Taylor & Francis
Taylor & Francis Group

🔓 OPEN ACCESS | Check for updates

RESEARCH ARTICLE

ALSgeneScanner: a pipeline for the analysis and interpretation of DNA sequencing data of ALS patients

ALFREDO IACOANGELI[1,2] (iD), AHMAD AL KHLEIFAT[2] (iD), WILLIAM SPROVIERO[2], ALEKSEY SHATUNOV[2], ASHLEY R. JONES[2], SARAH OPIE-MARTIN[2], ERSILIA NASELLI[2], SIMON D. TOPP[3], ISABELLA FOGH[2,4], ANGELA HODGES[2], RICHARD J. DOBSON[1,5,6], STEPHEN J. NEWHOUSE[1,5,6] AND AMMAR AL-CHALABI[3,7] (iD)

[1]*Department of Biostatistics and Health Informatics, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, UK;* [2]*Department of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, London, UK;* [3]*UK Dementia Research Institute, King's College London, London, UK;* [4]*Department of Neurology and Laboratory of Neuroscience, IRCCS Istituto Auxologico, Milan, Italy;* [5]*Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, London, UK;* [6]*National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Unit at South London and Maudsley NHS Foundation Trust, King's College London, London, UK;* [7]*Department of Neurology, King's College Hospital, London, UK*

Abstract
Amyotrophic lateral sclerosis (ALS, MND) is a neurodegenerative disease of upper and lower motor neurons resulting in death from neuromuscular respiratory failure, typically within two years of first symptoms. Genetic factors are an important cause of ALS, with variants in more than 25 genes having strong evidence, and weaker evidence available for variants in more than 120 genes. With the increasing availability of next-generation sequencing data, non-specialists, including health care professionals and patients, are obtaining their genomic information without a corresponding ability to analyze and interpret it. Furthermore, the relevance of novel or existing variants in ALS genes is not always apparent. Here we present ALSgeneScanner, a tool that is easy to install and use, able to provide an automatic, detailed, annotated report, on a list of ALS genes from whole-genome sequencing (WGS) data in a few hours and whole exome sequence data in about 1 h on a readily available mid-range computer. This will be of value to non-specialists and aid in the interpretation of the relevance of novel and existing variants identified in DNA sequencing data.

KEYWORDS: *ALS, genomics, NGS, bioinformatics, genome analysis*

Introduction

Amyotrophic lateral sclerosis (ALS) is a progressive neurodegenerative disease, typically leading to death within 2 or 3 years of first symptoms. Many gene variants have been identified that drive the degeneration of motor neurons in ALS, increase susceptibility to the disease or influence the rate of progression (1). The ALSoD webserver (2) lists more than 120 genes and loci which have been associated with ALS, although only a subset of these have been convincingly shown to be

ALS-associated (3), demonstrating one of the challenges of dealing with genetic data interpretation of findings. Next-generation sequencing provides the ability to sequence extended genomic regions or a whole-genome relatively cheaply and rapidly, making it a powerful technique to uncover the genetic architecture of ALS (4). However, there remain significant challenges, including interpreting and prioritizing the found variants (5) and setting up the appropriate analysis pipeline to cover the necessary spectrum of genetic factors, which includes expansions, repeats, insertions/deletions

(indels), structural variants and point mutations. For those outside the immediate field of ALS genetics, a group that includes researchers, hospital staff, general practitioners, and increasingly, patients who have paid to have their genome sequenced privately, the interpretation of findings is particularly challenging.

The problem is exemplified by records of *SOD1* gene variants in ALS. More than 180 ALS-associated variants have been reported in *SOD1* (2). In most cases, the basis of these variants being attributed to ALS is simply that they are rare and found in *SOD1*. Neither of these is sufficient for such a statement to be made. The p.D91A variant, for example, reaches polymorphic frequency in parts of Scandinavia, and yet has been convincingly shown to be causative of ALS. A few variants have been modeled in transgenic mice, shown to segregate with disease or have other strong evidence to support their involvement (6–10) but most do not have such support. Rare variation can be expected to occur by chance, and its existence in a gene is not evidence of relationship to a disease, making interpretation of sequencing findings difficult. Although various tools are available to predict the pathogenicity of a protein-changing variant, they do not always agree, further compounding the problem.

