Analysis of large deletions in human-chimp genomic alignments

Erika Kvikstad
BioInformatics I
December 14, 2004
Outline

• Mutations, mutations, mutations…
• Project overview
  – Strategy: finding, classifying indels
  – Analysis
• Future Goals
Why study mutations?

• sources for genetic disease
  – 68% nucleotide substitutions
  – 23% small indels
    from Human Gene Mutation Database

• genetic variation in natural populations

• stuff of molecular evolution
Variability in mutation rate

**Factors**
- GC content
- Recombination

**Scales**
- DNA sequence context
- within chromosomes
- between chromosomes
Why indels in particular?

- Indels responsible for more unmatched nucleotides between closely related populations:
  - Drosophila
  - Arabidopsis
  - Sea urchin
  - Primates
  - E. coli O157:H7

(Britten, PNAS, 2003, Vol 100)
Human vs Chimp

• Use human as reference
  – Most complete sequence from large-scale genome project (IHGSC, Nature, 2004, Vol.431)
    • assembly issues minimal
    • Alignment gaps infer true deletion event

• Compare to chimpanzee sequence
  – Nearest primate relative
    Divergence from human established from sequence
      – substitutions 1.2 to 1.4%
      – indel rate 5% (Britten, PNAS, 2002, Vol99)
Model to Identify Indels

Gaps in alignment of two sequences

Deletion in A?  Insertion in B?

How to determine history of this event?
Gaps in an Alignment of 2 sequences

Alignment

Assembly - missing sequence data

Indels: inferring ancestral state

Insertion vs Deletion

– Methods

• Use phylogenetic history: outgroup
• Use model: retrotransposition
  – 80%> repeat content of indel
    and repeat <5% divergence from consensus to infer insertions
  » Used in previous literature (e.g. Liu, Genome Research, 2003, Vol.13)
Transposition - Mechanisms

Figure 12.1 (a) Nonreplicative transposition. The element is transposed from the donor site to a target site. What will happen to the donor site is unclear. One model proposes that the ends of the donor DNA are not joined to each other and that the DNA molecule is destroyed. Another model proposes that the double strand break is repaired by the host repair system. (b) Replicative transposition. The element is replicated and a copy is inserted at a target site, while the other copy remains at the donor site. For a more detailed explanation of nonreplicative and replicative transposition, see Lewin (1994). (c) Retroposition. The element is transcribed into RNA, which is reverse-transcribed into DNA. The DNA copy is inserted into a host genome. For an example of retroposition, see Figure 12.4. Note that both transposition and retroposition create a short repeat (black box) at each end of the newly inserted element. From Li and Graur (1991).
# Interspersed Repeats

## Table 11: Number of copies and fraction of genome for classes of interspersed repeat

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of copies (x 1,000)</th>
<th>Total number of bases in the draft genome sequence (Mb)</th>
<th>Fraction of the draft genome sequence (%)</th>
<th>Number of families (subfamilies)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINEs</td>
<td>1,558</td>
<td>259.6</td>
<td>13.14</td>
<td>1 (~20)</td>
</tr>
<tr>
<td>Alu</td>
<td>1,090</td>
<td>200.1</td>
<td>10.60</td>
<td>1 (1)</td>
</tr>
<tr>
<td>MIR</td>
<td>393</td>
<td>60.1</td>
<td>2.20</td>
<td>1 (1)</td>
</tr>
<tr>
<td>MIR3</td>
<td>75</td>
<td>9.3</td>
<td>0.34</td>
<td>1 (1)</td>
</tr>
<tr>
<td>LINEs</td>
<td>868</td>
<td>558.8</td>
<td>20.42</td>
<td>1 (~55)</td>
</tr>
<tr>
<td>LINE1</td>
<td>516</td>
<td>462.1</td>
<td>16.99</td>
<td>1 (2)</td>
</tr>
<tr>
<td>LINE2</td>
<td>315</td>
<td>69.2</td>
<td>3.22</td>
<td>1 (2)</td>
</tr>
<tr>
<td>LINE3</td>
<td>37</td>
<td>8.4</td>
<td>0.31</td>
<td>1 (2)</td>
</tr>
<tr>
<td>LTR elements</td>
<td>443</td>
<td>227.0</td>
<td>8.29</td>
<td></td>
</tr>
<tr>
<td>ERV-class I</td>
<td>112</td>
<td>79.2</td>
<td>2.89</td>
<td>72 (132)</td>
</tr>
<tr>
<td>ERV(K)-class II</td>
<td>8</td>
<td>8.5</td>
<td>0.31</td>
<td>10 (20)</td>
</tr>
<tr>
<td>ERV(L)-class III</td>
<td>83</td>
<td>39.5</td>
<td>1.44</td>
<td>21 (42)</td>
</tr>
<tr>
<td>MalR</td>
<td>240</td>
<td>99.8</td>
<td>3.65</td>
<td>1 (31)</td>
</tr>
<tr>
<td>DNA elements</td>
<td>294</td>
<td>77.6</td>
<td>2.84</td>
<td></td>
</tr>
<tr>
<td>hAT group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MER1-Charlie</td>
<td>182</td>
<td>38.1</td>
<td>1.39</td>
<td>26 (50)</td>
</tr>
<tr>
<td>Zaphod</td>
<td>15</td>
<td>4.3</td>
<td>0.16</td>
<td>4 (10)</td>
</tr>
<tr>
<td>To-1 group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MER2-Tigger</td>
<td>57</td>
<td>28.0</td>
<td>1.02</td>
<td>12 (28)</td>
</tr>
<tr>
<td>Tc2</td>
<td>4</td>
<td>0.9</td>
<td>0.03</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Mariner</td>
<td>14</td>
<td>2.6</td>
<td>0.10</td>
<td>4 (5)</td>
</tr>
<tr>
<td>PiggyBac-like</td>
<td>2</td>
<td>0.5</td>
<td>0.02</td>
<td>10 (20)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>22</td>
<td>3.2</td>
<td>0.12</td>
<td>7 (7)</td>
</tr>
<tr>
<td>Total interspersed repeats</td>
<td>1,226.8</td>
<td>44.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 0.60.9.2000, sensitive settings, using RapdBase Update 5.08). In calculating the percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

