HilbertCurve: an R/Bioconductor package for high-resolution visualization of genomic data

Zuguang Gu[1,2], Roland Eils[1,2,3] and Matthias Schlesner[1,*]

[1]Division of Theoretical Bioinformatics, [2]Heidelberg Center for Personalized Oncology (DKFZ-HIPO), German Cancer Research Center (DKFZ), Heidelberg, Germany and [3]Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology (IPMB) and BioQuant, Heidelberg University, Heidelberg, Germany

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Abstract

Summary: Hilbert curves enable high-resolution visualization of genomic data on a chromosome- or genome-wide scale. Here we present the *HilbertCurve* package that provides an easy-to-use interface for mapping genomic data to Hilbert curves. The package transforms the curve as a virtual axis, thereby hiding the details of the curve construction from the user. *HilbertCurve* supports multiple-layer overlay that makes it a powerful tool to correlate the spatial distribution of multiple feature types.

Availability and implementation: The *HilbertCurve* package and documentation are freely available from the Bioconductor project: http://www.bioconductor.org/packages/devel/bioc/html/HilbertCurve.html

Contact:m.schlesner@dkfz.de

1 Introduction

Hilbert curves can be used to effectively visualize genomic data, because they can provide a global overview of genome-scale datasets while still revealing the spatial distribution of features at high resolution (Anders, 2009). Hilbert curve visualization has been utilized to compare the sequence difference between human and other primates (Wong, 2014), to show the spatial organization of different chromatin states (Kharchenko *et al.*, 2011), and to investigate the genome-wide distribution of histone modifications (Anders 2009, Henry *et al.*, 2012). The Hilbert curve (Hilbert, 1891) is a space-filling curve, i.e. a continuous mapping of one-dimensional axis to a two-dimensional area. It is constructed recursively by dividing the two-dimensional square (and in subsequent iterations the sub-squares) into four quadrants and folding the one-dimensional line into a U-shape to traverse all quadrants. After k iterations, the two-dimensional space is split into $2^k \times 2^k$ grids and the one-dimensional line folded into $4^k - 1$ segments. The iteration level k of the Hilbert curve controls the degree of folding, with a higher k meaning a longer but also more densely packed curve in a given space (Anders, 2009). For example, while in a linear visualization of human chromosome 1 of 1024 pixel length each pixel represents 243 kb, in a Hilbert curve visualization of 1024×1024 pixels each pixel represents only 238 bp, which greatly increases the resolution by 1000 times. Importantly, the Hilbert curve preserves the locality of data points, meaning that if two data points are in proximity on the one-dimensional axis, they are also close to each other in two-dimensional space (see demonstration in Supplementary File S1; it should be noted that the reverse is not always true and some points are close in the two-dimensional space while they are distant on the genome).

Here we present *HilbertCurve*, an R/Bioconductor package for genomic data visualization using Hilbert curves. *HilbertCurve* hides details of the curve construction and provides methods to add graphics only based on genomic positions. This enables smooth integration of *HilbertCurve* in genomic analysis and distinguishes the package from currently available tools such as *HilbertVis* (Anders, 2009). Further advantages of *HilbertCurve* over other tools are the capability to visualize multiple chromosomes simultaneously and

support of multiple-layer overlay, a powerful method to investigate correlations between different features. With these functionalities, the *HilbertCurve* package will greatly facilitate the visualization and interpretation of the ever increasing number of genome-wide datasets generated by next-generation sequencing and other omics techniques.

2 Implementation

The *HilbertCurve* package is implemented in an object-oriented manner. The package exports the classes `HilbertCurve` and `GenomicHilbertCurve` that encapsulate the structure of the curve and provide methods for mapping genomic positions onto the curve. From the users' perspective the curve serves as a virtual axis, and low-level graphics such as points and lines can be added by specifying corresponding genomic positions either by numeric vectors or `GRanges` objects (Lawrence *et al.*, 2013).

There are two operation modes that compromise between the level of details and visual complexity to address different visualization needs. The 'normal' mode is suitable for low-resolution visualization for which the structure of the curve is visible (Fig. 1A, B). Low-level graphic functions (e.g. `hc_points()`) can be applied to customize the curve. In 'pixel' mode, segments on the curve are degenerated as single pixels, which allows much higher resolution visualization (Fig. 1C–E). For example, one pixel represents approximately 238 bp when the full length of human chromosome 1 is mapped to a Hilbert curve with iteration level 10.

