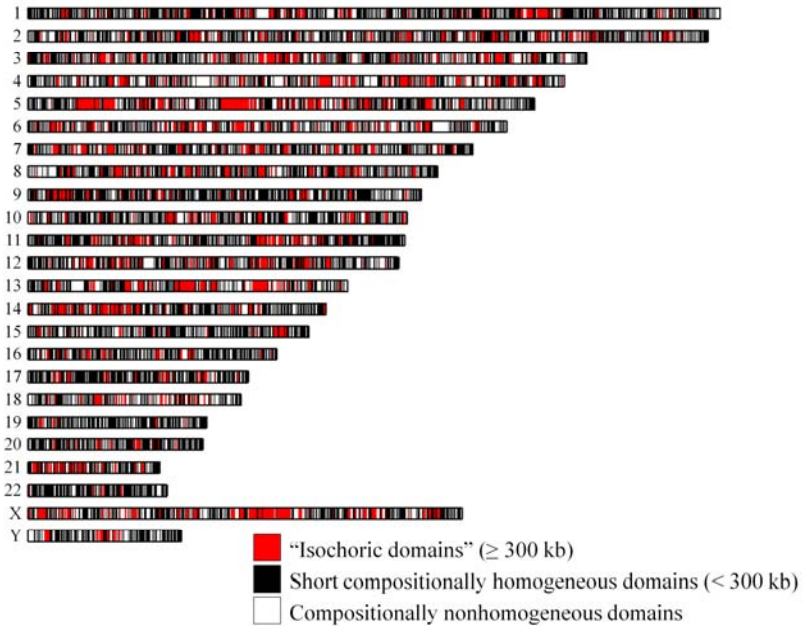
User Manual for IsoPlotter v2.3

Eran Elhaik

For Windows and Linux

January 21, 2013



Please do not hesitate to email all questions to: eelhaik@jhsph.edu

Contents

1. Introduction

IsoPlotter is a program for the inference of compositional domains and isochores. Compositional domains are genomic regions of DNA with a distinct guanine (G) and cytosine (C): G-C and C-G (collectively GC content). The homogeneity of compositional domains can be compared to that of the chromosome on which they reside labeling each compositional domain as homogeneous or nonhomogeneous domains. Compositionally homogeneous domains that are sufficiently long (≥ 300 kb) are termed isochores or "isochoric domains." For a given set of Fasta files, IsoPlotter infers the domains and outputs the range of each domain and related GC composition statistics.

For example, for chromosome 1 in Anopheles gambiae that starts with:
ATAGCTACATCGATCTCGCTAAGCTAGCATCGAGCTCAGCATCGACTAGCATCG…

The beginning of the output is:

Chromosome	Start	End	Size	Mean GC content	GC content STD	Is homogeneous
1	1	2720	2720	0.388	0.0985	0
1	2721	4832	2112	0.451	0.1181	1
1	4833	7104	2272	0.381	0.1151	0
1	7105	8672	1568	0.33	0.0922	0
1	8673	9728	1056	0.487	0.0888	0

Form the file: seg_no_ns_H.txt

2. Installation

IsoPlotter was developed in Matlab (R2011a). Source code is available as .m files under the Matlab_sourse_files folder. Executable files were compiled separately for Windows 32bit, Windows 64bit, and Linux 64bit.

2.1 If you have Matlab

Unzip the files to a folder and then add that folder to Matlab paths:
File->set path->add folder (or add with subfolders)->save->close

If you are running Matlab from the prompt, simply browse to that folder (using commands: 'ls','dir','cd' just like using terminal/cmd).

