EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics

Sophie Coutant, Chloé Cabot, Arnaud Lefebvre, Martine Léonard, Elise Prieur-Gaston, Dominique Campion, Thierry Lecroq, Hélène Dauchel

Coutant et al. BMC Bioinformatics 2012, 13(Suppl 14):S9
http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S14/S9

BMC Bioinformatics

RESEARCH Open Access

EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics

Sophie Coutant[1,2,3], Chloé Cabot[2,3], Arnaud Lefebvre[2,3], Martine Léonard[2,3], Elise Prieur-Gaston[2,3], Dominique Campion[1,3], Thierry Lecroq[2,3], Hélène Dauchel[2,3*]

From NETTAB 2011 Workshop on Clinical Bioinformatics
Pavia, Italy. 12-14 October 2011

Abstract

Background: Whole exome sequencing (WES) has become the strategy of choice to identify a coding allelic variant for a rare human monogenic disorder. This approach is a revolution in medical genetics history, impacting both fundamental research, and diagnostic methods leading to personalized medicine. A plethora of efficient algorithms has been developed to ensure the variant discovery. They generally lead to ~20,000 variations that have to be narrow down to find the potential pathogenic allelic variant(s) and the affected gene(s). For this purpose, commonly adopted procedures which implicate various filtering strategies have emerged: exclusion of common variations, type of the allelics variants, pathogenicity effect prediction, modes of inheritance and multiple individuals for exome comparison. To deal with the expansion of WES in medical genomics individual laboratories, new convivial and versatile software tools have to implement these filtering steps. Non-programmer biologists have to be autonomous combining themselves different filtering criteria and conduct a personal strategy depending on their assumptions and study design.

Results: We describe EVA (Exome Variation Analyzer), a user-friendly web-interfaced software dedicated to the filtering strategies for medical WES. Thanks to different modules, EVA (i) integrates and stores annotated exome variation data as strictly confidential to the project owner, (ii) allows to combine the main filters dealing with common variations, molecular types, inheritance mode and multiple samples, (iii) offers the browsing of annotated data and filtered results in various interactive tables, graphical visualizations and statistical charts, (iv) and finally offers export files and cross-links to external useful databases and softwares for further prioritization of the small subset of sorted candidate variations and genes. We report a demonstrative case study that allowed to identify a new candidate gene related to a rare form of Alzheimer disease.

Conclusions: EVA is developed to be a user-friendly, versatile, and efficient-filtering assisting software for WES. It constitutes a platform for data storage and for drastic screening of clinical relevant genetics variations by non-programmer geneticists. Thereby, it provides a response to new needs at the expanding era of medical genomics investigated by WES for both fundamental research and clinical diagnostics.

* Correspondence: helene.dauchel@univ-rouen.fr
[2]University of Rouen, LITIS EA 4108 Computer science, information processing and systems laboratory, 76821 Mont-Saint-Aignan cedex, France
Full list of author information is available at the end of the article

Background

Next-generation sequencing (NGS) technologies are widely used to answer key biological questions at the scale of the entire genome and with an unprecedented depth [1-4]. Whether determining genetic or genomic variations, cataloguing transcripts and assessing their expression levels, identifying DNA-protein interactions or chromatin modifications, surveying the species diversity in an environmental sample, all these tasks are now tackled with large-scale sequencing and require computer intensive bioinformatic analyses [5-7], although different.

Identification of genetic variations can be addressed by whole genome sequencing (WGS) or whole exome sequencing (WES) of single individuals. WGS is particularly attractive because it allows to access the full spectrum of genetic variations, i.e. coding and non coding Single Nucleotide Variations (SNV) and short insertion-deletion variants (indels), as well as Copy Number Variants (CNV) and Structural Variants (SV) [2,8]. In practice, out of major genome centers and *a fortiori* for the clinical routine translation, the development of this approach is still constrained by various difficulties such as the production organization, the yet expensive cost, the actual error rate of the technologies (~ 1 error per 100 kb; ~30, 000 erroneous variant calls for the whole genome), the sheer volume of data to store and to transfer, requiring intensive informatics infrastructures and robust bioinformatics and filter procedures to retain only clinically relevant variants [8,9]. As new genomes are sequenced, for example in the context of large projects like the 1000 Genomes Project [10], the number of expected variations may decrease. But, first complete individual constitutional genome sequencing studies reported 3-4 million of SNP per genome, 80-90% of which highly overlapped the National Center for Biotechnology Information public SNP database (dbSNP) [11], leaving anyway 0.5 million novel variations to sift per genome [8].

While WGS remains an appealing ultimate perspective, WES focusing on only the coding regions of the genome, has become in a few years the choice strategy to meet the challenge of identifying a coding allelic variant for rare human monogenic disorder [12]. Thanks to DNA enrichment techniques, targeted sequencing of coding regions decreases the cost and improves the efficiency of large-scale coding variations discovery compared with what would require the entire human genome. The human exome, made of ~180,000 exons for a size of ~30 Mbp, is 1.5% of the total human genome. Thereby, not only targeted selection strategy reduces the cost but also accelerates the discovery of coding genetic variants that cause rare Mendelian diseases. In 2009, Ng *et al.* [13], by using an intersection recurrence strategy, showed the proof of the concept that identifying a gene responsible for a rare dominantly inherited disorder (Freeman-Sheldon syndrome) was possible using WES of independant index cases. Since then, more and more papers confirmed the success of this strategy [14-17].

Up to now, classical approaches such as linkage analysis using genetic markers have been extensively used to identify the molecular basis for nearly 3,500 Mendelian disorders [18]. But for over 3,500 Mendelian disorders, the gene remains unknown [18,19]. The limited number of patients for rare diseases or the limited access to the related members of the family has been a frequent obstacle to conduct linkage analysis [14]. As the NGS technologies have emerged, the long and fastidious classical linkage analysis for human Mendelian disorders will be replaced by more direct identification of the causal variation(s) and the corresponding gene. Moreover, in numerous cases there are no caryotypic nor CGH-array anomaly or negative result with Sanger sequencing on known mutated genes or on neighbor genes in a pathway of interest, because of the low depth of this first generation sequencing technology [20]. So, the exome-scale sequencing approach generates a technological breakthrough in medical genetics history in fundamental research for disease gene discovery and consequently in terms of new diagnostic methods and personalized medicine [12,14,16,21].

