## **Summary of Mini Project in Bioinformatics**

**1. Introduction**

Next-generation sequencing (NGS) has made a significance change in science, particularly in biology and molecular biology. It is widely used in genetic, epigenetic and transcriptomic research.

The NGS RNA-Sequencing (RNA-Seq) enables transcriptomic analysis that enable quantifying the expression of the genes or other cellular molecules. Nowadays there are several commercial platforms available to perform RNA-Seq.

Studies based on RNA-Seq are usually published in two complementary databases of the NCBI: (a) Pubmed – a database of abstracts and manuscripts in life sciences and biomedical topics, and (b) GEO - a public repository that archives high-throughput functional genomics data submitted by the research community.

In order to complete the picture of data related to RNA-Seq, SRA was searched as an additional database. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms.

In this study we used database analysis tools and text mining in order to analyze the trends in recent years in transcriptome studies, specifically RNA-seq. We focused on different transcriptomic approaches, on diseases studied using RNA-Seq and on molecules sequenced in this technology.

This study may give a foresight of how RNA-Seq will be used in clinical and scientific researches in the coming years.

**2. Methods:**

We scanned the GEO and the Pubmed databases for manuscripts/studies conducted in the last decade (2010-2018).

The GEO was scanned for the types of experiments. All the GEO entries were downloaded manually and then text mined in order to search for different keywords. We focused on the ‘Experiment Type’ field in each GEO entry. The number of GEO entries was calculated using each category. Subcategories were merged into higher-class categories. A similar analysis was performed for the different platforms used in the studies.

Searching SRA might give data that is more reliable than the data of GEO since all the data related to high throughput sequencing has to be submitted to SRA and only some of it is uploaded to GEO, according to its submission processes.

The Pubmed was searched in two different ways: (a) We used the R package RISmed to contact Pubmed and download relevant abstracts in order to search for keywords, and (b) we created specific queries in our code using the R package rentrez, send them to the Pubmed and directly got the number of the relevant abstracts.

We looked at abstracts of RNA-seq related to:

(i) diseases - PubMed was searched for RNA-seq manuscripts containing words associated with 18 different disease categories. For the different diseases, we used the mineR package to extract the diseases and their associated words using text mining. The results were then manually filtered. The regression lines in the illustrating plot were added using the *lm* function. Univariate regression models were fitted to the number of diseases studied using RNA-Seq (using the year as the independent variable).

(ii) cancer types - The cancer types were classified by the type of cell that the tumor cells originated from. We extracted all cancer and RNA-Seq related abstracts and searched for the cancer types or the organs in which each cancer type mostly occurs.

(iii) cancers in different locations of the body - The division of body systems/locations into categories was made based on <https://www.cancer.gov/types/by-body-location>. For each category, we created a query based on all the subcategories, namely all the cancers related to a particular system/location. We send the query to Pubmed and counted the number of the relevant abstracts.

(iv) Cancers in population vs. in research - The percentages of each cancer type in the population was compared to the percentages of abstracts discussing RNA-Seq and the same cancer type.

(v) non-coding RNA

(vi) The relations between diseases and non-coding RNA were examined. For each non-coding RNA we searched which diseases are most studied in the context of this specific molecule and vice versa. Chi-square test was used for goodness of fit in order to compare the expected distribution of molecules related to diseases to the observed distribution. P-value was computed by Monte Carlo simulation.

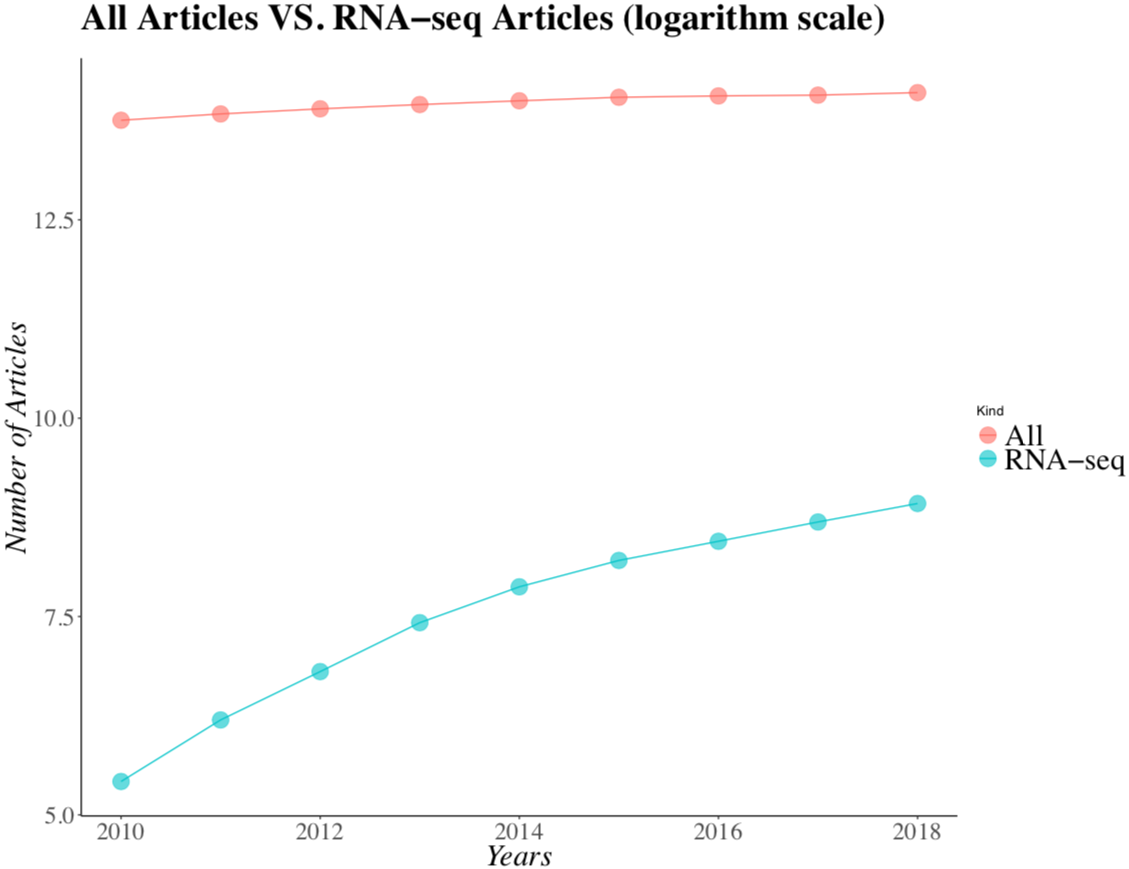
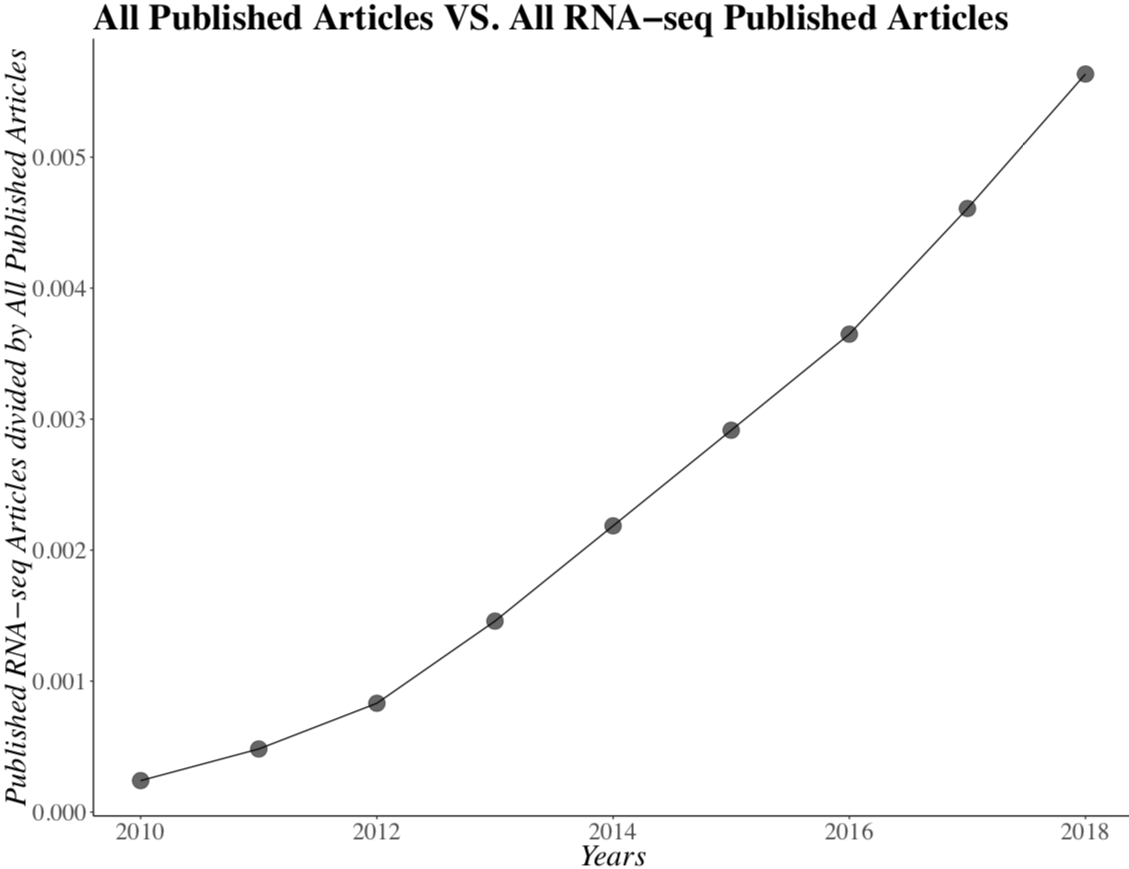
(vii) Additional uses and subcategories of RNA-Seq.

