Scalable Computational Frameworks for Next-Generation Sequencing Analysis and Gene Set Integration

By

MOHAMED ABUELANIN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHYLOSOPHY

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

C. Titus Brown, Chair

Fereydoun Hormozdiari

Megan Dennis

Committee in Charge

2024

Scalable Computational Frameworks for Next-Generation Sequencing Analysis and Gene Set Integration

Copyrights © 2024

by

Mohamed Abuelanin Hussien

*To my beloved wife, Shrouq, your unwavering support and love have been my rock. As we started our family during my PhD journey, welcoming our daughter Razan and soon another blessing, you kept us happy and united through long hours and busy days. Your presence has made this journey worthwhile, and I am forever grateful for your support and presence.*

*To my mom and dad, whose unwavering support and daily check-ins from overseas have been my lifeline. Your endless prayers and encouragement have made this journey possible. This achievement would not have been possible without your love and support throughout my life.*

*To my brothers, Ahmed and Hazem, for their unwavering support and encouragement throughout this journey. Your presence and belief in me have meant the world.*

## Acknowledgments

To Titus, who has been an incredibly understanding and patient PI. Throughout my PhD journey, you never rushed me but provided the space to learn and create. Your support extended beyond academics, offering personal encouragement that made all the difference. Your belief in my abilities and support for my work, even when it diverged from your initial ideas, created a healthy and stimulating environment. You fostered a lab culture that was friendly, stress-free, and built on trust, enabling me to innovate and explore my research passions. Your mentorship has been a cornerstone of my success; I am forever grateful for that.

To Tamer, who has been a pivotal part of my PhD journey. Since 2017, we have been working on numerous projects that have shaped my research mentality and built my skillset. Your research insights are incredible, and you are full of ideas. You spent countless hours working with me, supporting me, and paying attention to the finest details. Your patience and availability during your busiest times have been invaluable. Your dedication and expertise have been crucial to my success, and I am heartfelt in my gratitude for your unwavering support and cherished friendship.

To my lab friends Mostafa, Bry, Hannah, Gina, Luiz, Anne, Tessa, and Colton, thank you for your support. Your encouragement and shared passion for research have made this journey memorable and enriching.

To Sherif Bahriz, Saleh Salman, Ahmed Adel, Mahmoud Abdelghany, Bassel Elgharabawy, Mostafa Khater, and all my Davis friends, thank you for your friendship and support throughout this journey; your companionship has been invaluable.

# Abstract

The rapid growth in biomedical research has generated vast amounts of data, including genomic, molecular, imaging, and clinical information from humans and other species. Leveraging this data is essential for groundbreaking scientific discoveries and a deeper understanding of health and disease across different species. However, the complexity and volume of these datasets present significant computational challenges, limiting their potential.

This dissertation addresses two key challenges in biomedical data analysis: the efficient evaluation of sequencing data and the effective management and analysis of gene sets. By focusing on these areas, we develop innovative computational methods that enable the rapid, scalable, and accurate processing of large-scale biomedical data. For sequencing data, we create algorithms that enhance the speed and precision of data evaluation, making it feasible to manage the increasing volume of sequences generated by modern technologies. For gene sets, we devise tools for their efficient management and analysis, allowing researchers to draw meaningful insights from complex genetic information.

Through this research, we aim to contribute to the development of new analytical tools and methods, ultimately supporting the advancement of precision medicine and personalized healthcare for both human and veterinary applications.

# Chapter 1

# Background: Scalability of Computational Tools for Biomedical Data Analysis

We are surrounded by data but starved for insights.

Jay Baer

## Motivation

The biomedical data landscape includes genomic, molecular, imaging, and clinical data, such as electronic health records. Technological advances, such as next-generation and single-cell sequencing, have enhanced the resolution with which we study genes. At the same time, high-resolution imaging and electronic health records have improved disease detection and data integration, advancing personalized medicine and research.

However, the vast amount of available biomedical data presents a unique challenge (Dinov, 2016): its potential remains untapped without adequate tools and methods. Developing innovative approaches to harness this data is essential, as it can lead to significant scientific discoveries and a deeper understanding of human health and disease.

# Background

## Computational Techniques for Scalability

### Sequence Alignment and Alignment-Free Methods

The quest for efficient sequence alignment has driven dramatic advances in bioinformatics, with notable milestones in algorithmic innovation and implementation. The state-of-the-art BLAST tool (Altschul et al., 1990) introduced a heuristic approach for rapid local similarity searches, vastly accelerating database comparisons. Later, Bowtie (Langmead et al., 2009) and BWA (Li, 2013) leveraged the Burrows-Wheeler Transform and FM-index to align short DNA reads to large genomes with unprecedented speed and memory efficiency. These developments have been instrumental in analyzing next-generation sequencing data. More recently, DIAMOND (Buchfink et al., 2015) has pushed the boundaries further with double indexing, achieving BLAST-like speeds with reduced computational demands, which is particularly suited for protein alignment in metagenomics. Despite these advancements, large-scale bioinformatics applications pose significant computational challenges, necessitating ongoing innovation to optimize resource utilization, accuracy, and analysis time for the ever-growing volumes of genomic data (*Genomic Data Resources*, n.d.).

In contrast, alignment-free techniques offer a scalable alternative by estimating sequence similarity through statistical properties rather than explicit alignment (Vinga & Almeida, 2003) (Sims et al., 2009) (Zielezinski et al., 2017). For instance, k-mer counting facilitates the rapid comparison of sequences without the computational overhead associated with traditional alignment methods (Manekar & Sathe, 2018). These techniques enable faster data processing and support tasks like clustering and classification.

## Alignment-free Exact vs. Non-Exact Sequence Representation

In genomic research, inexact data representations offer a scalable solution for efficient analysis but can introduce inaccuracies, while exact representations ensure high precision at the cost of increased computational resources. The choice depends on research goals, with exact representations suited for applications requiring high accuracy and inexact representations suitable for scalable analyses.

### Exact Sequence Representation

Exact data (lossless) representations maintain the original sequence data without alteration, precisely capturing all its information. This approach is essential when studying specific genomic variants that can be a single SNP or even 1Kbp variation or identifying unique genetic markers in individuals or sub-populations (Chaung et

al., 2023). Pinpointing exact mutations within a patient's genetic makeup is crucial for personalized medicine applications, such as tailoring pharmacogenomic treatments or designing targeted gene therapies (Gasic et al., 2021). Similarly, in microbial genomics, distinguishing between closely related strains—each potentially harboring different pathogenic traits—relies on exact sequence data to accurately trace transmission pathways and resistance mechanisms (Hooper & Jacoby, 2015).

In DNA, exact data representation can be exemplified by k-mer full spectra, which rely on the complete set of k-mers in a genome or a sample to perform downstream analysis. This approach is crucial in genomics, as it enables researchers to capture the entire genomic landscape and gain a more comprehensive understanding of the underlying biology. To extract k-mers efficiently, high-performance tools like KMC (Kokot et al., 2017), Jellyfish (Marçais & Kingsford, 2011), and DSK are commonly employed, facilitating the rapid identification of k-mers that can then be utilized in various applications, such as sequence assembly (Koren et al., 2017), comparing datasets, contamination analysis (Wingett & Andrews, 2018), and reads binning in the metagenomics domain (Kawulok & Deorowicz, 2015). Notably, the full k-mer spectrum has recently been utilized to construct pan-genome full k-mer content, which has numerous applications, including investigating and studying structural variants at scale (Chaung et al., 2023). Moreover, the full k-mer spectrum offers a robust framework for exploring genomic data, allowing researchers to probe the complexities of genomic structure and function with precision, and by leveraging k-

mer full spectra, researchers can gain a deeper understanding of the intricate relationships between genomic elements and their role in shaping the evolution and diversity of species.

Despite the power of k-mer full spectra in genomics research, harnessing their potential at scale poses significant scalability challenges. One major hurdle is memory requirements, as the sheer volume of long k-mers in a single genome (e.g., k=31bp) can be staggering. For instance, the human genome alone contains approximately 3 billion k-mers, requiring around 24GB of memory (64bits x 3.2 billion k-mers). This becomes even more daunting when comparing hundreds of highly sequenced human samples, which would necessitate enormous computational resources and memory capacities. As such, researchers face significant technical barriers in scaling k-mer full spectra analysis to meet the demands of large-scale genomics studies, highlighting the need for innovative solutions to overcome these limitations and fully unlock the potential of k-mer full spectra in genomics research.

**Non-Exact Data Representations in Large-Scale Genomic Studies**

Sketching is a computational technique that transforms large datasets into compact, lossy representations called sketches (Rowe, 2019). Sketching algorithms, such as MinHash, offer a solution by providing a sublinear space representation of data.

Initially developed for document clustering and deduplication (Broder, 1997), MinHash uses hashing to guarantee query precision while balancing memory usage and accuracy. By converting documents into sets through shingling and reducing them to shorter, similarity-preserving signatures, MinHash efficiently estimates Jaccard similarity and containment between documents - crucial for comparing genomic datasets.

Recent innovations like Mash (Ondov et al., 2016a) have adapted MinHash for genomic data, employing fixed-size signatures to estimate similarity and introducing additional metrics for genomic complexity. CMash further enhances this approach by utilizing Bloom Filters for scalable containment estimates (Liu & Koslicki, 2022), supporting the comparison of diverse-sized datasets without excessive hash storage. Mash Screen streamlines containment score calculation by mapping distinct hashes from reference sketches to a query sequence, avoiding redundant data storage and processing (Ondov et al., 2019). Scaled MinHash (FracMinHash), introduced in sourmash (Pierce et al., 2019a) (Irber, Brooks, et al., 2022), refines this approach by adjusting hash selection based on a scaling parameter, combining the benefits of fixed and dynamic-sized sketches to suit large-scale, diverse, and comparing unequal-size complex datasets.

The FracMinHash is defined as:

$$\mathrm{FRAC}_s(S) = \{h(k) \mid k \in S \text{ and } h(k) \leq H_s\}$$

Where:

- S is a set of elements,
- h(k) is a hash function applied to each element $k$ in the set S
- $H_s$ is a threshold calculated as $\frac{2^{64}-1}{\text{scale}}$

**FracMinHash's Operational Steps:**

1. Hash Function Application: Every element k in the set S is processed through the hash function h, generating a hash value h(k).
2. Threshold Comparison: Each hash value h(k) is compared against the threshold $H_s$. If h(k) is less than or equal to $H_s$, it is included in the resultant hash set. Otherwise, it is excluded.
3. Subset Formation: The hash values that meet the threshold criterion form the FracMinHash of the set S, serving as a probabilistic representative of the set.

The scale parameter directly influences $H_s$, allowing precise control over the fraction of hash values included in the FracMinHash. A smaller $scale$ value includes a broader range of hash values, increasing the sensitivity and size of the FracMinHash, whereas a larger $scale$ value results in a more selective, smaller FracMinHash. By selectively including only a fraction of hash values, FracMinHash efficiently manages large datasets, reducing computational overhead while retaining the ability to estimate set similarities accurately.

The cornerstone of the FracMinHash is a good hash function. For instance, sourmash constructs the FracMinHash using the MurmurHash3 (*MurmurHash3 · Aappleby/Smhasher Wiki*, n.d.) hash function. MurmurHash3 is a robust and efficient non-cryptographic hashing function with the following properties:

- Simplicity: MurmurHash3 is computationally efficient, requiring a minimal number of assembly instructions.
- Excellent distribution: It passes rigorous chi-squared tests for various keysets and bucket sizes, ensuring a uniform hash distribution.
- Avalanche behavior: MurmurHash3 exhibits strong avalanche properties, with a maximum bias of 0.5%. This means that changing a single bit in the input will flip, on average, 32 bits (50%) of the 64-bit output hash. This ensures that even small input changes result in significantly different output hashes. That is why the Avalanche Effect is the primary reason behind FracMinHash's efficiency in representing DNA sequences, whether raw samples or

genomes.

## Balancing Exactness and Computational Efficiency

The choice between exact and non-exact data representations often hinges on the specific goals of the research and the computational resources available. Exact representations are indispensable for detailed studies requiring high-fidelity data, such as variant calling or strain differentiation. However, non-exact methods can significantly reduce computational demands and expedite the research process for exploratory studies to identify general trends or patterns.

Furthermore, the trend towards integrating exact and non-exact methods into hybrid approaches is gaining momentum. This strategy involves leveraging the efficiency of non-exact methods for initial screening, followed by precise methods for validation and refinement.

As a part of this PhD research, a hybrid decontamination workflow was developed that combines the strengths of both exact and approximate methods. This workflow begins with FracMinHash, which quickly compares query samples to thousands of microbial genomes (cite sourmash gather), identifying a list of potential contaminants within minutes. The workflow can then utilize the full k-mer content of these genomes to perform a more thorough read-level decontamination.

# Dissertation Objectives & Research Questions

This dissertation aims to advance bioinformatics by developing pioneering computational methodologies for processing and analyzing large-scale datasets. This research is dedicated to enhancing three fundamental aspects: sequence characterization, quality control, and exploring relationships among gene sets. This dissertation contributes to bioinformatics research by fulfilling these goals, empowering researchers to address complex biological queries and unravel the intricate webs of relationships that govern biological systems.

## Dissertation Objectives

The objective of this dissertation is to devise and implement innovative computational strategies for the analysis of massive-scale bioinformatics data, with a particular focus on the following two domains:

1. Sequence Characterization and Quality Control:
    a. Using lightweight sequence sketches, we aim to implement methods that would replace traditional sequence alignment to estimate the depth of sequencing and target sequence coverage and introduce new metrics to quantify the amplicon enrichment.
    b. Using a small subset of the raw sequencing samples, we can predict the coverage gain that more sequencing would achieve.
    c. We aim to distinguish between sequencing errors, contamination, and

novel sequence content.

2. Gene Set Relationship Exploration:
    a. Develop an algorithm that leverages the sparse nature of gene sets to perform efficient pairwise comparisons, reducing computational complexity and enabling scalable analysis.
    b. Implement a framework that would adapt the pairwise comparisons algorithm to apply algorithms designed to discover direct and indirect relationships among gene sets. This includes coexistences and associations among pathways, diseases, genetic variants, pharmaceuticals, and other gene set categories.

# Research questions

1. Can k-mer sketch-based methods replace sequence alignment with minor trade-offs in accuracy for estimating the primary alignment statistics in mammalian species?
2. Can large-scale analysis of thousands of unmapped reads be used to construct a comprehensive pan-genome k-mer content, and what new sequences can be revealed through a k-mer-based view of pan-genome content?
3. Can we utilize massive-scale analysis to create k-mer content that efficiently produces a reference genome k-mer content to perform reference-free downstream analysis?
4. Is it possible to predict the coverage gain that extra sequencing would achieve for the same biosample?
5. Gene sets are very sparse; would leveraging that sparsity in the data allow large-scale pairwise comparisons to connect different gene sets (pathways, diseases, variants, drugs, etc.) and find answers quickly for critical biological problems?

## Dissertation structure

**Chapter 2** introduces Snipe, a highly scalable set of methods for estimating alignment-based metrics (coverage, depth, mapping rate, amplicon enrichment) in mammalian space. Snipe was able to utilize approximately $\frac{1}{10{,}000}$ of the unique k-mer content in canine unassembled samples to estimate alignment metrics and perform extensive sequence-based quality control. It could also discriminate between new canine k-mer content (pan-genomic) and contaminants. Lastly, it was able to effectively predict the gain of coverage that extra sequencing efforts would achieve for a specific biosample, which significantly helps in quickly calculating the potential return on investment when considering deeper sequencing.

**Chapter 3** presents DBRetina, a set of methods to analyze and explore gene sets during large-scale pairwise comparisons. It can incorporate gene sets of different types (e.g., diseases, pathways, drugs, and pathways) to find the underlying connections and associations. For example, with a small set of commands, we could identify all the co-existent diseases with Alzheimer's disease. Another example is finding pathways similar to those found in multiple databases to a list of genes, which can improve the statistical significance of the gene set enrichment analysis.

**Chapter 4** discusses the significance of the introduced methods in the dissertation and provides future directions on how we can continue to improve highly scalable computational methods to leverage the exponential growth of biomedical data.