Numerous algorithms and software tools have been developed to efficiently manage terabytes of raw sequence variation data from WES. Commonly adopted variation discovery pipelines include successive bioinformatics steps for quality control of the short reads, alignment of the short reads to a reference sequence, variation calling and variation annotation [1,19,22-24]. Generally, ~20,000 variations per individual exome are obtained. The challenge remains in efficient filtering strategies to find the causal variant(s) and corresponding gene for a rare disease, among these thousands of candidates. With this aim, additional analytical procedures which implicate various heuristic filtering strategies have emerged [19,24]. Usually, wide range common variations (more than 90% of the total) are firstly excluded. This is done by comparison to publicy available databases of human genetic variations and privately available variants from other exome sequencing projects. To narrow down the search on remaining variations (often between 200 to 500), other filters take into account the type of variations (focus on presumed deleterious allelic variants, i.e. nonsynonymous, nonsense, stop loss, frameshift, splice site) and evaluate the functional effect of variations on gene products. Usually, various criteria are inspected for this task such as the physical properties of the wild-type and variant amino acids, the structural properties affecting protein dynamics and stability, the integrity of functional motifs and binding domains or sites implicated to posttranslational processing and cellular localization of proteins, evolutionary properties derived from a sequence alignment [21-24]. Beside these

molecular nature and effects of the alternative allelic variants, filtering strategies also have to take into account the mode of inheritance of the disorder suggested by pedigree (recessive or dominant model for Mendelian disorders or sporadic cases). Finally, taking advantages of multiple individuals, intersection or differential exome strategies can drastically reduce the remaining variations to several genes.

As the exome-scale sequencing is today positioned as a method of choice for disease gene discovery and personalized medicine, the success of the unavoidable filtering strategies of thousands variations lies in their implementation into convivial and versatile software tools. End users with no computational skill have to be autonomous to conduct and combine themselves different filtering approaches, depending on their assumptions and of their study design, leading them to extract a limited list of likely candidate genes underlying a genetic disease.

With this aim, in partnership with and for medical geneticists, we developed EVA (Exome Variation Analyzer), a user-friendly web-interfaced free software dedicated to filtering strategies for medical projects investigated with exome sequencing. EVA integrates the main filters dealing with common variations, molecular types, inheritance mode and multiple samples. Here we report a demonstrative case study with EVA that allowed to identify a new candidate gene related to a rare form of Alzeihmer disease [25]. We discuss our development choices and the position of EVA among other filtering tools recently published.

Methods

Implementation of EVA

ExomeDB was developed under MySQL (5.0). The main tables are Variation, Gene and Individual in which data are integrated from a list of variants (SNV, indel) associated with their annotations (*Cf.* Methods section, 'Data' subheadings). Currently, EVA works with the NCBI 37 (hg19) build version of the human genome but integrates an archive for the previous version (NCBI 36, hg18). Each new project is subject to a remote loading using an online *Variation integration* module that accepts TXT files and XLS files. The web interface was developed under PHP (5.3.2-1). For the implemented filtering strategies (*Cf.* Result section, 'Filtering strategy module' subheadings), a combination of criteria selected by the user, is transformed into an SQL query and sent to the ExomeDB database. Then, EVA's interface displays the remaining variations in table browsers (*Cf.* Result section, figures). The *Variation statistics* module proposes interactive bar and pie vizualizations of exome data implemented with the free JavaScript charting library Highcharts (Highsoft Solutions AS). To assure the confidentiality of the exome data, EVAs integrates an *Authentication* module requiring a login and a password given by an administrator. Each login/password is strickly project specific. Users can only see and manage their own exome projects. Some performance statistics are described in the Result section. At the time of this writing, EVA's interface is accessible at the web address http://plateforme-genomique-irib.univ-rouen.fr/EVA/index.php through the described authentication process. EVA's current and update versions will be freely available under a Creative Commons Attribution-Non-Commercial-NoDerivs 3.0 Unported License (CC-BY-NC-ND) and will be downloaded from the same web site.

Data

The input file (TXT file or XLS file) of the *Variation integration* module of EVA is a list of variants (SNV and indel) obtained from an independant variant calling procedure (briefly in this study: Solexa Illumina technology, base calling from raw image files with RTA1.8/SCS2.8, Illumina pipeline CASAVA 1.7 with ELAND v2) and then annotated (in this study: proprietary bioinformatics process from IntegraGen company, Genopole® Evry, France, [26]). Although actually, the format of these files is a proprietary format, it includes classical annotations for the detected variations and the affected genes. For the detected variations main information are: the chromosome and the genomic position, the number of the read bases for each nucleotide, the reference base and the modified base deduced from an allelic count procedure and annotated with the genotype homozygosis or heterozygosis status, the number of total sequenced bases and the number of used bases for the detection variant, the score of the variation depending on the quality and coverage, the type of variation (SNV or indel), the rs name if known in dbSNP (in this study dbSNP131 [11], HapMap [27]), CIGAR format and length for the indels. For the affected genes the main information are: the gene name (NCBI GeneID), the NCBI RefSeq accession number [28] for all mRNA variants expressed by the gene, the type of affected position (exons, introns (only variations in +/-20 regions are considered), 5' or 3' UTR) and the corresponding number of the exon or intron along the gene structure, the functional categories of variations (synonym, missense, stop loss and nonsense for SNV, frameshift or not for indel), the exon or intron start and stop positions included the variation and finally position of the variations in the corresponding protein sequence with the description of the codon and corresponding amino-acid for both the reference protein and the detected variations.

Case study: the Alzheimer disease

Thanks to a nationwide recruitment (Clinical Research Hospital Program from the French Ministry of Health (GMAJ, PHRC 2008/067)), exome sequencing was performed in fourteen autosomal dominant early-onset

Alzheimer disease (ADEOAD) unrelated index cases without mutation on known genes (*Amyloid precursor protein (APP), presenilin1* and *2 (PSEN1* and *2*)) and also without known copy number variants of *APP* gene and genes involved in Amyloid beta (Aβ) peptide processing or signaling. IntegraGen company (Genopole® Evry, France, [26]) performed exome sequencing. Three micrograms of genomic DNA from each individual, extracted from peripheral blood lymphocytes and sheared by sonication to obtain an average fragment size of 150-200 bp, were used for the construction of a shotgun sequencing library using paired-end adapters. Exome capture was performed using the SureSelect Human All Exon kits 38 Mb version 1 (Agilent) (n = 12) or SureSelect Human All Exon kits 44 Mb version 2 (Agilent) for a second batch (n = 2).

Sequencing was realised on an Illumina Genome Analyser GAIIx (n = 12) or on an Illumina HiSeq 2000 (n = 2). Raw image files were processed by using the Illumina pipeline (RTA1.8/SCS2.8 and CASAVA 1.7). For the genetics variant detection, the 76 bp sequencing reads were aligned to the NCBI human reference genome (NCBI (n = 12) or NCBI 37 (n = 2)), using ELANDv2. Means coverage were of 65-fold (n = 12) and 80-fold (n = 2) with a percentage of aligned reads ranging between 88% and 95%.

Only high quality variations having a QPhred threshold > 10 were conserved (86% of the targeted bases). The annotation procedure of the detected variations only concerned those included in the coordinates given by the exon kits capture extended to +/- 20 pb in the flanking intron. The description of the annotated files is explained in the Methods section, 'Data' subheadings. Each annotated file corresponding to the project (14 individuals) was integrated in ExomeDB using the *Variation integration* module of EVA.

