Gene duplication and loss

Matthew Hahn
Indiana University
mwh@indiana.edu
How many genes does a human have:

a. <1
b. 1-10,000
c. 10,000-100,000
d. >100,000
How many genes does a human have:

a. <1
b. 1-10,000
c. 10,000-100,000 (~23,000)
d. >100,000
How many genes does a fruitfly have:

a. <1
b. 1-10,000
c. 10,000-100,000
d. >100,000
How many genes does a fruitfly have:

a. <1
b. 1-10,000
c. 10,000-100,000 (~15,000)
d. >100,000
Gene Number Variation

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>40,000</td>
</tr>
<tr>
<td>Rice</td>
<td>33,000</td>
</tr>
<tr>
<td>Thale cress</td>
<td>28,000</td>
</tr>
<tr>
<td>Mouse</td>
<td>23,000</td>
</tr>
<tr>
<td>Human</td>
<td>23,000</td>
</tr>
<tr>
<td>Worm</td>
<td>19,000</td>
</tr>
<tr>
<td>Fruitfly</td>
<td>15,000</td>
</tr>
<tr>
<td>Yeast</td>
<td>6,000</td>
</tr>
<tr>
<td><em>M. genitalium</em></td>
<td>500</td>
</tr>
</tbody>
</table>
Gene duplication and loss result in genome size variation

<table>
<thead>
<tr>
<th>Domain</th>
<th>S. cerevisiae</th>
<th>C. elegans</th>
<th>D. melanogaster</th>
<th>H. sapiens</th>
<th>A. thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homeodomain</td>
<td>9</td>
<td>109</td>
<td>148</td>
<td>267</td>
<td>118</td>
</tr>
<tr>
<td>Zinc-finger</td>
<td>121</td>
<td>437</td>
<td>357</td>
<td>706</td>
<td>1049</td>
</tr>
<tr>
<td>Nuclear receptor</td>
<td>1</td>
<td>183</td>
<td>25</td>
<td>59</td>
<td>4</td>
</tr>
</tbody>
</table>

from Venter et al. (2001)
Similar genomes have similar numbers of genes

Insects: ~15,000
Mammals: ~23,000
Worms: ~19,000
Fungi: 6,000-10,000

Despite this, lots of variation in individual genes.
Gene duplication and loss

Table 1. Prevalence of gene duplication in all three domains of life

<table>
<thead>
<tr>
<th>Domain</th>
<th>Total number of genes</th>
<th>Number of duplicate genes (%) of duplicate genes</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Mycoplasma pneumoniae</em></td>
<td>677</td>
<td>298 (44)</td>
<td>[65]</td>
</tr>
<tr>
<td><em>Helicobacter pylori</em></td>
<td>1590</td>
<td>266 (17)</td>
<td>[66]</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>1709</td>
<td>284 (17)</td>
<td>[67]</td>
</tr>
<tr>
<td>Archaea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Archaeoglobus fulgidus</em></td>
<td>2436</td>
<td>719 (30)</td>
<td>[68]</td>
</tr>
<tr>
<td>Eukarya</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6241</td>
<td>1858 (30)</td>
<td>[67]</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>18 424</td>
<td>8971 (49)</td>
<td>[67]</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>13 601</td>
<td>5536 (41)</td>
<td>[67]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>25 498</td>
<td>16 574 (65)</td>
<td>[69]</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>40 580(^b)</td>
<td>15 343 (38)</td>
<td>[11]</td>
</tr>
</tbody>
</table>

\(^a\)Use of different computational methods or criteria results in slightly different estimates of the number of duplicated genes [12].

\(^b\)The most recent estimate is \(\sim 30\ 000\) [61].

