Received: 21 May 2022 | Accepted: 10 June 2022

DOI: 10.1002/mgg3.2016

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Array genotyping as diagnostic approach in medical genetics

Martina Witsch-Baumgartner[1] | **Gunda Schwaninger**[1] | **Simon Schnaiter**[1] | **Franziska Kollmann**[1] | **Silja Burkhard**[1] | **Rebekka Gröbner**[1] | **Beatrix Mühlegger**[1] | **Esther Schamschula**[1] | **Peter Kirchmeier**[2] | **Johannes Zschocke**[1]

[1]Institute of Human Genetics, Medical University Innsbruck, Innsbruck, Austria

[2]JSI Medical Systems GmbH, Ettenheim, Germany

Correspondence
Johannes Zschocke, Institute of Human Genetics, Medical University Innsbruck, Peter-Mayr-Str. 1, 6020 Innsbruck, Austria.
Email: johannes.zschocke@i-med.ac.at

Abstract

Genotyping arrays are by far the most widely used genetic tests but are not generally utilized for diagnostic purposes in a medical context. In the present study, we examined the diagnostic value of a standard genotyping array (Illumina Global Screening Array) for a range of indications. Applications included standalone testing for specific variants (32 variants in 10 genes), first-tier array variant screening for monogenic conditions (10 different autosomal recessive metabolic diseases), and diagnostic workup for specific conditions caused by variants in multiple genes (suspected familial breast and ovarian cancer, and hypercholesterolemia). Our analyses showed a high analytical sensitivity and specificity of array-based analyses for validated and non-validated variants, and identified pitfalls that require attention. Ethical-legal assessment highlighted the need for a software solution that allows for individual indication-based consent and the reliable exclusion of non-consented results. Cost/time assessment revealed excellent performance of diagnostic array analyses, depending on indication, proband data, and array design. We have implemented some analyses in our diagnostic portfolio, but array optimization is required for the implementation of other indications.

KEYWORDS

DNA array, genetic test, genotyping, screening, single nucleotide variants

1 | INTRODUCTION

The relevance of individual genetic analyses in clinical care is constantly increasing. The technological focus in most medical diagnostic centres has been on the expansion of massively parallel exome or genome sequencing resources that allow the comprehensive clarification of genetic factors in all types of disorders. However, despite impressive advances over the last decades, this approach requires an elaborate laboratory infrastructure as well as considerable bioinformatics and data storage resources. The costs of sequencing a single human genome has not fallen as much over the last years as may have been expected (Wetterstrand, 2020), and most health systems do not have the resources to offer comprehensive genetic analyses to all patients who may benefit from genetic information (Katsanis & Katsanis, 2013).

Genotyping with DNA arrays that target a large number of mostly single nucleotide variants ("SNP chips") has been developed as a cost-effective method for generating genetic information in a large number of individuals (Das et al., 2016; LaFramboise, 2009; Verlouw et al., 2021). The method has long been used for genome-wide association studies (Hirschhorn & Daly, 2005) and has recently become the method of choice for the imputation of polygenic scores (PGS), quantifying non-monogenic genetic risks for common conditions (Lakeman et al., 2020; Torkamani et al., 2018). Many pharmacogenetic assays employ arrays for typing single nucleotide target variants (Perreault et al., 2018), but the reliability of some results has been questioned (Lotta et al., 2021; Weedon, Wright, et al., 2021). Array genotyping is the method used for most direct-to-consumer (DTC) genetic tests, which often provide medically relevant data (e.g., on cancer risk or carrier status for recessive diseases) in addition to non-medical information such as ethnicity or individual non-medical traits. At the beginning of 2021, more than 30 million individuals had been genotyped with array technology by the two largest USA DTC genetic test companies alone (Lu et al., 2021). Various studies and recommendations have highlighted quality issues concerning the results of DTC tests with regard to both the reliability of variant calling and the communication of actionable genetic information (American College of & Gynecologists' Committee on, 2021; Horton et al., 2019; Tandy-Connor et al., 2018). A recent comparison of array genotyping with massive parallel sequencing data in the UK biobank found a high sensitivity, specificity, and positive/negative predictive values, for >100.000 common variants, but highlighted a very low reliability for genotyping very rare pathogenic variants (PVs). The authors strongly advise against using array genotyping data for health decisions without validation (Weedon, Jackson, et al., 2021).

In our current study, we took a different, more traditional approach to utilizing low-cost genotyping arrays for clinical purposes. Rather than harvesting a huge dataset for potentially relevant genetic information, we selectively interrogated a minute proportion of the available data with regard to specific indications in the diagnostic laboratory setting of a Human Genetics institute, ISO 15168-accredited for medical genetic testing. We make a distinction between diagnostic testing of specific variants validated for medical decisions, and screening for variants in candidate genes that require confirmation by a second method. Whereas validated variant testing by genotyping array may result in a stand-alone diagnostic report, the variant screening approach requires complementary investigations in many cases. An essential component of our strategy was the development of a dedicated software to facilitate analysis, and to limit the computation of array raw data to those variants that are clinically relevant and for which informed consent has been obtained from the tested individual.

2 | MATERIALS AND METHODS

2.1 | Study cohort

The total study cohort comprises 902 case samples referred to the Institute of Human Genetics, Medical University of Innsbruck, for 23 indications representing 30 different genes. All samples were analyzed in parallel by sequencing and genotyping array; the analyses in each case were restricted to the requested and consented genes or variants. DNA was extracted from peripheral EDTA blood samples in all individuals using standard methods. Sequence analyses of the coding regions as well as adjacent intron sequences of target genes were carried out either by standard massive parallel sequencing (Illumina, different instruments and enrichment kits) or by Sanger sequencing (3730xl DNA Analyzer, Applied Biosystems), combined with data analysis using the SeqNext or SeqPatient software (Sequence Pilot 5.2.0 Build 507, JSI, Medical Systems, Ettenheim, Germany).

