specific growth rates,
    termed viability selection
  - Fitness values are constants that do not vary with time, over space,
    or in the two sexes
  - Infinite population size, so there is no genetic drift (stochastics)
  - No population structure
  - No gene flow
  - No mutation

A GENERAL SELECTION MODEL

	Genotype		
	AA	Aa	aa
Generation t			
Initial frequency	p_t^2	$2p_tq_t$	q_t^2
Genotype-specific survival (absolute fitness)	ℓ_{AA}	ℓ_{Aa}	ℓ_{aa}
Relative fitness	$w_{AA} = \frac{\ell_{AA}}{\ell_{AA}}$	$w_{Aa} = \frac{\ell_{Aa}}{\ell_{AA}}$	$w_{aa} = \frac{\ell_{aa}}{\ell_{AA}}$
Frequency after natural selection	$p_t^2 w_{AA}$	$2p_tq_tw_{Aa}$	$q_t^2 w_{aa}$
Average fitness	$p_t^2 w_{AA} + 2p_tq_tw_{Aa} + p_t^2 w_{aa}$		
Generation $t+1$			
Genotype frequency	$\frac{p_t^2 w_{AA}}{\bar{w}}$	$\frac{2p_tq_tw_{Aa}}{\bar{w}}$	$\frac{q_t^2 w_{aa}}{\bar{w}}$
Allele frequency	$p_{t+1} = \frac{p_t(p_tw_{AA} + q_tw_{Aa})}{\bar{w}}$	$q_{t+1} = \frac{q_t(q_tw_{aa} + p_tw_{Aa})}{\bar{w}}$	
Change in allele frequency	$\Delta p = \frac{pq[p(w_{AA} - w_{Aa}) + q(w_{Aa} - w_{aa})]}{\bar{w}}$	$\Delta q = \frac{pq[q(w_{aa} - w_{Aa}) + p(w_{Aa} - w_{AA})]}{\bar{w}}$	

● Expected frequencies of three genotypes in Hardy-Weinberg equilibrium, after natural selection. The absolute fitness of the *AA* genotype is used as the standard of comparison when determining relative fitness.

● The general categories of relative fitness values for selection at a diallelic locus. The selection coefficients (*s*, *hs*, *t*) represent the decrease in viability of a genotype compared to the maximum fitness 1 (fitness = 1 – selection coefficient).

	w_{AA}	w_{Aa}	w_{aa}
Selection against a recessive phenotype	1	1	1 - *s*
Selection against a dominant phenotype	1 – *s*	1 – *s*	1
General dominance	1	1 – *hs*	1 - *s*
Heterozygote disadvantage (underdominance)	1	1 – *s*	1
Heterozygote advantage (overdominance)	1 - *s*	1	1 – *t*

Balancing selection is the often used term for heterozygote advantage.

● Change in allele frequency

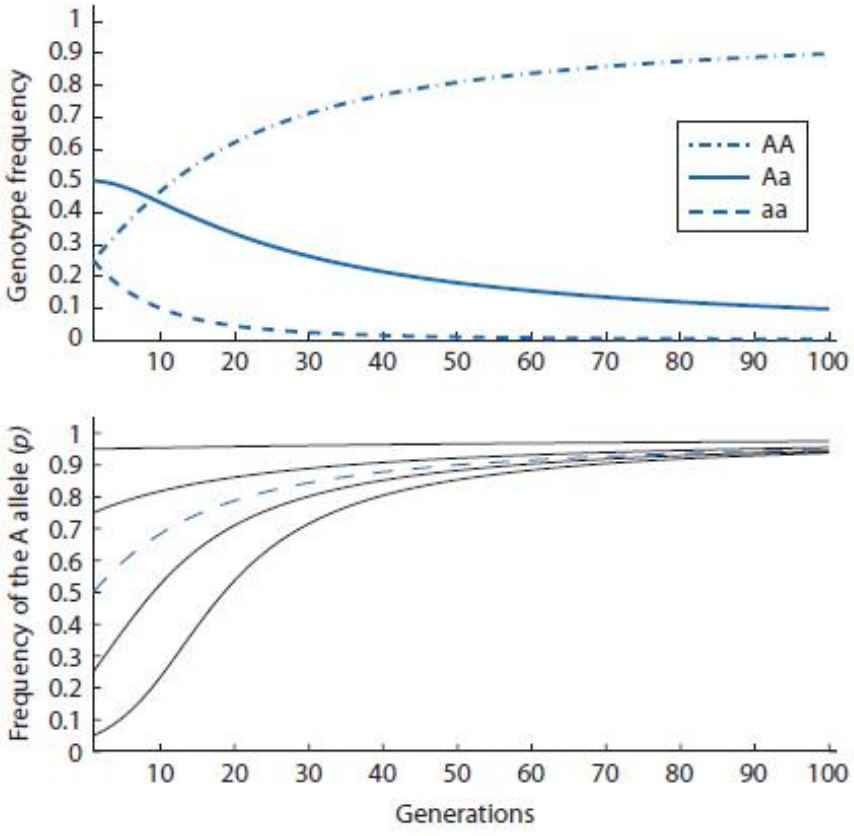
$$\Delta p = [\,(p^2 w_{AA} + pq\, w_{Aa}\,) / (p^2 w_{AA} + 2pq\, w_{Aa} + p^2 w_{aa}\,)] - p$$

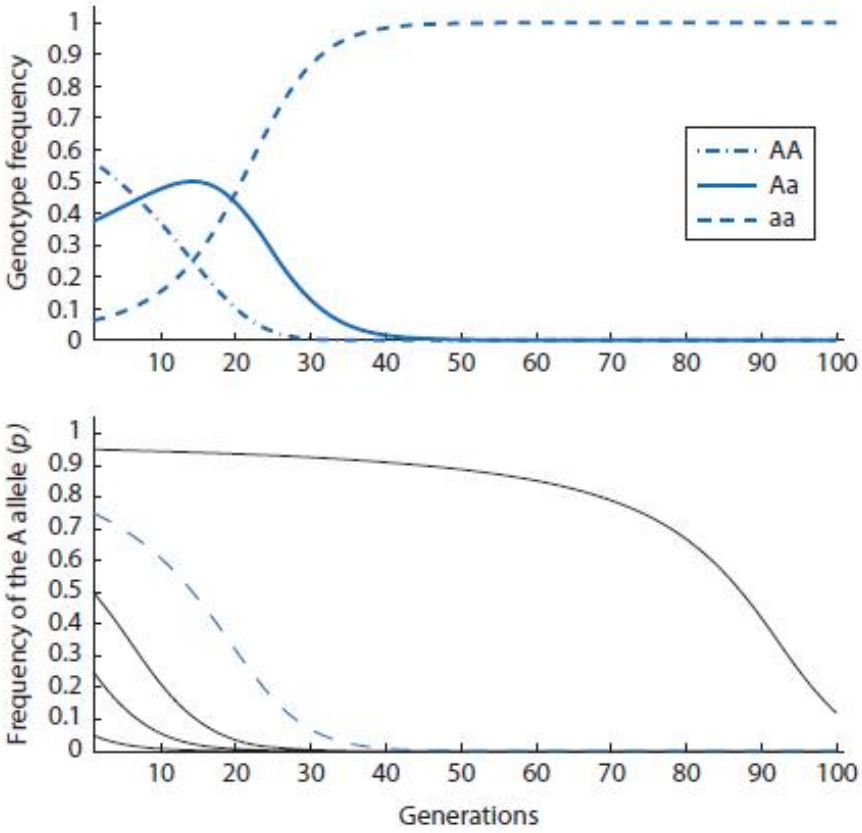
● Equilibrium $\Delta p = 0$.

ALLELE AND GENOTYPE FREQUENCY CHANGES UNDER SELECTION

■ Selection against recessive phenotype. Genotype *aa* has fitness 0.8. In the bottom figure five initial allele frequency conditions.