We, therefore, developed ALSgeneScanner, an ALS-specific framework for the automated analysis and interpretation of DNA sequencing data. The

Figure 1. ALSgeneScanner pipeline main steps. From sequencing data in fastq format to the report generation of the results.

tool is targeted for use by a wide audience which includes people with knowledge outside genetics.

Materials and methods

ALSgeneScanner is part of the DNAscan suite (11). Figure 1 shows the pipeline main steps. The pipeline accepts sequencing data in fastq and bam formats as well as DNA variants in vcf format. In the latter case, only the annotation, variant prioritization, and report generation steps are performed. A detailed description and benchmark of its analysis components have been previously published (11). ALSgeneScanner uses, among others, Hisat2 (12) and BWA-mem (13) to align the sequencing data to a reference genome, Freebayes (14) and GATK Haplotype Caller (15) to call SNVs and small indels, Manta (16) and ExpansionHunter (17) for the detection of large structural variants (bigger than 50 bps) and repeat expansions.

Software

ALSgeneScanner is available on GitHub (18) (https://github.com/KHP-Informatics/ALSgeneScanner). The repository provides detailed instructions for tool usage and installation. A bash script for an automated installation of the required dependencies is also provided as well as Docker (19) and Singularity (20) images for a fast and reliable deployment. A Google spreadsheet with the complete list of genes and loci used by ALSgeneScanner is publicly available to visualize and comment (see GitHub repository).

Gene and loci prioritization

ALSgeneScanner groups genes and loci associated with ALS into three classes: i) genes and loci identified by our manual scientific literature review to be associated with the disease or an influence on the phenotype in ALS (see Table 1), ii) genes in which variants of clinical significance have been reported on ClinVar (51) and for which no contradictory interpretation is present, and iii) genes for which any association evidence has been submitted to ALSoD (2). The union of these three sets of genes (available on GitHub) is used to restrict the genome analysis. However, ALSgeneScanner allows the user to use a custom list of genes.

Manual scientific literature review

The literature review was performed using several databases, including PubMed, MEDLINE, and EMBASE, to identify all articles reporting the contribution of genetic variations to the development of the disease or the modification of the phenotype in ALS from 1993, when *SOD1* was the first gene discovered to cause ALS (41), until the date of the last manuscript revision. Review articles were discarded. The resulting list of
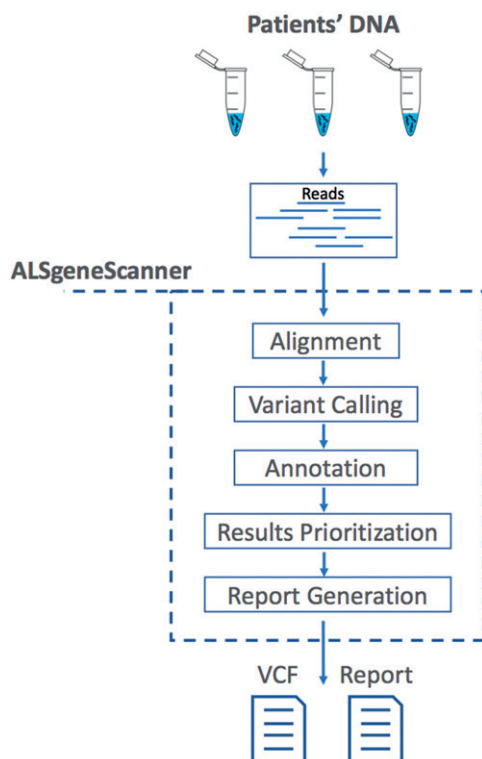
Table 1. List of *ALS* genes identified by literature review.

Gene	Associated ND	Phenotype influence	Key reference
ANG	ALS/PD		(21)
ANXA11	ALS		(22)
APOE		Longer survival	(23)
ATXN2	ALS		(24)
CAMTA1		Shorter survival	(25)
C21orf2	ALS		(26)
C9orf72	FTD/ALS	Primarily bulbar onset	(10)
CCNF	FTD/ALS		(27)
CHCHD10	FTD/ALS		(28)
DAO	ALS		(29)
DCTN1	ALS		(30)
EPHA4		Longer survival	(31)
FIG4	ALS		(32)
FUS	FTD/ALS	Early age of onset and shorter survival	(9)
HNRNPA1	ALS		(33)
IDE		Shorter survival	(25)
KIF5A	ALS		(34)
MATR3	ALS		(35)
MOBP	ALS		(26)
NEK1	ALS		(36)
NIPA1	ALS		(37)
OPTN	ALS		(38)
PFN1	ALS	Limb-onset	(39)
PGRN	FTD/ALS		(40)
SARM1	ALS		(26)
SCFD1	ALS		(26)
SOD1	ALS	Limb-onset, early age of onset and shorter survival	(41)
SPG11	ALS		(38)
SQSTM1	FTD/ALS		(42)
SETX	ALS		(43)
TAF15	ALS		(44)
TARDBP	FTD/ALS		(45)
TBK1	ALS		(26)
TUBA4A	FTD/ALS		(46)
UBQLN2	FTD/ALS		(47)
UNC13A	ALS	Shorter survival	(26)
VAPB	ALS		(48)
VCP	FTD/ALS		(49)
8p23.2	ALS		(26)
1p34-rs3011225		Late age of onset	(50)

