BIO00056I Summary of main topics & core concepts

Evolutionary Genetics & Population Genomics

Ferrari, Dasmahapatra, Harper, Jeffares

2026-09-17

Table of contents

1 Deterministic vs. Stochastic Evolutionary Processes

- **Definition of Evolution**: Fundamental change in allele or genotype frequencies in a population from one generation to the next.
- **Four Primary Mechanisms**: Allele frequency changes are driven by natural selection, genetic drift, mutation, and gene flow.
- **Natural Selection (Deterministic)**: Acts non-randomly on fitness differences between phenotypes, requiring phenotypic variation, a link between phenotype and fitness, and a genetic basis (heritability). Drives adaptation and typically reduces genetic variation.
- **Mutation (Stochastic)**: The ultimate random source of all new genetic variation, occurring at very low frequencies.
- **Genetic Drift (Stochastic)**: Random fluctuations in allele frequencies between generations due to sampling error, leading to a loss of genetic diversity—particularly in small populations.
- **Gene Flow / Migration**: Spatial movement between populations that reduces genetic divergence between subpopulations and restores variability.
- **Null Models**: Hardy-Weinberg equilibrium provides the baseline null model to predict expected genotype frequencies when no evolutionary forces act.

---

2 Molecular Mechanisms of Mutation and Substitution

- **Mutation vs. Substitution**: A mutation is any newly arising genetic change in an individual; a substitution is a mutation that has successfully spread to fixation across the population.
- **Point Mutations**: Base-pair substitutions in coding regions are classified as synonymous (silent) or non-synonymous (amino acid altering).
- **Indels & Reading Frames**: Insertions and deletions (indels) within coding sequence regions can alter the reading frame via frameshift mutations.
- **Structural & Chromosomal Alterations**: Gene duplications create multigene families and pseudogenes (e.g., human β-globin gene family). Larger structural mutations include transposons, duplications, rearrangements, inversions, translocations, fission, fusion, and ploidy changes.
- **Mutational Fates**: The fate of a new mutation (fixation, loss, or balancing) is governed by selection and drift; most new neutral or deleterious mutations are rapidly lost to drift or purifying selection.

3 Genetic Drift and Effective Population Size (N_e)

- **Fixation Rules**: For a neutral allele, the probability of eventual fixation equals its current initial allele frequency (p), and the time to fixation is directly proportional to population size.
- **Concept of N_e**: Effective population size (N_e) represents the size of an idealized population that experiences the same rate of inbreeding or loss of genetic diversity as the real population.
- **N_e vs. Census Size (N)**: N_e is almost universally smaller than N due to real-world deviations from ideal assumptions.
- **Factors Reducing N_e**:
  - **Reproductive Skew**: Unequal numbers of breeding females (N_f) and males (N_m) decrease N_e via:

$$N_e = \frac{4N_f N_m}{N_f + N_m}$$

- **Variance in Family Size**: High variance in individual reproductive output reduces N_e relative to N.
- **Population Fluctuations & Bottlenecks**: N_e across generations is dictated by the harmonic mean:

$$\frac{1}{N_e} = \frac{1}{t}\sum \frac{1}{N_i}$$

Bottlenecks dramatically reduce N_e and amplify genetic drift.

---

4 Natural Selection: Models, Dynamics, and Modes

- **Fitness Measures**:
  - **Absolute Fitness (W)**: Genotype-specific rate of increase predicting absolute population numbers.
  - **Relative Fitness (w)**: Fitness relative to the fittest genotype ($w_{\max} = 1$).

  - **Selection Coefficient (s)**: Measure of the strength of selection against a genotype ($w = 1 - s$).

- **General Selection Model**: The per-generation change in allele frequency (Δq) depends on initial frequencies (p, q) and relative fitness values (w_1, w_2, w_3):

$$\Delta q = \frac{pq[q(w_3 - w_2) - p(w_1 - w_2)]}{\bar{w}}$$

where population mean fitness is:

$$\bar{w} = p^2 w_1 + 2pqw_2 + q^2 w_3$$

- **Dominance Effects on Selection Dynamics**: Selection against a dominant allele is rapid initially; selection against a recessive allele slows down substantially as the allele becomes rare because it is "hidden" from selection in heterozygous carriers.
- **Modes of Selection**:
  - **Directional Selection**: Favors one extreme phenotype, eliminating deleterious alleles over time without reaching a stable polymorphic equilibrium.
  - **Stabilizing Selection**: Favors intermediate phenotypes, reducing phenotypic variance.
  - **Disruptive Selection**: Favors extreme phenotypes at both ends of the spectrum over intermediate ones.
- **Frequency-Dependent Selection (FDS)**: Occurs when genotype fitness depends on its frequency in the population. Negative FDS maintains polymorphism and leads to a stable equilibrium.