Zhang (2003)
Gene duplication and loss

- Homo sapiens (n = 336)
- Mus musculus (n = 225)
- Drosophila melanogaster (n = 462)
- Arabidopsis thaliana (n = 2671)
- Caenorhabditis elegans (n = 1933)
- Saccharomyces cerevisiae (n = 326)

Gene duplication and loss

The most common outcome of duplication is loss

Pseudogene
Gene duplication and loss

The most common outcome of duplication is loss
Big questions in gene duplication

“The main interest in duplications lay in ... identical genes which could subsequently mutate separately and diversify their effects.”

--Bridges 1918
THE BAR "GENE" A DUPLICATION

The nature of the Bar gene has been the subject of extensive investigation and speculation since February, 1913, when Tice found this red-eyed mutant as a single male in the progeny of normal-eyed parents. The eye-reduction behaves as a sex-linked dominant, with a locus at 67.0, and has been one of the most important of all the sex-linked characters of D. melanogaster. A remarkable peculiarity of the mutant is that occasionally the homozygous stock gives rise to a fly indistinguishable in appearance and genetic behavior from wild-type. More rarely the stock gives rise to an even more extreme reduction in eye-size, a type which was called Ultra-Bar by Zeleny, who found it.

Sturtevant and Morgan and Sturtevant found that these two-way changes were the result of a novel type of "unequal" crossing-over, by which the two genes originally present in the two parental chromosomes both emerged in the same chromosome (Bar-double) while the other resultant chromosome was without Bar gene. (See Figs. A, B, C, D, E, and F.)

The exact point of the insertion is ambiguous, for a reason which will appear below. The normal X in this region (see revised map in Fig. 1) shows in sub-sec-

Bridges (1935)
Big questions in gene duplication

Figure 1b

- Full function
- Dead function
- New function

Neofunctionalization
Subfunctionalization
Gene conservation

Hahn (2009)
Big questions in gene duplication

Figure 1a

- Full function
- Dead function
- New function

Neofunctionalization

Subfunctionalization

Gene conservation

Hahn (2009)
Gene conservation

Figure 3

Figure 3 estimates. Red (rich roots and tubers) differs among populations with high- and low-starch diets, we from fruit, honey and milk). To determine if (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, European Americans (for example, Biaka, Mbuti) and populations with traditional diets that include little or no starch (low-starch) (for example, Datog, Yakut). We found that mean diploid copy number was greater in high-starch populations (45), and Hadza hunter-gatherers who rely extensively on starch-rich diets instead emphasize proteinaceous resources (for example, meats and blood) and simple saccharides (for example, starchy foods. Such diets instead emphasize proteinaceous resources populations with traditional diets that incorporate relatively few substantial portion of the diet and the small fraction of 'low-starch' intake among human populations, a distinction can be made between agricultural populations, Eu...
Gene sharing by δ-crystallin and argininosuccinate lyase

(Jens proteins/evolution/gene expression/enzymes/urea cycle)


*Laboratory of Molecular and Developmental Biology, National Eye Institute, National Institutes of Health, Bethesda, MD 20892; and †Baylor College of Medicine, and Institute for Molecular Genetics, the Howard Hughes Medical Institute, Houston, TX 77030

Communicated by Donald D. Brown, February 1, 1988

ABSTRACT The lens structural protein δ-crystallin and the metabolic enzyme argininosuccinate lyase (ASL; L-argininosuccinate arginine-lyase, EC 4.3.2.1) have striking genome similarity. We have demonstrated that duck δ-crystallin share the enzyme activity, and although uricotelic, birds have some activity for ASL as well as for other enzymes of the urea cycle. δ-Crystallin is the dominant crystallin in lenses of birds and reptiles, but it is absent from lenses of mammals (E). Southern blot hybridization

Piatigorsky and Wistow (1991)
Big questions in gene duplication

Neofunctionalization, subfunctionalization, or conservation?