Results

Overview: ExomeDB and EVA web interface

For a given WES project, corresponding to several individuals and their respective variations, EVA manages data thanks to six modules (Figure 1). In input, after authentication (*Authentication module*), an online *Variation integration* module takes the variations files (for details of the format, Cf. Methods section, 'Data' subheadings) obtained from an independant variant calling bioinformatics pipeline. Annotated variations are stored in a relational database ExomeDB which main tables are the annotated variations, the corresponding genes and the individuals (Cf. Result section, 'performance' 'subheadings').

The web interface integrates four other modules for exome mining. The *Variation statistics* module allows through a guided mode selection of individuals, chromosomes, genomic regions, genes, genic region, type of SNV or indel, to summarize in tables and to graphically represent the global or selected distribution of variations of a given WES project (Figure 2). The *Table browser* module allows to precisely explore data by project, individual, gene or variation through rigorous and sortable categorized tables (Figure 3) (iii) *the Search* module can be used for a direct and quick access to a specific region, gene, variation for a given project, and finally (iv) the *Filtering strategy* module, which is the major element in exome mining to discover potential canditate genes, offers to combine filters for common variations, molecular types, inheritance mode and multiple samples to drastically narrow down variations (see details below). The selected combination is transformed into a SQL query and sent to the ExomeDB database.

Query results of the *Table browser* module, *Quick search* module and *Filtering strategy* module can be explored by five elements types presented in interactive tables: *'variation overview'* (Figure 3 and Figure 4*), 'gene list', 'gene details'* (Figure 5), *'variation list'* (Figure 6) and *'variation details'* (Figure 7).

In the 'variation overview' tables (Figure 3), the set of all the variations is divided in known and unknown variations according to the information in dbSNP. Due to the molecular process of the exome capture kit, most variations occur in exons but some detected variations also occur in splice sites. Even if ExomeDB integrates variations extended to +/- 20 pb in the flanking intron, we choose to show on the table only variations extended to +/- 2 pb in the intron, corresponding to the dinucleotide splicing site. Variations in exons can be SNV or indels. We categorized single variations into four functional classes: synonymous, miss sense, stop loss and non sense. For indels we classified into two categories: frameshift or non frameshift.

In output, EVA offers *export files (*CSV for tables, various graphical formats for the *Variation statistics* module). EVA also provides several cross-links with a selection of relevant external international databases and softwares for further functional and pathogenic effect inspection of the sorting variation and gene candidates (see details below and on Figure 5).

Filtering strategy module

The *Filtering strategy* module integrates the current main categories of filters based on common variations, molecular type of the variants, modes of inheritance, homozygous or heterozygous nature of the allelic variant and multiple individuals.

First, EVA compares the data to international catalogues of variations. In a constitutive sorting, the set of all the variations is divided in known and unknown variations according to the information in the dbSNP (Figure 3 and Figure 4). EVA also offers to reduce the number of variations by confronting them to the HapMap Project [27],

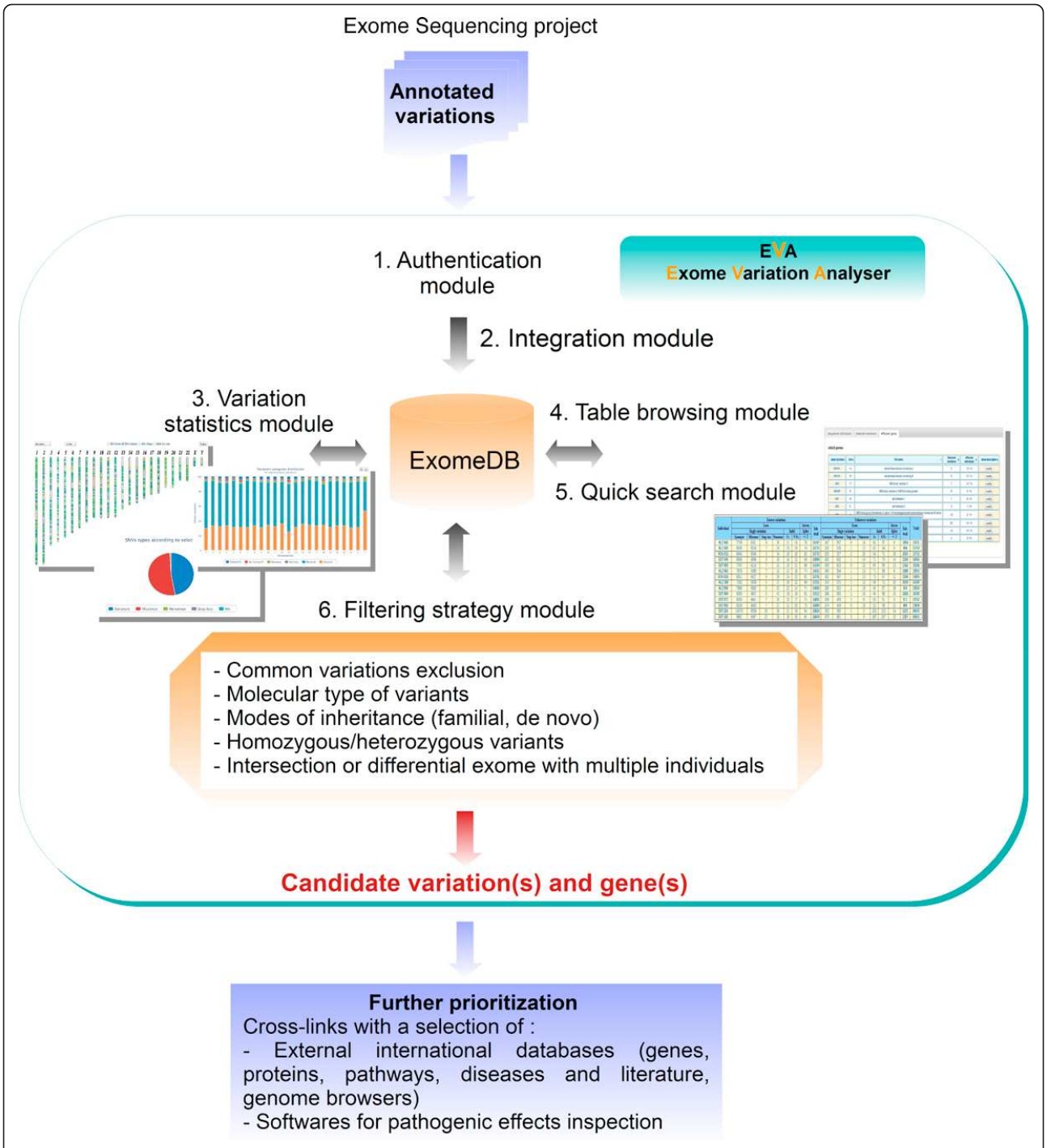

Figure 1 Overview of EVA with the six exome mining modules. To store and manage variation data, EVA consist of a database, ExomeDB and a friendly web interface. To date six modules permitted to authentify the user (1), to integrate data (2), to search and browse data with charts (3) or tables (4,5) and finally to conduct a customizable filtering strategy (6) by a combination of selected criteria dealing with common variations, molecular types of variants, inheritance modes and multiple samples intersection or difference.

the 1000 genomes Project [10], Complete Genomics public data [29], IntegraGen public data [26] or the Exome Sequencing Project [30]. In addition, other filters or table browsers offer to sift variations depending on their: (i) functional categories for SNV (synonymous, miss sense, stop loss and non sense) and indels (frameshift or non frameshift); (ii) genic region (UTR, CDS, intronic splice region) or genomic region; (iii) quality score and coverage.

