

This is an open access publisher version of an article that appears in:

PLOS ONE

The internet address for this paper is:

<https://publications.icr.ac.uk/13622/>

Please direct all emails to:

publications@icr.ac.uk

Institute of Cancer Research Repository

<https://publications.icr.ac.uk>

Published text:

M Scales, R Jager, et al (2014), *visPIG – a web tool for producing multi-region, multi-track, multi-scale plots of genetic data*, **PLOS One**, Vol. 9(9), e107497



visPIG - A Web Tool for Producing Multi-Region, Multi-Track, Multi-Scale Plots of Genetic Data

Matthew Scales^{1,2}, Roland Jäger¹, Gabriele Migliorini¹, Richard S. Houlston¹, Marc Y. R. Henrion^{1*}

1 Division of Genetics and Epidemiology, The Institute of Cancer Research, Surrey, United Kingdom, **2** Department of Mathematics, Imperial College London, London, United Kingdom

Abstract

We present Visual Plotting Interface for Genetics (visPIG; <http://vispig.icr.ac.uk>), a web application to produce multi-track, multi-scale, multi-region plots of genetic data. visPIG has been designed to allow users not well versed with mathematical software packages and/or programming languages such as R [1], Matlab®, Python, etc., to integrate data from multiple sources for interpretation and to easily create publication-ready figures. While web tools such as the UCSC Genome Browser [2] or the WashU Epigenome Browser [3] allow custom data uploads, such tools are primarily designed for data exploration. This is also true for the desktop-run Integrative Genomics Viewer (IGV) [4,5]. Other locally run data visualisation software such as Circos [6] require significant computer skills of the user. The visPIG web application is a menu-based interface that allows users to upload custom data tracks and set track-specific parameters. Figures can be downloaded as PDF or PNG files. For sensitive data, the underlying R [1] code can also be downloaded and run locally. visPIG is multi-track: it can display many different data types (e.g. association, functional annotation, intensity, interaction, heat map data,...). It also allows annotation of genes and other custom features in the plotted region(s). Data tracks can be plotted individually or on a single figure. visPIG is multi-region: it supports plotting multiple regions, be they kilo- or megabases apart or even on different chromosomes. Finally, visPIG is multi-scale: a sub-region of particular interest can be 'zoomed' in. We describe the various features of visPIG and illustrate its utility with examples. visPIG is freely available through <http://vispig.icr.ac.uk> under a GNU General Public License (GPLv3).

Citation: Scales M, Jäger R, Migliorini G, Houlston RS, Henrion MYR (2014) visPIG - A Web Tool for Producing Multi-Region, Multi-Track, Multi-Scale Plots of Genetic Data. *PLoS ONE* 9(9): e107497. doi:10.1371/journal.pone.0107497

Editor: Michael Edward Zwick, Emory University School Of Medicine, United States of America

Received: May 29, 2014; **Accepted:** August 12, 2014; **Published:** September 10, 2014

Copyright: © 2014 Scales et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All data files used in the paper can be downloaded from <http://vispig.icr.ac.uk>. The visPIG software source code can be downloaded from the same website.

Funding: This work is supported by grants from the European Union (FP7/207-2013) under grant 258236, FP7 collaborative project SYSCOL and COST Action BM1206, Cancer Research UK (C1298/A8362-Bobby Moore Fund) and Leukaemia Lymphoma Research (LRF05001 and LRF06002). R.J. was in receipt of an Erwin Schrödinger Fellowship from the Austrian Science Fund (FWF). M.S. was funded by an Institute of Cancer Research (ICR)/Engineering and Physical Sciences Research Council (EPSRC) summer student bursary. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: marc.henrion@icr.ac.uk

Introduction

In genetic research there is an increasing need to integrate multiple datasets, sometimes quite different in nature, for analysis or visualisation. For example, in genome-wide association studies (GWAS) association p-values are commonly annotated with evolutionary conservation data, RNAseq data, ChIP-seq data or functional information, such as ChromHMM [7] output. Often such datasets are on vastly different scales; for example a researcher might fine-map a particular association signal, often involving a single linkage-disequilibrium (LD) block of no more than a few hundred to a few thousand base-pairs, and then use C-based techniques such as 4C [8,9], Hi-C [10] or ChIA-PET [11] to identify potential long-range interactions spanning tens of kilobases, megabases or several chromosomes.

