MNG LABORATORIES

Neurogenetic Answers™

Protocols to keep NGS gene panels and annotation content current

Nicholas A. Rouse, Matthew J. Snyder, Thomas J. McMahan
William A. Langley, Peter L. Nagy

Introduction

With decreasing sequencing costs and growing numbers of out of the box bioinformatics solutions, the landscape of clinically relevant genes and variants is changing at an unprecedented rate. Staying at the forefront of clinical relevancy requires frequently scheduled curation and updating of both gene lists and annotation databases. A two-fold approach is taken to keep data interpretation up to date. Gene list classifications are updated quarterly and annotation databases are updated monthly, coinciding with ClinVar's monthly release. Gene classification is determined by the combined use of OMIM and ClinVar and review of the scientific literature. Newly pathogenic or reclassified genes are individually examined to confirm correct placement for each panel based on multiple lines of independent scientific evidence. This gene curation results in a quarterly custom capture redesign and NGS panel updates. To ensure that annotations are precisely defined and up to date, our bioinformatics solution processes variant and gene annotation from external databases monthly through a normalization pipeline. Frequently updating custom capture reagent in conjunction with updating annotation content increases diagnostic sensitivity in NGS panels.

Why is it Important to Keep Annotations Current?

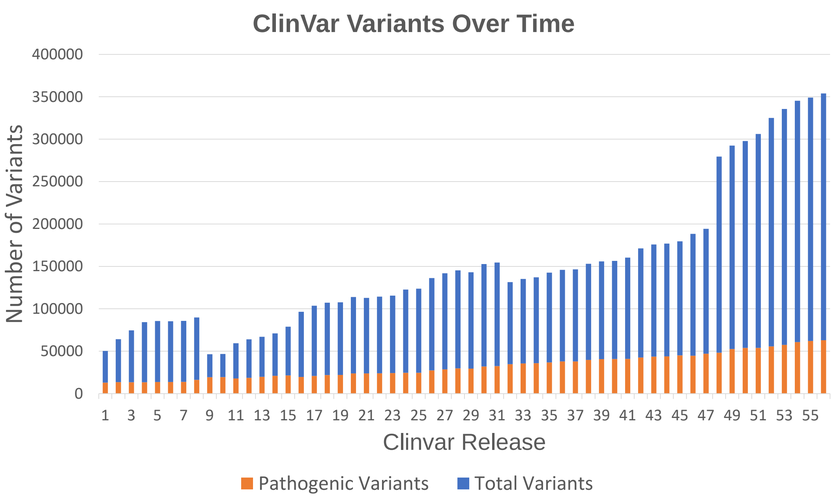
Human understanding of genetics is changing at an unprecedented rate, which is reflected in variant database growth. ClinVar, a database of reports of the relationships among human variations and phenotypes, sees an average increase of 1190 variants per release.

It is vital that next-generation sequencing and bioinformatic pipelines stay at the forefront of these new discoveries to ensure highest diagnostic quality.

This is accomplished with a two fold approach:

1. Monthly integration of new ClinVar annotation into bioinformatic pipelines.
2. Quarterly redesign of capture reagent and diagnostic test panel composition

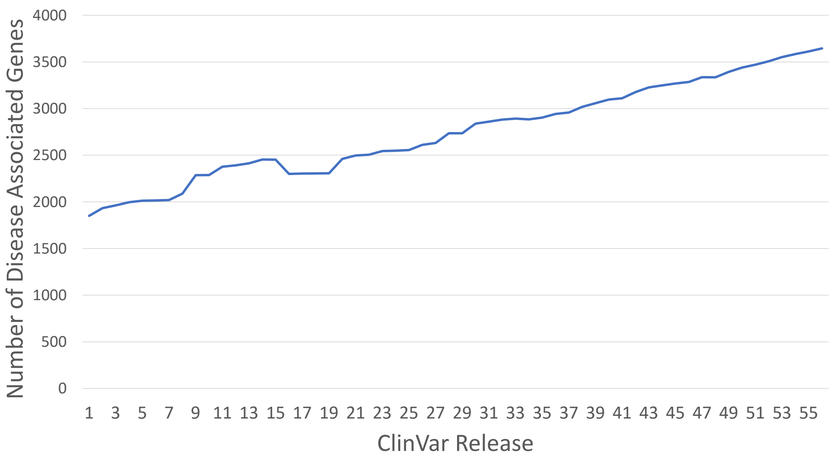
Data taken from 56 consecutive ClinVar vcf releases from: 06/2012 – 09/2017



On average per release:

- **New variants**: 4600
- **New Pathogenic variants:** 1190 (std=991)
- **Reclassified pathogenic variants as non- pathogenic:** 281(std=504)

Overall: Pathogenic variants 12944 --> 62960 (386 % increase)



On average per release:

- **New disease associated genes:** 41 (std=40)
- **Reclassified disease associated genes as non- pathogenic:** 8 (std=26)

Overall: Pathogenic genes 1850 --> 3645 (97 %increase)