---

5 Mechanisms Maintaining Genetic Variation

- **Mutation-Selection Balance**: Rare, severely disadvantageous alleles persist at an equilibrium frequency where their rate of removal by selection equals their rate of creation by recurrent mutation (μ).
  - **Dominant Deleterious Alleles**: $q \approx \frac{\mu}{s}$
  - **Recessive Deleterious Alleles**: $q \approx \sqrt{\frac{\mu}{s}}$
- **Heterozygote Advantage (Overdominance)**: Occurs when heterozygotes have higher fitness than either homozygote ($w_{AA} = 1 - s$, $w_{Aa} = 1$, $w_{aa} = 1 - t$), leading to a stable equilibrium $q = \frac{s}{s+t}$.

- **Segregational Load**: The persistent reduction in average population fitness below 1 due to continuous Mendelian emergence of lower-fitness homozygotes.
- **Genotype $\times$ Environment ($G \times E$) Interactions**: Spatial or temporal environmental heterogeneity where relative fitness rankings reverse across environments, preserving variation.
- **Genetic Correlations & Trade-offs**: Driven by pleiotropy or linkage disequilibrium, creating evolutionary trade-offs and co-adapted trait complexes.

---

6 Population Structure and Subgroup Differentiation

- **Panmixia vs. Population Structure**: Panmixia represents random mating with no allele frequency differences between groups. Reduced gene flow creates population structure, leading to divergence in allele frequencies between subpopulations.
- **Evolutionary Drivers**:
  - **Gene Flow**: Reduces genetic differentiation between populations.
  - **Genetic Drift**: Increases differentiation between isolated populations.
  - **Selection**: Increases differentiation under divergent environmental pressures, but decreases differentiation under universally adaptive selection.
- **Fixation Index (F_{ST})**: Measures genetic differentiation due to population subdivision, which reduces expected heterozygosity relative to total population expectation (Wahlund effect).

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

- H_S: Average expected heterozygosity across subpopulations.
- H_T: Total expected heterozygosity calculated from mean allele frequencies across all subpopulations.
- $F_{ST} = 0$: No allele frequency differences between populations (panmixia).
- $F_{ST} = 1$: Alternate alleles fixed in different populations.
- **Isolation by Distance (IBD)**: Pattern where genetic distance (F_{ST}) increases continuously with geographic distance (e.g., Florida black bears).
- **Locus-Specific Selection Signals**: Loci under divergent selection show significantly elevated F_{ST} compared to neutral background genome regions (e.g., wing pattern loci B/D and Yb in *Heliconius melpomene*).
- **Clustering & Multidimensional Methods**:
  - **Principal Component Analysis (PCA)**: Non-tree method to simplify and visualize high-dimensional genomic variation and genetic distance.

- **STRUCTURE Analysis**: Model-based Bayesian clustering method assigning individuals to population clusters to detect migration and admixture across physical barriers.

---

7 Neutral Theory and the Molecular Clock

- **Neutral Theory of Molecular Evolution**: Formulated by Motoo Kimura (1983), stating that most genetic variation within and between species is selectively neutral (having little or no effect on fitness).
- **Mathematical Derivation of the Molecular Clock**:
  - In a diploid population of size N, with a neutral mutation rate μ per gamete per generation, there are $2N\mu$ new mutations each generation.
  - The fixation probability u of any single neutral mutation by random genetic drift equals its initial frequency:

$$u = \frac{1}{2N}$$

- The substitution rate k (number of fixed substitutions per unit time) is:

$$k = 2N\mu \cdot u = 2N\mu \left(\frac{1}{2N}\right) = \mu$$

- **Core Result**: The substitution rate k equals the neutral mutation rate μ and is completely independent of population size N.
- **The Molecular Clock Concept**: Because $k = \mu$, neutral mutations become fixed at a constant rate over time, allowing sequence divergence to act as a clock.
- **Stochastic Nature**: Accumulation of substitutions is a stochastic process (contains inherent randomness).
- **Factors Influencing Clock Rates**:
  - **Mutation Rate (μ)**: Varies across organisms and viral types.
  - **Purifying Selection**: Removes harmful mutations, slowing the rate of substitution. Exons experience stronger purifying selection than introns, resulting in fewer observed substitutions.

