Beyond the molecular clock

Kelley Harris
University of Washington

Workshop on Population and Speciation Genomics
January 31, 2018
Forces that shape genomic diversity

- **Mutation**
- **Genetic Drift**
- **Selection**

Image co-artist: Natalie Telis
Mutations as a molecular clock
When clock breaks down (runs out of batteries?)

- Almost every population genetic method assumes that mutagenesis = a super boring clock like process

- This assumption works fine until it doesn’t

- The mutation process has cool, complex features that can trip you up if you aren’t looking out for them
Molecular clock 101

• Mutagenesis is more clock-like over short timescales compared to long time scales

• A simple branch length test can reveal whether mutagenesis is clock-ish in your data:

\[ D(A,O) = D(B,O) \]

Data can fail this test due to mutation rate variation, selection, or introgression
Violation of molecular clock over very long timescales

Drake 1991

The error threshold

• A simple model by Eigen & Schuster (1979) justifies Drake’s rule

• Consider a “master” virus with fitness 1+s and genome length $L$

• All mutant viruses have fitness 1

• The master sequence will die out due to Muller’s rachet/“error catastrophe” if and only if the mutation rate $\mu$ is below a threshold:

• $\mu < \log(s)/L$
Stable quasispecies vs error catastrophe

Lauring and Andino 2010

$\mu < \text{error threshold}$

$\mu > \text{error threshold}$

Lauring and Andino 2010
How might I gather some mutation rate data to test this weird theory?
Measuring mutation rates with mutation accumulation (MA) lines

Keightly and Charlesworth 2005
MA with a reporter gene

Inactive/broken promoter

Point mutations can restore promoter function

Reporter gene (e.g. encoding GFP or luciferase)

DNA

mRNA

A reporter protein
Amount is easily measured (e.g. GFP by fluorescence)
Mutation rate estimates vary enormously in quality

- Your PSMC results depend heavily upon a mutation rate number. Where might that number come from?

- MA experiment + whole genome sequencing ($$-$$$$$)

- MA experiment + reporter gene sequencing (cheap today, only game in town 10 years ago)

- Back-of-the-envelope calculation (substitutions / estimated divergence time)

- Whole-genome trio sequencing ($$$$$$$$$$$)
Drake’s rule driven mostly by viruses and bacteria

The mechanisms responsible for the discontinuity in scaling of mutational diversity with genome size are complex and likely involve a combination of factors that depend on the specific characteristics of the organisms involved.

Once past the critical effective population size of approximately $10^6$, the mutation rate and effective population size are negatively correlated. However, the drift-barrier hypothesis predicts that mutation rates should be independent of effective population size.

In general, the trend observed is that mutation rates are lower in species with larger effective population sizes. This is consistent with the idea that larger populations are more likely to maintain lower mutation rates due to the effects of genetic drift.

The data presented here are consistent with previous observations that mutation rates are lower in species with larger effective population sizes, although there are some exceptions to this pattern.

The relationship between mutation rate and effective population size is also affected by the nature of the mutation rate estimates. Some estimates are based on mutation rates from reporter constructs, which may not be fully representative of the mutation rates in the natural population. Additionally, some estimates may include biases due to selective pressures that affect the accumulation of mutations.

Overall, the data suggest that mutation rates are inversely proportional to effective population sizes, with the relationship being more pronounced in species with larger effective population sizes.
Why should effective population size affect mutation rate?

Why is the mutation rate what it is?
1. The Cost-of-Fidelity Model

Lynch *Trends in Genetics* 2010  
Sung, *et al.* *PNAS* 2012
2. The Drift-Barrier Hypothesis

![Graph showing the relationship between mutation rate and cost of fidelity for Drift Barrier and Biophysics Barrier Mutation Rates.](image)

- Drift Barrier Mutation Rate
- Biophysics Barrier Mutation Rate

- Excess Mutation Load: $\sim \frac{1}{Ne}$

Lynch *Trends in Genetics* 2010  
Sung, *et al.* *PNAS* 2012
Mutators can be favored in asexual organisms

- Expected extra load of deleterious mutations must not exceed the expected benefit of beneficial mutations
- Robustness to environmental change
- Stress-induced mutagenesis?
Stress-Induced Mutagenesis in Bacteria

Elevated Mutagenesis Does Not Explain the Increased Frequency of Antibiotic Resistant Mutants in Starved Aging Colonies

