DATA NOTE

REVISED The ICR142 NGS validation series: a resource for orthogonal assessment of NGS analysis [version 2; referees: 2 approved]

Elise Ruark[1], Anthony Renwick[1], Matthew Clarke[1], Katie Snape[1], Emma Ramsay[1], Anna Elliott[1], Sandra Hanks[1], Ann Strydom[1], Sheila Seal[1], Nazneen Rahman[1,2]

[1]Division of Genetics & Epidemiology, The Institute of Cancer Research, London, UK
[2]Cancer Genetics Unit, Royal Marsden NHS Foundation Trust, London, UK

v2 **First published:** 22 Mar 2016, **5**:386 (doi: 10.12688/f1000research.8219.1)
Latest published: 05 Sep 2018, **5**:386 (doi: 10.12688/f1000research.8219.2)

Abstract

To provide a useful community resource for orthogonal assessment of NGS analysis software, we present the ICR142 NGS validation series. The dataset includes high-quality exome sequence data from 142 samples together with Sanger sequence data at 704 sites; 416 sites with variants and 288 sites at which variants were called by an NGS analysis tool, but no variant is present in the corresponding Sanger sequence. The dataset includes 293 indel variants and 247 negative indel sites, and thus the ICR142 validation dataset is of particular utility in evaluating indel calling performance. The FASTQ files and Sanger sequence results can be accessed in the European Genome-phenome Archive under the accession number EGAS00001001332.

Keywords

Variant calling, next-generation sequencing, NGS, exome, indel, validation

This article is included in the Data: Use and Reuse collection.

Open Peer Review

Referee Status: ✔ ✔

	Invited Referees 1	2
REVISED **version 2** published 05 Sep 2018		
version 1 published 22 Mar 2016	✔ report	✔ report

1 **Richard Bagnall**, The University of Sydney, Australia

2 **Brad Chapman**, Harvard Public School of Health, USA
Oliver Hofmann, University of Glasgow, UK

Discuss this article

Comments (0)

Corresponding author: Nazneen Rahman (rahmanlab@icr.ac.uk)

Author roles: Ruark E: Conceptualization, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; **Renwick A**: Data Curation, Resources; **Clarke M**: Data Curation, Project Administration; **Snape K**: Resources; **Ramsay E**: Data Curation; **Elliott A**: Data Curation; **Hanks S**: Data Curation; **Strydom A**: Data Curation, Project Administration; **Seal S**: Resources; **Rahman N**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: We acknowledge NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and the ICR. This study was funded by the Institute of Cancer Research, London.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Ruark E, Renwick A, Clarke M *et al.* **The ICR142 NGS validation series: a resource for orthogonal assessment of NGS analysis [version 2; referees: 2 approved]** *F1000Research* 2018, **5**:386 (doi: 10.12688/f1000research.8219.2)

First published: 22 Mar 2016, **5**:386 (doi: 10.12688/f1000research.8219.1)

REVISED **Amendments from Version 1**

We recently released ICR142 Benchmarker, a tool that uses the ICR142 NGS validation series to generate standardised outputs and metrics to evaluate, optimise and benchmark variant calling algorithms. In the development of ICR142 Benchmarker we refined the ICR142 NGS validation series to maximise its utility, based on detailed re-review of all the data and user feedback.

The methods used in the data review are detailed in the amended paper and led to the exclusion of 26 sites such that the ICR142 NGS validation series now includes 704 sites: 416 sites with variants and 288 sites at which variants were called by an NGS analysis tool, but no variant is present in the corresponding Sanger sequence. This amended version of paper describes the updated dataset.

ICR142 Benchmarker is described in Ruark E, Holt E, Renwick A *et al.* ICR142 Benchmarker: evaluating, optimising and benchmarking variant calling using the ICR142 NGS validation series [version 1; referees: awaiting peer review]. Wellcome Open Res 2018, 3:108 (doi: 10.12688/wellcomeopenres.14754.1)

See referee reports

Introduction

Next-generation sequencing (NGS) approaches have greatly enhanced our ability to detect genetic variation. Over the past decade NGS hardware, software, throughput, data quality and analytical tools have evolved dramatically. Thorough evaluation of each new laboratory and analytical development is challenging but necessary to fully understand how pipeline modification can impact results. To fully assess performance, NGS analysis tools should ideally be run on samples with pre-determined positive and negative sites assessed through orthogonal experimentation such as Sanger sequencing.

