Gene duplication and loss
Part II

Matthew Hahn
Indiana University

mwh@indiana.edu

When genomes go bad

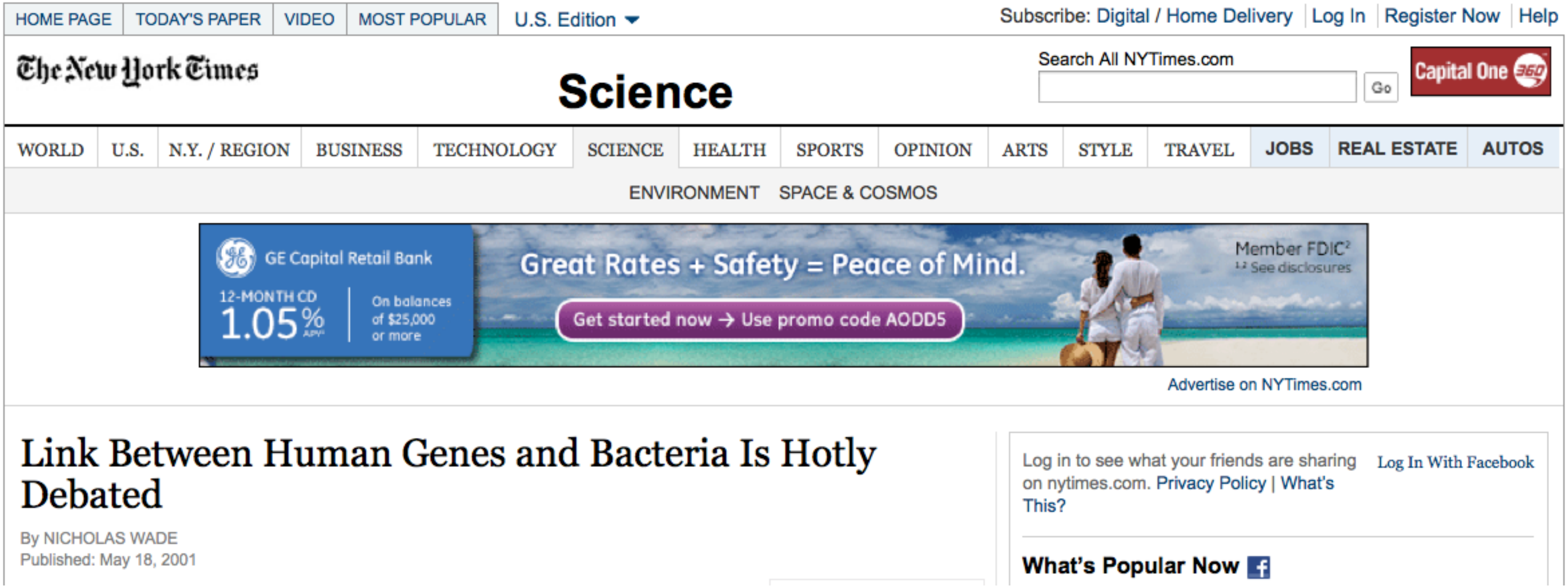
When genomes go bad

nature
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities

“At least 113 genes entered the vertebrate (or pre-vertebrate) lineage by horizontal transfer from bacteria”

International Human Genome Sequencing Consortium (2001)

When genomes go bad



Science

Link Between Human Genes and Bacteria Is Hotly Debated

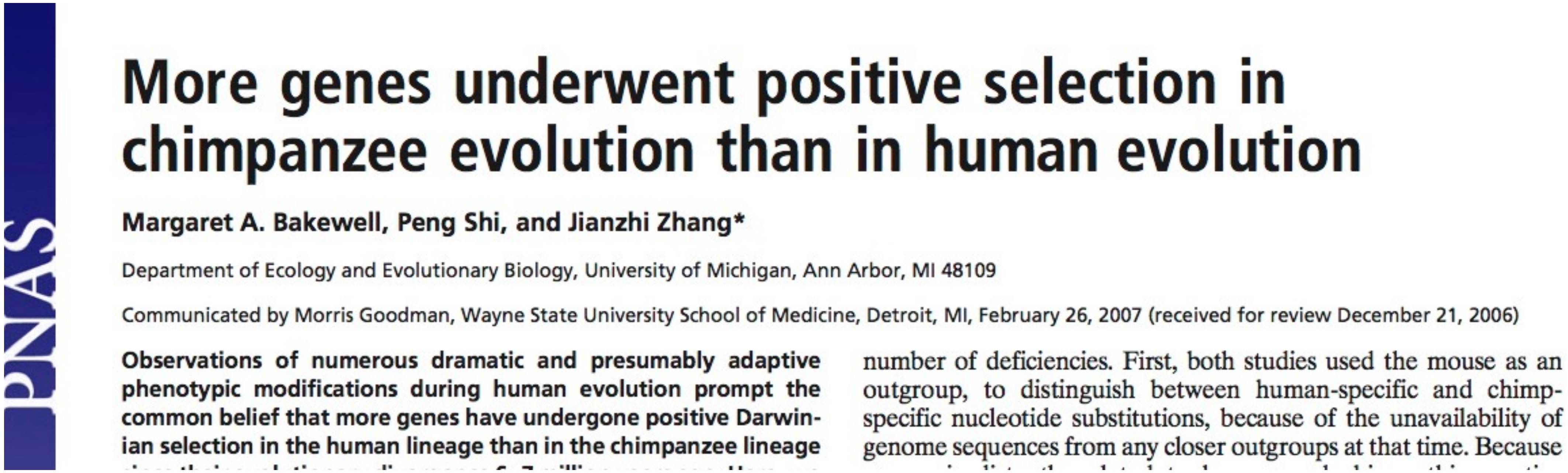
By NICHOLAS WADE
Published: May 18, 2001

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?

Log In With Facebook

What's Popular Now

When genomes go bad



More genes underwent positive selection in chimpanzee evolution than in human evolution

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

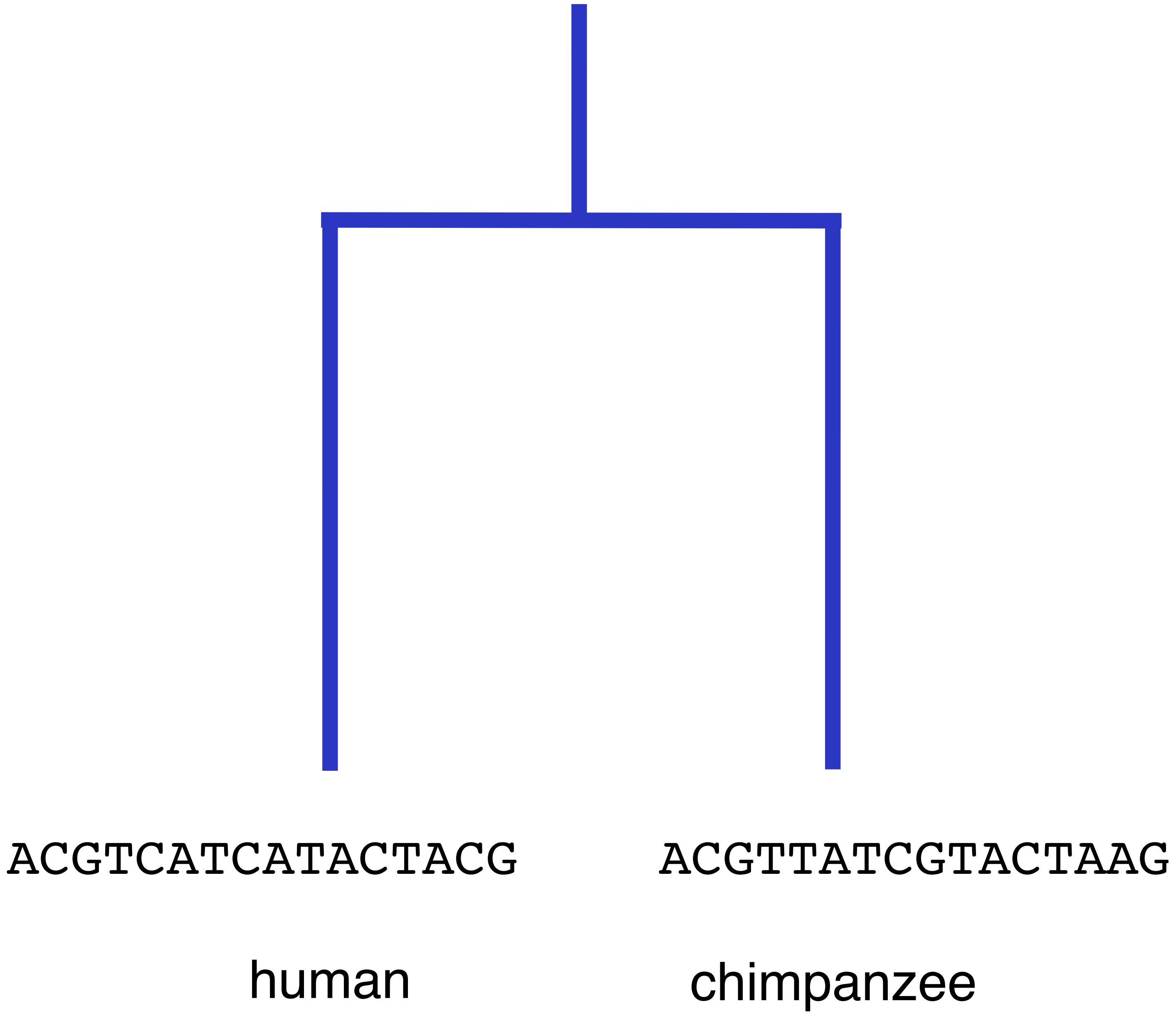
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

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

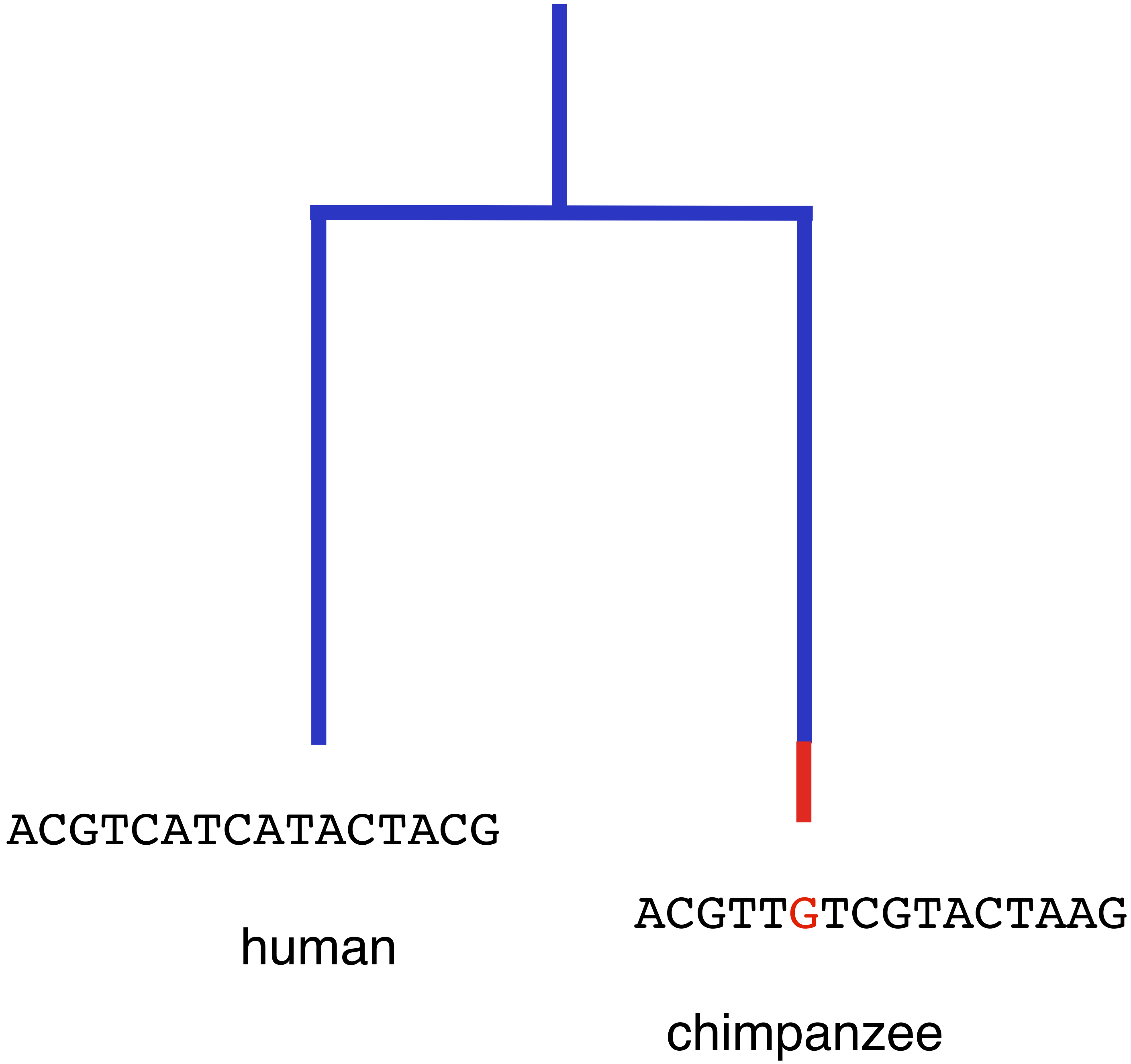
Observations of numerous dramatic and presumably adaptive phenotypic modifications during human evolution prompt the common belief that more genes have undergone positive Darwinian selection in the human lineage than in the chimpanzee lineage

number of deficiencies. First, both studies used the mouse as an outgroup, to distinguish between human-specific and chimp-specific nucleotide substitutions, because of the unavailability of genome sequences from any closer outgroups at that time. Because