Methodology: Keeping References Up to Date

1. **Automated download of ClinVar**
   - Software: crontab
   - Environment: Linux
   - Description: Automatic download and preprocessing script initialization scheduled daily to detect a new release of ClinVar
2. **Preprocessing of ClinVar**
   - Software: vt
   - Environment: Linux
   - Description: Variants are standardized in order for patient variants to be compared to external variant sources.
     i. Add allele ID vcf field to match normalized/decomposed variants with annotation
     ii. Decompose multiallelic variants into biallelic variants
     iii. Variant normalization consisting of:
       - parsimony, represent multi nucleotide polymorphisms in as few nucleotides as possible without an allele of length 0.
       - left alignment pertaining to the nature of a variant's length and position respectively.
3. **Process decomposed/normalized vcf in proprietary database (Genome MaNaGer®) containing patient variant data and various annotations**
   - Software: Microsoft SQL Server 2014
   - Environment: Windows Server 2012
   - Description: Updates variant and gene tables reflecting ClinVar release.
4. **ClinVar update report generated and distributed**
   - Software: Microsoft SQL Server 2014
   - Environment: Windows Server 2012
   - Description:
     i. Movement of pathogenic variants, reclassification
     ii. Movement of disease associated genes, reclassification
     iii. List of new pathogenic variants
     iv. List of patients with new pathogenic variants
     v. List of patients with re- categorized variants

Definitions

ClinVar VCF File Release:

- The ClinVar variant call format (VCF) file release dictates the variant and gene classifications. ClinVar also releases a tab delimited and HTML file that is inconsistently formatted and not entirely curated. The VCF release is curated through NCBI's dbSNP, all variants are required to have an rs identifier.

Disease Associated Gene:

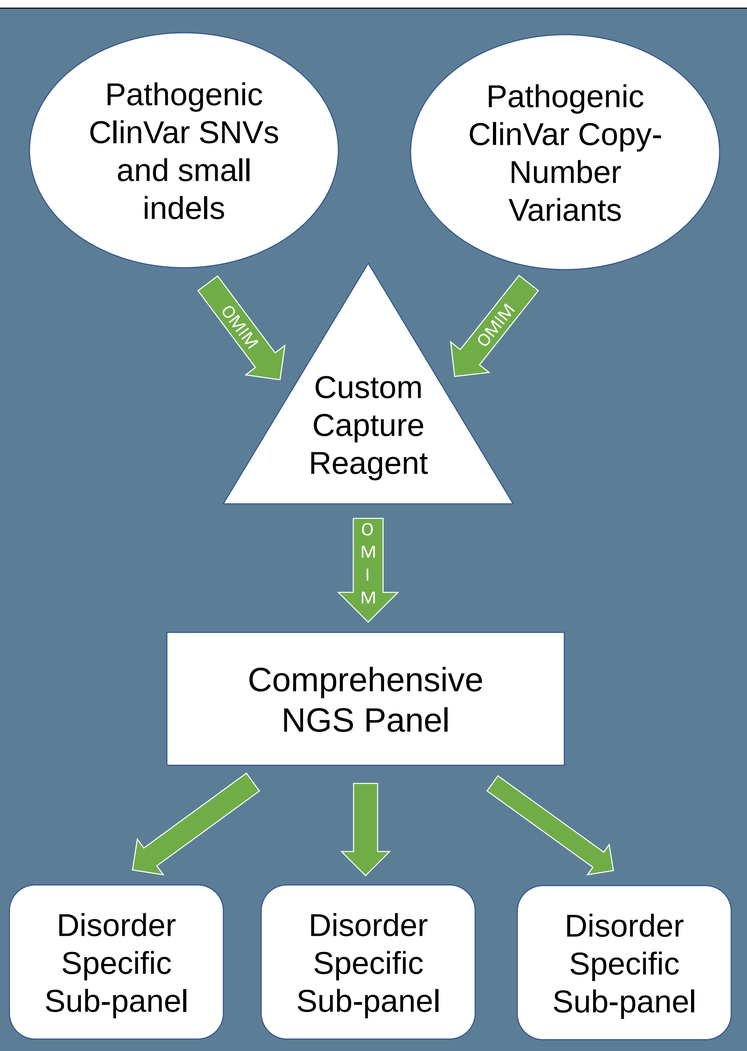
- Genes are given a binary classification of associated or not associated with disease, defined by the inclusion of at least one ClinVar pathogenic variant in the gene.

Pathogenic Variant:

- A pathogenic variant is defined as a variant that has a clinical significance flag of 4 (likely pathogenic) or 5 (pathogenic) in at least one variant submission (it is possible for variants to have multiple clinical significance flags).

Methodology: Keeping Test Panels Up to Date

1. **NGS test panels are phenotype-driven and composed of genes that fit certain criteria depending on the specific disorder being tested**
2. **List of pathogenic ClinVar genes filtered through OMIM database to pull out all genes with associated phenotypes covered by existing NGS test panels**
   – Use phenotype names and OMIM clinical synopsis information
3. **New custom reagent is designed by capturing all filtered ClinVar genes with pathogenic single nucleotide variants (SNV) and intragenic deletions/duplications**
   – Genes with SNV are captured using the Agilent SureSelect v6 Human All Exon probeset
   – Genes with intragenic deletions and duplications have 4x tiling intronic padding to detect single-exon resolution copy number changes
4. **Genes are then assigned to the correct phenotype-driven comprehensive panels by matching panel and gene information**
   – All gene assignments are double-checked manually by the scientific team before being finalized
5. **Genes that are assigned to comprehensive panels are then further designated to more specific sub-panels within the major disease categories targeted at specific disorders**



Summary

Keeping up to date with the most recent clinical variant and gene classification requires a set multifaceted methodology applied regularly. A two-fold approach is followed to achieve this goal. The first approach is accomplished by a monthly update of the ClinVar vcf file to include new variant and gene annotation in the Genome MaNaGer®. Secondly, the test panel capture, along with NGS test panel composition, is redesigned to include the most current knowledge of gene and disease relationships. By identifying key phenotype filtering criteria to each major comprehensive panel, newly-identified pathogenic genes can be seamlessly assigned to specific NGS panels on a regular basis. These two approaches with set methodologies ensure that our test panels and variant reporting functions are as up to date as possible.