A peer-reviewed version of this preprint was published in PeerJ on 20 January 2020.

Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, Chen J. 2020. *RIdeogram*: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Computer Science 6:e251
https://doi.org/10.7717/peerj-cs.251

RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms

Zhaodong Hao [1, 2], **Dekang Lv** [3], **Ying Ge** [3], **Jisen Shi** [1], **Dolf Weijers** [2], **Guangchuang Yu** [Corresp., 4], **Jinhui Chen** [Corresp. 1]

[1] Key Laboratory of Forest Genetics & Biotechnology of Ministry of Education, Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China

[2] Laboratory of Biochemistry, Wageningen University and Research, Wageningen, Netherlands

[3] Institute of Cancer Stem Cell, Dalian Medical University, Dalian, China

[4] Institute of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China

Corresponding Authors: Guangchuang Yu, Jinhui Chen
Email address: gcyu1@smu.edu.cn, chenjh@njfu.edu.cn

Background: Owing to the rapid advances in DNA sequencing technologies, whole genome from more and more species are becoming available at increasing pace. For whole-genome analysis, idiograms provide a very popular, intuitive and effective way to map and visualize the genome-wide information, such as GC content, gene and repeat density, DNA methylation distribution, etc. However, most available software programs and web servers are available only for a few model species, such as human, mouse and fly. As boundaries between model and non-model species are shifting, tools are urgently needs to generate idiograms for a broad range of species are needed to help better understanding fundamental genome characteristics.

Results: The R package *RIdeogram* allows users to build high-quality idiograms of any species of interest. It can map continuous and discrete genome-wide data on the idiograms and visualize them in a heat map and track labels, respectively.

Conclusion: The visualization of genome-wide data mapping and comparison allow users to quickly establish a clear impression of the chromosomal distribution pattern, thus making *RIdeogram* a useful tool for any researchers working with omics.

RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms

Zhaodong Hao[1,2], Dekang Lv[3], Ying Ge[3], Jisen Shi[1], Dolf Weijers[2], Guangchuang Yu[4] and Jinhui Chen[1]

[1]Key Laboratory of Forest Genetics & Biotechnology of Ministry of Education, Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing Jiangsu, China
[2]Laboratory of Biochemistry, Wageningen University, Wageningen, Haarlem, The Netherlands
[3]Institute of Cancer Stem Cell, Dalian Medical University, Dalian, Liaoning, China
[4]Institute of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, Guangdong, China

Corresponding Author:
Guangchuang Yu[4]
1023 ShaTaiNan Road, Guangzhou, Guangdong, 510515, China
Email address: gcyu1@smu.edu.cn
Jinhui Chen[1]
159 LongPan Road, Nanjing, Jiangsu, 210037, China
Email address: chenjh@njfu.edu.cn

Abstract

Background: Owing to the rapid advances in DNA sequencing technologies, whole genome from more and more species are becoming available at increasing pace. For whole-genome analysis, idiograms provide a very popular, intuitive and effective way to map and visualize the genome-wide information, such as GC content, gene and repeat density, DNA methylation distribution, etc. However, most available software programs and web servers are available only for a few model species, such as human, mouse and fly. As boundaries between model and non-model species are shifting, tools are urgently needs to generate idiograms for a broad range of species are needed to help better understanding fundamental genome characteristics.
Results: The R package *RIdeogram* allows users to build high-quality idiograms of any species of interest. It can map continuous and discrete genome-wide data on the idiograms and visualize them in a heat map and track labels, respectively.
Conclusion: The visualization of genome-wide data mapping and comparison allow users to quickly establish a clear impression of the chromosomal distribution pattern, thus making *RIdeogram* a useful tool for any researchers working with omics.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

Introduction

Recently, with the development of sequencing technologies, especially rapid advances in third generation sequencing (Pacific Biosciences and Oxford Nanopore Technologies), BioNano genome mapping and High-throughput chromatin conformation capture sequencing, many sequenced species have their genomes updated to chromosome level and more and more non-model species have their genomes sequenced (Jiao & Schneeberger 2017; Phillippy 2017). After the chromosome-level genome completion, an overview of some genome characteristics can help to better understand a species genome, such as gene and transposon distribution across the sunflower genome (Badouin et al. 2017).

