Building a Community for Development of Open Source Genomics Platform

Michael Bouzinier, Director of Informatics, Brigham genomics Medicine Program

Forome

association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

Why

- Whole genome sequencing (WGS) is rapidly becoming routine in clinical practice and in everyday life
- Processing and interpretation of genomic data requires more computational power and storage than any other task which an ordinary person is likely to come across in their life
- One can assume, it should lead to a massive software development effort
- But… most genomics platforms are either proprietary or academically developed and maintained

Ecosystem: Proprietary and Pseudo-Opensource

- Proprietary
  - Secret and closed source
- Nominally open source
  - But with the source that can be found nowhere
- Nominally open source tools with the code available on GitHub
  - Developed by a single academic lab
  - No real community
  - Often unmaintained
  - Work only on the cluster of the lab that has developed them

Ecosystem: True Open

- True open source tools for a specific task
  - E.g. parsers for special genetic file formats
  - Notable Example: Samtools & Bcftools by Heng Li
- Mix of pseudo-open source platform and true open source tools for a specific task
  - Tools developed by the Broad Institute of Harvard and MIT
- Global Alliance for Genomics and Health (GA4GH)
  - A strong community of people
  - Focused on the development of standards
  - Less focus on working tools

GA4GH

- We have been very much inspired by the work done by Cloud Workstream of GA4GH.
- The working group is focused on creating standards for portable workflows
  - Seems like exactly what we have been looking for
- But…
  - We need some software that works today and executes our and customers workflows
  - Cannot wait a few more years until all the standards will be in place

Apache: Nothing

- Looking at https://projects.apache.org/projects.html?category
  - More than 300 projects, ***nothing on genetics***

Forome

association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

Forome Goals

- Open and open source platform for analysis of whole genome
  - Build by International development community, used by community
  - Focus on collaboration between teams
- Support both clinical and research workflows, all flavors of data
  - Seamlessly transform research workflows into clinical guidelines
  - Using built-in integrated development environment for clinical rules
- Crowdsourcing support for solving difficult cases

Forome

association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

Platform

- Upstream pipeline
  - data from a sequencing lab → aligned BAMs → common variants
  - GATK best practices and haplotype callers
- Set of custom rare variant callers
  - identifies extremely rare and unknown variants in a pedigree-aware way
- QC Analysis
- Virus detection
- Third party plugins:
  - SvABA, Rufus
- Downstream annotation pipeline
  - injects a wide range of information related to functional analysis, population genetics, clinical knowledge, epigenetics, etc.
- Anfisa: a user interface for variant filtration, curation and interpretation

Forome

association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

FOROME ANFISA

Annotations for Variant Curation

"Open Collaboration by Design. Open Source by Nature"

Most variant curation tools …

- Easy to use, start with any VCF file
- Diverse annotations are gathered from various sources and combined in one place
- Annotations are used for
  - Interactive Variants Filtering
  - Variant interpretation

