OPEN ACCESS

Original research

Recommendations for laboratory workflow that better support centralised amalgamation of genomic variant data: findings from CanVIG-UK national molecular laboratory survey

Sophie Allen ,[1] Lucy Loong,[1] Alice Garrett ,[1,2] Bethany Torr ,[1] Miranda Durkie,[3] James Drummond,[4] Alison Callaway,[5] Rachel Robinson,[6] George J Burghel ,[7] Helen Hanson ,[1,2] Joanne Field,[8] Trudi McDevitt,[9] Terri P McVeigh,[10] Tina Bedenham,[11] Christopher Bowles,[12] Kirsty Bradshaw,[13] Claire Brooks,[14] Samantha Butler,[15] Juan Carlos Del Rey Jimenez,[2] Lorraine Hawkes,[16] Victoria Stinton,[17] Suzanne MacMahon,[18,19] Martina Owens,[12] Sheila Palmer-Smith,[20] Kenneth Smith,[21] James Tellez,[22] Mikel Valganon-Petrizan,[18,19] Erik Waskiewicz,[20] Michael Yau,[16] Diana M Eccles ,[23,24] Marc Tischkowitz,[25] Shilpi Goel,[26,27] Fiona McRonald,[26] Antonis C Antoniou ,[28] Eva Morris,[29] Steven Hardy,[26] Clare Turnbull [1,10]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jmg-2023-109645).

For numbered affiliations see end of article.

Correspondence to
Professor Clare Turnbull, Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, SM2 5NG, UK; turnbull.lab@icr.ac.uk

SA and LL contributed equally.

This work has previously been presented at the Curating the Clinical Genome conference held 10–12 July 2023.

Received 18 September 2023
Accepted 28 October 2023

To cite: Allen S, Loong L, Garrett A, *et al*. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmg-2023-109645

ABSTRACT

Background National and international amalgamation of genomic data offers opportunity for research and audit, including analyses enabling improved classification of variants of uncertain significance. Review of individual-level data from National Health Service (NHS) testing of cancer susceptibility genes (2002–2023) submitted to the National Disease Registration Service revealed heterogeneity across participating laboratories regarding (1) the structure, quality and completeness of submitted data, and (2) the ease with which that data could be assembled locally for submission.

Methods In May 2023, we undertook a closed online survey of 51 clinical scientists who provided consensus responses representing all 17 of 17 NHS molecular genetic laboratories in England and Wales which undertake NHS diagnostic analyses of cancer susceptibility genes. The survey included 18 questions relating to 'next-generation sequencing workflow' (11), 'variant classification' (3) and 'phenotypical context' (4).

Results Widely differing processes were reported for transfer of variant data into their local LIMS (Laboratory Information Management System), for the formatting in which the variants are stored in the LIMS and which classes of variants are retained in the local LIMS. Differing local provisions and workflow for variant classifications were also reported, including the resources provided and the mechanisms by which classifications are stored.

Conclusion The survey responses illustrate heterogeneous laboratory workflow for preparation of genomic variant data from local LIMS for centralised submission. Workflow is often labour-intensive and inefficient, involving multiple manual steps which introduce opportunities for error. These survey findings and adoption of the concomitant recommendations may support improvement in laboratory dataflows, better facilitating submission of data for central amalgamation.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The principles of genomic laboratory workflow have been described at high level in previous publications.

WHAT THIS STUDY ADDS

⇒ This first national survey of genomic data workflow conducted in 2023 reflects in detail practices in 17 National Health Service molecular genetics laboratories in England and Wales.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The survey responses illustrate laboratory workflow for preparation of genomic variant data for centralised submission that is frequently labour-intensive, highly manual and inefficient. These findings may be instructive to laboratories for improving dataflows which better enable downstream submission of data for national amalgamation endeavours.

INTRODUCTION

Over the last decade, there has been a substantial, national-level focus on expanding the role of genomics in routine National Health Service (NHS) care, with the goal of clinical services 'operating to national standards, specifications and protocols'.[1–3] To deliver this transformation, NHS England has reconfigured the 28 English molecular diagnostic

laboratories into seven Genomic Laboratory Hubs (GLHs) with development of a National Test Directory for each clinical indication (denoted by an 'R-code') determining germline test eligibility criteria, gene panels and constituent molecular analyses for each genomic test.[4]

Underpinning this transformation, and key to expansion of genomic testing capacity, will be the data systems by which the genomic data are generated, processed, analysed and stored. Upstream workflow can be wholly automated, with high throughput conversion of image data from next-generation sequencing (NGS) into output variant call format (VCF) files listing the called genomic variants, with structured fields for the genomic location, variant nomenclature and quality metrics. This workflow has now been widely and successfully implemented by clinical bioinformaticians across NHS molecular diagnostic laboratories.

However, downstream of this are unavoidably more manual processes requiring expert evaluation of the detected germline variants by experienced clinical diagnostic scientists for (1) technical veracity and (2) pathogenicity classification. To confirm the variant is truly present, the sequence data may require manual inspection and further molecular analysis using an orthogonal technology to validate the called variant (for example, Multiplex Ligation-dependant Probe Amplification to confirm an exon-level deletion). While pre-established filters can be applied to remove many variants that are likely benign, manual evaluation of pathogenicity must be undertaken for the remaining variants identified. This requires assembly of information from multiple sources, including variant effect prediction, *in silico* protein function effect prediction, population frequencies (eg, gnomAD, UK Biobank), functional assay results, case–control study data, familial co-segregation data, previous classifications (eg, ClinVar, the Human Gene Mutation Database, local records) and review of the literature for phenotypical case descriptions.[5–10] The assembled evidence must then be compared with the (continually evolving) generic and gene-specific protocols which dictate the scoring of evidence elements, along with rules defining combination of evidence scores, in order to produce a final classification.[11–15] Only variants for which there is sufficient evidence for classification as 'pathogenic' (class 5) or 'likely pathogenic' (class 4) will be included in the diagnostic laboratory report for return to the clinician (and often patient). The majority of manually reviewed variants will not have attained sufficient evidence and these 'variants of uncertain significance' (VUSs, class 3) will not typically be included in the clinical report, unless the variant is on the threshold of uplift to likely pathogenic and meets the specific national criteria for being reported (as per the Association of Clinical Genomic Science (ACGS) Best Practice Guidelines).[16] Because of the dynamic nature of the available evidence and associated guidance, after the elapse of a defined time period, new observations of previously evaluated variants require fresh review and (re)classification; national approaches have been agreed for national alerts and reissuing reports on clinically important reclassification of a variant.[17]

After variant classification and generation of the clinical report, the detected variants will typically then be transferred for long-term storage into the local Laboratory Information Management System (LIMS) and, variably, into the hospital-level Electronic Health Record (EHR) system. In an LIMS or EHR, the variant data are associated with patient identifiers (ie, NHS number, date of birth, name) and additional patient information (eg, test indication, ethnicity, age and sometimes phenotype). VCF files may also be stored but most often only the 'SampleID' and/or 'RunID' will be available within VCF files; typically there are no patient identifiers or patient information available in the VCF. The workflow through which variant data pass, and the stored data items and formats, are potentially critical to the readiness and fidelity by which these local data can be shared and amalgamated.

National amalgamation of genomic data is important to advance broader, global understanding of variant pathogenicity to reduce classifications of VUS. Assessing the frequency, familial segregation and phenotypical associations with which a variant is observed adds important evidence regarding the pathogenicity of a variant. Local amalgamations of accrued observations of a given variant may on occasion be informative, but it is typically only with national or international amalgamation of data on rare variants that we have sufficient instances by which to evaluate their association with clinical disease. Thus, along with ensuring robust local recording and reporting, LIMS/EHRs would ideally be designed to readily enable submission of local de-identified variant and phenotype data for national/international amalgamation.

