Analysis Tools for Connectomes

Stephen Plaza, Charlotte Weaver, Arjun Bharioke, Louis K. Scheffer
Howard Hughes Medical Institute

(Dated: February 18, 2015)

Obtaining connectomes is only the first step on the road to biological insight. By themselves, connectomes are relatively abstract directed graphs, and are not easily interpretable. We need tools to help biologists understand connectomes and integrate them with related information.

INTRODUCTION

Obtaining connectomes is only the first step - the next is obtaining insight into biological processes and circuits. By themselves, connectomes are merely a labelled directed graph, and are not easily interpretable. We need tools that present the information of connectomes in a manner easier to understand, to help biologists take advantage of connectomes and integrate them with related information.

A strong analogy exists with respect to the various genomes that have been acquired and the tools to interpret them. The genome itself is just a long string of the letters C, T, A, and G. The proteins and genes these encode are far more interesing to most biologists, so almost all interaction and searches that refer to genomes takes place using analysis software explicitly designed to help humans find biological insight in this string of characters.

PREVIOUS WORK

Most previous work on neuron-level connectomes has concentrated on extracting specific circuits, such as the basic motion detection circuit in *Drosophila*[1]. Studies such as these were designed to answer specific questions, and used analyses that were hand-coded and limited to the specific problem at hand. Nevertheless, these analyses are the intellectual ancestors of the analyses described and proposed here.

Although there has been relatively little work analyzing generic neuron level connectomes (not surprising since until recently very few even existed), considerable work on analyzing the higher level connectomes obtained from FMRI and similar methods[2][3][4][5] and gene expression networks[6].

There has been considerable work involving display of general graphs[7] and display of other graphs such as social networks[8].

At its heart, a connectome is just a directed graph. Since graphs are useful representations in many science and engineering tasks, there has been considerable research into specific tasks on graphs, such as partitioning[9][10] [11], clustering[12][13][14], finding cliques[15], finding patterns[16], finding small motifs[17] and so on. Only some of these techniques have been applied to connectomes, and it is not clear which, if any, can provide useful answers to practical biology problems.

One challenge with connectomes is that the connectomes are “fuzzy”, meaning every instance of a common sub-graph is slightly different. This means that some well-known graph and subgraph matching algorithms (such as [18]), particularly those based on graph invariants[19], may not work well when applied to connectomes. Conversely, algorithms designed to cope with errors, such as [20], are more likely to be applicable.

NAMING AND COORDINATES

Image analysis, and most reconstructions, work in orthogonal XYZ coordinates. But biologists think in terms of named organs, each with their own (non-Euclidean) coordinate system. Often this is first the name of the organ (such as medulla in the fruit fly), then location in two axes (dorsal-ventral and anterior-posterior), then a layer within the organ (M1 to M10 in the medulla). These layers merge into each other and hence are not mutually exclusive. An example is shown in Fig. 1.

We have created a tool for defining named regions, which we call ROIs (for Regions of Interest)[22]. ROIs can be of arbitrary shape, though they are typically simply connected. The user defines them by drawing contours on a (possibly sparse) set of layers. If there are unmarked layers between user-defined contours, the contours are interpolated from the layers above and below. ROIs are not mutually exclusive and the user can define as many as they believe useful. For example, in our 7 column medulla reconstruction, ROIs are defined both for the columns (approximately vertical cylinders) and layers (horizontal slices, layers M1-M10). Queries on the ROIs can be combined, so for example the user could ask for all synapses in column C, layer M2.

For computational efficiency, ROI membership is defined on 32x32x32 voxel cubes. Even with relatively large EM voxels (say 10 nm in each dimension) this corresponds to a volume 320 nm on a side. This is sufficient for ROI definition since the named biological structures are normally defined with respect to optical images, for which 320 nm is excellent resolution.

