illumina®

BaseSpace™ Knowledge Network

Variant interpretation is simplified with biomarker content curated from public databases.

Introduction

Estimates of human genome variants implicated in disease range in the millions, yet many variants found in next-generation sequencing (NGS) studies are of unknown function or impact. The ability to provide a relevant and contextualized report from NGS data increasingly relies on knowledge bases, or repositories of curated content for variants. To address this challenge, Illumina offers BaseSpace Knowledge Network, a private knowledge base where researchers can curate information about variants and attach detailed supporting evidence. The BaseSpace Knowledge Network compiles information from public databases to help labs get started with baseline content. Designed to allow a high degree of collaboration between team members within a workgroup and encourage curation best practices, the BaseSpace Knowledge Network can reduce the time for interpretation while increasing the accuracy and consistency of results.

The BaseSpace Knowledge Network is part of the Illumina BaseSpace Informatics Suite, which also includes BaseSpace Variant Intepreter (Beta). Together these products streamline prioritization, interpretation, and reporting of genomic variants in human samples (Figure 1). Exclusively integrated with BaseSpace Variant Interpreter (Beta), the BaseSpace Knowledge Network rapidly accelerates manual interpretation of variants by providing high-quality, curated content in the form of associations between genomic variations and phenotypes (Figure 2). BaseSpace Variant Interpreter (Beta) provides a complete variant analysis workflow that includes annotation, filtering, interpretation, and reporting.

Variant associations are sourced from the Illumina Knowledge Base, and the ClinVar database.[1] Illumina Knowledge Base contains expert-curated content from the Illumina Biomedical Informatics team, and ClinVar, a National Institute of Health (NIH) funded initiative to facilitate sharing of variant interpretations. These sources represent the first knowledge bases of the BaseSpace Knowledge Network, which connects individual knowledge bases to a standardized, high-performance network, and enables rapid sharing of variant interpretations. More knowledge bases will be added in the future. In this technical note, the curation process for each type of biomarker is described, with a summary of the total number of associations.

Applying Years of Expertise in Variant Interpretation

The Illumina Biomedical Informatics team, originally part of NextBio™ Clinical, has invested years of time and expertise into building a content set with hosted infrastructure, and a robust process to amass high-quality variant interpretations. The content consists of associations between genomic variants and phenotypes for somatic and germline variants. The adopted curation process at Illumina captures relevant information in standardized and structured ways to enable reporting. The purpose of curation is to harmonize the language, contextualize the value, and trace the evidence for each marker. In this initial release, the content curated by Illumina includes associations with single-nucleotide variants (SNVs), multinucleotide variants (MNVs), and small insertions/deletions (indels).

Ontologies for Phenotypes and Drugs

The phenotype and drug information included for each genetic variant is based on a modified Systematized Nomenclature of Medicine (SNOMED)[2] ontology for disease, and Medical Subject Headings (MESH)[3] for drugs.