# Chapter 2

# Snipe: Lightning Reference-Based Quality Control of Next-generation Sequencing Data


## Abstract

**Background:** Aligning sequencing data enables the calculation of valuable quality control metrics, but its high computational demands restrict scalability for large-scale analyses.

**Aim:** We aim to create a lightweight, alignment-free method with broad applications, including precise estimation of sequencing coverage and depth, return on investment in more sequencing, error rate, and possible contamination. This approach bypasses the constraints of traditional alignment methods, facilitating analysis on a petabyte scale.

**Methods:** We introduce Snipe, an alignment-free tool for thoroughly and efficiently evaluating sequencing datasets by comparing lightweight k-mer sketches of sequencing data with reference genomes and target amplicons at scale.

**Results:** Our approach has demonstrated exceptional efficiency, being at least 100-fold faster and more memory-efficient, with minimal disk space requirements compared to conventional alignment methods. Snipe was evaluated by analyzing ~19,000 canine SRA experiments, accurately estimating their alignment-based statistics. Moreover, it effectively identified duplications, detected mis-annotations, and pinpointed samples that could contribute novel content to the canine genome. Additionally, Snipe enabled the construction of pangenome k-mer content sketches, facilitating the identification of new sequence content and providing accurate predictions of return on investment (ROI) for sequencing experiments. A website has been developed to facilitate the search and visualization of results using SRA identifiers.

**Conclusions:** Our comprehensive evaluation of canine SRA datasets with Snipe enables the effective reuse of these valuable resources, enhancing the genomic research landscape. Researchers can now efficiently compare their new sequencing data against all available SRA samples, facilitating advancements in genomic studies through improved accessibility and insightful comparative analysis. Snipe's ability to predict ROI and construct pangenome sketches further empowers researchers to make informed decisions about sequencing experiments and identify new sequence content.

**Keywords:** dog – coverage - depth – pangenome – ROI – QC

# Introduction

Sequence alignment is a vital computational method in bioinformatics, with diverse

applications in high-throughput sequencing data, including quality control and contamination detection. Tools like Qualimap (García-Alcalde et al., 2012) and Samtools (Li et al., 2009) summarize alignment statistics to evaluate library quality, while VerifyBamID (Zhang et al., 2020) uses sequence alignment to identify contaminants. Moreover, sequence alignment-based tools like MarkDuplicates (Picard) (*Picard Tools - By Broad Institute*, n.d.) enable the quantification of duplicate reads, which helps predict the effectiveness of additional sequencing efforts. By leveraging sequence alignment, researchers can optimize their sequencing strategies, enhance data quality, and improve the overall efficiency of their workflows.

Despite the efficiency of short-read aligners like Bowtie (Langmead & Salzberg, 2012, p. 2) and BWA (Li, 2013), the computational intensity of sequence alignment methods still poses a significant bottleneck, exacting a heavy toll on processing power, memory resources, and storage capacity (Patro & Salmela, 2020). As of May 2024, the SRA has reached 91.2 PB of data, showing a 67.5% volume increase to 2022 **(Supplementary Figure 2.1).** This exponential growth of Next Generation Sequencing data has dramatically amplified this challenge, inundating researchers with an unprecedented deluge of complex sequencing data (Bansal et al., 2018). This has put public repositories like the SRA under immense pressure to manage the sheer scale of data and makes it challenging for researchers to identify, access, and utilize relevant datasets. Therefore, developing innovative solutions to address the challenges of sequence alignment scalability has become crucial to catering to the evolving needs of genomics research and enabling efficient analysis of massive-scale sequencing data.

K-mer-based alignment-free approaches offer an efficient alternative for analyzing genomic data, but they have limitations when dealing with thousands of samples. While techniques that utilize the full k-mer spectra, such as KMC (Kokot et al., 2017), Jellyfish (Marçais & Kingsford, 2011), and DSK (Rizk et al., 2013), have made significant progress, they still manifest a gap in enabling massive-scale quality control on thousands of samples. However, sketching methods have bridged this gap by compressing raw data into compact, lightweight representations using a subset of its k-mers (Rowe, 2019). Techniques like Mash (Ondov et al., 2016b) and sourmash (Pierce et al., 2019b) have made it possible to efficiently analyze, search, and compare large datasets, paving the way for discoveries in genomics research (Irber, Pierce-Ward, et al., 2022). FracMinHash sketching selects the k-mers whose hash values fall below a certain fraction of the maximum possible hash value, determined by a user-defined scaling factor (Irber, Brooks, et al., 2022). FracMinHash was proven reliable in the DNA representation of the prokaryotic space, accurately estimating genome containment, and is successfully applied in metagenomic taxonomic profiling (Irber, Brooks, et al., 2022). It was also used to estimate Average Nucleotide Identity (ANI) (Hera et al., 2023), showing promise in estimating other sequence alignment metrics.

This paper introduces ultrafast sequence assessment and quality control methods utilizing the FracMinHash sketching to the mammalian space for accurately estimating sequence-alignment-based statistics, including sequencing coverage, depth, mapping rate, and amplicon enrichment. Our approach enables rapid and informed decisions about resource allocation by predicting the gain in coverage additional sequencing would achieve,

thereby facilitating the calculation of return on investment (ROI). We showcase how our method can easily utilize sequencing datasets on a population scale to construct sketches for the whole pangenome k-mer content, enabling reference-free processing, discovery of new sequence content, and revealing new insights into genetic diversity.

# Materials and Methods

## SRA Data retrieval and sketching

All Illumina sequencing runs (n=19,914; 182 terabytes of compressed SRA files), either Whole-genome sequencing (WGS) and whole-exome sequencing (WXS), of *Canis lupus familiaris* were retrieved from the SRA repository using a target query: (txid9615[All Fields]) AND "illumina"[Platform] AND ("wgs"[Strategy] OR "wxs"[Strategy]) AND ("2000/01/01"[PDAT] : "2023/12/25"[PDAT]). Bioprojects PRJNA525883 and PRJEB22026, containing RNA and mitochondrial sequencing data, respectively, were excluded. The Bioprojects with metagenomic sequencing were kept to mimic the scenario of canine WGS samples with variable levels of bacterial contamination. Sequencing runs were sketched to FracMinHash sketches using sourmash v4.8.6 (Pierce et al., 2019b). Sketching was done using a k-mer size of 51 bp and a scale of 10,000, representing approximately every 10Kbp with a single k-mer. Duplicate sequencing runs were removed based on sourmash's MD5 checksums. SRA experiments with multiple runs were merged into a single sketch by summing k-mer abundances. Sketches with fewer than 100 genomic hashes and those with extreme k-mer-to-bases ratios (indicating inadequate representation) were excluded. The former represents tiny samples, usually less than one megabase pair of sequencing data, while the latter happens to samples with reads too short or have high N content and thus cannot generate enough k-mers. The data preprocessing ends with 18,067 sketches for SRA experiments in 215 bioprojects.

## Preparation of the reference materials

The Reference CanFam3.1 genome assembly (accession GCF_000002285.3) and the corresponding annotation GFF were downloaded from the NCBI Genome database. The exome sequences were extracted from the genome using a homemade script. The reference genome and exome sequences were sketched similarly to the SRA datasets using the same sourmash parameters.

## K-mer-based sequence assessment metrics

Unique k-mer count: The number of hashes in a sourmash sketch reflects the complexity of the sequencing library. The unique k-mer count in an experiment is expected to be close to the size of the target reference, but it increases with off-target sequencing, sample contaminations, and sequencing errors.

Total k-mer abundance: The sum of all hash abundances is a proxy for estimating the number of sequenced bases and is expected to rise with the increase in sequencing depth.

K-mer coverage Index: The coverage index is a ratio between the number of unique k-mers shared between sample and target sketches and the total number of unique k-mers in the sketch of this target reference. It quantifies the coverage percentage of the target sequence in a sequencing experiment.

K-mer mapping index: The mapping index is a ratio between the abundance of k-mers shared between sample and target sketches to the total k-mer abundance in this sample's sketch. It is an estimator of the mapping rate of sequencing reads.

k-mer mean abundance: The mean abundance of k-mers shared between sample sketches and a target reference is an estimate for the sequencing depth in this sample.

Return-on-investment (ROI): In sequencing, ROI is defined as the anticipated gain in coverage, reflected by an increase in the coverage index, with more sequencing. ROI can guide decisions on the necessity of additional sequencing.

Relative coverage: Calculated as the ratio of coverage indexes for an amplicon and a reference genome, quantifying the amplicon enrichment in a sequencing experiment. This ratio is typically higher in targeted sequencing experiments, such as WXS than in WGS. Median-trimmed relative coverage is a modified version of relative coverage, calculated after trimming k-mers at or below the median abundance to minimize the effect of shallow off-target sequences in WXS.

Relative mean abundance: Calculated as the ratio of k-mers mean abundances in a targeted amplicon region to that in non-targeted regions. Similar to relative coverage, it quantifies the amplicon enrichment but is based on the sequencing depth.

Amplicon score: A composite metric for quantifying amplicon enrichment in a sequencing experiment. It is calculated by multiplying median-trimmed relative coverage by relative mean abundance.

## Sample selection for benchmark analysis

For the benchmark of our k-mer-based sequence assessment metrics, 6,778 sequencing datasets were aligned by BWA v0.7.17-r1188 (Li & Durbin, 2009), and subsequent BAM QC was done by Qualimap v.2.2.2 (García-Alcalde et al., 2012) as a ground truth.

To select these datasets, 4,939 WGS and 1,896 WXS experiments from 91 and 38 canine BioProjects, respectively, were randomly selected from the SRA repository. The exome mapping rate was utilized to circumvent likely misannotation in SRA experiments **(Supplementary Figure 2.3)**. WGS experiments exhibiting atypical exonic sequence enrichment with an exome mapping rate > 8% and WXS experiments demonstrating insufficient enrichment with an exome mapping rate < 20% were eliminated from the dataset. This filtered 9 WGS and 36 WXS experiments from the analysis.

## Downstream analysis

For the detection of microbial contaminants, the Branchwater plugin of sourmash (specifically, the fastmultigather command) was used to compare all the canine SRA sketches against the sketches of all bacterial species in the Genome Taxonomy Database (GTDB) v214. Furthermore, we leveraged kSpider for pairwise comparisons and clustering the breeds' pan-genomic k-mers. We utilized iTOL (Interactive Tree of Life) (Letunic & Bork, 2021) to visualize the breed hierarchical clusters. Finally, we used Matplotlib and Seaborn for additional visualizations.

# Results

Snipe is a Python API with a command-line interface (https://github.com/snipe-bio/snipe). Additionally, it has a Java script/WebAssembly implementation for a web portal (https://snipe-bio.github.io/web). Snipe efficiently calculates several k-mer-based

statistics for sourmash sketches to enable alignment-free or even reference-free QC of sequencing datasets. To benchmark Snipe's computational performance, five sequencing experiments with varying depths were randomly selected to evaluate Snipe and BWA alignment statistics. Our comparison in Table 2.1 showed that alignment-free QC by snipe required, on average, 65 times less time, 83 times less memory, and 673 times less disk space than BWA Alignment, using a single core for Snipe, 32 cores for BWA, and a single core for qualimap BAMQC reports.

To showcase the power and versatility of Snipe in analyzing large-scale mammalian sequencing datasets, we sketched 182 terabytes of compressed SRA data into 67 gigabytes of sourmash sketches for 18,067 SRA experiments. Using Snipe, we calculate their proposed k-mer-based metrics to estimate their alignment-based coverage, depth, and mapping rate. Moreover, we showed how Snipe can be used to study the complexity of sequencing libraries to predict the possible gain in coverage if more sequencing was done. Finally, we present how Snipe can utilize the vast repositories of sequencing data to build pangenome sketches. Snipe used these sketches to calculate the k-mer-based metrics in a completely reference-free mode and to identify the common contaminants of the species. The precalculated statistics for all canine sequencing datasets are accessible through interactive visualizations on the Snipe web portal.

Table 2.1 Snipe's computational efficiency: Average computational statistics for data sketching and Snipe analysis versus BWA alignment and Qualimap assessment for randomly selected samples at different sequencing depths. Snipe requires sourmash, and Qualimap requires BWA's BAM file.

| Metric | Depth | BWA/Samtools | Qualimap | Sourmash | Snipe |
|---|---|---|---|---|---|
| **Wall time** | **1x** | 17.11 | 5.52 | 5.14 | 0.04 |
| | **5x** | 56.32 | 17.08 | 20.51 | 0.03 |
| | **10x** | 109.59 | 27.30 | 36.73 | 0.04 |
| | **20x** | 225.01 | 57.21 | 79.12 | 0.05 |
| **Memory** | **1x** | 11.4 GB | 3.8 GB | 104 MB | 64 MB |
| | **5x** | 11.3 GB | 3.4 GB | 120 MB | 76 MB |
| | **10x** | 12.2 GB | 3.5 GB | 129 MB | 87 MB |
| | **20x** | 12.4 GB | 3.3 GB | 155 MB | 119 MB |
| **Disk Space** | **1x** | 1.3 GB | 1.4 MB | 2.3 MB | 10 MB |
| | **5x** | 3.7 GB | 1.5 MB | 5.0 MB | 10 MB |
| | **10x** | 5.6 GB | 1.5 MB | 6.7 MB | 10 MB |
| | **20x** | 11 GB | 1.5 MB | 11 MB | 10 MB |

## K-mer sketching provides an efficient representation of the reference genome

FracMinHash sketching hash all the k-mers to of a given sequence or dataset and choose only those with hash values within a fraction of the total hash space. Sketching the Boxer's 2.4 Gb genome (CanFam3.1) comprises 231,482 k-mer hashes. Evaluating the efficiency of FracMinHash sketching in representing the Canine Reference Genome, we observed that k-mer distances follow a Poisson distribution with a lambda (λ) value of approximately 10,000 bp, which aligns with the algorithm's scale factor of 10,000 **(Figure 2. 1A, Supplementary Figure 2.2)**. This consistency confirms that the

sketching technique efficiently represents the genome despite challenges such as polyploidy and repetitive sequences. However, applying the same sketching scale to the exome sequences excluded 85% of canine genes, with only 15% retained, mostly represented by a single hash, as shown in **Figure 2.** 1**B**.

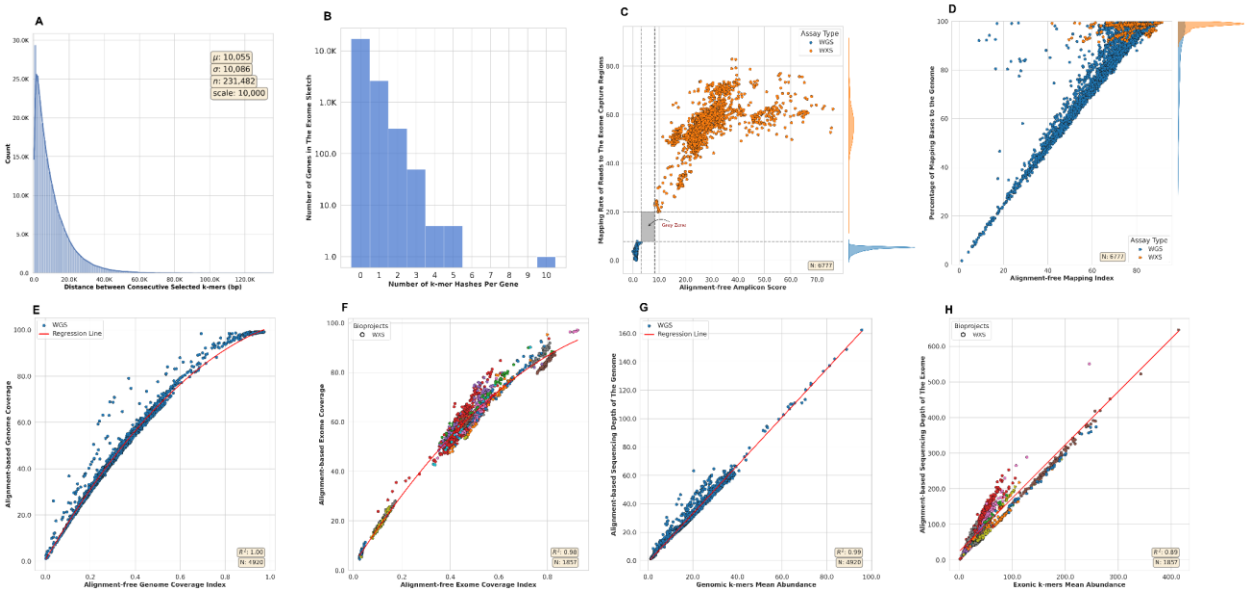



**Figure 2. 1 Benchmark of k-mer-based metrics: (A)** A histogram for the distances between selected k-mers from CanFam3.1 reference genome by FracMinhash sketching at a scale factor of 10,000. **(B)** A histogram for the counts of selected k-mers per gene from the canine exome by FracMinhash sketching at the same scale. **(C)** A scatter plot of the k-mer amplicon score (x-axis) versus the mapping rate of sequencing reads to the exome capture regions (y-axis). **(D)** A scatter plot of the k-mer mapping index (x-axis) versus the mapping rate of sequencing reads to the reference genome (y-axis). **(E, F):** Scatter plots of the k-mer-based coverage index (x-axis) versus the alignment-based coverages (y-axis) of the reference genome in WGS experiments (E) and the exome regions in WXS experiments (F). **(G, H)** Scatter plots of the k-mer abundance (x-axis)

versus the sequencing depths (y-axis) of the reference genome in WGS experiments (G) and the exome regions in WXS experiments (H). C and D are colored by the assay type, while F and H are colored by the SRA BioProject IDs.