Over the past five years, we have generated extensive data on thousands of samples using different NGS instruments, sequencing chemistry, gene panels, exome captures and variant calling tools. Fortuitously, during this process we have generated orthogonal validation data using Sanger sequencing for a core set of 142 samples that were included in the majority of our experiments. We now formally use these samples, which we call the ICR142 NGS validation series, to evaluate NGS variant calling performance after any change to experimental or analytical protocols. This series has proved an extremely useful resource for our assessment of NGS analysis in both the research and clinical settings. We believe that it may also have utility for others, and hence are making it available here.

Materials and methods

We used lymphocyte DNA from 142 unrelated individuals. All individuals were recruited to the BOCS study and have given informed consent for their DNA to be used for genetic research. The study is approved by the London Multicentre Research Ethics Committee (MREC/01/2/18).

Over the last five years we have generated data from the ICR142 validation series using different exome captures which we have analysed with multiple aligner/caller combinations[1–6]. To date we have generated Sanger sequence data for 704 sites amongst the 142 individuals. These sites include variants called by only one aligner and caller combination, increasing the representation of sites which can discriminate performance between methods.

To generate the Sanger sequence data, we performed PCR reactions using the Qiagen Multiplex PCR kit, and bidirectional sequencing of resulting amplicons using the BigDye terminator cycle sequencing kit and an ABI3730 automated sequencer (ABI PerkinElmer). All sequencing traces were analysed with both automated software (Mutation Surveyor version 3.10, SoftGenetics) and visual inspection.

To determine if a variant was present we visually inspected each Sanger sequence with Chromas software v2.13. For each site we selected an ENST from release 65 as the reference sequence. We reviewed at least 100 base pairs of sequence flanking each variant site to allow for position/annotation errors. We considered a base substitution to be confirmed if the correct variant was called at the exact position and the variant base signal was accompanied by a corresponding reduction in the reference base signal. There were 123 confirmed base substitution variants. We considered an indel variant to be confirmed if an indel variant was present in the region of interest and the indel variant allele signal was present along the complete length of the region of interest. There were 293 confirmed indel variants.

We considered a site negative for a base substitution if the specific base substitution was not present, resulting in 41 negative base substitution sites. We considered a site negative for an indel if no indel, of any kind, was detected in the 200 base pair region of interest, resulting in 247 negative indel sites (Figure 1). We annotated confirmed variants with the HGVS-compliant CSN standard using CAVA (version 1.1.0) according to the transcripts designated in Supplementary table 1[7].

We have also generated high-quality exome sequencing data for the ICR142 NGS validation series. We prepared DNA libraries from 1.5 µg genomic DNA using the Illumina TruSeq sample preparation kit. DNA was fragmented using Covaris technology and the libraries were prepared without gel size selection. We performed target enrichment in pools of six libraries (500 ng each) using the Illumina TruSeq Exome Enrichment kit. The captured DNA libraries were PCR amplified using the supplied paired-end PCR primers. Sequencing was performed with an Illumina HiSeq2000 (SBS Kit v3, one pool per lane) generating 2×101 bp reads. CASAVA v1.8.1 (Illumina) was used to demultiplex and create FASTQ files per sample from the raw base call files.

All of the 704 sites had at least 15× coverage in the exome data, defined as at least 15 reads of good mapping quality (mapping score ≥20). Because these sites are well covered, we can readily assess the variant calling performance of any software tool by applying the pipeline to the exome sequencing data and comparing the variant calls with the Sanger sequencing dataset.