2.2 | Array genotyping, data analysis and quality parameters

Array genotyping was performed using the commercially available Infinium Global Screening Array 24+ v.3 Kit (abbreviated GSA) with addition of the Multi-Disease Booster add-on content on the iScan array scanner system (Illumina), using procedures suggested by the manufacturer. This type of array is also used (with modifications) by the major DTC genetic test companies (Lu et al., 2021). The raw data were analyzed with the GenomeStudio 2.0 software (Illumina), using a cluster file kindly provided by P. Hoffmann (Institute of Human Genetics, University of Bonn, Germany) and the manifest file provided by the manufacturer. The manifest file records all available probes with their locations on the array, mapped against the human genome and aligned with the given NCBI unique identifiers (rs numbers, www.ncbi.nlm.nih.gov/snp/). Clustering algorithm version 3.0 generates the GenTrain score which reflects the general quality of a variant in all individuals tested. Genotyping quality for each variant in each individual case was determined using the GenCall score generated by the variant calling algorithm version 3.0. This score is relevant for both, presence or absence of this variant, and thus covers also the wild type sequence at the variant position. Variant analysis results that failed to reach a GenTrain score of 0.75, and a GenCall score of 0.2, were regarded as unreliable and denoted inconclusive.

Individual case samples were accepted for further processing if the initial call rate was 99.5% or greater. In order to identify and exclude low quality positions on the array, we decided to set stringent criteria for marking variants as reliably negative, and assessed this during an evaluation of at least 50 samples for each of the specific variants or target genes included in the current study. Variants that during evaluation were marked as inconclusive or false positive in ≥5% of samples were classified as unreliable. We either specifically examined them as such in the current study, or removed them from further analysis. In consequence, each variant position was denoted as homozygous wild type (WT), heterozygous variant, homozygous variant, or inconclusive.

For diagnostic variant genotyping, we adapted the Sequence Pilot software system for gene panel and exome analysis (JSI, Medical Systems, Ettenheim, Germany). Aim was to ensure easy and rapid analysis of the specific data generated by genotyping arrays, and to restrict the analysis to those variants or genes that are relevant for the individual indication; other non-targeted variants should not be recognizable in the diagnostic laboratory. This was regarded as essential to ensure that the analysis does not exceed consent given by the proband and prevents the generation of unwanted information.

Comprehensive massive parallel sequence analysis also entailed quantitative assessment of read numbers for the detection of large genomic rearrangements, as previously described (Povysil et al., 2017). Standard GSA data analysis using the adapted software solution did not include quantitative evaluation of the dataset for large deletions or duplications. However, quantitative analyses were carried out post hoc in all samples in which a large deletion or duplication had been detected by massive parallel sequencing, using the NxClinical (BioDiscovery, El Segundo, CA, USA) software.

3 | RESULTS

3.1 | Bioinformatics

In order to allow array genotyping for selected indications only, we developed a software solution (SeqArray, JSI Medical Systems, Ettenheim, Germany) which combines rapid and easy determination of specifically targeted variants with comprehensive quality data in the individual case. Relevant variants are identified based on candidate genes or gene panels. The software states variant genotypes as well as GenTrain and GenCall scores obtained from GenomeStudio 2.0. In addition it provides relevant background information such as transcript reference, c.HGVS, p.HGVS, dbSNP ID, classification from MutDB, and ClinVar, and allele frequencies from gnomAD, ExAC and 1000 Genomes, based on the manifest and cluster files. The results of quality assessments are visualized with different colors, allowing easy identification of low quality results for particular variants. Variants that do not fulfill the chosen quality thresholds (marked as "red" or "warning") are judged inconclusive with regard to both, variant or wild type genotypes at the particular position. The software therefore enables rapid evaluation of all relevant data limited to the specific consented indication.

3.2 | Testing for specific pathogenic variants

We identified 32 specific variants in 10 genes (OMIM numbers in parentheses) – *DPYD* (612779), *F2* (612309), *F5*, *LCT* (603202), *ALDOB* (612724), *FGFR3* (134934), *HFE* (613609), *MTHFR* (236250), *MEFV* (608107), and *SERPINA1* (608107) – that were potentially suitable for stand-alone GSA diagnostic testing (Table 1, Supplementary Table S1). The respective diseases were chosen because of their clinical relevance and the presence of a limited number of relevant disease-causing variants in our population. None of these target variants had to be removed from the study because of incorrect or inconclusive results during the evaluation phase. These variants were investigated by both, DNA sequencing and GSA analysis in indication-specific subsets of 3–35 cases (total of 212 samples). Positive, negative and inconclusive results were recorded for all variants, 23 of the 32 specific variants were observed, the other 9 variants were negative in all cases. Based on the GSA results that passed the relevant quality thresholds, there were no false negative and no false positive results, with tests denoted fully reliable (conclusive variant calls for all genotypes) in 211/212 cases investigated. Including the inconclusive result for a single variant genotype, the specificity for this type of analysis was 99.5% in our study.

