BMC Bioinformatics

SOFTWARE   **Open Access**

An automated graphics tool for comparative genomics: the Coulson plot generator

Helen I Field[1*], Richard MR Coulson[2] and Mark C Field[3*]

Abstract

Background: Comparative analysis is an essential component to biology. When applied to genomics for example, analysis may require comparisons between the predicted presence and absence of genes in a group of genomes under consideration. Frequently, genes can be grouped into small categories based on functional criteria, for example membership of a multimeric complex, participation in a metabolic or signaling pathway or shared sequence features and/or paralogy. These patterns of retention and loss are highly informative for the prediction of function, and hence possible biological context, and can provide great insights into the evolutionary history of cellular functions. However, representation of such information in a standard spreadsheet is a poor visual means from which to extract patterns within a dataset.

Results: We devised the Coulson Plot, a new graphical representation that exploits a matrix of pie charts to display comparative genomics data. Each pie is used to describe a complex or process from a separate taxon, and is divided into sectors corresponding to the number of proteins (subunits) in a complex/process. The predicted presence or absence of proteins in each complex are delineated by occupancy of a given sector; this format is visually highly accessible and makes pattern recognition rapid and reliable. A key to the identity of each subunit, plus hierarchical naming of taxa and coloring are included. A java-based application, the Coulson plot generator (CPG) automates graphic production, with a tab or comma-delineated text file as input and generating an editable portable document format or svg file.

Conclusions: CPG software may be used to rapidly convert spreadsheet data to a graphical matrix pie chart format. The representation essentially retains all of the information from the spreadsheet but presents a graphically rich format making comparisons and identification of patterns significantly clearer. While the Coulson plot format is highly useful in comparative genomics, its original purpose, the software can be used to visualize any dataset where entity occupancy is compared between different classes.

Availability: CPG software is available at sourceforge http://sourceforge.net/projects/coulson and http://dl.dropbox.com/u/6701906/Web/Sites/Labsite/CPG.html

Q1 Q2

Background

With a rapidly growing database of completed genomes and consequential improvements to the reconstruction of deep and broad phylogenetic relationships, it has become possible to consider the molecular origins of many complex cellular systems. Such analyses can reveal deep relationships between cellular functions, identify lineage-specific features and uncover evolutionary mechanisms [1-5], and are important in the identification of, for example,

pathogen-associated gene products, with potential for therapeutic intervention, as well as in attempts to understand how such systems arose. Further, falling costs of nucleotide sequencing are providing opportunities to generate genome sequences from even hard to culture organisms, making analysis of function in these taxa possible through comparison with tractable organisms. In short, the need to present comparative data is highly pressing and likely to remain an issue for some time.

While it is now comparatively trivial to generate vast datasets containing 100s to 1000s of query results using BLAST, HMMer and other sequence-based algorithms [6-10] these data constitute essentially gene lists, which only have value when processed and presented coherently

* Correspondence: imogen.field@lgcgenomics.com; mcf34@cam.ac.uk
[1]LGC Genomics Ltd, Pindar Road, Hoddesdon, Hertfordshire EN11 0WZ, UK
[3]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK
Full list of author information is available at the end of the article

[5,11-16]. The major biological added value within such analyses is the ability to rapidly compare the distributions of genes between multiple biological processes, i.e. protein complexes and pathways, and also across many taxa. This is quite challenging as these datasets can contain may hundreds/thousands of gene calls, and unless these data are represented graphically and in an easily comprehended manner, patterns are difficult to observe. In particular, spreadsheets do not lend themselves to browsing and fragmenting datasets into subgroups to reduce data complexity often removes much valuable comparative information. Production of comparison figures from developing datasets (works in progress) are invaluable during dataset production, and even for making decisions and developing hypotheses, but manual production of figures on the fly is unfeasible.

To address these needs we devised the Coulson plot, a matrix of colorized pie charts and which displays information in a clustered format, together with hierarchical taxonomic labels and a key to individual gene products. This plot we, and others, have used in multiple publications and which we have found to be highly useful and accessible to readers of these reports [3,17-24]. However, the manual construction of these plots is time consuming and, with hundreds of elements, error prone, and which precludes *on-the-fly* plots and possibly wider adoption of the format. Hence, to facilitate generic/automated production and adoption of the plot we developed a platform-independent application, the Coulson plot generator (CPG), to draw Coulson plots from structured data that uses standard spreadsheet file formats as input. CPG should be accessible to the vast majority of workers with only rudimentary computing skills and requires minimal post-plot manipulations to generate publication quality plots of considerable complexity.