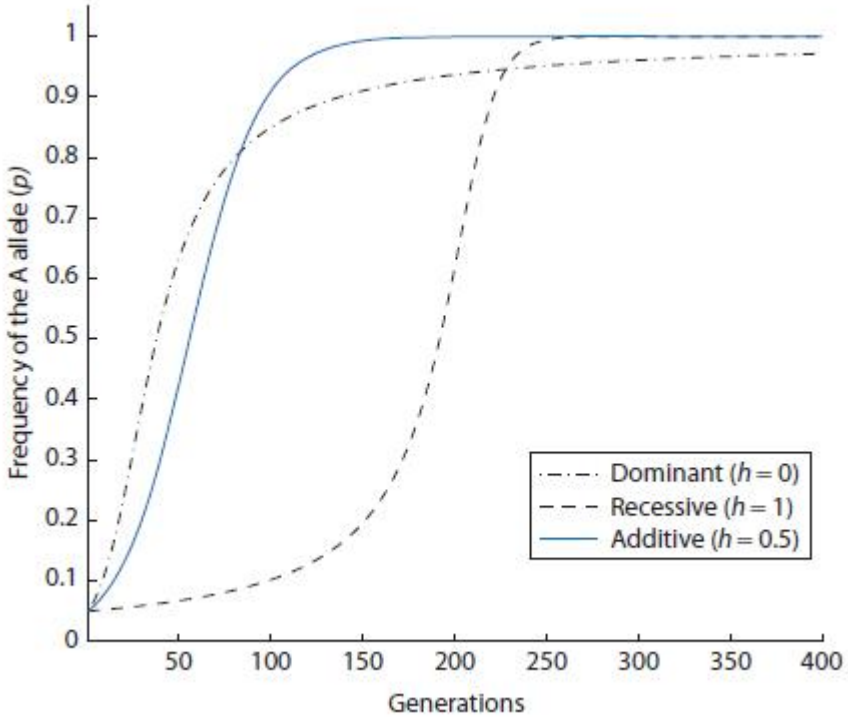
■ Selection against dominant allele. The dominant homozygote and heterozygote have fitness 0.8.



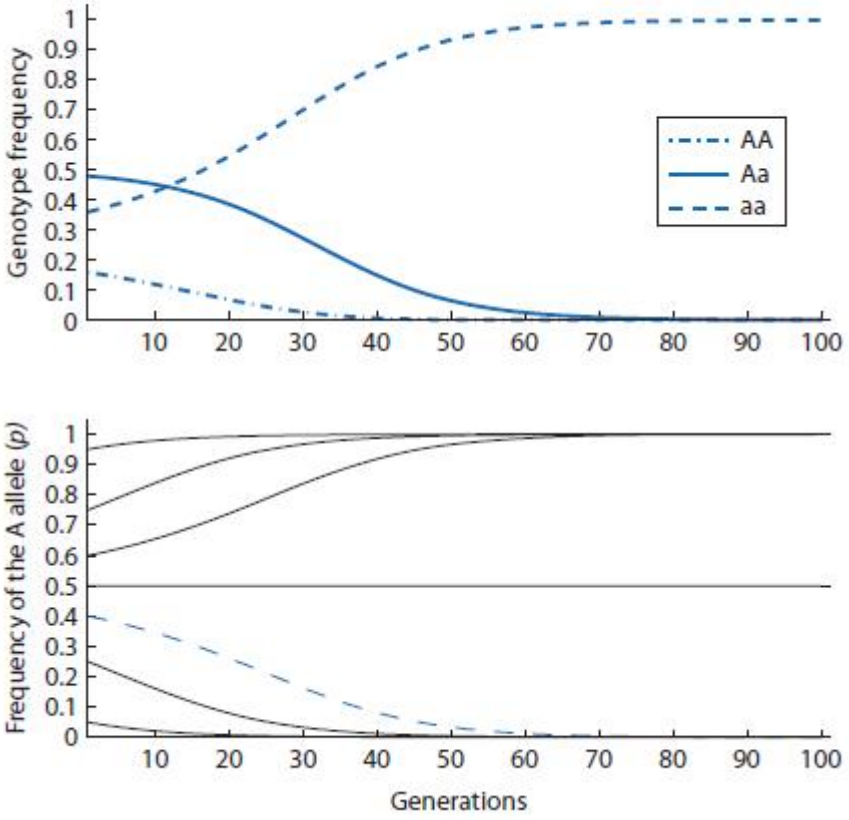


ALLELE AND GENOTYPE FREQUENCY CHANGES UNDER SELECTION

■ General dominance, three cases. In all cases the equilibrium allele frequency is fixation or near fixation of allele *A*.

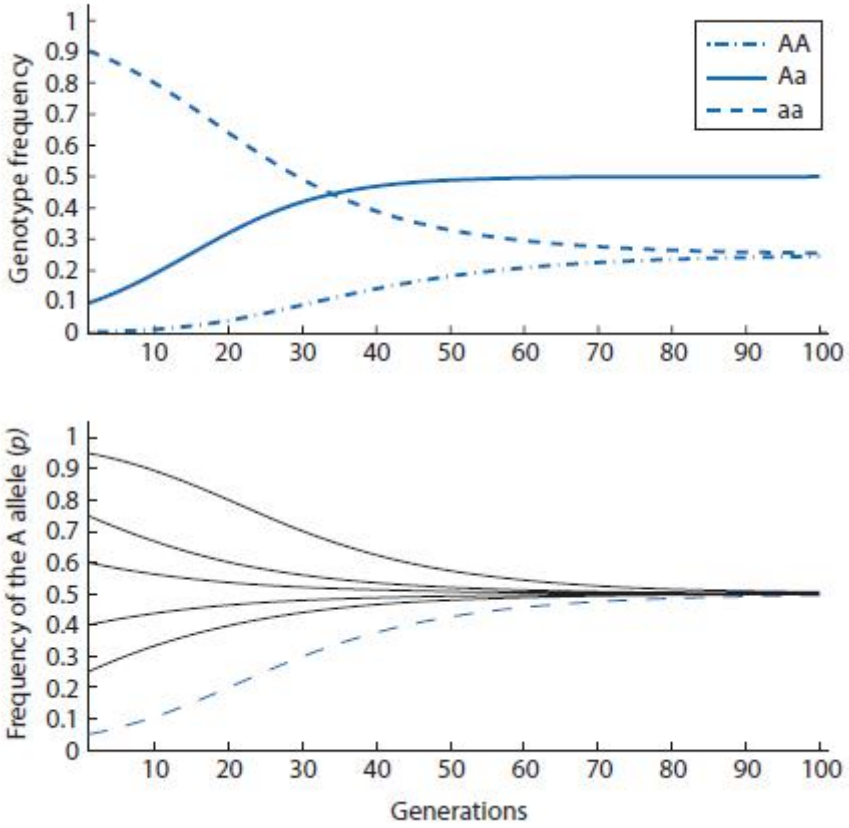


■ Heterozygote disadvantage. *Aa* has fitness 0.9.



ALLELE AND GENOTYPE FREQUENCY CHANGES UNDER SELECTION: OVERDOMINANCE, BALANCING SELECTION

- Overdominance. Heterozygote has fitness 1, and homozygotes have 0.9.



- The classical (simple) selection scheme which results in polymorphic equilibrium = polymorphism is maintained in a population.

- Frequency dependent selection models (fitness-values are functions of allele frequencies, not constants) can be shown to lead to polymorphism maintenance (a genotype becomes less fit as becomes more common). Differing fitness values, temporally or spatially (environment consists of patches, different genotypes are fittest in different patches) also form theoretically valid conditions for polymorphism maintenance.

- Balancing selection (although seldom proved), is a widely used framework for explaining polymorphism maintenance.

2.1. Derive the equilibrium allele frequency formulae for overdominance (see pages 7-8). First you need to construct Δp for this selection scheme and then find the solution for $\Delta p = 0$.

2.2. *A* is the "normal allele" at the the gene locus *β-globin,* but in malarial region of western Africa the *S* allele of this locus is present at a frequency of 0.2. Individuals with genotype *SS* have sickle-cell anemia and have only a 10% chance of surviving to reproductive age relative to heterozygous individuals with the *AS* genotype. Normal individuals with the *AA* genotype at this locus have an 85% cnahce of surviving to reproductive age, relative to *AS* individuals. Assume that at this locus, the genotypes of newborns are in Hardy-Weinberg proportions. If the relative fitness of *AS* is 1, what are the genotype frequencies among individuals of reproductive age? (The sickle-cell anemia is a classical example of a balanced polymorphims.)