In just such an endeavour, an informatic pipeline enabling local submission and national amalgamation of pseudonymised individual-level variant data for cancer susceptibility genes (CSGs) has been established by the NHS National Disease Registration Service (NDRS). Since 2018, all 16 English NHS laboratories undertaking CSG analyses have submitted locally held individual-level data on all germline CSG tests performed and variants detected. Data are then restructured and extracted at NDRS using bespoke laboratory-specific informatic processing algorithms developed by NDRS.[18] This endeavour has, for the first time, allowed successful assembly of nationally complete individual-level data on genetic tests and detected variants, dating back for some laboratories as far as 2002. These data provide opportunities for variant interpretation and also longitudinal study of CSG variant carriers via linkage to cancer registrations. However, this activity has revealed considerable heterogeneity across participating laboratories regarding (1) the structure, quality and completeness of submitted data, and (2) the ease with which those data could be assembled locally for submission.

The Cancer Variant Interpretation Group UK (CanVIG-UK) was established in 2017, at the directive of ACGS, to coordinate interpretation of variants in CSGs.[15] CanVIG-UK meets monthly and comprises >300 members, including clinical scientists and genetics clinicians from each of the seven English GLHs (in addition to those from the devolved nations and Ireland). CanVIG-UK has since inception worked in close partnership with NDRS to coordinate laboratory data submissions, and also in the analysis and dissemination of the NDRS nationally amalgamated variant data.

The combined bioinformatic and human workflow through which variant data in VCFs are evaluated, classified, transferred to local LIMS/EHRs and eventually stored has been demonstrated to be critical to the feasibility and utility of the NDRS national data amalgamation. To better understand this workflow, through CanVIG-UK, we conducted a survey of the 17 individual English and Welsh laboratories which perform molecular diagnostic germline testing in CSGs.

METHODS

The survey questions were designed and piloted by the CanVIG-UK Steering and Advisory Group, which comprises eight senior clinical scientists and three consultant clinical geneticists working in Cancer Genetics across the GLHs

(online supplemental table 1). The closed online survey was then sent to up to three CanVIG-UK clinical scientists undertaking germline cancer susceptibility genetic analyses from each of the 16 laboratories in which diagnostic CSG testing in the GLH network is performed, as well as one laboratory in Wales. Return of a single consensus response representing each laboratory was requested in June 2023 (see online supplemental methods). Complete responses were returned by 17 of 17 laboratories surveyed.

The survey comprised 18 questions relating to 'NGS workflow' (11), 'variant classification' (3) and 'phenotypical context' (4) (summary details of all questions and responses from responding laboratories are presented in online supplemental table 2); there were a further five questions relating to 'respondent details' and a separate component comprising 12 questions relating to very specific logistical aspects of 'GLH-NDRS centralised data submission'.[19]

This survey was designed to assess elements of the workflow relevant to NDRS data submission; we did not survey on the workflow used specifically for generation of the diagnostic clinical report.

RESULTS

Variant workflow from the NGS outputs into the LIMS

There was substantial variation in the workflow for managing and documenting the process of technical verification and variant classification upstream of entering the variants onto the LIMS. The laboratories reported workflow by which one, two or sometimes three generations of intermediary VCF-derived files were generated and stored (online supplemental figure 1A). Variants listed on the VCF were typically associated with a sample ID (15 of 17) which was usually also available in the LIMS, but rarely with a patient name (3 of 17), and never with a Date of Birth (0 of 17) or NHS number (0 of 17) (online supplemental figure 1B).

Methods by which selected variants are entered into the LIMS largely rely on manual processes (figure 1A). Eight of 17 laboratories reported manually typing out the variant details, while 4 of 17 laboratories reported a manual 'copy-and-paste' mechanism for entering the variants into the LIMS. In two laboratories, the variant data could be entered into the LIMS via an automated 'push-button' transfer from the NGS outputs (although only one of the systems serves both single nucleotide variants and CNVs).

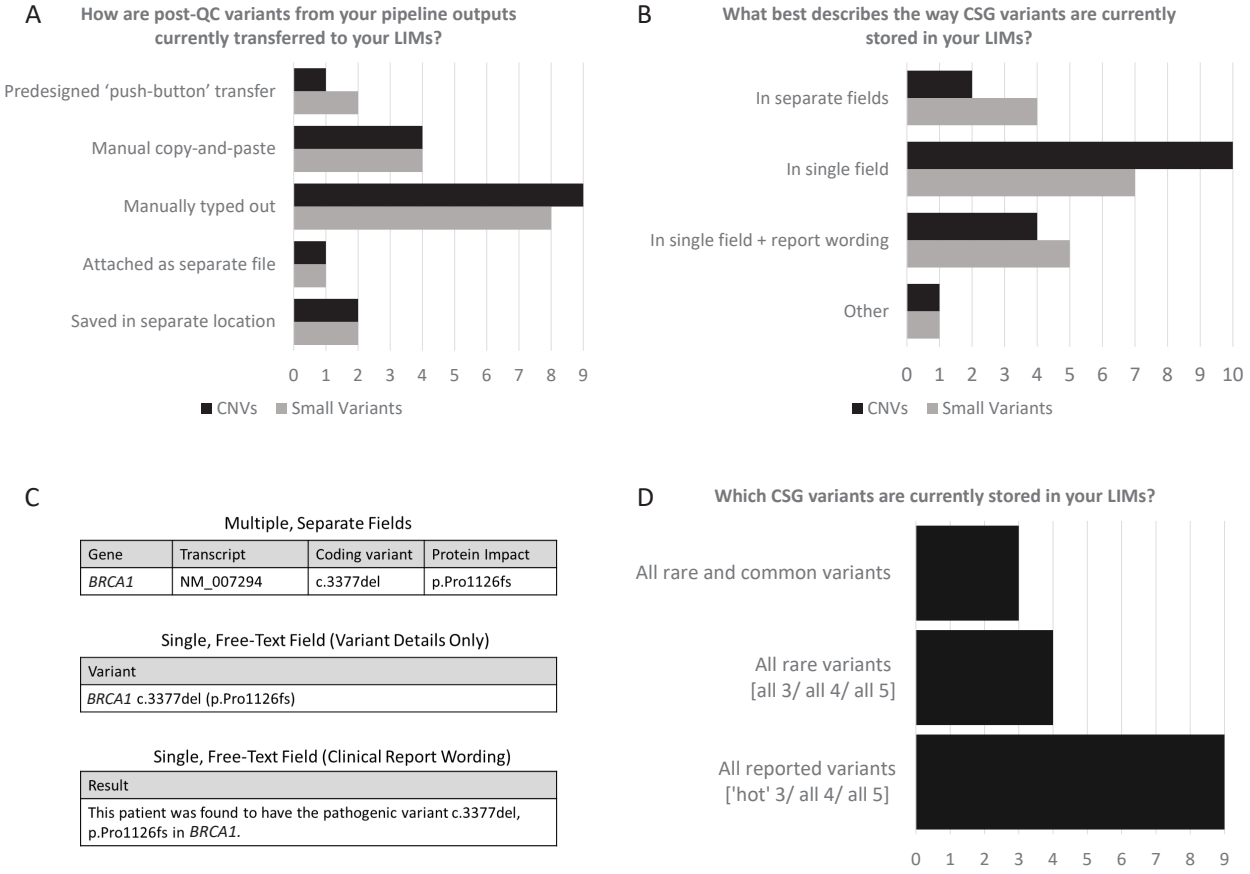


Figure 1 Responses from each laboratory for multiple choice questions pertaining to variant workflow from NGS pipelines to the LIMS. (A,B) All laboratory responses describing the method of transfer from NGS pipeline output to LIMS, and how this information is stored within the LIMS. (C) Visual description of differing LIMS storage formats. (D) All laboratory responses describing which detected variants are stored within an LIMS. CSG, cancer susceptibility gene; LIMS, Laboratory Information Management System; NGS, next-generation sequencing; QC, quality control.

