

## **EVA: Exome Variation Analyzer, a convivial tool for filtering strategies\***

S. Coutant<sup>1,2</sup>, A. Lefebvre<sup>2</sup>, M. Léonard<sup>2</sup>, E. Prieur-Gaston<sup>2</sup>, D. Champion<sup>1</sup>, T. Lecroq<sup>2</sup>, H. Dauchel<sup>2</sup>

<sup>1</sup>University of Rouen, INSERM U614, France

<sup>2</sup>University of Rouen, LITIS EA 4108, 76821 Mont-Saint-Aignan cedex, France

### **Abstract**

Next-generation sequencing technologies enable to obtain genomic data at an unprecedented scale, however mining these data requires powerful bioinformatics tools. In this article we describe EVA, an innovative software dedicated to the filtering of variations detected by exome sequencing projects in order to assist geneticists to select genes responsible for a specific disease. This is simply done by using different kind of filters. Furthermore, making medical geneticists autonomous to filter variations, EVA constitutes an efficient assistant software for the era of clinical genomics and personalized medicine, at both levels of diagnosis and prognosis. EVA has already been used to successfully identify a candidate gene in a specific form of Alzheimer disease. EVA can be accessed on a web site through an authentication process.

### **1 Introduction**

Next-generation sequencing technologies are currently used to answer key biological questions at the scale of the entire genome and with an unprecedented depth. 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 High Throughput Sequencing (HTS) and require computer intensive bioinformatic analyses (Zhang J. *et al.*, 2011, Voelkerding K.V. *et al.*, 2009, Shendure J. and Ji H., 2008), although different.

Identification of genetic variations can be addressed by whole-genome sequencing of single individuals. But thanks to DNA enrichment techniques, targeted sequencing of coding regions (full exome sequencing) decreases the cost and improves the efficiency of HTS compared with what would require the entire human genome. The human exome, made of 180,000 exons for a size of 30Mbp, is 1.5% of the total human genome. Hence, not only targeted selection strategy reduces the cost but also accelerates the discovery of genetic variants that cause rare genetic diseases. In 2009, Sarah Ng *et al.* showed the proof of the concept, using a recurrence strategy, that identifying a gene responsible for a Mendelian disorder was possible using whole exome sequencing. Since then, more and more papers confirmed the success of this strategy (Ku C.-S. *et al.*, 2011).

Despite numerous bioinformatics solutions and software tools that have been developed (Zhang J. *et al.*, 2011) to efficiently manage terabytes of raw data from whole exome sequencing (alignment of reads to a reference sequence, variation calling, variation annotation and functional prediction of Single Nucleotide Variation (SNV) and indel (microinsertion or microdeletion)), the real challenge for clinical bioinformatics remains in developing softwares that help filtering exome data by distinguishing apparently novel genetic variants present randomly in any single human exome from rare variations that are really causal of a disease.

---

\* This work has been partially supported by Grant PHRC GMAJ, Centre national de référence Malades Alzheimer jeunes

In this aim, we have developed EVA (Exome Variation Analyzer), a simple, convivial and efficient software dedicated to medical projects investigated with exome sequencing. It consists of a database, called ExomeDB, and of a web interface. EVA's purpose is twofold: storing and managing exome sequencing data, and assisting geneticists in filtering strategies to limit the list of likely candidate gene underlying a genetic disease. To our knowledge (Ku C.-S. *et al.*, 2011), EVA is the first public filtering software, holding this crucial position in the analysis process of exome projects, between variations detection/annotation tools and variations pathogenicity prediction tools.

## 2 Methods

### 2.1 Implementation of EVA

ExomeDB was developed under MySQL (5.0). The main tables are Variation, Gene and Individual. In ExomeDB, the integrated data are the lists of variants (SNV, indel) associated with their annotations (position, type of mutation, affected gene...). Each new project is subject to a remote loading using an online integration module. This module accepts .txt files and .xls files, it can feed ExomeDB from CASAVA-like analysis pipeline outputs. The web interface was developed under PHP (5.3.2-1). It was designed with and for geneticists, thus friendly to use. The web address is public (<http://bioinfo.litislab.fr/EVA/>) but EVA's access is permitted through an authentication process using a login and a password given by the administrator. Each login/password is specific to a project, assuring the confidentiality of the exome data. For the three filtering strategies (*Cf.* Result section 3.2), the combination of selected parameters by the geneticist, is transformed into an SQL query and sent to the ExomeDB database. Then, EVA's interface displays the remaining variations.