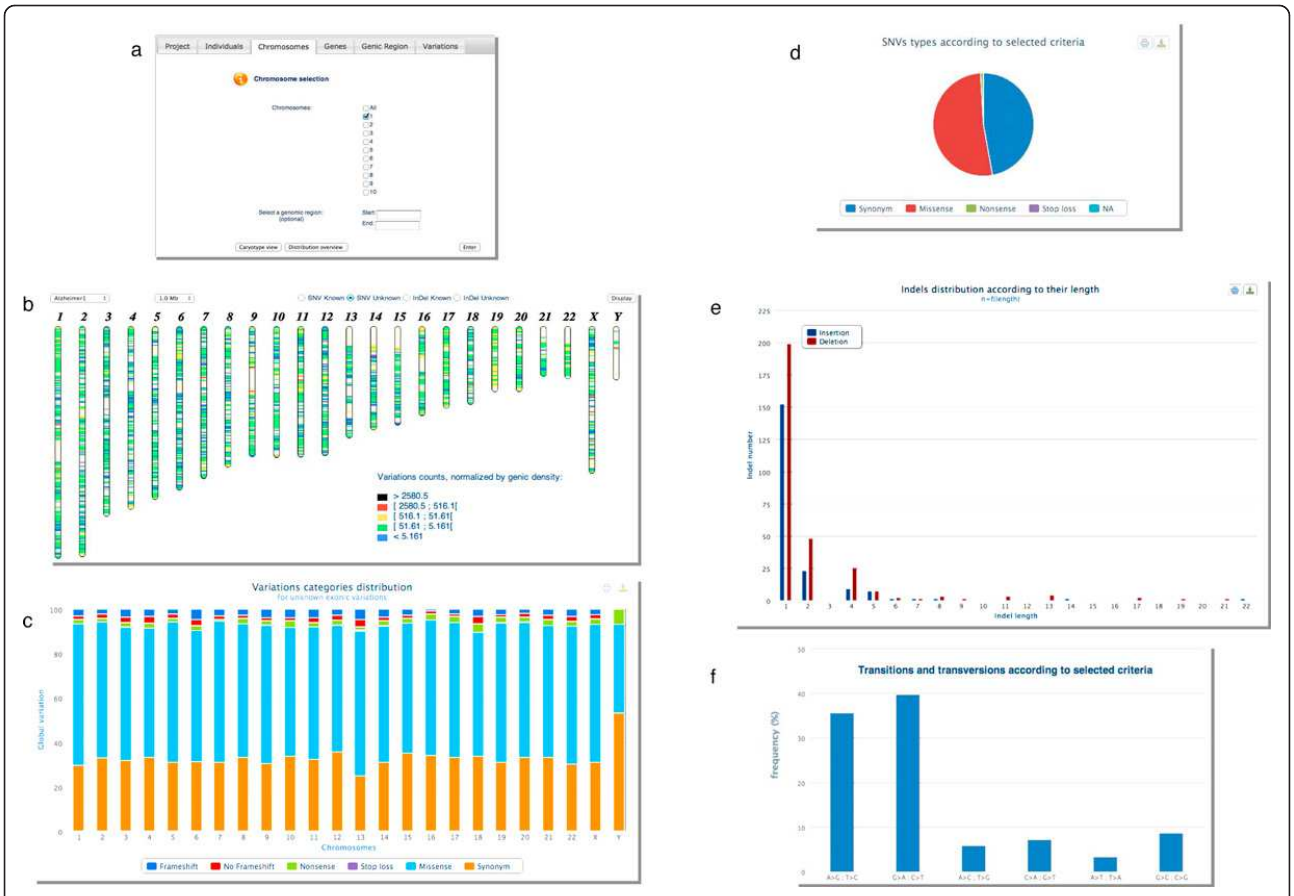

Figure 2 Examples of graphical visualization with the *Variation statistics* module of EVA. Through a guided mode selection (a), the distribution of variations for the 14 ADEOAD exome project are graphically represented by a caryotype view (b), a bar chart or a pie chart distribution (c, d). Useful analysis of indels (e), ratio of transition (purine/purine or pyrimidine/pyrimidine) substitutions to transversion (purine/pyrimidine) substitutions (f) and substitution matrix of amino-acid (not showed) are proposed.

Individual	Known variations								Unknown variations								Total
	Exon						Intron	Sub-total	Exon						Intron	Sub-total	
	Single variation				Indel		Splice		Single variation				Indel		Splice		
	Synonym	Missense	Stop loss	Nonsense	Fs	N Fs	+/- 2		Synonym	Missense	Stop loss	Nonsense	Fs	N Fs	+/- 2		
ALZ 049	7739	6301	9	30	21	19	78	14197	347	567	0	14	60	57	9	1054	15251
ALZ 426	8030	6534	7	30	20	19	76	14716	333	526	1	12	62	54	6	994	15710
ROU 632	8040	6540	3	34	19	18	82	14736	323	527	2	20	54	71	19	1016	15752
EXT 049	8060	6696	5	29	18	22	68	14898	382	602	1	14	71	74	14	1158	16056
EXT 055	7747	6210	7	32	19	21	68	14104	359	623	0	23	65	59	13	1142	15246
ALZ 062	7876	6385	7	33	22	18	74	14415	345	594	1	13	71	58	6	1088	15503
ROU 816	8011	6527	5	39	15	23	81	14701	362	587	1	13	73	57	11	1104	15805
ALZ 198	7282	5930	5	27	19	22	66	13351	314	575	1	12	56	71	10	1039	14390
ALZ 056	7860	6592	5	33	22	14	74	14600	280	522	2	15	40	57	18	934	15534
EXT 094	8300	6837	5	41	19	19	91	15312	338	563	2	10	49	56	10	1028	16340
EXT 077	8050	6641	7	39	23	17	74	14851	309	478	2	9	53	51	9	911	15762
EXT 050	8156	6683	7	27	21	20	75	14989	274	459	0	10	53	58	15	869	15858
EXT 220	10070	8558	26	36	21	19	94	18824	362	585	1	2	152	115	14	1231	20055
EXT 181	9981	8487	27	30	19	16	84	18644	373	681	3	4	167	107	22	1357	20001

Figure 3 Raw '*Variation overview*' in EVA for the 14 ADEOAD exome project. Both individuals **EXT 220** and **EXT 181** belong to batch #2 described in the section 2.2, all the others belong to batch #1.

