PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
SAAD DAHLAB UNIVERSITY Blida1 (USDB1)

Faculty of Natural and Life Sciences

Department of Biology

Final Project Graduation thesis

For the purpose of obtaining a Master's degree in the field of SNV

Sector: Biological Sciences

Specialty: Biochemistry

Tittle

Interest in Genomic study of Diabetes with Periodontitis: In silico study

Presented by:

- **Motjolopane Relebohile**

Graduated on :01/07/2025

Before the jury:

Mrs CHALAL	***MCA***	***USDB1***	***President***
Mrs TALEB	***MCA***	***USDB1***	***Examiner***
Mrs EDDAIKRA	***MCA***	***USDB1***	***Supervisor***

Academic Year : 2024/2025

Acknowledgements

Completing this research has been challenging and rewarding, and it would not have been possible without the support of many remarkable individuals.

I am deeply grateful to my supervisor, Dr Eddaikra Atika whose steady guidance, constructive feedback, and belief in my potential helped shape this research and kept me focused when the path ahead felt unclear.

I would also like to express my sincere gratitude to the members of the jury; **Mrs CHALAL**, President; **Mrs TALEB**, Examiner; and **Mrs EDDAIKRA**, Supervisor.

My sincere thanks go to all the teachers who have been part of my academic life since my first year. Each of you, in your way, contributed to my learning and growth, and I carry your lessons with me not only in this work but in life beyond the university walls.

To my family, thank you for being my anchor. The distance between us was never easy, but your love and encouragement made it bearable and gave me the strength to persevere.

Living and studying in Algeria was a challenge at first, but it quickly became a chapter I will always cherish, thanks to the incredible friends I met. To you, Sarafina, I wouldn't have made it without your constant concerns, your kindness, laughter, and support made this foreign place feel like home, and your presence made my academic journey so much smoother.

Finally, to everyone, whether mentioned by name or not, who helped me, inspired me, or stood by me through this journey: thank you from the bottom of my heart.

Dedication

This work is lovingly dedicated to my mother, whose faith in the power of education never wavered, even though she never had the chance to pursue it fully. Her strength, sacrifices, and constant words of encouragement, especially during the times I was away in Algeria, gave me the courage to keep going. Her daily check-ins reminded me that no matter how far I was from home, I was never truly alone.

I also dedicate this to my brother RABELE MOTJOLOPANE, who has been my quiet source of strength. Whenever I was unwell, overwhelmed, or in need of help, financial or emotional, he stood by me without hesitation or complaint. His actions have always reflected what it truly means to be family, and his support has carried me through more than he knows.

Abstract

Periodontitis is a complex inflammatory disease with well-established links to systemic conditions such as type 2 diabetes (T2D). The association of diabetes with periodontitis (PDDM) can cause very severe clinical complications. Our "*in silico"* study aims to identify the single nucleotide genetic polymorphism (SNPs) by comparing two groups of samples: patients with periodontitis (PD) and diabetic patients with periodontitis (PDDM), focusing on three genes: **MCUR1**, **FOS** and **RAP2A** in order to identify variants that may contribute to the pathogenesis and interactions of the disease.

Variant identification was performed by the FreeBayes tool, followed by filtering and annotation using the SnpEff and SNPnexus tools. Group-specific SNPs were analyzed for functional impact using Combined Annotation Dependent Depletion (CADD) and FunSeq2 scores and then visualized via the UCSC Genome Browser. An enrichment analysis was performed with g:Profiler to explore the biological processes associated with each group.

Our results revealed a significant number of non-coding regulatory variants with very high pathogenicity scores in PDDM patients. Notable results include the snip references: **rs3810196** in *MCUR1* (CADD: 15.68), **rs1046117** in *FOS* (3′ UTR), and **rs6115228** in *RAP2A* (CADD: 19.22), all exclusive to the PDDM group. Enrichment analysis showed that these PDDM-specific genes were associated with signal transduction, immune regulation, and metabolic stress, while PD-specific genes were linked to tissue remodeling processes and apoptosis. A transition/transversion ratio (Ts/Tv) of 2.22 confirmed the quality of the data.

Our study was able to highlight distinct SNP profiles between PD and PDDM, suggesting that diabetes modifies the genetic landscape of periodontitis through increased regulatory variation. The variants identified in *MCUR1*, *FOS* and *RAP2A* contribute to a better understanding of the mechanisms involved in mitochondrial dysfunction., inflammatory signaling and immune dysregulation, and provide a basis for the future development of biomarkers and targeted therapeutic strategies.

Keywords: Periodontitis; Diabetes; SNPs; MCUR1; FOS; RAP2A

Résumé

La parodontite est une maladie inflammatoire complexe, dont les liens avec des pathologies systémiques telles que le diabète de type 2 (DT2) sont bien établis. L'association du diabète à la parodontite (PDDM) peut être à l'origine de complications cliniques très sévères. Notre étude « *in silico* » vise à identifier le polymorphisme génétique mononucléotidiques (SNPs) en comparant deux groupes d'échantillons : patients avec parodontite (PD) et de patients diabétiques avec parodontite (PDDM), en se concentrant sur trois gènes : **MCUR1**, **FOS** et **RAP2A** afin d'identifier des variants susceptibles de contribuer à la pathogenèse et aux interactions de la maladie.

L'identification des variants a été réalisé par l'outil FreeBayes, suivi d'un filtrage et d'une annotation à l'aide des outils SnpEff et SNPnexus. Les SNPs spécifiques à chaque groupe ont été analysés pour leur impact fonctionnel à l'aide des scores Combined Annotation Dependent Depletion (CADD) et FunSeq2, puis visualisés via le UCSC Genome Browser. Une analyse d'enrichissement a été effectuée avec g: Profiler pour explorer les processus biologiques associés à chaque groupe.

Nos résultats ont révélé un nombre important de variants régulateurs non codants avec des scores de pathogénicité trés élevés chez le patient PDDM. Parmi les résultats notables, on retrouve les références snip : **rs3810196** dans *MCUR1* (CADD : 15,68), **rs1046117** dans *FOS* (3′ UTR), et **rs6115228** dans *RAP2A* (CADD : 19,22), tous exclusifs au groupe PDDM. L'analyse d'enrichissement a montré que ces gènes spécifiques au PDDM étaient associés à la transduction du signal, à la régulation immunitaire et au stress métabolique, tandis que les gènes propres au PD étaient liés aux processus de remodelage tissulaire et à l'apoptose. Un ratio de transition/transversion (Ts/Tv) de 2,22 a confirmé la qualité des données.

Notre étude a pu mettre en évidence des profils de SNPs distincts entre PD et PDDM, suggérant que le diabète modifie le paysage génétique de la parodontite par une variation régulatrice accrue. Les variants identifiés dans *MCUR1*, *FOS* et *RAP2A* contribuent à une meilleure compréhension des mécanismes impliqués dans le dysfonctionnement mitochondrial., la signalisation inflammatoire et la dérégulation immunitaire, et constituent une base pour le développement futur de biomarqueurs et de stratégies thérapeutiques ciblées.

Mots clés : Parodontite ; Diabète ; SNPs ; MCUR1 ; FOS ; RAP2A

Abstract (Arabic)

التهاب دواعم السن هو مرض التهابي معقد له روابط راسخة بالحالات الجهازية مثل مرض السكري من النوع 2 (T2D). يمكن أن يسبب ارتباط مرض السكري بالتهاب دواعم السن (PDDM) مضاعفات سريرية خطيرة للغاية. تهدف دراستنا *"في السيليكو"* إلى تحديد تعدد الأشكال الجيني للنوكليوتيدات المفردة (SNPs) من خلال مقارنة مجموعتين من العينات: المرضى الذين يعانون من التهاب دواعم السن (PD) ومرضى السكري الذين يعانون من التهاب دواعم السن (PDDM) ، مع التركيز على ثلاثة جينات: **MCUR1** و **FOS** و **RAP2A** من أجل تحديد المتغيرات التي قد تساهم في التسبب في المرض وتفاعلاته.

تم إجراء تحديد المتغير بواسطة أداة FreeBayes ، متبوعا بالتصفية والتعليق التوضيحي باستخدام أدوات SnpEff و SNPnexus. تم تحليل تعدد الأشكال الخاصة بالمجموعة من أجل التأثير الوظيفي باستخدام درجات الاستنفاد المعتمد على التعليقات التوضيحية (CADD) و FunSeq2 ثم تم تصورها عبر متصفح الجينوم UCSC. تم إجراء تحليل تخصيب باستخدام g: Profiler لاستكشاف العمليات البيولوجية المرتبطة بكل مجموعة.

كشفت نتائجنا عن عدد كبير من المتغيرات التنظيمية غير المشفرة ذات درجات عالية جدا من الإمراض في مرضى PDDM. تشمل النتائج البارزة مراجع القصاصة: **rs3810196** في *MCUR1* (CADD: 15.68) ، **rs1046117** في *FOS* (3 ′ UTR) ، و **rs6115228** في *RAP2A* (CADD: 19.22) ، وكلها حصرية لمجموعة PDDM. أظهر تحليل التخصيب أن هذه الجينات الخاصة ب PDDM كانت مرتبطة بنقل الإشارات ، وتنظيم المناعة ، والإجهاد الأيضي ، بينما ارتبطت الجينات الخاصة بمرض باركنسون بعمليات إعادة تشكيل الأنسجة وموت الخلايا المبرمج. وأكدت نسبة الانتقال/التحويل (Ts/Tv) البالغة 2.22 جودة البيانات.

تمكنت دراستنا من تسليط الضوء على ملامح SNP المميزة بين PD و PDDM ، مما يشير إلى أن مرض السكري يعدل المشهد الجيني لالتهاب دواعم السن من خلال زيادة التباين التنظيمي. تساهم المتغيرات المحددة في *MCUR1* و *FOS* و *RAP2A* في فهم أفضل للآليات التي ينطوي عليها الخلل الوظيفي في الميتوكوندريا.، والإشارات الالتهابية وخلل التنظيم المناعي ، وتوفير أساس للتطوير المستقبلي للمؤشرات الحيوية والاستراتيجيات العلاجية المستهدفة.

الكلمات الدالة: التهاب دواعم السن. داء السكري; تعدد الأشكال MCUR1; FOS. راب 2 أ

List of Abbreviation

Abbreviation	Definition
PD	Periodontitis
PDDM	Periodontitis and Diabetes Mellitus
SNPs	Single Nucleotide Polymorphisms
MCUR1	Mitochondrial Calcium Uniporter Regulator 1
FOS	FBJ Murine Osteosarcoma Viral Oncogene Homolog
RAP2A	Ras-related Protein Rap-2a
SnpEff	Genetic Variant Annotation and Effect Prediction Tool
CADD	Combined Annotation Dependent Depletion
UTR	Untranslated Region
CD4+ / CD8+	Cluster of Differentiation 4 and 8 (T-cell markers)
HLA	Human Leukocyte Antigen
TNF-α	Tumor Necrosis Factor Alpha
IL-1B	Interleukin 1 Beta
PGE2	Prostaglandin E2
RAGE	Receptor for Advanced Glycation End-products
AGEs	Advanced Glycation End-products
NF-κB	Nuclear Factor Kappa-light-chain-enhancer of Activated B Cells
CRP	C-Reactive Protein
GEO	Gene Expression Omnibus
DEGs	Differentially Expressed Genes
CD34	Hematopoietic Stem Cell Marker
EGR1	Early Growth Response 1
FMOD	Fibromodulin
IGF2	Insulin-like Growth Factor 2
NCF1	Neutrophil Cytosolic Factor 1
LRRC25	Leucine-rich Repeat Containing 25

ROS	Reactive Oxygen Species
M1 / M2	Macrophage Type 1 and Type 2
TH1 / TH17	T-helper Cell Subtypes 1 and 17
RUNX2	Runt-related Transcription Factor 2
GWAS	Genome-Wide Association Study
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
FDR	False Discovery Rate
PCA	Principal Component Analysis
QBWA-MEM	Query-Based Whole Alignment – Maximal Exact Match
PCR	Polymerase Chain Reaction
DNA	Deoxyribonucleic Acid
WGS	Whole Genome Sequencing
SRA	Sequence Read Archive
SRR	Sequence Read Run (accession code)
NCBI	National Center for Biotechnology Information
Galaxy NG	Galaxy's Next Generation Sequencing Platform
BAM	Binary Alignment Map
hg38	Human Genome Version 38
UCSC	University of California Santa Cruz Genome Browser
FunSeq2	Functional Sequencing Variant Prioritizer Tool
ANN	Annotation Field (SnpEff)
QUAL	Variant Quality Score
DP	Depth of Coverage
rsIDs	Reference SNP cluster IDs
OMIM	Online Mendelian Inheritance in Man
ClinVar	Clinical Variant Database
PolyPhen	Polymorphism Phenotyping
ENCODE	Encyclopedia of DNA Elements

TSV	Tab-Separated Values
Gbp	Gigabase Pairs
MAPQ	Mapping Quality
IGV	Integrative Genomics Viewer
K562	Human Leukemia Cell Line
IMR90	Human Lung Fibroblast Cell Line
CDS	Coding DNA Sequence
MF / BP / CC	Molecular Function / Biological Process / Cellular Component
AP-1	Activator Protein 1
eQTL	Expression Quantitative Trait Loci
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats / CRISPR-associated Protein 9

List of Figures

Summary

INTRODUCTION

INTRODUCTION

Periodontitis (**PD**) is a chronic inflammatory disease of the supporting structures of the teeth, characterized by progressive attachment loss, alveolar bone resorption, and the formation of periodontal pockets. Beyond its local oral manifestations, periodontitis is now widely recognized as a disease with systemic implications. One of the most extensively studied associations is its bidirectional relationship with type 2 diabetes mellitus (**T2DM**). Individuals with poorly controlled diabetes are at higher risk for more severe periodontal destruction, while the presence of periodontitis may negatively impact glycemic control and increase systemic inflammatory burden **(Lalla & Papapanou, 2011).**

Despite significant advances in understanding the microbial and immunological aspects of periodontitis, the genetic contribution to its pathogenesis, especially in the context of comorbid diabetes remains underexplored. Recent evidence suggests that single nucleotide polymorphisms (SNPs) may modulate host susceptibility, inflammatory response, and tissue repair mechanisms, potentially influencing disease onset, progression, and severity. However, few studies have addressed whether the genetic landscape of periodontitis differs when diabetes is present as a modifying factor.

In this study, we apply a genomic variant analysis approach to compare SNP profiles between individuals with periodontitis (PD) and those with periodontitis complicated by diabetes (PDDM). We focus on three biologically relevant genes, **MCUR1, FOS, and RAP2A**; each implicated in mitochondrial regulation, transcriptional control of inflammation, and immune cell migration, respectively. Using variant calling, functional annotation (SnpEff, SNPnexus), and pathway enrichment tools (g:Profiler), we identify and interpret group-specific genetic differences with potential mechanistic relevance.

This work aims to characterize the genomic distinctions between PD and PDDM, with the broader goal of uncovering SNPs that may serve as biomarkers of disease severity or therapeutic targets. By integrating bioinformatic evidence with known biological functions, we contribute to a more nuanced understanding of the genetic architecture underlying periodontal disease in the diabetic context, hence improving the diagnosis and treatment of the victims or patients struggling with these interrelated disorders, the insights gained from this study may help to precisely identify new therapeutic targets and develop accurate medicines.

LITERATURE REVIEW

Chapter 1: Literature Review

1.1. Generalities on Diabetes Mellitus and Periodontitis

1.1.1. Diabetes Mellitus (DM)

Recent studies define Diabetes Mellitus as a chronic metabolic disorder characterised by inefficient insulin secretion or insulin action **(Lalla & Papapanou, 2011; Ranbhise et al., 2025; Yameny, 2024)**. This leads to complications such as organ damage and failure, particularly affecting the retina, kidneys, nerves, heart, and blood vessels **(Alam et al., 2014)**.

It can also be defined as the inability or failure of the body to effectively utilize glucose for energy, while simultaneously producing high volumes or excessive glucose due to disrupted control of gluconeogenesis and glycogen breakdown, resulting in persistent high blood sugar levels.

DM is a rapidly growing disease, and from 1995 to 2025, there was an estimated predominant rise in diabetic cases geographically, with 42% in developed countries, from 51 million to 72 million, and a rise of 170% in developing countries, from 84 million to 228 million (King et al., 1998). However, with the recent studies, the prevalence of DM has already surpassed those projections, pushing the number of cases to 828 million within the range of 1990-2022 **Figure 1 (Zhou et al., 2024).**

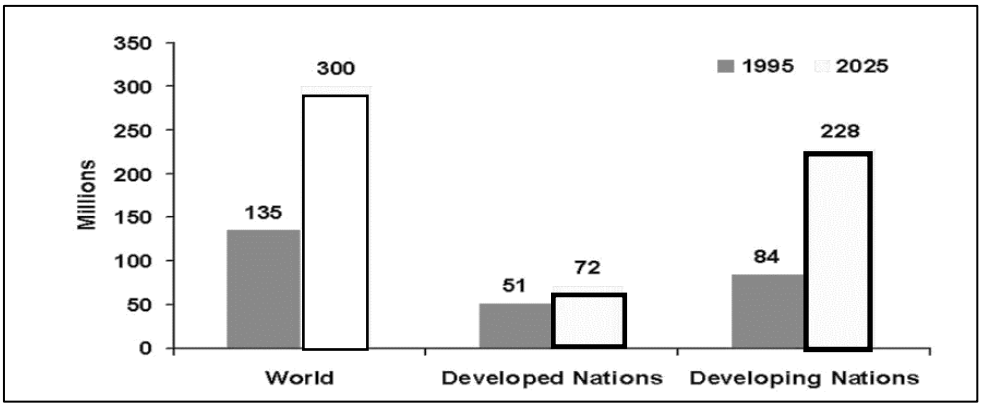


Figure 1: *Predicted cases of Diabetes Mellitus (King et al., 1998).*

1.1.1.1. Pathophysiology of diabetes mellitus (DM)

Pathophysiology of DM can be explored following its two main distinct categories or types, namely: Type 1 Diabetes Mellitus (T1DM) and Type 2 Diabetes Mellitus (T2DM).

1.1.1.1.1. Type 1 Diabetes Mellitus (T1DM)

T1DM is characterised by autoimmunity, or attack of the pancreas, implying the invasion of the islets by CD4+ and CD8+ T cells, along with macrophages; hence, the destruction of the insulin-producing beta cells (Insulin deficiency), **Figure 2 (Todd, 2010).**

According to (Mobasseri et al., 2020). Around 5%-10% of the population develops Type 1 diabetes mellitus, following the study conducted in Asia and Europe. T1DM is influenced by factors, including genetic influence, highlighting the role of the human leukocyte antigen (HLA) gene. The cell relies on HLA proteins present on its surface to distinguish the body's normal cells from foreign agents. Defects in these HLA proteins can trigger autoimmunity against beta cells.

1.1.1.1.2. Type 2 Diabetes Mellitus (T2DM)

Type 2 diabetes is the most common form of diabetes, representing an estimated 90 to 95% of all its cases. This disorder is largely associated with insulin resistance of the body. In type 1, immune cells destroy pancreatic β-cells, leading to insufficient insulin and hyperglycemia. In type 2, hyperglycemia, hyperlipidemia, cytokines, and amyloids damage the cells, resulting in relative insulin deficiency and hyperglycemia **Figure 2 (Khin et al., 2023),** which means that the body's cells do not respond to insulin, and insulin is produced, but blood glucose levels continue to rise. Instinctively, the pancreas compensates over the period of time this demand may result in chronic fatigue of the pancreas 'burnout' and undermined production of insulin, which worsens hyperglycaemia.

It is most often related to increasing age, obesity, and lack of physical activity, as well as bad eating habits; therefore, it is predominant in the elderly population. But because of alarming trends of obesity and lack of exercise, it is now being increasingly seen in children as well as adolescents and young adults **(Mahesh et al., 2022; Yameny, 2024).**

Evidence shows that while there is insulin deficiency, the predominant underlying defect is insulin resistance **(Alam et al., 2014).**

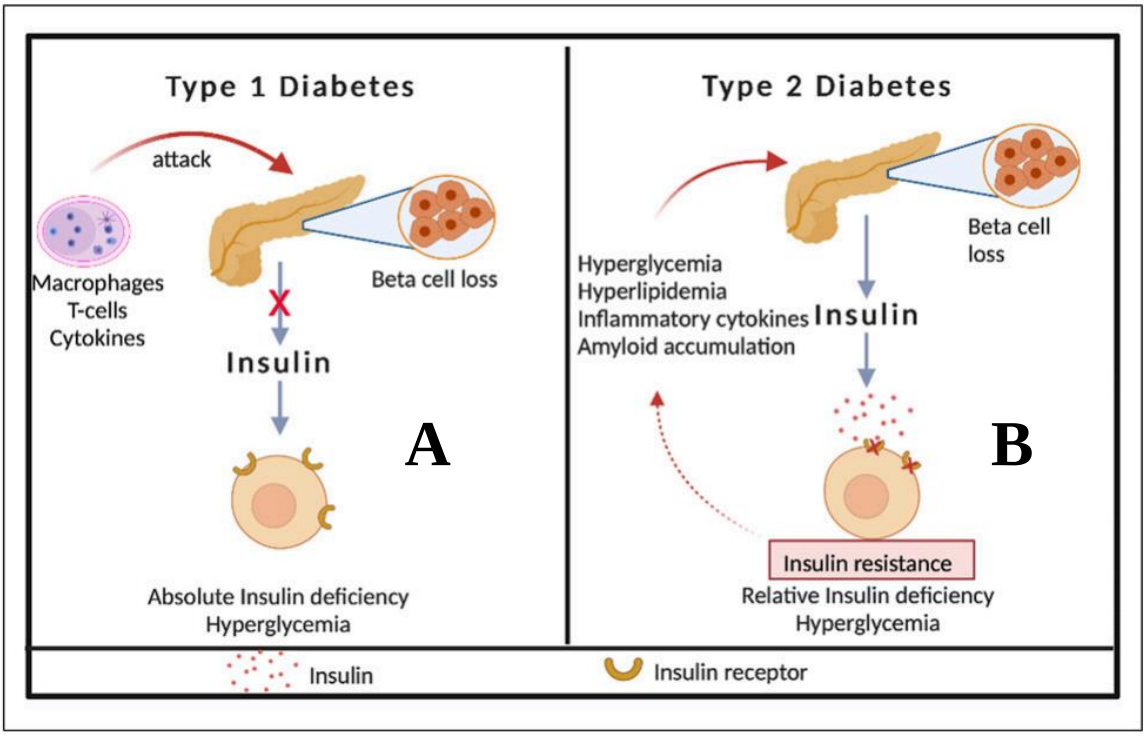


***Figure 2**: The development of type 1 and type 2 diabetes. (Khin et al., 2023).*

1.1.2. Periodontitis (PD)

Periodontitis is a long-term inflammatory condition that targets the structures supporting the teeth, such as the periodontal ligament and alveolar bone. It often develops from untreated gingivitis, which is a less severe condition and, if left unmanaged, can result in attachment loss, gum pocket formation, and eventually tooth loss.

This condition results from a combination of bacterial imbalance, mainly Gram-negative anaerobic bacteria such as Porphyromonas gingivalis, Tannerella forsythia, and Treponema denticola ("red complex"), and the body's inflammatory response. These pathogens trigger the body to release inflammatory molecules such as Tumor Necrosis Factor-alpha (**TNF-α**), Interleukin-1 beta (**IL-1β**), and Prostaglandin E_2 (**PGE2**), contributing to alveolar bone destruction and collagen fibre breakdown **(Lin et al., 2024).**

Periodontitis has also been identified as the sixth complication of diabetes mellitus, with research confirming that individuals with diabetes are at greater risk of developing severe forms of the disease **(Zhao et al., 2023).**

The study carried out by **(Barutta et al., 2022),** also noted the alarming **70%** greater chance a diabetic person would have their periodontitis worsen compared with a non-diabetic. Even after multiple factors were taken into account, the risk was estimated at 86% (RR 1.86 [95% CI 1.3-2.8]). The data affirms that diabetes and the progression of periodontal disease are intricately related, stressing the importance of proper glycaemic control **Figure 3**.

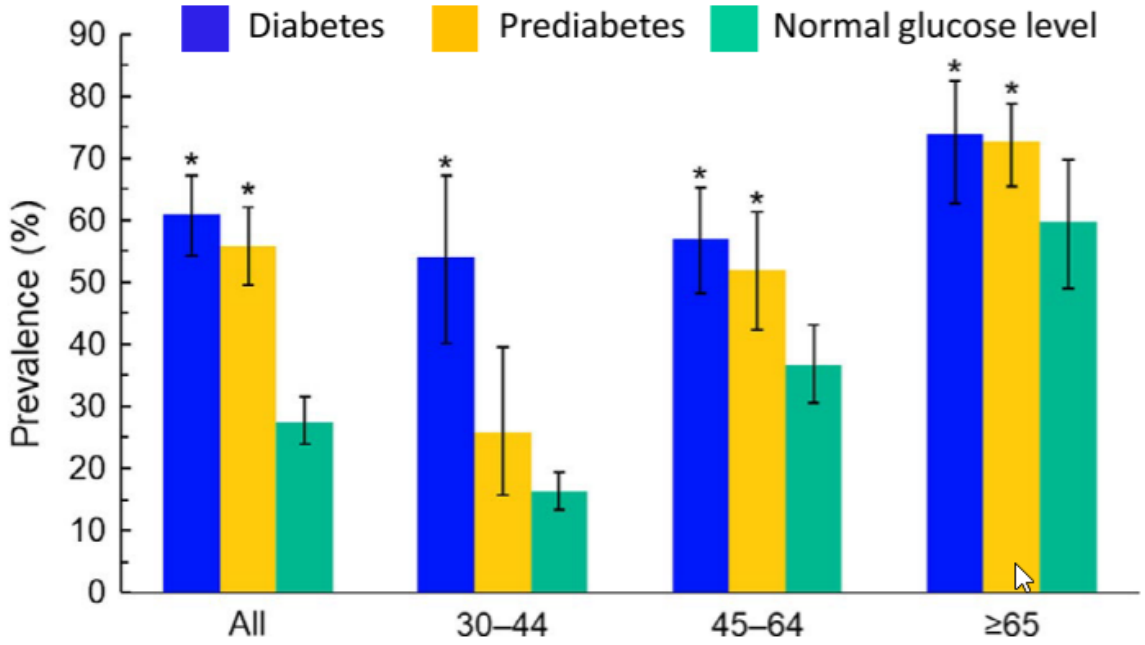


***Figure 3**: Prevalence of diabetes, prediabetes, and normal glucose levels by age group (Barutta et al., 2022).*

1.2. Bidirectional relationship between DM and PD

1.2.1. Impact of DM on PD

People with diabetes are more susceptible to the immune response to gum disease, leading to increased levels of inflammation and, thus, tissue destruction. The Receptor for Advanced Glycation End Products (RAGE) and its ligands contribute to this inflammatory response **(Figure 4**). Hyperglycaemia impairs collagen production, thereby compromising the structural integrity of the gums and inhibiting normal wound healing. In addition, it alters the bone formation, resorption balance, enhancing the degradation of bone and alveolar bone destruction, which is a characteristic of periodontitis.

Periodontal health is negatively impacted by several interconnected mechanisms that are triggered in the diabetic state, primarily through impaired immune function. High blood glucose concentrations disrupt immune cell performance, including neutrophils and

macrophages, which diminishes the host's capacity to fight bacterial pathogens in periodontal tissues. An increased likelihood emerges for enduring infections alongside periodontal tissue deterioration.

Individuals living with diabetes often present systemic conditions marked by elevated levels of pro-inflammatory mediators, which hasten the degradation of periodontal structures. The persistent inflammatory state contributes not only to tissue damage but also hinders the natural regenerative processes within the periodontium.

Chronic hyperglycaemia also promotes the formation of advanced glycation end-products (AGEs) that bind to receptors on immune and structural cells, intensifying inflammatory responses and altering the properties of the extracellular matrix. These collagen-related modifications result in stiffer connective tissue with diminished reparative capacity, ultimately impeding proper wound healing.

The persistent high blood sugar levels characteristic of chronic hyperglycaemia interferes with fibroblast function, along with the activity of essential regenerative cells. The reduced reparative ability causes periodontal tissues to heal wounds slowly while becoming more prone to chronic infections, which then accelerates periodontal disease progression.**(Lalla & Papapanou, 2011).**

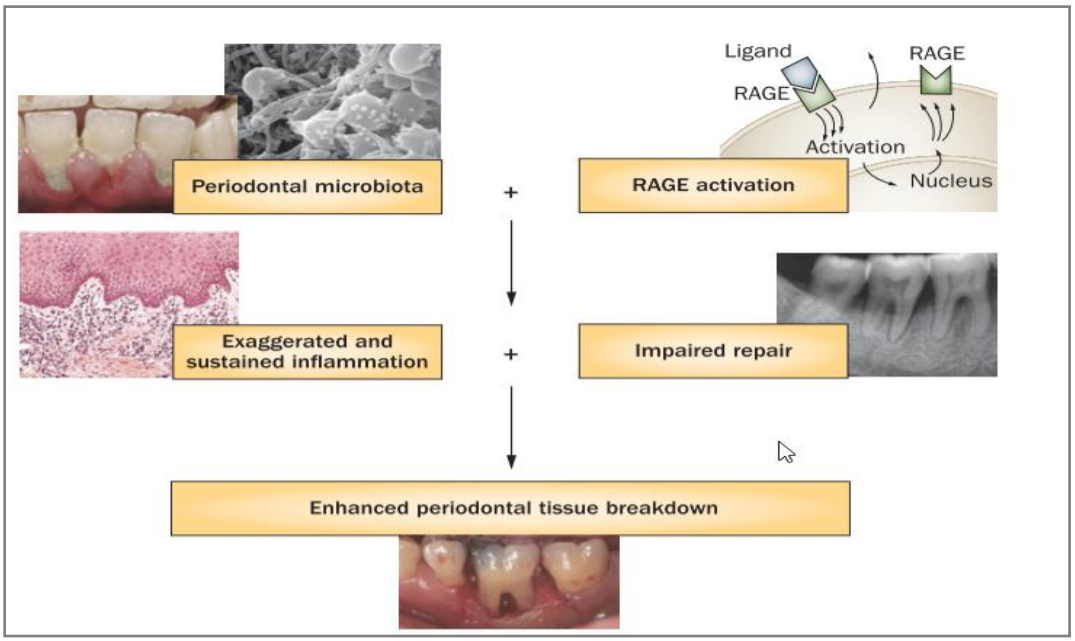


***Figure 4**: Schematic illustration of how RAGE activation, driven by periodontal microbiota and inflammation, contributes to impaired repair and enhanced periodontal tissue breakdown. Adapted from (Lalla & Papapanou, 2011).*

1.2.2. Impact of PD on DM

Through its systemic inflammatory effects, periodontal disease accelerates insulin resistance progression. The release of tumour necrosis factor-alpha (TNF-α) and interleukin-6 (IL-6) during periodontal infection disrupts insulin signalling pathways. The alteration of insulin sensitivity stands as a fundamental aspect within type 2 diabetes pathophysiology. Chronic periodontal inflammation triggers the nuclear factor kappa B (NF-κB) pathway, which then enhances cytokine production that disrupts glucose metabolism.

Periodontitis triggers oxidative stress, which, alongside inflammation, contributes to metabolic dysfunction. Infection-related reactive oxygen species production harms insulin receptors along with glucose uptake cellular components, which leads to increased insulin resistance.

The condition of microbial dysbiosis linked to periodontal disease serves as another contributing factor. Pathogenic entities like Porphyromonas gingivalis enter the bloodstream, where they initiate systemic immune reactions that elevate circulating markers such as C-reactive protein (CRP), which is associated with poor glycaemic management. Periodontal infections demonstrate the potential to produce systemic effects beyond oral boundaries, which may worsen metabolic disorders such as diabetes. **(Ranbhise et al., 2025).**

1.3. Molecular study with a genomic approach

The complex interplay between periodontitis (PD) and diabetes mellitus (DM) has been the subject of considerable interest, particularly since recent molecular and genomic studies shed light on the common biological processes under which these two diseases might share underlying mechanisms. Beyond clinical counterparts, newer in silico data support an understanding grounded on gene expression patterns, genetic structure, and immunologic pathways in common.