A idiogram, also known as a karyotype, is defined as the phenotypic appearance of chromosomes in the nucleus of a eukaryotic cell and has been widely used to visualize the genome-wide data since the first web server, *Idiographica*, came online in 2007 (Kin & Ono 2007). However, this web server (updated November, 2017) still only caters to four species: human, mouse, rat and fly. Recently, an R package called *chromoMap* was published in CRAN and this allows users to interactively visualize and map chromosome elements on the chromosome plot. In addition, there are two JavaScript libraries for chromosome visualization, one is Ideogram.js and the other is *karyotypeSVG*. We still lack the choice of available drawing tools for plotting idiograms conveniently and effectively with a wide range species, despite the widespread need to visualize features along entire chromosomes.

Scalable Vector Graphics (SVG) is a language for describing two-dimensional graphics applications and images. SVG graphics is defined in an eXtensible Markup Language (XML) text file which means that one can easily use any text editor or drawing software to create and edit SVG graphics. Most R graphics packages are built on two graphics systems, the traditional graphics system and the grid graphics system. Here, we developed an R package (*RIdeogram*) to draw high-quality idiograms without species limitations, that allows to visualize and map whole-genome information on the idiograms based on the SVG language.

Description

The package *RIdeogram* is written in R (R Core Team, 2018), one of the most popular programming languages widely used in statistical computing, data analytics and graphics. However, this new R graphics package is not built based on any existing graphics systems. We use the R environment to read the custom input files and calculate the drawing element positions in a coordinate system. Then, we use R to write all element information into a text file following the XML format which are used to define graphics by the SVG language. A list of the currently implemented commands is given in Table 1. In general, there are three main functions, *data*, *ideogram* and *convertSVG* implemented in the package *RIdeogram*. Users can use the function *data* to load the example data or the basic R function *read.table* to load the custom data from local files. Then, the function *ideogram* can be used to compute the information for all drawing elements based on the input files and generate a A4-sized SVG file containing a vector graphic which can be conveniently viewed and modified using the software Adobe Illustrator or

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

Inkscape. Alternatively, users can also use the function *convertSVG* to convert this SVG file into an adjustable image format (pdf, png, tiff, or jpg) with a user-defined resolution according to the practical requirements.

In general, there are two types of data, i.e., continuous and discrete data. For mapping and visualizing, *RIdeogram* considers the continuous data, such as gene density across the whole genome in 1-Mb windows, as overlaid features and maps them on the idiograms with dark/light colors representing high/low values. For the other data type that are scattered throughout the whole genome, such as the chromosomal distribution of members in one gene family, *RIdeogram* can add track labels next to the idiograms with three shapes (box, circle and triangle) available to represent different characteristics of these members, such as the subclade that one gene member belongs to. Users can also combine the shapes and colors to represent more than three distinct characteristic types.

RIdeogram is available through CRAN (https://cran.r-project.org/web/packages/RIdeogram/) and is developed on GitHub (https://github.com/TickingClock1992/RIdeogram). Further extensions in development and fixes can be seen in the issue listing page on the package's GitHub page. The new function that we are planning to implement in next version include, but are not limited to, linking genome regions on two adjacent idiograms with Bezier curves or straight lines for synteny visualization and enlarging the user-specified genome regions to display detailed characteristics, as we gather more from users.