2.2 If you do not have Matlab

We created executable files that were compiled to three operation systems: Windows 32bit, Windows 64bit, and Linux 64bit. To be able to run these executable files, a Matlab component called MCRInstaller need to be installed on your machine. This component is specific to (1) your

operation system and (2) to the Matlab version that compiled it, so it is important to install the version suitable for your operation system according to this table:

Operating system	MCR version	Obtain from	Executable folder
Windows 32bit	R2011a	IsoPlotter website	Linux
Windows 64bit	R2012a	Matlab's website	Windows_executables\32bit
Linux glnxa64	R2012a	Matlab's website	Windows_executables\64bit

For example, if you have Windows 32bit, you'll need to install the MCR version R2011a, available from IsoPlotter website. Once installed, make sure MCRInstaller exists on your machine by running the executable files. Matlab will look for the MCRInstaller program in the path.

For more instructions on the installation see MCR_instructions.pdf. Please note, the MCRInstaller is a Matlab component, so further help should be addressed to their customer support (support@mathworks.com)

3. Input File Format

IsoPlotter accepts files in Fasta format, for example:

```
>chrX
gcgagctgcattgcttcgcaaataacacagctgagtttcctttagttaag
aggaacgttaaaacctctttttaataagtgaatatttagtgaatatatca
aatttctacaggtttctattgctgaagcttaatttacgtctgtatcaagt
aaaaataaaagttttgatgaaaatacgcaaaatTTACGTATTAAAAGCAT
TCGATTGTCGCTTGCGATTGTAGTTTGGAGACCCCTGTGATAGATCAACA
AGTATCAGAATTTAAAACTTAGACTCTGAAATATTCTCATTCCCGGCTCT
CTGAAGTTTCTCTCTTAGGCCGAAATTACACGGACCGGAATTTCGCGGCC
GCGGAATTCCGCCTGCAGAAATTTCTATGGATATGACAGTTCAGAAAAGT
TGTTGTTGACAGTTTGTACATGGGTTTGACAGCAGGCGAAATTACGCTCC
```

Each Fasta file should have one Fasta sequence. If your Fasta sequence contains multiple entries, use the utility Split_fasta_file.pl to break them to multiple files.

4. The IsoPlotter Pipeline

Unlike in previous version, where you had to run each file separately, here we created a pipeline to take care of that for you. The pipeline should be called with the following parameters:

RunIsoPlotterPipeline(input_dir, output_dir, domain_min_size, ns_domain_min_size, win_size)

Input_dir – the folder where all (and only) your fasta files are.
Output_dir – your output folder. The pipeline will eventually create subfolders with temporary files and a final combined file.
domain_min_size – Optional. The smallest domain size. Default is 3008bp.
ns_domain_min_size – Optional. If N's islands exist in your fasta files, this is the size of N's islands that should not be included in the compositional domain, but rather be shown as a special domain in the segmentation file.

win_size – Optional. The number of nucleotides for which the GC content would be calculated. default is 32bp.

4.1 Running the pipeline from Matlab

RunIsoPlotterPipeline(input_dir, output_dir, domain_min_size, ns_domain_min_size, win_size)

For example:
RunIsoPlotterPipeline('C:\Input\Human\', 'C:\Output\Human\')

4.2 Running the pipeline from Windows as an executable file

RunIsoPlotterPipeline input_dir output_dir domain_min_size ns_domain_min_size win_size

For example:
RunIsoPlotterPipeline Z:\Genomes\Human Z:\IsoPlotter\Human

4.3 Running the pipeline from Linux as an executable file

./RunIsoPlotterPipeline.sh <Matlab mcr folder> input_dir output_dir domain_min_size ns_domain_min_size win_size

For example,
./run_RunIsoPlotterPipeline.sh /usr/local/matlab/R2012a/ './Example/' './Output/'

5. Output File Format

5.1 Genome files

The results include seg_no_ns_H.txt and seg_ns_H.txt files that exclude or include the Ns islands, respectively.

Chromosome	Start	End	Size	Mean GC content	GC content STD	Is homogeneous
1	1	2720	2720	0.388	0.0985	0
1	2721	4832	2112	0.451	0.1181	1
1	4833	7104	2272	0.381	0.1151	0
1	7105	8672	1568	0.33	0.0922	0
1	8673	9728	1056	0.487	0.0888	0

If your Fasta file includes Ns, they will be shown in the second file as domains with GC content of 0.