2.3. In November 1949 Linus Pauling (two Nobel prices) published in *Science* the paper *Sickle cell anemia, a molecular disease.* In this paper he (with collaborators) showed that hemoglobin from patients suffering from sickle cell anemia had a different electrical charge than that from healtly individuals. This paper was seminal in two ways. First, it showed that the cause of a disease could be traced to an alteration in the molecular structure of a protein, raising the possibility that many other diseases might also be explained in this way. Second, as this disease was known to be inherited, the paper argued that genes precisely determine the structure of proteins. In 1957 it was shown that a single amino acid difference between normal and sickle cell hemoglobin explained electrical charge differences and around that time population genetics captured the polymorphism, (co-existence of normal and sickle-cell alleles) in geographical areas with malaria, as a stable balanced polymorphism.

Malaria has been among the most conspicuous selection pressures acting on the human genome. In *Nature 2010 September 23; vol. 467(7314):420-5* something interesting was published about evolution of malaria. Explain briefly what.

2.4. Find out (use PubMed) other examples of balanced polymorphisms in humans and explain briefly what kind of arguments are given to support their maintenance as balanced polymorphisms. Hints: there are very little concrete evidence (i.e. sickle cell anemia is quite exceptional). There are, however, good reasons to vote for balancing selection as the maintaining evolutionary factor for HLA, ABO...

Answer need not be long, a few sentences are enough!

Using PubMed for screening papers is more practical within UH than at home:

- By performing PubMed-searches of your choice you get lists of publications of which you can usually open only the abstract if you are not at UH and also the full text within UH.

- If you are not at UH you can, of course, pick up a reference from PubMed, log in UH-library, e-journals, get the full text .

MUTATION

● Mutation frequencies range from 10^{-4} to 10^{-6} new mutations per gene per generation. This is the classical view. Estimation of mutation frequencies is extremely difficult, in practice.

● Though mutation rate is (always) very low, it can create many new alleles per genome at population level:
  ● Consider a population of size *N* diploid individuals. There are 2*N* copies of each gene, each of which can mutate in any generation.
  ● Mutation rate (probability of mutation), for example, 10^{-9} per nucleotide pair per generation.
  ● Each gamete, the DNA of which contains approximately $3x10^9$ nucleotide pairs in humans, would contain three new mutations in each generation =>
  each new zygote (the union of two gametes) would carry six new mutations. Human population size is ~7 billion =>
  42 billion new mutations that were not present one generation earlier.

● A single new mutant allele in a diploid population of size *N* has an initial frequency of 1/2*N.*

● If there is exactly one new mutation, then the mutant allele frequency increases 1/2*N,* 2/2*N,* 3/2*N,...* and if *N* is large, the mutant increases in frequency very slowly. Hence, the mutation pressure, changing allele frequencies, is a weak process.

● Recall Hardy-Weinberg (see lecture slides *Basic concepts: Probability, inheritance and population genetics*): Assumptions include: no mutation.
In the following, mutation is permitted.

MUTATION FREQUENCY AND FORWARD - REVERSE MUTATION EQUILIBRIUM

● Denote non-mutated allele is by *A* and mutated allele by *a*. Suppose that *A* (with frequency *p* in the population) mutates to *a* (frequency *q*) at the rate of μ mutations per *A* allele per generation. (Each *A* allele has a probability of μ mutating to *a* in any generation.)

● Allele frequency in generation *t*, $p_t = p_{t-1}(1 - \mu)$, including all the *A* alleles in generation *t* that did not mutate in that generation. By the same reasoning, p_{t-1} includes all *A* alleles in generation $t - 1$ that did not mutate in that generation, and so $p_{t-1} = p_{t-2}(1 - \mu)$ => $p_t = p_{t-2}(1 - \mu)^2$

$$p_t = p_0 (1 - \mu)^t \qquad (1)$$

● The approximation $p_t = p_0 (1 - \mu t)$ is quite accurate for small values of *t.*

● Suppose the rate v for *a* mutating back to *A* per generation. Thus an allele *A* in generation *t* can originate in either of two ways. It could have been an *A* allele in generation $t - 1$ that was not mutated to *a* (which happens with probabaility $1 - \mu$), or it could have been an *a* allele in generation $t - 1$ that mutated to *A* (which happens with probability v)

$$p_t = p_{t-1}(1 - \mu) + (1 - p_{t-1})\, v \qquad (2)$$

$$p_t - v / (\mu + v) = [\, p_{t-1} - v / ((\mu + v)]\,(1 - \mu - v), \ \ldots\ldots \quad = [\, p_0 - v / ((\mu + v)]\,(1 - \mu - v)^t \qquad (3)$$

● Consider equation (3) in the long run, when *t* is very large, for example 10^5 or 10^6 generations. Even though $1 - \mu - v$ is close to 1, the value of *t* eventually becomes so large that $(1 - \mu - v)^t$ becomes close to 0, and so p_t eventually attains a value that remains the same ($\Delta p = 0$). The equilibrium value

$$\hat{p} = v / (\mu - v) \qquad (4)$$

MUTATION – SELECTION BALANCE

- The scheme: selection against a recessive phenotype.

$$\Delta q_{selection} = -sq^2p / (1 - sq^2) \quad (6)$$

- Allele frequency change due to mutation is $\Delta q_{mutation} = \mu p$ (7)

  if we assume that the reverse (backward) mutation rate (ν) is low compared with the forward mutation rate (μ).

- Because the two forces, selection and mutation, have opposite effects on allele frequency, they "balance" each other, and thus, at some time point

$$\Delta q_{selection} + \Delta q_{mutation} = 0 \quad \text{or} \quad sq^2p / (1 - sq^2) = \mu p$$

If q^2 is small, the the denominator of the left-hand side of the expression is ≈ 1 and $q^2 = \mu / s$

- Equilibrium allele frequency at mutation-selection balance

$$\hat{q} = (\mu / s)^{1/2} \quad (8)$$

GENETIC DRIFT

● Next we permit stochastics, i.e. chance effects, i.e. random drift as relaxation of Hardy-Weinberg assumptions. This means the a population has a size, *N.*

● Example for illustrating chance effects = random genetic drift as an evolutionary factor.

● Consider a large population in HWE with alleles *A* and *a* at equal frequencies $p = q = ½$ The genotypes frequencies are thus (HWE assumption) ¼ *AA,* ½ *Aa,* ¼ *aa.* Suppose that something dramatic happens and only four randomly chosen individulals survive. Subsequent generations are based on this small sample from the original large population. It is possible, by chance, that the four survivors are all *AA* individuals.

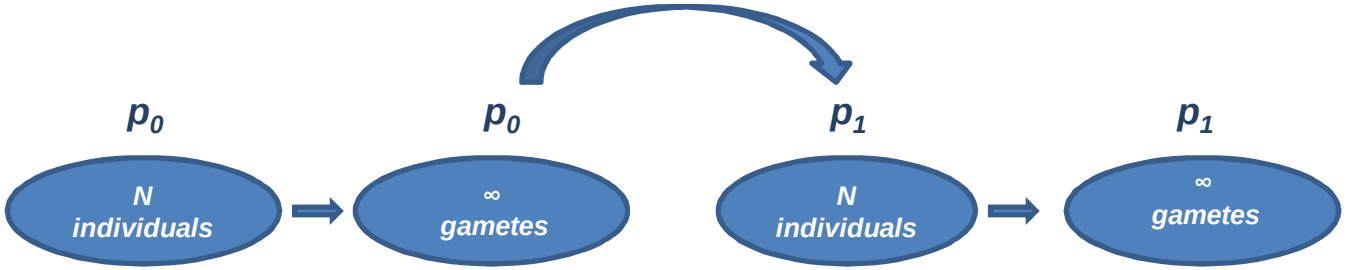
● Probability of this possibility is $(¼)^4 = 1/256$, similarly probabilities for other possibilites…. If the size of the new population remains at four individuals in each subsequent generation, this type of random sampling occurs repeatedly. In each generation the sampling process can cause large allele (and genotype) frequency changes and one consequence of random drift soon becomes true: the population has only *A* or *a* alleles and population reaches a fixation state. Only new mutations or migration from another population can reintroduce the polymorphism, which was: segregation of two alleles, *A* and *a* in the original large population. If mating takes places at random, sampling four diploid individuals as equivalent to sampling eight haploid gametes. In the example ($p = ½$) there are nine possible outcomes, having 0, 1, 2, 3, …. 8 copies of the *A* allele and the remaining copies being *a.*

BINOMIAL SAMPLING OF GENOTYPES AND ALLELES

● The probability of each of the nine possibilities is given by the binomial distribution, corresponding to the succesive terms in the expansion of $(\frac{1}{2} A + \frac{1}{2} a)^8$.