Individual	Known variations								Unknown variations							
	Exon						Intron	Sub-total	Exon						Intron	Total
	Single variation				Indel		Splice		Single variation				Indel		Splice	
	Synonym	Missense	Stop loss	Nonsense	Fs	N Fs	+/- 2		Synonym	Missense	Stop loss	Nonsense	Fs	N Fs	+/- 2	
ALZ 049	0	0	0	0	0	0	0	**0**	0	286	0	8	25	0	3	322
ALZ 426	0	0	0	0	0	0	0	**0**	0	250	0	7	16	0	2	275
ROU 632	0	0	0	0	0	0	0	**0**	0	258	1	16	18	0	9	302
EXT 049	0	0	0	0	0	0	0	**0**	0	344	0	9	25	0	5	383
EXT 055	0	0	0	0	0	0	0	**0**	0	360	0	17	14	0	4	395
ALZ 062	0	0	0	0	0	0	0	**0**	0	328	1	9	20	0	2	360
ROU 816	0	0	0	0	0	0	0	**0**	0	336	0	9	22	0	5	372
ALZ 198	0	0	0	0	0	0	0	**0**	0	288	0	7	17	0	6	318
ALZ 056	0	0	0	0	0	0	0	**0**	0	237	1	10	3	0	4	255
EXT 094	0	0	0	0	0	0	0	**0**	0	266	0	5	4	0	2	277
EXT 077	0	0	0	0	0	0	0	**0**	0	208	1	6	8	0	4	227
EXT 050	0	0	0	0	0	0	0	**0**	0	210	0	10	9	0	5	234
EXT 220	0	0	0	0	0	0	0	**0**	0	394	0	1	31	0	7	433
EXT 181	0	0	0	0	0	0	0	**0**	0	435	1	0	29	0	12	477

Figure 4 Primary screened '*variation overview*' after filtering strategy functionality of EVA for the 14 ADEOAD exome project. Both individuals **EXT 220** and **EXT 181** belong to batch #2 described in the section 2.2, all the others belong to batch #1.

Details of gene: NOTCH1

Gene Symbol: NOTCH1 (NCBI Entrez Gene) (Pubmed) (NCBI CCDS) (NCBI OMIM)
Full Name: notch 1
Build: 37.1
Position: chr 9 : 139388895-139440238 (Ensembl Viewer)
Pathway: KEGG
Expression profil: GeneCard - UniGene

RefSeq: NM_017617.3 ; Length: 9309 bp
Protein: NP_060087.3 ; Length: 2555 aa

Uncovered areas: 1 (See details)
Total uncovered size ~120bp hence: 1.29% of the transcript (mRNA) size.
This section provide informations on the unsequenced regions.
Each uncaptured area correspond to a bait: a 120 bases long segment.

Other link: SNPper - Polyphen 2 - Mutation Taster

Affected individuals and detected variations

- Numeric data
- Graphic data

23 Known variations & 13 Unknown variations:

Individual	Known variations									Unknown variations								
	Exon						Intron	UTR		Exon						Intron	UTR	
	Single variation				Indel					Single variation				Indel				
	Synonym	Missense	Stop loss	Nonsense	Fs	N Fs		5'	3'	Synonym	Missense	Stop loss	Nonsense	Fs	N Fs		5'	3'
Ind #1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Ind #2	2	0	0	0	0	0	4	0	0	0	0	0	0	0	0	1	0	0
Ind #3	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Ind #4	3	1	0	0	0	0	4	0	0	0	0	0	0	0	0	1	0	0
Ind #5	4	0	0	0	0	0	5	0	0	0	1	0	0	0	0	0	0	0
Ind #6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Ind #7	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
Ind #8	3	0	0	0	0	0	8	0	0	0	1	0	0	0	0	1	0	0
Ind #9	5	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
Ind #10	5	0	0	0	0	0	8	0	0	0	0	0	0	0	0	2	0	0
Ind #11	3	0	0	0	0	0	6	0	0	0	1	0	0	0	0	0	0	0
Ind #12	1	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0
Ind #13	2	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
Ind #14	4	0	0	0	0	0	9	0	0	0	1	0	0	0	0	0	0	0
TOTAL Project	8	2	0	0	0	0	13	0	0	1	7	0	0	0	0	5	0	0

Figure 5 '*Gene details*' table in EVA. For a given gene (here NOTCH1) it is possible: (top) to get information about its chromosomic laction, links to useful public databases (Entrez, Pubmed, CCDS, OMIM), areas not captured during the pre sequencing protocol and links to interpretation tools (SNPper, Polyphen 2, Mutation Taster); (bottom): the categorized variations located in that gene for all the individuals of a project.

Download variations list: txt file

Variation	Genomic Position	cDNA & aa Position	Class	Category	C.w -> C.m	aa.w -> aa.m	Gene symbol	Individuals	Genotype status	Integragen (%) het | hom	1000g & EVS frequencies	Description
EVA.9.NOTCH1.E.129204	9:139410160	c.1678 p.560	Exon	Missense	GGG -> TGG	G -> W	NOTCH1	Ind #14	het1.D	0 | 0	1000g: - EVS: -	modify
EVA.9.NOTCH1.E.15473	9:139411779	c.1500 p.500	Exon	Synonym	CTG -> CTA	L -> L	NOTCH1	Ind #1	het1	0 | 0	1000g: - EVS: -	modify
EVA.9.NOTCH1.E.49963	9:139400290	c.4058 p.1353	Exon	Missense	GGC -> GAC	G -> D	NOTCH1	Ind #11	het1.D	0 | 0	1000g: - EVS: -	modify
EVA.9.NOTCH1.E.57419	9:139405111	c.2734 p.912	Exon	Missense	CGG -> TGG	R -> W	NOTCH1	Ind #12	het2	0 | 0	1000g: - EVS: 0.00038	modify
EVA.9.NOTCH1.E.57420	9:139413097	c.1045 p.349	Exon	Missense	ACC -> CCC	T -> P	NOTCH1	Ind #12	het1.D	8 | 0	1000g: 0.2374 EVS: 0.42493	modify
EVA.9.NOTCH1.E.75718	9:139412640	c.1204 p.402	Exon	Missense	TCG -> CCG	S -> P	NOTCH1	Ind #3	het1	0 | 0	1000g: - EVS: -	modify
EVA.9.NOTCH1.E.84173	9:139405218	c.2627 p.876	Exon	Missense	AGC -> ACC	S -> T	NOTCH1	Ind #5	het2	0 | 0	1000g: - EVS: -	modify
EVA.9.NOTCH1.E.95021	9:139409148	c.2021 p.674	Exon	Missense	ATG -> AAG	M -> K	NOTCH1	Ind #8	het1	0 | 0	1000g: - EVS: -	modify