3.3 | First-tier array screening for causative variants in autosomal recessive diseases

We carried out parallel analyses by array genotyping and standard sequencing of the respective genes in individuals with a clinical/biochemical diagnosis of cystic fibrosis (CF, *CFTR* gene, OMIM 602421), phenylketonuria (PKU, *PAH* gene, OMIM 612349) and eight more infrequent autosomal

TABLE 1 GSA results for specifically targeted common pathogenic disease variants in 212 individuals

Gene	Cases	Target variants	Observed variants	Correct GSA results				GSA result incorrect	GSA results inconclusive
				Hom	Comp het	Het	Hom WT		
DPYD	26	4	2	0	0	9	17	0	0
F2	21	1	1	0	0	1	20	0	0
F5	35	1	1	2	0	7	26	0	0
LCT (MCM6)	19	1	1	4	0	11	4	0	0
ALDOB	22	3	1	0	0	1	21	0	0
FGFR3	3	2	2	0	0	3	0	0	0
HFE	22	2	2	2	2	8	10	0	0
MTHFR	15	1	1	6	0	8	1	0	0
MEFV	28	15	10	1	6[a]	8	13	0	1[b]
SERPINA1	21	2	2	2	1	10	8	0	0
Total	**212**	**32**	**23**	**17**	**9**	**66**	**120**	**0**	**1**

Notes: See Supplementary Table S1 for detailed variant and transcript information. 23 target variants were observed in 92 samples. Multiplying the number of SNP targets with the number of cases results in a total of 772 individual genotypes (homozygous, heterozygous or wild type at a particular position). All variant and wild type genotypes denoted conclusive by GSA analysis ($n = 771$) were identical with the genotypes identified by standard sequencing.

Bold font is used to highlight relevant (summary) information.

Abbreviations: Comp het, compound heterozygous; Het, heterozygous; Hom, homozygous.

[a] One sample contained three heterozygous variants, another contained a homozygous variant (c.442G > C) and two heterozygous variants.

[b] An inconclusive results was obtained at one position (*MEFV* variant c.2084A > G) in a single individual who did not carry this variant, but was homozygous for variant c.2082G > A (not covered by the GSA, identified by sequencing).

recessive metabolic diseases (Table 2, Supplementary Tables S2 and S3). Comparisons involved the assessment of all PVs identified by sequencing with regard to their GSA coverage, as well as the evaluation of positions of (likely or confirmed) pathogenic variants denoted positive or inconclusive on GSA analysis with regard to the true genotypes. The GSA provides two identical probes for the common European CF variant p.Phe508del under two different names: correct c.1521_1523delCTT and incorrect c.1520_1522delTCT. The latter was disregarded, but the combined calling of both probes can serve an internal quality control purpose. During evaluation, two supposedly GSA-covered *CFTR* PVs were excluded from array analysis as the respective probes showed frequent inconclusive results. One *PAH* PV with frequent inconclusive results in control samples (p.Arg408Gln) was kept in the analysis because it is a prevalent PKU PV.

Conventional sequencing in 33 individuals with CF revealed homozygous or compound heterozygous PVs in all cases (compound heterozygosity was either confirmed through analysis of parental samples, or was assumed in case of confirmed CF and no other pathogenic variant detected by full sequencing). Based on the GSA results that passed the relevant quality thresholds, and including the two deletions confirmed by quantitative GSA analysis, 60/66 CF alleles (91%) were correctly detected by GSA analysis (see Supplementary Table S2: 4 homozygous and 52 heterozygous variants), and 27/33 genotypes (82%) were fully characterized using this method. A similar result was obtained for 39 individuals with PKU. Based on the GSA results that passed the relevant quality thresholds, 73/78 PKU alleles (94%) were correctly detected by GSA analysis (see Supplementary Table S2: 9 homozygous and 55 heterozygous variants) and 33/39 PKU genotypes (85%) were fully and reliably characterized using this method.

Comparative sequence and array analysis of 8 additional autosomal recessive metabolic disease genes in 44 randomly chosen affected individuals identified 45 PVs, of which 27 (60%) were detectable by GSA analysis. The array reliably identified all detectable variants, with no false positive or false negatives GSA results. GSA array analysis as sole diagnostic method would have fully clarified the disease-causing genotype in 21 of these cases (48%).

3.4 | First-tier variant screening for inherited breast and ovarian cancer

The GSA-24+ v.3 assay with Multi-Disease Booster sadd-on content was designed for a total of 5524 pathogenic and non-pathogenic variants in *BRCA1* (2387 variants, OMIM 113705) and *BRCA2* (3137 variants, OMIM 600185). In order to evaluate the usefulness of using this array for the diagnosis of familial breast and ovarian cancer disposition syndrome, we compared GSA and sequencing results in a cohort of 181 randomly chosen individuals referred for diagnostic or predictive testing with this indication. We restricted the analysis to variants with a MAF <5% in our

TABLE 2 Pathogenic variants identified by GSA analysis in 116 individuals with cystic fibrosis (CF, $n = 33$), phenylketonuria (PKU, $n = 39$) or one of 8 other autosomal recessive disease ($N = 44$)

	Cases	Variants		Variants on GSA		Diagnostic GSA results[a]		incorrect[b]		inconclusive[c]	
Disease genes	**No.**	**No.**	**Alleles**	**No.**	**%**	**Cases**	**%**	**Variants**	**Alleles**	**Variants**	**Alleles**
CFTR small variants	32	22	62	18	82			0	0	2	2
CFTR del/dup	3	2	4	(2)[d]	(100)						
***CFTR* total**	**33**	**24**	**66**	**20**	**83**	**27**	**82**	**0**	**0**	**2**	**2**
PAH	**39**	**26**	**78**	**22**	**85**	**33**	**85**	**2**	**2**	**4**	**12**
ACADM	6	4	12	1	25	0	0	0	0	0	0
BTD	6	5	12	4	80	5	83	0	0	0	0
CBS	3	4	6	3	75	2	67	0	0	0	0
CPT2	3	2	6	2	100	3	100	0	0	0	0
FAH	4	4	8	2	50	2	50	0	0	0	0
GAA	6	5	12	1	20	1	17	0	0	0	0
GCDH	10	15	20	9	60	3	30	0	0	0	0
GALT	6	6	12	5	88	5	86	0	0	0	0
Total excl. CFTR/ PAH	**44**	**45**	**88**	**27**	**60**	**21**	**48**	**0**	**0**	**0**	**0**
Total	**116**	**95**	**232**	**69**	**73**	**81**	**70**	**2**	**2**	**6**	**14**

