Reviews in Computational Biology

5. Structuring & Outlining

James Smith

February 8th, 2012

Last week - Editing

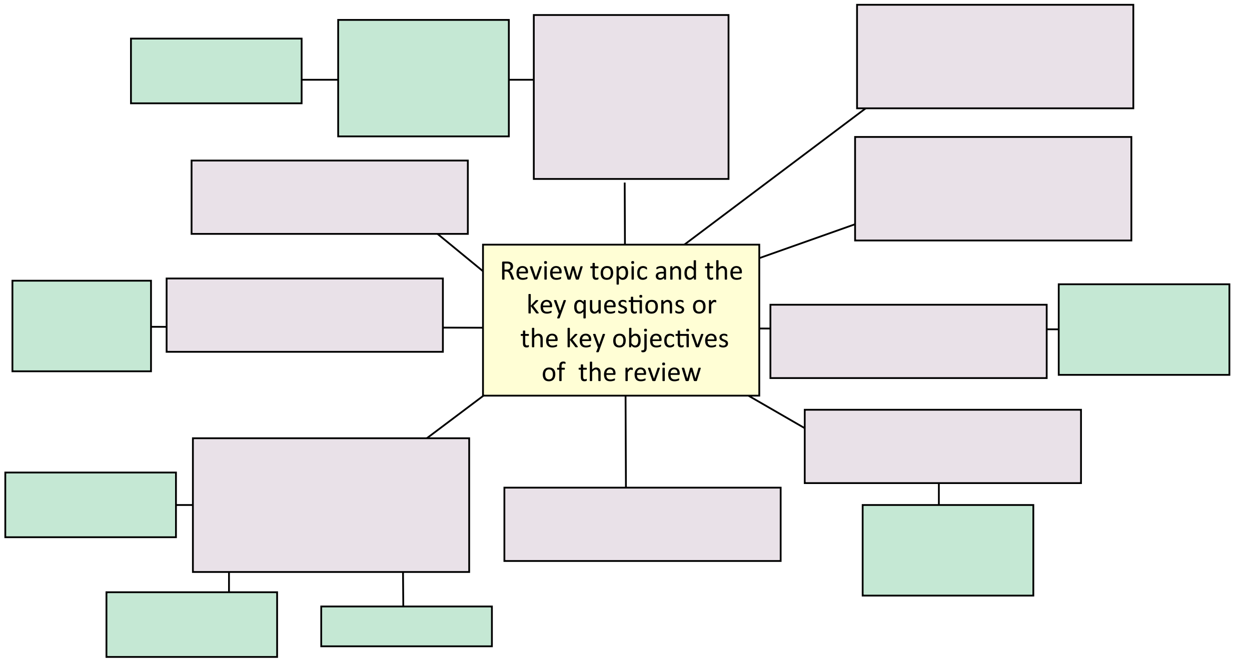
- Introduction, middle, conclusion
- Identifies (in)appropriate text
- Improves clarity

Structuring & Outlining

- Ideas/general concepts
- Navigates/signposts the narrative
- Not detail but a schema....

An example structure/outline

Review



Review topic and the key questions or the key objectives of the review

Grey - Sections of the narrative of the review

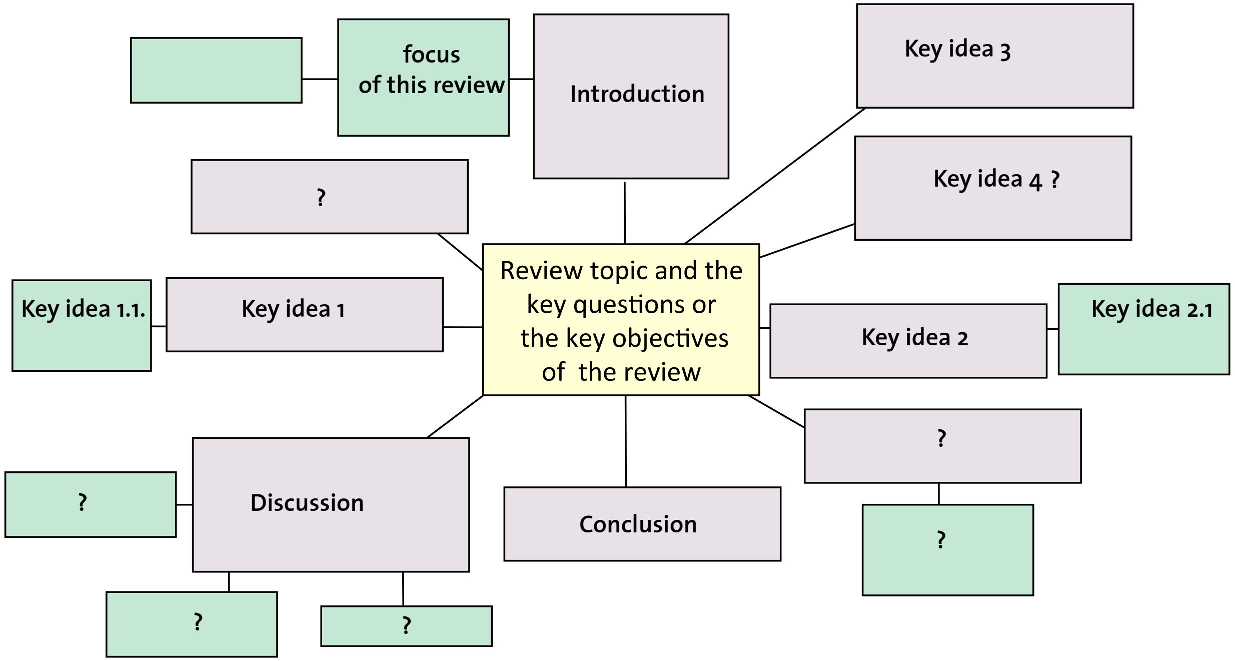
Green - Key points of importance - must stand out

Structuring & Outlining

- Does the reader notice?
- Does it keep the interest?
- Does it convey the arguments?

An example structure/outline

Review



focus of this review

Introduction

Key idea 3

Key idea 4 ?

?

Review topic and the key questions or the key objectives of the review

Key idea 1.1.

Key idea 1

Key idea 2

Key idea 2.1

?

Discussion

?

Conclusion

?

?

?