## Accurate estimation of alignment-based metrics

Benchmarking Snipe's k-mer metrics against Qualimap's alignment statistics in a dataset of 6,777 sequencing experiments demonstrated the robust estimation capabilities of k-mer metrics. Snipe's amplicon score, one of the k-mer-based metrics, combines depth and coverage data of targeted and non-targeted amplicon regions to quantify enrichment efficiency. **Figure 2. 1C** reveals a strong association between the Amplicon Score and the percentage of reads mapped to the exome, effectively discriminating between WGS and WXS sequencing types, with WGS experiments scoring less than 3 while WXS experiments scoring above 7. The k-mer mapping index is another Snipe metric that correlates with the alignment-based mapping rate of sequencing reads. However, this correlation weakens toward the high end (**Figure 2. 1D).** The Coverage Index (CI) estimates the coverage of a target sequence, whether a whole genome or a specific amplicon. **Figure 2. 1E and F** show a strong polynomial regression (quadratic) between CI and alignment-based coverage, validating its accuracy. Similarly, the k-mer mean abundance shows a strong linear correlation with alignment-based depth **(Figure 2. 1G and H)**. However, a batch effect associated with BioProject is observed mainly in WXS experiments, resulting in a systematic deviation from the regression line in **Figure 2. 1F and H**.

# Snipe Enables Quality Control of SRA-scale Sequencing Data

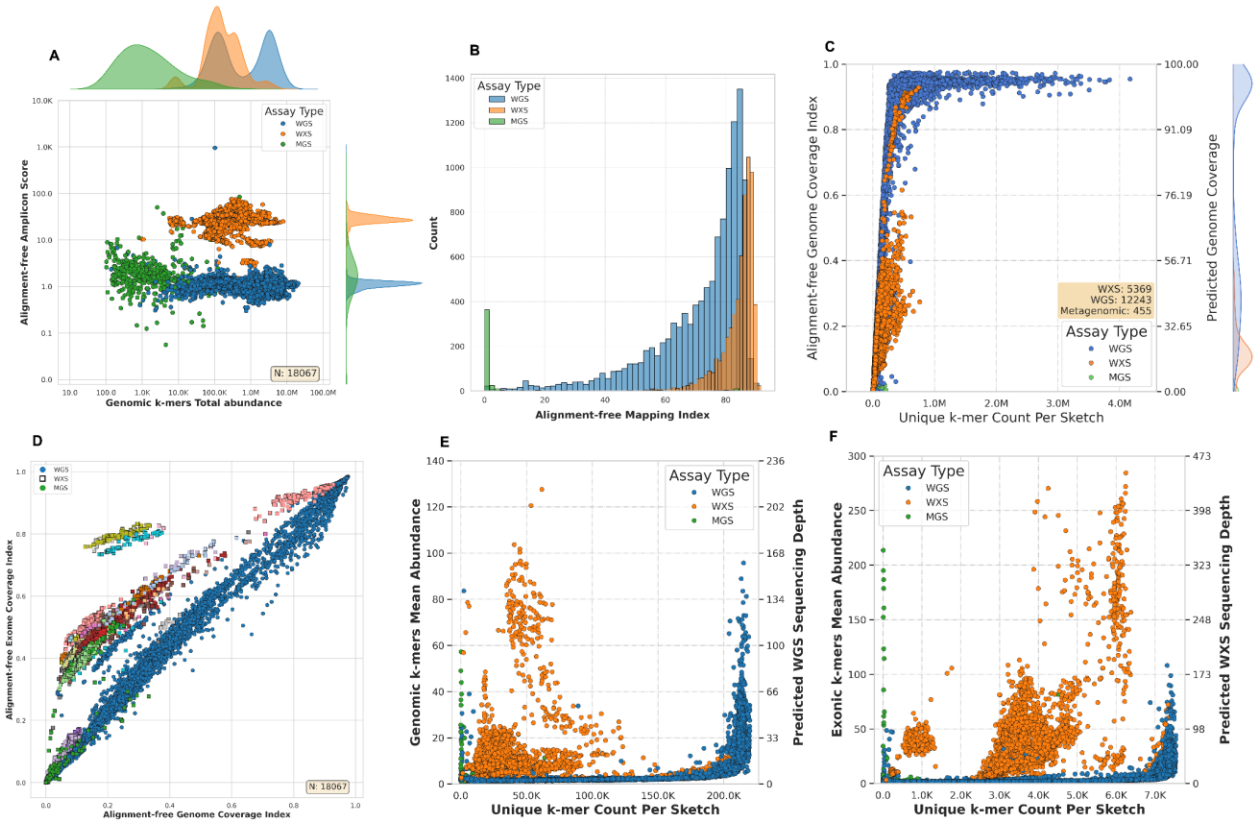



**Figure 2.2 Massive-scale quality-control analysis of sequencing datasets:** Snipe's analysis of all canine SRA experiments shows the distribution of many k-mer-based statistics, including **(A)** k-mer-based amplicon scores plotted in a scatter plot versus the total abundance of the k-mers recognized in the canine reference genome per sketch, **(B)** alignment-free mapping rates plotted in a histogram, **(C, D)** coverage indices of canine reference genome plotted in scatter plots versus the count of unique k-mers per sketch (C) and versus the corresponding coverage indices of canine exome (D), **(E, F)** mean abundances of the k-mers recognized in the canine reference genome (E) and the canine

exome (F) per sketch plotted in versus plotted in scatter plots against the corresponding unique k-mer count. All plots are stratified by the WGS, WXS, and MGS sequencing experiment type. In addition, the WXS experiments are colored by SRA Bioproject IDs in D. An extra y axes are added in C, E, and F for the predicted alignment-based statistics.

Calculating the amplicon score for the Canine SRA experiments revealed some cases of unexpected enrichment and/or misannotation **(Supplementary Figure 2.5).** Therefore, experiments were categorized based on their amplicon scores as follows: 1) Likely, if the score aligns with the SRA annotation; 2) Unlikely, if it contradicts the annotation; and 3) Ambiguous, if the score falls between 3 and 7 units, which corresponds to the grey zone in **Figure 1C**. The canine exome constitutes approximately 3.8% of the entire genome, representing the expected average mapping rate to exome capture regions if no enrichment occurs. In WXS experiments, successful enrichment is anticipated to elevate this rate. Therefore, we used sequence alignment to verify and correct unlikely annotations in subsequent references throughout the manuscript **(Supplementary Figure 2.6)**.

The computational efficiency of Snipe allowed quality control analysis of all WGS and WXS datasets in the SRA repository until the end of 2023, which resulted in several valuable insights. Most of the canine WXS experiments, even those with high depth of sequences, show high enrichment for exome sequences except for a subpopulation in the Bioproject PRJEB53653. On the other hand, almost all WGS experiments, after fixing

the likely misannotated experiments, show homogenous coverage of the genome manifested by low Amplicon Scores, while some metagenomic samples show higher scores (**Figure 2A**, **Supplementary Figures 5 and 6**). Moreover, the analysis shows that most WGS and WXS, unlike metagenomics, have high mapping rates **(Figure 2.2B)**. Snipe's analysis for the coverage of target sequences reveals an expected correlation between the total number of unique k-mers and the coverage of the genome but also shows that many samples have a high number of k-mers that don't contribute to that coverage **(Figure 2.2C).** Some of these novel k-mers are caused by genetic variance, especially in non-boxer breeds, while the major load of these k-mers is possibly due to sequencing errors and/or contamination. Focusing on WXS, most of these experiments have a genome coverage score of less than 0.2, corresponding to 32% predicted coverage, with a long tapering tail of experiments with high genome coverage **(Figure 2.2C)**. On the contrary, the WXS datasets show much more variable exome coverage than might be expected with different capture panels and the known bias of target sequencing. **Figure 2.2D** shows that most WXS samples achieve suboptimum coverage of the exome sequence, which is justifiable by targeting different genomic regions, panel design, and limitations of target sequencing. However, a few WXS Bioprojects at the top left corner of **Figure 2D** achieve a higher exome coverage index (~0.8) with a relatively low genome coverage index (less than 0.4), indicating a more efficient design for capturing exome sequences. This cluster consists of three bioprojects: PRJNA891496, PRJNA752630, and PRJNA701141. No WXS datasets can achieve near-full exome coverage without equivalently high full genome coverage consistent with poor enrichment like the WXS in PRJEB53653 at the top right corner of **Figure 2D**. Lastly,

WXS datasets in the bottom left cluster, primarily composed of PRJEB7540, PRJEB55865, and PRJEB55864, show shallow sequencing with minimal exome and genome coverage. **Figure 2.2E** and **F** illustrate the relationship between the number of unique k-mers and their mean abundance for WGS, WXS, and MGS experiments.

Surprisingly, many MGS experiments share k-mers with the genome or exome with very high sequencing dept, suggesting possible reference contamination or lateral gene transfer.

In **Figure 2.2E**, WGS shows a gradual increase in sequencing depth with more unique k-mers, while WXS shows significant variability in depth and coverage. MGS has a high mean abundance with few unique k-mers, likely due to low mapping rates **(Figure 2.2B)**. In **Figure 2.2F**, WGS maintains a gradual depth increase, and WXS continues to show variability, possibly due to different targeted regions or enrichment kits. The presence of multiple depths at the same coverage level suggests that higher depths do not always improve coverage. These figures highlight the variability in WXS sequencing compared to the consistent patterns in WGS.

# Predicting Sequencing Coverage and BioSample Reliability for Optimal Return-on-Investment

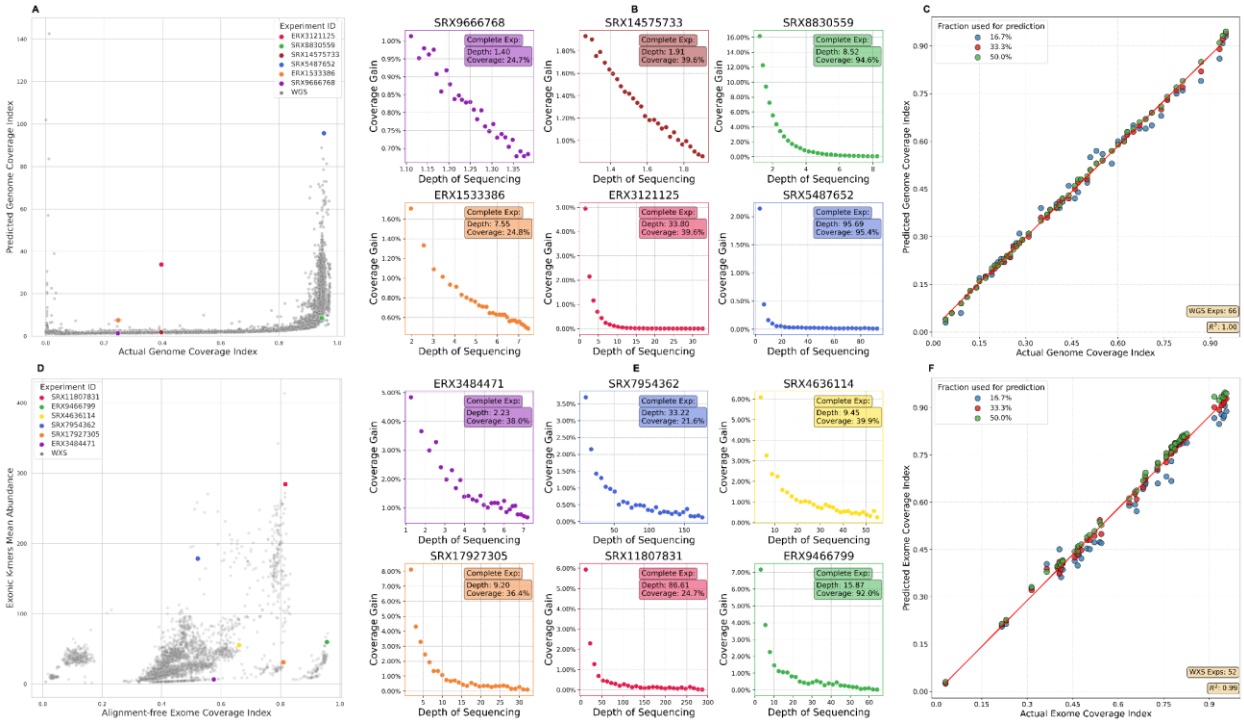



**Figure 2.3 Predicting coverage gain from small data fraction: (A)** Scatter plot of genomic coverage index versus genomic depth for all whole-genome sequencing (WGS) experiments, highlighting six experiments selected for further analysis. (**B)** Detailed analysis of the highlighted WGS experiments, divided into 30 sequential segments. Subplots display cumulative depth against delta coverage for each segment, color-matched to the corresponding experiments in panel **A**. Delta coverage is calculated as the difference in coverage between successive additions of segments. **C)** Correlation plots for WGS, comparing predicted genomic coverage using different data fractions (16.7%, 33.3%, and 50%) against actual coverage. Each point represents prediction accuracy for

each fraction, demonstrating the model's ability to predict total coverage from partial data. **(D-F)** Scatter plots and correlation analyses for WXS experiments, analogous to panels **A-C**, but displaying exon depth versus exon coverage index. Panels **D, E**, and **F** show the same analysis as panels **A, B**, and **C,** respectively, but for WXS experiments.

In **Figure 2.3A and D**, we selected six experiments from each assay type (WXS and WGS) that exhibit varying depths for similar coverage, indicating that increased depth does not necessarily lead to greater genomic or exonic coverage. To investigate this further, we split each experiment's raw data into 30 equal parts, cumulatively added successive parts, and plotted the combined depth against delta coverage **(Figure 2.3B and 3E).** The resulting curves show a high negative correlation when the biosample is suitable for further sequencing and a plateau when additional sequencing will not contribute to increased coverage. This is exemplified in **Figure 2.3B**, where SRX14575733 (dark red) exhibits a high negative correlation, indicating a potential need for further sequencing, while ERX3121125 (pink) has reached a plateau. In **Figure 2.3C** and **F**, we simulated a real-life application of Snipe's ROI calculation by predicting final coverage for 68 WGS and 54 WXS experiments using fractions of the original sequences (5/30, 10/30, and 15/30). Our predictions achieved a Pearson $R^2$ of 0.99 with the actual coverage, demonstrating the accuracy of Snipe's ROI calculation.

