

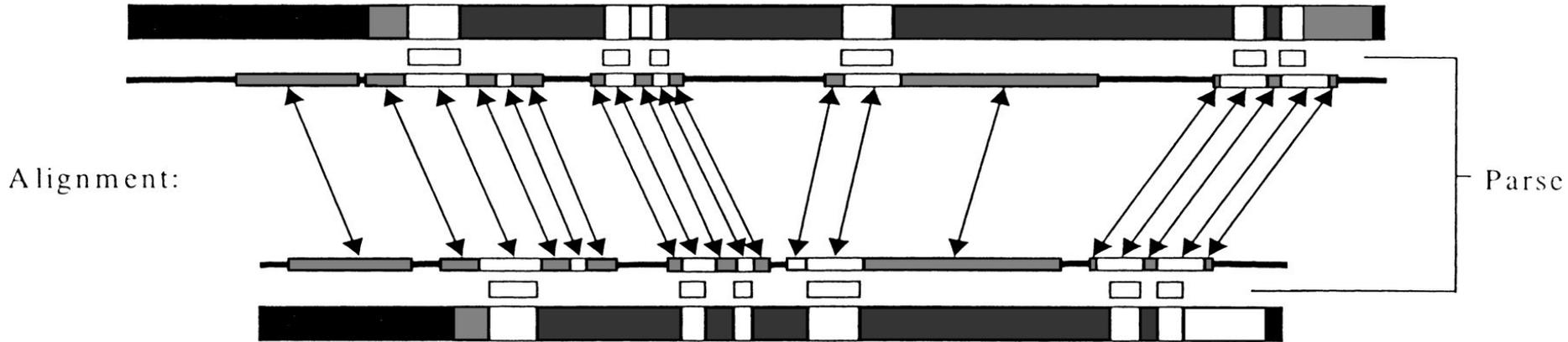
What is comparative genomics?

- Analyzing & comparing genetic material from different species to study evolution, gene function, and inherited disease
- Understand the uniqueness between different species

What is compared?

- Gene location
- Gene structure
 - Exon number
 - Exon lengths
 - Intron lengths
 - Sequence similarity
- Gene characteristics
 - Splice sites
 - Codon usage
 - Conserved synteny

Human Locus: HUMPCNA



Mouse Locus: MMPCNAG

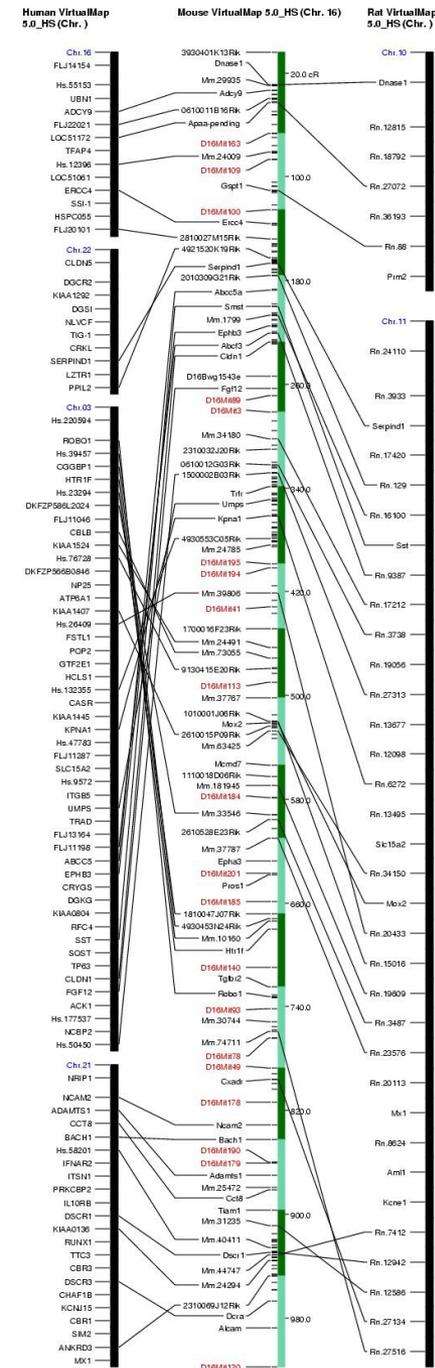
Figure 1 Regions of the human and mouse homologous genes: Coding exons (white), noncoding exons (gray), introns (dark gray), and intergenic regions (black). Corresponding strong (white) and weak (gray) alignment regions of GLASS are shown connected with arrows. Dark lines connecting the alignment regions denote very weak or no alignment. The predicted coding regions of ROSETTA in human, and the corresponding regions in mouse, are shown (white) between the genes and the alignment regions.

