Statistical inference

Ruth Gómez Graciani

30 de octubre de 2017

Biological question

Among the causes of the countless diseases that humans can suffer, genetic causes are ones of the most concerning because they are usually treatable but not fixable/curable and they can be transmitted to our offspring. However, genetic factors of diseases can vary from slight predispositions to totally letal mutations and rearrangements that lead to miscarriages even before the pregnancy is noticeable. When regarded in the Ohta's nearly neutral theory of evolution context (Casillas and Barbadilla 2017), it can be considered that these extremely deleterious mutations, which are hardly ever detected nor reported, are in the lower tail of the distribution, and that the mutations inducing slight predispositions are in the slightly deletrious zone. Therefore, what is known as 'genetic disease' must be around the center of the deleterious mutations distribution.

From now on in this work, there will be considered as 'mutations' only those that are deleterious or slightly deleterious, and for simplification it will be assumed that they affect a gene (although it is known that some deleterious mutations can affect other regions like enhancers or epigenetic targets).

The same mutation can be more or less severe depending on the gene that it affects, meaning that there are differences in the essentiality of the genes, and that any change can be more or less deleterious according to importance of the gene. Given two chromosomes of equal gene density, it would be more likely to have an extremely deleterious mutation in that with the larger number of important genes. That's why it is not expected to find chromosomes with a large proportion of important genes, but rather a random distribution of such genes.

In this context, the biological question is: are mutations causing diseases in fact evenly distributed across human chromosomes?

Hypotheses and variables

The biological hypotheses (random distribution of important genes vs. chromosomes with a higher deleterious potential) can be tested using as an estimator of important genes the number of diseases that have been reported for each of the chromosomes. It is expected to be proportional to the number of genes in each chromosome, meaning that both measures will follow the same distribution. Therefore, the null statistical hypothesis is that the distribution of diseases across the chromosomes follows the same as the genes, and the alternative hypothesis is that some chromosomes have more (or less) diseases associated than the expected, in other words, that the distributions are different.

Experimental design

A priori, the chromosomes included in the study are autosomal and sex human chromosomes. Their general properties are published in ("Homo Sapiens Genome, Ncbi" 2017). An unbeatable source for the number of diseases for each chromosome is the OMIM database("OMIM - Online Mendelian Inheritance in Man" 2017), because it has the information from the peer-reviewed biomedical literature updated and curated, resulting in reliable descriptions of human diseases and their most probably causing genes. It provides tools as genome coordinate searching and thesaurus-enhanced search term options.

A possibly confounding variable is the length of the genes, because there is more space to suffer mutations. To assess that this is not affecting the analysis, it can be analised the correlation between the number of genes and the total exon length in bp, obtained from ("Size of Human Transcriptome/Exome for Coverage Calculation - Seqanswers" 2017). If they are not significatively correlated, the study will have to be repeated for both measures (gene and exon length) and the results compared.

Other confounding variables that can not be controlled are those related to the diseases; for instance rare diseases tend to be less studied that those that affect a wide portion of the population, or the studies are less reliable due to the sample size. In addition, not all kind of mutations, genes and chromosomes are equally easy to study. These factors can be partially randomized by having into account diseases with different amounts of knowledge associated.

The statistical test

There is available a list of number of observations (diseases) for each category in a nominal variable (chromosome), enough data to take a large sample (>100 observations, >5 observations per category), and it has to be tested the adequation of this data to an expected distribution, so in this case the best statistical test would be a Chi-square test of goodness-of-fit("Chi-Square Test of Goodness-of-Fit - Handbook of Biological Statistics" 2017). It is also expected a fitness to the parametric assumptions but it's not mandatory in this test, so there's no need to confirm it.

As the data is available before the study, a post hoc power analysis can be performed after the data acquisition and cleaning to determine the best sample size.