genes and loci was filtered by keeping only the ones for which the link with ALS was shown in at least two independent studies (e.g. *SOD1*, *FUS*, *C9orf72*, etc.) or cohorts (e.g. *KIF5A*), or whose variants passed the genome-wide significance threshold in GWAS studies (e.g. CAMTA1). In the latter case, if a replication study was not yet available, to avoid spurious associations, we also required that these variants were surrounded by proxies in tight linkage disequilibrium (LD) that clearly indicated the presence of an associated haplotype block. The resulting list of ALS genes and loci is kept up to date by reviewing new articles as they become available. This list, as well as the complete list of reviewed articles, is available on GitHub (https://github.com/KHP-Informatics/ALSgeneScanner).

Variant prioritization

The pathogenicity prediction programs, SIFT (52), PolyPhen-2 HDIV and PolyPhen-2 HVAR (53), LRT (54), MutationTaster (55), MutationAssessor (56), Fathmm (57), PROVEAN (58), Fathmm-MKL coding (59), MetaSVM (60), and CADD (61) are used to prioritize variants. A variant is scored X where X is equal to the number of tools which predict it to be pathogenic. A higher priority is given to variants which are reported to be "likely pathogenic" or "pathogenic" on ClinVar. For each tool, we used the authors' recommendations for the categorical interpretation of the variants. For each variant, the score ranges between 0 and 11 according to the number of computational tools (11 in total) that predict it to be pathogenic. In order to leave the user free to customize the prioritization criteria, both our cumulative score and the categorical variant interpretations from the 11 tools are included in the final results.

Whole-genome sequencing

The whole-genome sequencing (WGS) sample used to assess the computational performance of ALSgeneScanner was sequenced as part of Project MinE (62). Venous blood was drawn from patients and controls and genomic DNA was isolated using standard methods. DNA integrity was assessed using gel electrophoresis. Samples were sequenced using Illumina's FastTrack services (San Diego, CA) on the Illumina Hiseq 2000 platform. Sequencing was 100 bp paired-end performed using PCR-free library preparation, and yielded ~40x coverage across each sample.

Whole-exome sequencing

To assess the computational performance of ALSgeneScanner we also used the Illumina Genome Analyzer II whole exome sequencing of NA12878 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20101201_cg_NA12878/NA12878.ga2.exome.maq.raw.bam).

VariBench and ClinVar datasets

To assess our variant prioritization approach, we used a set of non-synonymous variants from the VariBench dataset (63) for which the effect is known and all ALS-associated non-synonymous variants stored in ClinVar (71 benign and 121 pathogenic). The VariBench variants are not ALS genes specifically, but because they are all annotated depending on whether or not they are deleterious, the general principles of the method could be tested. The dataset includes VariBench protein tolerance dataset 1 (http://structure.bmc.lu.se/VariBench/tolerance_dataset1.php) comprising 23,683 human non-synonymous coding neutral SNPs and 19,335 pathogenic

missense mutations (64). None of the tools used in our pathogenicity score were trained on the VariBench dataset. However, it is possible that some VariBench variants were present in the training datasets. In order to minimize the overlap between training and evaluation sets, we derived a subset of variants (VariBenchFiltered) from the VariBench dataset by filtering out its overlap with HumVar (53), the CADD training dataset (61) and ExoVar (65), which are commonly used to train the tools (66). The resulting dataset comprising 5051 pathogenic and 14,077 neutral variants, was balanced by randomly subsampling 5051 neutral variants.

