Dating and splitting of my cousins

Celine Becquet
Dept. Human Genetics
Molly Przeworski Lab
My methodology (1)

Thinking: what the hell should I work on?
My methodology (2)

Thinking: what the hell should I work on?

Define a model

And /or get data files

>gi|66351419|emb|AJ968892.1| Medicago marina partial glu gene for NADH-dependent glutamate synthase
TCAGGGGCAATGTTAACTGGTATGTCCGTACTCCAGTTTTAATTAAACGACCCTTACAAATGAATACTAT
TACAAATTTAAACTGTGGTCATCATCTCCCACGTGGGGATGACATACAATTTGAAATGTAGACATTGTC

>gi|66351240|emb|AJ968810.1| Medicago sativa subsp. glomerata partial glu gene for NADH-dependent glutamate synthase
TCAGAGGCAATGTTAACTGGTATGTCTTGAAAAAGTTTTGCCGTACTCTAGTTTTAATTAAACGGCCCTT
ACAAATGAAATACTATGGACACTACAAATTTAAACGGTTATCATCCACACTGCTGGTGAAATGGATCAATA
AAACAGTAGACATTTGTTGGAGTTACACATGATTGTGTTCTCTCCCTCCTCTAAGTGAAGTGA
TCCAGTCCATTGTTCAACTGTGTTATTGTATTGATGCTAGGCTTAAACGCATTAAGTGGTTGTA

>gi|66351238|emb|AJ968809.1| Medicago sativa subsp. x tunetana partial glu gene for NADH-dependent glutamate synthase
TCAGGGGCAATGTTAACTGGTATGTCTTGAAAAAGTTTGCTACTCTAGTTTTAATTAAACCCACCTT
ACAAATGAAATAGTATGAGCTACAAATTTAAACGTGGTGATCCTCCACACTGCTGGTGAAATGGATCAAA
TAAATAGTAGACATTTGCTGTGGAGTTACACATGATTGTGTTCTCTCCCTCCCTCCTAAAGTGAAGAA
My methodology (3)

Thinking: what the hell should I work on?

Define a model

Write program, do simulations (i.e. make sure it works)

And/or get data files

Clean the data, find software to analyze the data,
Run the program I wrote and which works

>gi|66351419|emb|AJ968892.1| Medicago marina partial glu gene for NADH-dependent glutamate synthase
TCAGGGGCATTTAATGCTGTAATGGTCCGATCAAGTTAATATTTAAAGCGACCCCTACTACAATAGAATAT
TAGAAGTAAATTTGAATGTTCATATCACTCCGACTCCTGGTTGAAATGGATCAATAAAATAGTAGACATTGTC

>gi|66351240|emb|AJ968810.1| Medicago sativa subsp. glomerata partial glu gene for NADH-dependent glutamate synthase
TCAGAGGCAATGTTAACTGGTATGTCTTGAAAAAGTTTTGATCGATCAAGTTAATTTAAAGCCTGTCT
AACAATTGGAATATGGTACTGCTATACATCAAAATTTAAAGCGTTGATCTATGCTGGAATGATCATA
AAAAGATGAAGAATGGTTAATCGTGGAGTACACAGATGATGATGATCCTCCCTCCCCTGTAAGTAAGAGT
TCCAAGTGTCTGCTAAGCTGTCTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATG

>gi|66351238|emb|AJ968809.1| Medicago sativa subsp. x tunetana partial glu gene for NADH-dependent glutamate synthase
TCAGGGCCAATTTAATGCTGTAATGGTCCGATCAAGTTAATATTTAAAGCGACCCCTACTACAATAGAATAT
TAGAAGTAAATTTGAATGTTCATATCACTCCGACTCCTGGTTGAAATGGATCAATAAAATAGTAGACATTGTC

>gi|66351238|emb|AJ968809.1| Medicago sativa subsp. x tunetana partial glu gene for NADH-dependent glutamate synthase
TCAGGGCCAATTTAATGCTGTAATGGTCCGATCAAGTTAATATTTAAAGCGACCCCTACTACAATAGAATAT
TAGAAGTAAATTTGAATGTTCATATCACTCCGACTCCTGGTTGAAATGGATCAATAAAATAGTAGACATTGTC

>gi|66351238|emb|AJ968809.1| Medicago sativa subsp. x tunetana partial glu gene for NADH-dependent glutamate synthase
TCAGGGCCAATTTAATGCTGTAATGGTCCGATCAAGTTAATATTTAAAGCGACCCCTACTACAATAGAATAT
TAGAAGTAAATTTGAATGTTCATATCACTCCGACTCCTGGTTGAAATGGATCAATAAAATAGTAGACATTGTC
Thinking: what the hell should I work on?