Presenting all these datasets within a single, easy-to-interpret figure is challenging. Furthermore, it is generally not possible to plot each of the different datasets using the same software. In practice, many researchers produce separate plots, some generated by web resources such as the UCSC Genome Browser or the Broad Institute's SNAP web application [12], which are then

combined using image editing software such as Adobe® Photoshop® or Microsoft PowerPoint. This is not ideal as it is error-prone when multiple tracks are aligned.

To address these deficiencies we have developed the Visual Plotting Interface for Genetics, visPIG – <http://vispig.icr.ac.uk> which allows users to produce figures containing multiple (epi-) genetic data tracks, with options to plot multiple regions and zoom in on specific sub-regions. One key advantage of using visPIG to produce such multi-track, multi-scale and multi-region figures is that the tracks are aligned correctly down to a single base-pair.

Results and Discussion

visPIG has been developed to have three core capabilities: it can display multiple data tracks across multiple regions at multiple scales. Importantly, visPIG has been designed to be easy to use; specifically, users can access it via a web interface, selecting the files to be uploaded and adjusting plotting parameters through menu panels. Furthermore, at inception it was required that visPIG produces publication-grade graphs that need no subsequent image processing before inclusion in a scientific article.

3q26.2 & 8q24.21

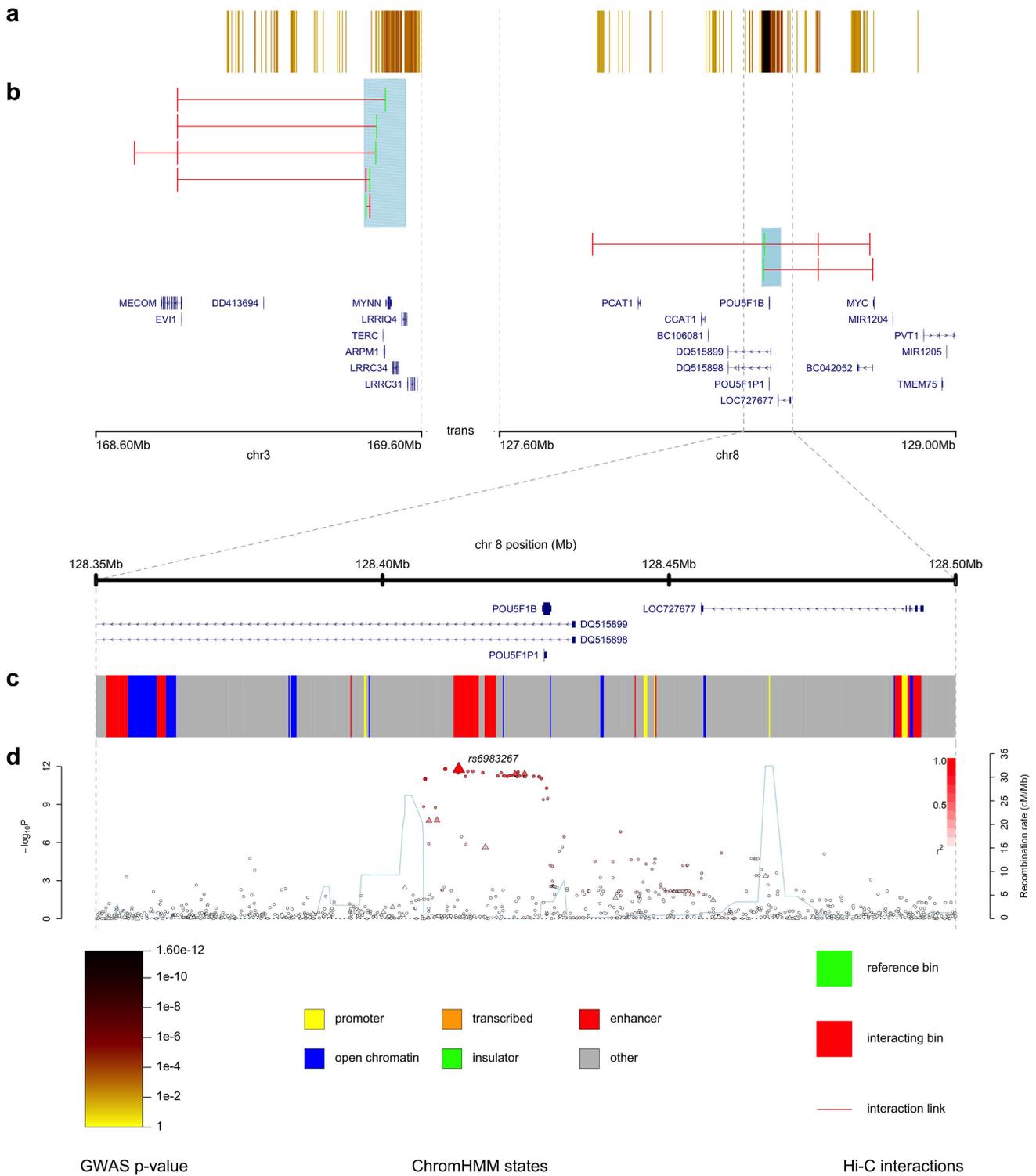


Figure 1. Annotated CRC GWAS p-values for 3q26.2 and 8q24.21, using data from [14] and [15]. The data tracks plotted across the two regions, 3q26.2 and 8q24.21, are **a.** association trend test p-values from a CRC GWAS as a colour intensity track, **b.** significant Hi-C interactions (3 kb resolution; determined from Hi-C experiments on 3 CRC cell-lines, LS174T, LoVo and Colo205), **c.** ChromHMM functional annotation, **d.** CRC GWAS trend test association p-values as a SNAP plot (with SNP type (imputed/typed), r^2 values and recombination rate). The figure includes a title, gene tracks for both scales and a legend.
doi:10.1371/journal.pone.0107497.g001

