Gene duplication and loss
Part II

Matthew Hahn
Indiana University

mwh@indiana.edu
When genomes go bad
“At least 113 genes entered the vertebrate (or pre-vertebrate) lineage by horizontal transfer from bacteria”

International Human Genome Sequencing Consortium (2001)
Link Between Human Genes and Bacteria Is Hotly Debated

By NICHOLAS WADE
Published: May 15, 2001
More genes underwent positive selection in chimpanzee evolution than in human evolution

Margaret A. Bakewell, Peng Shi, and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

Communicated by Morris Goodman, Wayne State University School of Medicine, Detroit, MI. February 26, 2007 (received for review December 21, 2006)

Observations of numerous dramatic and presumably adaptive phenotypic modifications during human evolution prompt the common belief that more genes have undergone positive Darwinian selection in the human lineage than in the chimpanzee lineage. We describe two recent studies, of human-specific and chimpanzee-specific nucleotide substitutions, that challenge this belief. Both studies use shared chimpanzee-mouse sequences, which can be used as an outgroup, to distinguish between human-specific and chimpanzee-specific nucleotide substitutions, because of the unavailability of genome sequences from any closer outgroups at that time. Because...
When genomes go bad

ACGTCATCATACTACG
human

ACGTTATCGTACTAAG
chimpanzee
When genomes go bad

ACGTCATCATACTACG

human

ACGTTGTCGTACTAAG

chimpanzee
When genomes go bad

"I know it’s wrong because we’re the ones reading the DNA"

-Paula Poundstone
How bad assemblies affect gene gain and loss
Genome assemblies are imperfect
Genome assemblies are imperfect

- genomes come in pieces
- there are gaps between pieces
- the order of pieces is not known
How bad assemblies add genes

alleles can be split, increasing number of genes
How bad assemblies add genes

genes can be fragmented by gaps, increasing number of genes
How bad assemblies add genes

genes can be over-predicted by software, increasing number of genes

(This is not due to error or incompleteness of assembly)
How bad assemblies remove genes

highly similar duplicates can be collapsed, decreasing number of genes
How bad assemblies remove genes

genes can be missing, decreasing number of genes
How bad assemblies affect gene gain and loss

Denton et al. (2014)
How bad assemblies affect gene gain and loss

v1.0: 4X coverage → v2.0: 6X coverage
How bad assemblies affect gene gain and loss

High-quality genome

-1 change

0 change

+1 change

Low-quality genome
Low-quality chimp assembly leads to errors

More genes in the lower-quality assembly:
Variation in error due to technology/coverage
Comparison among chicken genomes

High-quality reference genome

- 2X Sanger
- 12X 454
- 82X Illumina
Comparison among chicken genomes

2X Sanger vs. reference
Comparison among chicken genomes

12X 454 vs. reference

![Bar chart showing comparison between 12X 454 and reference.](chart.png)

Legend:
- **12X-Reference**
Comparison among chicken genomes

82X Illumina vs. reference

-3  -2  -1  0  1  2  3

16,000 12,000 8,000 4,000

82X-Reference
Variation in error due to technology/coverage

- 2X Sanger: very bad, vastly undercounts genes
- 12X 454: pretty bad, slightly overcounts
- 82X Illumina: bad, but equally over- and undercounts

The best of these (Illumina) still has ~40% of families with errors

(and don’t think your transcriptome assembly is any better!)
Phylogenetic inference of gene gain and loss
Phylogenetic inference of gene gain and loss

- Ks-based methods
- Species overlap methods
- Gene tree-Species tree reconciliation
- Count methods (e.g. CAFE)
Phylogenetic inference of gene gain and loss

Ks-based methods

Phylogenetic inference of gene gain and loss

Species overlap methods

The human phylome

Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo and Toni Gabaldón

Received: 30 November 2006  Accepted: 13 June 2007  Published: 13 June 2007
Phylogenetic inference of gene gain and loss

Gene tree-species tree reconciliation

"Reconciled" gene tree
Gene tree reconciliation

Want to:

- Count duplications and losses
- Identify when they occurred
- (Can be used for species tree inference)
Gene tree reconciliation

gene tree + species tree = reconciled gene tree
Least common ancestor (LCA) algorithm

