National Human Genome Research Institute

The Creation of a Relational Database for the Storage and Analysis of Structural Variation

Nicolas J. W. Fishman[1], Brian W. Davis[1], Danielle M. Karyadi[1], and Elaine A. Ostrander[1]

[1]Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

Comparative Genetics Section

Abstract

Structural variation (SV) within the genome is likely a primary mechanism leading to phenotypic diversity. This is particularly evident within the canine genome where strong human selection for a variety of traits has led to not only dog breeds of varying appearance and behavior patterns, but also differential disease susceptibility. Unlike single nucleotide variation (SNV), SV require a much more algorithmic and therefore inferential approach for variant identification and genotyping. While several methods have been developed to identify SVs, each has limitations and as of yet no gold standard has emerged. As such, a consensus-based approach combining algorithms is needed to increase variant detection and genotyping confidence. Each of the SV methods currently available produces a unique and non-standardized output, implements algorithm-specific support metrics, and predicts SVs from different categories that do not necessarily correspond one to another. Furthermore, the overlapping though not fully analogous techniques produce results with varying levels of within or across algorithm concordance. This is, in part, because current variant callers often specialize in interrogation of specific aspects of SV, for example read depth quantitation or identification of split reads.

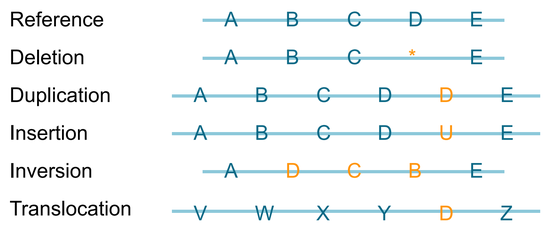
To overcome these limitations, we are developing SVStore, an integrated suite of tools for genomicists interested in interrogating SV within the genome. It is a relational database-driven tool that allows for the easy compilation of data from five SV detection algorithms: Delly, CNVnator, Lumpy, breakdancer, and pindel. First, a set of scripts converts the heterogeneous output of the most common SV detection algorithms into the standard Variant Call Format (VCF) to normalize algorithm output. Second, records are loaded into a relational database that utilizes the simplicity and utility of VCF, allowing for the centralized collection and curation of algorithm output. This database also allows users to leverage the power of SQL queries to search, sort and analyze the vast quantity and diversity of information present in contemporary genomic data sets. To simplify data interaction, we also created a graphical user interface to allow common queries to be executed without the need for a deep understanding of database structure. Results of any queries from the web interface can be viewed through a standard browser, or as downloadable output in VCF format. These three components make SVStore a powerful tool for enabling the analysis and study of SV.

1. Structural Variation

Any string of variants - minimum of 50 base pairs in size

- Play a large role in the development of disease in individuals
- Make up a significant amount of the variation between individuals

Types of Structural Variation



Structural variants can be difficult to detect

- Algorithms exist to identify variants each with strengths / weaknesses

Problematic to compare between algorithms

- Each algorithm provides different support metrics
- Output formats and constituent data also differs

Within an algorithm, nontrivial to compare between individuals

- Breakpoints specificity - support metrics - variation in genome assemblies

Delly | CNVnator | Lumpy | Pindel | Breakdancer

Current Canine Data

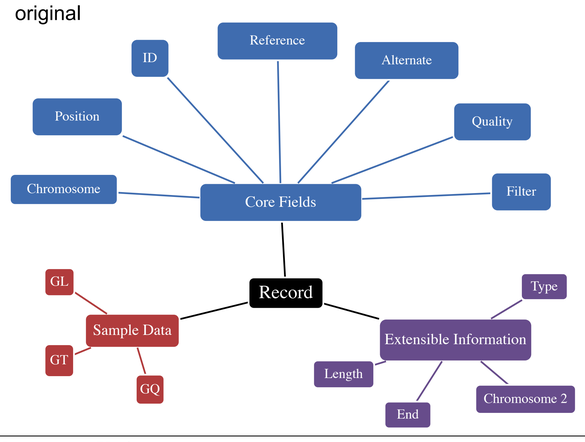
Large Deletions	1,605,765
Large Insertions	149,375
Duplications	274,522
Inversions	2,987,786
Translocations	3,972,081
LINE Insertions	62,312
SINE Insertions	226,593