For one laboratory, the variant data were not entered in the LIMS but attached in the LIMS as a separate file. For two laboratories, the variant details were not stored in the LIMS at all but in a separate location outside of the LIMS.

Formatting of variant data within the LIMS was another key area by which laboratories differed (figure 1B). In 4 of 17 laboratories, small variant details were stored in the LIMS using separate structured fields for gene, cDNA and protein change (figure 1C). For 7 of 17 laboratories, these details were stored in a single field, while for 5 of 17 laboratories, the variant details were embedded within free text (typically comprising the full report wording). For CNVs, separate structured fields were even less frequently used (2 of 17) to store details of the variants.

Most laboratories decided which of the detected variants were selected for storage in the LIMS dependant on if the variant was reported clinically (figure 1D). Three of 17 laboratories reported that all variants (rare and common) were stored in their LIMS, while 4 of 17 laboratories reported that all rare variants were stored in their LIMS (regardless of pathogenicity class). However, for 9 of 17 laboratories, only variants included in their clinical report were stored in their LIMS, which will therefore only comprise variants classified as likely pathogenic/pathogenic and occasional 'VUS' of particularly high suspicion (so-called 'hot' VUS).[16] Some of these laboratories reported parallel systems by which the other variants were stored against meaningful patient identifiers, for example, using an additional in-house Excel file or a separate database. When surveyed regarding their confidence that their local system would allow reliable variant retrieval in the event of a 'cold' VUS being upclassified into pathogenic, 13 were extremely or very confident, 3 were quite confident and 1 was not very confident (online supplemental figure 2).

Table 1 Responses for questions surrounding variant classification and storage of such information; for these questions, respondents could select multiple options

Interface for viewing variants requiring evaluation/classification	**No of labs**
Within a bioinformatic processing system/dedicated in-house variant system	8
In a spreadsheet (eg, VCF, VCF-derived file)	5
Other	4
Within the interface from which you view variants requiring interpretation, which description is most accurate?	**No of labs**
Most/many of the relevant data sources have been pre-imported	1
There are variant-specific links out to most/many of the relevant data sources	7
No or minimal annotations (eg, only population frequencies). Accessing of relevant data sources (Alamut, CanVar-UK, ClinVar, literature) requires manual interrogation (variant name is typed/pasted in)	9
Storage of variant evaluation/classification (laboratories may use more than one system)	**No of labs**
Dedicated in-house departmental variant data system	5
Commercial platform or software (eg, Congenica, Alamut)	4
LIMS (against specific patient)	3
Individual per-variant files. File is updated on each encounter of the variant	5
Individual per-variant files. New file is generated each time the variant is encountered	4
Individual per-patient episode files. May contain multiple variants	4
Per-gene files comprising multiple variants	1
Per-disease files comprising multiple genes (and multiple variants)	1

LIMS, Laboratory Information Management System; VCF, variant call format.

Variant classification

Also variable were the workflow and resources available to clinical scientists for variant classification, which is typically performed upstream of entry of the variant into the LIMS (table 1). In 8 of 17 laboratories, the variants requiring classification are viewed within a commercial or in-house 'variant system', while in 5 of 17 laboratories, the variants are viewed in a spreadsheet. Respondents from 9 of 17 laboratories reported that for the variants requiring classification, there would be minimal or no automatic annotations available in their workflow/system (beyond basic population frequencies), meaning that proactive manual accessing of multiple relevant data sources is required, such as Alamut, ClinVar and CanVar-UK. Respondents from 7 of 17 laboratories reported that their system provided variant-specific links out to most or many of the resources. Only one laboratory reported most/many of the relevant data resources being directly available within their variant system.

Regarding storage of detailed variant classification findings, the primary divide was between laboratories storing classification details as individual files versus those incorporating this information within a database (table 1). Most common (5 of 17 laboratories) was some form of in-house variant database (separate from the LIMS). Four laboratories store this information in a commercial platform (eg, Alamut, Congenica systems). Otherwise, there was a mix of approaches including dynamic per-variant files (updated on each new observation of the variant) or per-variant-per-patient files (generating a new file for each observation of a variant). These variant files were then stored in various locations including local drives or as attachments to the LIMS. Some laboratories reported using multiple storage methods (online supplemental table 2).

Phenotypical details and test context

In 15 of 17 laboratories, the details of the panel tested were specified in the LIMS (figure 2); this historically comprised series of gene names or local identifiers for their panels but since the introduction of the National Test Directory, 12 of 17 laboratories are routinely capturing the relevant clinical indication (R-code) in their LIMS. Only 11 of 17 laboratories record clinical details or phenotypical information in their LIMS (and this is only when information was provided on the request form). When small gene sets or single genes are reported from a larger panel (online supplemental table 2), a roughly equal number of laboratories reported listing the individual genes tested by name versus annotating with a subpanel name.

DISCUSSION

The responses from this survey describe the heterogeneous workflow used across the surveyed 17 laboratories for evaluating, classifying and storing germline variant data ahead of national centralised submission. Many of the reported elements in this workflow render potentially challenging and time-consuming the retrieval and centralised submission of variant data for national amalgamation. Workflow is often laborious and low-throughput, with manual steps which introduce opportunities for error. Given the commonality of many of these challenges across centres, we propose recommendations for workflow redesign that target the key challenges across workflow in variant data transfer, storage and retrieval.

Variant workflow

The survey showed that the highly structured VCF file output from NGS is frequently processed downstream using