When genomes go bad



When genomes go bad



ACGTCATCATACTACG

human

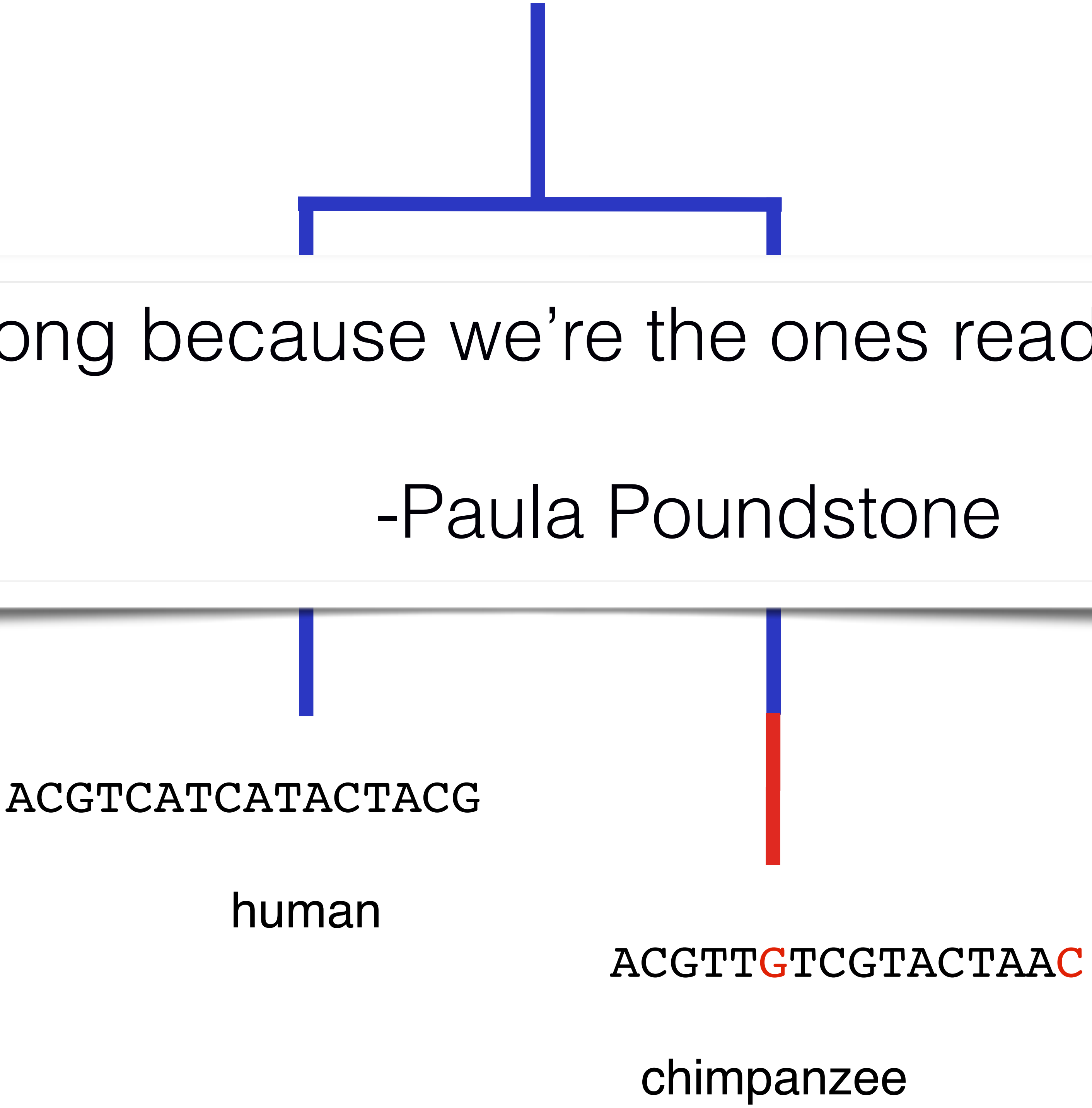
ACGTTGTCGTACTAAG

chimpanzee

When genomes go bad

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

-Paula Poundstone



ACGTCATCATACTACG

human

ACGTTGTCGTACTAAC

chimpanzee

How bad assemblies affect gene gain and loss

Genome assemblies are imperfect

nature

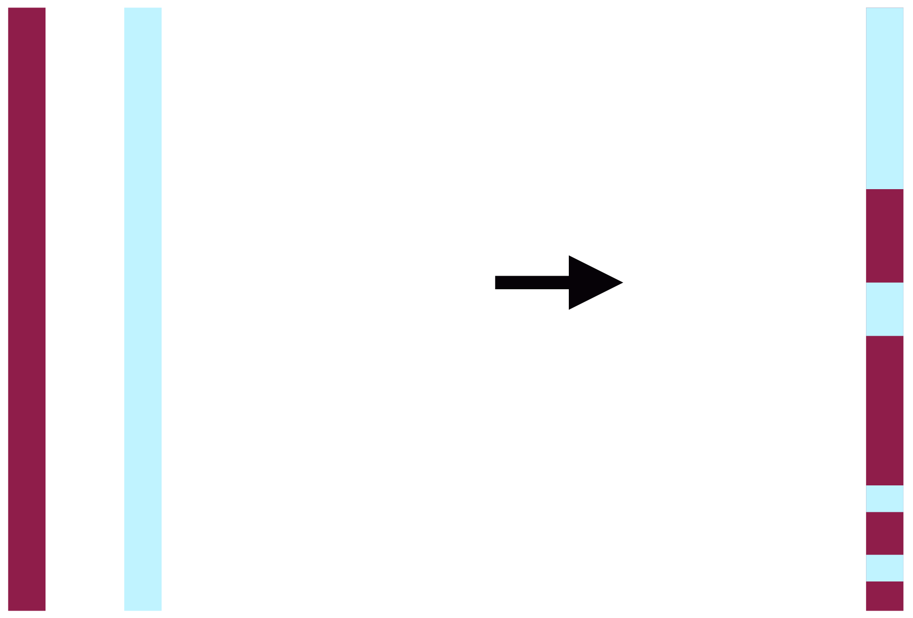
STAR FORMATION
A massive protostar unveiled

CANCER IMMUNOLOGY
How tumours dupe T cells

AIR POLLUTION
China's NO_2 build-up seen from space

NATUREJOBS
Membrane proteomics

THE CHIMPANZEE GENOME



Genome assemblies are imperfect

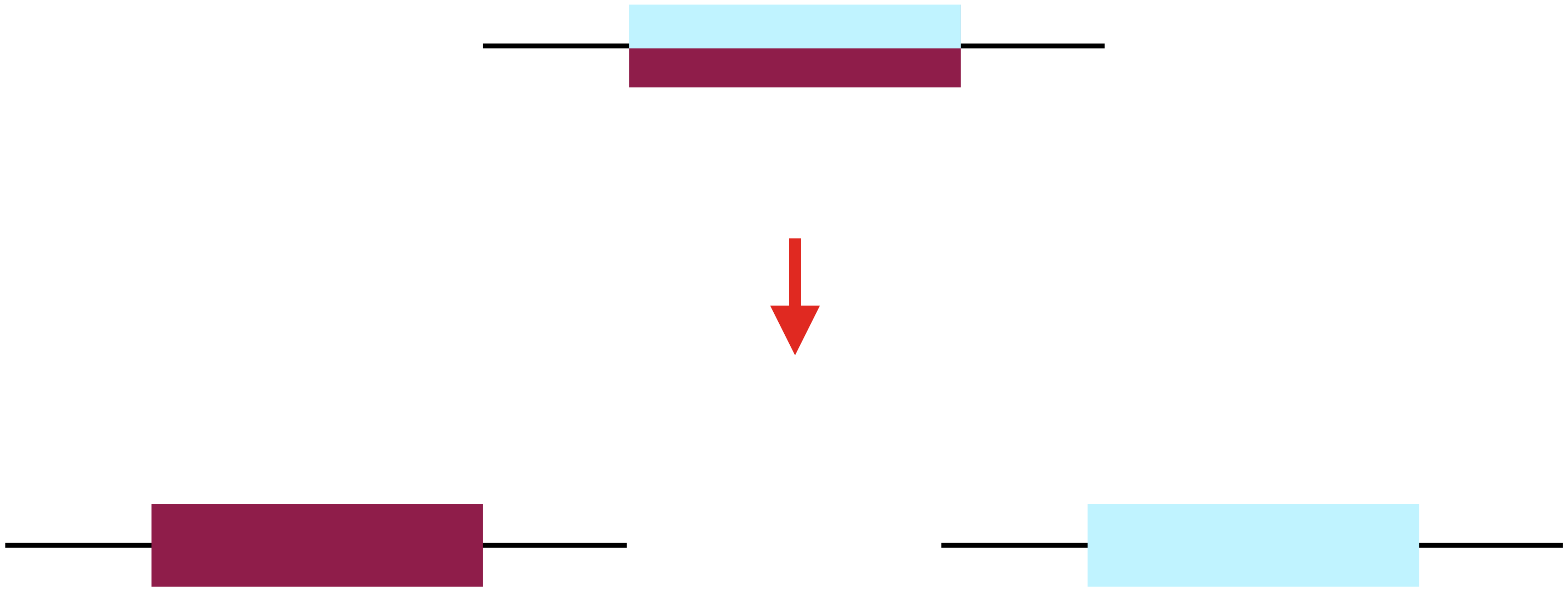
-genomes come in pieces

-there are gaps between pieces

-the order of pieces is not known

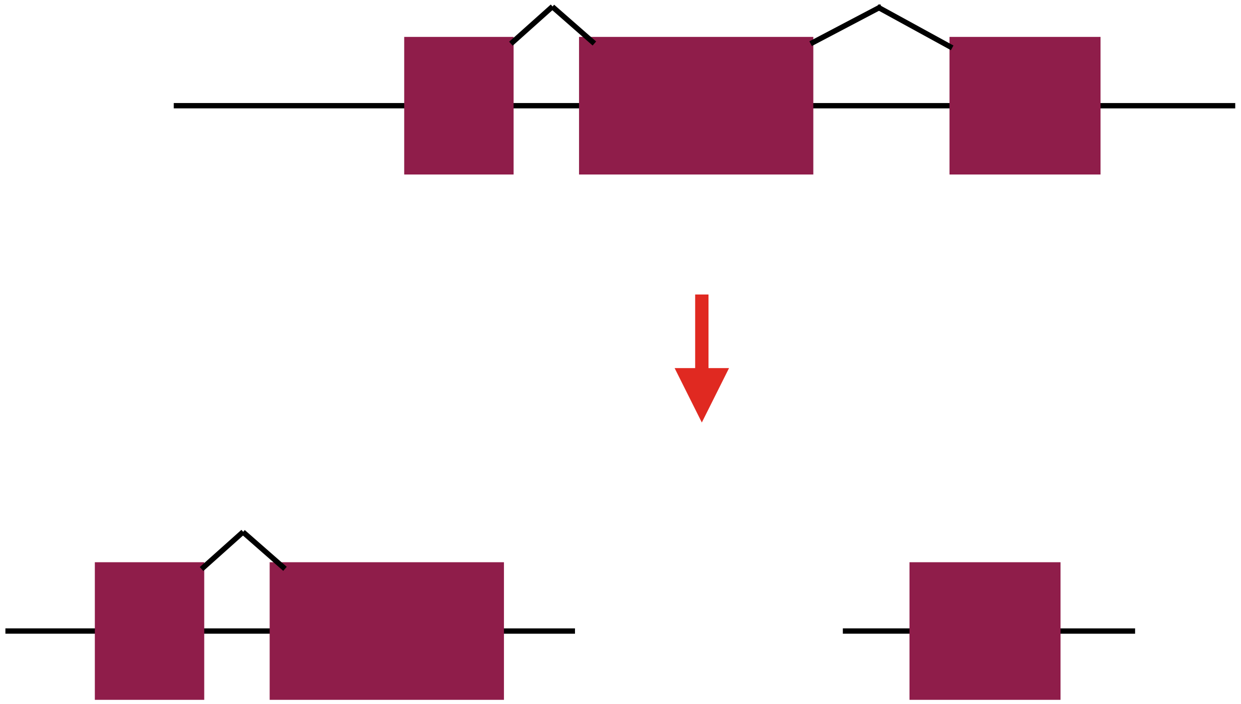
How bad assemblies add genes

alleles can be split, increasing number of genes



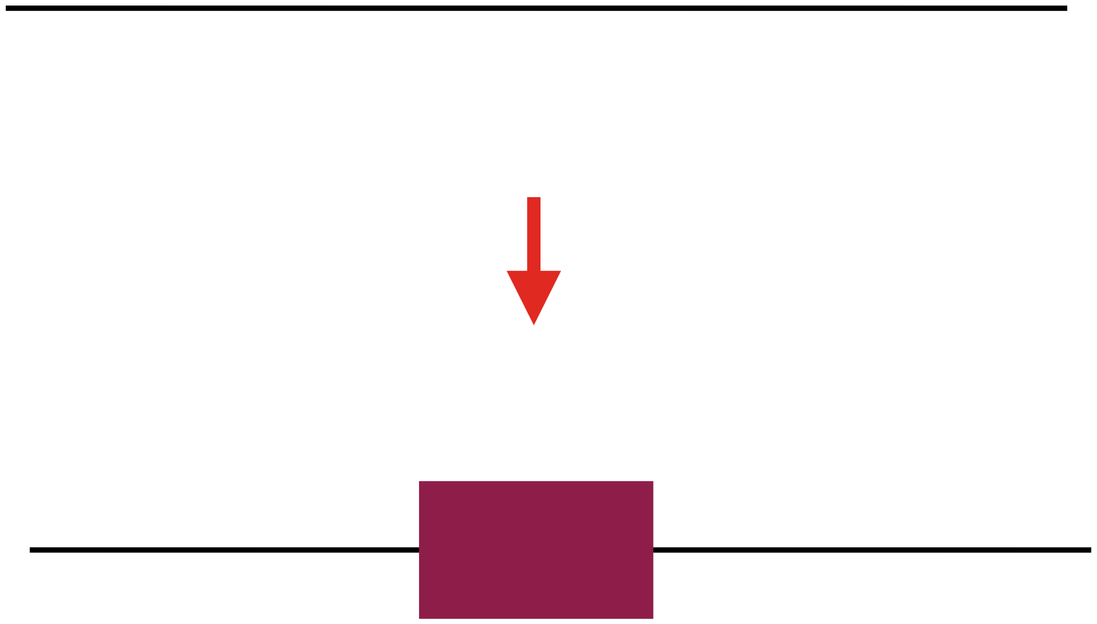
How bad assemblies add genes

genes can be fragmented by gaps, increasing number of genes



How bad assemblies add genes

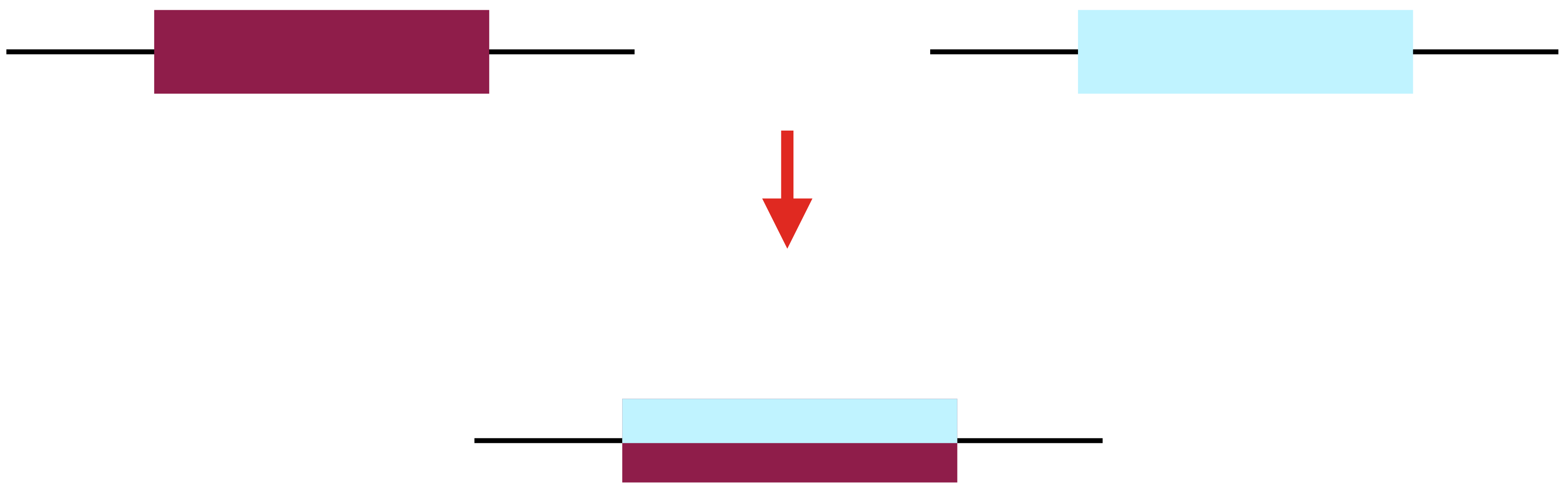
genes can be over-predicted by software, increasing number of genes



