Design and validation of a metabolic disorder resequencing microarray (BRUM1).

Christopher Bruce, Matthew Smith, Fatima Rahman, Zhi-Feng Liu, Dominic Mcmullan, Sarah Ball, Jane Hartley, Marian A Kroos, Lesley Heptinstall, Arnold Jj Reuser, et al.

Human Mutation

Design and validation of a metabolic disorder resequencing microarray (BRUM1).

Journal:	*Human Mutation*
Manuscript ID:	humu-2009-0430.R1
Wiley - Manuscript type:	Methods
Date Submitted by the Author:	23-Feb-2010
Complete List of Authors:	Bruce, Christopher; The University of Birmingham, Clinical and Experimental Medicine Smith, Matthew; Birmingham Women's Hospital, West Midlands Regional Genetics Laboratory and Clinical Genetics Unit Rahman, Fatima; The University of Birmingham, Clinical and Experimental Medicine Liu, Zhi-feng; Nanjing Medical University, Department of Digestory McMullan, Dominic; Birmingham Women's Hospital, West Midlands Regional Genetics Laboratory and Clinical Genetics Unit Ball, Sarah; Birmingham Children's Hospital Hartley, Jane; The University of Birmingham, Clinical and Experimental Medicine Kroos, Marian; Erasmus MC, Clinical Genetics Heptinstall, Lesley; Royal Manchester Children's Hospital, Willink Biochemical Genetics Unit Reuser, Arnold; Erasmus MC, Clinical Genetics Rolfs, Arndt; Universität Rostock, Albrecht-Kossel-Institut fur Neuroregeneration Hendriksz, Chris; Birmingham Children's Hospital, The Metabolic Unit; Birmingham Children's Hospital, The Metabolic Unit Kelly, Deirdre; Birmingham Children's Hospital, The Liver Unit Barrett, Timothy; The University of Birmingham, Clinical and Experimental Medicine Macdonald, Fiona; Birmingham Women's Hospital, West Midlands Regional Genetics Laboratory and Clinical Genetics Unit Maher, Eamonn; The University of Birmingham, Clinical and Experimental Medicine; Birmingham Women's Hospital, West Midlands Regional Genetics Laboratory and Clinical Genetics Unit Gissen, Paul; The University of Birmingham, Clinical and Experimental Medicine; Birmingham Children's Hospital, The Metabolic Unit

Design and validation of a metabolic disorder resequencing microarray (BRUM1).

CK Bruce[1], M Smith[2], F Rahman[1], ZF Liu[3], DJ McMullan[2], S Ball[4], J Hartley[1,5], MA Kroos[6], L Heptinstall[7], AJJ Reuser[6], A Rolfs[8], C Hendriksz[9], DA Kelly[5], TG Barrett[1,10], F MacDonald[2], ER Maher[1,2], P Gissen[1,9]

[1]School of Clinical and Experimental Medicine, The University of Birmingham;

[2]West Midlands Regional Genetics Service, Birmingham Women's Hospital;

[3]Department of Digestory, Nanjing Children's Hospital, Nanjing Medical University;

[4]Department of Clinical Chemistry, Birmingham Children's Hospital, Birmingham, UK;

[5]The Liver Unit, Birmingham Children's Hospital, Birmingham, UK;

[6]Department of Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands;

[7]Royal Manchester Children's Hospital, Manchester, UK;

[8]Albrecht-Kossel-Institut für Neuroregeneration, Universität Rostock, Rostock, Germany;

[9]The Metabolic Unit, Birmingham Children's Hospital, Birmingham, UK;

[10]The Diabetes Unit, Birmingham Children's Hospital, Birmingham, UK.

Corresponding Author: Dr. P. Gissen.

Tel: +44 121 415 8529

Email: p.gissen@bham.ac.uk

Abstract

The molecular genetic diagnosis of inherited metabolic disorders is challenging. The diseases are rare and most show locus heterogeneity. Hence testing of the genes associated with IMDs is time consuming and often not easily available. We report a resequencing array that allows the simultaneous resequencing of up to 92 genes associated with IMDs.

To validate the array, DNA samples from 51 patients with 52 different known variants (including point variants, small insertion and deletions (indels)) in 7 genes (*C14ORF133, GAA, NPC1, NPC2, VPS33B, WFS1, SLC19A2*) were amplified by PCR and hybridised to the array. A further patient cohort with 48 different mutations in *NPC1* were analysed blind. Out of 76 point variants, 73 were identified using automated software analysis followed by manual review. Ten insertion and deletion variants were detected in the extra tiling using mutation specific probes whilst 11 heterozygous deletions and 3 heterozygous insertions.

In summary, we identified 96% (95%CI 89-99%) of point variants added to the array, but the pickup rate reduced to 83% (95%CI 75-89%) when insertions/deletions were included. Whilst the methodology has strengths and weaknesses, application of this technique could expedite diagnosis in most patients with multi-locus IMDs.

Introduction

A number of inherited diseases present with a similar phenotype, but may be caused by mutations within different genes. For example, amongst the lysosomal storage disorders, patients with genetically distinct mucopolysaccharidoses have common clinical features (Neufeld and Muenzer, 2001). This also applies to the group of neuronal ceroidlipofuscinoses (Jalanko and Braulke, 2009). Further, the severe hepatic neurodegenerative disorder Niemann-Pick Type C is caused by mutations in two genes *NPC1* and *NPC2.* Although these forms are indistinguishable clinically, the management of patients with *NPC2* may include bone marrow transplantation as well as other therapeutic options available in this disease (Wraith et al., 2009). The neuromuscular and the mitochondrial diseases are two other groups of diseases in which the clinical phenotype leaves choices as to the underlying metabolic and causative genetic defects. Thus the availability of a rapid accurate molecular diagnostic platform could enhance acute clinical management as well as genetic counselling and prenatal diagnosis.

Current sequencing technologies employed in routine diagnostic laboratories rely on Sanger sequencing, and whilst providing the gold standard for the detection of small intragenic mutations, most diagnostic laboratories do not provide a service offering rapid analysis of multiple candidate genes (generally candidate genes are sequenced sequentially until a mutation is identified). In addition, there is a limit to how much sequence can be obtained from di-deoxysequencing; often, each exon has to be assayed as a single experiment. Second generation sequencing technologies can enable multiple gene testing in a single experiment, but it is not yet clear how suitable this approach is for routine testing in a clinical service laboratory. Microarray based resequencing is an evolving laboratory technique that offers the potential for rapid simultaneous mutation testing of multiple genes (Hacia, 1999). This technology has been used for pathogen identification (Malanoski et al.,

2006) bacterial genotyping (Corless et al., 2008; Zwick et al., 2008), mitochondrial DNA sequencing (Hartmann et al., 2009), and human gene mutation identification (Denning et al., 2007; Liu et al., 2007; Takahashi et al., 2008). One group has designed a resequencing chip, the Jaundice Chip, that is being offered as a diagnostic test to investigate inherited syndromes of intrahepatic cholestasis ((Liu et al., 2007) http://www.cincinnatichildrens.org/svc/alpha/m/molecular-genetics/jaundice-chip.htm). Microarray based resequencing could be a promising technology for many molecular genetic diagnostic laboratories as the equipment is often available in clinical and research laboratories and can be used for clinical purposes. However, the mutation detection rate, the reproducibility, and the accuracy of the resequencing technology have not been explored in detail. Therefore we used a custom designed resequencing microarray to address these questions and to elucidate areas of potential improvement for the resequencing technology and microarray analysis. The novel array (Birmingham ReseqUencing Microarray version 1(BRUM1)) includes 92 genes involved in metabolic pathways (or encoding proteins involved in various aspects of metabolism) and enables simultaneous resequencing of multiple genes that may each cause similar phenotypes within a group of disorders (eg hyperlipidaemia, glycogen storage disease and lysosomal storage diseases).

