Computer simulation of populations and data analysis to determine the effect of mutation parameters in the detection of a recent selection event

Universitat Autònoma de Barcelona

Adrià Mompart Palau - Bachelor's Degree in Genetics

Data Analysis - 2022/23

Facultat Biociències UAB

INTRODUCTION

In a **recent selection event (RSE)** :

- A mutation (pre-existing or not) becomes beneficial (selected) in a specific environment.
- The mutation rises in frequency.
- The mutation drags close variants in the genome with itself (hitchhiking), producing a **selective sweep (Fig. 1)**:
  - The haplotype that contains the mutation remains long and becomes more frequent than expected.
  - Genetic diversity is reduced.
- This signal produces **linkage disequilibrium,** but it diminishes with time (generations). Thus, mainly recent events can be detected. In humans, this can be seen in lactose tolerance (lactase non-persistance) and malaria resistance mutations.

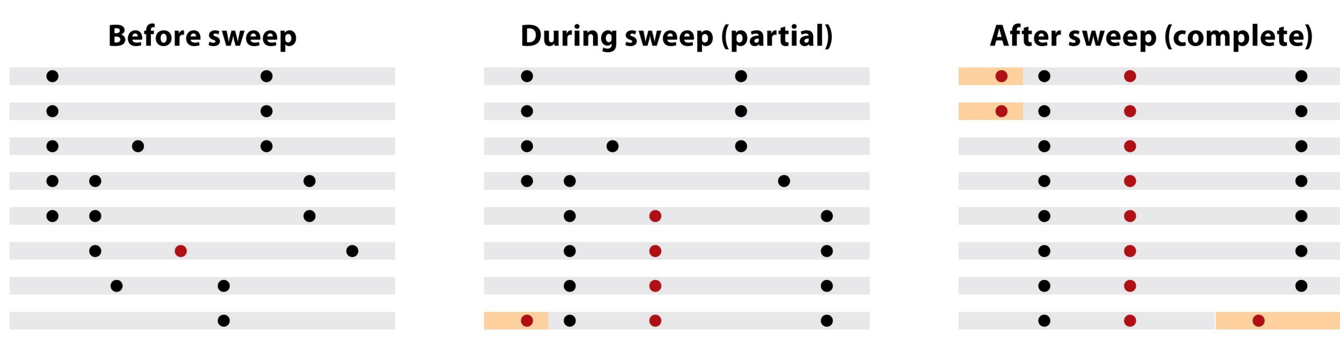


Fig. 1. General process of a selective sweep. Variants linked with the selected one are dragged and also rise in frequency [1]. In red, the selected mutation; in black, preexisting mutations that get dragged along; in orange, recombination events that alter haplotype length.

Integrated Haplotype Score (iHS)

-> **Statistical tool** to detect **recent selection events** (selective sweeps).
-> Based on **linkage disequilibrium** between genetic variants (SNPs).
-> Gives scores to every SNP: **Extreme scores** suggest a recent selection event.
-> Tends to be standardized to account for biases (**std-iHS**).

HYPOTHESIS

There must be some **parameters** that **affect** the behaviour of a **selective sweep** and the **linkage disequilibrium** of the selected variant and its adjacent ones, which can affect the **detection** of a **recent selection event**.

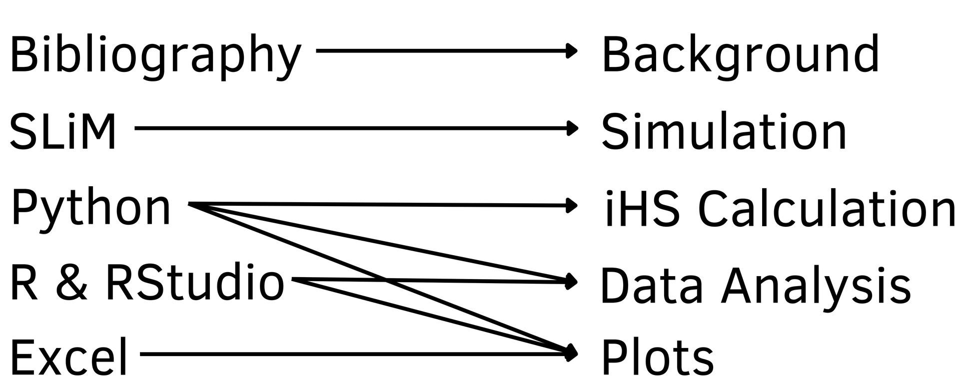
These parameters are likely to be the added fitness of the selected mutation (**selection coefficient**) and the frequency of the selected mutation at the point of the calculation of iHS (**selected mutation frequency**).

