Deliverable

D3.4: Publication on optimal pathway to obtain genetic diagnosis for new RD patients	
Version | Status	V1 | final
Work package	WP3
Lead beneficiary	24 - SRUMC
Due date	31.12.2023 (M72)
Date of preparation	17.04.2024
Target Dissemination Level	Public
Author(s)	Lisenka Vissers (SRUMC)
Reviewed by	Birte Zurek (EKUT), Aurore Pélissier (U Bourgogne)
Approved by	Gisèle Bonne (INSERM)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Publication on optimal pathway to obtain genetic diagnosis for new RD patients.

Content

Abstract:

The diagnostic trajectory for individuals with a genetic rare disease often still contains consecutive testing, in which for instance exome sequencing is supplemented by complementary targeted assays to overcome technical challenges from the NGS-based assay. Despite this strategy, still 40-60% of individuals remain genetically undiagnosed, thus questioning whether this is the best strategy to diagnose all. With technical advances still ongoing, making genomes as diagnostic test feasible, and increasing knowledge on non-coding variant interpretation, providing a basis for the use of genomes in clinic, we are at a cross road to evaluate which genome strategy would be best.

In this task, we first evaluated the potential for short read genome sequencing to serve as first-tier diagnostic test for all (germline-based) genetic rare disease (denominated Phase I). By assessing a series of 1000 samples with known clinically relevant variants, it was uncovered that >95% of all variants were identifiable from 30x short read GS (Illumina platform). The 5% remaining variants were not identifiable from short read genomes. The type of variants is relevant for 29% of referrals to this diagnostic laboratory, suggesting that for this centre, short reads are a useful first-tier test for a majority of all rare disease referrals, but not all. In Phase II we subsequently targeted the 5% of failed variants by long read genome sequencing. In a pilot study, including 100 samples, it was noted that PacBio HiFi long read genomes were able to detect >98% of these variants, thus providing higher potential as first-tier test for rare disease than short reads.

From the results obtained in phase I and II, it would be recommended to implement long read sequencing as first tier test for individuals with rare disease. Whereas health economic evaluation still remained to be performed to determine socio-economic feasibility, this assay is able to replace all routine germline-based workflows, thus yielding the maximum diagnostic yield in a single test. Moreover, with increasing knowledge on interpretation of non-protein coding variants, long read sequencing also provides an ultimate opportunity to enhance diagnostic yield beyond todays diagnoses.

Introduction:

Diagnostic approaches to detect the underlying genetic causes of rare (germline-based) genetic diseases (RD) require a broad spectrum of technologies, ranging from traditional approaches such as karyotyping, genomic microarrays, FISH, and Sanger sequencing, to more advanced technologies, such as exome sequencing and transcriptomics. Each of these technologies is dedicated to detecting one or multiple variant types. In clinical genomics, (*de novo*) single nucleotide and copy number variants (SNV/CNV) are the most found aberrations, but to a lesser extent aneuploidy, expansions of short tandem repeats (STR), and (copy-neutral) structural variants (SV) also contribute to disease. To molecularly diagnose a rare disease, multiple workflows are often used, as a single disease can often be caused by multiple variant types. Importantly, for diagnostic purposes, every technology needs to prove clinical, as well as analytical, validity.

Genome sequencing (GS) promises comprehensive variant calling of all variant types from a single experiment, allowing for all types of molecular diagnoses. This (potentially) not only leads to an increased diagnostic yield but also provides a higher efficiency for genetic diagnostic laboratories that would no longer need to maintain multiple workflows to capture the various variant types. So far, however, widespread implementation of GS is lagging as the increase in diagnostic yield has been limited, also largely depending on the RD type studied (see Solve-RD deliverable report D3.3) while incurring higher costs compared to routine workflows.

A less explored scenario for effective implementation of GS as a routine diagnostic test is the impact of GS replacing all currently used diagnostic workflows. For instance, in one of the

tertiary referral centers for genetic diagnostic testing within Solve-RD (e.g. Radboudumc, in collaboration with their strategic academic partner Maastricht UMC+) approximately 25,000 individuals with a rare disease are tested annually, requiring >10 molecular and cytogenetic workflows to capture all genetic variant types. Replacing these workflows with a single GS-based workflow would increase efficiency.

To determine the feasibility of transitioning to a generic short read GS diagnostic workflow, we performed short read GS on 1,000 individuals previously molecularly diagnosed with a rare genetic disease, representative of the myriad of genetic variant types identified across 10 different workflows and modeled the impact of a GS-first diagnostic strategy for rare disease in our centers (Phase I). For those workflows that potentially could not be replaced, or required further additional testing, we assessed whether long read genome sequencing would be more beneficial than short read sequencing (Phase II).

Of note, while from Solve-RD deliverable report D3.3 it is clear that optical genome mapping (OGM) provided a high diagnostic yield for individuals who reached the end of routine care, for the vast majority of RD both variants at nucleotide level as well as structural variation is of importance. As OGM can only provide structural variation, and no sequence level information, this approach is not taken into account as a potential first tier (germline-based) genetic test for rare disease.

Report:

PHASE I: SHORT READ GENOME SEQUENCING AS FIRST-TIER TEST FOR RD

Genome diagnostics and cohort demographics

We performed a local 1000 short read genome project included archival DNA samples of 505 males and 495 females who were genetically tested in the Radboudumc/Maastricht UMC+ laboratories using 10 different workflows (**Supplementary Figure S1-2**).

For 378 individuals, this included analysis of specific variants, a single gene or a few genes, whereas in 617 individuals, extensive gene panels or other genome-wide analyses were used. For the remaining five individuals, a combination of both approaches was employed (**Supplementary Figure S2**). A total of 1,271 diagnostically relevant variants were reported (**Supplementary Figure S2**). All variants were called complying to specifications of DRAGEN variant calling, grouping them in three categories: a category for small variants (n=860), including SNVs and indels up to 50bp in size, a second one for large variants (n=366), *i.e.,* CNVs and STRs, leaving a third category for all other variants (n=45), involving SVs and chromosome anomalies (CA) (**Supplementary Figure S2**). For our 1000 genomes we reached an average sequencing depth of 37x (**Supplementary Figure S3**).

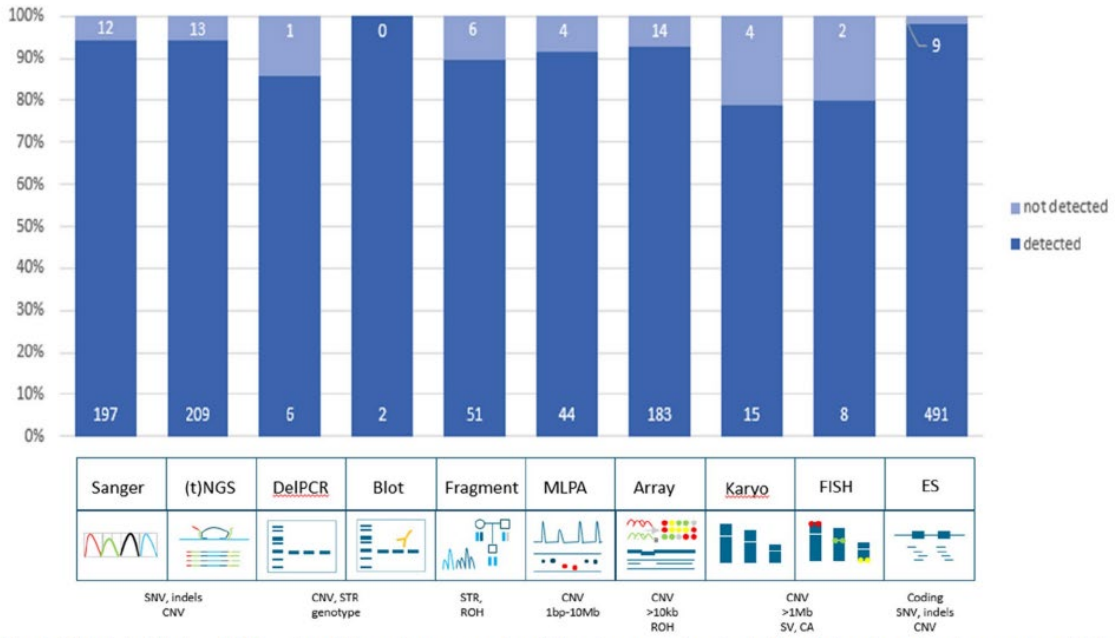


Fig. 1 Technical validation of 1271 variants. Schematic representation of detection rates of previously identified pathogenic variants across multiple different workflows. In total, 94.9% (1206/1271) of all variants were detected in GS data. The distribution of variants across the ten workflows shows a detection rate ranging between 79 and 100%. Abbreviations: targeted next-generation sequencing ((t)NGS), deletion polymerase chain reaction (DelPCR), multiplex ligation-dependant probe amplification (MLPA), fluorescence in situ hybridization (FISH), exome sequencing (ES), single nucleotide variants (SNV), copy number variants (CNV), short tandem repeat expansions (STRs), region of homozygosity (ROH), structural variants (SV), chromosome anomalies (CA)