Implementation
Graphical concept
We considered many of the formats commonly used in the published literature for the display of comparative genomic data, and found these frequently too complex or inelegant for the presentation of data in a manner that retains as much biological information as possible. Specifically, simple spreadsheets or dot plots are either difficult to read or lose information concerning complexes, which is especially critical to understanding evolutionary processes. In many instances dot blots also become very large, with moire effects and other issues emerging. Hence we designed a more sophisticated format that retains functional groupings, provides colors as keys to taxonomic relationships and also provides a key to subunit identity (Figure 1). We consider that the Coulson format retains more information than dot plots and provides this in an attractive and easy to comprehend manner.

Algorithm design
Originally, we generated Coulson plots using Perl with individual data structures for each diagram, with individual programming for each diagram, requiring considerable time and programming expertise to produce a basic figure [3]. In addition this is also potentially an error prone process. A Coulson plot generator (CPG) application was written as an open source, stand alone program developed in Java using Eclipse (http://www.eclipse.org) to execute on any machine running a Java Virtual Machine (v1.5.0 or greater). CPG takes as input a comma separated (.csv) text file of binary data recording subunit occupancy in multiple systems (Figure 2). Systems, e.g. multi-protein complexes, are labelled in the first column, with subunits in the second. Then the data itself (+/−) begins in subsequent columns. Use of either '+' and '-' or '1' and '0' for data occupancy are supported. Only one protein name is required per list of subunits, and column one is occupied only at the position of the first subunit of each group. The input table uses the four top rows and two columns for labeling. Kingdom or supergroup names [25] fall in rows one and two, while species names fall on row three, which must all be occupied and no gaps are permitted. The fourth row can be left unused, omitted completely or utilised for additional taxonomic annotation if desired. The input table uses the first two columns for labeling, with a protein/entity name in column one, and subunit names in column two. Only one protein name is required per list of subunits, and column one is occupied only at the position of the first subunit.