The keyword selection for the queries was fine-tuned manually.

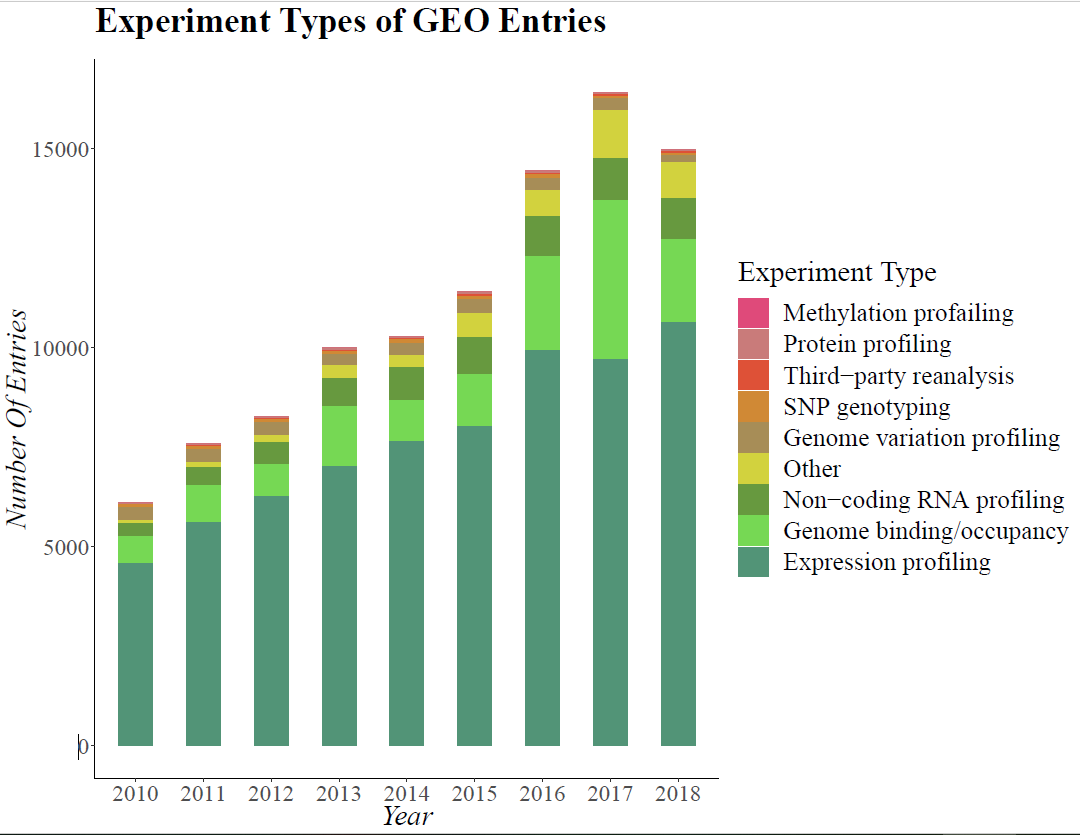
Data retrieval and statistical analyses were performed in R version 3.5.3. Graphs were created using the ggplot2 R package.

**3. Results:**

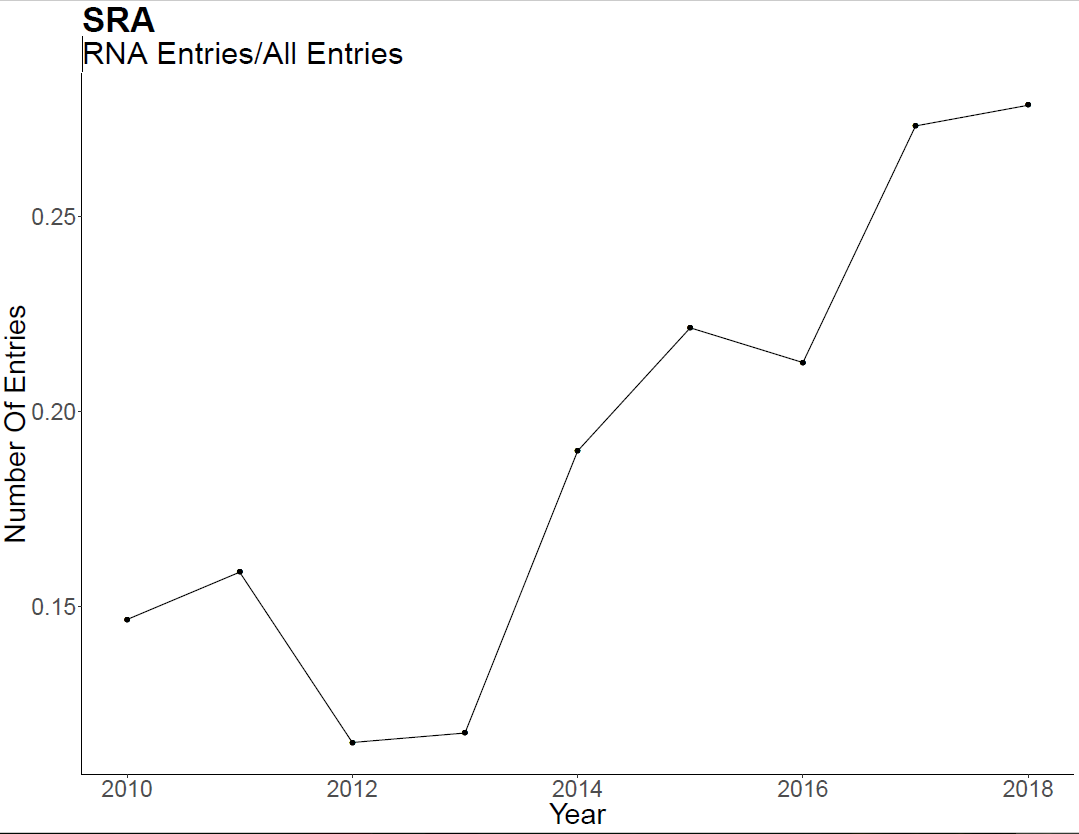
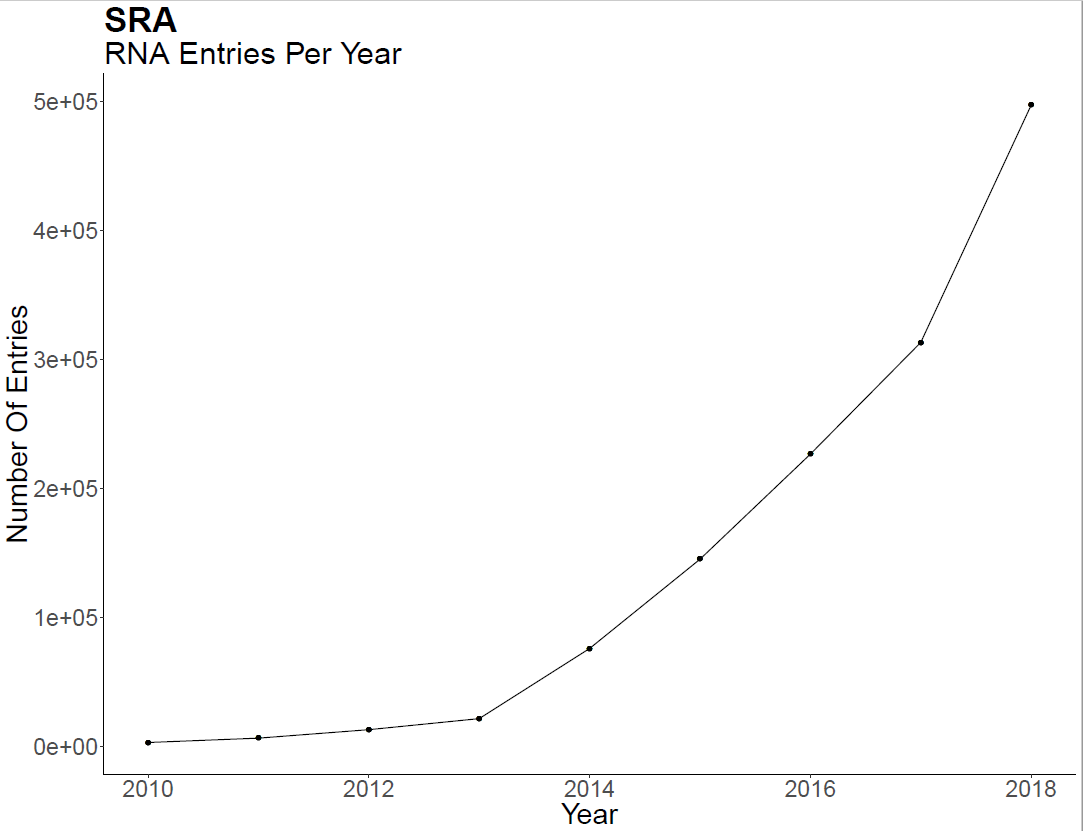
There was a dramatic increase in the number of papers published in Pubmed in the years 2010-2018 (Fig. 1a). However, the increase in the number of RNA-Seq papers published in Pubmed in these years was even higher (Fig. 1b). This is in accord with the significant increase in GEO entries of expression profiling (4615 in 2010 to 10664 in 2018, Fig. 1c), and rise in entries sourced in RNA uploaded to SRA(Fig. 1d and 1c)..

