GenomeVIP: a cloud platform for genomic variant discovery and interpretation

R. Jay Mashl,[1,2] Adam D. Scott,[1,2] Kuan-lin Huang,[1,2] Matthew A. Wyczalkowski,[1] Christopher J. Yoon,[1,2] Beifang Niu,[1] Erin DeNardo,[1] Venkata D. Yellapantula,[1,2] Robert E. Handsaker,[3,4] Ken Chen,[5] Daniel C. Koboldt,[1] Kai Ye,[1,2] David Fenyö,[6] Benjamin J. Raphael,[7] Michael C. Wendl,[1,8,9] and Li Ding[1,2,8,10]

[1]McDonnell Genome Institute, Washington University, St. Louis, Missouri 63108, USA; [2]Division of Oncology, Department of Medicine, Washington University, St. Louis, Missouri 63108, USA; [3]Stanley Center for Psychiatric Research, Broad Institute, Cambridge, Massachusetts 02142, USA; [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [5]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA; [6]Langone Medical Center, New York University, New York, New York 10016, USA; [7]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island 02912, USA; [8]Department of Genetics, Washington University, St. Louis, Missouri 63108, USA; [9]Department of Mathematics, Washington University, St. Louis, Missouri 63108, USA; [10]Siteman Cancer Center, Washington University, St. Louis, Missouri 63108, USA

Identifying genomic variants is a fundamental first step toward the understanding of the role of inherited and acquired variation in disease. The accelerating growth in the corpus of sequencing data that underpins such analysis is making the data-download bottleneck more evident, placing substantial burdens on the research community to keep pace. As a result, the search for alternative approaches to the traditional "download and analyze" paradigm on local computing resources has led to a rapidly growing demand for cloud-computing solutions for genomics analysis. Here, we introduce the Genome Variant Investigation Platform (GenomeVIP), an open-source framework for performing genomics variant discovery and annotation using cloud- or local high-performance computing infrastructure. GenomeVIP orchestrates the analysis of whole-genome and exome sequence data using a set of robust and popular task-specific tools, including VarScan, GATK, Pindel, BreakDancer, Strelka, and Genome STRiP, through a web interface. GenomeVIP has been used for genomic analysis in large-data projects such as the TCGA PanCanAtlas and in other projects, such as the ICGC Pilots, CPTAC, ICGC-TCGA DREAM Challenges, and the 1000 Genomes SV Project. Here, we demonstrate GenomeVIP's ability to provide high-confidence annotated somatic, germline, and de novo variants of potential biological significance using publicly available data sets.

[Supplemental material is available for this article.]

Understanding the relationship between genetics and disease is a central theme of biomedical research. Enabled by increasingly economical next-generation sequencing technologies, many projects have sought to characterize variation within and across populations and disease cohorts. Among these are the 1000 Genomes Project, The Cancer Genome Atlas (TCGA), UK10K, Pediatric Cancer Genome Project (PCGP), and the International Cancer Genome Consortium (ICGC), but numerous smaller-scale projects are also under way. Advances in sequencing technologies and further economies of scale are expected to increase the collective corpus of sequence data dramatically, particularly as clinical diagnostic sequencing becomes more prevalent while expanding across data types (methylation, mRNA, and miRNA) and especially if widespread screening of asymptomatic individuals is implemented. For example, the Precision Medicine Initiative (PMI) envisions a longitudinal collection of genomic data from more than 1 million individuals (Collins and Varmus 2015).

The vast amounts of data produced by today's sequencing projects impose logistical challenges in downloading and storing data prior to passing it through a pipeline of bioinformatics tools running on local high-performance computing resources, i.e., the traditional "bringing data to the tools" paradigm of variant analysis (Stein 2010). Cloud computing addresses both the computational and storage challenges associated with large data sets by enabling users to launch on-demand virtualized instances of computer systems with preinstalled tools and scripts that have the ability to import source data from and export processed data to cloud storage. Furthermore, the research community is transitioning to approaches that democratize access to genomic data (Heath et al. 2014; Stein et al. 2015), resulting in the creation of resources such as the National Cancer Institute's Genomic Data Commons (GDC; http://gdc.cancer.gov), a comprehensive cancer genomics repository. Cloud computing, where analysis is conducted by "bringing tools to the data," is viewed as having an

Corresponding author: lding@wustl.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.211656.116.

27:1450–1459 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

important role in future bioinformatics analysis (Dudley et al. 2010), which is especially critical for independent scientists and clinics that are unable to support a local high-performance computational infrastructure.

Here, we present the Genome Variant Investigation Platform (GenomeVIP), a system for performing variant discovery, annotation, and interpretation using cloud resources. Its intuitive, lightweight web interface enables users to detect genomic variants (single- and multinucleotide variants, short insertions and deletions [indels], complex indels, and structural alterations that include translocations, inversions, and tandem duplications) in whole-genome or exome sequence alignment files (BAMs). We highlight important functionalities of GenomeVIP and examine its capabilities within the larger context of genomic data processing systems, demonstrating GenomeVIP as an engine for future cancer and human genetic studies.

Results

GenomeVIP is an open-source, cloud-aware, multiuser platform with a web interface for performing discovery, annotation, and interpretation of genomic variation. Its sophisticated design brings powerful cloud resources to bear for task-specific bioinformatics analyses without any special cloud expertise required by the investigator. Users can direct GenomeVIP to perform germline, somatic, and de novo variant calling by selecting software from its palette of widely used bioinformatics tools and can specify provided best-practices discovery tool parameter sets or design a custom "execution profile" by adjusting online or uploading parameters through the web interface. By providing both high- and low-level access to tools and parameters, GenomeVIP provides flexibility for use by computational biologist power users as well as by users having expert knowledge or particular research needs. Its design and operational aspects aim to promote the reproducibility, transparency, and uniformity of the processing of genomic data. Although we focus on GenomeVIP's cloud capability, it runs equivalently on local high-performance computing clusters with local data with preinstalled versions of the named tools. The design and implementation of GenomeVIP and its usage for modern, high sample count cancer genomics and human genetics research also serves as a starting point for integration of additional tools and capabilities and expansion to other cloud-computing platforms.