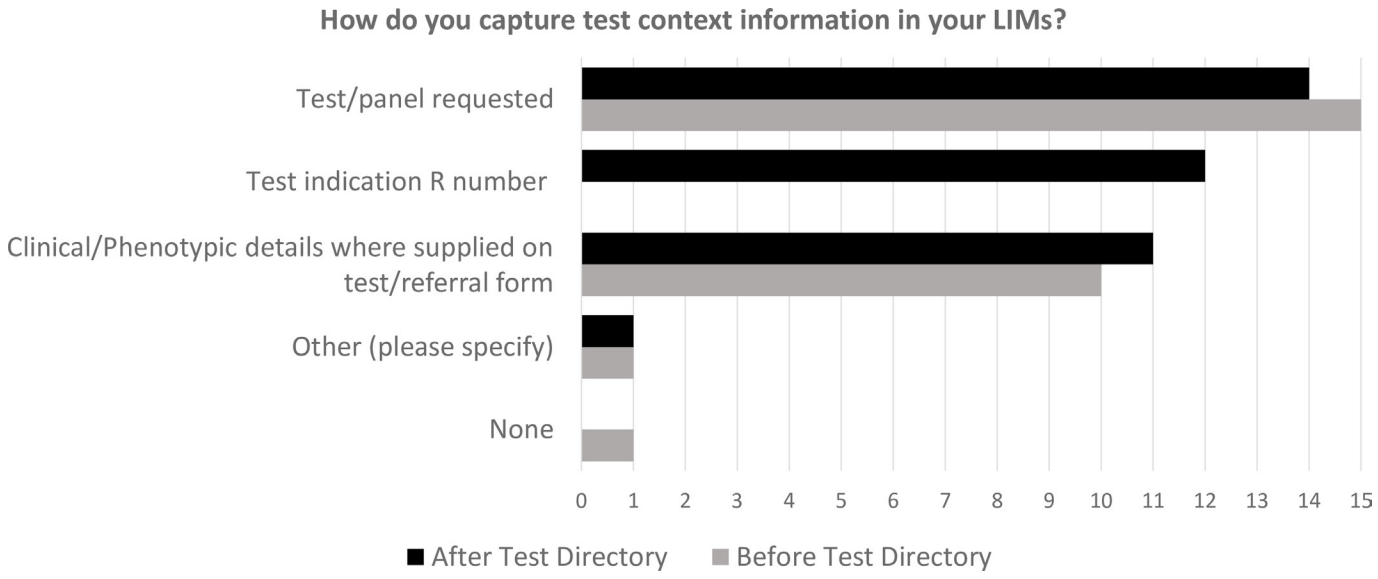


Figure 2 Comparison of testing context information captured by each laboratory before and after the National Test Directory was implemented. LIMS, Laboratory Information Management System.

non-automated approaches which generate multiple intermediary files, potentially introducing opportunity for human error with manual or copy–paste transcription of variants.

The existing laboratory processes were designed to ensure relevant results are included accurately on the clinical diagnostic report for each patient. They were not designed with large-scale national data amalgamation in mind. Consequently, the laboratories vary in whether they store rare variants (other than reported pathogenic and likely pathogenic variants) in their LIMS. Therefore, the majority of variants (especially rare VUS) may not universally be stored in the LIMS alongside patient identifiers and patient information. Instead, these variants might solely reside in stored VCF files which typically do not contain patient identifiers. Only with decryption using Sample and Run identifiers might these variant data be reunited with patient identifiers/phenotype information.

Our survey indicates that transformation of highly structured VCF variant data into the clinical information stored in an LIMS typically means that this final variant information (1) is for only a small subset of detected variants, (2) is in a variously less-structured format than a VCF (eg, free text) and (3) has potentially been subject to human transcription errors (although this has been shown to be infrequent).[18] These three factors have the following implications: (1) they impact the ability to search for variants (when they are encountered by a laboratory again, or to identify patients with a certain variant), (2) the varied structures make data amalgamation challenging and (3) variable capture in local LIMS means the true frequency of rare VUSs in amalgamated data may be underestimated.

From the survey responses, the distilling of voluminous technically unverified variant data from a VCF across into clinically classified variants in an LIMS/EHR would appear to be a universal challenge, and few laboratories appear to have informatic systems within which this variant workflow can be comprehensively automated, executed and documented.

Variant classification

The workflow reported for performing and capturing variant evaluation/classification was also highly heterogeneous and often labour-intensive. The reported workflow often necessitates not only accessing each of the required data sources manually on a variant-by-variant basis, but also then manual population of the findings in per-variant documents or files. Retrieval of per-variant files stored on local drives will also be subject to the inconsistencies of file-naming practices (especially where there is high staff turnover). Systems by which multiple per-patient-per-variant files for a given variant are maintained will also be potentially vulnerable to temporal inconsistencies.

Phenotypical context and gene panels

The survey also revealed that there is substantial variation in the extent of capture of phenotype or testing context (panel name or constituent genes) against the detected variant(s) in the LIMS. The personal and/or familial phenotype and the context of the gene panel requested are highly informative regarding the likely clinical significance of a detected rare variant. For example, the new evidence towards pathogenicity is much greater for a rare *TP53* variant detected in a context with high phenotypical specificity (for example, testing of only one gene, *TP53*, in an adolescent with rhabdomyosarcoma and a family history of young-onset cancers) compared with a context with low phenotypical specificity (for example, testing of a 40-gene panel in a 68-year-old woman with breast cancer and no family history).[19 20 21] Without data on the tested gene panel and phenotype, the two instances of the variant would not be distinguishable. The value of variant data for longer-term advancement of risk estimation and variant classification locally, nationally or globally is greatly diminished if the concomitant phenotype data and testing context are not captured. Use of clinical indications (R-codes) from the National Test Directory will give some indication of testing context. However, the associated eligibility requirements for each R-code are broad and both these and the gene sets for a given code have changed considerably since inception of the National Test Directory in 2018 (and are likely to continue to change).

Data sharing and national submission

There has been substantial global focus on improving national and international genomic data-sharing, with high-profile endeavours from the Global Alliance for Genomics and Health

(GA4GH), such as the GA4GH 'Beacon' Project and the Matchmaker Exchange designed to enable cross-identification across the world of other instances of a given variant.[22–25] In addition, the LOVD and ClinVar resources provide international portals by which clinical diagnostic laboratories can share observations and classifications of clinically observed variants.[7 26 27] While submission to ClinVar is undertaken by CanVIG-UK for their consensus variant classifications, participation by individual laboratories in these international endeavours is potentially limited by the LIMS data architectures.[15]

The UK national CSG genomic data amalgamation is a world first—a reflection of the challenging logistics and the complexity of related governance structures. However, this has been achieved largely despite of rather than because of the design of the laboratory data systems and workflow. The utility of sharing of the variant data between laboratories is complemented by the national interlaboratory discussion forum afforded by the CanVar-UK platform, which provides opportunity for discussing how the variant frequencies are applied for variant interpretation, for sharing other de-identified clinical information (eg, tumour testing) and for discussing consensus variant classifications.

Limitations of the survey

The laboratory workflow and practices are as reported from a single clinical scientist survey response for each laboratory. Although we contacted multiple clinical scientists per laboratory to facilitate consensus response, and we checked directly with respondents where there were any inconsistencies within the information supplied, there is opportunity for misinterpretation of questions or supply of erroneous responses. In addition, there is the possibility that questions or multiple choice options may be interpreted differently by different laboratories, or of difference in opinion between those working within the same laboratory. By piloting, we sought in our survey design to present for each question the most clear and relevant enumerations; nevertheless, additional information was provided as free text with many responses. This we have sought to share (where not identifiable).

RECOMMENDATIONS

Workflow redesign in a laboratory can be a sizeable undertaking, especially if this involves tighter integration of upstream sequencer VCF workflow into the LIMS and/or redesign of LIMS data storage/outputting. From our survey findings, we identified the following as priority recommendations to be considered in any laboratory data workflow redesign:

1. Transfer of variant data across the workflow from the VCF to the LIMS/EHR should be automated (informatic). *This will reduce the risk of errors in variant names (introduced by manual and/or copy-and-paste transcription) and will ensure consistent variant nomenclature.*
2. Variants (including CNVs) should be stored in a structured format, with separate fields for genome build, transcript, gene, cDNA and p. protein annotation (to standardised Human Genome Variation Society (HGVS) recommendations[28]). *This will support both individual variant querying and broader amalgamation/analyses.*
3. All rare variants (preferably all variants) should be stored against the patient record, either within the primary LIMS or within a linked data system which contains meaningful patient identifiers. *This will ensure that relevant historical patients can be identified if a variant is reclassified or prospectively introduced for clinical testing.*
4. All rare variants should be stored against details of the clinical indication for testing/gene set analysed and (where possible) details of patient phenotype. *This will ensure that the variant data are most informative for variant interpretation.*
5. Local variant classifications (and the contributory evidence) should be stored as unique entries within a single structured data system (or within the LIMS) rather than as individual files. *This will improve data retrieval and reduce the occurrence of multiple discordant entries of the same variant. Where possible, and avoiding the release of patient-identifiable information, local variant classifications should be shared nationally (eg, CanVar-UK) and internationally (eg, ClinVar).*

CONCLUSIONS

The reconfiguration of the NHS genomics laboratories, development of a National Test Directory, widening of testing indications and expansion of whole-genome sequencing offer potential for the UK to be a major contributor of national data to variant interpretation initiatives. However, this survey has revealed the workflow required for data amalgamation is frequently labour-intensive and potentially culminates in storage within the LIMS of variant data that may be incomplete, poorly structured, may incorporate rare manual transcription errors and lack corresponding phenotype data. Larger gene panels are being added to the National Test Directory and the volume of genetic testing in the NHS is increasing. With this volume increase, there is potential for current local and national data amalgamation processes to become compromised and for clinical diagnostic delivery to become increasingly burdensome, a particular concern given the current limited availability of trained clinical scientists.