● The probability of fixation of the *A* allele in the next generation corresponds to the probability of drawing eight copies of *A*, $(\frac{1}{2})^8 = 1/256$ (because each successive draw is considered independent and has a chance of ½ of yielding *A*.) The result is identical to the probability of drawing four *AA* genotypes (see above) and illustrates the principle that, with random mating, random sampling of diploid individuals is equivalent to random sampling of twice as many haploid gametes.

● The process of sampling gametes from a finite population. Sample 2N gametes:



● HWE expectations, with the exception of an assumption of finite population size: there are now *N* individuals, not infinitely large number of individuals. Note that the gamete pool in infinitely large. Sampling process yields a binomial distribution of all possible combinations of *A* and *a.*

● Example: A population of nine diploid individuals arises from a sample of 18 gametes, but the gametes can be thought of as being sampled from an infinite pool of gametes. Because small samples are not representative, an allele frequency in the sample may differ from that in the pool of gametes.

BINOMIAL SAMPLING OF GENOTYPES AND ALLELES

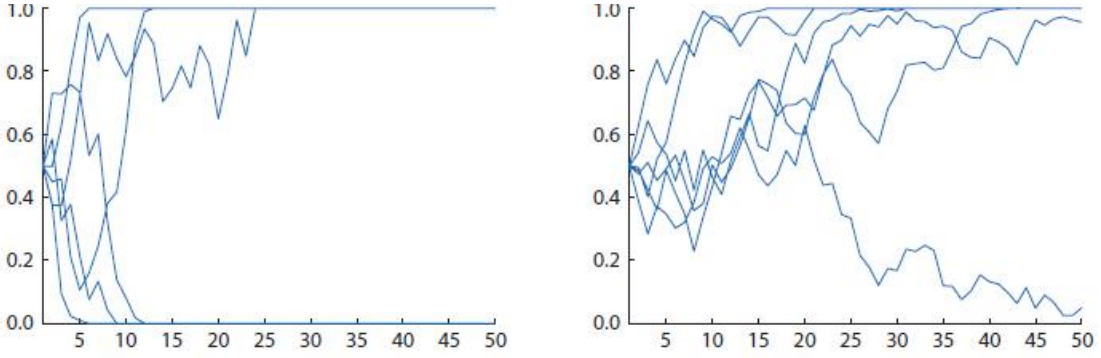
● Suppose that a pool contains *A* and *a* at frequencies *p* and *q* (*p* + *q* = 1). If 2*N* gametes are drawn at random to produce the zygotes of the next generation, the probabaility that the sample contains exactly *i* alleles of type *A* is the binomial probability

$$P_{(i=A)} = \binom{2N}{i} p^i q^{2N-i} \quad \text{where} \quad \binom{2N}{i} = \frac{(2N)!}{i!(2N-i)!} \tag{9}$$

i can take any integer value between 0 and 2*N*

● In the next generation the sampling process occurs anew according to this equation with *p* replaced by *p´* . Allele frequencies change at random from generation to generation.

● Random sampling simulations. Individual populations behave very erratically. In some populations the allele *A* becomes fixed (*p* = 1), in some lost (*p* = 0) during a small number of generations and in some populations both alleles remain unfixed, i.e. segregating. Left N=4, right N=20. x-axis is allele frequency (starting *p*=*q*=0.5 in all cases, y-axis is the number of generations.



WHAT HAPPENS TO A NEW MUTATION

- Consider the $2N$ alleles in generation t (cf. the N individuals in scheme at page 7)

- Each allele is assigned a unique label: α_1 , α_2 , α_3 , ..., α_{2N}
(at the moment we are not interested their status as *A* or *a*)

- In the gamete pool each labeled allele has a frequency of $1/(2N)$.

- Consider the genotypes in $t + 1$ formed by random sampling from the pool of gametes.

- By chance, the two alleles forming a genotype may be replicates of the same allele in the previous generation, for example $\alpha_i \alpha_i$ or they may come from different alleles in the previous generation, for example $\alpha_i \alpha_j$

- Random sampling from the gamete pool means that some alleles may be overrepresentated in generation $t + 1$, relative to their frequency in generation t, some underrepresentated. Any particular allele has also a chance of being unrepresented in $t + 1$ which means that the lineage of that allele is terminated.

- Each allele in generation t has a chance $e^{-1} = 0.368$ of not being represented in generation $t + 1$:

  - Consider the allele α_1
  - It´s frequency in the gamete pool is $1/(2N)$

  and the frequency of all other alleles together is $1 - 1/(2N)$

  - Because the genotypes in generation $t + 1$are formed by the random selection of $2N$ alleles from the pool of gametes, the distribution of the number of α_1 and non - α_1 alleles present in generation $t + 1$ is given by succesive terms in the binomial distribution

$$[(1 / 2N) \alpha_1 + (1 – ((1 / 2N)) \alpha]^{2N} \quad (10)$$

(α represents the collection of all alleles other than α_1)

- Hence, the probability that α_1 is not represented in generation $t + 1$ is

$$[1 – 1 / 2N]^{2N} \approx e^{-1} = 0.368 \quad (11)$$

- The approximation is quite good even when N is quite small. For example, when $N = 10$, the left-hand side of (11) is 0.358, and when N = 20, 0.363.

WHAT HAPPENS TO A NEW MUTATION

- The important implication is that an ancestral lineage of each allele has a substantial risk of extinction in each generation. As time goes on, the lineages progressively disappear, one or a few at a time.

- Eventually, a time is reached at which all lineages – except one – have become extinct.

- At that time, every allele in the population is identical by descent with a particular allele present in an ancestral population.

- The ultimate extinction of all but one lineage implies the answer to this: What is the probability that a single new mutation eventually becomes fixed in a population of size $2N$?

    - The answer is: $1 / (2N)$

- How to explain large amounts of genetic polymorphisms in populations?

    - In *Motoo Kimura´s* (1968) *neutral theory* the bulk of polymorphisms is modelled as a balance between the mutation pressure and random genetic drift.

    - Mutation introduces new alleles into a population, and random drift determines whether a neutral allele will ultimately be fixed or lost – loss being the usual outcome.

    - At equilibrium, there is a balance: on the average, each new allele gained by mutation in balanced against an existing allele that is lost.

EFFECTIVE POPULATION SIZE, N_e

- The effective population size of an actual population is the number of individuals in a *theoretically ideal population having the magnitude of random genetic drift as the actual population.*

- Three kind of effective population size definitions, based on how to measure the ”magnitude”

    - The change in probability of identity by descent.
    - The change in variance in allele frequency.
    - The rate of loss of heterozygosity.

- Derivations (classical population genetics) not given here. A couple of examples:

- Assume that a population is ideal with the exception that its size is not constant.

$$1/N_e = 1/t\ [\ 1/N_0 + 1/N_1 + \ldots + 1/N_{t\text{-}1}\ ] \qquad (12)$$

I.e. it can be shown that the ”correction term” for N_e from the actual numbers in generations 0, 1, …. is the harmonic mean.

- Assume that sex ratio is unequal, N_m males and N_f females.

$$N_e = 4\, N_m N_f \ /\ (N_m + N_f) \qquad (13)$$

MODELLING SUBSTITUTION DYNAMICS

● The probability that a particular allele will become fixed in a population depends on its frequency, its fitness advantage or disadvantage, i.e. (Darwinian) selection increasing or decreasing its frequency, the effective population size N_e which affects the sampling process (small population => more changes by mere chance, ´random drift´, larger population => selection (fitness differences) outcompetes chance effects.

● In classical population genetics (especially Motoo Kimura´s neutral theory of molecular evolution) the following selection scheme

(s is the selection advantage)

genotypes	A_1A_1	A_1A_2	A_2A_2
fitness	1	1+s	1+2s

● The probability of fixation of A_2 is

$$P = (1 - e^{-4Nesq}) / (1 - e^{-Nes}) \qquad (14)$$

where q is the initial frequency of allele A_2

- Since $e^{-x} \approx 1 - x$ for small values of x, the equation reduces to $P \approx q$ as s approaches 0.

- Thus, for a neutral allele, the fixation probability equals its frequency in the population. For example, a neutral allele with a frequency of 40% will become fixed in 40% of the cases and will be lost in 60% of the cases, fixation occurs by random drift, which facilitates neither allele.