GenomeVIP architecture

The functionality of GenomeVIP is provided through coordination among three central components: the user's web browser, the GenomeVIP server, and a cloud-computing resource (Fig. 1).

Web browser

The GenomeVIP web browser interface (Fig. 1, top) furnishes user controls for many tasks, including the following: loading sample file information; selecting samples; choosing parameters and tools for discovery, filtering, and annotation; selecting genomic regions; and managing cloud computational and storage resources. The interface is implemented using a combination of HTML, JavaScript, and cascading style sheets (CSS) and uses a jQuery JavaScript library to modify web page content and provide cross-browser compatibility. In addition, user-generated events and communications between the browser and the GenomeVIP server are handled by JSON-formatted AJAX requests, a standard jQuery feature.

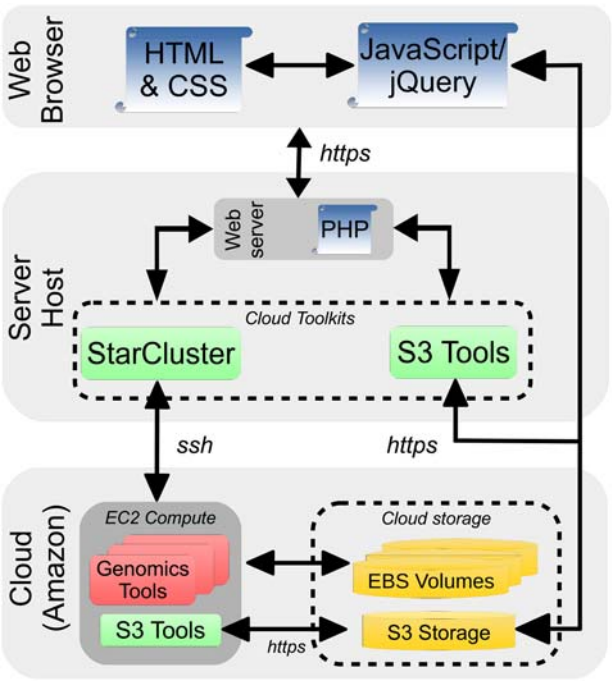

Figure 1. GenomeVIP platform. GenomeVIP consists of three components (web browser, server host, cloud), coordinated by various scripting languages (blue) and cloud toolkits (green). Interactive web pages, written in HTML (with CSS elements) and JavaScript, provide front-end functionality. JQuery is a JavaScript library providing methods to modify web page content with cross-browser compatibility. Server-side PHP modules utilize StarCluster and S3 Tools cloud toolkits to access EC2 Compute (gray) and storage resources (yellow) in the cloud. GenomeVIP creates within EC2 a virtual cluster, based on a machine image with preinstalled variant detection tools and supporting software (collectively, "Genomics Tools") (red), that can access sequence data on S3 and EBS (Elastic Block Storage) resources (yellow). Secure channels using HTTPS and secure shell (SSH) protocols allow communication between various components. Resulting variant call files stored in S3 are accessible via the GenomeVIP interface or the Amazon S3 Console.

Server host

The GenomeVIP server (Fig. 1, middle) runs a secure Apache HTTPS web server that may be installed and run locally as a real or virtual machine (VM) or that may be instantiated as a VM on Amazon Web Services (AWS). Server-side scripts written in PHP handle user selections and generate content for client-side interpretation and storage within the user's web browser (Fig. 1, top). To configure a computation to run on the AWS cloud, users provide their previously established AWS login credentials, specify genomic samples, and select one of the predefined analysis pipelines and parameter sets, which may be further customized. Users then specify storage and computing resources required, i.e., a VM instance type (giving a certain number of processors and memory) and the number of nodes or virtual cluster size. The server automatically builds the necessary configuration files for the StarCluster (http://star.mit.edu/cluster) and S3 Tools (http://s3tools.org) cloud toolkits to manage AWS computing and storage resources, respectively. After users submit a computation, the server performs key tasks: instantiating cloud computing resources; generating a master script that creates template-subordinate work scripts to perform the variant discovery, post-discovery processing, and storing of results in parallel fashion across genomic

regions and tools; and transmitting the work scripts to the targeted computing host for execution.

The cloud

Cloud resources (Fig. 1, bottom) consist of cloud computational and storage components as specified by users on the GenomeVIP server. Here, Amazon's Elastic Compute Cloud (EC2) computational infrastructure hosts a virtual cluster instantiated by GenomeVIP from a prebuilt run-time virtual machine (VM) image containing well-established variant discovery tools—VarScan (Koboldt et al. 2009, 2012, 2013), Pindel (Ye et al. 2009), GATK (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013), BreakDancer (Chen et al. 2009; Fan et al. 2014), Genome STRiP (Handsaker et al. 2011), and Strelka (Saunders et al. 2012)—with supporting genomics software and the S3 Tools cloud toolkit for interacting with the AWS Simple Storage Service (S3) cloud (Sarna 2011). GenomeVIP automatically mounts volumes from Amazon's Elastic Block Store (EBS) resource containing any user-specified BAM alignment and reference genome files to the cluster. Examples of mountable EBS volumes include encrypted volumes created by users, data providers, or collaborators under the given AWS account and any public, unencrypted volumes. (GenomeVIP does not address access control lists [ACLs] for data, but we note their use at the AWS account level may help to enforce regulatory data access and usage policies.) Files located in S3 are another source of input. When executed, the master script automatically submits work scripts as batch jobs to the local job resource manager (e.g., SGE on AWS and LSF on local clusters) for handling job concurrency and dependencies. As work units are completed, GenomeVIP transmits raw and final results to the specified storage location (i.e., S3 storage on AWS or a results directory on local clusters).