Evaluation of performance

Receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC) statistic were calculated using easyROC (67). Accuracy, precision, and sensitivity are defined as in equation below where T_p is true positives, F_p false positives, F_n false negatives, and T_n true negatives.

$$\text{Precision} = \frac{T_p}{T_p + F_p};\ \text{Sensitivity} = \frac{T_p}{T_p + F_n};$$
$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_n + F_p}$$

Hardware

All tests were performed on a single, mid-range, commercial computer with 16GB RAM and an Intel i7-670 processor.

Output

Resulting variants are reported in a tab-delimited format to favor practical use of worksheet software such as iWork Number, Microsoft Excel, or Google Spreadsheets.

Results

Manual literature review identified 486 articles describing a total of 127 genes and loci associated with ALS (the article and gene lists are available on GitHub), from which 38 genes and 2 loci (Table 1) with strong and reproducible supporting evidence of association with ALS or influence on phenotype were included. ClinVar reported SNVs and small indels in 44 genes and 4 structural variants ranging in size from 3 to 50 million base pairs. ALSoD reported variants in 126 genes and loci. The union of these sets of genes contained 149 genes and loci. The Venn diagram in Figure 2 shows the overlap between the three sets.

Using a midrange commercial computer (4 CPUs and 16 gigabytes of RAM) (Figure 3) ALSgeneScanner could analyze 40x WGS data of one individual in about 7 h using 12.8GB of RAM, and whole-exome sequencing data in 1 h and 20 min using 8.5GB of RAM.

We tested the computational score that the tool used to rank variants on three datasets. The VariBench dataset, the VariBenchFiltered dataset,
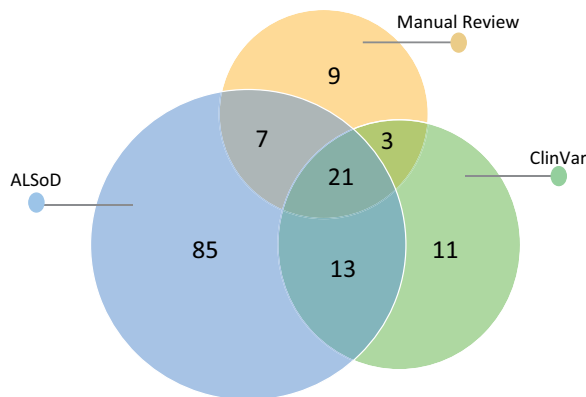

Figure 2. Venn diagram of the ALS related genes that we selected in our literature review, found in the ALSoD webserver and in the ClinVar database.
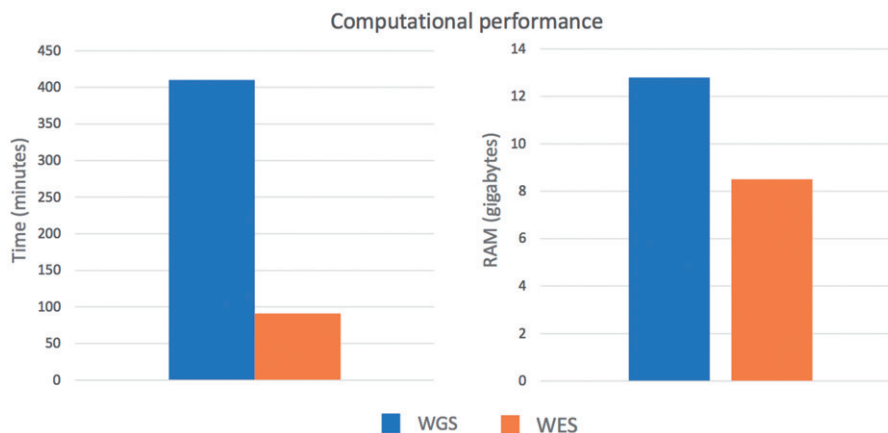

Figure 3. Computational performance of the pipeline to process whole-genome sequencing and whole exome sequencing data from fastq file to the generation of the final result report.