Materials and Methods

DNA Samples

A total of 51 genomic DNA samples harbouring disease causal gene mutations in *NPC1, NPC2, VPS33B, C14ORF133, GAA, SLC19A2,* and *WFS1* (Supp. Table S1) were obtained from the West Midlands Regional Genetics Service (Birmingham, UK),

Birmingham Children's Hospital (Birmingham, UK), Willink Biochemical Genetics Unit (Manchester, UK), and the Erasmus Medical Center (Rotterdam, NL). Genes and sample DNA were chosen based on the availability of DNA covering a wide range of mutation types. DNA from a human cell line MRC5VA was used as a control.

A further cohort of 28 DNA samples that harboured 48 mutations detected by direct sequencing within *NPC1* were obtained from Albrecht-Kossel-Institute for Neuroregeneration (Rostock, Germany). Only the exon number was known prior to testing the arrays but not the nature or precise location of the mutation.

Microarray Resequencing Chip Design

A total of 92 genes involved in various inherited disorders (Supp. Table S2) were selected to be sequenced on a 300kb resequencing microarray (Affymetrix). All coding exons with the addition of 25bp intronic sequence either side of the exon were selected. This allows for full sequencing of the exon, plus 13bp of intronic sequence allowing splice site mutations to be identified. The design of the array is such that there are a series of 25bp probes designed against the supplied reference sequences. As the array simultaneously sequences both forward and reverse strands, there are 8 probes for any given base; 4 for the forward and 4 for the reverse. The central base differs between the four probes allowing all variants to be identified. Sequencing commences at the 13thbp of the reference and completes 13thbp away from the end of the reference sequence.

Small insertions/ deletions and indels, especially in a heterozygous state are known limitations for resequencing arrays, but by including specific probes for known insertions and deletions on the array, these limitations can be overcome (Karaman et al., 2005; Kothiyal et al., 2009). Thus, 1343 known insertions/ deletions and indels (all of 5 or less nucleotides)

were identified from the Human Genome Mutation Database (www.hgmd.org) and included on the array. Thus this array is divided into two sections, the first herein termed main tiling is used to primarily detect single nucleotide variations, whilst the second section herein termed extra tiling, is used to detect insertion/ deletion variations. The sequences of all the genes were first analysed for repeat regions using repeat masker (http://repeatmasker.org/cgi-bin/webrepeatmasker) and any repeats or low complexity regions of >25bp were excluded. The sequences were submitted to Affymetrix for final array design and production. In total, 232 515bp can be sequenced on this array.

Microarray resequencing.

In order to efficiently test the arrays with 52 mutations, a combination of short and long-range PCR amplifying individual exons of 7 genes (*C14ORF133, GAA, NPC1, NPC2, SLC19A2, VPS33B* and *WFS1)* was undertaken. For exons without a known mutation, control DNA was used. Each fragment was independently amplified in triplicate allowing comparisons of the sequencing to be undertaken. Pooled PCR products were labelled PCR-A, PCR-B or PCR-C. Triplicate hybridisations for each pool were labelled PCR-A1, PCR-A2 etc. PCR was undertaken in 50µl volumes using 50ng of DNA, 0.5µM each primer (Sigma), 0.5mM dNTPs (Bioline) and 2 units of Taq DNA polymerase (Biomix Red, Bioline UK) in 1x PCR buffer. Primers were designed using Exon Primer (http://ihg2.helmholtz-muenchen.de/ihg/ExonPrimer.html; primers available on request) or had been previously published (Gissen et al., 2004; Cullinane et al., 2009; Cullinane et al., 2010). A 1kb control fragment was amplified using a template supplied by Affymetrix. The size range of amplicons used in this test ranged from 156bp to 4500bp with 250pmoles of each amplicon required for analysis. PCR product quantitation was undertaken using a picogreen assay, pooling, fragmentation, labelling and hybridisation were performed

according to the GeneChip Custom Resequencing Array Protocol V2.1. Arrays were washed and stained using a FS450 fluidics station before being scanned with a GCS3000 7G scanner.

BRUM1 Data Analysis.

Intensity files were produced using AGCC (Command Console V1.0) and processed in GSeq 4.1 (Affymetrix) that uses the resequencing algorithm version 2. Several parameters within this base calling can be altered which thus has an effect on the call rate and accuracy of the base calls (see Cutler et al., (2001) and Di and Cawley, (2005) for full details of the algorithm). Increasing the threshold on 2 parameters, the Quality Score Threshold (QST) and Base Reliability Threshold (BRT) will decrease the number of calls, thus making the data more stringent, whereas increasing the threshold on the sequence profile threshold and trace threshold will increase the number of calls, thus making the data less stringent. A further two parameters, the Max Signal to noise ratio and Modeltype have no effect on increasing or decreasing the call rate. The Modeltype allows the choice between haploid and diploid data. Haploid fits the data to 5 outcomes (A,G,C,T and N) whereas the diploid fits data to 11 outcomes (A,G,C,T,N,K,M,R,S,W, and Y) allowing heterozygous calls to be made. Altering the weak signal and no signal thresholds has an indeterminate effect on base calling. For the purpose of the analysis, base calling utilised the diploid model, whilst signal to noise ratio, no signal threshold, sequence profile threshold and weak signal fold threshold were left at default values. A range of quality score threshold, a measure of the reliability of the base call (QST; 2,3,6,9,12,30) was used. As increasing the QST, makes the data more stringent a base with a higher quality score is deemed to be more reliable and more accurate. The effect of changing the base reliability threshold (BRT) was also assessed, using either 0 or 0.5 (if 50% of samples no-called at this base, all samples would be no-called).

Di-deoxysequencing.

To compare the microarray based sequencing with direct sequencing, surplus PCR products were directly sequenced. PCR product were treated with ExoSap IT (GE Healthcare) before being directly sequenced using BigDye 3.1 (Applied Biosystems) Following precipitation and washing, sequencing products were HiDiformamide (Applied Biosystems) and analysed using a DNA analyzer3730xl (Applied Biosystems). The resulting sequence traces were analysed using Sequence Analysis 5.2.2 (Applied Biosystems).