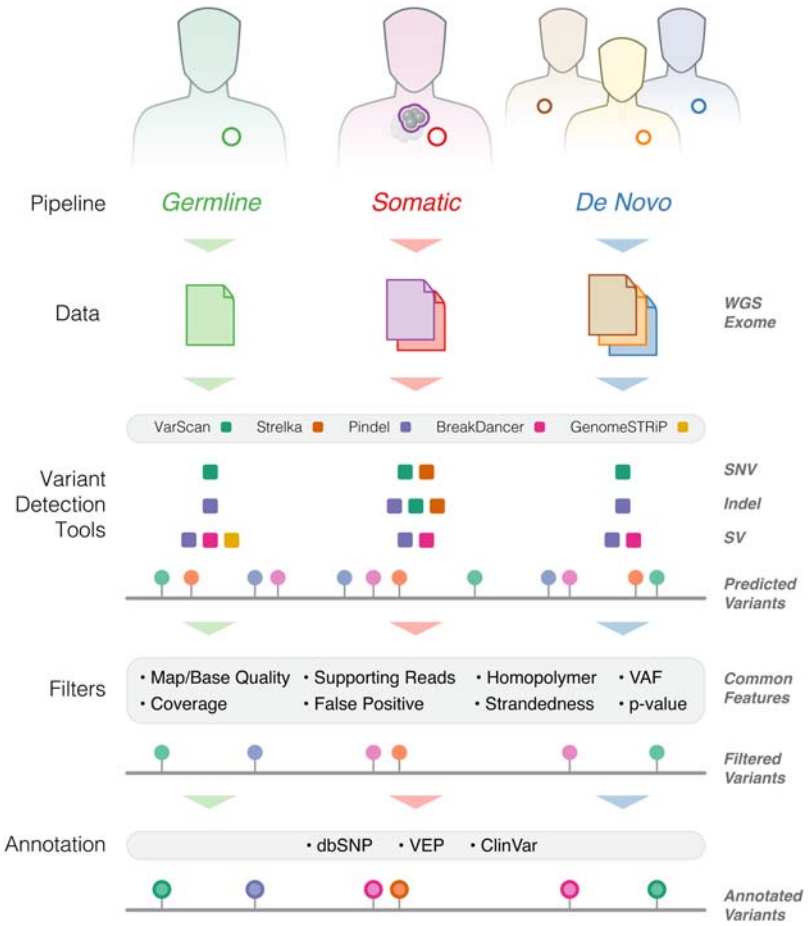

Figure 2. GenomeVIP workflows. Three variant-discovery pipelines (germline, somatic, and de novo) with predicted variant types, including single-nucleotide variants (SNVs), insertions and deletions (indels), structural variants (SVs); selected filtering features; and post-discovery annotation options provided by third-party software packages having knowledge of catalogs of genetic variation.

GenomeVIP components

The functional units within GenomeVIP collectively implement variant analysis of genomic data (Fig. 2). They encompass sets of public tools most trusted and most often relied upon by investigators for their respective tasks.

Variant detection

GenomeVIP deploys pipelines for germline, somatic, and de novo variant discovery with helper modules for performing filtering and annotation. The platform incorporates well-established, widely used tools, i.e., VarScan, GATK, Strelka, BreakDancer, Pindel, MuTect2, and Genome STRiP, for the detection of single-nucleotide variants (SNVs), small insertions and deletions (indels), and structural variants (SVs) that include inversions, copy-number variation, tandem duplications, and inter- and intra-chromosomal translocations. Results are reported in VCF format.

Variant filtering

GenomeVIP provides several user-adjustable filtering modules, which encapsulate various tool-specific methods or rule-based heuristics, to refine raw variant calls. Such filters include VarScan's own germline and somatic high-confidence filter as well as a new de novo/family trio filter we developed that considers the maximum total number of supporting reads in the parents. The BreakDancer filter we developed removes somatic calls having read support in the normal or de novo calls having read support in either parent. Pindel analysis incorporates a unified germline, somatic, and de novo filter we developed that considers read coverage, variant allele frequency, strandedness, read support, and homopolymer repeat length. Further details on these three new filters are provided in Methods. GenomeVIP also provides dbSNP (NCBI v.142, GRChr37 and GRCh38) and a false-positive filter as high-level filters. These methods are based in part on published (Xu et al. 2014) best practices for these tools, our results in the ICGC-TCGA DREAM Somatic Mutation Calling challenge

(http://www.synapse.org/#!Synapse:syn312572/wiki/58893), and our germline and somatic calling in the ICGC-TCGA Pan-Cancer Analysis of Whole Genomes Somatic Mutation Calling Challenge (Pilot-63) Validation project (http://www.synapse.org/#!Synapse:syn2875157). Finally, GenomeVIP can accept user-designated VCF files for filtering against a pan-normal data set.

Variant annotation

GenomeVIP furnishes several annotation methods: dbSNP, which provides various information on all known short sequence variation; the Ensembl Variant Effect Predictor system (VEP) (McLaren et al. 2010), including its interfaces to individual tools like SIFT (Ng and Henikoff 2003) and PolyPhen (Adzhubei et al. 2013), which provides information on the impact of variants on, e.g., genes, transcripts, protein sequences, and regulatory regions; and ClinVar (Landrum et al. 2016), which provides interpretations of the clinical significance of variants.

GenomeVIP dynamic interface

GenomeVIP session configuration spans the spectrum of running "out of the box" almost wholly on defaults to accepting customized user instructions for practically every step. Following modern user application programming practices, the graphical interface is highly menu-driven. A navigational menu summarizes the core steps of the configuration and execution process (Fig. 3, top). Activation of any of these items in the interface (via user mouse click) updates an adjacent display panel to show the corresponding options (Fig. 3A–F), each of which is described in more detail below.

Accounts

First, users provide their login credentials for Amazon Web Services (AWS) or a local high-performance compute cluster, depending on where the input data are located and thus where the computations are to be run (Fig. 3A). GenomeVIP employs semipersistent "sessions" to facilitate reuse of instantiated AWS computing resources. For local clusters, users may submit jobs to specific hosts and queues.

Select genomes

Next, users specify genomic data sources (Fig. 3B). GenomeVIP recognizes sequence alignment files in BAM format and reference genomes in FASTA or compressed FASTA format. On AWS, users may provide cloud storage volumes with prepared file lists or opt to use one of the prepared 1000 Genomes Project donor sample sets (i.e., pilot phase; phase 1 low-coverage or exome; or phase 3 low-coverage, high-coverage, or exome). On local clusters, GenomeVIP can obtain remote directory file listings directly (via SSH secure shell). The server parses the data sources and presents the alignment and reference files for selection. GenomeVIP notes any missing index or dictionary files and generates directives to create them at run time as necessary. This framework is sufficiently general to handle nonhuman genomes wherever supported by the variant detection tools.