Exercise 1

- “Reconstructing a Structure”
- 2 teams of “Technical Editors”
  - 1. Groups paragraphs in sections and subsections with headings - Draw a schema of the structure
  - 2. Provide 2 alternative structures that might work - Draw a schema of the structure

Exercise 1 - Structuring and Outlining

In this exercise you are now a “Technical Editor” and the Editor assigns you to (help the author) structure/re-structure this accepted manuscript.

The Editor says that the peer-reviewers were very favorable, however the resounding conclusion from them was that its structure could be improved.

“This review treats a question that has broad appeal (how many genes are there in humans). It is a good read but the structure is quite unusual or awkward, with quite a bit of going back and forth.”

The Editor thinks you can help and she thanks you very much…

“Technical Editor” Team 1 will suggest appropriate headings/subheadings following paragraphs 1 to 25 and are not allowed to reorder the paragraphs – the author’s original structure.

“Technical Editor Team 2 will suggest appropriate headings/subheading but can reorder the paragraphs. The Author might be resistant so you have to justify why it’s important to move the paragraphs around.

Hint for both teams: Do the references help your structure define your sub-headings?

Exercise 2   - Based on your structuring/re- structuring, the Editor asks you to suggest a brief abstract.

Between a chicken and a number of human genes

1

Ever since the discovery of the genetic code, scientists have been trying to catalog all the genes in the human genome. Over the years, the best estimate of the number of human genes has grown steadily smaller, but we still do not have an accurate count. Here we review the history of efforts to establish the human gene count and present the current best estimates.

2

The first attempt to estimate the number of genes in the human genome appeared more than 45 years ago, while the genetic code was still being deciphered. Friedrich Vogel published his 'preliminary estimate' in 1964 [1], based on the number of amino acids in the alpha- and beta-chains of hemoglobin (141 and 146, respectively). Knowing that three nucleotides corresponded to each amino acid, he extrapolated to compute the molecular weight of the DNA comprising these genes. He then made several assumptions in order to produce his estimate: that these proteins were typical in size (they are actually smaller than average); that nucleotide sequences were uninterrupted on the chromosomes (introns were discovered more than 10 years later [2,3]); and that the entire genome was protein coding. All these assumptions were reasonable at the time, but later discoveries would reveal that none of them was correct. Vogel then used the molecular weight of the human haploid chromosomes to correctly calculate the genome size as 3 × 109 nucleotides, and dividing that by the size of a 'typical' gene, came up with an estimate of 6.7 million genes.

3

Even at the time, Vogel found this number 'disturbingly high', but no one suspected in 1964 that most human genes were interrupted by multiple introns, nor did anyone know that vast regions of the human genome would turn out to contain seemingly meaningless repetitive sequences. Since Vogel's initial attempt, many scientists have tried to estimate the number of genes in the human genome, using increasingly sophisticated molecular tools. Over the years, the number has gradually come down, in a process that has been humbling at times, as we realized that many other species - even plants - are predicted to have more genes than we do (Figure 1). An estimate of 100,000 genes appeared in the 1990 joint National Institutes of Health (NIH)/Department of Energy (DOE) report on the Human Genome Project [4]; this was apparently based on a very rough (and incorrect) calculation that typical human genes are 30,000 bases long, and that genes cover the entire 3-gigabase genome.

4

Many people, including many geneticists, expected that we would have a definitive gene count when the human genome was finally completed, and indeed one of the main surprises upon the initial publication of the human genome in February 2001 [5,6] was that the number had again dropped, quite precipitously. However, as we shall see, the publication of the human genome did not come anywhere close to producing a precise gene list or even a gene count, and in the years since the number has continued to fluctuate. As a result, even today's best estimates still have a large amount of uncertainty associated with them.

5

In order to count genes, we need to define what we mean by a 'gene', a term whose meaning has changed dramatically over the past century. For our discussion, we will restrict the definition of gene to a region of the genome that is transcribed into messenger RNA and translated into one or more proteins. When multiple proteins are translated from the same region due to alternative mRNA splicing, we will consider this collec- tion of alternative isoforms to be a single gene. In this respect, our definition of a gene is equivalent to what may also be called a chromosomal locus. We will exclude non- protein-coding RNA genes (such as microRNAs (miRNAs) and small nuclear RNAs (snRNAs)), in part because of the even greater uncertainty surrounding their numbers. In recent years, as a result of the dramatic breakthroughs in our understanding of RNA interference [7] and miRNAs [8], the number and variety of known RNA genes has grown dramatically, and we expect that it will be many more years before we have a clear picture of how many of these non-coding genes exist in the human genome.

6

With the advent of automated DNA sequencing, it became possible to use sequencing methods to estimate the number of human genes more accurately. The most promising approach, which was used by many groups in the 1990s, was to capture mRNA transcripts in a cell by making use of the polyadenylated (poly(A)) 3' ends. Using poly(T) sequences as primers, researchers could use reverse transcription-polymerase chain reaction (RT-PCR) to capture and sequence large numbers of expressed genes in a cell. At a time when the human genome project was just getting under way, these expressed sequence tags (ESTs) represented a shortcut to capturing the protein- coding genes in the genome [9]. In 1995, one of the first large-scale surveys of human genes [10] used this approach to construct

300 complementary DNA (cDNA) libraries from 37 distinct organs and tissues, and constructed 87,983 distinct sequences, many of them assembled from multiple overlapping ESTs. This result was consistent with the NIH/DOE estimate of 100,000 genes in the human genome [11].

7

In the mid-1990s, a series of papers produced estimates based on ESTs that generally agreed on a human gene count of 50,000 to 100,000 genes (Figure 2). In 1993, Antequera and Bird [12] estimated that the human genome contained 45,000 CpG islands. These are stretches of genomic DNA with a relatively high density of CG dinucleotides. Combining this with their report that 56% of sequenced genes at that time (1993) were associated with CpG islands, they calculated a total human gene count of 80,000. The following year, Fields *et al*. [13] relied primarily on ESTs to produce an estimate of 64,000 genes, although this estimate relied critically on an uncertain estimate of the ‘redundancy’ of EST sequence databases, which they guessed to be 50%.

8

These two estimates, 64,000 and 80,000, reduced the expected gene count somewhat, but even in 1994 there was little agreement on which number was closer to the truth [14]. In a study that unified physical maps, genetic maps, and the sequence data available at the time, Schuler *et al*. [15] reported in 1996 that the genome held 50,000 to 100,000 genes, although their mapping effort only captured 16,000.