(This is not due to error or incompleteness of assembly)

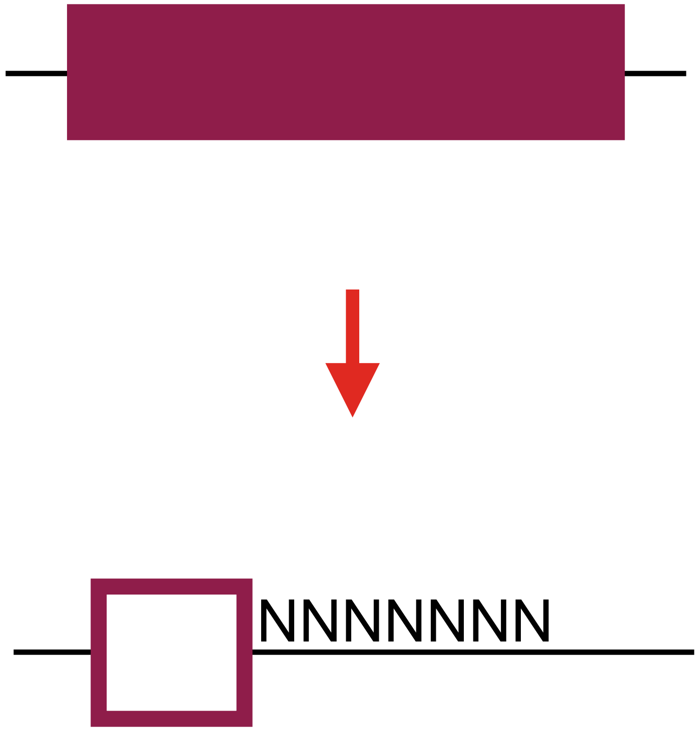
How bad assemblies remove genes

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

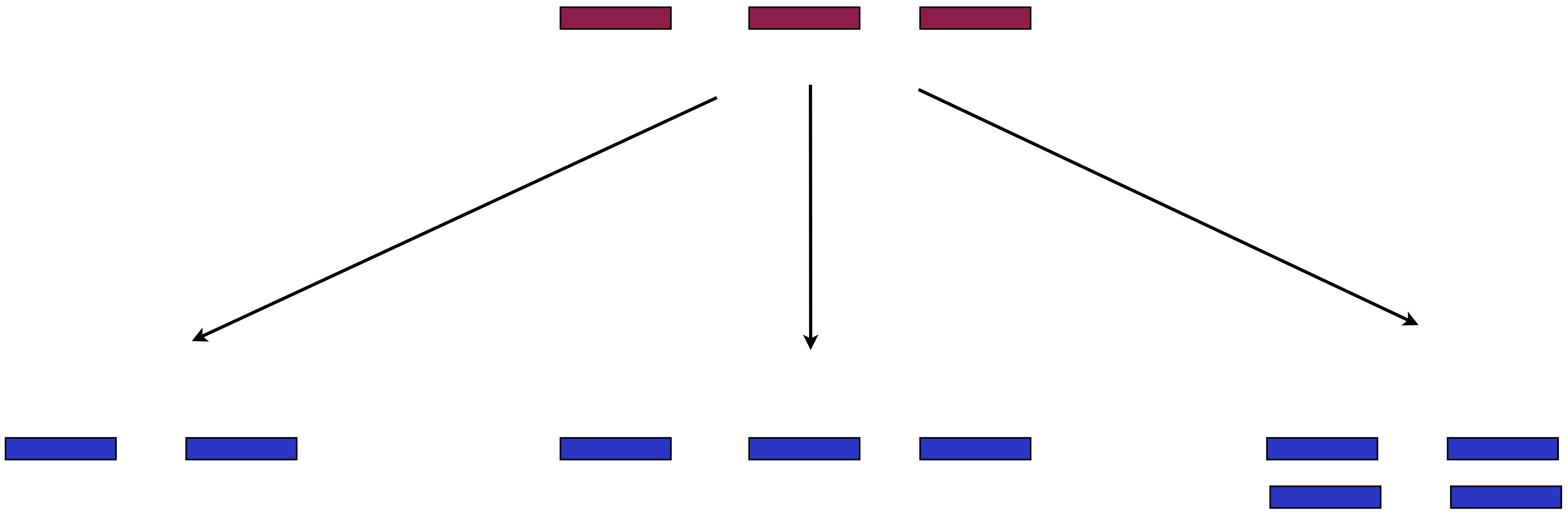
Denton et al. (2014)

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

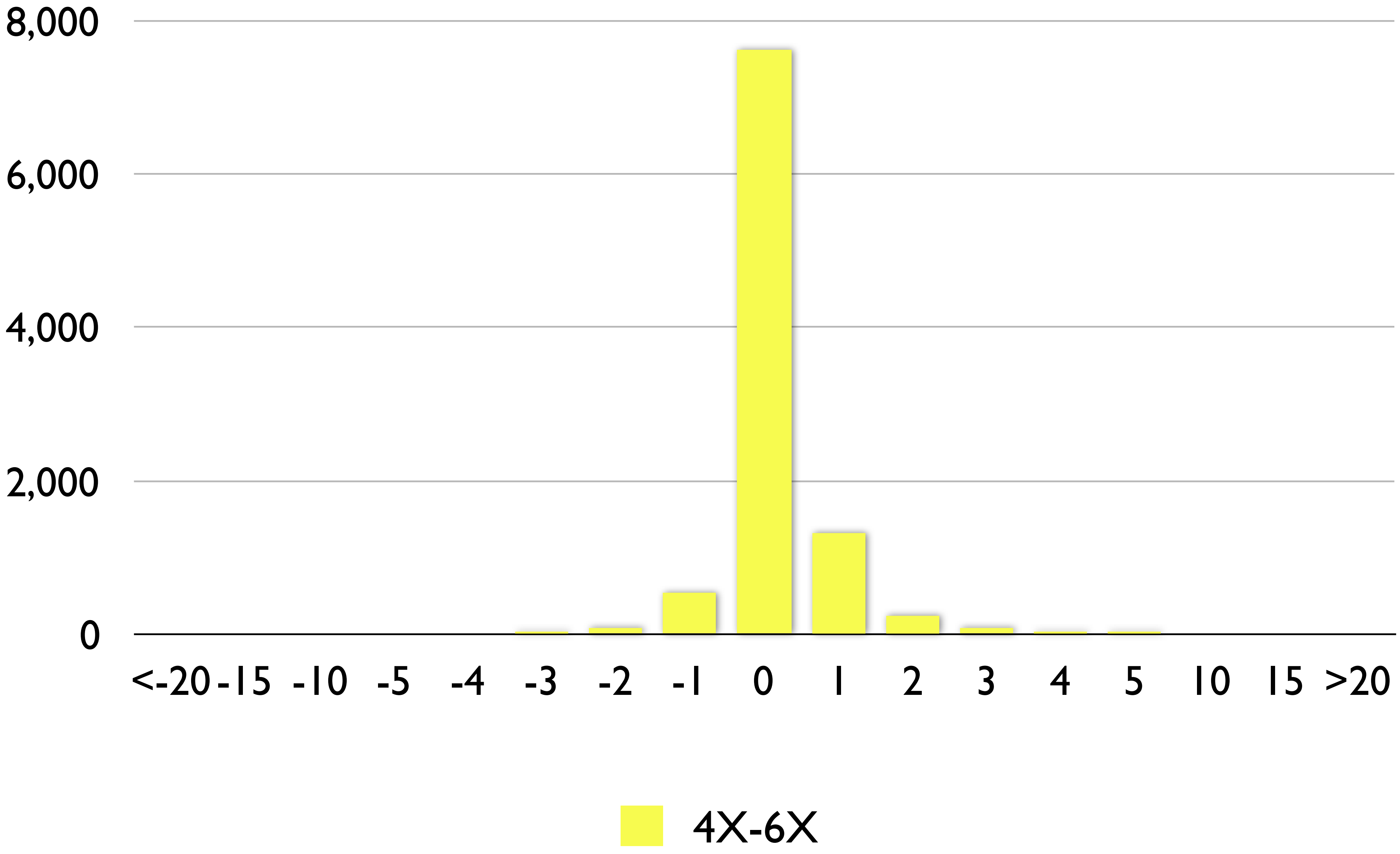
0 change

+1 change

Low-quality genome

Low-quality chimp assembly leads to errors

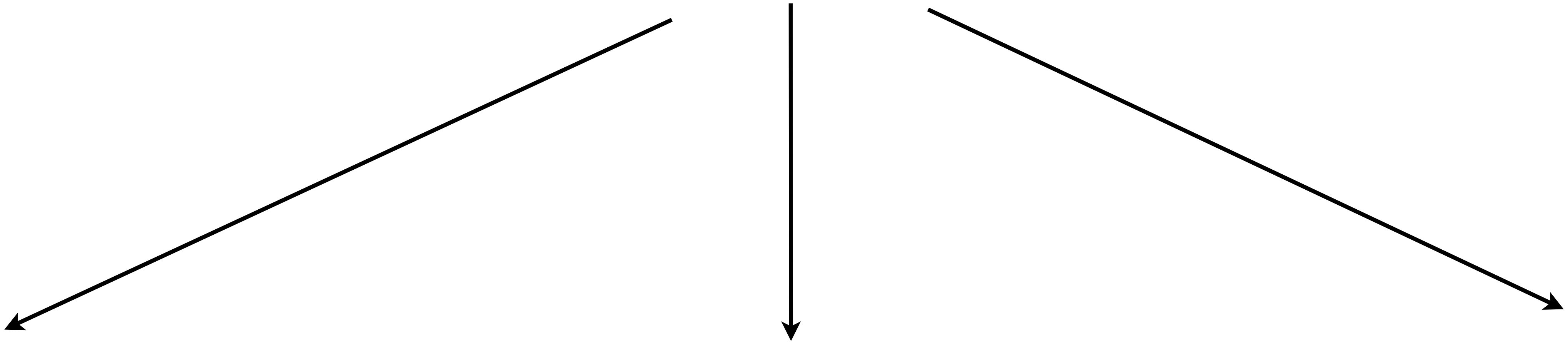
More genes in the lower-quality assembly:



Variation in error due to technology/coverage

Comparison among chicken genomes

High-quality reference genome



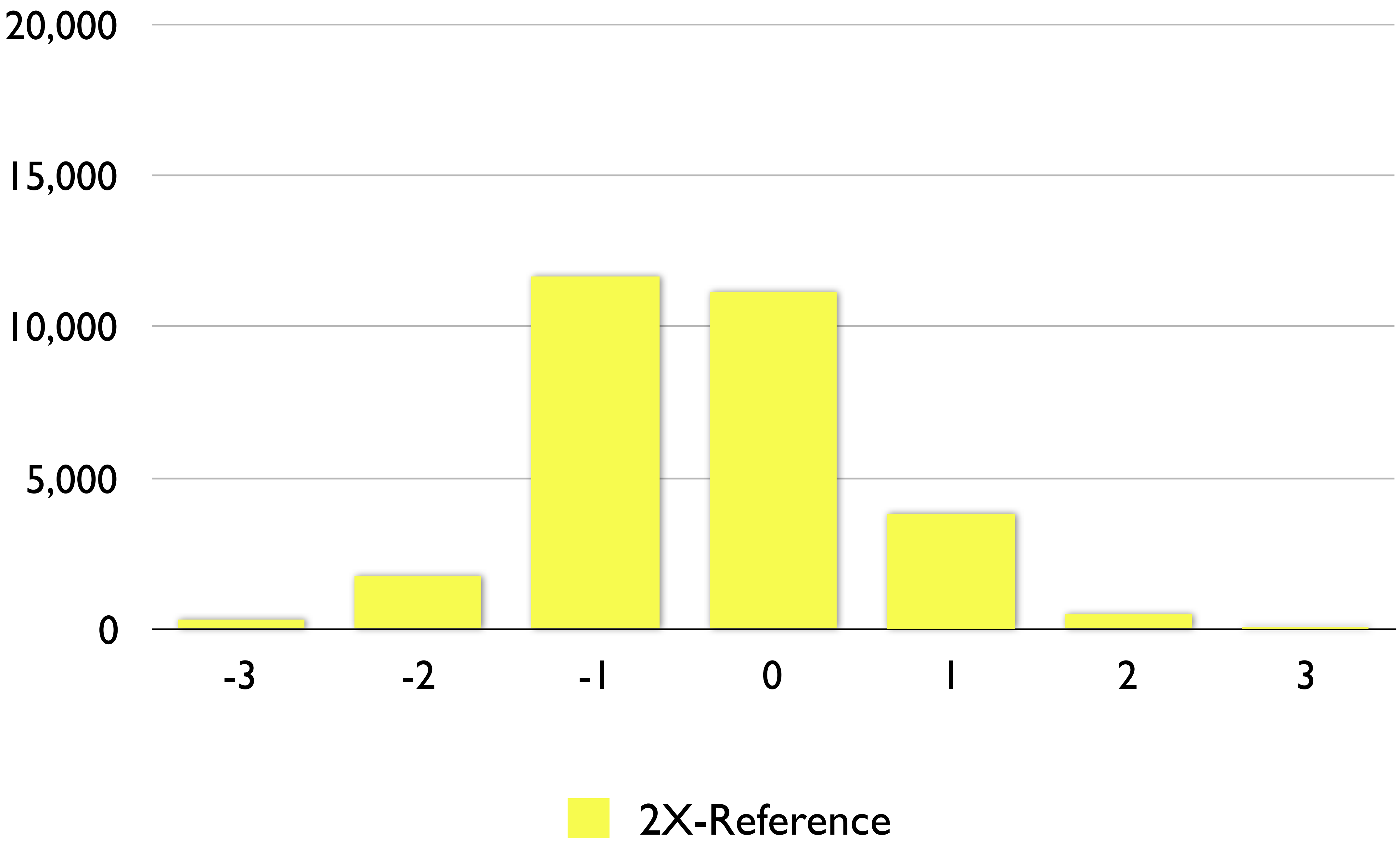
2X Sanger

12X 454

82X Illumina

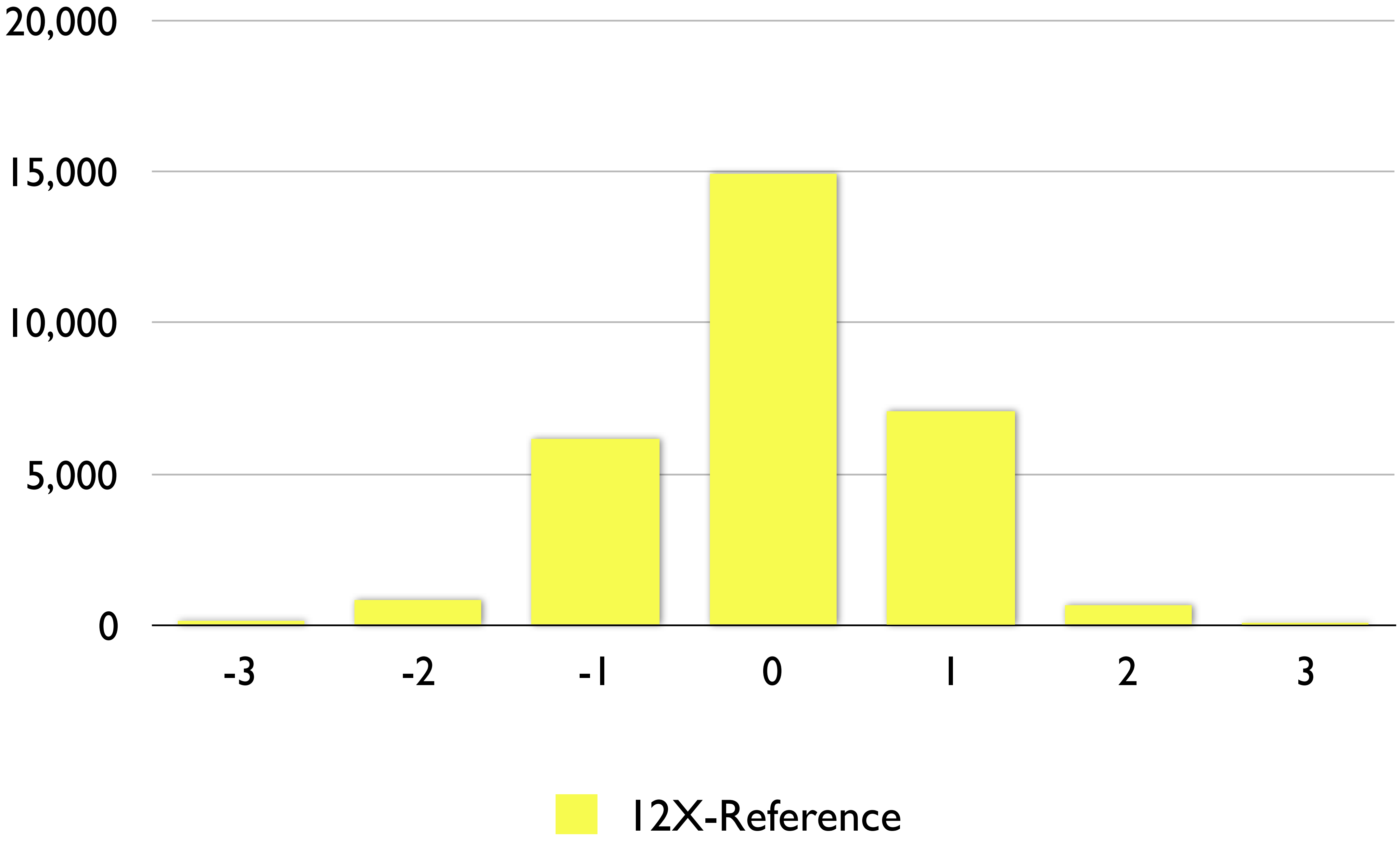
Comparison among chicken genomes

2X Sanger vs. reference



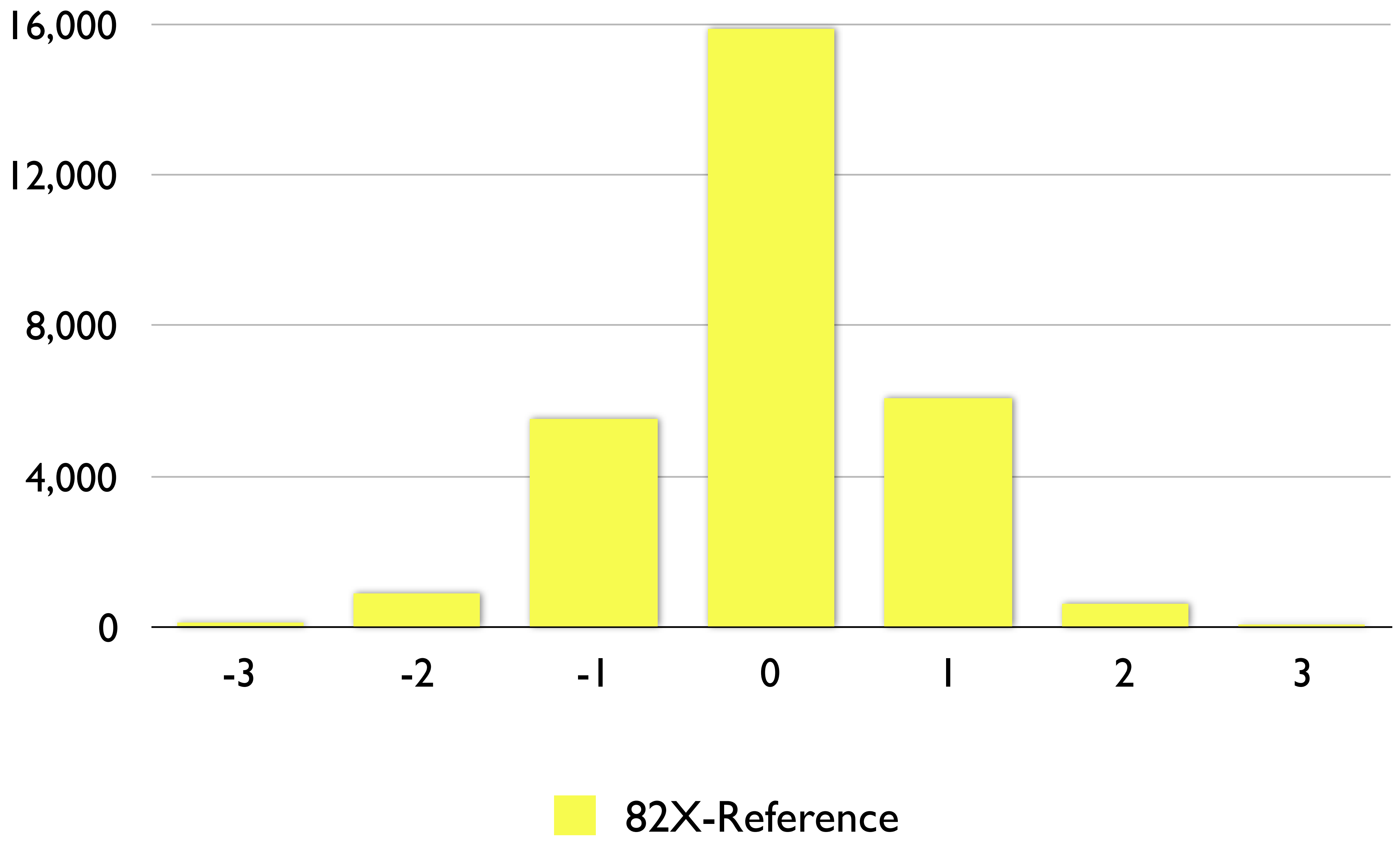
Comparison among chicken genomes

12X 454 vs. reference



Comparison among chicken genomes

82X Illumina vs. 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

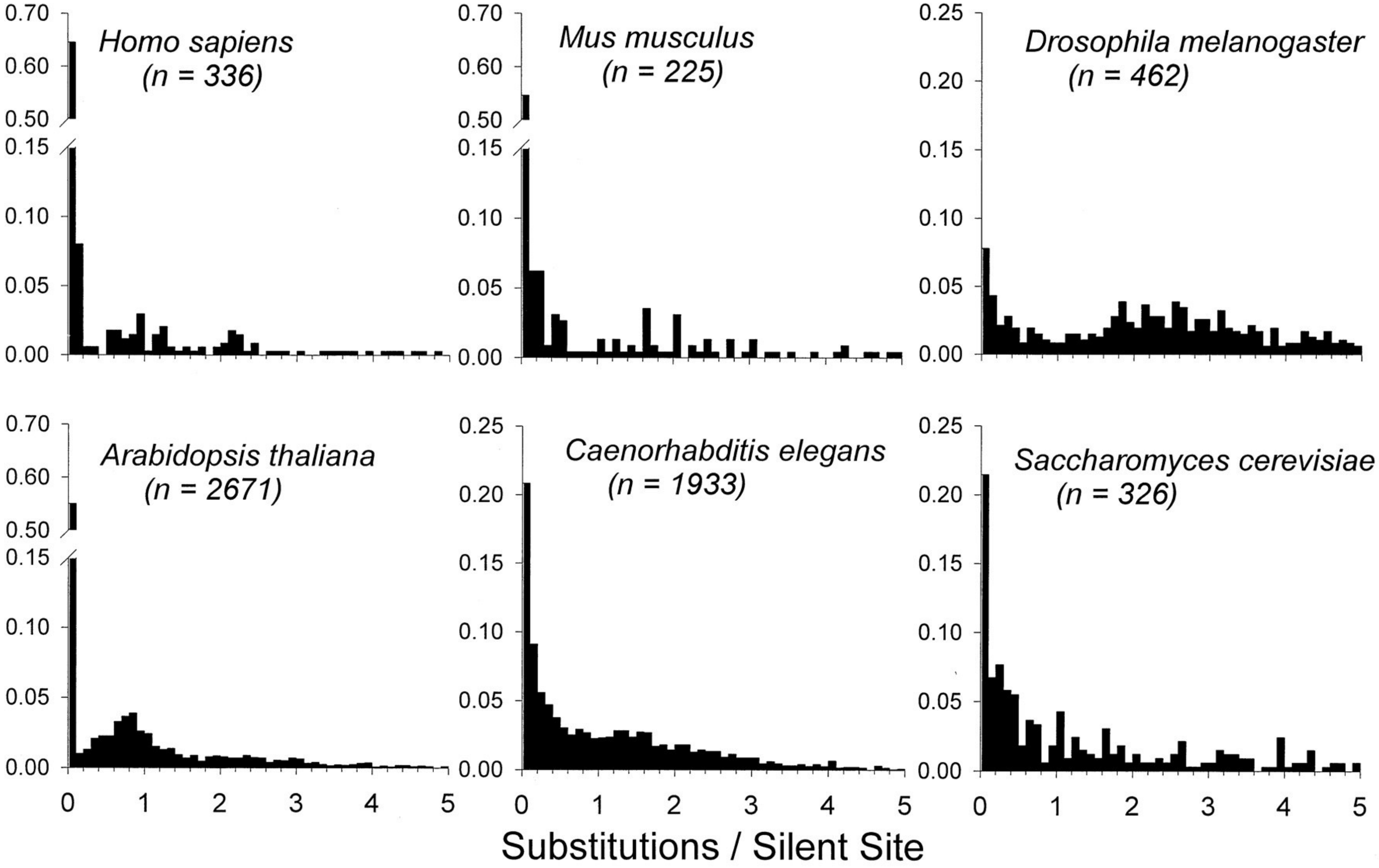
-Species overlap methods

-Gene tree-Species tree reconciliation

-Count methods (e.g. CAFE)

Phylogenetic inference of gene gain and loss

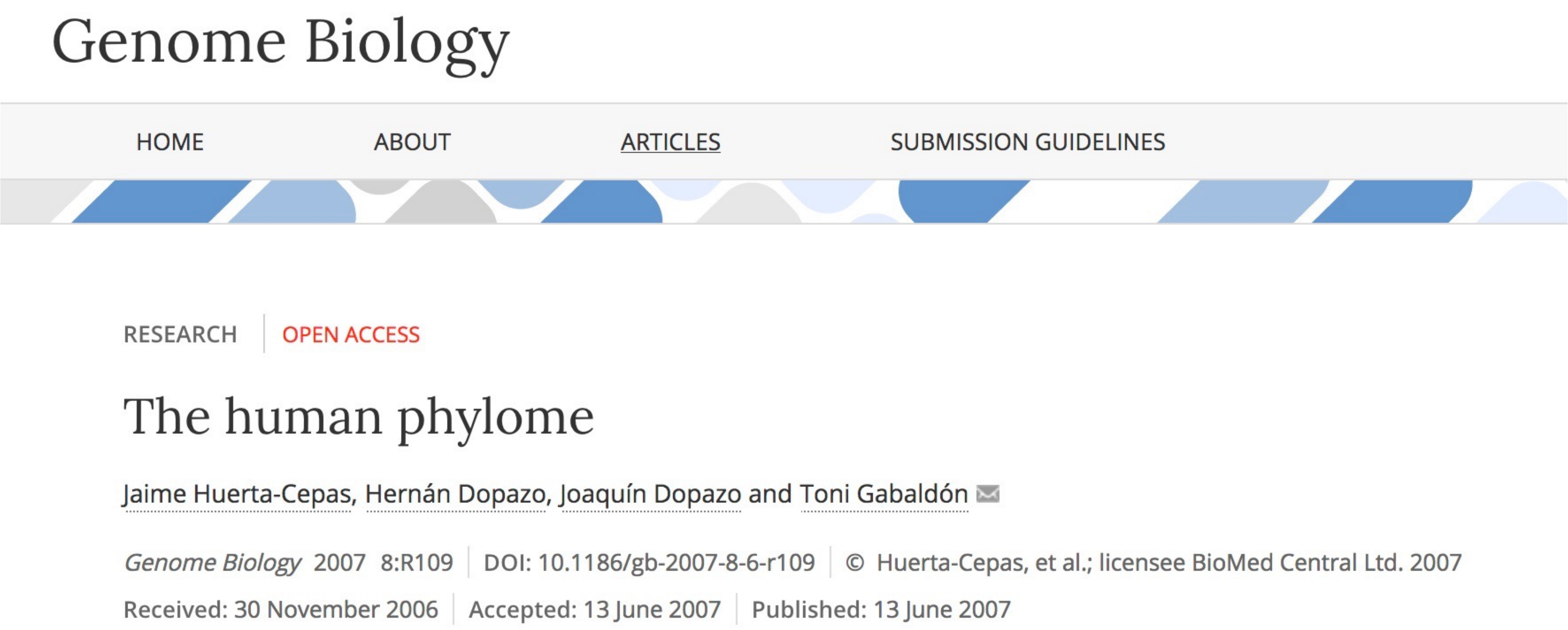
Ks-based methods



Lynch and Conery (2003)

Phylogenetic inference of gene gain and loss

Species overlap methods



Genome Biology

HOME ABOUT ARTICLES SUBMISSION GUIDELINES

RESEARCH | OPEN ACCESS

The human phylome

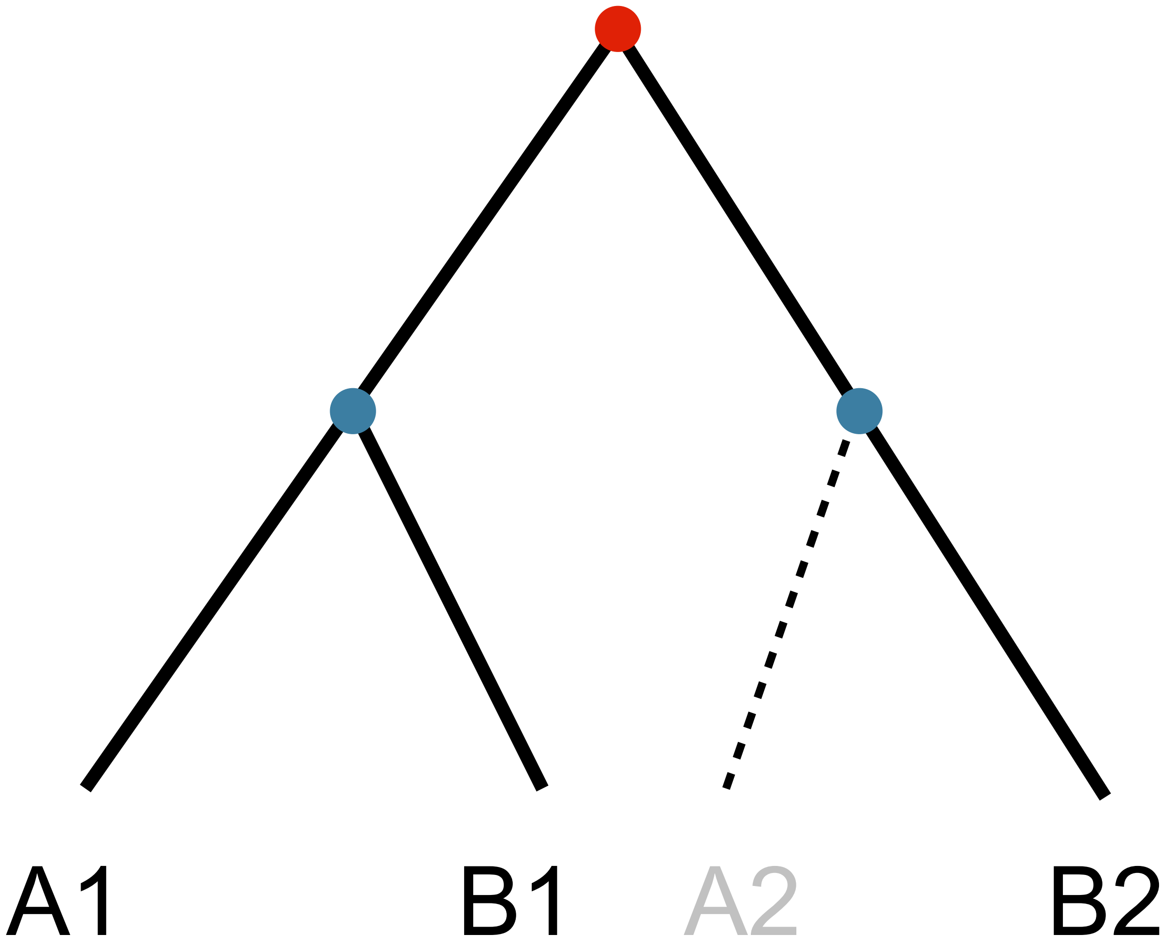
Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo and Toni Gabaldón

