Genomics England - Main Programme Data Release Note

Table of Contents

1. Purpose

This document provides a description of the Main Programme Data Release v6 dated 28th Feb 2019. Each progressive release incorporates new content, enhances existing content, and enables more effective use of the existing and new data.

This data is manifested within the current version of Genomics England Research Environment, accessed via the Inuvika virtual desktop interface and subject to all Genomics England data protection and privacy principles.

2. Release Overview

Release 6 provides clinical data for 94,285 participants, and 91,271 genomes from 78,165 of these participants. Of these genomes, 69,172 are rare disease genomes (from 67,874 participants) and 22,091 are cancer genomes (from 10,287 participants)[1].

Participants	
Rare Disease Participants	73,810
Cancer Participants	20,475
Participants Total	**94,285**

Genomes		
	Number of genomes	Participants
Cancer Germline	10,414	10,077
Cancer Tumour	11,473	10,276
Cancer Total	22,091	10,287
Rare Disease	69,172	67,874
Genomes Total	**91,271**	**78,165**

- Genomic data are manifested in file shares.
- Clinical data and secondary health data ("medical history") are manifested in LabKey.

This release seeks to include all variables that contain (or may contain in future) meaningful data whilst not compromising participant privacy.

Some genomic data are currently aligned against the reference genome version GRCh37 and some against version GRCh38. The alignments were also made using different versions of Illumina's alignment pipelines V2 and V4, reflecting the versions that were applicable at the time of sequencing. *All new genomic data added in the current data release (since July 2018) is aligned against the reference genome version GRCh38, using alignment pipelines V4*. The versions for each

[1] This excludes 77 TracerX genomes from 40 participants (refer to 5.4 for further information).

genome are identified in the Sequencing Report table. We intend to provide consistently realigned and recalled versions of all our genomes in the future.

3. Audience

The intended audience for this document is researchers that have access to the Genomics England Research Environment. This does not include taught students on the MSc Genomic Medicine, who have access to a small subset of Main Programme data.

4. Identifying this data release

The clinical data for this data release, and the paths to the applicable genome files, are found in the following LabKey folder:

main-programme /main-programme_ v6_2019-02-28

Subsequent releases will be identified by an incremental increase in the version number and the date of data release.

The main genome sequence files are found in the User's Inuvika Home Drive, organised by date. Some of the included genomic data produced by the Genomics England Bioinformatics pipeline (such as rare disease tiering, structural and copy-number variant reports for cancer genomes) are found in the Genomics England Data Resources (see Section 6.5).

5. Scope

5.1. In scope

Data that are in scope for this release:

- Cancer and rare disease data for the main programme participants with current consent. These data include:
    - Genomic data for participants when available
    - WGS family-based quality control for rare disease, reporting sex checks and pedigree checks
    - Outputs of the Genomics England Bioinformatics rare diseases interpretation pipeline
        - Tiering data – rare disease
        - GMC outcome data ("exit questionnaire data") – rare disease
        - Aggregated Illumina gVCF for germline genomes
        - Interpretation request data for rare disease up until 31/10/2018
    - Outputs of the Genomics England Bioinformatics cancer interpretation pipeline
        - Gold standard cancer genomes which have been through interpretation and passed quality checks
        - Tumour signature and mutational burden data - cancer
        - Annotation and tiering of small variant – cancer
            - Tiering, structural and copy number variant report

        - Cancer PCA stats
    - Primary clinical data, including formal pedigree data on rare disease participants where it is available; and
    - Secondary datasets (medical history) including:
        - Hospital Episode Statistics (HES), including HES Accident and Emergency, HES Admitted Patient Care, and HES Outpatient Care.
        - Diagnostic Imaging Dataset (DID).
        - Patient Reported Outcome Measures (PROMs).
        - Mental Health Minimum Data Set (MHMDS).
        - Office for National Statistics (ONS).
        - Systemic Anti-Cancer Therapy Data Set (SACT).

