HUGO J (2009) 3:41–49
DOI 10.1007/s11568-010-9137-y

RESEARCH ARTICLE

# Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing

**Hao Hu · Klaus Wrogemann · Vera Kalscheuer · Andreas Tzschach · Hugues Richard · Stefan A. Haas · Corinna Menzel · Melanie Bienek · Guy Froyen · Martine Raynaud · Hans Van Bokhoven · Jamel Chelly · Hilger Ropers · Wei Chen**

Received: 13 January 2010 / Revised: 24 February 2010 / Accepted: 12 March 2010 / Published online: 25 March 2010


**Abstract** Massive parallel sequencing has revolutionized the search for pathogenic variants in the human genome, but for routine diagnosis, re-sequencing of the complete human genome in a large cohort of patients is still far too expensive. Recently, novel genome partitioning methods have been developed that allow to target re-sequencing to specific genomic compartments, but practical experience with these methods is still limited. In this study, we have combined a novel droplet-based multiplex PCR method and next generation sequencing to screen patients with X-linked mental retardation (XLMR) for mutations in 86 previously identified XLMR genes. In total, affected males from 24 large XLMR families were analyzed, including three in whom the mutations were already known. Amplicons corresponding to functionally relevant regions of these genes were sequenced on an Illumina/Solexa Genome Analyzer II platform. Highly specific and uniform enrichment was achieved: on average, 67.9% unambiguously mapped reads were derived from amplicons, and for 88.5% of the targeted bases, the sequencing depth was sufficient to reliably detect variations. Potentially disease-causing sequence variants were identified in 10 out of 24 patients, including the three mutations that were already known, and all of these could be confirmed by Sanger sequencing. The robust performance of this approach demonstrates the general utility of droplet-based multiplex PCR for parallel mutation screening in hundreds of genes, which is a prerequisite for the diagnosis of mental retardation and other disorders that may be due to defects of a wide variety of genes.

**Keywords** Droplet-based multiplex PCR · Massive parallel sequencing · Mutation screening · X-linked mental retardation

**Electronic supplementary material** The online version of this article (doi:10.1007/s11568-010-9137-y) contains supplementary material, which is available to authorized users.

H. Hu · K. Wrogemann · V. Kalscheuer · A. Tzschach · H. Richard · S. A. Haas · C. Menzel · M. Bienek · H. Ropers · W. Chen (✉)
Max-Planck-Institute for Molecular Genetics, Berlin, Germany
e-mail: wei@molgen.mpg.de

K. Wrogemann
Department of Biochemistry & Medical Genetics, University of Manitoba, Winnipeg, MB, Canada

G. Froyen
Human Genome Laboratory, Centre for Human Genetics, VIB, K.U.Leuven, Leuven, Belgium

M. Raynaud
INSERM, U930; Centre Hospitalier Régional Universitaire de Tours, Service de Genetique, 37044 Tours, France

H. Van Bokhoven
Department of Human Genetics, Radboud University Medical Centre, Nijmegen, The Netherlands

J. Chelly
Faculté de Médecine Cochin, INSERM 129-ICGM, Paris, France

W. Chen
Max-Delbrück-Centrum für Molekulare Medizin, Berlin Institute for Medical Systems Biology, Berlin, Germany

## Introduction

The core problem of human and medical genetics is to identify genetic variants underlying specific phenotypes. This has been traditionally achieved by Sanger sequencing of PCR products, which is a tedious process and often economically formidable if a large set of genes needs to be

studied. The situation has recently been improved dramatically with the emergence of massive parallel sequencing technologies. Compared with the Sanger method, these so-called next-generation sequencing platforms can sequence DNA orders of magnitude faster and at much lower cost (Mardis 2008; Shendure and Ji 2008).

However, even with the dramatically improved efficiency, the current technology does not allow us to re-sequence the complete genome from a large number of human patients in an economically realistic manner (Olson 2007). Exploitation of the full potential of the current platforms requires a subset of the genome of medical interest to be isolated for targeted sequencing. To meet this need, a variety of methods have been developed in the last few years to carry out genome partitioning. Two recent strategies include capture by circularization (Porreca et al. 2007) and selection by hybridization (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007; Gnirke et al. 2009; Ng et al. 2009; Turner et al. 2009). With different strengths and weaknesses, nearly all of these strategies can be used to enrich megabase-scale target regions. These can be continuous genomic intervals or a full complement of protein-coding exons (Albert et al. 2007; Hodges et al. 2007; Okou et al. 2007; Bau et al. 2009; Gnirke et al. 2009; Ng et al. 2009; Turner et al. 2009).