5.2 Scaffold files

Alongside complete genomes, we included some unfinished genomes in which segmentation was performed on scaffolds (short genomic segments). Here, we included two type of output files. The first is in a similar format to that described in 5.1. All files will have "chromosome" set to 1 and be sorted so the first start and end domains would represent domains from different scaffolds. These files may be used for general analysis.

To analyze specific scaffolds, use the file with a name similar to that of the folder name. For example, the file ant_acromyrmex.txt under Scaffolds\Atta_acromyrmex includes the segmentation results for each scaffold.

Scaffold name	Start	End	Size	Mean GC content	GC content STD	Is homogeneous
C2021655	1	992	992	0.381	0.1401	0
C2021897	1	992	992	0.361	0.0818	0
C2022011	1	992	992	0.248	0.0721	0
C2022239	1	992	992	0.276	0.1194	0

6. Specific Pipeline Commands

You DO NOT need to run these commands separately, use the pipeline for that as explained above.

This summary is for users who wish to gain further understanding on how each program works. These commands should be used only from Matlab. We provided no compiled executable files for them.

6.1 Create the output directory format

1.List_ns
2.Seq_32bp
3.IsoPlotter_no_ns
4.IsoPlotter_no_ns_H
5.IsoPlotter_ns_H

6.2 First Step, Removing N's

N's are nucleotides that were not genotyped and they interfere with IsoPlotter's segmentation. So, first step is to remove them. It will create a frameshifts in the reported boundaries, but don't worry we will adjust those boundaries so there will be no frameshift.

MapN1(sequence_dir, output_dir_seqs, output_dir_32bp);
sequence_dir – input folder where all the sequences are.
output_dir_seqs – 1.List_ns – contains the mapping of the N's for each input file.
output_dir_32bp – 2.Seq_32bp – encodes the number of GC nucleotides in 32bp windows.

6.3 Prepare Input file

IsoPlotter accepts a sequence that contains GC counts for 32bp windows. This file can be created in two ways: Using the command MapN1 (see above). Or using CalculateGCWindow.n, however the later is per file not folder. To use it:
./CalculateGCWindow.n input_fasta.fa > output_file
input_fasta – source file
output_file – output file

The input file looks like that: 6 13 9 14 7 8 5 5 11 12 11 9 12 11 7 7

6.4 Segmentation Analysis

Partition sequence file (Xbp format) using IsoPlotter. This function implements IsoPlotter algorithm as described in (Elhaik et al. 2010).The algorithm recursively partition segments by maximizing the difference in GC content between adjacent subsequences, as measured by the

Jensen-Shannon divergence statistic (D_{JS}) (Lin 1991). The D_{JS} statistic is calculated over all possible partitioning points and the sequence is partitioned at the position of maximum D_{JS}. The process of segmentation is terminated when the maximal D_{JS} value is smaller than the dynamic threshold, calculated in real-time for each segment.

This is the main part of the analysis. IsoPlotter should be called as follows:
IsoPlotterSegmentation(Sequence_source_file, IsoPlotter_output_file, win_size, sizelim)
Sequence_source_file – The 32bp file
IsoPlotteR_output_file – the default is IsoPlotter_output.txt.
win_size – The default is 32.
Sizelim – the default is 3008.

6.5 Homogeneity Analysis

Here we test whether the domains inferred by IsoPlotter are more/less homogeneous than the sequence on which they reside. This function implements the homogeneity test for the domains inferred by IsoPlotter as described in (Elhaik et al. 2010). We used F-test with FDR correction to assay the relative homogeneity of each domain compared to that of the sequence. The function creates an updated segmentation file with an added column stating 1 (homogeneous domains) or 0 (nonhomogeneous domains).