5.2. Out of scope

Data that are out of scope for this release:

- Clinical and genomic data for participants that have withdrawn from the 100,000 Genomes Project.
- Participant data from the pilot phases of the project (i.e. not main programme).
- Sources of secondary data other than HES, DID, PROMs, MHMDS, ONS and SACT.

5.3. Quality Notes

- BAM and VCF genomic data files are as they have been delivered to us by our sequencing provider. These have all passed an initial QC check based on sequencing quality and coverage. They have, however, not all undergone our full in-house genetic checks and we therefore cannot guarantee against genetic versus reported sex and family relationship discrepancies.
- For Rare Disease genomes, it should be noted that all tiered genomes have passed through Genomics England in-house QCs and that all tiered genomes come from the pool of genomes that have had family checks applied to them, as a first step towards Genomics England tiering.
- For Cancer genomes, it should be noted that all gold standard genomes that have been through Genomics England interpretation and passed quality checks are found in the cancer quick view table cancer_analysis.
- Some rare disease families lack a proband due to the availability of data at the time of release. These families without probands will also lack a diagnosis unless there is a second affected individual in the family. The missing data will be made available in a future release.
- Clinical data and secondary data have been provided as submitted and have undergone limited validation.
- Human Phenotype Ontology (HPO) terms may be missing or incomplete for some participants. This will be updated in future releases.
- Formal pedigree data are only available for a subset of rare disease participants. This will be updated in future releases. Each participant's relationship to their family's proband is

available for all cases; this can be used to determine family relationships instead of formal pedigree data.

- WGS family selection quality checks are provided for rare disease genomes on GRCh38, reporting abnormalities of sex chromosomes, family relatedness, Mendelian inconsistencies and reported vs genetic sex summary checks (only sex checks are unpacked into individual data fields).

5.4. Conditions of Use

- It should be noted that participants identified as TracerX in the field normalised_consent_form in the participant table must not be used by commercial organisations.

6. Data Release Description

The Genomics England data are organised into data views (displayed within LabKey as tables) categorised into Quick View, Common, Rare Disease and Cancer. The Data Dictionary that describes the table structure and provides data definitions for this release can be found here.

6.1. Quick View

Data views that bring together data from several LabKey tables for convenient access.

Name of Table / Data View	Description
rare_disease_analysis	Data for all rare disease participants including: sex, ethnicity, disease recruited for and relationship to proband; latest genome build, QC status of latest genome, path to latest genomes and whether tiering data are available; as well as family selection quality checks for rare disease genomes on GRCh38, reporting abnormalities of the sex chromosomes, family relatedness, Mendelian inconsistencies and reported vs genetic sex summary checks. Please note that only sex checks are unpacked into individual data fields; a final status is shown in the "genetic vs reported results" column.
cancer_analysis	Data for all cancer participants whose genomes have been through Genomics England bioinformatics interpretation and passed quality checks, including: sex, ethnicity, disease recruited for and diagnosis; tumour ID, build of latest genome, QC status of latest genome and path to latest genomes; as well file paths to the genomes. This table includes information derived from laboratory_sample and cancer_participant_tumour. *Tumour Mutational Burden* The table includes the relative proportions of the different mutational signatures demonstrated by the tumour. Analysis of large sequencing datasets (10,952 exomes and 1,048 whole-genomes from 40 distinct tumour types) has allowed patterns of relative contextual frequencies of different SNVs to be grouped into specific mutational signatures. Using mathematical methods (decomposition by non-negative least squares) the contribution of each of these signatures to the overall mutation

	burden observed in a tumour can be derived. Further details of the 30 different mutational signatures used for this analysis, their prevalence in different tumour types and proposed aetiology can be found at the Sanger Institute Website. *Cancer PCA QC Statistics* The cancer analysis pipeline employs a sequencing quality control check which selects several important statistics associated with the sequencing returned by the sequencing provider, and uses them to check whether or not the sample in question is an outlier with respect to previous samples that have been run through the pipeline. It is, in effect, a safety net that can spot issues that have occurred at the tissue collection stage (i.e. at the GMC (Genomic Medicine Centre)) or at the library preparation step (i.e. at the sequencing provider), both of which may impact upon the final genomic analysis returned to the clinician.