Human Deletions Strategy

1. **Download sequence alignments**
   - human: hgs17
   - chimp: pantro1
   - from UCSC Genome database Golden Path

2. **Parse coordinates from alignment files**
   - **perl** script

3. **Create MySQL relational database**
   - human-chimp alignment coordinates

4. **Find potential "deletions" Human**
   - coordinates of alignments adjacent in human
   - separated at least 100bp in chimp

5. **Classify Indel event**
   - compare to chimp repeat content
   - RepeatMasker for pan tro 1 available UCSC
   - MySQL relational database
Results

Relative Frequency Human Deletions

Deletion length

Relative frequency (%)

<= 1 kb

Relative Frequency

0-100bp

100-200

200-300

300-400

400-500

500-600

600-700

700-800

800-900

900-1000

100-1000

1000-2000

2000-3000

3000-4000

4000-5000

5000-6000

6000-7000

7000-8000

8000-9000

9000-10000
...cont.

• Castresana (2002) reported significantly higher Ks for hum chr 19 than any other hum chr—based on human/mouse orthologous gene pairs

• Chr 19 unusual
  High GC content
  High gene density
  High expression levels


(Castresana, Nucl Acid Res, 2002, Vol 30)
Discussion

Compare to chimp chr22 paper results:
(analysis of indels 300-5,000bp range)

<table>
<thead>
<tr>
<th></th>
<th>Int’l ChimpChr22 Consort.</th>
<th>Deletion analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTR22 Inserted</td>
<td>25kb</td>
<td>7kb</td>
</tr>
<tr>
<td>HAS 21 deleted</td>
<td>39kb</td>
<td>268kb</td>
</tr>
<tr>
<td>Indels total</td>
<td>567</td>
<td>270</td>
</tr>
</tbody>
</table>

Future

• **Definition** of deletion
  – Allow for “wobble” alignment ends
    • +1, +5, +20 bp threshold for adjacent human sequence

• **Size** distributions
  – Different T thresholds
    • e.g. 10, 50, 100 kb etc. threshold for lower limit indel

• **Filter** data
  – Errors due to low coverage chimp sequence
    • e.g. RepeatMasker chimp
    • e.g. Bias certain chromosomes sequenced to better completion than others
      assembly - repetitive regions require finishing strategies
  – Compare to other alignment sources
Indel strategy

Whole Genome sequence

Finished BAC alignments  Genomic alignments  Individual read alignments

Identification of gaps

Human-specific events  Chimp-specific events

Confirmation
(comparison to outgroup)

Calculate Indel rates

Comparison:
X/Y
X/A
Y/A
Future

• **Definition** of deletion
  – Allow for “wobble” alignment ends
    • +1, +5, +20 bp threshold for adjacent human sequence

• **Size** distributions
  – Different T thresholds
    • e.g. 10, 50, 100 kb etc. threshold for lower limit indel

• **Filter** data
  – Errors due to low coverage chimp sequence
    • e.g. RepeatMasker chimp
    • e.g. Bias certain chromosomes sequenced to better completion than others
      assembly - repetitive regions require finishing strategies
  – Compare to other alignment sources

• Underlying **mechanisms**
  – Replication-driven vs Recombination-driven
    • Different sizes of indels
    • Chromosomal bias

• Other **mammals**
  – Complete genome sequence available for mouse and rat
Acknowledgements

People:
Kateryna Makova
Webb Miller

Alignments and tools:
UCSC Genome Browser (http://genome.ucsc.edu)
Chimp Genome Sequencing Consortium
International Human Genome Sequencing Consortium
Sequence Alignments

The scoring matrix used for blastz was:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-300</td>
<td>-150</td>
<td>-300</td>
<td></td>
</tr>
<tr>
<td>-300</td>
<td>100</td>
<td>-300</td>
<td>-150</td>
<td></td>
</tr>
<tr>
<td>-150</td>
<td>-300</td>
<td>100</td>
<td>-300</td>
<td></td>
</tr>
<tr>
<td>-300</td>
<td>-150</td>
<td>-300</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

with a gap open penalty of 400 and a gap extension penalty of 30.