Another way to obtain ROIs is to align EM data to optically defined standard brains. These brains are al-

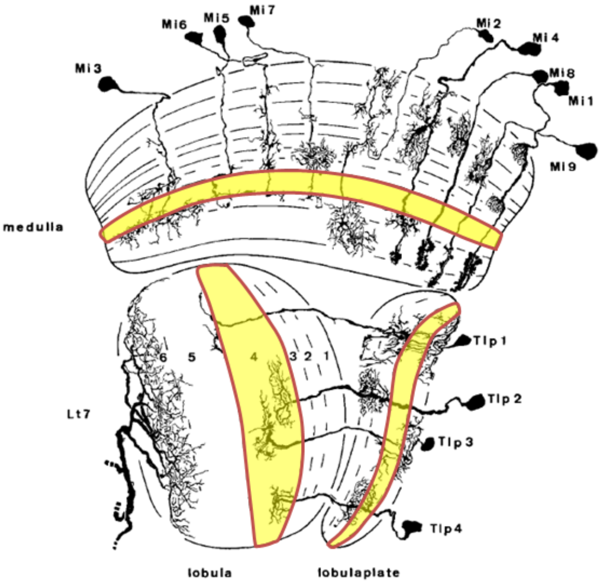


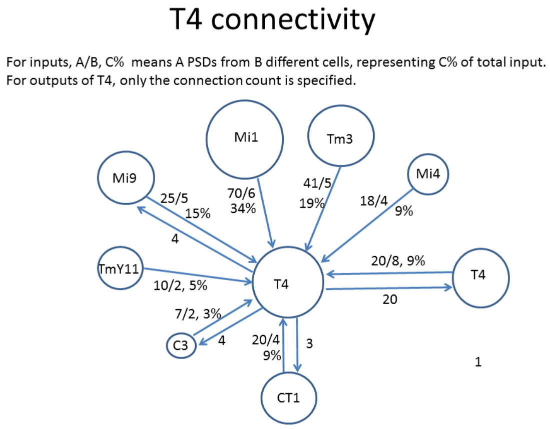
FIG. 1: Example of biological nomenclature. The different cell types (as defined by morphology) have different shapes. Three compartments - medulla, lobula, and lobularplate - are shown. Each is divided into layers. Shown highlighed are layer 7 of the medulla, layer 3 of the lobular plate, and layer 4 of the lobula. From [21].

ready marked with a number of named regions. Given a correspondence, these can then be imported into EM reconstructions. Finding such correspondence is not simple due to the large difference in scale mentioned above. Optical methods typically use a synapse stain such as nc82 to create an image of each brain for alignment. In EM we first identify the synapses (preferably automatically[23][24]), then blur these synapses to optical resolution, then use traditional optical alignment tools[22].

DISPLAYING CONNECTIVITY AS TABLES

The most basic form of output simply displays the connections to cells as tables. Each row represents one cell, and each column one connection. Each cell type is assigned an (arbitrary) color, so all connections to cells of a given type are displayed in the same color. Within each row, connections are ordered from strongest on the left to weakest on the right. Each entry lists the identity of the connected cell, they layer in which the centroid of the connection occurs (such as M2 in the medulla), the number of synapses in the connection, and (following a ':') the number of directly reciprocal synapses. There are two tables for each cell type - one for inputs and one for outputs.

Other similar tables emphasize different aspects of connections. A directional table shows the number of connections in each direction. (In the case of the medulla, there are 6 cardinal directions since the underlying structure is a hexagonal array.) Another table shows the connections in depth order, as opposed to by strength (so all they connections to layer M1 occur before all the connections to layer M2, and so on). In this case, if a cell has connections to another cell in many layers, these are regarded as separate connections and not combined (as they are in the strength table).


T4 connectivity

For inputs, A/B, C% means A PSDs from B different cells, representing C% of total input.
For outputs of T4, only the connection count is specified.

FIG. 2: Connectivity to and from each T4 cell, as shown by the 7 column reconstruction. The strengths are indicates as A/B C%, where A is the total number of synapses to all cells of that type, B is the number of cells connected, and C is the percentage of total input (output). The area of each circle is roughly proportional to the connection strength.