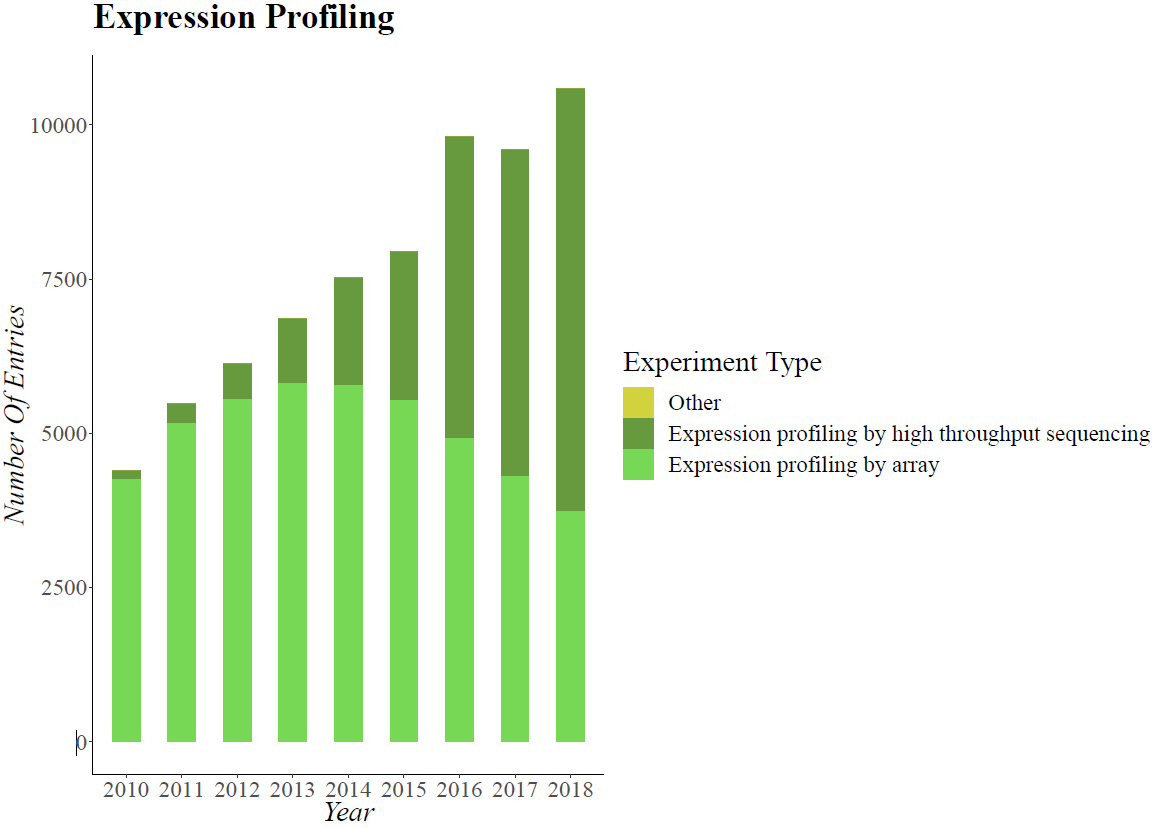
*Figure 1a Figure 1b*



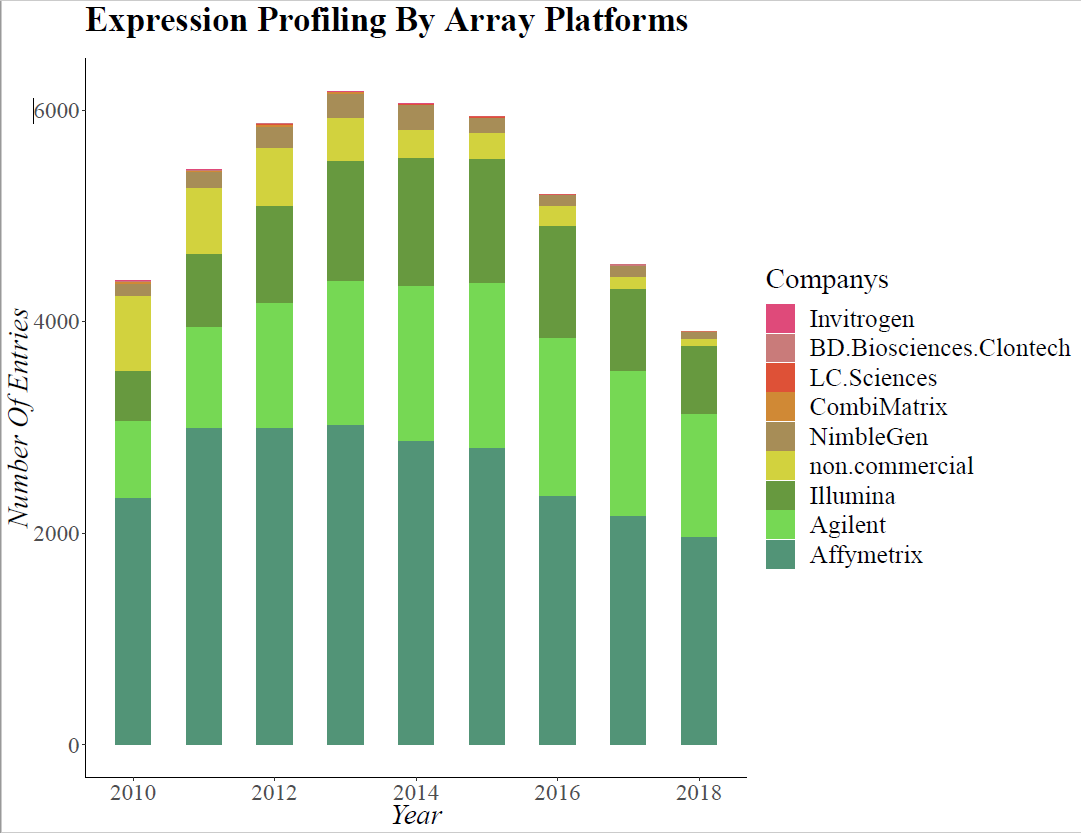
*Figure 1c* As mentioned above, this may not fully reflect the real picture of all published experiments, hence for example the SNP genotyping experiments are mostly updated to SRA and not to GEO. But yet it shows the rise of expression profiling.

*Figure 1d* *Figure 1e*

Analysis of the types of expression profiling reveals that throughout the last decade, high throughput sequencing has significantly increased while the use of arrays has decreases, but still play a significant role (Fig 2a). Focusing on the different array platforms reveals a gradual decrease in the use of Affymetrix© since 2013 and sharp decrease in non-commercial arrays (Fig 2b).