Notes: See Supplementary Tables S2 and S3 for detailed variant and transcript information. CFTR gene analyses in individuals with cystic fibrosis included a total of 22 pathogenic single nucleotide variants and small deletions on a total of 62 alleles; 18 of these variants were covered on the GSA. In addition, there were two single/multiple exon deletions on 4 alleles. GSA analysis correctly identified homozygous and heterozygous genotypes for all variants covered on the array, including (by targeted quantitative analysis) the two single/multiple exon deletions. There were no false negative results for GSA-covered PVs. Inconclusive results were called for two PVs in two cases; one of them was explained by a large deletion of the whole exon, the other was unexplained. Conventional *PAH* gene sequencing in individuals with phenylketonuria identified 26 different pathogenic single nucleotide variants including 2 single nucleotide deletions; there were no large deletions or duplications. Three PVs (heterozygous in single cases) were not identified as they are no GSA target variants; an additional PV not covered by the GSA triggered an incorrect positive result for a GSA-covered different variant at the same nucleotide. Of the remaining 22 variants identified in our cohort and covered by the GSA, 21 were correctly genotyped and one was denoted inconclusive in the single variant-positive sample. Variant-negative inconclusive results were mostly caused by the variant probe previously shown to be unreliable, or presence of a different variant in close vicinity. There were no false negative results, i.e., no GSA detectable variants denoted conclusive wild type. The high proportion of diagnostic GSA results in PKU is explained by the inclusion, on the GSA, of 27 of the 29 common PVs in Europe.[16] The allele detection rate of 94% (73/78 alleles, see Supplementary Table S2) fits nicely with the predicted GSA detection rate of 93% for PKU alleles, based on the combined frequency of GSA-covered variants in European PKU patients. In contrast, several relevant cystic fibrosis alleles have so far not been included in the GSA, explaining a somewhat lower allele characterization rate of 91% (60/66 alleles, see Supplementary Table S2) in our cohort. OMIM numbers: *CFTR* 602421, *PAH* 612349, *ACADM* 607008, *BTD* 609019, *CBS* 613381, *CPT2* 600,650, *FAH* 613871, *GAA* 606800, *GCDH* 608801, *GALT* 606999.

Bold font is used to highlight relevant (summary) information.

[a] GSA results were regarded as diagnostic when variants on both alleles were correctly identified, and there were no incorrect positives; inconclusive results were disregarded.

[b] Both samples with incorrectly called *PAH* gene variants contained a different pathogenic variant at the same position/region; in one sample the other variant was covered and correctly called by the GSA analysis.

[c] Inconclusive results were associated with presence of that variant in the respective sample (1 variant), presence of another variant at the same position/region (2 variants), wild type sequence at the position/region (2 variants), or poor performance of the variant also in control samples (1 variant).

[d] Copy number assessment was not part of the routine GSA analysis and was performed post hoc only in samples with respective abnormalities identified by massive parallel sequencing.

total cohort of samples tested on the GSA (>2000 samples), and excluded all variants with inconclusive results in the evaluation phase.

Sequencing identified 18 different PVs in *BRCA1* or *BRCA2* in 31 individuals. Three of these PVs in 3 samples (approx. 10% of all PV-positive individuals) were not present on the GSA and gave negative results on GSA analysis. From the remaining 15 PVs in 28 individuals, only 9 PVs in 21 individuals were unequivocally diagnosed by GSA analysis (Table 3). Four PVs in 5 cases were correctly identified together with 1–2 wrongly ascertained variants at the same or adjacent positions that were not present in the sample (given as variant clusters in Table 3), indicating that the differentiation of variants in close vicinity is incomplete. A reverse constellation was observed at four additional positions, where three non-PVs and one variant of unknown significance (VUS) were correctly identified together with 1–2 incorrectly

TABLE 3 *BRCA1* or *BRCA2* pathogenic variants and variant clusters identified by GSA analysis in 181 individuals with suspected inherited breast and ovarian cancer syndrome

Gene	Variant cluster	Cases	GSA-called variant	Class.	Confirmed variant	Class.	Comment
BRCA1 (NM_007294.3)	1	1	c.68_69dup c.64_65del c.65T>C	C5 C5 C5	c.68_69dupAG	C5	(2) Additional variants
	2	1	c.569_570insAACG	C5	c.570C>T[a]	C2	(4) Incorrect variant
	3	1	c.676del	C5	c.676delT	C5	(1) Correct positive
	4	2	c.843_846delCTCA	C5	c.843_846delCTCA	C5	(1) Correct positive
	5	1	c.1204del	C5	c.1204delG	C5	(1) Correct positive
	6	3	c.1687C>T	C5	c.1687C>T	C5	(1) Correct positive
	7	1	c.3296del	C5	c.3296delC	C5	(1) Correct positive
	8	2	c.3481_3491del c.3481G>T	C5 C5	c.3481_3491del	C5	(2) Additional variant
	9	2	c.3511A>T	C5	c.3511A>T	C5	(1) Correct positive
	10	8	c.4183C>T	C5	c.4183C>T	C5	(1) Correct positive
	11	1	c.4837A>G c.4834_4835del c.4838_4839insC	C1 C5 C5	c.4837A>G	C1	(5) Additional variants
	12	1	c.5057A>G c.5056_5057insC	C5 C5	c.5057A>G	C5	(2) Additional variant
	13	1	c.5212G>A c.5213del	C5 C5	c.5212G>A	C5	(2) Additional variant
BRCA2 (NM_000059.4)	14	1	c.1909+1G>A	C5	c.1909+1G>A	C5	(1) Correct positive
	15	2	c.3808_3809del	C5	c.3807T>C[a]	C1	(4) Incorrect variant
	16	2	c.4258del	C5	c.4258G>T[a]	C1	(4) Incorrect variant
	17	2	c.4440T>G	C5	c.4440T>G	C5	(1) Correct positive
	18	2	c.7544C>T c.7537_7538insA	C1 C4	c.7544C>T	C1	(5) Additional variant
	19	1	c.7565C>T c.7565_7568del	C3 C5	c.7565C>T	C3	(5) Additional variant
	20	1	c.8536G>T	C5	c.8535_8538delAGAG	C5	(3) Incorrect variant
	21	1	c.8583_8584insT	C5	c.8585dupT[a]	C5	(3) Incorrect variant
	22	3	c.9976A>T c.9981A>T	C1 C3	c.9976A>T	C1	(5) Additional variant
Total	22	40					