6.2. Common

Data views that are common to both the rare disease and the cancer domains. This data pertains to sample handling, genome sequencing, participant data and domain assignment.

Data Relating to Participants

Name of Table / Data View	Description
participant	Data on each individual participant in the 100,000 Genomes Project, e.g. personal information (such as relatives or ethnicity); points of contact with the Project (e.g. handling Genomic Medicine Centre or Trust); and a record of the status of their clinical review.
sequencing_report	For each participant in the 100,000 Genomes Project, this table contains data describing the sequencing of their genome(s) and associated output, as well as the sample type that the sequence is from.
domain_assignment	For each participant in the 100,000 Genomes Project, this table contains: data describing the disease type to which they were recruited; the disease panel applied to their genome; the GeCIP domain to which their genome has been assigned for the purposes of administering the GeCIP publication moratorium; as well as the end date of the GeCIP moratorium associated with their genome(s).
genome_file_paths_and_types	Data that specifies the genomic files and their folder locations for a given a participant.
death_details	Data on participant deaths submitted by GMCs, likely less complete than the data collected by ONS and NHSD
aggregate_gvcf_sample_stats	This table accompanies the aggregated Illumina gVCFs (/gel_data_resources/main_programme/aggregated_illumina_gvcf /GRCH38/20190228/). Individual sample QC data was retrieved from Genomics England openCGA data base. Most metrics are bam

	file statistics provided from Illumina or Genomics England WGS data processing pipeline BERTHA.

Data Relating to Samples

Name of Table / Data View	Description
clinic_sample	Data describing the taking and handling of participant samples at the Genomic Medicine Centres, i.e. in the clinic, as well as the type of samples obtained. Because of the complexities of handling and managing tumour tissues samples in a clinical setting, there are many fields that are cancer-specific.
clinic_sample_quality_check_result	Data describing the quality control of obtaining and handling participant samples at the Genomic Medicine Centres, i.e. in the clinic.
laboratory_sample	Data describing the handling of samples at the biorepository and in preparation for sequencing, as well as the type of sample.

6.3. Rare Diseases

Rare Disease data are presented at the level of Rare Disease families (families of probands), Rare Disease pedigrees and participants. Participants are individuals who have consented to be a part of the project with the expectation that a sample of their DNA will be obtained and their genome sequenced. Pedigree members are extended members of the proband's family, which will include some participants as well as a number of other individuals who will have no contact with the project, have not consented, but for whom a small amount of data are recorded to allow a full picture of the proband's extended family to be gathered.

All Rare Disease tables are prefixed by "Rare_diseases_" at the beginning of the table name.

Data at the Level of Rare Disease Families

Name of Table / Data View	Description
rare_diseases_family	Data describing the families of rare disease probands participating in the 100,000 Genomes Project, making family members participants of the Project. It includes the family group type, the status of the family's pre-interpretation clinical review and the settings that were chosen for the interpretation pipeline at the clinical review.
rare_diseases_pedigree	Data describing the Rare Disease participants, linking pedigrees to probands and their family members.
rare_diseases_pedigree_member	Data describing the Rare Disease pedigree members, similar to the data about each individual participant in the COMMON data view. It includes some additional data, such as the age of onset of predominant clinical features; data on links to other family members; as well as data collected only for Phenotypes.

Data at the Level of Rare Disease Participants

The data presented in these tables provides information on disease progression and pertinent medical history.