The package supports layers to visualize the relationship between multiple features simultaneously in one plot (Fig. 1C, E). In addition, self-defined color overlay allows changing colors for overlapping regions of two layers, to assist the visualization of co-localization of genomic features (Fig. 1C, E). Colors as fundamental graphical representation of input values on the curve can be flexibly defined by the user (see Supplementary File S2 for a demonstration).

Each graphic unit (pixel under 'pixel' mode or e.g. point under 'normal' mode) represents a genomic interval. *HilbertCurve* provides three averaging metrics to summarize the values associated with the interval to account for different characteristics of the plotted data (explained in detail in Supplementary File S3).

3 Application

Hilbert curve visualization of genomic data can be used to assess the spatial distribution, size distribution and co-localization of genomic features together in a single plot. Figure 1B illustrates sequence conservation for human chromosome 1 compared to zebrafish. The chromosome is mapped to a Hilbert curve with level 6 so that each point represents a window of 61 kb. The plot reveals large highly conserved regions that alternate with regions of low conservation and non-conserved regions. For comparison, Supplementary File S4 also illustrates conservation between mouse and human.

Histone modifications (e.g. methylation, acetylation) affect the chromatin–DNA interaction and thus play a role in transcription regulation. Figure 1C visualizes the H3K36me3 modification in human chromosome 1. The curve is initialized with level 9 and illustrated under 'pixel' mode. While H3K36me3 is indicated in red, gene bodies are illustrated as additional layer in grey. Overlapping regions are adjusted in purple to visualize the strong enrichment of

