OpEx - a validated, automated pipeline optimised for clinical exome sequence analysis

Elise Ruark1, Márton Münz2, Matthew Clarke1, Anthony Renwick1, Emma Ramsay1, Anna Elliott1, Sheila Seal1, Gerton Lunter2 & Nazneen Rahman1,3

We present an easy-to-use, open-source Optimised Exome analysis tool, OpEx (http://icr.ac.uk/opex) that accurately detects small-scale variation, including indels, to clinical standards. We evaluated OpEx performance with an experimentally validated dataset (the ICR142 NGS validation series), a large 1000 exome dataset (the ICR1000 UK exome series), and a clinical proband-parent trio dataset. The performance of OpEx for high-quality base substitutions and short indels in both small and large datasets is excellent, with overall sensitivity of 95%, specificity of 97% and low false detection rate (FDR) of 3%. Depending on the individual performance requirements the OpEx output allows one to optimise the inevitable trade-offs between sensitivity and specificity. For example, in the clinical setting one could permit a higher FDR and lower specificity to maximise sensitivity. In contexts where experimental validation is not possible, minimising the FDR and improving specificity may be a preferable trade-off for slightly lower sensitivity. OpEx is simple to install and use; the whole pipeline is run from a single command. OpEx is therefore well suited to the increasing research and clinical laboratories undertaking exome sequencing, particularly those without in-house dedicated bioinformatics expertise.

Exome sequencing is becoming a standard method used by increasingly diverse research and clinical laboratories. Unfortunately, easy-to-use, open-source exome analytical pipelines are not readily available, and most laboratories therefore implement bespoke pipelines1. There is wide variability in these underlying approaches and performance of analytical tools, and different versions of individual tools are often concurrently available, impeding standardisation and cross-validation. This is further exacerbated by the limited pipeline performance validation typically undertaken, particularly with respect to indel calling and specificity of pipelines, which are both vitally important for clinical exome sequencing. A further challenge for the clinical setting is the requirement to analyse large numbers of samples for optimal performance under the joint calling approach recommended by the most widely used software suite, GATK2. This is often not possible for clinical pipelines, where the need for fast test results necessitates analyses of smaller numbers of samples.

OpEx Description

To address these issues we created the OpEx (Optimised Exome) analysis tool. OpEx includes a fixed implementation of tools for read alignment, variant calling and annotation optimised for individual or multiple exome sequencing analysis, outputting data to clinical standards. We specifically focused on making OpEx simple to install and to use; the whole pipeline is run with a single command. It is therefore well suited to the increasing number of research and clinical laboratories without in-house dedicated exome bioinformatics expertise. An overview of OpEx is given below. Full details of OpEx are given in the Supplementary Appendix and are available at www.icr.ac.uk/opex.

OpEx takes the raw FASTQ files generated by paired-end Illumina sequencing as its input, which are mapped to the human reference genome using Stampy with BWA for pre-mapping3,4. A custom Python script then performs BAM sorting and indexing. Duplicate reads are marked with Picard (http://picard.sourceforge.net). We developed CoverView to provide the quality control information required in clinical laboratories. CoverView

1The Institute of Cancer Research, London, Division of Genetics & Epidemiology, Sutton SM2 5NG, UK. 2Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. 3Royal Marsden NHS Foundation Trust, Cancer Genetics Unit, Downs Road, Sutton, SM2 5PT, UK. Correspondence and requests for materials should be addressed to N.R. (email: rahmanlab@icr.ac.uk)
provides coverage, base quality and mapping quality metrics. CoverView utilises the user-supplied BED file and Ensembl transcripts to output the target regions which failed to achieve good quality coverage and annotates these regions according to their location in the gene transcript (e.g. c.1500–1550 of BRCA1). This seamless transcript-based annotation is particularly useful for the many laboratories that use orthogonal methods, such as Sanger sequencing, to ‘fill-in’ areas that fail to reach the required coverage threshold.

Within OpEx we provide a default transcript set which includes the Locus Reference Genomic (LRG) sequences to facilitate clinical utility. However, users can specify any Ensembl transcripts, including the entire set of available transcripts, if preferred. For the default transcript we used BioMart to obtain a list of all known protein-coding transcripts. There were 83,331 different protein-coding transcripts corresponding to 20,228 genes. We selected the 19,922 genes present on chromosomes 1–22, X, or Y, that had both a start and stop codon, and were not known pseudogenes. 74% of these genes (n = 14,830) had more than one transcript. We selected a single transcript for these genes with preference for a CCDS transcript according to Supplementary Figure 1. The specific transcripts selected are given in Supplementary Table 1.