and on the ALS associated ClinVar entries. Figure 4 shows the results on the three datasets and Table 2 precision, sensitivity and accuracy of the method in function of the chosen threshold. The ROC curve for the VariBench and VariBenchFiltered dataset (Figure 4, AUC = 0.90 and 0.81) suggests a cutoff equal to 9 which maximizes the accuracy (0.83 and 0.73) however, a lower or higher cutoff can be chosen to reach a better precision or sensitivity according to the user's needs. For example, for diagnostics a higher sensitivity is generally required and a cutoff equal to 5 would increase the sensitivity to 0.90 (Table 2). The ROC curve for the ClinVar variants suggests a cutoff equal to 7. The AUC for such variants is 0.82 (Figure 4) and the accuracy for the ideal cutoff is 0.75 (Table 2). The better performance on the VariBench dataset can be partially explained by the fact that some of its variants were used for training the tools used by our cumulative
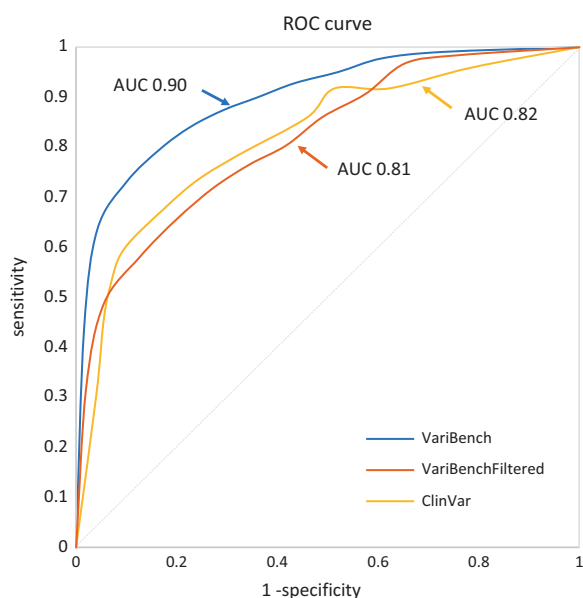
score. However, other factors can contribute to the performance drop on the VariBenchFiltered and ClinVar ALS datasets: first the uncertainty in the ClinVar entries. ClinVar provides the community with an infrastructure to allow researchers to store their clinical observations, but the quality checks are very limited and the only filter we have adopted in this study to select the variants was the absence of contradictory entries. A similar effect is also likely for the VariBenchFiltered dataset. Indeed, filtering out all variants present in the other datasets might increase the proportion of misclassified variants. Also, the different definitions of pathogenicity and neutrality used in the different benchmark/training datasets could contribute to this effect (66). The second is the difficulty that available computational tools have in assessing the effect of ALS related variants (3,36), in part because the mechanism of ALS is unknown, and in part because at least some of the variants result in a toxic gain of function that is difficult to understand or model.

Correlation analysis was performed to investigate the correlation between the 11 tools used by our score, using the categorical results of each individual tool on the VariBenchFiltered dataset. Supplementary Table 1 shows the results of this analysis. The average correlation was 45% and the standard deviation 14%. Only PolyPhen-2 HDIV and PolyPhen-2 HVAR showed a strong correlation (83%). PolyPhen-2 HDIV differs from PolyPhen-2 HVAR in the training dataset which only included Mendelian disease variants. These tools can provide the user with complementary useful information.

Figure 4. ROC curve of the performance of ALSgeneScanner on the three datasets.

Discussion

We have developed ALSgeneScanner, a fast, efficient, and complete pipeline for the analysis and interpretation of DNA sequencing data in ALS, targeted for use by a wide audience including non-geneticists. The method is able to distinguish

Table 2. ALSgeneScanner variant prioritization performance.

Score	VariBench			VariBenchFiltered			ClinVar ALS variants		
	Precision	Sensitivity	Accuracy	Precision	Sensitivity	Accuracy	Precision	Sensitivity	Accuracy
0	0.430	1	0.430	0.500	1	0.500	0.612	1	0.613
1	0.507	0.990	0.581	0.560	0.984	0.606	0.659	0.957	0.670
2	0.549	0.978	0.644	0.592	0.968	0.651	0.707	0.949	0.728
3	0.580	0.950	0.682	0.609	0.907	0.663	0.745	0.949	0.770
4	0.618	0.928	0.721	0.634	0.860	0.682	0.754	0.889	0.754
5	0.653	0.900	0.751	0.657	0.804	0.693	0.798	0.812	0.759
6	0.692	0.875	0.779	0.687	0.766	0.708	0.832	0.761	0.759
7	0.736	0.841	0.801	0.721	0.719	0.720	0.863	0.701	0.749
8	0.783	0.796	0.817	0.762	0.657	0.726	0.911	0.615	0.728
9	0.845	0.731	0.827	0.817	0.582	0.726	0.926	0.538	0.691
10	0.919	0.635	0.819	0.895	0.486	0.714	0.931	0.462	0.649
11	0.954	0.436	0.748	0.937	0.315	0.647	0.925	0.316	0.565