### 2.2 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 (CNVs) of *APP* gene and genes involved in Amyloid beta (A $\beta$ ) peptide processing or signaling. IntegraGen society (Genopole® Evry, France) 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 150-200bp, 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 (CASAVA 1.7). For the genetics variants detection, the 76bp sequencing reads were aligned to the NCBI human reference genome (NCBI36.3 ( $n=12$ ) or NCBI 37 ( $n=2$ )), using ELANDv2. Means coverage were of 65-fold ( $n=12$ ) and 80-fold ( $n=2$ ). Only high quality variations having a QPhred threshold  $>10$  were conserved (86 % of the targeted bases). The annotation of the detected variations includes the report of the corresponding ID gene (NCBI RefSeq), the references of previously known variations (dbSNP131 (Sherry S.T. *et al.*, 2001), HapMap (The International HapMap Consortium, 2003)), the genome position, the description and categorization of

variations (*Cf.* Results section 3.1).

### **3 Results**

#### **3.1 Overview: ExomeDB and EVA web interface**

EVA consists of a database called ExomeDB and of a web interface. ExomeDB is a relational database which main tables store exome sequencing data with the information about the variations, the genes and the individuals. Not only the variations are stored in ExomeDB but also their annotations (genome position, corresponding gene, mutation type, genotype status, quality score). EVA can manage different projects. Each new project is subject to a remote loading using an online integration module. Each project can manage several patients (or individuals). Each individual contains variations. Due to the molecular process of the exome capture kit, most variations occur in exons but some detected variations also occur in splice sites. Variations in exons can be SNV or indels. Single variations have been categorized into four functional classes: synonymous, miss sense, stop loss and non sense. Indels have been classified into two categories: frameshift or non frameshift. The set of all the variations is divided in known and unknown variations according to the information in dbSNP.

The web interface proposes three query modalities for exome mining: (i) the Browser functionality to explore data by project, individual, gene or variation, (ii) the Search functionality for a direct and quick access to a specific gene or variation for a given project, and finally, (iii) the Filtering Strategy functionality, which is the major element in exome mining to limit potential candidate genes. Query results can be browsed by five elements types: '*variations overview*', '*genes list*', '*variations list*', '*gene details*', and '*variation details*'.

#### **3.2 Filtering Strategy functionality**

In three main steps, EVA proposes to geneticists to obtain a limited list of likely candidate genes. The first step consists in selecting the project to analyse. The second step consists in applying filters (primary screening). Filters can be combined at will by geneticists to address different kind of questions. Users can choose to keep or disregard variations that are: known (in dbSNP) or found in other exome projects (EVA projects, HapMap projects, IntegraGen society projects); synonymous; indel (frameshift and/or not frameshift); affected splice sites; homozygous or heterozygous genotype; low quality. The combination is transformed into a SQL query and sent to the ExomeDB database. Then, the tool displays the remaining variations.

The third step consists in a secondary screening of the remaining variations. In this aim, the EVA's filtering functionality is designed according to three strategies depending on the genetic disorder case: recurrence, familial and *de novo*. The recurrence strategy can be applied on dominant or recessive pathologies: EVA's filters select genes the most affected by remaining variations among a specified number of non related individuals. For the familial strategy, EVA's filters extract genes with remaining common variants among selected related individuals. Conversely, for the *de novo* strategy, EVA's filters select genes with remaining variations found in a diseased child but not in the two healthy parents (sporadic case, trio-family exome sequencing). Thus, one crucial step before considering to use EVA, is the design of the exome sequencing project. Are the affected samples coming from several families? Or, on the contrary, are they related to each other? Are they all affected or are we searching for a *de novo* mutation? Is the pathology dominant or recessive? Depending on the design of the project the user will not select the same filtering strategy in EVA.

For each strategy the displayed result is a list of potential candidate genes associated with the number of affected individuals. Then, the user can get '*gene details*' containing useful external international database links (such as NCBI Gene, NCBI OMIM, KEGG, GeneCard, UniGene, Ensembl browser...), external functional and pathogenicity interpretation tools (SNPper (Riva A., Kohane I.S., 2002), Polyphen 2 (Adzhubei I.A. *et al.*, 2010), MutationTaster (Schwarz, J.M. *et al.*, 2010), containing information about not captured regions during the pre-sequencing protocol, and containing '*variations overview*', '*variations list*', and '*variation details*'.

### 3.3 Case study: Alzheimer disease

The recurrence strategy has been applied with EVA in fourteen autosomal dominant early-onset Alzheimer disease (ADEOAD) unrelated index cases without mutation on known genes. Exome sequencing, variations detection and annotation were performed by IntegraGen society (*Cf.* Methods section). Table 1 corresponds to the raw '*variations overview*' of this exome project integrated in EVA and is obtained with the Browser functionality. 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 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 indel. 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).

**EVA**  
Exome Variations Analyzer

Basic filters

Variations Genes

Choose a project to display the details of its variations grouped by category and localisation. Alzheimer1

39791 Known variations & 16774 Unknown variations:

Individual	Known variations							Unknown variations							Total		
	Exon				Intron			Exon				Intron					
	Single variation				Indel	Splice	Sub-total	Single variation				Indel	Splice	Sub-total			
	Synonym	Missense	Stop loss	Nonsense	Fs	NFs	+/- 2		Synonym	Missense	Stop loss	Nonsense	Fs	NFs	+/- 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

Table 1: Raw '*Variations 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.