Define a model

Write program, do simulations (i.e. make sure it works)

And/or get data files

Write the results and publish

My methodology (4)

Introducing...

Standard transaminases recognize two types of transamination. They promote the exchange of an amine group (or acceptor) for the ketone group (or donor) in a transamination. They have three main roles:

1. They catalyze the exchange of amino groups in the synthesis of amino acids.
2. They are involved in the detoxification process of ammonia in plants.
3. They are involved in the regulation of enzyme activities, such as the regulation of urea cycle enzymes.

Recent studies have shown that transaminase enzymes are involved in the regulation of enzyme activities, such as the regulation of urea cycle enzymes. This suggests that they may play a role in the regulation of enzyme activities in plants.

Genetic Structure of Chimpanzee Populations

Considering the genetic structure of chimpanzees, it is important to understand the genetic diversity within and between species. This information can be utilized to inform conservation strategies and to better understand the evolutionary history of chimpanzees.

Results/Discussion

We analyzed data from the chimpanzee genome to understand the genetic diversity within and between species. Our results provide insights into the evolutionary history of chimpanzees and highlight the importance of conserving genetic diversity for future research.
My methodology (5)

Thinking: what the hell should I work on?

Define a model
Write program, do simulations (i.e. make sure it works)

And/or get data files

Clean the data, find software to analyze the data, or run the program I wrote and which works

Write the results and publish

Do this 3 times, staple the papers, celebrate my PhD, get the hell out of here, get a life!???
The cousins

- Our closest living relatives
- 95–99% of DNA shared between human and chimps
- Living models to understand the evolution of biology and culture of humans
- Endangered environments and species
West African chimpanzees (P. troglodytes verus)
Phylogenetically unclear at present (P. troglodytes)
E Nigeria / W Cameroon chimpanzees (P. troglodytes vellerosus)
Central African chimpanzees (P. troglodytes troglodytes)
East African chimpanzees (P. troglodytes schweinfurthii)
Bonobos (P. paniscus)

Won & Hey 2005
Shared trait between Bonobo and chimpanzee

knuckle walkers
Shared trait between Bonobo and chimpanzee

philopatric males within groups
While the young females disperse
Shared trait between Bonobo and chimpanzee

philopatric males within groups
While the young females disperse
Shared trait between Bonobo and chimpanzee

fission-fusion societies
Shared trait between Bonobo and chimpanzee

multi-male mating → sperm competition
Shared trait between Bonobo and chimpanzee

multi-male mating → sperm competition
Human is not a monogamous organism!
Does the size matter? (1)
Does the size matter? (2)

Gorilla penis is 3 cm long, whereas a chimpanzee penis reaches about 8 cm → Human males have the larger penis of the great apes! I love you guys 😊

- **Evolutionary advantage related to female orgasm?**
  - If you get the female to enjoy sex, she’ll ask for more, you’ll get more chance to reproduce
- **Social status of males related to their penis size?**
  - Since you guys seem to be so obsess about its size, maybe hierarchy between males is dependent of it? The higher the status, the higher mate choices and reproductive chances?
Differences - environment

Chimps: Montane forest - dry savana - woodland

Bonobos: Dense tropical forest
Differences - diet

Chimps eat mostly fruits, leaves + meat & hunting

Bonobos are Strict vegetarian
Differences - locomotion

Chimps are mostly terrestrial

Bonobos are mostly arboreal when terrestrial bipedal
Differences - social group

- Chimps = male dominant + politics
- ♀ submit to ♂
- ♀ do not associate
Differences - social group

- Bonobos: females dominant + politics
- ♂ submit to ♀
- ♀ associate together
Differences - mum/son relationships

mother-son association until adolescence

mother-son association throughout life
Differences - tool use

Chimps use rudimentary tool

Bonobos can, but do not in the wild
**Differences** - female stuffs

- visible ovulations
- 8 months pregnancies

- concealed ovulations
- 8.5 months pregnancies
Differences - behavior

Chimps are aggressive - intercommunity warfares

Zoo chimps rip off man’s face

BY SUZY AUSTIN

A zoo visitor was yesterday having surgery to rebuild his face after being mauled by two chimps at a tea party.