For instance, **(Liu et al., 2023)**, examined genetic crosstalk between type 1 diabetes (T1D) and periodontitis using comparisons of Gene Expression Omnibus (GEO) datasets. They found that 59 DEGs were common to both conditions. Of these, hub genes such as CD34, EGR1, FMOD, and IGF2 were recognized for their function in cellular signaling, lipid metabolism, and tissue development processes relevant to both immune response and tissue degeneration in PD and T1D.

Following this genomic analysis, more recent studies were conducted employing a Mendelian randomisation (**MR**) approach to define causality between Periodontitis disease (PD) and type 2 diabetes (T2D). The analysis revealed 79 shared DEGs enriched in inflammatory signalling pathways like interleukin-17 and chemokine receptor binding (**Figure 5**). Notably, genes RAP2A, MCUR1, and FOS were shown to be significant contributors to T2D risk in PD patients, which points towards an active predisposition to systemic glucose dysregulation by chronic oral inflammation (**Wei et al., 2024)**.

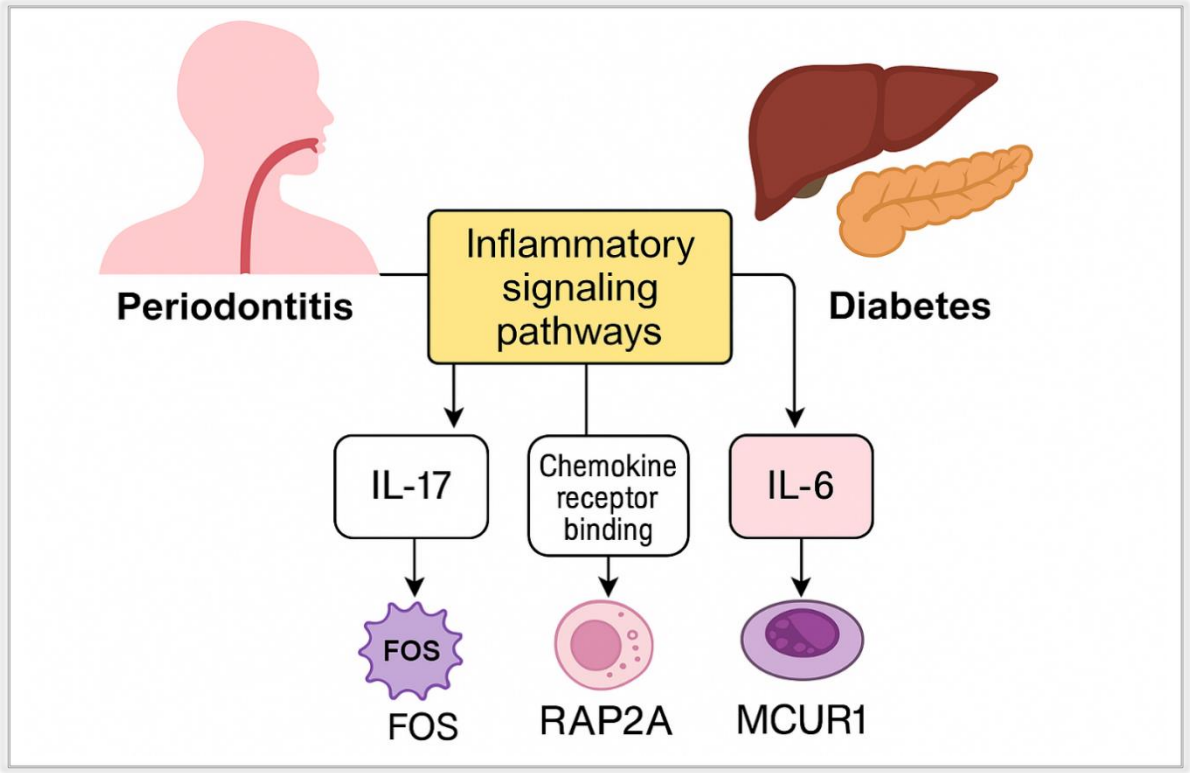


Figure 5: *Illustration of IL-17 signaling in the link between diabetes and periodontitis (Wei et al., 2024).*

Adding yet another layer, **(Kang et al., 2024)** also carried out a multi-omics analysis and discovered a widespread epigenetic change, namely hypomethylation of the pathway of Fc-gamma receptor-mediated phagocytosis (**Figure 6**). These were highly expressed in monocytes, implying a heightened and perhaps ongoing immune response in diabetic patients, which might exacerbate periodontal damage.

In the broader synthesis of clinical and experimental evidence, hyperglycaemia induces oxidative stress, immune cell profile alterations, and increased production of inflammatory cytokines such as TNF-α and IL-6 increase alveolar bone resorption. This work builds on the

idea that diabetes is not a comorbidity but an increasing factor in periodontal disease progression **(Liu et al., 2023).**

The study by **(Wu et al., 2024a)** , also complemented this notion by performing a genome-wide cross-trait analysis that identified 62 pleiotropic loci connecting PD with glycaemic traits like fasting glucose and insulin resistance. Interestingly, some of these genetic correlations were independent of BMI, indicating an unconfounded direct genetic overlap and not due to confounding by metabolic syndrome. This makes the hypothesis of an intrinsic genomic interdependence between Periodontitis and Diabetes Mellitus (PD and T2D) stronger.

Genomic variants influence lipid mediators such as **prostaglandins and leukotrienes**, which are involved in **metabolic pathways**. These lipid mediators amplify the inflammatory response in the co-existence of diabetes mellitus (DM) and periodontitis (PD), leading to tissue damage and systemic complications.

Genes involved in lipid metabolism and immune response regulate the biosynthesis and signaling of these mediators, adding a metabolic-genomic dimension to the interplay between Diabetes Mellitus and Periodontitis (PD and DM).

According to **(Zou & Yang, 2024)** Single-cell RNA sequencing combined with Mendelian randomisation and colocalization analysis was used to identify therapeutic targets at a more precise level. Their study found that classical and intermediate monocytes are critical cell types, and that the genes **NCF1** and **LRRC25** play significant roles in both periodontitis and type 2 diabetes. Recognising these targets not only fills an important knowledge gap but also opens the door to **precis ion therapies** focused on regulating monocyte-driven inflammation.

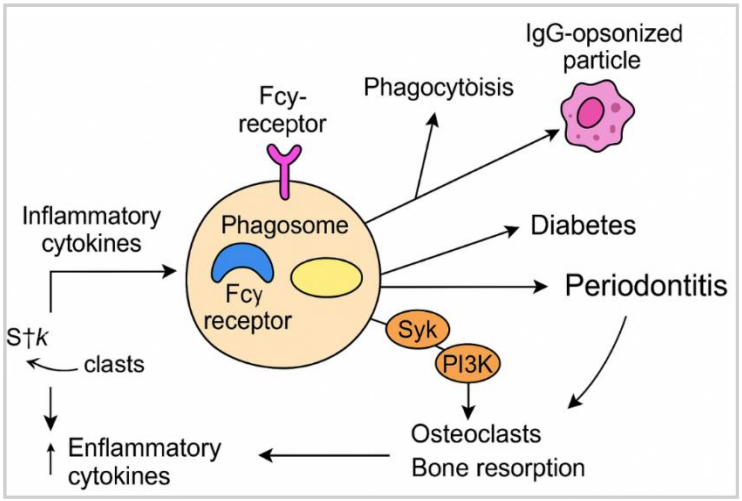


***Figure 6**: Fcγ Receptor-Mediated Phagocytosis and Inflammatory Signaling in Diabetes-Associated Periodontitis: The Syk-PI3K Axis and Osteoclast Activation (Kand et al,. 2024).*

1.4. Molecular insights from experimental and immunological studies

One of the most striking contributions of diabetes to periodontal pathology is the dysregulation of immune cell function. Neutrophils in diabetic individuals are overactivated, releasing excessive reactive oxygen species (**ROS**), which increases tissue damage.

However, their bactericidal function is paradoxically impaired, leading to insufficient resolution of microbial threats. Likewise, macrophages become polarized toward the M1 phenotype, amplifying pro-inflammatory signaling, while the M2 reparative subset is diminished. This polarization results in sustained cytokine production and chronic tissue destruction.

In the adaptive immune system, dendritic cells under diabetic conditions favor the expansion of Th1 and Th17 lymphocytes, both of which contribute to osteoclast activation and alveolar bone loss. Simultaneously, the population of regulatory T cells is reduced, weakening the immune system's ability to suppress destructive inflammation **(Graves et al., 2020)**.
From a cellular repair standpoint, diabetes also disrupts bone homeostasis and periodontal ligament (PDL) integrity through:

- Downregulation of key osteogenic transcription factors such as RUNX2 and C-fos.
- Apoptosis of osteoblasts, osteocytes, and PDL fibroblasts, mediated by increased caspase-3, -8, and -9 activity.
- Elevated **RANKL** expression and a higher RANKL/OPG ratio, fostering osteoclastogenesis and bone loss.
- The accumulation of **AGEs** in periodontal tissues, which bind to RAGE and trigger NF-κB activation, thereby increasing inflammatory mediator production and further impairing bone regeneration.

1.5. Bioinformatics Pipeline

Several studies have employed comprehensive bioinformatics pipelines to investigate the molecular mechanisms linking type 2 diabetes (T2D) and periodontitis (PD). These approaches typically integrate quality control, alignment, variant detection, annotation, and pathway enrichment tools to decode the genetic underpinnings of these interrelated conditions.

In cross-trait genome-wide association studies (GWAS), researchers have used tools such as **LDSC**, **GNOVA**, and **SUPERGNOVA** to estimate both global and local genetic correlations between PD and glycemic traits. These tools help identify shared genomic regions that may influence both diseases. To uncover pleiotropic SNPs driving such comorbidities, **CPASSOC** has been used for cross-phenotype meta-analyses.

For downstream functional interpretation, platforms such as **g: Profiler** have played a critical role in pathway enrichment analysis. By leveraging Gene Ontology (GO) terms and KEGG pathways, while applying False Discovery Rate (FDR) control, gProfiler has enabled researchers to detect significant biological pathways and gene networks.

To process gene expression data, particularly from single-cell RNA sequencing (scRNA-seq), tools like **Cell Ranger** are widely utilized for pre-processing, while **Seurat** is applied for data normalization, clustering, and dimensionality reduction techniques such as PCA and UMAP. Seurat's FindMarkers function is often used for identifying differentially expressed genes across cell populations **(Wu et al., 2024b)**.

For deeper enrichment analysis, the **clusterProfiler R package** has been extensively used in literature to identify significantly enriched GO terms and KEGG pathways via its enrichGO and enrichKEGG functions **(Silva et al., 2022)**. Complementing this, Enrichr, a powerful web-based platform. offers interactive visualizations and curated gene set libraries to explore the biological significance of gene clusters **(Guan et al., 2024)**.

In parallel, other studies have adopted toolsets to handle raw genomic data. The **Galaxy platform** (https://usegalaxy.org/) has been highlighted for its accessibility and reproducibility. As a web-based, open-source interface, Galaxy enables users to execute complex genomic workflows without programming expertise, and it integrates seamlessly with other platforms like Jupyter and RStudio (**Afgan et al., 2018; Jalili et al., 2021**).

For initial data quality assessment, **FastQC** is often used to evaluate metrics such as per-base sequence quality, GC content, and adapter contamination, providing researchers with visual diagnostics for raw sequence data (**Andrews, 2010**). After quality checks, **Trimmomatic** is

commonly employed to remove sequencing adapters and trim low-quality reads, enhancing the accuracy of downstream results (**Bolger et al., 2014**) .

Cleaned reads are typically aligned to the human reference genome (*hg38*) using **BWA-MEM**, a fast and accurate alignment tool that utilizes a seed-and-extend algorithm based on Maximal Exact Matches (MEMs) (**Li, 2013**). Post-alignment, **MarkDuplicates** from the Picard toolkit (v3.1.1.0) is used to flag PCR duplicates. This step has been validated in various genomic studies as essential for reducing false positives and maintaining data integrity **(Karcı & Kafkas, 2024).**

Variant calling is then performed using **FreeBayes**, a haplotype-based Bayesian genetic variant detector. FreeBayes offers the advantage of modeling multi-allelic sites and complex variant structures, which makes it suitable for diverse datasets (**Baraja-Fonseca et al., 2024; Garrison & Marth, 2012a**). The output VCF files are typically filtered using **VCFfilter**, which excludes low-quality variants based on depth and Phred scores **(Danecek et al., 2011; GATK Team, 2025; Jia et al., 2012; Li, 2011)**.

Finally, to predict the functional impact of variants, researchers use **SnpEff**. This tool annotates variants by classifying them into categories such as HIGH, MODERATE, LOW, or MODIFIER, based on genomic context and gene effect **(Cingolani et al., 2012**). These annotated variants can then be cross-referenced with known disease pathways for further interpretation.

Taken together, the use of these integrated tools across different studies has significantly advanced our understanding of the shared genomic and molecular pathways that underlie diabetes and periodontitis. They provide a framework for dissecting complex traits and pave the way for more targeted therapeutic approaches.

MATERIALS AND METHODS

Chapter 2: Materials and Methods

This study compares genomics profiles, focusing on DNA sequence variations and potential epigenetic changes, between individuals with periodontitis (**PD**) and those with type 2 diabetes mellitus combined with periodontitis (**PDDM**), relative to the normal human genome.

2.1. Materials and Methods

2.2. Biological Materials

2.2.1. Biological data

The whole study was entirely based on publicly available Whole Genome Sequencing (**WGS**) data retrieved from the:

- NCBI Gene Expression Omnibus. **(GEO)** *https://www.ncbi.nlm.nih.gov/geo/* and
- Sequence Read Archive **(SRA)**. *https://www.ncbi.nlm.nih.gov/sra*.

Table I represents the dataset used in this study. It includes the details or metadata of each corresponding dataset: Periodontitis (PD), Periodontitis with diabetes mellitus (PDDM), and Healthy control (HC).

Table I: Characteristics of the samples

Metadata	PDDM (Periodontitis + Diabetes Mellitus)	PD (Periodontitis only)
Organism	*Homo sapiens*	*Homo sapiens*
Tissue	Periodontium	Periodontium
Average Age of donor	80 years	58 years
Donor Sex	Female	Female
Library Layout	Paired-end	Paired-end
Geographic location	South Korea	South Korea
Sequencing Platform	Illumina HiSeq 4000	Illumina HiSeq 4000
Experiment Type	Whole Genome Sequencing (WGS)	Whole Genome Sequencing (WGS)
BioProject	**PRJNA1116808**	**PRJNA1116808**
BioSample	**SAMN42309066**	**SAMN42309062**
SRA Run Accession	**SRR29729000**	**SRR29729004**

2.3. Non-biological materials

The open-source tools used for the analysis of the data in our study are listed in Table II below.

***Table II**: The tools used in this genomic data.*

Tool	Location (Galaxy Platform)	Purpose in Workflow
FastQC	Galaxy (https://usegalaxy.org/)	Quality control check of raw sequencing reads.
Trimmomatic	Galaxy (https://usegalaxy.org/)	Adapter trimming and quality filtering of reads.
BWA-MEM	Galaxy (https://usegalaxy.org/)	Alignment of cleaned reads to the human reference genome.
MarkDuplicates	Galaxy (https://usegalaxy.org/)	Removal of duplicate reads from BAM files.
FreeBayes	Galaxy (https://usegalaxy.org/)	Variant calling (identification of SNPs and indels).
BCFtools	Galaxy (https://usegalaxy.org/)	Filtering and refinement of variant calls.
SnpEff	Galaxy (https://usegalaxy.org/)	Functional annotation of variants.
g: Profiler	External (https://biit.cs.ut.ee/gprofiler/)	Pathway enrichment analysis of annotated genes.
UpSetR	R/Bioconductor (https://cran.r-project.org/package=UpSetR)	Visualization of shared and unique variants (intersection analysis).

2.3.1. Methods

The following bioinformatic workflow or pipeline was followed to carry out this study effectively:

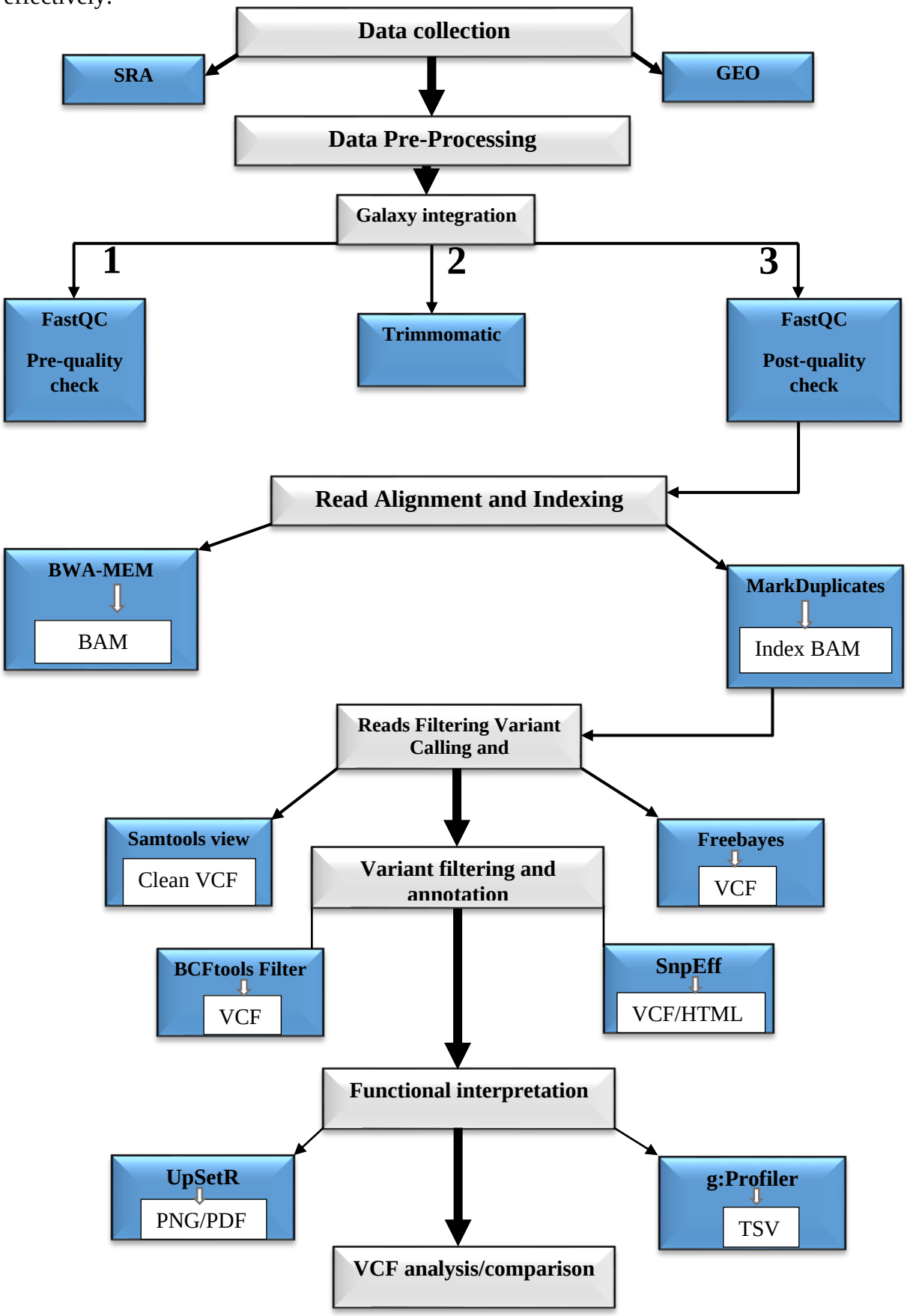


***Figure 7**: The Bioinformatic Workflow*

2.3.1.1. Data Collection

Publicly available human whole genome sequencing (WGS) datasets were systematically identified and retrieved using a stepwise search strategy.

2.3.1.1.1. Search strategy and keywords

Initial searches were conducted in the NCBI GEO and ENA repositories using combinations of keywords including “Diabetes”, “Periodontitis”, “Homo sapiens”. This strategy aimed to capture datasets relevant to periodontal disease (PD), type 2 diabetes mellitus (T2DM), and their comorbidity. https://www.ncbi.nlm.nih.gov/.

The results below in (**Figure 8a**) were identified correlating to studies investigating the human genome (*Homo sapiens*) related to diabetes and periodontitis. Among the retrieved entries were whole genome sequencing (WGS) datasets such as **WGS: PD20** (Accession: SRX5231162), and **WGS: PD19** (Accession: SRX5231161), each generated using the Illumina HiSeq 4000 platform with paired-end reads and genome strategy. The results also indicated a wide number of 488 available accessions under similar search parameters for better refinement and filtering.

Clicking on one of the Bioproject (**WGS: PD20**) in (**Figure 8a)** below led to the **SRA Run page** https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP513973&o=acc_s%3Aa, detailing the specific experiment (**SRX5231162**) associated with the dataset. This page provided metadata including the sequencing platform (Illumina HiSeq 4000), library design (MGI FS DNA library), submitting institution (Pusan National University), and linked studies such as **PRJNA1116808** and **SRP513973**, alongside the associated sample (**SAMN42309061**) classified under *Homo sapiens* as seen below (**Figure 8b**).

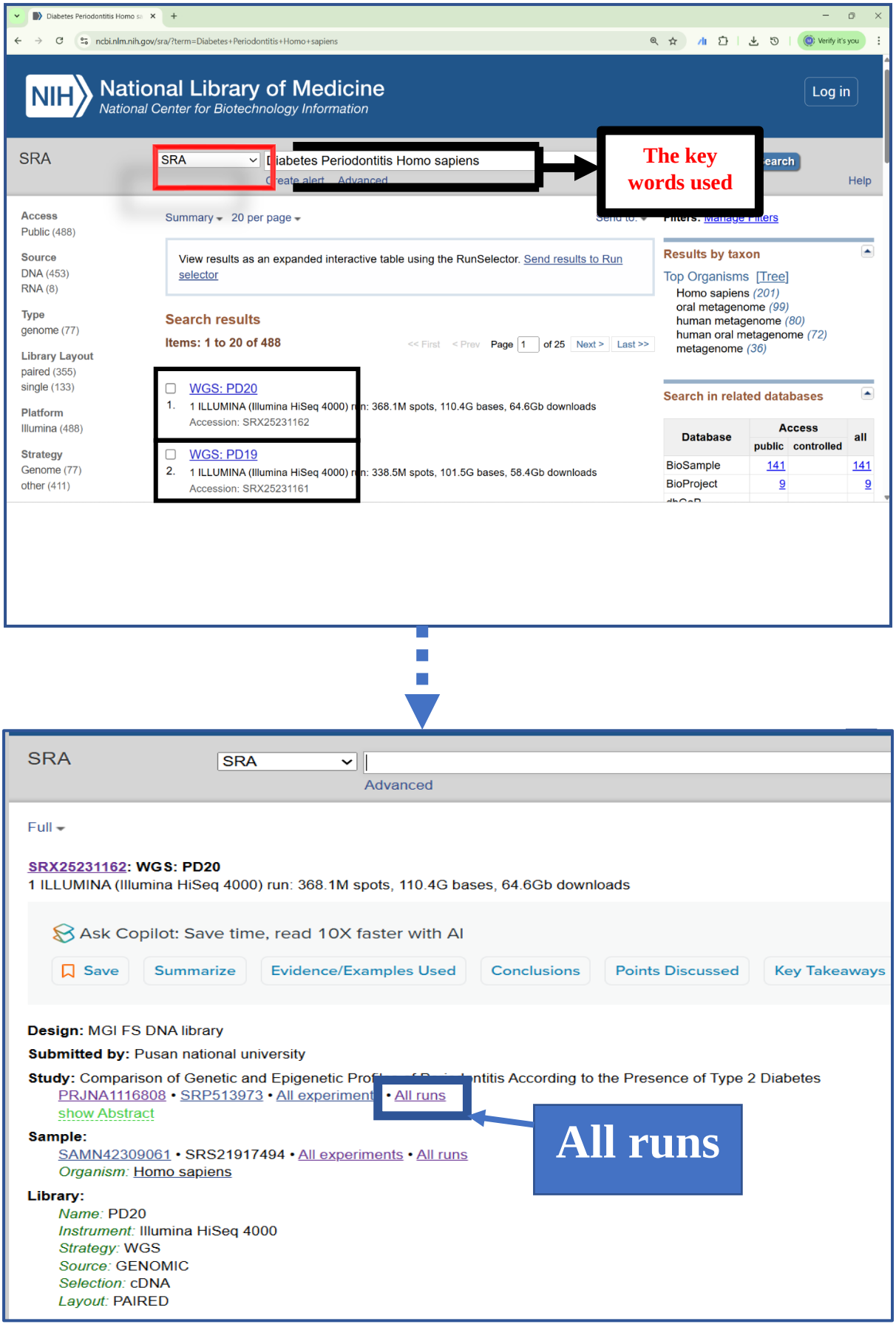


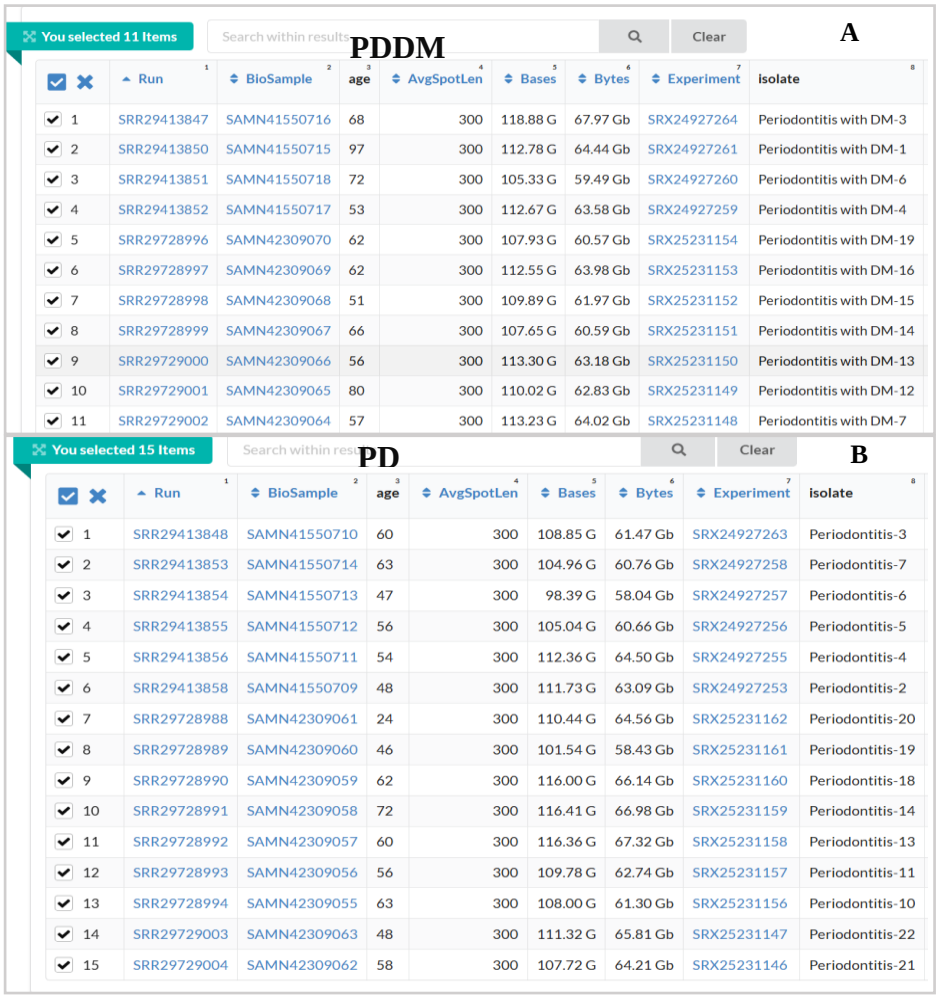
Figure 8: *SRA run page with experiment details.*

2.3.1.1.2. Identification of relevant studies

Relevant studies were shortlisted from the searches. A recent key BioProject identified was PRJNA1109994, titled "*Comparison of Genetic and Epigenetic Profiles of Periodontitis According to the Presence of Type 2 Diabetes*" by (**Kang et al., 2024**).The study consisted of patient groups diagnosed with periodontitis (PD), and periodontitis with diabetes mellitus (PDDM) below.

The next step was to sort and filter the dataset based on the relevance (PDDM, PD), and for that, the "**All Runs**" link on (***9***) above was clicked to allow access to all runs relating to the study.

The first step in filtering was to select all the Runs related to PDDM (**11 Runs**), followed by the ones related to PD (**15 Runs**), as shown below in). Full details on **Appendix A**.



	Run	BioSample	age	AvgSpotLen	Bases	Bytes	Experiment	isolate
1	SRR29413847	SAMN41550716	68	300	118.88 G	67.97 Gb	SRX24927264	Periodontitis with DM-3
2	SRR29413850	SAMN41550715	97	300	112.78 G	64.44 Gb	SRX24927261	Periodontitis with DM-1
3	SRR29413851	SAMN41550718	72	300	105.33 G	59.49 Gb	SRX24927260	Periodontitis with DM-6
4	SRR29413852	SAMN41550717	53	300	112.67 G	63.58 Gb	SRX24927259	Periodontitis with DM-4
5	SRR29728996	SAMN42309070	62	300	107.93 G	60.57 Gb	SRX25231154	Periodontitis with DM-19
6	SRR29728997	SAMN42309069	62	300	112.55 G	63.98 Gb	SRX25231153	Periodontitis with DM-16
7	SRR29728998	SAMN42309068	51	300	109.89 G	61.97 Gb	SRX25231152	Periodontitis with DM-15
8	SRR29728999	SAMN42309067	66	300	107.65 G	60.59 Gb	SRX25231151	Periodontitis with DM-14
9	SRR29729000	SAMN42309066	56	300	113.30 G	63.18 Gb	SRX25231150	Periodontitis with DM-13
10	SRR29729001	SAMN42309065	80	300	110.02 G	62.83 Gb	SRX25231149	Periodontitis with DM-12
11	SRR29729002	SAMN42309064	57	300	113.23 G	64.02 Gb	SRX25231148	Periodontitis with DM-7

	Run	BioSample	age	AvgSpotLen	Bases	Bytes	Experiment	isolate
1	SRR29413848	SAMN41550710	60	300	108.85 G	61.47 Gb	SRX24927263	Periodontitis-3
2	SRR29413853	SAMN41550714	63	300	104.96 G	60.76 Gb	SRX24927258	Periodontitis-7
3	SRR29413854	SAMN41550713	47	300	98.39 G	58.04 Gb	SRX24927257	Periodontitis-6
4	SRR29413855	SAMN41550712	56	300	105.04 G	60.66 Gb	SRX24927256	Periodontitis-5
5	SRR29413856	SAMN41550711	54	300	112.36 G	64.50 Gb	SRX24927255	Periodontitis-4
6	SRR29413858	SAMN41550709	48	300	111.73 G	63.09 Gb	SRX24927253	Periodontitis-2
7	SRR29728988	SAMN42309061	24	300	110.44 G	64.56 Gb	SRX25231162	Periodontitis-20
8	SRR29728989	SAMN42309060	46	300	101.54 G	58.43 Gb	SRX25231161	Periodontitis-19
9	SRR29728990	SAMN42309059	62	300	116.00 G	66.14 Gb	SRX25231160	Periodontitis-18
10	SRR29728991	SAMN42309058	72	300	116.41 G	66.98 Gb	SRX25231159	Periodontitis-14
11	SRR29728992	SAMN42309057	60	300	116.36 G	67.32 Gb	SRX25231158	Periodontitis-13
12	SRR29728993	SAMN42309056	56	300	109.78 G	62.74 Gb	SRX25231157	Periodontitis-11
13	SRR29728994	SAMN42309055	63	300	108.00 G	61.30 Gb	SRX25231156	Periodontitis-10
14	SRR29729003	SAMN42309063	48	300	111.32 G	65.81 Gb	SRX25231147	Periodontitis-22
15	SRR29729004	SAMN42309062	58	300	107.72 G	64.21 Gb	SRX25231146	Periodontitis-21

Figure 9*:Two selected different datasets, PDDM and PD with respect to the number of RUNS per dataset.*

2.3.1.1.3. Data download and Preparation

The selected sequencing runs were downloaded directly from the NCBI SRA repository. Files were prepared and uploaded to the Galaxy platform for further quality control, cleaning, and downstream analysis.