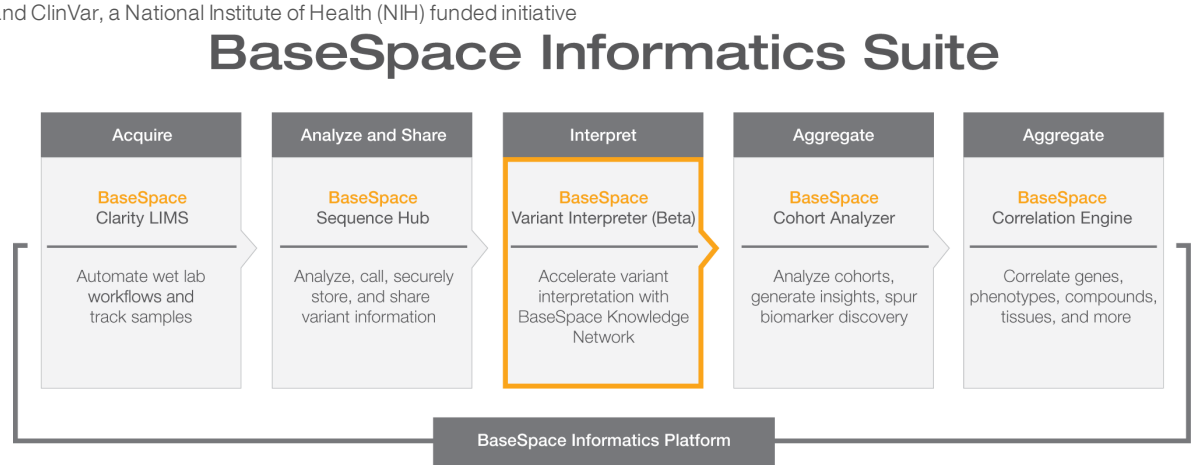

Figure 1: BaseSpace Informatics Suite—A comprehensive series of genomics solutions offering continuous support for researchers from sample collection to final reports. Beta versions of BaseSpace Variant Interpreter (Beta) and BaseSpace Knowledge Network are now available.

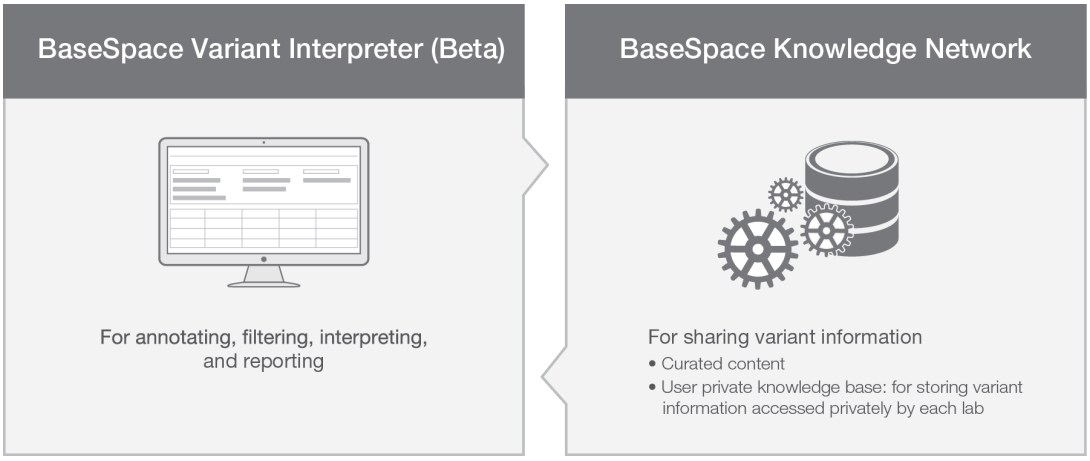

Figure 2: BaseSpace Variant Interpreter (Beta) and BaseSpace Knowledge Network—Both software platforms are integrated to decrease time and effort required for associating variants with data curated from previous studies.

Somatic Biomarker Curation Process

Somatic mutations are acquired genetic alterations in somatic tissue that are not transmitted to offspring. These mutations can be linked to a specific condition or be the underlying cause for cancer (driver mutation). Evidence-based summaries for somatic alterations are harvested from publications, guidelines, drug labels, clinical trials, and approved companion tests. The curation process uses an established workflow that identifies both emerging and established variants to add to the Illumina Knowledge Base. Internal tracking and formal review steps ensure that each variant interpretation is consistent with established protocols. Curated information is structured around 4 major components: genetic variants, tumor types/drugs tested, evidence summary, and evidence types (Figure 3).

Interpretation

Somatic variant associations in cancer are categorized according to their utility (Table 1):

- Predictive associations describe variants potentially responsive to specific drug treatments, usually compared to standard therapies.
- Prognostic associations suggest that variants affect the rate of disease progression over time.
- Classification associations suggest that variants are related to disease subtyping; these associations are often found in the pathological correlations sections of clinical research papers.
- Clinical trial associations are established when variants have been mentioned in inclusion criteria or as part of a trial design. Variants can be specified (gene/codon) or implied (activity or drug sensitivity). All implied variants require additional evidence, such as a functional study establishing the effect of the alteration.

Details

Phenotype/s	Glioblastoma multiforme
Drugs Tested	Vemurafenib
Mechanism	None specified
Validation Level	Clinical Studies
Confidence	Not Available
Direction	Improved Response
Gene	BRAF
Transcript	NM_004333.4
Exon	15/18
cDNA Change	c.1799T>A
Amino Acid Change	p.Val600Glu
Consequence	Missense variant
Submitted Date	Feb 20, 2015

Curator Summary
Curator summary is not available.

