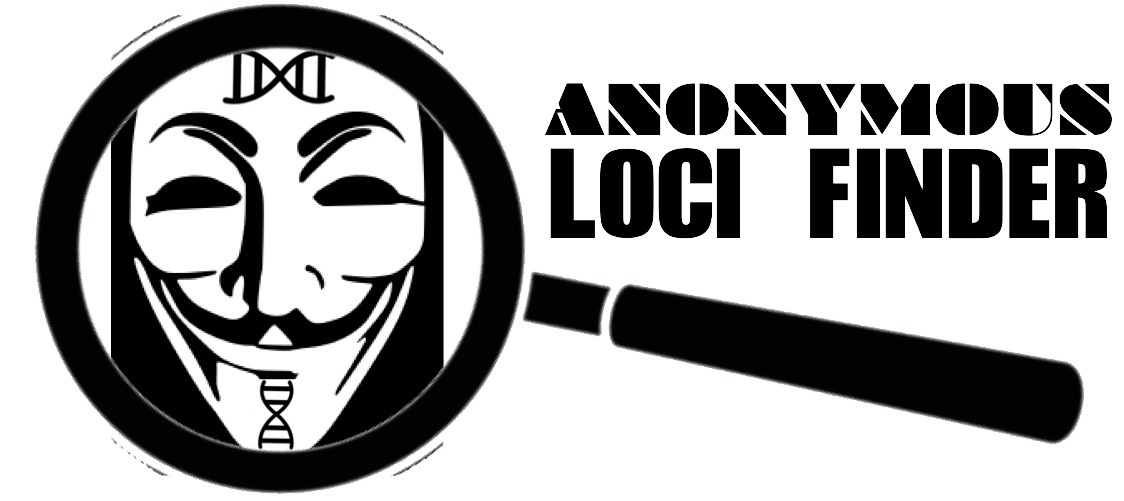
****

**Alfie 1.0**

**a package for nuclear anonymous loci prediction and phylogenomic analysis using complete genomes**

Igor Rodrigues da Costa

William Bryan Jennings

Francisco Prosdocimi

LAMPADA

LAboratório Multidisciplinar Para Análise de Dados

Universidade Federal do Rio de Janeiro

Instituto de Bioquímica Médica Leopoldo de Meis

Rio de Janeiro, 2015

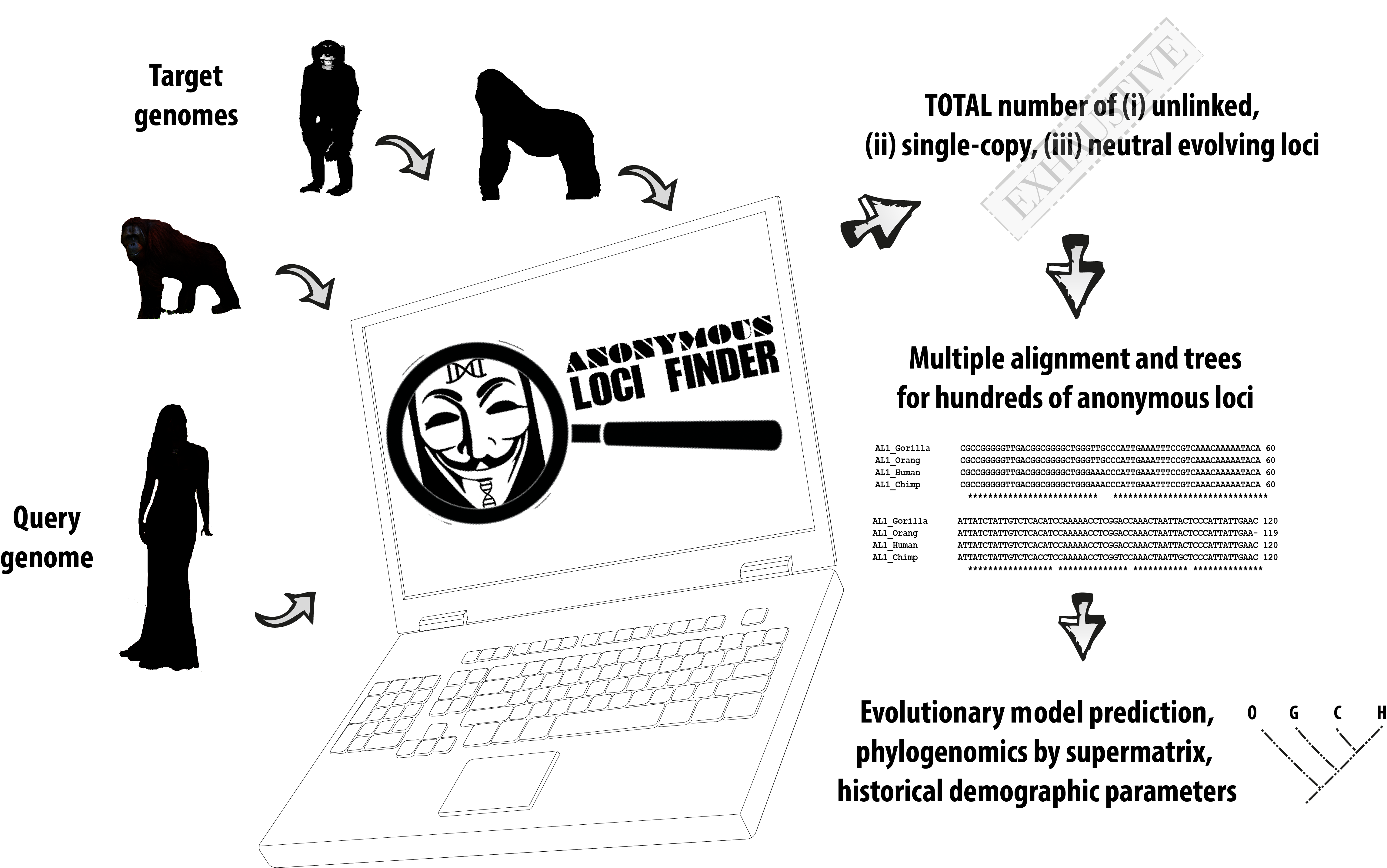
**SUMMARY**

1. Introduction3
   1. What is alfie?3
   2. How it works? 4
   3. What to use alfie for?4
2. Installation4
   1. Downloading4
   2. Installation requirements4
   3. Testing your instalation4
3. Quick Run5
4. Advanced Parameters5
5. Standalone Modules6
6. Citing Us6

**Thank you for downloading Alfie !**

1. **Introduction**
   1. **What is alfie?**

Alfie is an open-source python package containing a pipeline that searchs for single-copy, neutral evolving and non-linked anonymous loci. In other words, it looks for non-coding loci with a set of characteristics which makes them perfect for **phylogenomics** and **populational genetic** analyses. The goal of alfie is to be a complete, user-friendly and flexible anonymous loci predictor.



* 1. **How it works?**

Alfie utilises a sequential modular approach, with each step of the pipeline being stored in a single module that can be excecuted with greater flexibility in standalone mode. Many of these steps require external programs, such as BLAST or CLUSTALW. To reduce the complexity of the steps the whole pipeline can be run automatically using the afie.py script.

Using as input a single query genome and its associated GTF file, alfie finds all anonymous regions distant from genomic features (>200kb distance, by default) to avoid the effect of genetic linkage. After the initial anonymous regions have been found, they are split in small pieces (default 1kb). Each piece is searched against subject and query genomes to filter duplications and find orthologous regions in all genomes.

Alfie performs an exhaustive search, *i. e*., it finds every possible anonymous regions for the specified parameters in the genomes analyzed. Using default parameters and closelly related vertebrate genomes, alfie will usually find 100-1000 anonymous loci. Each ortholog group of anonymous region is written in NEXUS, PHYLIP, ALN and FASTA formats. Alfie will also concatenatenate all the alignments into a large file for phylogenomic analysis. The package comes with support for running phylogenomics and population genetics software.

* 1. **What to use alfie for?**

Alfie was developed to find hundreds to thousands of independent, nuclear markers among whole complete genomes and facilitate population genomics studies. An usual user case would use four to ten closely related genomes (we recommend no more than 20 million years of divergence to allow precise ortholog assignment). As an example, we were able to find about 300 anonymous loci with ortologues in all four hominoid genomes used (human, chimpanzee, gorilla and orangutan).