- A new mutant arising as a single copy in a population of size N individuals has an initial frequency of $1 / (2N)$ (diploid individuals). The probability of fixation of a particular mutant allele is thus obtained by replacing q with $1 / (2N)$ in the previous equation. When $s \neq 0$,

$$P = (1 - e^{-(2Nes/N)}) / (1 - e^{-4Ns}) \quad (15)$$

For a neutral mutation, $s = 0$, the equation becomes $P = 1 / 2N$ (16)

- If the population size is equal to the effective population size

$$P = (1 - e^{-2s}) / (1 - e^{-4Ns}) \quad (17)$$

- If the absolute value of s is small $P = 2s / (1 - e^{-4Ns})$ (18)

- For positive values of s and large values of N, $P \approx 2s$ (19)

● Thus: if an advantageous mutation arises in a large population and its selective advantage over the rest of the alleles is, say up to 5%, the probability of its fixation is approximately twice its selective advantage. For example, if a new mutation with s = 0.01 arises in a population, the probability of its eventual fixation is 2%. The message: Advantageous mutations do not always become fixed in the population. In fact, 98% of all the mutations with the selective advantage of 0.01 will be lost by chance. Deleterious mutations have a finite probability of becoming fixed in a population, albeit a small one. The fact that a deleterious allele *may* become fixed in a population at the expense of better alleles illustrates the importance of chance events in determining the fate of mutations during evolution.

● Considering larger population, the chance effects become smaller. For example, if N = 10 000, then the fixation probabilities become 0.005%, 2% and ~ 10^{-20}.

● Thus, while the fixation probability for the advantageous mutations remains approximately the same, that for the neutral mutation becomes smaller, and that for the deleterious allele becomes indistinguishable from zero.

● The fixation time. Consider fixation and loss separately and restrict consideration to those mutants that will eventually become fixed in the population. This is called the *conditional fixation time.*

●In the case of a new mutation whose initial frequency is $q = 1/(2N)$, the mean conditional fixation time (theory by Kimura in 1960´s) for a neutral mutation is approximated by

$$t = 4N \text{ generations} \quad (20)$$

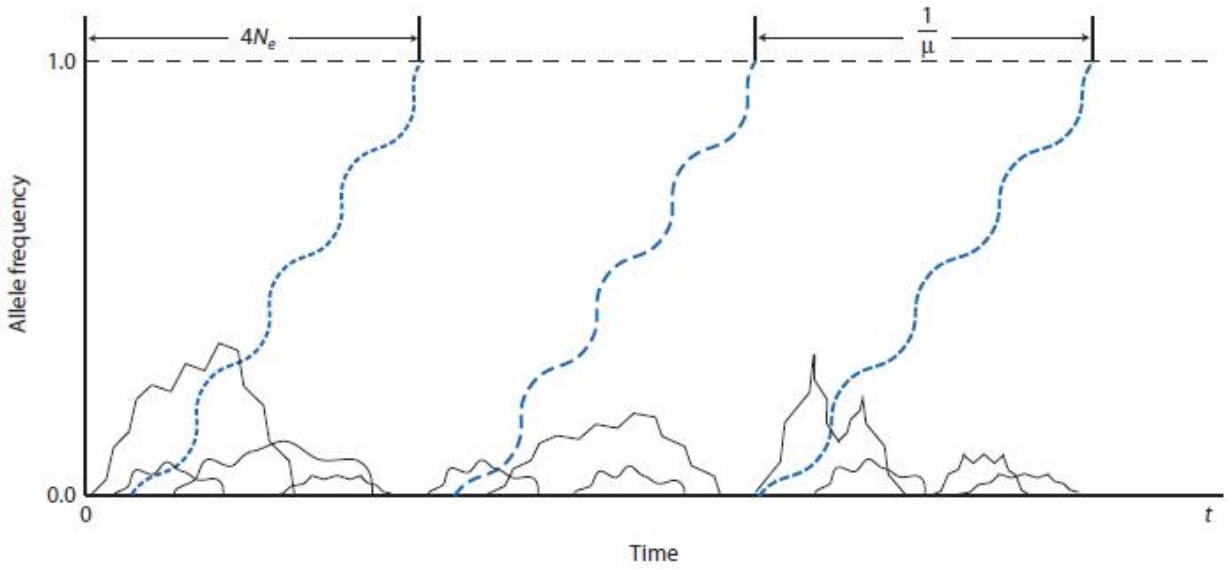
● For a mutation with a selective advantage of s

$$t = (2/s)\ln(2N) \text{ generations} \quad (21)$$

MODELLING SUBSTITUTION DYNAMICS

- Let´s assume a species which has an effective population size of about 10^6 and a mean generation time 2 years.
- Under these conditions, it will take a neutral mutation, on average 8 million years to become fixed in the population.
- A mutation with a selective advantage of 1% will become fixed in the population in 5800 years.
- The conditional fixation time for a deleterious allele with a selective disadvantage –s is the same as that for an advantageous allele with a selective advantage s (theory by Kimura in the 1970´s).
- This is intuitively understandable *given the high probability of loss for a deleterious allele. That is, for a deleterious allele to become fixed in a population, fixation must occur very quickly.*

THE FATE OF SELECTIVELY NEUTRAL MUTATIONS

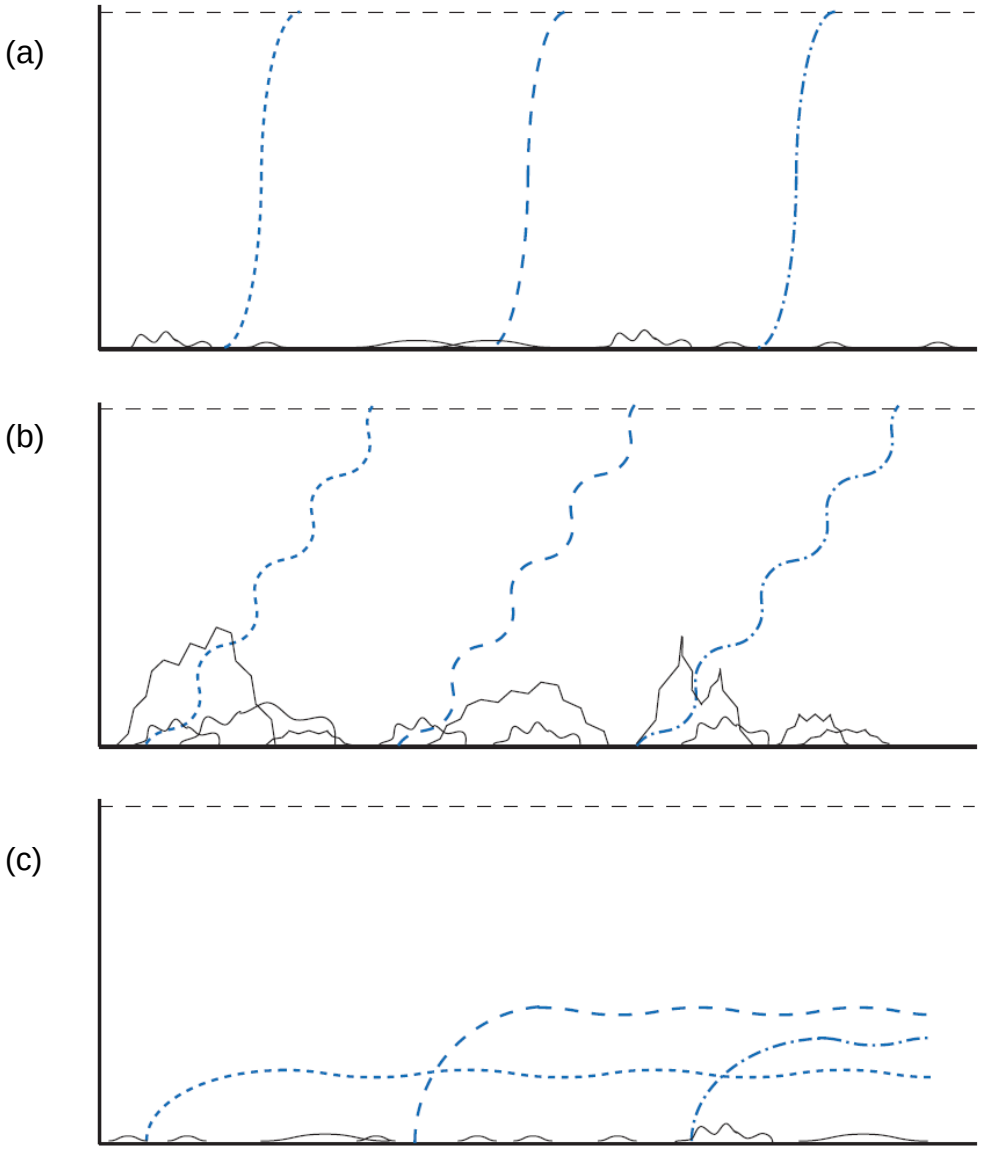


New mutations enter the population at rate μ and an initial frequency of $1/2N$. Allele frequency is a random walk (random drift). The time that a new mutation segregates in the population, or the dwell time of a mutation, depends on the effective population size. However, the chance that a new mutation goes to fixation (equal to its initial frequency) is also directly related to the effective population size.

