Genomic studies of speciation and gene flow
Why study speciation genomics?

Long-standing questions (role of geography/gene flow)

How do genomes diverge?

Find speciation genes
Genomic divergence during speciation

1. Speciation as a by-product of physical isolation

2. Speciation due to selection – without isolation
Genomic divergence during speciation

1. Speciation as a bi-product of physical isolation

2. Speciation due to selection – without isolation

Cline theory - e.g. Barton and Gale 1993
1. Speciation as a bi-product of physical isolation

2. Speciation due to selection – without isolation
Stage 1 - one or few loci under disruptive selection

Gene under selection

Genome

\[ F_{ST} \]

Feder, Egan and Nosil TiG
Stage 2 - Divergence hitchhiking

Genome

\[ F_{ST} \]

Feder, Egan and Nosil TiG
Stage 2b - Inversion

Inversion links co-adapted alleles

Genome

$F_{ST}$

Feder, Egan and Nosil TiG
Stage 3 - Genome hitchhiking

Feder, Egan and Nosil TiG
Stage 4 - Genome wide isolation

Genome

$F_{ST}$

Feder, Egan and Nosil TiG
Some sub-species clearly in stage 1

Wing pattern “races” of *Heliconius melpomene*

*Heliconius melpomene*

![Graphs showing frequency distribution for Heliconius melpomene races](image)

*H. melpomene amaryllis*  
*H. melpomene aglaope*
Some sub-species clearly in stage 1

Wing pattern “races” of *Heliconius melpomene*

Some sub-species clearly in stage 1

Carrion and hooded Crows

And an example with multiple islands?

Other species have islands... but are they real?

Collared and Pied Flycatchers

Other species have islands...but are they real?

*Anopheles gambiae* and *A. coluzzi*

Formerly *M* and *S* forms of *A. gambiae*

Clarkson et al. 2014 Nature Communications
Other species have islands... but are they real?
**Aa** Parapatric races: *H. m. amaryllis* (Per) versus *H. m. aglaope* (Per)

**Ab** Allopatric races: *H. m. rosina* (Pan) versus *H. m. melpomene* (FG)

**Ac** Sympatric species: *H. cydno* (Pan) versus *H. m. rosina* (Pan)

**Ad** Allopatric species: *H. cydno* (Pan) versus *H. m. melpomene* (FG)

Seehausen et al., Nature Reviews Genetics, 2014
What do patterns of $F_{st}$ really mean?

- $F_{st}$ measures relative divergence
  \[ F_{ST} = \frac{H_T - H_S}{H_T}, \]

- Peaks indicate regions of higher than expected between population divergence, given the within population divergence

- Peaks can therefore result from reduced diversity within species

- This could be due to lower $Ne$ within species (selective sweeps, background selection)

- So peaks NOT NECESSARILY due to reduced gene flow
Note that sometimes sweeps within species = speciation genes
Sweeps across the species barrier can also lead to Fst peaks

Double peaks??

Nicolas Bierne, Daniel Berner and others
REVIEW

Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species

MAF Noor and SM Bennett
Biology Department, Duke University, Durham, NC, USA

Anopheles M-S divergence

Relative divergence higher in low recombination regions - not significant for absolute divergence

see also: Charlesworth 1998 MBE Measures of divergence...
More recently see papers by Reto Burri
INVITED REVIEWS AND SYNTHESSES

Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow

TAMI E. CRUICKSHANK* and MATTHEW W. HAHN*†
*Department of Biology, Indiana University, Bloomington, IN 47405, USA, †School of Informatics and Computing, Indiana University, Bloomington, IN 47405, USA
No evidence for higher Dxy in wing pattern loci

One further issue with interpreting the data from these two races is whether this comparison relates to speciation at all. There is strong geographic structure involving the wing colour patterns that define these morphs as races, largely due to selection determined by colour morphs in the Müllerian mimic, *H. erato* (Mallet *et al*. 1990). But the races are not separate species: they do not show evidence of hybrid sterility or inviability and appear to be randomly mating in the narrow zone where the colour morphs overlap (Mallet *et al*. 1990). This raises the possibility that the colour-patterning loci contain locally adapted alleles within a largely panmictic (or at least continuously distributed) population and that gene flow outside of these regions represents nothing more than the normal movement of alleles within a species. In this case, there should be
Suggestion that we use absolute measures of divergence?
No single statistic will capture the complex history of mutation, migration and selection.