*Figure 2a*



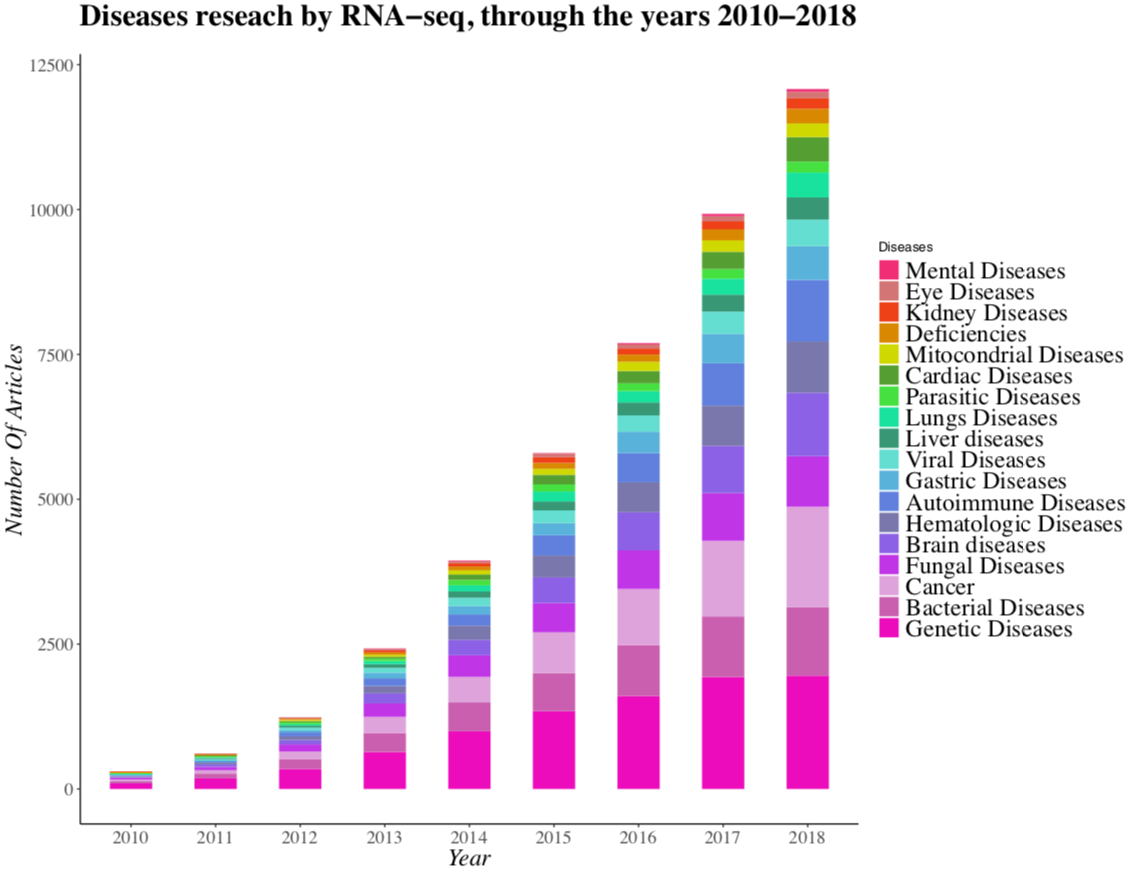
*Figure 2b*

Diseases and cancers studied by RNA-Seq

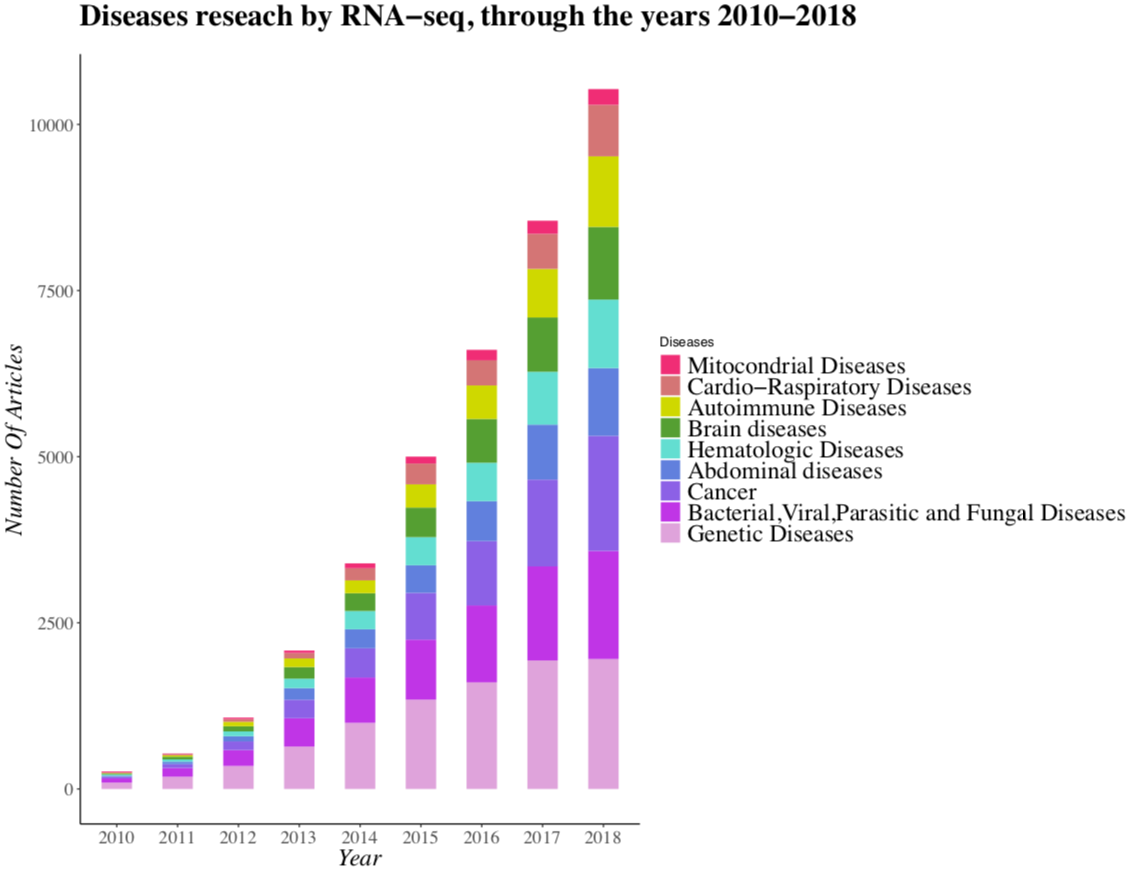
RNA-Seq has become highly popular in scientific and clinical research. It has become a major tool in the research of a variety of diseases, mostly genetic diseases, cancers, brain diseases and bacterial induced diseases (Fig 3a). Additionally, after combining some of the disease classifications (e.g. ‘Cardiac’ and ‘Lungs’ into ‘Cardio-Respiratory’ group), the order of magnitude almost hasn't changed - the Genetic, Bacterial and Cancer diseases are still the most studied in RNA sequencing (Fig 3b).

Quantifying the dependencies between number of publications of diseases based on RNA-Seq and the year revealed high betas (Fig 3c).