Statistics.

In order to assess the reliability and reproducibility of the microarray, Fleiss Kappa statistic was used to compare base calls between samples and also to compare the base calls against di-deoxysequencing (Fleiss, 1971). Fleiss Kappa allows comparisons between multiple raters (here each BRUM1 chip) to assess their agreement. This was implemented in Microsoft Excel (King, 2004). Confidence intervals for proportions were calculated at Measuring Usability (http://www.measuringusability.com/wald.htm) using the Adjusted Wald method (Sauro and Lewis, 2005)

Results

Chip Assembly.

Analysing the sequences prior to BRUM1 array assembly, there were eight regions that proved to have repetitive DNA sequences of >25bp. These were located in *RET, IDUA, NAGLU, ARSB, PTPN11, EIF2AK3, APOB* and *PCSK9.* Repeats shorter than 25bp can be safely tiled onto the array (GeneChip Custom Resequencing Array Design Guide). However, this means that the above genes cannot be fully sequenced, resulting in 30bp of *RET* (CTG repeat), 72bp of *IDUA* (GC rich low complexity), 147bp of *NAGLU* (GC rich, low

complexity), 42bp of *ARSB* (GC rich, low complexity), 31bp of *PTPN11* (CAAAAA repeat), 75bp of *EIF2AK3* (CTG and CGG repeats), 38bp of *APOB* (CTG repeat) and 38bp of *PCSK9* (CTG repeat) being excluded. Further during the design phase, all sequences were compared to each other to predict cross hybridisation that resulted in 927 submitted sequences having a homology of greater than 10% (range 10.07%-100%). All the sequences that demonstrated homology were between the parental sequences and extra tiling included to detect insertion and deletion variants or between the extra tiling itself. For example, NPC1#01 (NPC1 exon 1) and NPC1#M21 (used to detect NPC1 46-47delTG) have 15% homology. However, this observation was to be expected.

The final BRUM1 chip comprises 1.8million 25bp probes generated as 1.525 micron features and has the capabilities of sequencing 232 515 bp in a single experiment.

Call Rate and accuracy of the BRUM1.

The call rate was assessed under several conditions with and without the BRT (Table 1). The overall call rate for the BRUM1 array was low, ranging from 0.05% for a QST of 30 and BRT of 0.5 to 31% for a QST of 2 and BRT of 0. However, only a portion of the BRUM1 array was used, approximately 15 890 out of the total 232 515 which reflects in this low call rate. Assessing the average call rate for the genes sequenced, an average call rate of 0.25% for a QST of 30 with a BRT of 0.5 was observed, rising to 94.25% for a QST of 2 and BRT of 0. This equates to 15815bp that could not be assigned unambiguously at a QST of 30, dropping to 912bp at a QST of 2.

The accuracy of base calling was better at a higher QST being 100% with a QST of 30. However, using a QST of 2, the accuracy of base calling dropped to 99.45%. This indicates that 87bp out of the 15 890bp sequenced may be called incorrect. Using a QST of 2

without the BRT gave the highest call rate with only a marginal increase in incorrect base calls. Therefore these settings were used in all subsequent analysis. Quality scores ranged from 0-45.

Reproducibility.

In order to assess the reproducibility of the assay, triplicate PCR's and triplicate hybridisation experiments were performed. Each hybridisation challenge comprised the same fragments either from independent PCR's (A,B and C) or from the same PCR pool (A1, A2, A3; B1, B2, B3; C1, C2, C3). Base calling was undertaken with the above settings and the resulting sequences exported. Resulting sequences from arrays comprising either PCR-A, PCR-B or PCR-C were aligned to the reference sequences and compared using Fleiss Kappa statistics (Table 2). In total 15880bp were compared between each array for the triplicate PCR pool. PCR-A had the lowest agreement with 91.7% bases in agreement (Kappa=0.895 CI 0.891-0.898). PCR-B and PCR-C were very similar with 96.8% (Kappa=0.959 CI 0.955-0.962) and 97.3% (Kappa=0.966 CI 0.962-0.970) bases in agreement. Analysing the same 15 880bp between all nine arrays, 94.8% (Kappa=0.932 CI 0.933-0.935) were in agreement. However, it must be noted that two fragments failed to hybridise efficiently from PCR-A, which is reflected in the lower agreement between the repeats.

Consistent regions of no-calls.

Certain probes, for example those containing increased numbers of Gs, can be problematic for resequencing arrays resulting in decreased signal and an increased chance of being no-called (Cutler et al., 2001), To assess whether there were any consistent regions of no-calls, all fragments were aligned. Any base or region no-called in all 9 BRUM1s was defined as a region of consistent no-call. Table 3 gives an overview of the consistent regions

of no-call. Interestingly, the most common base consistently no-called was T accounting for 19.5% of all no-called regions. The most common multiple base no-called region were regions containing multiple C's accounting for 12.7% of these regions. Analysis of the nucleotide content of the probes showed that probes with higher than average GC content were more likely to be no-called. Further as the proportion of G bases increased in a single probe, there was an increased chance of the base being no-called.

Comparison of BRUM1 array based sequencing to di-deoxysequencing.

Direct sequencing of the DNA used on the microarray was undertaken to determine whether there was any difference when the two methods were compared. For the purpose of automated analysis of di-deoxysequencing, a heterozygous call was made if the lower peak height was >40% of the peak height of the main peak. A total of 8147bp from 58 different fragments were compared to BRUM1 array sequencing with an overall agreement of 93.1% (Kappa =0.91 CI 0.90-0.92). There were 39 discordant calls between the two methodologies, 385 were no-called in either methodology with 11 no-calls being common between both methodologies. A further 30 were a heterozygous call in one methodology, but with a related homozygous call in the other. There was an overall agreement of 93% between automated analysis of di-deoxysequencing to the reference sequence.

However, following manual review of both di-deoxysequencing and BRUM1 sequencing, 112 bases were removed from the analysis due to low quality di-deoxysequencing. These low quality bases were confined at the beginning or end of the di-deoxysequencing. Comparison of the remaining 8035 bases demonstrated an agreement of 99.7% (Kappa 0.99 CI 0.98->0.99). There were no discordant calls, 5 bases were no-called in either methodology and 22 bases were a heterozygous call in one methodology but related homozygous in the other. Comparison of the reviewed di-deoxysequencing to the reference

sequence there was an agreement of 98.9% (Kappa 0.99 CI 0.97->0.99). There were 8 discordant calls, 8 no calls and 21 heterozygous calls.

Detection of Variation: single nucleotide variations.