9

In 2000, shortly before the human genome was published, several additional estimates appeared: Roest *et al*. [16] estimated 28,000 to 34,000 genes using alignments to pufferfish, and two new EST-based estimates reported 35,000 [17] and 57,000 [18] genes. This set the stage for the human genome paper, which was soon to appear.

10

To better understand the source of this continuing uncertainty about the gene count, it is instructive to mention a few of the most significant advances in computational gene prediction. (For a more compre- hensive review of gene structure prediction methods, the interested reader can consult several recent reviews [19-21].)

11

One of the oldest and most reliable ways to identify a gene in a newly sequenced genome is by locating a highly similar protein-coding sequence in another organism. Together with EST and cDNA alignments, gene finding by homology is the first step in all the major annotation pipelines. But even the most thorough EST sequencing projects fail to capture many exons and genes. The dis- covery of these genes is still dependent, at least in part, on *de novo* gene finders that only require information inherent in the DNA sequence itself.

12

Computational gene recognition began about 30 years ago, when it was observed that statistical analysis could detect differences between protein-coding and non-coding nucleotide sequences [22-24]. Early gene-predic- tion programs attempted to identify relatively few properties of genes, such as the signals around splice sites, and they made simplifying assumptions to make the problem more tractable [25]. With the development of gene-finding systems designed to predict any number of complete gene structures transcribed from either strand of the genome, automated methods made a significant step forward. The most successful framework for these systems was the generalized hidden Markov model (GHMM) approach. Thanks to their modularity and to their capability to model variable-length features, GHMMs are well suited to modeling the statistical properties of genes. Genscan [26] was one of the first of these, in 1997, and it was also the first *de novo* gene predictor to reach 80% exon-level accuracy on a human benchmark set. Despite its performance on coding exons, Genscan's gene-level accuracy (the proportion of genes for which it correctly predicts every exon) on the human genome was only about 10%. One reason for the low gene-level accuracy is that typical human genes contain 5 to 10 exons, and even at 80% accuracy per exon, the likelihood of getting all the exons correct for any particular gene is low.

13

Although later gene finders would improve on Genscan's results, the next real leap in accuracy came with the development of comparative gene finders. Comparative gene finders use patterns of conservation between two related species, such as human and mouse, to predict the location and structure of protein-coding genes. They can also use the GHMM framework. The biggest effect of using two genomes at once was to reduce the number of false-positive predictions: using human- mouse alignments, Twinscan [27], a dual-genome gene finder, predicted 25,600 human genes versus 45,000 predicted by Genscan [19].

14

Until 2007, GHMMs were the dominant framework for *de novo* gene finders, but this changed when conditional random fields (CRFs), a new class of discriminative models, were introduced as a means of using more than two genomes simultaneously. Unlike GHMMs, which are trained by maximum likelihood to generate sequences statistically similar to actual DNA sequences, CRFs are trained to discriminate between genomic elements of interest in order to maximize annotation accuracy. In addition, they are capable of utilizing external evidence and submodels that are not inherently probabilistic [28]. Through the use of 11 informant genomes, CONTRAST [29] predicted the exact exon-intron structure of 59% of known human protein-coding genes, compared to 25 to 35% from the best previous methods. This is a very strict measure of accuracy: if even one splice site from a multi- exon gene is incorrect, the entire gene is considered to be wrong. But also note that all *de novo* methods have a significant false-positive rate, predicting many exons (and genes) that do not appear to be genuine. Pseudogenes are one source of false predictions, although the precise reasons for high false positive rates have never been fully determined.

15

Despite a steady increase in accuracy over the years, *de novo* gene predictors are still not accurate enough to rely on for the definitive human gene list. Much greater gains in accuracy have been made through advances at the level of integrative evidence-based methods, such as those employed by JIGSAW [30]. By effectively combining multiple forms of evidence generated from a diverse set of sources, including gene finders, protein sequence alignments, EST and cDNA alignments, and splice-site predictions, JIGSAW's predictions are exactly correct for approximately 75% and partially correct for 97% of human genes [31]. Similar integrated methods are used to generate the gene lists at Ensembl [32] and the National Center for Biotechnological Information (NCBI), which uses the Gnomon system [33].

16

The release of the draft human genome sequence in 2001 revealed a much lower human gene count than expected [6,34]. The paper published by the public consortium estimated 30,000 to 40,000 protein-coding genes. This number was in rough agreement with the count in the private consortium's paper, which reported 26,588 protein-coding genes with 'strong' evidence, and an additional 12,000 computationally predicted genes with weaker evidence. Strong evidence included similarity to previously known proteins, homology to another mammal, and EST evidence. Weak genes were those with homology to mouse, but lack of other supporting evidence. After 3 years of detailed finishing work, a much more complete draft genome was published in 2004 [35], and along with this more complete sequence, the public consortium announced a new, much lower, estimate of human protein-coding genes, only 20,000 to 25,000. This low number - lower even than the model plant *Arabidopsis thaliana* - was surprising to scientists across a wide range of fields, who had expected that the number of genes to be a measure of organismal complexity. Furthermore, the imprecision of the estimate raised questions about the validity of many predicted genes [36].

17

Although the near-finished human genome sequence now covers 99% of the euchromatic (or gene-containing) genome at 99.999% accuracy, the exact number of human genes is still unknown. The two leading repositories of genome annotation, relied on by most researchers looking for genes, are the databases at Ensembl and NCBI. At present, Ensembl lists 22,619 human protein-coding genes, which is 286 higher than the 22,333 protein-coding genes in NCBI's RefSeq database [37]. This Ensembl total excludes 1,002 genes mapped onto alternative MHC regions in chromosome 6. The gene count from NCBI includes all protein-coding genes in RefSeq that either have been manually curated or that have supporting cDNA evidence, and that map onto the current human reference assembly (GRCh37). Another popular resource, the University of California at Santa Cruz (UCSC) genome browser [38], lists 21,814 'known' protein-coding genes [39]. The 'known' genes list was created by mapping human RefSeq mRNA sequences to the genome.