CPG parses the input file and breaks it up into an array of tab-delimited strings. These are processed and displayed as a table, where diagram settings can then be selected (Figure 3B). For creating the figure, the CPG algorithm takes each row, representing the occupancy data across species for one single subunit of a protein and tracks which protein/entity the subunit belongs to. When all the subunits have been collected for a protein/entity, a row of pies is plotted. Species can be separated by Kingdom or supergroup in the same way, and colors are allocated for each supergroup. Pie data are stored in vectors containing 1 or 0 to show pie occupancy which is converted into a graphic (Figures 1 and 3). The collection of graphics for all the pies are laid out with labels. The number of segments for each pie depends on the number of rows associated with a complex: the number of pies depends on the number of species (indicated by columns). Once plotted, the image is stored in memory and resized by zooming in or out, and may be saved in various image formats as well as editable images (svg in PDF or svg format).
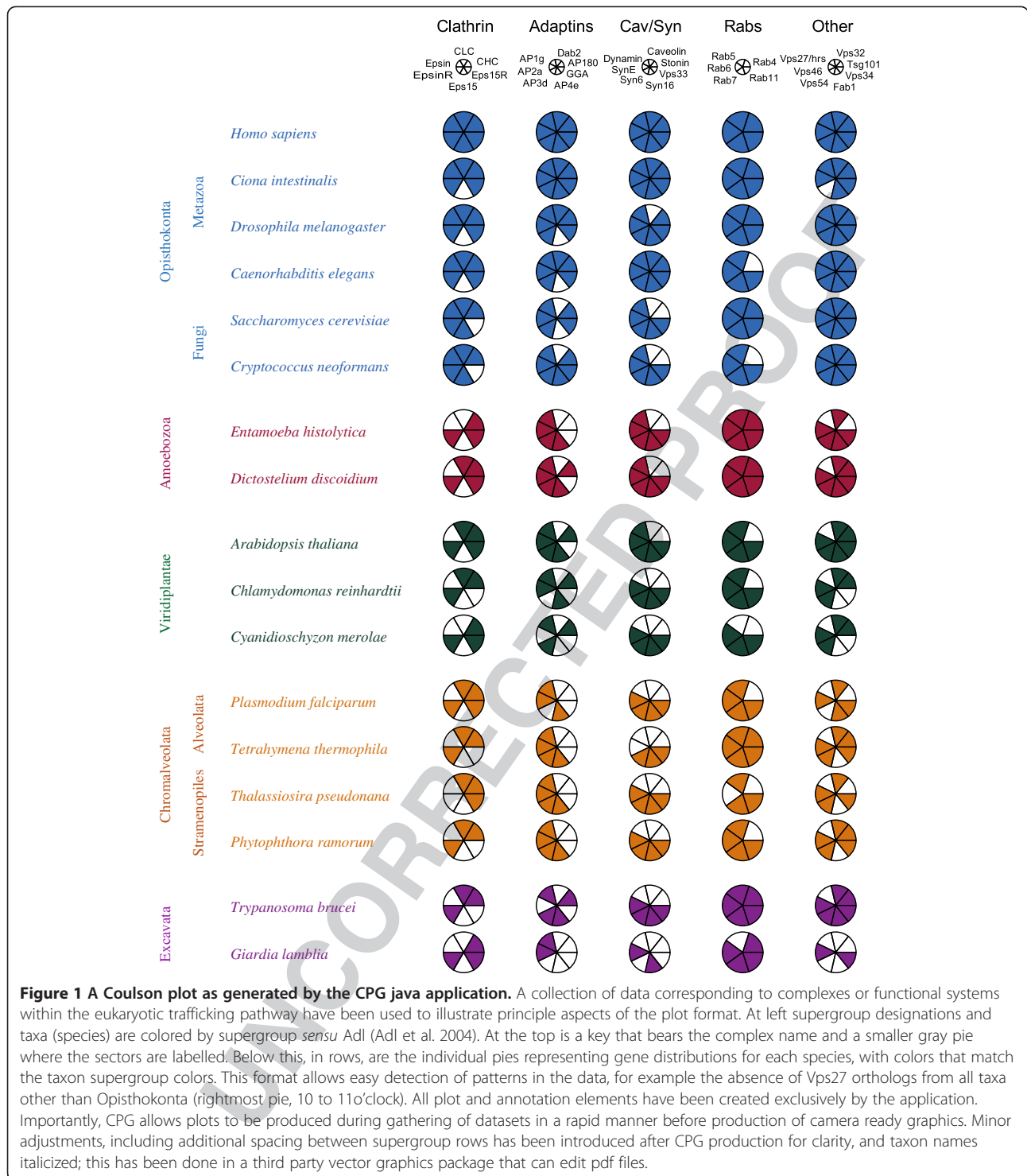
Program operation
The CPG application opens with three tabs (Figure 3A). The first allows the user to select an input data file, the

Figure 1 A Coulson plot as generated by the CPG java application. A collection of data corresponding to complexes or functional systems within the eukaryotic trafficking pathway have been used to illustrate principle aspects of the plot format. At left supergroup designations and taxa (species) are colored by supergroup *sensu* Adl (Adl et al. 2004). At the top is a key that bears the complex name and a smaller gray pie where the sectors are labelled. Below this, in rows, are the individual pies representing gene distributions for each species, with colors that match the taxon supergroup colors. This format allows easy detection of patterns in the data, for example the absence of Vps27 orthologs from all taxa other than Opisthokonta (rightmost pie, 10 to 11o'clock). All plot and annotation elements have been created exclusively by the application. Importantly, CPG allows plots to be produced during gathering of datasets in a rapid manner before production of camera ready graphics. Minor adjustments, including additional spacing between supergroup rows has been introduced after CPG production for clarity, and taxon names italicized; this has been done in a third party vector graphics package that can edit pdf files.