Figure 6 *'Variation list'* table in EVA. For a given gene (here NOTCH1) it is possible to get the list of all the variations found in that gene for all the individuals of a project. For each variation the following information is listed: genomic position, gene position, functional class, category, wild codon vs modified codon, exome frequencies (heterozygous, homozygous), gene symbol, individual and genotype status.

dbSNP: yes Direct link	HapMap: 0	IntegraGen: het: 10 | hom: 15	1000 genomes: no	EVS exomes: G=6405/A=4115 0.39115

Position	Gene
Chromosome 9	NOTCH1
genomic: 139418260 cDNA: 313 ATG: 312 Protein: 104	strand: -
	NM_017617
	Exon 3

Variation		
reference	alternative allele	
T	C	Exon
AAT	AAC	Synonym
N	N	

Coverage & Quality	
EXT CAL 049 002	
A: 1	Total: 9
C: 0	
G: 4	Score: 13.90
T: 0	
Used: 5	
EXT PAR 094 001	
A: 0	Total: 11
C: 0	
G: 8	Score: 29.60
T: 0	
Used: 8	
EXT MAN 181 001	
A: 14	Total: 31
C: 0	
G: 7	Score: 47.70
T: 0	
Used: 21	

Figure 7 *'Variation details'* table in EVA. For a given variation, it is possible to get information such as: links to dbSNP, counts in HapMap, 1000 genomes and EVA, genomic position, gene information, coverage and quality per individual.

Finally, one of the strengths of EVA is the implementation of inheritance filters considering intersection or conversely differential exome strategies: (i) recurrence strategy for dominant or recessive independent familial cases (filters select the genes the most affected by remaining variations among a specified number of non related individuals.); (ii) filters for homozygous, heterozygous or composite cases in intra-familial studies (filters extract genes with remaining common variants among selected related individuals); (iii) and *de novo* strategy for sporadic cases (filters select genes with remaining variations found in a diseased child but not in the two healthy parents (sporadic case, trio-family).

For each strategy the displayed result is a list of potential candidate genes associated with the number of affected individuals (*'genes list'*). Again, it consists on an interactive table that could be readily explored. The user can get '*gene details*' (Figure 5) containing interactive links to other tables *'variation overview'* (Figure 3), *'variation list'* (Figure 6), and '*variation details*' (Figure 7). To ensure a rapid execution of EVA (Cf. 'Performance' subheadings) implemented in priority to focus on filtering strategies, we made the choice not to include variant effect prediction functionalities. Nevertheless, to facilitate the further prioritization of remained variations and genes, external functional and pathogenicity interpretation tools (SNPper [31], Polyphen 2 [32], MutationTaster [33]) are cross-linked as well as useful external international databases of genes, proteins, pathways, diseases and literature and genome browsers.

Case study: Alzheimer disease

After screening more than one hundred autosomal dominant early-onset Alzheimer disease (ADEOAD) families for known mutations (*Cf.* Methods section, 'Case study' subheadings) the molecular basis of this rare disorder still remained unexplained in several of them. Moreover, the lack of DNA for affected relatives precluding a linkage analysis in these cases, a full exome sequencing strategy was decided to identify new candidate gene(s) with unknown mutations. Exome sequencing, variation detection and annotation were performed by IntegraGen company (Cf. Methods section, 'Case study' subheadings) for fourteen ADEOAD unrelated index cases. The annotated variant files were subjected to ExomeDB to a remote loading using the online *Variation integration module* of EVA. Then, the intersection recurrence filtering strategy was applied with EVA. Here the main steps of our filtering procedure are summarized:

Firstly we displayed the full project data. Figure 1 corresponds to the raw *'variation overview'* of this exome project integrated in EVA and is obtained with the *Table browser* module. In this interactive table, variations are displayed by individuals and divided into two groups on the dbSNP131 referencing basis. '*Known*' means variations referenced in dbSNP, while '*unknown*' means variations not referenced in dbSNP. Within those groups the variations are rigourously and usefully displayed by two functional classes '*Exon*' and '*Intron*' *(*only two intronic base pairs before and after exons (*'+/-2'*)). Exonic variations are classified into six sub-categories, '*Synonym*', '*Missense*', *'Stop loss*' and '*Nonsense*' for SNV and Frameshift (*'Fs'*) and No Frameshift (*'Nfs'*) for indels. In total, 14,390 (batch #1) to 20,055 (batch #2) genetic variants were identified *per* exome according to the capture protocol (15,600 in average for batch #1 and 20,028 in average for batch #2). Among these, 6.6% in average are unknown variations (1028 in average for batch #1 and 1294 in average for batch #2).

Secondly, thanks to the *Filtering strategy* module we applied a stringent primary screening based on [common variations + molecular type of variants + heterozygous nature]. Figure 2 corresponds to the *'variation overview'* after this one: variations retained were previously '*unknown*' *(filtered against db SNP31)* but then filtered against HapMap exome projects, and against 42 IntegraGen exome projects from unrelated individuals with non-neurodegenerative diseases, the other filters parameters were '*non-synonym*' *SNV, 'frameshift coding'* indels, *'splice acceptor and donor site'* and *'heterozygous'*. Finally, the number of unknown variations by individual drastically decreases from 1028 in average for batch #1 and 1294 in average for batch #2, to 310 and 455 respectively. So, remaining unknown variations after this primary screening with EVA represented only 2% of total genetic variants identified *per* exome *versus* 6.6% in the raw data.

Thirdly, a secondary screening of the remaining variations based on the inheritance mode assumption of the disease was applied with an intersection recurrence procedure. Table 1 summarizes the number of genes harboring at least one of these variants classified according to their recurrence in the patient sample. The 14 patients did not have in common a single altered gene, indicating that, within this sample, the disease was genetically heterogeneous. Nevertheless, we observed that the number of candidate genes drastically decreased with the increasing number of concerned individuals. So, EVA enabled geneticists to focus further investigations on the affected genes shared by a minimum of 5 patients, representing a short list of less than 10 genes.

Finally, after wet investigations (Sanger resequencing verifications, family co-segregation analysis, genotyping of each variant in 1500 control individuals, RT-PCR expression analysis) combined with *in silico* analysis (predicted functional impact of each variation, comparison to the data set from the 1000 genomes project [10], and from Complete Genomics [29]), one gene (*SORL1*)

Table 1 Secondary screening obtained thanks to the 'recurrence' filtering Strategy functionality of EVA for the 14 ADEOAD exome project.