Patterns need to be interpreted in the specific context of the study species.
Much better to use explicit tests for gene flow

Need to design sampling so the expectations in the absence of gene flow are clear and testable

The key is to identify ‘control’ populations that are not influenced by admixture
Explicit tests for gene flow: Neanderthal genome

- Isolated DNA from bones 38,000 yrs old in Croatia
- We diverged from Neanderthals around 270-440,000 yrs ago
- Evidence for gene exchange with humans (1-4% of genome?)

Green et al., 328:710 Science 2010
Explicit tests for gene flow: ABBA-BABA test

**EXPECT:**
50% ABBA
50% BABA

**OBSERVE:**
103612 ABBA
94029 BABA

\[
D(p_1, p_2, p_3, o) = \frac{\sum C_{ABBA}(i) - C_{BABA}(i)}{\sum C_{ABBA}(i) + C_{BABA}(i)}
\]

Green et al. 2010 Science 328:710-722
Explicit tests for gene flow: ABBA-BABA test
Explicit tests for gene flow: ABBA-BABA test
Explicit tests for gene flow:
Combining multiple signals

1) Derived alleles at high frequency shared with Neanderthal
2) High divergence to Africa but low to Neanderthal
3) Long haplotype blocks

The genomic landscape of Neanderthal ancestry in present-day humans - Sankararaman et al. Nature 2014
Sampled 10 complete high coverage genomes per population
Explicit tests for gene flow: *Heliconius* butterflies

Many sources of reproductive isolation:

- Female hybrids are sterile
- Different host plant use
- Different habitat preference
- Strong assortative mating
Using Simon Martin's Twisst method to characterise relationships
ABBA-BABA statistics

melG  melW  cyd  outgroup

Simon Martin

Martin et al., MBE 2015
300+ offspring for each cross type

H. cydno
Panama × H. cydno
Panama

H. m. rosina
Panama × H. cydno
Panama

H. m. rosina
Panama × H. m. rosina
Panama

PstI RAD sequencing
(site every ~10kb)

Linkage maps built
with Lep-MAP

John Davey

Davey et al., Evol Letters 2017
Pop gen vs actual estimates of recombination rate
Recombination rate strongly correlated with admixture proportion
Short chromosomes have more admixture:
And chromosome ends have more admixture:
Sequenced 20 individuals per population at 20x coverage
**Supplemental Table S4.** ABBA-BABA tests for gene flow. Populations/species among which the test indicates gene flow are highlighted in bold.

<table>
<thead>
<tr>
<th>1. Inner</th>
<th>2. Inner</th>
<th>1. Outgroup</th>
<th>Mean(D)</th>
<th>SE(D)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>collared Italy</td>
<td>collared CZ</td>
<td>pied CZ</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.3344</td>
</tr>
<tr>
<td>pied Spain</td>
<td>pied CZ</td>
<td>collared CZ</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.4186</td>
</tr>
<tr>
<td><strong>pied Spain</strong></td>
<td>Atlas</td>
<td><strong>collared Italy</strong></td>
<td>-0.1648</td>
<td>0.0027</td>
<td>&lt;10^{-4}</td>
</tr>
<tr>
<td><strong>pied Spain</strong></td>
<td>Atlas</td>
<td><strong>semicollared</strong></td>
<td>-0.0108</td>
<td>0.0016</td>
<td>&lt;10^{-4}</td>
</tr>
<tr>
<td>pied Spain</td>
<td><strong>collared Italy</strong></td>
<td>semicollared</td>
<td>0.1162</td>
<td>0.0018</td>
<td>&lt;10^{-4}</td>
</tr>
<tr>
<td>Atlas</td>
<td><strong>collared Italy</strong></td>
<td>semicollared</td>
<td>0.1242</td>
<td>0.0016</td>
<td>&lt;10^{-4}</td>
</tr>
</tbody>
</table>
An alternative is to take an explicit modelling approach.
Population-genomic inference of the strength and timing of selection against gene flow

