Molecular Applications in Zoology

X.
Functional genetic variability:
From SNP to selection

Michal Vinkler
Charles University
Department of Zoology
e-mail: michal.vinkler@natur.cuni.cz
European Molecular Biology Laboratory (EMBL)
European Bioinformatics Institute (EBI)

http://www.ebi.ac.uk/training/
Outline of the lecture

• Brief introduction to genetic variability
• Basic methods of genetic variability detection
• Examples of polymorphism at different types of sites with different effects
• Recombination
• Natural selection – principles and methods of detection in molecular data
Molecular variation

- Phenotypic variability
  - observed traits
  - interaction of genotype & environment

- Genotypic variability
  - germline encoded differences in NA sequences

- Somatic variability
  - somatic mutations
  - germline encoded sequences rearranged in somatic cells
  - increase of pre-existing variability
Heritability

- how much of the variation in a trait is due to variation in genetic factors
- \( H^2 = \frac{V_G}{V_P} \)
  \( V_P \) = proportion of phenotypic variation
  \( V_G \) = proportion of variation due to genetic factors
Heritability in host-parasite interactions

Slash pine (*Pinus elliotii*) - Fungal rust (*Cronartium quercuum*)

$h^2 = 0.21$

Corn borer (*Ostrinia nubilalis*) - bacteria *Bacillus thuringiensis*

$h^2 = 0.31$

Human (*Homo sapiens*) - various viral and bacterial diseases

32-65%
Molecular variation

• Interspecific vs. Intraspecific
  – Principally the same, difference in gene flow

• Interpopulation vs. Intrapopulation

• Polymorphism
  Definition:
  – „The condition in which the DNA sequence shows variation between individuals in a population.“ (Patthy 2008)
  – Convention: genetic variability with minor allele frequency > 0.01

How many common SNPs diversify living humans?
Polymorphism

Genotypic variability ~ Genetic polymorphism

- Human vs. Chimpanzee – 1.2% divergence
- Human vs. Neanderthal – 0.50% divergence
- Humans vs. Human – variability in 0.10% positions (frequency > 1%)
  ➔ All humans are from 99.9% genetically identical

But human genome >3‘000‘000‘000 bp
  ➔ 0.10% > 3‘000‘000 bps are commonly variable in humans
  ➔ many more are variable with frequency < 1%
  - mostly no phenotypic effect – 3%-5% SNPs functional
  ➔ ca. 100‘000 common functional SNPs
Polymorphism

Markers of genetic variability:

- **Single nucleotide polymorphism (SNP) and short indels**
  - > 99.9% of variants
  - Human genome – ca. 5 million common SNPs (8M over 5%; >80M known)
  - Every 6kb on average, linkage between neighbouring loci

- **Short tandem repeats (STRs) = microsatellites**
  - short (usually 2-5 bp) sequences repeated in genome
  - highly variable in length
  - usually neutral

In a typical genome contains ~ 2500 structural variants:

- Insertions (Alu, L1, SVA): ~ 1000
- **Copy number variants (CNVs):** 160; 4.8–9.7% of the human genome
  - larger than 50 bp, often longer sequences (1kb-1Mb) and genes
- **Large deletions:** 1 000
- **Inversions:** 10
Polymorphism

Markers of genetic variability:

- **Single nucleotide polymorphism (SNP) and short indels**
  - > 99.9% of variants
  - Human genome – ca. 5 million common SNPs (8M over 5%; >80M known)
  - Every 6kb on average, linkage between neighbouring loci
Basic methods of SNP detection:
- **Sequencing** – Sanger / 2nd generation / 3rd generation
Techniques of SNP detection

Presetly 1 genome for ~ $1000
Techniques of SNP detection

Basic methods of SNP detection:

- **Single strand conformation polymorphism (SSCP)** and related approaches (e.g. Reference strand conformation analysis, RSCA)
Techniques of SNP detection

Basic methods of SNP detection:

- **High resolution melting (temperature) analysis (HRMA)**
  - PCR → warming from 50°C up to 95°C → real-time fluorescent detection of double-stranded DNA

![Normalized Melting Curves](image-url)
Techniques of SNP detection

Basic methods of SNP detection:

- **SNaPshot**
- primer extension–based method
- multiplexing capability (up to 10-plex)
- sensitive allele-frequency detection (typically 5%)
- analysis - GeneMapper
Techniques of SNP detection

Basic methods of SNP detection:
- SNP microarrays (SNP chip)
Techniques of SNP detection