Execution profile

In this series of tabbed panels (Fig. 3C,D), users design an "execution profile" consisting of their choice of variant detection, filtering, and annotation tools. The "Quick Setup" tab (Fig. 3C) provides access to several built-in execution profiles, based on best practices and our own experience that can afford high-quality calls. A Run Mode field sets the study type (i.e., germline, somatic, or de novo) and filtering and annotation options, and a Parameters field accounts for the sequence data type (i.e., whole-genome or exome), depth of coverage (i.e., low or targeted), and palette of applied filters. Users may also upload execution profiles through their web browser, enabling them to readily reuse settings from prior computations to ensure consistency across multiple runs. Users can select from predefined chromosome sets or specify a custom list of chromosomes and/or regions by entering details directly or by uploading a list file of regions. Options within the individual tools' tabs or the Post-discovery Analysis tab allow customization of the current profile (Fig. 3D). For example, the user can include or exclude particular discovery tools from the analysis or modify the more commonly altered discovery parameters. GenomeVIP also utilizes collapsible panels (e.g., Fig. 3D, "Options") to provide users with access to other command line parameters offered by particular tools. Modules for performing false-positives filtering, including filtering by a user-supplied panel-of-normals VCF file, and annotation by dbSNP, VEP (McLaren et al. 2010), and ClinVar (Landrum et al. 2016) are available.

Submit

In the final steps of preparing a computation, users select the computing resource and finalize the execution profile (Fig. 3E). On AWS, users select whether to launch a new virtual computing cluster or reuse an existing cluster and specify the top-level destination (S3 "bucket") for storing results and supporting output (see below). On local clusters, users designate the working directory in which the computations are executed. Buttons to preview, validate (or error-check), or download the current execution profile are available. These functions also are available to users throughout the configuration process to assist them in preparing their job so that it executes as intended. Submitting the computation causes a final validation check to be performed and, if successful, a summary of the submission is displayed with a jobID for identifying the computation.

Results

GenomeVIP places various outputs in the location specified by the user during the Submit step. As shown in Figure 3F, the main results are placed in a "Results" folder, and the raw outputs and intermediate results are stored in folders corresponding to individual variant tools. As the time needed to complete an analysis is expected to vary significantly depending on the type of analysis and the complexity of the sample, GenomeVIP uses a "status" subdirectory with sentinel files indicating which discrete jobs in the workflow are unfinished. The main location contains a copy of workflow script (file: *.sh) and the execution profile (file: *.ep) serving as a record of the computation. Execution profiles may be uploaded to GenomeVIP in future computations to ensure uniformity in processing across multiple runs. Users may obtain results and job status information for AWS computations either by providing the S3 storage location and jobID of the computation to a running GenomeVIP server, or by logging into the AWS Console ("dashboard") to navigating to the S3 storage service. Either of these methods allows users to download files through the browser using secure HTTPS protocols. Access to results and job status for computations run on local clusters can be obtained in an analogous manner to AWS by providing the working