From 500 individuals:
- 125 breeds
- Feral populations
- Wolf
- Coyote

2. Heterogeneity of Data

Variant Call Format (VCF)

- The standard for storing variant call data
- Creates a homogenous, searchable dataset
- Has flexibility to accommodate all information from original



Tool	Variant
Standard VCF	CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
Pindel	@HWI-EAS255_8291_FC30GRN_PE:3:61:366:1255/2 CTGAAACTTGAGTTTTTTTTTTCTTTTTAATCACCATCACAGTGAAGGATACGTTG + 20 10023140 60 200 COLO-829-BL
Breakdancer	1 30659 56+55- 1 31539 52+116- DEL 804 99 17 /data4/users/vjaganna/dog/BROAD/FASTQ/LE0048 LE0048.cleaned.dedup.baml3 1.35
CNVnator	Genotype chr1:1-13000 5173T.root 0.377524 0.368896
Delly2	Chr1 2747 DUP00000001 N <DUP> 0 LowQual IMPRECISE;SVTYPE=DUP; SVMETHOD=EMBL.DELLYv0.7.3;CHR2=chr1;END=3722;INSLEN=0;PE=62;MAPQ=5; CT=5to3;CIPOS=-40,40;CIEND=-40,40 GT:GL:GQ:FT:RCL:RC:RCR:CN:DR:DV:RR:RV 0/0:0,-1000,-1000:10000:PASS:7522:158994:82224:4:4894:89:0:0

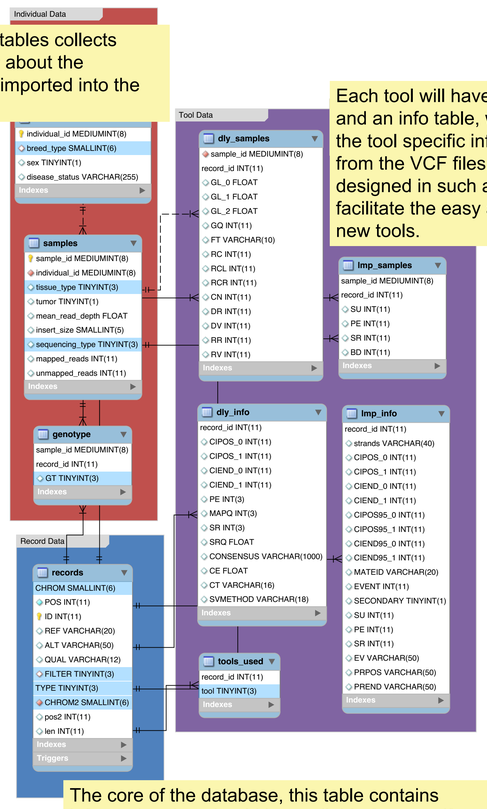
Importation Toolset

- VCF allows for the collection of diverse data in a standard file type
- Solves the problem of loading different data types into a database
- Informs structure for the design of the relational structure itself.

3. Database Design

This set of tables collects information about the individuals imported into the database.

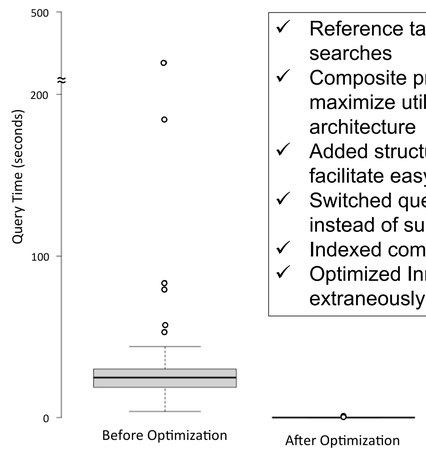
Each tool will have a sample and an info table, which store the tool specific information from the VCF files. This is designed in such a way as to facilitate the easy addition of new tools.



The core of the database, this table contains information that should be available for most if not all records put into the system.

