

Discover exons with highest SNP density

Edgar G. Nogales

10/28/2016

1 Introduction

In this case we will work with Galaxy tool. It's an open source, website platform where we can perform queries in a multiple bio-medical data-sets.

Differentiation is now a technique taught to mathematics students throughout the world. In this document I will discuss some aspects of differentiation.

2 Exercise

In this case we will work with 2 main databases. In the database "UCSC Main on Human: knownGene (genome)" we will find all the exons from the human genome. And in the data-set "UCSC Main on Human: snp147Common (genome)" will give us the SNPs of the human genome.

1.Chrom	2.Start	3.End	4.Name
chr1	69890	70008	uc001aa1.1_cds
chr1	925941	926813	uc057axs.1_cds
chr1	938154	938336	uc057axs.1_cds
chr1	931838	931889	uc057axs.1_cds
chr1	935771	935896	uc057axs.1_cds

1.Chrom	2.Start	3.End	4.Name	5	6
chr1	17372	17373	rs750111615	0	+
chr1	17374	17375	rs755771866	0	+
chr1	17378	17379	rs754322362	0	+
chr1	17384	17385	rs201535981	0	+
chr1	17397	17398	rs200784459	0	+

Once we have all the data we need, it's time to start working with it. Using the exon's table, we will calculate the size subtracting the position of the exons and adding 1. To perform the operation we need to click on "Text Manipulation" \rightarrow "Compute an expression on every row" and we will fill the fields like in the image.

Compute an expression on every row (Galaxy Version 1.1.0)

Add expression

$c3 - c2 + 1$

as a new column to

29: Cut on data 28

Dataset missing? See TIP below

Round result?

NO

Execute

And the result will be like:

1	2	3	4	5	6	7
chr1	69090	70008	uc001aal.1_cds_0_0_chr1_69091_f	0	+	919.0
chr1	925941	926013	uc057axs.1_cds_1_0_chr1_925942_f	0	+	73.0
chr1	930154	930336	uc057axs.1_cds_2_0_chr1_930155_f	0	+	183.0
chr1	931038	931089	uc057axs.1_cds_3_0_chr1_931039_f	0	+	52.0
chr1	935771	935896	uc057axs.1_cds_4_0_chr1_935772_f	0	+	126.0
chr1	939039	939129	uc057axs.1_cds_5_0_chr1_939040_f	0	+	91.0
chr1	939274	939291	uc057axs.1_cds_6_0_chr1_939275_f	0	+	18.0
chr1	925941	926013	uc057axt.1_cds_1_0_chr1_925942_f	0	+	73.0
chr1	930154	930336	uc057axt.1_cds_2_0_chr1_930155_f	0	+	183.0
chr1	931038	931089	uc057axt.1_cds_3_0_chr1_931039_f	0	+	52.0
chr1	935771	935793	uc057axt.1_cds_4_0_chr1_935772_f	0	+	23.0

Next step will be to count how many SNPs we find for each Exon. For that we will use the "Group by" command and then "count" the repetitions.

1	2
uc001aal.1_cds_0_0_chr1_69091_f	5
uc001abw.2_cds_10_0_chr1_942559_f	2
uc001abw.2_cds_13_0_chr1_943908_f	1
uc001abw.2_cds_2_0_chr1_930155_f	1
uc001abw.2_cds_4_0_chr1_935772_f	1
uc001abw.2_cds_6_0_chr1_939275_f	1
uc001abz.5_cds_10_0_chr1_953175_r	2
uc001abz.5_cds_11_0_chr1_953782_r	1
uc001abz.5_cds_3_0_chr1_946173_r	1
uc001abz.5_cds_4_0_chr1_946402_r	1
uc001abz.5_cds_9_0_chr1_952412_r	1
uc001aca.3_cds_10_0_chr1_964349_f	1
uc001aca.3_cds_11_0_chr1_964963_f	3
uc001aca.3_cds_3_0_chr1_961826_f	1
uc001aca.3_cds_4_0_chr1_962355_f	1

Next step will be to join both tables. Here we find some problems because we used the command "Join two Datasets side by side on a specified field" and it generates too much data and Galaxy was not able to work with it. Finally we find that we need to use another command "Join the intervals of two datasets side-by-side"

Input Parameter	Value
Join	33: SNPs by Exons
using column	1
with	28: Size calc
and column	4
Keep lines of first input that do not join with second input	No
Keep lines of first input that are incomplete	No
Fill empty columns	no_fill

To calculate the density of our data, we need to divide the number of SNPs by the size of the Exons

1	2	3	4	5
chr1	uc001aal.1_cds_0_0_chr1_69091_f	5	919.0	0.00544069640914
chr1	uc001abw.2_cds_10_0_chr1_942559_f	2	501.0	0.00399201596806
chr1	uc001abw.2_cds_13_0_chr1_943908_f	1	247.0	0.00404858299595
chr1	uc001abw.2_cds_2_0_chr1_930155_f	1	183.0	0.00546448087432
chr1	uc001abw.2_cds_4_0_chr1_935772_f	1	126.0	0.00793650793651
chr1	uc001abw.2_cds_6_0_chr1_939275_f	1	187.0	0.00534759358289
chr1	uc001abz.5_cds_10_0_chr1_953175_r	2	115.0	0.0173913043478
chr1	uc001abz.5_cds_11_0_chr1_953782_r	1	112.0	0.00892857142857
chr1	uc001abz.5_cds_3_0_chr1_946173_r	1	115.0	0.00869565217391
chr1	uc001abz.5_cds_4_0_chr1_946402_r	1	145.0	0.00689655172414
chr1	uc001abz.5_cds_9_0_chr1_952412_r	1	190.0	0.00526315789474
chr1	uc001aca.3_cds_10_0_chr1_964349_f	1	183.0	0.00546448087432
chr1	uc001aca.3_cds_11_0_chr1_964963_f	3	230.0	0.0130434782609
chr1	uc001aca.3_cds_3_0_chr1_961826_f	1	223.0	0.00448430493274

Finally, we need to sort the result to have the highest density in the upper rows. For that we will use the "sort" function.

1	2	3	4	5
chr4	uc062voj.1_cds_1_0_chr4_22346823_r	1	2.0	0.5
chr14	uc059dbv.1_cds_19_0_chr14_73264010_f	4	11.0	0.363636363636
chr1	uc010omg.3_cds_8_0_chr1_46665647_r	1	3.0	0.333333333333
chr1	uc057gfy.1_cds_9_0_chr1_46665647_r	1	3.0	0.333333333333
chr4	uc062yyc.1_cds_28_0_chr4_108920996_r	1	3.0	0.333333333333
chr6_GL000255v2_alt	uc063zpx.1_cds_5_0_chr6_GL000255v2_alt_3784117_r	32	113.0	0.283185840708
chr6_GL000253v2_alt	uc063wpx.1_cds_5_0_chr6_GL000253v2_alt_3993939_r	35	132.0	0.265151515152
chr1	uc057ggq.1_cds_7_0_chr1_46815073_f	1	4.0	0.25
chr12	uc001sjf.4_cds_0_0_chr12_56042167_f	1	4.0	0.25
chr12	uc058pgp.1_cds_1_0_chr12_56042167_f	1	4.0	0.25
chr12	uc058pxv.1_cds_3_0_chr12_57517376_r	1	4.0	0.25

We can see that the highest density is in chr4 Exon uc062voj because we can find 4 SNPs in just 2 bases.
After export the workflow, we have a flux like that