It is strongly recommended that at least one of the genomes have been extensively studied and annotated, presenting a comprehensive GTF file that describes the precise location of the main features targeted by natural selection, such as genes and regulatory elements.

1. **Installation**
   1. **Downloading alfie**

Users can either download and install the entire source code from github at the address <https://github.com/igorrcosta/alfie/archive/master.zip> or download the docker container with all the dependencies included (available soon). To install, simply extract the donwloaded repository file in a folder.

* 1. **Installation requirements**

Working versions for each required program can be found on the following links:

Blast: [http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\_TYPE=Blast Docs&DOC\_TYPE=Download](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=Blast%20Docs&DOC_TYPE=Download)

ClustalW: <http://sourceforge.net/projects/mira-assembler/>

PhyML: <http://www.atgc-montpellier.fr/phyml/binaries.php>

ModelTest: <https://code.google.com/p/jmodeltest2/>

Biopython**:** <http://biopython.org/wiki/Download>

* 1. **Testing your installation**

You can test your installation by running alfie on the human and chimpanzee Y chromossomes’ test case. To do that, go to the alfie folder and run the following command:

$> python alfie.py -i test/homo\_y.fasta test/pan\_y.fasta -g test/y.gtf

This command might take more than a minute to finish, depending on your computer speed. A successful run will output 42 candidate loci in the test.fasta file, of which only loci 11, 27 and 40 will be selected and aligned.

1. **Quick Run**

To run alfie, you will need a reference genome and gtf file (there are several available at <http://www.ensembl.org/info/data/ftp/index.html>) and some genomes to compare against the reference.

Example command line:

$> python alfie.py –i \*reference\_genome.fasta\* \*genome2.fasta\* \*genome3.fasta\* -g \*annotation\_file.gtf\* -o \*output\_folder\*

Here the \*.fasta\* represents the path to the genome files in FASTA format, \*annotation\_file.gtf\* is the path to the gtf file and \*output\_folder\* is where the loci will be saved.

This command will find several candidade loci in the reference genome, which will be stored at a test.fasta file. This candidade loci will be blasted against all genomes and the final loci will be saved in several formats.

1. **Advanced parameters**

While alfie can be executed with good results without any additional configuration, there are several options to flexibilize your analysis:

|  |  |
| --- | --- |
| -i, --genomes | Path to the genome files in the FASTA format. The first genome file inserted will be considered the reference genome. |
| -g, --gtf | Path to the gtf file relative to the reference genome. |
| -o, --outpath | Path where all output files will be saved. |
| -f, --skip\_formatdb | Skip making BLAST databases, use databses from the last run. Can significantly improve analysis speed. |
| --locus\_length | Length of the anonymous loci (default 1000bp). You may increase this length if you are planning to predict primers for these loci. |
| --max\_n | Maximum percentage of N's in the AL sequence (default 0%). Increase this if you want to find more loci in a low quality reference genome. |
| --inter\_distance | Minimum distance between ALs (defaut 200000bp). |
| --gene\_distance | Minimum distance between ALs and genes (default 200000bp). You can use negative numbers to find loci close to the gene regions, for example, -2000 will find loci between 0-2000bp from genes |
| --end\_distance | Minimum distance between ALs and the telomeres (default 10000bp). |
| --gene\_locus | Use this flag to find loci inside the gene regions (will ignore the gene\_distance flag). |
| --cds | Only considers the CDS features of GTF files, ignoring all pseudogens, miRNA, etc. |
| --duplication\_cutoff | ALs with 2 hits with identity higher than this will be considered duplicated (default: 50%). |
| --identity\_cutoff | ALs with a identity higher than this will be considered homologous (default: 90%). |
| --coverage\_cutoff | BLAST hits must have at least this much %coverage to be considered hits (default: 90%). |
| --chromossomes | Chromossomes to be excluded. We recommend excluding all sex chromossomes. |
| --min\_align | Minimum final alignment length (default 900bp). This will exclude loci that have many Ns in genomes other than the reference. |
| --remove\_gaps | Remove gaps from the final alignment. |

1. **Standalone modules**

dasd

1. **Citing us**

Costa IR, Prosdocimi F, Jennings WB (2015). *in silico* Phylogenomics Comes of Age: Using Bioinformatic Algorithms to Discover Evolutionary Markers in Whole Genome Datasets. Manuscript in preparation.