Short read GS technical validation and feasibility assessment of replacing workflows by short read GS

In total, 94.9% (1,206/1,271) of all variants were detected with short read GS (**Figure 1**). Small variants were detected in 96.1% (826/860), large variants (123 bp – 72.8Mb) in 93.2% (341/366), and other variants in 86.7% (39/45) (**Supplementary Figure S4**). Subdividing the cohort by the variants we expected to readily identify (n=1,148) and those that we would not (n=123), indeed confirmed the prior knowledge of the technical challenges in detecting mosaic variants and variants located in homologous regions or genes with short-read 30x GS: 1,134 of 1,148 variants (98.9%) were detected as expected, whereas only 72/123 (58.5%) of challenging variants were identified (Fisher's exact test $p<0.001$). Of note, the detection limit of small mosaic variants was 13%.

We next reconstituted the 1,271 variants to their original workflows to determine the overall performance of detection of different variant types per workflow, which ranged from 79% for karyotyping to 100% for Southern blots (**Figure 1; Table 1**). Subsequent analysis of the TPR per workflow revealed that all workflows, except repeat length analysis, karyotyping and FISH, were determined to have a TPR>98%.

Table 1

*Excluded indications: Adenomatous polyposis coli, Chronic lymphocytic leukemia, PTEN Hamartoma tumor syndrome (diagnostic referrals that are under suspicion of harboring mosaic variants and/or added to include mosaic variants although not primarily aimed at germline testing); Excluded variants: mosaic variants <20%, variants in the *CYP21A2*, *SMN1*, *OTOA*, *STRC* or OPSIN genes.
TPR≥98% indicated by grey marking

# variants	1) Technical validation				2) Technical validation + exclusion expected false negatives*			
Workflow	positive	false negative	total	TPR	positive	false negative	total	TPR
Sanger	197	12	209	94.3%	178	0	178	100.0%
(t)NGS	209	13	222	94.1%	193	0	193	100.0%
DelPCR	6	1	7	85.7%	4	0	4	100.0%
Blot	2	0	2	100.0%	2	0	2	100.0%
Fragment	51	6	57	89.5%	51	6	57	89.5%
MLPA	44	4	48	91.7%	33	0	33	100.0%
Array	183	14	197	92.9%	168	2	170	98.8%
Karyo	15	4	19	78.9%	15	1	16	93.8%
FISH	8	2	10	80.0%	8	2	10	80.0%
ES	491	9	500	98.2%	486	3	489	99.4%
Total	**1206**	**65**	**1271**	**94.9%**	**1138**	**14**	**1152**	**98.8%**
Type variant								
SNV, indels	827	34	861	96.1%	789	3	792	99.6%
STR	52	6	58	89.7%	52	6	58	89.7%
ROH	26	1	27	96.3%	24	0	24	100.0%
CNV	262	18	280	93.6%	239	2	241	99.2%
CA	28	2	30	93.3%	23	0	23	100.0%
SV	11	4	15	73.3%	11	3	14	78.6%
Total	**1206**	**65**	**1271**		**1138**	**14**	**1152**	

Abbreviations: targeted next generation sequencing ((t)NGS), deletion polymerase chain reaction (DelPCR) multiplex ligation-dependent probe amplification (MLPA), fluorescence in situ hybridisation (FISH), exome sequencing (ES), single nucleotide variants (SNV), short tandem repeat expansions (STRs), regions of homozygosity (ROH), copy number variants (CNV),), chromosome anomalies (CA), structural variants (SV)

In silico extrapolation of detection rates to 58,393 variants and 4,266 disease genes

Assessing the available coverage data of 794 detected SNVs in our cohort showed that 99.1% had a ≥10x coverage (**Supplementary Figure S5**). We next leveraged the observations onto a larger *in silico* data set of variants. Hereto, we obtained 58,393 genomic coordinates from variants known in the VKGL and/or ClinVar databases to cause autosomal dominant/recessive disease and determined the sequence coverage for those positions across 35 genomes. For 99.5% of variants, the minimal coverage across 35 genomes was ≥10x (**Supplementary Figure S5**). Generation of similar coverage statistics for all coding bases of 4,266 disease-associated genes showed that the average coverage was 45x (**Supplementary Figure S5**), with 88.1% of genes (3,759/4,266) having a coverage of ≥10x for all protein-coding bases (**Supplementary Figure S5**).

Modeling the impact of short read GS implementation in clinical practice

We next set out to model the impact of short read GS implementation on everyday practice in our clinical centers, from both the clinical point of view, as well as from the laboratory point of view. In addition, we determined the impact on overall diagnostic yield obtained from a GS-first perspective.

In 2022, our tertiary genetic diagnostic laboratory received 30,514 diagnostic referrals to identify the primary germline DNA defect in 24,570 individuals with rare disease (**Figure 2**;

Supplementary Figure S6). In total, 883 different reasons for referral were observed, with the top 10 ranking clinical indications being responsible for 21% of all referrals. On average, per individual 1.24 referrals were noted, and 82% of individuals were referred only once (**Supplementary Figure S6**). Of note, for 966 individuals, the diagnostic referral (n=2,072) consisted of reanalysis of existing exome data and did not require the generation of novel experimental data. For the other 28,442 referrals, 36,633 wet lab experiments were performed using 11 different workflows (**Figure 2**).

From a clinical point of view, 750 of 883 (85%) clinical reasons for referral could be addressed via short read GS (**Figure 3**). The remaining 133 could not be performed via short read GS for various reasons, of which somatic variant detection (53%) and detection of variants in homologous regions (13%) are the most prominent (**Figure 3**). From a laboratory point of view, this short read GS-first strategy would not only fully replace the exome workflow and all Southern blots but would also considerably reduce the use of other workflows, such as Sanger sequencing (by 89%), MLPA (by 80%) and targeted NGS approaches (by 70%; **Figure 3**). Importantly, applying these observations to the diagnostic trajectory of all individuals shows that short read GS can be used as first-tier test for 16,777 (68%; **Figure 3**) of individuals.

Finally, we modeled the impact on the overall diagnostic yield. In 2022, a conclusive molecular diagnosis was obtained in 2,652 of 24,570 individuals (10.79%), and for another 3,597 (14.64%) a possible diagnosis was identified. Extrapolation of TPRs for individuals whose diagnostic trajectory would include short read GS, resulted in an anticipated conclusive diagnosis in 2,643 individuals (10.76%) and a possible diagnosis in 3,589 (14.61%; **Supplementary Figure S7**). Collectively, a generic short read GS-first strategy would thus possibly negatively impact the diagnostic outcome for 17 (0.07%) individuals (FN=17), translating to a possible false negative diagnostic rate of 0.3%.