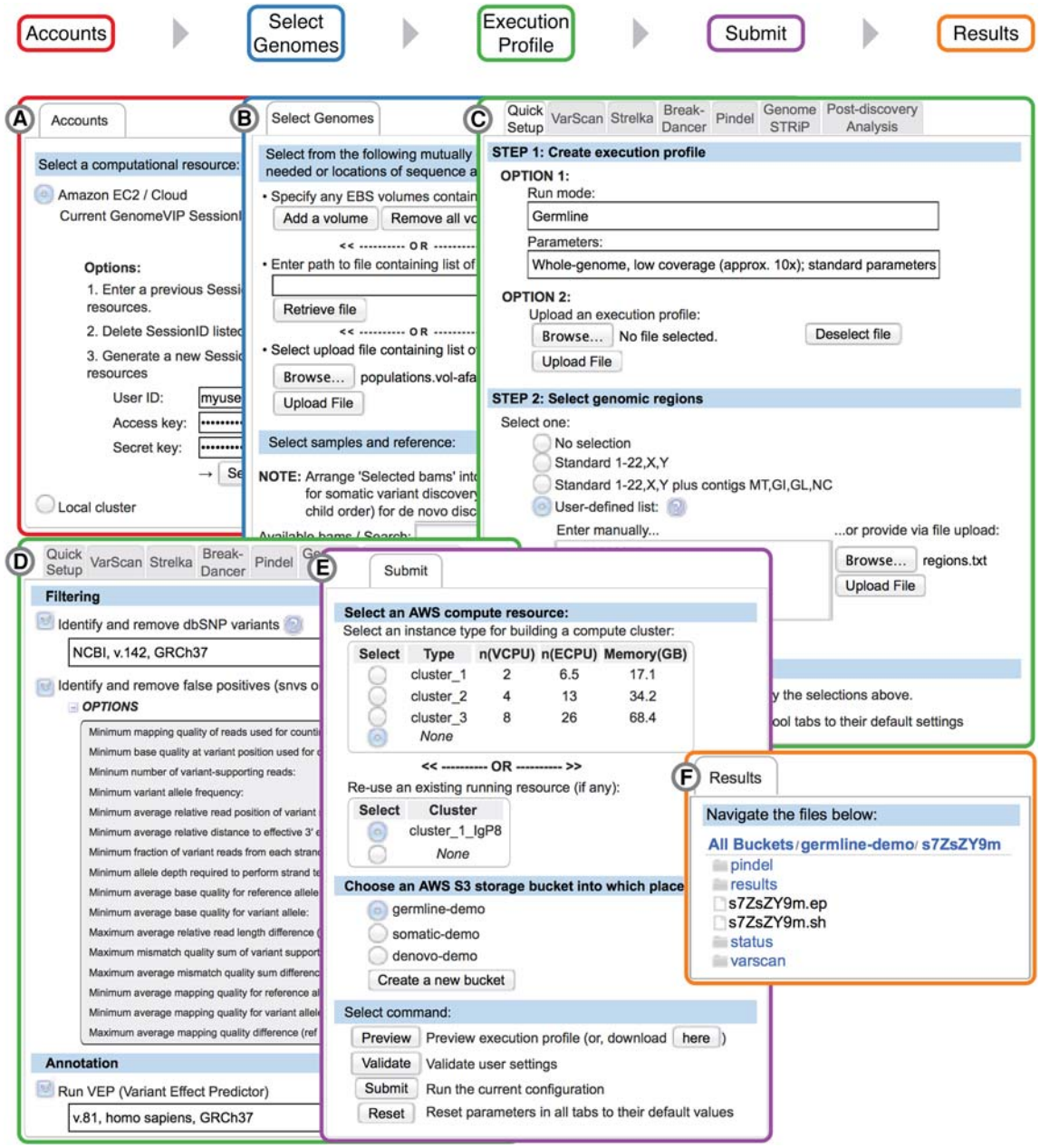

Figure 3. GenomeVIP screenshots. (*A*) Accounts. Presentation of the user's valid Amazon Web Services causes GenomeVIP to generate a semipersistent sessionID used to store or recall previous cloud resource configurations. (*B*) Select Genomes. A user-uploaded file listing sequence alignment, reference, and index files is parsed and displayed for item selection. (*C*) Quick Setup tab configuration for loading a built-in execution profile with predefined tools and parameters (Step 1, option 1); a profile may alternatively be uploaded via the interface (Step 1, option 2). Predefined genomic regions may be selected or uploaded via the interface (Step 2). Clicking the Apply Profile button (Step 3) configures tools listed under the other tabs (gray) with the current predefined profile and regions, which may be subsequently modified manually under the other tabs. (*D*) Post-discovery Analysis. Selection of filters and annotation as part of the execution profile, showing the expanded false-positives filter panel (gray) for customization. (*E*) Submit. Resource management options are provided to create new or reuse existing computing instances and cloud storage location. Buttons to preview, download, or error-check the current execution profile, or to submit it as a computation, are available. (*F*) Results. An Amazon cloud storage file listing showing folders for tools' outputs, job status, and results. Files .sh and .ep represent the master script describing the computation's workflow and the execution profile, respectively.

directory of the computation to a running GenomeVIP server or via command-line login to the user's cluster account.

Options

Features in this panel (not shown) allow users to manage several advanced features of Amazon cloud operations, such as deactivating the use of encrypted streams for data in transit to and from AWS S3 cloud storage and terminating instantiated AWS clusters. Users may launch an updated GenomeVIP runtime machine image instead of the server's built-in default by providing the corresponding Amazon machine image (AMI) identification tag.

Discussion

Use with big data

In addition to numerous smaller-scale projects, GenomeVIP has been used in several major projects, two of which are listed in Table 1: the International Cancer Genome Consortium (ICGC) Somatic Pilot and the TCGA PanCan Germline projects. In the ICGC project, GenomeVIP invoked the native somatic calling capability of VarScan and Strelka and directed Pindel to generate calls for each matched pair in cohort mode. Generating these calls consumed more than 9000 CPU-hours, with the pairs requiring on average 11.7, 40.8, and 135 CPU-hours for VarScan, Strelka, and Pindel, respectively. In the TCGA PanCan Atlas germline analysis project, more than 8000 samples across 22 tumor types were processed individually with VarScan, GATK, and Pindel using our in-house pipeline. Comparison of these local-based calls to those generated more recently on cloud resources have produced excellent recall rates for insertions and deletions (~96%) and for single-nucleotide variants (>99%), the difference being largely attributable to the use of more recent tool versions. The tool-specific CPU usage here was 8.2, 5.4, and 10.2 CPU-hours for VarScan, GATK, and Pindel, respectively.

Comparison to existing cloud pipelines

Genomic sequencing has long been supported by automated processing of raw data (Wendl et al. 1998); as a result, a variety of systems, many of which are deployable on cloud computing resources, have become available with varying levels of capability, user convenience, and sophistication. Early entries only targeted specific parts within the larger process, e.g., applying cloud technologies to the genome alignment process (Langmead et al. 2009; Schatz 2009; Wall et al. 2010), but recent work has progressed toward solutions that are more complete. The positioning of GenomeVIP within this milieu has been guided by its underlying design goal of furnishing an intuitive, graphics-based system to the nonspecialist biomedical investigator for harnessing well-established, task-specific, external tools to analyze WGS/WES data for somatic, germline, and de novo variants using on-demand cloud resources. A comparison of features across a selection of comparable systems is presented in Table 2.

Intuitive, web–based control

Special bioinformatics skills related to transferring large files, database creation, programming, or algorithms are required to use systems such as Atlas2 Genboree (Evani et al. 2012), the COSMOS library (Gafni et al. 2014), or the CloudBioLinux tool set (Krampis et al. 2012), and familiarity with Unix/Linux command-line functionality is needed to easily run tools like TREVA (Li et al. 2014), TOGGLE (Monat et al. 2015), HugeSeq (Lam et al. 2012), GotCloud (Jun et al. 2015), or Churchill (Kelly et al. 2015). Many systems also have more subtle aspects of designing a calculation, for example, manually discretizing genomic regions for parallelization (Afgan et al. 2010). GenomeVIP enables complete specification of job execution entirely by web-interface menu prompting, making it easier for users without considerable bioinformatics experience to undertake genomic analyses.

Vetted tools

Some approaches rely on native methods for processing, e.g., Atlas-SNP2 and Atlas-Indel2 in the Atlas2 Cloud and Mercury (Reid et al. 2014) systems, and almost all lack capability for managing the full array of biomedically relevant variant types. For example, most systems (Evani et al. 2012; Li et al. 2014; Jun et al. 2015; Kelly et al. 2015) cannot handle SVs, and none treat any type of complex indels. GenomeVIP follows a UNIX-like philosophy of recruiting only highly vetted, task-specific external tools and porting them for cloud compatibility. GenomeVIP runs each tool with fully specified parameter lists to guard against inadvertent changes of defaults. It also supports the reuse of custom execution profiles for repeating an analysis, which may be helpful for performing longitudinal studies on the same patient or for simply dividing a large data set into smaller computational sets.