Examples

Our first example use the data contained in this package. After the completion of genome sequencing, assembly and annotation, *RIdeogram* can be used to give some idea of how genes are distributed across the whole genome. The example data contained numbers of protein-coding genes calculated in 1-Mb windows which can be considered as continues data and positions of 500 random selected non-coding RNAs, including ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and microRNAs (miRNAs), which can be considered as discrete data. *RIdeogram* maps the gene density information on the idiograms as overlaid features in a heat map and adds track labels next to the idiograms with green boxes, purple circles and orange triangles representing rRNAs, tRNAs and miRNAs, respectively (Fig. 1). Obviously, inter- and intra-chromosomal gene distributions are non-uniform. For instance, the chromosomal regions adjacent to the centromeres are gene-poor in chromosome 1, 9 and 16 while those are gene-rich in chromosome 11, 14 and 17. This function can be applied to many different situations, such as single nucleotide polymorphism (SNP) density and candidate markers (Fig. S1 & Data S1, original data see Li et al. 2019), DNA methylation dynamics and potential activated genes (Fig. S2 & Data S2, original data see Huang et al. 2019) and transcription factor (TF) binding sites and candidate target genes (Fig. S3 & Data S3, original data see Shamimuzzaman & Vodkin 2013).

Besides visualizing some specific genome characteristics across the whole genome at the chromosome level as showed in Fig. 1, *RIdeogram* can also be used to compare two relevant genome features, such as gene and repeat density, which will provide some important

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

implications for better understanding the relevance of chromosomal distribution patterns of these two features. The example data implemented in this package also contained the information of long terminal repeat (LTR) distribution across the human genome. Since the transposable elements have been suggested to have a potential detrimental effect on gene expression (Hollister & Gaut 2009), the distributions of gene and LTR are supposed to be opposite across the whole genome as a result of natural selection. As expect, the region that has a relatively high gene content usually has a relatively low LTR density and vice versa (Fig. 2), indicating that LTR seems to avoid inserting in the regions with a high gene content in the genome. This similar phenomenon was also observed in the sunflower genome explained using two idiogram graphics, one showing the gene distribution and the other showing the LTR distribution (Badouin et al. 2017). Using *RIdeogram*, users can integrate these two graphics into one, much easier for researchers to interpret and readers to understand. Apart from the differences, this function can also be used to show the similarities, like the similar genetic diversity patterns across the whole genome between two geographical groups of the same species (Fig. S4 & Data S4, original data see Chen et al. 2019).

Conclusion

The *RIdeogram* package provides an efficient and effective way to build idiograms with no species limitations and map genome-wide information on the idiograms for better visualizing and understanding the chromosomal distribution patterns of some particular genomic features. Meanwhile, this package is user-friendly and accessible for biologists without extensive computer programming expertise. In addition, *RIdeogram* can generate two types of images, a vector graphic or a bitmap file, both in high-quality and meeting conventional requirements for direct use in presentations or journal publications.

Acknowledgements

We thank Dr. Zhongjuan Zhang for her comments on the manuscript.

References

Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Briere C, Owens GL, Carrere S, Mayjonade B, Legrand L, Gill N, Kane NC, Bowers JE, Hubner S, Bellec A, Berard A, Berges H, Blanchet N, Boniface MC, Brunel D, Catrice O, Chaidir N, Claudel C, Donnadieu C, Faraut T, Fievet G, Helmstetter N, King M, Knapp SJ, Lai Z, Le Paslier MC, Lippi Y, Lorenzon L, Mandel JR, Marage G, Marchand G, Marquand E, Bret-Mestries E, Morien E, Nambeesan S, Nguyen T, Pegot-Espagnet P, Pouilly N, Raftis F, Sallet E, Schiex T, Thomas J, Vandecasteele C, Vares D, Vear F, Vautrin S, Crespi M, Mangin B, Burke JM, Salse J, Munos S, Vincourt P, Rieseberg LH, and Langlade NB. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 546:148-152. DOI: 10.1038/nature22380

Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y, Xu H, Xie H, Long X, Zhang J, Wang Z, Shi M, Lu Y, Liu S, Guan L, Zhu Q, Yang L, Ge S, Cheng T, Laux T, Gao Q, Peng Y, Liu N, Yang S, and Shi J. 2019. Liriodendron genome sheds