Number of individuals	Number of genes with remaining variations
14/14	0
13/14	0
12/14	0
11/14	0
10/14	0
09/14	0
08/14	1
07/14	3
06/14	3
05/14	7
04/14	31
03/14	112
02/14	542
01/14	2730

containing unknown mutations in 5/14 exomes (nonsense (n = 1) or missense (n = 4)) has become a new strong candidate gene for the ADEOAD [25].

Performance

To date, ExomeDB stores WGS projects (multiple unrelated cases, duo or trio cases) corresponding to a total of 23 individuals and contains also targeted resequencing projects corresponding to 5 genes for 25 individuals. As showed on Table 2, the size of ExomeDB is about 400 Mb, mainly due to the tables Variation (112.25 Mb) and Individual_Variation (271.09 Mb). Tests of EVA have been performed on the ADEOAD exome project (14 individuals) with one user logged in. The server is running Linux with four 3 GHz processors, 5 GB RAM and 150 GB HD. We use the "mysqli::prepare" mode which speeds up the request time once the first request has been treated. Table 3 shows request times in both cases. While it is clear that performances depend on the number of users logged in simultaneously and on the number of variants in the database (177,303 currently), EVA works with a reasonable time of execution compatible with the regular needs of a medical genetics laboratory.

Conclusions

EVA is developed to be a user-friendly, versatile, efficient-filtering and free assisting software for whole exome sequencing, providing a response to new needs at the expanding era of medical genomics investigated by these targeted next-generation sequencing technologies, for fundamental research, clinical diagnostics and personalized medicine [12,14-16,19,21,24]. Interfacing various now commonly adopted filtering criteria and strategies on whole exome data, EVA thereby makes non-programmer medical geneticists autonomous to pinpoint themselves among ~20,000 variations per individual exome, few candidate variations and genes related to a rare disease, depending of their specific assumptions and study design.

Table 2 Performance of EVA: Tables size of ExomeDB

Tables	Size (Mb)
Gene	7.03
IG_NoCouv	5.55
Individual	0.02
Individual_Variation	271.09
Project	0.02
Project_Individual	0.05
User	0.02
User_Project	0.06
Variation	112.25

EVA constitutes a platform for exome sequencing data storage and for drastic screening of clinical relevant genetics variations. Thanks to different modules (i) it integrates and stores annotated exome variation data as strictly confidential to the project owner, (ii) for the analytical process, it proposes to combine the main filters dealing with common human variations (various international external public data [10,11,26,27,29,30], molecular types and functional categories (synonym, missense, stop loss and nonsense for SNV, frameshift or not for indel; genic region i.e. UTR, CDS, splice site), homozygous or heterozygous nature of the allelic variant, inheritance modes and multiple samples considering intersection or conversely differential exome strategies (independent familial cases, intra-familial studies, sporadic cases), quality of the variations (iii) it offers quick searching or advanced browsing of annotated data and filtered results thanks to various interactive categorized or sortable tables and useful graphical visualizations (iv) finally it offers export files and cross-links to external relevant databases and softwares for further functional effects inspection [31-33] of the small subset of sorted candidate variations and genes.

EVA has been used to successfully identify a new candidate gene, *SORL1*, related to a rare form of Alzeihmer

Table 3 Performance of EVA: Running times of EVA modules

Request	Time (1st time)	Time (after)
Table Browse	15 s	3 s
Quick search	9 s	1 s
Filters loading	13 s	2 s
Filters execution	7 s	1 s

Tests of EVA have been performed on the ADEOAD exome project (14 individuals) with one user logged in.

Disease (ADEOAD), despite a genetics heterogeneity [25]. *SORL1* encodes the Sortilin-related receptor LR11/SorLA, a protein involved in the control of amyloid beta peptide production, the same pathway as previously known genes *APP*, and *presenilin 1 and 2*. In this case study, the primary screening with EVA (based on the mutation types and common human variations) reduced unknown variations to only 2% (330 on average) of total genetic variants identified per exome. The secondary screening implementing the intersection recurrence strategy led to a short list of genes (< 10) on which geneticists focused for further *in silico* and wet experiments and among which they discovered one. In 5 patients of the 14 independant index cases investigated, we found that the *SORL1* gene harbored unknown nonsense (n = 1) or missense (n = 4) mutations.

Performance tests showed that EVA run with a reasonable time of execution compatible with the regular needs of a medical genetics laboratory. For the case study it takes between 21 s (1st time) to 3 s (after) to load and execute the selected filters (server with four 3 GHz processors, 5 GB RAM and 150 GB HD, and with one user logged in) from the currently 400 MB size of ExomeDB.

The commonly assumption for WES mining is that causal variants related to a Mendelian disorder under investigation will not be present in public databases of genetic variations or other exome sequencing projects [1,13,14,17,19,22-24]. That is why, the more variation data available the more the filtering strategies in exome mining would be successful. To enhance its filtering performances, EVA confronts exome data currently to 6 external public data [10,11,26,27,29,30] and will be regularly updating as new large-scale variations data will be published.

Some polymorphisms of these ressources (dbSNP) are not associated with their allelic frequency and lack experimental annotation of their functional impact. So, projects like the SNP database of effects (SNPdbe) [34], storing computationally annotated functional impacts of non synonymous SNPs or the annotation of 1000 top human cancer genes frequently mutated [35] could be of interest for EVA improvement.

Alternative tools designed for the similar task as EVA have been recently published [35-38]. Varsifter [36] is a graphical Java program for desktop computers. It is designed to read exome-scale variation data in either a tab-delimited text file with header, or an uncompressed VCF file. It proposes numerous filtering options but doesn't propose graphic visualization nor statistical summaries of a WES project. SVA [36] is largely based on a genome browser to deal with WGE as well as WES and sifts small and large variants. While it proposes many manipulations of data, it is not clear if inherihance filtering are implemented. More, SVA is a JAVA program requiring a recommend hardware equipped with at least 48 GB of RAM and 1TB of free hard disk, which are substantial computational resources, in practice not very compatible to all individual laboratory. Finally, VAR-MD [37] is a family based tool. It analyzes WGS and WES variants exclusively in small human pedigrees with Mendelian inheritance excluding the scope of the differential exome analysis.

As perspectives are concerned, the input format for EVA for the *Variation integration* module, which is currently a proprietary format will be soon standardized in order to offer a wide use of this tool; we retained the Variant Call Format (VCF) format, generated by the 1000 Genomes Project. The *Variation integration* modulewill also allow the annotation of the raw variations by both Annovar [39] and the Variant Effect Predictor Ensembl API [40]. Regular updates are made concerning build version of the human genome, international variation catalogues and improvement of filtering functionalities as well as organization of results tables and graphics. Future developments include a graphical representation of a candidat gene with its variations and a more specific filtering strategy for somatic mutations.