Genome Biology 2007 8:R109 | DOI: 10.1186/gb-2007-8-6-r109 | © Huerta-Cepas, et al.; licensee BioMed Central Ltd. 2007

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

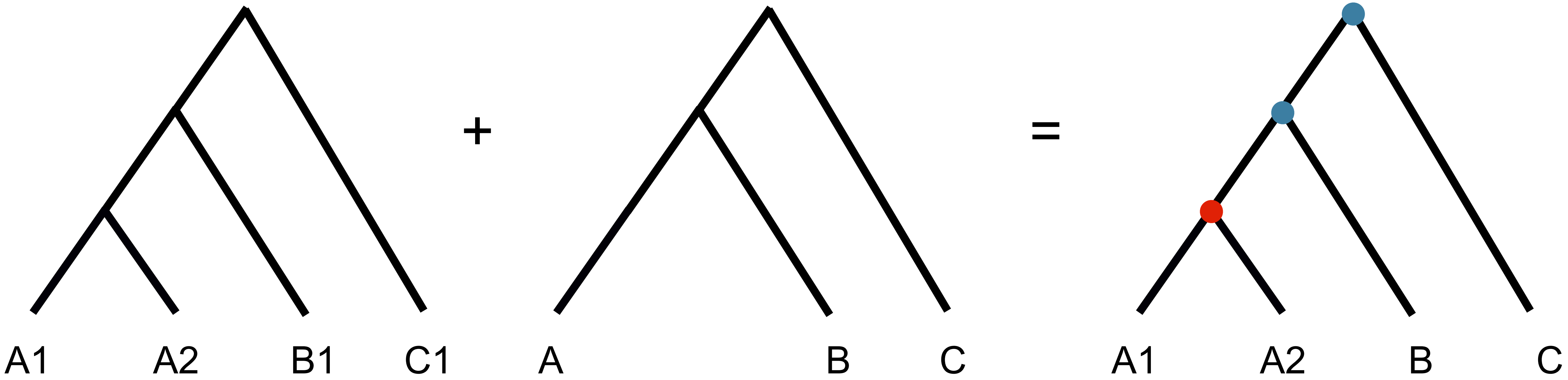
Want to:

-Count duplications and losses

-Identify when they occurred

-(Can be used for species tree inference)

Gene tree reconciliation



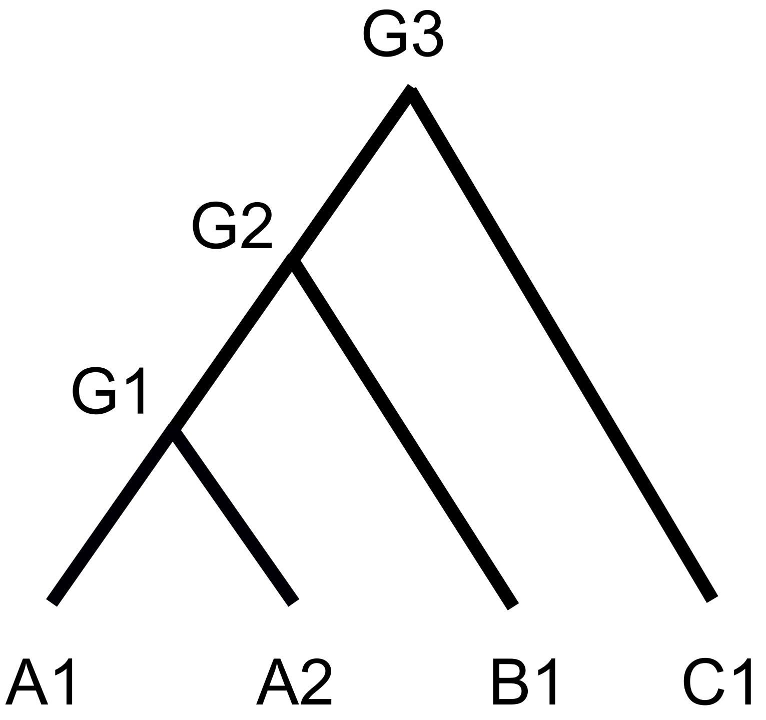
gene tree

species tree

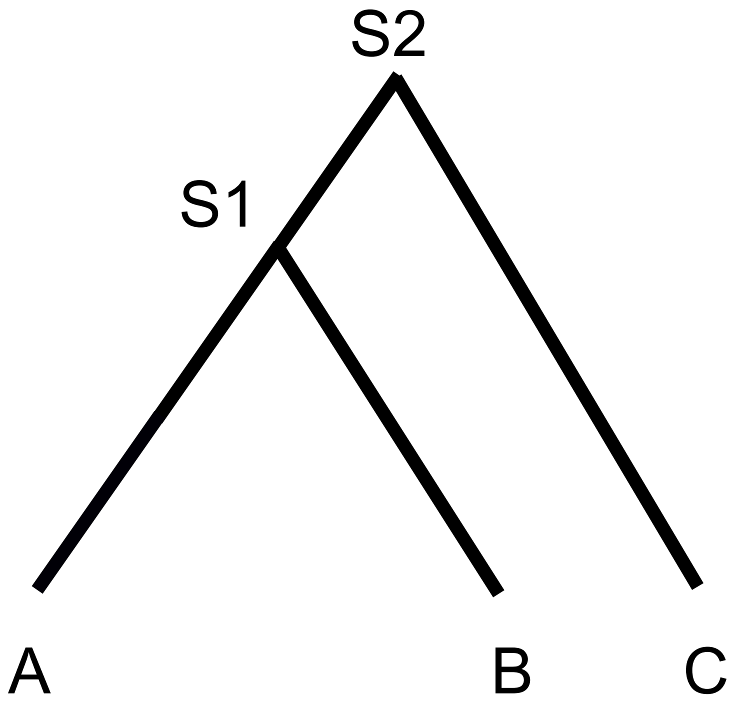
reconciled gene tree

Least common ancestor (LCA) algorithm

1. Label internal nodes



gene tree

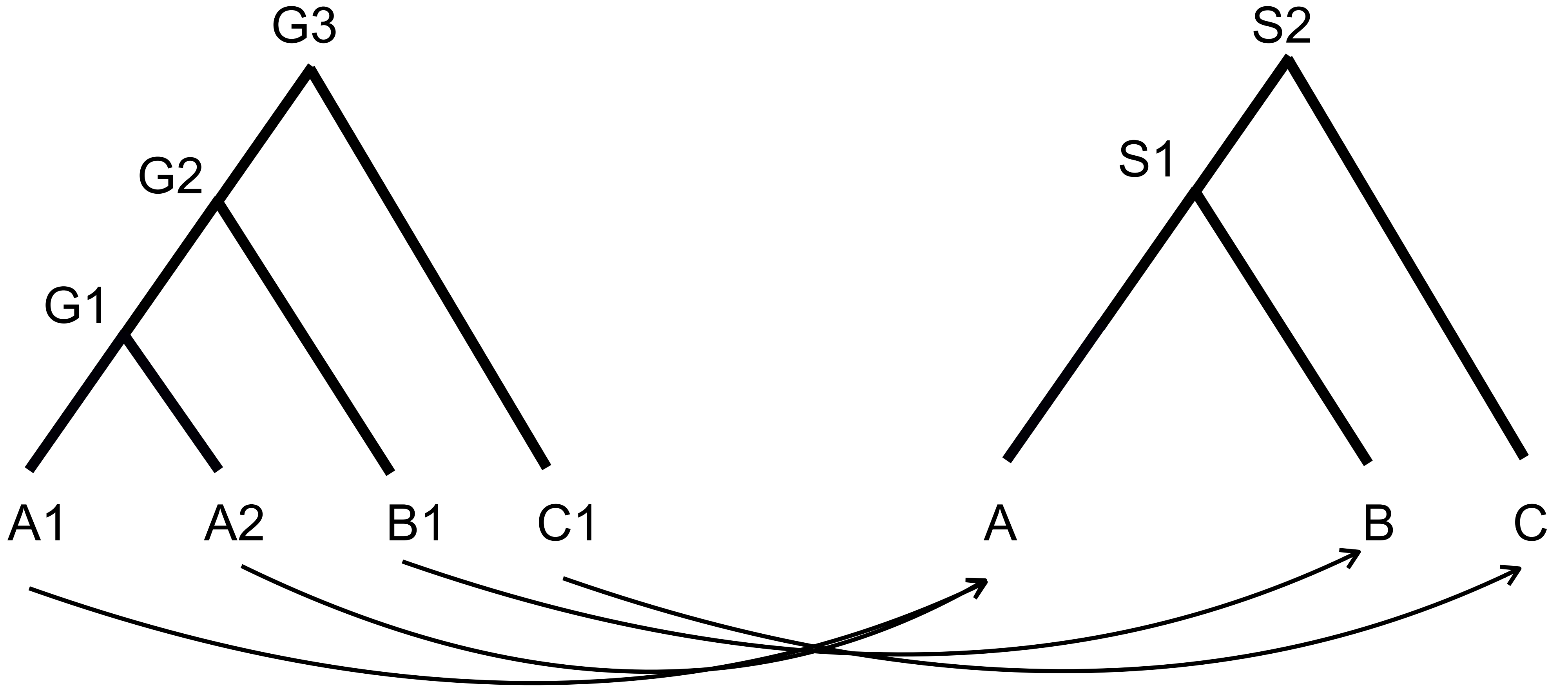


species tree

Goodman et al. (1979)