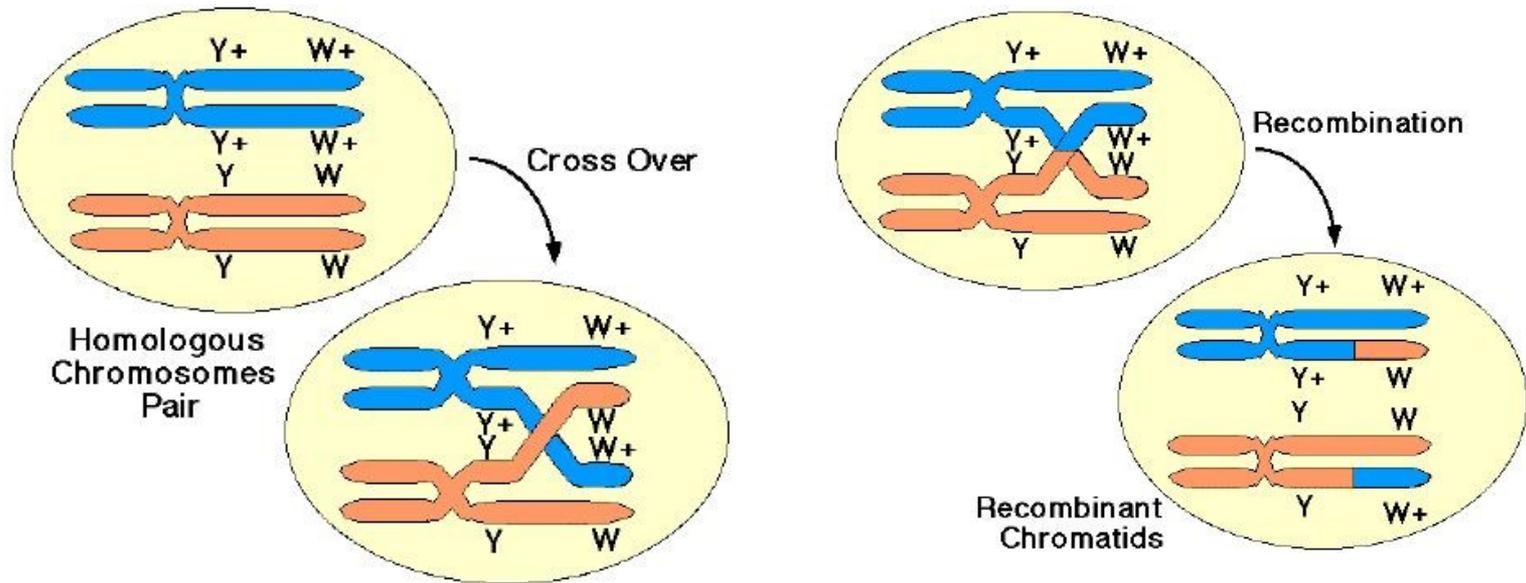
Comparison of mouse chromosome 16 and the human genome

- Mural et al., Science, 2002, 296:1661
- Celera group
- Synteny with human chr.'s 3,8,12,16,21,22 and rat chr.'s 10,11

Q: Why more breakpoints in mouse-human than in mouse-rat?

Q: Why more conserved genes in human than in rat?