Other tables show the overall strongest connections, summed over all cells of a given type. So if a cell gets input from 5 cells of type A, each with a strength of 3, then this yields a total strength of 15. The cells are sorted in order of this combined strength.

Still another table shows the cardinality of the various connections between a cell and all others of a given type. So the connection between a cell and all cells of type T4 (for example) might be listed as 15,5,2,1,1,1 for a total strength of 25, which represents 18% of all the input synapses on that cell.

DISPLAYING CONNECTIVITY GRAPHICALLY

For humans, it is sometimes easier to visualize connections as a graph rather than a table. One often requested form is the connections to just one cell type, as shown in Fig 2. In general only the stronger connections are of interest. Also, for some purposes the connections to one instance of a cell are wanted, but in other cases it might be the average connectivity to all cells of the same type.

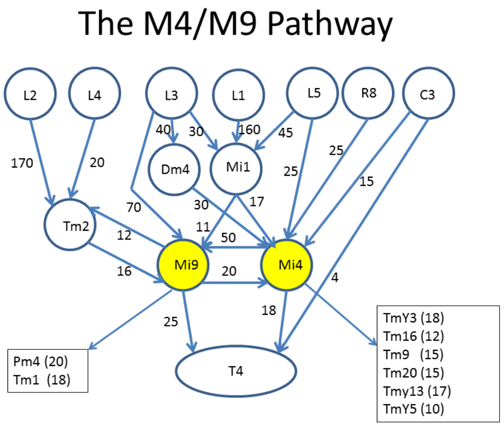


FIG. 3: Circuits leading to Mi4 and Mi9. To facilate human understanding, the signal flow is largely uni-directional (top to bottom in this case), there are relatively few line crossings, and the edges are annotated with weights. This diagram was drawn manually, but automated and semi-automated tools to create such diagrams would be helpful.

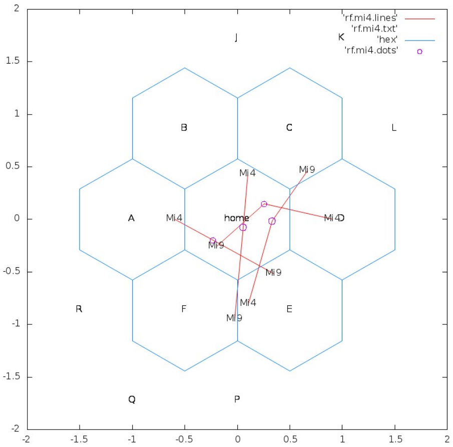


FIG. 4: Receptive fields of Mi1, Mi4, and Mi9, onto the 4 cells of type T4 that are most connected to the central (home) column. The dot indicates the synapse weighted centroid of the columns of the Mi1 connections. The labels Mi4 and Mi9 are similarly computed centroids for Mi4 and Mi9. The lines connect the centroids of each cell.

Circuits from input to output

One of the main reasons to draw a graph, rather than a table or list, is to enable human understanding of circuit operation. It is therefore important that the display diagram be designed not only to be technically correct, but to show the information flow in a way that is easy for humans to understand. Programs that do this for arbitrary electronic circuits[25] and directed graphs[26] have long existed. These could perhaps be mined for ideas helpful for drawing biological networks.

An example of what is desired is shown in Fig 3. This diagram was created (manually) to highlight the role of Mi4 and Mi9 from the medulla, cells that have strong cross-connections. Between them, they receive inputs from many cells from the lamina. The diagram is organized with inputs at the top and the T4 cell at the bottom. Only strong connections are shown, and other inputs to the T4 are ignored in this diagram.

INPUT AND OUTPUT CORRELATIONS

If there are N cell types, for each cell type we can represent the inputs(outputs) as a vector of size N, where the elements are how many inputs (outputs) this cell type gets from that cell type. Then we can look at the correlation of these vectors.

Two cells whose input vectors are highly correlated means they are getting their inputs from the same cell types, in the same proportions. The user might want to look at these cell types, since perhaps they differ primarily in temporal response or receptive fields.

