PATHONEXT

Custom-based analyses for oncogenomics using the programming language R

Y. Gager[1], *I. Vogl*[1], *K. Neumann*[2] and *J. Gabert*[1]

[1]Pathonext GmbH, Leipzig, Germany
[2]Städtisches Klinikum Dessau, Dessau, Germany

Introduction

The field of **oncogenomics** is tackling **cancer** as a major health problem. The **high-througput-sequencing technologies** currently available allow for parallel sequencing of numerous genes. Large genomic data sets from different cancer tissues can be obtained within a few days.

However, the **analysis of these datasets remains a major challenge**. Corporate products are available for different analyses (e.g. variant calling) but they are often expensive and limited in their applicability. Here we present R - a free and open source programming language - for **manipulating, analysing and visualizing** complementary aspects of **cancer genomic data** including amplification, comparison between different samples or different sequencing events.

Methods

Sequencer: Miseq (Illumina)
Analyses: Biomedical Genomics Workbench (Qiagen), Rv3.4.3, R studio v1.1.383 and diverse R packages

Example 1: Gene amplification
Material: FFPE blocks of tumors (Städtisches Klinikum, Dessau)
Gene panel: TruSight Tumor 15 (Illumina)
Parameter: Copy number variation (standardized value z)

Example 2: Comparing two sequencing events
Material: Glioblastom tumor (Klinik für Neurologie, Leipzig)
Gene Panel: SureSelectQXT (Agilent Technologies)
Parameter: Frequency difference between similar variants

Results

Our results are saved as **HTML files** and can be explored using **action buttons to display/hide information** and **mouseover text to visualize metadata** of choice for each data point.

For a given patient and a selection of genes, we can visualize the **standardized value z for the copy number variation** and the average per gene (horizontal line, Figure 1). In our example, we can identify a clear **gene amplification** in the gene *ERBB2/HER2*but not in *TP53.*

For a given patient, we can also track down the **frequency difference for similar variants between two sequencing events** per chromosome (Figure 2). For example, one variant from the chromosome 7 increased from 60% from one run to the next (red point). Different **categories of variants** (SNV, Replacement, Insertion, Deletion and MNV) are displayed with a different type of symbols. Additionally, extra information for each variant can be visualized with the mouseover text. Such information concern the chromosomal region, the type of gene and the specific allele (orange box).

Our two examples comprise just a selection of how we can extract useful data relevant to the **diagnosis and evolution of the tumor** as well as to the **preconization of personalized treatments**.

Fig. 1

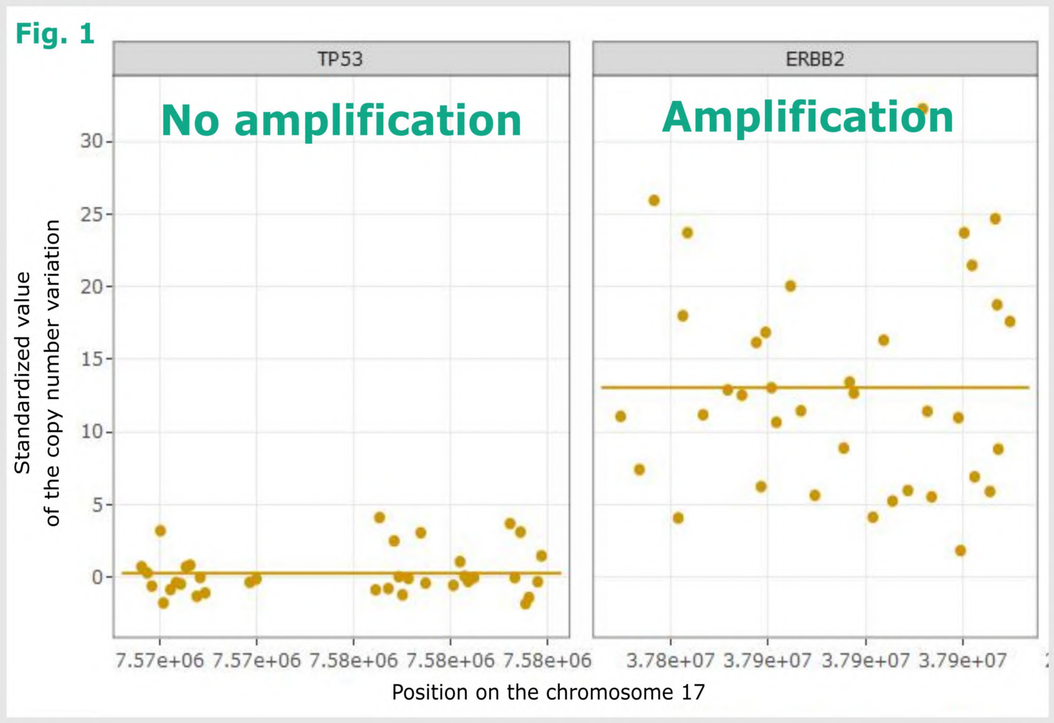
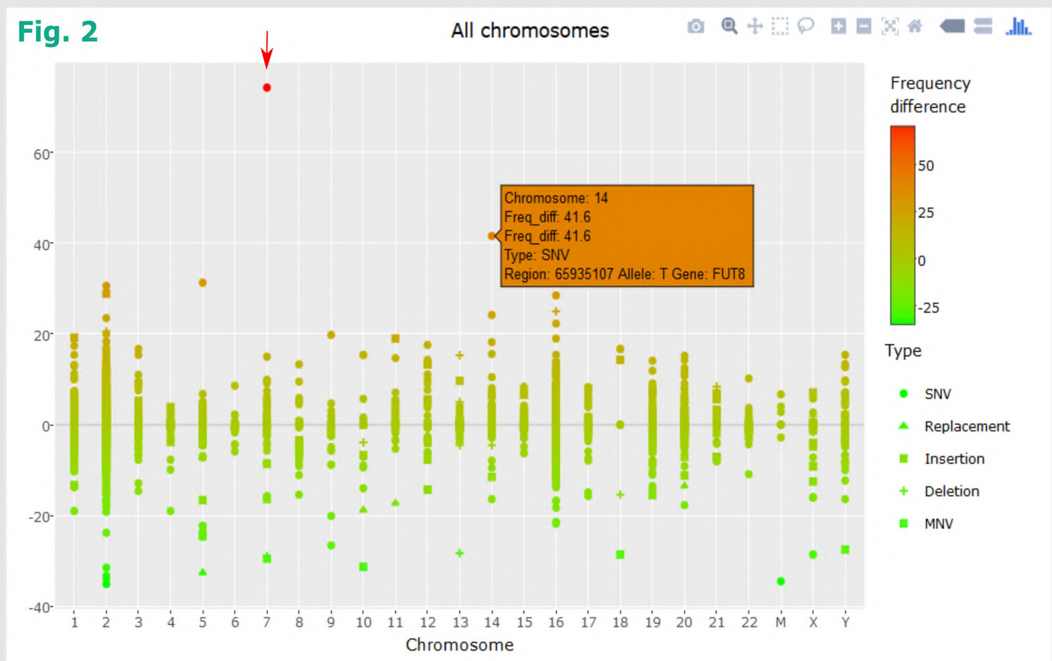


Fig. 2



Discussion

The **programming language** R appears to have many advantages. One of them is the **manipulation, analysis and visualization of data** within the same environment thanks to the numerous packages available (more than 4800). Another advantage is the possibility to import data directly from Excel and export data in the format of JPG, PDF or HTML.

The main challenge when using R is the **steep learning curve**. However, this process can be accelerated using **online communities** such as Stack Overflow as well as collaborating and exchanging R **scripts**.