Saint James Davis had most of his face torn off and his testicles and a foot severed in the attack.

He and his wife, LaDonna, were injured as they celebrated the birthday of a chimp they had rescued from Africa.

Two other apes flew into a jealous rage – the attack lasted only when the animals were shot dead.

The couple were visiting Moe for his 59th birthday at a wildlife sanctuary in California.

Mrs Davis, 61, said they were giving him a cake when two chimps, Ollie and Buddy, escaped from their cages and attacked.

She said: "I turned around and they started charging."

Mrs Davis was pushed by one of the chimps and had her thumb bitten off, she said. ‘James saw that, pushed me behind a table and took the brunt of everything else,” she added.

Two keepers at the sanctuary shot the apes dead.

Moe, a 1m tall chimpanzee weighing 60kg, was banished to the sanctuary in September 1999 after biting off a woman’s finger.

“It looks as though Mr Davis did something to rile the chimps,” a spokeswoman for the sanctuary said. ‘It is very rare for a chimp to launch a sustained attack where they will not stop until they have killed.’

Mr and Mrs Davis, of California, rescued Moe during a visit to Tanzania in the 1960s. They raised the chimp as their son — teaching him to wear clothes, take showers and use the toilet.
Differences - behavior

Bonobos: sexual display - pacific intercommunity
Back to science: 1st story

Genetic Structure of Chimpanzee Populations

Celine Becquet, Nick Patterson, Anne Stone, Molly Przeworski and David Reich. 2007, PloS Genetics 3 (4): 617-626
Genetic Structure of Chimpanzee Populations

- **Data**
  - 84 individuals (16 Central, 7 Eastern, 41 Western, 11 Unknown, 3 Hybrids, 6 Bonobos)
  - 310 Microsatellites

- **Goals of the study**
  - Do the subspecies labels correspond to well-defined genetic populations?
  - Is there evidence of gene flow between them?
  - How many markers do we need to assign individuals? What about the individuals of unknown origins?
STRUCTURE with four clusters

- Central
- Eastern
- Western
- Unknown
- Bonobos
- Hybrids

Wild-caught hybrids
Conclusion

- **Labels = genetics structure**
  - Found only 3 populations of common chimps
  - Little gene flow in the wild
  - 30 markers enough for assigning individuals to pop.

- **Differentiation between**
  - Eastern/Central ~ human populations
  - Western/central+eastern ~ 3 x human pop.
  - Bonobo/ common chimp ~ 6x human pop.

- **Eastern and Central chimpanzees are phylogenetically most closely related**
Some real science now: 2nd story
Dating the splits in apes

A new approach to estimate parameters of speciation models, with application to apes

Celine Becquet, Molly Przeworski, Genome Research, in review
Allopatry vs. parapatry

How often does speciation occur in the presence of gene flow between diverging populations?

Adapted from Wu & Ting 2004
Why else do we care?

- **Demographic inference**
  - When did two populations start diverging?
  - Did gene flow occur since their split?