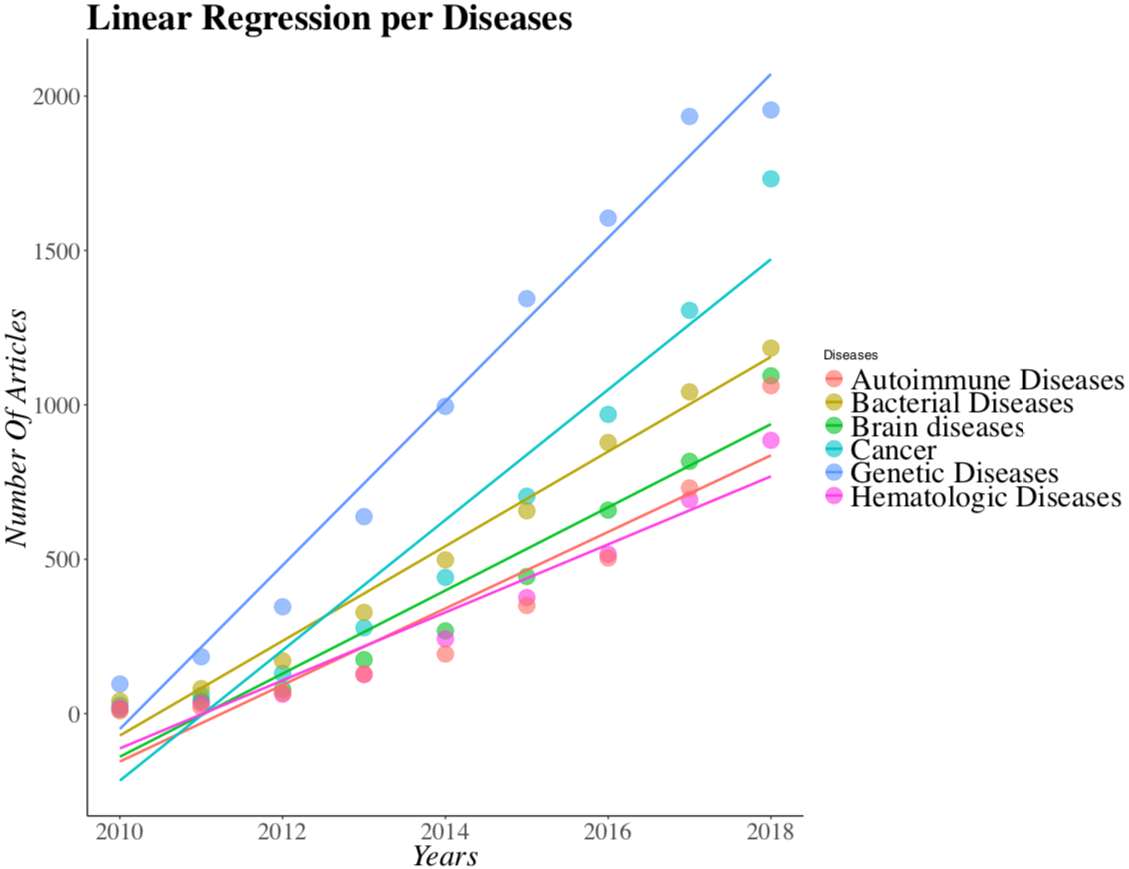
Focusing on the different types of cancers and body systems reveals that most of the cancer publications based on RNA-Seq are of carcinomas, which is the most common type of cancer in human (Fig 3d) .The most studied cancered organ (by RNA-Seq) is the breast (Fig 3e).



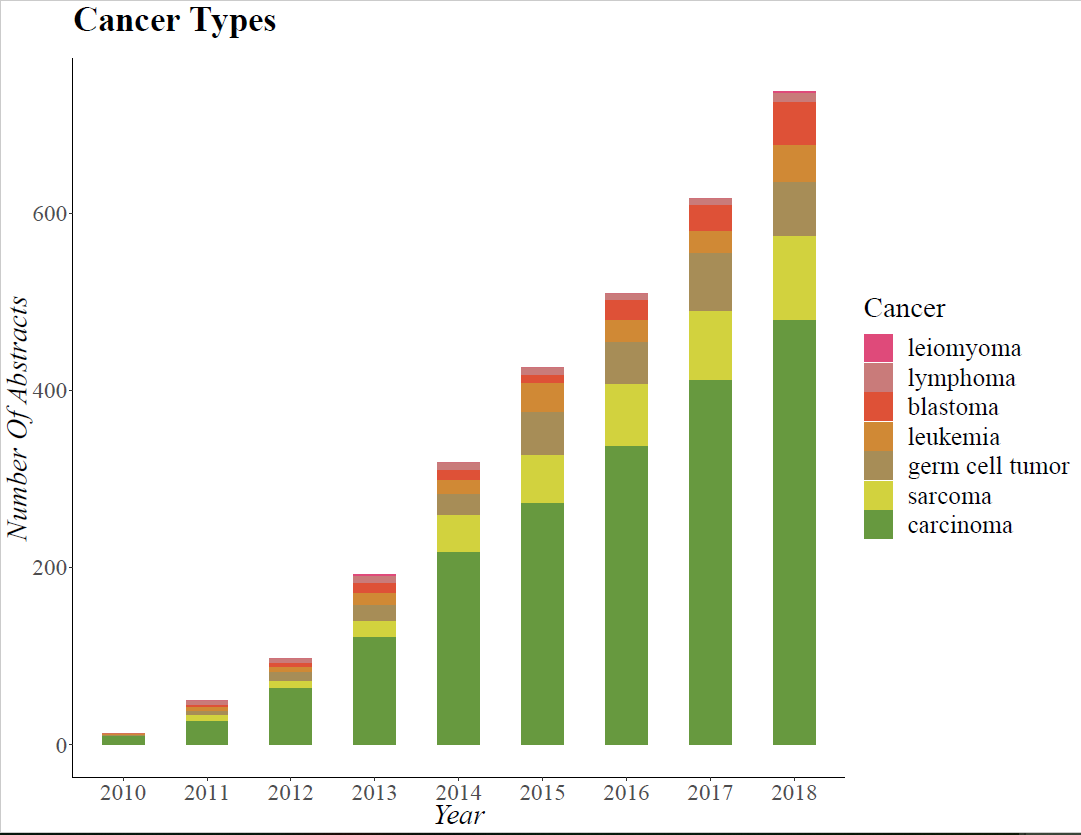
*Figure 3a*

**

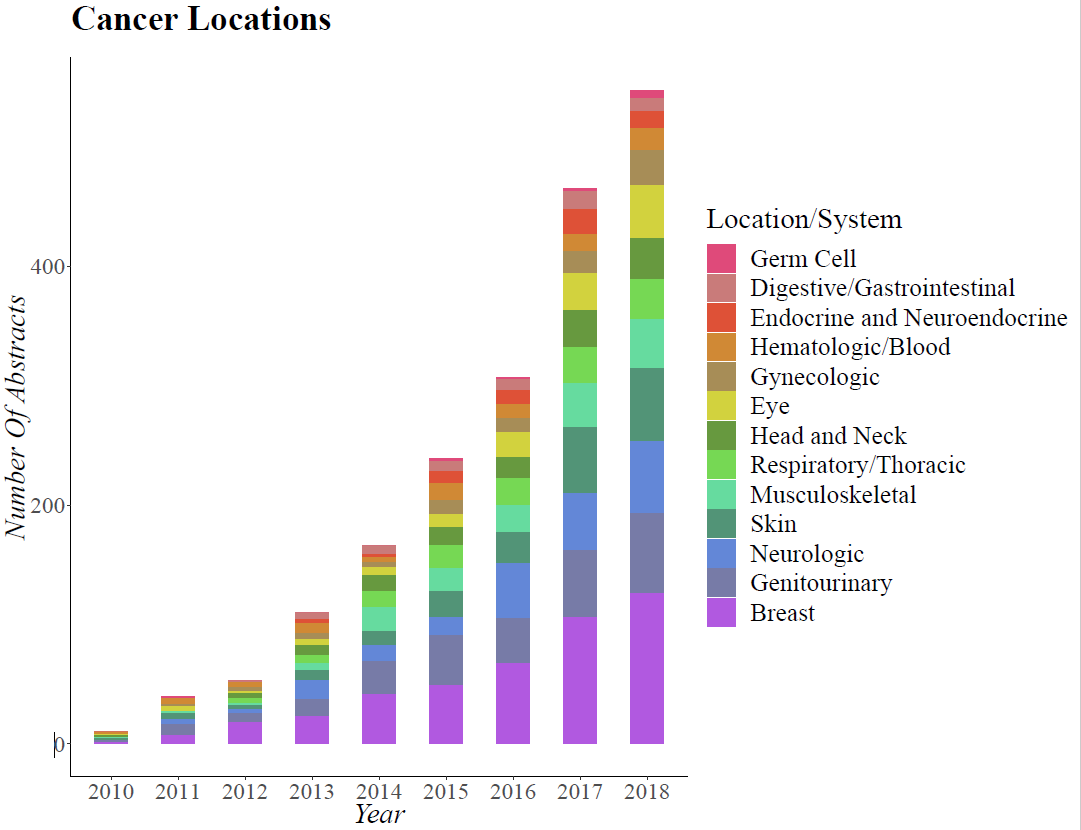
*Figure 3b*



*Figure 3c:* R-squared for each category of disease: Cancer – 0.9247, Genetic Diseases – 0.9789, Brain Diseases – 0.9265, Bacterial Induced Diseases – 0.9811, Hematological Diseases – 0.9291, Immune Diseases – 0.8753.