Rate measurements from whole genome sequencing

Reported gene mutation rate estimates
Selection against mutator alleles is weak in sexual organisms.
Other factors affecting the mutation rate

Environmental Mutagens

Life history
Male mutation bias

Hurst and Ellegren 1998
Paternal age effect

Amster and Sella 2016

Expected number of mutations

Age at reproduction (years)

Slope during spermatogenesis ($D_M/\tau$)

Onset of puberty ($P$)

Males

Females

$C_M$

$\mu_F$

Mutation rate per year

Average gen. time (Y)

X-to-autosome ratio

Average gen. time (Y)

Catarrhines

Mammals

$10^{-9}$
Branch length ~ number of substitutions
Label = Estimate of
(male mutation rate)
/(female mutation rate)
Two additional *de novo* mutations per year of paternal age

A small but significant maternal age effect (0.5 muts/year)

Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16

Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16.

Other causes of mutation rate variation along the genome

- Replication timing
- Transcription-associated-mutagenesis (TAM) and transcription-coupled-repair (TCR)
- Non-B-DNA structures and other DNA repeats
- Chromatin state
Replication timing


Francioli, et al. 2015
Replication and transcription induce strand asymmetry

Excess of G+T over A+C on coding strand of most genes

Green, et al. 2003
Measuring the human mutation rate

Human

Chimpanzee

ATCCAGTGCG
ATGCAGTCCCG

2.5e-8 mutations per site per gen

1.0e-8 mutations per site per gen

Parent-child trios

1000 Genomes Consortium 2010

Nachman and Crowell 2001
DNA mutation clock proves tough to set

Geneticists meet to work out why the rate of change in the genome is so hard to pin down.

Ewen Callaway
10 March 2015

- What is the real human mutation rate?
- Has the mutation rate slowed down during recent human history?
The Hominoid Mutation Rate Slowdown

Too large to explain by paternal age effect alone (Scally and Durbin 2012)

Have genetics and environment played a role?

Adapted from http://www.bio.indiana.edu/graduate/multidisciplinary/GCMS/trainees/thomas_gregg.php

“The” mutation rate encompasses a menagerie of mutation types

Point Mutations

Multinucleotide Mutations

Small indels

Large Copy Number Changes

Transitions

Transversions

Other Transversions

GC-conservative Transversions
CpG Mutations

- Many species (incl humans, not incl *Drosophila*) methylate C when it’s next to G (C-phosphate-G)

- CpG methylation regulates gene expression
CpG sites are hypermutable

- On average, CpG sites have a 30-fold higher mutation rate than other C’s in the human genome.

- 70-80% of CpGs are methylated in mammals; most unmethylated CpGs are part of CpG islands.

- Fewer than 1% of dinucleotides in the human genome are CpGs, although the expected frequency is $0.21 \times 0.21 = 0.0441 = 4.41\%$. 
CpG transitions are somewhat more clocklike than other mutations

In a tree of 19 mammals, CpG mutations yield a more clocklike tree than mutations occurring in other contexts.

Hwang and Green 2004

CpG mutations also appear more clocklike than other mutations in great ape tree.

Limits to clock-like behavior of CpGs

Embryonic development

Bulk sequencing of tumor samples

Single-cell sequencing of normal neurons

CpG mutation count

Age of patient at diagnosis

Fast renewing tissue

Slowly renewing tissue

Number of somatic mutations

A (15yr, F)

B (17yr, M)

C (42yr, F)

Others

C>T

Highly efficient repair

Time between two consecutive cell divisions
“Mutational signatures” of types of DNA damage in cancer

<table>
<thead>
<tr>
<th>Tobacco exposure</th>
<th>Error-prone Polymerase ε activity</th>
<th>Off-target DNA editing by APOBEC enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>C → A</td>
<td>TCT → TAT</td>
<td>TCT → TTT</td>
</tr>
<tr>
<td>CC → AA</td>
<td>AAA → ACA</td>
<td>TCA → TTA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCT → TGTT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TCA → TGTA</td>
</tr>
</tbody>
</table>
APOBEC / AID deaminases

• APOBEC attacks RNA viruses, mutating TCA and TCT by deamination

• Its homologue AID hypermutates T cell receptors for proper immune function

• Both cause off-target germline mutations, especially in endogenous retroviral sequences

• APOBEC is erroneously switched on in many cancers (esp cervical), associated with poorer outcomes
Mutations that involve different classes of nucleotides; that is, purine-to-pyrimidine or pyrimidine-to-purine mutations.