OBJECTIVES

- Understand the fundamentals of recent selection events and selection statistics.
- **Simulate** a **population** with a genome in which a recent selection event takes place.
- **Calculate iHS** and analyse the data to **detect the selection event**.
- Determine the effect of (i) the **selected mutation frequency** and (ii) **selection coefficient** in the calculation of iHS and the detection of a recent selection event.
- To design an **experiment** in which these effect of the **parameters** can be **compared**.

METHODOLOGY

The methodology of the project is a **Pipeline** that **(Fig. 2)** uses the following:



The **RSE** is simulated using **SLiM**, a **forward genetic simulator** used to simulate the genome of a population and let it evolve with time.

Each **situation** (a specific value for the parameters of study) and a standard to compare are **simulated 10 times (Table 1).**

Population data is extracted from a **sample** of the **simulation**, iHS is calculated and the **K Value** of the situation is obtained.

The **K Value** is used to measure the effect of the parameters of study (higher K value = better RSE detection) and compare situations. It is defined as: proportion of high values (|std-iHS > 2|) in a window of ~100kb comprising the most extreme values of the simulation.

Table 1. Standard value and range of simulated values of the parameters of study.

Variable	Standard Value	Iteration Range
selection coefficient	0.01	0.001-0.018
selected mutation frequency	0.7	0.3 - 0.9