Microarrays

- **Probes** (or reporters or oligos) = picomoles ($10^{-12}$ moles) of a specific DNA sequence – synthesised and attached covalently to the chip surface
- tens of thousands of probes per chip
  → hybridize cDNA or cRNA (~ anti-sense RNA) sample (**Target**)
  → detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target → quantification
Techniques of SNP detection

One-channel (colour) detection
- relative abundance when compared to other samples (on the same slide)
- aberrant samples cannot affect raw data

Two-channel (colour) detection
- two different fluorophores:
  - Laser → e.g. Cy3 (570 nm = orange) and Cy5 (emission of 670 nm = red)
  - control probes → normalization
Techniques of SNP detection

Usage
- Gene expression microarrays – control vs. treatment
- SNP microarray (SNP chip) – allele A vs. allele B
- Comparative genomic hybridization
- Alternative splicing (Exon arrays)

Applicable only to species with complete genome (model species)
Techniques of SNP detection

Exome sequencing
Target-enrichment strategies:
- Array-based capture
- In-solution capture
Techniques of SNP detection

Molecular Inversion Probes (MIP)

Target-enrichment

→ Resequencing (NGS)

Niedzicka et al. 2016
Techniques of SNP detection

ChIP-seq
- determines how transcription factors and other chromatin-associated proteins influence phenotype-affecting mechanisms
- combines chromatin immunoprecipitation (ChIP – antibodies attached to beads) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins
Well, and what now?

Which questions may I now ask?
Well, and what now?
Well, and what now?
Well, and what now?
Congenic animals and strains
- animals / strains in which a specific and defined part of the genome from one inbred strain (strain A) is introgressed on the genetic background of second inbred strain (strain B)
Quantitative trait loci (QTLs)

**Quantitative traits** - polygenic effects on phenotype

- **QTL mapping** – effect of SNPs
- **Mus musculus** (e.g. C57BL/6)
  - susceptible to *Y. pestis* (10^2 CFU → <8% survival)
- **Mus spretus** (SEG/Pas)
  - resistant to *Y. pestis* (10^2 CFU → >70% survival)

Blanchet et al. 2011
Quantitative trait loci (QTLs)

QTL mapping in mice using 322 backcrosses

- B6xSEG F1 females + B6 males → F2 progeny
  - 721 polymorphic markers covering the entire genome

- Quantitative trait loci (QTLs) on chromosomes 3, 4 and 6, with dominant SEG protective alleles:
  - Y. pestis resistance loci (Yprl1-3)
  - Each QTL contributes with ~20% → 67% in total
  - Large chromosomal segments (between 50 and 84 Mb)

- Candidate genes:
  - Yprl1: Pglyrp3, Pglyrp4 (AMPs), IL-6α
  - Yprl2: Tlr4, IFNs
  - Yprl3: Nod1, IL-17R subunits, IL-23R

Blanchet et al. 2011
Indigenous Ethiopian chicken ecotype Horro

Infectious bursal disease (IBDV) antibody titres
Affx-51084536 – missense variant within the *XK-related protein 8 (XKR8)* gene

Affx-51878048 - region for IBDV response on chromosome 9: two putative candidate genes -
- protein tyrosine phosphatase nonreceptor type 1 (*PTPN1*)
- nuclear factor of activated T-cells cytoplasmic calcineurin-dependent 2 (*NFATC2*)

Most significant SNPs were located in intergenic or intronic regions!

Psifidi et al. 2016
Polymorphism in coding regions

What applies to non-synonymous substitutions?
Where is the detected polymorphism localised?

Polymorphism types:
- Single nucleotide polymorphism (SNPs)
- Indels – insertions & deletions
- Rearrangements

Look at position:
- Non-coding – more common
- Coding – synonymous (silent) vs. non-synonymous (missense & nonsense) ~ 1:1
- Regulatory regions

Patthy 2008
Polymorphism in regulatory regions

Promoter sequence
- combination of SNP database (NCBI) and expression microarrays

Human antioxidant response elements (AREs)
- cis-acting enhancer sequences found in the promoter regions of many genes that encode antioxidant and Phase II detoxification enzymes/proteins
Polymorphism in regulatory regions

Promoter sequence
- combination of SNP database (NCBI) and expression microarrays

Human antioxidant response elements (AREs)

Hemoglobin epsilon 1 (HBE1)

rs7130110, -4732 bp
Polymorphism in coding regions

Synonymous vs. non-synonymous substitutions:
- Translate to amino acid sequence

Software:
- e.g. Genious, BioEdit, web tools: ExPASy → geneious
- show variable sites (FaBox) → count
Polymorphism in coding regions