3q26.2 & 8q24.21

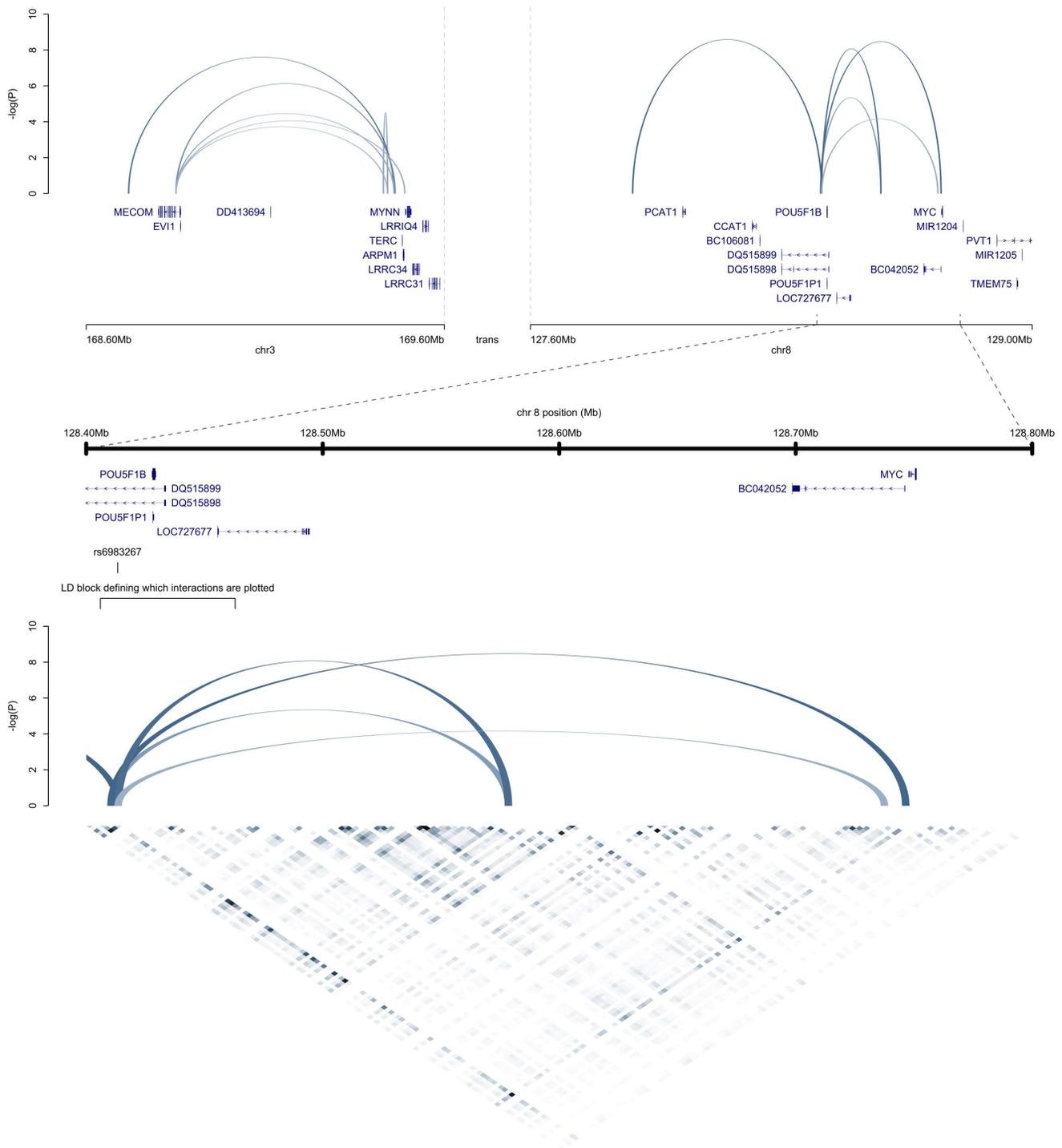


Figure 2. Hi-C interactomes of 3q26.2 and 8q24.21, using data from [15]. This figure shows the same two regions as Figure 1, 3q26.2 and 8q24.21. The data tracks plotted are the significant 3 kb Hi-C interactions (determined from 3 CRC cell-lines, LS174T, LoVo, Colo205), displayed as arches this time, as well as the corresponding, un-normalised Hi-C reads from the LS174T cell-line as a heat map. Also shown are some manual annotations to highlight the LD block within which one end of the significant interactions have to lie to be plotted, as well as the top CRC risk associated SNP in the LD block, rs6983267.

doi:10.1371/journal.pone.0107497.g002

Table 1. Comparison of visPIG with other genetic data visualisation tools.

application	multi-track	multi-region	multi-scale	type ^d	interactive	format ^b	legend ^c	graphics ^d
visPIG	✓	✓	✓	w,c	✓ ^e	pdf, png	✓	v,r
WashU Epigenome Browser	✓	✓ ^f	✓ ^f	w	✓	svg		v
IGV	✓		✓ ^g	w,g	✓	png		r
UCSC Genome Browser	✓		✓ ^g	w	✓	pdf, png		v,r
Circos	✓	✓	✓	c		svg, png		v,r
SNAP plot				w, c		pdf, png	✓	v,r

The comparison has been restricted to the capabilities of visPIG. Some of the other tools have additional features that are not listed here.

^a: w = web application, g = locally run graphical user interface, c = locally run from a terminal command line

^b: figure output format

^c: ability to add a legend onto the output figure

^d: v = vector graphics, r = raster graphics

^e: possible for some features (e.g. all graphical parameters, the type and order of tracks displayed, the zoom view), but not fully interactive; e.g. to change the plotted regions, the region file needs to be edited and re-uploaded

^f: multiple panels can be juxtaposed, which can be used to either show multiple regions and/or zoom in on a specific sub-region; difficult to get different regions on exactly identical scales; we were unable to indicate the location of the zoomed panel on the full view