Likewise, cells whose outputs are highly correlated means their information is going to same places. This probably indicates the target cell (or cells) is doing some sort of comparison.

RECEPTIVE FIELD COMPUTATIONS

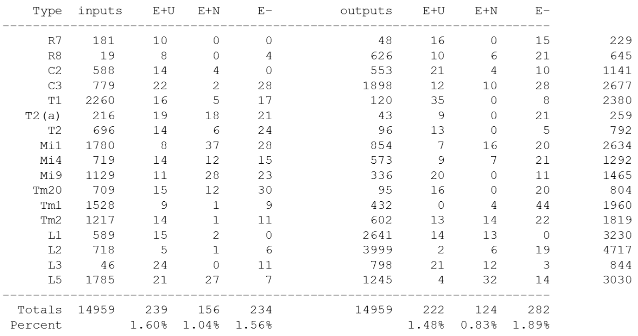
A visual receptive field defines which pixels of an optical input contribute to a computation, and the weights of such pixels. Receptive fields apply mostly in the visual system, where many common computions (such as motion detection, or feature recognition) rely on such fields.

Such a calculation starts with input cells, where we can assign a physical location that defines which location in the visual field drives that cell. Then the cells that get their input from these layer 1 cells can be given a location, typically the weighted centroid of all the inputs. The next layer is computed in a similar manner, and so on.

Receptive fields are often drawn in a 2D form corresponding to a visual field. An example is shown in Fig 4.

COMPARING CONNECTOMES

Comparing connectomes is helpful in many cases - comparing the two halves of the same animal, comparing the sub-units of a connectome (such as columns of the medulla), or comparing the results of two different experiments (such as two reconstructions of a medulla column done by two different techniques).



```
    Type  inputs     E+U   E+N     E-         outputs    E+U    E+N     E-
-----------------------------------------------------------------------------
      R7     181      10     0      0              48     16      0     15        229
      R8      19       8     0      4             626     10      6     21        645
      C2     588      14     4      0             553     21      4     10       1141
      C3     779      22     2     28            1898     12     10     28       2677
      T1    2260      16     5     17             120     35      0      8       2380
   T2(a)     216      19    18     21              43      9      0     21        259
      T2     696      14     6     24              96     13      0      5        792
     Mi1    1780       8    37     28             854      7     16     20       2634
     Mi4     719      14    12     15             573      9      7     21       1292
     Mi9    1129      11    28     23             336     20      0     11       1465
    Tm20     709      15    12     30              95     16      0     20        804
     Tm1    1528       9     1      9             432      0      4     44       1960
     Tm2    1217      14     1     11             602     13     14     22       1819
      L1     589      15     2      0            2641     14     13      0       3230
      L2     718       5     1      6            3999      2      6     19       4717
      L3      46      24     0     11             798     21     12      3        844
      L5    1785      21    27      7            1245      4     32     14       3030
-----------------------------------------------------------------------------
  Totals   14959     239   156    234           14959    222    124    282
 Percent           1.60% 1.04%  1.56%                  1.48%  0.83%  1.89%
```

FIG. 5: Example of report of discrepancies between subcircuits.

For small circuits, perhaps containing no more than 30-40 neurons, circuits can be compared by superimposing the two weighted adjacency graphs, with nodes that differ significantly highlighted with the use of color. For larger circuits this becomes unwieldy, and new approaches will need to be developed.

Estimating error rates

Comparing multiple copies of the same subcircuit is among the best ways we have to estimate both biological and reconstruction error rates. The first step is to find a subcircuit that is repeated several times in the same reconstruction, both in terms of the cells involved and the connections between them. To date, these subsets have been identified manually. In the longer run, with connectomes from less well studied organisms, it would be helpful to try to identify such subsets automatically.

Given several sub-circuits, the differences can be classified into several different types. A connection found in most (but not all) of the instances may be deemed to be missing, an error we call 'E-'. Extra connections can also exist and can be divided into two classes. One is a connection to a type not connected in the other instances, called 'E+U', where U stands for an Unexpected type. This represents a failure of the biological mechanism that determines the legal partners of connections. The other is a connection to a cell of the correct type, but the wrong instance of such a cell. For example, a cell may normally connect to a cell of type Mi1, but only to the one in its own column. If it connects to the same type of cell in another column, that's an error we call 'E+N', where N stands for a Normally connected type.