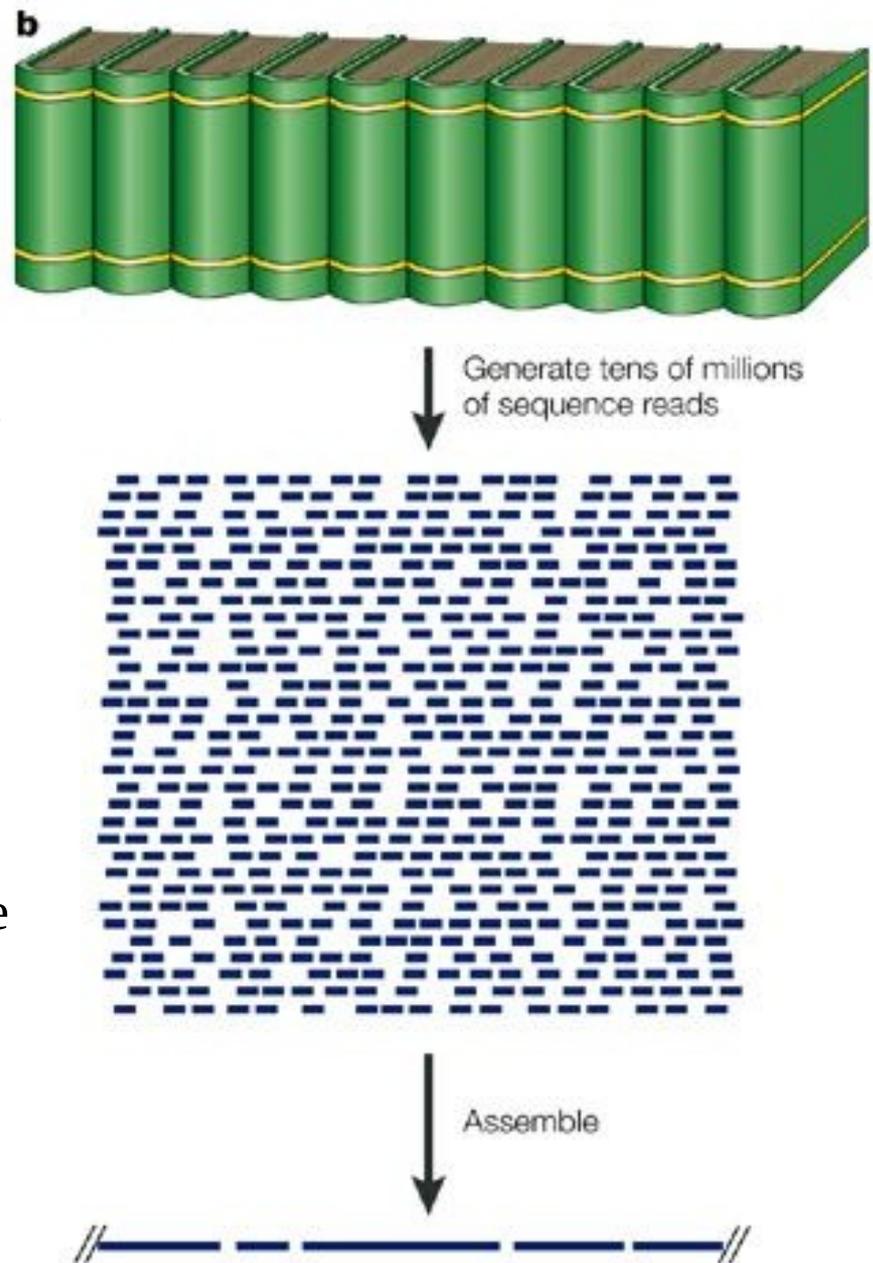
- This also can occur between chromosomes
- The longer the divergence time between 2 species, the more recombination has occurred
- 100 million years since human-mouse divergence
- 40 million years since rat-mouse divergence

Whole-genome shotgun sequencing:

1. Genome is cut into small sections
2. Each section is hundreds or a few thousand bp of DNA
3. Each section is sequenced and put in a database
4. A computer aligns all sequences together (millions of them from each chromosome) to form contigs
5. Contigs are arranged (using markers, etc) to form scaffolds

Q: What are the advantages of this over the traditional method?

Q: What are the potential sources of error?