^g: multi-scale in the sense that one can interactively zoom in on a region, but never more than one scale displayed at a given time.

doi:10.1371/journal.pone.0107497.t001

Interactivity, as required for data exploration, was of secondary importance.

Figures 1 and 2, which have been generated using only visPIG without additional image processing, showcase the utility of the software.

Figure 1 shows association p-values from a colorectal cancer (CRC) GWAS across two regions (3q26.2 and 8q24.21), annotated with Hi-C and ChromHMM data. The figure demonstrates all three main features of visPIG: i) ability to plot two distinct regions from two different chromosomes, ii) display of multiple data types and iii) zooming in on one specific sub-region. Furthermore, the software allows a title, a legend and track letters for easily referencing the individual data tracks in the caption. Figures 1a & 1b show how visPIG allows direct comparison of the GWAS p-values and the 3kb Hi-C interactomes of 3q26.2 and 8q24.21. Note that the two regions are shown on exactly the same scale. Zooming in on the region directly surrounding the highest CRC association signal, rs6983267, makes it possible to study the association data in more detail for the LD block containing rs6983267 (Figure 1d), and to highlight the functional features in this region (Figure 1c).

All input files and parameter values necessary to produce Figure 1 can be downloaded from the example section of the visPIG website.

Figure 2 shows the same two regions as in Figure 1, but displays only Hi-C data. The interactions are plotted in a different style and visPIG's heatmap and feature annotation tracks are showcased. Plotting Hi-C interactions as arches, allows to display the strength of individual interactions; here the p-values of the interactions determine the height and colour intensity of the arches. Zooming in on a sub-region shows the raw, binned Hi-C data that has been used to determine which interactions are significant.

