Fast and Intuitive Epigenetic Data Analysis

Felipe Fernandes Albrecht

Motivation

Functional genomics and epigenomics consortia, like BLUEPRINT[1] (http://www.blueprint-epigenome.eu/) are generating resources of tremendous value for studying epigenetic gene regulation.

Unfortunately, accessing and analyzing these datasets remains cumbersome. We explored this challenge developing a tool that gives to users the possibility of analysing the Epigenetic data in a fast and intuitive way.

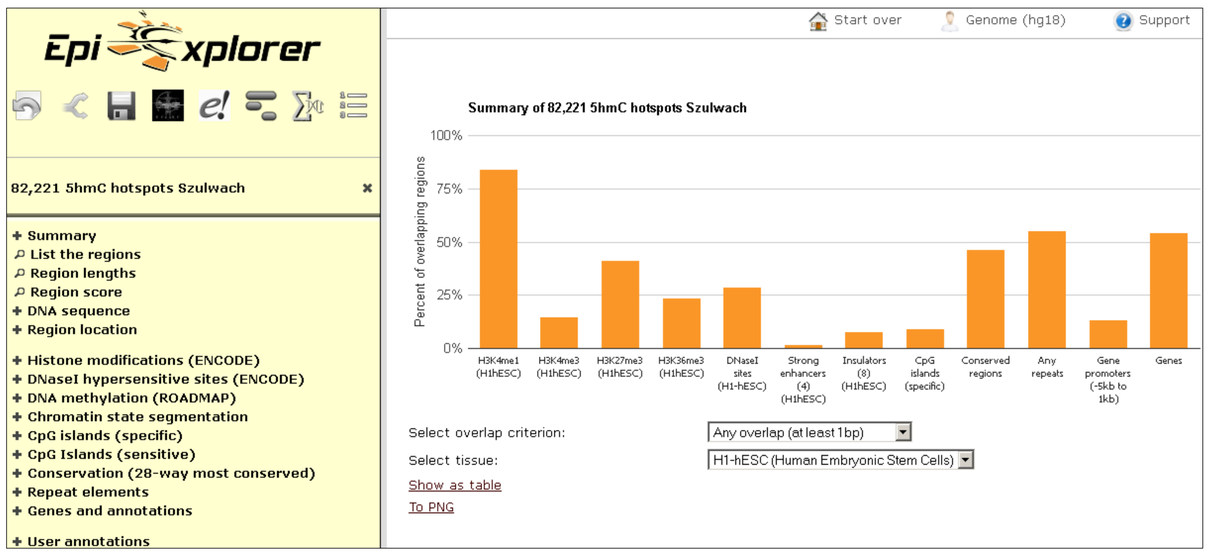
Methods

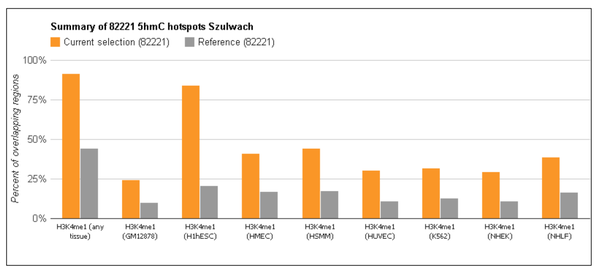
EpiExplorer (http://epiexplorer.mpi-inf.mpg.de) is a web tool for exploring genome and epigenome data on a genomic scale.

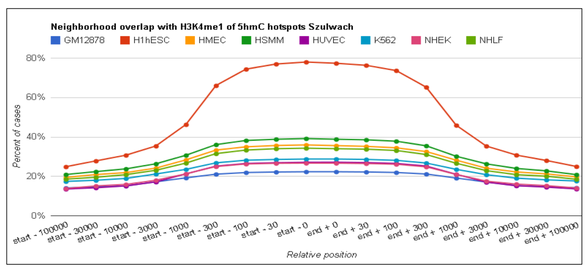
All EpiExplorer analyses are performed dynamically within seconds, using an efficient and versatile text indexing scheme, called CompleteSearch[1] that we introduce to bioinformatics.

As a result, we provide a web server interface that allows users to dynamically and interactively explore the genomic and epigenomic properties of sets of genome region. Also EpiExplorer was designed to scale to high user load and to be readily extensible with additional datasets.