OpEx uses Platypus for variant calling. Usefully, Platypus provides equivalent performance irrespective of the number of exomes being analysed. Analysis of many samples can thus be parallelised with no reduction in calling performance. Platypus also calls base substitutions and indels simultaneously. This rapid independent variant calling approach enables analysis to keep pace with the sequencing output, allowing early assessment of potential sample quality issues which are not detectable through the CoverView coverage metrics, e.g. a contaminated sample showing excess heterozygosity. It also provides early opportunities to assess and act on variant data, for example allowing early identification of a disease-causing variant in a plausible candidate gene. These give OpEx a useful advantage over callers that need to call across all samples together to provide optimal performance and thus cannot be performed until all laboratory work for the exome sequences is completed.

For the ICR142 NGS validation series, which includes exome sequence FASTQ files for 142 samples together with Sanger sequencing data at 730 sites, OpEx uses Platypus v0.1.5. However, it cannot be assumed that other versions will have exactly the same performance because the OpEx code is freely available, users have the flexibility to use different versions of any OpEx software component, including other versions of Platypus. Moreover, the OpEx code can be run until all laboratory work for the exome data is completed, with the additional benefit that a single run of the OpEx pipeline can process a larger number of samples. This is particularly useful for laboratories that use multiple sequencing technologies.

CoverView coverage evaluation. To evaluate OpEx performance we first analysed the ICR142 NGS validation series, which includes exome sequence FASTQ files for 142 samples together with Sanger sequencing data at 730 sites. OpEx was run using one CPU per exome. Because samples are analysed independently, runtimes are limited only by the number of CPUs available and the size of the FASTQ file. The 142 samples had an average of 70 M reads in the input FASTQ files (range 30 M–159 M), resulting in an overall average runtime of 35 hours (range 17–101 hours). All 142 were run simultaneously, thus the actual runtime for the series was 101 hours. Faster runtimes would be possible with multithreading. Alignment constituted the majority of the runtime (69–88%), followed by the CoverView coverage evaluation (6–18%), though this could be reduced if evaluation is restricted to select genes, a common approach in clinical exome analysis. Variant calling and annotation comprised a very small proportion of runtime (0.4–1.6%) due to the rapid performance of Platypus and CAVA. The ICR142 NGS validation series contains Sanger sequencing data for both positive and negative sites, with a particular focus on indels, allowing assessment of specificity as well as sensitivity for indels. Four of the 730 variant sites with Sanger data were in transcripts not included in the default OpEx database and were excluded. All of the remaining 726 sites had at least 15x coverage in the exome data, defined as at least 15 reads of good mapping quality (mapping score ≥ 20). We classified base substitutions as detected if the exact base change was called by OpEx. For indels, we also included the seven complex indels that were called by OpEx but annotated slightly differently from the Sanger call as having been detected. The evaluated sites are given in Supplementary Table 2. Performance of runtime (0.4–1.6%) due to the rapid performance of Platypus and CAVA.

The ICR142 NGS validation series contains Sanger sequencing data for both positive and negative sites, with a particular focus on indels, allowing assessment of specificity as well as sensitivity for indels. Four of the 730 variant sites with Sanger data were in transcripts not included in the default OpEx database and were excluded. All of the remaining 726 sites had at least 15x coverage in the exome data, defined as at least 15 reads of good mapping quality (mapping score ≥ 20). We classified base substitutions as detected if the exact base change was called by OpEx. For indels, we also included the seven complex indels that were called by OpEx but annotated slightly differently from the Sanger call as having been detected. The evaluated sites are given in Supplementary Table 2. Performance of runtime (0.4–1.6%) due to the rapid performance of Platypus and CAVA.

The ICR142 NGS validation series contains Sanger sequencing data for both positive and negative sites, with a particular focus on indels, allowing assessment of specificity as well as sensitivity for indels. Four of the 730 variant sites with Sanger data were in transcripts not included in the default OpEx database and were excluded. All of the remaining 726 sites had at least 15x coverage in the exome data, defined as at least 15 reads of good mapping quality (mapping score ≥ 20). We classified base substitutions as detected if the exact base change was called by OpEx. For indels, we also included the seven complex indels that were called by OpEx but annotated slightly differently from the Sanger call as having been detected. The evaluated sites are given in Supplementary Table 2. Performance of runtime (0.4–1.6%) due to the rapid performance of Platypus and CAVA.