---

8 Phylogenetic Analysis and Evolutionary Trees

- **Definition & Core Purpose**: Phylogenetics is the study of evolutionary histories among species or populations, represented via branching network graphs.
- **Homology**: Fundamental requirement that sequence or phenotypic characters must be derived from a common ancestor.
- **Tree Components & Terminology**:
  - **Nodes**: Represent internal common ancestors or speciation/split events.
  - **Branches (Edges)**: Connect nodes; length represents time or inferred evolutionary change.
  - **Tips / Leaves / Taxa**: Observed extant or sampled entities.
  - **Clade**: A monophyletic group consisting of a common ancestor and all its descendants.
  - **Topology**: The branching shape and organization.
- **Tree Types**:
  - **Cladograms**: Equal branch lengths showing common ancestry without indicating evolutionary time or change.
  - **Phylograms**: Branch lengths proportional to inferred evolutionary change (mutations per site).
  - **Time Trees**: Dated phylograms where branch lengths represent absolute dates (years or mya).
- **Tree Building Methods**:
  - **Neighbor-Joining (NJ)**: Fast distance-based method without explicit sequence change models.
  - **Maximum Likelihood (ML)**: Calculates the most likely tree topology given an explicit statistical model of sequence change.
  - **Bayesian Inference**: Iteratively calculates posterior probabilities using explicit sequence change models.
- **Tree Rooting Methods**: Outgroup rooting, midpoint rooting, and dated tips.
- **Epidemic & Evolutionary Patterns Inferred from Topology**: Demonstrates viral dynamics and lineage turnover in pathogens (e.g., Influenza A, Dengue, SARS-CoV-2).

---

9 Conservation Genetics and Molecular Forensics

- **The Extinction Vortex**: Small population size leads to increased drift and inbreeding, causing loss of genetic diversity and inbreeding depression, which further reduces population size and increases extinction risk.

- **Inbreeding Depression**: Reduced fitness in offspring from related matings, caused by unmasking deleterious recessives and loss of overdominance.
- **Genetic Rescue**: Introduction of outside individuals into an inbred population to restore genetic diversity and eliminate inbreeding depression.
- **Evolutionary Significant Units (ESUs)**: Populations prioritized for conservation based on historical isolation, defined by reciprocal monophyly for mtDNA and significant nuclear allele frequency divergence.
- **DNA Barcoding**: Standardized species identification using specific short gene sequences (~650 bp of mitochondrial COI in animals; *rbcL* and *matK* in plants). Based on the assumption that interspecific sequence divergence is significantly greater than intraspecific variation ("barcoding gap").

---

10 Integrative Real-World Applications & Quantitative Methods

- **Medical Genetics & Pathogen Tracking**: Mapping disease loci, quantifying pathogen substitution rates, and tracking transmission chains using time trees.
- **Agriculture & Pest Management**: Detecting selection on pesticide resistance (F_{ST} outliers) and modeling crop pest adaptation.
- **Conservation Management**: Incorporating N_e, F_{ST}, ESUs, and genetic rescue to halt extinction vortices in threatened populations.
- **Genomics & Computational Tools**: Synthesis of theoretical population models (N_e, F_{ST}, $k = \mu$) with high-throughput sequencing, non-tree clustering (PCA, STRUCTURE), and phylogenetic software (MEGA, IQ-TREE, RAxML-NG).

---

11 Population Genomics and Genome-Wide Selection Dynamics

- **Population Genomics Definition**: The expansion of population genetics from single/few loci to high-throughput, whole-genome sequencing across hundreds to thousands of individuals within a species.
- **Sequencing Technologies & Applications**:
  - **Illumina**: Primary short-read platform for population resequencing (5×–40× depth), RNA-seq, ChIP-seq, and metagenomics.
  - **Pacific Biosciences (PacBio)**: Long, accurate reads suitable for *de novo* reference assemblies.

  - **Oxford Nanopore**: Generates ultra-long read lengths; ideal for rapid field sequencing and complex structural variant resolution.
  - **Sanger Sequencing**: ABI-based lower-throughput method used for plasmid verification and targeted small-scale surveys.

- **Genome-Wide Summary Statistics**:
  - **Nucleotide Diversity (π)**: Average number of nucleotide differences per site between any two randomly chosen sequences.
  - **Segregating Sites (S) & Watterson's Estimator (θ_W)**: Metric of diversity scaled by the number of mutating sites.
  - **Allele Frequency Metrics**: Minor Allele Frequency (MAF) and Derived Allele Frequency (DAF).