In coding regions may influence protein structure:
- **Electrostatic forces**
  - charges – within protein, surface electrostatic potential
Polymorphism in coding regions

TLRs-ligand:
- Dimerisation
  - Homodimerisation
  - Heterodimerisation
- G- LPS → MD-2/TLR4
- Bound based on Lipid A

Botos et al. 2009
Pseudomonas aeruginosa

- opportunistic bacterium
- during infection
  - Down-regulates the flagellin expression
  - Increases $\rightarrow$ decreases the acylation state of LPS lipid A

Polymorphism in coding regions

Hajjar et al. 2002
Polymorphism in coding regions

- Lipid IVa = precursor in Lipid A synthesis
- **agonist** in horse and mouse but an **antagonist** in humans and cat
- TLR4: R385G in the glycan-free flank of the horse TLR4 solenoid confers the ability to signal in response to lipid IVa

Walsh et al. 2008
Polymorphism in coding regions

In coding regions may influence protein structure:

- **Van der Waals interactions**
  - charged groups induce dipole → dipole-dipole interaction

- **Disulphide bonds**
  - oxidation of the sulfhydryl groups on cysteine

- **Hydrogen bonds**
  - two electronegative atoms compete for the same H atom

- **Hydrophobic interactions**
  - non-polar groups cannot interact with polar groups & water → keep together

- **Role of posttranslational modifications**
  - functional groups – e.g. acetate, phosphate, various lipids and carbohydrates
Polymorphism in coding regions

Protein structure prediction

Software:

- SMART – domain architecture

- Specialised domain prediction tools (SignalP, LRR-finder, DAS-Tmfilter, etc.)

- 3D modelling
  - AlphaFold2
  - Modeller
  - I-TASSER

Next best
Polymorphism in coding regions

Vinkler et al. 2014

Surface electrostatic potential in avian-mammal TLR4s

Software: PDB2PQR Server → visualisation in Jmol
Polymorphism in coding regions

Surface electrostatic potential

Software:
- PDB2PQR Server
- PIPSA

Electrostatic Distance $D_{a,b} = \sqrt{2 - 2S_{a,b}}$
Polymorphism in coding regions

Těšický et al. 2020

CHARLES UNIVERSITY
Alleles

Software:

• DnaSP-PHASE

• FaBox – haplotype collapsed

CHARLES UNIVERSITY
Genetic recombination

Two DNA molecules exchange genetic information $\rightarrow$ new DNA variant
• Neighbouring SNPs in linkage = haplotypes

• Linkage disequilibrium
  $=$ degree of genotype (combination of alleles on several loci) frequency deviation from the expected independent assortment
  – lowers with distance (probability of recombination increases)
  – 1cM (=1% frequency of crossover) $\approx$ 1Mb, but not uniform
  – considered in association studies

  – **Recombination hot spots** = regions (1-2kb) with 10x higher recombination than surroundings – humans ca. 50’000 hot spots
  $\rightarrow$ genomes in blocks (haplotype blocks), 5-100kb

• **Tag SNPs** – SNPs selected to identify individual haplotypes
Genetic recombination

Software:


<table>
<thead>
<tr>
<th>BPs</th>
<th>AIC_c</th>
<th>Δ AIC_c</th>
<th>Segments</th>
<th>Δ AIC_c Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8694.58</td>
<td></td>
<td>1-2526</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8537.02</td>
<td>157.556</td>
<td>1378</td>
<td>1755</td>
</tr>
<tr>
<td>2</td>
<td>8453.61</td>
<td>83.4092</td>
<td>675</td>
<td>1755</td>
</tr>
<tr>
<td>3</td>
<td>8424.32</td>
<td>29.2955</td>
<td>675</td>
<td>1293 1755</td>
</tr>
</tbody>
</table>

Two of the gametes are parental types

Two of the gametes are recombinant types

CHARLES UNIVERSITY
Genotype evolution: neutral or adaptive?

- **Genetic drift**: Random changes in allele frequencies
- **Mutation**: Evolution of adaptive traits through increased fitness of better adapted individuals
- **Migration**
- **Inbreeding**
- **Selection**

Gene flow
Polymorphism in coding regions

Which processes decrease genetic variation in a population?
Genotype evolution: neutral or adaptive?

