Introduction

- Instroduction about SNAPShot

- The process of genotype calling

- Vary each run time

- Benefit of cloud base system

Star-allele calling from genotyping results is crucial for PGx implementation in

clinical practice. Additionally, a laboratory-developed test (LDT) panel for PGx

implementation is also critical and cost-effective for a specific population; however,

the star-alleles calling tools for the panel need to be developed. In this work,

therefore, we developed a star-allele calling tool applied for the developed testing

panel. The tool also was integrated into the system and website application.

Taken all into consideration, we develop a cloud base system that support user can call genotype and matching with our knowledge database.

We have developed modules to call genotype (star allele) from results derived from the Gene Mapper machine (gc.spmed.kr). This system is currently not integrated with the PDSS system.

1. Since the intensity of each run can be different, therefore, the peaks might occur outside of the bin standard range.
2. Users must adjust the bin range and re-analysis this sample to export new results.
3. Those processes made it inconvenient for users (suggest copy-paste and uploading in different systems).

**We would like to develop a module that supports:**

* **Users to call and adjust bin range and height directly on our web application system from a raw intensity file (FSA file)**
* **The application is expected to integrate with the PDSS system and cover processes from finding peaks from intensity data to genotype calling and integrate those data to make the report, including the QC plot.**
* **This is also a core function that we can provide a web application for our customers who bought our KIT**

Methods

**Loading FSA data**

Raw data from fragment analysis sofware were loaded using package Biopython (version) package. The intensisty data were stored in different predefined chanel name such as DATA105 for refrernce chanel, for other chanels defined for 4 nucleotites (DATA1: A; DATA2:C; DATA3:G and DATA4:T). Those intensisty data were primary used in peak detection process.

**Reference peak detection**

As mentioned aboved, the DATA105 chanel as refrence intensity was used to detect the reference peak. Usually, the information of reference peak depends on the experimen desgin. In this study, we used ??? GeneScan 120 Lize dye size standard (CYP2D6 kit developed by SPMED Co. Ltd. (GTR link)).

We used peak finding function of scipy (version) with default height 800, width of peak, ... However, the setting information can be adjusted by experties.

The number of detected peaks should be equal to number of reference size standard. In case of LIZ 120, there are 9 reference sizes, therefore the number of peaks should be 9 peaks corresponding with the reference size. Only qualifed peak detection was used to develop Least Square model as the next step, otherwise, it need to be adjusted by user to detect the correct reference peaks.

**Sizing model development**

As described in the manual of gene mapper software, there were three method to interfer length of DNA based on intensity points (Local Southern, Global Southern and Least Square). In the work, we used Least Square as advance method to develop a model for sizing identification (Ref).

The model was developed using numpy (version) python package. Users can choose the order of least sequare method (second-order or third-order). The performance of this model was evaluated using R square index from actual size and precited size. Based on R square, we found that third-order resulted in a slightly higher than second order, therefore, thrid order was set as default of the model. The model development process will be done by each input FSA file.

**Peak detection of raw data**

As mentioned before, raw data includes 4 chanels (DATA1, DATA2, DATA3, DATA4). Each chanel theorically defines for a nucleotite. The nucleotide can be change base on primer design in forward or reverse trand of the DNA.

The data chanel for peak detection process was selected based on which base and which the direction of the primer (Table). The direction of each marker was preloaded following kit information. (???Kit information, certificate)

**Star allele calling**

**Performance validation**

We validated our software using XXX sample including XXX cell lines and XXX real sample.

To compare the performance of the tools with experts, the only intensity plot of those sample were randomly distribute to 2 PGx experts. All other information of those sample were hidden. The manual calling result from two experts then gathered and compared. if any miss-matched between result of two persons, we invited the third expert to decided the genotype of this sample. The consuming time was also recorded.

The final call obtained by experts were compared with our tool's calling results.

Workflow

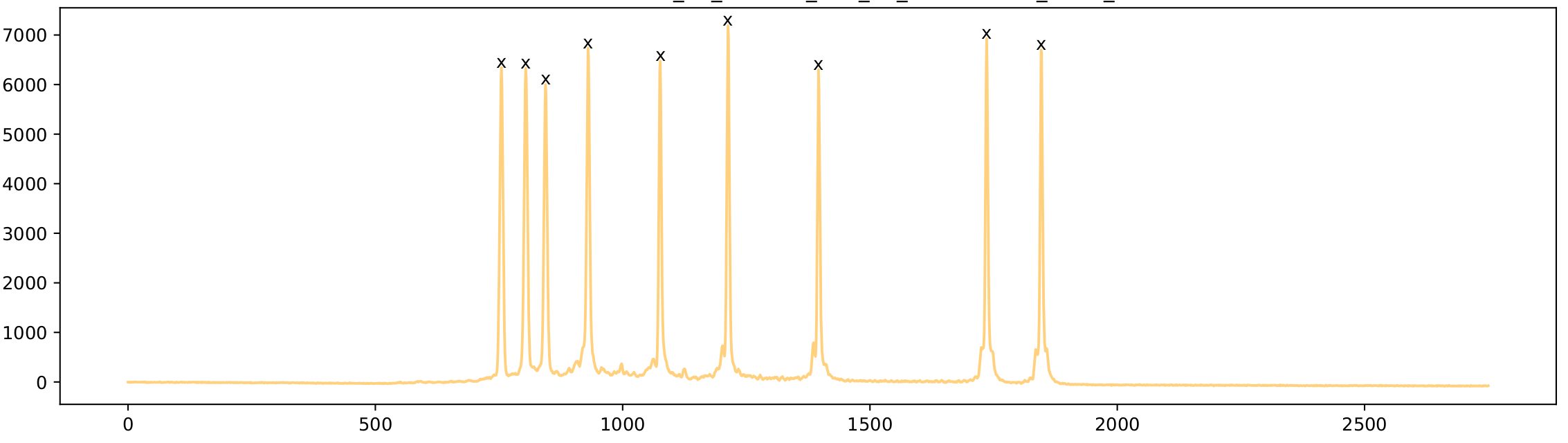
[https://app.diagrams.net/#G1HzMMyRkVhr8Y\_M1wMmXwmZagIEpak1Bb](https://app.diagrams.net/" \l "G1HzMMyRkVhr8Y_M1wMmXwmZagIEpak1Bb)