# Snipe enables pan-genomic analysis of all SRA content

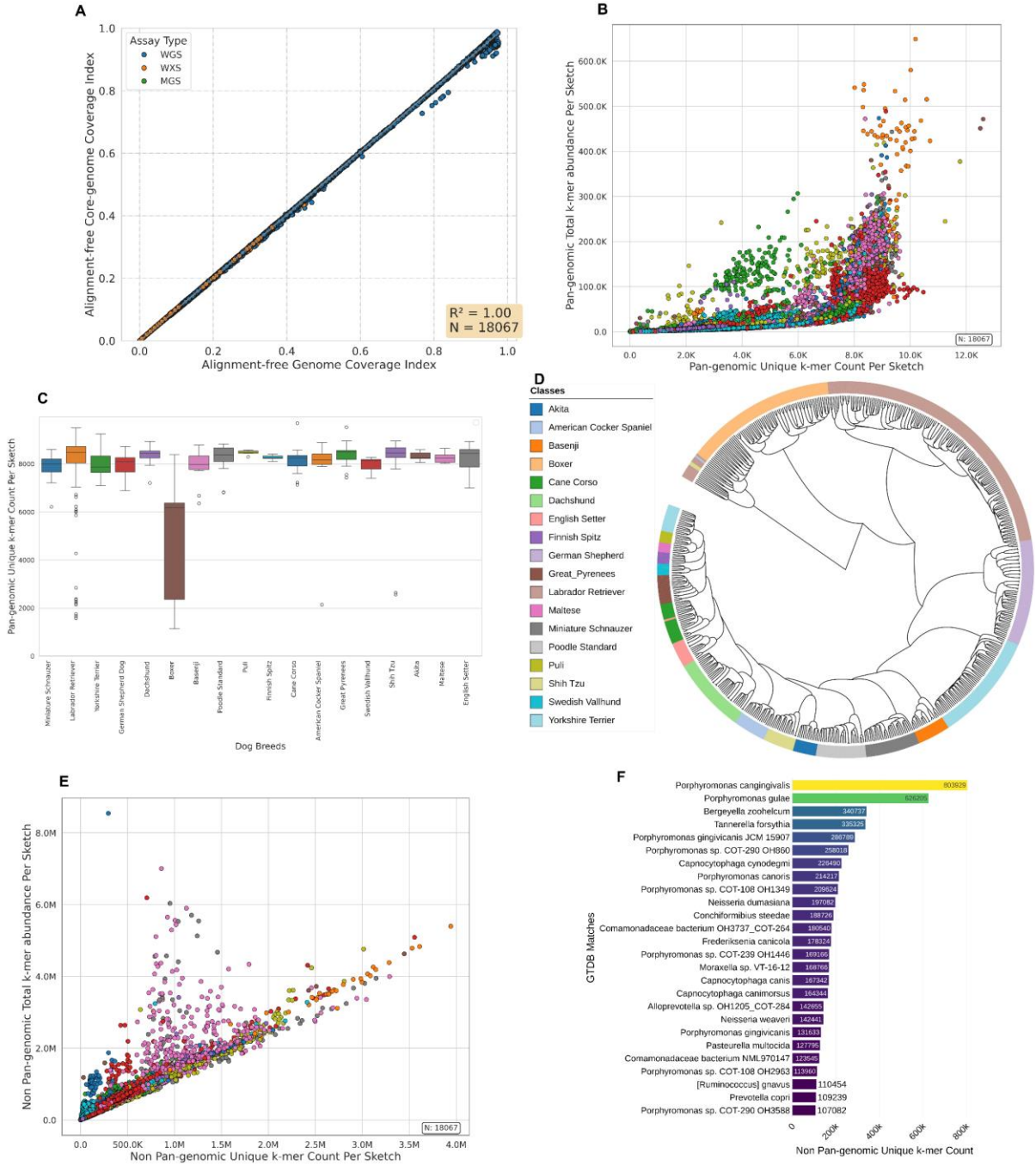



**Figure 2.4 Canine genomics and pangenome exploration: (A)** A scatter plot of 18,067 experiments, colored by Assay Type. X-axis: reference genome coverage index. Y-axis: constructed core-genome coverage index. **(B)** A scatter plot of the non-genomic k-mers

shared with the pangenome k-mer content with the number of unique k-mers on the x-axis and its total abundance on the y-axis, colored by BioProject ID. **(C)** A boxplot showing the unique k-mer count distribution selected from Plot B sketches across 18 randomly selected breeds. (**D)** Hierarchical clusters of Plot C k-mer content colored by breed. **(E)** Scatter plot of non-canine k-mers absent in the reference genome and the pangenome content colored by Bioproject ID. **(F)** A barplot representing the aggregated number of unique k-mers for the top 26 contaminants identified in Plot **E** using the GTDB database.

We selected the best candidate signatures from 2,216 experiments for constructing core genome and pangenome k-mer content sketches **(Supplementary Figure 2. 7)**. K-mers present in at least 50% of these sketches were retained as core genome k-mers, resulting in the retention of 95.6% of the CanFam3.1 reference genome k-mers.

The core genome coverage index was calculated **(Figure 2.4A)** and perfectly correlated to the k-mer-based genomic coverage in **Figure 2.2C**. A pangenome k-mer content was constructed to investigate non-reference canine sequences by retaining k-mers in at least 1% of the sketches. The pangenome k-mer content intersected with k-mers absent in the reference genome (**Figure 2.4B**).

Seventeen dog breeds from different clades were randomly selected from Parker et al. (2017) (Parker et al., 2017), and their pangenome k-mer content was analyzed (**Figure 2.4C**). The results revealed unrepresented genetic diversity in all breeds, with the least diversity in the Boxer breed (the reference genome's breed). Clustering these new k-mers

formed distinct breed-specific clusters (**Figure 2.4D**), with no major Bioprojects batch effects (**Supplementary Figure 2.9**), validating the method for constructing the pangenome k-mer content.

K-mers absent from the reference genome and pangenome k-mer content (**Figure 2.4E**) displayed a diagonal trend of increasing total abundance with a unique k-mer count, consistent with sequencing errors. However, vertical surges in total abundance suggested possible contamination, confirmed by compositional analysis against the GTDB v214 database, identifying contaminants linked to canines, notably oral bacteria, corroborating findings by Ruparell et al (2020) (Ruparell et al., 2020)

# Discussion and Conclusion

Sequence alignment plays a crucial role in sequencing quality control by validating data accuracy through the BAMQC metrics, thus enabling reliable bioinformatics analysis. However, traditional sequence alignment methods, such as BWA, struggle to scale with the vast data outputs and computational demands of next-generation sequencing. In response, we introduce Snipe, a novel alignment-free tool that is both lightweight and highly scalable. Snipe efficiently estimates essential sequence alignment metrics used in sequence content quality control and, for the first time, facilitates SRA-wide exploration of sequence content. This capability significantly enhances the scalability of meta-analysis and pangenome studies, which is critical for understanding the complex relationships between genomes and their associated biological phenomena.

FracMinHash, a key component of Snipe, efficiently represents prokaryotic communities and estimates Average Nucleotide Identity (ANI) using lightweight sketches, maintaining high resolution with reduced computational demands. Although its application to complex mammalian genomes remains underexplored, our results demonstrate its reliability in representing the canine genome, as shown in **Figure 2. 1**, suggesting promising applications for other vertebrates. Snipe efficiently estimates genome coverage, sequencing depth, and mapping rate through the k-mer-based metrics and calculates additional metrics to quantify amplicon enrichment in targeted sequencing. These metrics are computed in under a minute using a minimal memory footprint, making Snipe a valuable tool for quality control and annotation of sequencing experiments at scale.

Snipe's efficiency enables the rapid processing of thousands of sequencing experiments on standard laptops, making it ideal for deployment as a lightweight client-side web application. This web application provides an intuitive interface for visually comparing experiments with public sequencing data, large-scale meta-analyses, and discovering new k-mer content. Snipe also generates error profiles, facilitating contamination analysis, PCR duplicate identification, and differentiation from known sequencing errors. Furthermore, the command-line tool and API offer flexibility for custom integrations, allowing developers to extend Snipe's functionality and adapt it to emerging technologies and methodologies.

Snipe's ability to predict experiment coverage and the effectiveness of additional sequencing significantly transform how ROI is calculated **(Figure 2.3)**. This feature enables researchers to make informed decisions about the cost-effectiveness of additional sequencing, maximizing the value of their research investment. Consequently, Snipe's Fast Mode is implemented to rapidly estimate alignment-based metrics from a small fraction of the raw data, reducing sketching time and enhancing scalability.

Because of our method's scalability, we utilized all the SRA experiments for canines to identify breed-specific k-mers, demonstrating the reliability and potential of our approach for analyzing complex genomes. Dog breeds are genetically complicated due to their recent origins, rapid evolution, and extensive artificial selection, resulting in high genetic variation and admixture between breeds (Streitberger et al., 2012). Despite these challenges, our method successfully constructed a canine pangenome k-mer content and detected unrepresented genetic diversity in 17 randomly selected breeds, with breed-

specific clusters emerging from the analysis as demonstrated in **Figure 2.4**. This achievement highlights the potential of our approach for extending to other complex genomes, enabling the discovery of novel genetic variations and a deeper understanding of genomic diversity across species. Our approach also enabled the detection of contamination and error profiling in the sequencing data, and by removing genomic and pan-genomic k-mers from sketches and focusing on the still highly abundant k-mers, we were able to separate contaminants from new sequence content. This allowed us to distinguish between contaminants and new k-mers that might contribute to the canine pangenome.

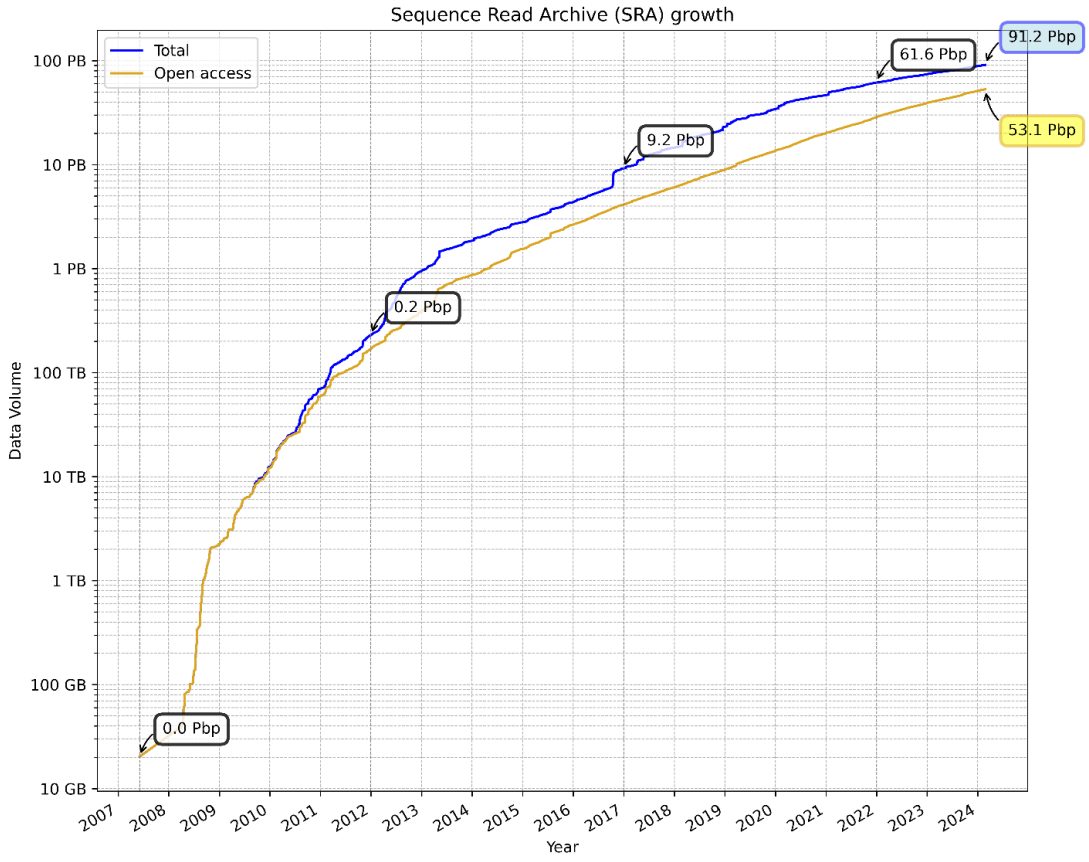
The core genome k-mer content constructed demonstrated a remarkable concordance with reference genome coverage, exhibiting a perfect correlation. This highlights a crucial aspect of our method: its capacity for reference-free applicability. By harnessing k-mer content, our approach can be seamlessly extended to non-model organisms, enabling the analysis of genomes without a pre-existing reference. This feature facilitates the exploration of genetic diversity in previously understudied species, providing valuable insights into their evolutionary trajectories.

While FracMinHash sketching effectively selects representative k-mers, it faces difficulties with short coding genes, resulting in significant gene loss. Despite this, the remaining genes were adequate for the analysis. This highlights the primary limitation of our method: it cannot provide accurate calculations for sequencing experiments with read lengths shorter than the k-mer size or reads that mostly contain Ns, which makes the k-mers invalid **(Supplementary Figure 2. 8)**. In such scenarios, Snipe alerts users with

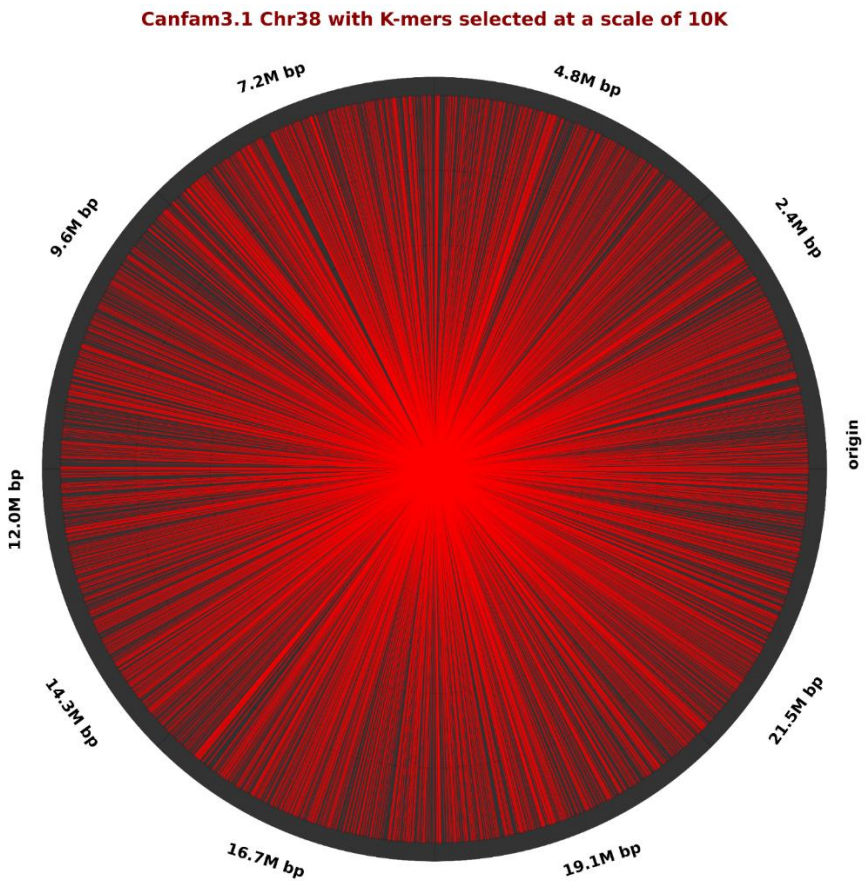
warnings, ensuring they are aware of these potential limitations during analysis. Note that we retained the metagenomic samples (green) in **Figure 2.2** to demonstrate how samples with contamination and low genomic coverage will behave.

In conclusion, our development of Snipe has enabled the efficient and accurate estimation of coverage, depth, and sequence content. Through benchmarking against BWA, we demonstrated the reliability of our metrics, which were further validated through quality control on a large scale using ~19,000 canine SRA experiments. Additionally, Snipe's ability to predict the ROI, facilitate reference-free analysis, and construct pangenome k-mer content sketches enables researchers to make informed decisions about sequencing experiments and identify new sequence content extending to non-model species.