These two effects of the effective population size cancel each other out for neutral alleles.

The neutral theory then predicts that the rate of fixation is μ and therefore the expected time between fixations is $1/\mu$ generations. For that subset of mutations that eventually fix, the expected time from introduction to fixation is $4N_e$ generations.

MUTATIONS DRIVEN BY SELECTION VS. NEUTRAL MUTATIONS



The dwell time for new mutations is different if fixation and loss is due to genetic drift or natural selection.

With neutral mutations (b), most mutations go to loss fairly rapidly and a few mutations eventually go to fixation. For eventual fixation or loss of neutral mutations the path to that outcome is a random walk, implying that the time to fixation or loss has a high variance.

For mutations that fix because they are advantageous (a), directional selection fixes them rapidly in the population. Therefore under directional selection alleles segregate for a shorter time and there is less polymorphism than with neutrality.
For mutations that show overdominance for fitness, natural selection favoring heterozygote genotypes maintains several alleles in the population indefinitely. Therefore balancing selection (see below, page 27) in greatly increases the segregation time of alleles and increases polymorphism compared to neutrality.

Both cases of natural selection (a and c) are drawn to show negative selection acting against most new mutations. If new mutations are deleterious then the time to loss is very short and there is very little random walk in allele frequency since selection is nearly deterministic.

KIMURA´s MODEL ON SUBSTITUTION DYNAMICS: THE RATE OF MOLECULAR EVOLUTION

● The rate of gene substitution. Definition: the number of mutants reaching fixation per unit time. If *neutral mutations* occur at a rate of u per gene per generation, then the number of mutants arising at gene locus in a population of size N is $2Nu$ per generation.

● Since the probability of fixation for each of these mutations is $1 / (2N)$, the rate of substitutiton of neutral alleles is obtained by multiplying the total number of mutations by the probability of their fixation

$$K = 2Nu\,(1/2N) = u \qquad (22)$$

● The rate of substitution is thus equal to the rate of mutation.

● Intuitively: in a large population the number of mutations arising every generation is high, but the fixation probability is low. In a small population the number of mutations arising every generation in low, but the fixation probability of each mutation is high. As a consequence, the rate of substitution for neutral mutations is independent of population size.

● For *advantageous mutations* the rate of substitution can also be obtained by multiplying the rate of mutation by the probability of fixation for advantageous alleles as given above ($P \approx 2s$). For selection with $s > 0$

$$K = 4Nsu \qquad (23)$$

KIMURA ´s MODEL ON SUBSTITUTION DYNAMICS: THE RATE OF MOLECULAR EVOLUTION

● The rate of substitution depends on the population size, selective advantage and mutation rate. The inverse of *K* is the mean time between two consequtive fixation events (see the figures above)

● The formulae for *K* have been very important in the field of molecular evolution.
They define two very different predictions for genetic polymorphisms in populations. Are the polymorphisms (for example sequence polymorphisms of certain genes) outcomes from *neutral evolution* or from *evolution dictated by selection??* Two opposite schools since ~1970´s.

● Today, when sequence information is extensive, the question has turned to: How to identify genes from, for example, human sequence databases, which bear signatures of positive selection?

● In phylogenetic studies such genes have evolved faster => biased, “too long” branch lengths in phylogenetic trees. During the evolution of the human lineage, as compared with our phylogenetic relatives (the great apes), positively selected genes might be responsible for some important human specific phenotypic traits.

ASSIGNMENT SET 2 continues

2.5. The bacterium *Salmonella enterica* has a genetic switching mechanims that regulates the production of alternative forms of a protein component of the cellular flagella. There are two alleles: *A* (for the ”specific-phase” flagellar property) and *a* (for the ”group-phase” flagellar property. Switching back and forth between *A* and *a* takes places rapidly enough that equation (3) can be applied.
The transition from *A* to *a* has a rate of $\mu = 8.6 \times 10^{-4}$ per generation and that of *a* to *A* has a rate of $v = 4.7 \times 10^{-3}$ per generation.
These rates are orders of magnitude larger than mutation rates typically observed.
In fact, these changes do not result from mutations in the conventional sense, but from intrachromosomal recombination events.
Formally, however, the system can be treated as one with mutation – reverse mutation.

In a classical expriment (Stocker 1949) it was found that the frequency, initially $p_0 = 0$ (for *A*) , increased to $p = 0.16$ after 30 generations and to $p = 0.85$ after 700 generations.
In experiments initiated with $p_0 = 1$, the frequency decreased to 0.88 after 388 generations and to 0.86 after 700 generations.

How do these valueas agree with those calculated with equation (3) using the estimated mutation rates? What is the predicted equilibrium frequency of *A* ?

ASSIGNMENT SET 2 continues

2.6. Cystic fibrosis (CF) is a Mendelian recessive disease of humans caused by defects in ion transport. Until the 1950s, when antibiotics were first used to treat CF patients, most newborns with CF died at an early age. Yet CF is relatively common in Caucasians, with a frequency at birth of 1/2500, which implies that the frequency of CF-causing mutations is about 0.02 – a surprisingly high frequency of an allele that is lethat to homozygotes. There is no agreement on the reason for this high frequency.

a) Suppose that an allele that causes CF is maintaned by mutation-selection balance. What would be the mutation rate necessary for that allele to have a frequency of 0.02?

b) Suppose that an allele that that causes CF is maintaned by heterozygote advantage. In order for the equilibrium frequency to be 0.02, what would the difference between the viabilities of the homozygote and the heterozygote have to be?

2.7. Equation (8) gives the mutation-selection equilibrium allele frequency for a recessive trait. Derive the corresponding equation for a general dominance case.

2.8. Suppose a population went through a bottleneck as follows: $N_0 = 1000$, $N_1 = 10$, $N_2 = 1000$. Calculate the effective population size.

2.9. What is the effective population size in a population of African lions, *Panthera leo,* in which ecah breeding male controls a harem of five females and the total population consists of 200 males and 200 females?

2.10. A new mutant arises in a population of 1000 individuals. For simplicity, assume that $N_e = N$ (=1000). What is the probability that this allele will become fixed in the population if

(a) it is neutral,

(b) it confers a selective advantage of 0.01,

(c) it has a selective disadvantage of 0.001?

Simulate these cases by using http://www.radford.edu/~rsheehy/Gen_flash/popgen/

INTRO TO (HUMAN) DNA-POLYMORPHISMS

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

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

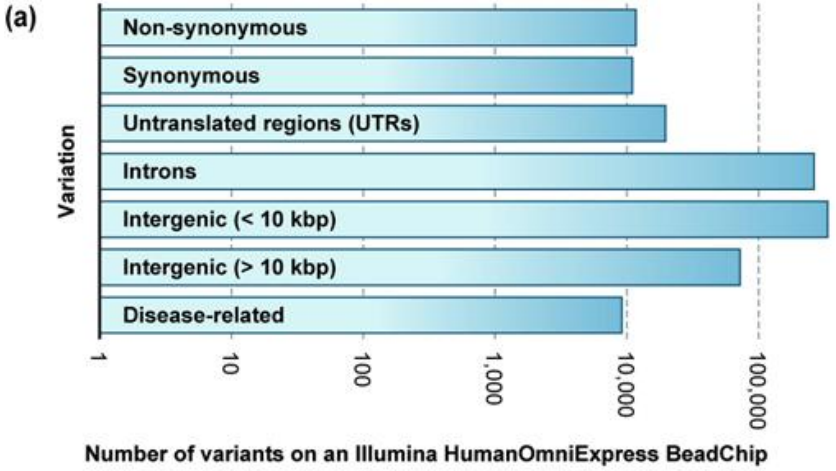
● General public have begun to gain access to their genetic variation profiles by using direct-to-consumer DNA tests available from commercial vendors, which profile hundreds of thousands of genomic markers for low costs. Through this genetic profiling, individuals hope to learn about not only their ancestry, but also genetic variations underlying their physical characteristics and predispositions to diseases.