1. Label internal nodes

Gene tree

Species tree

Goodman et al. (1979)
Least common ancestor (LCA) algorithm

2. Initialize map of gene tree tip nodes to species tree tip nodes
Least common ancestor (LCA) algorithm

3. Map gene tree internal nodes to species tree nodes:
   this is done to least common ancestor that includes the same lineages

best algorithm: Zmasek and Eddy (2001)
Least common ancestor (LCA) algorithm

Summary of map:

<table>
<thead>
<tr>
<th>gene tree</th>
<th>species tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>G3</td>
<td>S2</td>
</tr>
<tr>
<td>G2</td>
<td>S1</td>
</tr>
<tr>
<td>G1</td>
<td>A</td>
</tr>
<tr>
<td>A1</td>
<td>A</td>
</tr>
<tr>
<td>A2</td>
<td>A</td>
</tr>
<tr>
<td>B1</td>
<td>B</td>
</tr>
<tr>
<td>C1</td>
<td>C</td>
</tr>
</tbody>
</table>

If the map of a parent node is the same as a child, it is labeled as a duplication.
Least common ancestor (LCA) algorithm

4. Label nodes such that parent nodes sharing a map with at least one of their children are duplication nodes.
Least common ancestor (LCA) algorithm

Once duplication nodes have been identified, all others are speciation nodes.

gene tree + species tree = reconciled gene tree
Least common ancestor (LCA) algorithm

What about gene losses?
Least common ancestor (LCA) algorithm

What about gene losses?

![Gene tree and species tree diagram](image)
### Least common ancestor (LCA) algorithm

What about gene losses?

<table>
<thead>
<tr>
<th>gene tree</th>
<th>species tree</th>
<th>depth of species tree node</th>
</tr>
</thead>
<tbody>
<tr>
<td>G2</td>
<td>S3</td>
<td>1</td>
</tr>
<tr>
<td>G1</td>
<td>S2</td>
<td>2</td>
</tr>
<tr>
<td>A1</td>
<td>A</td>
<td>4</td>
</tr>
<tr>
<td>C1</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>D1</td>
<td>D</td>
<td>2</td>
</tr>
</tbody>
</table>
Least common ancestor (LCA) algorithm

Counting the depth of a node

.species tree
Least common ancestor (LCA) algorithm

Counting losses

\[ L(b_X) = [(\text{depth of daughter}) - (\text{depth of parent}) - 1] + \text{IsDup}(0,1) \]

is the parent node a duplicate? 
no=0
yes=1

depth of node it maps to in species tree
Least common ancestor (LCA) algorithm

\[ L(b_1) = (4 - 2 - 1) + 0 = 1 \]
\[ L(b_2) = (3 - 2 - 1) + 0 = 0 \]
\[ L(b_3) = (2 - 1 - 1) + 0 = 0 \]
\[ L(b_4) = (2 - 1 - 1) + 0 = 0 \]
Problems with reconciliation

- gene tree error
- biological discordance
- gene conversion
- polyploidy
Error in gene trees

If your gene tree is inferred incorrectly, reconciliation can result in extra duplications and losses.

species tree  +  gene tree  =  reconciled gene tree
If your gene tree is discordant (e.g. due to ILS), reconciliation can result in extra duplications and losses.
If there is gene conversion, reconciliation can result in extra duplications and losses.

Gene tree (before conversion)

Gene tree (after conversion)

Reconciled gene tree
If there is allopolyploidy, reconciliation can result in extra duplications and losses.
Solutions!

- gene tree error
- biological discordance
- polyploidy
- gene conversion
Error in gene trees

Use bootstrap cut-offs to rearrange nodes with low support

A1  C1  B1

A1  B1  C1

gene tree  corrected gene tree

implemented in Notung (Chen et al. 2000)
Error in gene trees

a) Mammals

b) Drosophila

Bootstrap cut-off

- gains
- losses

Hahn (2007)
Biological discordance due to ILS

Reconcile to a non-binary species tree

Vernot et al. (2008)
Discordance due to ILS or error

The human phylome

Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo and Toni Gabaldón


Received: 30 November 2006  Accepted: 13 June 2007  Published: 13 June 2007
Allopolyploidy

Reconcile to a multiply-labeled (MUL-) tree

Thomas et al. (biorxiv)
Gene conversion

What to do about gene conversion?

“Count” methods!

gene tree (before conversion)  gene tree (after conversion)