Any discrepancies between nominally identical subcircuits are normally double checked, since they often result from reconstruction errors. The ones that remain, however, represent real failures of the biological systems that determine wiring accuracy. Both human and biological errors are of interest, so it is helpful if the analysis program counts, reports, and keeps statistics on such errors. An example of such a report is shown in Fig. 5.

INFORMATION FROM RELATED METHODS

Some information, such as the identity of neurotransmitters used by a synapse, are of great interest to those who study biological circuits, but cannot be determined from the EM pictures. These must be found by other techniques, such as antibody labelling or RNA sequencing.

It is not yet clear how this should be displayed. Most likely theorists would like to multiply the synapse count by the effectiveness of the transmitter-receiver combination that is used. This would result in a signed, real number for the strength, instead of the currently displayed integer. In addition, the time constant (mostly of the receptor) would be good to display.

ANOMALOUS FEATURES

One helpful analysis for connectomes is the simple detection of anomalies. This is useful since odd features are often the result of reconstruction errors, so re-examinng suspicious points is a simple and effective way to find errors in reconstructions. For such features, output in both reconstruction and biological coordinates is helpful.

Information that has been used for this has been quantitative (such as synapses per unit area or volume), or connectivity driven (such as looking for nodes that output onto themselves). One that has worked very effectively is detailed comparison of circuits that are reasonably expected to be similar, as explained in the section on error rates. However this may only be applicable in specific cases, since it is not yet biologically clear how many such cases exist.

The features found by anomaly analysis may also be biological in origin. For example, we found that most autapses (a synapse where a cell connects to itself) were indeed reconstruction errors. However, there were two cell types where the synapses appeared consistently.

LIBRARIES OF MOTIFS

Current connectome projectes usually concentrate on well studied problems in well known organisms. The mammalian retina[27][28] and the *Drosophila* optic lobe[1], recent targets of connectomics, have already been studied for more than half century. Soon, however, we will be in a position to recreate connectomes of systems that are less well understood. At this point, it will be helpful to find in the connectome circuits which parallel those in systems that have already been studied. This is an exact parallel to the current situation in genomics, where unknown sequences are compared to a database of known sequences as a first step towards understanding their function.

Therefore we will need to figure out how a library of motifs might best be stored, annotated, and searched for.

DISCUSSION

In the long run, it almost sure that most uses of a connectome will be mediated by some tool devised for that purpose. This is exactly what happened with the various genome projects.

Improvements to analysis tools can offer very high leverage, since a better tool increases the usefullness of all connectomes by both making insights easier to obtain and by making it available to a wider audience. This pattern is clear from genome analysis, where the popular sequence analysis tool BLAST[29] has more than twice as many cites as the human[30], mouse[31] and *Drosophila*[32] genomes combined.

CONCLUSIONS

We should work on better connectome analysis tools. Some of this work is underway, but much remains to be done.

[1] Shin-ya Takemura, Arjun Bharioke, Zhiyuan Lu, Aljoscha Nern, Shiv Vitaladevuni, Patricia K Rivlin, William T Katz, Donald J Olbris, Stephen M Plaza, Philip Winston, et al. A visual motion detection circuit suggested by *Drosophila* connectomics. *Nature*, 500(7461):175–181, 2013.

[2] Olaf Sporns. Graph theory methods for the analysis of neural connectivity patterns. In *Neuroscience databases*, pages 171–185. Springer, 2003.

[3] Bin He, Yakang Dai, Laura Astolfi, Fabio Babiloni, Han Yuan, and Lin Yang. econnectome: A MATLAB toolbox for mapping and imaging of brain functional connectivity. *Journal of neuroscience methods*, 195(2):261–269, 2011.