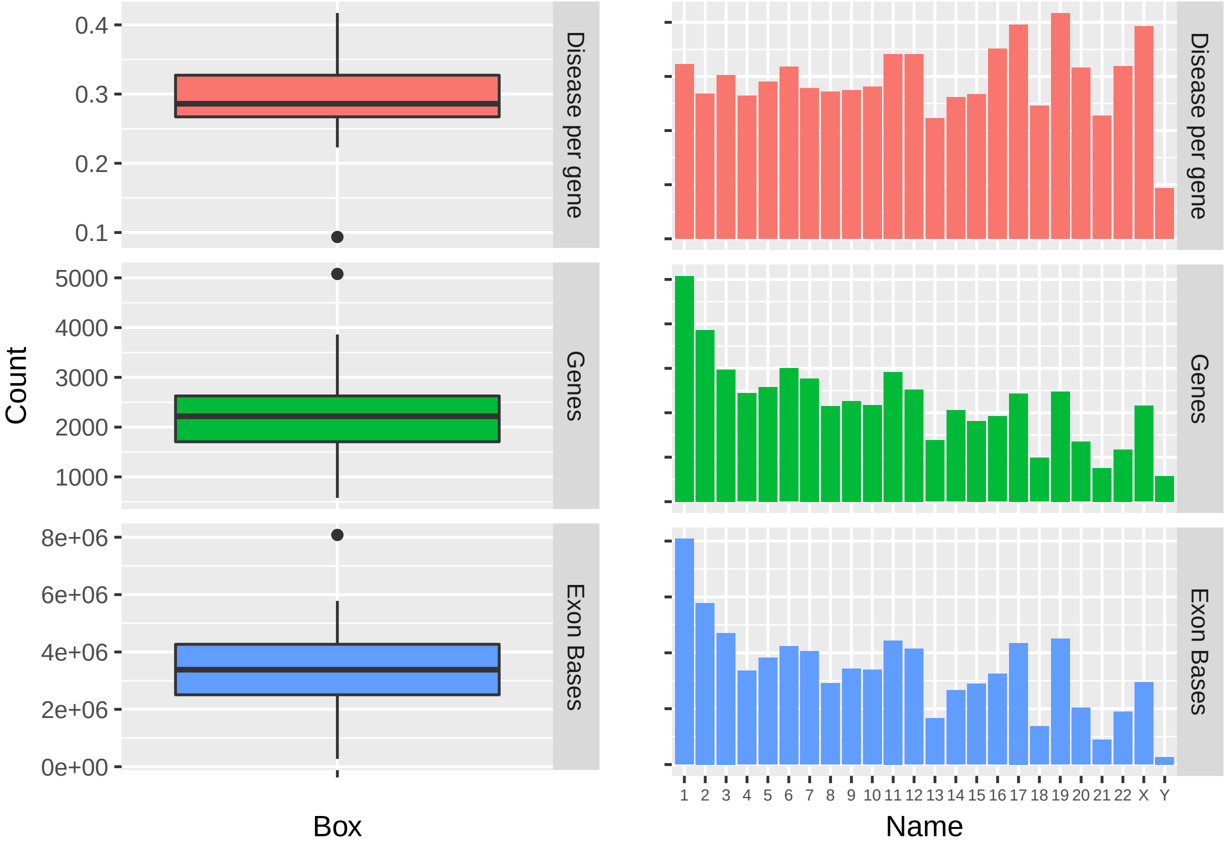
Data preparation

All the data was incorporated to the R environment. After that, some descriptive analysis were performed:

```
sapply(t[2:5], summary)
```

```
##         Size_(Mb) Gene Number_Diseases Exon_Bases
## Min.        46.71  577            54.0     271500
## 1st Qu.     82.54 1706           471.2    2508000
## Median     133.50 2218           660.5    3382000
## Mean       128.70 2242           687.4    3379000
## 3rd Qu.    162.20 2627           912.2    4268000
## Max.       249.00 5078          1638.0    8079000
```

In addition, the plot below summarizes all the numerical data with boxplots and barplots. They can be observed several outlier points corresponding to chromosome y for number of diseases and 1 for Genes and Exon Bases.