The main purpose of the resequencing microarray was to detect mutations. In order to assess the ability of the array to detect mutations, we initially amplified the DNA known to contain 39 point mutations/ polymorphisms, which was then hybridised to the array. Using automated analysis, 32 out of 39 point variations were detected demonstrating a pickup rate of 82.1% (95%CI 67.0-91.3%). Of the changes not picked up by automated base calling, 6 were manually called increasing the pickup rate to 97.4% (95%CI 85.6-99.9%). In addition, 4 false positive calls and 5 known polymorphisms were identified (Table 4).

Detection of Variation: insertions and deletions.

As insertions and deletions are known problems for resequencing arrays, the majority of the known insertions and deletions were tiled separately on the array (Karaman et al., 2005; Kothiyal et al., 2009). To test how well the technology detects insertions and deletions, 3 samples with known insertions and 10 samples with known deletions were amplified and hybridised to the array. Analysing the main tiling, six deletions that were in a heterozygous state and one deletion in a homozygous state were not detected. However, 3 other homozygous deletions were detected and manifested as regions of no-call. Furthermore, 1 homozygous insertion manifested as a compound heterozygote but no other insertions were identified. However, when analysing the extra tiling all insertions and 6 of the 10 deletions were easily identified. Three heterozygous deletions were not identified. Overall, 48 out of 52 variations were detected representing a pickup rate of 92.3% (95%CI 84.0-97.2%)

Detection of Variation: Blind test of BRUM1 sequencing NPC1.

To assess how the array performed when the primary investigators had no prior knowledge of the mutations (except for exonic location), 28 NPC1 samples were analysed for 48 variations (37 were single nucleotide changes, 7 were novel heterozygous deletions, 3 were novel heterozygous insertions and 1 was an extra tiled deletion). Of these variations 35 were detected representing a pickup rate of 72.3% (95% CI 58.9-83.5%). Of these changes we could detect 35/37 (94.6%) of point mutations, however, we were unable to detect any insertion or deletions.

Discussion

This paper describes the development and validation of a novel microarray based resequencing platform that has been designed to sequence a number of genes involved in inherited metabolic disorders simultaneously. In order to validate the BRUM1 array, 7 genes were chosen based on the availability of samples with known mutations. A range of mutations were chosen such that a broad spectrum of mutation types would be tested, i.e. homozygous and heterozygous point mutations, homozygous insertions and heterozygous/homozygous deletions. Whilst this would give an indication of mutation pick up rate, and which mutations type would cause technical problems, it was also important to determine what GSEQ algorithm parameters gave the best base call with minimal impact on accuracy, and further investigate reproducibility and comparison to di-deoxysequencing.

Array design.

During the design phase of the array, it was noted that several sequences had repetitive or low complexity stretches of DNA that could not be safely tiled on the array and

thus had to be excluded. However, the removal of these sequences means that the genes in question could not be completely analysed. In these patients with negative microarray resequencing results could then be analysed by di-deoxysequencing of these fragments.

Base Calling.

As is evident from Table 1, altering the algorithm parameters drastically alters the call rates. Every base is given a quality score based on a comparison of signal intensity across the 8 probes for a single base. The higher the score is, the more reliable the call is. Hence, with a QST of 30, all bases are called with extremely high degree of accuracy. However, this also meant that only 0.2% of bases were called. Using this parameter would mean extensive operator input to manually base call. Given the number of bases sequenced in a single experiment, this would prove impractical. Lowering the QST did increase the call rate. Using a QST of 2 increased the call rate to 94%. This is in broad agreement with other custom resequencing microarrays that demonstrate call rates of 93.5% (Liu et al., 2007) and 93.6% (Denning et al., 2007). However, by lowering the QST, there is a concomitant reduction in base call accuracy. Comparing the QST of 3 and 2, there is a reduction in base call accuracy from 99.53% to 99.45%. This equates to the potential of 74 inaccurate bases rising to the potential of 87 incorrect bases by changing the QST from 3 to 2.

Using automated analysis, false positive and negative results were observed. The false negative base calling regions were primarily located within NPC1 and with the exception of 1 base, manifested as regions of no-call. Interestingly, 3 of these false negative bases increased the length of a run of identical bases to 4 or more. All the false negatives manifesting as no-calls were picked up during manual review of the no-called regions. However, one heterozygous change could not be detected. Increasing the stringency of analysis and using a QST of 6, this base was consistently called as wild type sequence. Analysing the peak heights

at this region, the base was called as a heterozygote from the forward sequencing, but homozygous wild type called in the reverse (Figure 1). The base calling algorithm uses a set of rules to assign a final base call. One rule states "If one call is homozygote, the other is heterozygote and the homozygote allele is one of the two in the heterozygote call, assign homozygote call and its quality score to that base" (Di and Cawley, 2005). Consequently, as this base was wild type called, it would not have been picked up during manual review of the no called regions and is a true false negative. Five false positive results were detected of which only one could not be eliminated following manual review. Comparison of this remaining false positive to di-deoxysequencing proved that it was a true false positive result. The overall false positive rate was calculated to be <0.01% (Suojanen, 1999). Thus we felt the algorithm parameters used were satisfactory and are in agreement with other studies (Kothiyal et al., 2009).

It must also be noted, that only 9 BRUM1 arrays have been used to analyse these data. The algorithm performs best with large number of samples(Di and Cawley, 2005). Thus one would expect the base calling and accuracy to improve with greater numbers of BRUM1 arrays. Indeed analysing a single sample gave an average call rate of 85.4% whilst the same BRUM1 array analysed as a batch increased the call rate to 94.8%. An improvement in call rate was also observed by (Liu et al., 2007) when multiple samples were analysed as a batch. Accuracy also improved marginally from 99.6% to 99.7%.

Reproducibility.

In order to reliably sequence DNA, it is important that the sequencing is reproducible. In order to test this, three independent PCRs were performed for every single amplicon tested. Further, to test cross chip reliability, identical products were loaded across three BRUM1 arrays. This would give an indication of interchip variation. Overall there was a 94%

agreement between samples. This varied between 90.3% (kappa=0.877, CI 0.874-0.881) to 97% (kappa=0.968, CI 0.964-0.971) depending which combination of BRUM1 arrays used. The likely reason for the 7% variation is that BRUM1-A3, giving the lowest agreement, also had three problematic exon hybridisations. Further, the eighth exon of WFS1 proved to be problematic for all BRUM1 arrays. Whilst the PCRs were very clean, the call rate (84%) for this particular exon was not efficient. Assessing the signal intensities for this exon, it is evident that saturation has been reached for a large number of the bases in either the forward or reverse sequence. As a result, the base calling algorithm no-called 16% of these bases.

No-Called regions.