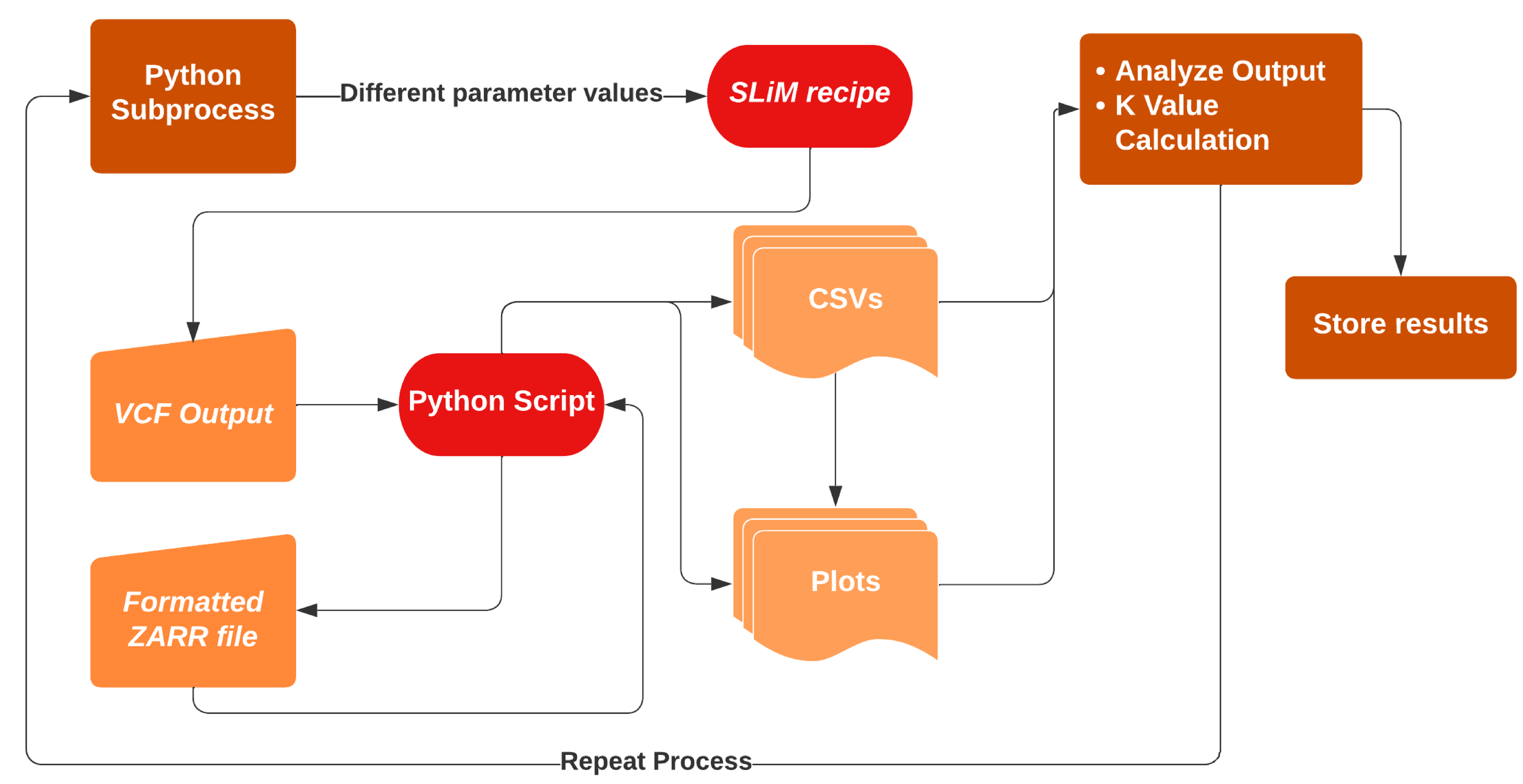


Fig. 2. Summarized process of obtaining and analyzing the data used in this project. VCF, ZARR and CSV are different file formats.