pathogenic from nonpathogenic variants with high accuracy and reports findings in a simple format, able to be exported for further analysis. With the decreasing costs and increasing availability of next-generation sequencing, health care professionals and motivated patients are progressively more likely to have WGS data available, without the tools to interpret findings. An automated system to provide a meaningful report, therefore, has a potentially important part to play in giving patients ownership of their data and arming them with the knowledge to understand it, but this should always be interpreted with the assistance of a specialized genetic counselor.

Omictools (68), a web database where available bioinformatics tools are listed and reviewed, lists over 7000 such tools for next-generation sequencing, including more than 100 pipelines; given the great interest in this field, new tools are frequently released. As a result, designing a bioinformatics pipeline for the analysis of next-generation sequencing data, keeping the system simple to use on a standard computer and translating the output into a format that is easily understood, is not trivial, and requires specialized expertise. The computational effort and the informatics skills required to use typical pipelines can dramatically limit the use of next-generation sequencing data. Adequate high-performance computing facilities and staff specialized in informatics are not always present in medical and research centers. Furthermore, the use of cloud computing facilities, which could theoretically provide unlimited resources, is not always possible due to privacy and ownership issues, cost and the expertise required for their use. To this end, ALSgeneScanner is computationally light as it can run on a midrange commercial computer. Performing the same analyses with other widely used pipelines, e.g. SpeedSeq (69) and GATK Best Practice Workflow (15), would require high-performance facilities (HPC) and about 3–10 times more computational resources than for ALSgeneScanner (11). It is easy to use since it performs sophisticated analyses using only a few command lines (see Figure 5) and is comprehensive, including the necessary analyses to identify all known ALS associated genetic factors. Finally, a tab-delimited output, in which the analysis results are enriched with information from several widely used databases such as ClinVar, ALSoD, our manual literature review, pathogenicity scores and the graphical visualization utilities (see Supplementary Material) integrated in the pipeline as part of DNAscan (11), favor an easily accessible interpretation of the results. No other currently available pipeline provides the user with such a comprehensive end-to-end analysis framework.

Our table of sensitivity, specificity, and accuracy (Table 2) means that the appropriate cutoff can be used to interrogate data, depending on whether the aim is the exclusion of potentially harmful variants, or the detection of definitely harmful variants.

ALSgeneScanner puts a powerful bioinformatics tool, able to exploit the potentialities of next-generation sequencing data in the hands of patients, ALS researchers, and clinicians.

Acknowledgments

Fast deployment and basic usage instructions for Linux based systems

Deployment: After downloading DNAscan* (https://github.com/KHP-Informatics/DNAscan) and Annovar (http://www.openbioinformatics.org/annovar/annovar_download_form.php), you can deploy ALSgeneScanner by running one simple script that you can find in /path/to/DNAscan/scripts:

```
$ bash /path/to/install_ALSgeneScanner.sh /path/to/DNAscan /path/to/annovar
```

This script will download and install all dependencies as well as create the necessary index and reference files.

Usage: To run the whole ALSgeneScanner pipeline on paired-end reads in fastq format stored in data1.fq.gz and data2.fq.gz, you can run the following command running the main DNAscan scrip in /path/to/DNAscan/scripts :

```
$ python3 /path/to/DNAscan/DNAscan.py -alsgenescanner -in data1.fq.gz -in2 data2.fq.gz \
-out /path/to/outdir
```

* ALSgeneScanner is part of the DNAscan analysis framework

Figure 5. Deployment and usage instructions.

