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 above, the reference intensity (DATA105 data chanel) was used to detect the reference peak. Usually, the information of reference peak depends on the experiment design. 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 expertise.

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.

PGx data collection

PharmGKB as manual curation database is a standard database for PGx interaction. In this work, we collect all data about clinical guidelines, PGx drug label and clinical annotation data to build as a PGx knowledge database.

Workflow

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

**Results**

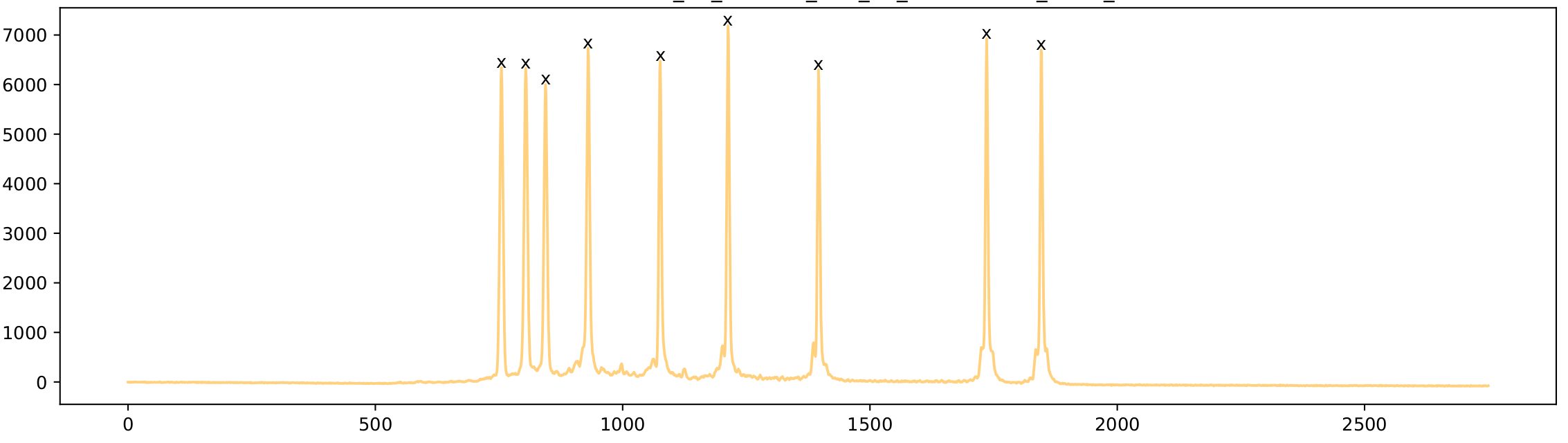
**Reference peak detection**

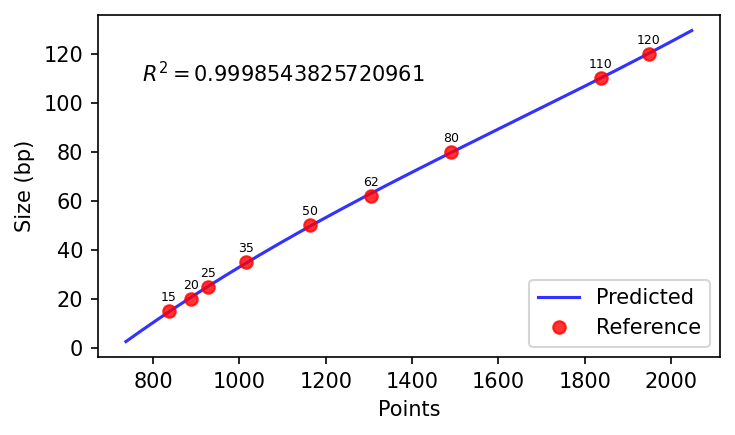
we could detect correctly all reference peak of 101 samples (202 FSA files).