Name of Table / Data View	Description
rare_diseases_ participant_disease	Data describing the Rare Disease participants' rare diseases. This is as for pedigree_member_diseases_level_data, with the addition of a date of diagnosis.
rare_diseases_ participant_phenotype	Data describing the Rare Disease participants' phenotypes. For each Rare Disease participant in the 100,000 Genomes Project, there are data about whether a phenotypic abnormality as defined by an HPO term is present and what the HPO term is, as well as the age of onset, the severity of manifestation, the spatial pattern in the body and whether it is progressive or not.
rare_diseases_ gen_measurement	For Rare Disease participants in the 100,000 Genomes Project, this table contains general measurements relevant to the disease, alongside the date that the measurements were taken on.
rare_diseases_ early_childhood_observation	For Rare Disease participants in the 100,000 Genomes Project, this table contains measurements and milestones provided by the GMCs, related to childhood development.
rare_diseases_imaging	For Rare Disease participants in the 100,000 Genomes Project, this table contains various data and measurements from past scans, alongside the date of the scans.
rare_diseases_ invest_genetic	For a proportion of Rare Disease participants in the 100,000 Genomes Project, this table contains information on any genetic tests carried out. Data characterising the genetic investigation is recorded alongside records of the sample tissue source and the type of testing laboratory.
rare_diseases_ invest_genetic_test_result	For a proportion of Rare Disease participants in the 100,000 Genomes Project, this table contains the results of any genetic tests carried out. Following on from the rare_diseases_invest_genetic table, a summary of the results is presented and contextualised by testing method and scope.
rare_diseases_ invest_blood_laboratory_ test_report	For a proportion of Rare Disease participants in the 100,000 Genomes Project, this table contains the results of any blood tests carried out. Over 400 blood values are recorded alongside type and technique of testing and the status of the participating patient in the care pathway.

Data output from the Genomics England interpretation pipeline

panels_applied	For each participant of the 100,000 Genomes Project, this table contains the name and version of the panel(s) that was applied to his or her genome.
tiering	For each participant of the 100,000 Genomes Project who has been through the Genomics England interpretation pipeline, this table contains data describing the variants identified and their pathogenicity.

tiered_variants_frequency	This table contains the frequencies of each tiered variant for every Project participant for whom we provide tiered variants
gmc_exit_questionnaire	Data reporting back from the Genomic Medicine Centres, for variants reported to them by Genomics England, to what extent a family's presenting case can be explained by the combined variants reported to them (including any segregation testing performed); confidence in the identification and pathogenicity of each variant; and the clinical validity of each variant or variant pair in general and clinical utility in a specific case (only the most recent update will be shown and only one questionnaire per report).

6.4. Cancer

Cancer data are presented for either the patient level cancer diagnosis or "disease type" or the tumour specific sample details of participants in the Cancer arm of the 100,000 Genomes Project.

Data Relating to Cancer Participants

Name of Table / Data View	Description
cancer_participant_disease	For each cancer participant in the 100,000 Genomes Project, this table includes data about their cancer disease type and subtype.
cancer_participant_tumour	For each cancer participant's tumour in the 100,000 Genomes Project, this table contains data that characterises the tumour, e.g. staging and grading; morphology and location; recurrence at time of enrolment; and the basis of diagnosis.
cancer_participant_tumour_ metastatic_site	For each cancer participant in the 100,000 Genomes Project, this table contains the site of their metastatic disease in the body (if applicable) at diagnosis.
cancer_care_plan	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains information from their NHS cancer care plan on their treatment and care intent, in particular outcomes of MDT meetings and coded connected data (e.g. diagnoses from scans).
cancer_surgery	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains details of what surgical procedures were had, as well as the specific location of the intervention.
cancer_risk_factor_general	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains data on general cancer risk factors, namely smoking status, height, weight and alcohol consumption. This table was compiled with input from GeCIP members.
cancer_risk_factor_ cancer_specific	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains data on specific risk factors related to particular cancer types. This table was compiled with input from GeCIP members.
cancer_invest_imaging	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains: coded data on imaging investigations characterising the scan, its modality, anatomical site and outcome; as well as the outcome of the imaging report in free text form.