Also, as Genes and Exon Bases are significatively correlated, the Chi square test can be performed using only Genes as a predictor (see Extra Plot 1).

```
cor.test(t$Gene, t$Exon_Bases)
```

```
##
##  Pearson's product-moment correlation
##
## data:  t$Gene and t$Exon_Bases
## t = 23.722, df = 22, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9558909 0.9918818
## sample estimates:
##       cor
## 0.9810078
```

Finally, the post hoc power test suggested that the a good sample size would be of around 800 diseases. This allows to have a power different from 1, which would be problematic because Chi-square tests are sensitive to the excess of data, giving false positives.

```
P0<-t$Gene/sum(t$Gene)
P1<-t$Number_Diseases/sum(t$Number_Diseases)

effect.size <- ES.w1(P0, P1)

pwr.chisq.test(w=effect.size,  N=NULL,  df=23,power=0.8,  sig.level=0.05)
```

```
##
```

```
##      Chi squared power calculation
##
##               w = 0.1712296
##               N = 754.3725
##              df = 23
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

Statistical test

First, a sample was acquired in order to have the right N for this test. It can be seen in that 800 individuals can be representative of the total (see Extra Plot 2). However, to mantain the low effect size, the original sample will be multiplied by 0.05 instead of sampling 800 individuals.

After this,the new was also calculated to adjust the significance level.

```
  chisq.test(x=Observed, p =Expected_Frq)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Observed
## X-squared = 24.186, df = 23, p-value = 0.3936
```

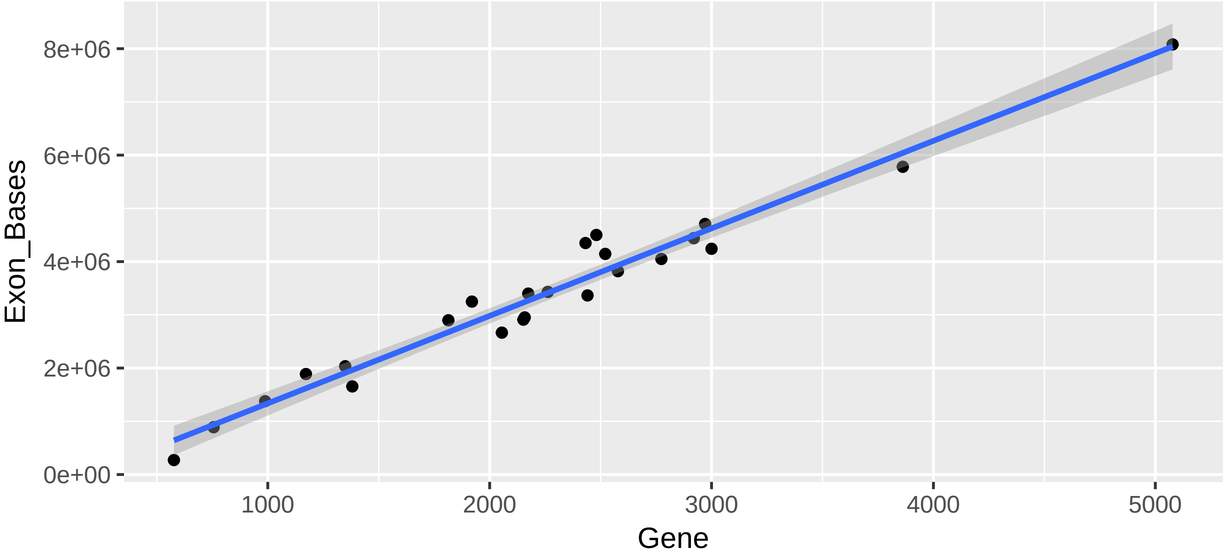
```
# NEW POWER

  P0<-Expected_Frq
  P1<-Observed/sum(Observed)

  effect.size <- ES.w1(P0, P1)

  pwr.chisq.test( w=effect.size,N=sum(Observed), df = 23, power= 0.8,   sig.level = NULL)
```