160 second, to choose custom pie colors, and the third
161 tab provides the Help/Manual and change log (and
162 licensing). The fourth tab provides a process log, and
163 information to assist with input file formatting (appropriate
164 error messages if your input is not acceptable). The 'Plot'
165 button is not enabled unless the input is correct; clicking

166 on 'Plot' generates the figure. By returning to the first,
167 tabbed window, multiple plots may be created from
168 different inputs, and different versions of a figure may
169 be created from the second window and viewed all
170 together. A default color set is supplied (text file and
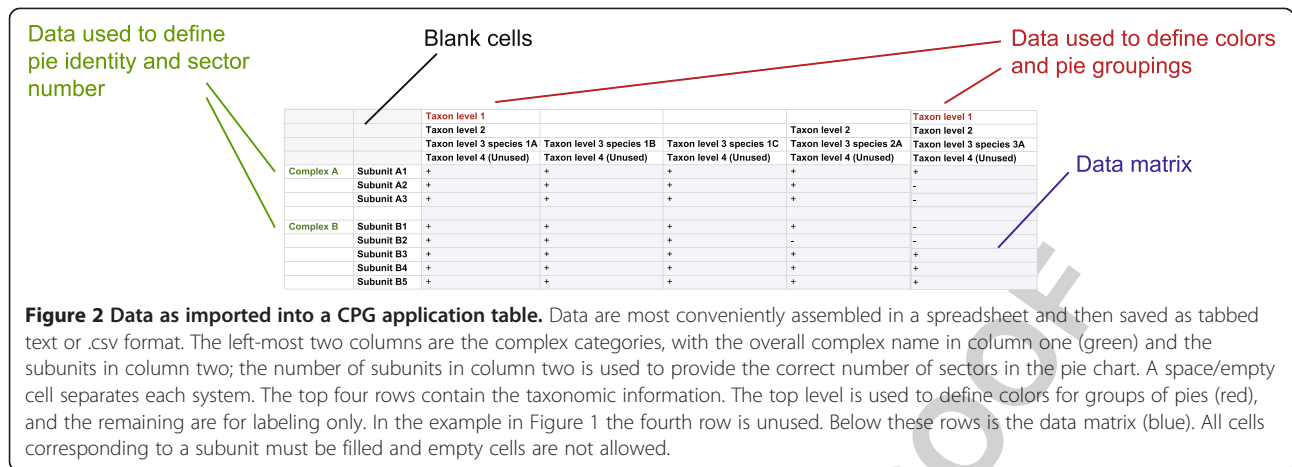171 hard coded) (Figure 3B). After selecting an input file,

Figure 2 Data as imported into a CPG application table. Data are most conveniently assembled in a spreadsheet and then saved as tabbed text or .csv format. The left-most two columns are the complex categories, with the overall complex name in column one (green) and the subunits in column two; the number of subunits in column two is used to provide the correct number of sectors in the pie chart. A space/empty cell separates each system. The top four rows contain the taxonomic information. The top level is used to define colors for groups of pies (red), and the remaining are for labeling only. In the example in Figure 1 the fourth row is unused. Below these rows is the data matrix (blue). All cells corresponding to a subunit must be filled and empty cells are not allowed.