18

In an effort to identify a core set of human genes that are universally agreed upon, the collaborative consensus coding sequence project (CCDS) tracks identical protein annotations that are consistently represented at NCBI, Ensembl, and the UCSC Genome Browser [40]. As of January 2010, CCDS contained 18,173 human genes that are shared by all three browsers (counting alternative splice variants, where one gene is represented by two or more loci, it lists 23,739 protein-coding loci). Because CCDS takes an extremely conservative strategy, its gene list represents a lower bound on the total number of human genes. Indeed, in its original incarnation in 2005, it listed only 13,142 genes, and the total has steadily grown since then.

19

Currently, the average number of genes listed in the human gene catalogs appears to be somewhere around 22,500, with an uncertainty of around 2,000 genes. One recent report claims that this number is much too high: Clamp *et al.* [41] used a conservation-based method, relying on similarity to the mouse and dog genomes as well as other techniques, to reduce it to about 20,500 ‘valid’ protein-coding genes. They discarded as invalid genes that appeared to be retroposons, pseudogenes, and other miscellaneous artifacts, as well as ‘orphan’ DNA sequences. These orphans have many features of protein- coding genes, but are not conserved in other mammalian genomes, including those of chimpanzees and macaques. Because there were a relatively large number of orphans compared with the otherwise very small gene differences between humans and chimps, Clamp *et al*. rejected as implausible the alternative hypothesis that the orphans are human-specific genes.

20

Recently, the Mammalian Gene Collection (MGC), a multi-year effort to produce full-length cDNA clones for all human genes, reported the completion of its work [42]. This report describes 18,877 human protein-coding genes ‘with curated RefSeq transcripts’, of which MGC has produced clones for 17,421 (92%). The same report noted that recent efforts using comparative sequence data and computational gene finding, followed by confirmation with RT-PCR, had confirmed 563 distinct genes that were missing from the cDNA-based RefSeq and Vega collections [43] at the time. The MGC also excluded the transcripts of many single-exon genes and genes shorter than 100 amino acids, in order to avoid including pseudogenes, although their own report found that out of a set of 351 ‘likely’ single-exon genes, 198 (57%) were confirmed via RT-PCR [42]. Thus, although the 18,877 number is substantially lower than the total in Ensembl and RefSeq, at least some of the discrepancy is due to the conservative strategy used to identify protein- coding genes by the MGC.

21

Comparative genome analysis suggests that the numbers of protein-coding genes are not expected to differ very much from mammal to mammal [41]. When new genes arise in a species, most such cases are the result of duplications of previously existing genes, followed by neofunctionalization [44]. However, entirely novel genes must arise at some point, although the rate of gene ‘birth’ is not precisely known. Interestingly, a recent study provides the first evidence for the *de novo* origin of human protein-coding genes, which evolved from non-coding DNA after the divergence of humans and chimpanzees. In this study, Knowles and McLysaght [45] identified three entirely novel genes, all of which have strong mRNA expression evidence supporting transcrip- tion, and peptide matches from proteomics databases supporting translation. The orthologous DNA sequence exists in other primate genomes - chimp, macaque, gorilla, gibbon, and orangutan - but in the other primates, the DNA has disabling mutations that disrupt the reading frame. By extrapolating their findings to the whole human genome, the authors estimate that 18 genes are likely to have arisen *de novo* in humans since our divergence from chimps.

22

In addition to the ongoing uncertainty about the precise number of protein-coding genes, recent evidence has emerged that makes it clear that different humans have slightly different individual gene sets. A major source of such differences is variation in the number of segmental duplications scattered across the genome. Sebat *et al*. [46] looked at 20 individuals for copy-number polymor- phisms, and found 70 different genes included in regions with variable copy numbers. Iafrate *et al*. [47] found more than 100 gene-containing regions that varied in copy number among individuals. Most recently, Alkan *et al*. [48] estimated, on the basis of three sequenced human genomes, that gene counts vary by 73 to 87 genes between any two individuals.

23

In another recent finding, Li *et al*. [49] sequenced and assembled two human genomes, one from Africa and one from Asia, and compared them with the reference human genome at NCBI. They identified around 5 Mb of novel sequence in each of the new genomes, and they estimate that the human 'pangenome', which would include all the DNA of every individual human, should have up to 40 Mb of sequence additional to the reference genome, including an unknown number of genes. This additional potential sequence is 1.3% of the genome, which suggests that the eventual gene count might grow by about that same amount.

24

We aligned all human genes from NCBI's RefSeq database to the Ensembl gene set in an attempt to explain the differences, but although the total counts differ by less than 300, there are several thousand genes in each set that do not map cleanly onto the other, many of them representing genes of unknown function. Our personal best guess for the total number of human genes is 22,333, which corresponds to the current gene total at NCBI. We prefer this to the slightly higher Ensembl gene count both because the NCBI annotation is slightly more conservative, and because recent compelling arguments support an even lower gene total [41,42]. This number could easily shrink or grow by 1,000 genes in the near future. However, recent analyses make it clear that even if we agree on a complete list of human genes, any particular individual might be missing some of the genes in that list. The genome sequence is complete enough now (although it is not yet finished) that few new genes are likely to be discovered in the gaps, but it seems likely that more genes remain to be discovered by sequencing more individuals.

Additional discoveries are likely to make our best estimates for this basic fact about the human genome continue to move up and down for many years to come.

Acknowledgements

References

1. Vogel F: A preliminary estimate of the number of human genes. Nature 1964, 201:847.

2. Chow LT, Gelinas RE, Broker TR, Roberts RJ: An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. Cell 1977, 12:1-8.

3. Berget SM, Moore C, Sharp PA: Spliced segments at the 5’ terminus of adenovirus late mRNA. Proc Natl Acad Sci USA 1977, 74:3171-3175.

4. US Department of Health and Human Services, US Department of Energy: Understanding our Genetic Inheritance, The U.S. Human Genome Project: The First Five Years, Fiscal Years 1991-1995.
[http://www.ornl.gov/sci/techresources/Human_Genome/project/5yrplan/summary.shtml]

5. The International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: The sequence of the human genome. Science 2001, 291:1304-1351.

7. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 1998, 391:806-811.

8. Lee RC, Feinbaum RL, Ambros V: The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993, 75:843-854.

9. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC: Complementary DNA sequencing: expressed sequence tags and human genome project. Science 1991, 252:1651-1656.

10. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH,

Kirkness EF, Weinstock KG, Gocayne JD, White O, Sutton G, Blake JA, Brandon RC, Chiu MW, Clayton RA, Cline RT, Cotton MD, Earle-Hughes J, Fine LD, FitzGerald LM, FitzHugh WM, Fritchman JL, Geoghagen NSM, Glodek A, Gnehm CL, Hanna MC, Hedblom E, Hinkle PS Jr, Kelley JM, Klimek KM, et al.:Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 1995, 377:3-174.

11. Goodfellow P: A big book of the human genome. Complementary endeavours. Nature 1995, 377:285-286.

12. Antequera F, Bird A: Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA 1993, 90:11995-11999.

13. Fields C, Adams MD, White O, Venter JC: How many genes in the human genome? Nat Genet 1994, 7:345-346.

14. Antequera F, Bird A: Predicting the total number of human genes. Nat Genet 1994, 8:114.

15. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, et al.: A gene map of the human genome. Science 1996, 274:540-546.

16. Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quétier F, Saurin W, Weissenbach J: Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet 2000, 25:235-238.

17. Ewing B, Green P: Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet 2000, 25:232-234.

18. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet 2000, 25:239-240.

19. Brent MR: Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 2008, 9:62-73.

20. Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigo R: Identifying protein-coding genes in genomic sequences. Genome Biol 2009, 10:201.

21. Jones SJ: Prediction of genomic functional elements. Annu Rev Genomics Hum Genet 2006, 7:315-338.

22. Erickson JM, Altman GG: A search for patterns in the nucleotide sequence of the MS2 genome. J Math Biol 1979, 7:219-230.

23. Shulman MJ, Steinberg CM, Westmoreland N: The coding function of nucleotide sequences can be discerned by statistical analysis. J Theor Biol 1981, 88:409-420.

24. Fickett JW: Recognition of protein coding regions in DNA sequences. Nucleic Acids Res 1982, 10:5303-5318.

25. Claverie JM: Computational methods for the identification of genes in vertebrate genomic sequences. Hum Mol Genet 1997, 6:1735-1744. 26. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997, 268:78-94.

27. Korf I, Flicek P, Duan D, Brent MR: Integrating genomic homology into gene structure prediction. Bioinformatics 2001, 17 Suppl 1:S140-S148. 28. Majoros H: Methods for Computational Gene Prediction. Cambridge: Cambridge University Press; 2007.

29. Gross SS, Do CB, Sirota M, Batzoglou S: CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol 2007, 8:R269.

30. Allen JE, Salzberg SL: JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 2005, 21:3596-3603.

31. Allen JE, Majoros WH, Pertea M, Salzberg SL: JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODEregions. Genome Biol 2006, 7 Suppl 1:S9.

32. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf̈ S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kaḧäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, et al: Ensembl’s 10th year. Nucleic Acids Res 2010, 38(Database issue):D557-D562.

33. NCBI Gnomon [http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml] 34. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

35. ENCODE Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004, 306:636-640.

36. Stein LD: Human genome: end of the beginning. Nature 2004, 431:915-916.

37. Pruitt KD, Tatusova T, Klimke W, Maglott DR: NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 2009, 37(Database issue):D32-D36.

38. Karolchik D, Hinrichs AS, Kent WJ: The UCSC Genome Browser. Curr Protoc Bioinformatics 2009, Chapter 1:Unit 1.4.

39. UCSC Genome Table Browser [http://genome.ucsc.edu/cgi-bin/hgTables]

40. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, et al.: The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 2009, 19:1316-1323.

41. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA 2007, 104:19428-19433.

42. MGC Project Team: The completion of the Mammalian Gene Collection (MGC). Genome Res 2009, 19:2324-2333.

43. Siepel A, Diekhans M, Brejová B, Langton L, Stevens M, Comstock CL, Davis C, Ewing B, Oommen S, Lau C, Yu HC, Li J, Roe BA, Green P, Gerhard DS, Temple G, Haussler D, Brent MR: Targeted discovery of novel human exons by comparative genomics. Genome Res 2007, 17:1763-1773.

44. Long M, Betran E, Thornton K, Wang W: The origin of new genes: glimpses from the young and old. Nat Rev Genet 2003, 4:865-875.

45. Knowles DG, McLysaght A: Recent de novo origin of human protein-coding genes. Genome Res 2009, 19:1752-1759.

46. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M: Large-scale copy number polymorphism in the human genome. Science 2004, 305:525-528.

47. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet 2004, 36:949-951.

48. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE: Personalized copy number and segmental duplication maps using next- generation sequencing. Nat Genet 2009, 41:1061-1067.

49. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, Zhou G, Zhu X, Wu H, Qin J, Jin X, Li D, Cao H, Hu X, Blanche H, Cann H, Zhang X, Li S, Bolund L, Kristiansen K, Yang H, Wang J, Wang J: Building the sequence map of the human pan-genome. Nat Biotechnol 2010, 28:57-63.

50. International Chicken Genome Sequencing Consortium: Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004, 432:695-716.

51. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 2007, 449:463-467.

① INTRO: NO OF GENES IN HUMANS

1.

5. → ?

② EXPERIMENTAL METHODS

(2. 3. 4.) 6. 7. 8. 9. 17?

③ COMPUTATIONAL METHODS

10.