Comparisons between the samples demonstrated several regions of consistent no-call. These were mainly confined to areas surrounding repeat regions and runs of a single base, in particular C. Analysing the GC content of the no-called probes, there is a skew towards probes having a higher GC content. Further, assessing the individual nucleotide content of the probes, there was a decrease in fluorescence for those with an increased number of C bases. This phenomenon has been observed previously (Cutler et al., 2001), but we have no explanation for this observation. Despite these no-called regions, when assessing the no-called regions manually, it is evident that the majority of these regions can be manually called. Interestingly, the most common no-called single base was a T. It is important to note, that despite regions of no-call, mutations that reside in these no-called areas can be detected. However, if the mutation increases the length of a run of bases that are no-called, the mutated base is also likely to be no-called. But, if no-called, these changes are likely to get picked up during manual review of the sequences.

From our data, an average of 55 (0.35%) no-calls remained following manual review. The majority of these no-calls were confined to *WFS1*. We are exploring recommendations

by (Kothiyal et al., 2009) to tile each sequence in triplicate but altering the interrogation base of each set of probes to assess whether this tiling strategy will decrease the no-calls observed. At present however, if a base is no-called and cannot be assigned through manual review and no mutations are identified elsewhere in the gene, this sequence will need to be di-deoxysequenced to ascertain the true nature of this base.

Variation detection.

The main goal of the custom resequencing array is the ability to detect mutations in patients with a suspected disease to pinpoint the cause and give a better diagnosis and more informed prognosis. To determine how well the BRUM1 array faired at detecting sequence variations, a broad spectrum of changes were tested. A pick up rate of 97.4% (95%CI 85.6-99.9%) was found for point mutations, 100% for known tiled deletions, 25% for unknown or not tiled deletions and 100% for known tiled insertions. For the point variations, 18 changes were in a heterozygous state, whilst 20 were in a homozygous state. As expected, deletions were the main problem for the array. Only 1 out of 4 untiled deletions were detected in the main tiling, and this was in a homozygous state (Supp. Figure S1a). Of the remaining three deletions, one of which was a large 12bp deletion that was felt sufficiently large to be detected by other means. The other two deletions had not been included in the HGMD since they were unknown at the time of the array design. Furthermore, analysing the peak intensity at these deletion sites in the main tiling, no obvious reduction in signal intensity was observed, thus these deletions would have been missed (Supp. Figure S1b).

In our blind test, an overall pickup rate of 72.9% was observed. Whilst this pickup rate is low, stratifying the mutation types, we noticed all 35 detected variations were single nucleotide variations suggesting a pickup rate of 94.3% (95% CI 81.4-99.4%) for this variation type. However, this also meant that no insertion or deletion mutations were

identified. After the analysis was undertaken, the insertion and deletion mutations were compared against the extra tiling and it transpires that only 1 of the deletions had been present in the HGMD at the time of the array design and had been tiled. All other insertions and deletions blind tested had not been tiled and must therefore be classed as novel. It would be expected that any novel heterozygous insertion and deletion would be missed by the array technology. Indeed in our blind analysis, we failed to detect any heterozygous insertions or deletions. However, novel homozygous deletions and insertions should be detected. We did detect an untiled 5bp deletion in *C14ORF133* that was in a homozygous state. Deletions will manifest as no-called regions, whilst insertions may appear as compound heterozygous samples with heterozygous calls a few base pairs apart. Consequently, consideration must be given during the design phase of the microarray resequencing chip to include all known insertions and deletions irrespective of the size of the change to allow adequate detection. Further, it would be prudent to update the resequencing chip with newer insertion and deletions when any re-order is performed.

Comparison to di-deoxysequencing.

To see how the BRUM1 array compares to traditional di-deoxysequencing, surplus products from microarray analysis were sequenced by direct sequencing. Using automated analysis, 93% of sequenced bases were in agreement with the BRUM1 array sequencing. There were 39 bases that were truly discordant between the two technologies. Comparison of the BRUM1 sequencing and di-deoxysequencing to the reference sequence indicated that these discordant bases were concordant using BRUM1 sequencing, but discordant with di-deoxysequencing indicating that the di-deoxysequencing was wrong. Interestingly, these bases were confined to the start of the sequencing traces when di-deoxysequencing doesn't perform well. The primary cause of this was the primers used were not designed with di-

deoxysequencing as the primary analysis technique. Thus the location of the primers may be too close to the start of the sequence of interest to obtain clean sequence. There were also heterozygous calls made in the di-deoxysequencing that were homozygous calls in BRUM1 array sequencing. These incorrect heterozygous calls were due to influence of neighbouring bases raising the secondary peak height above the threshold limit.

Following manual review of the data, the agreement between the di-deoxysequencing and BRUM1, and di-deoxysequencing and reference sequence increased to 99% and 98% respectively. Whilst no discordancy was observed between BRUM1 and di-deoxysequencing, discordancy was observed when di-deoxysequencing was compared against the reference sequence. However this was to be expected as 28 mutations resided in these fragments sequenced.

In conclusion, we performed a stringent analysis of the BRUM1 array using DNA of patients with known mutations and found that it had an overall 96% pick up rate for single base changes, 50% for homozygous insertions and deletions and 39% pick up rate for heterozygous insertion/ deletions. In general we acknowledge that the current resequencing technology will not be able to detect a significant proportion of novel insertions and deletions. These problems can be partly overcome by incorporating all known insertion and deletions into the extra tiling and regularly updating the extra tiling with recently identified insertion/ deletions. Clinically, this could mean that a patient with novel heterozygous insertion/ deletions would be missed. BRUM1 array probe analysis was undertaken using GSeq4.1. A number of other genes have been sequenced using this array (data not shown), with comparable call rates and accuracy and positive identification of known mutations. Thus we feel that the statistics generated would translate to the rest of the array. There are other

software packages available to analyse Affymetrix resequencing arrays that are currently being assessed for performance (Schroeder et al., 2009).

We feel the BRUM1 array can be recommended as a cost-effective, fast, transferrable and reliable screening tool in patients whose phenotype can be caused by mutations in one of several genes. We acknowledge that before the technology can be successfully implemented as a screening tool further evaluation needs to be performed in the clinical setting. This is currently being undertaken. Our estimates suggest that the price of the additional gene sequencing performed on the same array will not be significantly increased compared with the basic price. This price and time reduction is largely achieved by using long range PCR to amplify the genomic DNA fragments. In addition, biochemical testing such as tissue enzyme diagnosis could be avoided if pathogenic mutations are detected. Therefore sequencing of several genes per patient and also using the array for several patients (in whom different genes will be sequenced) will make this method extremely cost-effective. For example we estimate that screening 7 genes by microarray resequencing will be significantly cheaper and faster than sequencing two genes by the conventional di-deoxysequencing methods. Thus, this methodology can be used in clinical practice but it is extremely important to bear in mind the limitations in relation to the detection of novel insertions and deletions. Further, regular updates to the extra tiling to facilitate the detection of recently identified insertion and deletions would be advantageous.

Acknowledgements.

We wish to thank Birmingham Children's Hospital Research Foundation, The Children's Liver Fund and WellChild for funding this research. We would also like to

acknowledge educational grants from Actelion and Ferring Pharmaceuticals. PG is a GSK Clinician Scientist Fellow.