1. Assembly of Mmu16

- Total size: 99Mbp
 1. Not one contiguous sequence (contig)
 2. 8,635 contigs on 20 “scaffolds”
 3. Average scaffold size: 10Mbp
 4. Number of gaps: 8615
 5. Total size of gaps: ~6Mbp
 6. Total coverage: ~93Mbp

Q: Why are there so many gaps?

Q: Is there a great danger of missing genes that might be in the gaps?

2. Identify genes in Mmu16

1. Scaffolds of >10kbp were examined (scaffolds larger than 1Mbp were chopped)
2. Regions with repeat motifs were ignored using RepeatMasker
3. Several gene prediction engines use (GenScan, Grail, Fgenes)
4. Amino acid sequences from open reading frames searched against nr protein db (NCBI)
5. Nucleotide searchers (using DNA from across scaffolds) performed against:
 1. Celera's gene clusters
 2. Mmu, Rno, & Hsa EST db's
 3. NCBI's RefSeq mRNA db
 4. Celera's dog genomic db
 5. Public pufferfish genomic db

Q: Why dog and pufferfish? Why not just rat/mouse/human?

Q: Why not just use DNA from the predicted genes for the db searches?

Q: Why were only regions >10kbp used?

2. Identify genes in Mmu16

6. 1055 genes with high & medium confidence were predicted
7. Other efforts have identified 1142 genes
8. After visual annotation inspection, psuedogenes and annotation errors removed, leaving 731 homologues genes
9. The genes found were mostly orthologues because they were reciprocal best matches by BLAST searches.

Q: Why did they find fewer than 1142 genes?

Q: Why did they visually inspect the results?

Q: Does their “reciprocal best match” test ensure they they are dealing with orthologues? How else can they provide evidence for orthology?

3. Identify regions of conserved synteny between Mmu16 and Hsa

- Regions of conserved synteny predicted by sequence similarity and by protein comparisons
- **Synteny based on sequence comparisons:**
 1. Syntenic anchors were located - regions with high (80%) similarity over short distances (~200bp or more).
 2. Average distance between anchors is 8kbp, but there are gaps as large as 707kbp in the mouse and 3.4Mbp in the human

Q: Why look at anchors – why not just look at known gene locations?

Q: Why are the anchors not uniformly distributed?

3. Identify regions of conserved synteny between Mmu16 and Hsa

3. 56% of anchors were in mouse genes - exons mostly
4. 44% in intergenic regions
5. Relatively density is independent of coding/noncoding - making the anchors an important marker of synteny (in addition to genes)

Human chr.	Mmu len.	Hsa len.	No. anchors	bad anch. (% incon.)	Orthologues
16	10,461	12,329	1,429	21 (1.5)	87
8	1,284	1,491	121	1 (0.8)	6
12	363	306	31	3 (9.7)	3
22	2,081	2,273	418	8 (1.9)	30
3q27-29	13,557	16,461	1,714	18 (1.0)	107
3q11.1-13.3	41,660	46,493	5,485	63 (1.1)	165
21	22,327	28,421	2,127	27 (1.3)	111

Q: Why were most of the anchors in genetic regions?

Q: Why were some of the anchors not in or near genes?

Q: Why is there not a clean correlation between the number of orthologues and the length of the syntenous regions?

3. Identify regions of conserved synteny between Mmu 16 and Hsa

- **Synteny based on gene comparisons:**
 1. Search for matches b/n human and mouse proteins with mutually best BLAST scores
 2. Syntenic blocks found - each block composed of dozens of genes (of the 731 genes)
 3. 99% of Mmu16 can be mapped to single, unique human chromosome segments
 4. Of the 731 genes, 509 paired with a homologue that is most likely orthologous
 5. For 164 genes, this assertion is weaker - low similarity, not best reciprocal match
 6. For 14 genes - no homologue could be identified

Q: Why predict synteny with gene comparisons?