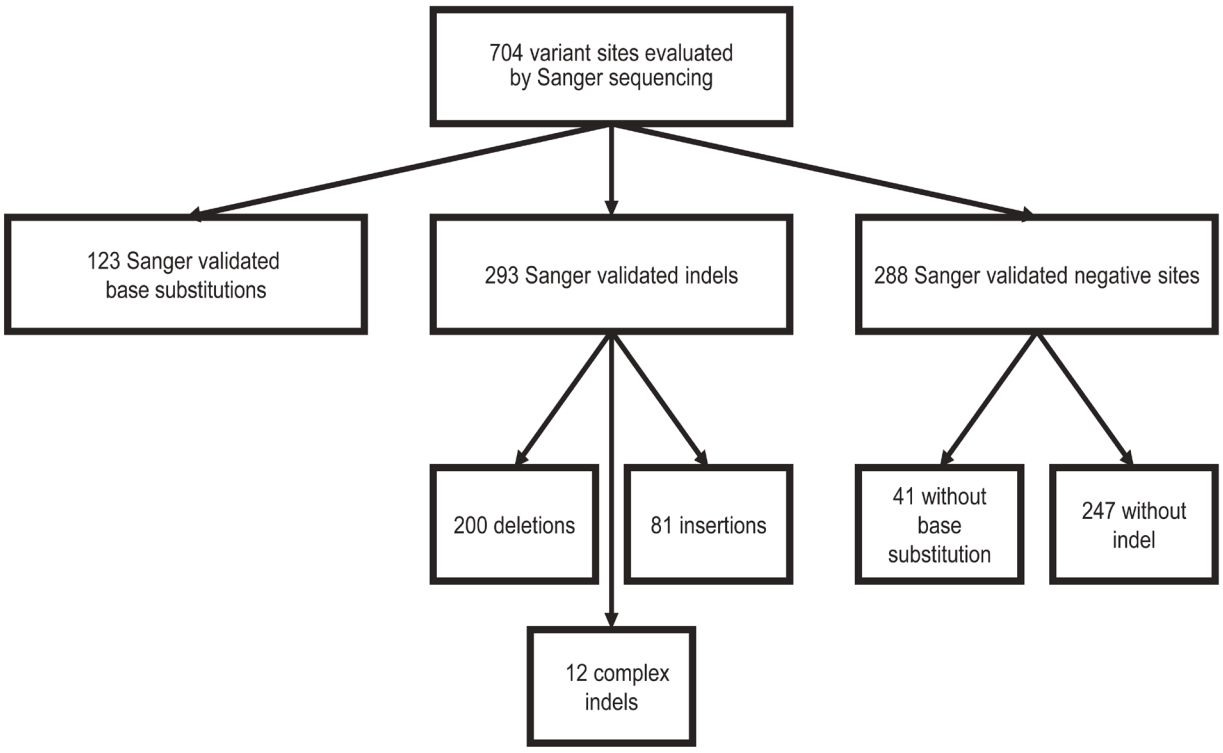

Figure 1. Description of variant sites evaluated by Sanger sequencing in the ICR142 NGS validation series.

Data availability

We have deposited the FASTQ files for all 142 individuals in the European Genome-phenome archive (EGA). The accession number is EGAS00001001332.

Researchers and authors that use the ICR142 NGS validation series should reference this paper and should include the following acknowledgement: "This study makes use of the ICR142 NGS validation series data generated by Professor Nazneen Rahman's team at The Institute of Cancer Research, London".

Author contributions

N.R. and E.Ru. designed the experiment. A.R., E.Ra. and SH generated the exome data. E.Ru. and A.E. undertook data management, S.S., A.R., and K.S. undertook sample management and Sanger validations. M.C. and A.S. undertook the data and administrative management required for data to be accessible. E.Ru. and N.R. wrote the manuscript. All authors contributed to the final manuscript.

Grant information

We acknowledge NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and the ICR. This study was funded by the Institute of Cancer Research, London.

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are grateful to the Scientific Computing Team at the Institute of Cancer Research for provision of HPC services. We are grateful to Peter Humburg, Andy Rimmer, Manuel Rivas and Peter Donnelly for undertaking some of the aligner/caller comparisons.

Supplementary material

Supplementary table 1. Sanger sequencing results for 704 sites in the ICR142 NGS validation series. Confirmed variants are annotated according to the designated transcript by CAVA using CSN[7].

Click here to access the data.