RESULTS

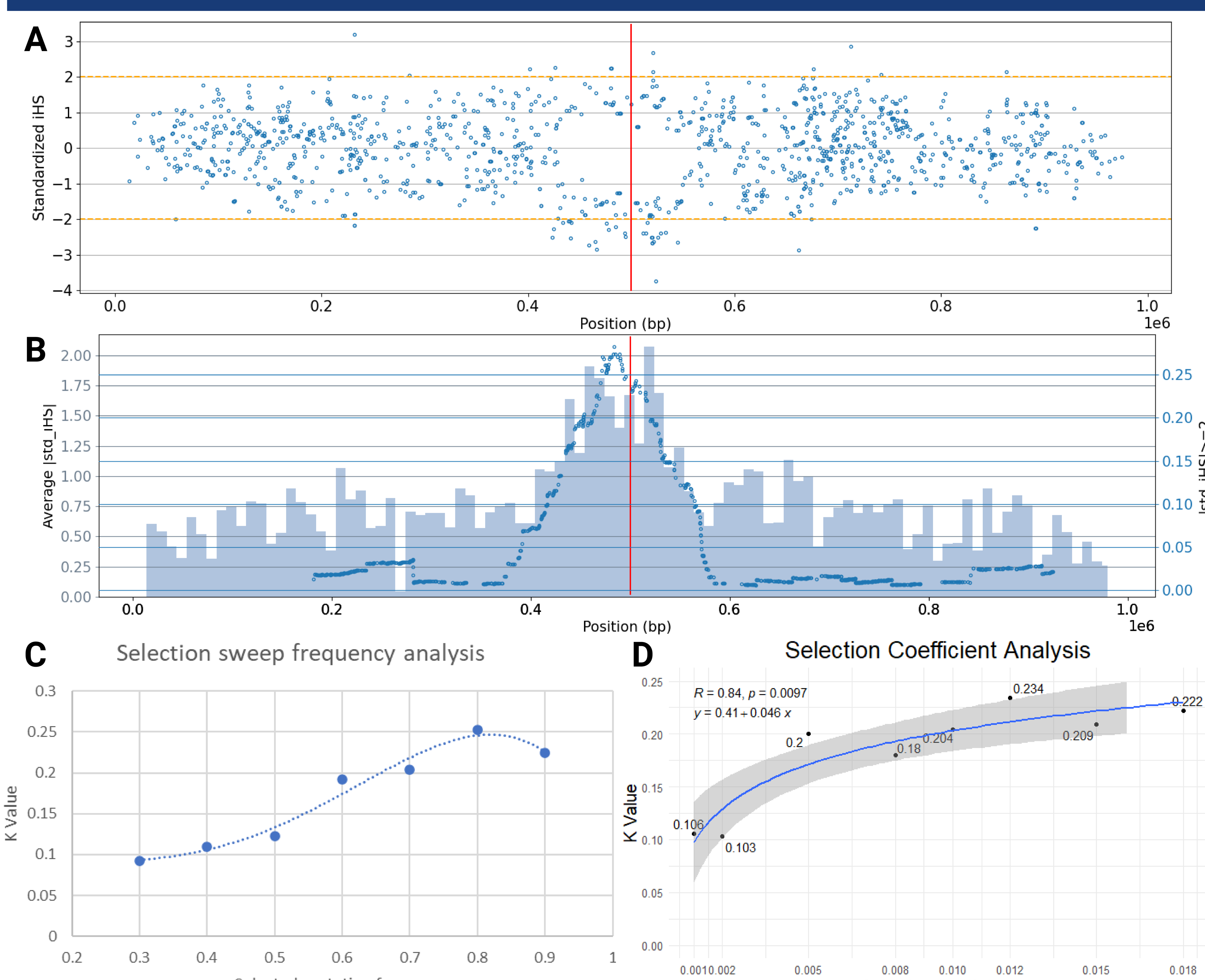


Fig. 3. Top (A), representation of the std-iHS of all SNPs algong the genome. Each dot: std-iHS for a different SNP. Red vertical line: Selected mutation site. Orange dotted lines: Threshold for high values. **Center (B)**, representation of the average |std-iHS| value in 10kb windows (bar blot) and proportion of high values in overlapping 100kb windows (scatter plot). **Bottom left (C)**, average K Value against the selected mutation frequency, with the trend curve. **Bottom right (D)**, average K Value against the selection coefficient, with the regression curve.

A clear lack of low values and and a higher number of **extreme values** can be seen at the **site of the mutation**, in which the RSE has happened **(Fig. 3A)**.

The signals of selection can be appreciated by the **peak** in both **average |std-iHS|** and **proportion of high values**, but the contrast is greater in the latter **(Fig. 3B)**.

The **detection** of the selective sweep is **influenced** by both the **selected mutation frequency** and the **selection coefficient (Fig. 3C and 3D)**, but their **dinamics differ**.

For the first, the best values for RSE detection are between 0.7 and 0.9 in terms of frequency, with a decay K value for both directions.

For the second, the higher the values, the better the sweep detection, following a **logarithmic function** (y = ln(x)), with a **really significant** p-value.

In both cases, values in the lower spectrum give the lowest K Values.

CONCLUSIONS

- **SLiM** and **Python** are useful tools to obtain and **analyse genomic data**.
- The **simulation** was **successful**, allowing to see the effect of a selective sweep using iHS selection statistic.
- Both **selected mutation frequency** and **selection coefficient** show an **influence** in the calculation of the K Value, and thus the recent selection event **detection**.
- For the **selected mutation frequency**, values between **0.7-0.9** gave the best results. For the **coefficient of selection, higher values** gave higher results, following a logarithmic curve.
- **Low values** show **low power** for detecting RSE.
- Results are **only aplicable** to the **simulation**.
- The simulation is **too simple** for the results to be **extrapolated** to real situations.

REFERENCES

[1] Wenqing Fu and Joshua M. Akey. Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 14:467–489, 9 2013. ISSN 15278204. doi: 10.1146/ANNUREV-GENOM-091212-153509.