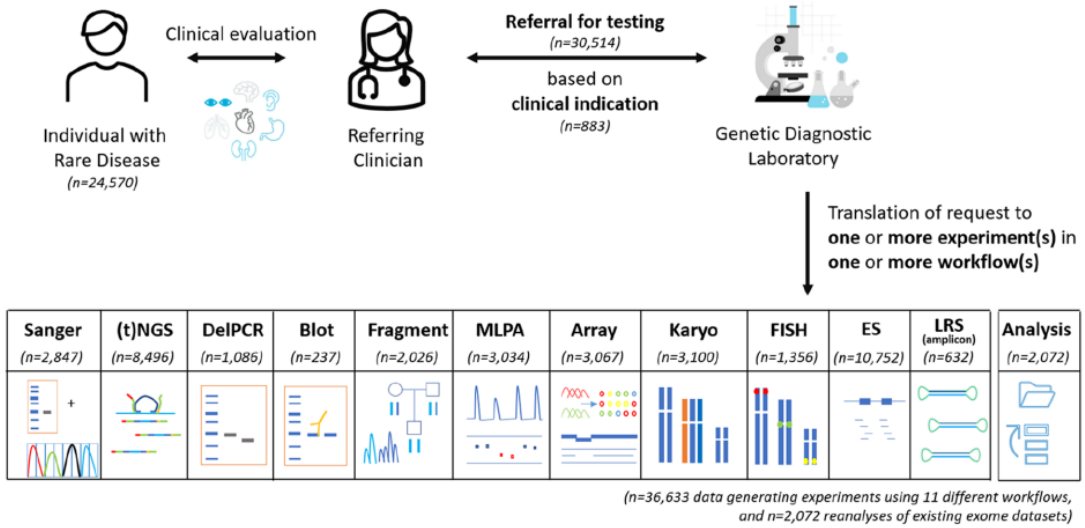


Fig. 2 Diagnostic referrals for genetic testing in 2022. In total, 24,570 individuals were referred, together requiring 36,633 data-generating experiments (in 23,604 individuals) in 11 different workflows, and 2072 reanalyses of existing (exome) datasets (in 966 individuals). Abbreviations: targeted next-generation sequencing ((t)NGS), deletion polymerase chain reaction (DelPCR), multiplex ligation-dependant probe amplification (MLPA), fluorescence in situ hybridization (FISH), exome sequencing (ES), long-read sequencing (LRS)

Discussion on short read genomes as potential first tier test

Over the last decade, the use of short read GS as a routine diagnostic test has been debated in the context of a higher potential diagnostic yield by interpreting non-coding DNA variants, as well as the potential to diagnose individuals with rare disease more efficiently, as short read GS allows the identification of virtually all genetic variants in a single experiment. Widespread diagnostic implementation has however been hampered by the costs involved with short read GS, given that the anticipated higher diagnostic yield has so far not materialized. An increased

diagnostic yield is however still expected for unexplained rare genetic disease, especially when looking beyond SNV and CNV detection in the exome only. To ultimately benefit from the advantages of GS, costs need to be reduced for a generic genetic diagnostic laboratory. In this study, we focused on the potential for GS as generic diagnostic rare disease test, replacing the full spectrum of workflows available in a genetic diagnostic laboratory. With our cohort of 1000 genomes, representative of 10 different workflows and a multitude of genetic variant types, we found that GS detected >95% of all pathogenic variants, albeit with variable efficacy across variant types and workflows. We also modeled the impact of a transition to a generic GS workflow for our diagnostic laboratories and conclude that for 68% of individuals diagnostically referred to our departments a generic GS workflow would be possible.

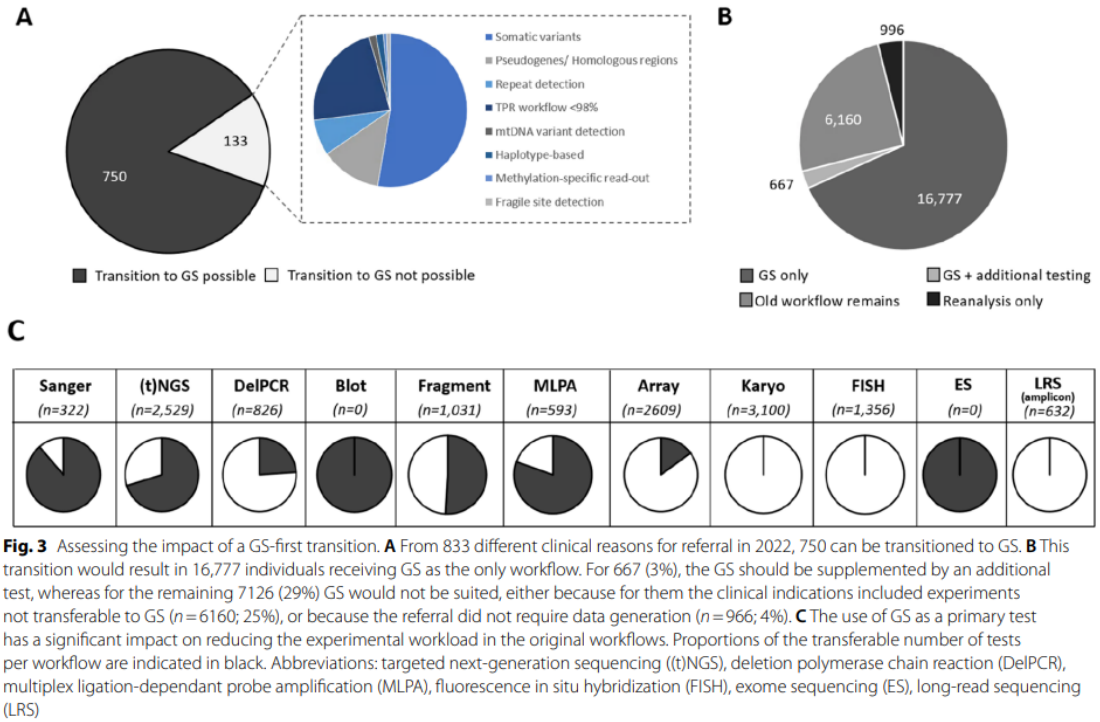


Fig. 3 Assessing the impact of a GS-first transition. **A** From 833 different clinical reasons for referral in 2022, 750 can be transitioned to GS. **B** This transition would result in 16,777 individuals receiving GS as the only workflow. For 667 (3%), the GS should be supplemented by an additional test, whereas for the remaining 7126 (29%) GS would not be suited, either because for them the clinical indications included experiments not transferable to GS ($n = 6160$; 25%), or because the referral did not require data generation ($n = 966$; 4%). **C** The use of GS as a primary test has a significant impact on reducing the experimental workload in the original workflows. Proportions of the transferable number of tests per workflow are indicated in black. Abbreviations: targeted next-generation sequencing ((t)NGS), deletion polymerase chain reaction (DelPCR), multiplex ligation-dependant probe amplification (MLPA), fluorescence in situ hybridization (FISH), exome sequencing (ES), long-read sequencing (LRS)

In our series of 1,000 samples, we noted differences in the detection of different variant types; 96.1% of small variants (<50bp) were detected, whereas only 93.3% of large variants, and 86.7% of other variants were recovered from short read GS. Interestingly, one of the arguments generally used as benefit from short read GS is its ability to better detect structural variation compared to ES. Conceptually, this is true from having a more uniform coverage across the genome. In addition, we, and others, have previously shown that additional diagnoses obtained via short GS compared to routine care, not only are often SV, but also that the resolution of SV complexity identified, often (far) exceeds that of other technologies. However, our data now show that the capture of SNVs/indels from short GS is more complete than of structural variants (Fisher's exact, p=0.006). Another striking observation was the recovery of 72 of 123 variants that we *a priori* expected to be beyond the technical limitations of 30x short read GS. These included variants located in highly homologous regions such as *STRC* and *OTOA*, as well as variants present in mosaic state (>14%). For the mosaic variants, increasing GS sequence depth may be the only way to recover all clinically relevant variation, especially if present at low variant allele fractions. For capturing variants in homologous regions bioinformatic solutions are under development, allowing the retrieval of (likely) pathogenic variants in these complex genomic regions. Currently, such dedicated callers exist, e.g., we successfully used in our analyses for the SMA (*SMN*) and *CYP21A2* loci, and for other paralogous region, suggesting that in the near future more (likely) pathogenic variants in such regions can be recovered.