Evidence

Publication (1)

Publication PMID: 24725538

Evidence Summary	In a case study, a pediatric grade IV glioblastoma multiforme patient harboring BRAF V600E mutation was associated with a complete response to vemurafenib monotherapy after 4 months of treatment, which was sustained at 6 month of therapy; the patient was undergoing the 7th cycle of therapy at the time of publication of this study.
Support Type	Not Selected
Evidence Strength	Not Selected
Publication Type	Case Report
Publication Id	24725538

Figure 3: Variant Entry—Example of an entry for a *BRAF* variant.

Table 1: Number of Associations for Somatic Variants, Categorized by Clinical Utility

Category	Value
Predictive associations	1800
Prognostic associations	4000
Classification associations	4000
Clinical trial associations	20,000

Somatic Biomarker Content

The Illumina Knowledge Base currently contains approximately 52,000 potential variant associations with 206 unique drugs, 133 diseases, and > 1000 identifiers in the PubMed database (PMIDs). Predictive variant associations are classified with companion tests, clinical studies, case reports, and experimental validation levels, with 13 unique companion tests (Table 2). Curated content supporting predictive variants currently includes:

- Drug label information curated for FDA-approved drugs, specifying predictive variants in the disease indication section of drug labels, or as inclusion criteria.
- Guidelines for non-small cell lung cancer, breast cancer, colon cancer, rectal cancer, soft tissue sarcoma, melanoma, and dermatofibrosarcoma protuberans, including treatment recommendations and specified tests for relevant variants.
- In collaboration with the American Society of Clinical Oncology (ASCO), the Targeted Agent and Profiling Utilization Registry Study (TAPUR)[4] curates information about 15 targeted anticancer drugs (cetuximab, erlotinib, vemurafenib, crizotinib, dasatinib, ado-trastuzumab emtansine, sunitinib, temsirolimus, palbociclib, bosutinib, vismodegib, axitinib, olaparib, regorafenib, and pembrolizumab). Evidence has been sourced from PMID-indexed studies to identify variants associated with resistance or sensitivity to each drug.

The Actionable Genome Consortium[5] compiled a subset of clinical trials that were added to the Illumina Knowledge Base. All trials relevant to biomarkers were manually curated, and were last updated on June 1, 2015 (Table 2).

Table 2: Number of Associations for Predictive Variants, Based on Clinical Utility

Category	Target Region/Sequence (Exon)	Unique Associations
Companion test	299	13
Clinical trials	40,600	180
Clinical studies	1334	
Case reports	167	
Experimental	357	

Evidence Types

Different types of evidence can substantiate an association between variants and utility. Evidence supporting an interpretation can be obtained from various sources, such as PubMed, FDA, and clinical trial websites.[6-8] The Illumina Knowledge Base ranks evidence types based on credibility and significance of the source of evidence.

Companion Test—The variant has been identified as a biomarker correlated to a specific drug, used for treatment of a specific tumor type. Companion test entries are considered 1 of the highest levels of evidence in the knowledge base.

Clinical Studies—Cohort analysis is used to determine the potential role of the variant in predicting tumor response to drug treatment or association with outcome endpoints. Studies are included in the database only when statistically significant results between cohorts are reported. Results have not been validated by a regulatory authority.

Case Reports—Published results for individual patients can indicate potential variant associations with treatment outcomes, such as partial response. Alternatively, case series are also included when a paper identifies a variant associated with response, or lack of response, to a drug in a subset of tumors. Results have not been validated by a regulatory authority, nor analyzed within a statistically significant study.

Experimental—Published studies that examine variant associations with the response to a drug treatment within the context of tissue culture (*in vitro*) or animal models (*in vivo*) are considered experimental.