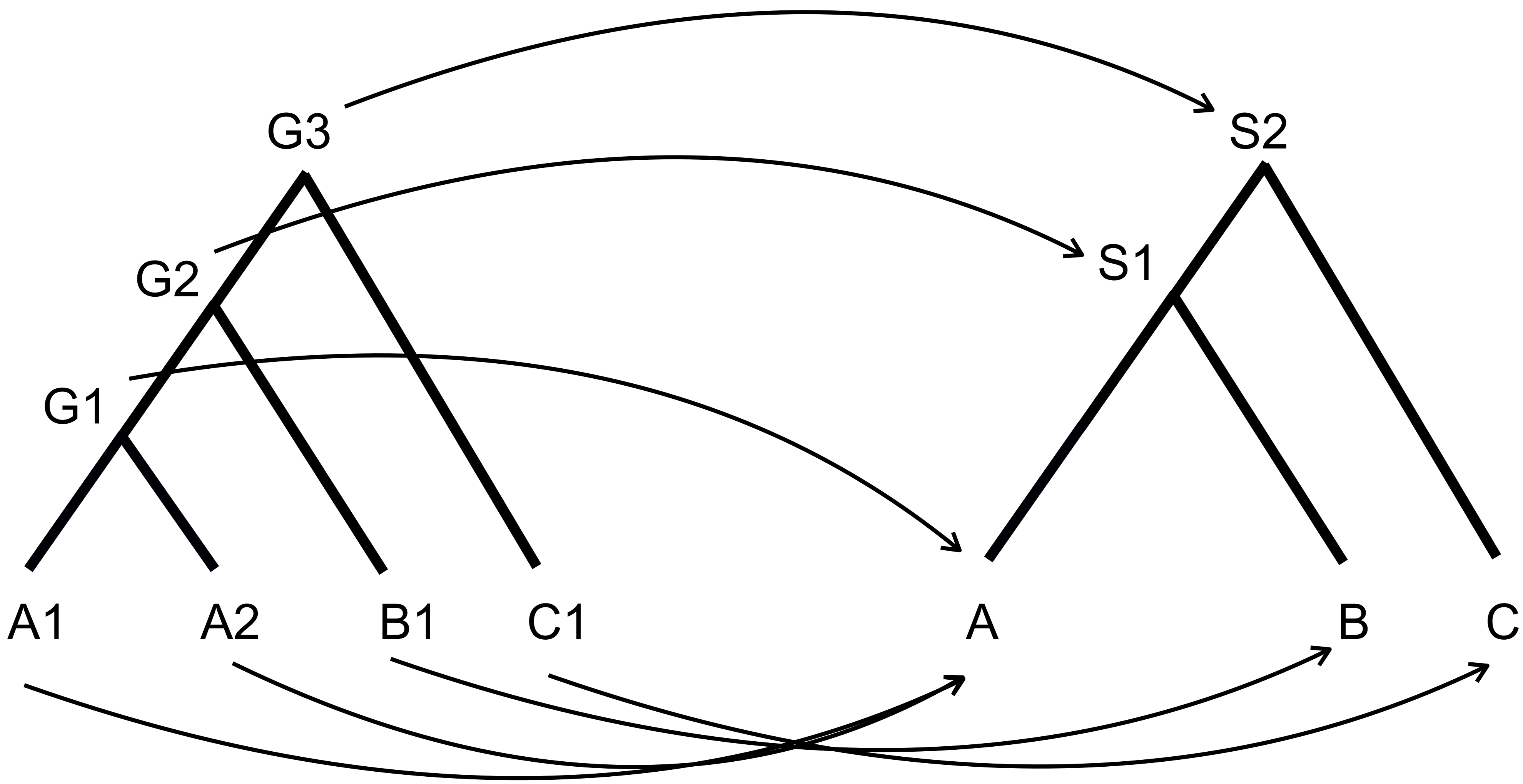
Least common ancestor (LCA) algorithm

2. Initialize map of gene tree tip nodes to species tree tip nodes



Least common ancestor (LCA) algorithm

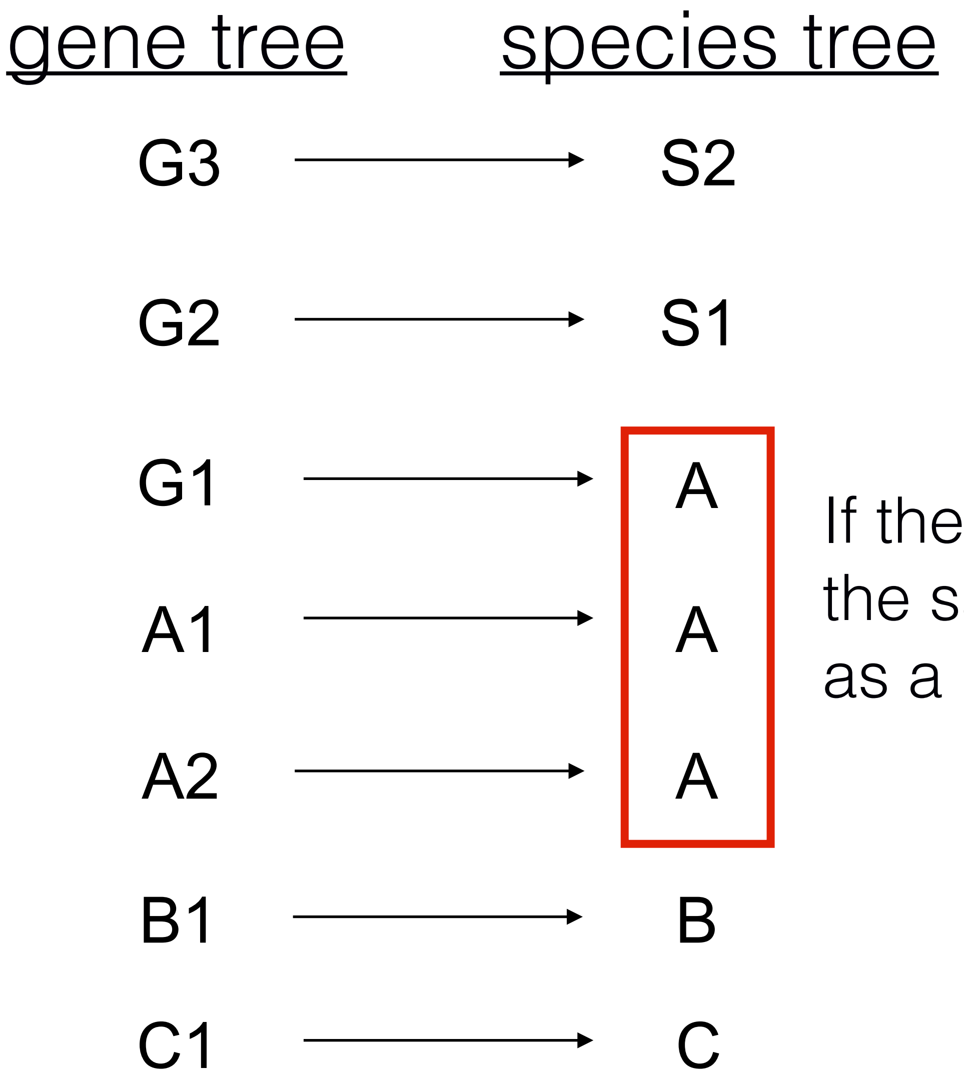
3. Map gene tree internal nodes to species tree nodes:
   this is done to least common ancestor that includes the same lineages



best algorithm: Zmasek and Eddy (2001)

Least common ancestor (LCA) algorithm

Summary of map:

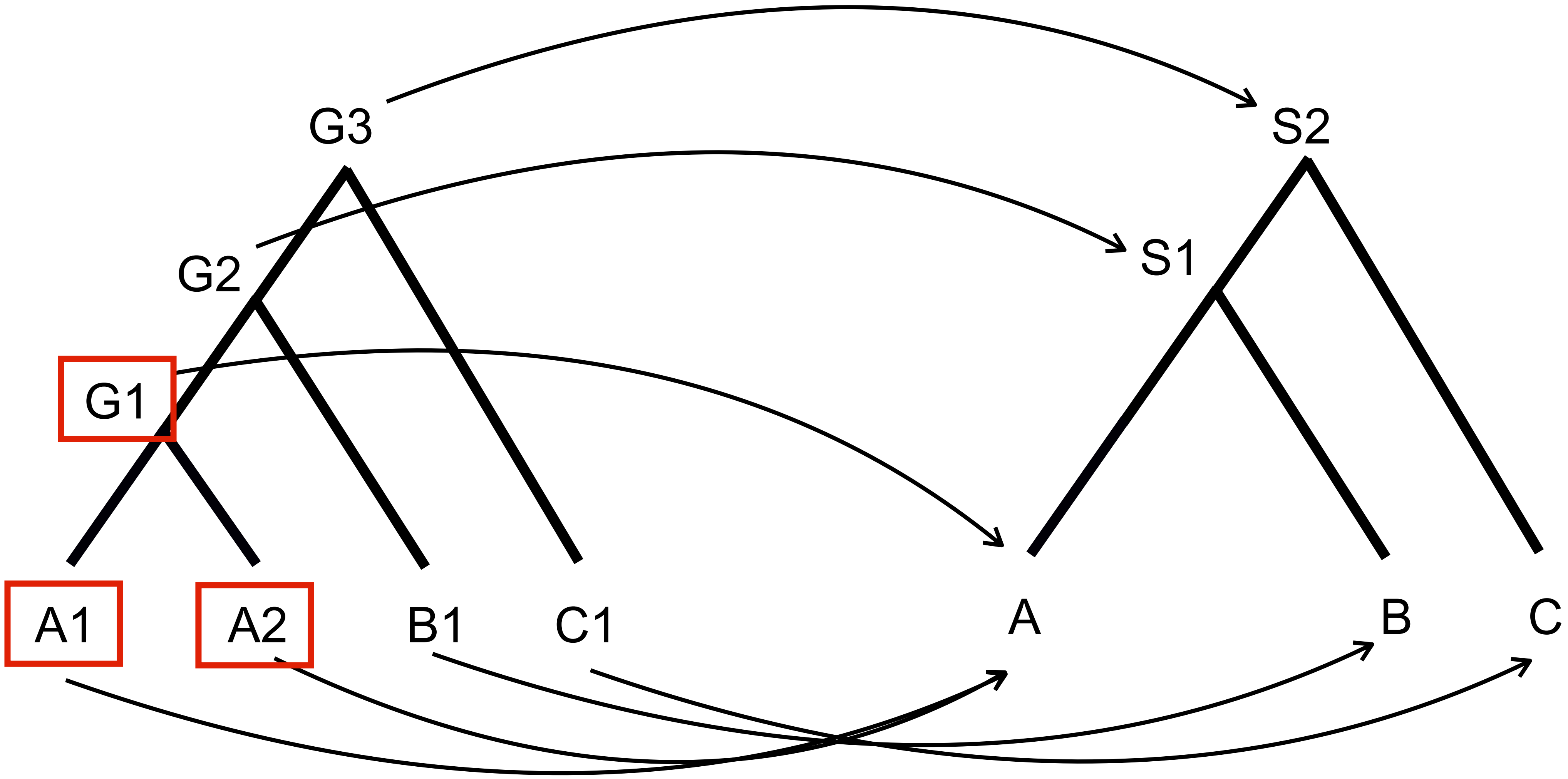


gene tree		species tree
G3	→	S2
G2	→	S1
G1	→	A
A1	→	A
A2	→	A
B1	→	B
C1	→	C

If the map of a parent node is the same as a child, it is labeled as a duplication.

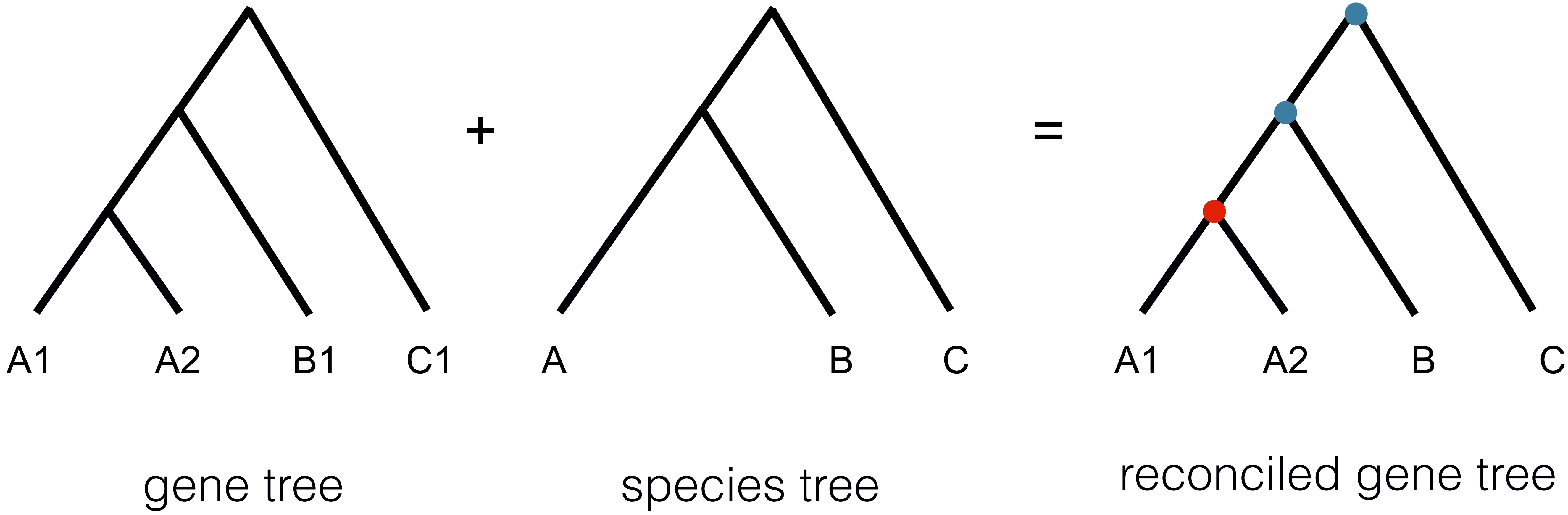
Least common ancestor (LCA) algorithm

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

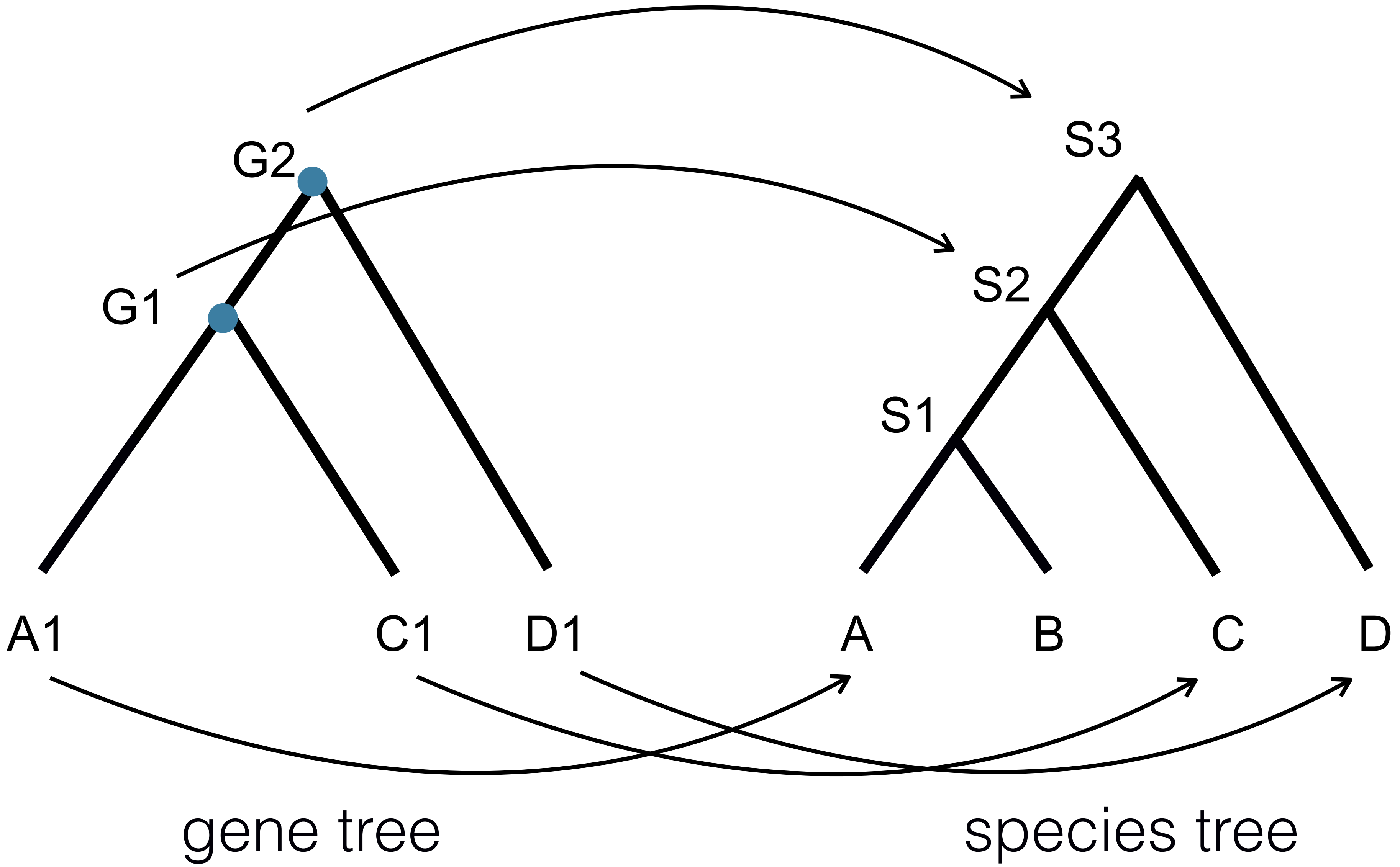


Least common ancestor (LCA) algorithm

What about gene losses?

Least common ancestor (LCA) algorithm

What about gene losses?