The description of the column headings are given below:

Sample – sample name in the ICR142 series

Gene – HGNC symbol
SangerCall – the most 3' representation annotated with CSN
Type – "bs", "del", "ins", "complex", or "indel" for base substitutions, simple deletions, simple insertions, complex indels, or negative indel sites, respectively
Transcript – the ENST ID from Ensembl v65 used to annotate the Sanger call
CHR – chromosome
EvaluatedPosition – evaluated hg19 site position, centre of designed amplicon
POS – the left-aligned position in hg19 coordinates for variants called in exome data by Platypus v0.1.5
REF – the reference allele in hg19 for variants called in exome data by Platypus v0.1.5
ALT – the alternative allele in hg19 for variants called in exome data by Platypus v0.1.5
Zygosity – "heterozygous" a variant that is present on one allele, "homozygous" a variant that is present on both alleles.
SiteID – numerical ID (1-704) of sites in the ICR142 NGS validation series

References

1. Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res.* 2011; **21**(6): 936–9.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–60.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Rivas MA, Beaudoin M, Gardet A, *et al.*: **Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.** *Nat Genet.* 2011; **43**(11): 1066–73.
PubMed Abstract | Publisher Full Text | Free Full Text
4. McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–303.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Rimmer A, Mathieson I, Lunter G, *et al.*: **Platypus: An Integrated Variant Caller.** 2012.
Reference Source
6. SOFTGENETICS: **NextgGENe® software for Next Generation (NGS) sequence analysis.**
Reference Source
7. Münz M, Ruark E, Renwick A, *et al.*: **CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting.** *Genome Med.* 2015; **7**(1): 76.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 03 May 2016

doi:10.5256/f1000research.8841.r13013

✔ **Brad Chapman** [1], **Oliver Hofmann** [2]

[1] Department of Biostatistics, Harvard Public School of Health, Boston, MA, USA

[2] Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Glasgow, UK

The authors describe ICR142, a publicly available set of fastq files and confirmed true and false variants for validating analysis pipelines. This is an incredibly useful community resource that complements existing efforts like the Genome in a Bottle project by providing a set of validated, difficult regions to evaluate variant detection tools. I appreciate the efforts to make these test sets public; instead of having validation sets like these developed internally at clinical laboratories, we can collaborate and improve them publicly.

In collaboration with Oliver Hofmann at the Wolfson Wohl Cancer Research Center (https://twitter.com/fiamh) we obtained access to the data and were able to run a validation using bcbio variant calling (http://bcbio-nextgen.readthedocs.io). In doing this, we tried to address a couple of challenges for other users wanting to make immediate use of this data in their own in hour validation work:

- The truth sets are not easy to plug into existing validation frameworks. Most validation tools like rtg vcfeval and hap.py work from VCF format files, while this truth set is in a custom spreadsheet format with a mixture of methods for describing changes. You can use Platypus positions for many but need to use CSN descriptions or evaluated position for the remainder.

- The truth sets don't appear to describe if we expect calls to be homozygous or heterozygous calls at each position.

- Many existing validation approaches expect a single (or few) samples so coordinating checking and validation for all these samples can be a challenge.

- As part of this review, we generated a set of configuration files and scripts to help make running validations with ICR142 easier (https://github.com/bcbio/icr142-validation).

This comparison work also includes a set of comparisons with common callers (GATK HaplotypeCaller, FreeBayes and VarDict). Several of the Sanger validated regions without variants are false positives in at least 2 of the callers tested, so this dataset exposes some common issues with calling and filtering. It would be useful to hear the author's experience with validating callers using this benchmark set and if they have additional filters used to avoid these problems. Knowing a baseline expectation for results would help ensure that the users understand how correctly they've setup the validation resources.

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 21 April 2016

doi:10.5256/f1000research.8841.r13347

Richard Bagnall
Agnes Ginges Centre for Molecular Cardiology, Centenary Institute, Sydney Medical School, The University of Sydney, Sydney, NSW, Australia

A myriad of software tools have been developed for the alignment of next generation sequencing data to a reference genome and for the subsequent genotyping of DNA variants. Evaluating the specificity and sensitivity of a variant calling framework can be achieved with a dataset containing validated genotypes. Ruark et. al. provide the 'ICR142 NGS validation series' exome sequence fastq files of 142 individuals, and a large set of corresponding Sanger sequencing validated variant sites and sites where variants were called by an NGS tool, but no variant was found with the corresponding Sanger sequencing.

I found the NGS dataset to be easily accessible, on request, from the European Genome-phenome archive and it comprises paired end fastq sequencing files generated by an Illumina sequencing system on the stated 142 individuals. The Sanger sequencing dataset is available as supplementary table 1 of the manuscript. This is a useful resource for evaluating variant calling pipelines.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.