*Figure 3d*

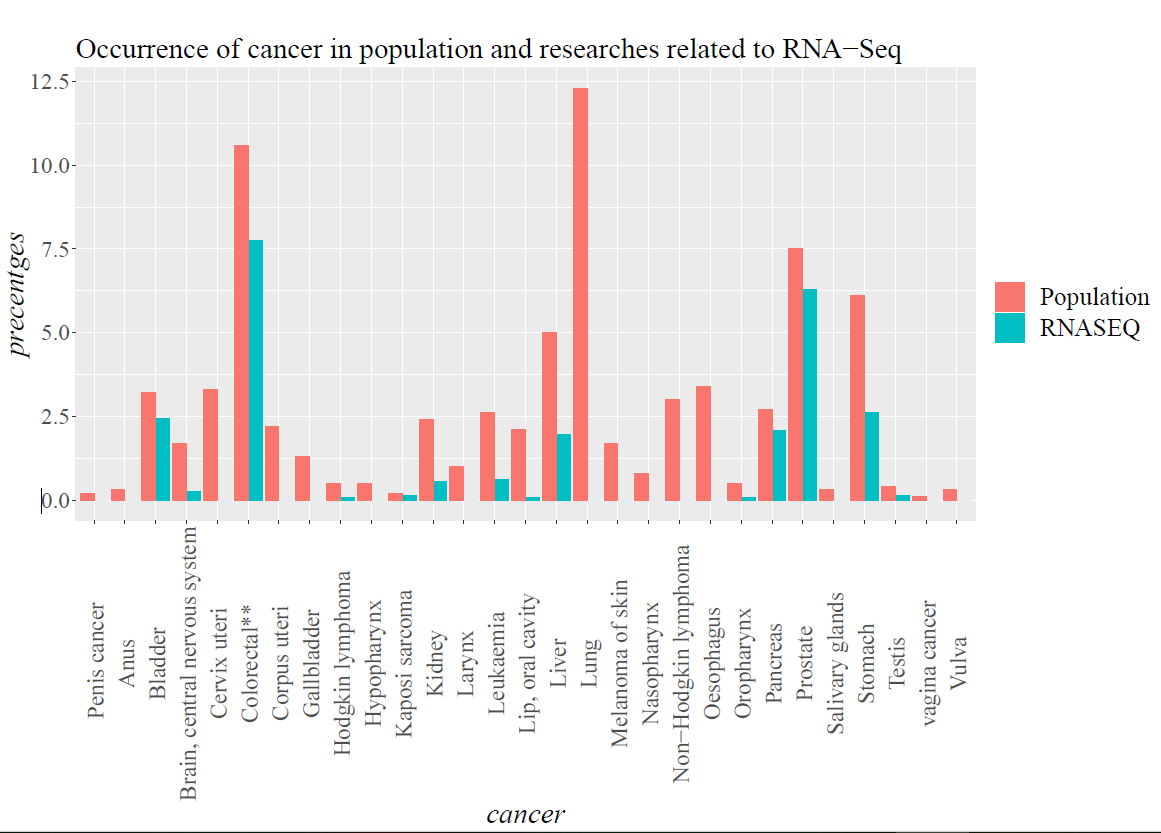


*Figure 3e*

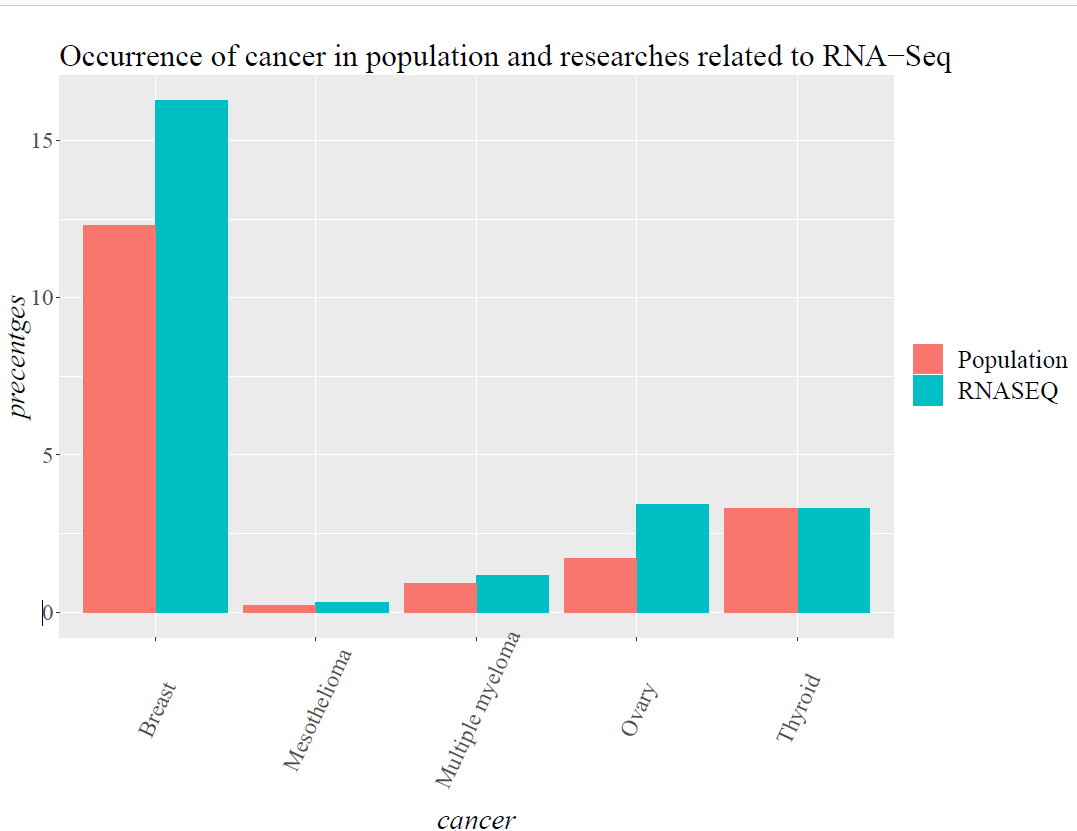
Table 1a and Figures 3e, 3f show the relation between the occurrence of different cancer types in the population and the frequency of the same types in researches related to RNA-Seq.

|  |  |  |
| --- | --- | --- |
| Cancer Types | RNA-Seq abstracts of cancer type (in % to other cancers) | Cancer occurence in population (in % to other cancers) |
| Lung | 17.881 | 12.3 |
| Breast | 16.289 | 12.3 |
| Thyroid | 3.307 | 3.3 |
| Ovary | 3.429 | 1.7 |
| Multiple myeloma | 1.164 | 0.9 |
| Mesothelioma | 0.306 | 0.2 |
| Colorectal cancer | 7.777 | 10.6 |
| Prostate | 6.307 | 7.5 |
| Stomach Cancer | 2.633 | 6.1 |
| Liver cancer | 1.96 | 5 |
| Oesophagus cancer | 0 | 3.4 |
| Cervix uteri cancer | 0 | 3.3 |
| Bladder | 2.449 | 3.2 |
| Non-Hodgkin lymphoma | 0 | 3 |
| Pancreas | 2.082 | 2.7 |
| Leukaemia | 0.612 | 2.6 |
| Kidney Tumors | 0.551 | 2.4 |
| Corpus uteri cancer | 0 | 2.2 |
| Lip and Oral Cavity | 0.061 | 2.1 |
| Brain Cancer / central nervous system cancer | 0.245 | 1.7 |
| Melanoma of skin | 0 | 1.7 |
| Gallbladder cancer | 0 | 1.3 |
| Larynx cancer | 0 | 1 |
| Nasopharynx cancer | 0 | 0.8 |
| Oropharynx cancer | 0.061 | 0.5 |
| Hypopharynx cancer | 0 | 0.5 |
| Hodgkin lymphoma | 0.061 | 0.5 |
| Testis cancer | 0.122 | 0.4 |
| Salivary glands cancer | 0 | 0.3 |
| Anus cancer | 0 | 0.3 |
| Vulva cancer | 0 | 0.3 |
| Kaposi sarcoma | 0.122 | 0.2 |
| Penis cancer | 0 | 0.2 |
| vaginal cancer | 0 | 0.1 |

*Table 1a.* The green background in the table shows the cancer types which have a higher percentage in research than occurence in the population. The red background shows the types that are more common than researched, relative to other cancer types.

