The Journal of Molecular Diagnostics, Vol. 17, No. 4, July 2015

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

REVIEW

Annotation of Sequence Variants in Cancer Samples

Processes and Pitfalls for Routine Assays in the Clinical Laboratory

Lobin A. Lee, Kevin J. Arvai, and Dan Jones

From the Department of Pathology, Quest Diagnostics Nichols Institute, Chantilly, Virginia

CME Accreditation Statement: This activity ("JMD 2015 CME Program in Molecular Diagnostics") has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education (ACCME) through the joint providership of the American Society for Clinical Pathology (ASCP) and the American Society for Investigative Pathology (ASIP). ASCP is accredited by the ACCME to provide continuing medical education for physicians.

The ASCP designates this journal-based CME activity ("JMD 2015 CME Program in Molecular Diagnostics") for a maximum of 36 *AMA PRA Category 1 Credit(s)*™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

CME Disclosures: The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Accepted for publication March 23, 2015.

Address correspondence to Lobin A. Lee, M.S., Department of Molecular Oncology, Quest Diagnostics Nichols Institute, 14225 Newbrook Dr, Chantilly, VA 20151. E-mail: lobin.a.lee@questdiagnostics.com.

As DNA sequencing of multigene panels becomes routine for cancer samples in the clinical laboratory, an efficient process for classifying variants has become more critical. Determining which germline variants are significant for cancer disposition and which somatic mutations are integral to cancer development or therapy response remains difficult, even for well-studied genes such as *BRCA1* and *TP53*. We compare and contrast the general principles and lines of evidence commonly used to distinguish the significance of cancer-associated germline and somatic genetic variants. The factors important in each step of the analysis pipeline are reviewed, as are some of the publicly available annotation tools. Given the range of indications and uses of cancer sequencing assays, including diagnosis, staging, prognostication, theranostics, and residual disease detection, the need for flexible methods for scoring of variants is discussed. The usefulness of protein prediction tools and multimodal risk-based or Bayesian approaches are highlighted. Using *TET2* variants encountered in hematologic neoplasms, several examples of this multifactorial approach to classifying sequence variants of unknown significance are presented. Although there are still significant gaps in the publicly available data for many cancer genes that limit the broad application of explicit algorithms for variant scoring, the elements of a more rigorous model are outlined. *(J Mol Diagn 2015, 17: 339—351; http://dx.doi.org/10.1016/j.jmoldx.2015.03.003)*

Germline Polymorphisms and the Risk of Cancer Development

Genetic differences in individuals include single nucleotide polymorphisms (SNPs), intragenic insertion and deletion polymorphisms (indels), and structural variants, such as copy number variations. These factors contribute to risk of cancer development and responses to therapy (pharmacogenomics). During tumor development, there is a complex interplay between somatic or acquired mutations in oncogenes, tumor suppressors, and epigenetic regulators and germline, or inherited, genetic variation.

Initial work on localizing genetic variants associated with cancer susceptibility focused on well-defined clinical syndromes, such as alterations of *TP53* in Li-Fraumeni syndrome,

Disclosures: All authors were employees of Quest Diagnostics at the time this review was composed. Quest Diagnostics offers sequencing assays commercially.

Current address of D.J., Ohio State University, Columbus, OH; of K.J.A., GeneDX, Gaithersburg, MD.