TestingHomogeneityStatistics(Sequence_source_file, IsoPlotter_output_file, win_size, Homogeneity_output_file)
Sequence_source_file – The 32bp file.
IsoPlotteR_output_file – IsoPltoter's output file.
win_size – The default is 32.
Homogeneity_output_file – The default is "H_"+ IsoPlotter_source_file

6.6 Put Back the N's

If you removed the N's your sequence became shorter and the inferred domains do not represent the actual coordinates. While this is an acceptable strategy for the most part, you may be interested in reporting the real coordinates. For that, we will map the N's back and adjust the coordinates.

IsoPlotterAddNsH(IsoPlotter_output_dir, IsoPlotteR_output_file, n_map_files, n_cutoff, Output_dir)

IsoPlotter_output_dir – Output folder of IsoPlotter, e.g., 4.IsoPlotter_no_ns_H
IsoPlotteR_output_file – IsoPltoter's output file.
n_map_files – an output file of MapN1.
n_cutoff – the number of N's that should be ignored for the adjustment. For example, using a cutoff of 5, means that an island of over 5 N's would readjust the domain borders and anything less than 5 would be ignored. In any case, there will be no frameshift in the final coordinates.
Output_dir – the output file of domains after the adjustment to N's.

6.7 Put it All Together

Combine IsoPlotter output files from two folders into a single file.

OrganizeIsoPlotter(source_dir, output_dir)
source_dir – IsoPlotter output folder for the unadjusted files (e.g., 4.IsoPlotter_no_ns_H)
output_dir – the final output directory for the combined file

7. Utilities

Split_fasta_file.pl – Split Fasta file to multiple Fasta files.

8. Graphic tools

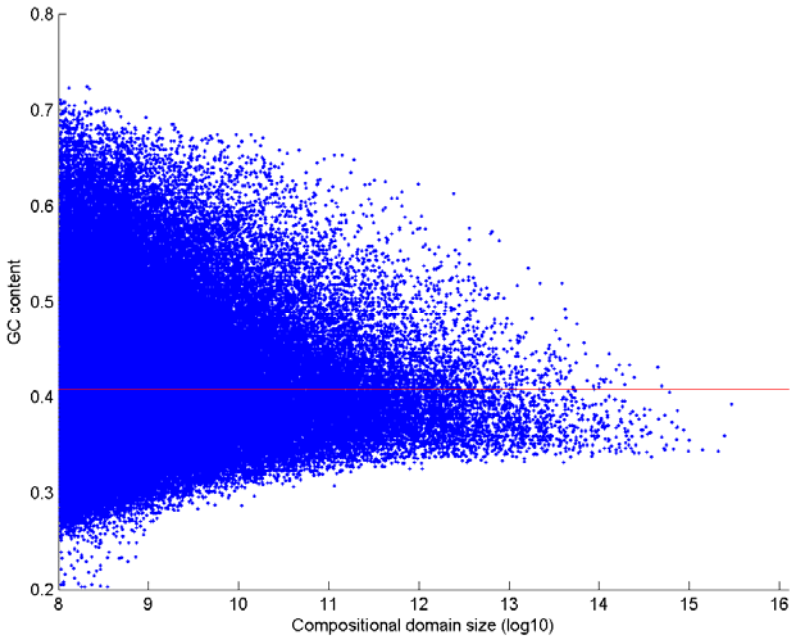
Three scripts are available to help you visualize the genome. Here, we demonstrate how to run them from Matlab. As explained before, we provided executable files for different operation systems. See section 4.2 and 4.3 on how to run them from each one.