Data privacy protection

The GenomeVIP server uses secure HTTPS by default, with encryption specifications consistent with HIPAA standards, to communicate with the user's web browser. The StarCluster toolkit uses secure shell (SSH) encrypted protocol for communicating with and transferring files to/from Amazon. The S3 Tools toolkit is configured by GenomeVIP by default to use HTTPS when transmitting data between computing resources and S3 cloud storage and to request server-side encryption (AES-256 protocol) be applied to new data stored at rest in S3. When used with a local cluster, GenomeVIP uses SSH when accessing the remote file system and when transferring files to/from the cluster.

Example comparison

We compared GenomeVIP directly to another cloud-based analysis system, namely GotCloud (Jun et al. 2015). As the latter does not readily handle somatic or trio analysis, we have limited the comparison to germline variants. In particular, we used GotCloud to repeat the germline analysis of the nine cases from the 1000 Genomes Project discussed above. Downstream analysis revealed that dbSNP concordances of GotCloud-generated calls were 99.0%–99.7% for SNVs and 88.5%–93.0% for indels, both of whose ranges are comparable to the results from GenomeVIP (Fig. 4F).

Table 1. Examples of large-scale projects utilizing GenomeVIP

		Computational resources (CPU-h)				Variants (millions)		
Project	**Samples**	**VarScan2**	**GATK**	**Strelka**	**Pindel**	**SNVs**	**Indels**	**SVs**
ICGC Somatic Pilot-50	50 WGS pairs (tumor/normal)	583	N/A	2041	6770	44×10^5 [a]	0.94×10^5 [a]	1.0×10^5 [a]
TCGA Germline	8695 WXS samples	71,081	47,305	N/A	88,333	4.3×10^9 [b]	0.9×10^9 [b]	N/A

[a]Unique, filtered.
[b]Nonunique, raw.

Table 2. Brief comparison of variant discovery frameworks

		Pipeline				Variants							
Software	Version	Som	Germ	De novo	SNV	Indel	SV	CNV	Anno	Web GUI	Cloud API-Aware	Machine image available	User-installed required
GenomeVIP	v1.2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓(AWS)	optional
TREVA	v.1	✓	✓		✓	✓		✓	✓			✓	
HugeSeq	v2.0		✓		✓	✓	✓	✓	✓				✓
Atlas2 Genboree	website, accessed 2016-01		✓		✓	✓			✓	✓			
Mercury	v3.2.1		✓		✓	✓			✓	✓(DNAnexus)	✓(DNAnexus)		optional
Churchill	v1.8		✓		✓	✓			✓				✓
GotCloud	v1.14.4, ami-6ae65e02		✓		✓	✓			✓			✓(AWS)	
TOGGLE	v0.2 (2015-10-08)		✓		✓	✓			✓			✓	optional

Software packages are compared according to the following: the types of pipelines (Som, somatic; Germ, germline; De novo, de novo) and callable variants (SNV, single-nucleotide polymorphism; Indel, short insertions and deletions; SV, structural variants; CNV, copy number variation) available; built-in annotation (Anno) options; presence of native or supported web browser graphical user interface (Web GUI) and built-in cloud resource management tools (Cloud API-Aware); availability of ready-to-run, pre-built machine images; and requirement for manual installation of the software package itself and/or supporting genomics software.

Methods

GenomeVIP genomic applications

GenomeVIP includes tools to perform germline, somatic, and de novo variant discovery and annotation. We illustrate these capabilities with three examples: (1) germline variant discovery from exome samples from a large cohort; (2) somatic variant detection on a synthetic matched tumor/normal sample pair; and (3) de novo analysis of a well-studied family trio from the 1000 Genomes Project. Parameters used are available in Supplemental Information. In each case, local and cloud deployments of GenomeVIP produced identical raw calls. We also confirmed the parallelizability of genomic regions by performing cloud computations over entire chromosomes and local computations by processing 10-Mb windows spanning the entire genome. Performance and compute statistics are reported in Figure 4D and Supplemental Information.

Germline

We selected nine unrelated donors from the 1000 Genomes Project (Abecasis et al. 2010), three each from three populations (CHB, FIN, YRI) and directed GenomeVIP to launch VarScan and Pindel variant callers to perform SNV and indel discovery on Chromosome 20 using discovery parameters used in a previous germline analysis (Kanchi et al. 2014). Raw calling performance, as measured by dbSNP concordance, revealed SNV concordances ranging from 97.0% to 98.1% and indel concordances ranging from 92.2% to 96.2%. To validate the germline variants, we then conducted a downstream principal component analysis of the SNVs using PLINK (Chang et al. 2015) and found that analysis faithfully recapitulated the population structures of these three ethnic groups (Fig. 4A).

Somatic

We called somatic mutations in the matched tumor/normal synthetic DREAM-3 samples, the most complex of the open-access data sets from the ICGC-TCGA DREAM Somatic Mutation Calling (SMC) Challenge (Ewing et al. 2015). We directed GenomeVIP to generate raw calls using VarScan and Strelka with limited filtering to enable downstream exploration of the effect of selected false-positive filtering parameters, namely, the number of supporting reads and read mapping qualities (Methods). Calculations of true-positive (TP) and false-positive (FP) rates, based on unmasked regions of a known synthetically generated tumor's "truth" set (Ewing et al. 2015), for multiple combinations of parameter sets for the individual, intersected, and combined filtered call sets, is plotted on receiver operating characteristic (ROC) curves (Fig. 4B). Comparison of unfiltered, novel calls reveals comparable TP but significantly different FP rates for VarScan and Strelka callers. Although Strelka's TP rates (ranging from 0.784 to 0.789) and FP rates (ranging from 0.011 to 0.016) did not appreciably change across the false-positives parameter landscape investigated, increasing the number of required variant-supporting reads (VSR) to four for filtering VarScan calls dramatically decreased FP rates while also modestly decreasing TP rates. This contrast in behavior is a result of VarScan having produced manifold more putative calls than Strelka in the "unmasked" genomic regions targeted by the evaluator script. Finally, the intersections of calls from the two callers were found to be largely invariant (TP, ~62.7%; FP ~0.94%), supporting the view that combining calls from multiple callers may be an effective strategy to identify a core set of high-quality calls while mitigating against a significant portion of false positives (Cantarel et al. 2014).