Unique to Anfisa

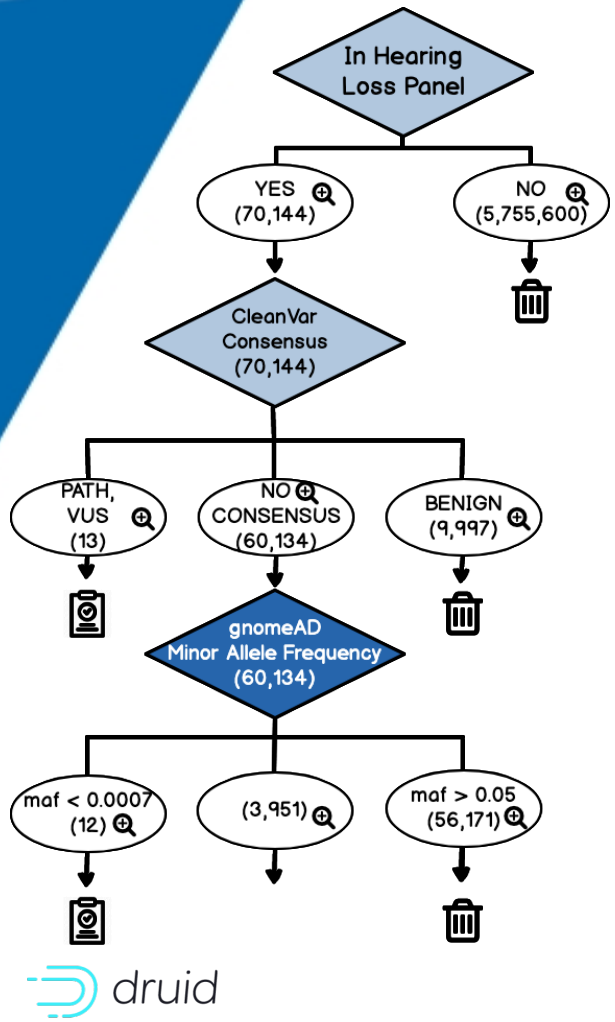
- Genetic analysis belongs to the same class of analytical problems as data warehousing and business intelligence
  - relatively static data is processed by OLAP
  - Traditional DBMS balance efficient updates with fast access
  - OLAP tools specifically focus on achieving the maximum performance for data querying and information retrieval
- OLAP approach is proven with other verticals
  - financial analysis, sales forecasting etc.
- **Data warehousing** principles work for genomics
  - Integrating data from multiple heterogeneous sources
  - Data cleaning, data integration, data consolidations
  - Analytical reporting, structured and ad hoc queries
  - Decision support

Forome

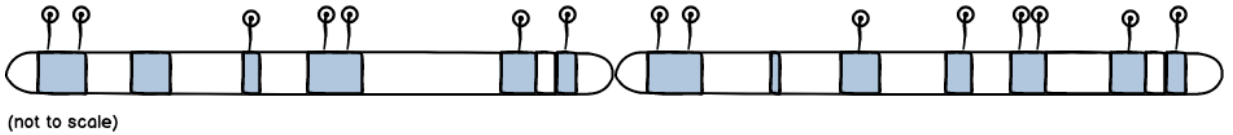
association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

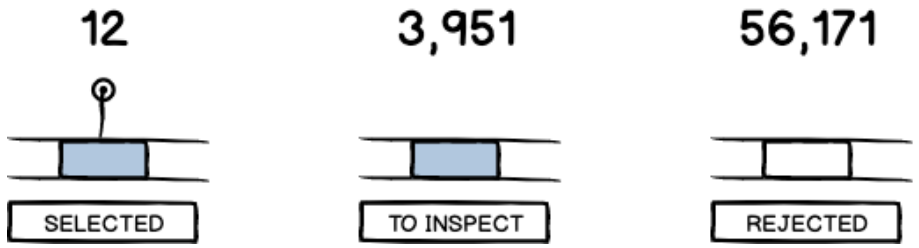
Integrated Development Environment for Mendelian Genomics



druid



(not to scale)