Germline Biomarker Curation Process

Germline mutations are heritable genetic alterations found in germline tissue. These mutations are responsible for genetic diseases including some types of cancer and rare undiagnosed genetic disorders. Significance of germline alterations is obtained from publications and public databases. Disease prevalence and mode of inheritance are also included for each germline variant association. Pathogenicity assertions for variant–disease associations are based on collective evidence following American College of Medical Genetics and Genomics (ACMG) guidelines published in 2015.[9] Interpretations are standardized for optimal readability, summarizing the collective evidence used to determine pathogenicity. Evidence supporting the interpretation can be obtained from a variety of sources, including PubMed. This detailed section also provides scope, study design, and relevant statistics when available.

Germline Biomarker Content

There are approximately 34,578 germline variants currently present in the Illumina Knowledge Base, representing 1663 unique diseases and > 1000 PMID indexes. Curated biomarkers represent a total of > 800 genes (Table 3), with 95% of variants from the 6 ACMG genes (*CFTR*, *MSH2*, *NF2*, *LDLR*, *PMS2*, and *PCSK9*) that were curated in early 2015.

Table 3: Associations for Germline Variant Categorized by Pathogenicity

Pathogenicity Level	No. of Associations
Pathogenic	1129
Likely pathogenic	1682
Variant of unknown significance, suspicious	2154
Variant of unknown significance	23,637
Likely benign	1620
Benign	4356

ClinVar: Resource for Aggregation and Variant Interpretation Reports

ClinVar is a public repository of clinically relevant variants, funded by the National Institutes of Health.[1] The ClinVar database contains over 100,000 clinically significant variant–phenotype associations. When any organization submits a record to ClinVar, the record is assigned a submitted clinical variant ID (SCV). The SCV record consists of a genomic variant and a clinical phenotype (disease) with comments and evidence supporting the relationship between variant and phenotype. After review, the ClinVar team assigns a reference clinical variant ID (RCV) to an SCV. Additional SCVs with unique identifiers can be added to existing RCVs, and the RCV may apply to 1 or more SCV records with equivalent variants and phenotypes. The ClinVar team adds additional features to the RCV record, such as review status and phenotype/variant normalization.

ClinVar content is available in BaseSpace Variant Interpreter (Beta) as a source of content on the BaseSpace Knowledge Network (Figure 4). Original submissions are displayed at the SCV level, and parsed from XML and VCF files available on the ClinVar FTP site.[10] Variant submissions with incomplete or misannotated information are filtered out and variants are further validated using the Illumina Annotation Engine 1.4.1. Data are limited to germline variants for the first release of BaseSpace Knowledge Network.

The following fields are parsed for ClinVar variants: Variant, RCV ID, SCV ID, name of submitter, date of submission, date of update, mode of inheritance, level of confidence, phenotype ontology term, ontology ID, and review status (SupplementaryTable 4/1). Because each SCV may have multiple components with different evidence types, they are mapped and aggregated into the knowledge network data model.

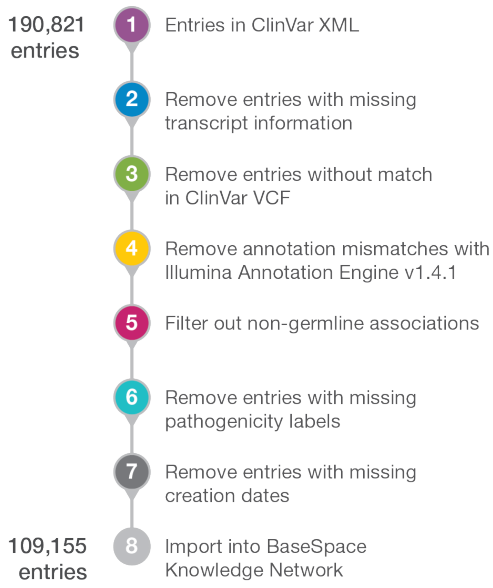

Figure 4: Process for Ingestion of ClinVar Version—Multiple SCVs may be a part of a single RCV and filtered at the RCV or SCV level.

Summary

BaseSpace Variant Interpreter (Beta) and BaseSpace Knowledge Network are software platforms that reduce time and effort required to transform genomic information into biological insight. The BaseSpace Knowledge Network curates variant information from numerous databases and organizes clinically relevant associations into an easily accessible and readable interface. The simple user-friendly interface of BaseSpace Variant Interpreter (Beta) integrates with BaseSpace Knowledge Network, delivering an intuitive framework for nonexpert users to annotate, filter, and interpret variant data easily. Together these tools provide a flexible, reliable, and efficient method for enriching variant information with biological context.