The alignments were done with blastz, which is available from Webb Miller's group at PSU. Each chromosome was divided into 10000000 base chunks with 10000 bases of overlap.

The axtNet alignments were processed with chainNet, netSyntenic, and netClass from Jim Kent at UCSC.

(http://genome.ucsc.edu)
Sequence Assemblies:
human/chimp alignments made using the May 2004 human assembly (hg17) vs. the Nov 2003 chimp assembly (panTro1) produced by the Chimp Genome Sequencing Consortium.

Repeats
RepeatMasker of Nov 2003 chimp assembly performed at the -s sensitive setting.

(http://genome.ucsc.edu)
### Summary Statistics

<table>
<thead>
<tr>
<th>hum_chrom</th>
<th>nt deleted</th>
<th># deletions</th>
<th># chimp insertions</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr2</td>
<td>2,282,021</td>
<td>3,237</td>
<td>160</td>
</tr>
<tr>
<td>chr1</td>
<td>2,431,061</td>
<td>3,073</td>
<td>144</td>
</tr>
<tr>
<td>chr4</td>
<td>1,973,522</td>
<td>2,592</td>
<td>108</td>
</tr>
<tr>
<td>chr3</td>
<td>1,845,042</td>
<td>2,559</td>
<td>148</td>
</tr>
<tr>
<td>chr5</td>
<td>1,600,512</td>
<td>2,319</td>
<td>69</td>
</tr>
<tr>
<td>chr7</td>
<td>1,706,268</td>
<td>2,265</td>
<td>88</td>
</tr>
<tr>
<td>chr6</td>
<td>1,567,469</td>
<td>2,242</td>
<td>132</td>
</tr>
<tr>
<td>chr8</td>
<td>1,556,725</td>
<td>2,081</td>
<td>98</td>
</tr>
<tr>
<td>chr10</td>
<td>1,563,384</td>
<td>2,000</td>
<td>82</td>
</tr>
<tr>
<td>chr11</td>
<td>1,390,982</td>
<td>1,735</td>
<td>82</td>
</tr>
<tr>
<td>chr9</td>
<td>1,056,252</td>
<td>1,528</td>
<td>69</td>
</tr>
<tr>
<td>chrX</td>
<td>1,166,142</td>
<td>1,498</td>
<td>29</td>
</tr>
<tr>
<td>chr12</td>
<td>1,001,185</td>
<td>1,433</td>
<td>64</td>
</tr>
<tr>
<td>chr16</td>
<td>946,623</td>
<td>1,300</td>
<td>37</td>
</tr>
<tr>
<td>chr13</td>
<td>1,014,162</td>
<td>1,238</td>
<td>58</td>
</tr>
<tr>
<td>chr19</td>
<td>992,812</td>
<td>1,212</td>
<td>22</td>
</tr>
<tr>
<td>chr14</td>
<td>815,757</td>
<td>1,120</td>
<td>63</td>
</tr>
<tr>
<td>chr15</td>
<td>918,207</td>
<td>1,111</td>
<td>31</td>
</tr>
<tr>
<td>chr17</td>
<td>895,303</td>
<td>1,091</td>
<td>17</td>
</tr>
<tr>
<td>chr18</td>
<td>663,871</td>
<td>1,019</td>
<td>53</td>
</tr>
<tr>
<td>chr20</td>
<td>625,877</td>
<td>919</td>
<td>31</td>
</tr>
<tr>
<td>chr22</td>
<td>407,996</td>
<td>612</td>
<td>17</td>
</tr>
<tr>
<td>chr21</td>
<td>321,388</td>
<td>572</td>
<td>17</td>
</tr>
<tr>
<td>chrY</td>
<td>326,637</td>
<td>210</td>
<td>6</td>
</tr>
</tbody>
</table>

**Total <= 10 kb**

<table>
<thead>
<tr>
<th>n</th>
<th>sum(bp)</th>
<th>min</th>
<th>max</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>38,966</td>
<td>29,069,198</td>
<td>98</td>
<td>9,980</td>
<td>746.0144</td>
</tr>
</tbody>
</table>

**<= 300b**

n = 19,928 => 51%