CPG will parse the data and if successful it will convert the data to a table (Figure 3C). Clicking 'Figure' will display a Coulson plot of the data in a new scrollable window. An example dataset used for testing is shown (Figure 3B) from which a small portion was taken for early development. The text file was produced using Microsoft Excel, with data entry in the table as described (Figure 2). Data from Excel were exported as comma separated files (.csv). The output file is an editable PDF or SVG file which can be opened and manipulated with Inkscape (http://inkscape.org/) or Adobe Illustrator (http://www.adobe.com/products/illustrator.htm). We selected this option as more efficient than attempting to build sophisticated editing tools into CPG as the precise choices and requirements of users and datasets are difficult to predict.
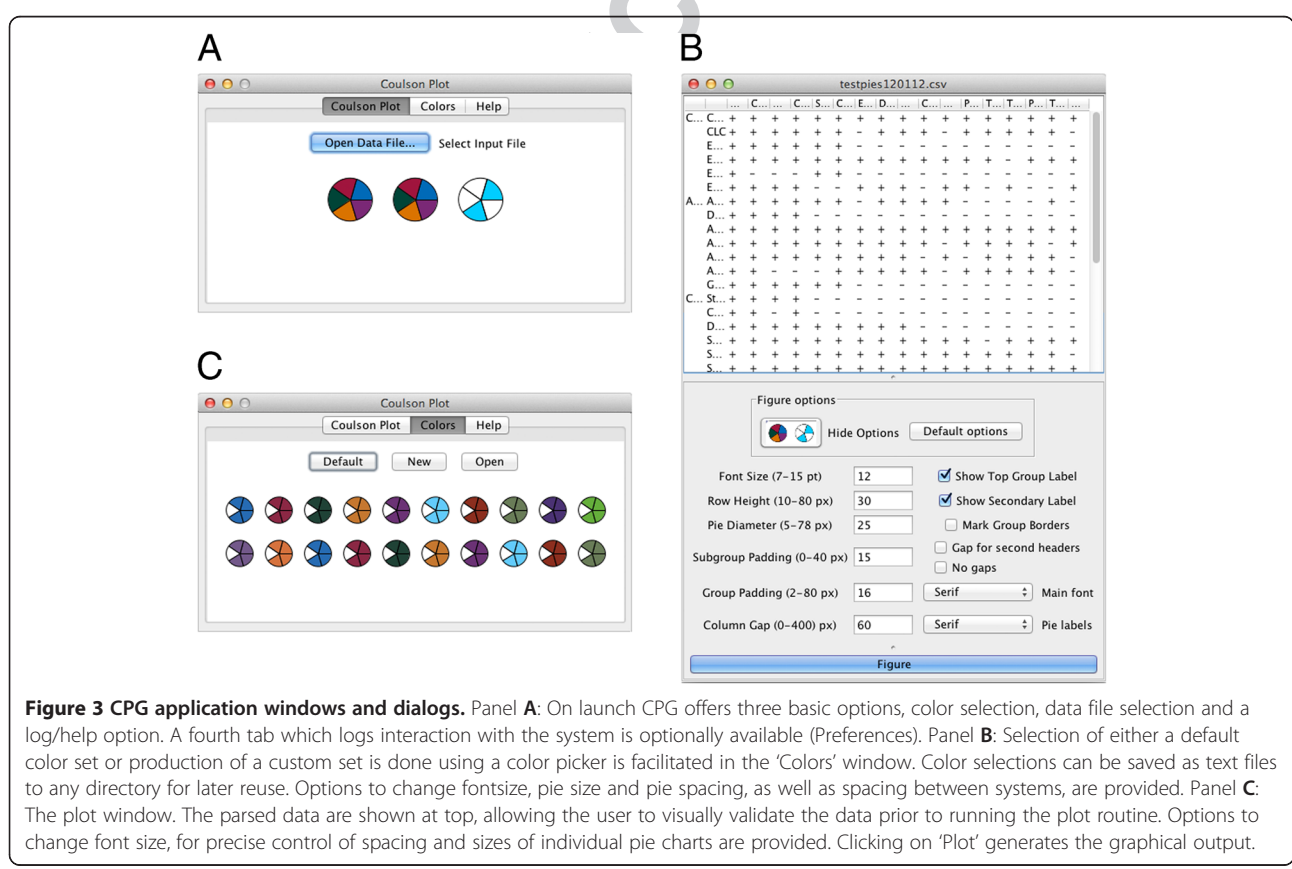

Figure 3 CPG application windows and dialogs. Panel **A**: On launch CPG offers three basic options, color selection, data file selection and a log/help option. A fourth tab which logs interaction with the system is optionally available (Preferences). Panel **B**: Selection of either a default color set or production of a custom set is done using a color picker is facilitated in the 'Colors' window. Color selections can be saved as text files to any directory for later reuse. Options to change fontsize, pie size and pie spacing, as well as spacing between systems, are provided. Panel **C**: The plot window. The parsed data are shown at top, allowing the user to visually validate the data prior to running the plot routine. Options to change font size, for precise control of spacing and sizes of individual pie charts are provided. Clicking on 'Plot' generates the graphical output.

