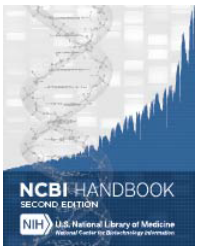
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Maglott D, Barrett T, Murphy T, et al. Genes and Gene Expression. 2013 Nov 7. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Genes and Gene Expression

Donna Maglott, PhD,[1] Tanya Barrett, PhD,[1] Terence Murphy, PhD,[1] Michael Feolo, PhD,[1] Lukas Wagner, PhD,[1] and Richa Agarwala, PhD[1]

Created: November 7, 2013.

Scope

Gene Overview

NCBI maintains information about genes primarily in two contexts. One context is defined by public sequence information, such as annotation of RefSeqs (see RefSeq chapter) or linking with records in the International Nucleotide Sequence Database Consortium or INSDC (see Genome Reference Consortium chapter). Connection of sequence information to a GeneID or a UniGene cluster identifier is critical to any analysis of gene expression. The second context for defining a gene is by mapped phenotype. GeneIDs are not assigned to mapped loci for all taxa, but when they are, the expectation is that the genes will eventually be connected to sequence as the molecular basis for the phenotype is defined.

Once an identifier is assigned to the concept of a gene, multiple databases connect information to that concept. Within NCBI, these databases include Gene, for primary data about the gene and portals to information about its expression, products, homologs, and phenotypes; BioSystems for pathways involving its products; GEO (see GEO chapter) and UniGene for information about expression; Bookshelf, PubMed and PubMedCentral for publications; dbGaP, PheGenI, MedGen and OMIM for phenotypes; HomoloGene for homology; dbSNP, dbVar, and ClinVar for variation; and Taxonomy for information about the organism. In other words, there are many resources at NCBI that maintain information abut genes, but this section focuses on these:

- Gene
- HomoloGene
- UniGene

Expression Overview

Several resources at NCBI maintain primary data about the tissues, health states, and developmental stage or age in which genes are expressed, or the sequence variation that affects their expression. The data archives reflect the primary methods by which these data have been generated, starting with sampling cDNA sequences from non-normalized libraries in UniGene, through arrary or RNAseq based approaches in GEO, to association data in GTex. These resources also maintain tools to analyze the datasets, which can be quite large.

Author Affiliation: 1 NCBI.

History

Genes

Representation of genes as objects with stable identifiers began at NCBI in 1995 with the clustering of 3' untranslated regions from GenBank release 88 into gene-specific sets as UniGene. (1) . When the RefSeq project got underway in 1998, a collaboration developed among the human nomenclature committee (now HGNC), OMIM, and the RefSeq team to aggregate gene-specific information for tracking. This developed into LocusLink (2) which evolved into Gene in 2004(3). In the late 1990's and into the early 2000's, most eukaryotic genes identified by sequence information were characterized by sequences assumed to represent gene expression, namely cDNAs. Now, however, many genes are first identified computationally rather than by direct sequence evidence, namely by gene prediction software that may use direct experimental results, but may also calculate which regions of a genomic sequence are likely to be a gene based on comparison to related species or analysis of predicted proteins.

HomoloGene

Examination of a gene's function is facilitated by evaluation across multiple species, HomoloGene (see HomoloGene chapter), launched as a distinct resource in 2000, is designed to facilate these analyses by grouping genes according to homology, providing tools for comparisons, and aggregating data for these homology groups.

Gene Expression

Methods of identifying regions of the genome that are transcribed have changed over the years. Large-scale cDNA projects such as I.M.A.G.E. (4) and the Mammalian Gene Collection (MGC) (5) determined what sequences were expressed based on cDNA cloning and sequencing of those clones. Given those sequences, array-based techniques were used to compare expression of sequences under different experimental conditions. Now, with the power of RNAseq, expression can be analyzed qualitatively and quantitatively without requiring a cloning step.

Sequence variation that affects gene expression is also being evaluated (6). PheGenI provides a window to these data.

Data Model

When a gene is defined by sequence, map location, or a nomenclature group, it is assigned a stable identifier for tracking, the GeneID. The connection between the GeneID and nucleotide or protein sequence is used by many of NCBI's databases to represent their data in the context of a gene. Thus you will see links to Gene, or GeneIDs annotated, in numerous sites at NCBI. HomoloGene, by grouping genes computed or curated to be homologs, is one of those sites.

Dataflow

Establishment of records in Gene, and evaluation of sequence expression, occur independently. In other words, there are many sequences in GEO, Nucleotide, Protein, or UniGene that do not cross reference a record in Gene. For genomes annotated by NCBI (see Eukryotic and Prokaryotic genome annotation chapters) connections between sequence and GeneIDs are evalutated with each Annotation Release, thus narrowing some of the gaps.

References

1. Boguski MS, Schuler GD. ESTablishing a human transcript map. Nat Genet. 1995 Aug;10(4):369–71. PubMed PMID: 7670480.

2. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res. 2000 Jan 1;28(1):126–8. PubMed PMID: 10592200.
3. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D54–8. PubMed PMID: 15608257.
4. Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics. 1996 Apr 1;33(1):151–2. PubMed PMID: 8617505.
5. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnerch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J; MGC Project Team. The status,quality, and expansion of the NIH full-length cDNA project: the Mammalian GeneCollection (MGC). Genome Res. 2004 Oct;14(10B):2121–7. Erratum in: Genome Res.2006 Jun;16(6):804. Morrin, Ryan [corrected to Morin, Ryan]. PubMed PMID: 15489334.
6. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5. PubMed PMID: 23715323.