- Galaxy platform integration

Principle: The Galaxy platform (*https://usegalaxy.org/*) is a widely used, web-based, open-source resource that enables researchers to perform complex biomedical data analysis without any programming expertise.

This platform provides a graphical interface for uploading data, selecting tools, and executing bioinformatics workflows, ensuring computational workflows or analysis are accessible to scientists of all backgrounds. Galaxy's focus on reproducibility, automatically recording each analytical step to promote transparency and facilitate collaboration, is regarded as one of its most important features. Over time, Galaxy has expanded its capabilities to support cloud computing, interactive visualization tools, and interoperability with platforms such as Jupyter and RStudio. Thousands of tools covering a wide range of applications, including genomics, transcriptomics, and epigenomics, are now integrated into the platform. Its global community reinforces Galaxy's mission to promote open science and collaborative **(Afgan et al., 2018; Jalili et al., 2021)** by ensuring ongoing development, updates, and training resources.

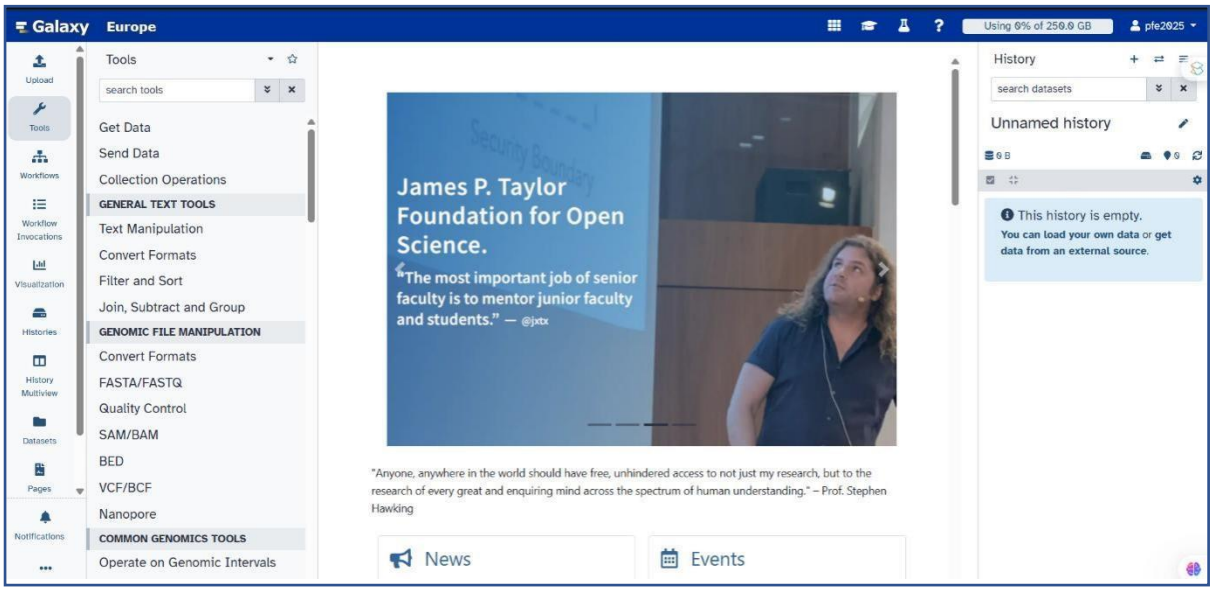


***Figure 10**: Galaxy Europe Platform Interface Demonstrating Toolset and Workflow Environment.*

Galaxy workflow

a-Data importation:

After multiple trials on different datasets, the SRR29729000: PDDM dataset (age: 80, sex: Female), and SRR29729004: PD dataset (age: 58, sex: Female), were chosen among other samples based on their representative ages, gender (same gender), and because of their data volume (bases and bytes) fell within the median range, avoiding extreme outliers in sequencing depth or quality.

This ensured the selected runs would be both representative and manageable for initial workflow optimization and testing. The raw FASTQ files corresponding to three datasets (Accession IDs) were imported using the Galaxy platform.

The two datasets were processed separately in different Galaxy accounts (**Galaxy EU and US**) due to the storage quota of Galaxy, which allows only 250GB per user, so due to our voluminous datasets, we processed them separately and later transferred the processed outputs to one account. continue with further analysis.

Process

- Opened the **Galaxy platform** and accessed the *search tools* panel.
- Navigated to **Get Data** → selected *Download and Extract Reads in FASTQ format from NCBI SRA* **(1).**
- The output files format was chosen as **gzip-compressed FASTQ (2).**
- Clicked the **Run Tool (3)** button to initiate download and extraction from the **SRA database**.
- Upon completion, the outputs **paired-end data and single-end data files (4)** were ready for downstream analysis.

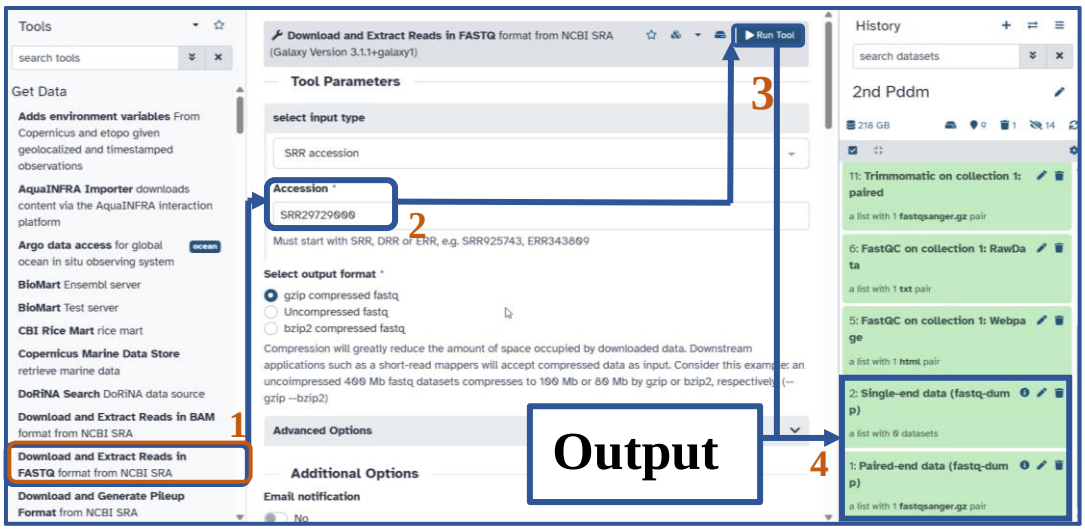


***Figure 11**: Retrieval of raw sequencing data (SRR297290900) using Galaxy's SRA download tool.*

b-Quality assessment

Fast-QC pre-quality assessment: FastQC is a quality assessment application that provides quick visual assessments and metrics for sequence data from high-throughput platforms such as Illumina. FastQC generates detailed reports on per-base sequence quality, GC content, adapter contamination, sequence duplication levels, and overrepresented sequences (Andrews, 2010). This allows researchers to identify potential issues in their datasets and determine if additional pre-processing steps.

Process:

- Search for **"FastQC"** **(1)** in the tool panel.
- Select the *Paired-end data* **(2)** as the dataset collection.
- Leave the contaminant list and adapter list (default).
- Click **"Run Tool"** **(3)** to initiate the quality check.
- Generated outputs **(4)**: **RawData** and **Webpage report** appear in the history panel.

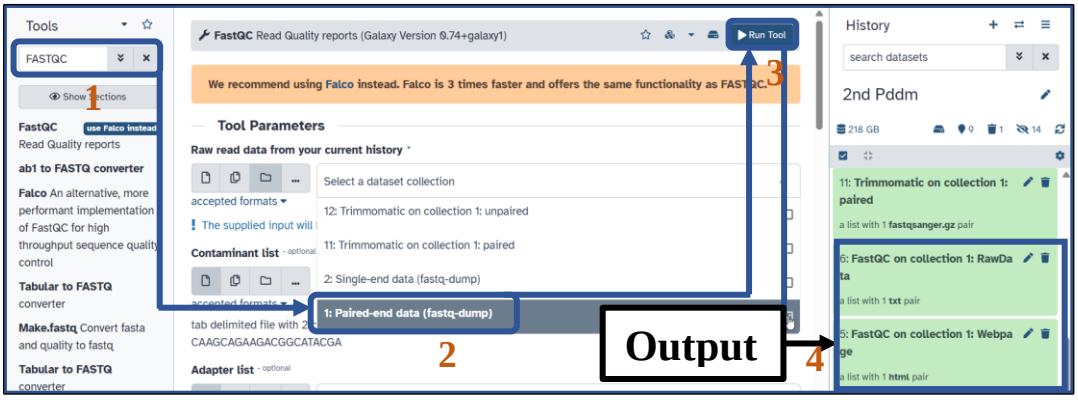


***Figure 12**: FastQC quality assessment of imported paired-end FASTQ data,*

c- Quality control

-Trimmomatic tool

Principle: a widely used tool for trimming low-quality bases and removing sequencing adapters (**Bolger et al., 2014**). This step ensures that only high-quality data proceeds to downstream analyses, improving accuracy and reducing bias.

Process:

- Open the Galaxy platform and locate the tool search bar.
- Type **"Trimmomatic"** (1) to locate the tool under "All Tools".
- Click on **Trimmomatic**.
- In the tool parameters, under **"Single-end** or **paired-end reads"**, change the default option from **Single-end** to **Paired-end** (2) to ensure both ends are trimmed.
- Select a **paired-end FASTQ dataset** (3) already uploaded in the initial steps (Download data) from the history (*Paired-end data (fastq-dump)*).
- Click **"Run Tool"** (4) to start trimming.
- The output will appear in the history panel as two separate entries (5): one for **forward** and one for **reverse-trimmed** reads.

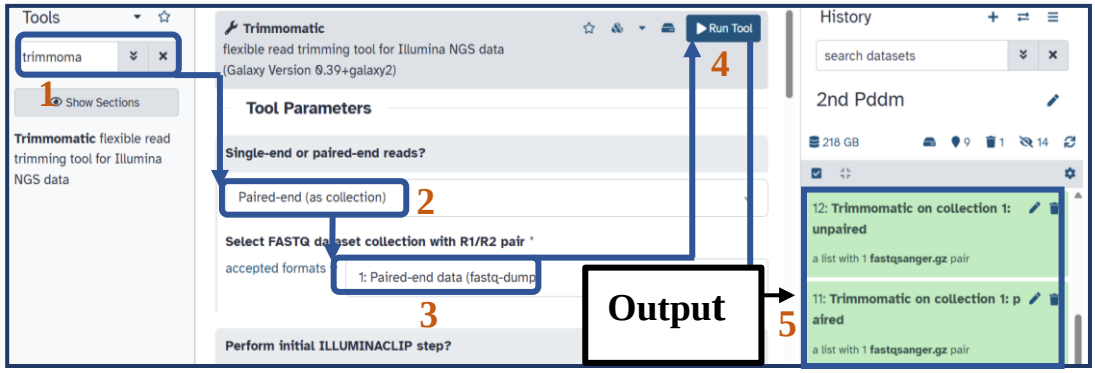


***Figure 13**: Reads trimming with Trimmomatic to ensure equal distribution of bases within the sequence.*

d-Fast-QC post-quality assessment

Principle: Following trimming with Trimmomatic, a second round of quality assessment was conducted using FastQC to evaluate the effectiveness of the trimming process. This post-quality check followed the same steps previously used in the initial FastQC assessment **Figure 12**, including tool selection, dataset collection input, and execution.

Important note: It is essential to choose the correct file(s) format while dealing with bioinformatic tools **(1)**. By reapplying the same workflow, this ensured consistency in quality metrics and verified that low-quality sequences and adapters had been effectively removed before proceeding to downstream analyses.

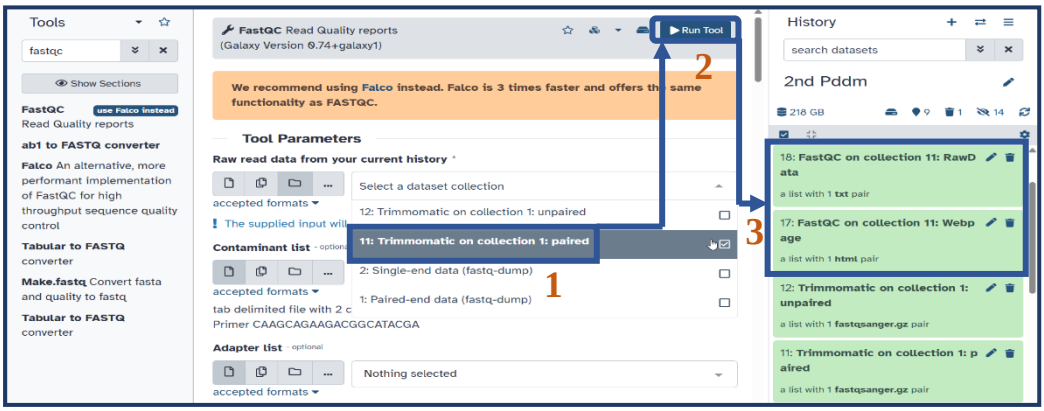


Figure 14: *FastQC post-quality analysis.*

The above principle or steps were applied to all two different datasets downloaded from: *https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP513973&o=acc_s%3Aa* (PDDM, PD), which included: Data collection (***NCBI SRA***), pre-quality analysis (***FastQC***), quality control (***Trimmomatic***), and post-quality analysis (***FastQC***).

e-Reads alignment with BWA-MEM

Principle: BWA-MEM (Burrows-Wheeler Aligner - Maximal Exact Matches) is a widely used aligner optimized for speed and accuracy in mapping short DNA fragments generated by next-generation sequencing. It employs a seed-and-extend approach, identifying maximal exact matches (MEMs) as seeds and extending them using the Smith-Waterman algorithm. This facilitates efficient and accurate alignment of sequencing reads to large reference genomes (**Li, 2013)**.

Process:

Cleaned and high-quality paired-end reads were aligned to the Homo sapiens reference genome (hg38) using the BWA-MEM algorithm within the Galaxy platform (**Figure 15**).

The tool was accessed through Galaxy's NGS: Mapping section. The trimmed forward and reverse FASTQ reads from all 2 datasets were selected as input, and the default alignment parameters were used. It is important to choose the correct reference human genome, the hg38, which is already indexed from Galaxy's built-in genome list **(1)**, and selecting appropriate file extensions or reads for both Forward (R1 paired) and reverse (R2 paired) **(2)** is crucial. The output consisted of a sorted BAM file **(4)** and alignment summary statistics.

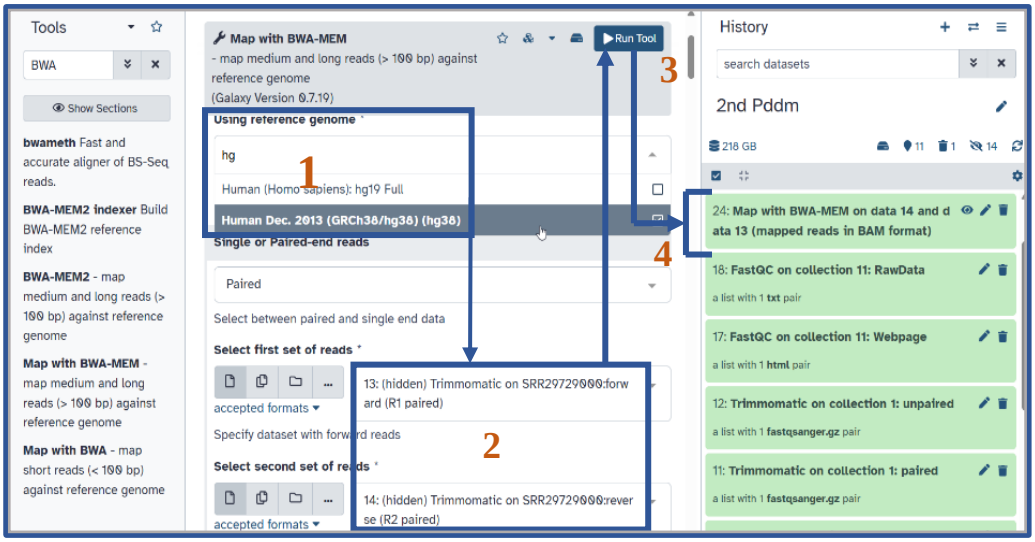


***Figure 15**: BWA-MEM interface for PDDM (SRR29729000) paired reads.*

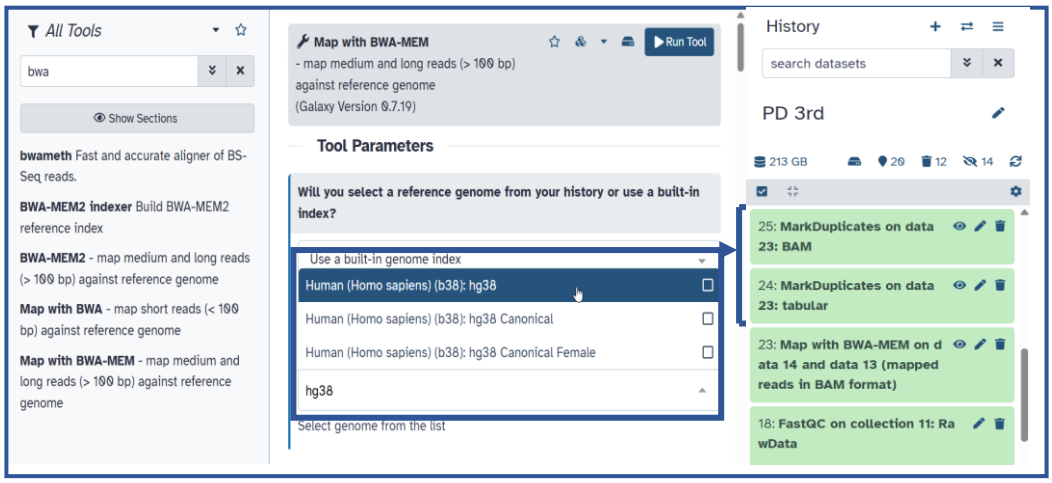


***Figure 16**: BWA-MEM interface for PD (SRR29729004) paired reads.*

f- Post-alignment processing

-Duplicate marking and BAM indexing

Principle: Duplicate marking is based on base quality scoring, ensuring that the most reliable reads are retained for downstream analysis. The importance of this step was highlighted in a study by **(Karcı & Kafkas, 2024)**, which showed that skipping the double marker resulted in the loss of over 42,000 true negative loci and the inclusion of over 10,000 false positives in variant calling in Pistacia vera. These findings highlight the critical role of duplicate marking in preserving data integrity and improving the reliability of variant calls in high-throughput sequencing.

Process:

PCR duplicate reads were identified and marked using the MarkDuplicates tool (Picard, v3.1.1.0) in Galaxy using default parameters to reduce the number of false-positive variant calls and increase the accuracy of variant detection. Rather than removing duplicates, the tool marks reads that originate from the same DNA fragment, often a result of PCR amplification, which can artificially inflate read depth and introduce misleading variant signals.

Choosing the correct input (BAM file from the alignment by BWA-MEM in (**Figure 16**) is crucial in this phase as seen below in (**Error! Reference source not found.**; **(1)**. Upon c ompletion, t wo files are outputted (Indexed BAM file and tabular); **(2).**

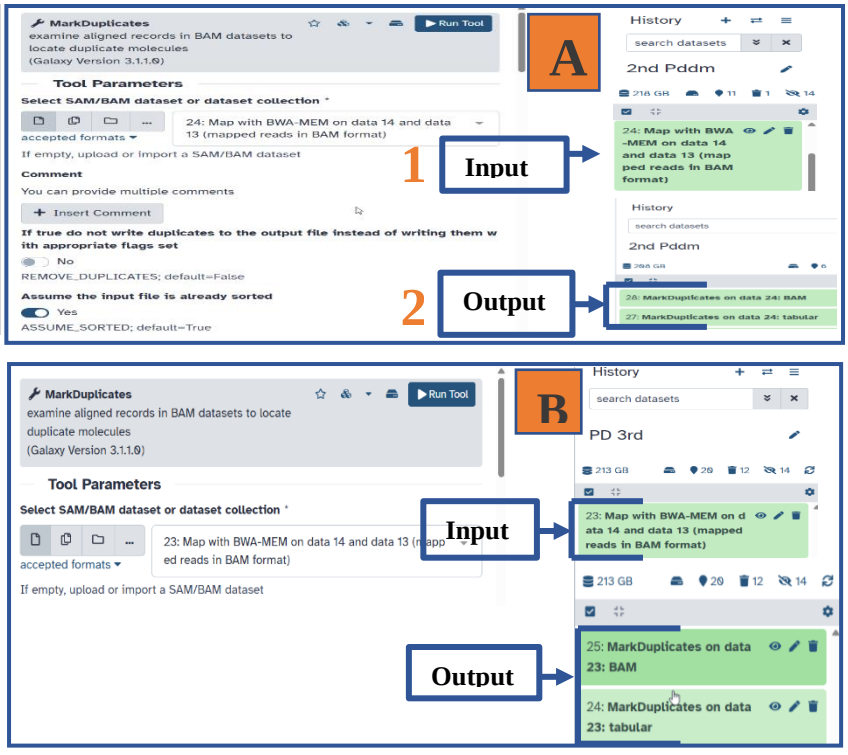


***Figure 17**: Removing duplicates reads from PDDM (A) and PD (B).*

g-Variant calling

Principle: FreeBayes is a Bayesian-based genetic variant detector designed to call polymorphisms, including SNPs, indels, and complex events, by directly modelling haplotype structure from aligned sequencing reads **(Garrison & Marth, 2012a)**. Unlike position-based callers, FreeBayes uses a probabilistic framework to identify likely genotypes, making it particularly suitable for both diploid and polyploid organisms. Its implementation has been widely adopted in large-scale genomic studies due to its ability to call multi-allelic variants and incorporate mapping qualities, base qualities, and alignment complexity into genotype likelihood calculations. Studies such as **(Baraja-Fonseca et al., 2024)** have compared FreeBayes with other tools (e.g., GATK and SAMtools) and reported its strength in handling large and complex genomes with consistent performance.

Process

- Pre-processed and deduplicated BAM files were used as input for variant calling.
- The tool **FreeBayes:** Bayesian genetic variant detector (Galaxy version) was accessed via the usegalaxy.us server.
- The reference genome was set to **Human Dec. 2013 (GRCh38/hg38)**.
- Variant calling was configured under the following settings:
- Read alignment input: Deduplicated and indexed BAM files
- Ploidy: 2 (diploid organism)
- Target region restriction: Not applied (whole genome)
- Output mode : Default VCF format
- Run mode: Individual (non-batch)

• All other parameters were set to **default**, optimized for germline variant calling in human datasets.

• The resulting VCF file contained SNPs and short indels, which were used for downstream filtering and annotation.

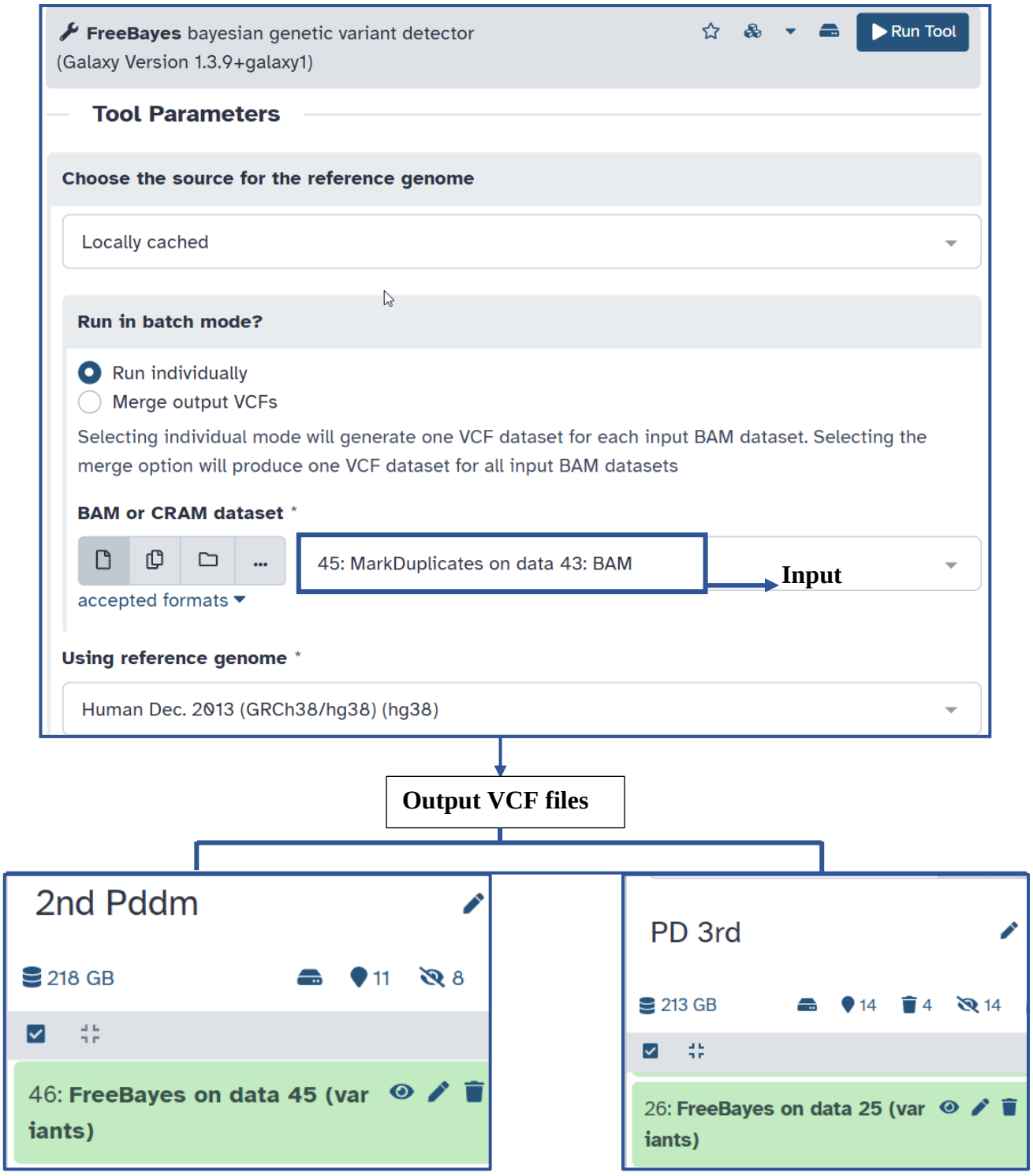


Figure 18: *Variant Calling with FreeBayes and Sample-specific Outputs.*

h-Manual SNP Inspection and Cross-validation with External Tools

Principle: To validate whether key genes of interest from literature (such as *MCUR1*, *RAP2A*, and *FOS*) exhibited variation in our samples, we first performed genome-wide variant calling using FreeBayes. Then, we applied a manual inspection strategy using two tools UCSC Genome Browser and SNPnexus; to visually confirm and biologically annotate the existence and potential pathogenicity of variants found in these genes.

Process:

Step 1: UCSC Genome Browser

FreeBayes-filtered VCFs were uploaded to the UCSC Genome Browser, specifically exploring chromosomal regions corresponding to *MCUR1*, *FOS*, and *RAP2A*.

Zoomed in to specific gene loci to visually confirm the presence of SNPs, Checked for overlaps with known dbSNP entries to validate variant authenticity, and Confirmed rsIDs (e.g., *rs3810196* in *MCUR1*) when available.

Step 2: SNPnexus Functional Annotation

Chromosomal positions extracted from UCSC-confirmed variants were batch submitted to SNPnexus (https://www.snp-nexus.org/).

Assessed **gene consequence** (3'UTR, intronic, missense, etc.), Captured **non-coding impact scores** (e.g., CADD, FunSeq2), Retrieved **regulatory element overlaps,** Used these annotations to prioritize variants with potential pathogenic or regulatory roles, especially within PDDM group.This pre-annotation inspection allowed us to correlate raw SNP positions with known disease-associated variants, increasing confidence in biological interpretation before further filtering and enrichment.

➢ Variant annotation using Single nucleotide polymorphism Effect (SnpEff)

Principle: In this study, we used SnpEff, a well-established bioinformatics tool for functional annotation of variants (Cingolani et al., 2012b). It was used to annotate the variants that were detected in WGS data sets (PDDM, PD,) by assigning predicted effects.

1. **Input configuration:** Ensure input is a **VCF file** (VCF output from Freebayes (**Figure 18**).
2. **Output configuration:** Match the input format (VCF format)
3. **Run** the tool to produce summary stats and fully annotated SnpEff VCF file extension.

Process:

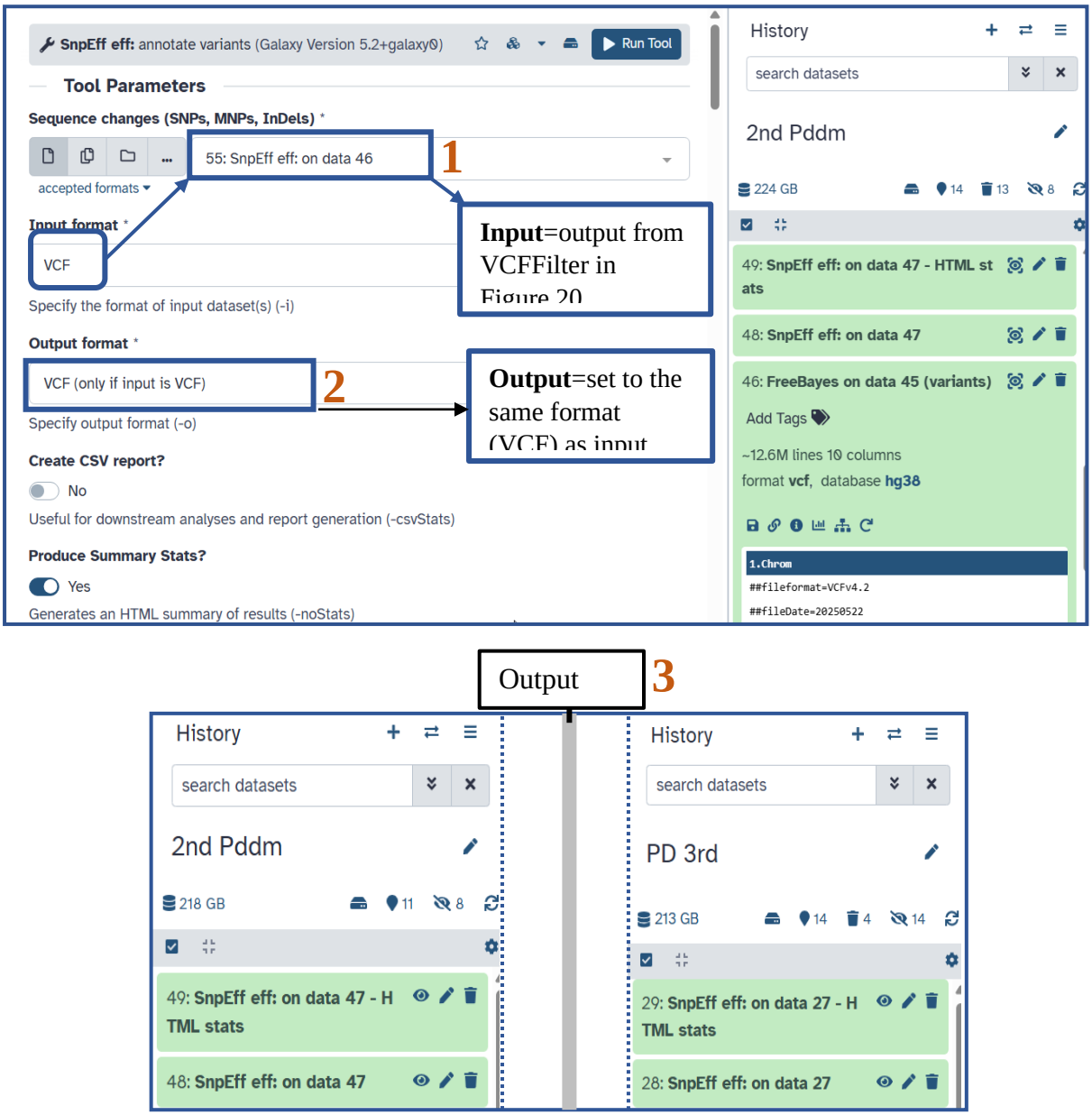


Figure 19: *Illustration and the stepwise process of using SnpEff in Galaxy.*

- **Filtering high and moderate impact Variants using SnpSift**

Due to compatibility issues with the Galaxy version of SnpSift, specifically its inability to parse ANN fields introduced by recent SnpEff annotations, we encountered multiple execution errors when attempting direct functional filtering (e.g., by IMPACT or GENE) from. This was crucial for our study, as we needed the **ANN** field to isolate HIGH and MODERATE impact variants for downstream comparative gene analysis.