[a] Variant not on GSA. Class. = pathogenicity classification: C1 benign, C2 likely benign, C3 variant of uncertain significance (VUS), C4 likely pathogenic, C5 pathogenic. Comments: (1) = PV correctly identified. (2) = PV correctly identified together with 1–2 wrongly ascertained variants at the same or adjacent positions. (3) = PV incorrectly ascertained as different PV. (4) = non-pathogenic variant incorrectly ascertained as PV. (5) = non-pathogenic variant or VUS correctly identified together with 1–2 wrongly ascertained (pathogenic) variants at the same or adjacent positions.

assigned (often pathogenic) variants that represented false positives. Four variants not covered on the GSA – including one PV – generated incorrect positive calls for other (pathogenic) variants in the vicinity, representing a significant risk for false positive results. Finally, in one case, one pathogenic 4 nt deletion supposedly detectable by the GSA was not called (false negative), but a different PV at the same position was assigned instead (false positive). Post hoc analysis of the data showed that the probe for the missed deletion failed to reach quality criteria (GenCall score) probably due to competitive binding, and therefore was denoted inconclusive. 141 individuals in our *BRCA1/2* cohort had normal GSA results, and the probability of a PV in these individuals was 2–3% (3/141).

3.5 | Array genotyping for familial hypercholesterolemia

Finally, we compared the results of sequence analysis and array genotyping in 392 individuals (♀ 223, ♂ 169) with elevated blood cholesterol concentrations and a suspected diagnosis of autosomal dominant familial hypercholesterolemia (FH). Target genes were the FH genes *LDLR* (OMIM 606945), *APOB* (OMIM 107730), and *PCSK9* (OMIM 607786), and the autosomal recessive hypercholesterolemia-related genes *ABCG5* (OMIM 605459), *ABCG8* (OMIM 605460), *LDLRAP1* (OMIM 605747), and *LIPA* (OMIM 613497). In addition, we sequenced parts of the *APOE* gene (OMIM 107741) for variant c.500_502del, which is not covered on the GSA. Comparisons again entailed the assessment of all PVs identified by sequencing with regard to their GSA coverage, as well as the evaluation of positions of (likely) pathogenic variants denoted positive or inconclusive on GSA analysis with regard to the true genotypes. In the GSA analysis, variants with a MAF of >5% in our total cohort of samples tested on the GSA (>2000) were disregarded. Four *LDLR* variants that were repeatedly detected as inconclusive or false positives in the array evaluation stage and were removed from the analysis and were regarded as not detectable by GSA.

All 231 samples that contained no PV in the genes studied were correctly marked as negative in the GSA analysis; 68 heterozygous PVs that cause autosomal dominant FH were present in 160 individuals (Table 4, Supplementary Tables S4 and S5). Excluding structural alterations, 38/62 FH-related PVs in our cohort (61%) were covered by the GSA; 5 other variants were detectable through false positive calls of other GSA-covered PVs. GSA-based genotyping including quantitative analysis correctly identified PVs in 107/160 PV-positive FH cases (67%), and would have led to a conclusive diagnostic result in 107/392 individuals with hypercholesterolemia (27%). In 47 individuals, 24 *LDLR* variants not present on the GSA led to wild type or inconclusive genotypes. Combined with the single *LDLRAP1*-associated hypercholesterolemia patient, variant-negative results were obtained in 279 cases, with a false negative diagnostic rate of 48/279, i.e., 17%. There was no case in which a positive variant call was associated with wild type sequence at this position.

As with the other indications, copy number analyses were not carried out routinely for the hypercholesterolemia individuals because of cost reasons. Quantitative massive parallel sequence analysis identified six single/multiple-exon deletions and one two-exon duplication in the *LDLR* gene which were confirmed by multiplex ligation-dependent probe amplification (data not shown). Post hoc quantitative analysis of GSA data detected four deletions in five individuals but failed to identify one deletion (promoter region and exon 1) and the duplication (exons 11–12). With regard to recessive causes of hypercholesterolemia, one individual in our cohort was homozygous for a PV in *LDLRAP1* not covered on the GSA, whereas four heterozygous GSA-covered variants in *LIPA*, *ABCG4* and *ABCG5* were correctly identified in six cases but were judged non-diagnostic.