There are a number of visualisation programs and toolkits available. These include the WashU Epigenome Browser [3], the Integrative Genomics Viewer (IGV) [4,5], the UCSC Genome Browser [2], Circos [6] and the Broad Institute's SNAP regional association plotting interface [12]. Table 1 lists the capabilities of these applications. While some applications, notably IGV, the WashU Epigenome Browsers and the UCSC Genome Browser, are better suited for data exploration, visPIG is the only application that can produce publication-ready graphs with the above mentioned capabilities. Circos has similar capabilities (multi-track, -region & -scale), but has no interface, making it less accessible to users not familiar with executing programs from a terminal, legends have to be added manually and all figures produced by Circos are circular, which may not be appropriate for all situations. The WashU Epigenome Browser is an extremely powerful data exploration and visualisation tool and also has similar capabilities than visPIG. It is multi-track and multiple panels can be juxtaposed, which allows plotting several regions at different scales. In visPIG's case the different regions are always plotted at exactly the same scale, with only zoomed regions being at a different scale. The Epigenome Browser is more flexible, but in practice it is difficult to get different regions to be plotted at exactly the same scale, and we were unable to indicate where the zoomed panel is located on the full view. The panels will also share the same tracks, whereas for a zoomed view one might want to highlight different data. Legends are available, but are not included on the output figure, onto which they need to be added manually.

We conclude that apart from directly programming one's own toolkit using R, Matlab®, Python or similar, we are not aware of any other software that combines visPIG's three core capabilities

(plotting multiple datasets for multiple regions at multiple scales), can add legends directly onto the output figure and has an easy-to-use graphical user interface.

Materials and Methods

The main visPIG code has been written in R. The complete R code can be downloaded from the visPIG website and does not require installation of any additional R libraries. The R code is intended to be run locally from the command line (i.e. without any graphical user interface as provided on the webpage). Running visPIG locally is required for users with sensitive data that cannot be uploaded to the visPIG server, or with very large data files. The R code is provided freely under a GPLv3 license.

The web application is generated by the R library shiny [13], and hosted on a linux server running the server software shiny-server, which has been developed for hosting R shiny applications. The menu layout has been designed using html and css.

References

1. R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
2. Kent W, Sugnet C, Furey T, Roskin K, Pringle T, et al. (2002) The human genome browser at UCSC. *Genome Research* 12: 996–1006.
3. Zhou X, Maricque B, Xie M, Li D, Sundaram V, et al. (2011) The human epigenome browser at washington university. *Nature Methods* 8: 989–990.
4. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative genomics viewer (IGV): highperformance genomics data visualization and exploration. *Briefings in Bioinformatics* 14: 178–192.
5. Robinson JT, Thorvaldsdóttir H, Winckler W, Gutman M, Lander ES, et al. (2011) Integrative genomics viewer. *Nature Biotechnology* 29: 24–26.
6. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: An information aesthetic for comparative genomics. *Genome Research* 19: 1639–1645.
7. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9: 215–216.
8. Zhao Z, Tavosoidana G, Sjolinder M, Gondor A, Mariano P, et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics* 38: 1341–1347.
9. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, et al. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics* 38: 1348–1354.
10. Lieberman-Aiden E, van Berkum N, Williams L, Imakaev M, Ragozy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289–293.
11. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 462: 58–64.
12. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a webbased tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24: 2938–2939.
13. RStudio, Inc (2013) shiny: Web Application Framework for R. URL <http://CRAN.R-project.org/package=shiny>. R package version 0.6.0.
14. Whiffin N, Dobbins S, Hosking F, Palles C, Tenesa A, et al. (2013) Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Human Molecular Genetics* 22: 5075–5082.
15. Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, et al. (2014) Capture Hi-C (cHi-C) identifies the chromatin interactome of colorectal cancer risk loci. Submitted.

Acknowledgments

We wish to thank the ICR Scientific Computing Team for funding the visPIG server. Specifically we want to thank Jon Lockley, Peter Thompson and Mark Allen for their assistance with building and configuring the server hosting visPIG.

The visPIG SNAP track makes use of modified R code from the Broad Institute's SNAP regional association plotting script, freely available online (<http://www.broadinstitute.org/mpg/snap/ldplot.php>).

Author Contributions

Contributed reagents/materials/analysis tools: MH MS RJ GM RSH. Contributed to the writing of the manuscript: MH. Design of visPIG: MH. R code: MH MS. Web interface: MS MH. Further input and feedback: GM RJ RSH. Obtaining of funding: RSH. Project supervision: MH.