The laboratory workflow taking VCFs into clinically processed variant data is inherently complex, with unavoidable requirement for human variant review. There is further complexity where integrated pathology reporting necessitates downstream integration of molecular genetics findings with results from other pathology disciplines (for example, in cancer reporting). Redesign of this workflow and LIMS architectures is therefore far from straightforward and will only be successful where consultative design involves substantial dedicated time from experienced clinical scientists and clinical bioinformaticians (who must be released from service delivery). However, considered investment in the redesign of this workflow will be of high value in empowering laboratory scientists for the proposed expansions in genomic analyses. It will also position the UK clinical genetic testing community to make the best use of the data generated to contribute to national and international initiatives in data amalgamation, thus supporting improved variation interpretation for our patients.

Author affiliations
[1]Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, UK
[2]Department of Clinical Genetics, St George's University Hospitals NHS Foundation Trust, London, UK
[3]Sheffield Diagnostic Genetics Service, NEY Genomic Laboratory Hub, Sheffield Children's NHS Foundation Trust, Sheffield, UK
[4]East Anglian Medical Genetics Service, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[5]Wessex Regional Genetics Laboratory, University Hospital Southampton NHS Foundation Trust, Southampton, UK
[6]Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Leeds, UK
[7]Manchester Centre for Genomic Medicine and NW Laboratory Genetics Hub, Manchester University Hospitals NHS Foundation Trust, Manchester, UK
[8]Genomics and Molecular Medicine Service, Nottingham University Hospitals NHS Trust, Nottingham, UK
[9]Department of Clinical Genetics, CHI at Crumlin, Dublin, Ireland
[10]Cancer Genetics Unit, The Royal Marsden NHS Foundation Trust, London, UK

[11]West Midlands, Oxford and Wessex Genomic Laboratory Hub, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[12]Department of Molecular Genetics, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK
[13]East Midlands and East of England Genomics Laboratory, Nottingham University Hospitals NHS Trust, Nottingham, UK
[14]North Thames Genomic Laboratory Hub, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK
[15]Central and South Genomic Laboratory Hub, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK
[16]South East Genomics Laboratory Hub, Guy's Hospital, London, UK
[17]North West Genomic Laboratory Hub, Manchester Centre for Genomic Medicine, Manchester, UK
[18]Centre for Molecular Pathology, Institute of Cancer Research Sutton, Sutton, UK
[19]Department of Molecular Diagnostics, The Royal Marsden NHS Foundation Trust, London, UK
[20]Institute of Medical Genetics, Cardiff and Vale University Health Board, University Hospital of Wales, Cardiff, UK
[21]South West Genomic Laboratory Hub, University Hospitals Bristol NHS Foundation Trust, Bristol, UK
[22]North East and Yorkshire Genomic Laboratory Hub, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
[23]Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK
[24]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, UK
[25]Department of Medical Genetics, National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK
[26]NHS England, National Disease Registration Service, London, UK
[27]Health Data Insight CIC, Cambridge, UK
[28]Department of Public Health and Primary Care, University of Cambridge Centre for Cancer Genetic Epidemiology, Cambridge, UK
[29]Nuffield Department of Population Health, University of Oxford, Oxford, UK

Twitter George J Burghel @BurghelG and Helen Hanson @Helen_Hanson1

Contributors CT, LL, SA and AG designed the survey questions with input from MD, JD, AC, RR, GJB, HH, FMR, SG and SH. MD, JD, AC and RR piloted the survey. MD, JD, AC, RR, GJB, TB, CBo, KB, CBr, SB, JCDRJ, LH, VS, SM, MO, SP-S, KS, JT, MV-P, EW and MY provided responses to the full survey on behalf of their laboratories. SA collated responses and generated figures for presentation. CT, LL, SA and AG reviewed survey responses and prepared draft recommendations. MD, JD, AC, RR, GJB, HH, JF, TM and TPM (CStAG members) finalised recommendations. SA and CT drafted the manuscript with detailed input from LL, MD, ACA and EM. All authors contributed to review and editing of the manuscript. CT, DME and MT obtained funding. SA and BT provided project management. SA and LL contributed equally to this paper. CT accepts official responsibility for the overall integrity of the manuscript as guarantor.

Funding SA, LL, AG, HH and BT are supported by CRUK Catalyst Award CanGene-CanVar (C61296/A27223). MT is supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312).

Competing interests MD and TM have received honoraria from AstraZeneca and MSD for contributions as expert assessors in the GenQA/EMQN GTACT schemes: Ensuring accurate classification of BRCA1, BRCA2 and other HRR gene variants. ACA is the creator of BOADICEA, licensed by Cambridge Enterprise, and receives royalties from Cambridge University. TPM is a council member for the UK Cancer Genetics Group, and has received honoraria from AstraZeneca and Novartis, and consulting fees from Roche as an Expert Advisor for the National Molecular Tumour Board (Ireland). DME was co-applicant on an AstraZeneca Research Grant (2021–2023), is on the University of Southampton Executive Board and is the Non-Executive Director of UHS NHS Foundation Trust. CT has received honoraria from AstraZeneca and MSD for educational activities and scientific boards, which are donated in full to charity.

Patient consent for publication Not required.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

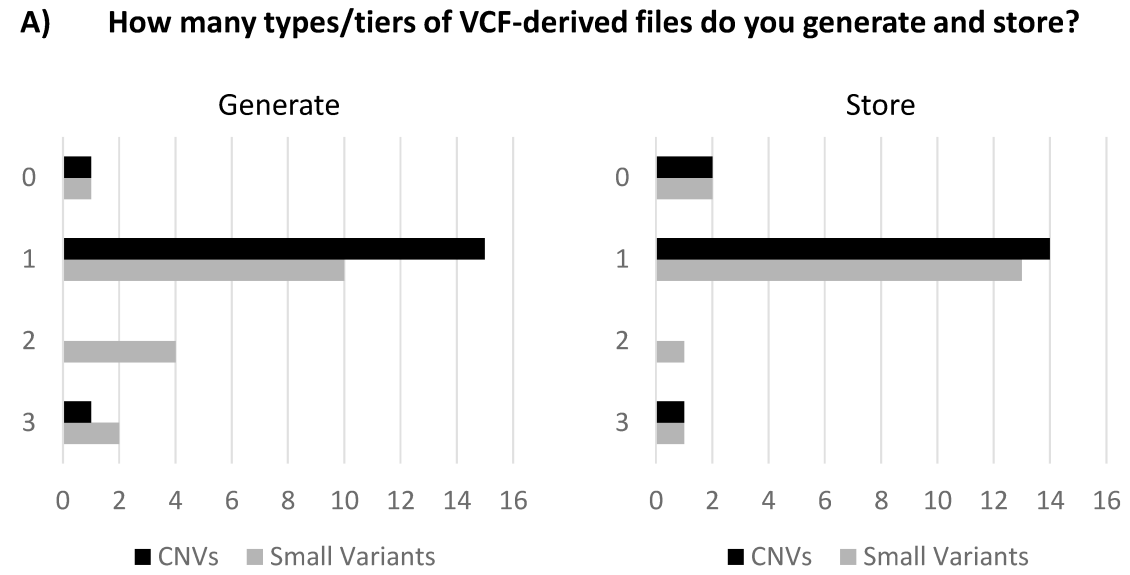
Data availability statement Data are available in a public, open access repository. Results from the GLH-NDRS component of the full survey are available in the Zenodo repository at https://doi.org/10.5281/zenodo.8340397.