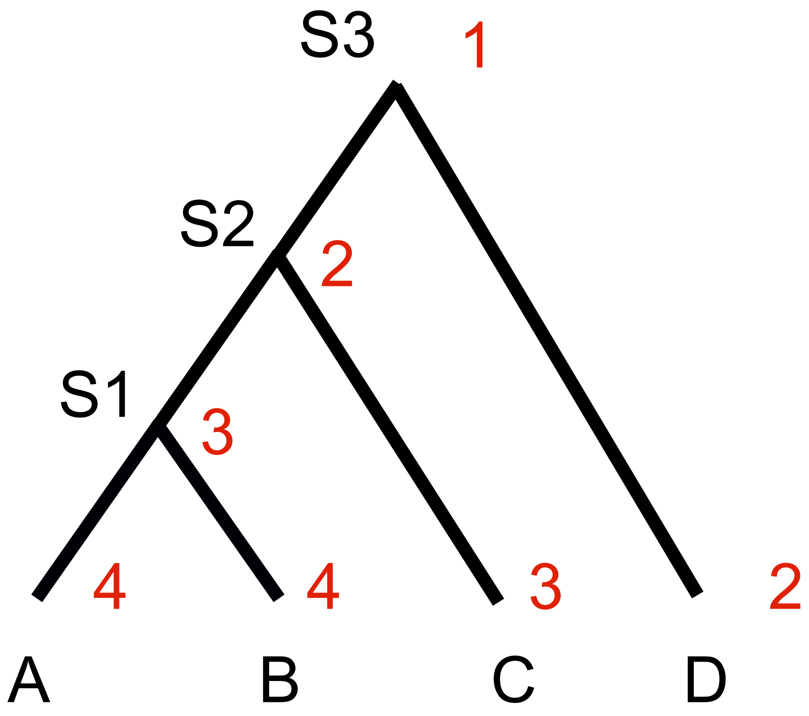
Least common ancestor (LCA) algorithm

What about gene losses?

gene tree	species tree	depth of species tree node
G2	S3	1
G1	S2	2
A1	A	4
C1	C	3
D1	D	2

Least common ancestor (LCA) algorithm

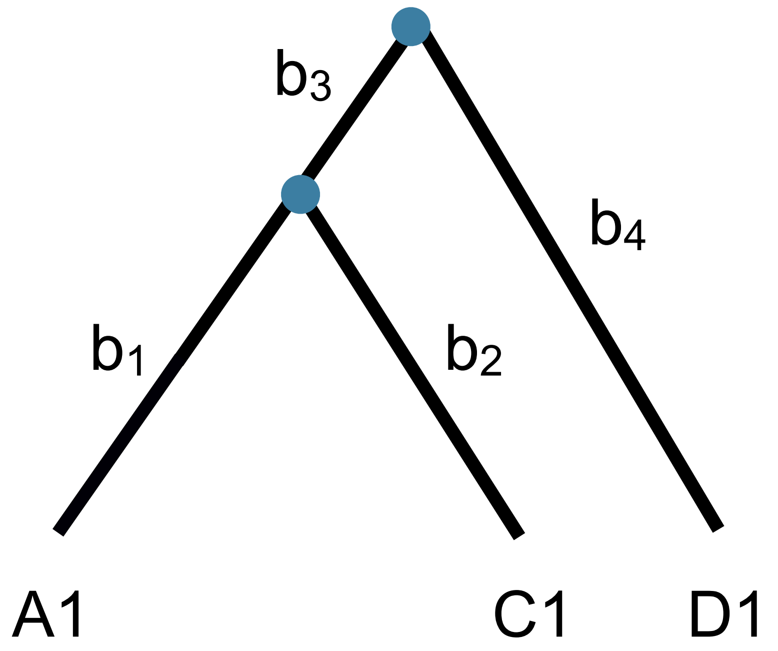
Counting the depth of a node



species tree

Least common ancestor (LCA) algorithm

Counting losses



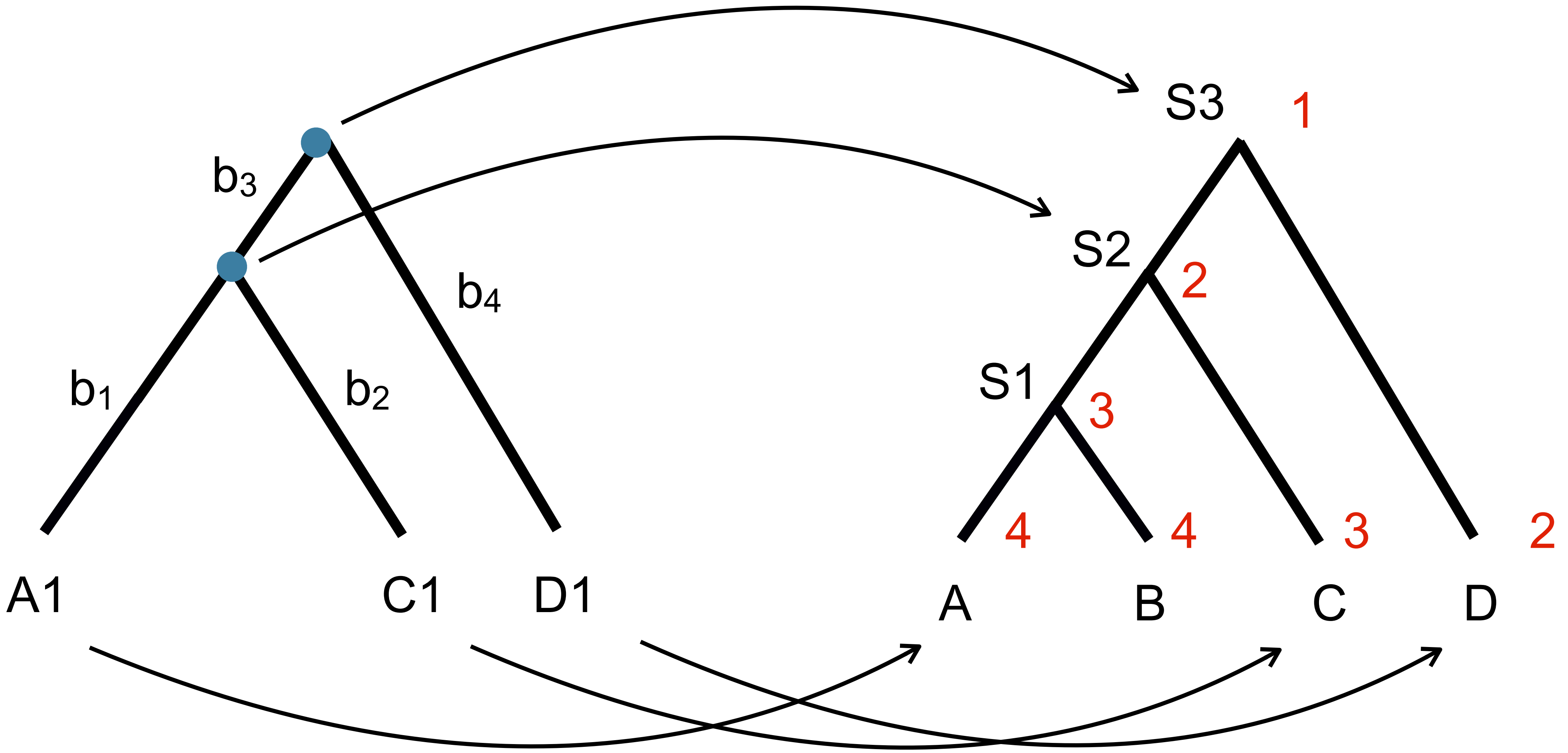
gene tree

$L(b_X)$ = [(depth of daughter)- (depth of parent)-1] + 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



$L(b_1) = (4 - 2 - 1) + 0 = 1$ ← loss on b_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

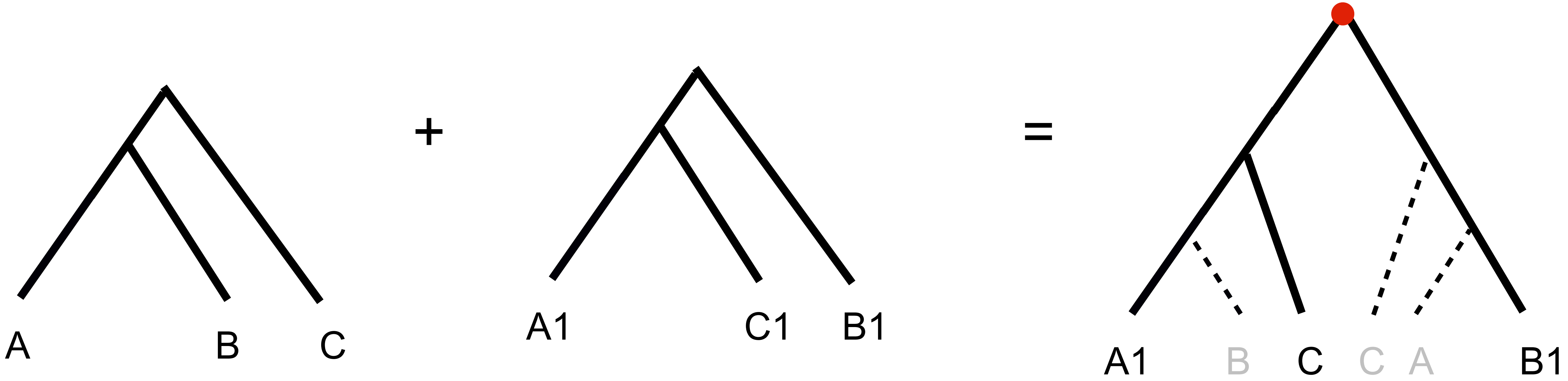
-biological discordance

-gene conversion

-polyploidy

Error in gene trees

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



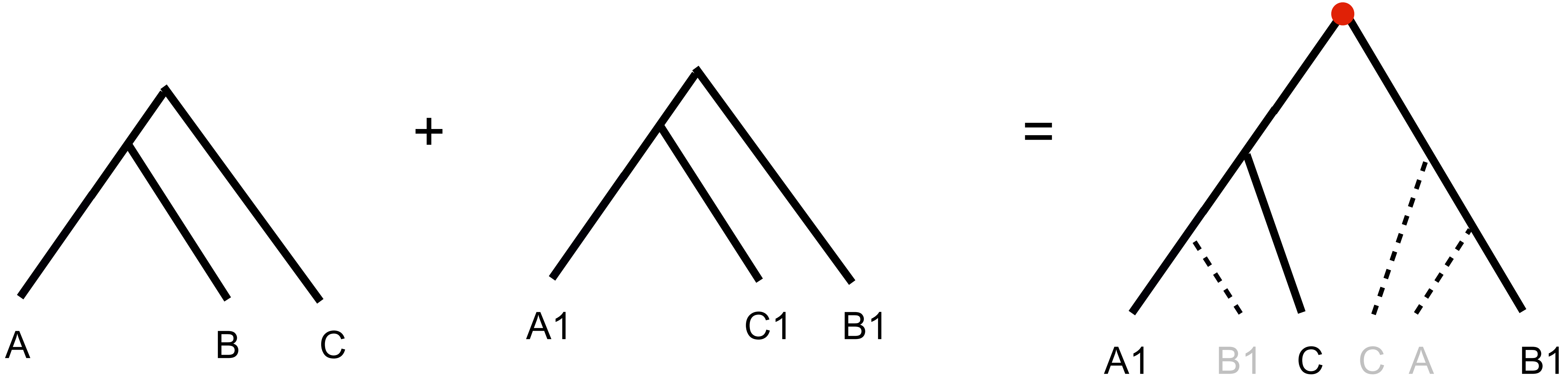
species tree

gene tree

reconciled gene tree

Biological gene tree discordance

If your gene tree is discordant (e.g. due to ILS), reconciliation can result in extra duplications and losses



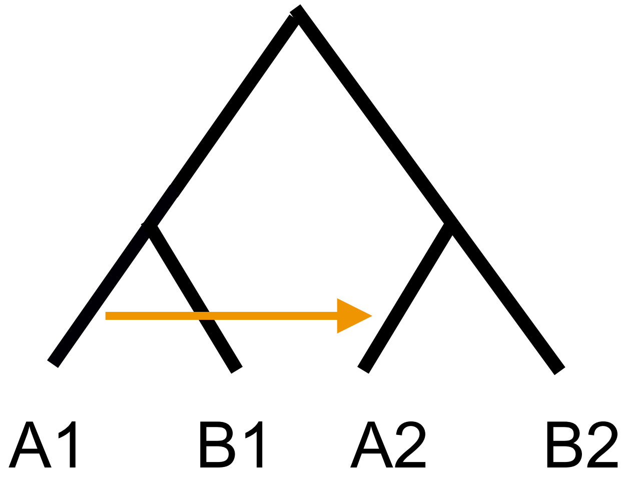
species tree

gene tree

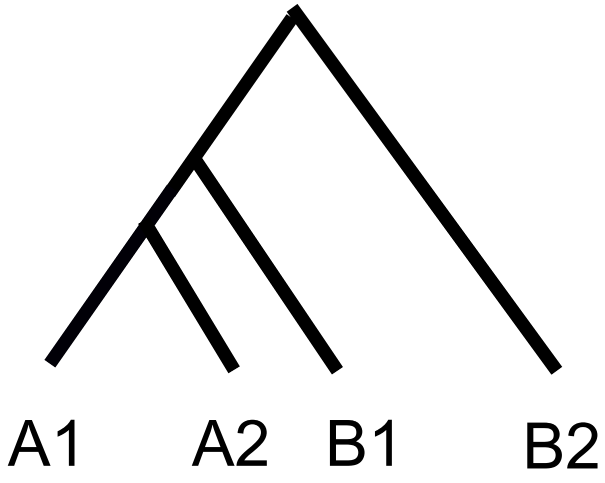
reconciled gene tree

Gene conversion

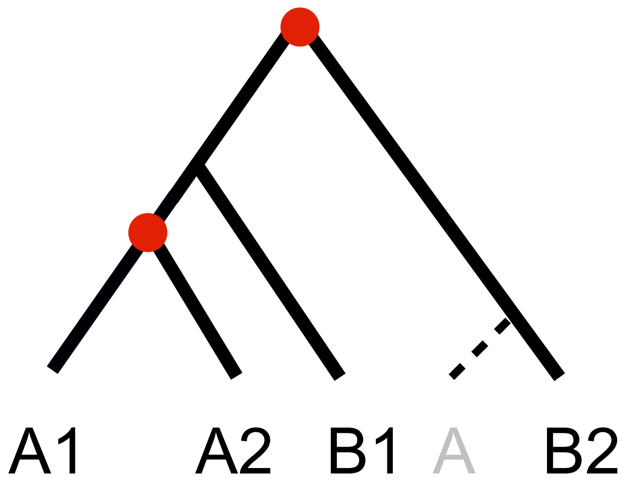
If there is gene conversion, reconciliation can result in extra duplications and losses



gene tree
(before conversion)



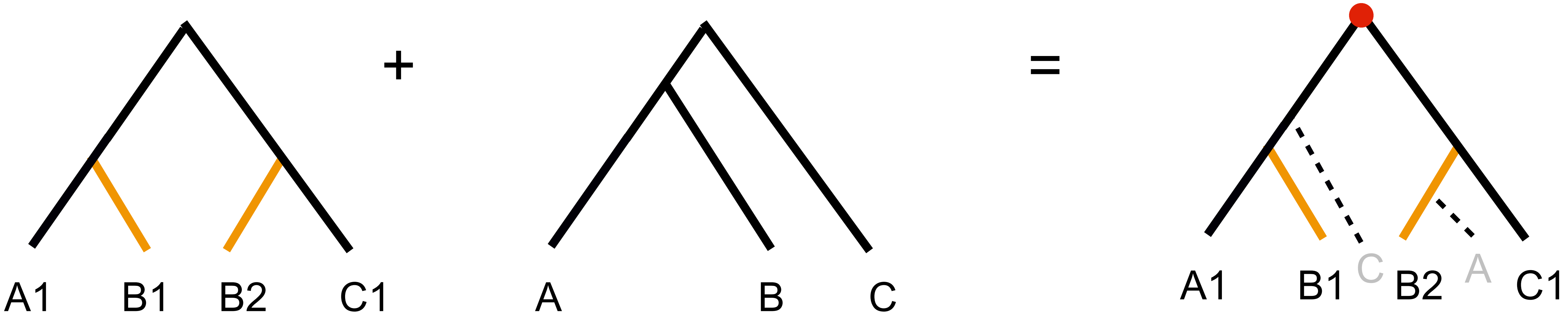
gene tree
(after conversion)



reconciled gene tree

Allopolyploidy

If there is allopolyploidy, reconciliation can result in extra duplications and losses



gene tree

species tree

reconciled gene tree

Solutions!