To resolve this, we transitioned the filtering and extraction of annotated genes to an offline environment using PowerShell command with SnpSift v5.2 installed (**Figure 21** and **Figure 22**). In Galaxy, we continued using SnpSift only to perform basic quality filtering (QUAL > 100 and DP > 10) (**Figure 20)** prior to UCSC and SNPnexus investigations.

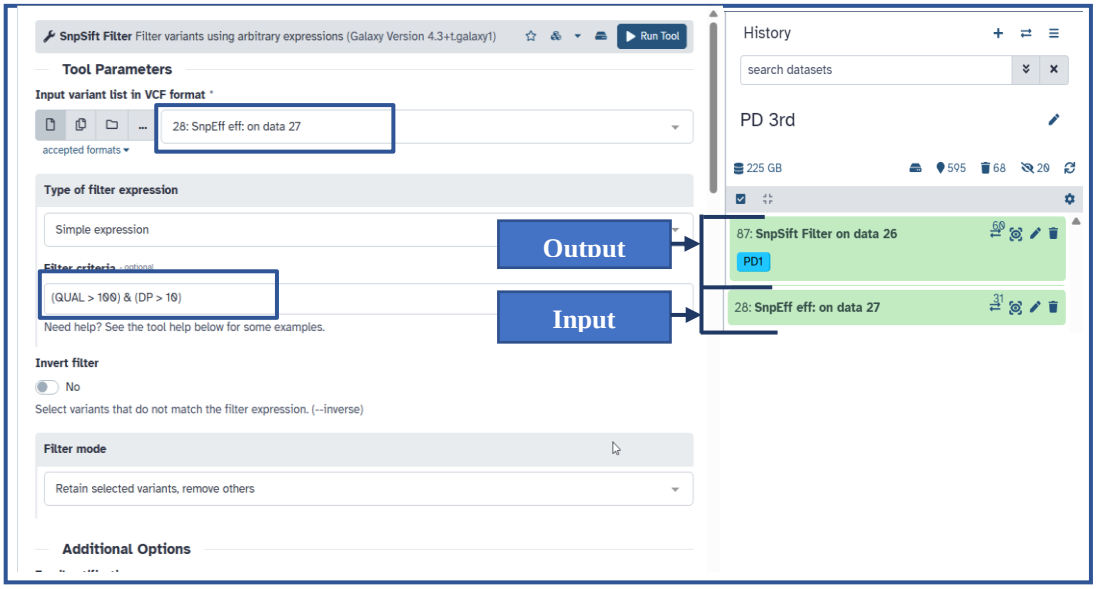


Figure 20: *Quality filtering using SnpSift; QUAL: Quality, DP: Depth.*

➢ Offline variant filtering using Powershell script

Principle: SnpSift is an extension of SnpEff capable of filtering the annotated VCF files by the SnpEff. For this study, variants were filtered to retain only those annotated as having a **high** or **moderate** impact, as these are the most likely to affect protein structure or expression. This step refines the gene list to include only those variants that are most relevant to disease mechanisms, such as immune dysregulation or epithelial degradation, which are seen in PD and T2DM complications (Cingolani et al., 2012c).

We downloaded two variant datasets (**Figure 21**), one for combined Periodontitis and Type 2 Diabetes **(PDDM)** and the other for Periodontitis-only **(PD)**, which were pre-annotated using **SnpEff** on the Galaxy platform (**Figure 19**). These files were in VCF format and contained functional annotations, including predicted variant impact levels. However, due to persistent syntax errors when attempting to filter **HIGH** and **MODERATE** impact variants using Galaxy's SnpSift interface (version mismatch), we opted to perform the filtering and downstream analysis offline via **PowerShell**.

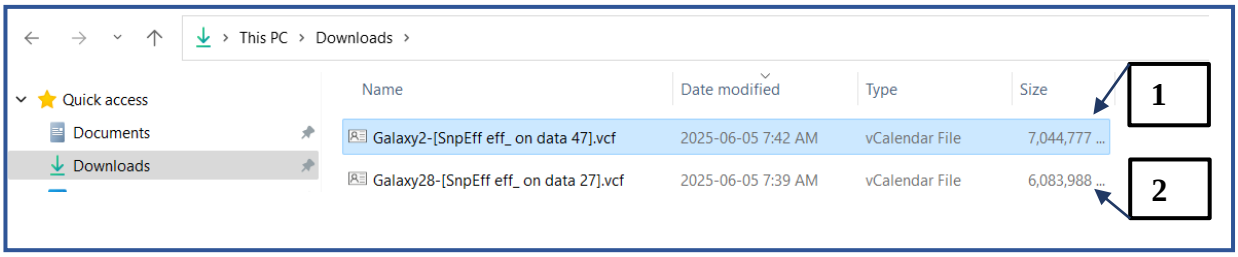


Figure 21: *The two datasets downloaded from the galaxy/history/SnpEff output VCF files and stored under local machine.*

Process:

To start filtering, we installed the necessary environment on a local Windows machine, including Java and the standalone **SnpEff/SnpSift v5.2** package. We then used PowerShell to execute a series of commands **Table III** that filtered variants (**High/Moderate impact**), extracted genes, cleaned the data, and compared the gene lists between PD and PDDM groups. This generated 4 separate files (vcf file containing filtered high and moderate impacts variants and txt file with a list of associated genes for each group) (**Figure 22**); the first one for PDDM having variants of both high and moderate impact, then the second one for PD having also high and moderate impact variants in the text format **(.txt).**

Table III:*The PowerShell workflow for SnpSift.*

Step	Purpose	PowerShell Command
1	Navigate to the working directory containing SnpSift.jar and the genome database	cd C:\Users\DAVE.MOTJOLOPANE\Downloads\snpEff_latest_core\snpEff
2	Filter PDDM dataset to retain only HIGH and MODERATE impact variants	java -Xmx4g -jar SnpSift.jar filter "(ANN[*].IMPACT = \'HIGH\') | (ANN[*].IMPACT =
3	Filter PD dataset for HIGH and MODERATE impact variants (from within the PD	java -Xmx4g -jar ..\SnpSift.jar filter "(ANN[*].IMPACT = \'HIGH\') | (ANN[*].IMPACT =
4	Extract gene names from the filtered PDDM VCF	java -Xmx4g -jar SnpSift.jar extractFields PDDM_filtered.vcf ANN[*].GENE >
5	Extract gene names from the filtered PD VCF	java -Xmx4g -jar ..\SnpSift.jar extractFields PD_filtered.vcf ANN[*].GENE >
6	Clean and deduplicate the PDDM gene list (tab-split, remove blanks, sort,	Get-Content PDDM_genes.txt | ForEach-Object { $_ -split "`t" } | Where-Object {
7	Clean and deduplicate the PD gene list	Get-Content PD_genes.txt | ForEach-Object { $_ -split "`t" } | Where-Object {
8	Identify genes shared between PDDM and PD groups	Compare-Object -ReferenceObject (Get-Content .\PDDM\PDDM_genes_clean.txt)
9	Identify genes unique to the PDDM group	Compare-Object -ReferenceObject (Get-Content .\PDDM\PDDM_genes_clean.txt)
10	Identify genes unique to the PD group	Compare-Object -ReferenceObject (Get-Content .\PDDM\PDDM_genes_clean.txt)

***Figure 22**: Two .txt files generated by PowerShell with fully annotated and filtered datasets.*

The vcf filtered files were then re-uploaded to galaxy (**Figure 23**) .

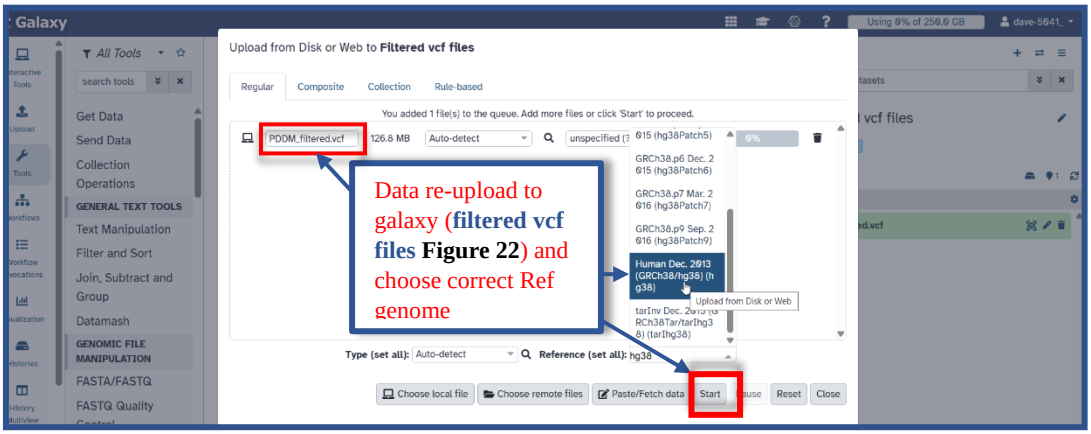


Figure 23: *Re-upload of cleaned and filtered SnpEff vcf files.*

The re-uploaded dataset files (vcf) were then integrated into UCSC genome browser to cross check the existence of common SNPs (**rs**IDs).

Principle: UCSC, ***https://genome.ucsc.edu/*** is a Genome Browser allowing researchers to view the 22 autosomes and the sex chromosomes (chrY and chrX) at any genomic level of detail. This includes everything from the complete chromosome to a single base pair. The browser compiles data from scientists and research conducted at UCSC and elsewhere into an interactive, easy-to-navigate graphical interface. The browser also contains genomes for more than a hundred other organisms.

Figure 24*: Interface pf UCSC Genome Browser*

Process:

Data import via galaxy server into UCSC

- Navigate to galaxy history panel (**A**).
- Click on the (eye icon) (**1**) corresponding to vcf file (newly uploaded **Figure 23**).
- Galaxy will open a new panel (**2**) on the left with a list of suggest tools to visualize the data, and choose the one suitable for making visual cross link between the variants found in the genes of interest and the analysed data (in this case we chose UCSC) then search for a gene of interest in the search bar (**B**).

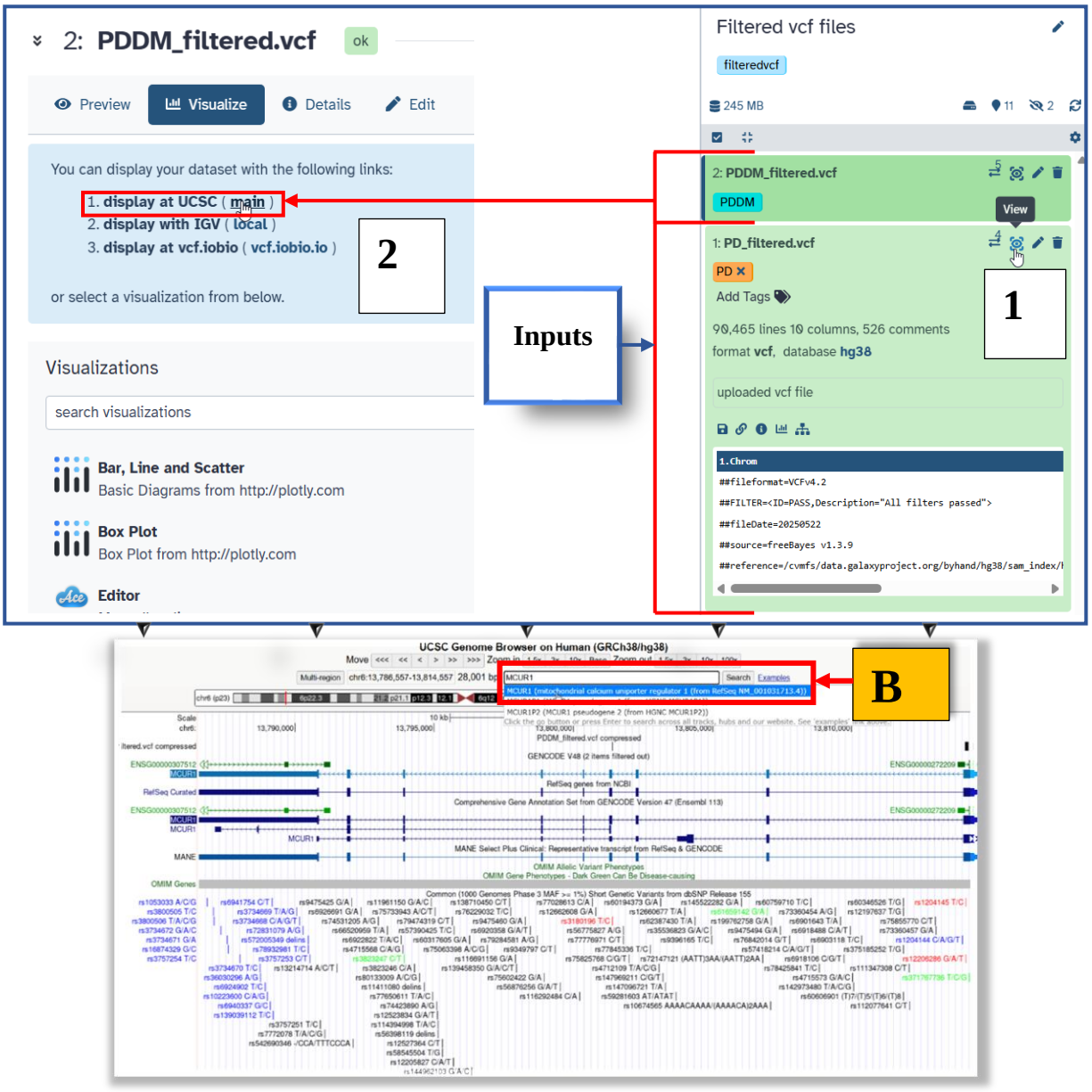


***Figure 25**: Importing galaxy data to UCSC for cross checking (variants in gene of interest and our data).*

➢ **Enabling of the Custom Tracks with UCSC**

Principle: Custom tracks which is basically composed of our dataset samples is by default enabled upon opening of the UCSC. But it must be customized based on the relevance of the study. For example; **OMIM and ClinVar** are enabled for evaluation of the clinical relevance and disease association. After choosing the required customs, click on the refresh to add it to the top main interface (welcome screen **Figure 25**).

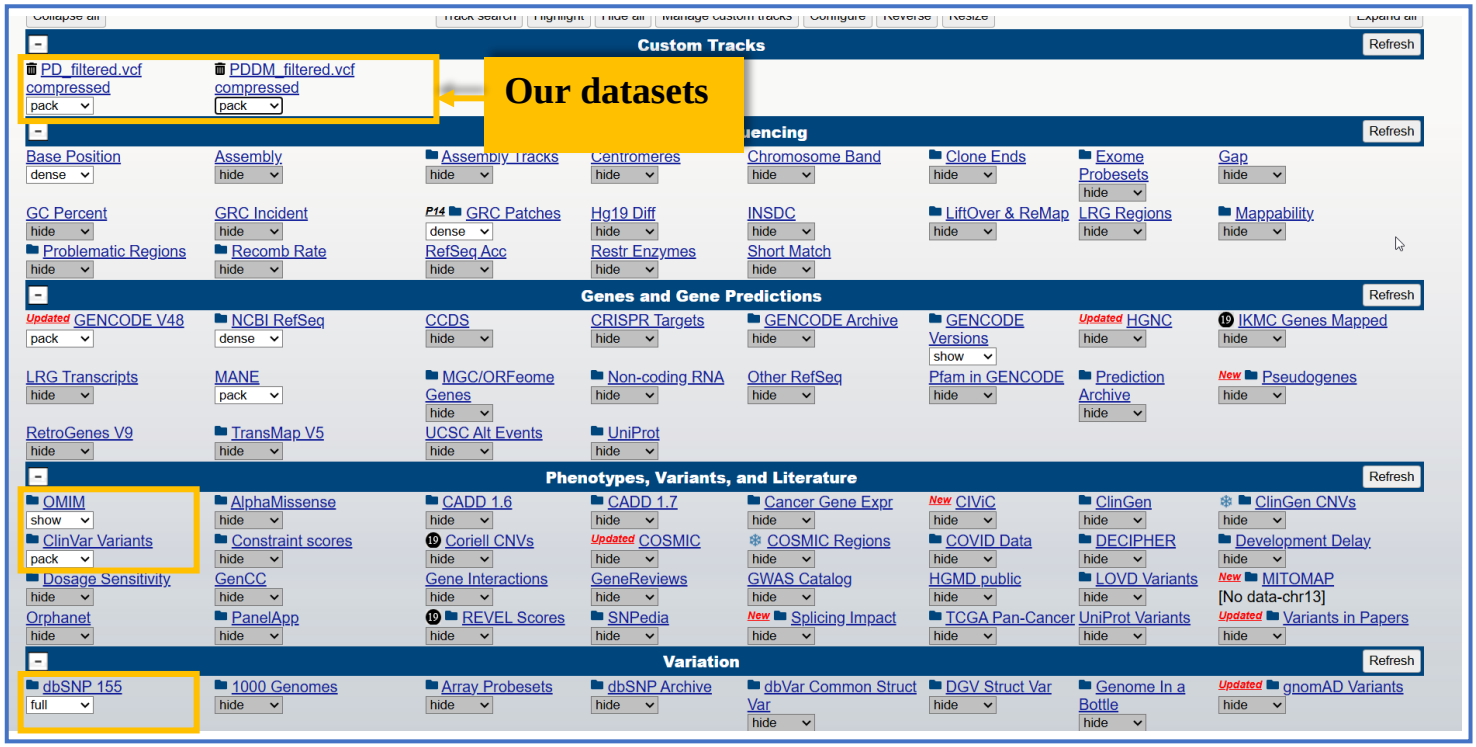


Figure 26: *Custom Tracks within UCSC genome browser.*

The objective for steps above in (**Figure 25**) was to visually identify overlaps between SNPs in the filtered datasets (the 2 vcf files uploaded into galaxy **Figure 23**) and three candidate genes of interest, MCUR1, FOS, and RAP2A; based on evidence from prior literature. For each gene, we manually examined the genomic coordinates to determine whether SNPs in our datasets aligned within or in proximity to these genes. Where alignment was confirmed, the chromosome position, reference, and alternate alleles of the matching SNPs were recorded.

For MCUR1 variants (**rs Ids**) , we chose those ones overlapping with our base change in our data T>C and those lying within the close range in order to maximize the quantity of variants.

Manual inspection of the overlapping common variants (our dataset mapped with RefSeq and the gene of interest) was done by:

Process:

- Visually track or trace down (to the common dbSNPs, **B**) the colored vertical pairs (marked in red boxes below, **A**), this will reveal to which variant/**rs** ID position each pair corresponds to in the common or known variants in (**Figure 27**), **B**. This manual visualization was consistently applied for all three genes (MCUR1, FOS, RAP2A) with both filtered VCF data files (**PD_filtered_vcf** and **PDDM_filtered_vcf**).

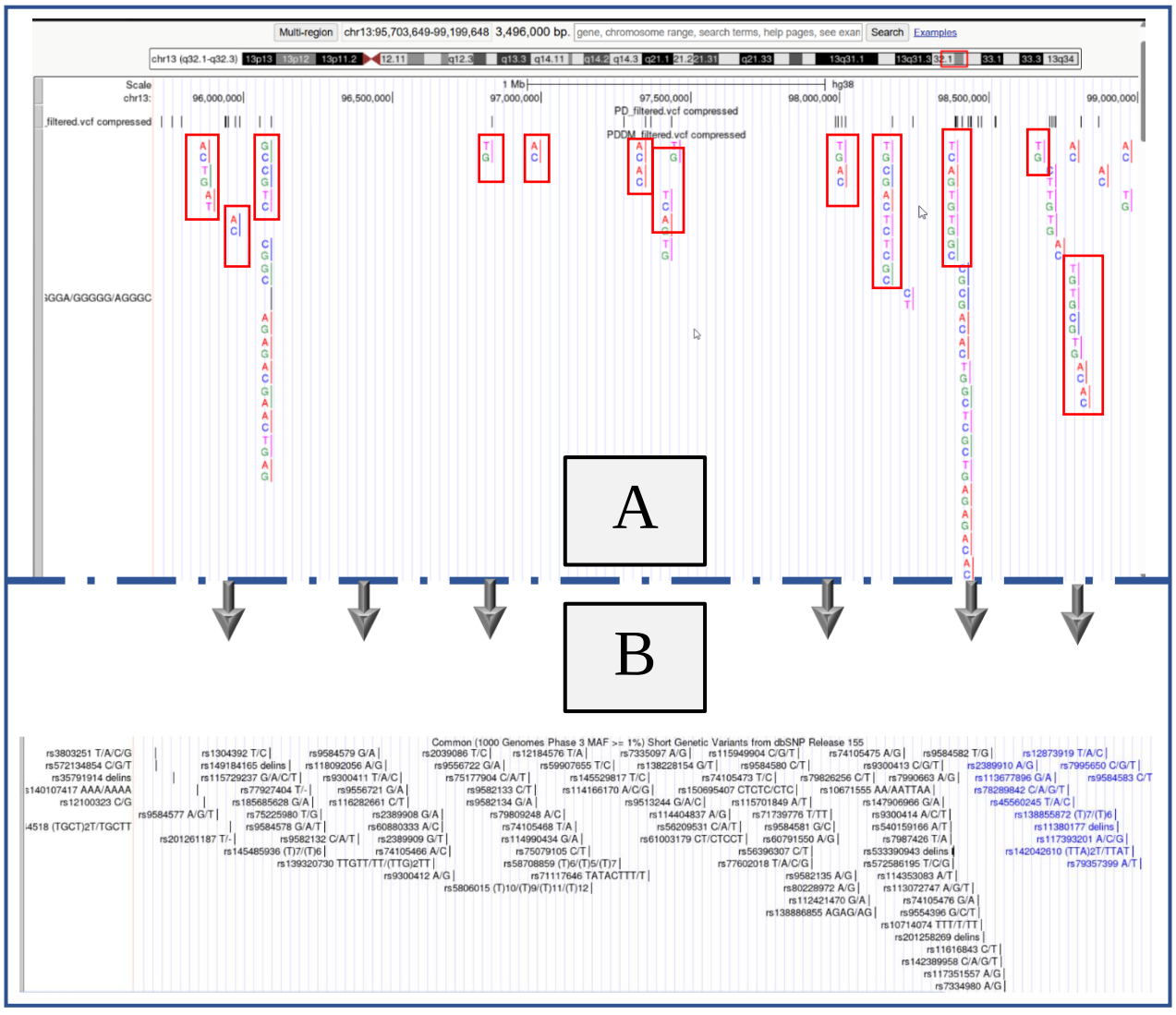


***Figure 27**: Visual inspection of common variants in UCSC genome browser.*

These verified SNP positions, corresponding to regions of interest identified in UCSC, were extracted as simple coordinate sets and prepared in tabular format (**CHROM, POS, REF, ALT**). This curated list formed the input for the subsequent SNPnexus analysis, where further functional, regulatory, and pathogenic annotation was performed.

All visually observed rsIDs that aligned well with our custom tracks (PD and PDDM vcf data files) were recorded for each of the 3 genes separately, then uploaded to SNPnexus Figure 28 (**B**), which is basically a web-based platform that supports comprehensive annotation of genomic variants across multiple biological contexts, to functionally characterize and prioritize single-nucleotide polymorphisms (SNPs) based on the following criteria Figure 29:

- Gene-centric consequences via **RefSeq, Ensembl**, and **UCSC.**
- Pathogenicity predictions via **CADD, SIFT,** and **PolyPhen**.
- Regulatory element overlaps from **ENCODE** and **Ensembl Regulatory Build**.
- Clinical associations from **ClinVar, OMIM**, and **GWAS**. Its ability to annotate both coding and non-coding variants from the GRCh38 human genome assembly makes SNPnexus highly suitable for disease-related variant prioritization workflows (Dayem Ullah et al., 2018).

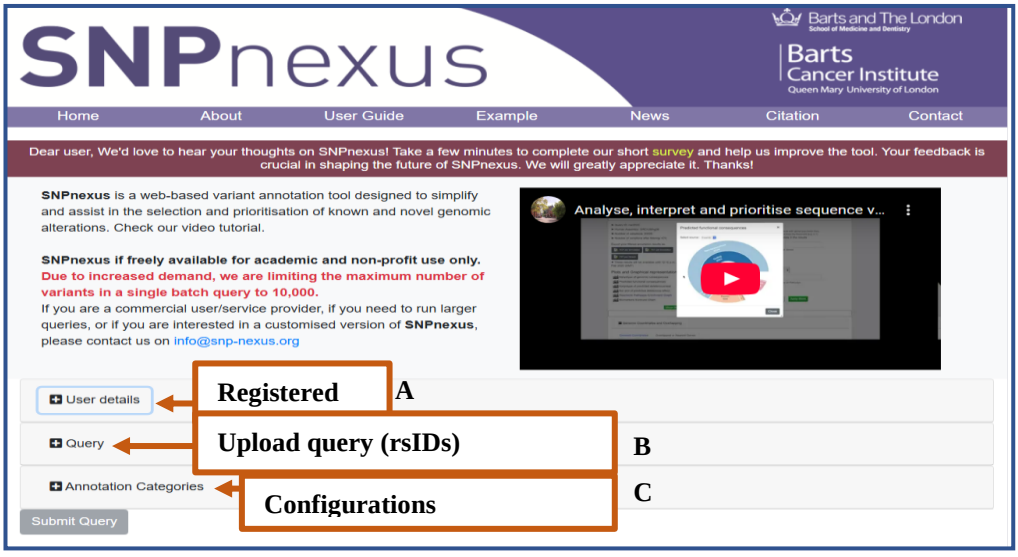


Figure 28: SNPnexus interface.

Process:

- Upload of query (rsIDs identified at (**Figure 27**) and we chose a minimum of 6/7 per gene.
- Configure Annotation category as shown below in (Figure 29).
- Then click submit query.

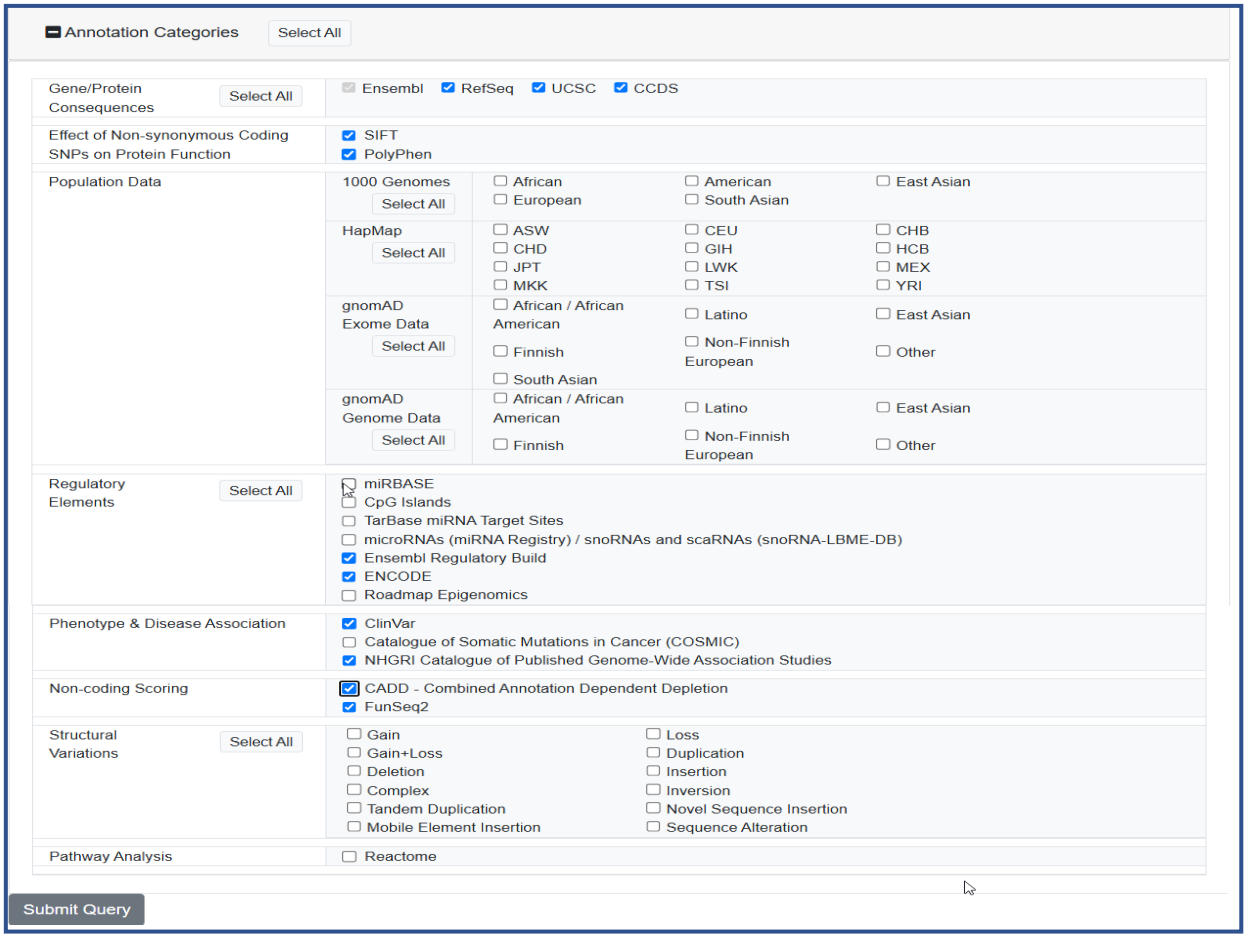


Figure 29: SNPnexus query configuration.

- **Gene set comparison via Venn Diagrams (Preparation for biological interpretation)**

Principle: To explore the biological convergence and divergence between PD-only and PDDM gene sets generated at (**Figure 22**), we used 2-set Venn diagram analysis **(Lam et al., 2016)**. This visual representation aids in identifying shared genetic factors and those exclusive to each condition, providing insights into potential common pathways and distinct mechanisms underlying T2DM and PD comorbidity. *https://bioinformatics.psb.ugent.be/cgi-bin/liste/Venn/calculate_venn.htpl*.