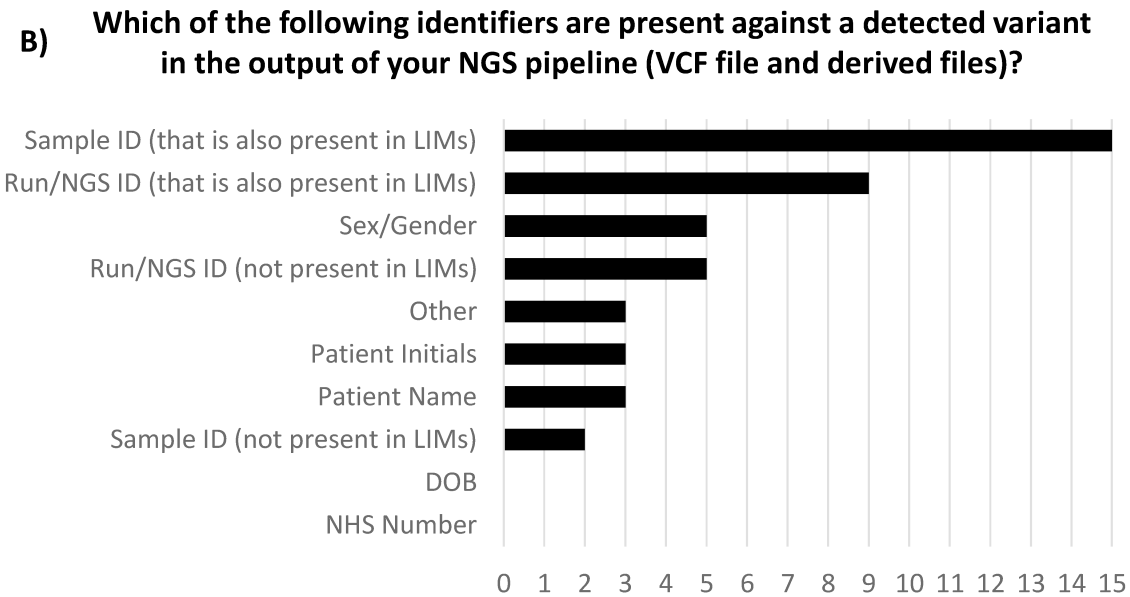
ORCID iDs
Sophie Allen http://orcid.org/0000-0003-4928-2240
Alice Garrett http://orcid.org/0000-0001-8942-283X
Bethany Torr http://orcid.org/0000-0003-3487-9749
George J Burghel http://orcid.org/0000-0001-9360-8194
Helen Hanson http://orcid.org/0000-0002-3303-8713
Diana M Eccles http://orcid.org/0000-0002-9935-3169
Antonis C Antoniou http://orcid.org/0000-0001-9223-3116
Clare Turnbull http://orcid.org/0000-0002-3797-7398

REFERENCES

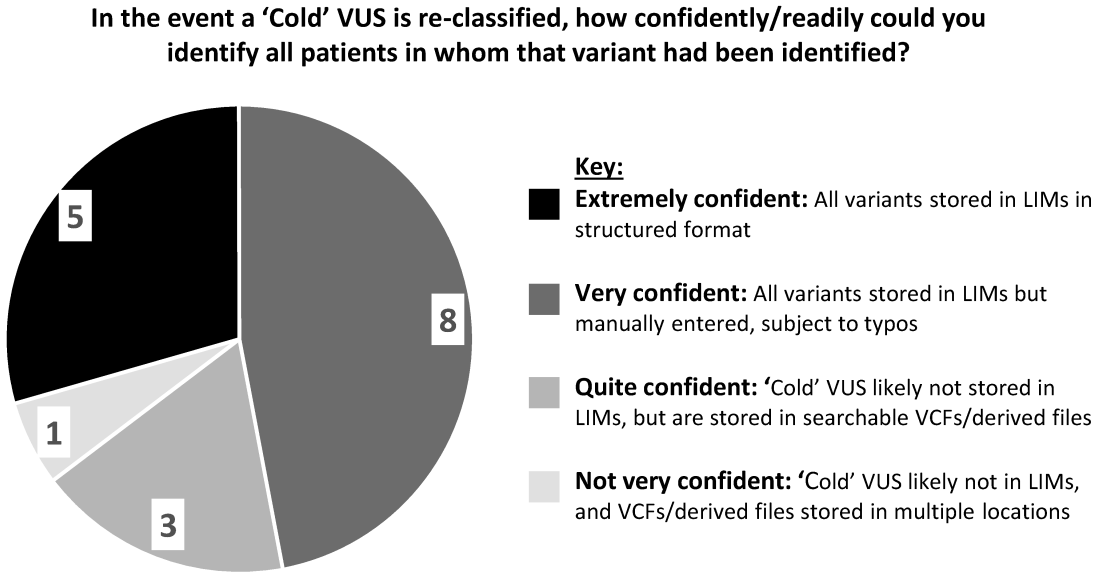
1 Annual report of the chief medical officer 2016: generation genome. Office of the Chief Medical Officer, Department of Health; 2017.
2 Accelerating genomic medicine in the NHS. A strategy for embedding genomics in the NHS over the next 5 years. NHS England, 2022.
3 Landmark strategy launched to cement UK's position as global leader in genomics. Department of Health and Social Care, 2020.
4 NHS-England. National genomic test directory. 2023.
5 Karczewski KJ, Francioli LC, Tiao G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Genomics* [Preprint].
6 Backman JD, Li AH, Marcketta A, *et al*. Exome sequencing and analysis of 454,787 UK biobank participants. *Nature* 2021;599:628–34.
7 Landrum MJ, Lee JM, Riley GR, *et al*. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;42:D980–5.
8 Stenson PD, Mort M, Ball EV, *et al*. The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;136:665–77.
9 Loong L, Cubuk C, Choi S, *et al*. Quantifying prediction of pathogenicity for within-codon concordance (PM5) using 7541 functional classifications of BRCA1 and MSH2 missense variants. *Genet Med* 2022;24:552–63.
10 Cubuk C, Garrett A, Choi S, *et al*. Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genet Med* 2021;23:2096–104.
11 Richards S, Aziz N, Bale S, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–24.
12 Tavtigian SV, Harrison SM, Boucher KM, *et al*. Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. *Hum Mutat* 2020;41:1734–7.
13 Tavtigian SV, Greenblatt MS, Harrison SM, *et al*. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet Med* 2018;20:1054–60.
14 Garrett A, Durkie M, Callaway A, *et al*. Combining evidence for and against pathogenicity for variants in cancer susceptibility genes: CanVIG-UK consensus recommendations. *J Med Genet* 2021;58:297–304.
15 Garrett A, Callaway A, Durkie M, *et al*. Cancer variant interpretation group UK (CanVIG-UK): an exemplar national subspecialty multidisciplinary network. *J Med Genet* 2020;57:829–34.
16 Ellard S, Baple EL, Owens M, *et al*. ACGS best practice guidelines for variant classification 2018. Association for Clinical Genetics Science (ACGS), 2018.
17 Loong L, Garrett A, Allen S, *et al*. Reclassification of clinically-detected sequence variants: framework for genetic clinicians and clinical scientists by CanVIG-UK (cancer variant interpretation group UK). *Genet Med* 2022;24:1867–77.
18 Loong L, Huntley C, McRonald F, *et al*. Germline mismatch repair (MMR) gene analyses from English NHS regional molecular genomics laboratories 1996-2020: development of a national resource of patient-level genomics laboratory records. *J Med Genet* 2023;60:669–78.
19 Allen S, Loong L, Garrett A, *et al*. GLH-NDRS centralised data submission survey results. *Zenodo* 2023;13.
20 Fortuno C, Cipponi A, Ballinger ML, *et al*. A quantitative model to predict pathogenicity of missense variants in the TP53 gene. *Hum Mutat* 2019;40:788–800.
21 Fortuno C, Lee K, Olivier M, *et al*. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum Mutat* 2021;42:223–36.
22 Philippakis AA, Azzariti DR, Beltran S, *et al*. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat* 2015;36:915–21.
23 Buske OJ, Schiettecatte F, Hutton B, *et al*. The matchmaker exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles. *Hum Mutat* 2015;36:922–7.
24 Chatzimichali EA, Brent S, Hutton B, *et al*. Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Hum Mutat* 2015;36:941–9.