Figure 4 (See legend on next page.)
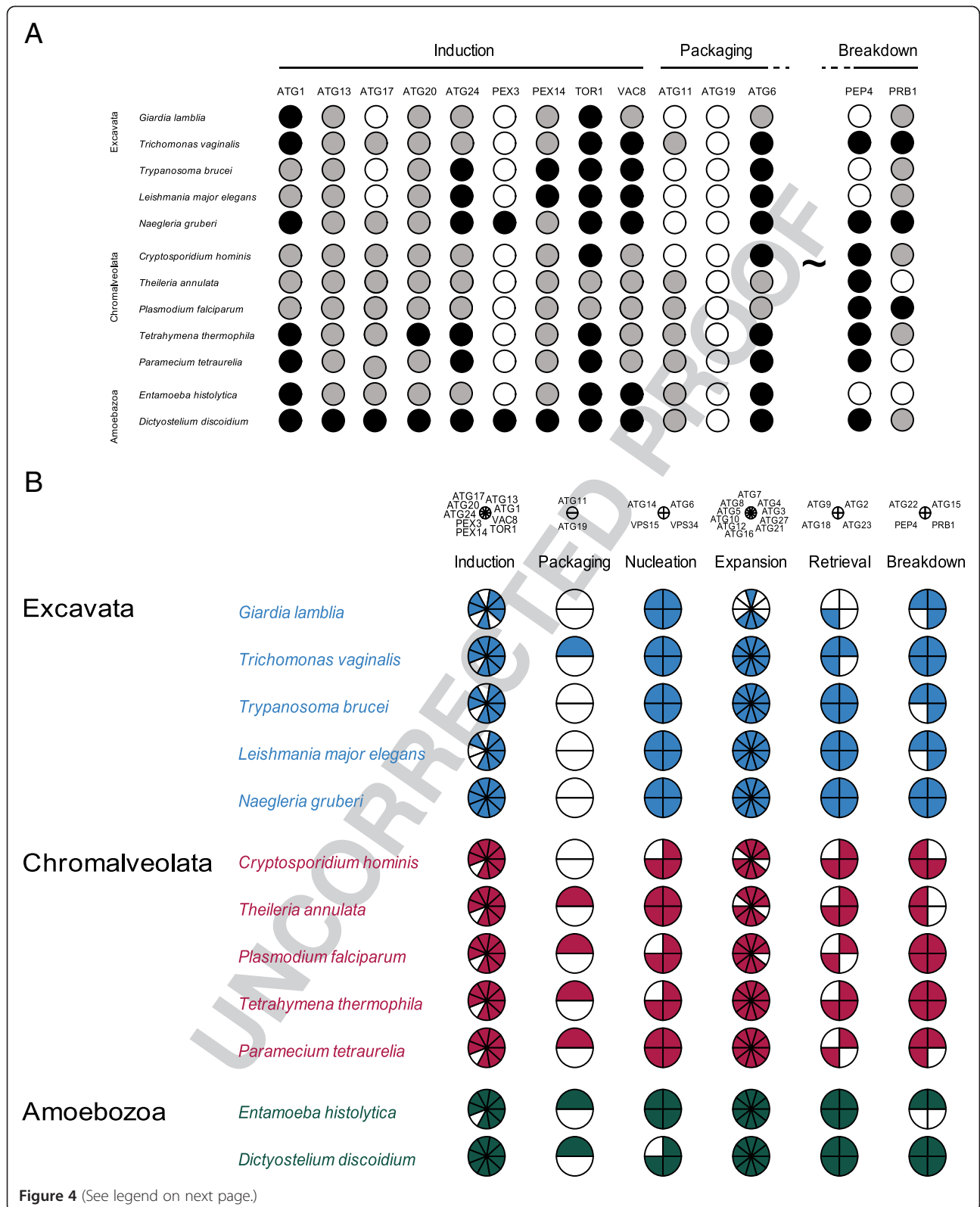
(See figure on previous page.)

Figure 4 Two distinct display formats that can be produced using CPG, based on the autophagy pathway in select protists. Panel **A**: A traditional three-state dot plot representation, similar to that presented in the original publication [25]. White designates not found or absent, black a confident assignment, and gray a lower confidence assignment, in this case lacking phylogenetic support. The plot has some appeal, is clean, but is also large, and lacks functional groupings, making assessment of differential levels of occupancy of the distinct complexes or processes within the autophagy pathway difficult to comprehend, while the plot is also not very compact. Tilda and dotted lines indicate many columns omitted for space reasons. Panel **B**: Standard CPG format, with each complex represented as a single pie, and the higher order taxon membership colorized. Note that the plot is considerably more compact, and complex occupancy and/or subunit retention clear, even for complexes containing many subunits. Data are taken from ref 24 for illustrative purposes only. Both figures generated with CPG with manipulations in Adobe Illustrator. Note that three state occupancy is not allowed in CPG at present so that gray circles in panel A were manually colored.

