Leveraging distant relatedness to quantify human mutation and gene conversion rates

Pier Palamara
Harvard T. H. Chan School of Public Health
ASHG 2015.10.10
Methods for inferring the mutation rate

generations

Methods for inferring the mutation rate

Generations

100,000s - Phylogenetic methods

[image: Pääbo, Nature 2003]
Methods for inferring the mutation rate

- Phylogenetic methods

10,000s

- Trios

1 generation


[Image: Šešelj et al., Nature 2003]
Methods for inferring the mutation rate

Different estimates: $2.4 \times 10^{-8}$ vs $1.2 \times 10^{-8}$


Methods for inferring the mutation rate

1,000s

Deep genealogical relationships

100,000s

Phylogenetic methods

1

Trios

e.g. [Lipson et al. PLOS Gen. 2015 (in press)]
[Image: Tishkoff and Verrelli, 2003]
Methods for inferring the mutation rate

- 100,000s: Phylogenetic methods
- 1,000s: Deep genealogical relationships
- 10s: Recent genealogical relationships
- 1: Trios

This work
Identity By Descent

see e.g. [Browning & Browning, Annual Review of Genetics 2012]
Identity By Descent

see e.g. [Browning & Browning, Annual Review of Genetics 2012]
Inferring mutation rate in “unrelated” individuals

- *tMRCA regression*: Regress IBD sequence mismatching rate on age of segments.
Inferring mutation rate in “unrelated” individuals

- *tMRCA regression:* Regress IBD sequence mismatching rate on age of segments.
Inferring mutation rate in “unrelated” individuals

- *tMRCA regression*: Regress IBD sequence mismatching rate on age of segments.
Inferring mutation rate in “unrelated” individuals

- *tMRCA regression*: Regress IBD sequence mismatching rate on age of segments.

\[
\text{IBD mismatching rate} = 2 \times \text{IBD segment age}
\]

\[
\text{slope} = \text{mutation rate}
\]
Inferring mutation rate in “unrelated” individuals

- **tMRCA regression**: Regress IBD sequence mismatching rate on age of segments.

![Graph showing the relationship between IBD mismatching rate and 2 × IBD segment age. The slope represents the mutation rate, and the intercept is approximately the genotype error.]
Inferring the age of IBD segments

Unknown TMRCA
Infer from demographic history

[Palamara et al. AJHG 2012]
[Ralph & Coop, PLOS Bio. 2013]
Dealing with non-crossover gene conversion

- Gene conversion occurs at a rate proportional to recombination
- When it occurs, an existing SNP may be copied on IBD haplotypes
Dealing with non-crossover gene conversion

- Gene conversion occurs at a rate proportional to recombination
- When it occurs, an existing SNP may be copied on IBD

\[ 2 \times T_{\text{MRCA}} \]

\[ 2 \times \text{IBD segment age} \]

... with probability proportional to number of generations
Dealing with non-crossover gene conversion

• Gene conversion occurs at a rate proportional to recombination
• When it occurs, an existing SNP may be copied on IBD

... with probability proportional to number of generations and variant frequency...
Non-crossover gene conversion: MaAF regression

- Solution: perform a second regression, now using threshold on maximum MAF variants in sequence
Non-crossover gene conversion: MaAF regression

- Solution: perform a second regression, now using threshold on maximum MAF variants in sequence

![Graph showing relationship between maximum MAF and inferred mutation rate. Dashed line indicates variants with MAF < 0.4.](image)
Non-crossover gene conversion: MaAF regression

- Solution: perform a second regression, now using threshold on maximum MAF variants in sequence

![Graph showing inferred mutation rate vs. maximum MAF]

- Gene conversion-corrected estimate
Non-crossover gene conversion: MaAF regression

- Solution: perform a second regression, now using threshold on maximum MAF variants in sequence

If population heterozygosity is known, can infer rate of gene conversion
• Results: simulation and real data
tMRCA regression is robust to genotyping error

![Graph showing the relationship between simulated error rate and inferred mutation rate. The graph indicates that the tMRCA regression remains robust across a range of simulated error rates.](image-url)
IBD approach is more efficient than trio approach
Real data: the Genome of the Netherlands

• ~250 trios\(^1\)
• ~13x coverage (~26x on transmitted haplotype)
• Trio-phased using MVNcall\(^2\)
• IBD detected using GERMLINE\(^3\) (+ filtering)
• Demographic history (piece-wise expansion) inferred using DoRIS\(^4\)

1: [Francioli et al., Nat. Gen. 2014]
2: [Melanou & Marchini, Bioinformatics 2013]
3: [Gusev et al., Gen. Res. 2009]
4: [Palamara et al., AJHG 2012]
When gene conversion correction is applied, for segments $> 1.6\text{cM}$, $\mu = 1.66 \times 10^{-8}$, s.e. $0.04 \times 10^{-8}$