Declaration of interest

Funding

ORCID

Alfredo Iacoangeli http://orcid.org/0000-0002-5280-5017

Ahmad Al Khleifat http://orcid.org/0000-0002-7406-9831

Ammar Al-Chalabi http://orcid.org/0000-0002-4924-7712

References

1. Al-Chalabi A, van den Berg LH, Veldink J. Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. Nat Rev Neurol. 2017;13:96–104.
2. Abel O, Powell JF, Andersen PM, Al-Chalabi A. ALSoD: a user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. Hum Mutat. 2012;33:1345–51.
3. Brown RH, Al-Chalabi A. Amyotrophic Lateral Sclerosis. N Engl J Med. 2017;377:162–72.
4. Morgan S, Shatunov A, Sproviero W, Jones AR, Shoai M, Hughes D, et al. A comprehensive analysis of rare genetic variation in amyotrophic lateral sclerosis in the UK. Brain. 2017;140:1611–8.
5. Sayitoğlu M. Clinical interpretation of genomic variations. Turk J Haematol. 2016;33:172.
6. Liu J, Lillo C, Jonsson PA, Vande Velde C, Ward CM, Miller TM, et al. Toxicity of familial ALS-linked SOD1 mutants from selective recruitment to spinal mitochondria. Neuron. 2004;43:5–17.
7. Andersen PM, Al-Chalabi A. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? Nat Rev Neurol. 2011;7:603–15.
8. Liscic RM, Breljak D. Molecular basis of amyotrophic lateral sclerosis. Prog Neuropsychopharmacol Biol Psychiatry. 2011;35:370–2.
9. Vance C, Rogelj B, Hortobagyi T, De Vos KJ, Nishimura AL, Sreedharan J, et al. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. Science. 2009;323:1208–11.
10. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011;72:245–56.
11. Iacoangeli A, Al Khleifat A, Sproviero W, Shatunov A, Jones A, Dobson R, et al. DNAscan: a fast, computationally and memory efficient bioinformatics pipeline for the analysis of DNA next-generation-sequencing data. BioRxiv. 2018;267195.
12. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
13. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-prints [online], 1303. Available at: http://adsabs.harvard.edu/abs/2013arXiv1303.3997L. Accessed March 1, 2013.
14. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. ArXiv e-prints [online], 1207. Available at: http://adsabs.harvard.edu/abs/2012arXiv1207.3907G. Accessed July 1, 2012.
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
16. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
17. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27:1895–903.
18. Dabbish L, Stuart C, Tsay J, Herbsleb J. Social coding in GitHub: transparency and collaboration in an open software repository. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, Seattle, Washington, February 11–15, 2012, 1277–1286.
19. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014; 2014:2.

20. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12: e0177459.

21. Bradshaw WJ, Rehman S, Pham TT, Thiyagarajan N, Lee RL, Subramanian V, Acharya KR. Structural insights into human angiogenin variants implicated in Parkinson's disease and Amyotrophic Lateral Sclerosis. Sci Rep. 2017; 7:41996.

22. Smith BN, Topp SD, Fallini C, Shibata H, Chen HJ, Troakes C, et al. Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. Sci Transl Med. 2017;9:pii: eaad9157.

23. Lacomblez L, Doppler V, Beucler I, Costes G, Salachas F, Raisonnier A, et al. APOE: a potential marker of disease progression in ALS. Neurology. 2002;58:1112–4.

24. Elden AC, Kim HJ, Hart MP, Chen-Plotkin AS, Johnson BS, Fang X, et al. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature. 2010;466:1069.

25. Fogh I, Lin K, Tiloca C, Rooney J, Gellera C, Diekstra FP, et al. Association of a locus in the CAMTA1 gene with survival in patients with sporadic amyotrophic lateral sclerosis. JAMA Neurol. 2016;73:812–20.

26. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet. 2016;48:1043.

27. Williams KL, Topp S, Yang S, Smith B, Fifita JA, Warraich ST, et al. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. Nat Commun. 2016;7:11253.

28. Bannwarth S, Ait-El-Mkadem S, Chaussenot A, Genin EC, Lacas-Gervais S, Fragaki K, et al. A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. Brain. 2014;137:2329–45.

29. Mitchell J, Paul P, Chen HJ, Morris A, Payling M, Falchi M, et al. Familial amyotrophic lateral sclerosis is associated with a mutation in D-amino acid oxidase. Proc Natl Acad Sci. 2010;107:7556–61.

30. Münch C, Sedlmeier R, Meyer T, Homberg V, Sperfeld AD, Kurt A, et al. Point mutations of the p150 subunit of dynactin (DCTN1) gene in ALS. Neurology. 2004;63: 724–6.

31. Van Hoecke A, Schoonaert L, Lemmens R, Timmers M, Staats KA, Laird AS, et al. EPHA4 is a disease modifier of amyotrophic lateral sclerosis in animal models and in humans. Nat Med. 2012;18:1418.

32. Osmanovic A, Rangnau I, Kosfeld A, Abdulla S, Janssen C, Auber B, et al. FIG4 variants in central European patients with amyotrophic lateral sclerosis: a whole-exome and targeted sequencing study. Eur J Hum Genet. 2017; 25:324.