Data derived from or relating to tumour samples

Name of Table / Data View	Description
cancer_invest_sample_ pathology	For a proportion of cancer participants in the 100,000 Genomes Project, this table contains full pathology reports and other related data on and from their tumour samples around diagnosis and characterisation of the cancer. Please note that much of this information is also found in the clinic_sample and cancer_participant_tumour tables.
cancer_specific_ pathology	For a proportion tumours from cancer participants in the 100,000 Genomes Project, this table contains pathology data specific to that

	participant's cancer type. This may provide additional data to the cancer_invest_sample_pathology and cancer_participant_tumour tables.
cancer_systemic_ anti_cancer_therapy	For a proportion tumours from cancer participants in the 100,000 Genomes Project, this table contains details the regimen and intent of the patients' chemotherapy.
cancer_invest_circulating _tumour_marker	For a proportion tumours from cancer participants in the 100,000 Genomes Project, this table contains biomarker measurements specific to particular cancer types.

6.5. Medical History

Data Relating to Medical History

Table Name	Table Description
APC	Historic records of admissions into secondary care of GeL main programme participants
CC	Historic records of admissions into critical care of GeL main programme participants
OP	Historic records of outpatient attendances of GeL main programme participants
AE	Historic records of A&E attendances of GeL main programme participants
DID	Historic diagnostic Imaging records of GeL main program participants
DID_Bridge	Linking file of participants to DID records
PROMS	Patient Reporting Outcome Measures of GeL main programme participants
MHMD_v4_Record	Historic records of MH related admissions of GeL main programme participants
MHMD_v4_Event	Historic records of MH related admissions of GeL main programme participants
MHMD_v4_Episode	Historic records of MH related admissions of GeL main programme participants
MH_Bridge	Linking file of participants to MHMD records
CEN	Cohort Event Notification for GeL main programme participants
ONS	Office of National Statistics' cause of death records for the GeL main programme participants
SACT	Systemic Anti-Cancer Therapy (chemotherapy detail) data for cancer participants from PHE.

6.6. Genomics England Data Resources

Genomics England Data Resources are available in the following locations:

From the Inuvika Desktop:

~/gel_data_resources/

From the HPC:

/gel_data_resources/

The data resources available here are:

Tiering data for rare disease: Tiering data are available for rare disease participants who have been through the Genomics England interpretation platform. These data provide information on the pathogenicity of variants that have been identified in the proband's genome. Tiering data for rare disease probands can also be found in the designated LabKey table outlined above.

GMC exit questionnaires for rare disease: Outcomes questionnaire for interpreted genomes generated by Genomics England and Clinical Interpretation Providers.

Interpretation request data for rare disease: The following information can be found within the interpretation request JSON file: Family Pedigree and Other Family History, Analysis Panels & versions, Specific Disorder, Tiered Variants and Tiering version, HPO terms, Workspace (NHS GMC or LDP site code), Gene Panel Coverage, Disease Penetrance, Variant Classification.

Tiering, structural, and copy-number variant reports for Cancer: Annotated in JSON format. The file paths are available in the Quick View titled cancer_analysis.

Aggregated gVCF dataset:

This is a set of multi-sample gVCF files containing germline genomic data from 59464 participants from Release 5.1. The file contains samples from both the rare disease and the cancer programs, but only genomes on build GRCh38 were included. All included samples have passed a set of basic QC metrics

- cross-contamination <5%
- mapping rate >75%
- mean sample coverage >20
- insert size <250).

These QC metrics are given in the LabKey table `aggregate_gvcf_sample_stats`.

The aggregated dataset is split into 912 pieces for easier handling, due to its large size. No variant QC filters were applied in the dataset, but the VCF filter was set to PASS for variants which passed GQ, DP, missingess, allelic imbalance, and Mendel error filters. We recommend only using variants that have PASS in the filter column in your analyses. The data set alongside with a more detailed description is stored here: `/gel_data_resources/main_programme/aggregated_illumina_gvcf/GRCH38/20190228/`

7. Contact and Support

For all queries relating to this data release please contact the Genomics England Service Desk portal: www.bit.ly/ge-servicedesk. The Service Desk is supported by dedicated GeCIP team members for all relevant questions.