- **Implications for**
  - Speciation
  - Natural history
  - Conservation biology
The Isolation-Migration model

- Population mutation rates:
  - $\theta_1 = 4N_1\mu$
  - $\theta_2 = 4N_2\mu$
  - $\theta_A = 4N_A\mu$

- Time of divergence: $T$ (in generations)

- Migration rate: $M = 4N_1 m$
Data
### Existing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Fast</th>
<th>Approach</th>
<th>Full data</th>
<th>Multiple loci</th>
<th>Gene flow</th>
<th>Rec</th>
<th>Genotype</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wakeley &amp; Hey 1997</td>
<td>✓</td>
<td>Method of moments estimator</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>Low accuracy S_s &amp; S_f &gt; 0</td>
</tr>
<tr>
<td>Nielsen &amp; Wakeley 2001</td>
<td>—</td>
<td>MCMC</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Hey &amp; Nielsen 2004</td>
<td>—</td>
<td>MCMC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>Locus-specific gene flow</td>
</tr>
<tr>
<td>Leman et al. 2005</td>
<td>✓</td>
<td>Approximate Bayesian Computation</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
</tr>
<tr>
<td>MIMAR</td>
<td>—</td>
<td>MCMC</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Mutation are neutral along genealogy $G$:

- $S_k \sim \text{Poisson}(L_k \theta)$
- $k \in [1,4]$ refers to the branch type
- $\theta = \text{locus-specific pop. mutation rate}$

$S_1 = S_{\text{unique to pop}1=1}$
$S_2 = S_{\text{unique to pop}2=1}$
$S_3 = S_{\text{shared by both pop}=1}$
$S_4 = S_{\text{fixed in either pop}=1}$

Wakeley and Hey, 1997; Leman et al., 2005
Goals of the method - MIMAR

- **MCMC** estimation of the **Isolation-Migration model** Allowing for **Recombination**
  - *mimar verbo transitivo*
    - *(acariciar)* to caress, to fondle
    - *(consentir)* to pamper, to spoil
    - *(tratar con cariño)* to fuss over

- **Estimate split time, ancestral population size and detect gene flow**
  - **Bayesian Monte Carlo approach**
  - **Uses summary of polymorphism data**
    - \( D_y = \{ S_{y1}, S_{y2}, S_{y3}, S_{y4} \} \) at locus \( y \)
    - From \( Y \), independent loci surveyed in two closely related populations
  - **Allows for intra-locus recombination**
  - **Uses MCMC** to explore parameter space efficiently
Build your own talk!

- Pick among the following options

1. Ignore the next 20 slides
   - Good reasons
     - It’s full of maths and complex statistical methods and you really HATE maths
     - I’ll certainly mess up the explanations because it’s Friday and I HATE maths as well
   - Bad reasons
     - You’ll never know how smart I am and what I need to put up with every day
     - You’ll miss a chance to understand some pop. genetics

2. Go through the next 20 slides
   - Good reasons
     - You’ll discover how smart I am and what I need to put up with every day
     - You’ll get a chance to understand some pop. genetics
   - Bad reasons
     - It’s full of maths and complex statistical methods and you really HATE maths
     - I’ll certainly mess up the explanations because it’s Friday and I HATE maths as well
Estimation method

- Likelihood framework

- Bayesian approach - Posterior distribution of interest:

\[
\pi(\Theta \mid D) \propto p(D \mid \Theta)p(\Theta)
\]

- \( D=(D_1,\ldots,D_Y) \) for \( Y \) loci
- \( \Theta = (T, \theta_A, M, \theta_1, \theta_2, P) \)
  - \( P=(\rho_1,\ldots,\rho_Y) = \text{nuisance parameter} \)
  - \( r_y=\rho/\theta_1 \sim \text{exp}(1/\lambda) \), chosen independently between loci and steps of MCMC
1st step: chose parameter from prior distributions (1)

Chose the parameters from uniform prior distributions
1st step: chose parameter from prior distributions (2)

Chose the parameters from uniform prior distributions

Generate a genealogy for a locus

Compute the likelihood of the data | G

\[
p(\mathbf{D}_y \mid G, \Theta_1) = \prod_{k=1}^{4} Po(S_k \mid L_k \theta_{locus})
\]
1\textsuperscript{st} step: chose parameter from prior distributions (3)

Chose the parameters from uniform prior distributions

Generate a genealogy for a locus

Posterior: prob. of parameter|data

\[
\pi(T|D) \quad \pi(M|D)
\]

\[
\pi(T|D) \quad \pi(M|D)
\]

\[
\ln(M)
\]

\[
\ln(M)
\]