Drift is random

→ Neutral theory is a null hypothesis

Adaptive variability

Selection is deterministic

Lack of variability
Fluctuations of allele frequencies in time

**Great reed warbler (Acrocephalus arundinaceus)**

- MHC I
- Comparison of frequency changes in 23 MHC alleles and 23 microsatellite alleles in time
- Non-random fluctuation of frequencies of 2 alleles in time

→ evidence for variability in selection

Westerdahl et al. 2004
Selection on host genes:

- **Negative** = Stabilising = Purifying
  - Elimination of deleterious
  - Shallower genealogy

- **Disruptive** = Diversifying
  - Polymorphism maintenance
  - Deeper genealogy

- **Positive** = Directional
  - Fixation of advantageous
  - Shallower genealogy

---

**Table 10.5 Effects of various evolutionary processes on genetic variation (modified after Nielsen 2005).**

<table>
<thead>
<tr>
<th>Process</th>
<th>Degree of genetic variation</th>
<th>Frequency spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within species</td>
<td>Between species</td>
</tr>
<tr>
<td>Mutation accumulation</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Negative directional selection</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Positive directional selection</td>
<td>+ or -</td>
<td>+</td>
</tr>
<tr>
<td>Balancing selection</td>
<td>+</td>
<td>+ or -</td>
</tr>
<tr>
<td>Selective sweeps</td>
<td>-</td>
<td>No effect, or (+)</td>
</tr>
</tbody>
</table>
Tajima’s D test

• Compares observed **nucleotide heterozygosity** (θπ) and observed **number of polymorphic sites** corrected for sample size (θw)

At equilibrium θπ = θw

• \( D = \frac{θπ - θw}{\text{Var } D} \)
  – \( D>0 \) if θπ>θw
  – \( D<0 \) if θπ<θw

• Excess of low-frequency alleles
  → more SNP sites than heterozygosity (θπ<θw)
  → \( D<0 = \text{positive or negative selection} \) (or population expansion)

• Intermediate allele frequencies
  → High heterozygosity per SNP
  → \( D>0 = \text{balancing selection} \) (or population contraction)

• Roughly +2<D or D<-2 is likely to be significant
Ratio Polymorphism: Divergence is the same for different loci under neutrality

**HKA-test** (Hudson-Kreitman-Aquadé)
- Expected levels of divergence and polymorphism vs. observed levels of divergence and polymorphism at several loci (at least 2)
- Distinguish locus-specific selection from population-level effects
- Due to linkage disequilibrium the selection may not be on the tested site but somewhere in the neighbourhood

**MK-test** (MacDonald-Kreitman)
- Comparison of dN and dS within species and among species
  = compares two types of sites within the same locus! (same genealogy)
- $X^2$ test or Fisher’s exact test
Human variability vs. chimpanzee

- exon-specific PCR amplification of 11,624 genes in 39 humans and one chimp
- Variability in 10,767 genes (92.6%), 8,292 had >1 NS SNP or fixed difference
- 9.0% under rapid amino acid evolution
- 13.5% with low divergence between human and chimp
- Negative selection in cytoskeletal proteins, vesicle transport and related functions
- Positive selection in transcription factors, mRNA transcription, receptors and immunity

Bustamante et al. 2005
CD5

- is a lymphocyte surface co-receptor
- scavenger receptor cysteine-rich (SRCR) superfamily
- Poorly known function:
  - cell-to-cell immune interactions
  - recognition of fungal β-glucans
- 27 polymorphic sites:
  - comprising 17 intronic, 3 synonymous, 7 nonsynonymous substitutions
  - selection for A471V substitution in cytoplasmatic region
    → differences in MAPK cascade activation
    → higher IL-8 in V471

Population differentiation

Carnero-Montoro et al. 2012
Selective sweep

CD5

rs2229177

Carnero-Montoro et al. 2012
Selective sweep

Haplotype network
→ Low diversity in East Asia = recent selective sweep

Carnero-Montoro et al. 2012
Fixation index ($F_{ST}$)

- measure of population differentiation due to genetic structure
- based on the variance of allele frequencies between populations / probability of Identity by descent

Evolution of Darwin’s finches’ beaks

Geospiza
1. G. magnirostris_G
2. G. difficilis_W
3. G. difficilis_P

Tiaris bicolor

ALX1 genotype

Lamichhaney et al. 2015
Genome-wide association studies (GWAS).