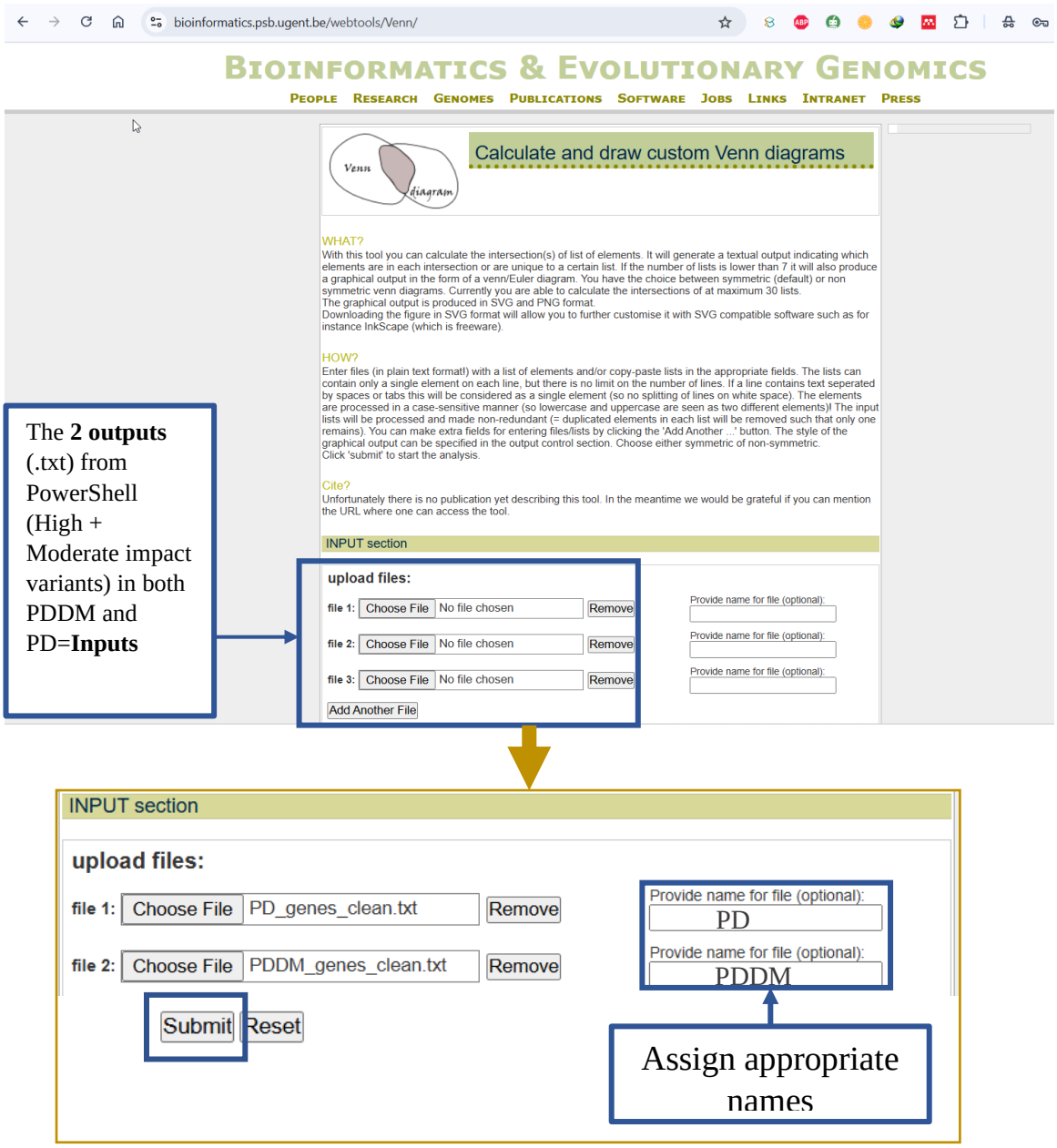


***Figure 30**: Interface of Webtools Venn for Bioinformatics.*

The venn diagram generated 1 output gene list in .txt format named (**venn_result1749299305304532341.txt**); separated by their unique naming, (PDDM + PD, PDDM only, PD only). The gene list from Venn diagram contained some unrecognized formatting or characters (ॐ) and spaces, so in order to prepare them ready for biological interpretation, we cleaned them using the PowerShell prompt (**Figure 31**), generating 1 input file named (**venn_cleaned_genes.txt**) which we split into 3 (according to the corresponding gene list) gene list with correct formatting and named differently (**genes_shared, genes_pddm_unique, genes_pd_unique**) in order to distinguish them from the first output (**Figure 32**).

```powershell
cd "$HOME\Downloads"

Get-Content "venn_result1749299305304532341.txt" |
    ForEach-Object { $_ -replace "[^\x20-\x7E]", "" } |
    Where-Object { $_ -match "^[A-Za-z0-9_-]+$" } |
    Set-Content "venn_cleaned_genes.txt"
```

***Figure 31**:PowerShell command to clean and remove the unknown formatting.*

```
cd "$HOME\Downloads"

# Load all lines
$lines = Get-Content "venn_cleaned_genes.txt"

# Create empty arrays
$shared = @()
$pddm = @()
$pd = @()

# Detect blocks by header
$currentList = ""

foreach ($line in $lines) {
    switch -Regex ($line) {
        "^PDDM\s+PD"   { $currentList = "shared"; continue }
        "^PDDM\b"      { $currentList = "pddm"; continue }
        "^PD\b"        { $currentList = "pd"; continue }
        default {
```

```
            if ($line.Trim() -ne "") {
                switch ($currentList) {
                    "shared" { $shared += $line }
                    "pddm"   { $pddm += $line }
                    "pd"     { $pd += $line }
                }
            }
        }
    }
}

# Save each group to a file
$shared | Set-Content "genes_shared.txt"
$pddm   | Set-Content "genes_pddm_unique.txt"
$pd     | Set-Content "genes_pd_unique.txt"
```

Figure 32: *The PowerShell command for splitting the cleaned venn diagram gene list into their 3 correspondance groups.*

i-Visualization using UpsetR in Rstudio

Principle: The R and Rstudio was already installed in our local machine downloaded from CRAN's official website https://cran.r-project.org/. With the power and integration of UpsetR; a super powerful open-source statistical programming language in bioinformatics (Conway et al., 2017), we analyzed data, visualize and interpret with its user-friendly interface for coding, plotting, and managing packages (**A, B, C**).

Process:

Firstly, we ran the following command to install UpsetR in Rstudio:

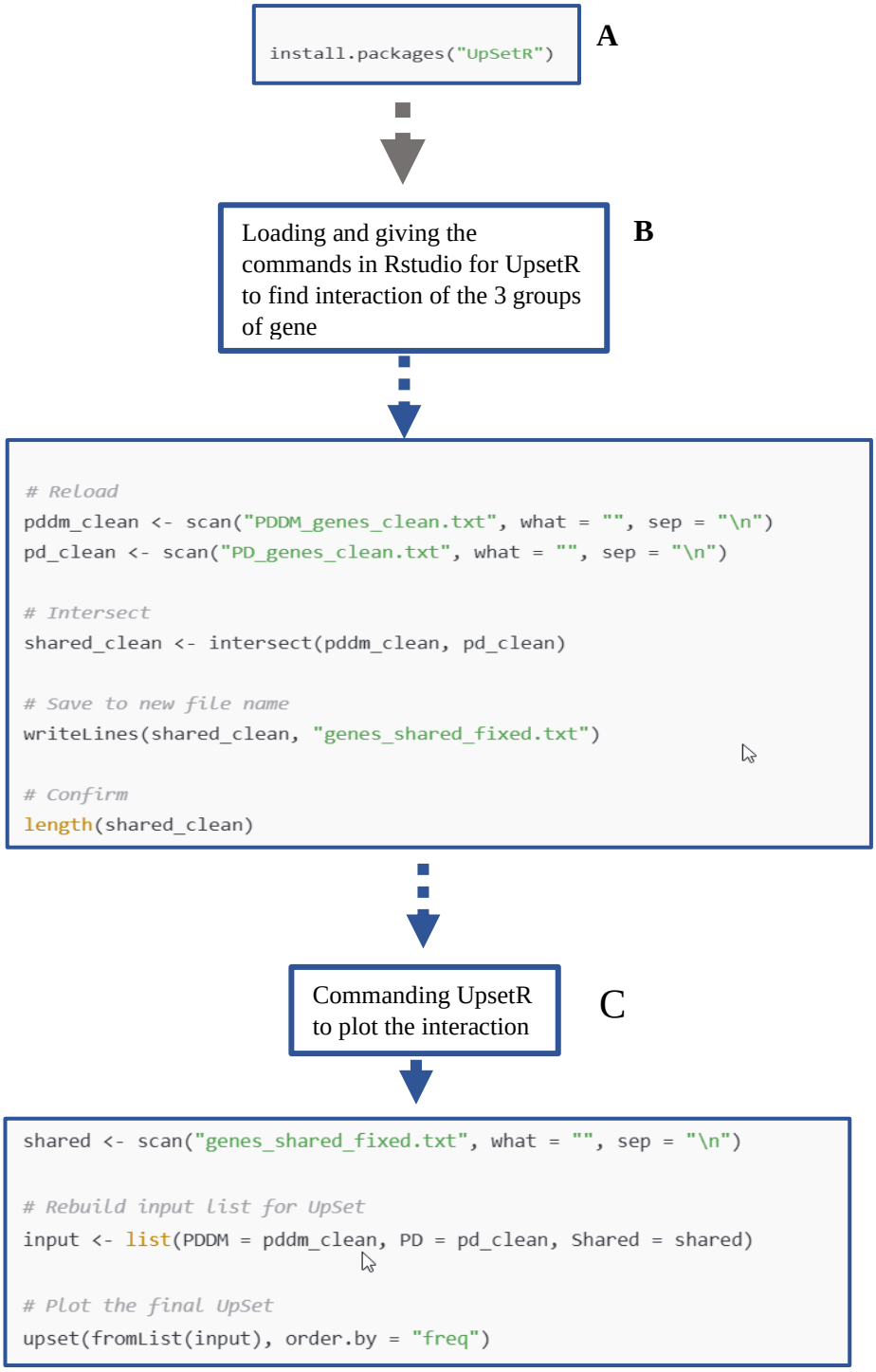
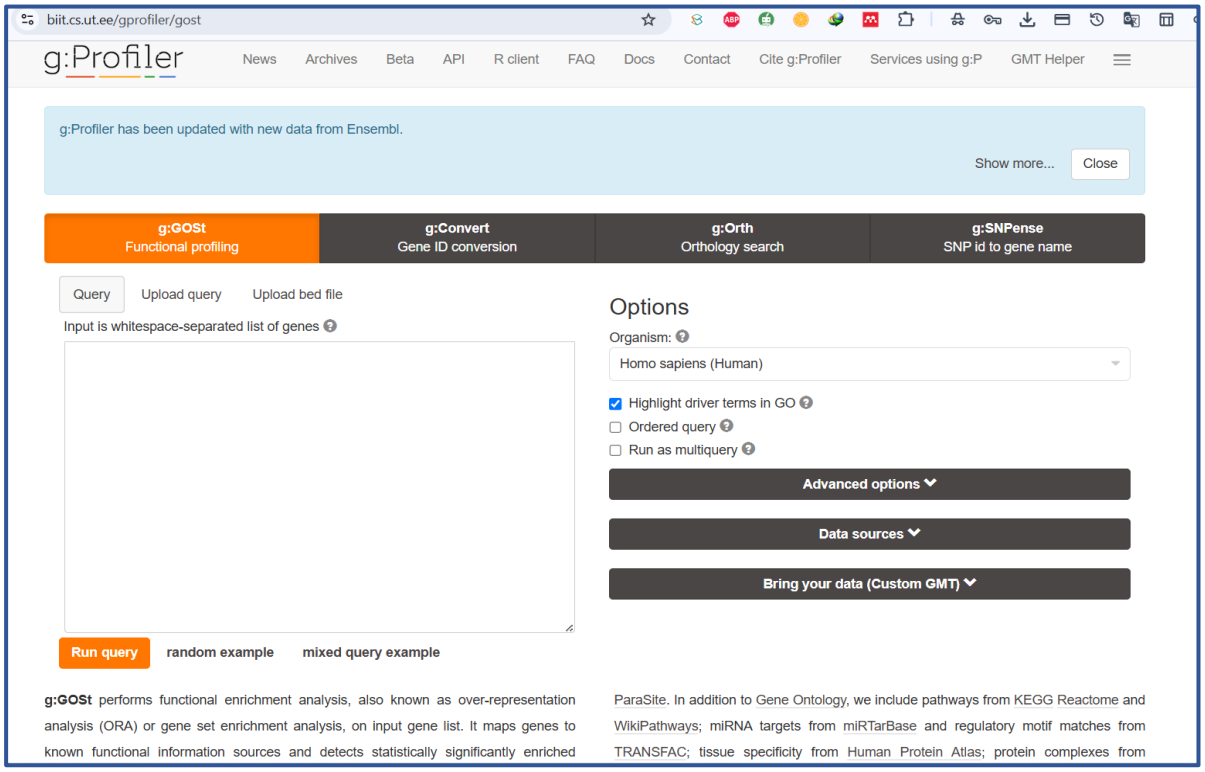


Figure 33: *The UpsetR workflow in Rstudio to find the interaction between the 3 groups of genes (PDDM, PD only and Shared).*

k- Functional interpretation with gProfiler

Principle: A functional enrichment analysis was conducted using g:Profiler, https://biit.cs.ut.ee/gprofiler/gost ,(**Figure 34**), which is a web service that automatically detects statistically significant (overrepresented) Gene Ontology (GO) biological and biochemical pathways (KEGG, Reactome) and regulatory elements within gene lists (**Raudvere et al., 2019**).



c

***Figure 34**:Web interface of the g: Profiler.*

Process:

The input data consisted of gene sets obtained from the cleaned Shared, PDDM-unique, and PD-unique VCF files in .txt format. The analysis was run with default settings granting FDR correction using the Benjamini, Hochberg method, and the outcomes were saved in TSV files for further analysis and visualization.

a) **Input gene lists** (shared, PDDM-only, PD-only) were uploaded in the query after being cleaned using PowerShell.
b) Used **default settings**, including:
    a. Benjamini–Hochberg FDR correction
    b. Only annotated genes for statistical scope.
    c. Highlight driver GO terms.

c) Ran the query and **exported results** as TSV for interpretation.

d) Identified key enriched pathways and GO terms relevant to diabetes and periodontitis, such as:

   a. Biological regulation,
   b. Signal transduction,
   c. Cell communication

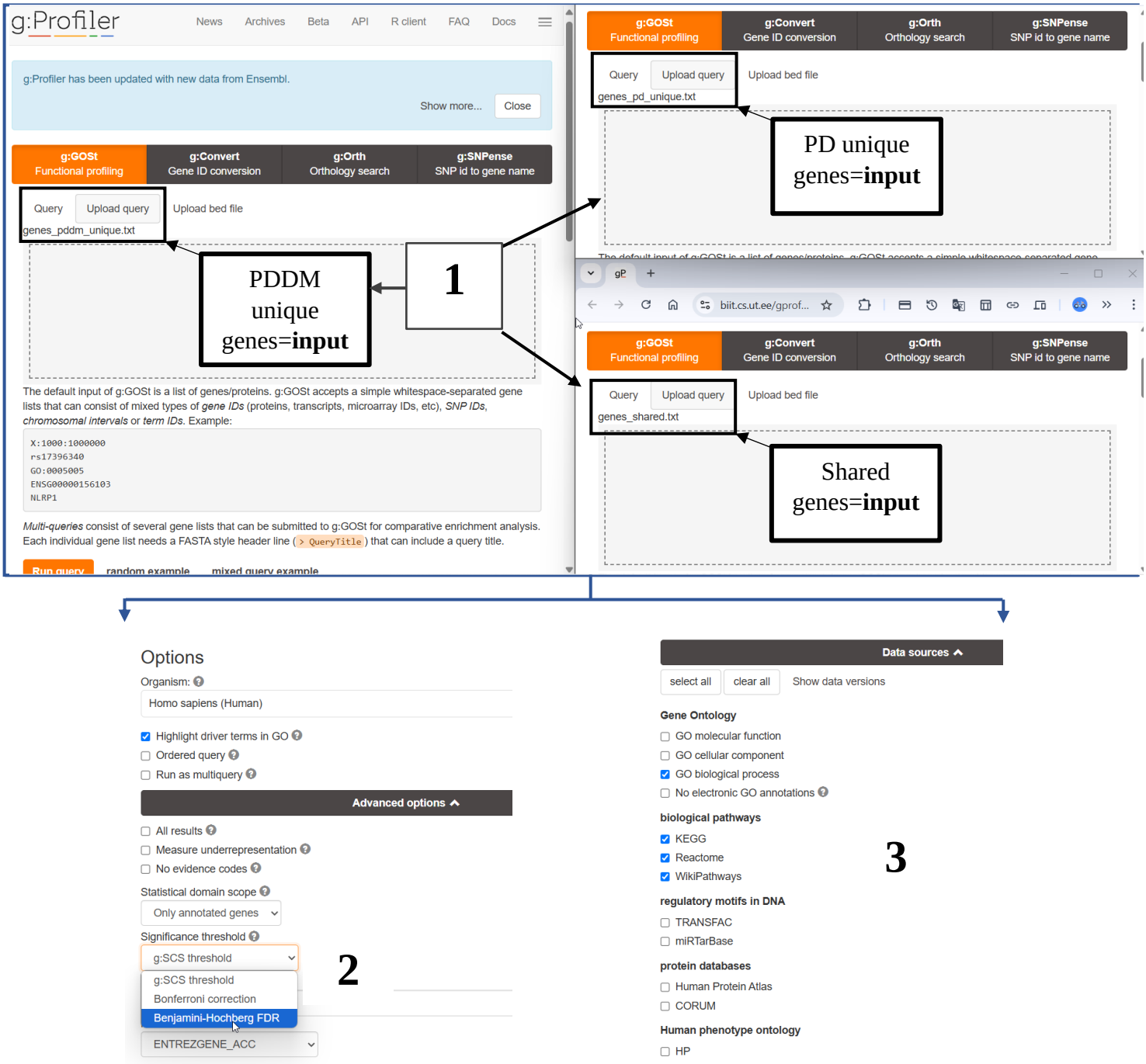


Figure 35: *The g: Profiler genomic analysis workflow.*

RESULTS

Chapter 3: Results

3.1. Results of the FASTQC on PD dataset

A preliminary FastQC assessment of the periodontitis dataset (SRR29729004) confirmed that both forward and reverse reads exhibited high sequencing quality, with the majority of base positions scoring above Q30. As typical with Illumina data, a slight quality decline was noted at the 3′end of the reverse **Appendix B**-(**Top)** and **Appendix C**-(**Top).** To address this, Trimmomatic was applied to remove adapters and low-quality bases. Post-trimming FastQC analysis demonstrated a clear improvement in read quality across all positions, with the majority of bases falling within the high-quality zone and no sequences flagged **Appendix B**-(**Bottom**) and **Appendix C**-(**Bottom**).

(**Figure 36**)summarizes the basic statistics for forward reads before and after trimming. The total number of bases decreased from 53.8 Gbp to 49.2 Gbp, indicating that approximately 4.6 Gbp of low-quality or adapter-contaminated bases were trimmed. The GC content slightly increased from 42% to 43%, reflecting the removal of biased regions. These results confirm the effectiveness of the pre-processing step and validate the dataset's readiness for downstream analysis.

3.1.1. Results of FASTQC on PDDM dataset

Basic Statistics

Measure	Value
Filename	forward.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	359060063
Total Bases	53.8 Gbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	42

Basic Statistics

Value	Measure
forward.gz	Filename
Conventional base calls	File type
Sanger / Illumina 1.9	Encoding
336229430	Total Sequences
49.2 Gbp	Total Bases
0	Sequences flagged as poor quality
30-150	Sequence length
43	%GC

***Figure 36a**: Basic statistics for PD dataset (forward reads) before (left) and after (right) trimming.*

Basic Statistics

Measure	Value
Filename	reverse.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	359060063
Total Bases	53.8 Gbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	43

Basic Statistics

Value	Measure
reverse.gz	Filename
Conventional base calls	File type
Sanger / Illumina 1.9	Encoding
336229430	Total Sequences
44.7 Gbp	Total Bases
0	Sequences flagged as poor quality
30-150	Sequence length
42	%GC

***Figure 37b**: Basic statistics for PD dataset (reverse reads) before (left) and after (right) trimming.*

Initial FastQC analysis of the forward reads from the Periodontitis with Diabetes Mellitus (PDDM) dataset revealed high-quality sequencing data, with most base positions showing Phred scores above Q30 **Appendix D**-(**Top**) and **Appendix E**-(**Top**). A minor decline in base quality was observed toward the 3′ end, a common feature in Illumina sequencing. Trimmomatic was used to trim low-quality bases and remove adapter sequences. The post-trimming FastQC analysis confirmed enhanced base quality and the successful elimination of trailing low-quality segments **Appendix D**-(**Bottom**) and **Appendix E**-(**Bottom**).

As shown in (**Figure 38**), the total number of bases decreased from 56.6 Gbp to 53.1 Gbp after trimming, indicating the removal of ~3.5 Gbp of low-quality or adapter-contaminated sequences (≈6.2%). The GC content decreased slightly from 41% to 40%. These results support the effectiveness of the pre-processing step in improving data quality for downstream genomic analysis.

Basic Statistics

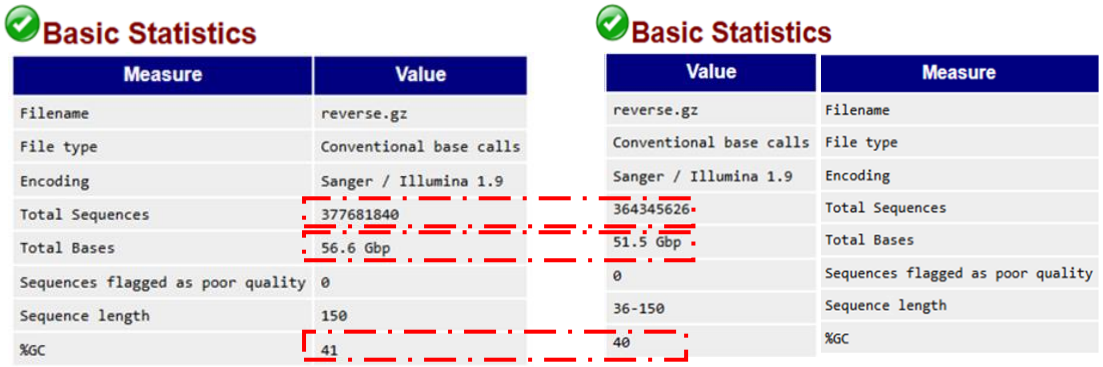
Measure	Value
Filename	forward.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	377681840
Total Bases	56.6 Gbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	41

Basic Statistics

Value	Measure
forward.gz	Filename
Conventional base calls	File type
Sanger / Illumina 1.9	Encoding
360282086	Total Sequences
53.1 Gbp	Total Bases
0	Sequences flagged as poor quality
50-150	Sequence length
40	%GC

***Figure 38**: Basic statistics for PDDM dataset (forward reads) before (left) and after (right) trimming.*

The reverse reads for the PDDM dataset were subjected to the same processing pipeline and displayed similar quality trends, as reflected in the corresponding FastQC results and **Figure 39**.



Basic Statistics

Measure	Value
Filename	reverse.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	377681840
Total Bases	56.6 Gbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	41

Basic Statistics

Value	Measure
reverse.gz	Filename
Conventional base calls	File type
Sanger / Illumina 1.9	Encoding
364345626	Total Sequences
51.5 Gbp	Total Bases
0	Sequences flagged as poor quality
36-150	Sequence length
40	%GC

***Figure 39**: Basic statistics for PDDM dataset (forward reads) before (left) and after (right) trimming.*

3.2. Results of the BWA-MEM alignment tool and deduplication (MarkDuplicates)

3.2.1. PD and PDDM dataset BWA-MEM results

A condensed summary of key alignment metrics is shown in **Table IV**, while additional plots and metrics are detailed in **Appendix F**.

This dataset also demonstrated high-quality alignment:

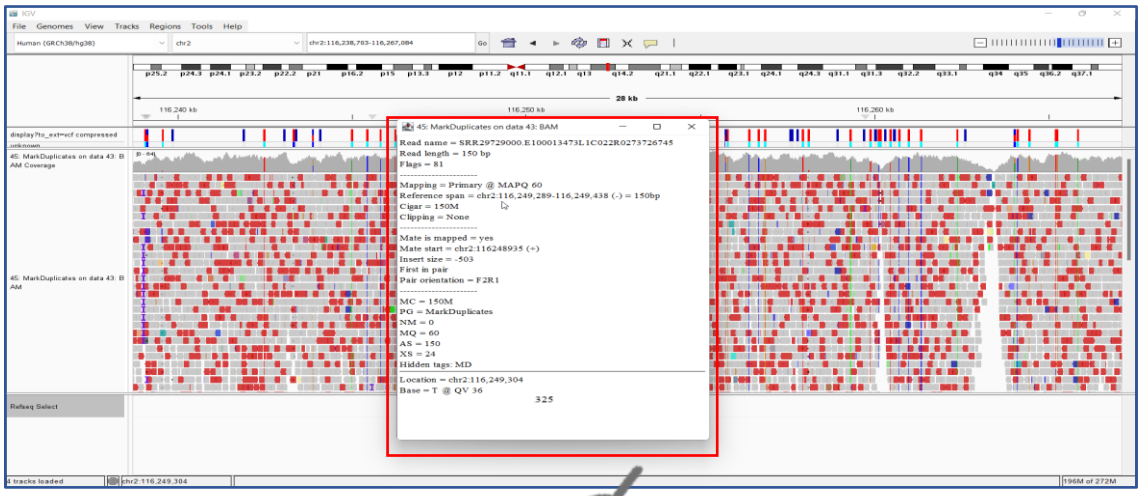
***Table IV**: BWA-MEM alignment results.*

Metric	PD Dataset	PDDM Dataset
Total input reads	679278372	730755529
Mapped reads (%)	99.5%	99.74%
Properly paired reads (%)	62.81%	69.35%
Duplicate reads (%)	0	0
Read length	150 bp	150 bp
Mapping quality (MAPQ)	60 (high)	60 (high)
Reference genome	GRCh38/hg38	GRCh38/hg38
Visual confirmation tool	IGV (Integrative Genomics Viewer + vcf.iobio.io)	IGV (Integrative Genomics Viewer + vcf.iobio.io)

As visualized in **Appendix F**, chromosomal alignment coverage was well distributed across all GRCh38 chromosomes. The balance between forward and reverse strand alignment was nearly **50:50**, indicating proper orientation.

In **(Figure 40-42)**, the mapping quality distribution is illustrated. The majority of reads have a MAPQ score of 60, signifying uniquely and confidently mapped reads. A smaller proportion maps with low confidence (NM = 0), possibly reflecting reads that map to repetitive or random regions. The BWA-MEM score is represented below at (**Figure 40**) with the colour coding.

IGV (Integrative Genomics Viewer) ***https://igv.org/*** ***(*****Figure 40** *and* **Figure 41*)*** was used to visually validate alignment quality and detect variant-rich regions. It enables real-time inspection of read coverage, mismatch profiles, and mapping quality across reference loci **(Robinson et al., 2011)** .



45: MarkDuplicates on data 43: BAM

Read name = SRR29729000.E100013473L1C022R0273726745
Read length = 150 bp
Flags = 81

Mapping = Primary @ MAPQ 60
Reference span = chr2:116,249,289-116,249,438 (-) = 150bp
Cigar = 150M
Clipping = None

Mate is mapped = yes
Mate start = chr2:116248935 (+)
Insert size = -503
First in pair
Pair orientation = F2R1

MC = 150M
PG = MarkDuplicates
NM = 0
MQ = 60
AS = 150
XS = 24
Hidden tags: MD

Location = chr2:116,249,325
Base = C @ QV 36

Color panel

Gray = Match to reference
Red = Mismatch
Blue = Indel

Figure 40*: IGV detailed view of an individual read alignment PD on chromosome 2 (GRCh38/hg38), (SRR29729000) on position 150 bp in length and mapped as a primary alignment with a mapping quality (MAPQ) of 60.*

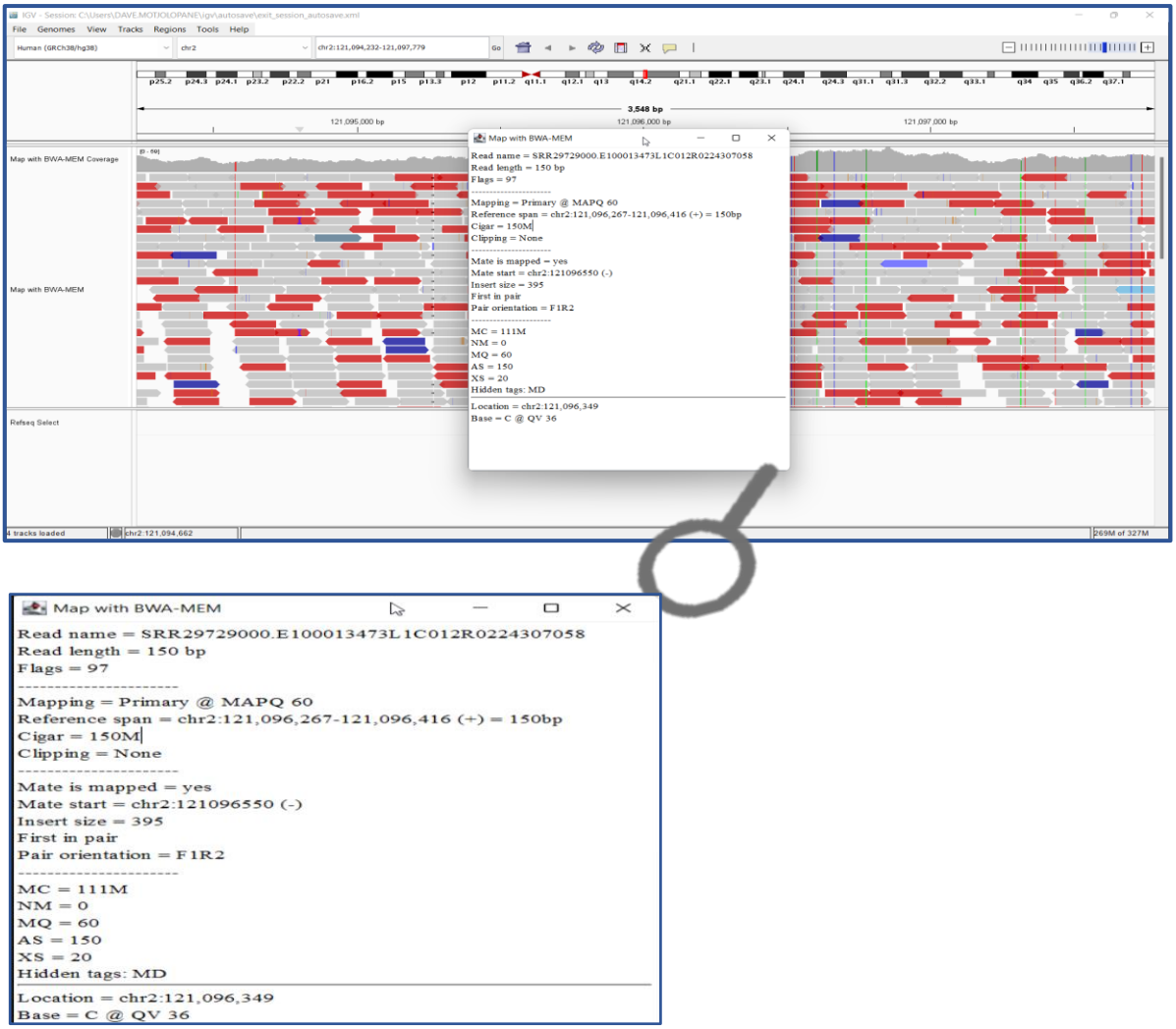


***Figure 41**: IGV visualization of PDDM read alignments to GRCh38 at chr2:121,096,267–121,096,416. The selected read is 150 bp, mapped with MAPQ 60, and paired correctly. Red and grey bars indicate aligned reads with minimal mismatches.*

3.2.2. Variant calling results

The quality score distribution of our variants shows a clear shift toward higher values, with most variants scoring well above 100. As shown in the first plot, the average quality is around 600, and many variants fall between 500 and 900 (**Figure 42**). This indicates that the majority of variants are high confidence with majority of SNPs distributed across all chromosomes (The transition/transversion (Ts/Tv ratio was **2.22 and 2.20** respectively, **Appendix F**) indicating high-quality variant calls, which is important for reliable downstream analysis. Similar thresholds (above 30–50, or 100 for stricter filters) have been used in other studies to reduce false positives **(Depristo et al., 2011; Garrison & Marth, 2012b; Van der Auwera et al., 2013)**.

The second plot shows that most of the variants are SNPs, while insertions, deletions, and others make up only a small portion. This is typical in variant calling, and combined with the strong quality scores, suggests that our variant calls are solid and ready for further analysis.