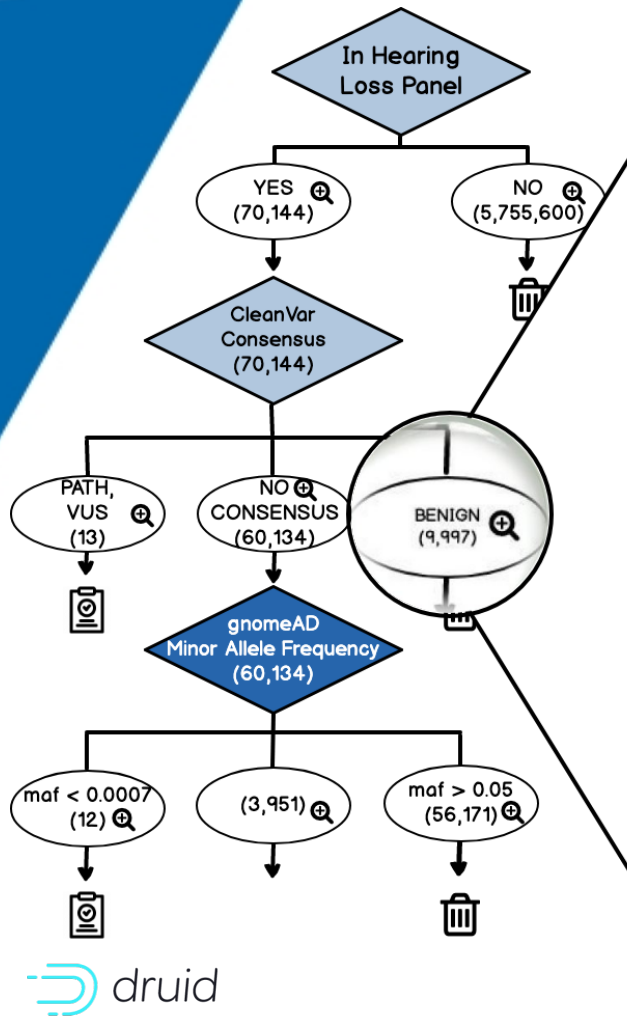
"Open Collaboration by Design. Open Source by Nature"