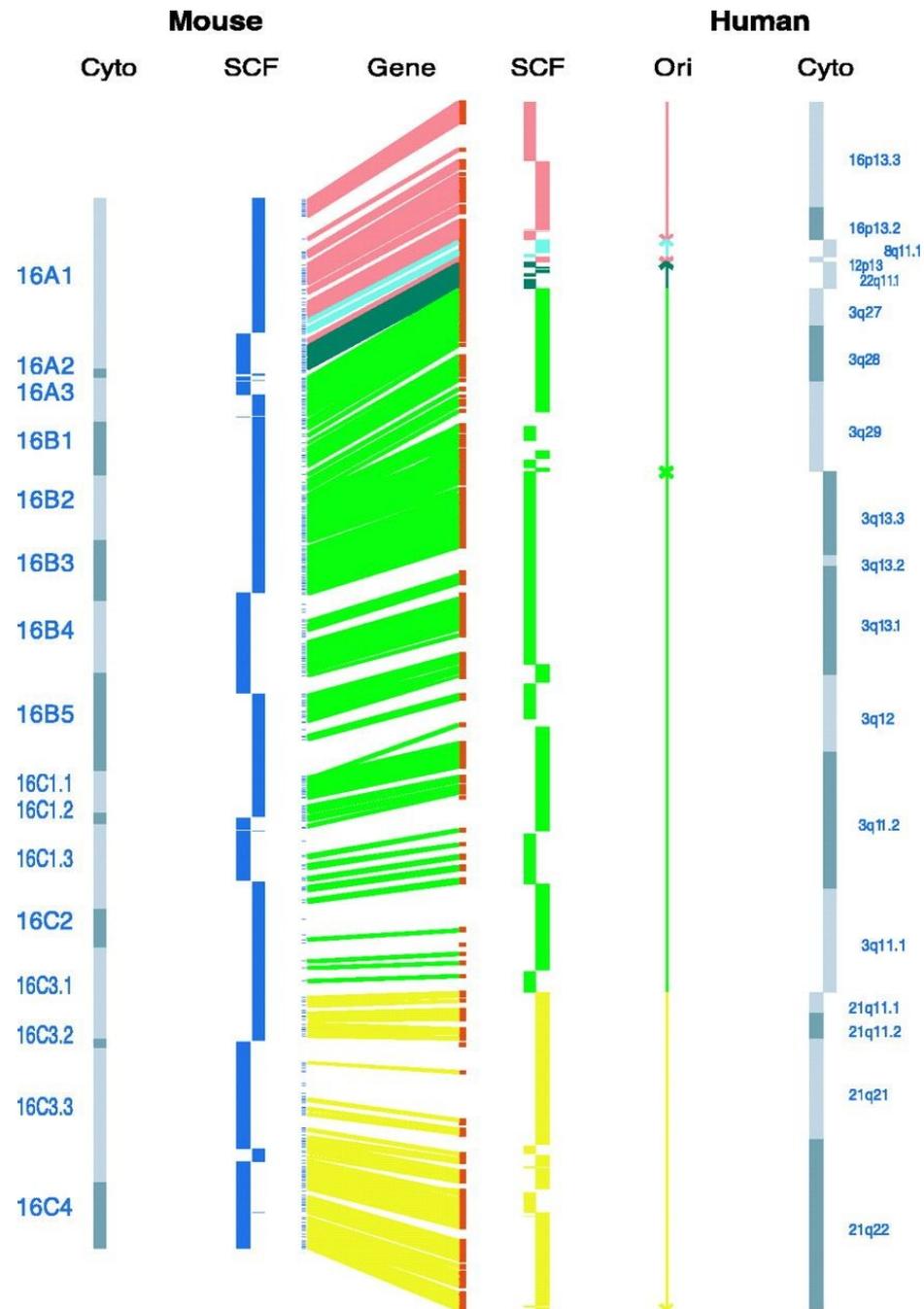
Q: Why do genes occur in “blocks”?