25 Kirkpatrick BE, Riggs ER, Azzariti DR, *et al*. Genomeconnect: matchmaking between patients, clinical laboratories, and researchers to improve genomic knowledge. *Hum Mutat* 2015;36:974–8.
26 Fokkema IFAC, Taschner PEM, Schaafsma GCP, *et al*. LOVD V.2.0: the next generation in gene variant databases. *Hum Mutat* 2011;32:557–63.
27 Landrum MJ, Lee JM, Benson M, *et al*. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;46:D1062–7.
28 den Dunnen JT, Dalgleish R, Maglott DR, *et al*. HGVS recommendations for the description of sequence variants: 2016 update. *Human Mutation* 2016;37:564–9.





Supplementary Figure 1: Additional questions pertaining to variant workflows. A: Approaches to storing intermediary files generated as part of the workflow. B: The range of patient identifiers available within the workflow upstream of entry in the LIMs.



Supplementary Figure 2: Confidence of responding laboratories in ability to reliably retrieve 'cold' VUS variant information in the event of reclassification.

SUPPLEMENTARY METHODS

Initial creation of survey (Pilot Phase 1)

The initial survey questions and format were iterated and reviewed through discussion with the CanVIG-UK Steering and Advisory Group (CStAG), an expert panel supporting Cancer Variant Interpretation Group UK (CanVIG-UK) activities, and consisting of 8 senior clinical scientists and 3 consultant clinical geneticists from across the Genomic Laboratory Hubs.

After initial discussions, the survey was piloted by 4 CStAG members to assess the survey clarity and content. Following integration of comments during the pilot, a post-pilot version was confirmed by consensus, which contained 29 multiple choice questions split across 6 categories in total. Broadly, these categories can be summarised as 'Respondent Details', 'NGS Workflow', 'Variant Classification', 'Phenotypic Context', 'LIMs Status' and 'Central Submission'. These questions were added to an electronic survey format and user-tested for functionality by the CanVIG-UK leadership team prior to questionnaire release.

Participant Identification

Participants were identified from a list of persons previously involved with cancer susceptibility gene data submissions to Public Health England (PHE) or NHS Digital (NHSD), which was shared from the PHE data team with the CanVIG-UK leadership team. These persons were contacted to confirm that they were previously involved in one or more of the below activities:

- Variant Data Restructuring and Submission
- Design of Laboratory Information Management Systems (LIMs) or data flow between NGS outputs and LIMs
- Design of database structure for capture and storage of variant classifications

Involvement in these activities was deemed baseline for sufficient knowledge in completing the survey. The CanVIG-UK leadership also contacted the full CanVIG-UK membership to identify further individuals meeting these criteria, and in total 51 individuals were contacted to complete the survey. Individuals at the same laboratory were asked to co-ordinate and submit a single submission from their lab rather than submit individually in order to generate an overall answer to represent their lab workflow and prevent duplicate responses.

Web Survey (Pilot Phase 2)

The survey link was sent out to the selected pool of 51 individuals using SurveyMonkey (http://www.surveymonkey.com) and was live between February 2022 and April 2022. In the invitation email, participants were provided with details of the purpose behind the survey, how all their responses would be utilised as part of feedback to NHS England (with confirmation that individual participant or laboratory identifiers would not be included in any presentation of the data, including through any additional comments made). Incentives were not offered for participation. The investigators for the survey were confirmed as Prof Clare Turnbull, the CanVIG-UK team, and the PHE team, and participants were provided an approximate length of time the survey would take (around 10 minutes). Personal details (name, role, email address) were collected for the purposes of re-contacting respondents in order to deliver feedback in the first instance, to identify unique visitors or duplicate submissions, and to clarify responses given. This information was downloaded from SurveyMonkey and held for the duration of study analysis on a secure drive in a folder with access provided only to the investigators.

Questions were not randomised, and were split across 7 pages, one for each category plus an additional page for centralised data submissions. Number of questions per page varied as a result, between 1 and 11. All questions had a comment box and an 'Other' option. Most questions were required, however one question was reliant on specific selection from the

previous answer and was therefore optional to answer. All questions provided an 'other' or 'unsure' option. Respondents were able to save their progress and return to the survey to change answers or complete the survey at any point before the survey closed. If the same individual submitted multiple times, the most recent responses were retrieved and all other responses removed. Completion was reviewed after the questionnaires were submitted.

The web survey received a total of 27 responses, 18 of which were complete and from unique participants (17 responses from across all 16 submitting labs in England, and 1 response from the All Wales Medical Genomics Service). Following survey closure, a feedback meeting was held in May 2022 to discuss findings from this version of the survey and gather detailed feedback, which involved CanVIG-UK leadership, those involved with informatics at several Genomic Laboratory Hubs, and the survey respondents. This meeting highlighted that updates to several laboratory data management systems were ongoing, as well as responses which did not fit the multiple choices provided.

Email Survey (Final responses)

The survey was updated following the Web Survey to incorporate some common answers provided as 'Other' (for example, the addition of 'Patient Initials' and 'Patient Sex' as workflow identifiers in Question 1). One additional question was also included relating to NHS centralised data submission of variants of uncertain significance, bringing the total questions asked to 30.

To allow time for implementation of management system updates, the final iteration of the survey was sent out one year after the web survey, and was open between April and June 2023. To assist those responding, the final survey was sent to all 18 previous respondents as a pre-filled Microsoft Word document containing their previous answers. Any new questions or options were highlighted. Respondents were asked to review their answers in the context of new LIMs updates, and update or add to any comments made from the previous year. The invitation to complete this survey, in addition to the consent provided for the Web Survey, also described the investigators intent to publish responses from this final survey iteration.

By June 2023, 17 of the previous respondents had replied with completed questionnaires and had updated their answers as appropriate (94.4% response rate). 1 respondent did not reply as a member of their laboratory had already provided responses for the email survey. Responses were received as Microsoft Word documents, which were manually collated into a 'master' Microsoft Excel database for review.