Learn More

To learn more about BaseSpace Variant Interpreter (Beta) and BaseSpace Knowledge Network, visit www.illumina.com/informatics/research/biological-data-interpretation/variant-interpreter.html

References

1. ClinVar. www.ncbi.nlm.nih.gov/clinvar. Accessed August 3, 2016.
2. Systematized Nomenclature of Medicine-Clinical Terms. bioportal.bioontology.org/ontologies/SNOMEDCT. Accessed August 3, 2016.
3. US National Library of Medicine: Medical Subject Headings. www.nlm.nih.gov/mesh/. Accessed August 3, 2016.
4. The Targeted Agent and Profiling Utilization Registry (TAPUR) Study. www.tapur.org/. Accessed August 3, 2016.

5. Actionable Genome Consortium to guide NGS in cancer. *Nature Biotechnology*. 2014;32:965. doi:10.1038/nbt1014-965d. Accessed August 3, 2016.
6. PubMed. www.ncbi.nlm.nih.gov/pubmed. Accessed August 3, 2016.7.
7. US Food and Drug Administration. www.fda.gov/default.htm. Accessed August 3, 2016.
8. Clinical Trials Database. clinicaltrials.gov/. Accessed August 3, 2016.
9. Hampel H, Bennett RL, Buchanan A, et al. A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: referral indications for cancer predisposition assessment. *Genet Med*. 2015;17(1):70-87.
10. ClinVar FTP site. ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/. Accessed August 3, 2016.

Table 4: ClinVar XML Root Paths

ClinVar Extraction Root Path	User Interface Display	Example
ReleaseSet/clinvarset/ clinvarassertion/observedin/observeddata	Evidence summary, PMID	Fields et al. (2002) demonstrated that the Fin(major) *USH3A* mutation in exon 3 of the *USH3A* gene, which had been identified by Joensuu et al. (2001) as 300C-T (TYR100TER), should be referred to as 528T-G, resulting in a tyr176-to-ter substitution. Joensuu et al. (2001) had identified homozygosity for this mutation in a Finnish family segregating Usher syndrome type IIIA (*USH3A*; 276902) and found it in a further 52 Finnish patients. Fields et al. (2002) found this mutation in 11 of 28 mutated alleles from affected individuals of Finnish and other northern European ancestry.
ReleaseSet/clinvarset/clinvarassertion/ observedin/sample/origin	Not displayed	Germline
ReleaseSet/clinvarset/clinvarassertion/ clinicalsignificance/comment	Curator summary	The study set was not selected for affection status in relation to any cancer. Pathogenicity categories were based on literature curation. See PubMed ID:22703879 for details.
ReleaseSet/clinvarset/clinvarassertion clinicalsignificance/description	Pathogenicity	Uncertain significance
ReleaseSet/clinvarset/ClinVarAssertion/ ClinicalSignificance/ReviewStatus	Review status	Reviewed by expert panel
ReleaseSet/clinvarset/referenceclinvarassertion/ attributeset/attribute	Mode of inheritance	Autosomal recessive inheritance
ReleaseSet/clinvarset/ClinVarAssertion/ TraitSet	Phenotype	Usher syndrome, type 3
ReleaseSet/clinvarset/ClinVarAssertion/ ClinVarSubmissionID	Submitted by	OMIM
ReleaseSet/clinvarset/ReferenceClinVarAssertion/ ClinVarAccession	RCV	RCV000004642.1
ReleaseSet/clinvarset/ClinVarAssertion/ ClinVarAccession	SCV	SCV000024816.1
ReleaseSet/clinvarset/referenceclinvarassertion/ measureset/name/elementvalue	Not displayed, used to join on IAE	NM_001195794.1:c.567T>G
ReleaseSet/clinvarset/ClinVarAssertion/ ClinVarSubmissionID	Submitted date	"2015-01-29"
ReleaseSet/clinvarset/ClinVarAssertion/ ClinVarAccession	Updated date	"2015-01-30"