Diagnostic efficacy can be enhanced by reducing the complexity of sample handling and the number of workflows. In our laboratory set-up, one clinical referral is often translated into experiments in multiple workflows; for example, to molecularly diagnose CHARGE syndrome, caused by *CHD7* haploinsufficiency, both Sanger sequencing and MLPA analysis are needed to allow the detection of SNV/indels as well as of (partial) gene deletions. The introduction of a generic short read GS workflow would allow for calling both SNV/indels, CNVs and other SVs affecting *CHD7* from a single experiment. For other disorders, for instance those caused by the expansion of short tandem repeats, it might be more challenging, as short read sequencing technologies may be unable to capture the full length of the extension. However, our data shows that although for some repeats the exact length cannot be obtained, a generic short read GS workflow is able to identify those individuals with repeat lengths outside of the normal range. This result can be followed with dedicated tests to determine the size of the repeat. From an efficacy point of view, one may argue that a second workflow is still required. While this is a valid point, in a generic short read GS workflow, the subsequent use of a second workflow is much more efficient, as it will only be used for those individuals with a high *a priori* chance of a positive outcome (given their abnormal short read GS results).

Whether or not it is efficient for laboratories to make a transition towards a generic GS workflow may depend on lab-specific factors, including size of the lab, number of workflows in use, and type of diagnostic referrals received. From our series of 1,000 genomes tested, we showed that ES can technically be replaced by GS (TPR>98%), in line with previous reports on comparing diagnostic outcomes of ES and GS. Hence, diagnostic laboratories, whose expertise is to only perform ES, could easily move towards GS with the benefit of a faster workflow as enrichment is no longer needed. Yet, for laboratories specialized in the use of karyotyping (TPR<98%) for the detection of somatic copy number changes, routine 30x short read GS might not be sufficient. The results of our study should therefore be carefully examined and extrapolated to local infrastructure and clinical expertise. Of note, a site-specific (early) health economic impact analysis is also recommended prior to large-scale implementation, in which cost-effectiveness evaluations are gaining increasing awareness. These studies are mostly performed in the context of proving that an early diagnosis also has a beneficial impact on overall health care cost expenditure. In light of implementing a generic short read GS workflow, a micro-costing study could, however, be more relevant. These latter studies would allow to weigh possible cost-reductions from phasing out workflows and changes in workforce against potential increase of per-sample sequencing costs, as well as differences in (ease of) clinical data interpretation.

Here, we report on our laboratories, which together maintain >10 workflows, representative for most core technologies used in genetic testing, and enabling detection of all variant types. The scenario models for our centers showed that 750/883 (85%) diagnostic referrals can be completed using GS, which would result in 68% of all individuals referred to our diagnostic laboratory making use of a single workflow and a single experiment, and 3% needing additional testing, suggesting that for 71% of individuals a short-read GS-first strategy would be beneficial. Whereas this analysis did not include a full micro-costing study, a generic short read GS-first workflow for such volume of samples might become within reach, especially with prices announced for germline short read GS in the range of 100 to 200 dollars per genome. For the 15% of clinical indications not transferable to short read GS (responsible for 29% of individuals referred), we noted trends, such that most of these required somatic structural variant detection, currently assayed via karyotyping, FISH and/or arrays, or variants that were located in complex regions of the genome, currently assessed by amplicon-based long read sequencing strategies. Based on the results obtained in this study, we could maintain these workflows to be primarily used for these diagnostic referrals. Alternatively, technological innovations specifically targeting these variant types would constitute a worthwhile investment. For somatic variant detection via karyotyping, FISH and/or arrays, optical genome mapping could replace these workflows as a second major generic assay, available in parallel to GS, but used for mutually exclusive clinical referrals. Similarly, a more generic use of long read genomes may provide a costs-effective strategy for diagnostic referrals involving variants in complex regions

in the genome, or where variant size exceeds those detectable from short reads (such as for repeat expansions).

The implementation of a novel technology requires careful balancing of the pros and cons. For short read GS, our study has highlighted advantages related to laboratory efficiency, but also showed that not all previously detected (likely) pathogenic germline variants were also identifiable from GS. Hence, if a generic GS workflow was to be used, it is to be expected that some individuals who would receive a conclusive diagnosis with the old diagnostic test strategy, would no longer do so with the implementation of a generic GS. In our objective quantification of the false negative rate from GS, using all diagnoses obtained by the current diagnostic strategy as the gold standard, we modeled that the transition to a generic GS in our laboratory might result in an additional diagnostic false negative rate of 0.3%. Whereas this is undesirable for the individual patient, previous experience has shown that there may be trade-offs. For instance, with the introduction of genomic microarrays at the expense of karyotyping, no longer detecting apparently balanced chromosomal rearrangements had to be accepted. Further, with the introduction of ES as replacement for Sanger sequencing for genetically and clinically heterogeneous disorders, one lost sensitivity at base pair level while gaining in mutation target size. Both innovations changed diagnostic testing, because despite losing out on a few positive diagnoses, they still improved the overall diagnostic yield. So far, the overall diagnostic advantage of short read GS is still limited (Solve-RD deliverable report D3.3). Disease-specific evaluations of diagnostic yield of short read GS have, however, reported on an increase in diagnostic yield, ranging from 1.3% for neurodevelopmental disorders to 17% for congenital limb malformation. Additionally, it has been reported that cytogenetically found apparently balanced chromosomal rearrangements appear to be genomic imbalances in ~1/3 of patients with *de novo* translocations and inversions, and that ~2/3 of balanced chromosomal abnormalities are involved in pathogenic mechanisms. With growing experience in detecting and interpreting structural variants in GS data, we also expect to identify more inversions, translocations, and other structural variants as underlying causes of human genetic disease. The use of GS over current workflows would provide an added value for which individuals with rare disease would immediately benefit, thus potentially compensating for the 0.3% diagnostic loss from introducing a generic short read GS workflow. Finally, our study is designed as technical benchmarking, which did not include an evaluation of variant prioritization. We and others, have, however, recently shown in prospective parallel and randomized short read GS studies that the similar variants and diagnostic yield is be obtained when comparing GS to current (non-GS) standard-of-care diagnostic workflows. In light of this, it is also worthwhile to underscore that even though analytically, a full genome sequence is provided, a targeted interpretation of variants, in line with the clinical request would still be pursued. That is, initially variants in single genes can be prioritized using *in silico* enrichment strategies when the short read GS is performed instead of a Sanger test, or, alternatively, only CNVs can be visualized when otherwise a karyotype would have been generated. If negative, a more agnostic approach for interpretation of genetic variation can performed where the existing and already available short read GS data provide a valuable resource for efficient reanalysis and reinterpretation strategies. We note that 6.8% of our referrals (n=2,072) involved reanalysis of existing exome data. With increasing knowledge on the role of (rare) non-coding variants in relation to disease and improvement in the bioinformatic detection of variants in complex regions of the genome from short reads, the availability of short read GS provides more flexibility in adapting reanalysis strategies towards these loci and variant types in the near future.

PHASE II: LONG READ GENOME SEQUENCING AS FIRST-TIER TEST FOR RD

Cohort collection and Long Read Genome Sequencing

Phase I revealed the challenges in the detection of variants in short read genomes, including the identification of structural variants, sequencing repetitive regions, phasing of alleles and distinguishing highly homologous genomic regions. Long read sequencing may overcome these challenges. We therefore next set out to evaluate the possibility for long read genome sequencing to replace routine genetic testing. The use of long read sequencing could additionally even further increase diagnostic yield (as highlighted in the Solve-RD deliverable report D3.3).

To determine the clinical utility, we performed LRS for 100 samples. LRS was performed using HiFi genomes on Pacific Biosciences (PacBio) Revio instrument at ~30-fold coverage. These 100 samples collectively contained 128 variants of known clinically significance, but which are challenging or impossible to identify by short-read sequencing (e.g. heavily biased towards the 133 clinical indications in Figure 3A which cannot be replaced by short read GS). In more detail, these included 25 short tandem repeats (STRs), 9 indels (<50bp), 33 single nucleotide variants (SNVs) in complex homologous regions, 55 structural variants (SVs), 3 regions of homozygosity (ROH) and 3 for methylation defects.