Higher than pedigree-based $\mu$
Inferring gene conversion rate in real data

- When gene conversion correction is applied
  $\mu = 1.66 \times 10^{-8}$, s.e. $0.04 \times 10^{-8}$

- Gene conversion rate of $5.99 \times 10^{-6}$, s.e. $0.69 \times 10^{-6}$

(Matches estimate of Williams et al. eLife 2015)
Inferring indel rate in real data

• When gene conversion correction is applied
  \[ \mu = 1.66 \times 10^{-8}, \text{s.e.} 0.04 \times 10^{-8} \]

• Gene conversion rate of \(5.99 \times 10^{-6}, \text{s.e.} 0.69 \times 10^{-6}\)

• Same method can be applied to estimate rate of short indels
  \[ \mu_{\text{indel}} = 1.26 \times 10^{-9}, \text{s.e.} 0.06 \times 10^{-9} \]

(Compatible with Besenbacher et al. Nat. Comm. 2015)
Recombination ↔ Mutation

- Rec. and mut. rates strongly correlated ($p<10^{-5}$)
- After controlling for gene conversion, no association ($p=0.17$)
B statistic closely reflects local IBD sharing ($p<10^{-6}$)
But no impact on mutation rate estimate ($p=0.19$)

B statistic: [McVicker et al. PLOS Gen. 2009]
Other analyses

- Mismatching variants on IBD enriched for deleterious variation
- No evidence for enrichment/depletion of mutation rate in several genomic annotations

Chromatin marks identify critical cell types for fine mapping complex trait variants

Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu & Soumya Raychaudhuri

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium
Conclusions and future work

• New method to infer mutation and gene conversion rates
  – $\mu = 1.66 \times 10^{-8}$ (higher than pedigree studies)
    • Agrees with recent estimate of Lipson et al. PLOS Gen. 2015 (in press)
  – No effects of recombination/selection on estimate
  – No enrichment/depletion in functional annotations

• Use in multi-generation pedigree data
Acknowledgements

*Palamara et al. AJHG 2015 (in press) available on BioRxiv*

(Harvard University) Alkes Price, John Wakeley, Shamil Sunyaev, Alexander Gusev, Peter Wilton, Hilary Finucane, Sriram Sankararaman

(University Medical Center Utrecht) Laurent Francioli, Paul de Bakker

(Columbia University) Itsik Pe’er

(Broad Institute) Giulio Genovese

Genome of the Netherlands consortium

Paul de Bakker
Cisca Wijmenga
Clara C Elbers
Sara L Pulit
Androniki Menelaou
Laurent Francioli
Abdel Abdellaoui
Albert Hofman
Alexandros Kanterakis
Andre G Uitterlinden
Anton JM de Craen
Ben Oostra
Bruce H Wolffenbuttel
Cornelia M van Duijn
Dorret Boomsma
Eka HD Suchiman
Eline P Slagboom
Fernanodo Rivadeneira
Freerk van Dijk
Gert-Jan van Ommen
Gonneke Willemsen
Heorhiy Byelas
Hongzhi Cao
Jeanine Houwing-Duistermaat
Itsik Pe’er
Jeroen FJ Laros
Jessica van Setten
Johan den Dunnen
Jouke Jan Hottenga
Jun Wang
Kai Ye
Karol Estrada
Lennart C Karssen
Marian Beekman
Martijn Dijkstra
Martijn Vermaat
Mathijs Kattenberg
Morris A Swertz
Ning Li
Paz Polak
Peter de Knijff
Pier Palamara
Pieter B Neerincx
Qibin Li
Ruoyan Chen
Shamil Sunyaev
Suje Cao
Victor Guryev
Vyacheslav Koval
Wigard Kloosterman
Yingrui Li
Yuanping Du

Funding

**NIH** R01 MH101244, U54 CA121852-06; **NSF** 08929882, 0845677
Genotyping error is captured by intercept.
IBD approach is more efficient than trio approach

![Graph showing standard error of estimate across different samples for Trio, GoNL, and MASAI methods.]
No effects of background selection on inference
Time (generations)

Effective size

- Ashkenazi
- European
- Masai
- Dutch
Derived allele frequency
Probability

- alpha = 0.01
- alpha = 0.5
- alpha = 1

![Graph](image-url)
Minimum IBD segment length (cM)

- 1.0
- 1.5
- 2.0
- 2.5

Inferred mutation rate

- 1.5e−08
- 2.0e−08
- 2.5e−08
- 3.0e−08

No correction
With gene conversion correction

Graph showing the relationship between Minimum IBD segment length (cM) and Inferred mutation rate.
\[ q = -2.719 \times 10^{-05} + 1.480 \times 10^{-03} F \]

\[ r^2 = 0.9964 \]

\[ q = -2.436 \times 10^{-06} + 1.170 \times 10^{-04} F \]

\[ r^2 = 0.9966 \]
Average recombination rate in region

Inferred gene conversion rate per bp

Average recombination rate in region

Inferred gene conversion rate per bp