Table 2 corresponds to the '*variations overview*' obtained thanks to the Filtering Strategy functionality of EVA, after applying a stringent primary screening: variations retained were previously '*unknown*' (*filtered against db SNP31*) but then filtered against HapMap exome projects, and against 42 IntegraGen exomes 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 the batch #1 and 1294 in average for the batch #2, to 310 and 455 respectively. So, remaining unknown variations after this primary screening with EVA represent only 2% of total genetic variants identified *per* exome *versus* 6.6% in the raw data.

**EVA**  
Exome Variations Analyzer

Basics filters

Variations Genes

Choose a project to display the details of its variations grouped by category and localisation. Alzheimer1

Show

0 Known variations & 4260 Unknown variations:

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

Table 2: Primary screened 'variations overview' obtained thanks to the 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.

A secondary screening of the remaining variations corresponding to the recurrence Filtering Strategy functionality of EVA was then applied. Genes harboring at least one of these variants were grouped according to their recurrence in the patient sample. The 14 patients did not have a single altered gene in common, 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 (from more than 2500 genes to only one). So, EVA enabled geneticists to focus their further investigations on the affected genes shared by a minimum of 5 patients, representing a short list of less than 10 genes (publication submitted).

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 (1000genomes.org), and from Complete Genomics (completegenomics.com)), one gene containing unknown mutations in 7/14 exomes has become a strong likely candidate gene for the ADEOAD (publication submitted).

## 4 Discussion

The list of variants from the exome sequences of individuals must be screened to narrow the list of likely candidate gene underlying a rare disorder. Hence, with an average of 17,000 variations *per* exome, the real challenge of the exome sequencing for medical genomics lies not only in efficient bioinformatics filtering strategies to exclude common variations, but also lies in their implementation into a convivial tool enabling clinicians to be autonomous to test different filtering approaches. In this aim, we have developed EVA in collaboration with and for medical geneticists.

Firstly, EVA proposes a rigorous storage of exome projects (ExomeDB) accessible *via* an online integration module. Secondly, EVA offers, through a convivial web interface, not only classical functionalities of Browsing or Quick search, but, more innovatively, offers an efficient assistance to ensure a drastic screening of genetics variations. Thanks to its Filtering Strategy functionality, EVA has been used to successfully identify a candidate gene in a rare form of Alzheimer Disease, despite a genetics heterogeneity. In this case study, the primary screening with EVA (based on the mutation types and the comparison to data from dbSNP, HapMap, IntegraGene's independent exome projects) reduced unknown variations to only 2% (330 on average) of total genetic variants identified *per* exome. The secondary screening implementing the 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 strong likely candidate gene for the ADEOAD.

Among the *in silico* analysis, there was a supplementary comparison against the 1000 genomes and Complete Genomics variations data. This enabled to filter few more variations and to reinforce that the more variations data available, the more the filtering strategies in exome mining will be successful. This case study encourages us to implement in EVA the 1000 genomes and Complete Genomics variations data to enhance its filtering performances.

## References

- Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., Kondrashov A.S., Sunyaev S.R. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249 (2010).
- Ku C.-S., Naidoo N., Pawitan Y. Revisiting Mendelian disorders through exome sequencing, *Hum. Genet.* **129**, 351-370 (2011).
- Ng S.B., Buckingham K., Lee C., Bigham A., Tabor H., Dent K., Huff C., Shannon P., Jabs E., Nickerson D., Shendure J., Bamshad M. Exome sequencing identifies the cause of a Mendelian disorder, *Nature Genetics* **42**, 30-35 (2010).
- Ng S.B., Turner E., Robertson P., Flygare S., Bigham A., Lee C., Shaffer T., Wong M., Bhattacharjee A., Eichler E., Bamshad M., Nickerson D., Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes, *Nature* **461**, 272-276 (2009).
- Riva A., Kohane I.S. SNPper: retrieval and analysis of human SNPs. *Bioinformatics* **8**, 1681-1685 (2002)
- Schwarz, J.M., Rodelsperger, C., Schuelke, M., Seelow, D. MutationTaster evaluates disease causing potential of sequence alterations. *Nat. Methods* **7**, 575-576 (2010).
- Shendure J., Ji H. Next-generation DNA sequencing, *Nature Biotechnology* **26**, 1135-1145 (2008).
- Sherry S., Ward M., Kholodov M., Baker J., Phan L., Smigielski E., Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).
- The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796 (2003).

Voelkerding K., Dames S., Durtschi J. Next-generation sequencing: from basic research to diagnostics. *Clin Chem.* **55**, 641-58 (2009).

Zhang J., Chiodini R., Badr A., Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics* **38**, 95-109 (2011).