This exercise also serves as a case study, showing the creation of expert knowledge and experience that investigators may wish to capture for subsequent analyses, such as maximizing sensitivity for discovery or specificity for diagnostic purposes. Users can then provide these optimal parameter values as previously described to perform a complete run.

A similar approach was used to process WGS tumor/normal matched pairs from more than 10 different cancer types for the ICGC Pan-Cancer project. In the initial "Pilot 50" project, GenomeVIP produced 4.4 million SNV, 94,000 indel, and 100,000 SV filtered calls (Table 1) that we submitted for evaluation for selection for validation experiments. The comparison between GenomeVIP's predictions and those from eight other standard pipelines showed that GenomeVIP produces results highly concordant with noncloud pipelines (Fig. 4C), suggesting high-quality, reproducible analysis.

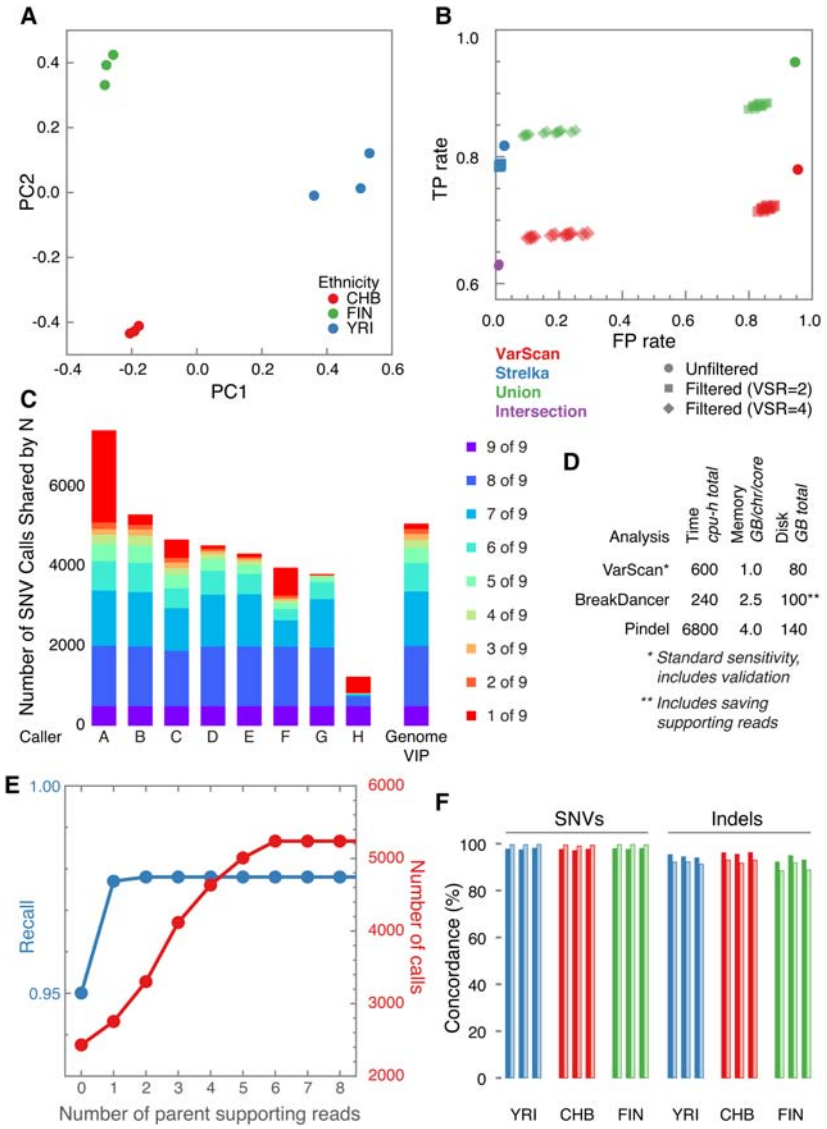

Figure 4. Applications of GenomeVIP. (*A*) Principal component analysis of germline SNV and indel predictions for nonrelated 1000 Genomes Project Phase 1 samples from three populations: (red) CHB; (green) FIN; (blue) YRI. (*B*) True-positive (TP) and false-positive (FP) rates for somatic SNV calls novel to dbSNP. Performance of VarScan and Strelka callers individually (red, blue) and in combination (green, purple) are evaluated before and after exploratory false-positives filtering using multiple parameter combinations, in which VSR is the minimum number of variant-supporting reads. (*C*) GenomeVIP performance on ICGC Pan-Cancer Pilot-50 somatic mutation calling for one matched sample pair, in which the colors correspond to the number of pipelines predicting the same variant. (*D*) Performance statistics. (*E*) De novo recall performance (blue), as compared to published experimental validation results, and filtered call set size (red) for SNV calling in NA12878 as a function of PVSR, the number of variant-supporting reads in parental genomes NA12891 and NA12892. (*F*) dbSNP concordances of germline SNVs and indels, as called by GenomeVIP (darker shading) and GotCloud (lighter shading), for the samples described in *A*.

De novo SNV mutations in family trios

We analyzed the NA12891-NA12892-NA12878 family trio samples for de novo single-nucleotide variants using VarScan and associated filtering modules provided by GenomeVIP and compared the results to experimentally validated germline and somatic de novo variants (Conrad et al. 2011). GenomeVIP attained 97.8% recall (979 of 1001 experimentally validated sites) after predicted false positives were removed. The additional effect of applying a variant read-support filter for the parental genomes showed that exclusion of variant support in the parents (PVSR = 0) yielded the smallest GenomeVIP call set ($N = 2431$) with a recall rate of 95.0% (Fig. 4E, blue curve); in this set, 41 calls were validated experimentally as false positives. Increasing PVSR resulted in larger GenomeVIP variant call sets by at least 13% (Fig. 4E, red curve), while improving recall rates only marginally. For example, at PVSR = 2, at which the recall rate has plateaued, GenomeVIP made 3302 final calls, of which 42 validated as false positives. In this example, the highest balanced accuracy is likely to be obtained for values of PVSR at or near zero.