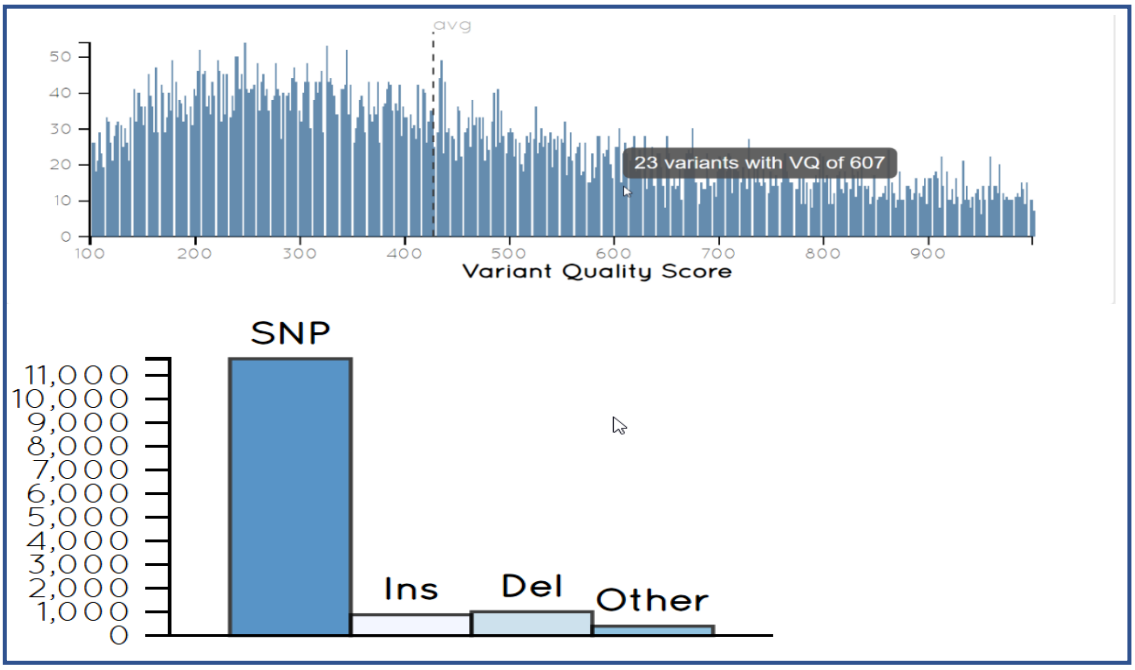


Figure 42: *Variant calling results on samples (PD, PDDM).*

3.2.3. Results of variants annotation

In PD dataset **(A)**, a total of **9,446,021 variants** were processed from 9,446,254 entries, with **19,782,199 effects** identified and a variant rate of **1 per 338 bases**. In PDDM dataset **(B)**, **10,863,256 variants** were processed out of 10,863,515, with **23,068,177 effects** and a variant rate of **1 per 294 bases**. Both runs reported **zero errors** and over **100,000 warnings**, with **no known variants (ID = 0)** and over **180,000 multi-allelic entries** in each case. These results confirm successful annotation and a high density of identified variants across the genome in both datasets.

***Table V**: Summary of SnpEff annotation results for PD dataset (A) and PDDM dataset (B).*

Summary — **A**

Genome	hg38
Date	2025-05-22 12:46
SnpEff version	SnpEff 5.2 (build 2023-09-29 06:17), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /corral4/main/jobs/067/839/67839377/outputs/dataset_4aa6734c-c06b-4098-ba5f-163dad60fcc1.dat hg38 /corral4/main/objects/3/4/9/dataset_34950d43-750c-48a9-bdfe-bfbe832a7985.dat
Warnings	100,004
Errors	0
Number of lines (input file)	9,159,503
Number of variants (before filter)	9,446,254
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	9,446,021
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	189,427
Number of effects	19,782,199
Genome total length	3,247,448,939
Genome effective length	3,199,083,047
Variant rate	1 variant every 338 bases

Summary — **B**

Genome	hg38
Date	2025-05-22 22:27
SnpEff version	SnpEff 5.2 (build 2023-09-29 06:17), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /data/jwd07/main/083/842/83842270/outputs/dataset_5a07f64d-f9cd-4e9a-9c38-8834b0e8705a.dat hg38 /data/dnb11/galaxy_db/files/a/0/c/dataset_a0cd945c-6fb1-4038-aa79-038cc8c38934.dat
Warnings	100,051
Errors	0
Number of lines (input file)	10,641,832
Number of variants (before filter)	10,863,515
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	10,863,256
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	201,520
Number of effects	23,068,177
Genome total length	3,247,448,939
Genome effective length	3,200,555,980
Variant rate	1 variant every 294 bases

3.2.4. UCSC genome browser report

Visual inspection of the *MCUR1*, *FOS*, and *RAP2A* gene regions was conducted using the UCSC Genome Browser, with custom tracks displaying filtered PD and PDDM VCF datasets overlaid on the common dbSNP track (build 155, hg38). For each gene, variants were selected based on clear vertical alignment between sample stacks and labelled dbSNP entries.

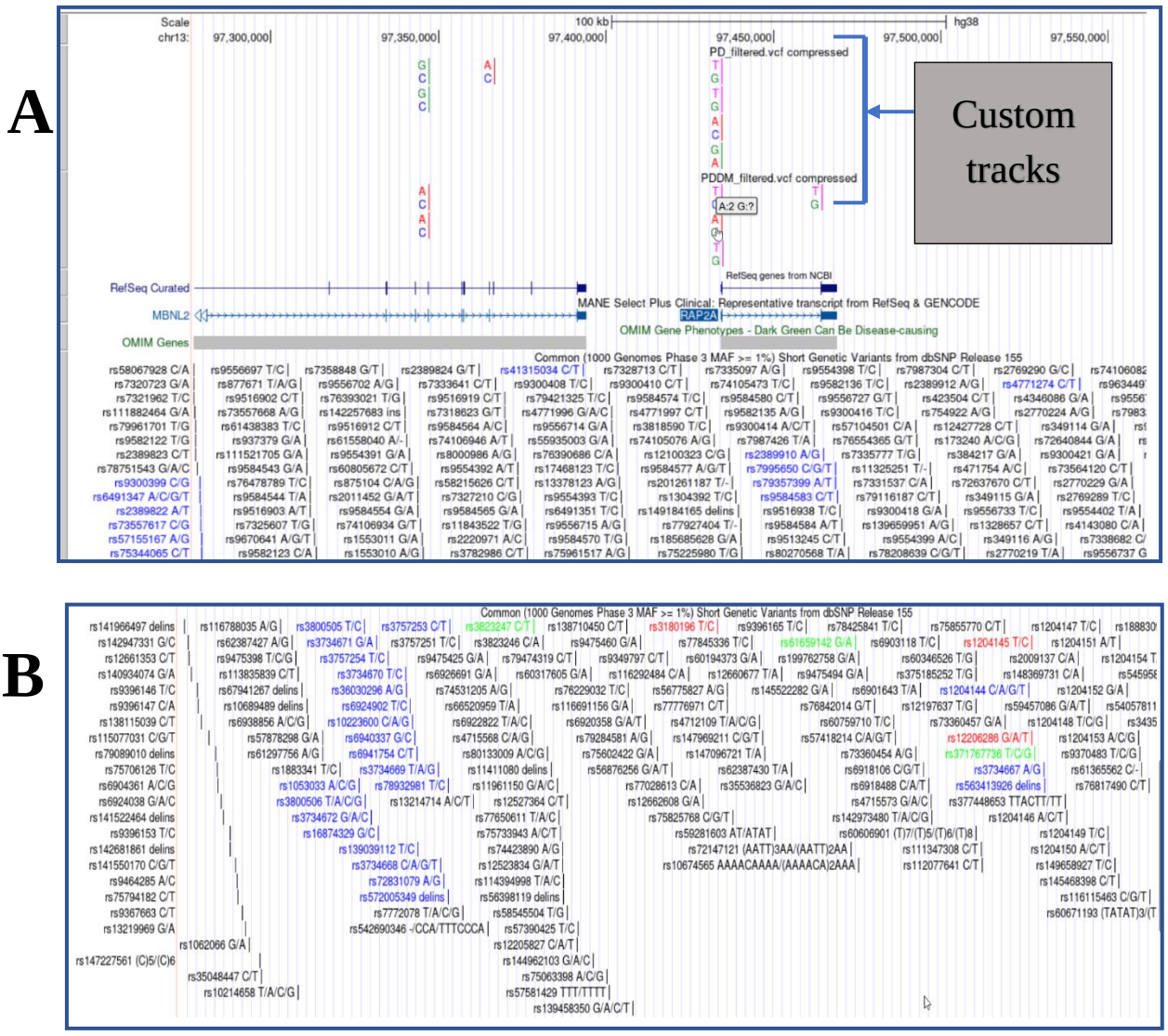


***Figure 43**: FOS (A) and MCUR1 (B) UCSC visuals on common SNPs aligned with samples (PD, PDDM).*

A total of **7** SNPs were confirmed in MCUR1 Error! Reference source not found., including a known missense variant (**rs3180196**) and several intronic or UTR variants overlapping transcript regions. In *FOS*, **7** SNPs were observed **Table VII**, primarily clustered in the 5′ UTR and synonymous exonic sites, all shared between PD and PDDM datasets. For *RAP2A*, **6** clearly aligned SNPs, **Table VIII** were selected from within intronic and promoter-flanking regions; one additional SNP (rs9289822) was excluded due to chromosomal mismatch. These UCSC-confirmed variants were used for downstream functional annotation. Refer to **Figure 43 (A)** for full visualization.

***Table VI**: UCSC genome browser results for MCUR1 (combined samples: PD and PDDM).*

dbSNP ID	Nucleotide Change	Justification
rs3180196	T>C	Seen in red (ClinVar), overlaps PDDM and PD
rs3757254	T/C	Found in transcript, previously used
rs3734670	T/C	Within annotated region
rs3800507	T/C	Matches filtered SNP
rs78932981	T/C	Clearly visible, overlaps gene
rs78425841	T>C	Confirmed on right-hand side of gene track
rs60759710	T>C	Mid-gene, visible and close to variant stack

***Table VII**: UCSC genome browser result for FOS (PD and PDDM).*

dbSNP ID	Position (hg38)
rs4645853	chr14:75279191
rs2239615	chr14:75278848
rs7101	chr14:75278923
rs76195634	chr14:75279924
rs4645855	chr14:75280166
rs1046117	chr14:75279987
rs2229024	chr14:75281127

***Table VIII**: UCSC genome browser result for RAP2A (PD and PDDM)*

dbSNP ID (Chr13)	Nucleotide Change
rs57155167	A>G
rs9300399	C>G
rs73557617	C>G
rs75344065	C>T
rs12016692	C>T
rs12019957	T>C
rs9289822	T>C (excluded due to chr3)

3.2.5. Pathogenic assessment results (SNPnexus)

- ***MCUR1 SNPnexus***

The SNPnexus annotation for MCUR1 highlighted a key missense variant (rs3180196) that was consistently present in both PDDM and PD datasets and confirmed by UCSC Genome Browser alignment. This variant caused an amino acid substitution and was annotated across all major transcript models (RefSeq, Ensembl, UCSC). PolyPhen predicted it as tolerated, and no pathogenic annotations were found in ClinVar or OMIM.

Among the non-coding variants, rs3800507 stood out with a CADD PHRED score of 7.11 and a FunSeq2 score of 0.797, the highest among all MCUR1 variants. This SNP also overlapped active regulatory regions (ENCODE datasets: K562, IMR90, etc.). Other variants such as rs60759710 and rs78425841 showed lower scores and were located in intronic regions, with no functional annotation reported.

***Table IX**: SNPnexus results on MCUR1(PD, PDDM).*

SNP ID	Consequence	CADD PHRED	FunSeq2 Score	PolyPhen	Regulatory Notes
rs3180196	Missense (T>A, pos 216)	15.68	–	Tolerated (0.360)	CDS of all transcript models
rs3757254	3′UTR / downstream	3.51	0.011	–	Found in 3′UTR (UCSC, RefSeq)
rs3734670	3′UTR / downstream	0.245	0.011	–	Located downstream
rs3800507	Promoter flank	7.11	0.797	–	Overlaps active ENCODE elements (K562 etc.)
rs60759710	Intronic	2.01	0.167	–	Non-coding, low impact
rs78425841	Intronic	4.73	0.011	–	Mid-gene, non-coding

➢ *FOS SNPnexus*

All seven SNPs were non-coding:

- **rs7101** and **rs2239615** were located in the **5′ UTR**
- **rs1046117**, **rs76195634**, and **rs2229024** were **synonymous variants**
- No SNPs scored high on CADD or FunSeq2

These were consistently annotated across RefSeq and Ensembl and confirmed by UCSC tracks, with green-labeled exonic matches suggesting non-pathogenicity.

***Table X**: SNPnexus results on FOS (PD, PDDM).*

SNP ID	Location	CADD PHRED	FunSeq2 Score	Effect Type	Regulatory Notes
rs2239615	5′UTR	Low	–	UTR (non-coding)	Found in Ensembl + RefSeq
rs7101	5′UTR	Low	–	UTR (non-coding)	RegElement overlap
rs1046117	Synonymous exon	–	–	Coding (no AA change)	Green-labeled (benign)
rs76195634	Synonymous exon	–	–	Coding (no AA change)	Seen in UCSC exon + variant stack
rs4645853	Intronic	–	–	Deep intronic	Not linked to regulatory sites
rs4645855	Downstream	–	–	Intergenic	No pathogenic annotation
rs2229024	Synonymous exon	–	–	Coding (silent)	CDS-labeled, green in UCSC

➢ *RAP2A SNPnexus*

Six SNPs were retained. The standout variant was:

- **rs12019957**, located in a **promoter-flanking region**
    - **CADD = 10.09**
    - **FunSeq2 = 0.167**
    - Overlapped regulatory chromatin features (histone marks)
    - Documented **GWAS association** with age at menarche

Another variant, **rs73557617**, had a **FunSeq2 score = 0.79**, supporting regulatory impact.

***Table XI**. SNPnexus Results on RAP2A (PD, PDDM).*

SNP ID	Region	CADD PHRED	FunSeq2 Score	GWAS / ClinVar	Regulatory Notes
rs57155167	Intronic	3.017	0.172	–	Intron of MBNL2
rs9300399	Intronic	0.832	0.167	–	Weak regulatory support
rs73557617	Intronic	1.406	0.790	–	High FunSeq2, retained
rs75344065	Intronic	1.240	0.167	–	Low score, visually aligned
rs12016692	Promoter/Histone	0.535	0.178	–	Histone mark and promoter flank
rs12019957	Promoter/Histone	10.090	0.167	**GWAS: Menarche (p = 9E-9)**	Regulatory + disease link

3.3. Results of bioinformatic Venn diagram and genomic profiling using g: profiler.

3.3.1. Shared and Unique Genes Between PDDM and PD

After filtering for HIGH and MODERATE impact variants using SnpSift **Table III** , we compared gene lists from the PDDM and PD datasets with a Venn diagram. This revealed unique genes in each group and a large set of shared genes; including MCUR1, FOS, and RAP2A, which were selected for deeper analysis. For gene list please refer to **Appendix J**.

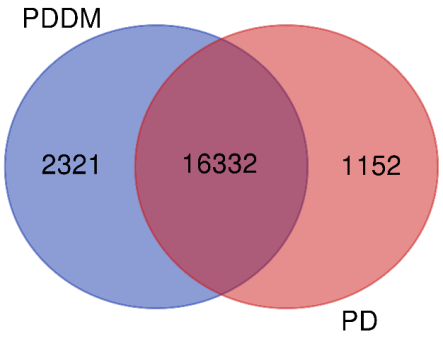


***Figure 44:** Venn diagram bioinformatic representation of intersection of PD and PDDM.*

3.3.2. g: Profiler results on PD sample

To understand the biological relevance of PD-specific genes, we conducted enrichment analysis using g: Profiler through different pathways such as **KEGG**. The analysis revealed that these genes are significantly enriched in **tissue remodeling**, **biological regulation**, and **metabolic processes**, including *organelle organization*, *apoptotic pathways*, and *DNA metabolic process*. Notably, key molecular functions such as **protein binding**, **catalytic activity**, and **molecular carrier activity** were also overrepresented.

Figure 45 shows **e**ach dot represents a GO term coloured by domain, with circle size proportional to gene set size and vertical axis denoting statistical significance (**–log₁₀ *p*-value**). The accompanying table summarizes enriched terms with adjusted *p*-values and source classifications.

ID	Source	Term ID	Term Name	p_{adj} (query...
1	GO:MF	GO:0005515	protein binding	4.747×10^{-57}
2	GO:MF	GO:0003824	catalytic activity	1.390×10^{-5}
3	GO:MF	GO:0140104	molecular carrier activity	4.466×10^{-2}
4	GO:BP	GO:0065007	biological regulation	4.139×10^{-37}
5	GO:BP	GO:0006996	organelle organization	5.001×10^{-9}
6	GO:BP	GO:0006793	phosphorus metabolic process	1.155×10^{-8}
7	GO:BP	GO:1901135	carbohydrate derivative metabolic process	6.976×10^{-5}
8	GO:BP	GO:0006915	apoptotic process	9.116×10^{-5}
9	GO:BP	GO:0042592	homeostatic process	8.888×10^{-4}
10	GO:BP	GO:0042254	ribosome biogenesis	1.677×10^{-3}
11	GO:BP	GO:0006259	DNA metabolic process	5.127×10^{-3}
12	GO:BP	GO:0072521	purine-containing compound metabolic process	1.255×10^{-2}
13	GO:BP	GO:0051168	nuclear export	3.115×10^{-2}
14	GO:CC	GO:0005737	cytoplasm	3.496×10^{-69}
15	GO:CC	GO:1902494	catalytic complex	1.059×10^{-8}
16	GO:CC	GO:0140535	intracellular protein-containing complex	2.481×10^{-5}
17	GO:CC	GO:1904813	ficolin-1-rich granule lumen	8.216×10^{-3}
18	GO:CC	GO:0005813	centrosome	4.551×10^{-2}

version	e112_eg59_p19_25aa4782
date	6/7/2025, 3:42:27 PM
organism	hsapiens

g:Profiler

Figure 45: g :Profiler enrichment map and statistical table illustrating top Gene Ontology (**GO**) terms across Molecular Function (**MF**), Biological Process (**BP**), and Cellular Component (CC) categories for PD-specific genes.

3.3.3. g: Profiler results on PDDM sample

The analysis identified significant enrichment in Gene Ontology Biological Process (GO:BP) categories, with top terms including **biological regulation**, **regulation of biological process**, and **multicellular organismal process**. Additional enriched pathways involved **response to stimulus**, **signal transduction**, **cell differentiation**, and **positive regulation of metabolic processes**. These results indicate that the PDDM-associated genes are predominantly linked to regulatory and signaling mechanisms across immune, metabolic, and cellular communication pathways.

In Figure 47, the plot visualizes the **$-\log_{10}$ *p*-values** of enriched terms, while the table highlights top-ranked biological processes based on adjusted statistical significance, revealing functional clustering in regulation, signalling, and differentiation pathways.

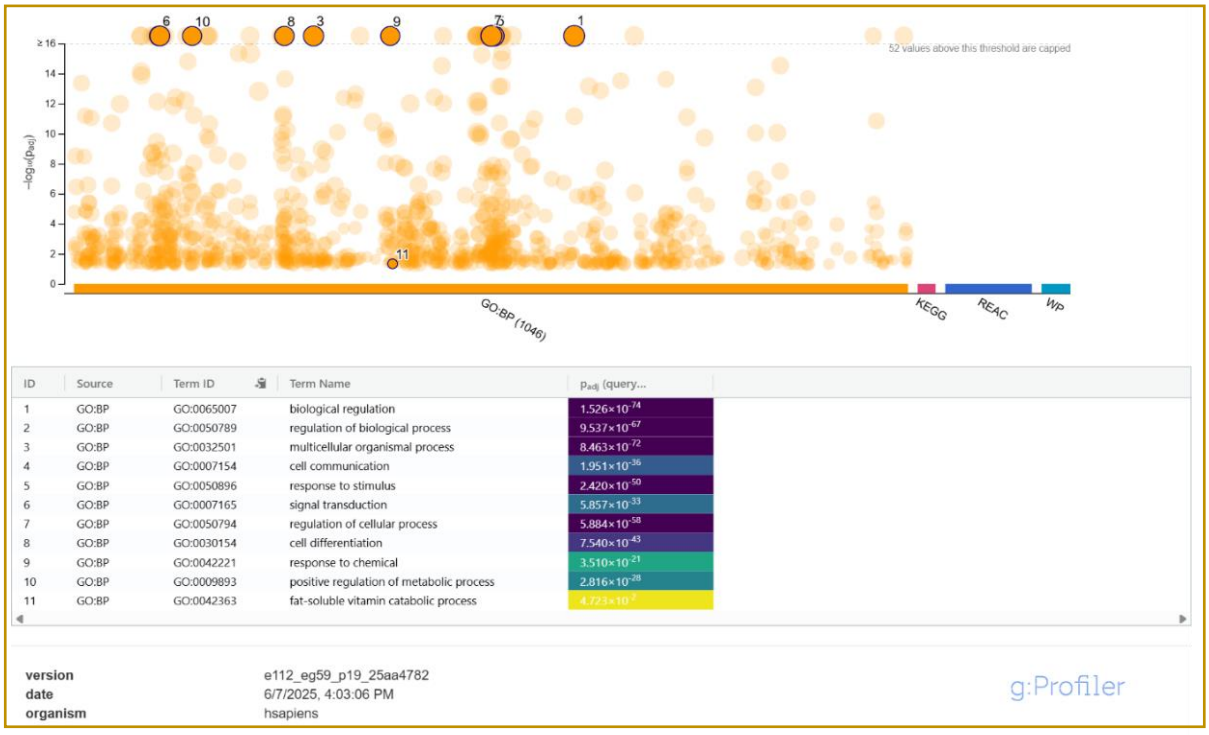


Figure 46: Dot plot and enrichment summary from g: Profiler showing the distribution of significantly enriched Gene Ontology Biological Process (**GO:BP**) terms among PDDM-specific genes.

DISCUSSION

Chapter 4: DISCUSSION

This study aimed to identify and compare SNPs (variants), more precisely, pathogenic variants that links or relates individuals with periodontitis (PD) and those with periodontitis complicated by type 2 diabetes mellitus (PDDM), focusing on some highly cited genes, MCUR1, FOS, and RAP2A, to genetically study these two diseases.

Variant calling via FreeBayes, followed by SnpEff and SNPnexus annotation, revealed group-specific and shared SNPs with variable predicted functional effects. Importantly, we observed that PDDM-specific variants were more often associated with regulatory features and higher predicted pathogenicity, as measured by CADD and FunSeq2 scores.

In MCUR1, we identified **rs3810196,** a non-coding SNP located downstream of the gene and exclusive to the PDDM group. This variant overlapped active transcriptional marks in UCSC and carried a CADD **score of 15.68** with FunSeq2 classification as 'regulatory'.

MCUR1 encodes a component of the mitochondrial calcium uniporter complex, which governs calcium signaling and redox balance. Disruption of this process can induce oxidative stress, mitochondrial permeability changes, and ATP depletion, particularly under hyperglycemic conditions. These mechanisms are well-documented in the context of diabetic tissue stress and align with our finding that rs3810196 may contribute to the exacerbation of periodontitis in diabetic individuals Click or tap here to enter text. **(Mallilankaraman et al., 2012).**

For FOS, several variants were observed in both datasets, but **rs1046117** and rs7101 were unique to PDDM and overlapped regulatory regions. Both SNPs lie within the 3′-UTR and show moderate CADD scores (around 12–14) and evidence of enhancer overlap in ENCODE data. FOS is part of the AP-1 transcription factor complex, a major regulator of inflammatory

cytokine genes. Following a study by **(Zhao et al., 2023)**, AP-1 activity is amplified under high glucose conditions, promoting IL-6, TNF-α, and bone-resorbing mediators. The location of these SNPs within regulatory UTRs may influence transcript stability or microRNA binding, thereby potentially inducing AP–1–driven inflammation in diabetic periodontitis **(Wei et al., 2024).**

In RAP2A, the PDDM group showed unique variants, including **rs6115228** and **rs2298771**, with both mapping to downstream regions annotated as active in regulatory chromatin states. The CADD score of **rs6115228** was **19.22**, indicating strong deleterious potential.

RAP2A is a small GTPase involved in actin cytoskeleton regulation. It is also involved in epithelial cell signaling. Its role in immune cell trafficking is especially relevant in barrier tissues such as the periodontium. A recent multi-omics analysis confirmed that RAP2A is differentially expressed in both diabetes and periodontitis models **(Wei et al., 2024b)**, suggesting that our observed variants may contribute to impaired epithelial integrity or exaggerated immune cell recruitment.

Beyond gene-specific effects, functional enrichment analysis using g: Profiler revealed that PDDM-specific genes were enriched in signal transduction (GO:0007165), response to stimulus (GO:0050896), and metabolic regulation (GO:0019222), contrasting with PD-only enrichment in tissue remodeling (GO:0048771) and organelle organization (GO:0006996). These findings support the interpretation that diabetes shifts the inflammatory profile of PD toward systemic dysfunction, not only amplifying local inflammation but reprogramming stress, signaling, and metabolic pathways **(Wei et al., 2024c)**.

The variant dataset was technically promising: a **Ts/Tv ratio of 2.22 (PD)** and **2.20 (PDDM)** indicated high biological quality, and SNPs were evenly distributed across chromosomes, including sex chromosomes, confirming the uniformity of variant detection.

Although this study is limited by its in-silico design, it provides compelling SNP-level evidence of genomic differentiation between PD and PDDM. Specific variants such as rs3810196 in MCUR1, rs1046117 in FOS, and rs6115228 in RAP2A show both functional prediction and biological relevance, making them strong candidates for future biomarker validation.

But still, to confirm these hypotheses, we may need to conduct additional studies such as eQTL analysis, allele-specific expression, or CRISPR/Cas9 mutagenesis, which could be experimentally beneficial.

CONCLUSION

While our study successfully met its objective of identifying group-specific SNPs and regulatory signatures, we observed that the variants found, although functionally relevant; did not always align with those most commonly reported in previous studies. This discrepancy may reflect differences in methodology, dataset origin, or population structure, but it also underlines the **unique contribution of our approach**, which prioritized regulatory and non-coding SNPs often overlooked in traditional association studies.

Notable findings included **rs3810196 (MCUR1)**, **rs1046117 (FOS)**, and **rs6115228 (RAP2A)**, all exclusive to the PDDM group and linked to regulatory chromatin regions.

Enrichment analysis further highlighted distinct biological themes: PD-specific genes were linked to structural tissue remodeling and apoptosis, while PDDM-specific genes were more involved in immune signaling and metabolic regulation. This supports the interpretation that diabetes modifies the genomic background of periodontitis, even if the relationship is subtler than expected based on prior literature.

The study was not without limitations. One key challenge was the **processing and cleaning of large genomic datasets**, which demanded meticulous attention at each stage to ensure accuracy and time consuming. Furthermore, as this work was entirely computational, experimental validation is required to confirm the true functional impact of the identified variants.

Finally, although our findings do not replicate all previously reported gene-disease links, they provide a valuable alternative view of the genomic landscape of PDDM. This work lays the groundwork for further investigation into the regulatory roles of these genes and offers a basis for future biomarker discovery or personalized intervention strategies in the context of diabetes-related periodontitis.

BIBLIOGRAPHY REFERENCES

1. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379
2. Alam, U., Asghar, O., Azmi, S., & Malik, R. A. (2014). General aspects of diabetes mellitus. In *Handbook of Clinical Neurology* (Vol. 126, pp. 211–222). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53480-4.00015-1
3. Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Baraja-Fonseca, V., Arrones, A., Vilanova, S., Plazas, M., Prohens, J., Bombarely, A., & Gramazio, P. (2024). *Benchmarking of Low Coverage Sequencing Workflows for Precision Genotyping in Eggplant*. https://doi.org/10.1101/2024.10.24.619843
5. Barutta, F., Bellini, S., Durazzo, M., & Gruden, G. (2022). Novel Insight into the Mechanisms of the Bidirectional Relationship between Diabetes and Periodontitis. In *Biomedicines* (Vol. 10, Issue 1). MDPI. https://doi.org/10.3390/biomedicines10010178
6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
7. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695
8. Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/BIOINFORMATICS/BTX364
9. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

10. Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., & Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research*, *46*, 109–113. https://doi.org/10.1093/nar/gky399
11. Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. https://doi.org/10.1038/ng.806
12. Garrison, E., & Marth, G. (2012a). *Haplotype-based variant detection from short-read sequencing*. http://arxiv.org/abs/1207.3907
13. Garrison, E., & Marth, G. (2012b). *Haplotype-based variant detection from short-read sequencing*.
14. GATK Team. (2025). *How to filter variants either with VQSR or by hard-filtering*. Broad Institute / GATK Documentation. https://gatk.broadinstitute.org/hc/en-us/articles/360035531112
15. Graves, D. T., Ding, Z., & Yang, Y. (2020). The impact of diabetes on periodontal diseases. In *Periodontology 2000* (Vol. 82, Issue 1, pp. 214–224). Blackwell Munksgaard. https://doi.org/10.1111/prd.12318
16. Guan, H., Zhao, S., Tan, Y., Fang, X., Zhang, Y., Zhang, Y., Miao, R., Yin, R., Yao, Y., & Tian, J. (2024). Microbiomic insights into the oral microbiome's role in type 2 diabetes mellitus: standardizing approaches for future advancements. In *Frontiers in Endocrinology* (Vol. 15). Frontiers Media SA. https://doi.org/10.3389/fendo.2024.1416611
17. Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., Taylor, J., & Nekrutenko, A. (2021). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research*, *48*(W1), W395–W402. https://doi.org/10.1093/NAR/GKAA434
18. Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., & Zhao, Z. (2012). Consensus rules in variant detection from next-generation sequencing data. *PLoS ONE*, *7*(6). https://doi.org/10.1371/journal.pone.0038470

19. Kang, J., Lee, H., Joo, J. Y., Song, J. M., Kim, H. J., Kim, Y. H., & Park, H. R. (2024). Comparison of genetic and epigenetic profiles of periodontitis according to the presence of type 2 diabetes. *MedComm*, *5*(7). https://doi.org/10.1002/mco2.620
20. Karcı, H., & Kafkas, S. (2024). Impact of Mark Duplicate Reads During Variant Calling in Next Generation Sequencing (NGS) Data of Pistacia vera L. *Ereğli Tarım Bilimleri Dergisi*. https://doi.org/10.54498/etbd.2024.29
21. King, H., Aubert, R. E., & Herman, W. H. (1998). Global Burden of Diabetes, 1995-2025 Prevalence, numerical estimates, and projections. In *Diabetes Care* (Vol. 21). http://diabetesjournals.org/care/article-pdf/21/9/1414/586598/21-9-1414.pdf
22. Lalla, E., & Papapanou, P. N. (2011). Diabetes mellitus and periodontitis: A tale of two common interrelated diseases. In *Nature Reviews Endocrinology* (Vol. 7, Issue 12, pp. 738–748). https://doi.org/10.1038/nrendo.2011.106
23. Lam, F., Lalansingh, C. M., Babaran, H. E., Wang, Z., Prokopec, S. D., Fox, N. S., & Boutros, P. C. (2016). VennDiagramWeb: A web application for the generation of highly customizable Venn and Euler diagrams. *BMC Bioinformatics*, *17*(1), 1–8. https://doi.org/10.1186/S12859-016-1281-5/TABLES/1
24. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509
25. Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. http://arxiv.org/abs/1303.3997
26. Lin, B., Fan, Y., Yang, X., Pathak, J. L., & Zhong, M. (2024). MMP-12 and Periodontitis: Unraveling the Molecular Pathways of Periodontal Tissue Destruction. In *Journal of Inflammation Research* (Vol. 17, pp. 7793–7806). Dove Medical Press Ltd. https://doi.org/10.2147/JIR.S480466
27. Liu, J., Zhang, B., Zhu, G., Liu, C., Wang, S., & Zhao, Z. (2023). Discovering genetic linkage between periodontitis and type 1 diabetes: A bioinformatics study. *Frontiers in Genetics*, *14*. https://doi.org/10.3389/fgene.2023.1147819
28. Mahesh, T. R., Kumar, D., Vinoth Kumar, V., Asghar, J., Mekcha Bazezew, B., Natarajan, R., & Vivek, V. (2022). Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/4451792
29. Mobasseri, M., Shirmohammadi, M., Amiri, T., Vahed, N., Fard, H. H., & Ghojazadeh, M. (2020). Prevalence and incidence of type 1 diabetes in the world: A systematic

review and meta-analysis. In *Health Promotion Perspectives* (Vol. 10, Issue 2, pp. 98–115). Tabriz University of Medical Sciences. https://doi.org/10.34172/hpp.2020.18