8. Change Summary

The change summary below summarises the changes in this release.

Data Release	Description

main_programme_v6_2019-02-28	• Date fields have been added to the following, tables:
	  ○ cancer_surgery
	  ○ rare_diseases_invest_blood_laboratory_test_report
	  ○ rare_diseases_invest_genetic
	  ○ cancer_participant_tumour
	  ○ cancer_risk_factor_general
	  ○ cancer_invest_imaging
	  ○ rare_diseases_participant_phenotype
	• In rare_diseases_pedigree, pedigree_family_id was renamed rare_diseases_family_id, and in rare_diseases_pedigree_member both member_participant_id and member_participant_sk were renamed participant_id and participant_sk accordingly
	• In participant table, duplicated_participant_id was added to highlight instances where a single person has been recruited under multiple participant_ids
	• A new table, death_details, was added. It contains death data received from GMCs only
	• In the participant table both mother_affected and father_affected have been changed to Yes/No/Unknown values
	• A new table, plated_sample, has been created to accommodate plated sample-level data from the laboratory sample table, specifically:
	  ○ platekey
	  ○ well_id
	  ○ plate_id
	  ○ biorepository_dispatch_datetime
	  ○ illumina_qc_datetime
	  ○ dna_amount (renamed illumine_dna_amount)
	  ○ illumina_delta_cq
	  ○ illumina_qc_status
	  ○ illumina_sample_concentration
	  ○ illumina_sequence_gender
	  ○ matched_dna_germline_laboratory_sample_sk (which is now accommodated in matched_sample_type and matched_sample_ids)
	• Column mydob has been removed from apc, op, ae tables
	• Column cdsuniqueid has been removed from ae table
	• SACT table with 38 fields covering details of chemotherapy regimens recorded by PHE for cancer patients has been added.
	• The sequencing_report table now contains the column
	  ○ lab_sample_id
	• The sequencing_report table has the following columns removed
	  ○ No

	○ Bam date ○ Bam size ○ Status ○
main-programme_v5.1_2018-11-20	• cancer_analysis – 8 new columns • hes_ae – 55 new columns, 2 columns removed: lsoa01, oacode6 • hes_apc - 64 new columns, 1 column removed: oacode6 • hes_op - 52 new columns, 2 columns removed: lsoa01, pctorig02
main-programme_v5_2018-10-31	• This release provides clinical data for 85,070 participants, and 71,860 genomes from 62,487 of these participants. Of these genomes, 54,456 are rare disease genomes (from 54,138 participants) and 17,404 are cancer genomes (from 8,349 participants) ○ 15,545 families with Tier 1, 2 and 3 variants from the interpretation pipeline; 2,470 families with GMC exit questionnaires • The LabKey table domain_assignment has been updated to include Moratorium end dates for genomes associated with participants in this table • File paths to tiering and structural variants from cancer genomes added to cancer quick view • New clinical LabKey tables with information on progression and medical history: cancer_surgery; cancer_risk_factor_cancer_specific; cancer_specific_pathology; cancer_systemic_anti_cancer_therapy; cancer_care_plan; cancer_invest_circulating_tumour_marker; as well as rare_diseases_imaging; rare_diseases_gen_measurement and rare_diseases_early_childhood_observation. • A new table tiered_variants_frequency was added between Main Programme Data Release V4 and this one (V5.1) • Multiple data fields were added, removed and renamed in cancer_invest_sample_pathology: ○ The following were added: tumour_id; sample_pathology_id; topography_icd_code; topography_snomed_ct_code; topography_snomed_rt_code; topography_snomed; topography_snomed_version; sample_receipt_date; sample_taken_date; vascular_or_lymphatic_invasion_cancer; event_date ○ The following were removed: topography_id; sample_details_id; vascular_or_lymphatic_invasion_cancer_id ○ The following were renamed: preoperative_therapy_id renamed to preoperative_therapy; vascular_or_lymphatic_invasion_cancer_id renamed to vascular_or_lymphatic_invasion_cancer • cancer_invest_imaging now includes free imaging report texts (report_text) and multiple other data fields were added to this table: cancer_invest_imaging; tumour_id; imaging_modality;