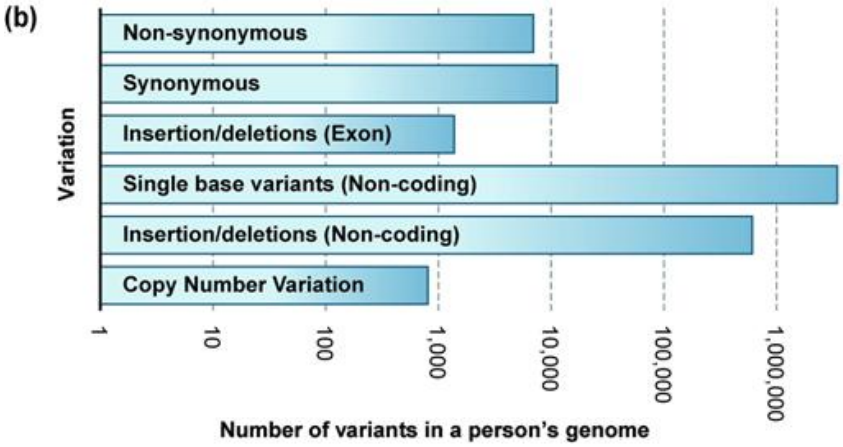
● Majority of the known disease-associated variants are found within protein-coding genes with genome-wide association studies beginning to reveal also thousands of non-coding variants. Proteins are encoded in genomic DNA by exon regions, which comprise just ~1% of the genomic sequence, Exome. This is best understood part: how DNA blueprint sequence relates to function, and is arguably the best chance to connect genetic variations with disease pathophysiology.

A person's exome carries about 6,000 – 10,000 amino-acid-altering nonsynonymous SNVs, nSNVs, known to be associated with more than a thousand major diseases .

Based on Kumar et al. 2011, Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations Trends in Genetics 27: 377-386

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING



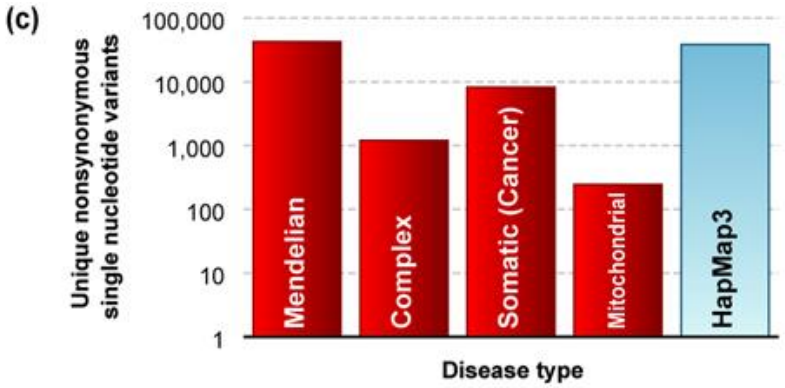


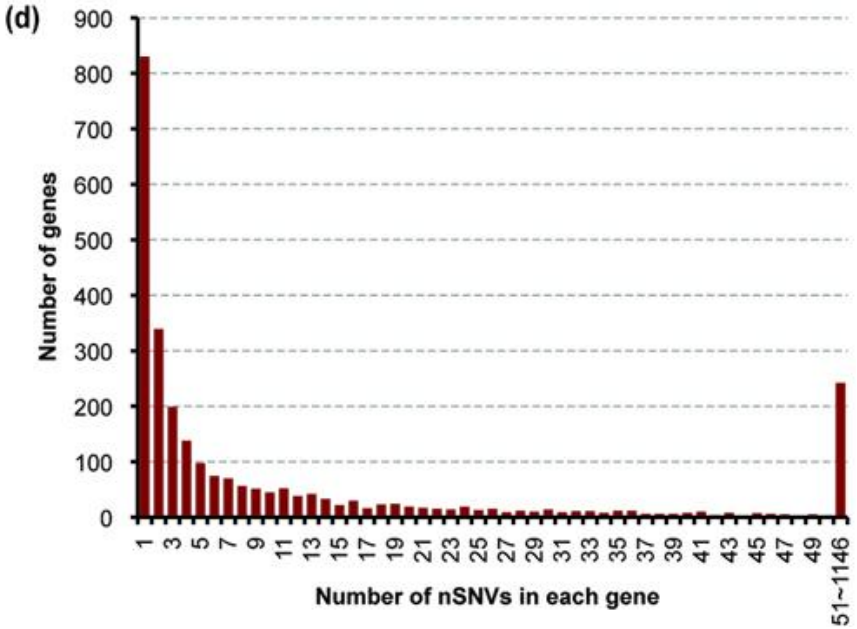
Profiles of personal and population variations.

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

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

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING





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

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

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

nSNV´s IN HUMAN GENOMES – INDIVIDUAL GENETIC PROFILING

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

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

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

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

Mendelian (monogenic) diseases

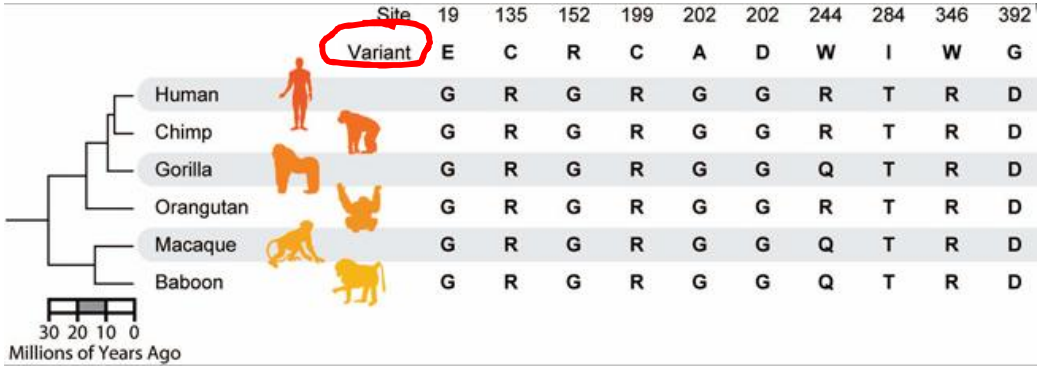
● For centuries it has been known that particular diseases run in families, notably in some royal families where there was a degree of inbreeding. Once Mendel's principles of inheritance became widely known in the early 1900s it became evident from family genealogies that specific heritable diseases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