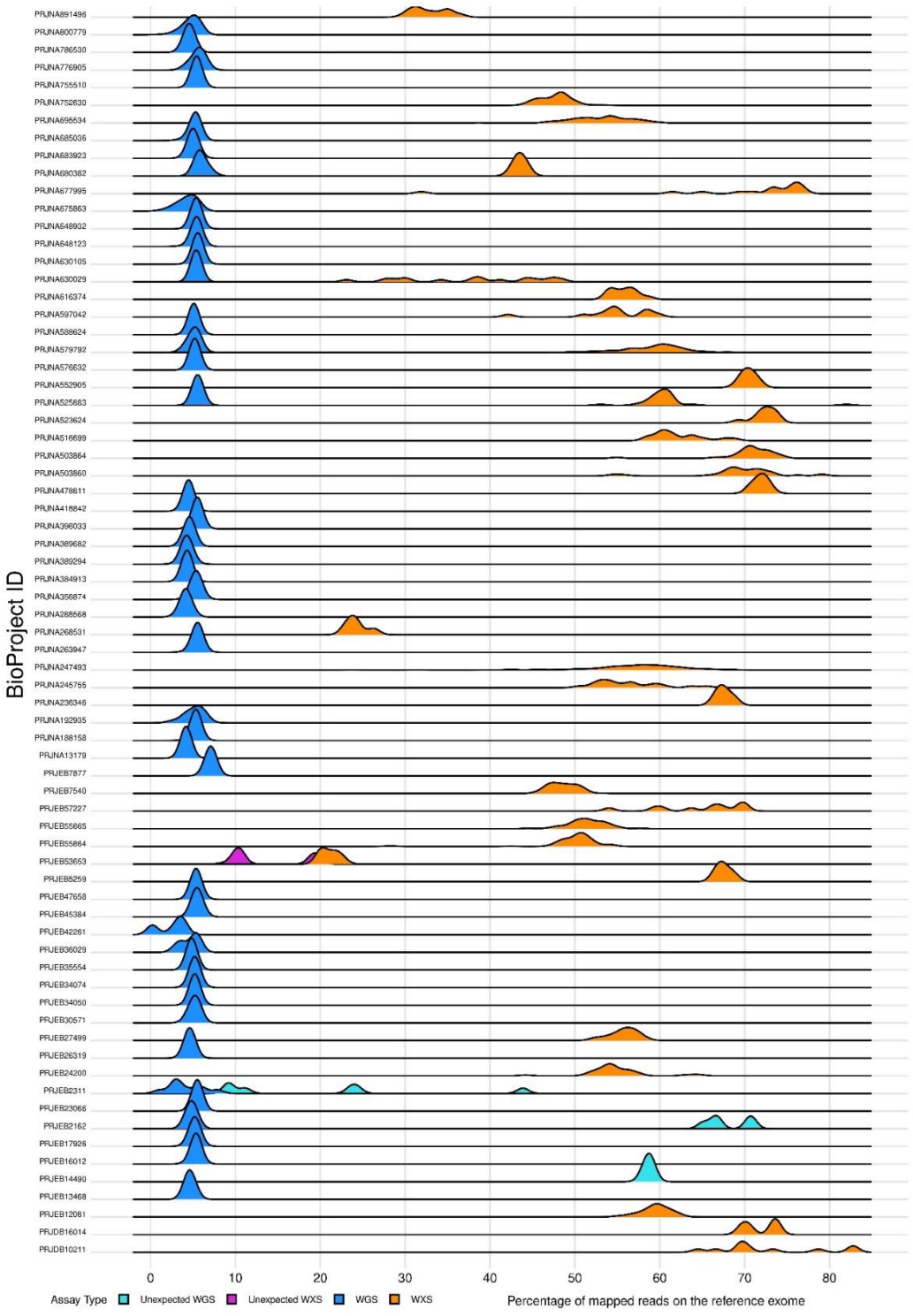
# Supplementary Figures and Tables




**Supplementary Figure 2.1 The growth of the data volume in the Sequence Read Archive (SRA) from 2007 to 2024:** A line chart where the data volume is shown on a logarithmic scale on the y-axis, ranging from 10 GB to 100 PB, with years marked on the x-axis. The blue line indicates the total data volume in the archive, while the yellow line represents the open-access data volume. Key milestones in data volume are annotated on the graph. This plot is updated monthly and is available on GitHub: https://www.mr-eyes.com/sra_size_plot/.

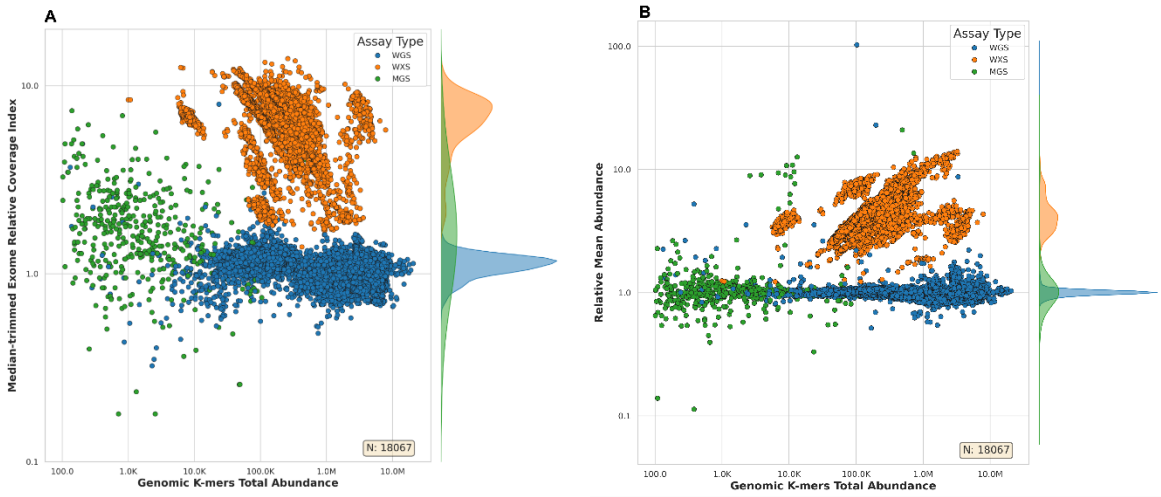



**Supplementary Figure 2.2 Uniform distribution of k-mers selected from the genome by the FracMinHash function:** A circular plot illustrating the distribution of selected k-mers along chromosome 38 of the CanFam3.1 genome assembly. Each red line represents a selected k-mer, while each black line indicates skipped k-mers. The lines radiate from a central origin point, and their positioning around the circle corresponds to specific locations on the chromosome, marked by base pair (bp) distances from the origin. The FracMinHash scale used in this experiment is 10,000, meaning the approximate selection of a single k-mer for each 10,000 k-mer. This is the same scale used for data sketching throughout the manuscript.



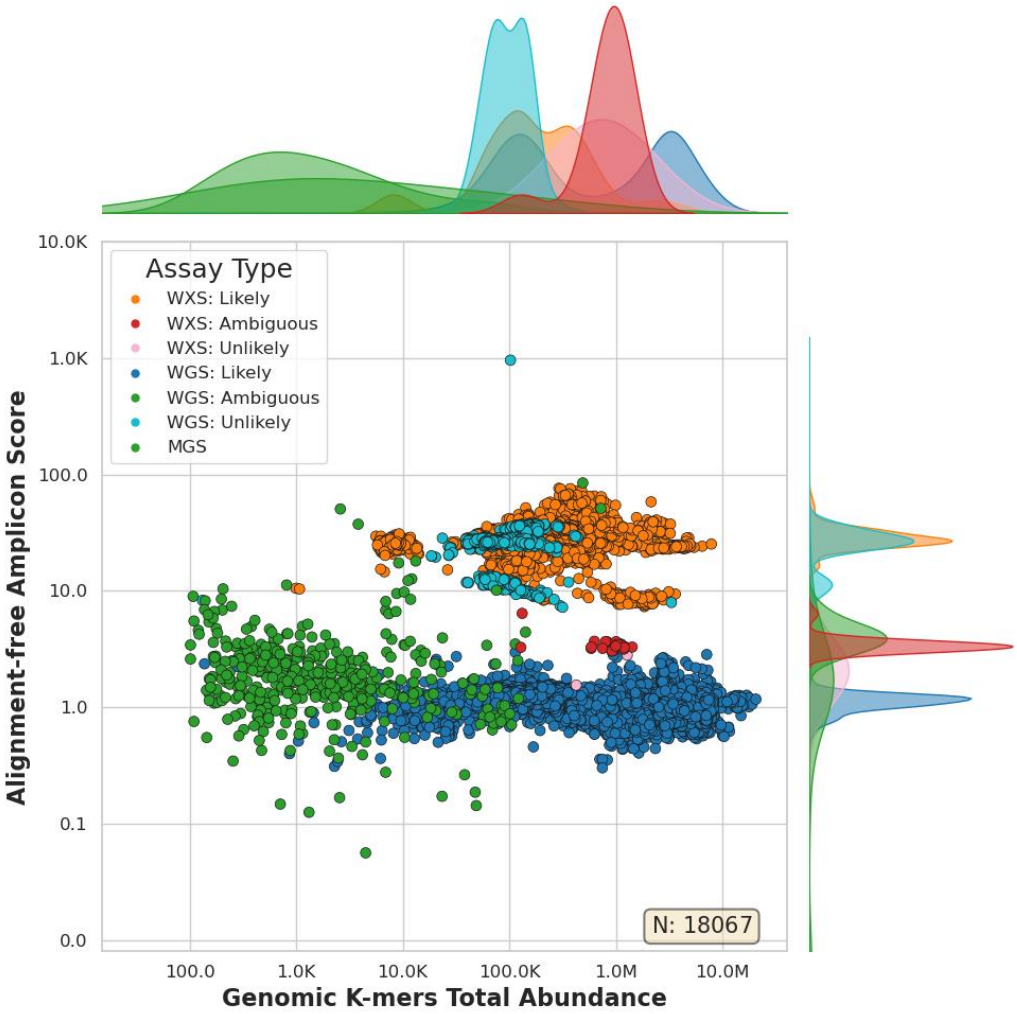

**Supplementary Figure 2.3 Assay-Specific Density of Mapped Reads to the Exome**

Density ridgeline plot with BioProjects on the y-axis and percentage of mapped reads to the exome extracted from Qualimap on the x-axis. Ridge colors differentiate assay types: dark blue for Whole Genome Sequencing (WGS), dark orange for Whole Exome Sequencing (WXS), and introduces color codes for anomalies—cyan for "Unexpected WGS" when the percentage is greater than or equal to 8, and magenta for "Unexpected WXS" when the percentage is less than or equal to 20. These anomalous categories highlight deviations from typical distributions.



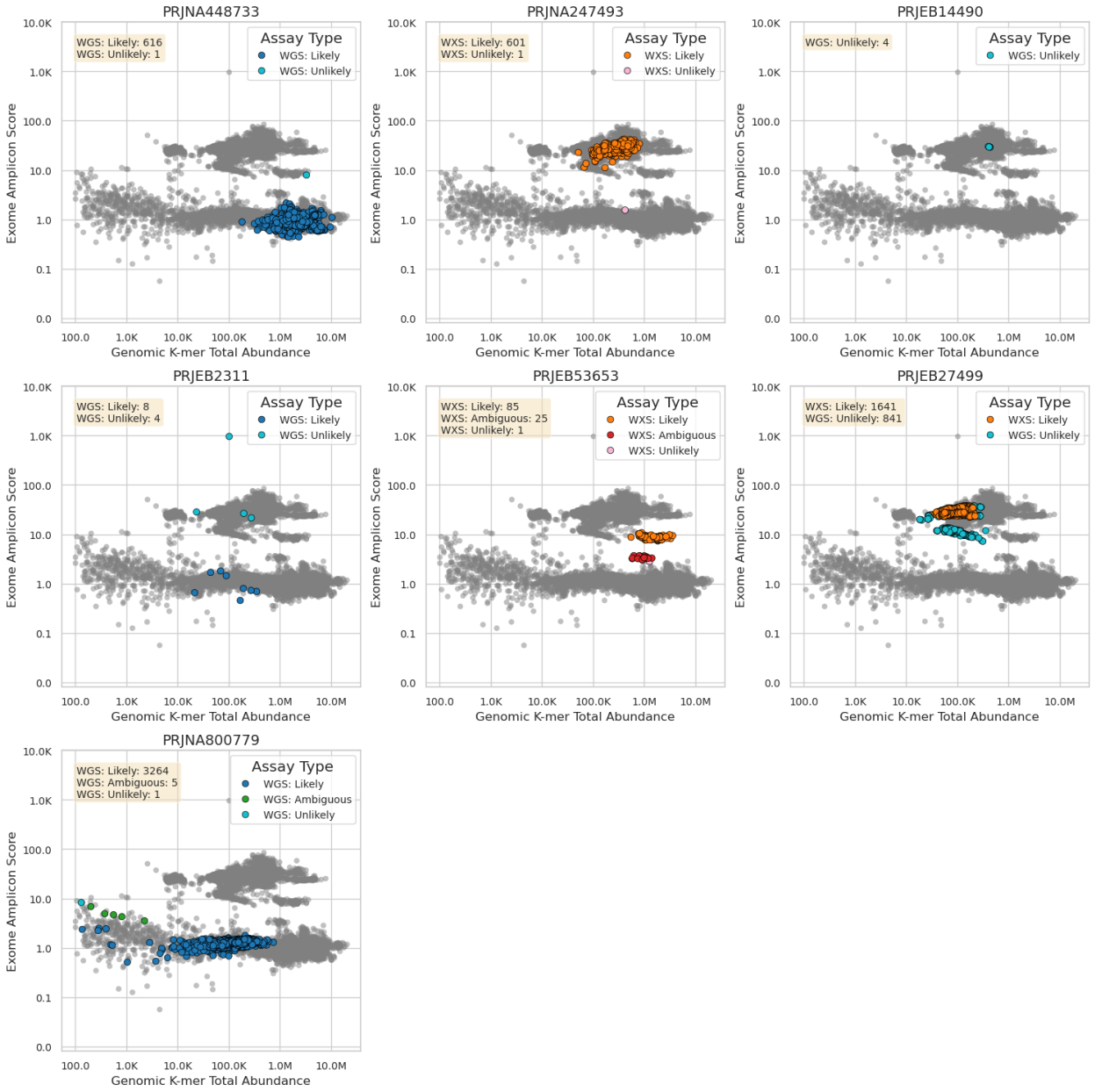

**Supplementary Figure 2.4 Exome Relative Coverage and Abundance in WXS and WGS Assays** (**A)** This plot displays median-trimmed exome relative coverage, where we trim k-mers at or below the median to remove abundant genomic k-mers and isolate exonic k-mers. Dividing exonic by non-exonic k-mers in both WXS and WGS experiments reveals that WXS yields a high relative coverage. In contrast, WGS yields a low relative coverage, allowing for clear differentiation between the assay types. **(B)** shows relative mean abundance by dividing the exome k-mers mean abundance over the genomic mean abundance, and that will also yield a differentiation between both assay

types since WXS has a relatively higher abundance than genomic mean abundance.