[4] Trygve B Leergaard, Claus C Hilgetag, and Olaf Sporns. Mapping the connectome: multi-level analysis of brain connectivity. *Frontiers in neuroinformatics*, 6, 2012.

[5] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[6] Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek. Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.

[7] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G Tollis. Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry*, 4(5):235–282, 1994.

[8] Linton C Freeman. Visualizing social networks. *Journal of social structure*, 1(1):4, 2000.

[9] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2):291–307, 1970.

[10] Alex Pothen, Horst D Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990.

[11] George Karypis and Vipin Kumar. Multilevel K-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, 48(1):96–129, 1998.

[12] Erez Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Information processing letters*, 76(4):175–181, 2000.

[13] Ulrik Brandes, Marco Gaertler, and Dorothea Wagner. *Experiments on graph clustering algorithms.* Springer, 2003.

[14] Scott White and Padhraic Smyth. A spectral clustering approach to finding communities in graph. In *SDM*, volume 5, pages 76–84. SIAM, 2005.

[15] Martin G Everett and Stephen P Borgatti. Analyzing clique overlap. *Connections*, 21(1):49–61, 1998.

[16] Michihiro Kuramochi and George Karypis. Finding frequent patterns in a large sparse graph. *Data mining and knowledge discovery*, 11(3):243–271, 2005.

[17] Shalev Itzkovitz and Uri Alon. Subgraphs and network motifs in geometric networks. *Physical Review E*, 71(2):026117, 2005.

[18] Julian R Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):31–42, 1976.

[19] Derek G. Corneil and David G. Kirkpatrick. A theoretical analysis of various heuristics for the graph isomorphism problem. *SIAM Journal on Computing*, 9(2):281–297, 1980.

[20] Bruno T Messmer and Horst Bunke. A new algorithm for error-tolerant subgraph isomorphism detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(5):493–504, 1998.

[21] K-F Fischbach and APM Dittrich. The optic lobe of *Drosophila melanogaster.* I. A Golgi analysis of wild-type structure. *Cell and tissue research*, 258(3):441–475, 1989.

[22] Ting Zhao, Shinya Takemura, Gary Huang, Jane Anne Horne, William Katz, Kazunori Shinomiya, Louis Scheffer, Ian Meinertzhagen, Pat Rivlin, and Stephen Plaza. Large-scale EM analysis of the *Drosophila* antennal lobe with automatically computed synapse point clouds. *To be published*, 2015.

[23] Anna Kreshuk, Christoph N Straehle, Christoph Sommer, Ullrich Koethe, Marco Cantoni, Graham Knott, and Fred A Hamprecht. Automated detection and segmentation of synaptic contacts in nearly isotropic serial electron microscopy images. *PloS one*, 6(10):e24899, 2011.

[24] A Kreshuk, Christoph N Straehle, Christoph Sommer, Ullrich Koethe, Graham Knott, and Fred A Hamprecht. Automated segmentation of synapses in 3D EM data. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 220–223. IEEE, 2011.

[25] Yeu-Shen Jehng, Liang-Gee Chen, and Tai-Ming Parng. ASG: Automatic schematic generator. *INTEGRATION, the VLSI journal*, 11(1):11–27, 1991.

[26] Emden R Gansner, Eleftherios Koutsofios, Stephen C North, and K-P Vo. A technique for drawing directed graphs. *Software Engineering, IEEE Transactions on*,

19(3):214–230, 1993.

[27] James R Anderson, Bryan W Jones, Carl B Watt, Margaret V Shaw, Jia-Hui Yang, David DeMill, James S Lauritzen, Yanhua Lin, Kevin D Rapp, David Mastronarde, et al. Exploring the retinal connectome. *Molecular vision*, 17:355, 2011.

[28] Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

[29] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[30] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[31] Asif T Chinwalla, Lisa L Cook, Kimberly D Delehaunty, Ginger A Fewell, Lucinda A Fulton, Robert S Fulton, Tina A Graves, LaDeana W Hillier, Elaine R Mardis, John D McPherson, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[32] Mark D Adams, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, Richard F Galle, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195, 2000.