References

Corless CE, Kaczmarski E, Borrow R, Guiver M. 2008. Molecular characterization of Neisseria meningitidis isolates using a resequencing DNA microarray. J Mol Diagn 10:265-271.

Cullinane AR, Straatman-Iwanowska A, Zaucker A, Wakabayash Y, Bruce C, Guanmei L, Rahman F, Gurakan F, Utine E, Ozkan TB, Denecke J, Vukovic J, Rocco MD, Mandel H, Cangul H, Matthews RP, Thomas SG, Rappoport J, Arias IM, Wolburg H, Knisely AS, Kelly DA, Mueller F, Maher ER, Gissen P. 2010. Mutations in VIPAR cause an arthrogryposis, renal dysfunction, and cholestatsis syndrome phenotype with defects in epithelial polarisation. Nat Genet In Press.

Cullinane AR, Straatman-Iwanowska A, Zaucker A, Wakabayash Y, Bruce C, Rahman F, Rappoport J, Arias IM, Wolburg H, Knisely AS, Kelly DA, Mueller F, Maher ER, and Gissen P. 2009. Mutations in polarin (PLRN) cause an ARC syndrome phenotype and defects in epithelial polarisation and apical junction complex formation (Abstract 290). 59th Annual Meeting of The American Society of Human Genetics. Honolulu, Hawaii.

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A. 2001. High-throughput variation detection and genotyping using microarrays. Genome Res 11:1913-1925.

Denning L, Anderson JA, Davis R, Gregg JP, Kuzdenyi J, Maselli RA. 2007. High throughput genetic analysis of congenital myasthenic syndromes using resequencing microarrays. PLoS One 2:e918.

Di X and Cawley S. 2005. Alternative base calling method for resequencing microarrays. Conf Proc IEEE Eng Med Biol Soc. 2809 2812.

Fleiss JL. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76:378-382.

Gissen P, Johnson CA, Morgan NV, Stapelbroek JM, Forshew T, Cooper WN, McKiernan PJ, Klomp LW, Morris AA, Wraith JE, McClean P, Lynch SA, Thompson RJ, Lo B, Quarrell OW, Di RM, Trembath RC, Mandel H, Wali S, Karet FE, Knisely AS, Houwen RH, Kelly DA, Maher ER. 2004. Mutations in VPS33B, encoding a regulator of SNARE-dependent membrane fusion, cause arthrogryposis-renal dysfunction-cholestasis (ARC) syndrome. Nat Genet 36:400-404.

Hacia JG. 1999. Resequencing and mutational analysis using oligonucleotide microarrays. Nat Genet 21:42-47.

Hartmann A, Thieme M, Nanduri LK, Stempfl T, Moehle C, Kivisild T, Oefner PJ. 2009. Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. Hum Mutat 30:115-122.

Jalanko A, Braulke T. 2009. Neuronal ceroid lipofuscinoses. Biochim Biophys Acta 1793:697-709.

Karaman MW, Groshen S, Lee CC, Pike BL, Hacia JG. 2005. Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. Nucleic Acids Res 33:e33.

King JE. 2004. Software solutions for obtaining a kappa-type statistic for use with multiple raters. Annual Meeting of the Southwest Educational Research Association. Dallas, Texas.

Kothiyal P, Cox S, Ebert J, Aronow BJ, Greinwald JH, Rehm HL. 2009. An overview of custom array sequencing. Curr Protoc Hum Genet 61:17.1-17.11.

Liu C, Aronow BJ, Jegga AG, Wang N, Miethke A, Mourya R, Bezerra JA. 2007. Novel resequencing chip customized to diagnose mutations in patients with inherited syndromes of intrahepatic cholestasis. Gastroenterology 132:119-126.

Malanoski AP, Lin B, Wang Z, Schnur JM, Stenger DA. 2006. Automated identification of multiple micro-organisms from resequencing DNA microarrays. Nucleic Acids Res 34:5300-5311.

Neufeld EF, Muenzer J. 2001. The Mucopolysaccharidoses. In: Scriver CR, Beaudet AL, Sly WS, Valle D, editors. The metabolic and molecular bases of inherited disease. New York: Mcgraw-Hill. p 3241-3251.

Sauro J and Lewis JR. 2005. Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. PROCEEDINGS of the HUMAN FACTORS AND ERGONOMICS SOCIETY 49th ANNUAL MEETING. 2100 2104.

Schroeder C, Stutzmann F, Weber BH, Riess O, Bonin M. 2009. High-throughput resequencing in the diagnosis of BRCA1/2 mutations using oligonucleotide resequencing microarrays. Breast Cancer Res Treat.

Suojanen JN. 1999. False false positive rates. N Engl J Med 341:131.

Takahashi Y, Seki N, Ishiura H, Mitsui J, Matsukawa T, Kishino A, Onodera O, Aoki M, Shimozawa N, Murayama S, Itoyama Y, Suzuki Y, Sobue G, Nishizawa M, Goto J, Tsuji S. 2008. Development of a high-throughput microarray-based resequencing system for neurological disorders and its application to molecular genetics of amyotrophic lateral sclerosis. Arch Neurol 65:1326-1332.

Wraith JE, Baumgartner MR, Bembi B, Covanis A, Levade T, Mengel E, Pineda M, Sedel F, Topcu M, Vanier MT, Widner H, Wijburg FA, Patterson MC. 2009. Recommendations on the diagnosis and management of Niemann-Pick disease type C. Mol Genet Metab 98:152-165.

Zwick ME, Kiley MP, Stewart AC, Mateczun A, Read TD. 2008. Genotyping of Bacillus cereus strains by microarray-based resequencing. PLoS One 3:e2513.

Figure Legends

Figure 1: False negative call in *NPC1*.

Forward strand shows a heterozygous call with peaks of C and T being equal and prominent.

Reverse strand shows a homozygous calls with a peak of A.

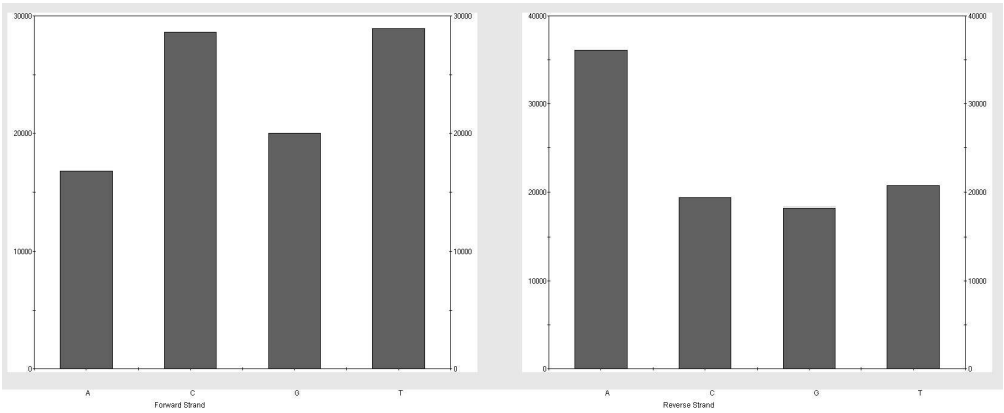


Figure 1: False negative call in NPC1.
Forward strand shows a heterozygous call with peaks of C and T being equal and prominent.
Reverse strand shows a homozygous calls with a peak of A.