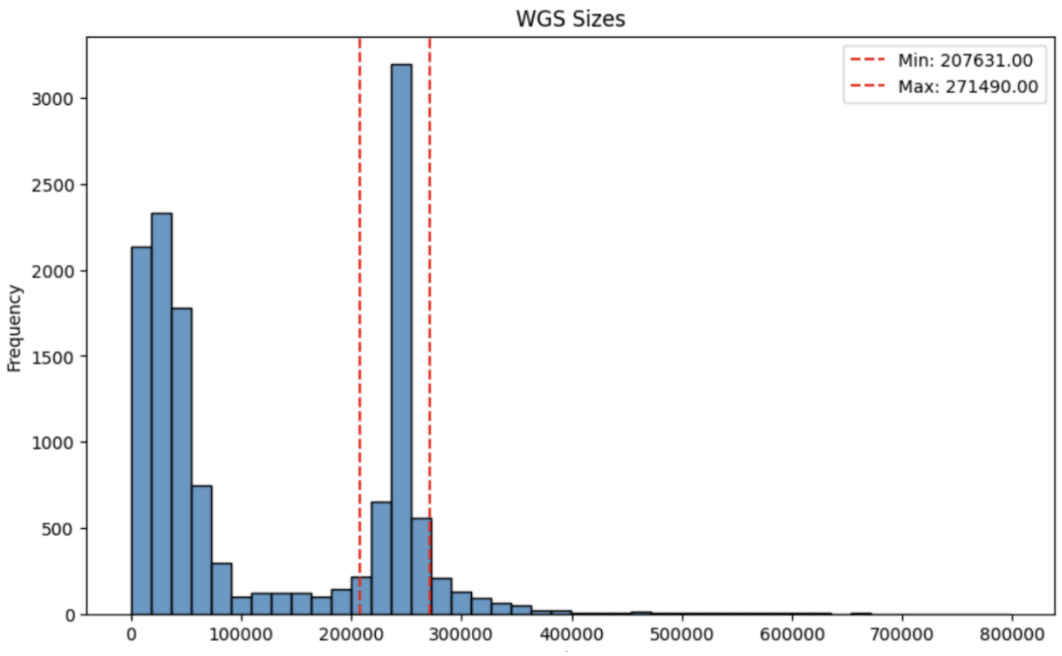

**Supplementary Figure 2.5 Unexpected amplicon scores for multiple experiments** A log-log scatter plot revealing the relationship between Genomic K-mer Total Abundance and Exome Amplicon Score, colored by assay type prediction. Decision thresholds at Exome Amplicon Scores of 3 and 7 categorize assays based on target amplicon detectability. Notably, "WGS: Unlikely" designates data points with Amplicon Scores >

7, characteristic of WXS, while "WXS: Unlikely" represents data points with scores < 3, resembling WGS. Data points between these thresholds are classified as "WXS: Ambiguous" or "WGS: Ambiguous", reflecting categorization uncertainty. Data points convincingly meeting each sequencing type's criteria are labeled "WGS: Likely" or "WXS: Likely".

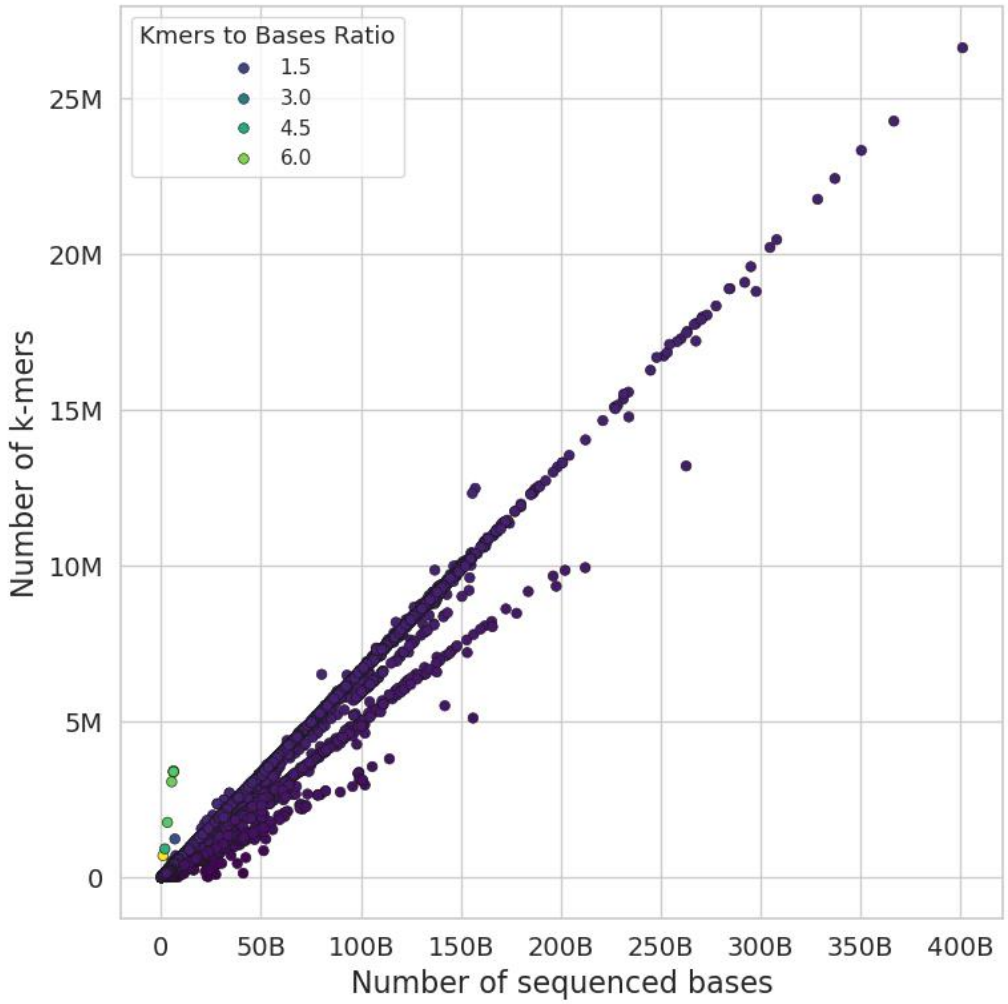



**Supplementary Figure 2.6 Bioproject-Level Analysis of Assay Type QC by Amplicon Score** Scatter plots for each Bioproject showing three categories of experiments: (1) Likely, confirming the original annotation; (2) Unlikely, that we think are showing characteristics of the opposite assay type; and (3) Ambiguous, that we don't have enough data to support a decision regarding them. Assay types of the experiments

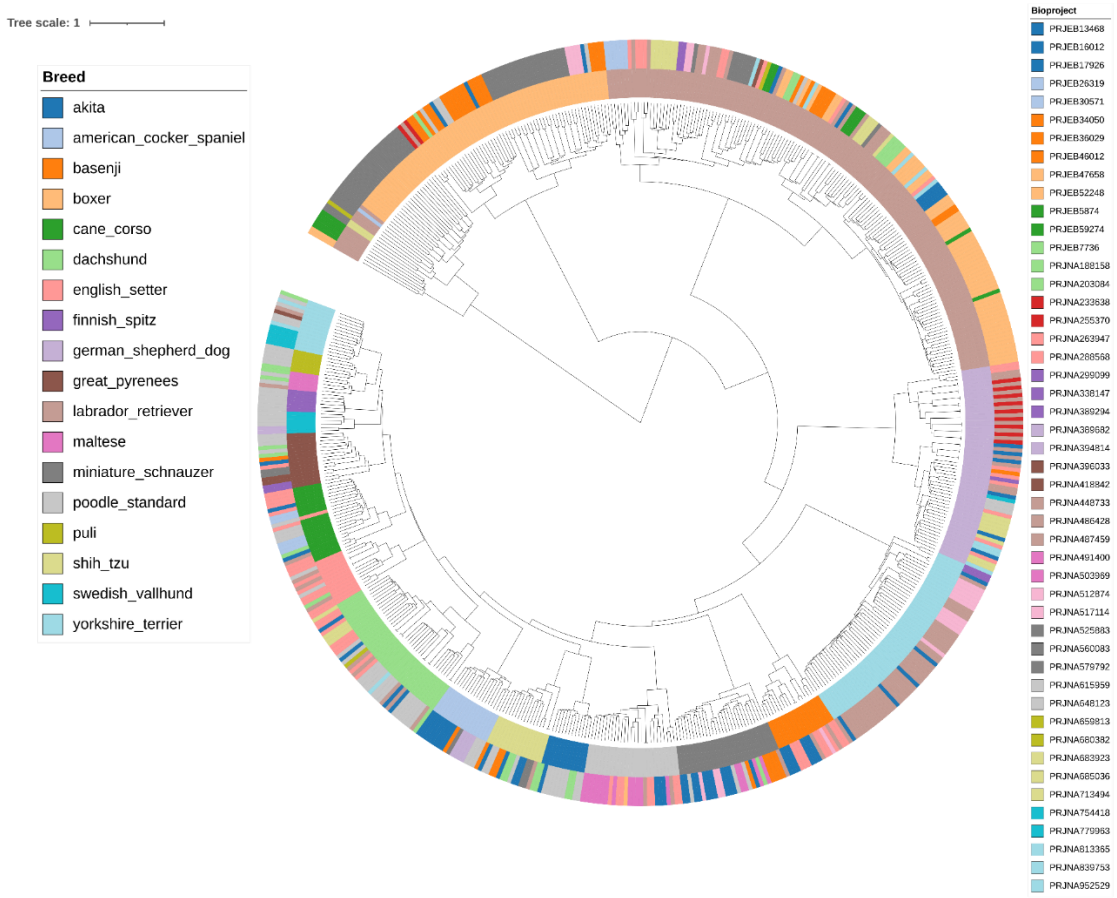
categorized as Unlikely were changed to reflect the opposite one. The Bioproject PRJNA48733 on the bottom left, collected from saliva, might be why some of its experiments overlap with MGS experiments due to contamination. The MGS Bioprojects were not included in this analysis, and they are (PRJEB31756, PRJEB34360, PRJEB38078, PRJEB66438, PRJEB66439, PRJNA407973, PRJNA471557, PRJNA473018) and were confirmed to contain metagenomic sequences from their description on the SRA.




**Supplementary Figure 2. 7 The distribution of Unique k-mer Counts in WGS Experiments:** A histogram of the unique k-mer counts of the WGS experiments, showing the selection boundaries of signatures selected in the pangenome k-mer content construction.




**Supplementary Figure 2. 8 Sketching efficiency assessment with k-mers-to-bases ratio:** with A scatter plot showing the number of bases obtained from the SRA metadata on the x-axis and the total number of k-mers obtained from the sourmash signatures on the y-axis, colored by k-mers-to-bases ratio = $(\frac{k-mers\ total\ abundance}{number\ of\ bases} \times scale)$. Snipe utilizes this metric to assess sketching efficiency, as a low k-mers-to-bases ratio will yield a significant loss in reads and potentially cause inaccurate calculations.




**Supplementary Figure 2.9 Hierarchical clusters of the pan-genomic k-mers showing their breeds and Bioprojects:** This plot is identical to Figure 2.4D, with the addition of a BioProject legend, demonstrating the absence of Bioproject batch effects.

# Chapter 3

# DBRetina: A Scalable Framework for Gene-based Networks Analysis

## Abstract

Gene-set analysis is a fundamental bioinformatics method, but publicly-available gene sets suffer from high duplication rates, hindering robust statistical analyses. Existing databases are fragmented and inconsistent, each containing unique gene sets designed for specific purposes. This fragmentation limits the ability to merge, deduplicate, and connect gene sets through an integrated network.

To address these challenges, we introduce DBRetina, a comprehensive framework for managing and analyzing gene sets and pathway databases. By leveraging gene name hashing, DBRetina creates extensive similarity networks that integrate multiple databases, enabling large-scale comparative analysis and providing a scalable solution for merging and deduplicating thousands of gene sets. To our knowledge, no publicly-available tool offers this comparative analysis and integration level, making DBRetina a valuable resource for bioinformatics research.

# Introduction

Gene set enrichment analysis (GSEA) has become a cornerstone in bioinformatics for interpreting large-scale genomic data. Traditional GSEA methods, introduced by Subramanian et al. in 2005, involve a single dataset and a single gene set database to determine whether a predefined set of genes shows statistically significant, concordant differences between two biological states (Subramanian et al., 2005). While effective, this approach is limited in scope and flexibility, restricting the analysis to a narrow context.

Recent advancements have aimed to overcome these limitations. For instance, the PAGER (Pathway, Annotated-list, and Gene-signature Electronic Repository) web application allows GSEA across multiple gene-set databases, constructing pathway-annotated graphs (PAGs) and establishing new PAG-PAG relationships to facilitate a network-based understanding of gene sets and pathways (Chen et al., 2006, 2006; Yue et al., 2015). However, PAGER still confines users to a single gene set as the query, limiting the flexibility and scope of the analysis.

Beyond PAGER, several other tools have made significant contributions to the field. The DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool integrates functional annotation with enrichment analysis, offering a more comprehensive interpretation of gene lists (Huang et al., 2007). Enrichr, developed by

Kuleshov et al., provides an interactive and user-friendly interface for GSEA, incorporating a wide range of gene set libraries and visualization options (Kuleshov et al., 2016). GeneMANIA, by Warde-Farley et al., predicts gene function by integrating gene networks and functional associations, allowing for a more holistic understanding of gene interactions and their biological implications (Warde-Farley et al., 2010).

Studies have emphasized the importance of integrating multiple types of genomic data for a more comprehensive understanding of biological processes. Huang et al. highlighted the need for tools that can dynamically integrate new gene set databases as they become available, ensuring that analyses remain current and accurate (Huang et al., 2009). Similarly, Liberzon et al. discussed the importance of tools that can keep pace with the growing complexity and volume of genomic data, integrating multiple functionalities to provide a seamless environment for enrichment analysis (Liberzon et al., 2015). Platforms such as GSEA-MSigDB and Cytoscape have made strides in this direction, combining multiple functionalities to offer robust environments for enrichment analysis and network visualization (Shannon et al., 2003; Subramanian et al., 2005).

Despite these advancements, significant gaps remain in the current landscape of GSEA tools. Firstly, most existing tools do not allow for the integration of custom gene set databases, limiting the flexibility of the analysis. Secondly, few tools can utilize multiple gene sets simultaneously to perform GSEA, which is crucial for statistically accurate interpretation for large-scale studies (Lu et al., 2018). Lastly, current tools often lack

advanced downstream analysis functionalities, such as clustering, network analysis, and flexible gene network visualization, which are essential for studying the connectivity between gene sets and target databases. These limitations prevent a comprehensive understanding of the data and reduce the potential insights that can be drawn from complex genomic datasets.

The inability to integrate custom gene set databases limits researchers' ability to tailor their analyses to specific datasets or research questions. Moreover, the lack of frequent updates in online tools like PAGER results in outdated data, affecting the accuracy and relevance of the analysis. Additionally, the absence of advanced downstream analysis features hinders researchers' ability to perform comprehensive meta-analyses, integrating multiple datasets and identifying robust patterns and insights that may not be apparent from individual studies.

This chapter introduces DBRetina, an efficient command-line tool designed to fill these gaps by offering flexibility and control over gene set downstream processing. DBRetina allows users to integrate custom databases, apply various search methods, and utilize downstream functionalities (e.g., querying, filtering, and building targeted gene networks). To demonstrate the capabilities of DBRetina, we will apply it to DisGeNET, a comprehensive repository of disease-gene associations, hosting over 1 million relationships between 20,000+ genes and 30,000+ diseases (Piñero et al., 2015).

# Methods

## Gene sets indexing and pairwise similarity calculations


**Figure 3.1 Schematic representation of the indexing and pairwise similarity calculations.** It starts by indexing the gene sets to create a color table for them. Each gene is associated with a single color, which maps to one or more sources. Finally, a sparse distance matrix is constructed by iterating over the color count table to count the number of shared genes between each two gene sets.

DBRetina indexes two key input formats: the Gene Matrix Transposed (GMT) format and the Association TSV file. The GMT format is standard, encompassing the gene set name in the first column, descriptive metadata in the second, and individual genes in the remaining columns. Conversely, the Association TSV file, with its two-column structure, features the gene set name in the first column and a single gene in the second. The indexing process involves a single iteration over the input data, creating two hash tables. The first hash table maps a gene key to a color, while the second links each color to a combination of associated gene sets, thus enabling a coherent connection between each gene, its color, and its source sets (**Figure 3.1**). The index information is utilized for pairwise comparisons and query functions.

