Bioinformatics 2018-2019

Date: DD Month YYYY

Project report

|  |
| --- |
| Bioinformatics@Data Science A.Y. 2018-2019  An analysis over the Y Chromosome  Silvia Basile1, Tommaso Lanciano1  1Group no. 11 Abstract Among all the chromosomes, the Y chromosome is certainly one of the most discussed in science, due to its importance over the determination of the sex. Here we conduct a simple gene analysis, starting from a list of seed genes, and detecting their interactions with other genes. Identifying genes on each chromosome is an active area of genetic research. Because researchers use different approaches to predict the number of genes on each chromosome, the estimated number of genes varies. The Y chromosome likely contains 50 to 60 genes that provide instructions for making proteins. Because only males have the Y chromosome, the genes on this chromosome tend to be involved in male sex determination and development.  You can find all the files and the codes involved in this work at: <https://github.com/tlancian/BI_Homeworks> |

**Introduction**

The***Y chromosome***, along with the X chromosome, is responsible of sex determination of an offspring in all mammals, including humans, and for this reason they are known as the *sex chromosomes*.

Each person normally has one pair of sex chromosomes in each cell: XX for females, XY for males. Hence, the Y chromosome is transmitted only from a father to a son. It is composed of more than 59 million base pairs, representing approximately the 2% of the total DNA in a male cell.

The Y chromosome expresses (according to the HGNC) 45 unique proteins, some associated with sex and fertility, and others associated with non reproductive functions, including ribosomal proteins, transcription factors, histone methylation enzymes, and cell adhesion molecules. Apart from individual genes, the Y chromosome also houses multiple repetitive sequences and many multicopy gene arrays within palindromes.

Seed genes

Thus, our analysis started considering these 45 unique proteins (seed genes). We will refer to them with the official gene symbols assigned by the HGNC (HuGO Gene Nomenclature Committee), since no misinterpretation have been found in this phase.

Our first step, consisted in getting more information about the seed genes, building a table with the following information:

* *Uniprot Accession Number*: an alphanumeric identifier taken from UniProt Knowledgebase, considering only humans and only records coming from reviewed version (Swiss-Prot).
* *Entrez Gene ID*: an identifier given by the National Center for Biotechnology Information.
* *Protein Name*: the name of the protein, taken by the UniProt Knowledgebase.
* *Function*: description of its function, taken by the UniProt Knowledgebase.

We were able to scrape all this information, using the Python library *bioservices* that contains several API to interact with the websites mentioned above. Hence, once scraped all this information, results has been reported in Table 1.

The main difference we have found collecting data from the different sources, is about the number of unique entries found. In fact, for the UniProt AC we gathered 37 unique entries, and this is due to the merging/splitting of two genes. Meanwhile, for the Entrez Gene ID we have only found a couple of genes with the same ID: TSPY4, TSPY8. Both of them are testis specific proteins, and this analogy is totally reasonable, because they derive from the same ancestral gene.