Mutational processes

Biological activities that generate mutations; each of these processes comprises both a DNA repair component and a DNA repair component. In this hypothetical cancer genome, arrows indicate the duration and intensity of exposure to a mutational process. The final mutational portrait is the sum of all of the different mutational processes (A–D) that have occurred in a lifetime.

The pattern of mutations arises from the insertion of another in DNA. A type of mutation in which one base is replaced by another in DNA.

For example, mutations associated with smoking-related tumour type and related to exogenous carcinogens. Mutational processes can be ongoing or historical depending on the mutational processes are no longer active. Signature A represents deamination of methylated cytosines, which could be used as prognostic indicators, as predictors of therapeutic sensitivity or as targets of disease control. The total genetic changes that is, the sum of all observed in a cancer genome; the mutational signatures that is, the sum of all mutational processes that have occurred in a lifetime.

The total genetic changes that is, the sum of all mutational processes that have occurred in a lifetime. The final mutational portrait is the sum of all of the different mutational processes (A–D) that have occurred in a lifetime.

Mutational processes

Biological activities that generate mutations; each of these processes comprises both a DNA repair component and a DNA repair component. In this hypothetical cancer genome, arrows indicate the duration and intensity of exposure to a mutational process. The final mutational portrait is the sum of all of the different mutational processes (A–D) that have occurred in a lifetime.

The pattern of mutations arises from the insertion of another in DNA. A type of mutation in which one base is replaced by another in DNA.

For example, mutations associated with smoking-related tumour type and related to exogenous carcinogens. Mutational processes can be ongoing or historical depending on the mutational processes are no longer active. Signature A represents deamination of methylated cytosines, which could be used as prognostic indicators, as predictors of therapeutic sensitivity or as targets of disease control. The total genetic changes that is, the sum of all observed in a cancer genome; the mutational signatures that is, the sum of all mutational processes that have occurred in a lifetime.
Mutation signature analysis

Nonnegative matrix factorization

Mutation counts in 96 triplet contexts across cancers
Effect of BRCA germline mutations on breast cancer mutation distribution
Mutational signatures in the germline?

Image co-artist: Natalie Telis
Hypothesis: different germline signatures have different evolutionary histories.
Private European SNPs are enriched for a mutational signature of unknown origin.
A signature of elevated mutagenesis in the European germline

Harris *PNAS* 2015
Visualizing differences between mutation spectra

More TCC-to-TTC transitions in Europeans
More CG-to-TG transitions in Africans

Mutation type odds ratio
Europe vs Africa

Harris and Pritchard eLife 2017
Beyond 3-mers to 7-mers

Aggarwala and Voight 2016

Carlson, et al. 2017
Genes mirror geography within Europe

John Novembre¹,², Toby Johnson⁴,⁵,⁶, Katarzyna Bryc⁷, Zoltán Kutalik⁴,⁶, Adam R. Boyko⁷, Adam Auton⁷, Amit Indap⁷, Karen S. King⁸, Sven Bergmann⁴,⁶, Matthew R. Nelson⁸, Matthew Stephens²,³ & Carlos D. Bustamante⁷
TCC-to-TTC transitions are enriched in South Asia as well as Europe.

A-to-T and AC-to-CC transversions are enriched in East Asia.
TCC-to-TTC transitions are enriched in South Asia as well as Europe.

A-to-T and AC-to-CC transversions are enriched in East Asia.

How and when did this mutation spectrum variation arise?
Hypothetical Signature of a TCC-to-TTC mutation rate increase

TCC-to-TTC mutation rate

Present  Past

Time

TCC-to-TTC mutation fraction

Low  High

Allele frequency
Pulse replicates in the UK10K data

Harris and Pritchard eLife 2017
A pulse of TCC-to-TTC mutations in Europe and South Asia?
Expected TCC fraction as a function of allele frequency

- Partition time into discrete intervals
- \( A(k, i) = \text{the total branch length subtending } k \text{ lineages between times } T_i \text{ and } T_{i-1} \)
- \( r_i \sim \text{the rate of TCC mutations between } T_i \text{ and } T_{i-1} \)

Expected TCC fraction as a function of allele frequency is

\[
E[f(k)] \sim \frac{\sum_i A(k, i) \cdot r_i}{\sum_i A(k, i)}
\]
Inference of a mutation pulse lasting from 15,000 to 2,000 years ago
A younger Japanese mutation pulse

Harris and Pritchard eLife 2017
Great ape mutation spectrum evolution

Within-species SNPs from 79 great ape whole genomes (Prado-Martinez, et al. 2013)

Harris and Pritchard eLife 2017
Future direction: are mutation pulses the relics of lost mutator alleles?