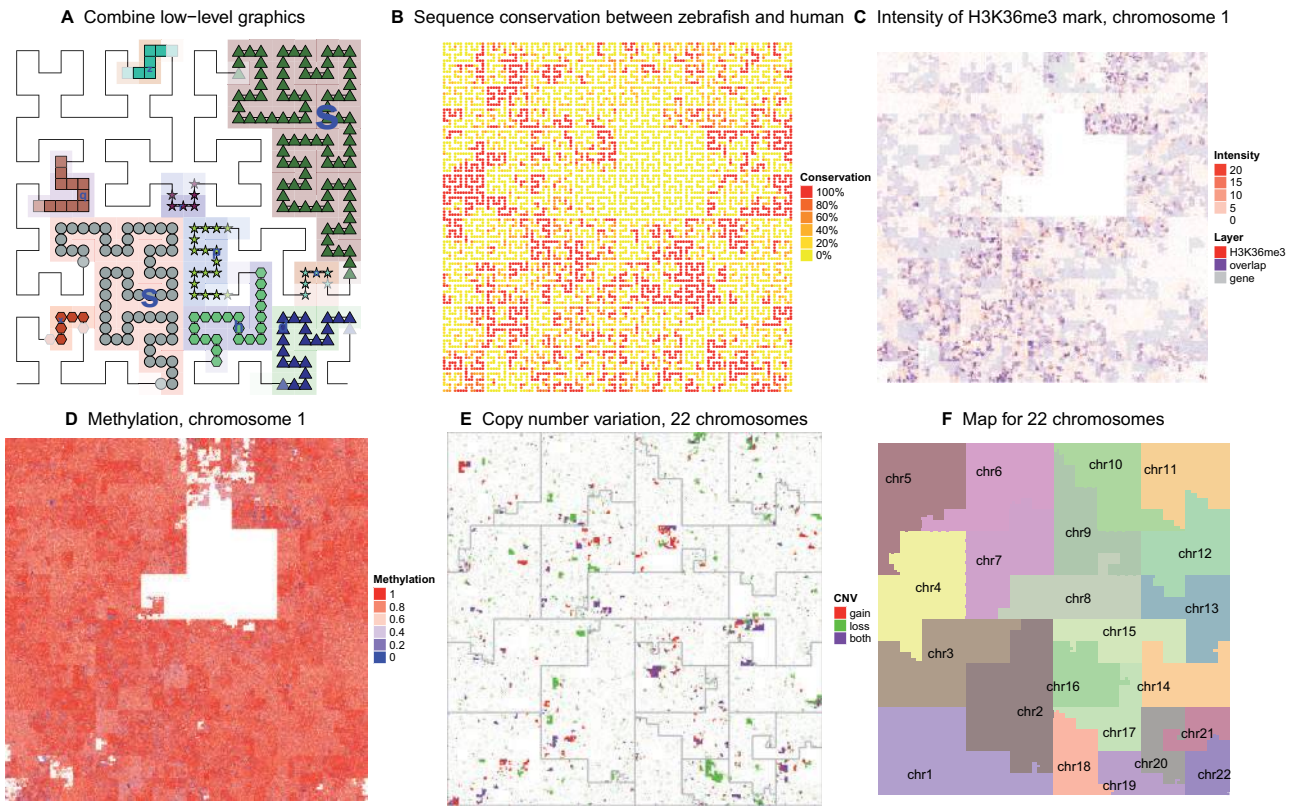

Fig. 1. Example plots generated with *HilbertCurve*. (**A**) Combining low-level graphics to construct a complex plot. (**B**) Sequence conservation between human chromosome 1 and the zebrafish genome. (**C**) H3K36me3 modification on human chromosome 1 overlaid with gene regions. To indicate co-localization, the coloring for the overlapping regions was changed to purple. (**D**) DNA methylation on human chromosome 1. (**E**) Copy number gain and loss for human chromosome 1–22. (**F**) A map for E showing the areas of the 22 chromosomes in the curve. Data sources and R code for all plots can be found in Supplementary File S6

H3K36me3 in gene bodies. Supplementary File S5 contains similar visualization for H3K27ac, H3K4me3 and H3K9me3 modifications, which reveal the different distribution of these marks, indicating different regulation patterns.

Figure 1D gives a high-resolution visualization of DNA methylation in human chromosome 1. With a curve of level 9, each pixel represents approximately 950 bp which is sufficient to visualize even small CpG islands. Most of the chromosome is methylated (red and light red areas) and no large unmethylated regions exist. However, a large number of narrow unmethylated regions (blue dots) are spread over the chromosome. Additionally, the chromosome contains a high proportion of partially methylated domains (PMDs, visible as light red areas). PMDs are relatively long regions of intermediate methylation, which are related to the formation of repressive histone states and suppress gene expression (Hon *et al.*, 2012). Interestingly, most of the narrow unmethylated regions are located in otherwise highly methylated areas and not in PMDs.

Figure 1E illustrates copy number variations (CNVs) for the 22 human autosomes (Zarrei *et al.*, 2015). The figure is accompanied by a map indicating the chromosome areas (Figure 1F). Several large regions that are affected by CNVs become apparent, and many of them are affected by both gains and losses (highlighted in purple by overlap color adjustment). In addition, there are many small regions of loss, but only few small regions of gain or of loss and gain.

In conclusion, the *HilbertCurve* package offers an easy way to provide a global view on genomic data while still preserving details at high resolution. It hereby greatly assists discoveries from genomics and other omics datasets.

Funding

This work was supported by the German Cancer Research Center-Heidelberg Center for Personalized Oncology (DKFZ-HIPO) and the BMBF-funded de.NBI HD-HuB network (#031A537A, #031A537C).

Conflict of Interest: none declared.

References

Anders,S. (2009) Visualization of genomic data with Hilbert Curve. *Bioinformatics*, **25**, 10.

Henry,G. *et al.* (2012) Cell type-specific genomics of Drosophila neurons. *Nucleic Acid Res.*, **40**, 19.

Hilbert,D. (1891) Über stetige Abbildungen einer Linie auf ein Flächenstück. *Math. Annal.*, **38**, 459.

Hon,H.C. *et al.* (2012) Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.*, **22**, 2.

Kharchenko,P. *et al.* (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, **471**, 7339.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, 8.

Wong,K. (2014) Tiny genetic difference between humans and other primates pervade the genome. *Sci. Am.*, **311**, 3.

Zarrei,M. *et al.* (2015) A copy number variation map of the human genome. *Nat. Rev. Genet.*, **16**, 3.