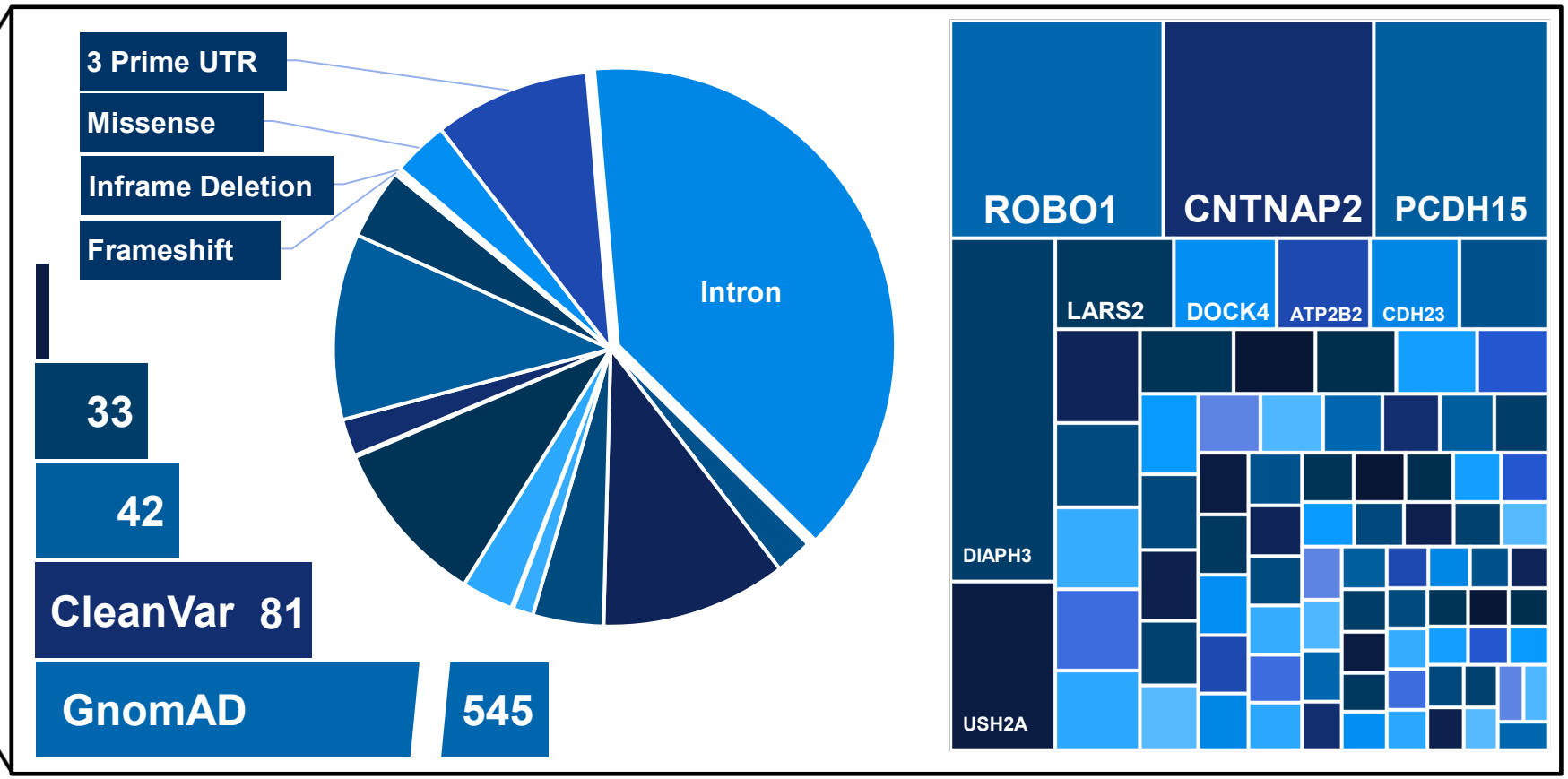
Forome

association &
open source genomic
platform

BRIGHAM HEALTH
BWH BRIGHAM AND WOMEN'S HOSPITAL

Integrated Development Environment for Mendelian Genomics





druid

"Open Collaboration by Design. Open Source by Nature"

Implemented Protocols and Built-in Filters

- Undiagnosed Patients Solution Pack
  - Diagnostics through Gene Discovery
  - Two families
    - BGM rare variants: use the output of the BGM rare variant callers
    - Mendelian rare variants: can use arbitrary VCF
  - Each family include:
    - Autosomal dominant
    - Homozygous Recessive
    - X-Linked
    - Compound Heterozygous
- Phenotype specific filters
  - Hearing Loss
  - ACMG59
  - developed together with SEQaBOO team.

Compound Heterozygous filter is used for a different kind of recessive analysis and shows tuples of heterozygous variants in all affected samples, where at least one damaging variant is inherited from a heterozygous mother, while the father is homozygous reference and at least one other damaging variant is inherited from heterozygous father while the mother is homozygous reference for this variant

Available Today

- Download from GitHub:
  - Backend and internal UI client: https://github.com/ForomePlatform/anfisa
  - Modern UI Client : https://github.com/ForomePlatform/Anfisa-Front-End
  - Annotations pipeline: https://github.com/ForomePlatform/Anfisa-Annotations
  - Analytical pipeline: https://github.com/ForomePlatform/pipeline
  - Variant Callers: https://github.com/ForomePlatform/variant_callers
- Beta release v.0.5.13
  - For a gene panel: Client Installation: 10 minutes
  - For whole genome:
    - Requires Druid: http://druid.io/
    - Installation including annotation pipeline and databases ~72 hours
    - Some functionality is only available through internal UI and REST

Current Users

- SEQaBOO: SEQuencing a Baby for an Optimal Outcome
  - A clinical research project aimed at integrating high-throughput clinical-grade genomic analysis into routine newborn screening
  - Best-practice clinical variant filtration algorithm for the phenotype of congenital deafness
    - Typically selects 10 to 30 variants for further review from a whole genome
- Brigham Genomic Medicine (BGM) / Harvard UDN Clinical Site
  - uses a phenotype-agnostic approach to analyze rare disease cases refractory to clinical analysis for novel genetic mechanisms of disease
- Research cohort analysis of purpura fulminans (PF) patients (BIDMC)
  - Uses the cohort analysis capabilities within Forome Anfisa
  - A population of 481 samples is divided into three cohorts: PF patients, sepsis patients without PF, and control cohort
  - Filters are based on a variant being present in specific cohort and combination of its frequencies within the cohorts

Next steps

- Annotation Service: upload your VCF, get back annotated JSON, feed JSON into Anfisa
  - Preview: http://anfisa.forome.dev/annotation-dev/#/annotation-dev
- Integration with FAVOR:
  - Comprehensive Annotation Database
  - More than 3000 annotations for a variant
- Support for ML in Clinical Rules

Thank you!