List of abbreviations used

ADEOAD: Autosomal Dominant Early-Onset Alzheimer Disease; WGS: whole genome sequencing; WES: whole exome sequencing; SNV: single nucleotide variation.

Acknowledgements and funding

This work has been partially supported by Grant PHRC GMAJ, Centre national de référence Malades Alzheimer jeunes. The authors thank the LITIS for the host of EVA on one of its server.

This article has been published as part of *BMC Bioinformatics* Volume 13 Supplement 14, 2012: Selected articles from Research from the Eleventh International Workshop on Network Tools and Applications in Biology (NETTAB 2011). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/13/S14

Author details

[1]University of Rouen, INSERM U1079 Molecular genetics of cancer and neuropsychiatric diseases, 76183 Rouen cedex, France. [2]University of Rouen, LITIS EA 4108 Computer science, information processing and systems laboratory, 76821 Mont-Saint-Aignan cedex, France. [3]Institute of Research and Biomedical Innovation (IRIB), Haute-Normandie, France.

Authors' contributions

DC expressed the need of the tools, supervised the project and its design, oriented the filtering strategy functionality and realized the case study. SC realized the implementation of EVA, co-realized the case study, contributed to the manuscript. AL, ML, EPG and TL co-supervised the implementation of ExomeDB and the EVA interface and contributed to the manuscript. CC implemented the *Variation statistics* module and contributed to the manuscript. HD co-supervised the project and its design, oriented the categorization of variations and the different views of results on interface. HD and TL coordonned the collaboration between the two teams INSERM U1079 (clinicians) and TIBS-LITIS (bioinformaticians) and wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Published: 7 September 2012

References

1. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.
2. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends in Genetics* 2008, **24(3)**:133-141.
3. Mardis ER: **A decade's perspective on DNA sequencing technology.** *Nature* 2011, **470(7333)**:198-203.
4. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11(1)**:31-46.
5. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics.** *J Genet Genomics* 2011, **38**:95-109.
6. Voelkerding K, Dames S, Durtschi J: **Next-generation sequencing: from basic research to diagnostics.** *Clin Chem* 2009, **55**:641-58.
7. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, **26**:135-1145.
8. Koboldt DC, Ding L, Mardis ER, Wilson RK: **Challenges of sequencing human genomes.** *Brief Bioinform* 2010, **11(5)**:484-98.
9. Cooper GM, Shendure J: **Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.** *Nat Rev Genet* 2011, **12(9)**:628-40.
10. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-73.
11. Sherry S, Ward M, Kholodov M, Baker J, Phan L, Smigielski E, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308-311.
12. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N: **What can exome sequencing do for you?** *J Med Genet* 2011, **48(9)**:580-9.
13. Ng SB, Turner E, Robertson P, Flygare S, Bigham A, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler E, *et al*: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461**:272-276.
14. Ku C-S, Naidoo N, Pawitan Y: **Revisiting Mendelian disorders through exome sequencing.** *Hum Genet* 2011, **129**:351-370.
15. **Exome sequencing special issue.** In *Genome Biology* Garvey C, Cosgrove A, Attar N, Bilsborough G, Creavin T, Shendure J 2011, **12(9)**.
16. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nat Rev Genet* 2011, **12(11)**:745-55.
17. Singleton AB: **Exome sequencing: a transformative technology.** *Lancet Neurol* 2011, **10(10)**:942-6.
18. **Online Mendelian Inheritance in Man.** [http://omim.org/].
19. Stitziel NO, Kiezun A, Sunyaev S: **Computational and statistical approaches to analysing variants identified by exome sequencing.** *Genome Biology* 2011, **12(9)**:227-237.
20. Rovelet-Lecrux A, Legallic S, Wallon D, Flaman JM, Martinaud O, Bombois S, Rollin-Sillaire A, Michon A, Le Ber I, Pariente J, *et al*: **A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease.** *Eur J Hum Genet* 2011, doi: 10.1038/ejhg.2011.225.
21. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB: **Bioinformatics challenges for personalized medicine.** *Bioinformatics* 2011, **27(13)**:1741-8.
22. Van Oeveren J, Janssen A: **Mining SNPs from DNA sequence data. computational approaches to SNP discovery and analysis.** *Methods Mol Biol* 2009, **578**:73-91.
23. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data.** *Nat Rev Genet* 2011, **12(6)**:443-51.
24. Ku CS, Cooper DN, Polychronakos C, Naidoo N, Wu M, Soong R: **Exome sequencing: dual role as a discovery and diagnostic tool.** *Ann Neurol* 2012, **71(1)**:5-14.
25. Pottier C, Hannequin D, Coutant S, Rovelet-Lecrux A, Wallon D, Rousseau S, Legallic S, Paquet C, Bombois S, Pariente J, *et al*: **High frequency of potentially pathogenic *SORL1* mutations in autosomal dominant early-onset Alzheimer disease.** *Mol Psychiatry* 2012, **AOP, 3 April 2012, doi:10.1038/mp.2012.15.**
26. **IntegraGen company.** [http://www.integragen.fr].
27. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426**:789-796.
28. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, , **37 Database:** D32-6.
29. **Complete genomics.** [http://www.completegenomics.com].
30. **Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA.** [http://evs.gs.washington.edu/EVS/].
31. Riva A, Kohane IS: **SNPper: retrieval and analysis of human SNPs.** *Bioinformatics* 2002, **8**:1681-1685.
32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**:248-249.
33. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease causing potential of sequence alterations.** *Nat Methods* 2010, **7**:575-576.
34. Schaefer C, Meier A, Rost B, Bromberg Y: **SNPdbe: constructing an nsSNP functional impacts database.** *Bioinformatics* 2012, **28**:601-602.
35. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39**: e118.
36. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP, Zhu M, Singh A, Allen AS, Goldstein DB: **SVA: software for annotating and visualizing sequenced human genomes.** *Bioinformatics* 2011, **27**:1998-2000.
37. Teer JK, Green ED, Mullikin JC, Biesecker LG: **VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer.** *Bioinformatics* 2012, **28**:599-600.
38. Sincan M, Simeonov DR, Adams D, Markello TC, Pierson TM, Toro C, Gahl WA, Boerkoel C: **VAR-MD: A tool to analyze whole exome-genome variants in small human pedigrees with mendelian inheritance.** *Hum Mutat* 2012, **33**:593-598.
39. Wang K, Li M, Hakonarson H: **ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38(16)**:e164.
40. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: **Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.** *Bioinformatics* 2010, **26(16)**:2069-70.

doi:10.1186/1471-2105-13-S14-S9
Cite this article as: Coutant *et al.*: **EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics.** *BMC Bioinformatics* 2012 **13**(Suppl 14):S9.