33. Kim HJ, Kim NC, Wang YD, Scarborough EA, Moore J, Diaz Z, et al. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. Nature. 2013;495:467.

34. Nicolas A, Kenna KP, Renton AE, Ticozzi N, Faghri F, Chia R, et al. Genome-wide analyses identify KIF5A as a novel ALS gene. Neuron. 2018;97:1268–83. e06.

35. Johnson JO, Pioro EP, Boehringer A, Chia R, Feit H, Renton AE, et al. Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. Nat Neurosci. 2014; 17:664.

36. Kenna KP, van Doormaal PTC, Dekker AM, Ticozzi N, Kenna BJ, Diekstra FP, et al. NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. Nat Genet. 2016;48:1037.

37. Tazelaar GH, Dekker AM, van Vugt JJFA, van der Spek RA, Westeneng HJ, Kool LJBG, et al. Association of NIPA1 repeat expansions with amyotrophic lateral sclerosis in a large international cohort. Neurobiol Aging. 2018;pii: S0197–4580:30336–1.

38. Cirulli ET, Lasseigne BN, Petrovski S, Sapp PC, Dion PA, Leblond CS, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. Science. 2015;347:1436–41.

39. Wu CH, Fallini C, Ticozzi N, Keagle PJ, Sapp PC, Piotrowska K, et al. Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. Nature. 2012;488: 499.

40. Bonifati V, Rizzu P, van Baren MJ, Schaap O, Breedveld GJ, Krieger E, et al. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. Science. 2003;299:256–9.

41. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature. 1993;362:59.

42. Fecto F, Yan J, Vemula SP, Liu E, Yang Y, Chen W, et al. SQSTM1 mutations in familial and sporadic amyotrophic lateral sclerosis. Arch Neurol. 2011;68: 1440–6.

43. Chen YZ, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, et al. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). Am J Hum Genet. 2004;74:1128–35.

44. Couthouis J, Hart MP, Shorter J, DeJesus-Hernandez M, Erion R, Oristano R, et al. A yeast functional screen predicts new candidate ALS disease genes. Proc Natl Acad Sci. 2011;108:20881–90.

45. Sreedharan J, Blair IP, Tripathi VB, Hu X, Vance C, Rogelj B, et al. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science. 2008;319: 1668–72.

46. Smith BN, Ticozzi N, Fallini C, Gkazi AS, Topp S, Kenna KP, et al. Exome-wide rare variant analysis identifies TUBA4A mutations associated with familial ALS. Neuron. 2014;84:324–31.

47. Deng HX, Chen W, Hong ST, Boycott KM, Gorrie GH, Siddique N, et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. Nature. 2011;477:211.

48. Nishimura AL, Mitne-Neto M, Silva HC, Richieri-Costa A, Middleton S, Cascio D, et al. A mutation in the vesicle-trafficking protein VAPB causes late-onset spinal muscular atrophy and amyotrophic lateral sclerosis. Am J Hum Genet. 2004;75:822–31.

49. Johnson JO, Mandrioli J, Benatar M, Abramzon Y, Van Deerlin VM, Trojanowski JQ, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron. 2010;68:857–64.

50. Consortium A. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiol Aging. 2013;34:357.e7–19.

51. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.

52. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016; 11:1.

53. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248.

54. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19: 1553–61.

55. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575.

56. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8:R232.

57. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C, et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34:511–3.

58. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31:2745–7.

59. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.

60. Kim S, Jhong JH, Lee J, Koo JY. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10:2.

61. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.

62. Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics* 2018;26(10):1537.

63. Nair PS, Vihinen M. VariBench: a benchmark database for variations. Hum Mutat. 2013;34:42–9.

64. Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. Hum Mutat. 2016;37:1013–24.

65. Li MX, Kwan JS, Bao SY, Yang W, Ho SL, Song YQ, et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet. 2013;9:e1003143.

66. Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015; 36:513–23.

67. Goksuluk D, Karaagaoglu E, Korkmaz S, Zararsiz G. easyROC: an interactive web-tool for ROC curve analysis using R language environment. R J. 2016;8: 213–30.

68. Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ, Desfeux A. OMICtools: an informative directory for multi-omic data analysis. Database (Oxford). 2014;2014:pii: bau069.

69. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015; 12:966–8.