Compute the likelihood of the data|\(G\)

\[
p(D_y | G, \Theta_1) = \prod_{k=1}^{4} Po(S_k | L_k \theta_{\text{locus}})
\]
2\textsuperscript{nd} step: Update parameters using MCMC (1)

Chose parameters from distributions around previous value

\[ q(T_2) \]

\[ q(M_2) \]

\[ T_1 \]

\[ T_2 \]

\[ M_1 \]

\[ M_2 \]
2nd step: Update parameters using MCMC (2)

Chose parameters from distributions around previous value

Generate a genealogy for a locus

Compute the likelihood of the data $|G$

$$p(D_y | G, \Theta_2) = \prod_{k=1}^{4} Po(S_k | L_k \theta_{locus})$$

$$h = \min\left(1, \frac{p(D | G', \Theta_2)}{p(D | G, \Theta_1)}\right)$$
2nd step: Update parameters using MCMC (3)

Chose parameters from distributions around previous value

Generate a genealogy for a locus

Compute the likelihood of the data|G

\[
p(D \mid G, \Theta_2) = \prod_{k=1}^{4} P_0(S_k \mid L_k \theta_{locus})
\]

Move to \((\Theta_2)\) with probability \(h\) or else record \((\Theta_1)\)
Lots of steps: Update parameters using MCMC

Repeat a lot of times (wait ~20 days)

Hopefully, at the end you obtain posterior distribution of the parameters given the data
Conclusions from simulated data

- Precise/accurate estimates of $\theta_A$, time of split
- Reasonable power to detect gene flow
- Comparable to full likelihood approach
  - On wide range of parameters
  - Summary statistics carry much of the information for short loci

Advantage of MIMAR

- Flexible demographic model
- Easily extendable to consider other summaries
  - As long as $p(D_y|G_{jx}, \Theta)$ can be readily calculated
MIMAR application to apes data sets

  Resequencing data for:
  - 69 loci in
    - 9 bonobos
    - 10 (5) Central chimpanzee
    - 10 (6) Western chimpanzees
  - 26 loci in
    - 10 Eastern chimpanzees
  - 15 loci in
    - 16 Western gorillas
    - 3 Eastern gorillas
  - 19 loci in
    - 6 Sumatran orangutans
    - 10 Bornean orangutans
Results – Bonobo and Common chimpanzee

Bonobo: ~725-850ky
Common chimpanzee: ~35,000

Common chimpanzee: ~10,000-20,000
Bonobo: ~10,000

~725-850ky
Results - Common Chimpanzees

~300–600ky

~220ky

~0.9

~10,000

~20,000

~16,000–30,000

~25,000

~0.3

~0.4

~10,000

~6,000

~300–600ky

Western

Central

Eastern
My older cousins, the gorillas
Geographical range of Gorillas

Clifford et al. (2004)
Results - Gorilla populations

- Western population: ~6,000
- Eastern population: ~4,500
- Total estimated population: ~15,000

Time since divergence: ~110ky
The other cousins, the orangutans
Geographical range of Orangutans

Sumatran

Bornean

VAN SCHAIK et al. 2003
Results – Orangutan populations

- Bornean: ~31,000
- Sumatran: ~17,000
- ~9,000

~220ky
Goodness of fit of isolation-migration models

- The distribution of the summaries of the data simulated under the estimated models fit the observed values from the ape data sets.

- Exception for Central/Eastern divergence
  - Poor fit to $F_{ST}$, $\pi$ and Tajima’s $D$ from Central chimpanzees
  - Isolation-migration model not appropriate?
    - Low $F_{ST}$ → isolation by distance model?
    - Little evidence of gene flow?
  - Missing important demographic features?
    - Known complex demographic history for Central chimpanzees
Conclusion & Future directions

- We have developed a new method to estimate parameters of an isolation-migration model, which allows for intra-locus recombination.
- When applied to great ape populations/species, we find a good fit to the isolation-migration model (with the exception of Central/Eastern?)
- We plan to apply the method to polymorphism data from a wide range of species, and test the prevalence of parapatry.
- Investigate possible demographic events that lead to high $N_A$ and more generally, robustness of *in silico* inferences regarding allopatry vs. parapatry.
Thanks
Thanks