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

light on angiosperm phylogeny and species-pair differentiation. *Nature Plants* 5:18-25. DOI: 10.1038/s41477-018-0323-6

Hollister JD, and Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research* 19:1419-1428. DOI: 10.1101/gr.091678.109

Huang H, Liu R, Niu Q, Tang K, Zhang B, Zhang H, Chen K, Zhu JK, and Lang Z. 2019. Global increase in DNA methylation during orange fruit development and ripening. *Proceedings of the National Academy of Sciences of the United States of America* 116:1430-1436. DOI: 10.1073/pnas.1815441116

Jiao WB, and Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology* 36:64-70. DOI: 10.1016/j.pbi.2017.02.002

Kin T, and Ono Y. 2007. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* 23:2945-2946. DOI: 10.1093/bioinformatics/btm455

Li X, Singh J, Qin M, Li S, Zhang X, Zhang M, Khan A, Zhang S, and Wu J. 2019. Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (Pyrus). *Plant Biotechnology Journal* 17:1582-1594. DOI: 10.1111/pbi.13085

Phillippy AM. 2017. New advances in sequence assembly. *Genome Research* 27:xi-xiii. DOI: 10.1101/gr.223057.117

Shamimuzzaman M, and Vodkin L. 2013. Genome-wide identification of binding sites for NAC and YABBY transcription factors and co-regulated genes during soybean seedling development by ChIP-Seq and RNA-Seq. *BMC Genomics* 14:477. DOI: 10.1186/1471-2164-14-477

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

Table 1(on next page)

Table 1: Functions and data contained in the package *RIdeogram*.

Table 1. Functions and data contained in the package *RIdeogram*.

Function/Data name	Description
data(human_karyotype)	Load a data frame of the human karyotype information including the length information of each chromosome and the position information of each centromere.
data(gene_density)	Load a data frame of the gene number calculated in 1-Mb windows across the whole human genome.
data(LTR_density)	Load a data frame of the LTR number calculated in 1-Mb windows across the whole human genome.
data(Random_RNAs_500)	Load a data frame of the positions of 500 random selected RNAs throughout the whole human genome.
ideogram	Map and visualize the genome-wide data on the idiograms
convertSVG	Convert the output file from the SVG format to the format users chose.
svg2tiff	Convert the output file from the SVG format to the TIFF format.
svg2pdf	Convert the output file from the SVG format to the PDF format.
svg2jpg	Convert the output file from the SVG format to the JPG format.
svg2png	Convert the output file from the SVG format to the PNG format.

Figure 1

Figure 1: Gene distribution across the whole human genome.

The overlaid heatmap shows the gene density and the tack labels refer to 500 random selected RNAs consisted of rRNAs (green boxes), tRNA (purple circles) and miRNA (orange triangles) locus across the human genome. Annotation information was downloaded from the GENCODE website (https://www.gencodegenes.org).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27928v1 | CC BY 4.0 Open Access | rec: 29 Aug 2019, publ: 29 Aug 2019