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

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

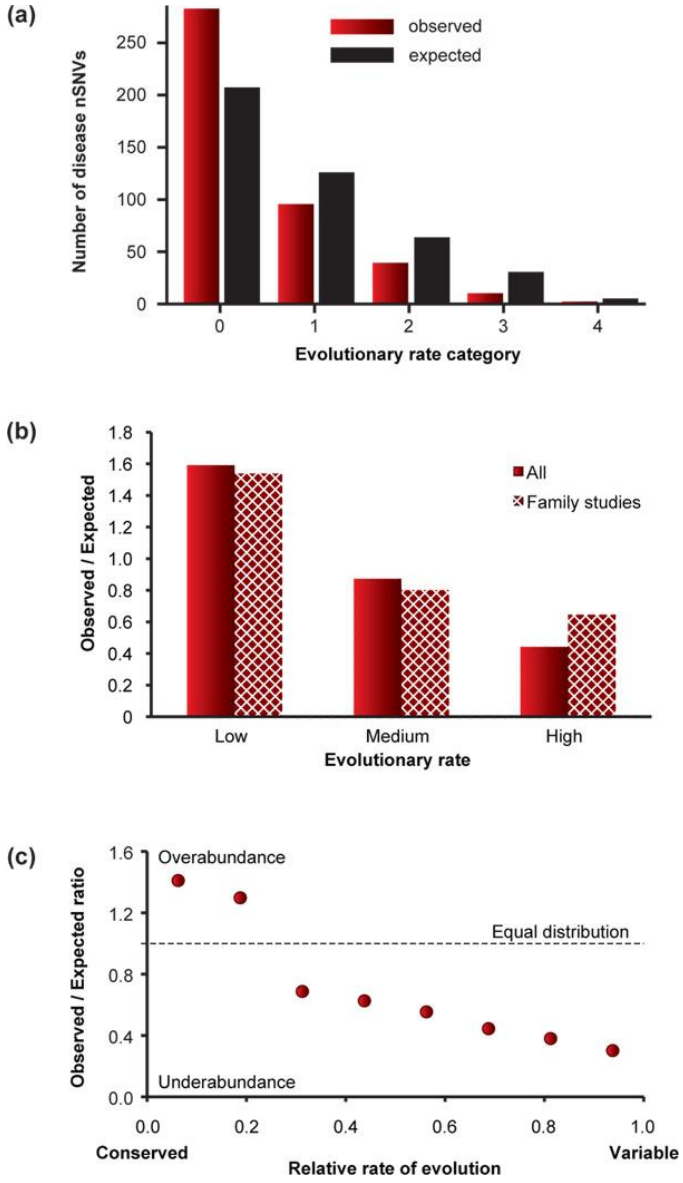
An example

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



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

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



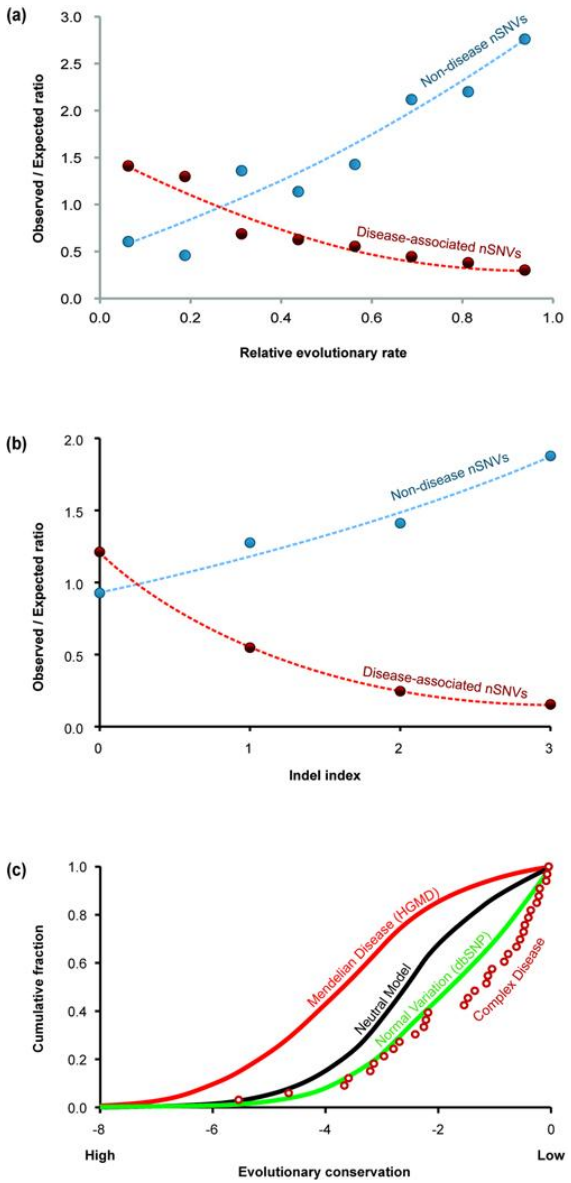
Evolutionary properties of positions afflicted with disease-associated nonsynonymous single nucleotide variants (nSNVs).

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

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

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

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES



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

(a) evolving with different rates and

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

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

nSNV´s IN HUMAN GENOMES – MENDELIAN DISEASES

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

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

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

Identification of deleterious mutations within three human genomes

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

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

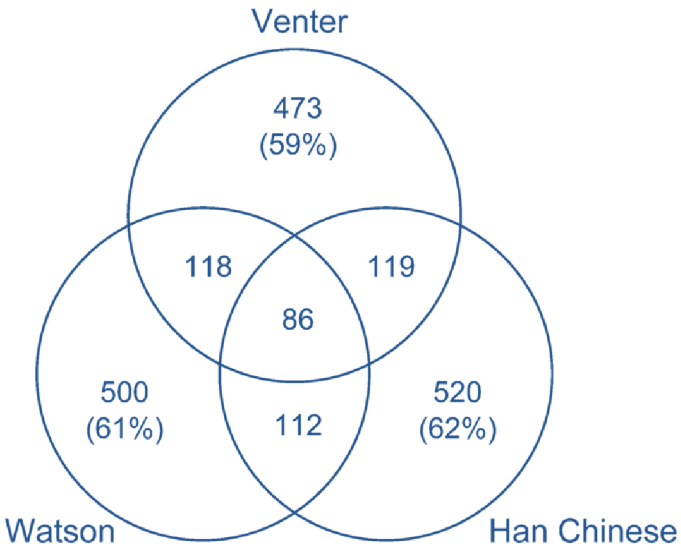
Each human carries a large number of deleterious mutations. Together, these mutations make a significant contribution to human disease. Identification of deleterious mutations within individual genome sequences could substantially impact an individual's health through personalized prevention and treatment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

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

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES

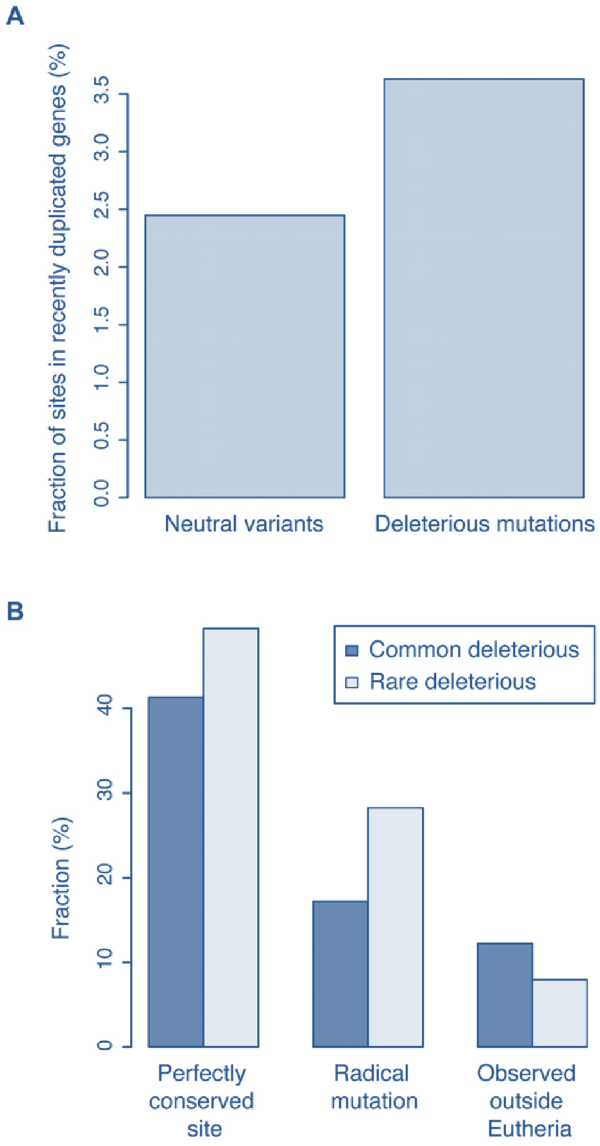
- The paper has data from a complete catalog of SNPs from J. Craig Venter (Google who he is if you don´t know), from a Han Chinese male from their respective websites (http://www.jcvi.org/cms/research/projects/huref/ and http://yh.genomics.org.cn), and for James D. Watson (from "Watson – Crick")

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



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

DELETERIOUS MUTATIONS IN THREE HUMAN GENOMES



Characteristics of deleterious mutations.

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

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

EXAMPLE 1

OPEN ACCESS Freely available online

PLoS GENETICS

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

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

Abstract

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

PLoS Genetics | www.plosgenetics.org

May 2008 | Volume 4 | Issue 5 | e1000083

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

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

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

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

1562 Genome Research
www.genome.org

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

UNDERSTANDING EVOLUTIONARY PATTERNS OF MUTATIONS