**Table 1.**Seed Genes Table. It contains for each one of the seed genes the information listed above. In the Github repository you can find the complete table, with also the function that has each protein.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene | Uniprot AC | Protein Name | HGNC Symbol | Entrez Gene ID |  |
| AMELY | Q99218 | Amelogenin, Y isoform | AMELY | 266 |  |
| BPY2 | O14599 | Testis-specific basic protein Y 2 | BPY2 | 9083 |  |
| BPY2B | O14599 | Testis-specific basic protein Y 2 | BPY2B | 442867 |  |
| BPY2C | O14599 | Testis-specific basic protein Y 2 | BPY2C | 442868 |  |
| CDY1 | Q9Y6F8 | Testis-specific chromodomain protein Y 1 | CDY1 | 9085 |  |
| CDY1B | Q9Y6F8 | Testis-specific chromodomain protein Y 1 | CDY1B | 253175 |  |
| CDY2A | Q9Y6F7 | Testis-specific chromodomain protein Y 2 | CDY2A | 9426 |  |
| CDY2B | Q9Y6F7 | Testis-specific chromodomain protein Y 2 | CDY2B | 203611 |  |
| DAZ1 | Q9NQZ3 | Deleted in azoospermia protein 1 | DAZ1 | 1617 |  |
| DAZ2 | Q13117 | Deleted in azoospermia protein 2 | DAZ2 | 57055 |  |
| DAZ3 | Q9NR90 | Deleted in azoospermia protein 3 | DAZ3 | 57054 |  |
| DAZ4 | Q86SG3 | Deleted in azoospermia protein 4 | DAZ4 | 57135 |  |
| DDX3Y | O15523 | ATP-dependent RNA helicase DDX3Y | DDX3Y | 8653 |  |
| EIF1AY | O14602 | Eukaryotic translation initiation factor 1A, Y-chromosomal | EIF1AY | 9086 |  |
| HSFY1 | Q96LI6 | Heat shock transcription factor, Y-linked | HSFY1 | 86614 |  |
| HSFY2 | Q96LI6 | Heat shock transcription factor, Y-linked | HSFY2 | 159119 |  |
| KDM5D | Q9BY66 | Lysine-specific demethylase 5D | KDM5D | 8284 |  |
| NLGN4Y | Q8NFZ3 | Neuroligin-4, Y-linked | NLGN4Y | 22829 |  |
| PCDH11Y | Q9BZA8 | Protocadherin-11 Y-linked | PCDH11Y | 83259 |  |
| PRORY | Q9H606 | Proline-rich protein, Y-linked | PRORY | 1,01E+08 |  |
| PRY | O14603 | PTPN13-like protein, Y-linked | PRY | 9081 |  |
| PRY2 | O14603 | PTPN13-like protein, Y-linked | PRY2 | 442862 |  |
| RBMY1A1 | P0DJD3 | RNA-binding motif protein, Y chromosome, family 1 member A1 | RBMY1A1 | 5940 |  |
| RBMY1B | A6NDE4 | RNA-binding motif protein, Y chromosome, family 1 member B | RBMY1B | 378948 |  |
| RBMY1D | P0C7P1 | RNA-binding motif protein, Y chromosome, family 1 member D | RBMY1D | 378949 |  |
| RBMY1E | A6NEQ0 | RNA-binding motif protein, Y chromosome, family 1 member E | RBMY1E | 378950 |  |
| RBMY1F | Q15415 | RNA-binding motif protein, Y chromosome, family 1 member F/J | RBMY1F | 159163 |  |
| RBMY1J | Q15415 | RNA-binding motif protein, Y chromosome, family 1 member F/J | RBMY1J | 378951 |  |
| RPS4Y1 | P22090 | 40S ribosomal protein S4, Y isoform 1 | RPS4Y1 | 6192 |  |
| RPS4Y2 | Q8TD47 | 40S ribosomal protein S4, Y isoform 2 | RPS4Y2 | 140032 |  |
| SRY | Q05066 | Sex-determining region Y protein | SRY | 6736 |  |
| TBL1Y | Q9BQ87 | F-box-like/WD repeat-containing protein TBL1Y | TBL1Y | 90665 |  |
| TGIF2LY | Q8IUE0 | Homeobox protein TGIF2LY | TGIF2LY | 90655 |  |
| TMSB4Y | O14604 | Thymosin beta-4, Y-chromosomal | TMSB4Y | 9087 |  |
| TSPY1 | Q01534 | Testis-specific Y-encoded protein 1 | TSPY1 | 728403 |  |
| TSPY2 | A6NKD2 | Testis-specific Y-encoded protein 2 | TSPY2 | 64591 |  |
| TSPY3 | P0CV98 | Testis-specific Y-encoded protein 3 | TSPY3 | 728137 |  |
| TSPY4 | P0CV99 | Testis-specific Y-encoded protein 4 | TSPY4 | 728395 |  |
| TSPY8 | P0CW00 | Testis-specific Y-encoded protein 8 | TSPY8 | 728395 |  |
| TSPY10 | P0CW01 | Testis-specific Y-encoded protein 10 | TSPY10 | 1E+08 |  |
| USP9Y | O00507 | Probable ubiquitin carboxyl-terminal hydrolase FAF-Y | USP9Y | 8287 |  |
| UTY | O14607 | Histone demethylase UTY | UTY | 7404 |  |
| VCY | O14598 | Testis-specific basic protein Y 1 | VCY | 9084 |  |
| VCY1B | O14598 | Testis-specific basic protein Y 1 | VCY1B | 353513 |  |
| ZFY | P08048 | Zinc finger Y-chromosomal protein | ZFY | 7544 |  |

Interaction data

Once we gathered initial information about the seed genes, we can procede collecting all binary protein-protein interactions (PPI). For this point the references are:

* *BioGRID:* Biological General Repository for Interaction Datasets, version 3.5.167;
* *Integrated Interactions Database* (IID), selecting all human tissues from experimental results. For this database we will compare the outputs obtained considering queries on both the gene symbol of the seed genes and the Uniprot Accession Number.

