Zentrum für Kardiovaskuläre Genetik und Gendiagnostik
Das Genetikzentrum der Stiftung für Menschen mit seltenen Krankheiten

Chances and Challenges of High-Throughput Sequencing of Mendelian Disorders

Meienberg J,[1] Kopps A,[1] Plüss M,[1,2] Caspar S,[1] Dubacher N,[1] Matyas G[1,3]

[1]Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Schlieren-Zurich, Switzerland, [2]Institute of 4D Technologies, University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland, [3]Zurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland

Contact: meienberg@genetikzentrum.ch, kopps@genetikzentrum.ch, matyas@genetikzentrum.ch

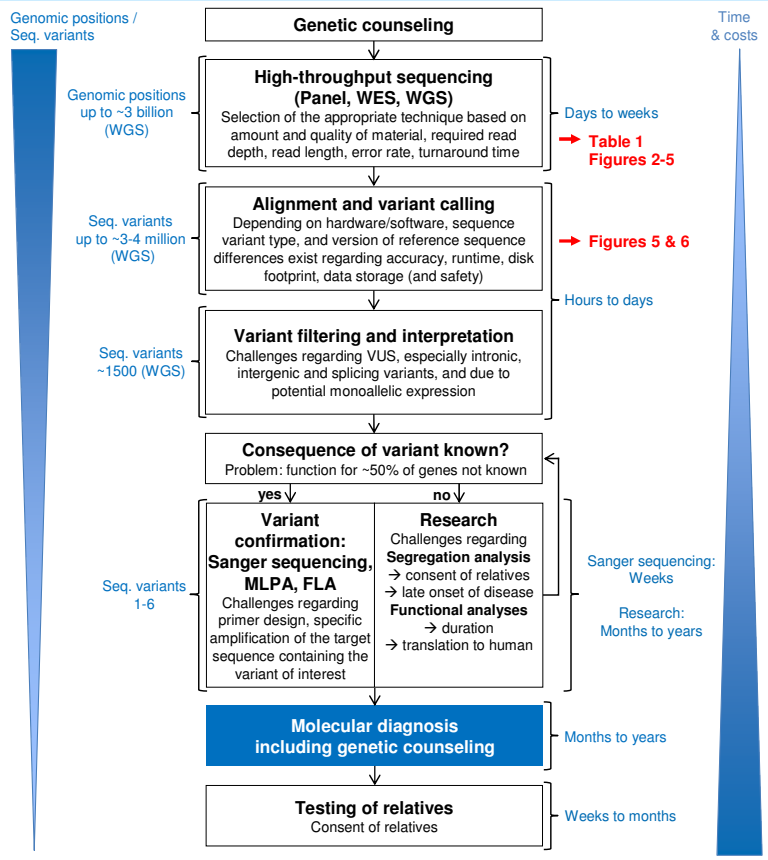


Figure 1. Challenges during the process from sequencing to diagnosis. FLA, fragment length analysis; MLPA, multiplex ligation-dependent probe amplification; Seq. variants, sequence variants; VUS, variants of unknown significance; WES, whole-exome sequencing; WGS, whole-genome sequencing [1].

Selection of appropriate sequencing method

Table 1. Comparison of widely used sequencing applications and platforms [1].

		Short-read[a]			Long-read (real) WGS	
	Sanger	TS	WES	PCR-free WGS	PacBio	ONT
Read length (bp)	Max: 500-1000	~300	~150	~150	Up to template length[b]	Up to template length[b]
Typical read depth	Not applicable	200-1000×	~100×	30-60×	10-30×	10-30×
Raw-read error rate (%)	0.001	0.1	0.1	0.1	10-15	12-17
Costs per sample ($)[c]	15-20	200-1000	500-1000	1000-2500	7000-20000	2750-8250
Disk footprint (GB) / ($)[d]	<0.1 / <0.01	<1 / <0.1	6-13 / <1	90-400 / 4-20	45-130 / 2-7	75-220 / 4-11
Chances **Advantage**	High accuracy	High read depth, easy interpretation, cost-efficiency, short turnaround time	Additional sequence information compared to TS, cost-efficiency	Uniform, GC content independent coverage of the genome	Coverage of repetitive and homologous genomic regions, detection of large SVs, discovery of novel isoforms, DNA/RNA base modifications, phasing	
Challenges **Limitation**	Low throughput	Incomplete coverage due to high GC-content, missing enrichment probes, and regions with mappability <1	Incomplete coverage due to high GC-content, missing enrichment probes, and regions with mappability <1	Incomplete coverage in regions with mappability <1	High first-pass (raw-read) error rate, low cost-efficiency	
Amplification step prior to sequencing	Yes	Yes	Yes	No	No	No

[a]Parameters of short-read sequencing are adapted to Illumina MiSeq v3 system (TS) and Illumina HiSeq X Ten system (WES, WGS); [b]Maximal read length only limited by length of the fragments sequenced (template); [c]Costs calculated according to most frequently used sequencing systems, library preparation kits, and reagents for the respective application, considering "typical read depth"; [d]Calculated for files like FASTQ, BAM, and VCF using corresponding in-house and publicly-available data. Costs were calculated considering disk footprint for backup as well. For TS, disk footprint was calculated for 100 average-sized genes with 2.5-kb coding region per gene; Avg, average; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; TS, targeted sequencing; WES, whole-exome sequencing; WGS, whole-genome sequencing; SV, structural variation.