30. Ranbhise, J. S., Ju, S., Singh, M. K., Han, S., Akter, S., Ha, J., Choe, W., Kim, S. S., & Kang, I. (2025). Chronic Inflammation and Glycemic Control: Exploring the Bidirectional Link Between Periodontitis and Diabetes. In *Dentistry Journal* (Vol. 13, Issue 3). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/dj13030100
31. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/NAR/GKZ369.
32. Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011, January). *Interactive Genome Viewer*. Nature Biotechnology.
33. Silva, D. N. de A., Casarin, M., Monajemzadeh, S., Bezerra, B. de B., Lux, R., & Pirih, F. Q. (2022). The Microbiome in Periodontitis and Diabetes. *Frontiers in Oral Health*, *3*. https://doi.org/10.3389/froh.2022.859209.
34. Todd, J. A. (2010). Etiology of Type 1 Diabetes. In *Immunity* (Vol. 32, Issue 4, pp. 457–467). https://doi.org/10.1016/j.immuni.2010.04.001.
35. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *SUPL.43*. https://doi.org/10.1002/0471250953.bi1110s43
36. Wei, X., Zhang, X., Chen, R., Li, Y., Yang, Y., Deng, K., Cai, Z., Lai, H., & Shi, J. (2024). Impact of periodontitis on type 2 diabetes: a bioinformatic analysis. *BMC Oral Health*, *24*(1). https://doi.org/10.1186/s12903-024-04408-1
37. Wu, K. C. H., Liu, L., Xu, A., Chan, Y. H., & Cheung, B. M. Y. (2024a). Shared genetic architecture between periodontal disease and type 2 diabetes: a large scale genome-wide cross-trait analysis. *Endocrine*, *85*(2), 685–694. https://doi.org/10.1007/s12020-024-03766-8
38. Wu, K. C. H., Liu, L., Xu, A., Chan, Y. H., & Cheung, B. M. Y. (2024b). Shared genetic architecture between periodontal disease and type 2 diabetes: a large scale genome-wide cross-trait analysis. *Endocrine*, *85*(2), 685–694. https://doi.org/10.1007/s12020-024-03766-8

39. Yameny, A. A. (2024). Diabetes Mellitus Overview 2024. In *Journal of Bioscience and Applied Research* (Vol. 10, Issue 3, pp. 641–645). Society of Pathological Biochemistry and Hematology. https://doi.org/10.21608/jbaar.2024.382794
40. Zhao, M., Xie, Y., Gao, W., Li, C., Ye, Q., & Li, Y. (2023). Diabetes mellitus promotes susceptibility to periodontitis—novel insight into the molecular mechanisms. In *Frontiers in Endocrinology* (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fendo.2023.1192625
41. Zhou, B., Rayner, A. W., Gregg, E. W., Sheffer, K. E., Carrillo-Larco, R. M., Bennett, J. E., Shaw, J. E., Paciorek, C. J., Singleton, R. K., Barradas Pires, A., Stevens, G. A., Danaei, G., Lhoste, V. P., Phelps, N. H., Heap, R. A., Jain, L., D'Ailhaud De Brisis, Y., Galeazzi, A., Kengne, A. P., … Ezzati, M. (2024). Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. *The Lancet*, *404*(10467), 2077–2093. https://doi.org/10.1016/S0140-6736(24)02317-1
42. Zou, M., & Yang, J. (2024). Identify novel therapeutic targets for type II diabetes and periodontitis: insights from single-cell analysis and Mendelian randomization analysis. *Frontiers in Endocrinology*, *15*. https://doi.org/10.3389/fendo.2024.1410537.
43. Mallilankaraman, K., Cárdenas, C., Doonan, P. J., Chandramoorthy, H. C., Irrinki, K. M., Golenár, T., Csordás, G., Madireddi, P., Yang, J., Müller, M., Miller, R., Kolesar, J. E., Molgó, J., Kaufman, B., Hajnóczky, G., Foskett, J. K., & Madesh, M. (2012). MCUR1 is an essential component of mitochondrial Ca2+ uptake that regulates cellular metabolism. *Nature Cell Biology*, *14*(12), 1336–1343. https://doi.org/10.1038/ncb2622

Appendices

Appendix A

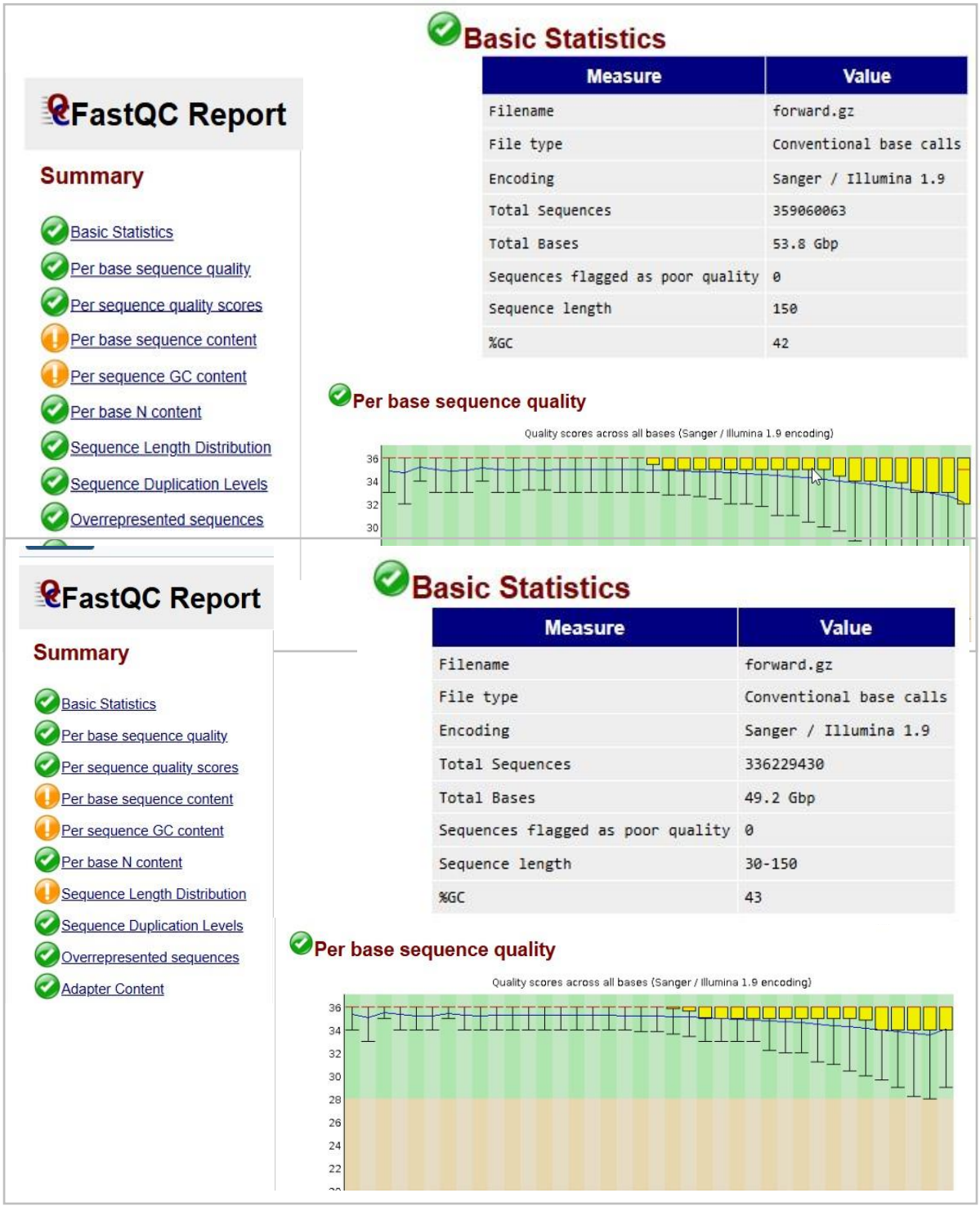
SRA Run selector tool results

Found 36 Items | Search within results | Clear

	Run	BioSample	AGE	AvgSpotLen	Bases	Bytes	Experiment	isolate	Library Name	LibraryLayout	ReleaseDate	create_date	Sample Name	sex
1	SRR29413843	SAMN41550707	64	300	115.73 G	65.73 Gb	SRX24927268	Healthy-6	Healthy6	PAIRED	2024-06-15	2024-06-14 15:11:00Z	Healthy-6	male
2	SRR29413844	SAMN41550706	20	300	105.45 G	59.69 Gb	SRX24927267	Healthy-5	Healthy5	PAIRED	2024-06-14	2024-06-14 13:51:00Z	Healthy-5	male
3	SRR29413845	SAMN41550705	33	300	115.55 G	65.76 Gb	SRX24927266	Healthy-4	Healthy4	PAIRED	2024-06-15	2024-06-14 15:08:00Z	Healthy-4	male
4	SRR29413846	SAMN41550704	20	300	101.71 G	57.38 Gb	SRX24927265	Healthy-3	Healthy3	PAIRED	2024-06-15	2024-06-14 14:18:00Z	Healthy-3	male
5	SRR29413847	SAMN41550716	68	300	118.88 G	67.97 Gb	SRX24927264	Periodontitis with DM-3	PDDM3	PAIRED	2024-06-15	2024-06-14 13:58:00Z	PDDM-3	male
6	SRR29413848	SAMN41550710	60	300	108.85 G	61.47 Gb	SRX24927263	Periodontitis-3	PD3	PAIRED	2024-06-15	2024-06-14 14:50:00Z	PD-3	female
7	SRR29413849	SAMN41550703	28	300	107.28 G	61.27 Gb	SRX24927262	Healthy-2	Healthy2	PAIRED	2024-06-14	2024-06-14 14:58:00Z	Healthy-2	male
8	SRR29413850	SAMN41550715	97	300	112.78 G	64.44 Gb	SRX24927261	Periodontitis with DM-1	PDDM1	PAIRED	2024-06-15	2024-06-14 14:56:00Z	PDDM-1	male
9	SRR29413851	SAMN41550718	72	300	105.33 G	59.49 Gb	SRX24927260	Periodontitis with DM-6	PDDM6	PAIRED	2024-06-15	2024-06-14 15:11:00Z	PDDM-6	female
10	SRR29413852	SAMN41550717	53	300	112.67 G	63.58 Gb	SRX24927259	Periodontitis with DM-4	PDDM4	PAIRED	2024-06-15	2024-06-14 15:12:00Z	PDDM-4	male
11	SRR29413853	SAMN41550714	63	300	104.96 G	60.76 Gb	SRX24927258	Periodontitis-7	PD7	PAIRED	2024-06-14	2024-06-14 13:33:00Z	PD-7	female
12	SRR29413854	SAMN41550713	47	300	98.39 G	58.04 Gb	SRX24927257	Periodontitis-6	PD6	PAIRED	2024-06-15	2024-06-14 15:04:00Z	PD-6	female
13	SRR29413855	SAMN41550712	56	300	105.04 G	60.66 Gb	SRX24927256	Periodontitis-5	PD5	PAIRED	2024-06-14	2024-06-14 15:07:00Z	PD-5	male
14	SRR29413856	SAMN41550711	54	300	112.36 G	64.50 Gb	SRX24927255	Periodontitis-4	PD4	PAIRED	2024-06-14	2024-06-14 13:33:00Z	PD-4	female
15	SRR29413857	SAMN41550708	25	300	114.67 G	65.78 Gb	SRX24927254	Healthy-7	Healthy7	PAIRED	2024-06-14	2024-06-14 13:35:00Z	Healthy-7	male
16	SRR29413858	SAMN41550709	48	300	111.73 G	63.09 Gb	SRX24927253	Periodontitis-2	PD2	PAIRED	2024-06-14	2024-06-14 13:54:00Z	PD-2	male
17	SRR29413859	SAMN41550702	26	150	52.78 G	33.17 Gb	SRX24927252	Healthy-1	Healthy1	SINGLE	2024-06-14	2024-06-14 13:04:00Z	Healthy-1	male
18	SRR29728988	SAMN42309061	24	300	110.44 G	64.56 Gb	SRX25231162	Periodontitis-20	PD20	PAIRED	2024-07-07	2024-07-07 22:28:00Z	PD-20	female
19	SRR29728989	SAMN42309060	46	300	101.54 G	58.43 Gb	SRX25231161	Periodontitis-19	PD19	PAIRED	2024-07-08	2024-07-07 23:54:00Z	PD-19	male
20	SRR29728990	SAMN42309059	62	300	116.00 G	66.14 Gb	SRX25231160	Periodontitis-18	PD18	PAIRED	2024-07-07	2024-07-07 22:52:00Z	PD-18	female
21	SRR29728991	SAMN42309058	72	300	116.41 G	66.98 Gb	SRX25231159	Periodontitis-14	PD14	PAIRED	2024-07-07	2024-07-07 22:10:00Z	PD-14	female
22	SRR29728992	SAMN42309057	60	300	116.36 G	67.32 Gb	SRX25231158	Periodontitis-13	PD13	PAIRED	2024-07-07	2024-07-07 22:18:00Z	PD-13	male
23	SRR29728993	SAMN42309056	56	300	109.78 G	62.74 Gb	SRX25231157	Periodontitis-11	PD11	PAIRED	2024-07-07	2024-07-07 22:17:00Z	PD-11	female
24	SRR29728994	SAMN42309055	63	300	108.00 G	61.30 Gb	SRX25231156	Periodontitis-10	PD10	PAIRED	2024-07-07	2024-07-07 22:08:00Z	PD-10	male
25	SRR29728995	SAMN42309054	62	300	121.10 G	68.74 Gb	SRX25231155	Healthy-10	Health10	PAIRED	2024-07-07	2024-07-07 22:13:00Z	Healthy-10	male
16	SRR29413858	SAMN41550709	48	300	111.73 G	63.09 Gb	SRX24927253	Periodontitis-2	PD2	PAIRED	2024-06-14	2024-06-14 13:54:00Z	PD-2	male
17	SRR29413859	SAMN41550702	26	150	52.78 G	33.17 Gb	SRX24927252	Healthy-1	Healthy1	SINGLE	2024-06-14	2024-06-14 13:04:00Z	Healthy-1	male
18	SRR29728988	SAMN42309061	24	300	110.44 G	64.56 Gb	SRX25231162	Periodontitis-20	PD20	PAIRED	2024-07-07	2024-07-07 22:28:00Z	PD-20	female
19	SRR29728989	SAMN42309060	46	300	101.54 G	58.43 Gb	SRX25231161	Periodontitis-19	PD19	PAIRED	2024-07-08	2024-07-07 23:54:00Z	PD-19	male
20	SRR29728990	SAMN42309059	62	300	116.00 G	66.14 Gb	SRX25231160	Periodontitis-18	PD18	PAIRED	2024-07-07	2024-07-07 22:52:00Z	PD-18	female
21	SRR29728991	SAMN42309058	72	300	116.41 G	66.98 Gb	SRX25231159	Periodontitis-14	PD14	PAIRED	2024-07-07	2024-07-07 22:10:00Z	PD-14	female
22	SRR29728992	SAMN42309057	60	300	116.36 G	67.32 Gb	SRX25231158	Periodontitis-13	PD13	PAIRED	2024-07-07	2024-07-07 22:18:00Z	PD-13	male
23	SRR29728993	SAMN42309056	56	300	109.78 G	62.74 Gb	SRX25231157	Periodontitis-11	PD11	PAIRED	2024-07-07	2024-07-07 22:17:00Z	PD-11	female
24	SRR29728994	SAMN42309055	63	300	108.00 G	61.30 Gb	SRX25231156	Periodontitis-10	PD10	PAIRED	2024-07-07	2024-07-07 22:08:00Z	PD-10	male
25	SRR29728995	SAMN42309054	62	300	121.10 G	68.74 Gb	SRX25231155	Healthy-10	Health10	PAIRED	2024-07-07	2024-07-07 22:13:00Z	Healthy-10	male
26	SRR29728996	SAMN42309070	62	300	107.93 G	60.57 Gb	SRX25231154	Periodontitis with DM-19	PDDM19	PAIRED	2024-07-07	2024-07-07 22:08:00Z	PDDM-19	male
27	SRR29728997	SAMN42309069	62	300	112.55 G	63.98 Gb	SRX25231153	Periodontitis with DM-16	PDDM16	PAIRED	2024-07-07	2024-07-07 22:04:00Z	PDDM-16	female
28	SRR29728998	SAMN42309068	51	300	109.89 G	61.97 Gb	SRX25231152	Periodontitis with DM-15	PDDM15	PAIRED	2024-07-07	2024-07-07 22:10:00Z	PDDM-15	male
29	SRR29728999	SAMN42309067	66	300	107.65 G	60.59 Gb	SRX25231151	Periodontitis with DM-14	PDDM14	PAIRED	2024-07-07	2024-07-07 22:11:00Z	PDDM-14	male
30	SRR29729000	SAMN42309066	56	300	113.30 G	63.18 Gb	SRX25231150	Periodontitis with DM-13	PDDM13	PAIRED	2024-07-07	2024-07-07 22:08:00Z	PDDM-13	female
31	SRR29729001	SAMN42309065	80	300	110.02 G	62.83 Gb	SRX25231149	Periodontitis with DM-12	PDDM12	PAIRED	2024-07-07	2024-07-07 22:06:00Z	PDDM-12	female
32	SRR29729002	SAMN42309064	57	300	113.23 G	64.02 Gb	SRX25231148	Periodontitis with DM-7	PDDM7	PAIRED	2024-07-07	2024-07-07 22:09:00Z	PDDM-7	male
33	SRR29729003	SAMN42309063	48	300	111.32 G	65.81 Gb	SRX25231147	Periodontitis-22	PD22	PAIRED	2024-07-08	2024-07-07 23:26:00Z	PD-22	male
34	SRR29729004	SAMN42309062	58	300	107.72 G	64.21 Gb	SRX25231146	Periodontitis-21	PD21	PAIRED	2024-07-07	2024-07-07 22:30:00Z	PD-21	female
35	SRR29729005	SAMN42309053	53	300	103.47 G	59.20 Gb	SRX25231145	Healthy-9	Health9	PAIRED	2024-07-07	2024-07-07 22:26:00Z	Healthy-9	male
36	SRR29729006	SAMN42309052	42	300	109.70 G	62.31 Gb	SRX25231144	Healthy-8	Health8	PAIRED	2024-07-07	2024-07-07 22:05:00Z	Healthy-8	male

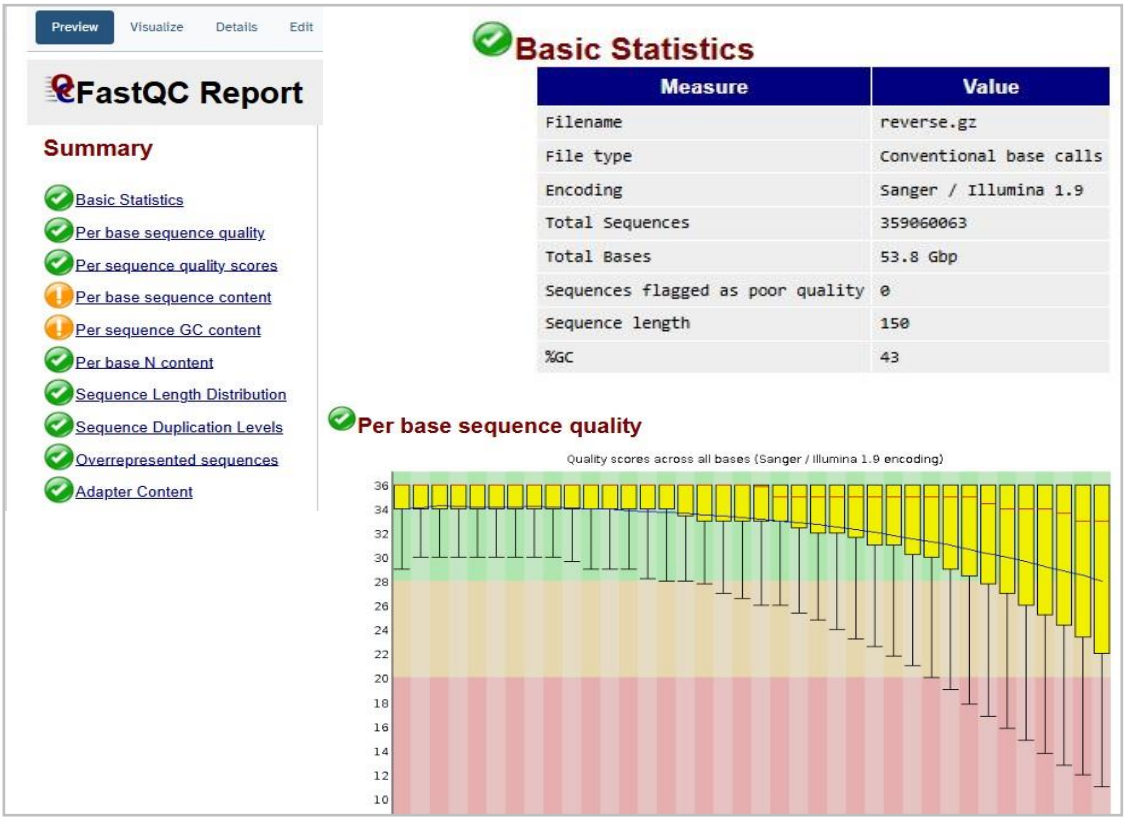
Appendix B

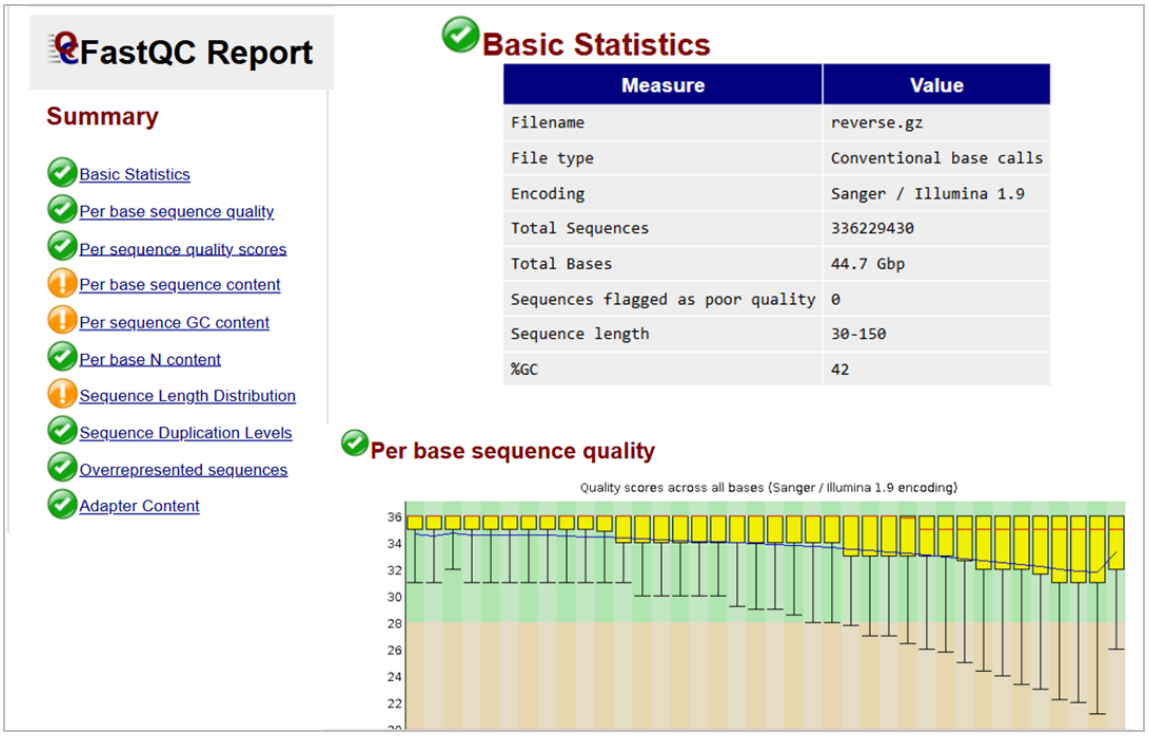
FASTQC Pre-Quality Analysis for PD Dataset (Forward)-**Top**. FASTQC Post-Quality Analysis for PD Dataset (Forward)- **Bottom**.



Appendix C

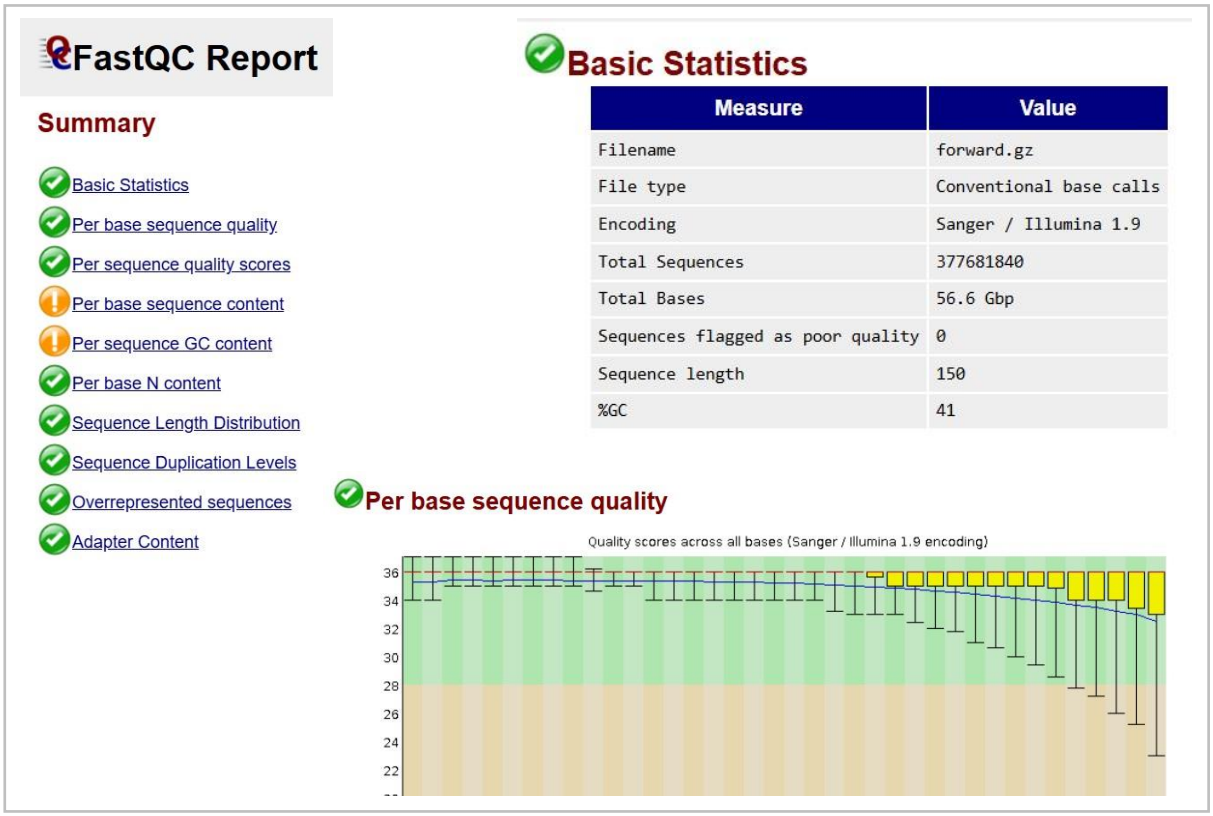
FASTQC Pre-Quality Analysis for PD Dataset (Reverse)-**Top**, and FASTQC Post-Quality Analysis for PD Dataset (Reverse)-**Bottom**.