Preliminary Results

A fully automated PacBio-based in-house pipeline was developed for quality control and sequence alignment, as well as the detection and phasing of all variant types including SNV/indels, STRs, SVs and methylation alterations. LRS could readily identify the vast majority (95%) of known pathogenic variants. Of these, 87% were automatically called, including CNVs, translocations, inversions, STR expansions, *de novo* methylation defects and SNV/indels in homopolymer stretches and/or homologous sequences; 8% required manual curation such as the inspection of the aligned sequencing reads. A minority of variants (5%) posed systematic challenges, which included variants in very long AG-rich repeats and cases with cytogenetic aberrations affecting the repeat-rich regions of the Y-chromosome and/or acrocentric parms. In conclusion, LRS can identify the vast majority of pathogenic variants that are most challenging to detect with short-read technologies. Although our study identified some specific pitfalls, we expect that these can likely be resolved with further (bioinformatic) optimization. Most importantly, we show the potential to use a single technology to accurately identify all types of medically relevant genome variants, opening avenues to work towards a single generic test for germline testing in the future.

Overall Conclusions:

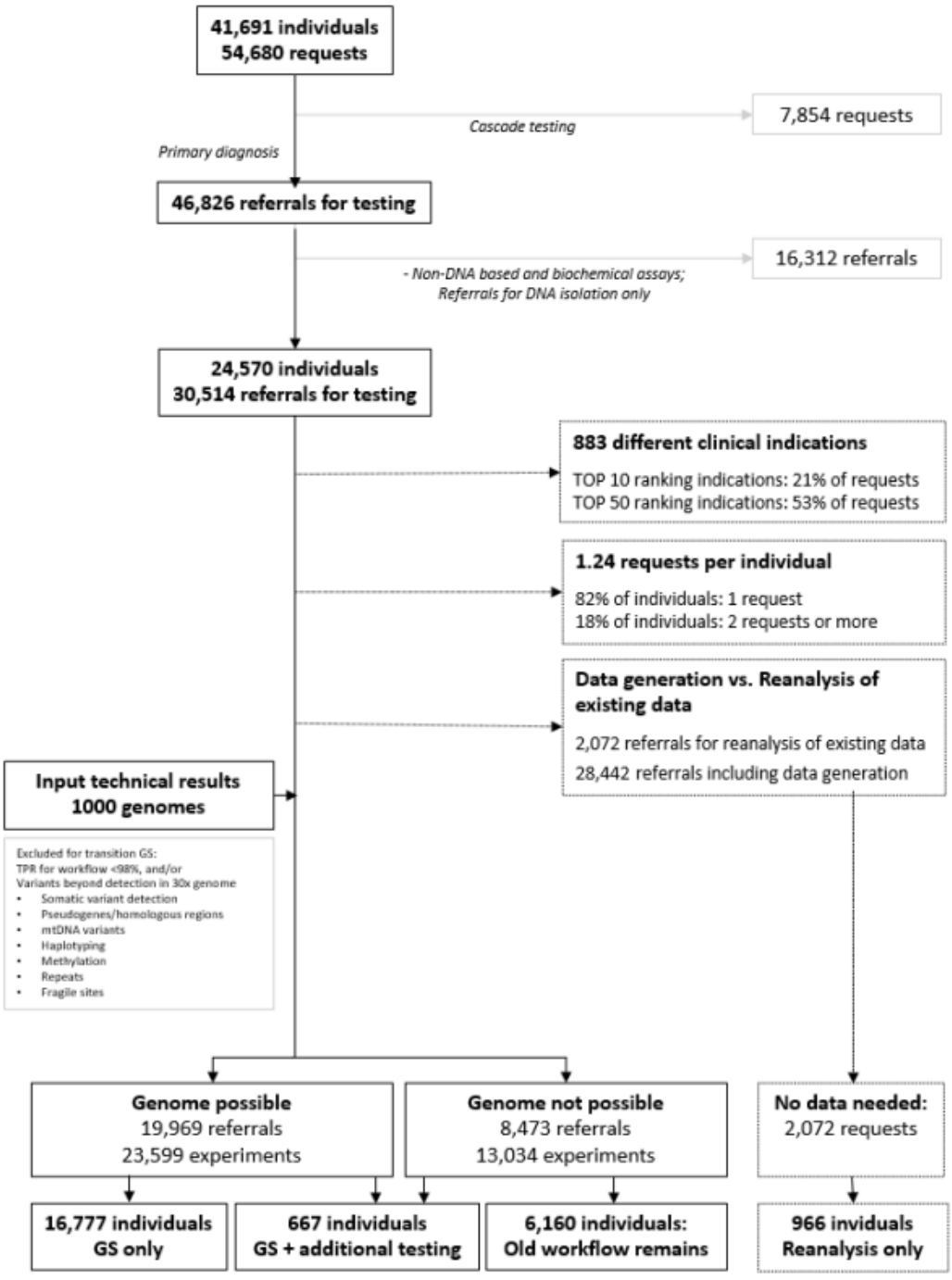
In this task we set out to find the best pathway to diagnose genetic rare disease. Ideally, this path consists of a single test that is able to detect (virtually) all clinically relevant variants. Whereas each diagnostic lab may serve different types of Rare Disease cohorts, from the novel -omics (see Solve-RD deliverable report D3.3), we noted that both short- and long read sequencing provided the highest likelihoods of being that 'one-test-fits all', while simultaneously increasing diagnostic yield.

Phase I, in which a systematic analysis of short read sequencing was performed to replace all other test, we noted that short read genome sequencing cannot identify all clinically relevant germline variants. Those that failed detection all were the consequence of known technological challenges of short read sequencing techniques. When, in Phase II, long read genome sequencing was performed and challenged with those variants that failed detection in short reads, we noted that long read sequencing can identify the vast majority of these.

Whether laboratories would first implement short read sequencing as first tier test because of higher-throughput testing in addition to lower per sample sequencing costs than currently achievable for long read genome sequencing, likely depends on the type of samples a laboratory receives (*e.g.*, which workflows to replace) and the socio-economic feasibility of implementing genome-based sequencing. From a diagnostic point of view, the data from our project would undoubtfully favour the use of long read sequencing, provided that the platform chosen has the accuracy and robustness to also detect *de novo* mutations. To date, the most promising technology to achieve these requirements would be HiFi sequencing using PacBio Revio systems.

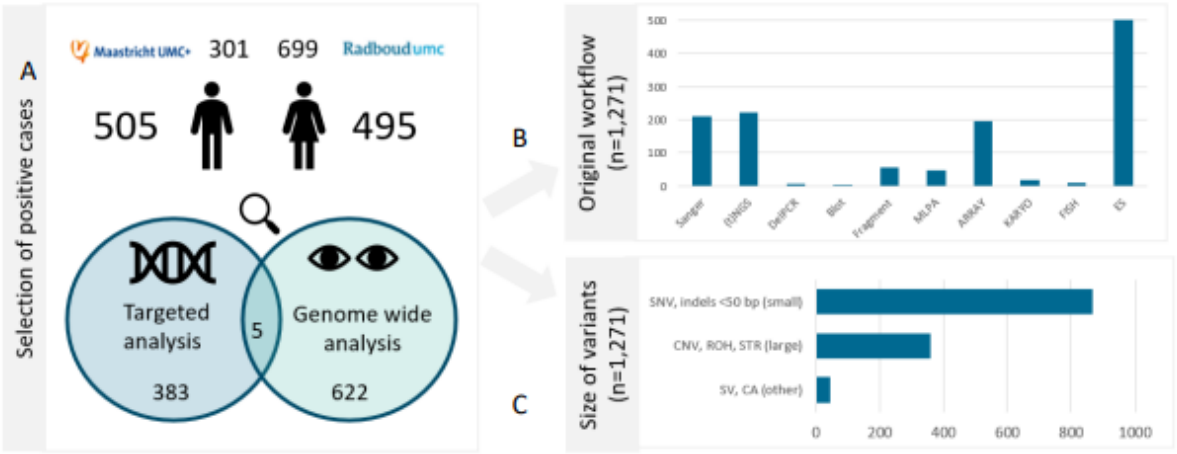
The data presented under Phase I are published by Schobers et al. as online first publication in Genome Medicine on February 14, 2024, under DOI 10.1186/s13073-024-01301-y.