- Fixed differences between great ape spectra
  - European TCC mutation pulse
  - Japanese mutation pulse
How mutator alleles could promote rapid mutation spectrum turnover

Sawyer and Malik *PNAS* 2006
Positive selection in DNA repair genes and other housekeeping genes

- BRCA1 & BRCA2 are under positive selection in primates
- 5 Nonhomologous end joining genes experienced positive selection during primate evolution, incl XRCC4 which has been under selection in Europeans
- Iron-uptake receptor TfR1 evolves under positive selection to avoid facilitating viral entry

Demogines, et al. 2010
Demogines, et al. 2013
A case study of a mutational process that complicates population genetics

Multinucleotide mutations (MNM) are nearby SNPs that appear in the same generation

AAAGTTTAGCCGACAC

↓

AAAGATAAACCAGACAC

Schrider, et al. 2011

Harris and Nielsen 2014
Effect of MNMs in the distribution of tracts of identity by state

Ancient demographic history

Recent demographic history
Direct evidence for MNMs

• Most methods assume that all SNPs arise from rare, independent mutation events

• MA experiments and trio sequences show that *de novo* mutations are too clustered for this to be true

**Independent mutation hypothesis**

**Multinucleotide mutation**

**Observed: Excess correlation between *de novo* mutations**

**“Mutator” yeast strains: Some abnormal polymerases generate clustered mutations at a higher rate**
MNMs could accelerate evolution across fitness valleys.
Multinucleotide mutation should create pairs of SNPs in *perfect linkage disequilibrium (LD)*
(derived alleles occur in the same set of individuals)

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Perfect LD

```
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Not Perfect LD
Independent mutations at neighboring sites can also create SNPs in perfect LD.
Compared to theoretical predictions, the 1000 Genomes Phase I data (1,092 humans from Africa, Europe, Asia, and the Americas) has excess close-together SNPs in perfect LD.
SNPs in perfect LD are enriched for transversions

- 66% of human mutations are transitions (A>G, G>A, C>T, T>C)

- Pairs of SNPs in perfect LD are enriched for transversions, suggesting a different balance of mutational signatures
Transversion-enrichment as a function of the distance between linked SNPs
A candidate mechanism: error-prone translesion synthesis

Northam et al., Nucleic Acids Res. 2014
Matching mutational signatures between human variation and laboratory yeast

Research Article

DNA Polymerase zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in Saccharomyces cerevisiae

Jana E. Stone, Scott A. Lujan, and Thomas A. Kunkel*
Laboratory of Molecular Genetics and Laboratory of Structural Biology, National Institute of Environmental Health Sciences, NIH, DHHS, North Carolina

- Stone, et al. created yeast deficient in nucleotide excision repair machinery and observed a high MNM rate
- Mechanism: increased translesion synthesis by Pol Zeta

Environmental and Molecular Mutagenesis 53:777–786 (2012)
A matching dinucleotide mutational signature

![Graph showing frequency of di-nucleotide transitions](image-url)
Further characterization of the Pol zeta mutational signature

- GC>AA mutations are concentrated in late-replicating regions of the genome
- Usually occur in GCG context, triggered by CpG deamination followed by polymerase stalling
- CpG deamination is triggered by transcription; usually occurs on transcribed strand
- Some genes contain GC>TT mutation hotspots, including HRAS where the mutation causes the Mendelian disorder Costello Syndrome

Seplyarskiy, et al. 2015
More on the weirdness of Costello Syndrome

- A high penetrance Mendelian disease caused by a nonsynonymous point mutation in the HRAS oncogene
- Causes developmental delay and early childhood tumors
- Most commonly caused by a GC>TT mutation with a mutation rate of $10^{-5}$ per generation (normal mutation rate is $10^{-8}$ per site per generation)
- Biggest risk factor is paternal age
HRAS mutations experience selfish selection within the testis

Goriely and Wilkie 2012
The Harris Lab is recruiting