Simon Aeschbacher\textsuperscript{1,a}, Jessica P. Selby\textsuperscript{2}, John H. Willis\textsuperscript{2}, and Graham Coop\textsuperscript{1}
The effect of background selection on introgression in humans

Admixture is less in gene rich regions supporting this model.....

Harris and Nielson Genetics 2016, Juric, Aeschbacher and Coop 2016
Population and speciation genomics: Conclusions

• Great power to detect subtle signals of selection and gene flow
• Can make more general observations about genes and regions involved in adaptation
• BUT genomic processes complicate the picture

• Best approaches combine multiple signals to infer process

• Eventually we need to combine background selection, recombination, positive selection
And finally a plug....
Adaptive introgression
\( G \)-test: \( G = 7.25, \) d.f. = 1, \( p = 0.007 \)

Peaks of divergence correspond to wing pattern genes

van Belleghem et al., Nature Ecol Evol
Wing pattern controlled almost entirely by large effect loci
Richard Wallbank
Adaptive introgression

Heliconius Genome Consortium Nature 2012
Phylogenies across $B/D$

ML tree based on 50,000 bp
Phylogenies across B/D

ML tree based on 50,000 bp
Phylogenies across $B/D$

ML tree based on 50,000 bp
Phylogenies across B/D

ML tree based on 50,000 bp
Phylogenies across \( B/D \)

ML tree based on 50,000 bp
Okay, so introgression causes mimicry

But mimicry is weird, right?
Novelty can arise through introgression and recombination

Camilo Salazar
Mavarez et al., Nature 2006
Generate dated trees using this node as a reference point
Characterizing distinct haplotypes/allelic lineages at two outlier loci

Distribution of allelic lineages among species/populations at 20 outlier loci

Known colour genes

Additional autosomal outlier loci

- L. castaneothorax
- L. flavipryrna
- L. nevermanni
- L. stygia
- L. caniceps
- L. castaneothorax
- L. grandis
- L. spectabilis
- L. spectabilis
- L. melaena
- L. forbesi
- L. melaena
- L. hunsteinii
- L. nigerina
- L. grandis
- L. castaneothorax
- L. spectabilis
- L. melaena
- L. forbesi
- L. melaena
- L. hunsteinii
- L. nigerina

Autosomal (k = 4, m = 5)
What about behaviour?
H. melpomene × H. cydno
Backcross design:
Difference between approaches to cydno and melponemene.

Genotype

Richard Merrill
Significant QTL detected on three linkage groups

5% genome-wide significance threshold

Richard Merrill unpub.
Together explain ~ 50% measured differences between *H. melpomene* and *H. cydno*
Heliconius cydno cordula

Heliconius melpomene melpomene

Heliconius heurippa

Mavarez et al., Nature 2006
ALX1 associated with beak shape
Most of these studies use phenotype associations to identify introgressed loci. But can we identify them a priori using the ABBA-BABA method?
Green et al. 2010 Science 328:710-722
$D$ is not at all good at detecting outlier windows.

Where $s$ is numerator from the $D$ equation
$f$ is the fraction of introgression compared to maximum possible
But Martin’s F is quite good at finding the introgression outliers
Smith and Kronforst argued that introgression could be inferred where ABBA-BABA outliers showed lower Dxy compared to genome-wide average.
• Be wary of window based D statistics

• F is better than D…

• Sampling design is very important!
Implications for tree-thinking

The tree of life is reticulated
Implications for tree-thinking
The true phylogeny
showing bifurcation events as well as all introgression events
Okay, so what have we learnt and where do we go from here?