SUPPLEMENTARY FIGURES



Supplementary Figure 1: Overview of genetic requests in 2022.

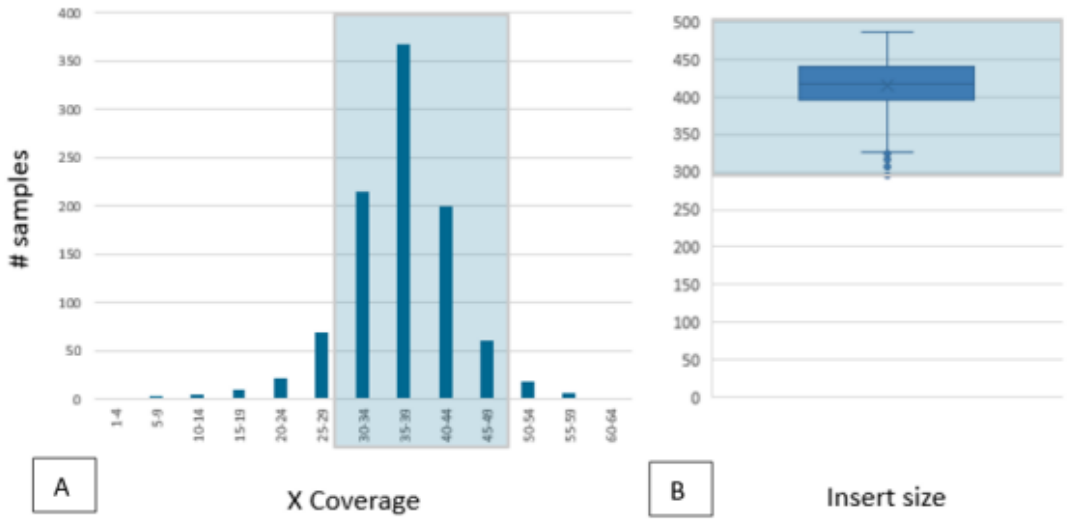
Figure S2: A cohort of 1000 cases with clinically relevant variants spanning the broad range of genome diagnostics.



A The 1000 genomes cohort consisted of 505 males and 495 females, who were genetically diagnosed in the Radboudumc or Maastricht UMC+ in 2018. The assays that were performed to find these diagnoses were either targeting specific variants and single (or a small set of) genes or complete gene panels or chromosomes were analyzed based on the patient's phenotype. **B** In these cases, a total of 1,271 variants were identified, requiring >10 different workflows to diagnose them. **C** The variants were grouped in small (<50 bp), large (50 bp and up), and other variants (SVs and CA).

Abbreviations: targeted next generation sequencing ((t)NGS), deletion polymerase chain reaction (DelPCR), multiplex ligation-dependent probe amplification (MLPA), fluorescent in situ hybridisation (FISH), exome sequencing (ES), single nucleotide variants (SNV), short tandem repeat expansions (STRs), regions of homozygosity (ROH), copy number variants (CNV), structural variants (SV), chromosome anomalies (CA)

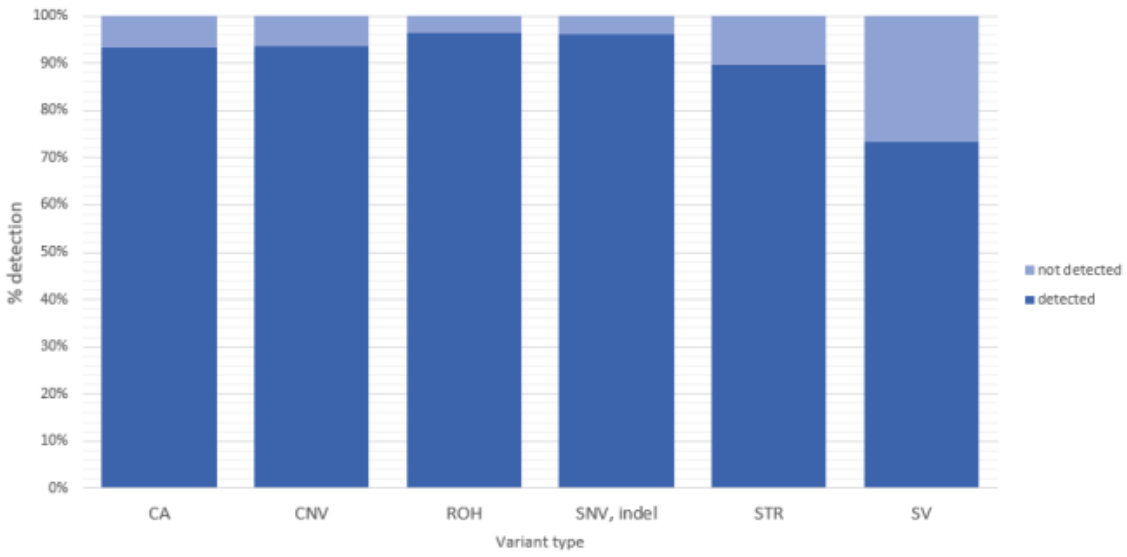
Figure S3: The average output of 1000 genomes.



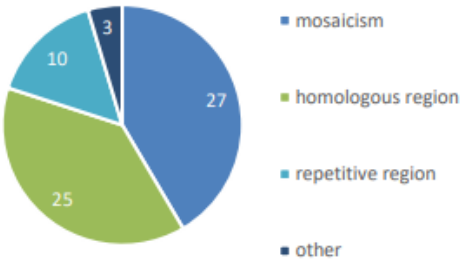
A As multiple observations per base are needed to come to a reliable base call, the recommended sequencing depth for genome sequencing is 30x to 50x. **B** Insert sizes are also important for the sequencing. For efficient sequencing, small insert sizes (risk of overlapping paired sequences) as well as larger fragments (decrease of cluster efficiency) must be avoided. We therefore aimed for a 300-500bp range for our 2x150bp paired-end sequencing. In this project we reached an average sequencing depth of 37x and an insert size of around 400-450bp.

Figure S4: GS Technical validation by variant type and assessment of why variants were not identified

