DNA Insertions and Deletions in the Human Genome

Philipp W. Messer

Max Planck Institute for Molecular Genetics
Genetic Variation

1. Nucleotide mutations
2. Genomic rearrangements
3. DNA insertions / deletions (indels)
Outline

1. Origin and characteristics of indels
2. Indels in protein-coding regions
3. Duplications and genomic correlations
4. Duplications and alignment scores
Part 1

Origin and Characteristics of Indels in the Human Genome

Identifying Insertions/Deletions

Human   Chimp   Rhesus

~5 Myr

High-quality flanks

H: ..ATACCTCGTACAAATAGCGCTGGGTACAGATA.. 
C: ..ATACCTCGTACAAAT--CGCTGGGTACAGGTA.. 
R: ..ATACCTCGTACAAAT--CGCTGCTACAGATA.. 

Insertions can be distinguished from deletions (parsimony)
**Insertion/Deletion Statistics**

**Insertions:**
- ~250 000 events
- 15% SSR

**Deletions:**
- ~450 000 events
- 5% SSR
Molecular Mechanisms

1. Unequal Crossing Over (UCO)

   nonhomologous recombination

2. Replication Slippage (RS)

   slipped-strand mispairing
Indel Signatures for UCO and RS

**Insertions:**
Tandem duplications of preexisting duplicates

**Deletions:**
Remove one copy of preexisting duplicates
Indel Trace Extensions

a) $d > l$
   UCO, RS (insertion)

b) $0 < d < l$
   UCO, RS (deletion)

c) $d = l$
   Tandem duplication

d) $d = 0$
   Random indel
Measured Trace Extensions ($l=8$ bp)
1. Indels occur preferentially in the male germline
2. Indels are not recombination-mediated
Indel Characteristics

1. The majority of insertions are tandem duplications

2. Long preexisting duplicates are often missing

3. Indels occur preferentially in the male germline

4. Indels are not recombination-mediated
Nonhomologous End Joining

- DNA break
- End joining
- Small or no homology required
- Filling in of single strands
Part 2

Indels in Protein-coding Regions of the Human Genome

[Chaux, Messer, Arndt, *BMC Evol Biol* (submitted)]
Indel Rates in Coding Regions
# Genetic Code

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU Phe</td>
<td>UCU Ser</td>
<td>UAU Tyr</td>
<td>UGU Cys</td>
</tr>
<tr>
<td>UUC Phe</td>
<td>UCC Ser</td>
<td>UAC Tyr</td>
<td>UGC Cys</td>
</tr>
<tr>
<td>UUA Leu</td>
<td>UCA Ser</td>
<td>UAA Stop</td>
<td>UGA Stop</td>
</tr>
<tr>
<td>UUG Leu</td>
<td>UCG Ser</td>
<td>UAG Stop</td>
<td>UGG Trp</td>
</tr>
<tr>
<td>CUU Leu</td>
<td>CCU Pro</td>
<td>CAU His</td>
<td>CGU Arg</td>
</tr>
<tr>
<td>CUC Leu</td>
<td>CCC Pro</td>
<td>CAC His</td>
<td>CGC Arg</td>
</tr>
<tr>
<td>CUA Leu</td>
<td>CCA Pro</td>
<td>CAA Gln</td>
<td>CGA Arg</td>
</tr>
<tr>
<td>CUG Leu</td>
<td>CCG Pro</td>
<td>CAG Gln</td>
<td>CGG Arg</td>
</tr>
<tr>
<td>AUU Ile</td>
<td>ACU Thr</td>
<td>AAA Lys</td>
<td>AGU Ser</td>
</tr>
<tr>
<td>AUC Ile</td>
<td>ACC Thr</td>
<td>AAG Lys</td>
<td>AGC Ser</td>
</tr>
<tr>
<td>AUA Ile</td>
<td>ACA Thr</td>
<td>AGG Lys</td>
<td>AGG Arg</td>
</tr>
<tr>
<td>AUG Met</td>
<td>GCU Ala</td>
<td>GAU Asp</td>
<td>GGU Gly</td>
</tr>
<tr>
<td></td>
<td>GCC Ala</td>
<td>GAC Asp</td>
<td>GGC Gly</td>
</tr>
<tr>
<td></td>
<td>GCA Ala</td>
<td>GAA Glu</td>
<td>GGA Gly</td>
</tr>
<tr>
<td></td>
<td>GCG Ala</td>
<td>GAG Glu</td>
<td>GGG Gly</td>
</tr>
</tbody>
</table>
Physico-chemical Properties