The ICR142 NGS validation series contains Sanger sequencing data for both positive and negative sites, with a particular focus on indels, allowing assessment of specificity as well as sensitivity for indels. Four of the 730 variant sites with Sanger data were in transcripts not included in the default OpEx database and were excluded. All of the remaining 726 sites had at least 15x coverage in the exome data, defined as at least 15 reads of good mapping quality (mapping score ≥ 20). We classified base substitutions as detected if the exact base change was called by OpEx. For indels, we also included the seven complex indels that were called by OpEx but annotated slightly differently from the Sanger call as having been detected. The evaluated sites are given in Supplementary Table 2. Performance of runtime (0.4–1.6%) due to the rapid performance of Platypus and CAVA.
The overall sensitivity of all calls was 97%, specificity was 90% and FDR was 8%. Using only the high-quality calls improved the specificity (97%) and FDR (3%) with a marginal reduction of sensitivity (95%). Sensitivity for base substitutions was excellent in both the all call set and the high-quality call set, but the specificity was markedly improved in the high-quality call set (from 60% to 96%). The indel calling performance was good, with overall sensitivity of 96%, specificity of 95% and FDR of 6% in the all call set. Using only high-quality calls, specificity (97%) and FDR (4%) of indel calls were improved, but there was some reduction in sensitivity (93%) (Table 1).

The eleven missed indels were explicable by the sequence context or quality: four were deletions in long homopolymer tracts, four were in regions with poor mapping quality, two had >25% of reads with poor base quality, and one was in a simple repeat (Supplementary Table 3). Thus it is possible to predict, and potentially flag, regions in which indel calling may be suboptimal, due to well-recognised limitations of short-read NGS data. Importantly, the nine false indel calls in the high-quality call set are all present in both the Exome Variant Server and ExAC databases, suggesting they may represent pipeline-independent artefacts of NGS analysis.

Dependong on the individual performance requirements the OpEx output allows one to optimise the inevitable trade-offs between sensitivity and specificity. For example, in the clinical setting one often tolerates a higher FDR and lower specificity to maximise sensitivity. In contexts where experimental validation is not possible, minimising the FDR and improving specificity is often an acceptable trade-off for slightly lower sensitivity.

We next applied OpEx to a large-scale dataset, the ICR1000 UK exome series, to better evaluate OpEx performance in detection of rare variation. We selected 116 sites at which OpEx detected a variant in only one individual in the ICR1000 UK exome series, to evaluate performance at rare variation. Variant sites were validated by Sanger sequencing. The sites included 31 base substitutions, 36 deletions and 49 insertions. The 31 base substitutions included 31 base substitutions, 36 deletions and 49 insertions. The 31 base substitutions were excellent in both the all call set and the high-quality call set, but the specificity was markedly improved in the high-quality call set (from 60% to 96%). The indel calling performance was good, with overall sensitivity of 96%, specificity of 95% and FDR of 6% in the all call set. Using only high-quality calls, specificity (97%) and FDR (4%) of indel calls were improved, but there was some reduction in sensitivity (93%) (Table 1).

Finally we tested OpEx performance in a proband-parent trio exome sequencing dataset to detect de novo variation, one of the commonest uses of clinical exome sequencing. We used exome data from a trio in which the proband has an overgrowth syndrome due to a de novo indel mutation c.933_934insTCTT in DNMT3A. We used a Python script to first identify all variant sites in the proband which were not present in either parent and then generated the coverage at these sites in both parents using CoverView. A variant was a candidate de novo variant if the quality flag was “high”, the TC value was at least 15x, and the coverage was at least 15x in both parents. Protein-altering variants were defined as those with CAVA class NSY, EE, IF, ESS, SG, or FS and those variants with CAVA class ES, SG, or FS were further defined as protein-truncating. Comparison of the proband and parental data highlighted 27 candidate de novo protein-altering variants, of which only two were protein-truncating variants, one of which was the DNMT3A mutation. The analysis was completed in 33 hours.