4 | DISCUSSION

Our proof-of-principle study shows that a simple low-cost non-optimized genotyping array may be successfully used for reliable quality-controlled genetic analyses in a clinical setting, provided that it is integrated into an appropriate diagnostic consent and laboratory work-flow. GSA analysis reliably detected or excluded 32 specific target variants in 10 different genes in a combined total of 212 individuals. Array-based variant screening in 10 autosomal recessive conditions reliably identified 69 different variants in 116 affected individuals, with only two false positive results caused by a different pathogenic variant in the vicinity. In individuals with hypercholesterolemia, approximately two thirds of cases with molecularly confirmed autosomal dominant FH were correctly identified by array analyses. Also, the majority of PVs in individuals with suspected breast and ovarian cancer syndrome were detected by GSA although there was a considerable number of false positive calls in the vicinity of true variants in various samples. Careful assessment of variants with incorrect and inconclusive genotypes provided by GSA analysis identified technical causes of incorrect results

TABLE 4 GSA analysis results in 392 individuals with hypercholesterolemia

	Cases	Variants	Variants on GSA		Diagnostic GSA results		Incorrect PV calls[a]		Inconclusive PV calls[b]	
Disease genes	**No.**	**No.**	**No.**	**%**	**Cases**	**%**	**Variants**	**Cases**	**Variants**	**Cases**
LDLR small variants	139	60	36	58	93	67	5	6	5	16
LDLR del/dup	7	6	(4)	(66)	(5)	(71)	(2)	(2)	n.a.	n.a.
APOB	13	1	1	100	13	100	0	0	0	0
PCSK9	1	1	1	100	1	100	0	0	0	0
APOE	–	–	–	–	–	–	–	–	–	–
Total FH (aut. Dom.)	**160**	**68**	**38**	**56**	**107**	**67**	**5 (+2)**	**6 (+2)**	**5**	**16**
Recessive[c]	1	1 (+4)	(4)	(80)	0	–	0	0	0	0
No variants	231	–	–	–	–	–	–	–	–	–
Total	**392**	**69**	**38**	**56**	**107**	**27**	**5 (+2)**	**6 (+2)**	**5**	**16**

Notes: See Supplementary Table S4 for additional variant information. The GSA-covered well-known *APOB* variant p.Arg3527Gln in 13 individuals and the *PCSK9* variant p.Asp129Asn in one individual were correctly genotyped; there were no false positive or inconclusive PV results for these genes. 36 GSA-covered small *LDLR* variants (single nucleotide variants or small deletions/insertions) were correctly identified in 93 individuals. 5 PVs in 16 additional cases were called inconclusive; 4 of these variants in 12 cases were present in the sample, as shown by sequencing. These variants thus are not reliably detectable by standard GSA analysis focusing on conclusive genotypes. Five PVs not covered on the GSA gave rise to incorrect positive calls for other PVs at the same positions in six individuals. In total, 93/139 (67%) cases were correctly genotyped with a diagnostic PV. The combination of GSA analysis with targeted sequencing of all PV calls (excluding inconclusive PVs) would have led to correct PV identification in 99 cases.

Bold font is used to highlight relevant (summary) information.

Abbreviations: *LDLR* del/dup, large whole exon deletions or duplications; 4/6 copy number changes were correctly identified by post hoc quantitative GSA analysis; n.a., not applicable; PV, pathogenic variant.

[a] All incorrect results were due to variants not covered on the GSA that were incorrectly assigned as other (GSA-covered) variants at the same or adjacent positions. There were no false negative results, i.e., wild type sequence called conclusive at the position of a GSA-covered pathogenic variant.

[b] Inconclusive GSA results for pathogenic variants were observed in 16 cases. In 12 cases, they were associated with presence of the respective variant (4 different variants), in one case there was WT sequence at the respective position. In three cases with the 4 bp duplication c.1415_1418dup, the GSA analysis gave also inconclusive calls for the adjacent missense variant c.1414G > T not present in the samples.

[c] Recessive disease genes analyzed: *LDLRAP1*, *ABCG5*, *ABCG8*, *LIPA*. Only one sample showed a homozygous pathogenic variant for *LDLRAP1*, which was not covered by the GSA. Four other recessive variants (GSA covered) were coincidentally detected on 6 alleles; as these variants were non-diagnostic, the respective figures are given in parentheses in the table and disregarded in the summary total.

(Supplementary Text S6) which should be taken into consideration in the systematic design of genotyping arrays for diagnostic purposes.

Our study allows some general conclusions with regard to the accuracy and reliability of genotyping arrays in a diagnostic setting:

- *Validated prevalent single nucleotide variants can be reliably tested with a standard genotyping array*. We show that this approach may replace other methods for specific testing of well-characterized high- or moderate-impact variants that are relevant in medical practice. Provided that adequate quality assurance measures are observed and inconclusive results are recognized, this type of analysis should have a high sensitivity and specificity not inferior to other genotyping approaches.
- *No incorrect wild type results for validated single nucleotide variants*. 154 out of 160 PVs in 902 individuals of our study were correctly identified by standard GSA analysis, 6 variants were not tested positive because they failed quality requirements. An unambiguous quality-controlled wild type designation appears to reliably exclude the presence of the respective variant in that sample.
- *Distinguishing variants at the same or adjacent positions is challenging*. All incorrect variant calls in our study were caused by the presence of another variant at the same or neighboring nucleotide. Identification of a rare genetic variant thus generally requires independent verification if there is a significant probability for another variant at the same position. Specific array design with several probes for the same variant may further reduce the probability of false positive results.
- *Testing many adjacent variants in close proximity causes incorrect results*. Abnormal *BRCA1/2* genotyping results in 8/22 positions in our study represented clusters of 2–3 array-positive variants of which only one variant was correct. No false positive result was associated with the wild type sequence in the tested samples. False positive results, therefore, may not lead to diagnostic errors as long as all positive results – including presumably benign variants – are confirmed by another method.