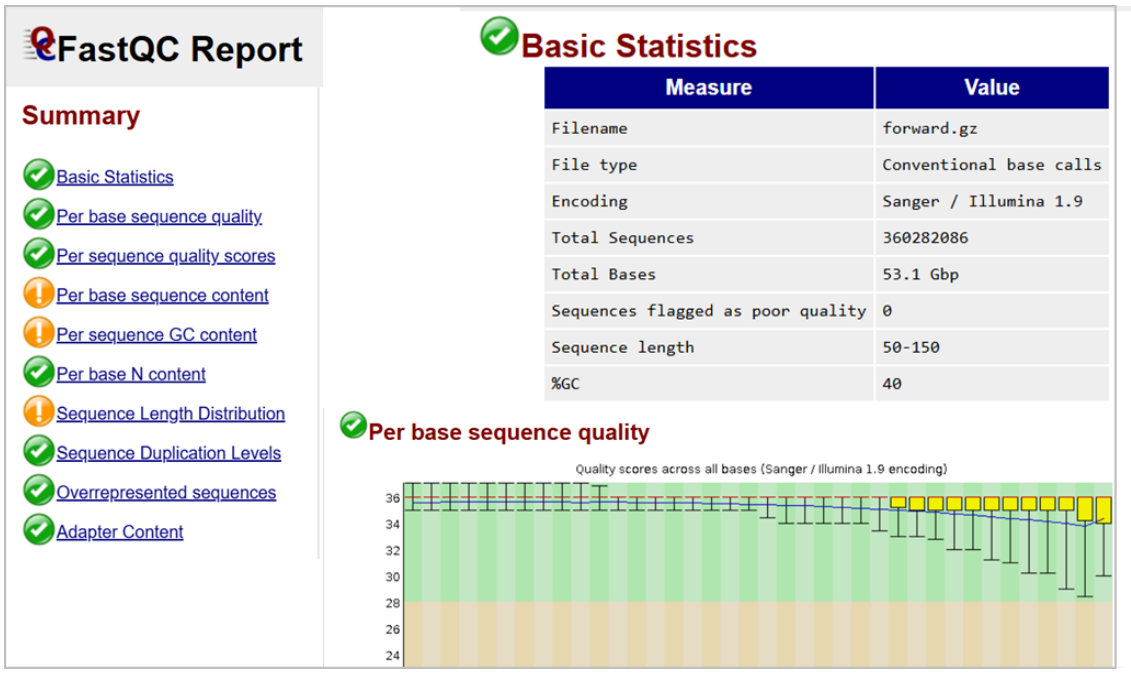




Appendix D

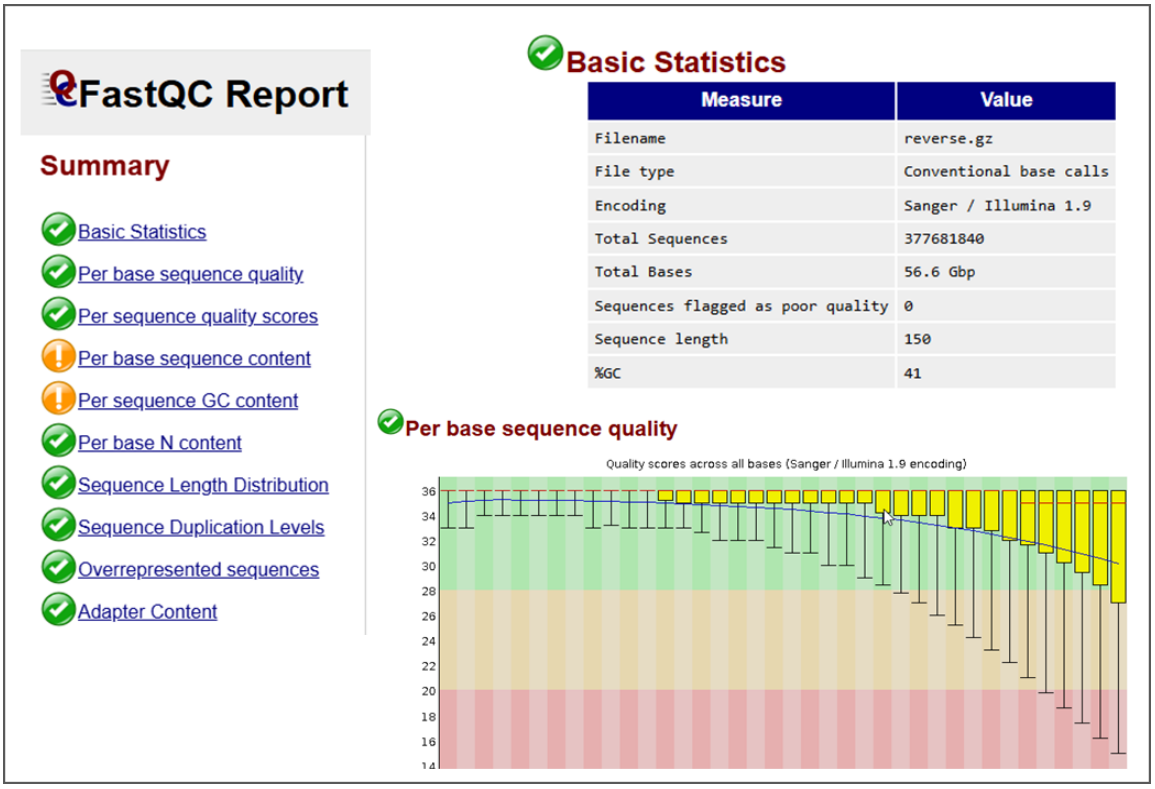
Results of FASTQC Pre-Quality Analysis for PDDM Dataset (Forward)-**Top**, and FASTQC Post-Quality Analysis for PDDM Dataset (Forward)-**Bottom**

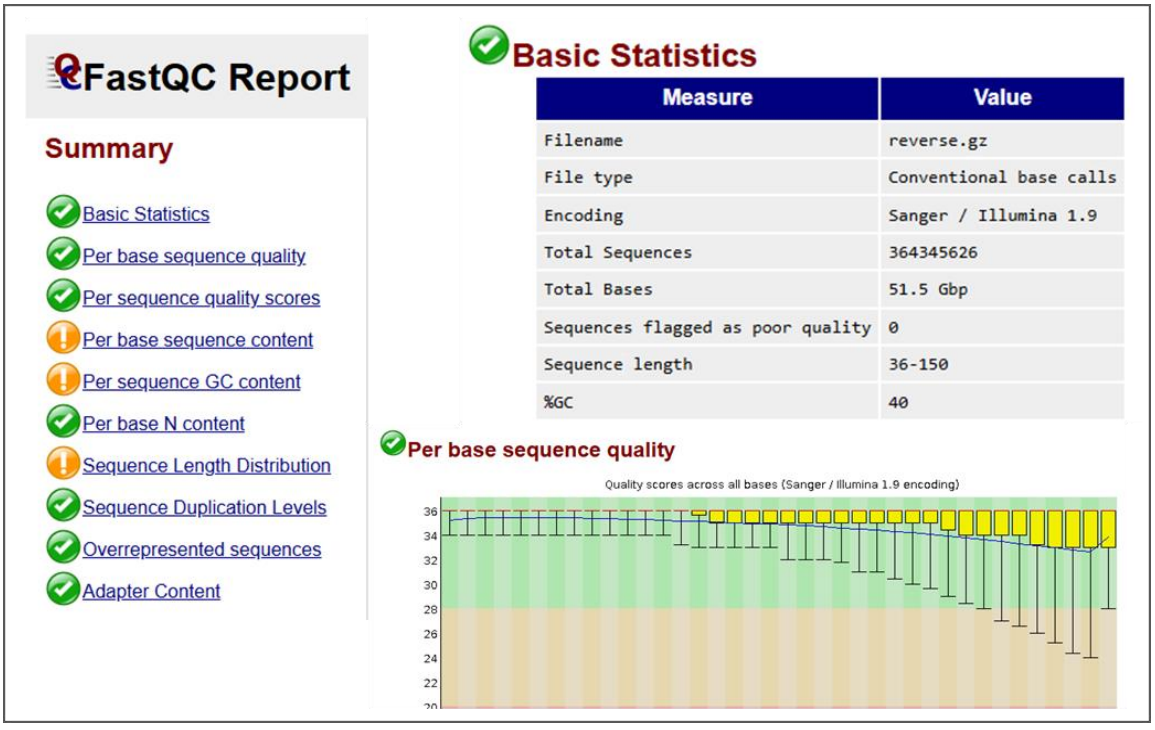




Appendix E

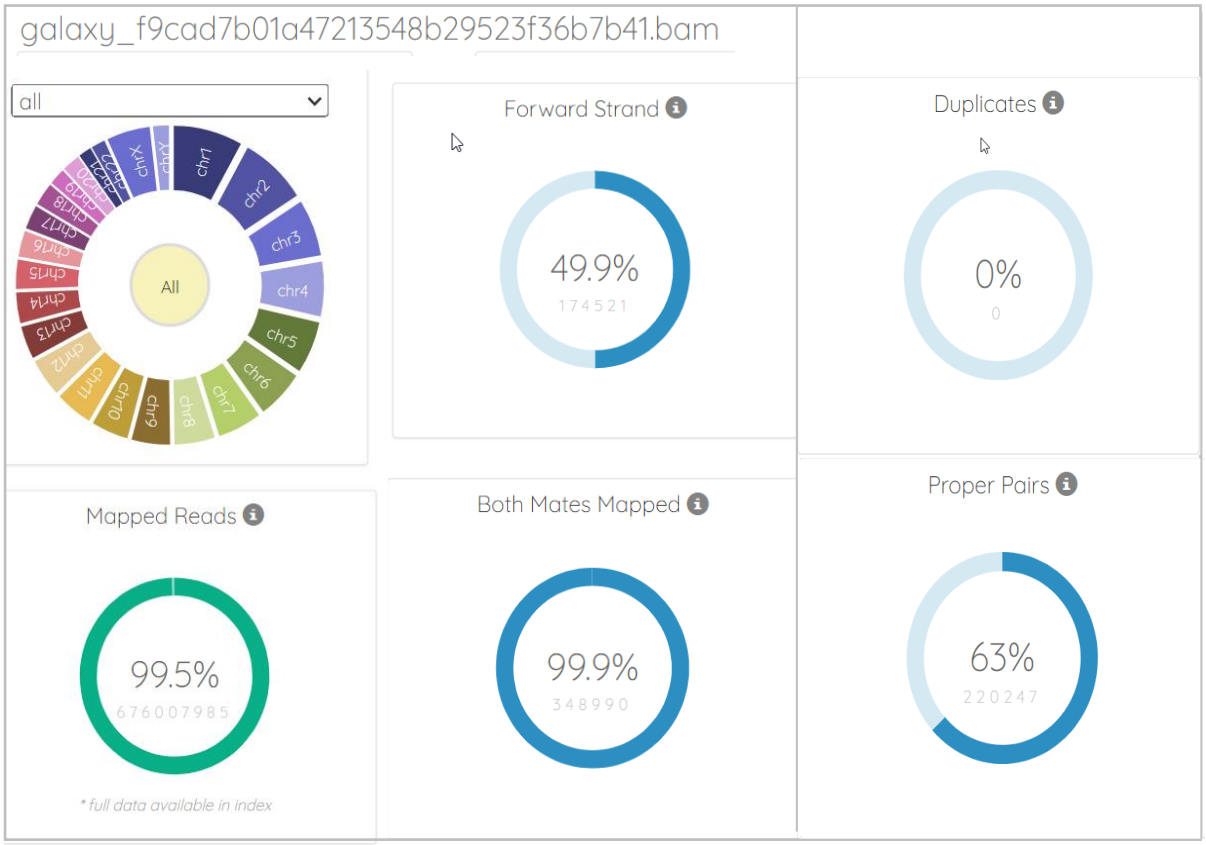
FASTQC pre-quality analysis for PDDM dataset (reverse) and FASTQC post-quality analysis for PDDM dataset (reverse).

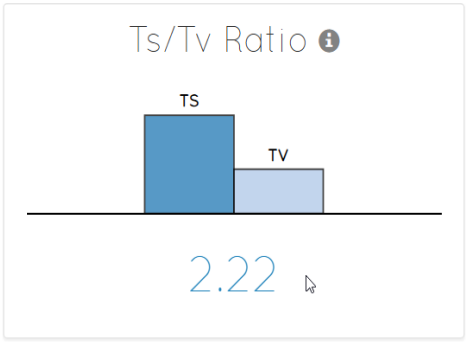


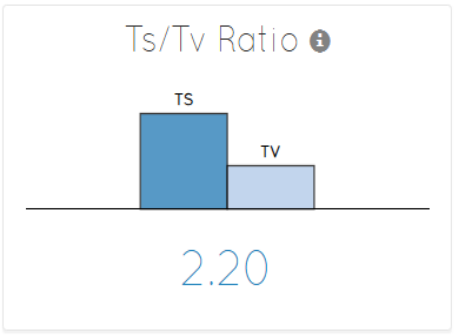


Appendix F

BWA-MEM results in vcf.iobio.io-**Top** and Ts/Tv Ratio PDDM (left) and PD (right)-**Bottom**

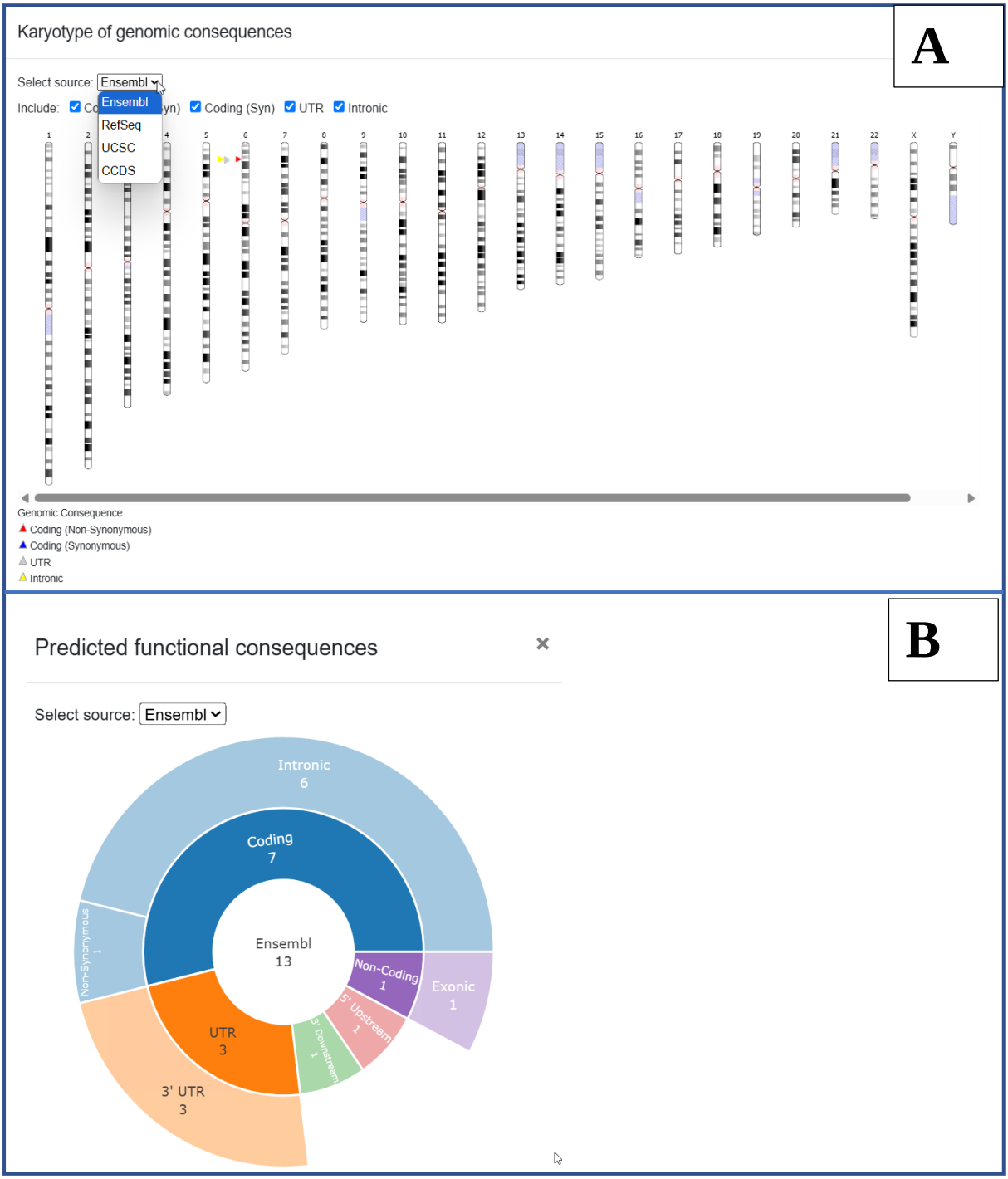






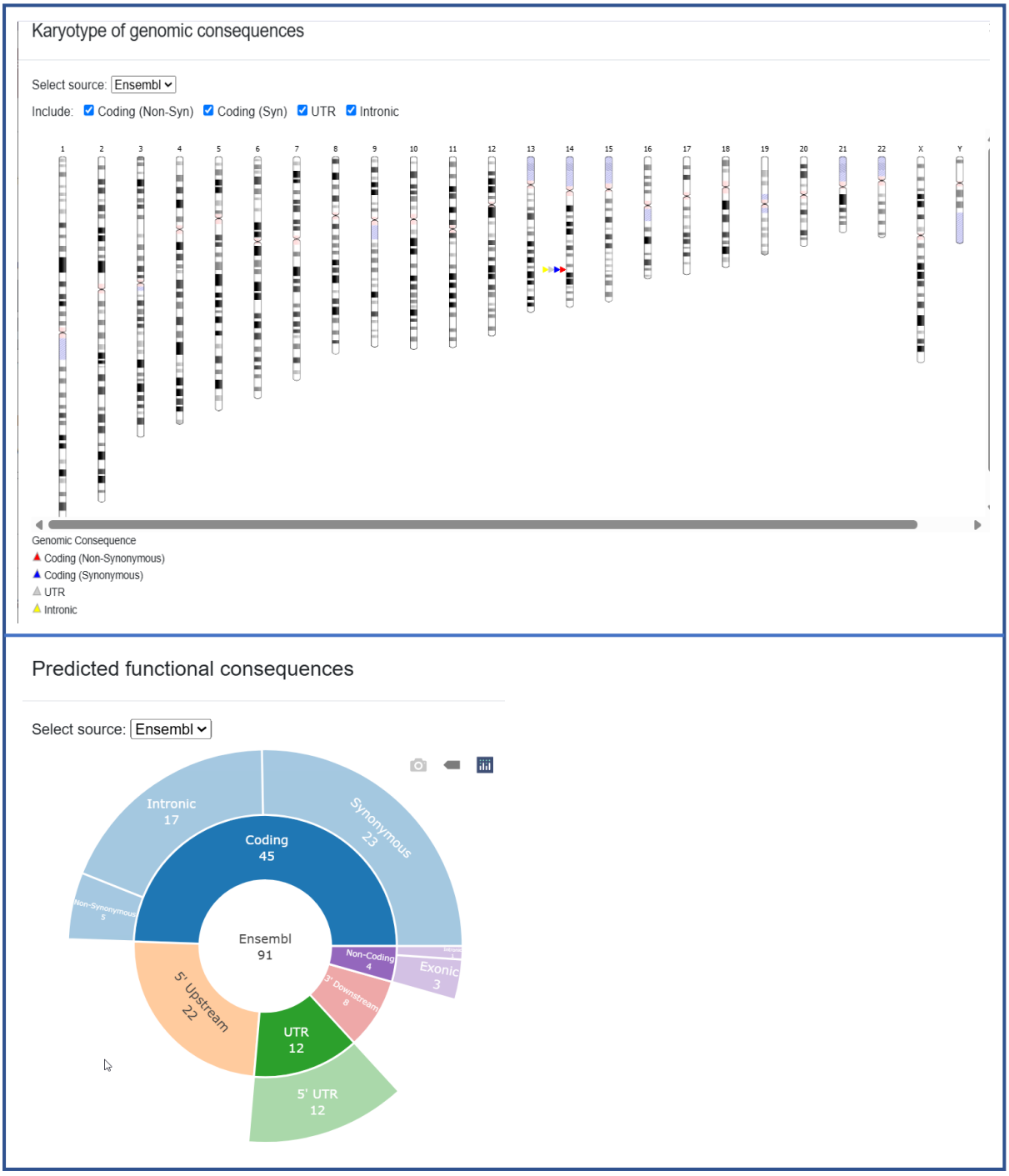
Appendix G

Predicted Genomic and Functional Consequences of MCUR1 Variants Based on Ensembl Annotations from SNPnexus.



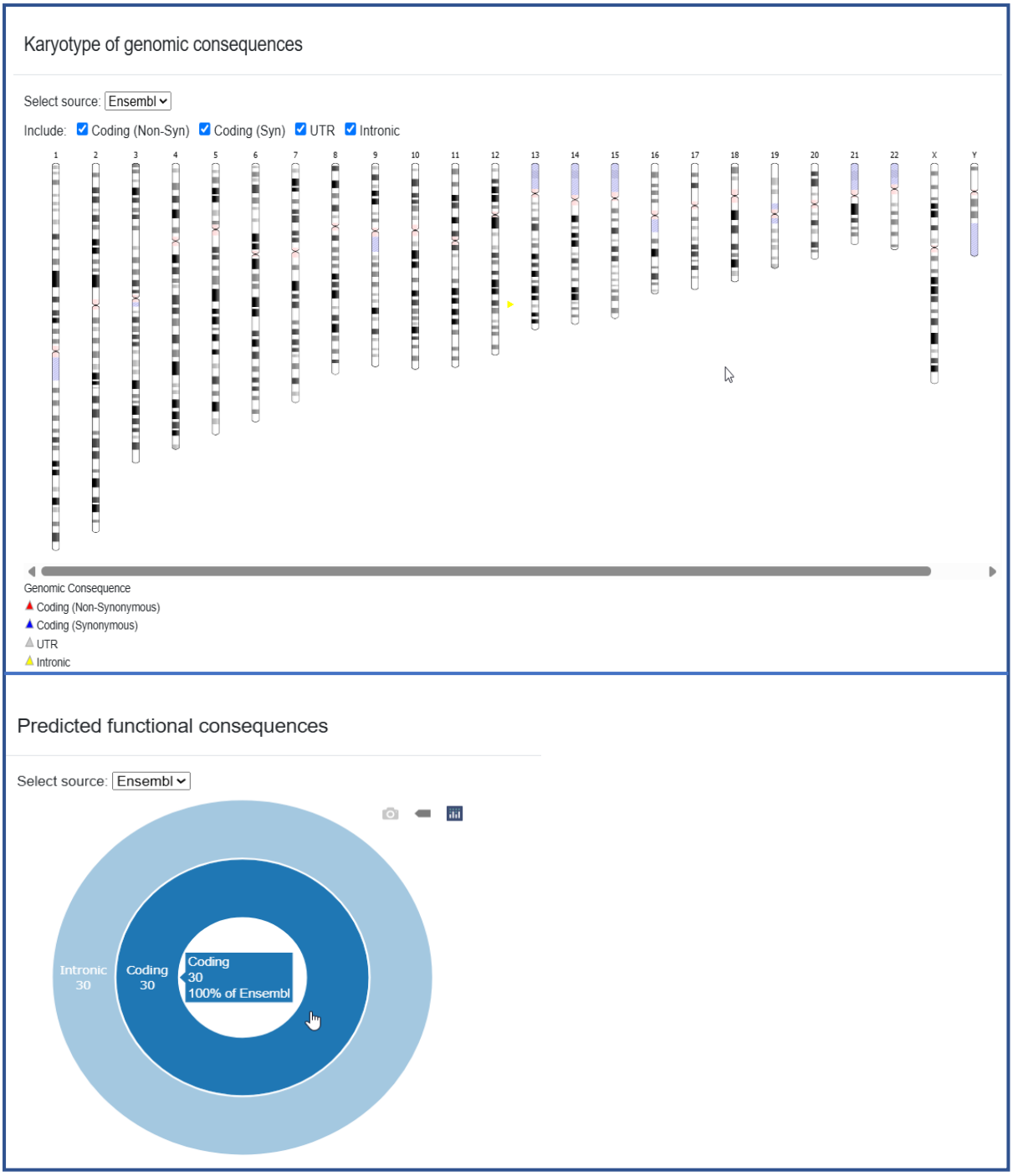
Appendix H:

Predicted Genomic and Functional Consequences of FOS Variants Based on Ensembl Annotations from SNPnexus.



Appendix I

Predicted Genomic and Functional Consequences of FOS Variants Based on Ensembl Annotations from SNPnexus.



Appendix J

Overall Venn diagram composed of specific gene list (PD) and (PDDM).

PD	1152	HPAT5 YWHAZ MAPRE2 C15orf48 H3-3A NOSTRIN C5orf34 GAGE12C IGFL3 HBZ HEPACAM2 RAD18 RHEX CFHR3 TSSC2 PROS1 CXorf21 SEPTIN14 TOB1-AS1 PPP1R14A LINC01802 COPS3 HLA-DQB1-AS1 APOC1 ANGEL2 ZBTB41 LOC100507156 DNASE2B CTXN1 CYP4F29P SLC30A7 UQCRC2 DPH3 CHROMR PSMD14 LINC01788 DOCK8-AS1 LOC100101148 AZI2 RAB3B MMP1 SLC46A3 DDX58 DTX2P1-UPK3BP1-PMS2P11 LINC00443 NPM1 SPART RAP1GDS1 SMAD1-AS1 KRTAP5-5 CFB ZWILCH LINC00917 MMD2 TRA2B CLIC2 AKAP7 PROK2 FAM9C ACRV1 MRPL44 PCBD2 RDX RAB3C BBS12 OIP5-AS1 ANKRD19P KCNMA1-AS3 TMPRSS11E GPR22 SSR3 MFSD14C CEP19 MNX1-AS2 MPPED1 LOC541473 LXN SLAMF6 DLGAP1-AS4 NCRNA00250 VLDLR-AS1 MAGEA6 NECAP1 TRAPPC3 C2orf76 EEF1AKMT2 CFHR2 STMP1 TMEM117 UBALD2 MAP3K2 MIR1227 ANAPC1P1 SNRPD1 TEX30 NMB OR5M3 CMC4 LINC01891 MANEA COX14 LOC102723809 POT1-AS1 MIR202HG CERS6-AS1 MDM2 PJVK PPIC TCF12 ARL6 ARG1 ADAM32 LOC100286906 RBX1 PPIAL4A SLC4A10 TPP2 NUP43 TBCE OLA1 ANKRD20A11P NMD3 ZNF213-AS1 F8A2 MS4A1 PLAC8 SPDYE11 SKP1 ALKBH3 VSTM2A-OT1 INTS7 OR4K3 NAT1 UBXN2B CLRN3 EGFEM1P SERINC3 C1orf54 H2AB3 LINC00210 STRADB RPP21 SF3B1 MEF2C-AS1 RIOK3 HSD3B2 SOSTDC1 GRXCR1 LOC79999 SLC25A33 CECR7 C17orf67 DDR1 GTF2B CROT BRK1 LILRA6 PPP3CA C16orf87 DYNLT3 FSBP RAD17 SPRR2E LINC00207 SH3YL1 SPINT1-AS1 RPH3AL LINC02692 SPRR2B LILRB3 IRF2 ZFP2 LINC02198 IGSF6 SH3GL2 CLDND1 C17orf75 IDI1 APCS SSX8P ATP6V0D2 MTERF3 BTG4 DPY19L2P1 PKIB VAMP8 FAM111A ART4 HDAC8 LINC00471 AMY2A H2AB1 TRHR GOLGA8EP LOC112577592 LLPH LOC105373805 RECQL RABL3 MCEMP1 LINC00355 CPSF7 C5orf66-AS2 FAM241B RPS6 HNMT BTG1 LOC102724770 GAS2 LINC02425 COMMD9 DEFB134 ANKRD40CL PSMC3IP RAB24 PLSCR4 GSTM5 MIR6789 LOC339166 GUSBP3 HLA-DPB1 ISCA1 CHI3L2 LOC647070 LPAR1 MMAA LINC02362 CCDC18-AS1 DENR SCART1 DEFA1 SMCP RBM41 SAGE1 ATP5PD ZSCAN16 MIR3150A KLRB1 LRR1 COPB1 C1orf162 TAFA3 CHST9 TSPAN31 GMNC IL31 F8A3 OGFR-AS1 SP140L TH2LCRR GPR1-AS DUSP21 OR5L1 LINC01869 SPATA31D3 MAPK8 LINC02338 AS3MT HSF2 JKAMP XXYLT1-AS1 H2BC5 LRRC3B MED4 ST8SIA4 MIR5193 ZNF732 TTLL7 SLC17A3 SCGB2A1 LINC02618 ELP3 LINC01945 HLA-DRB5 PSMD7 GABRB1 UBE2K MST1P2 EML4-AS1 LOC101927620 RNF24 NDRG3 KIN SSB RAB28 MRPL32 PSG10P LOC101926948 SGTB ZBTB6 ZNF490 LINC01118 STX8 MIR3652 GOLT1B BUD31 FAM118B CPEB2-DT DNAJB14 OR1G1 RPL39L LINC00367 MAP1LC3B LINC02622 CT45A10 USP17L22 SPATA6 SELENOI MANSC4 RAP2C-AS1 HMGN2 SMAD5-AS1 RNMT MS4A7 PTPRG-AS1 CUL4B LINC00868 ARL14EPL LINC01048 CLDN16 PATE2 TFAM H2BC13 TRIM13 OSCAR LDHAL6A GABRP ETDA SNAP23 IFT27 NSUN6 RAD21 DDX53 LOC100505622 SCEL-AS1 SMIM15 TTR LOC339975 OR4K14 SPINK4 CRISP2 MDH1B GABRR3 CEACAM8 LOC102723729 ARHGDIB TMEM263 ARL13A STRIT1 LOC284950 DNAJC2 OR4F21 CNIH2 VTCN1 AUNIP RPL32 LTO1 SDR16C6P ZNF461 SPCS1 TMEM68 MAL2-AS1 GNAQ SPRR2G APOBEC3H PPP1R10 LINC01646 ANAPC13
PDDM	2321	SPATA17 UBE2L5 LOC107986163 S100A14 NHLRC3 LINC02453 TMCO5B EEF1A1 TRAF6 MIR4683 RPS27L FETUB MRPL21 LINC02167 GAP43 SLC26A7 YPEL2 MORN4 RYBP ZFP57 CYB5R4 MTERF1 LINC02014 RAB11A LOC105375166 LSM5 DSTN CHIC2 VAMP2 PPP2CA SIKE1 VPS72 MDGA2 OCM CFAP206 HMGA2-AS1 CPHXL GTF2A1 DDIT4L ZNF709 LINC01331 CPLX4 RTCA-AS1 OR2AP1 APOF ALG14 SCAMP1 MRGBP CSMD2-AS1 STAT1 GLT8D1 LNCNEF HAO1 PDCD10 NUP35 HPGD KLK3 GPC6-AS2 EGFR-AS1 LINCMD1 NRADDP LOC100128076 LOC101928402 LCMT1-AS2 PTER C20orf197 COL5A1-AS1 LOC101929479 RNF144B FAM209A ELOVL3 SMIM28 ATP8B5P LINC01356 PRRC2A GEMIN8 AMBN CALR4P ARMC8 ITIH4-AS1 GTF2H2C CLEC4O ZRANB2 PTCSC2 NEURL1-AS1 CDK17 AOX3P-AOX2P KPNA5 FKBP7 CRK GJD2 ARMC1 SPACA4 KCNAB3 AFM CALCB PFDN4 KRTAP24-1 DARS2 ANXA10 CLCC1 CAB39 SMIM31 USP32P3 CCNG2 CD1D DNAJC5G MBD3L3 EQTN DEFB129 TMEM211 TMED10 RAMAC ZNF560 RACK1 POM121L1P TRAPPC6B RPL36A LOC100130950 MC3R THRB-AS1 PBLD BMPR1B TATDN1 LOC101928123 ZNF300 RRN3P2 LINC01924 LOC105379176 AQP4 SLC30A4 NEIL2 ATL2 RPS5 ALDH8A1 GPR180 DNM1P46 CCL4L2 UBE2V2 SLC7A9 PITPNC1 AIG1 RPSA ZNF70 SOS1 CMTM4 RPL26 BPIFA1 ZNF655 TAB3-AS1 BANF1 TGDS SH3BGRL2 LGALS8 E2F8 LINC01529 DEFB109B SLC22A3 LOC101929208 SPTY2D1OS LINC02383 KRT222 CDC6 ALG10 TMEM236 RFC4 ZNF561 APPBP2 ANKRD20A8P LOC100287497 HMG20A TAGLN3 TSC22D3 GTF2H2B TSPAN3 SEC31A TM4SF19-AS1 DEFB116 LOC100506885 GAL CHCHD4 SLMO2-ATP5E MEGF11 MS4A18 NBPF20 COA3 ACTR1A INHBA-AS1 LSM10 ADCY10P1 CXorf56 RBM5-AS1 ACTR2 FAM107B HPRT1 LOC101928797 DDX18 OR52J3 LOC100996724 ATP5ME NDUFB5 RNF170 CCDC28B GJA9-MYCBP SCIMP CST8 ZCCHC18 LINC01578 SMNDC1 BCL2A1 TBK1 SRSF3 CFL1P1 MPP1 TRPC4 RPL6 C11orf65 LINC01412 UXT-AS1 RBM45 LINC00537 STMN4 LOC101927809 MEIG1 PFN1 SMKR1 SESN3 LINC00466 PSMC1 ZFAND2A MAGEB1 FAM229B FAM151B DIO3OS LINC01375 ATP8A2 CSNK1A1P1 ARRDC3-AS1 GABARAPL2 POLB FAAHP1 GABPB1 LINC02595 LINC01282 PSMA3 QKI TMEM237 LOC389765 ATXN10 FGD5-AS1 DOK6 SMIM6 S100A6 H2BE1 MID1IP1-AS1 BTBD3 SREK1IP1 GCNT4 ERG OR4A5 ZCRB1 LETM2 CNPY4 LINC01451 LINC01102 ABHD5 SLC3A1 XKRX FBXO9 CHRAC1 OR2B2 BIRC3 SH2D1B GRID2 FAM200A BIVM GPAT3 PDE11A-AS1 RMDN2-AS1 UBE2D2 ANXA5 ORM2 HAUS3 CAMKMT H3C8 MIR555 VPS13A-AS1 CHMP2A C11orf94 CALML5 LOC105376453 LINC00940 TMEM170B ANKRD37 LOC108783645 C1orf226 KIAA0930 XPO1 LOC101927391 RBM15-AS1 TRIM23 ZCCHC10 LOC102724152 LOC100129203 BRMS1L LINC00598 DUSP18 GNG5 LGALS1 LOC100131626 SRP54 LINC01877 GCHFR RNF11 N4BP1 ARPC5 SNX6 PHGDH LINC00271 ANKRD26P3 ARMCX1 XAGE5 MIR10394 GHITM CNIH3 HNRNPCL3 MRPL42 ST3GAL6 FAM72D CLVS2 ZFP69 EIF4EBP2 SLC50A1 PIP LINC02527 CEP95 TXNDC17 POLR3K MYL12A ATP2B1 LINGO2 TBC1D7-LOC100130357 SELENOS TTC3-AS1 TOLLIP DYNC1I2 DNASE1L3 IFI44 PGPEP1 LOC101927078 PSMB5 LOC105376194 NDUFA12 H3Y1 ZDHHC17 GRIFIN THYN1

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
SAAD DAHLAB UNIVERSITY Blida1 (USDB1)

Faculty of Natural and Life Sciences

Department of Biology

Final Project Graduation thesis

For the purpose of obtaining a Master's degree in the field of SNV

Sector: Biological Sciences

Specialty: Biochemistry

Tittle

Interest in Genomic study of Diabetes with Periodontitis: In silico study

Presented by:

- **Motjolopane Relebohile**

Graduated on :01/07/2025

Before the jury:

Mrs CHALAL	*MCA*	*USDB1*	*President*
Mrs TALEB	*MCA*	*USDB1*	*Examiner*
Mrs EDDAIKRA	*MCA*	*USDB1*	*Supervisor*