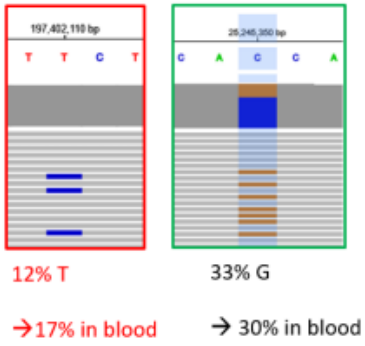
A



B



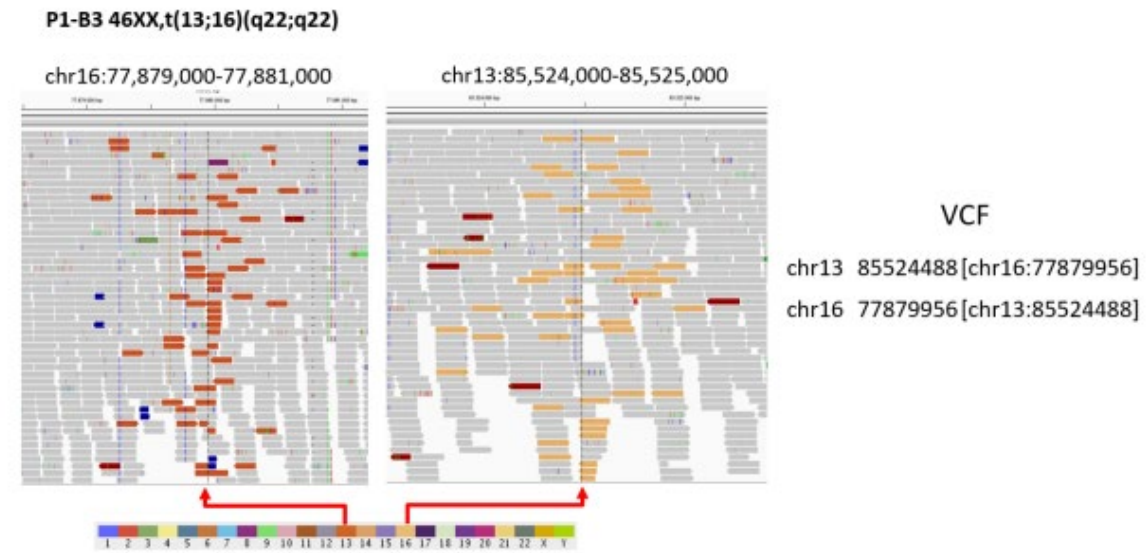
C



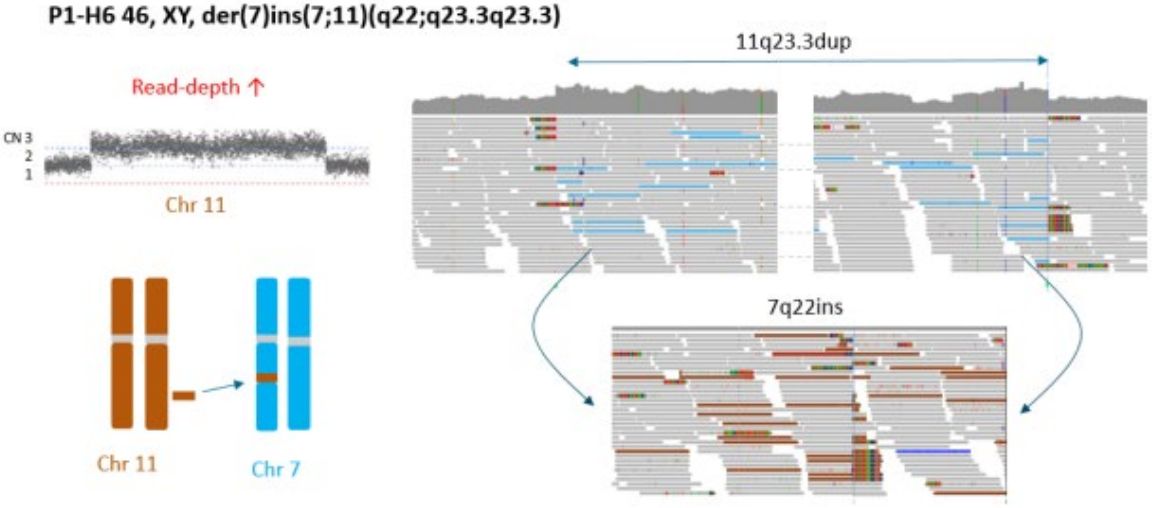
A In total, 94.9 % (1,206/1,271) of all variants were detected with GS. Small variants (<50bp) were detected in 96.1% (833/867), large variants (123 bp - 72.8Mb) in 93.3% (334/359), and other variants in 86.7% (39/45). The total list of variants and whether they were present in the GS data ('detected' vs. 'not detected') can be found in **Supplementary Table S2**. **B** In the 5% undetected variants (N=65), we identified common themes that are attributable to short-read 30x GS and downstream analysis. Undetected variants were mostly found in mosaic cases (n=27, 2.4-20%), homologous regions (n=25), i.e. pseudogenes or paralogues genes, or likewise in repetitive regions (n=10), i.e. repeats, telomeres or centromeres, and 3 others. **C** A mosaic variants in the *SF3B1* gene (Chr2(GRCh38):g.197402110T>C), which was originally detected with a targeted NGS approach in 17% of the blood sample, was present in 6/50 (12%) of the reads and not present in the VCF file of the GS data. A mosaic variant in the *KRAS* gene (Chr12(GRCh38):g.25245350C>G, originally detected with a targeted NGS approach in 30% of the blood sample, was present in 15/46 (33%) of the reads and in the VCF file of the GS data.

Figure S5: Examples of comprehensive GS

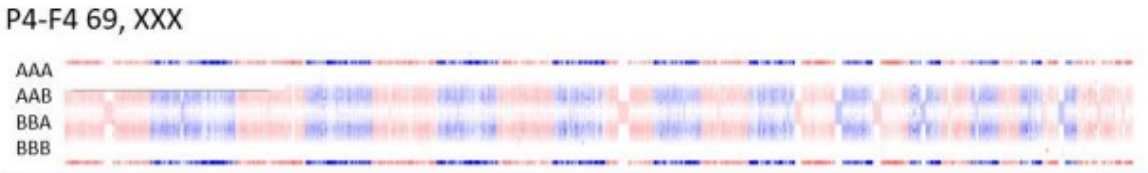
A



B



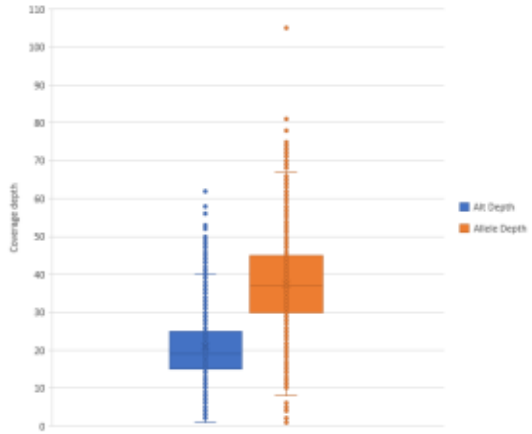
C



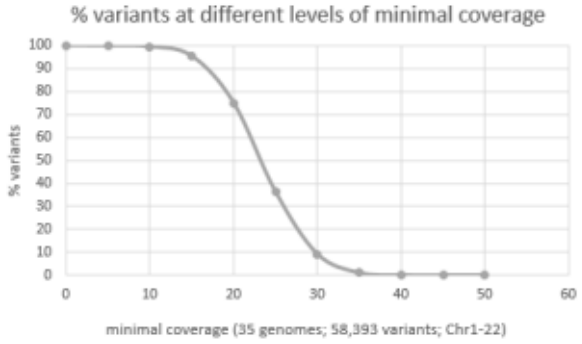
A Based on visual inspection and targeted manual search of the variant calling format (VCF) file, we could identify the previously detected translocation between chromosomes 13 and 16. **B** Likewise, we detected a copy-number gain on chromosome 11, which translocated to chromosome 7. In the diagnostic trajectory this derivative chromosome was detected with a targeted FISH analysis performed subsequent to an array analysis, in which only the gain was identified. **C** GS B-allele frequency plots can identify triploidies.