4.1 | Using genotyping arrays in a clinical setting

The strengths and limitations of array genotyping identified in our study have consequences for the utilization of this method for different clinical indications:

- *Testing for specific genetic risk variants*. Special care must be taken if array genotyping is used as a stand-alone test for specific variants. False negative (wild type) results appear to be exceptionally unlikely as long as true positive and true wild type results are reliably distinguished from inconclusive results that must be further investigated or disregarded. With single probes, the probability of false positive results depends on the likelihood of other variants in the vicinity, which is highly variable and differs between populations (as an example see thrombophilia variant *F5* "Leiden", Supplementary Text S7).
- *Diagnosis of monogenic diseases*. Array genotyping of prevalent gene variants may be a useful screening method in individuals with a suspected monogenic diseases, provided that it is integrated into a comprehensive multi-step diagnostic concept. Without improved probe design, all variant-positive results need to be confirmed e.g., by Sanger sequencing, and non-diagnostic results should be followed with complete gene sequencing to avoid false negative results.

Diagnostic genotyping for monogenic diseases with the non-optimized GSA array worked best when there is a limited number of relatively frequent pathogenic variants, as e.g., in PKU (Zschocke, 2003). We fully characterized the disease genotype in >80% of individuals with CF or PKU. Full sequencing of GSA negative cases would only have been required in 15–20% of affected individuals in our population, representing a substantial cost saving. The approach is less well suited for the genetic diagnosis of more heterogeneous conditions with a low a priori probability of a monogenic cause. Covering a large number of variants increases the probability of probe interference, i.e., there are more false positives and negatives when many probes (and target SNV) are located in proximity to each other.

- *Monogenic CNV analysis*. The GSA array has not been optimized for the reliable detection of deletions or duplications of single or few exons in particular genes. This application is also hampered by the inclusion of a very large number of probes for variants that are unlikely to be present in a given sample. Nevertheless, standard CNV analysis of the GSA data correctly identified 6/7 deletions present in the samples studied. Optimization of probes for all target exons should further improve this type of analysis also for the detection of small duplications.
- *Bioinformatic and ethical considerations*. For effective integration in diagnostic routine, it is mandatory to limit the visible data to specifically targeted variants that are relevant for the investigated individual, and for which informed diagnostic consent has been obtained. This approach – i.e., the exclusion of non-warranted laboratory results from assessment – is well established e.g., for acylcarnitine analysis in newborn screening in many countries (Gemeinsamer Bundesausschuss, 2013). In this context, abnormal concentrations of biomarkers that are not relevant for the target diseases remain concealed to avoid irritating or adverse screening outcomes. In line with European recommendations regarding opportunistic genomic screening (de Wert et al., 2021) we would argue that for diagnostic purposes it is neither appropriate nor necessary to generate large lists of clinically significant variants without immediate relevance to be handed out to an individual. Sufficient genetic counseling resources must be available to translate the results of genetic tests to the investigated person, who must also have had the opportunity to refrain from testing prior to the analysis.

4.2 | Weaknesses of our study

This proof-of-principle study evaluated only a minute proportion of the data generated on the array used, and the number of samples assessed for each individual variant was limited. Considering the clear separation of positions marked as wild type or heterozygous variant in our study, we believe that the array results are likely reliable also for other variants on the array. Nevertheless, this requires confirmation by examination of a large number of individual variants. In addition, there were only very few samples with rare variants in a homozygous state, thus the evidence for a reliable separation of heterozygous and homozygous variants is limited. Finally, we used a simple non-optimized array which has not been adapted for diagnostic purpose; improved variant-specific probe design should allow reliable genotyping also for challenging variant constellations.

5 | CONCLUSION

Genotyping arrays can be integrated into a quality-controlled diagnostic workflow. The technological advantage of array genotyping – compared to sequencing approaches – is the parallel analysis of numerous genetic variants covering a broad range of genes and indications at low cost, with limited bioinformatics and data storage requirements. It is essential to confirm non-optimized array results with an independent second method unless the variant genotypes have been thoroughly validated. Care must be taken to limit the analyses to relevant indications for which informed consent has been obtained. Rapid indication-based read-out of previously generated data combined with semi-automated reporting with medical overview make the approach suitable for high-volume low-cost clinical testing. Optimization of medically relevant array content and probe design should substantially improve the usefulness of array diagnostics. Ethical issues regarding data storage, access, and protection, in real-life constellations remain to be addressed.

AUTHOR CONTRIBUTIONS

Conceptualization, supervision, validation: Johannes Zschocke and Martina Witsch-*Baumgartner; laboratory investigations*: Simon Schnaiter, Franziska Kollmann, Silja Burkhard, Rebekka Gröbner, Beatrix Mühlegger, Esther Schamschula; *data curation*: Martina Witsch-Baumgartner, Simon Schnaiter, Franziska Kollmann, Silja Burkhard, Rebekka Gröbner, Beatrix Mühlegger, Esther Schamschula, Gunda Schwaninger, Johannes Zschocke; *software development*: Peter Kirchmeier; *writing-review & editing*: all authors.

ACKNOWLEDGMENTS

The data related to familial breast and ovarian cancer syndrome was generated in the group of Prof. Katharina Wimmer at the Institute of Human Genetics Innsbruck. Results of the molecular investigations in familial hypercholesterolemia were included in the Medical Diploma thesis of F.K. (Medical University Innsbruck, 2021). We wish to thank Prof. T. Meitinger, Munich, for helpful comments on an earlier draft of the manuscript.

FUNDING INFORMATION

The study was carried out with funds of the Institute of Human Genetics, Medical University Innsbruck, Austria.

CONFLICT OF INTEREST

M.W.B., G.S., S.S., F. K., S.B., R.G., B.M., E.S. and J.Z. declare no conflict of interest with regard to the current study. P.K. is employee of JSI Medical Systems, Ettenheim, Germany, which offers the software solution developed in the current study as commercially available SeqArray module complementing other genetic analysis software.

ETHICS STATEMENT

The data used in the study were generated in the standard diagnostic process with full informed consent from the investigated individuals. For study purposes, variant and genotype data were pseudonymized by removing individual personal, clinical and other information, and fully anonymized after completion of the comparative analyses. The analyses were approved by the Research Ethics Committee of the Medical University Innsbruck (votum 1320/2020).