-gene tree error

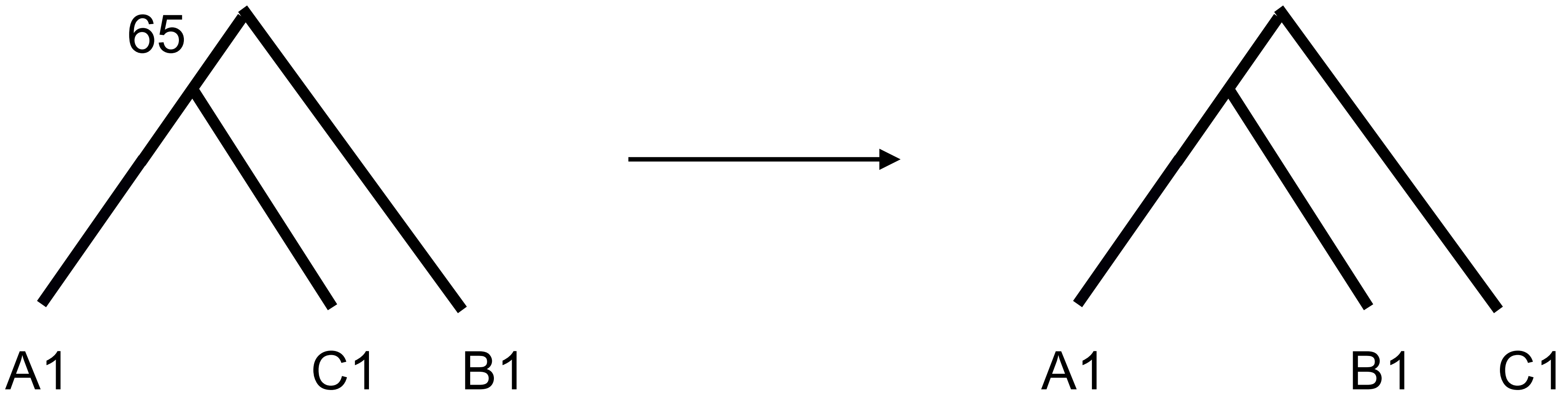
-biological discordance

-polyploidy

-gene conversion

Error in gene trees

Use bootstrap cut-offs to rearrange nodes with low support

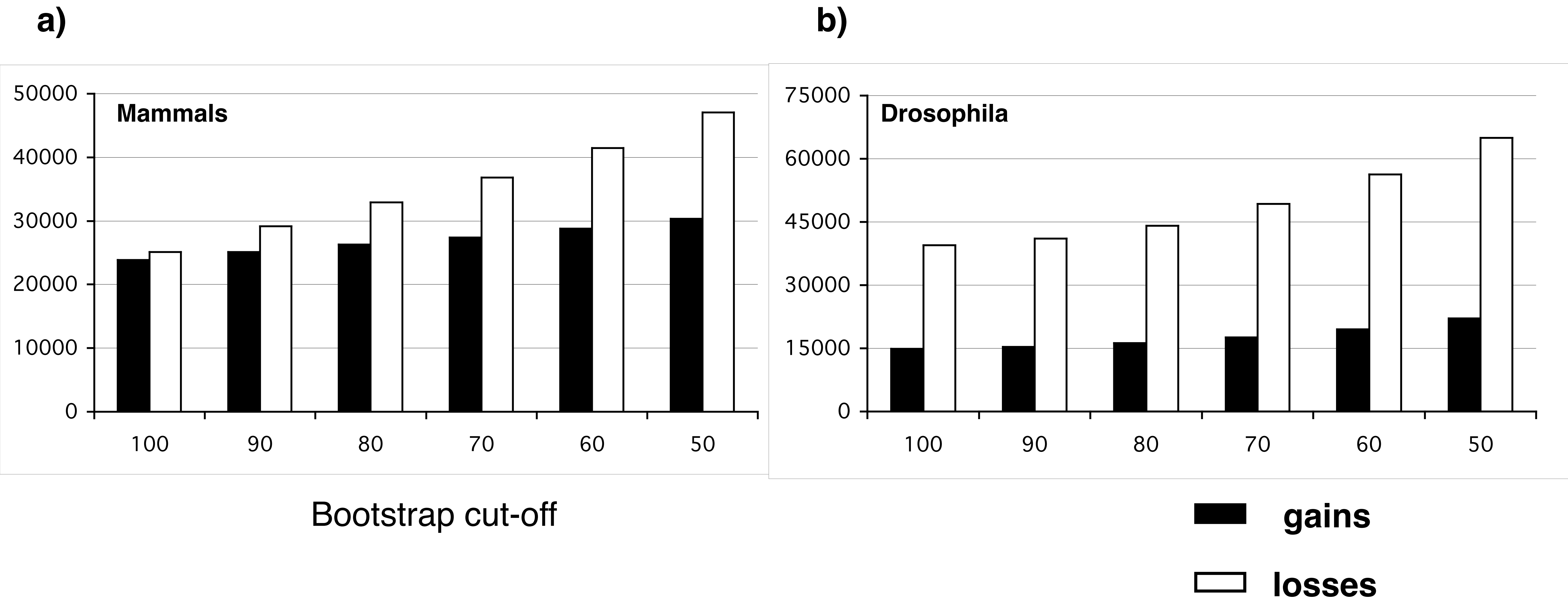


gene tree

corrected gene tree

implemented in Notung (Chen et al. 2000)

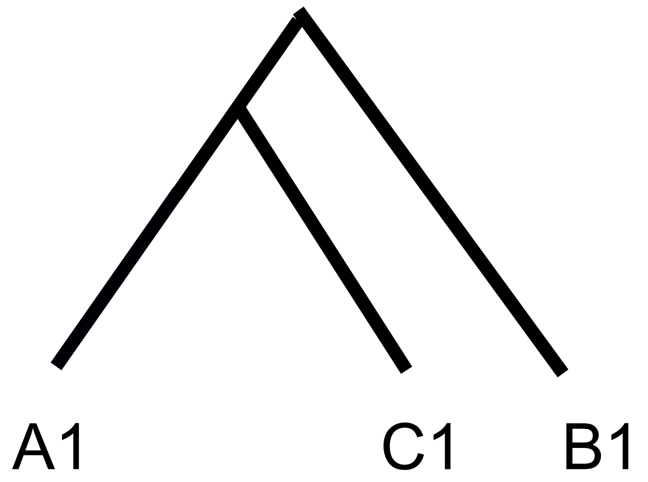
Error in gene trees



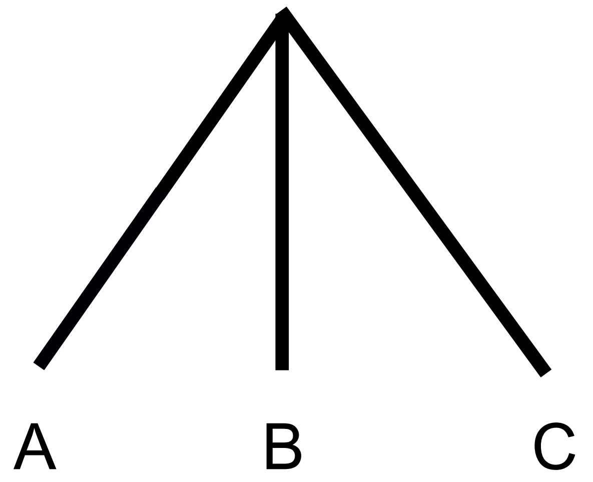
Hahn (2007)

Biological discordance due to ILS

Reconcile to a non-binary species tree



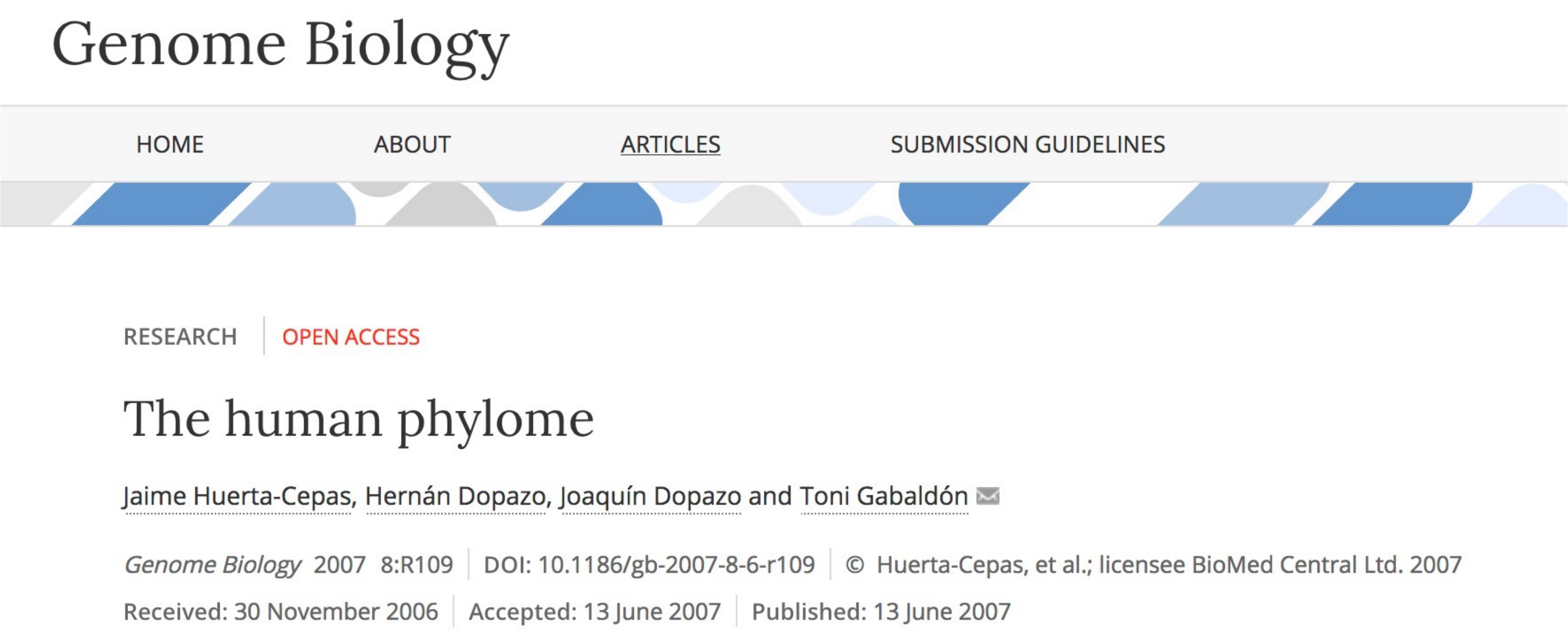
gene tree



species tree

Vernot et al. (2008)

Discordance due to ILS or error



Genome Biology

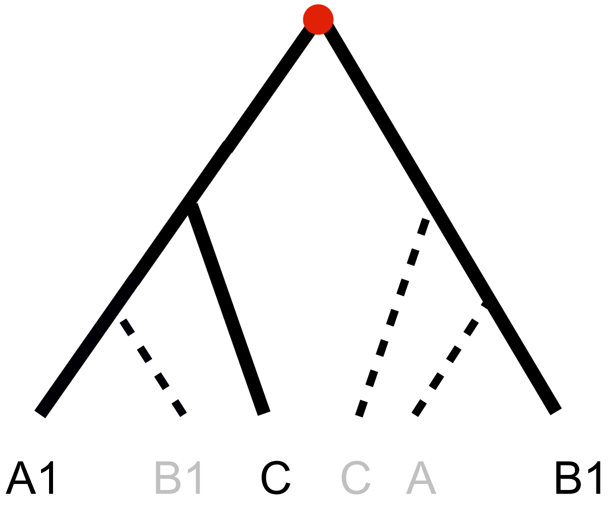
RESEARCH | OPEN ACCESS

The human phylome

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

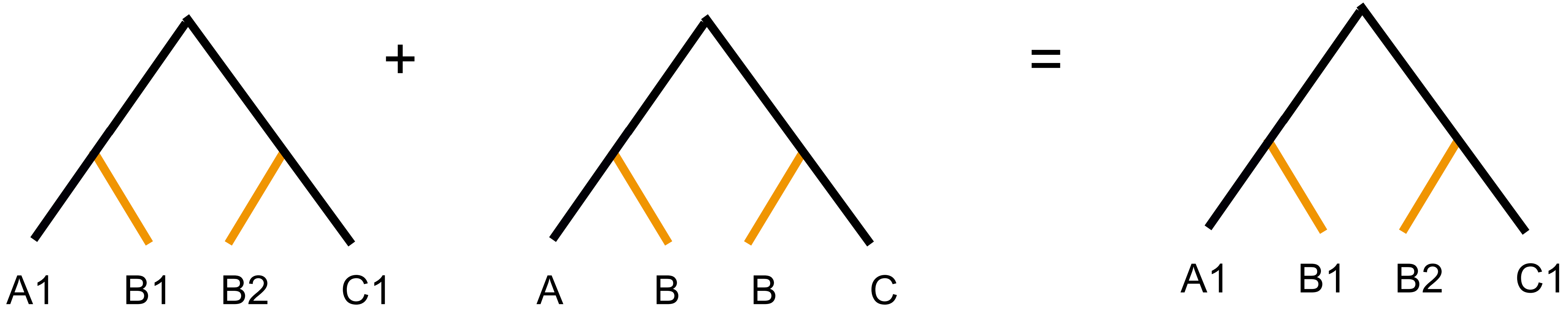
Genome Biology 2007 8:R109 | DOI: 10.1186/gb-2007-8-6-r109 | © Huerta-Cepas, et al.; licensee BioMed Central Ltd. 2007

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



Allopolyploidy

Reconcile to a multiply-labeled (MUL-) tree



gene tree

MUL-tree

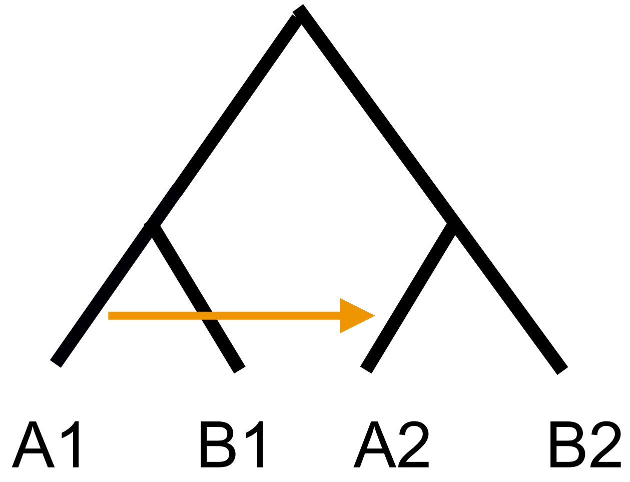
reconciled gene tree

Thomas et al. (biorxiv)

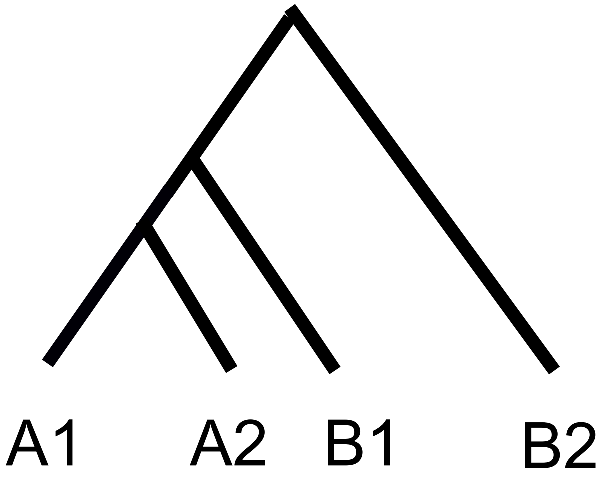
Gene conversion

What to do about gene conversion?

"Count" methods!



gene tree
(before conversion)



gene tree
(after conversion)