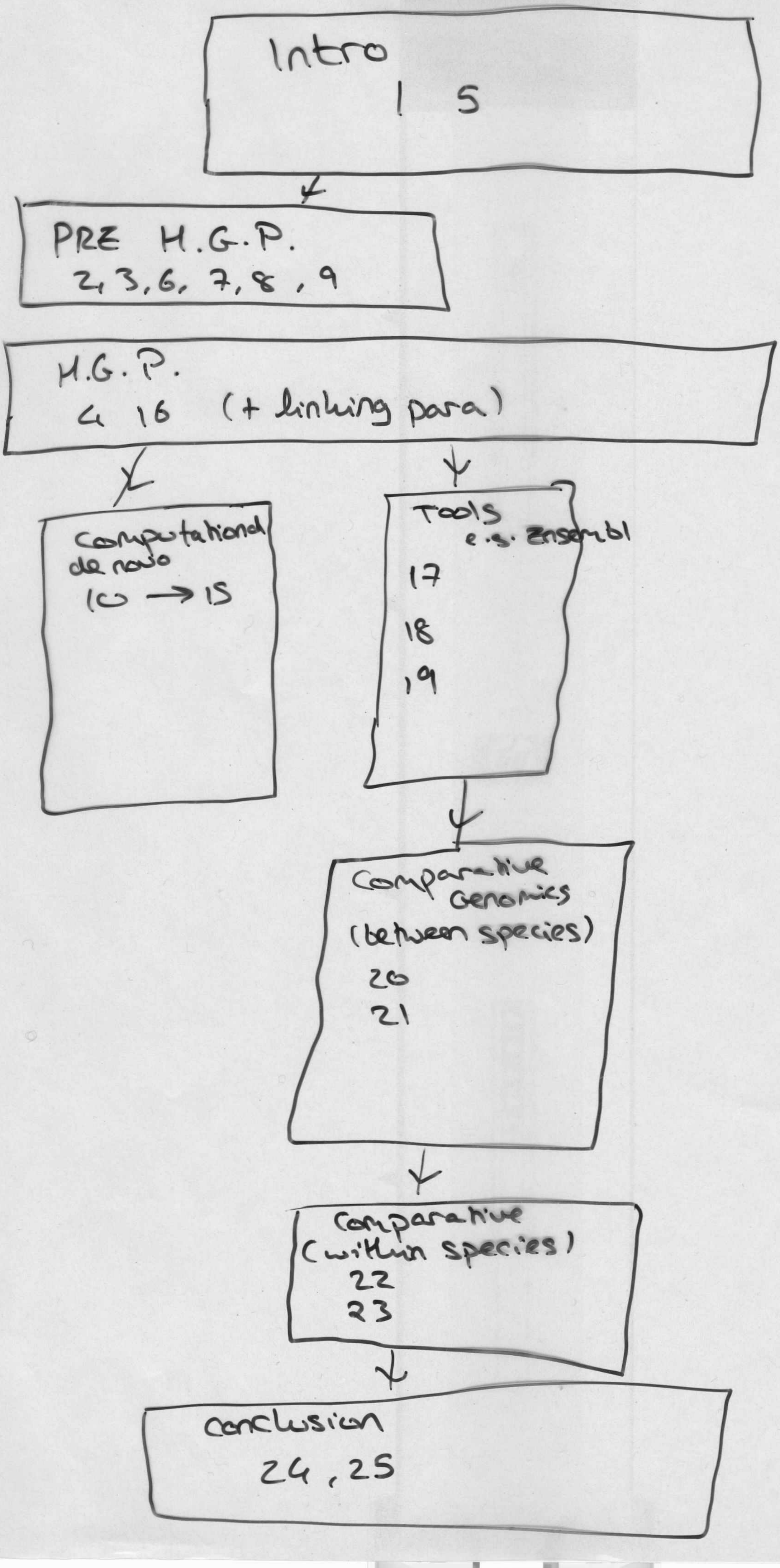
④ DATABASES („RESULTS") (SUB-PAR. OF 2?)

16,? 17, 16, 19

⑤ COMPARISON TO OTHER SPECIES; GENE NO. VARIATION BETWEEN INDIVIDUALS („DISCUSSION")

20, 21, 22, 23, 24 → 5a) COMP. GENOMICS

→ 5b) POP. GENOMICS


Intro
1 5
PRE H.G.P.
2,3,6, 7,8 , 9
H.G.P.
4 16 (+ linking para)
computational
de novo
10 → 15
Tools
e.g. Ensembl
17
18
19
Comparative
Genomics
(between species)
20
21
Comparative
(within species)
22
23
conclusion
24 , 25

Discussion

- What are the shortcomings?

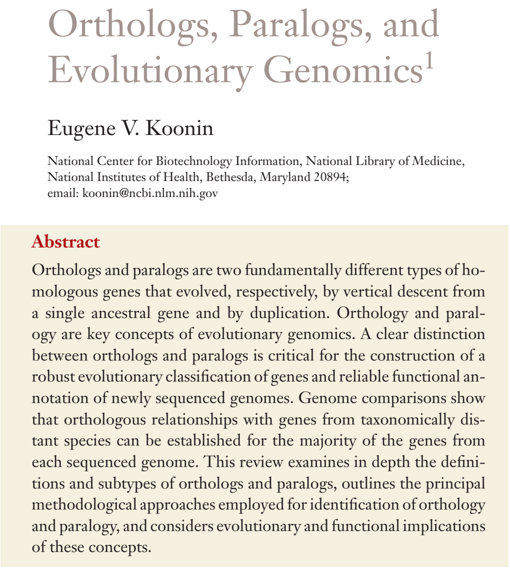
For your own reviews

- 1) Introductory paragraph or sentence(s)
- 2) 3 Sections (ideas) for your manuscript
- 3) Conclusion paragraph or sentence(s)

Reviews with clear structure & outline

- The Annual Reviews Series
- Structure is presented as a Table of Contents

The Abstract reflects the Table of Contents



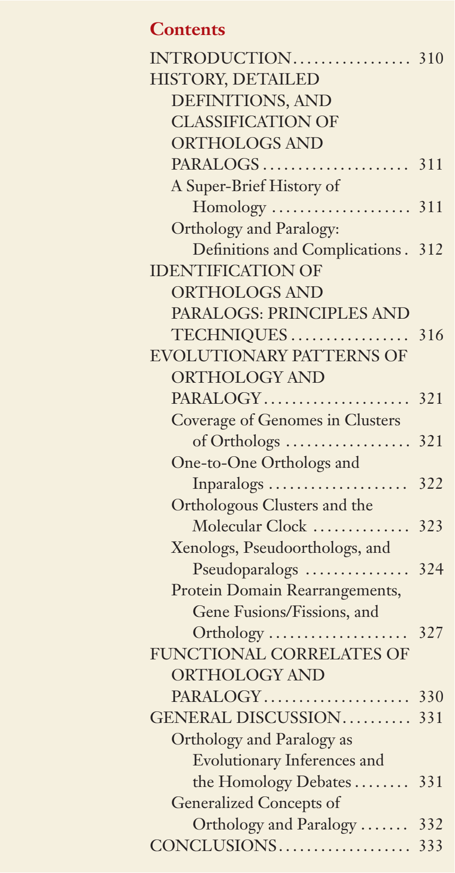
Orthologs, Paralogs, and Evolutionary Genomics[1]