Easy and Intuitive Interface







Annotation Process

Step	Description	Representation

Upload — The user uploads a set of genomic regions (in standard BED format)

Region	Chromossome	start	end
Region 1	chr1	1000	4240
Region 2	chr2	500	1545
Region 3	chr1	8300	8850
Region 4	chr5	3100	3400

Annotate — Each genomic region is annotated with a broad range of genomic attributes

Region	Chrom.	Length	CpG Freq.	CGI overlaps	Distance to nearest CGI
Region 1	chr1	3240	0.07	34%	0
Region 2	chr2	1045	0.02	0%	521
Region 3	chr1	550	0.05	5%	0
Region 4	chr5	300	0.16	80%	0

Convert to text — Every region is represented as a text document and its annotations are translated into words

Region 1	Region 2	Region 3	Region 4
chr1 length:3240 frequency:CG:07 overlapratio:CGI:34 overlap:CGI	chr2 length:1045 frequency:CG:02 distanceTo:CGI:521	chr1 length:0550 frequency:CG:05 overlapratio:CGI:05 overlap:CGI	chr5 length:0300 frequency:CG:16 overlapratio:CGI:80 overlap:CGI

Sort — Words and documents are sorted & assigned unique identifiers

Doc ID	Document
D1	Region 1
D2	Region 2
D3	Region 3
D4	Region 4

Word ID	Word
W1	chr1
W2	chr2
W3	chr5
W4	distanceTo:CGI:521
W5	frequency:CG:02
W6	frequency:CG:05
W7	frequency:CG:07
W8	frequency:CG:16

Word ID	Word
W9	length:0300
W10	length:0550
W11	length:1045
W12	length: 3240
W13	overlap:CGI
W14	overlapratio:CGI:05
W15	overlapratio:CGI:34
W16	overlapratio:CGI:80

Create index — Sorted lists are stored in memory such that blocks correspond to ranges of word IDs and contain all pairs of document/word IDs in a given range

Block	Word ID range	Corresponding words	document-word pairs
B1	W1 - W3	chr1, chr2, chr5	(D1,W1) (D2,W2) (D3,W3) (D4,W1)
B2	W4 - W8	distanceTo:CGI:521, frequency:CG:02, frequency:CG:05, frequency:CG:07, frequency:CG:16	(D1,W7) (D2,W4) (D2,W5) (D3,W6) (D4,W8)
B3	W9 - W12	length:0300, length:0550, length:1045, length:3240	(D1,W13) (D1,W15) (D3,W13) (D3,W14) (D4,W13) (D4,W16)
B4	W13- W16	overlap:CGI, overlapratio:CGI:05, overlapratio:CGI:34, overlapratio:CGI:80	(D1,W13) (D1,W15) (D3,W13) (D3,W14) (D4,W13) (D4,W16)

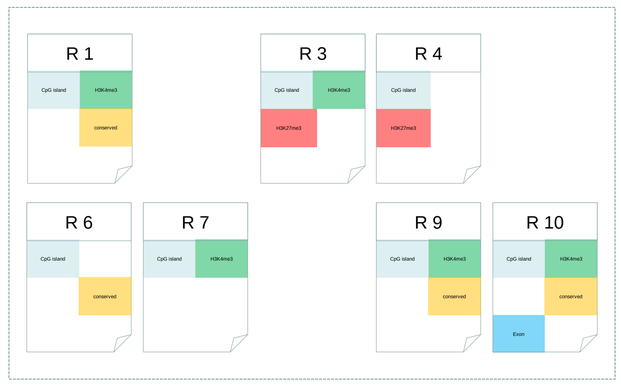
Querying Process

Query:
How many regions overlap with CpG islands?

Query: **overlap:CGI**

```
<result>
  <query>overlap:CGI</query>
  <time unit="msecs">510</time>
  <completions >
     <c dc="7">overlap:CGI</c>
  </completions>
  <hits total="7">
     <hit><title>R1</title></hit>
     <hit><title>R3</title></hit>
     <hit><title>R4</title></hit>
     <hit><title>R6</title></hit>
     <hit><title>R7</title></hit>
     <hit><title>R9</title></hit>
     <hit><title>R10</title></hit>
  </hits>
</result>
```

Response:
Seven regions overlap with CpG islands: R1, R3, R4, R6, R7, R10



Access and use now:

http://epiexplorer.mpi-inf.mpg.de

Conclusions

The Epigenome Explorer implements a novel way for genome data integration based on an indexing structure developed for text-search.

It demonstrates that despite the size and diversity of the data, genome and epigenome annotations can be explored in an interactive and advantageous way.

References

[1] Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., Bock, C., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, *30*(3), 224-226.

[2] *H. Bast and I. Weber (2007) The CompleteSearch Engine: Interactive, Efficient, and Towards IR & DB integration (CIDR)*

max planck institut informatik

UNIVERSITÄT DES SAARLANDES

International Max Planck Research School for Computer Science