While for BioGRID we needed to download the whole datasets, and to make the queries by our self using Pandas library in Python, for the IID datasets we exploited the tool provided on their website. For each dataset we have selected the interactions of all seed genes, and subsequently we also included the interactions among non-seed genes that interact with at least one seed gene. Main results are summarized in Table 2.

**Table 2.**Summary of interaction data**.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **BioGRID** | **IID** | **IID (AC)** |
| **Total interactions** | 1439 | 1621 | 1676 |
| **Seed Genes Involved** | 29 | 25 | 25 |
| **Genes Involved** | 245 | 243 | 243 |

Before getting the results described in Table 2, we have performed some operations over the datasets. In particular:

***BioGRID:*** we took from the whole datasets the interactions that involved at least a seed gene, and then the ones among the genes linked at least once with a seed gene. After that, once performing Part 4, we noticed that many interactions were involving genes not related to humans. Indeed, we have performed a query over the Uniprot DB, asking for the Uniport AC and where the organism involved was human. By this, we have deleted all the interactions where there was no result in the Uniprot query.

***IID:*** The IID website offered many possibilities for making a query. In order to make everything like the BioGRID, that report only interactions published in papers, we asked the IID only for interactions where there was an experimental evidence on it. In order to confirm our results, we have performed the same query using the related Uniprot AC obtained in Part 2. Comparing the two outputs, we noticed a different number of interactions. But after some reasonings, we observed that for genes that had the same Uniprot AC, the IID system was returning the same result for each one. Once deleted these duplicates, we obtained the same interactions for both the datasets.

From the results we have obtained, we can see that there are few discrepancies, thus our research seems consistent.

# Interactomes data

In this section, we report the different interactomes that we have built. They are three:

* *Seed Genes Interactome*: it contains only interactions among seed genes, by both datasets;
* *Union Interactome*: it contains interactions that involve at least one seed genes, by both datasets;
* *Intersection Interactome*: it contains interactions that involve at least one seed genes, and that are present in both datasets.

In order to obtain this three interactomes, we used the two datasets mentioned in the previous section and processed them with Pandas library in Python. A further information needed is the UniProt AC for the BioGRID interactome. Since it was not provided in the whole dataset, we retrieved it through the library *bioservices* and stored all data in the Biogrid interactome dataset*.*

We have set all the constraint mentioned above and the results are summarized in Table 3.

**Table 3.**Summary of interactome data**.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Seed genes** | **Union** | **Intersection** |
| **Total interactions** | 13 | 513 | 211 |
| **Total genes** | 10 | 288 | 199 |
| **Seed genes** | 10 | 31 | 22 |
| **Non-seed genes** | 0 | 257 | 177 |
| **BioGRID interactions** | 5 | 255 | - |
| **IID interactions** | 8 | 258 | - |

As we can see from the table above, only 13 seed genes are involved in interactions with other seed genes. Most of the times the genes interact with themselves while only three times there are interactions among two different seed genes.

Completely different scenario is for both the union and the intersection interactomes in which, even if we consider interactions with at least one seed genes, there are only 11% of seed genes in both interactomes.

As last observation, we can notice that interactions belong to BioGRID or IID datasets in an equivalent way, especially in the union interactome.

*You can find the results for this section in the folder.*

# Enrichment analysis

In order to perform an enrichment analysis, we have used InnateDB. Thus, once obtained the interactomes, we’ve extract from them, the list of the unique genes involved in each one of the three interactomes. Once obtained this information, we have exploited the tools provided by InnateDB, that directly online can perform an Over-Represented Analysis (ORA), both for the Gene Ontology side and for the Pathway.

The main interesting thing to say, is that we couldn’t get any result by the Pathway ORA of the seed genes list, due to the small number of genes involved.

*All the results are stored in the folder.*

*- dire alla fine che si può consultare la cartella, senza specificarlo in ogni sezione (?)*