8.1 Viewing the GC content versus domain length

PlotSegLengthGC(input_file, output_file)

For example:
PlotSegLengthGC('d:\seg_no_ns_H.txt', 'd:\PlotSegLengthGC_results')

This script plots the GC content of each domain against its size (log10). For example:

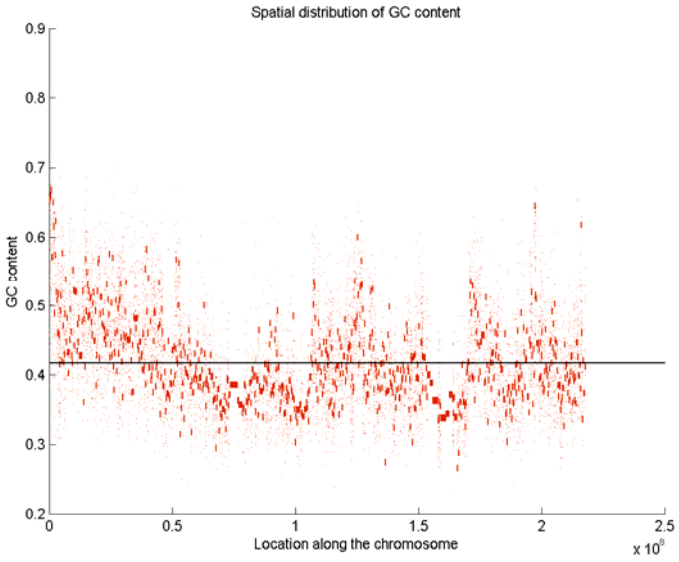


8.2 Viewing the spatial organization of a chromosome

PlotSpatialGC (input_file, chromosome #, output_file)

For example:
PlotSpatialGC('seg_no_ns_H.txt', 1, 'PlotSpatialGC.tif')

This script plots the compositional domains along a chromosome. The black line represents the mean GC content of the genome.



8.3 Viewing the three ideograms of compositional domains

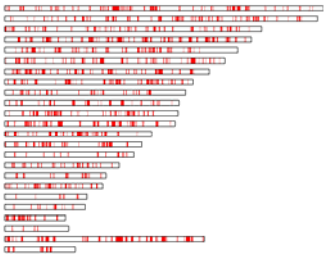
PlotGenome (input_file, output_file)

For example:
PlotGenome (‘d:\seg_no_ns_H.txt’, ‘d:\ PlotGenome.tif’)

In Linux, please print to files with eps extension:
PlotGenome('./Example/seg_no_ns_H.txt', './Output/PlotGenome.**eps**')

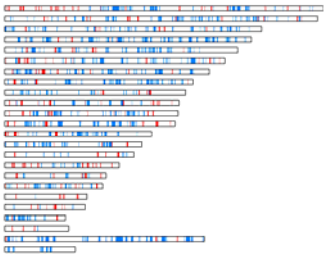
This script plots three ideograms of compositional domains inferred by IsoPlotter and mapped to chromosomes. The ideograms uncover the compositional patterns of long homogeneous domains (‘isochoric’) in three layers (from top to bottom): (a) compositionally homogeneous domains and nonhomogeneous domains; (b) long compositionally homogeneous domains (>300 kb), short compositionally homogeneous domains (<300 kb) and compositionally nonhomogeneous domains; and (c) long compositionally homogeneous domains (>300 kb) color coded by their mean GC content and all other domains (short compositionally homogeneous and nonhomogeneous domains). For more details see (Elhaik et al. 2010).

a.



b.

c.



9. FAQ

9.1 How does IsoPlotter works?

IsoPlotter is an improved recursive segmentation algorithm that employs a "dynamic threshold" that takes into account the composition and size of each segment (Elhaik et al. 2010). IsoPlotter calculates the D_{JS} statistic over all possible cutting points and compares its maximum to a dynamic threshold. In contrast to the standard D_{JS} algorithm, the threshold is not determined a priori, but separately for each segment to be partitioned. The length and standard deviation of GC content averaged over small windows along the segment are used to determine the dynamic threshold. If the maximum D_{JS} statistic exceeds this dynamically determined threshold, the segment is partitioned and segmentation continues recursively.