140x57mm (300 x 300 DPI)

	Average Call Rate		Average Accuracy	
QST	BRT0	BRT 0.5	BRT0	BRT0.5
2	94.25% (914)	93.55% (1025)	99.45% (87)	99.51% (78)
3	90.70% (1478)	89.42%(1681)	99.53% (75)	99.58% (67)
6	74.52% (4049)	71.71% (4495)	99.72% (44)	99.76% (38)
9	53.67% (7362)	49.01% (8102)	99.82% (29)	99.85% (24)
12	33.83% (10 514)	29.23% (11 245)	99.89% (17)	99.91% (14)
30	0.42% (15 223)	0.22% (15 540)	100% (0)	100% (0)

Table 1 Call rate and accuracy of the sequencing.

Figures in parenthesis refer to the actual numbers of no-called bases for Average Call rate or the accurate numbers of incorrect bases for Average Accuracy; QST =quality score threshold, BRT=base reliability threshold.

	PCR A	PCR B	PCR C	All Chips
pA	0.917	0.968	0.973	0.948
Kappa	0.895	0.959	0.966	0.934
LCI	0.891	0.955	0.962	0.933
UCI	0.898	0.962	0.970	0.935
No of bases different between samples	1319	508	429	826

Table 2: Reproducibility

The output from each BRUM1 array was compared within each PCR and between all samples. Fleiss Kappa was used to compare the outputs. pA is the level of agreement between the samples, whilst LCI and UCI are the lower and upper 95% confidence intervals.

No Called base/ region	A	G	C	T	AT	>2CC	>2GG	Other*
Frequency	71	58	64	89	58	58	17	41
% of all no-calls	15.6%	12.7%	14.0%	19.2%	12.7%	12.7%	3.7%	9.0%

Table 3: Consistent regions of No-Call

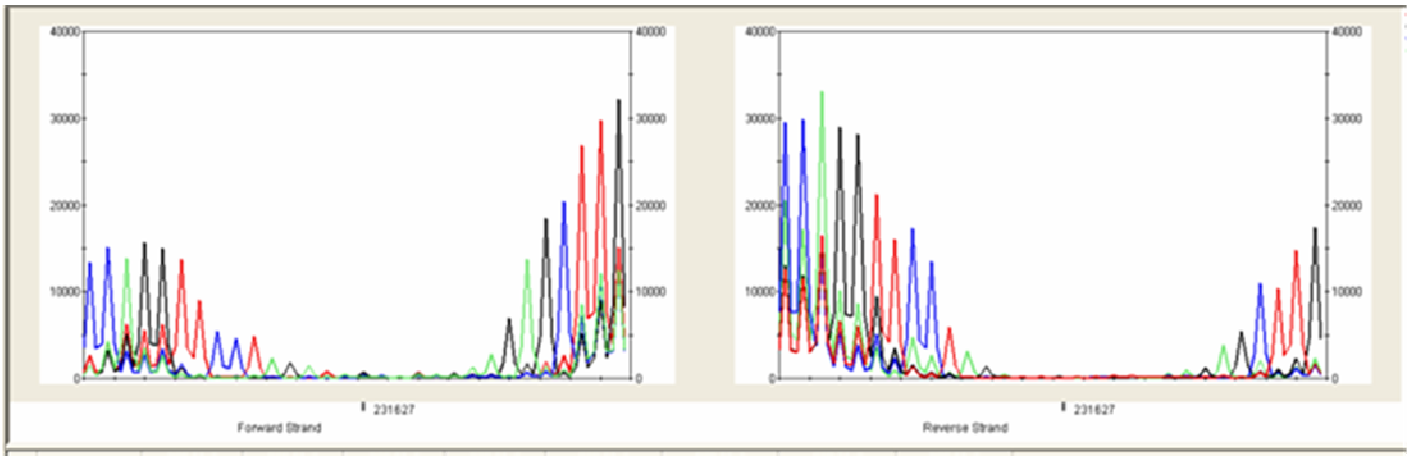
*Other refers to no-called regions observed less than twice.

	Point Changes	Insertions	Deletions
Number of changes analysed	39 (76)	3 (6)	10 (18)
Number of changes detected	32 (65)	0 (0)	4 (4)
False Positives	3 (3)	-	-
False Positive rate	<0.01% (<0.01%)	-	-
False Negatives	7 (9)	3 (6)	5 (13)
False Positive following manual review	1 (3)	-	-
False negative following manual review	1 (3)	0 (3)	3 (11)
Novel changes	5		
Pickup Rate	97% (96%)	100% (50%)	70% (39%)
95% CI	82-99% (88.5-99.1)	47-100% (19-81%)	39-90% (20-61%)

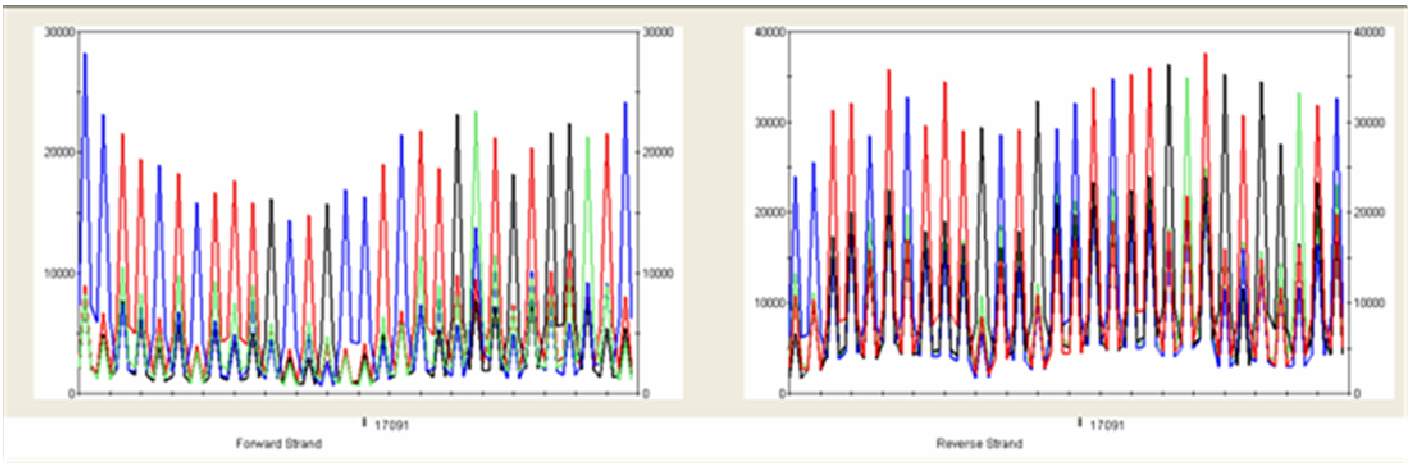
Table 4: Summary of Changes

False positives were not calculated for insertions and deletions. Any no-called base has to be classed as a false positive deletion and thus needs manual review. Figures in parenthesis are combined data including additional *NPC1* variations used in the blind study. Novel changes were known polymorphisms that were detected but were not expected.