Software availability

The GenomeVIP source code and associated scripts are freely available for academic use and are available through GitHub (https://github.com/ding-lab/GenomeVIP/) and as Supplemental Code. Users having Amazon Web Services (AWS) login credentials can launch a GenomeVIP server on the AWS cloud by instantiating server images located in the AWS public repository. GenomeVIP executes computations using a public run-time AWS image providing a nearly complete set of the required genomics software. Software packages on these images carry their own licensing and usage terms. For example, GenomeVIP requires users to provide the location of their own licensed copy of GATK (version 3.5 and higher supported). Installation of GenomeVIP on a local web server allows users to design and execute computations on a local high-performance compute cluster using local data or on Amazon's cloud using data stored at AWS; furthermore, GenomeVIP's internal configuration files may be edited manually to point to local versions of the tools on which GenomeVIP depends. GenomeVIP orchestrates many tools that are all upgraded independently, and these updates will be passed through to GenomeVIP users in the following ways: Tool updates not requiring user interface modification will be distributed in updated runtime images, but those requiring such modification will require an updated server image. Tool versions will accumulate rather than be replaced to preserve backward compatibility. Database updates will be managed similarly, although the user can alternatively specify the location of public or custom annotation VCFs that GenomeVIP can retrieve via FTP/HTTP or from Amazon cloud storage. Documentation, support, and further information is available through GitHub and the GenomeVIP home page at our Turnkey Variant

Analysis Project website (http://tvap.genome.wustl.edu/tools/genomevip/).

Acknowledgments

This work was supported by the National Human Genome Research Institute grant U01 HG006517 to L.D. and National Cancer Institute grants R01CA180006 and R01CA178383 to L.D. and R01CA172652 to K.C. We thank Seva Kashin and David Larson for useful comments on tool usage and Kimberly Johnson, Michael McLellan, Matthew Bailey, and Mingchao Xie for helpful discussion. We thank the ASHG for hosting a workshop where we introduced a prerelease version GenomeVIP.

Author contributions: Software design and development (R.J.M., A.D.S.); genomics application design (K.Y., R.J.M., M.C.W., L.D.); genomics data analysis (R.J.M., C.J.Y., V.D.Y., K.H., D.C.K.); filtering methodologies (V.D.Y., D.C.K., K.Y., R.J.M., B.N.); tool implementation (A.D.S., B.N., R.J.M., K.C., R.E.H., D.C.K.); figure preparation (C.J.Y., K.H., R.J.M., M.A.W.); web interface design (M.A.W., R.J.M.); manuscript text (R.J.M., K.H., C.J.Y., M.C.W., D.F., B.J.R., L.D.); interface and software testing (R.J.M., A.D.S., E.D.); project direction and supervision (L.D.).

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **76:** 7.20.1–7.20.41.

Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. 2010. Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics* **11**(Suppl 12): S4.

Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J. 2014. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics* **15:** 104.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4:** 7.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6:** 677–681.

Collins FS, Varmus H. 2015. A new initiative on precision medicine. *N Engl J Med* **372:** 793–795.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43:** 712–714.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498.

Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. 2010. Translational bioinformatics in the cloud: an affordable alternative. *Genome Med* **2:** 51.

Evani US, Challis D, Yu J, Jackson AR, Paithankar S, Bainbridge MN, Jakkamsetti A, Pham P, Coarfa C, Milosavljevic A, et al. 2012. Atlas2 Cloud: a framework for personal genome analysis in the cloud. *BMC Genomics* **13**(Suppl 6): S19.

Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, et al. 2015. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* **12:** 623–630.

Fan X, Abbott TE, Larson D, Chen K. 2014. BreakDancer: identification of genomic structural variation from paired-end read mapping. *Curr Protoc Bioinformatics* **45:** 15.6.1–15.6.11.

Gafni E, Luquette LJ, Lancaster AK, Hawkins JB, Jung JY, Souilmi Y, Wall DP, Tonellato PJ. 2014. COSMOS: Python library for massively parallel workflows. *Bioinformatics* **30:** 2956–2958.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43:** 269–276.

Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, Bandlamudi C, McNerney ME, White KP, Grossman RL. 2014. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *J Am Med Inform Assoc* **21:** 969–975.

Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25:** 918–925.

Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, et al. 2014. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun* **5:** 3156.

Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, Nordquist RD, Newsom DL, White P. 2015. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol* **16:** 6.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25:** 2283–2285.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22:** 568–576.

Koboldt DC, Larson DE, Wilson RK. 2013. Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr Protoc Bioinformatics* **44:** 15.14.11–15.14.17.

Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson KE. 2012. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC Bioinformatics* **13:** 42.

Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, et al. 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol* **30:** 226–229.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44:** D862–D868.

Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. 2009. Searching for SNPs with cloud computing. *Genome Biol* **10:** R134.

Li J, Doyle MA, Saeed I, Wong SQ, Mar V, Goode DL, Caramia F, Doig K, Ryland GL, Thompson ER, et al. 2014. Bioinformatics pipelines for targeted resequencing and whole-exome sequencing of human and mouse genomes: a virtual appliance approach for instant deployment. *PLoS One* **9:** e95217.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26:** 2069–2070.

Monat C, Tranchant-Dubreuil C, Kougbeadjo A, Farcy C, Ortega-Abboud E, Amanzougarene S, Ravel S, Agbessi M, Orjuela-Bouniol J, Summo M, et al. 2015. TOGGLE: toolbox for generic NGS analyses. *BMC Bioinformatics* **16:** 374.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814.

Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundquist A, English A, Bainbridge M, White S, Salerno W, Buhay C, et al. 2014. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15:** 30.

Sarna DEY. 2011. *Implementing and developing cloud computing applications*. CRC Press, Boca Raton, FL.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28:** 1811–1817.

Schatz MC. 2009. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25:** 1363–1369.

Stein LD. 2010. The case for cloud computing in genome informatics. *Genome Biol* **11:** 207.

Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO. 2015. Data analysis: create a cloud commons. *Nature* **523:** 149–151.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43:** 11.10.11–11.10.33.

Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, Patil P, Tonellato PJ. 2010. Cloud computing for comparative genomics. *BMC Bioinformatics* **11:** 259.

Wendl MC, Dear S, Hodgson D, Hiller L. 1998. Automated sequence preprocessing in a large-scale sequencing environment. *Genome Res* **8:** 975–984.

Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. 2014. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **28:** 244.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25:** 2865–2871.

Received June 21, 2016; accepted in revised form May 3, 2017.