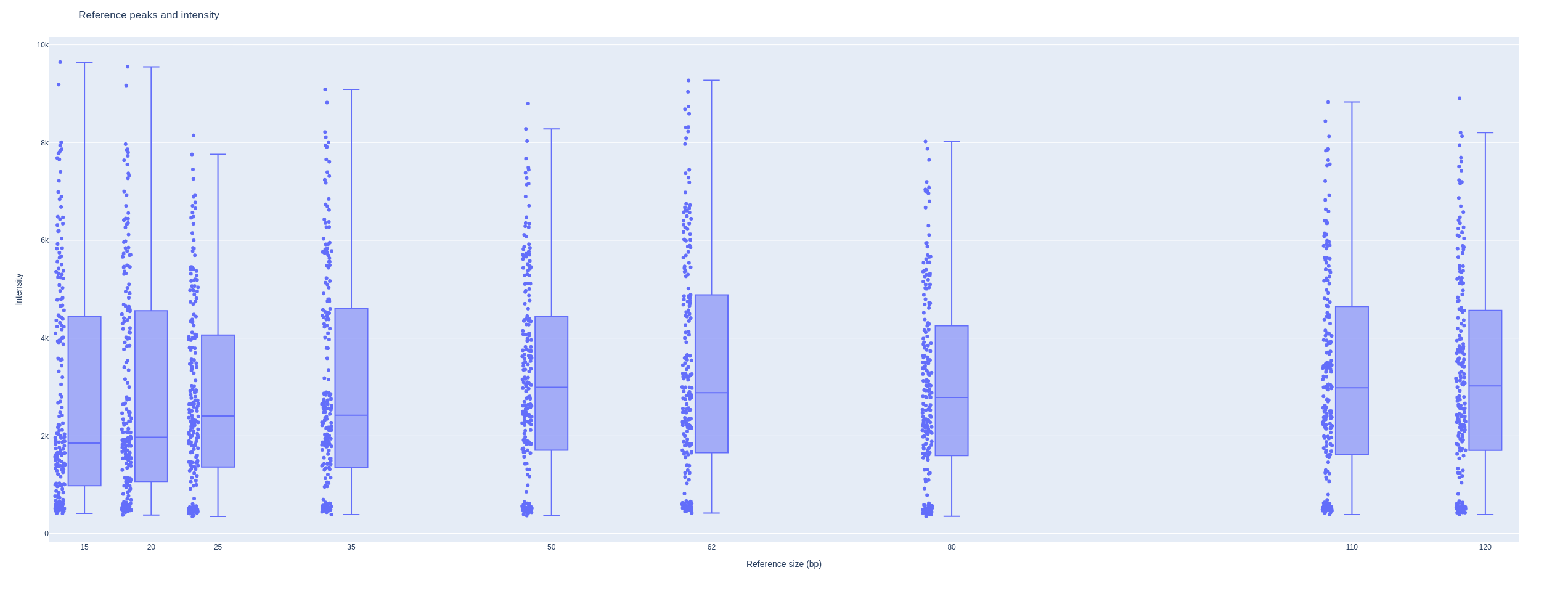
With minimum 300 intensity,

A function to build Least square model supporting both 2nd order or 3rd order options was developed. In the development process, we found that 3rd resulted in a slightly higher R square score. Therefore, we used 3rd order as defaulted in this study.

The sizing model was built for each FSA file from 202 FSA files of 101 samples. As a result, R square is higher than 0.999863 ± 0.000011 (Figure)

**Peaks detection result**

****



**Table 1. Information of reference peak’s detection**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Reference sizes (bp)** | **Data points (cycles)** | | | **Intensity** | | |
| **Median** | **Min** | **Max** | **Median** | **Min** | **Max** |
| **15** | 756 | 619 | 973 | 1854 | 416 | 9643 |
| **20** | 806 | 667 | 1025 | 1973 | 383 | 9549 |
| **25** | 846 | 703 | 1067 | 2409 | 353 | 8146 |
| **35** | 934 | 788 | 1157 | 2424 | 392 | 9088 |
| **50** | 1077 | 925 | 1307 | 2992 | 372 | 8796 |
| **62** | 1215 | 1056 | 1451 | 2882 | 423 | 9270 |
| **80** | 1395 | 1228 | 1636 | 2786 | 358 | 8023 |
| **110** | 1732 | 1549 | 1983 | 2984 | 390 | 8828 |
| **120** | 1841 | 1654 | 2094 | 3022 | 393 | 8906 |

**Validation results**

**Table 2: Status of calling samples**

|  |  |  |
| --- | --- | --- |
| **Index** | **Value** | **Note** |
| Total sample | 101 |  |
| No. called sample | 83 |  |
| No. sample no result | 3 | Ask PGx team, need to be confirmed |
| No. sample corrected call | 80 | After adjust intensity and bin range; Among those 2 samples could not be detected by both manual and automatically |
| No. sample not call | 18 | Samples have very low intensity or noise or lack some variant testing signal; or need to validated the result of the calling. |

**Accuracy analysis**

The performance evaluation of the algorithm were compare the genotype calling results from from algorithms and manual call from PGx team. In this analysis we excluded 3 samples without manual call from PGx team and 2 sample could not be called by PGx team. The total samples we used in this analysis is 96 samples.

* **True Positives (T.P.)**: All those genotype that are matched from both algorithm and manual call set.
* **False Positives (F.P.)**: All those genotypes that are in algorithm call set but no in manual call set. This value is 0 in this case.
* **False Negatives (F.N.)**: All those variants that are not matched between algorithm and manual call set.

The values are computed the following three metrics

* Recall = TP/(TP + FN)
* Precision = TP/(TP + FP)
* F-measure = 2 \* Precision \* Recall / (Precision + Recall)

|  |  |  |  |
| --- | --- | --- | --- |
| Indexes | | **Manual called by PGx team** | |
| Positive | Negative |
| **Algorithm** | True | 78 | N/A |
| False | 0 | 18 |

**Recall = 78/96 = 0.8125**

**Precision = 78/78 = 1**

**F-measure = 2 \* 0.8125 \* 1 / (0.8125 + 1) = 0.897**

*It will have a bias in those metrics because:*

* *This algorithm was designed to support PGx user can review raw data and adjust the intensity and bin range of each variant in the panel before going to star allele matching process*
* *Some samples are quite very noise, low intensity or lacking test variant signal such as (\*5, \*XN, or \*49), therefore it required more efforts (such as confirmed by other experiments or based on experiences) for calling those samples. These data could not be seen by the algorithm and therefore, those are considered as the false negative (18 samples)*