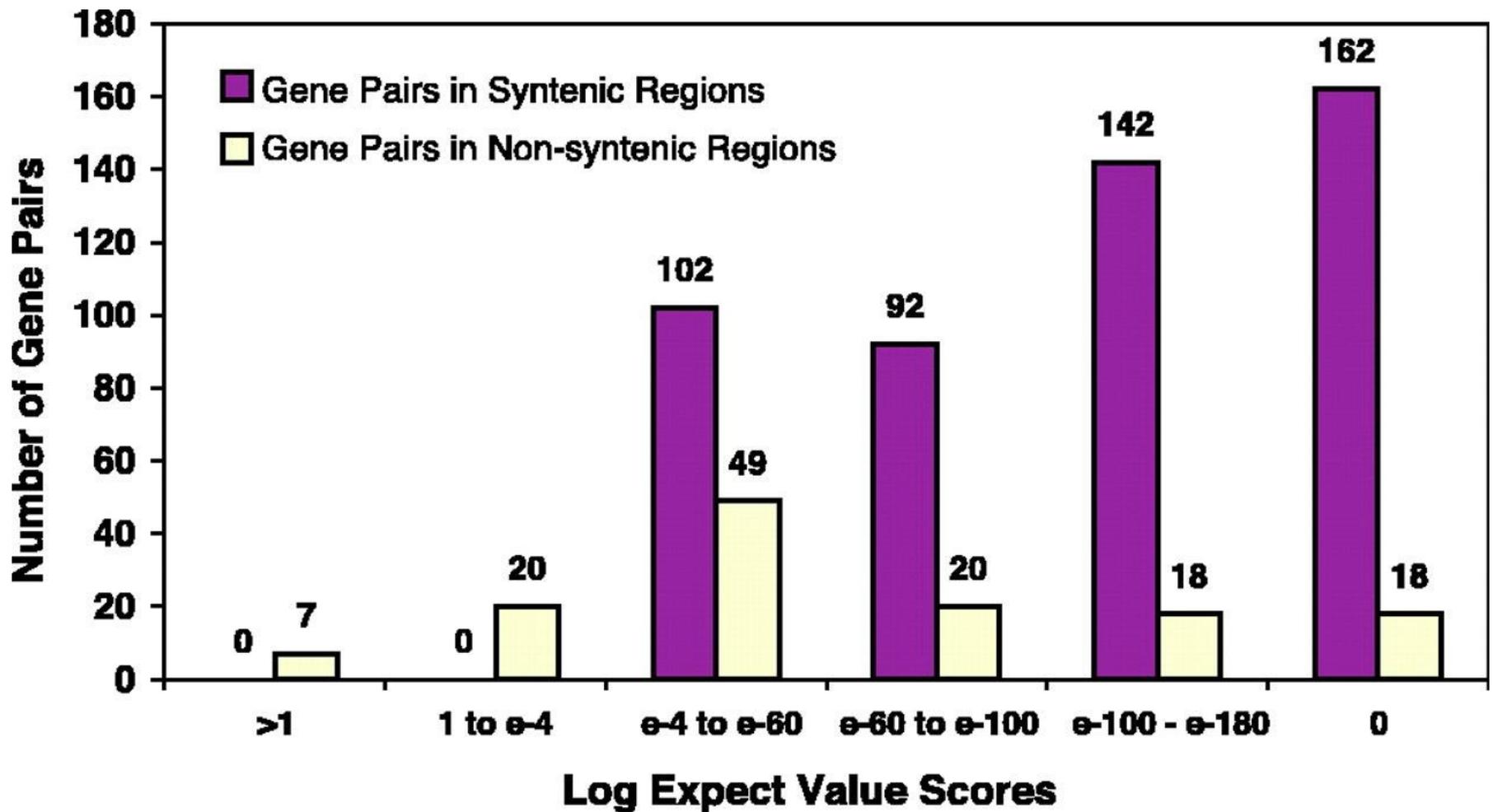
Q: What are the genes for which no homologue could be found?