*figure 3e*

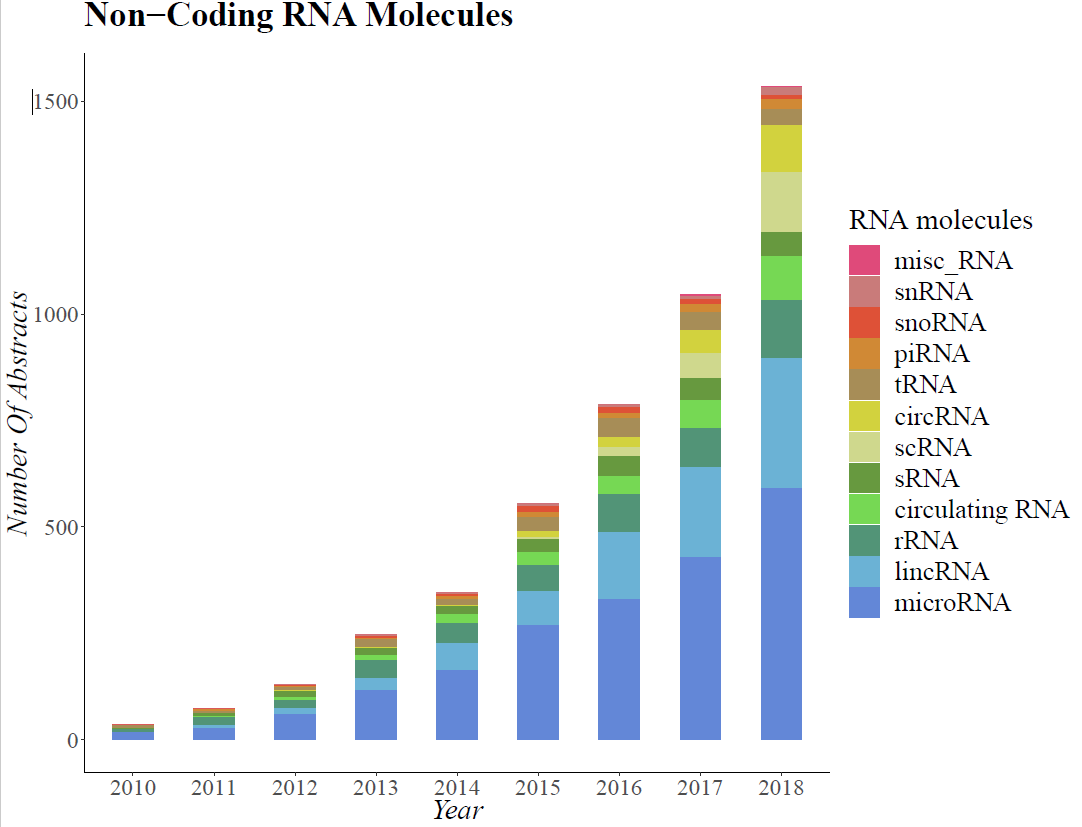
Cancers with highest percentages in population.

*figure 3f*

Cancers with higher percentages in RNA-Seq.

Molecules studied by RNA-Seq

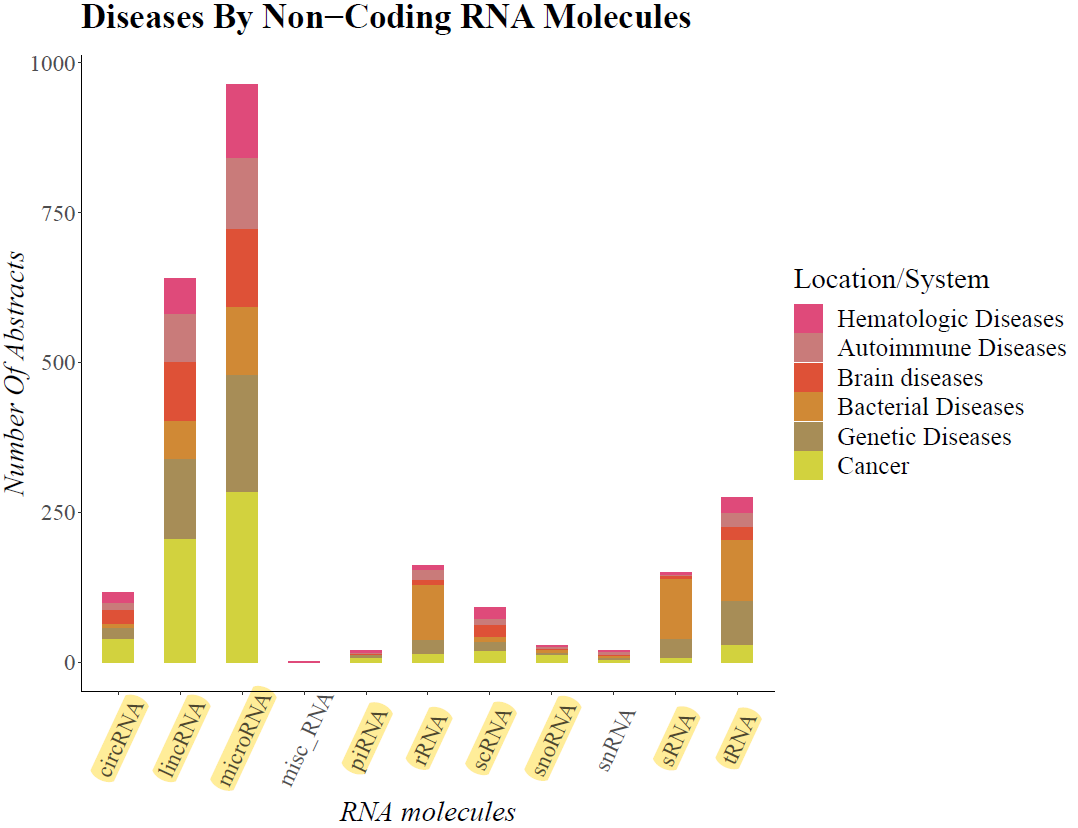
We looked at which molecules were studied using RNA-Seq. As can be seen in Fig. 4a, miRNAs are the most studied non-coding RNAs, and there is a dramatic increase in the studies that used these molecules (11 in 2010 to 290 in 2018). In addition, more and more non-coding molecules have been discovered since 2010 and are also studied in these technologies.



*Figure 4a*

Diseases/Cancers and non-coding RNA in RNA-Seq

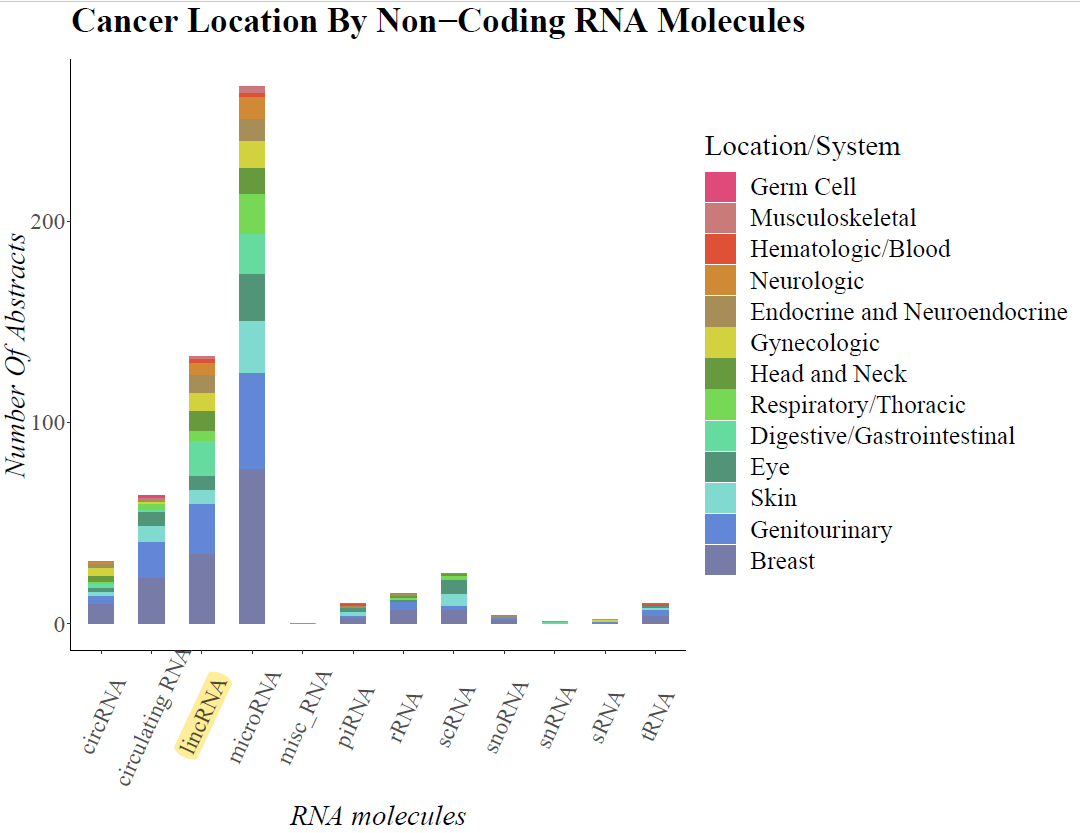
We analyzed the interplay in RNA-Seq studies between diseases and molecules: which diseases are studied using expression quantification of which molecules. We compared these findings to the distribution of molecules used in all RNA-Seq studies (not necessarily for disease research) in order to get the statistical significance of the results.



*Figure 5a.* The molecules marked in yellow are highly significant.