This follows a similar philosophy to FigTree, a popular phylogenetics tree graphics package which also generates editable graphics requiring a small amount of finessing prior to publication (http://tree.bio.ed.ac.uk/software/figtree/). A Coulson plot with more than 200 pies can be produced satisfactorily.

Results

We developed the Coulson plot to display and compare data on gene representation grouped by gene product complex or pathway membership and to display this information across multiple taxa (Figure 1). An array of gene product components from multiple species with each complex is displayed as a pie chart comprising a variable number of components (sectors), the number of which matches the number of protein subunits in a functional complex, process (i.e. pathway) or other functional group. Pie charts are arranged by phylogenetic hierarchy to allow evaluation of evolutionary trends and the rapid identification of gene losses, specializations or expansions. Several such systems may be compared, so that an array of systems is represented for each species. Using colors, it is possible to separate groups of systems with excellent visual clarity.

One of the more flexible aspects of the CGP is that the user can decide quickly how best to group data. For example, complexes or pathways with many components may be difficult to visualize in the individual pie charts, with the result that clarity is lost. However, CPG allows pies to be set up that have only one subunit, for example mimicking the more standard dot blot format, or to subdivide the data into subcomplexes with biological relevance, to improve clarity or increase the amount of data that may be logically compared (Figure 4). A second flexible feature is the ability to manually edit the plot to improve spacing, add additional annotation and change fonts, which allows the user to control the graphic and maintain consistency with additional elements in a figure. Overall, we have found that CPG improves workflow and reduces data transcription errors. Finally, the program is lightweight, making minimal impact on CPU resources and runs without issue on most major platforms.

Conclusions

We have found the Coulson plot to be highly valuable for presentations of comparative genomic data, and that the lucid display of patterns within datasets more than offset the time required to manually produce these plots. However, we are aware that the skills required and potentially the effort needed acted as a barrier to adoption of a broadly potentially useful graphing format, and which is not available as part of commercial graphing packages as far as we are aware. We therefore developed a plotting tool that manages the vast majority of the plot functionality, leaving the user a format that can be subjected to final editing as appropriate for individual requirements.

A great many datasets have been used to test CPG [3,17-24]. We find the software is stable on OS X (10.5.8 to 10.8.2), Microsoft Windows (XP, 7 and 8) and multiple versions of Linux. Creation of .csv output files from Microsoft Excel, Apple Numbers or open source office suites that can be read by CPG is routine, and the PDF and SVG output successfully imported to Adobe Illustrator or Inkscape as an editable graphic. A diagram with more than 200 pies and over 600 individual elements can be routinely produced, allowing publication quality figures to be generated in one hour. The ability to rapidly generate plots from dissimilar datasets *on-the-fly*, allowing hypothesis-driven composition of datasets, is a distinct advantage, and we hope that the Coulson plot will become a more generally exploited format, and that the use of this plot beyond comparative genomics will be facilitated with the provision of CPG.

Availability and requirements

CPG is a Java application and requires Java 1.5.0 or higher for the JVM. CPG source code and binaries are available from sourceforge: http://sourceforge.net/projects/coulson as a jar file or disc image for Mac OS X. Project home page: http://dl.dropbox.com/u/6701906/Web/Sites/Labsite/CPG.html and http://sourceforge.net/projects/coulson. The software is licensed under GNU Artistic license 2.0.

Author details
[1]LGC Genomics Ltd, Pindar Road, Hoddesdon, Hertfordshire EN11 0WZ, UK. [2]Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [3]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, UK.