9.2 Which papers used IsoPlotter used?

IsoPlotter and the D_{JS} segmentation approach became the primary tool to study genomic structure in newly published genomes (Sodergren et al. 2006a; Sodergren et al. 2006b; Richards et al. 2008; Elsik et al. 2009; Kirkness et al. 2010; Werren et al. 2010; Smith et al. 2011a; Smith et al. 2011b; Suen et al. 2011).

9.3 I have no access to Matlab, what should I do?

Matlab Installer is available for free and should be installed on your OS. See here for more details.

9.4. Why Using IsoPlotter and not some other algorithm?

There are many algorithms and many implementations available in the literature. Before choosing the appropriate algorithm, please read these papers (Elhaik, Graur, and Josić 2010; Elhaik et al. 2010). In (Elhaik, Graur, and Josić 2010) we showed how different algorithms that use one or more thresholds give completely different and mostly incorrect results. In the follow up paper (Elhaik et al. 2010), we compared IsoPlotter to D_{JS} and showed the magnitude of this bias in D_{JS} and why the parameter-free approach of IsoPlotter is better. In other words, IsoPlotter is a parameter-free unbiased algorithm that should be preferred over the alternatives.

9.5 I have no chromosomes, only scaffolds, can I still use IsoPlotter?

Certainly, simply run the algorithm for each scaffold file (one sequence per fasta file). If your fasta file contains multiple sequences, use the perl script Split_fasta_file.pl to split it into multiple files (see more documentation in the file).

9.6 What is the minimum domain size that IsoPotter? finds? Can I change this size?

The minimum domain size is set to 3,008bp, which is the commonly used size in the literature. If your genome is particularly small, you can change this size by changing sizelim. Because the algorithm adjusts the threshold automatically based on a linear formula, there will be no bias in the calculations.

9.7 I have some N's islands in my files, can IsoPlotter handles N's?

The algorithm adjusts for N's by first saving their coordinates, then by running the segmentation on the remaining sequence and then integrating back the N's islands. In post publications, N's are ignored. IsoPlotter's will produce two files, one with N's islands and one without. You need to decide which N's island size you wish to keep? For example, in the human genome, it is usually 50k. So I you have 2 N's in the middle of a 100kb domain, they would be included in the domain. If you have N's island of 50kb, the 100kb domain would be split into 2 domains with the 50kb N's domain in the middle.

9.8 I am using the MCRInstaller and I got an error

The most common errors are due to mismatch between the MCRInstaller version and the version of the operation system. Please read subsection 2.2.

10. References

Elhaik E, Graur D, Josić K. 2010. Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol. Biol. Evol. 27:1015-1024.

Elhaik E, Graur D, Josic K, Landan G. 2010. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acids Res. 38:e158.

Elsik CG, Tellam RL, Worley KC, et al. [co-authors]. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324:522-528.

Kirkness EF, Haas BJ, Sun W, et al. [co-authors]. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc. Natl. Acad. Sci. USA. 107:12168-12173.

Lin J. 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory. 37:145-151.

Richards S, Gibbs RA, Weinstock GM, et al. [co-authors]. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature. 452:949-955.

Smith CD, Zimin A, Holt C, et al. [co-authors]. 2011a. Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). Proc. Natl. Acad. Sci. USA. 108:5673-5678.

Smith CR, Smith CD, Robertson HM, et al. [co-authors]. 2011b. Draft genome of the red harvester ant Pogonomyrmex barbatus. Proc. Natl. Acad. Sci. USA. 108:5667-5672.

Sodergren E, Weinstock GM, Davidson EH, et al. [co-authors]. 2006a. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 443:931-949.

Sodergren E, Weinstock GM, Davidson EH, et al. [co-authors]. 2006b. The genome of the sea urchin Strongylocentrotus purpuratus. Science. 314:941-952.

Suen G, Teiling C, Li L, et al. [co-authors]. 2011. The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 7:e1002007.

Werren JH, Richards S, Desjardins CA, et al. [co-authors]. 2010. Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science. 327:343-348.