- MySQL-based relational database, InnoDB tables
- Table design inspired by VCF data structure
  - Core: normalized, searchable data in the *RECORDS* table
  - Tool-specific information in the *_INFO* and *_SAMPLES* tables
- High level individual and sample data maintained
- Modular design facilitates easy addition of new tools

Optimization



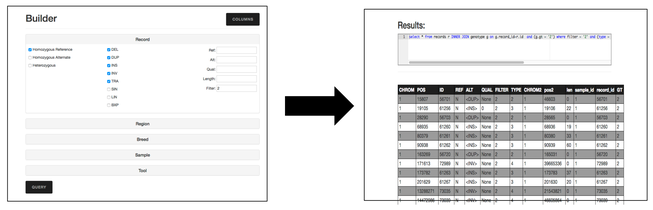
- ✓ Reference tables to avoid text searches
- ✓ Composite primary keys to maximize utility of InnoDB tree architecture
- ✓ Added structured joins to facilitate easy joins
- ✓ Switched queries to joins instead of subqueries
- ✓ Indexed common searches
- ✓ Optimized InnoDB settings for extraneously large tables

4. User Interfaces

Graphical Interface

- Simplified SQL query builder with specialized functions
- Output option in VCF file

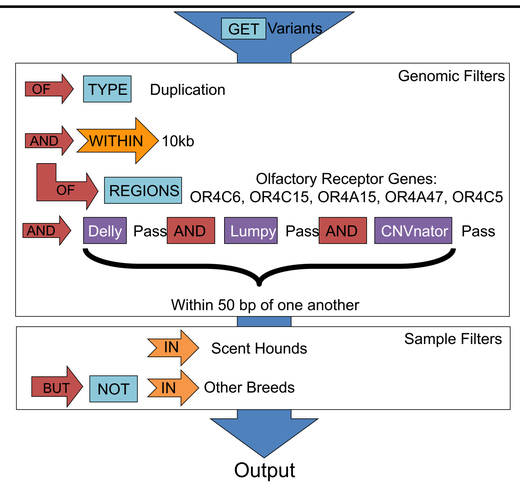
Example of query built with web tool:



Command Line Interface

- Fast support for unix terminal
- Fully implemented API with support for developing external tools and processing functions

5. Querying

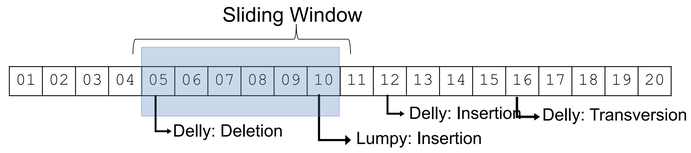


6. Post Processing

False Positives in Structural Variant Callers

- Structural Variant callers all tend to high false positive rates
- Trade-off: always choose to include everything, rather than miss anything
- Makes research nearly impossible, necessitates filtering

Sliding Window Analysis



- Basic clustering technique to limit number of variants included in analyses and to mitigate false positive rate
- Allows researchers granular control of window size

Variant Impact Querying

- Allows filtering based on desired coding and non-coding variation
  - Genic (exon, intron and UTR)
  - Exonic only
  - Updatable to included new designations like ChIP-Seq promoter/enhancer regions
- Allows filtering based on gene and gene pathways

7. Conclusions

- SVStore is a tool to collate the heterogeneous output of structural variant detection tools
- SVStore facilitates easy general queries, but also allows for complex queries which compares varied data types across tools and individuals for a powerful data analysis

8. Continuing Development

Clustering Solution

- A clustering algorithm run on records from the same individuals or breeds but from different tools could identify records which have a high likelihood of referring to the same variant.
- Store this information in a meta structure within SVStore.

Diverse Species Applications

- Although this database was designed to assist in the analysis of structural variation in dogs, it can be used for any species.
- Upon completion, a tool will be made public which will generate a database and a website to allow for the hosting of SV records.
- Thoroughly documented to allow for an easy open source release when that becomes feasible.

Additional Variant Types

- This database will be extended to other variant types such as SNVs, small indels, LTRs, retroviral insertions, and more to create a comprehensive storage solution for variation.

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA
NIH National Human Genome Research Institute