Eugene V. Koonin

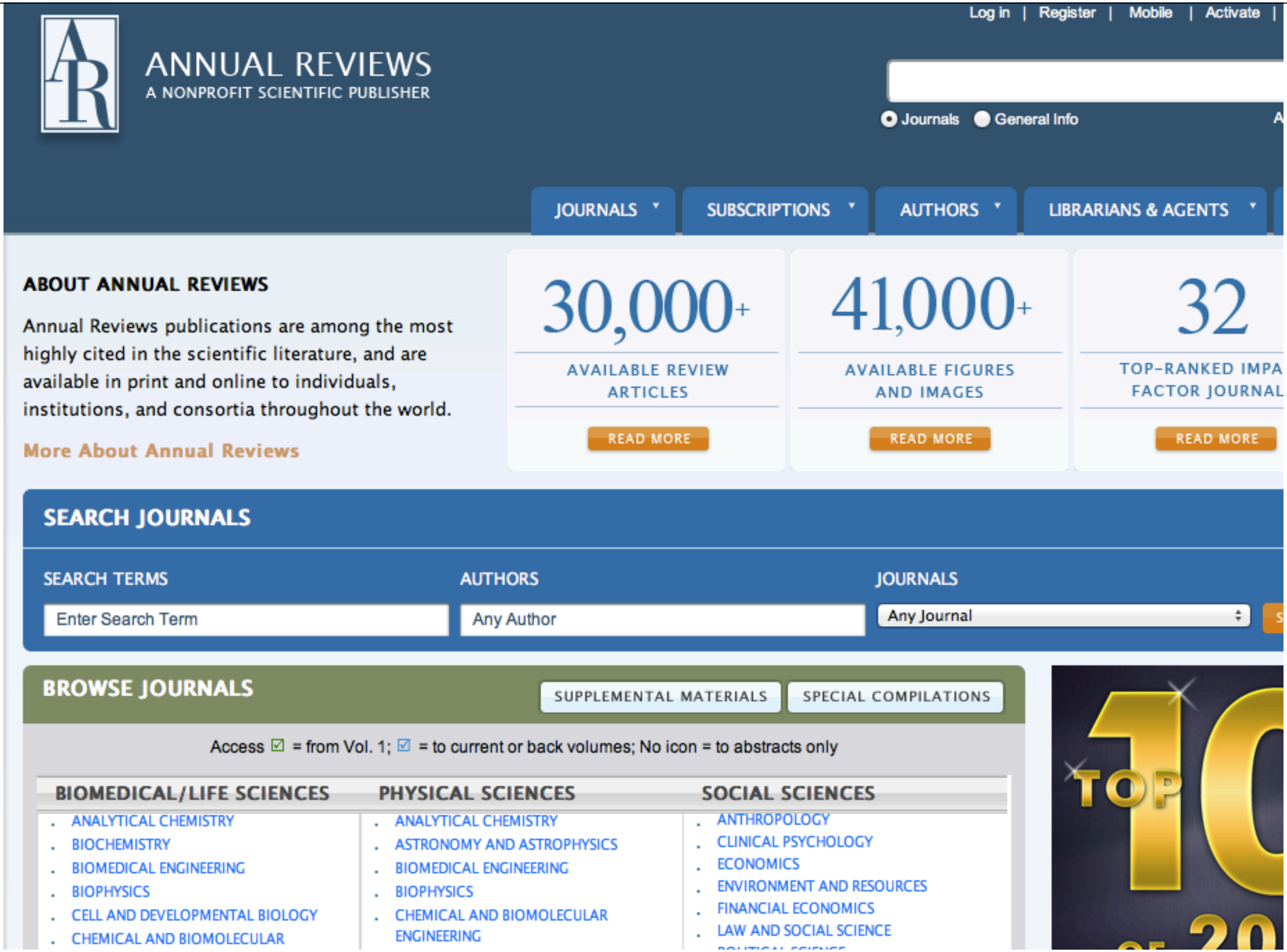
National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, Maryland 20894;
email: koonin@ncbi.nlm.nih.gov

Abstract

Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Orthology and paralogy are key concepts of evolutionary genomics. A clear distinction between orthologs and paralogs is critical for the construction of a robust evolutionary classification of genes and reliable functional annotation of newly sequenced genomes. Genome comparisons show that orthologous relationships with genes from taxonomically distant species can be established for the majority of the genes from each sequenced genome. This review examines in depth the definitions and subtypes of orthologs and paralogs, outlines the principal methodological approaches employed for identification of orthology and paralogy, and considers evolutionary and functional implications of these concepts.



Contents



Log in | Register | Mobile | Activate |

ANNUAL REVIEWS
A NONPROFIT SCIENTIFIC PUBLISHER

● Journals ○ General Info

JOURNALS ▾ SUBSCRIPTIONS ▾ AUTHORS ▾ LIBRARIANS & AGENTS ▾

ABOUT ANNUAL REVIEWS

Annual Reviews publications are among the most highly cited in the scientific literature, and are available in print and online to individuals, institutions, and consortia throughout the world.

More About Annual Reviews

30,000+
AVAILABLE REVIEW ARTICLES
READ MORE

41,000+
AVAILABLE FIGURES AND IMAGES
READ MORE

32
TOP-RANKED IMPA
FACTOR JOURNAL
READ MORE

SEARCH JOURNALS

SEARCH TERMS: Enter Search Term

AUTHORS: Any Author

JOURNALS: Any Journal

BROWSE JOURNALS

SUPPLEMENTAL MATERIALS | SPECIAL COMPILATIONS

Access ☑ = from Vol. 1; ☑ = to current or back volumes; No icon = to abstracts only