- hydrophobic
- aliphatic
- aromatic
- small
- tiny
- proline
- polar
- charged
- positive
- negative

Legend:
- Insertions
- Deletions
- Background
Protein Secondary Structure

suppressed
Part 3

Tandem Duplications and Genomic Correlations

[Messer, Arndt, *Nucleic Acid Res* 2006]
Sequence Evolution Model

**Mutation**

A C A

\[ \text{rate } \mu \]

A T A

**Segmental Duplication**

A G T C C A

\[ \text{length } l \]

A G T C C G T C C A

\[ \text{rate } \delta_l \]

**Random Insertion**

A T

\[ \text{rate } \gamma_l^+ \]

A A G A C T

\[ \text{length } l \]

**Segmental Deletion**

G T C C

\[ \text{length } l \]

T A

\[ \text{rate } \gamma_l^- \]
The Correlation Function

Measures likelihood of finding two G/C base pairs separated by a distance $r$ along the genome

$$C(r) = P_{G/C}(x, x+r) - P_{G/C}^2(x)$$

**Position:** $x$ $x+r$
Types of Correlation Behavior

Random sequence

\[ C(r) \approx 0 \]
- Noise fluctuations

Local correlations

\[ C(r) \propto \exp\left(-\frac{r}{r_0}\right) \]
- Characteristic scale
- Generated e.g by Markov-processes

Long-range correlations

\[ C(r) \propto r^{-\alpha} \]
- Scale free, fractal
- Generated by a non-trivial dynamical model
Calculation of $C(r)$ for Our Model

Approach: continuous time Master Equation formalism

$\downarrow$

Exact equation for the dynamics of $C(r)$

$\downarrow$

Stationary solution in a continuum limit:

$$C(r) \propto r^{-\alpha} \quad \text{with} \quad \alpha = \frac{4\mu_{\text{eff}}}{\lambda}$$
Correlations in Genomic DNA

\[ C(r) \propto r^{-\alpha} \]
CorGen measuring and generating long-range correlations for DNA sequence analysis

Your uploaded sequence was **1000000 bp long** and has a **GC content of 0.457**. A power-law has been fitted to the correlation function in the range 10-10000. The decay exponent is **0.369** and the amplitude (at distance 10 bp) is **0.02340**.

GC profile and the correlation function of the submitted sequence:

A sequence with the same correlation parameters has been generated (and can be downloaded [here](#)). Its GC profile and the correlation function are shown below:

You can get an independently sampled sequence [here](#).

It is also possible to retrieve independent samples using non-interactive network clients, e.g. using:

```
wget -q -O - http://corgen.imagen.mpg.de/cgi-bin/corgen.cgi?seqonly=1&len=1000000&gc=0.457&alpha=0.35332&dist=10&c=0.02340
```
Part 4

Tandem Duplications and Alignment Score Statistics

[Messer, Bundschuh, Vingron, Arndt, RECOMB 2006]
[Messer, Bundschuh, Vingron, Arndt, JCB 2007]
Alignment Score Statistics

Sequence alignment
  ↓ Significance

P-values of scores
  ↓ Requires DNA null model

Standard iid model
  ↓ Problem

Correlations in DNA
  ↓ Incorporate into null model

P-values change

Seq1: ACCTAGTGCTA
Seq2: ATCTAGTGATA
Gaussian Approximation

Analytic approach to calculate alignment score statistics for null models with LRC sequences, i.e. $C(r) \propto r^{-\alpha}$

\[ \lambda = \frac{-2\langle s \rangle}{\sigma^2 + c\zeta(2\alpha)} \]

vanishes for iid sequences

LRC’s increases probability of finding high alignment scores by chance
Numerical Verification

Score distribution

Decay parameter $\lambda$

Gaussian approximation captures qualitative behavior
Biological Significance

Alignment of random sequences with same correlation parameters as human chr. 22

Difference in $\lambda$ is approx. 16%, p-value increase $> 10 \times$
Summary

1. The majority of short DNA insertions are tandem duplications

2. Amino acid insertions/deletions are less deleterious than substitutions

3. Tandem duplications cause long-range correlations in genomic base composition

4. These correlations have profound impact on the statistics of sequence alignment scores
Acknowledgments

Peter Arndt
Nicole de la Chaux
Paz Polak
Federico Squartini

Ralf Bundschuh
Michael Lässig
Martin Vingron

Max Planck Institute for Molecular Genetics

International Max Planck Research School for Computational Biology and Scientific Computing