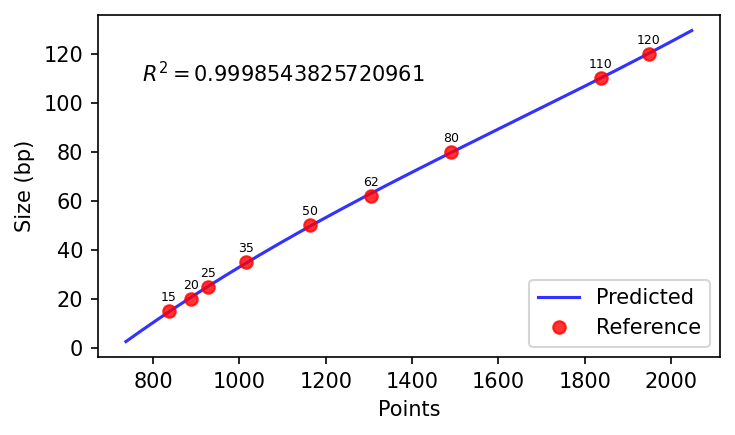
**Results**

**Reference peak detection**

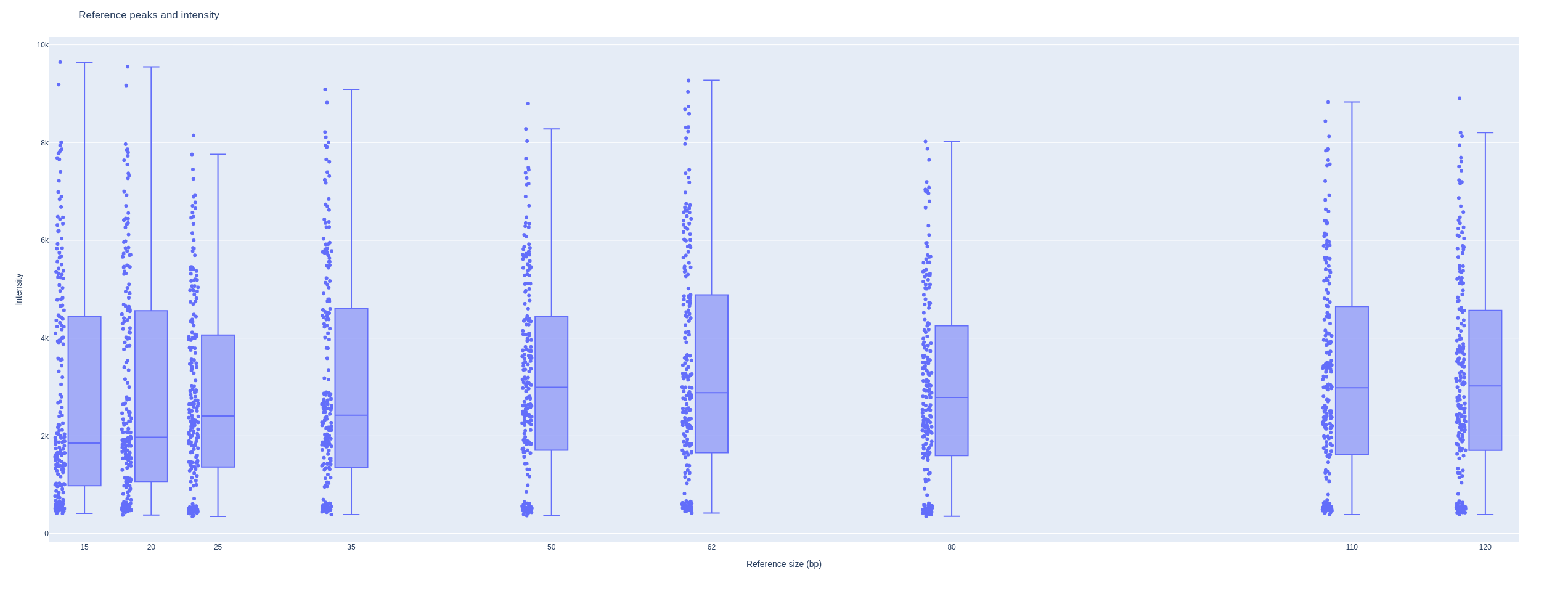
With minimum 300 intensity, we could detect correctly all reference peak of 101 samples (202 FSA files).

Least square third order were use to sizing model development, and we found that sizing model resulted in R square higher than 0.99XX (Figure)

**Peaks detection result**

****

**Need a plot for R square of all samples**



**Table**