BIOMEDICAL/LIFE SCIENCES

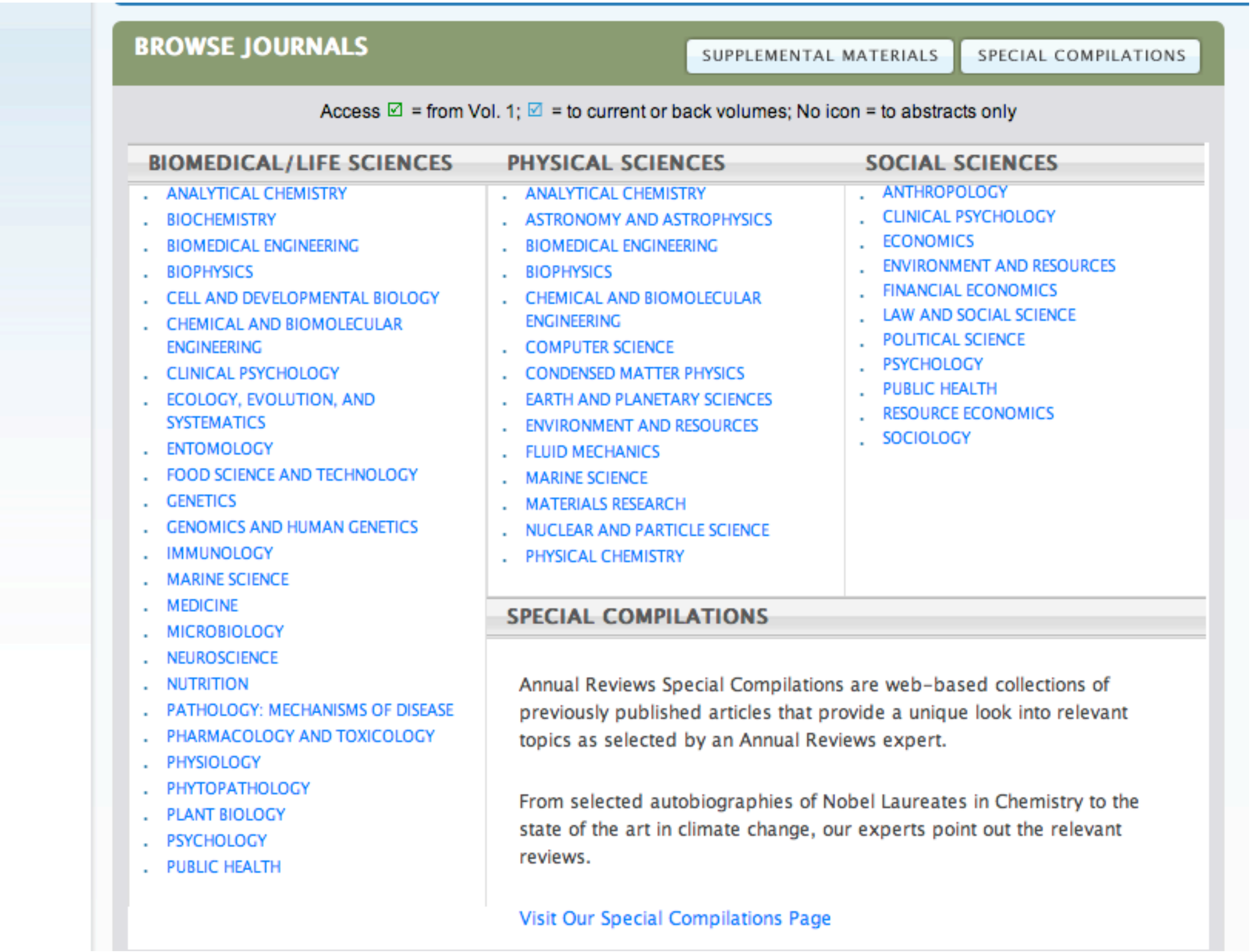
- ANALYTICAL CHEMISTRY
- BIOCHEMISTRY
- BIOMEDICAL ENGINEERING
- BIOPHYSICS
- CELL AND DEVELOPMENTAL BIOLOGY
- CHEMICAL AND BIOMOLECULAR

PHYSICAL SCIENCES

- ANALYTICAL CHEMISTRY
- ASTRONOMY AND ASTROPHYSICS
- BIOMEDICAL ENGINEERING
- BIOPHYSICS
- CHEMICAL AND BIOMOLECULAR ENGINEERING

SOCIAL SCIENCES

- ANTHROPOLOGY
- CLINICAL PSYCHOLOGY
- ECONOMICS
- ENVIRONMENT AND RESOURCES
- FINANCIAL ECONOMICS
- LAW AND SOCIAL SCIENCE



BROWSE JOURNALS

Access ☑ = from Vol. 1; ☑ = to current or back volumes; No icon = to abstracts only

BIOMEDICAL/LIFE SCIENCES

- ANALYTICAL CHEMISTRY
- BIOCHEMISTRY
- BIOMEDICAL ENGINEERING
- BIOPHYSICS
- CELL AND DEVELOPMENTAL BIOLOGY
- CHEMICAL AND BIOMOLECULAR ENGINEERING
- CLINICAL PSYCHOLOGY
- ECOLOGY, EVOLUTION, AND SYSTEMATICS
- ENTOMOLOGY
- FOOD SCIENCE AND TECHNOLOGY
- GENETICS
- GENOMICS AND HUMAN GENETICS
- IMMUNOLOGY
- MARINE SCIENCE
- MEDICINE
- MICROBIOLOGY
- NEUROSCIENCE
- NUTRITION
- PATHOLOGY: MECHANISMS OF DISEASE
- PHARMACOLOGY AND TOXICOLOGY
- PHYSIOLOGY
- PHYTOPATHOLOGY
- PLANT BIOLOGY
- PSYCHOLOGY
- PUBLIC HEALTH

PHYSICAL SCIENCES

- ANALYTICAL CHEMISTRY
- ASTRONOMY AND ASTROPHYSICS
- BIOMEDICAL ENGINEERING
- BIOPHYSICS
- CHEMICAL AND BIOMOLECULAR ENGINEERING
- COMPUTER SCIENCE
- CONDENSED MATTER PHYSICS
- EARTH AND PLANETARY SCIENCES
- ENVIRONMENT AND RESOURCES
- FLUID MECHANICS
- MARINE SCIENCE
- MATERIALS RESEARCH
- NUCLEAR AND PARTICLE SCIENCE
- PHYSICAL CHEMISTRY

SOCIAL SCIENCES

- ANTHROPOLOGY
- CLINICAL PSYCHOLOGY
- ECONOMICS
- ENVIRONMENT AND RESOURCES
- FINANCIAL ECONOMICS
- LAW AND SOCIAL SCIENCE
- POLITICAL SCIENCE
- PSYCHOLOGY
- PUBLIC HEALTH
- RESOURCE ECONOMICS
- SOCIOLOGY

SPECIAL COMPILATIONS

Annual Reviews Special Compilations are web-based collections of previously published articles that provide a unique look into relevant topics as selected by an Annual Reviews expert.

From selected autobiographies of Nobel Laureates in Chemistry to the state of the art in climate change, our experts point out the relevant reviews.

Visit Our Special Compilations Page