PCR-free WGS overcomes limitations of GC-rich regions

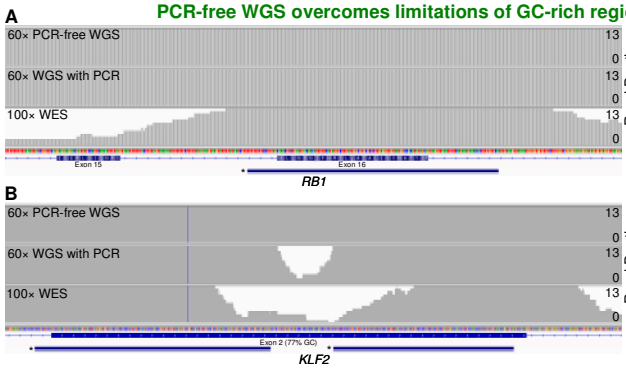


Figure 2. Coverage difference between whole-exome (WES) and whole-genome sequencing (WGS) exemplified for two genes. **A**, Not all exons are captured leading to low/insufficient read depth in WES in contrast to WGS. **B**, GC-rich (77%) exon 2 of *KLF2* is completely covered by PCR-free WGS and almost completely by WGS with PCR, while a large gap in coverage is present in WES. Coverage tracks are visualized by the Integrative Genomics Viewer (IGV, broadinstitute.org/igv). *Designed target region of Agilent SureSelect v5+UTR [1-3].

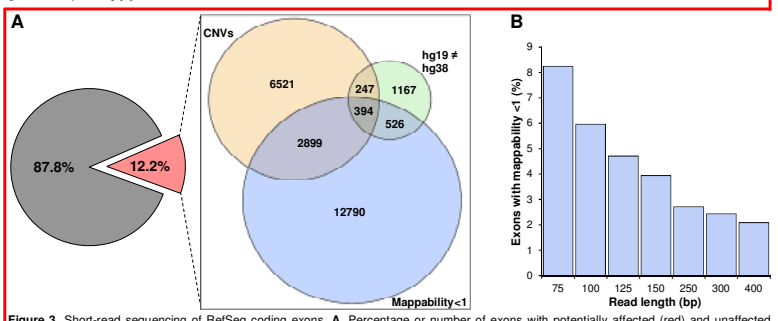


Figure 3. Short-read sequencing of RefSeq coding exons. **A**, Percentage or number of exons with potentially affected (red) and unaffected (gray) read depth alignment and/or variant calling in short-read whole-genome sequencing due to ambiguous 75-mer mappability (mappability <1), the presence of common copy number variations (CNVs), and the difference between the GRCh37 and GRCh38 reference genomes (hg19 ≠ hg38). **B**, Percentage of exons with mappability <1 (calculated using GEM version GEM-binaries-Linux-x86_64-20100419-003425 with m=2 like the UCSC mappability tracks). Y-chromosomal exons were excluded [1].

Repetitive regions can be analyzed using long-read sequencing

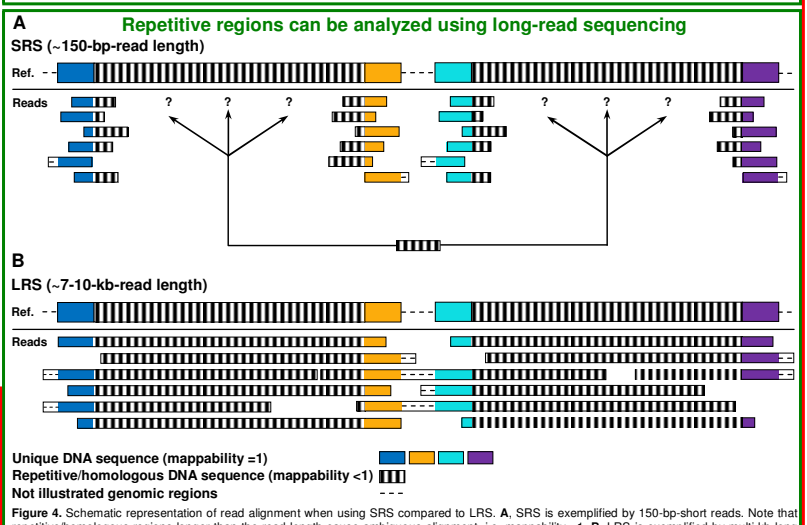


Figure 4. Schematic representation of read alignment when using SRS compared to LRS. **A**, SRS is exemplified by 150-bp-short reads. Note that repetitive/homologous regions longer than the read length cause ambiguous alignment, i.e. mappability <1. **B**, LRS is exemplified by multi-kb-long reads. Note that long reads can cover unique DNA sequences flanking repetitive/homologous regions, enabling unambiguous alignment, i.e. mappability =1. LRS, long-read sequencing; Ref., reference genome; SRS, short-read sequencing [1].