Fig. 5a: The significant molecules with their p-values

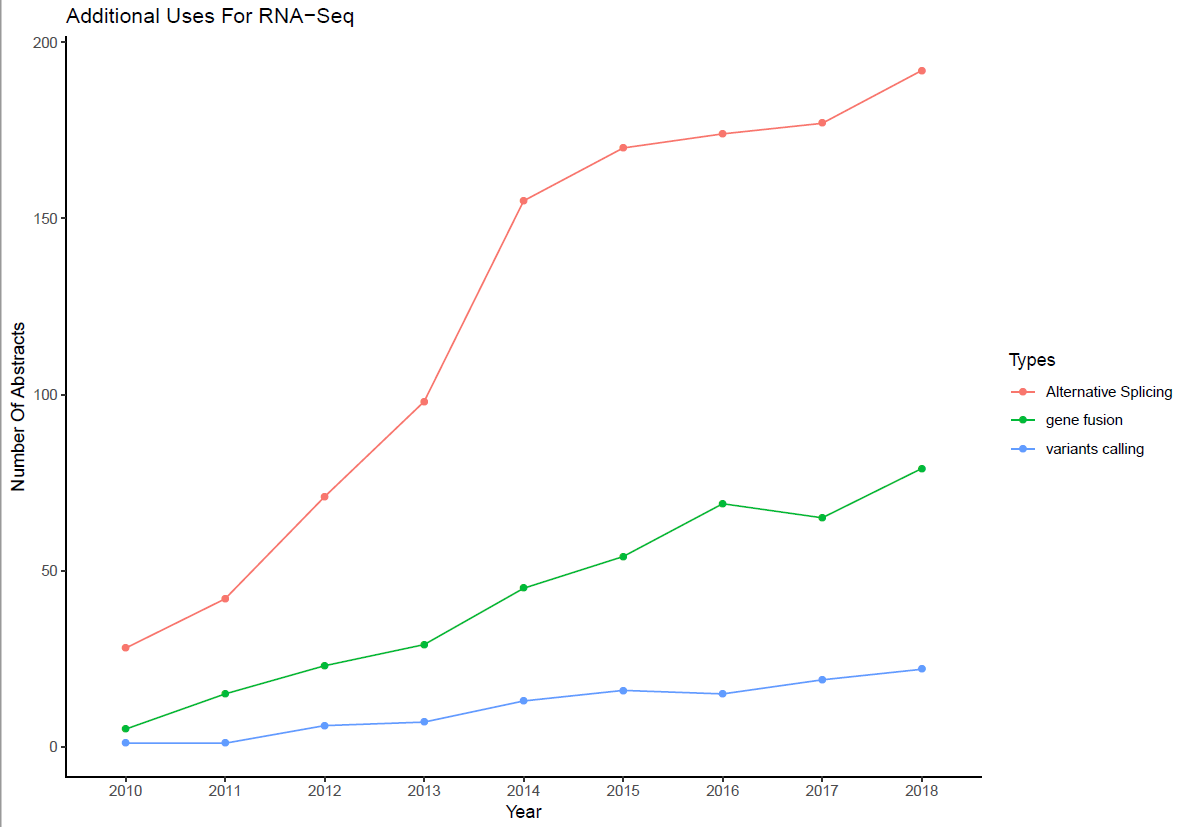
rRNA – 0.0005, tRNA – 0.0005, siRNA – 0.0005, lincRNA – 0.0005, microRNA – 0.0005, scRNA – 0.001, sRNA - 0.0005, snoRNA – 0.03, circRNA – 0.0005, piRNA = 0.02,



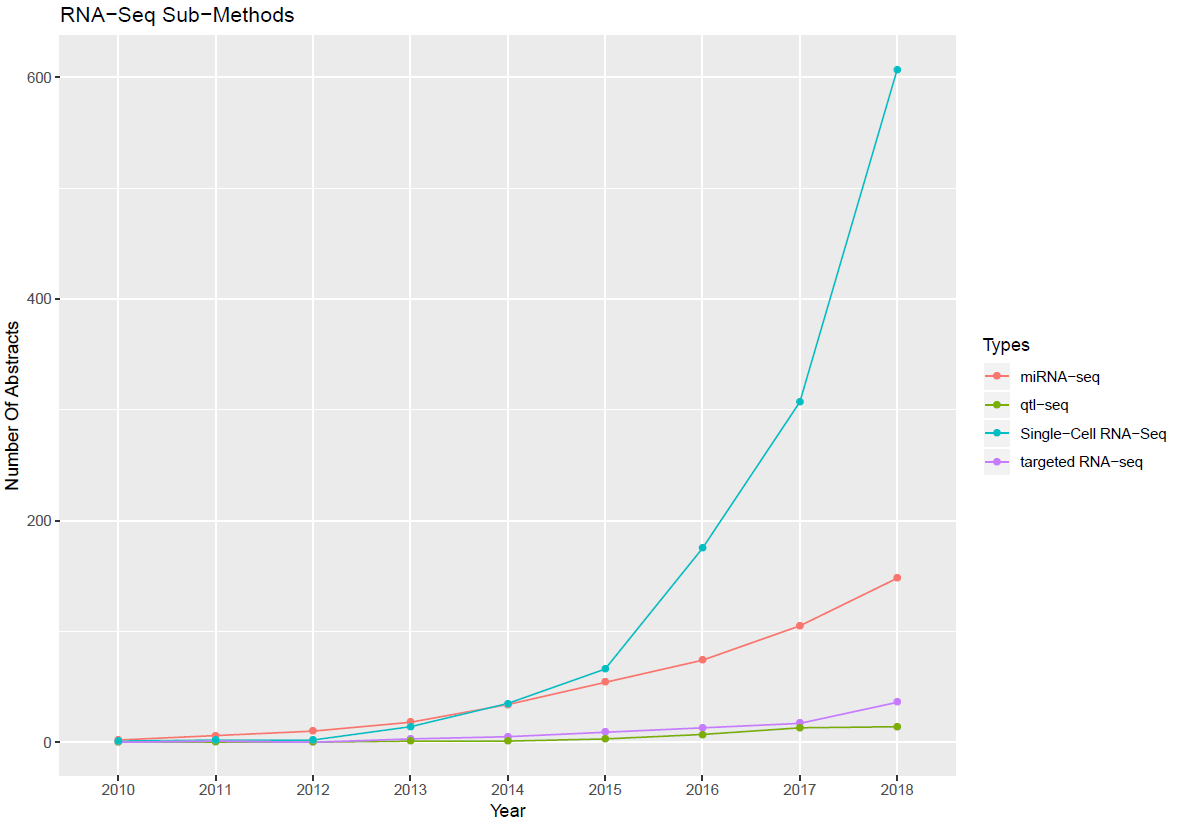
*Figure 5b* The molecule marked in yellow was statistically significant (using chi-square test for goodness of fit). Significant results: lincRNA with p-value 0.0005

Additional uses and subcategories of RNA-Seq:

There are many more fields where RNA-Seq can be applied to. Searching 3 types of these fields (Alternative Splicing, Gene Fusion, Variants Calling) showed a rise in the number of articles in the last decade, though not as expected.



*Figure 6a*

**

*Figure 6b*

**4. Discussion**

In this study we show a significant rise in the use of RNA-seq, this was significantly observed both in Pubmed and in GEO. This rise originates from the rapidly progressing development in technology and from the decrease in expenses of sequencing.

Surprisingly, there is still a vast use in RNA arrays for Expression Profiling…

In this work we show that several cancer types, such as breast cancer are widely studied using RNA-Seq…

One of the significant advantages in the use of next-generation sequencing is the fact that these technologies are unbiased and enable finding novel molecules. Therefore, as the use in RNA-Seq increases, more and more novel molecules are being discovered and enhance the additional RNA-Seq studies of these molecules (positive feedback loop).

Study limitations

1. Some of the analyses were based on text mining. Since the Pubmed was searched using certain keywords such as RNA-Seq the results consist of all the abstracts containing this keyword although the case might be that the word was only mentioned and not actually studied in the research, or if an important key word is not mentioned in the abstract but only in the article.

*Additional thoughts and analyses:*

1. *Calculate some of these parameters for 2005, 2000?*
2. *Molecules - it is important to calculate the baseline, meaning what was the total of paper (not necessarily RNA-Seq) for each of the molecules (miRNA, piRNA etc…)*
3. *Track sub-technologies / approaches:*
   1. *targeted RNAseq*
   2. *miRNA-seq*
   3. *qtl-seq*
   4. *Single-Cell RNA-Seq*
   5. *DGE-Seq (Digital Gene Expression)*
4. *Additional uses of RNA-Seq*
   1. *variants calling*
   2. *fusion gene*
   3. *Alternative Splicing*
   4. *denovo transcriptome*

**5. Acknowledgments**

We thank Dr. Gigli Friedlander for her helpful comments.