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Johannes Zschocke https://orcid.org/0000-0002-0046-8274

REFERENCES

American College of Obstetricians and Gynecologists. (2021). Consumer testing for disease risk: ACOG Committee Opinion, Number 816. *Obstetrics and Gynecology*, *137*(1), e1–e6. https://doi.org/10.1097/AOG.0000000000004200

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

de Wert, G., Dondorp, W., Clarke, A., Dequeker, E. M. C., Cordier, C., Deans, Z., van El, C. G., Fellmann, F., Hastings, R., Hentze, S., Howard, H., Macek, M., Mendes, A., Patch, C., Rial-Sebbag, E., Stefansdottir, V., Cornel, M. C., Forzano, F., & European Society of Human Genetics. (2021). Opportunistic genomic screening. Recommendations of the European Society of Human Genetics. *European Journal of Human Genetics*, *29*(3), 365–377. https://doi.org/10.1038/s41431-020-00758-w

Gemeinsamer Bundesausschuss. (2013). Richtlinie der Gendiagnostik-Kommission (GEKO) für die Anforderungen an die Inhalte der Aufklärung bei genetischen Untersuchungen zu medizinischen Zwecken gemäß 23 Abs. 2 Nr. 3 GenDG. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, *56*(2), 326–331. https://doi.org/10.1007/s00103-013-1675-8

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108. https://doi.org/10.1038/nrg1521

Horton, R., Crawford, G., Freeman, L., Fenwick, A., Wright, C. F., & Lucassen, A. (2019). Direct-to-consumer genetic testing. *BMJ*, *367*, l5688. https://doi.org/10.1136/bmj.l5688

Katsanis, S. H., & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews. Genetics*, *14*(6), 415–426. https://doi.org/10.1038/nrg3493

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. https://doi.org/10.1093/nar/gkp552

Lakeman, I. M. M., Rodriguez-Girondo, M., Lee, A., Ruiter, R., Stricker, B. H., Wijnant, S. R. A., Kavousi, M., Antoniou, A. C., Schmidt, M. K., Uitterlinden, A. G., van Rooij, J., & Devilee, P. (2020). Validation of the BOADICEA model and a 313-variant polygenic risk score for breast cancer risk prediction in a Dutch prospective cohort. *Genetics in Medicine*, *22*, 1803–1811. https://doi.org/10.1038/s41436-020-0884-4

Lotta, L., Langenberg, C., Wareham, N. J., & Farooqi, I. S. (2021). Reply to unreliability of genotyping arrays for detecting very rare variants in human genetic studies: Example from a recent study of MC4R. *Cell*, *184*(7), 1652–1653. https://doi.org/10.1016/j.cell.2021.03.014

Lu, C., Greshake Tzovaras, & B., Gough, J..(2021). A survey of direct-to-consumer genotype data, and quality control tool (GenomePrep) for research. Computational and Structural Biotechnology Journal, *19*(7), 3747–3754. https://doi.org/10.1016/j.csbj.2021.06.040

Perreault, L. P. L., Zaid, N., Cameron, M., Mongrain, I., & Dube, M. P. (2018). Pharmacogenetic content of commercial genome-wide genotyping arrays. *Pharmacogenomics*, *19*(15), 1159–1167. https://doi.org/10.2217/pgs-2017-0129

Povysil, G., Tzika, A., Vogt, J., Haunschmid, V., Messiaen, L., Zschocke, J., Klambauer, G., Hochreiter, S., & Wimmer, K. (2017). panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. *Human Mutation*, *38*(7), 889–897. https://doi.org/10.1002/humu.23237

Tandy-Connor, S., Guiltinan, J., Krempely, K., LaDuca, H., Reineke, P., Gutierrez, S., Gray, P., & Tippin Davis, B. (2018). False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genetics in Medicine*, *20*(12), 1515–1521. https://doi.org/10.1038/gim.2018.38

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x

Verlouw, J. A. M., Clemens, E., de Vries, J. H., Zolk, O., Verkerk, A. J. M. H., Zehnhoff-Dinnesen, A. A., Medina-Gomez, C., Lanvers-Kaminsky, C., Rivadeneira, F., Langer, T., van Meurs, J. B. J., van den Heuvel-Eibrink, M. M., Uitterlinden, A. G., & Broer, L. (2021). A comparison of genotyping arrays. *European Journal of Human Genetics*, *29*, 1611–1624. https://doi.org/10.1038/s41431-021-00917-7

Weedon, M. N., Jackson, L., Harrison, J. W., Ruth, K. S., Tyrrell, J., Hattersley, A. T., & Wright, C. F. (2021). Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ*, *372*, n214. https://doi.org/10.1136/bmj.n214

Weedon, M. N., Wright, C. F., Patel, K. A., & Frayling, T. M. (2021). Unreliability of genotyping arrays for detecting very rare variants in human genetic studies: Example from a recent study of MC4R. *Cell*, *184*(7), 1651. https://doi.org/10.1016/j.cell.2021.03.015

Wetterstrand, K. A. (2020). *DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP)*. National Human Genome Research Institute (NHGRI). Retrieved from www.genome.gov/sequencingcostsdata

Zschocke, J. (2003). Phenylketonuria mutations in Europe. *Human Mutation*, *21*(4), 345–356. https://doi.org/10.1002/humu.10192

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Witsch-Baumgartner, M., Schwaninger, G., Schnaiter, S., Kollmann, F., Burkhard, S., Gröbner, R., Mühlegger, B., Schamschula, E., Kirchmeier, P., & Zschocke, J. (2022). Array genotyping as diagnostic approach in medical genetics. *Molecular Genetics & Genomic Medicine*, *10*, e2016. https://doi.org/10.1002/mgg3.2016