Selection of appropriate analysis pipeline

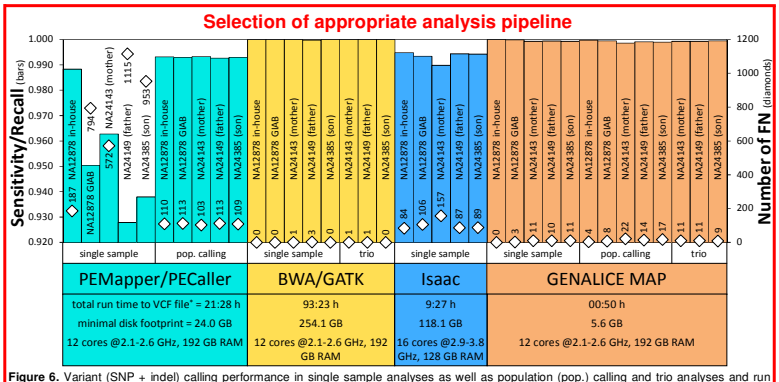
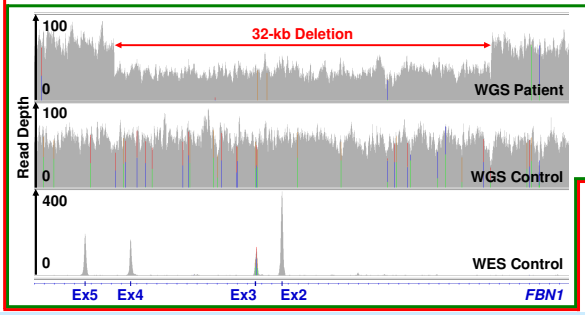


Figure 6. Variant (SNP + indel) calling performance in single sample analyses as well as population (pop.) calling and trio analyses and run time, disk footprint, and hardware specifications of four pipelines. Sensitivity/Recall (TP/[TP+FN]; TP = true-positive and FN = false-negative calls) and number of FN. WGS (Illumina HiSeq 2500, PE150, PCR-free) FASTQ files for NA12878 (NA12878 GIAB) and the Ashkenazim trio (NA24143, NA24149, and NA24385) were downloaded (ftp-trace.ncbi.nlm.nih.gov/giab, 300×) and downsampled to ~60×. In addition, we analyzed our in-house NA12878 WGS data (NA12878 in-house) sequenced at ~60× (Illumina X Ten, PE150, PCR-free). For population calling, the focal sample was analyzed together with 96 additional WGS data sets (sequenced like "NA12878 in-house") from our Caucasian (Swiss) patient cohort. The number of NIST-GIAB high-confidence benchmarking TP calls were 15,990 (NA12878), 15,345 (NA24143), 15,458 (NA24149), and 15,366 (NA24385). *, Run times shown are for single sample analyses of the downloaded NA12878 GIAB data [4].



Detection of SNVs and CNVs in one assay

Figure 5. Copy number variation (CNV) detection by HTS. Read depth (coverage tracks) of 60× PE150 PCR-free WGS data of the index patient and a control as well as 100× PE100 whole-exome sequencing (WES) data of a control for the deleted and flanking genomic regions displayed in IGV [2, 5].

References
1. Caspar and Dubacher *et al.* Clin Genet 93:508-19, 2018
2. Meienberg and Zerjavic *et al.* Nucleic Acids Res 43:e76, 2015
3. Meienberg *et al.* Hum Genet 135:359-62, 2016
4. Plüss and Kopps *et al.* Proc Natl Acad Sci U S A 114:E8320-2, 2017
5. Benke *et al.* J Thorac Dis 10:2456-60, 2018

Acknowledgement: This work was supported by following foundations (Stiftungen): Bangerter-Rhyner, Blumenau-Léonie Hartmann, COFRA, Ebnet, Ernst Göhner, FERNE HORIZONTE, Gebauer, Gemeinnützige Stiftung der ehemaligen Sparkasse Limmattal, Hirzel-Callegari, Jubiläumsstiftung Swiss Life, NOMIS, Palatin, Schäppi-Jecklin, Spendenstiftung Bank Vontobel, Suyana, and Wohlfahrtsstiftung des Vereines Zürcher Brockenhaus.