	cns_imaging_radiological_number_of_lesions; cns_imaging_radiological_lesion_size; cns_imaging_radiological_lesion_location; cns_imaging_radiological_largest_lesion_features; cns_imaging_principal_diagnostic_imaging_type; breast_imaging_mammogram_result • All new genomic data added in the current data release (since July 2018) are aligned against the reference genome version GRCh38, using alignment pipelines V4 • The following normalised diseases were renamed to match the official terms: *Cytopaenia and pancytopaenia* was renamed *Cytopenia and pancytopenia*; *Early onset dementia (encompassing fronto-temporal dementia and prion disease)* was shortened to *Early onset dementia* • The rare_disease_analysis quick view table now provides WGS family selection quality checks for rare disease families with genomes on build GRCh38, reporting abnormalities of the sex chromosomes, family relatedness and Mendelian inconsistencies, as well as reported vs genetic sex summary status (this contains an overall status – only sex checks are unpacked into individual data fields) • New outputs from the Genomics England Bioinformatics pipeline: The cancer_analysis quick view table now contains gold standard cancer genomes that have been through Genomics England Bioinformatics interpretation and passed quality checks
main-programme_v4_2018-07-31	• This release provides clinical data for 71,331 participants, and 55,681 genomes from 49,303 of these participants. Of these genomes, 43,997 are rare disease genomes (from 43,570 participants) and 11,684 are cancer genomes (from 5,715 participants). • New LabKey tables: panels_applied, rare_diseases_invest_genetic, rare_diseases_invest_genetic_test_result, rare_diseases_invest_blood_laboratory_test_report, panels_applied, cancer_invest_sample_pathology, cancer_invest_imaging, cancer_risk_factor_general, cancer_PCA_QC_stats, tumour_MB_signatures • LabKey tables removed: family_members • "Relationship to proband" field moved from family_members to rare_disease_analysis • Multiple data fields from cancer_participant_tumour and laboratory_sample added to cancer_analysis • "Disease" field changed to "disease or panel" in domain_assignment; an "origin" field has been added to domain_assignment to indicate whether the GeCIP domain applied to each participant is based on the disease they were recruited for or the panel applied to their genome • "Panel name" and "panel version" fields moved from tiering to panels applied

main-programme_v3_2018-04-30	• The dataset now includes 44,067 genomes. • Clinical data are also provided for participants with *and* without a sequenced genome, for a total of 61,554 • New LabKey tables are: family_members, genome_file_paths_and_types, rare_disease_analysis, tiering_data. • LabKey tables removed: rare_diseases_pedigree_member_disease, rare_diseases_pedigree_member_hpo_term. • Changes to LabKey tables including the new fields in the clinic_sample_level data and participant_level_data tables. • A new field genome_build was added to the sequencing_report table. This specifies 37 when the Delivery Version is V2 or before, and 38 when it is V4. • Removal of some ID fields where a human readable description of the value is available. • A new column named normalised_consent_form has been created in the participant table, assigning the free text values in consent_form to sensible categories • Pedigree diagnosis and phenotype data were removed from the research dataset
main-programme_v2_2018-01-30	• The dataset includes 31,384 genomes – an increase of 11,519 genomes from the first release. • Clinical data are also provided for participants with *and* without a sequenced genome, for a total of 53,190 participants. • A far broader set of clinical data are provided for participants, comprising 16 tables in LabKey. • In addition to Hospital Episode Statistics (HES), the secondary datasets Diagnostic Imaging Dataset (DID), Patient Reported Outcome Measures (PROMs) and Mental Health Services Data Set (MHSDS) are included in the release. • There have been significant changes to the data structure of the LabKey tables. Refer to the Data Dictionary that accompanies this release for further details.
main-programme_v1_2017-10-11	This data release represents the baseline for subsequent releases.