- Large regions conserved (1/3 of Mmu16 on Hsa16 & Hsa21; the rest in 5 other regions)
- Content of genes in regions preserved
- Order of genes preserved (only a couple exceptions)
- 99% of anchors conserved in order and orientation
- All 509 putative orthologues are consistent in their location and order
- About 2% of Mmu genes are unique to mouse (relative to human) – no homologue found for 14 of 731 genes

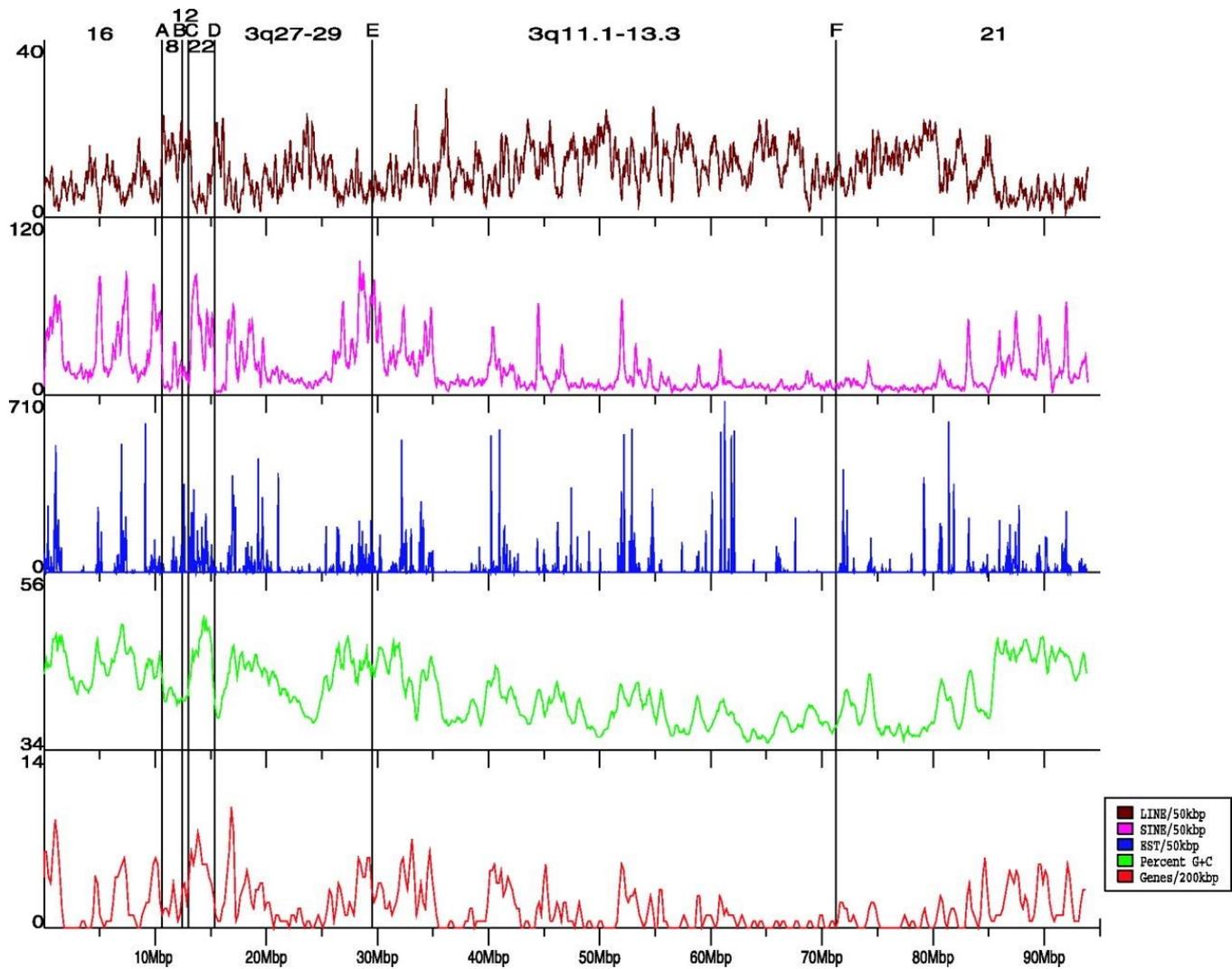
Q: Why is the order of genes and anchors preserved?

Q: If no homologue is found in human for those 14 genes, where should we look?





Q: Why do gene pairs in syntenic regions have more significant E scores?



Q: What general patterns can be seen?

Q: Why do some of the factors correlate w/ gene density?