**Summary.** OpEx is simple to install and can be run with a single command and input FASTQ files. The software is open source and freely available for download at www.icr.ac.uk/opex or at https://github.com/RahmanTeam/OpEx. The individual components of the pipeline are also freely available. Our detailed

| Variant type | All calls | High quality calls |
|--------------|-----------|--------------------|
| **Sensitivity** |           |                    |
| Overall | 378/389 (97%) | 368/389 (95%) |
| Base substitutions | 120/120 (100%) | 117/120 (98%) |
| All indels | 274/286 (96%) | 266/286 (93%) |
| Short indels | 258/269 (96%) | 251/269 (93%) |
| Long indels | 16/17 (94%) | 15/17 (88%) |
| **Specificity** |           |                    |
| Overall | 288/320 (90%) | 309/320 (97%) |
| No base substitution present | 27/45 (61%) | 43/45 (96%) |
| No indel present (short only) | 261/275 (95%) | 266/275 (97%) |
| No indel present (short and long) | 245/275 (89%) | 260/275 (95%) |
| **False detection rate** |           |                    |
| Overall | 32/410 (8%) | 11/379 (3%) |
| Base substitutions | 18/138 (13%) | 2/119 (2%) |
| All indels | 32/307 (10%) | 15/281 (5%) |
| Short indels | 15/273 (6%) | 9/260 (4%) |
| Long indels | 16/32 (50%) | 6/21 (29%) |

Table 1. OpEx variant calling performance in ICR142 NGS validation series. *Overall numbers exclude long indels.
documentation, the availability of the OpEx source code, and the availability of the ICR142 and ICR1000 datasets allow users of other open-source frameworks to fully replicate and evaluate our pipeline in-house. However, we believe the primary advantage of OpEx is as a fully developed validated pipeline requiring minimal user input and no specialised informatic expertise. Due to its excellent performance and suitability for clinical and research settings, OpEx is now our standard pipeline for all exome analyses.

References
1. Lelieveld, S. H., Veltman, J. A. & Gilissen, C. Novel bioinformatic developments for exome sequencing. Hum Genet, doi: 10.1007/s00439-016-1658-6 (2016).
2. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43, 11.10.11-33, doi: 10.1002/0471250953.bi111043 (2013).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760, doi: 10.1093/bioinformatics/btp324 (2009).
4. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21, 936–939, doi: 10.1101/gr.111120.110 (2011).
5. MacArthur, J. A. et al. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res 42, D873–D878, doi: 10.1093/nar/gkt1198 (2014).
6. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46, 912–918, doi: 10.1038/ng.3036 (2014).
7. Muñiz, M. et al. CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med 7, 76, doi: 10.1186/s13073-015-0195-6 (2015).
8. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6, 80–92, doi: 10.4161/fly.19695 (2012).
9. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070, doi: 10.1093/bioinformatics/btp330 (2010).
10. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164, doi: 10.1093/nar/gkq603 (2010).
11. Ruark, E. et al. The ICR142 NGS validation series: a resource for orthogonal assessment of NGS analysis. F1000Res 5, 886, doi: 10.12688/f1000research.8219.1 (2016).
12. Exome Aggregation Consortium (ExAC), Cambridge, MA (URL: http://exac.broadinstitute.org) Accessed: December 2014.
13. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) Accessed: December 2014.
14. Ruark, E. et al. The ICR1000 UK exome series: a resource of gene variation in an outbred population. F1000Res 4, 883, doi: 10.12688/f1000research.7049.1 (2015).
15. TDDD. Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223–228, doi: 10.1038/nature14135 (2015).
16. Tatton-Brown, K. et al. Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nat Genet 46, 385–388, doi: 10.1038/ng.2917 (2014).

Acknowledgements
We are grateful to the Scientific Computing Team at the Institute of Cancer Research for provision of HPC services, to Chey Loveday, Shawn Yost and Tara Mills, for their helpful comments and Ann Strydom for assistance in preparing the manuscript. We acknowledge support from the NIHR RM/ICR Specialist Biomedical Research Centre for Cancer. This study was funded by Wellcome Trust (Grant reference: 200990/Z/16/Z) and the Institute of Cancer Research, London.

Author Contributions
N.R., E. Ruark, M.M., M.C. and G.L. participated in the design of OpEx, the code for which was written by M.M. with input from E. Ruark and M.C. A.R., E. Ramsay, S.S. and A.E. generated the sequencing data. The manuscript was written by E. Ruark and N.R. All authors read and approved the final manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Ruark, E. et al. OpEx - a validated, automated pipeline optimised for clinical exome sequence analysis. Sci. Rep. 6, 31029; doi: 10.1038/srep31029 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2016