**Algorithm:**Pairwise Comparison in Sparse Datasets
**Require:**Color-to-Datasets Hashtable $H$, Dataset Array $D$
**Ensure:**Sparse Pairwise Matrix $M$
Initialize $M$ as an empty sparse matrix
for each color $c$ in $H$ do
  Let $D_c$ be the array of datasets associated with color $c$
  for each pair $(D_i, D_j)$ in $D_c$ do
    $M[D_i, D_j]$ += ColorCount$(c)$
  end for
end for
return $M$

Algorithm 3.1 Pairwise Distance Calculation

In the development of the pairwise comparison algorithm, we focused on the computational challenges associated with processing large sparse hash sets. A brute-force approach typically suffers from quadratic time complexity ($O(n^2)$) due to exhaustive pairwise comparisons. Our algorithm, however, capitalizes on the sparsity of the data to significantly enhance computational efficiency. The algorithm's key operation involves an iterative process over a color-datasets hash table. Each color in the hash table represents a combination of datasets, and the shared hashes between each pair of datasets are computed by aggregating the counts of colors they share. This is formalized as (SharedHashes($D_i$, $D_j$) = $\sum_{c \in Ci\ j}$ ColorCount(c)) where ($D_i$ and $D_j$) represent pairs of datasets, ($C_{i\ j}$) denotes the set of colors shared by these datasets, and (ColorCount(c)) is

the count associated with each color. The resulting sparse pairwise matrix efficiently encapsulates the shared hashes across all dataset pairs. **Figure 3.1** and **Algorithm 3.1** illustrate the core steps for the pairwise comparisons of highly sparse datasets.

This algorithm allowed us to compare only gene sets with shared genes, streamlining efficiency by achieving one billion comparisons in under 10 minutes on a standard laptop. The algorithm incorporates multi-processing execution for enhanced performance.

## Gene set similarity calculations

Let A and B be two gene sets. We calculate similarity metrics as follows:

- Containment: measures the proportion of shared genes between A and B, normalized by the minimum number of genes in
  either set containment = $(100 \times |A \cap B|) / \min(|A|, |B|)$
- Jaccard distance: a measure of dissimilarity based on the size of the intersection and the sum of the set sizes: Jaccard = $100 \times (|A \cap B| / (|A| + |B| - |A \cap B|))$
- Ochiai distance: an alternative measure of similarity that is less sensitive to set size differences compared to Jaccard distance because it uses the geometric mean of the set sizes, rather than the arithmetic mean, in its calculation:
  Ochiai = $100 \times (|A \cap B| / \sqrt{(|A| \times |B|)})$. This makes Ochiai distance less affected by differences in set size and more focused on the intersection size, compared to Jaccard distance.

We implemented statistical tests to evaluate the significance and strength of associations between gene sets. P-values are calculated using the hypergeometric distribution to determine the statistical significance of overlaps between gene sets. Additionally, we compute odds ratios using Fisher's exact test to quantify the strength of associations. By combining these metrics, we gain an improved understanding of the results.

## Querying pairwise comparisons

Querying the DBRetina gene sets pairwise comparisons can be done using a combination of similarity thresholds and statistical tests to filter the results. It also allows for finding direct and indirect connections among a list of gene sets. **Figure 3. 2** demonstrates the network extension functionality. In **Figure 3. 2A**, a query on gene sets (1, 2, 5) returns their direct connections, while in **Figure 3. 2B**, the `extend` flag retrieves the second layer of connections. This functionality is crucial for identifying related diseases that might indirectly influence the query gene set.

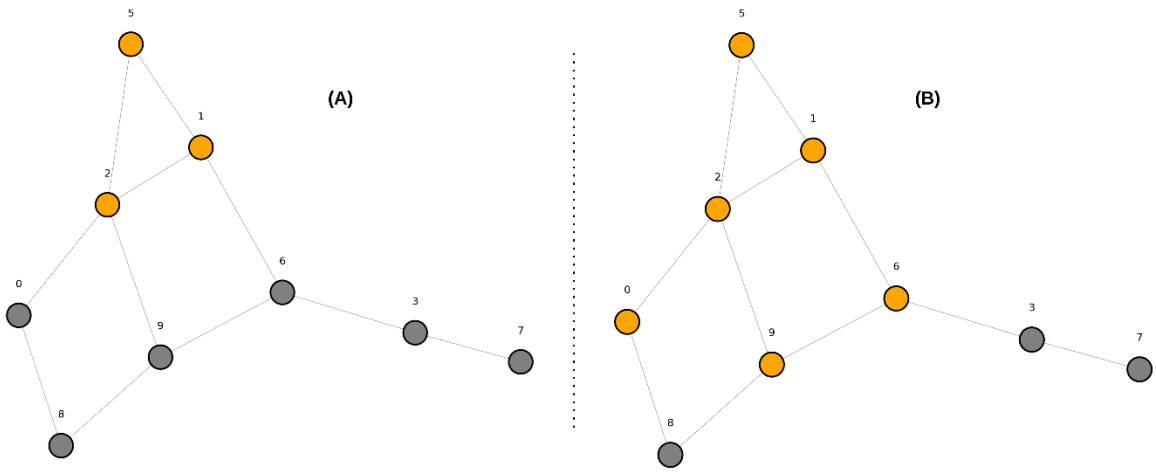



**Figure 3. 2 Demonstration of the query functionality in DBRetina. (A)** The selected nodes (1, 2, 3) are highlighted in orange, demonstrating the node query feature. **(B)** The query is expanded to retrieve the next layer of connected nodes (0, 9, 6), showcasing the ability to fetch additional related nodes.

## Building an interactome

DBRetina builds an interactome graph to visualize the intricate network of gene-gene interactions within a specific study. In this graph, nodes represent genes, and edges represent their interactions, such as regulatory relationships, co-expression patterns, pathways, and other relationships. Building the interactome network with a specific list of gene sets enables focusing on specific gene interactions within a targeted study.

## Bipartite Connections

DBRetina implements a Bipartite graph between a query and a target gene set, and it provides significant insights into the co-occurrence of gene sets. For example, it can find diseases that coexist with sickle cell disease, such as malaria anemia.

# Results and Discussion

DBRetina's capabilities were applied to DisGeNET, demonstrating its potential. A target module of diseases related to Alzheimer's disease was identified, along with the primary causative genes for both Alzheimer's disease and autism spectrum disorder (ASD). This showcases DBRetina's ability to uncover novel disease relationships and provide new insights into disease mechanisms.

## Enhancing disease-disease similarity analysis with DBRetina

DisGeNET provides a comprehensive collection of disease-gene associations, however, its approach to calculating disease-disease similarity has limitations. Specifically, it treats all disease-gene associations equally without accommodating user-defined filters or weights to prioritize or validate them. Furthermore, the search output does not provide insight into the similarity between matching diseases, hindering the identification of meaningful disease clusters.

In contrast, DBRetina offers a more robust and scalable approach to disease-disease similarity analysis. By applying user-defined filters and weights to disease-gene associations, DBRetina enables the identification of high-priority disease relationships. Additionally, DBRetina's visualization capabilities facilitate the exploration of disease clusters and networks.

## DBRetina identified diseases closely related to Alzheimer's disease

In **Figure 3.3**, the heatmap elucidates the disorders sharing many causative genes with several types of Alzheimer's diseases. Unlike traditional search methodologies, the pairwise distance approach organizes the matching hits into distinct clusters. The heatmap reveals three primary clusters: the first encompasses multiple subtypes of Alzheimer's disease, the second is enriched with mental disorders, and the third includes a group of metabolomic and vascular disorders, many of which are recognized as comorbidities of Alzheimer's disease (Craft, 2009; Heilman & Nadeau, 2022). This clustering highlights the genetic overlap and potential shared mechanisms among these disorders, providing insights into the complex interplay between Alzheimer's disease and other related diseases.

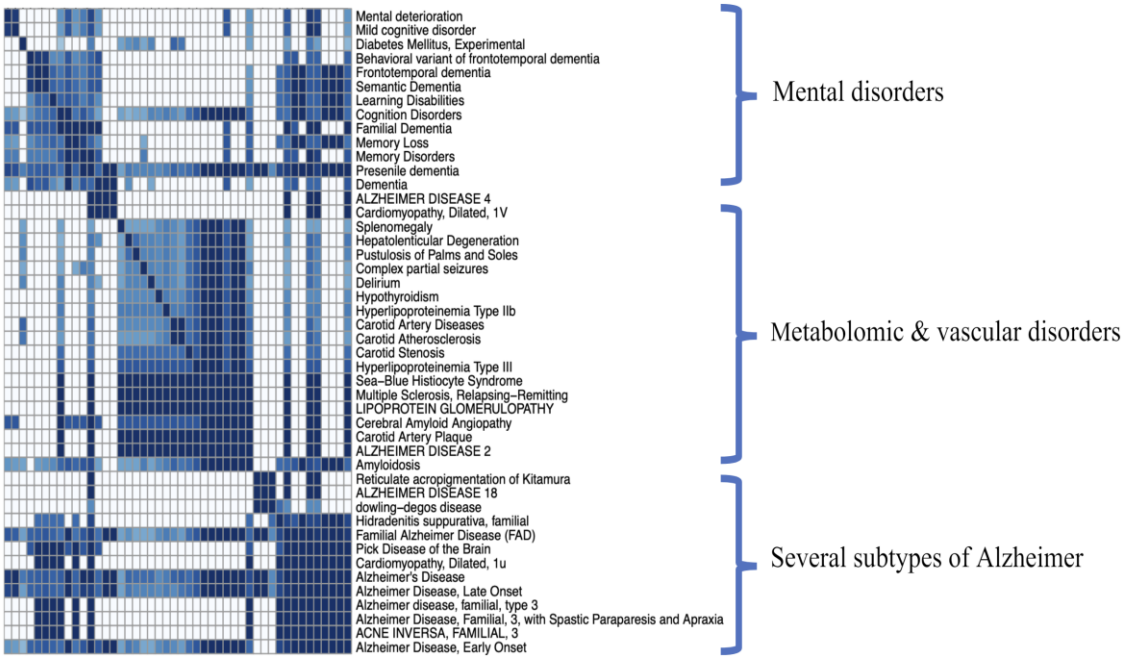


**Figure 3.3 Heatmap displaying clusters of diseases sharing significant causative genes with Alzheimer's subtypes**. Clusters include Alzheimer's disease subtypes, mental disorders, and metabolomic and vascular disorders.

## Causal Genes Identification from Multi-Gene Sets

Identifying causative genes for specific diseases through their gene sets provides valuable insights into the complex relationships between genetic disorders. By examining the gene sets associated with a disease, researchers can determine which genes are responsible for the disease's clinical presentation, treatment response, and potential complications. This process can reveal that a single gene might contribute to multiple diseases, or a particular disease could indicate a genetic predisposition to others. In DBRetina, we highlight this functionality by focusing on two neurological disorders: Alzheimer's disease and ASD. For that study, we used DisGeNET to find the primary causative genes.

In Figure 3.4, we extracted the gene set names that have the word “Alzheimer”, then we queried the DisGeNet pairwise connections to create a network of Alzheimer's disease, followed by creating an interactome of the genes that shows their co-occurrence. We used the 4-path score for nodes to control their size. Three genes were distinguished by that operation (*APOE*, *PSEN1*, and *ADAM10*). *APOE* and *PSEN1* show the strongest co-occurrence in the graph and frequently occur with the *ADAM10* gene (Yuan et al., 2017) (Raulin et al., 2022).

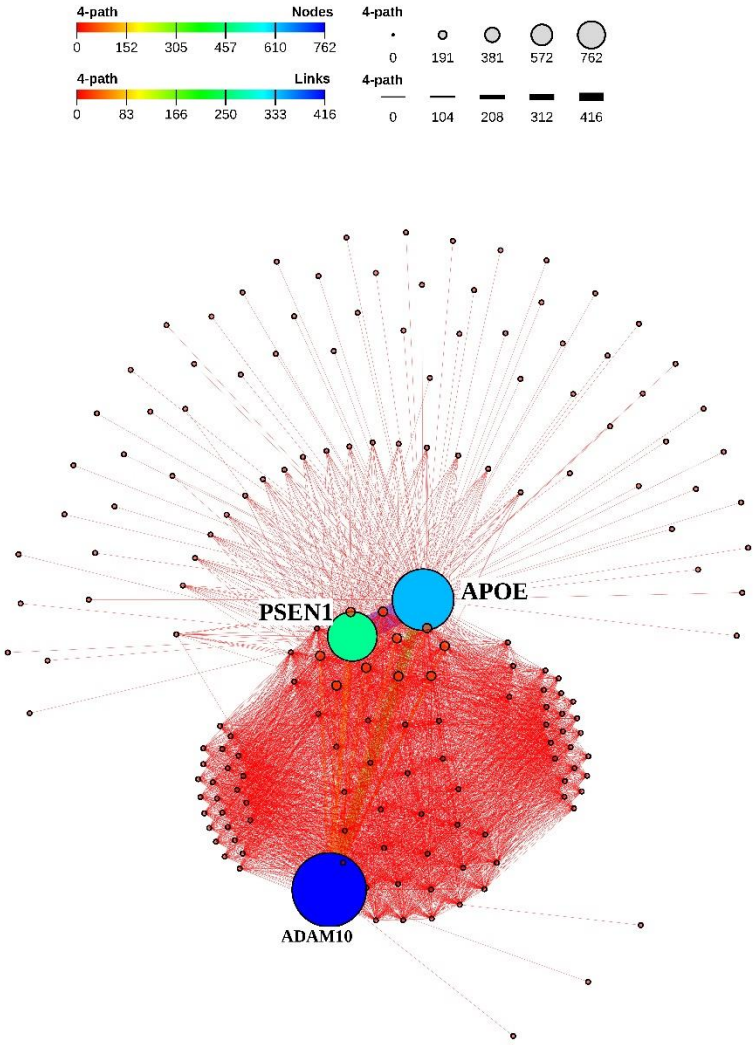



**Figure 3. 4 Network visualization of Alzheimer's disease-related genes.** Nodes represent genes, sized by the number of 4-paths (sequences of four edges connecting five nodes). Links represent interactions, with thickness indicating the number of 4-paths. Key genes.

In **Figure 3. 5**, we visualized the network of autism-related genes by querying the

DisGeNet pairwise connections to identify co-occurrences. The gene interactome graph highlighted key genes based on their betweenness centrality. Nodes are sized according to their betweenness centrality, indicating their importance within the network.

Prominent genes *PTEN*, *CHD8*, *NLGN3*, *NLGN4X*, *SHANK2*, and *PTCHD1* exhibited the highest centrality scores, suggesting their pivotal roles in the network. These six genes are equally connected to the other genes in the network, indicating their widespread influence and significant involvement in autism-related pathways. The thick blue links between these genes denote strong interactions and frequent co-occurrence, emphasizing their interconnected nature.