Supplementary Table 1: Current CStAG Membership and representation across the UK GLHs

Genomic Laboratory Hub (GLH)	Number of CStAG Representatives
Central and South GLH	1
East GLH	2
North West GLH	1
North Thames GLH	2
South East GLH	2
South West GLH	0
North East and Yorkshire GLH	2
N/A - Ireland (Dublin)	1

Supplementary Table 2: Full response breakdown for the 18 questions pertaining to laboratory workflows. To preserve anonymity, free-text comments are not provided.

Question		Options	Number of Responses
1	Which of the following identifiers are present against a detected variant in the output of your NGS pipeline (VCF file and derived files)?	Hospital Trust ID	0
		NHS Number	0
		DOB	0
		Sample ID (not present in LIMs)	2
		Patient Name	3
		Patient Initials	3
		Run/NGS ID (not present in LIMs)	5
		Sex/Gender	5
		Run/NGS ID (that is also present in LIMs)	9
		Sample ID (that is also present in LIMs)	15
		Other	3
2	How are the QC'ed small variants (eg SNVs and indels) from your NGS pipeline outputs (VCF or VCF-derived file) CURRENTLY transferred to your LIMs?	Saved in separate location	2
		Attached as separate file	1
		Manual copy-and-paste	4
		Automatic and invisible (no action required from scientist)	0
		Manually typed out	8
		Predesigned 'push-button' transfer	2
		Other	0
3	For small variants: How many types/tiers of VCF-derived files do you generate?	>3	0
		3	2
		2	4
		1	10
		0	1
4	For small variants: How many types/tiers of VCF-derived files do you store?	>3	0
		3	1
		2	1
		1	13
		0	2
5	How are the QC'ed CNVs (eg exon-level deletion) from your pipeline outputs (from VCF, VCF-derived or MLPA workflow output) CURRENTLY transferred to your LIMs?	Saved in separate location	2
		Attached as separate file	1
		Manual copy-and-paste	4
		Automatic and invisible (no action required from scientist)	0
		Manually typed out	9
		Predesigned 'push-button' transfer	1
		Other	0
6	For CNVs: How many types/tiers of VCF-derived	>3	0

	files do you generate?	3	1
		2	0
		1	15
		0	1
7	For CNVs: How many types/tiers of VCF-derived files do you store?	>3	0
		3	1
		2	0
		1	14
		0	2
8	What best describes the way CSG small variants are CURRENTLY stored on your LIMs?	In a single free text/unstructured field (which also includes additional clinical report wording)	5
		In a single free text/unstructured field (just contains variant name)	5
		As a single field of strict formal hgvs notation, ie: gene + transcript + genomic location + coding (c.) change + protein (p.) change	2
		As separate fields for gene/transcript/genomic location/coding (c.) change/protein (p.) change	4
		Stored outside of a LIMs system	1
9	What best describes the way CSG CNVs are CURRENTLY stored on your LIMs?	In a single free text/unstructured field (which also includes additional clinical report wording)	4
		In a single free text/unstructured field (just contains variant name)	9
		As a single field of strict formal hgvs notation, ie: gene + transcript + genomic location + coding (c.) change + protein (p.) change	1
		As separate fields for gene/transcript/genomic location/coding (c.) change/protein (p.) change	2
		Stored outside of a LIMs system	1
10	Which CSG variants are CURRENTLY stored in your LIMs?	All variants detected (rare and common)	3
		All rare variants detected (all class 3 and above, may also include rare class 1/2)	2
		All rare variants detected (all class 3 and above, no rare class 1/2)	2
		Most rare variants of relevance (all 4/5 and most interesting VUS)	5
		Only variants included in the clinical report	4
		No variants stored in the LIMs / No LIMs system	1
11	In the event of a variant re-classification (for example, up-classification of a previous cold class 3), how confidently/readily could you identify all patients in whom that variant had been identified (since inception of your current system)? (if not using a LIMs system for this process, please specify which system would instead be used to identify all patients)	5: extremely: all variants are stored in LIMs/current storage system in accurate structured format	5
		4: very: all variants are stored in LIMs/current storage system but have been manually entered, so subject to typos	8
		3: quite: cold class 3 variant likely not in LIMs/current storage system. Would require a comprehensive search of various historic bioinformatics systems/VCFs/derived files but these are stored so as to be easily searchable	3
		2: not very: cold class 3 variant likely not in LIMs/current storage system. Would require a comprehensive search of various historic bioinformatics systems/VCFs/derived files which are stored in multiple locations	1
		1: poorly: cold class 3 variant likely not in LIMs/current storage system. Would require a comprehensive search of various historic bioinformatics systems/VCFs/derived files which are not readily accessible	0
12	From what type of interface are the variants requiring interpretation viewed?	Within a bioinformatics processing system/dedicated in-house variant system	8
		In a spreadsheet (eg VCF, VCF-derived file.)	5
		Other (please specify)	4
13	Within the interface from which you view variants requiring interpretation, which description is most accurate?	Most/many of the relevant data sources have been pre-imported	1
		There are variant-specific hyperlinks to most/many of the relevant data sources.	7
		No/minimal annotations (eg only population frequencies). Accessing of relevant data sources (Alamut, CanVar-UK, ClinVar, literature) requires manual interrogation (variant name is typed/pasted in).	9

		Other	0
14	Following interpretation of a CSG variant, where do you store your updated detailed findings/pathogenicity classification (eg scoring on ACMG sub-elements)?	Dedicated in-house departmental variant datasystem	5
		Individual per-VARIANT files (excel, word, other document). Updated for recurrent viewings of variant.	5
		Individual per-VARIANT files (excel, word, other document). New file time a variant is encountered.	4
		Individual per-PATIENT EPISODE files (excel, word, other document) (may contain multiple variants)	4
		Commercial platform or software (eg Congenica, Alamut)	4
		LIMs (against specific patient)	3
		Individual per-GENE files (excel, word, other document).	1
		Individual per-DISEASE files (excel, word, other document).	1
		LIMs (against variant)	0
		Dedicated single departmental excel/spreadsheet	0
		Variant classifications are only documented on the patient report and not elsewhere stored	0
		Other (please specify)	0
15	How did you capture test context information in your laboratory patient-level datastorage system (LIMs) prior to the GMS test directory?	Clinical details/Phenotype information from test request or referral form	10
		Test/panel requested	15
		None	1
		Other (please specify)	1
16	How do you capture test context information in your laboratory patient-level datastorage system (LIMs) since the GMS test directory implemented?	Clinical details/Phenotype information from test request form or referral form	11
		Test/panel requested	14
		Test indication R number (since publication of the National Genomic Test Directory)	12
		None	0
		Other (please specify)	1
17	Following implementation of the GMS test directory, please estimate what % of CSG test requests now contain some clinical details/phenotype information (in addition to test requested/R number).	0-24%	0
		25-49%	2
		50-74%	8
		75-100%	6
18	When a single gene/small gene set is reported from a larger panel/exome, how is the gene set which is analysed captured in your LIMs?	Genes analysed are listed by name in LIMs against patient (free text entry eg list with commas)	4
		Name of subpanel(s) listed in LIMs against patient (structured entry)	4
		Genes analysed are listed by name in LIMs against patient (structured entry eg selected from a drop-down list, imported from a separate portal)	2
		Name of subpanel(s) listed in LIMs against patient (free text entry)	2
		Mixture of genes and subpanels (free text entry)	2
		Genes/panels analysed not documented in LIMs; requires reference to other files within NGS workflow	1
		Exported as a list from commercial platform and attached to LIMs record	1
		Other (please specify)	1