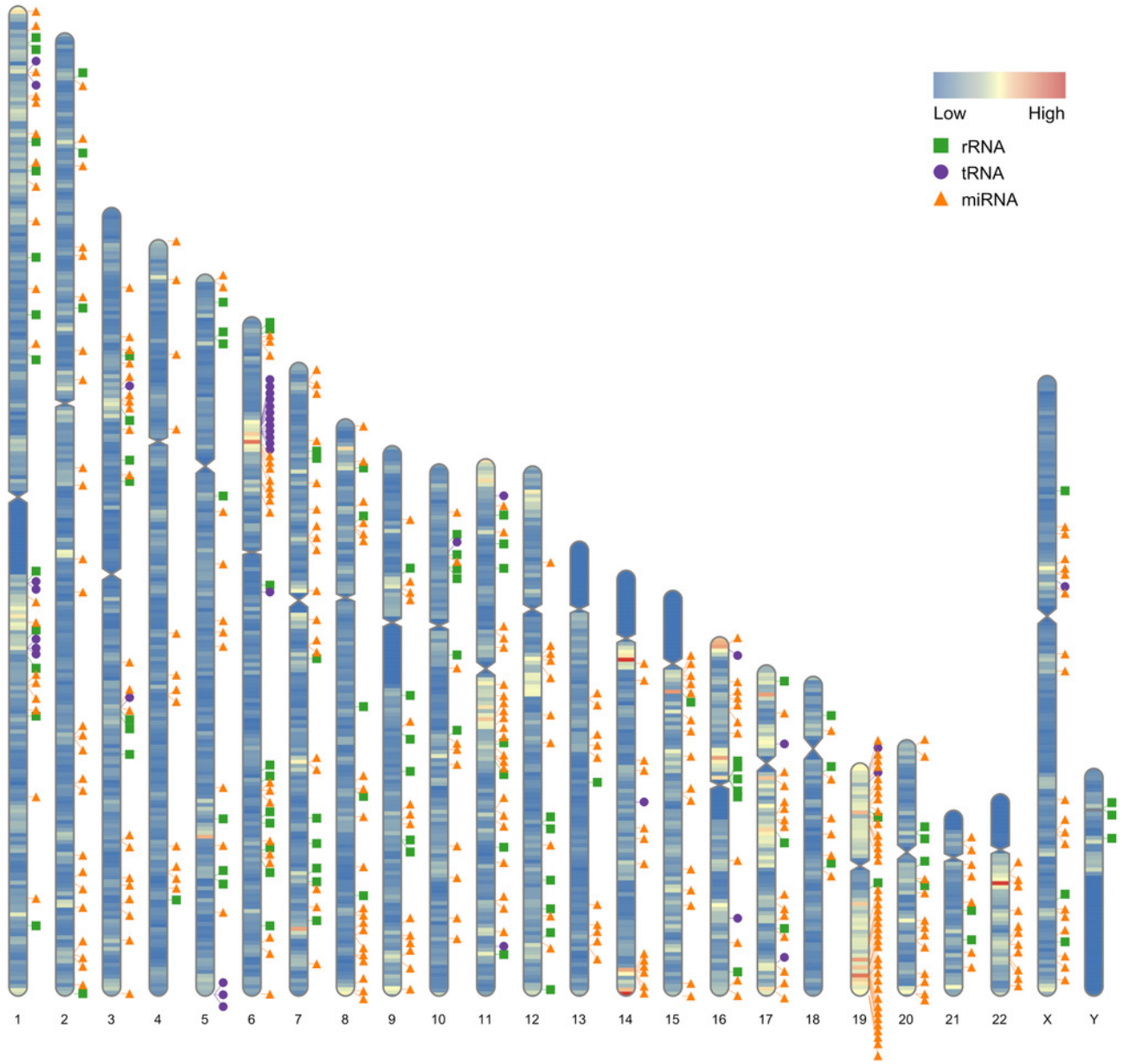

Figure 1. Gene distribution across the whole human genome.

The overlaid heatmap shows the gene density and the tack labels refer to 500 random selected RNAs consisted of rRNAs (green boxes), tRNA (purple circles) and miRNA (orange triangles) locus across the human genome. Annotation information was downloaded from the GENCODE website (https://www.gencodegenes.org).

Figure 2

Figure 2: A comparison of chromosomal distribution of genes and LTRs in the human genome.

The gene number and LTR number are both counted in a 1-Mb window. Red color represents the gene number (range 0–135 per Mb) and blue color represents the LTR number (range 0–606 per Mb). The light and dark colors represent a low and high content, respectively. This plot shows that gene and LTR have an opposite distribution pattern along the human chromosomes.