Figure S6: *In silico* coverage statistics at variant level and disease genes

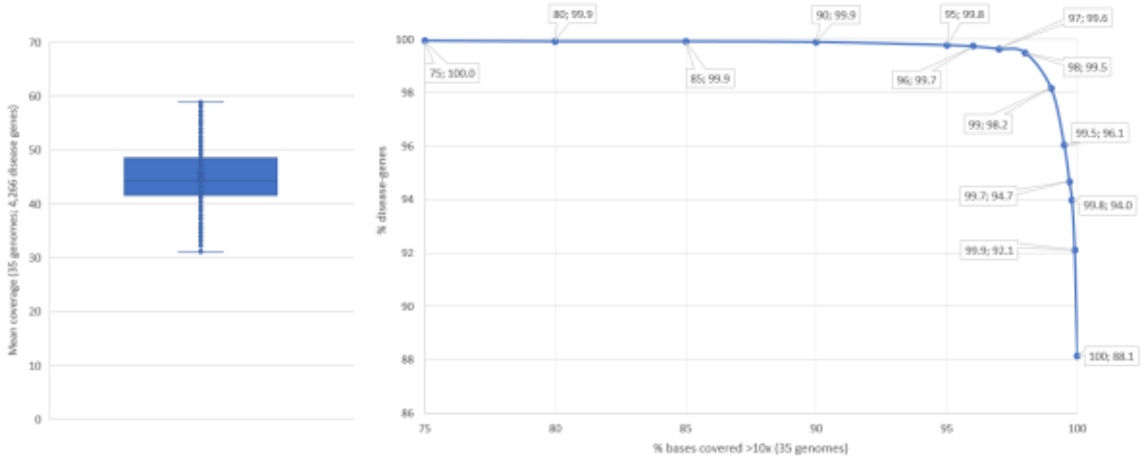
A:Coverage statistics for 794 detected SNVs from the 1000 Genomes



B: Coverage statistics of 58,393 ClinVar and VKGL variants



C-D: Coverage statistics of 4,266 disease genes



A Coverage data of 794 detected SNVs in our cohort, where allele depth ranged from 1-105, and (variant) alternative allele depth ranged from 1-62, with a 13-100% variant range. **B** Sequence depth at genomic positions that are known to harbor (likely) pathogenic variation and **C** Mean coverages for all coding positions of genes with well-established rare disease associations were calculated from 35 randomly selected genomes. **D** The fraction of genes versus the percentage of bases of the gene with ≥10x coverage.

Figure S7: Schematic representation of referrals to Radboudumc and MUMC+ in 2022

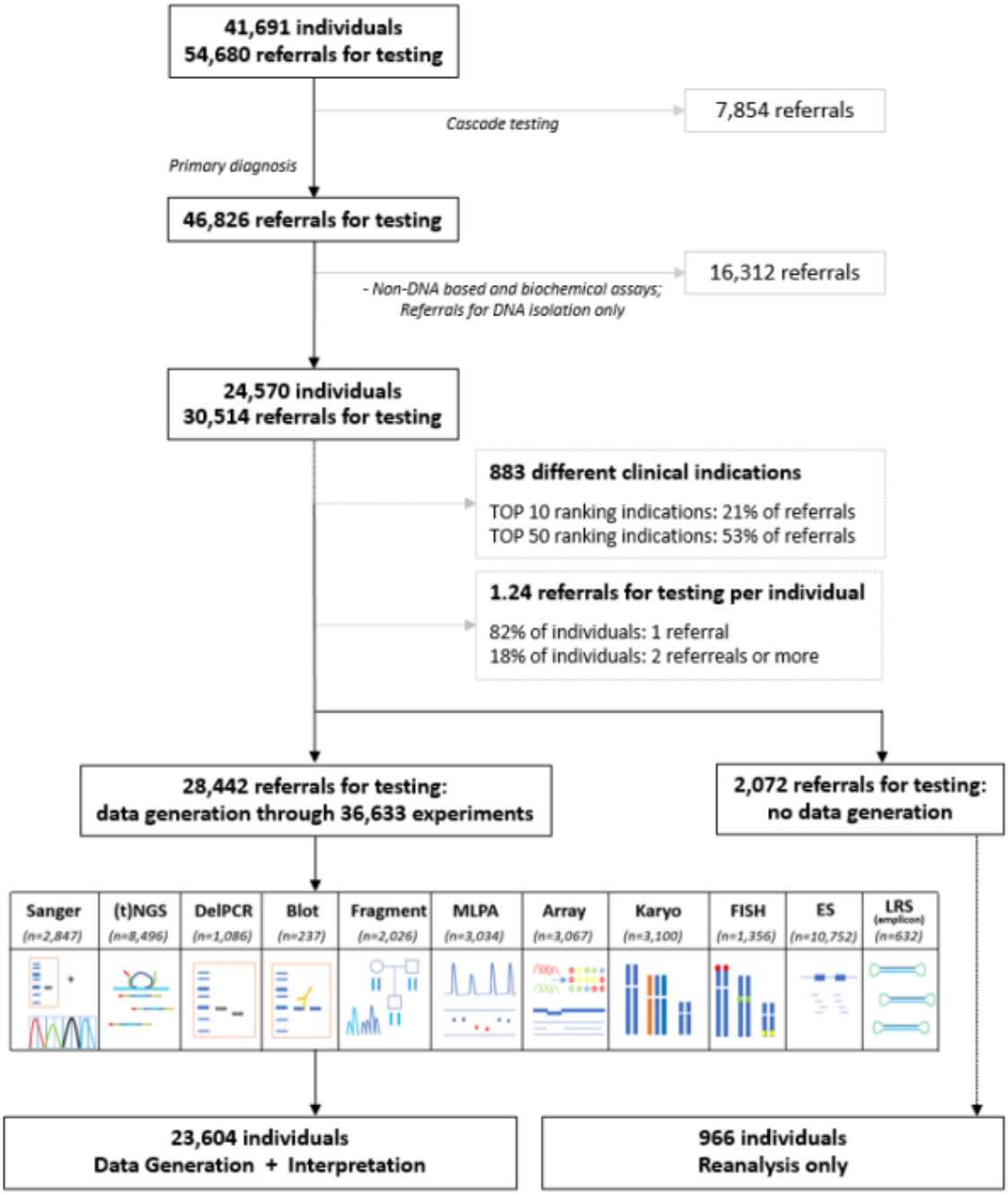
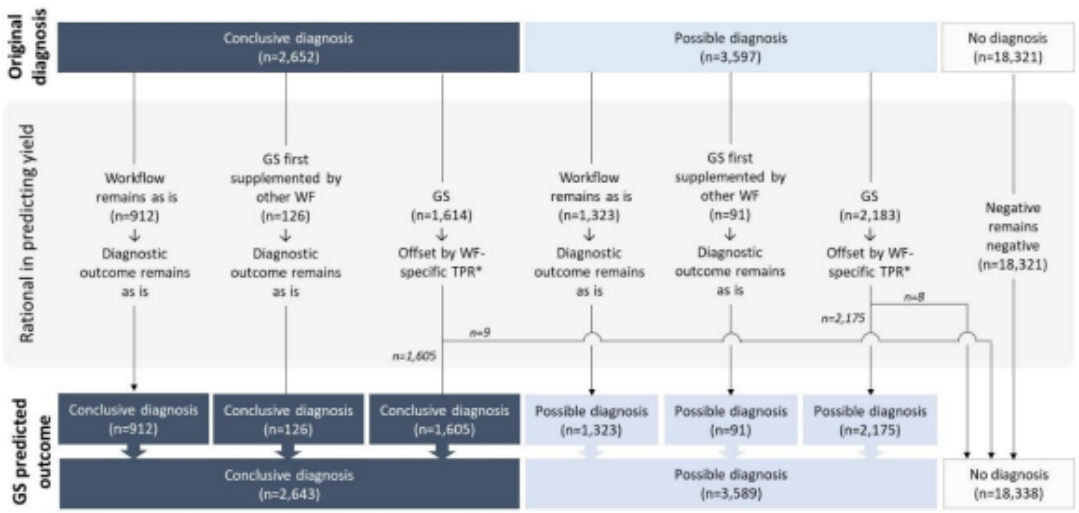


Figure S8: Schematic overview of assumptions made to evaluate the impact on diagnostic yield from transition to a generic GS approach

A



B

Workflow leading to (possible) diagnosis	Conclusive	Possible	TPR	Conclusive	Possible	Delta
(t)NGS	197	662	100,00%	197	662	-
ES	1253	1260	99,40%	1245	1253	15
Sanger	74	176	100,00%	74	176	-
Array	16	17	98,80%	15	16	2
Fragment	61	7	100,00%	61	7	-
MLPA	4	2	100,00%	4	2	-
DelPCR	9	24	100,00%	9	24	-
Blot	0	35	100,00%	0	35	-

A) Based on clinical referrals being transferable to generic GS, the impact on diagnosis was evaluated for all 24,570. Top row shows original diagnosis per individual, where '*n*' refers to number of individuals; *Offset with workflow specific TPRs are provided in **B**. Assuming all negative diagnoses remain negative, this translates to a possible false negative diagnostic rate of 0.3% (17/6232).