Positive selection or genetic drift?
Big questions in gene loss

McCutcheon and Moran (2012)
Loss of *Myh16* associated with cranial enlargement

Stedman et al. (2004)
Molecular mechanisms of gene duplication

DNA- or RNA-based

Multiple genes, single genes, partial genes
Molecular mechanisms of gene duplication

“Unequal crossing-over”
Molecular mechanisms of gene duplication

“Unequal crossing-over”:

- Requires repeated elements to be present
- Is generally due to NAHR
- Does not always result in tandem duplicates
Molecular mechanisms of gene duplication

Hypothesis: the more TEs there are in a genome, the farther apart the duplicates are

McGrath et al. (2009)
Molecular mechanisms of gene duplication

- Head-to-tail
- Head-to-head
- Tail-to-tail

McGrath et al. (2009)
Molecular mechanisms of gene duplication

DNA-based mechanisms

RNA-based mechanisms

“Retrotransposition”
Molecular mechanisms of gene duplication

Retrotransposition:

- Results in a daughter copy without introns
- Brings along (almost) no flanking sequence
- Can only copy one gene at a time
Molecular mechanisms of gene duplication

Weird hybrid case: “chimeric” gene duplicates

Long and Langley (1993)
Molecular mechanisms of gene duplication

Polyploidy

Unreduced gametes
Molecular mechanisms of gene duplication

Polyplody:

- Doubles the entire content of the genome
- Most genes subsequently return to single-copy
- Two types (at approximately equal frequency in nature): Autopolyploidy and allopolyploidy
Molecular mechanisms of gene duplication

Autopolyploidy:

Doubling the number of chromosomes, where both parents are from the same species (or same individual)
Molecular mechanisms of gene duplication

Allopolyploidy:

Doubling the number of chromosomes, where the parents are from *different* species
Molecular mechanisms of gene loss

- Nonsense mutation
- Frame shifting indel
- Complete deletion (often due to NAHR)
Genealogical relationships among genes

**Paralogs**: genes (loci) whose most recent common ancestor is a duplication node.
Genealogical relationships among genes

A1

A2

B

Paralogs
Genealogical relationships among genes

**Paralogs:** genes (loci) whose most recent common ancestor is a duplication node

**Orthologs:** genes (loci) whose most recent common ancestor is a speciation node
Genealogical relationships among genes

Orthologs (co-orthologs)
Genealogical relationships among genes

A1

B2

B1

Orthologs

Paralogs
Genealogical relationships among genes

Paralogs
(out-paralogs wrt the A-B speciation event)
Genealogical relationships among genes

In-paralogs wrt the A-B speciation event
More genealogical relationships among genes

Co-orthologs

A1

A2

B
More genealogical relationships among genes

“Positional orthologs” (Dewey 2011)
More genealogical relationships among genes

“Positional orthologs”
(Dewey 2011)
More genealogical relationships among genes

Retrotransposition

Primary/parent ortholog  Secondary/daughter ortholog
Polarized duplicates and adaptation

Han et al. (2009)
More genealogical relationships among genes

A1

A2

B

C

Xenologs (orthologs)

HGT
More genealogical relationships among genes

Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes

Todd J. Treangen, Eduardo P. C. Rocha

Published: January 27, 2011  •  http://dx.doi.org/10.1371/journal.pgen.1001284
More genealogical relationships among genes

Autopolyploidy

Ohnologs (paralogs)
More genealogical relationships among genes

Allopolyploidy

Homoeologs (orthologs)
More genealogical relationships among genes

Allopolyploidy
More genealogical relationships among genes

Glover et al. (2016)
More genealogical relationships among genes

- A1
- B1
- B2
More genealogical relationships among genes

“Reconciled” gene tree
More genealogical relationships among genes

A1  B1  A2

B2

pseudo-orthologs
(“hidden paralogy”)
Divergence between paralogs
Divergence between paralogs
Divergence between paralogs

How divergent are new paralogs?

\[ 4N_e \mu \]

Figure 9-9 Essential Cell Biology 3/e (© Garland Science 2010)