```
##
##      Chi squared power calculation
##
##               w = 0.1712296
##               N = 824.9
##              df = 23
##       sig.level = 0.03297974
##           power = 0.8
##
## NOTE: N is the number of observations
```

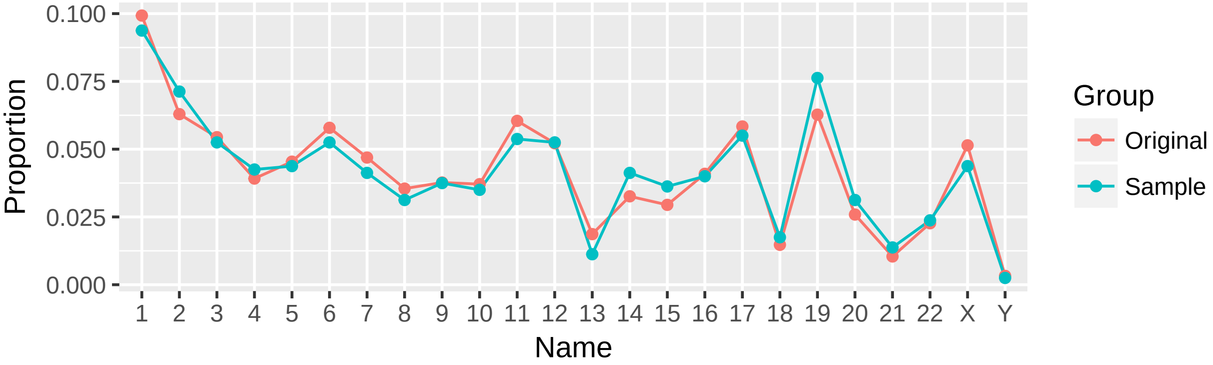
Afther these analysis, we can't reject the null hypothesis, meaning that both observed and expected distributions are the same (See Estra Plot 3) and taht there's no apparent presence of a chromosome that causes more diseases than expected.

Extra plots

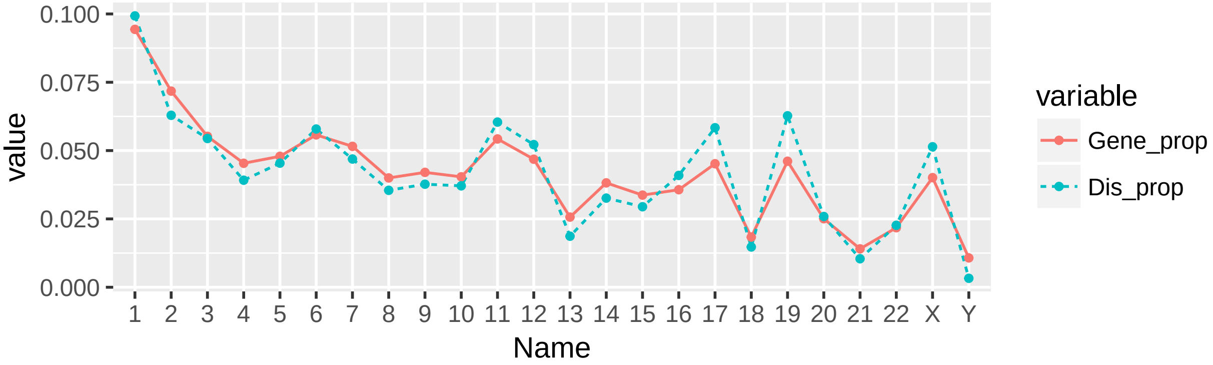
Extra Plot 1



Extra Plot 2



Extra Plot 3



Bibliography

Casillas, Sònia, and Antonio Barbadilla. 2017. "Molecular Population Genetics." *Genetics* 205 (3). Genetics: 1003–35. doi:10.1534/genetics.116.196493.

"Chi-Square Test of Goodness-of-Fit - Handbook of Biological Statistics." 2017. Accessed October 31. http://www.biostathandbook.com/chigof.html.

"Homo Sapiens Genome, Ncbi." 2017. Accessed October 31. https://www.ncbi.nlm.nih.gov/genome/51?genome_assembly_id=214366.

"OMIM - Online Mendelian Inheritance in Man." 2017. Accessed October 31. https://www.omim.org/.

"Size of Human Transcriptome/Exome for Coverage Calculation - Seqanswers." 2017. Accessed October 31. http://seqanswers.com/forums/showthread.php?t=5298.