a)



b)



Supp. Figure S1: Deletion detection: 30bp of sequence incorporating two known deletions were captured in GSeq4.1. When deletions are in a homozygous state (panel a; 5bp deletion in *C14ORF133*) there is a drop in signal intensity on both strands indicating a deletion. However, when deletions are in a heterozygous state (panel b; a 12bp heterozygous deletion in *VPS33B*) no reduction in signal intensity is observed.

Gene		Gene		Gene		Gene		Gene		Gene		Gene	
C14ORF133	Patient	GAA	Patient	NPC1	Patien	NPC2	Patient	SLC19A2	Patient	VPS33B	Patient	WFS1	Patient
1	ARC-18	2	GAA-33	1	MRC5VA	1	NPC-9	1	SLC-1	1	MRC5VA	2	MRC5VA
2	MRC5VA	3	GAA-24	2	MRC5VA	2	MRC5VA	2	SLC-2	2	ARC-8	3	WFS1-1
3	MRC5VA	4	GAA-36	3	MRC5VA	3	MRC5VA	3	MRC5VA	3	ARC-21	4	MRC5VA
4	MRC5VA	5	MRC5VA	4	MRC5VA	4	MRC5VA	4	MRC5VA	4	ARC-9	5	WFS1-2
5	MRC5VA	6	GAA-31	5	MRC5VA	5	MRC5VA	5	MRC5VA	5	ARC-10	6	WFS1-3
6	MRC5VA	7	GAA-3	6	MRC5VA			6	MRC5VA	6	ARC-11	7	MRC5VA
7	ARC-2	8	MRC5VA	7	MRC5VA					7	ARC-12	8a	WFS1-4
8	MRC5VA	9	GAA-15	8	NPC-5					8	ACR-13	8b	WFS1-4
9	ARC-3	10	GAA-21	9	NPC-2					9	MRC5VA	8c	WFS1-5
10	ARC-4	11	GAA-21	10	MRC5VA					10	MRC5VA	8d	WFS1-6
11	MRC5VA	12	GAA-16	11	MRC5VA					11	MRC5VA	8e	MRC5VA
12	ARC-5	13	GAA-27	12	NPC-20					12	ARC-14		
13	MRC5VA	14	GAA-18	13	NPC-20					13	ARC-15		
14	MRC5VA	15	GAA-8	14	NPC-21					14	MRC5VA		
15	MRC5VA	16	MRC5VA	15	MRC5VA					15	ARC16		
16	ARC-6	17	MRC5VA	16	MRC5VA					16	ARC16		
17	ARC-7	18	GAA-14	17	NPC-20					17	MRC5VA		
18	MRC5VA	19	GAA-4	18	NPC-20					18	ARC-17		
19	MRC5VA			19	NPC-7					19	ARC-1		
				20	NPC-6					20	ARC-19		
				21	NPC-6					21	ARC-20		
				22	NPC-12					22	ARC-20		
				23	NPC-11					23	MRC5VA		
				24	MRC5VA								
				25	MRC5VA								

Supp Table S2: Source of DNA used for the array

Lysosomal Storage	Lysosomal Storage	Glycogen Storage	Cholestasis	Iron-deposition/
NPC1 (607623) ENST00000269228	*NAGLU* (609701)	*G6PC* (232200)	*VPS33B* (608552) ENST00000333371	*PANK2* (606157)
NPC2 (601015) ENST00000238633	*HGSNAT* (610453)	*SLC37A4* (232220)	*ATP8B1* (602397)	*PLA2G6* (603604)
PSAP (607939)	*GNS* (607664)	*SLC2A1* (138140)	*ABCB11* (603201)	
SUMF1 (176801)	*GALNS* (612222)	*SLC2A2* (138160)	*ABCB4* (171060)	
SMPD1 (607608)	*ARSB* (611542)	*AGL* (610860)	*JAG1* (601920)	
GBA (606463)	*GUSB* (611499)	*GBE1* (607839)	*ATP7B* (606882)	
GLB1 (611458)	*HYAL1* (607071)	*PYGM* (608455)	*SLC25A13* (603859)	
HEXB (606873)	*DYM* (607461)	*PYGL* (608455)	*C14ORF133* ENST00000327028	
HEXA (606689)	*AGA* (208400)	*PFKM* (610681)		
ASAH1 (228000)	*MAN2B1* (609458)	*PHKA2* (306000)		
GNPTG (607838)	*MANBA* (609489)	*PHKB* (172490)		
GNPTAB (607840)	*ABHD5* (604780)	*PHKG2* (172471)		
MCOLN1 (605248)	*PNPLA2* (609059)	*GYS2* (138571)		
SLC17A5 (604322)	*LAMP2* (309060)	*PRKAG2* (602743)		
ARSA (607574)	*NEU1* (608272)	*ALDOC* (103870		
GLA (300644)				
GAA (606800) ENST00000302262				
IDUA (252800)				
IDS (309900)				
SGSH (605270)				

Supp. Table S2; Genes included in the array. Figures in parenthesis are OMIM reference numbers. ENST numbers are Ensembl accession numbers (www.ensembl.org).

Endocrine Malignancy	Growth	Diabetes	Obesity	Lipid Disorders
VHL (608537)	*GH1* (139250)	*WFS1* (606201) ENST0000022670	*LEP* (164160)	*LDLR* (606945)
SDHB (185470)	*IGF1* (147440)	*CISD2* (604928)	*LEPR* (601007)	*APOB* (107730)
SDHC (602413)	*IGF2* (147470)	*SLC19A2* (603941) ENST00000236137	*MC3R* (155540)	*OLR1* (602601)
SDHD (602690)	*PTPN11* (176876)	*EIF2AK3* (604032)	*GHRL* (605353)	*PCSK9* (607786)
RET (164761)	*RAF1* (164760)		*CD36* (173510)	*LDLRAP1*(600073)
				APOE (107741)
				USF1 (191523)
				CETP (118470)
				ABCA1 (600046)
				ABCG5 (605459)
				ABCG8 (605460)
				CYP27A1 (606530)
				MTTP (157147)

Supp. Table S2 continued.