References
1. Carvalho-Santos Z, Azimzadeh J, Pereira-Leal JB, Bettencourt-Dias M: **Evolution: tracing the origins of centrioles, cilia, and flagella.** *J Cell Biol* 2011, **194**(2):165–175.
2. Wickstead B, *et al*: **Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton.** *BMC Evol Biol* 2010, **10**:110.
3. Field MC, Gabernet-Castello C, Dacks JB: **Reconstructing the evolution of the endocytic system: insights from genomics and molecular cell biology.** *Adv Exp Med Biol* 2007, **607**:84–96.
4. Barnes RL, Shi H, Kolev NG, Tschudi C, Ullu E: **Comparative genomics reveals two novel RNAi factors in** *Trypanosoma brucei* **and provides insight into the core machinery.** *PLoS Pathog* 2012, **8**(5):e1002678.
5. Serpeloni M, Vidal NM, Goldenberg S, Avila AR, Hoffmann FG: **Comparative genomics of proteins involved in RNA nucleocytoplasmic export.** *BMC Evol Biol* 2011, **11**:7.
6. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39**(Web Server issue):W29–W37.
7. Price MN, *et al*: **FastBLAST: homology relationships for millions of proteins.** *PLoS One* 2008, **3**:e3589.
8. Grant JR, Arantes AS, Stothard P: **Comparing thousands of circular genomes using the CGView Comparison Tool.** *BMC Genomics* 2012, **13**(1):202.
9. Koumandou VL, Klute MJ, Herman EK, Nunez-Miguel R, Dacks JB, Field MC: **Evolutionary reconstruction of the retromer complex and its function in** *Trypanosoma brucei.* *J Cell Sci* 2011, **124**(Pt 9):1496–1509.
10. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV, Ortho DB: **the hierarchical catalog of eukaryotic orthologs in 2011.** *Nucleic Acids Res* 2011, **39**(Database issue):D283–D288.
11. Elias M, Brighouse A, Gabernet-Castello C, Field MC, Dacks JB: **Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases.** *J Cell Sci* 2012, **125**(Pt 10):2500–2508.
12. Kloepper TH, Kienle CN, Fasshauer D: **An elaborate classification of SNARE proteins sheds light on the conservation of the eukaryotic endomembrane system.** *Mol Biol Cell* 2007, **18**(9):3463–3471.
13. Iyer LM, Anantharaman V, Wolf MY, Aravind L: **Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes.** *Int J Parasitol* 2008, **38**(1):1–31.
14. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran AN, Wortman JR, Alsmark UC, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC,
Kummerfeld SK, Pereira-Leal JB, Nilsson D, Peterson J, Salzberg SL, Shallom J, Silva JC, Sundaram J, Westenberger S, White O, Melville SE, Donelson JE, Andersson B, Stuart KD, Hall N: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309**(5733):404–409.
15. O'Reilly AJ, Dacks JB, Field MC: **Evolution of the karyopherin-β family of nucleocytoplasmic transport factors; ancient origins and continued specialization.** *PLoS One* 2011, **6**(4):e19308.
16. Neumann N, Lundin D, Poole AM: **Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor.** *PLoS One* 2010, **5**(10):e13241.
17. Koumandou VL, Dacks JB, Coulson RM, Field MC: **Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins.** *BMC Evol Biol* 2007, **7**:29.
18. Leung KF, Dacks JB, Field MC: **Evolution of the multivesicular body ESCRT machinery; retention across the eukaryotic lineage.** *Traffic* 2008, **9**(10):1698–1716.
19. Nevin WD, Dacks JB: **Repeated secondary loss of adaptin complex genes in the Apicomplexa.** *Parasitol Int* 2009, **58**(1):86–94.
20. Dokudovskaya S, Waharte F, Schlessinger A, Pieper U, Devos DP, Cristea IM, Williams R, Salamero J, Chait BT, Sali A, Field MC, Rout MP, Dargemont C: **A conserved coatomer-related complex containing Sec13 and Seh1 dynamically associates with the vacuole in** *Saccharomyces cerevisiae.* *Mol Cell Proteomics* 2011, **10**(6):M110.006478.
21. Herman EK, Walker G, van der Giezen M, Dacks JB: **Multivesicular bodies in the enigmatic amoeboflagellate** *Breviata anathema* **and the evolution of ESCRT 0.** *J Cell Sci* 2011, **124**(Pt 4):613–621.
22. Hirst J, Barlow LD, Francisco GC, Sahlender DA, Seaman MN, Dacks JB, Robinson MS: **The fifth adaptor protein complex.** *PLoS Biol* 2011, **9**(10):e1001170.
23. Klute MJ, Melançon P, Dacks JB: **Evolution and diversity of the Golgi.** *Cold Spring Harb Perspect Biol* 2011, **3**(8):a007849. 1.
24. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MF: **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol* 2005, **52**(5):399–451.
25. Brennand A, Gualdrón-López M, Coppens I, Rigden DJ, Ginger ML, Michels PA: **Autophagy in parasitic protists: unique features and drug targets.** *Mol Biochem Parasitol* 2011, **177**(2):83–99.