- **Linkage Disequilibrium (LD) & Sweep Signatures**:
  - **LD Decay**: Non-random association of alleles at distinct loci, broken down over time via recombination.
  - **Selective Sweeps**: Positive selection drives a beneficial mutation to high frequency, sweeping adjacent linked neutral variants along with it.
  - **Signatures**: Localized drop in nucleotide diversity (π), marked increase in LD, excess of rare alleles, and high Extended Haplotype Homozygosity (EHH).

11.0.1 Genomic Patterns Across Modes of Selection

Selection Mode	Diversity (π, θ_W) Effect	Allele Frequency Spectrum	Haplotype Structure
Purifying (Negative)	Reduced in functional regions (exons, promoters)	Excess of rare, low-frequency alleles (negative Tajima's D)	Baseline decay
Adaptive (Positive)	Locally eliminated surrounding the swept allele	Rapidly fixed/high DAF for beneficial allele	Long, conserved haplotypes (high EHH / nS_L)
Balancing	Elevated surrounding target locus	Excess of intermediate-frequency common alleles	Divergent persistent haplotypes

- **Polygenic Adaptation & Distribution of Fitness Effects (DFE)**:
  - **DFE Profile**: Most new mutations are neutral or slightly deleterious; strongly advantageous mutations ($s > 0$) are extremely rare.

- **Polygenic Architecture**: Most complex traits adapt through subtle, simultaneous frequency shifts across many loci.

- **Population Genomic Case Studies**:
  - **Drug Resistance Sweeps**: Pyrimethamine selection on *dhfr* in *P. falciparum*.
  - **Haplotype-Based Selection Tests**: nS_L statistic applied to *P. vivax* targets like the $AP2$ transcription factor.

---

12 Comparative Genomics, Evolutionary Rates, and Divergence

- **Diversity vs. Divergence**:
  - **Diversity**: Polymorphic variation present within a species population.
  - **Divergence**: Fixed sequence differences accumulated between isolated species over time.

- **Comparative Genomics Pipeline**: Assembly, annotation/homology search, synteny analysis, and multi-species reference alignments.
- **Determinants of Evolutionary Rates**:
  - **Functional Essentiality**: Critical genes evolve significantly slower due to intense purifying selection.
  - **Expression Level**: Highly expressed genes exhibit lower evolutionary rates.
  - **Protein Structural Constraints**: Core residues evolve slowly; surface-exposed residues evolve rapidly.

- **Quantifying Evolutionary Constraint**:
  - Conserved non-coding regions and coding exons show reduced divergence across multi-species alignments due to purifying selection.

- **Sequence Divergence Tests for Selection**:
  - **Synonymous Rate (dS or Ks)**: Amino acid-preserving substitutions; selectively neutral baseline.
  - **Nonsynonymous Rate (dN or Ka)**: Amino acid-altering substitutions; subject to functional selection.
  - $\frac{dN}{dS}$ **Ratio Interpretation**:
    - $\frac{dN}{dS} < 1$: Purifying selection.
    - $\frac{dN}{dS} = 1$: Neutral sequence divergence.
    - $\frac{dN}{dS} > 1$: Adaptive positive selection.

- **The McDonald-Kreitman (MK) Test**:

- Compares nonsynonymous-to-synonymous polymorphism within species (P_n/P_s) against divergence between species (D_n/D_s).
- **Adaptive Divergence**: $D_n/D_s > P_n/P_s$.
- **Balancing Selection**: $P_n/P_s > D_n/D_s$.

---

13 Continuous Variation, Phenotypic Components, and Heritability

- **Continuous vs. Discontinuous Traits**: Continuous quantitative traits are determined by multiple genetic loci interacting with environmental factors.
- **Components of Phenotypic Variance (V_P)**:

$$V_P = V_G + V_E + V_{G\times E}$$

- **Components of Genetic Variance (V_G)**:

$$V_G = V_A + V_D + V_I$$

- **Additive Genetic Variance (V_A)**: Cumulative phenotypic effects of individual allele substitutions (primary basis for parent-offspring resemblance).
- **Dominance Genetic Variance (V_D)**: Results from allelic interactions at the same locus.
- **Epistasis Genetic Variance (V_I)**: Arises from inter-locus interactions.
- **Broad-Sense vs. Narrow-Sense Heritability**:
  - **Broad-Sense Heritability (H^2)**: $H^2 = \frac{V_G}{V_P}$
  - **Narrow-Sense Heritability (h^2)**: $h^2 = \frac{V_A}{V_P}$
- **Estimating h^2 via Regressions**:
  - Mid-parent regression slope b gives $h^2 = b$.
  - Single-parent regression slope b gives $h^2 = 2b$.
- **The Breeder's Equation**:

$$R = h^2 S$$

- Selection Differential ($S = M_S - M$) and Response to Selection ($R = M' - M$).