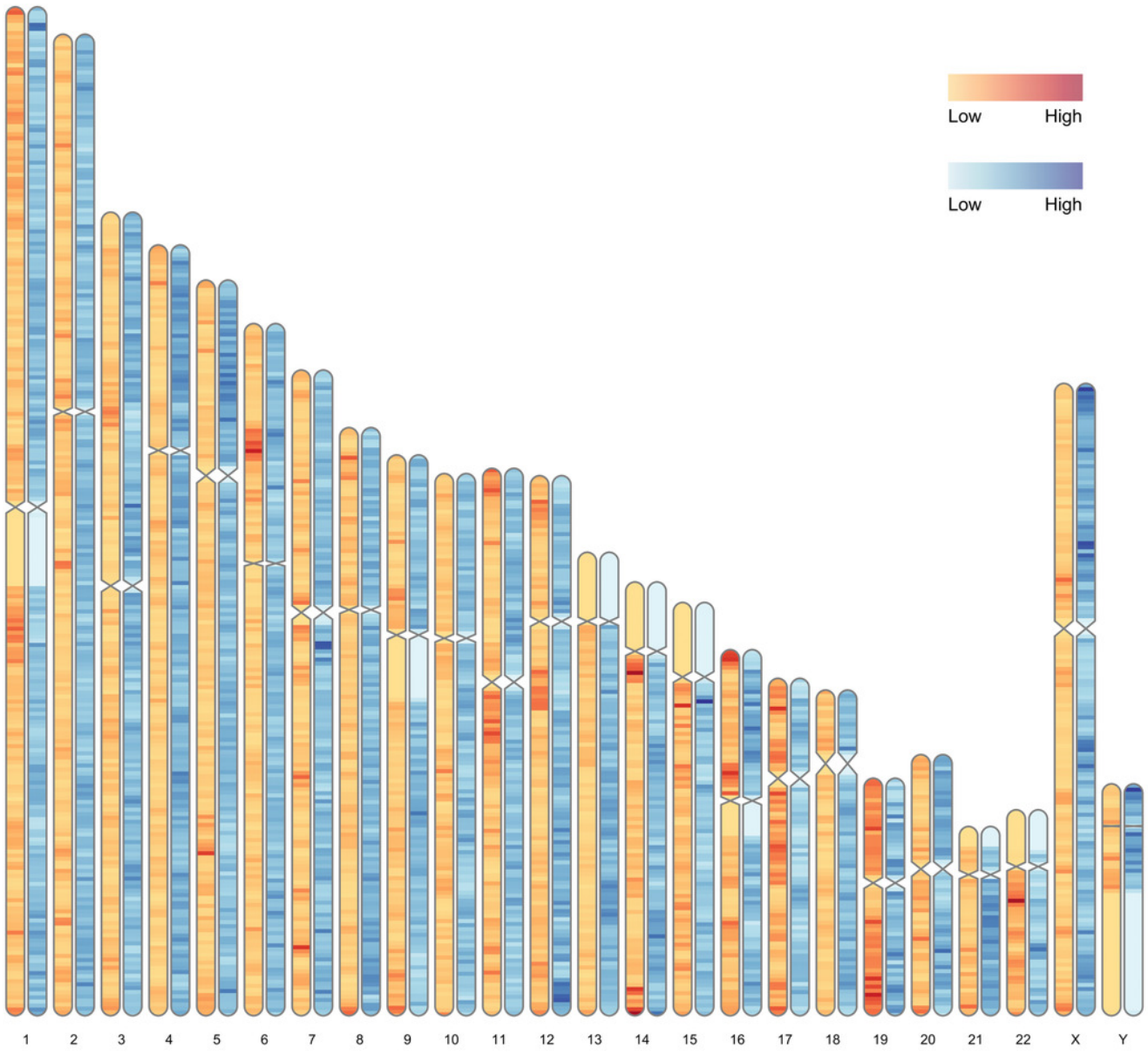

Figure 2. A comparison of chromosomal distribution of genes and LTRs in the human genome.

The gene number and LTR number are both counted in a 1-Mb window. Red color represents the gene number (range 0–135 per Mb) and blue color represents the LTR number (range 0–606 per Mb). The light and dark colors represent a low and high content, respectively. This plot shows that gene and LTR have an opposite distribution pattern along the human chromosomes.