http://dx.doi.org/10.1016/j.jmoldx.2015.03.003

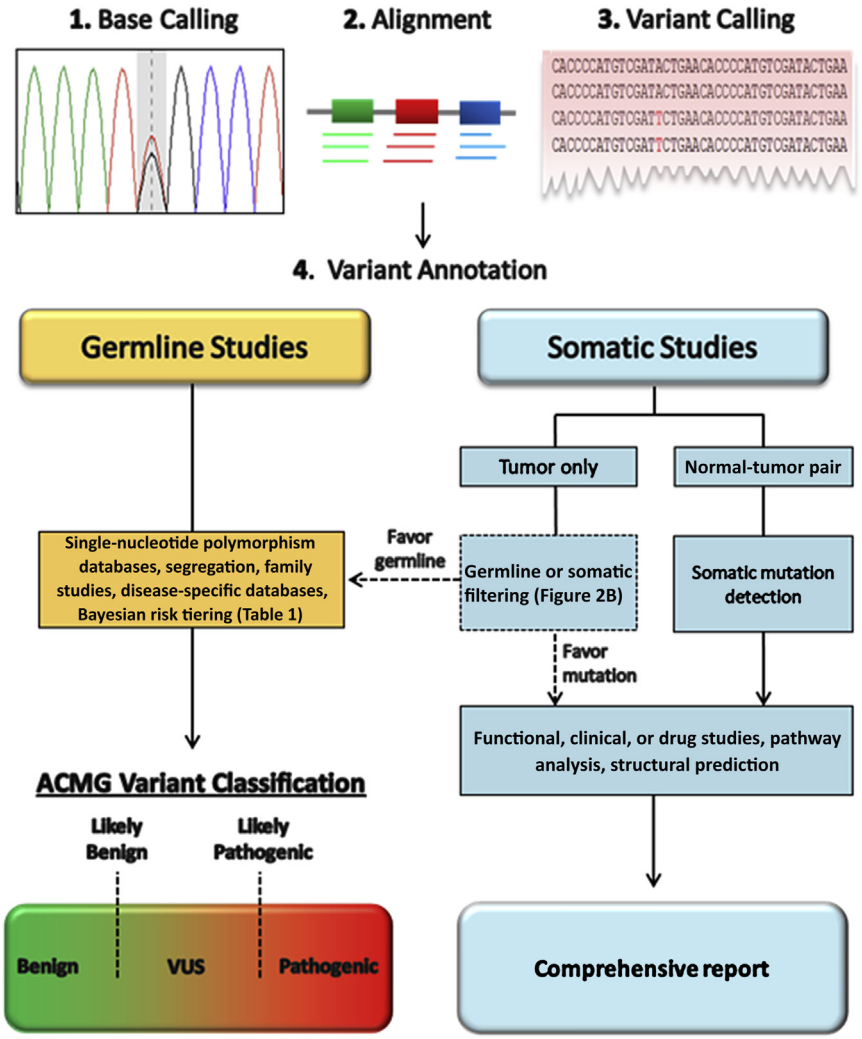

Figure 1 Variant analysis pipeline comparing germline and somatic annotation. Base-calling, alignment, and variant calling (steps 1 to 3) typically use a standard toolset, such as Samtools, Genome Analysis Toolkit (*https://www.broadinstitute.org/gatk*), Bowtie (*http://bowtie-bio.sourceforge.net*), and Burrows-Wheeler Aligner (*http://bio-bwa.sourceforge.net*). All websites were last accessed October 2, 2014. For variant annotation (step 4), the toolset and analysis parameters are less standardized. For germline studies, the American College of Medical Genetics and Genomics (ACMG) variant classification system provides guidance for interpretation. For somatic mutations in cancer samples, several tools are available for germline filtering if a normal and nonneoplastic reference sample is available for comparison, including MuTect (*http://www.broadinstitute.org/cancer/cga/mutect*), Varscan (*http://varscan.sourceforge.net*), and SomaticSniper.[7] When no reference sequence is available, rules for trimming germline calls must be applied before the somatic calls can be annotated for significance. VUS, variant of unknown significance.

BRCA1 and *BRCA2* in hereditary breast and ovarian cancer,[1] and the mismatch repair genes in Lynch syndrome.[2] The genes involved, typically tumor suppressors, were initially localized using linkage analysis and targeted DNA sequencing. The genetic changes observed in affected individuals include frameshift and nonsense mutations as well as inactivating mutations with loss of function linked to tumor initiation.

More recently, genome-wide association studies (GWASs), which compare variant profiles of diseased versus healthy individuals, have accelerated the rate of discovery of cancer-associated variants.[3] GWASs have identified many more cancer-associated variants in well-characterized cancer genes than family studies, including many missense mutations that have subtle or undetermined effects. They have also identified recurrent germline variants in genes whose association with carcinogenesis was not previously known. These include genes that serve core cellular functions, such as energy metabolism, chromatin maintenance, and protein translation.

Connecting a given variant to its phenotypic effect(s) is more difficult for GWASs compared with classic genetic analyses. Even when published data are available to support interpretation of a particular mutation call, GWASs have often had low reproducibility attributable to weak or incomplete phenotypic penetrance, bias against effects due to more common variants and covariants, and inadequate statistical power.[4] As a result, only approximately one-third of germline variant associations reach statistical significance across multiple studies.[5] Unsurprisingly, GWASs, massively parallel exome sequencing projects, and routine targeted next-generation sequencing (NGS) in clinical laboratories have resulted in many more variants of undetermined significance (VUS) in cancer-associated genes.[6]

The Central Role of the Analysis Pipeline

The first step in ensuring robust variant calling relies on accurate base-calling and alignment. To accomplish this goal, raw data from high-throughput sequencers are moved through an analysis pipeline to sequentially accomplish sequence alignment, read filtering, variant calling, and variant