---

14 Quantitative Trait Loci (QTL) Mapping and GWAS

- **Quantitative Trait Loci (QTL)**: Specific chromosomal regions statistically associated with variation in a quantitative trait.
- **Bi-Parental Mapping Populations**: F_2, Backcross (BC), Recombinant Inbred Lines ($RILs$), and Near Isogenic Lines ($NILs$).
- **QTL Mapping Methods**: Single Marker Analysis ($ANOVA$), Interval Mapping (IM), and Composite Interval Mapping (CIM).
- **Genome-Wide Association Studies (GWAS)**: Scans genome-wide SNPs across natural populations to detect marker-trait associations using historical recombination.
- **Linkage vs. Linkage Disequilibrium (LD)**: Linkage refers to physical co-localization on a chromosome; LD represents statistical non-random association of alleles across a population.
- **Factors Governing LD Decay**: Recombination rate, population size (N_e), mating system (selfing vs. outcrossing), natural selection, and population structure.
- **Methodological Controls & Missing Heritability**: Mixed linear models and PCA correct for population structure. Missing heritability describes the gap between pedigree h^2 and GWAS SNP-based heritability (h^2_{SNP}).

---

15 Genome Evolution and Structural Dynamics

- **Definition and Complexity (The C-Value Paradox)**:
  - **Genome Size vs. Complexity**: Prokaryote genome size correlates linearly with complexity. Eukaryotes show no correlation between genome size and morphological complexity (**C-value paradox**).
  - **Diversity**: Ranges from small eukaryotic genomes to massive genomes (*Homo sapiens* 3.3 Gbp, *Ambystoma mexicanum* 32 Gbp, *Paris japonica* 149 Gbp).

15.0.1 Structural Organization: Prokaryotes vs. Eukaryotes

Feature	Prokaryotic Genomes	Eukaryotic Genomes
Chromosome Structure	Typically single circular chromosome	Multiple linear chromosomes
DNA Packaging	Condensed in nucleoid via supercoiling	Condensed in nucleus via histone complexes

Feature	Prokaryotic Genomes	Eukaryotic Genomes
Ploidy & Gene Copy	Haploid (single gene copy)	Mostly diploid (two copies per gene)
Expression Coupling	Simultaneous transcription and translation	Spatially separated (nucleus vs. cytoplasm)
Extrachromosomal Elements	Common nonessential plasmids	Plasmids rare (mostly organellar genomes)
Genome Compactness	Highly compact; minimal noncoding DNA	Extensive non-coding introns, regulatory regions, TEs

- **Prokaryotic Genome Evolution and Lifestyle**:
  - Free-living prokaryotes maintain larger genomes (e.g., *Sorangium cellulosum* 13 Mb); obligate symbionts undergo gene decay/streamlining (e.g., *Carsonella ruddii* 0.16 Mb).
  - **Horizontal Gene Transfer (HGT)**: Transformation, Transduction, and Conjugation.
- **Organellar Origins**:
  - Endosymbiotic theory: Mitochondria (α-proteobacteria) and chloroplasts (cyanobacteria).
- **Eukaryotic Repetitive Elements and Transposable Elements (TEs)**:
  - **Class I (Retrotransposons)**: "Copy-and-paste" via RNA intermediate (LINEs, SINEs, LTRs).
  - **Class II (DNA Transposons)**: "Cut-and-paste" mechanism.
- **Spliceosomal Introns & Alternative Splicing**:
  - Alternative splicing expands protein diversity (e.g., *Drosophila Dscam1* generating thousands of isoforms).
  - Spliceosomal introns originated from ancestral bacterial Group II self-splicing introns.
- **Linear Chromosomes, Sex Chromosomes, & Polyploidy**:
  - Telomerase resolves end-replication problems on linear chromosomes.
  - Suppression of recombination around sex-determining loci leads to heteromorphic sex chromosomes (X/Y, Z/W).
  - **Whole Genome Duplication (WGD)**: Autopolyploidy and Allopolyploidy (e.g., *Triticum aestivum*) drive instant speciation and gene duplication.