When performing over-representation analysis, these genes were significantly associated with autism, highlighting their relevance to the disorder. This extensive connectivity underscores the potential shared mechanisms and interactions among these key genes, providing insights into the genetic architecture underlying autism. The relatedness of these genes is supported in the literature, with recent studies showing some of the gene's connection to autism. (Chatterjee et al., 2023; Cummings et al., 2022; Lai et al., 2024; Nguyen et al., 2020)

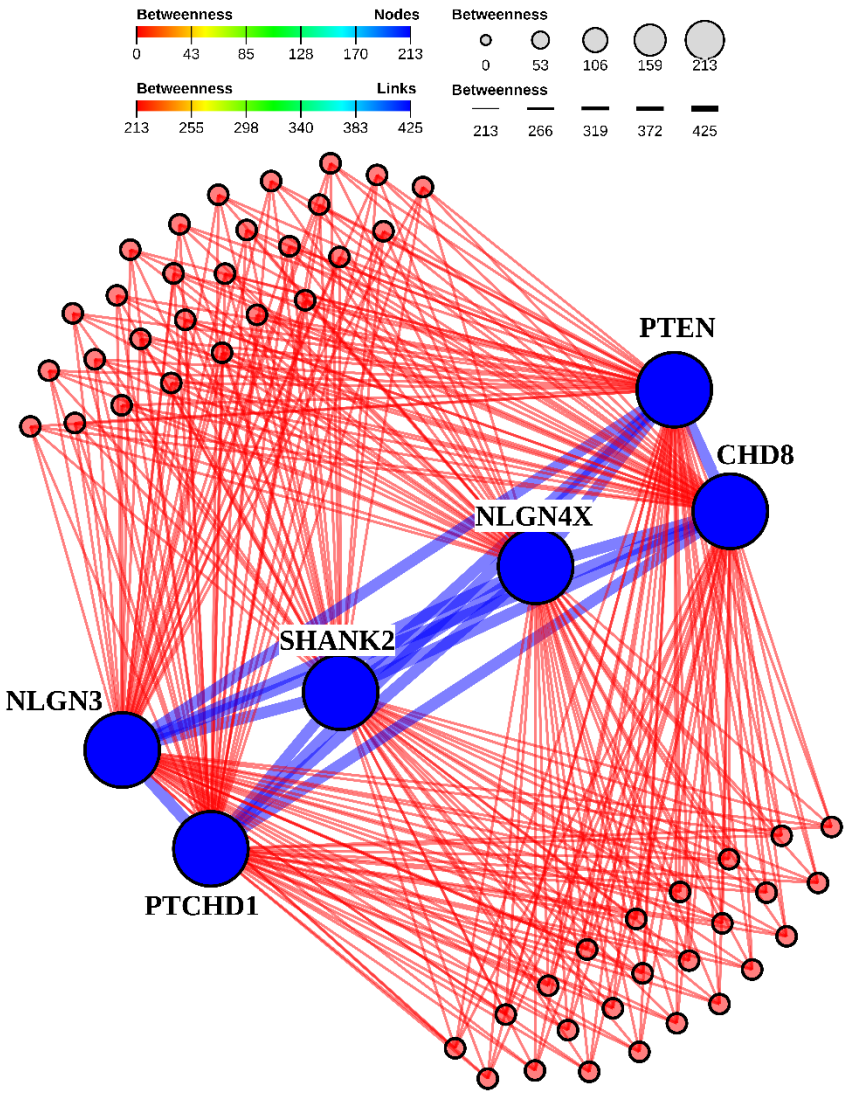



**Figure 3. 5 Network visualization of autism-related genes.** Nodes represent genes, sized by their betweenness centrality. Links represent interactions, with thickness indicating the betweenness centrality of the connections. Key genes, including *PTEN*, *CHD8*, *NLGN3*, *NLGN4X*, *SHANK2*, and *PTCHD1*, are highlighted, showing their central roles. The color gradient from red to blue reflects the increasing betweenness centrality for nodes and links.

# Conclusion

This chapter introduced DBRetina as a flexible command-line tool for processing large-scale gene sets. Although DBRetina offers a wide array of functionalities, we focused on demonstrating its capabilities in querying and filtering pairwise connections and constructing gene interactomes. DBRetina has proven its effectiveness by identifying diseases strongly related to Alzheimer's disease and pinpointing core genes contributing to neurodegenerative diseases such as Alzheimer's disease and ASD.

We believe that DBRetina has the potential to significantly improve how scientists conduct meta-analysis studies at a scale. By integrating diverse datasets—including expression profiles, drug databases, variant databases, and various other gene-based datasets—DBRetina can uncover new direct and indirect relationships among these datasets. This capability positions DBRetina as a powerful tool for advancing our understanding of complex biological systems and diseases.

# Software availability

DBRetina is open-source and available on GitHub (https://github.com/DBRetina/DBRetina).

# Chapter 4

# Conclusion and Future Directions

Every day, massive amounts of biomedical data are generated from various sources such as laboratories, hospitals, research studies, and sequencing facilities. To unlock new research possibilities, it is crucial to have the right tools to analyze and process these large datasets efficiently, using minimal computational resources.

Researchers often rely on publicly available data to validate their experiments, investigate findings, or conduct meta-analysis studies. However, public databases like NCBI do not provide detailed statistics on the quality of uploaded sequencing data due to the computationally intensive sequence alignment process. This lack of information hinders other researchers from fully utilizing this public data.

As introduced in Chapter 2, Snipe offers a lightweight sequence representation that replaces traditional sequence aligners for estimating coverage, depth, and other essential statistics. By integrating Snipe into bioinformatics pipelines as a quality control step, we expect to save significant time and effort in selecting the best candidate experiments, all while minimizing resources. Additionally, Snipe can predict the extra sequencing coverage gain using a small fraction of the raw data, saving time and cost associated with

resequencing the same biomaterial.

We envision Snipe's capabilities extending beyond its current applications to benefit a broader range of species, including humans, farm animals, and other mammals. By utilizing Snipe to construct pan-genome k-mer content and differentiating it from contamination, researchers can lay the groundwork for performing population-level analysis, like finding breed-specific k-mers or common phenotypes in certain populations. Furthermore, Snipe can facilitate investigations into population-level phenotypes.

In Chapter 3, we introduced DBRetina, a framework designed to facilitate the analysis and integration of gene sets through multiple approaches. We envision DBRetina's utility in refining the focus of certain clinical investigations by identifying specific pathways or genes for targeted research. By consolidating and linking diverse gene sets from various sources, DBRetina offers a time-saving advantage, enabling researchers to efficiently navigate complex gene relationships and prioritize experimental design.

As a future direction, we intend to develop a comprehensive database integrating symptoms, pathways, diseases, MeSH terms, expression profiles, and drug databases, leveraging the DBRetina framework. This database will be designed as a readily queryable network, facilitating the exploration of complex biological relationships.

Furthermore, we aim to harness the advancements in large language models (LLMs) and artificial intelligence to enhance accessibility for scientists and biologists without expertise in network queries. To achieve this, we plan to incorporate an AI-driven system capable of translating research questions into network queries, summarizing findings, and visualizing connections, thereby streamlining the discovery process.

By providing our introduced tools as easily usable and flexible methods to biologists and bioinformaticians, we aim to empower researchers to tackle complex biomedical questions more efficiently and effectively, bridging the gap between data generation and knowledge discovery. By leveraging scalable computational methods, we ultimately aim to contribute to the betterment of human lives and the world at large by accelerating the pace of scientific progress and its translation into tangible benefits for society.

# References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bansal, G., Narta, K., & Teltumbade, M. R. (2018). Next-Generation Sequencing: Technology, Advancements, and Applications. In A. Shanker (Ed.), *Bioinformatics: Sequences, Structures, Phylogeny* (pp. 15–46). Springer. https://doi.org/10.1007/978-981-13-1562-6_2

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Chatterjee, I., Getselter, D., Ghanayem, N., Harari, R., Davis, L., Bel, S., & Elliott, E. (2023). CHD8 regulates gut epithelial cell function and affects autism-related behaviors through the gut-brain axis. *Translational Psychiatry*, *13*(1), 305. https://doi.org/10.1038/s41398-023-02611-2

Chaung, K., Baharav, T. Z., Henderson, G., Zheludev, I. N., Wang, P. L., & Salzman, J. (2023). SPLASH: A statistical, reference-free genomic algorithm unifies biological discovery. *Cell*, *186*(25), 5440-5456.e26.

https://doi.org/10.1016/j.cell.2023.10.028

Chen, J. Y., Shen, C., Yan, Z., Brown, D. P. G., & Wang, M. (2006). A systems biology case study of ovarian cancer drug resistance. *Computational Systems Bioinformatics. Computational Systems Bioinformatics Conference*, 389–398.

Craft, S. (2009). The Role of Metabolic Disorders in Alzheimer Disease and Vascular Dementia: Two Roads Converged. *Archives of Neurology*, *66*(3), 300–305. https://doi.org/10.1001/archneurol.2009.27

Cummings, K., Watkins, A., Jones, C., Dias, R., & Welham, A. (2022). Behavioural and psychological features of PTEN mutations: A systematic review of the literature and meta-analysis of the prevalence of autism spectrum disorder characteristics. *Journal of Neurodevelopmental Disorders*, *14*(1), 1. https://doi.org/10.1186/s11689-021-09406-w

Dinov, I. D. (2016). Volume and Value of Big Healthcare Data. *Journal of Medical Statistics and Informatics*, *4*, 3. https://doi.org/10.7243/2053-7662-4-3

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*(20), 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Gasic, V., Pavlovic, D., Stankovic, B., Kotur, N., Zukic, B., Pavlovic, S., Gasic, V., Pavlovic, D., Stankovic, B., Kotur, N., Zukic, B., & Pavlovic, S. (2021). Pharmacogenomics and Pharmacotranscriptomics of Glucocorticoids in Pediatric

Acute Lymphoblastic Leukemia. In *Corticosteroids—A Paradigmatic Drug Class*. IntechOpen. https://doi.org/10.5772/intechopen.98887

*Genomic Data Resources: Curation, Databasing, and Browsers | Learn Science at Scitable*. (n.d.). Retrieved May 12, 2024, from https://www.nature.com/scitable/topicpage/genomic-data-resources-challenges-and-promises-743721/

Heilman, K. M., & Nadeau, S. E. (2022). Emotional and Neuropsychiatric Disorders Associated with Alzheimer's Disease. *Neurotherapeutics*, *19*(1), 99–116. https://doi.org/10.1007/s13311-021-01172-w

Hera, M. R., Pierce-Ward, N. T., & Koslicki, D. (2023). Deriving confidence intervals for mutation rates across a wide range of evolutionary distances using FracMinHash. *Genome Research*, *33*(7), 1061–1068. https://doi.org/10.1101/gr.277651.123

Hooper, D. C., & Jacoby, G. A. (2015). Mechanisms of drug resistance: Quinolone resistance. *Annals of the New York Academy of Sciences*, *1354*(1), 12–31. https://doi.org/10.1111/nyas.12830

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. (2007). DAVID

Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research*, *35*(suppl_2), W169–W175. https://doi.org/10.1093/nar/gkm415

Irber, L., Brooks, P. T., Reiter, T., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., & Brown, C. T. (2022). *Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers* (p. 2022.01.11.475838). bioRxiv. https://doi.org/10.1101/2022.01.11.475838

Irber, L., Pierce-Ward, N. T., & Brown, C. T. (2022). *Sourmash Branchwater Enables Lightweight Petabyte-Scale Sequence Search* (p. 2022.11.02.514947). bioRxiv. https://doi.org/10.1101/2022.11.02.514947

Kawulok, J., & Deorowicz, S. (2015). CoMeta: Classification of Metagenomes Using k-mers. *PLOS ONE*, *10*(4), e0121453. https://doi.org/10.1371/journal.pone.0121453

Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics*, *33*(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z.,

Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90-97. https://doi.org/10.1093/nar/gkw377

Lai, W., Zhao, Y., Chen, Y., Dai, Z., Chen, R., Niu, Y., Chen, X., Chen, S., Huang, G., Shan, Z., Zheng, J., Hu, Y., Chen, Q., Gong, S., Kang, S., Guo, H., Ma, X., Song, Y., Xia, K., … Shi, L. (2024). Autism patient-derived SHANK2BY29X mutation affects the development of ALDH1A1 negative dopamine neuron. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-024-02578-6

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Liu, S., & Koslicki, D. (2022). CMash: Fast, multi-resolution estimation of k-mer-based Jaccard and containment indices. *Bioinformatics*, *38*(Supplement_1), i28–i35. https://doi.org/10.1093/bioinformatics/btac237

Lu, W., Wang, X., Zhan, X., & Gazdar, A. (2018). Meta-Analysis Approaches to Combine Multiple Gene Set Enrichment Studies. *Statistics in Medicine*, *37*(4), 659–672. https://doi.org/10.1002/sim.7540

Manekar, S. C., & Sathe, S. R. (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*, *7*(12), giy125. https://doi.org/10.1093/gigascience/giy125

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

*MurmurHash3 · aappleby/smhasher Wiki*. (n.d.). Retrieved May 13, 2024, from

https://github.com/aappleby/smhasher/wiki/MurmurHash3

Nguyen, T. A., Lehr, A. W., & Roche, K. W. (2020). Neuroligins and Neurodevelopmental Disorders: X-Linked Genetics. *Frontiers in Synaptic Neuroscience*, *12*. https://doi.org/10.3389/fnsyn.2020.00033

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biology*, *20*(1), 232. https://doi.org/10.1186/s13059-019-1841-x

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016a). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016b). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Parker, H. G., Dreger, D. L., Rimbault, M., Davis, B. W., Mullen, A. B., Carpintero-Ramirez, G., & Ostrander, E. A. (2017). Genomic analyses reveal the influence of geographic origin, migration and hybridization on modern dog breed development. *Cell Reports*, *19*(4), 697–708. https://doi.org/10.1016/j.celrep.2017.03.079

Patro, R., & Salmela, L. (2020). Algorithms meet sequencing technologies – 10th edition of the RECOMB-Seq workshop. *iScience*, *24*(1), 101956. https://doi.org/10.1016/j.isci.2020.101956

*Picard Tools—By Broad Institute*. (n.d.). Retrieved May 6, 2024, from https://broadinstitute.github.io/picard/

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019a). *Large-scale sequence comparisons with* sourmash (8:1006). F1000Research. https://doi.org/10.12688/f1000research.19675.1

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019b). *Large-scale sequence comparisons with* sourmash (8:1006). F1000Research. https://doi.org/10.12688/f1000research.19675.1

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., & Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database: The Journal of Biological Databases and Curation*, *2015*, bav028. https://doi.org/10.1093/database/bav028

Raulin, A.-C., Doss, S. V., Trottier, Z. A., Ikezu, T. C., Bu, G., & Liu, C.-C. (2022). ApoE in Alzheimer's disease: Pathophysiology and therapeutic strategies. *Molecular Neurodegeneration*, *17*(1), 72. https://doi.org/10.1186/s13024-022-00574-4

Rizk, G., Lavenier, D., & Chikhi, R. (2013). DSK: K-mer counting with very low

memory usage. *Bioinformatics*, *29*(5), 652–653. https://doi.org/10.1093/bioinformatics/btt020

Rowe, W. P. M. (2019). When the levee breaks: A practical guide to sketching algorithms for processing the flood of genomic data. *Genome Biology*, *20*(1), 199. https://doi.org/10.1186/s13059-019-1809-x

Ruparell, A., Warren, M., Staunton, R., Deusch, O., Dobenecker, B., Wallis, C., O'Flynn, C., McGenity, P., & Holcombe, L. J. (2020). Effect of feeding a daily oral care chew on the composition of plaque microbiota in dogs. *Research in Veterinary Science*, *132*, 133–141. https://doi.org/10.1016/j.rvsc.2020.05.001

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sims, G. E., Jun, S.-R., Wu, G. A., & Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, *106*(8), 2677–2682. https://doi.org/10.1073/pnas.0813249106

Streitberger, K., Schweizer, M., Kropatsch, R., Dekomien, G., Distl, O., Fischer, M. S., Epplen, J. T., & Hertwig, S. T. (2012). Rapid genetic diversification within dog breeds as evidenced by a case study on Schnauzers. *Animal Genetics*, *43*(5), 577–586. https://doi.org/10.1111/j.1365-2052.2011.02300.x

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Vinga, S., & Almeida, J. (2003). Alignment-free sequence comparison—A review. *BIOINFORMATICS*, *19*(4), 513–523. https://doi.org/10.1093/bioinformatics/btg005

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, *38*(suppl_2), W214–W220. https://doi.org/10.1093/nar/gkq537

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2

Yuan, X.-Z., Sun, S., Tan, C.-C., Yu, J.-T., & Tan, L. (2017). The Role of ADAM10 in Alzheimer's Disease. *Journal of Alzheimer's Disease: JAD*, *58*(2), 303–322. https://doi.org/10.3233/JAD-170061

Yue, Z., Kshirsagar, M. M., Nguyen, T., Suphavilai, C., Neylon, M. T., Zhu, L., Ratliff,

T., & Chen, J. Y. (2015). PAGER: Constructing PAGs and new PAG–PAG relationships for network biology. *Bioinformatics*, *31*(12), i250–i257. https://doi.org/10.1093/bioinformatics/btv265

Zhang, F., Flickinger, M., Taliun, S. A. G., Consortium, I. P. G., Abecasis, G. R., Scott, L. J., McCaroll, S. A., Pato, C. N., Boehnke, M., & Kang, H. M. (2020). Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Research*, *30*(2), 185–194. https://doi.org/10.1101/gr.246934.118

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *GENOME BIOLOGY*, *18*. https://doi.org/10.1186/s13059-017-1319-7