Karlsson et al. 2014
Pattern of variability in sequences

nsSNP/ssSNP ratio → dN/dS

Lys  Ala  Asp  Cys

TTCCGGGCTAACA

AAGGCCCGATTGT
AAGGCCGGACCTGT
AAAGCTGATTGT
AAGGCTGTGTTGG
AAAGACGCCTGT
AAAGCGGACGTGC
AGAGAGGAGTGC

\[
\frac{dN}{dS} \sim \frac{8}{8} = 1
\]
Pattern of variability in sequences

nsSNP/ssSNP ratio → dN/dS

Lys  Ala  Asp  Cys

TTCCCTAACACGG

AAGGCCGGATTGT

AAGGCCGGACCTGT

AAAGCTGATTGT

AAGGCTGGTTGC

AAAGACGATTGT

AAAGCGGACCTGC

AGAGCGGACCTGC

\[
\frac{dN}{dS} \sim \frac{3}{9} < 1
\]
Pattern of variability in sequences

nsSNP/ssSNP ratio → dN/dS

**Lys** **Ala** **Asp** **Cys**

TTCCTTAACACGG

AAGGCCGATTGT

AAGGGCGATTGT

AAGGGTGATTGT

AAGGCTGGTTGG

AAAGACGCGTGT

AAAGAGGAGTGC

AGAGAGGAGTGC

\[
\frac{\text{dN}}{\text{dS}} = \frac{8}{3} > 1
\]
Pattern of variability in sequences

Ratio of non-synonymous (\(d_N=Ka\)) substitutions to synonymous (\(d_S\sim Ks\)) substitutions

- \(Ka/Ks = \omega\)
  - per locus
  - per site
  - Interspecific & intraspecific
  - neutral: \(Ka/Ks=1\), positive: \(Ka/Ks>1\), negative: \(Ka/Ks<1\)
  - little power to detect weak positive selection
  - power increases with sample size (species number)

Software:
- PAML
Selection in different domains

- **MHC I and MHC II**
  - The $d_N/d_S$ was higher in antigen-binding sites (ABS) than in non-ABS
  - Differentiating selection on antigen binding features

<table>
<thead>
<tr>
<th>MHC</th>
<th>Region</th>
<th>Sites</th>
<th>$d_N/d_S$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>Exon 2</td>
<td>ABS</td>
<td>2.87</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-ABS</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>1.60</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Exon 2</td>
<td>ABS</td>
<td>1.20</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-ABS</td>
<td>0.15</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>0.38</td>
<td>0.02</td>
</tr>
<tr>
<td>Class II-DRB</td>
<td>Exon 2</td>
<td>ABS</td>
<td>1.40</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-ABS</td>
<td>0.31</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>0.69</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Castro-Prieto et al. 2011
Pattern of variability in sequences

- lineage-specific effects and codon-specific effects (MEME)

Fornuskova et al. 2013
Ka/Ks usage for pathogen elicitor detection

- **G- bacteria core genome**
  - genes represented in all studied species (6)
  - 1‘322 orthologous
- **Ka>Ks**
  - 35% of the core genes had at least one positively selected
  - 56 proteins exhibited significant signatures of positive selection
- **comparison with selection in soil-inhabiting bacteria**
  - 48 proteins positively selected in pathogenic and not in non-pathogenic
- **positively selected sites in clusters**
  - 45 loci

→ Functional testing in *A. thaliana*
Footprints of selective events in genes

Advantages:

• Detection of events with small but biologically relevant selection coefficients
• Selection on evolutionary not ecological time scale (= selection in the past)
• Testing selection in genes without knowledge of the phenotype
• May allow detection of genes responsible for emergence of novel traits

Disadvantages:

• Demographic changes may give similar results as selection if only 1 gene observed
• When selection is found, linking with phenotype is difficult

→ We know about selection but we do not understand its functional consequences
→ Linking detected selection with phenotype is a major challenge
Florida panthers (*Puma concolor coryi*)

Inbred population in Florida

→ in 1995 translocation of 8 female pumas from Texas (*P. c. stanleyana*)

→ Fitness improvement

Two pre-1995 groups (CFP and EVG) TX females → TX-backcross (TX-BC) admixed Florida panthers (AdmFPs)

Johnson et al. 2010
• There are multiple techniques to detect genetic variation

What can I do with sequence data to reveal functional differences between individuals / populations / species?

1. Check the position of your SNPs
2. Model protein structures and predict functional effects of substitutions
3. Identify alleles and non-synonymous protein variants and assess their frequencies in distinct populations
4. Detect recombination
5. Detect selection and recognize adaptive evolution
The latest urging e-mail has been sent today by Zuznana Starostová

→ ACTION REQUIRED !!!!!

• Fill in information on the preference of the training time:

• 2 May:
  - Morning?
  - Afternoon?