Multiplex PCR is an option for enrichment of a set of sequencing targets. However, given the challenges, including primer dimer formation, mispriming, and non-uniform amplification, the scale is often limited to a few dozen targets in one reaction (Edwards and Gibbs 1994). To circumvent these limitations, several derivative methods, such as Megaplex PCR and Nested Patch PCR have been reported, where $\sim$100 targets were simultaneously amplified and sequenced by using the Roche 454 platform (Meuzelaar et al. 2007; Varley and Mitra 2008). Compared with capture by hybridization or circularization these attempts have achieved higher specificity and uniformity, although their scales cannot match the ever-increasing throughput of new sequencing technologies.

Very recently, a novel multiplex PCR method, which uses emulsion PCR and a microfluidic chip to compartmentalize the PCR reactions by single primer pairs has been developed. In their proof-of-principle experiment, Tewhey et al demonstrated that up to 3,976 amplicons could be simultaneously amplified (Tewhey et al. 2009).

Here, we applied this droplet-based PCR platform to screen for mutations in 86 genes which are known to cause X-linked mental retardation (XLMR) if mutated (Ropers 2007, 2008). We designed 1,912 amplicons to cover the coding regions and flanking intronic regions. After sequencing the PCR products obtained from 24 unrelated patients with XLMR using the Illumina/Solexa platform, three known causative mutations were confirmed and another seven potential deleterious mutations were detected, including five missense changes, one frameshift mutation and one 3 bp deletion. All these variants were confirmed by traditional Sanger sequencing.

## Results

In order to search for mutations in 86 known XLMR genes, we applied a novel droplet-based multiplex PCR method to simultaneously amplify their coding regions and splice sites. In total, 1,315 target regions with a total length of 552,930 bp were identified in silico. After filtering the regions for which suitable primers could not be found, 1,912 primer pairs were successfully designed to cover 1,198 target regions with a total length of 515,598 bp. The size of the amplicons ranges from 201 to 600 bp, with an average of 509 bp and a total length of 783,183 bp. Each of the primer pairs was separately encapsulated into droplets. These droplets were then pooled into a droplet primer library. With a microfluidic platform (Tewhey et al. 2009) the primer droplets were merged with droplets containing fragmented genomic DNA, DNA polymerase and dNTPs and dispensed into a single tube for PCR amplification.

The amplification products obtained from 24 patients with XLMR were sequenced using the Illumina GAII platform. More specifically, the products from each patient were recovered by breaking the emulsion, then concatenated, and finally shotgun sequenced on one lane of the Illumina flowcell. Single-end 36 nucleotide (nt) reads generated from the shotgun library were aligned to the human reference genome (NCBI build 36.1). Given our alignment setting (see "Materials and methods"), 474,504 of the 515,598 bp target sequences can be unambiguously mapped using 36 nt reads.

For each of the 24 patients, an average of 402 Mb could be unambiguously mapped to the reference genome sequence (NCBI build 36.1), out of which 67.9% map within the amplicons. The median sequence depth per base within the target regions ranges from 92 to 445. To eliminate PCR artefacts introduced during the sequencing library construction, we discarded all but one reads that mapped exactly to the same region and on the same strand. After this process, the median sequence depth ranges from 10 to 66. Between 94.1 and 99.6% of the target bases were covered at least once and 87.9–99.5% were covered sufficiently (base quality score $\geq$30; $\geq$2$\times$ coverage) to call hemizygous mutations on the X chromosome from male patients (see Table 1).

Uniformity of enrichment, together with specificity determines the total amount of sequencing required to achieve a coverage sufficient for reliable variation detection. The higher the difference in abundance between

**Table 1** Summary of sequencing results from 24 patients

| Patient ID | Percentage of reads mapped on human genome | Percentage of reads mapped on the target regions | Median sequencing depth of target bases | 10th Percentile sequencing depth of target bases | Target bases with depth ≥1 | Target bases with non-redundant depth ≥2, quality score ≥30 | No. known SNVs | No. Novel SNVs | No. known indels | No. novel indels | No. non-recurrent novel synonymous variants | No. non-recurrent novel nonsynonymous variants |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MPI01 | 57 | 14.4 | 127 | 22 | 469,202 | 457,338 | 63 | 3 | 6 | 1 | 0 | 0 |
| MPI02 | 71.4 | 41.6 | 214 | 32 | 471,763 | 464,524 | 64 | 8 | 7 | 0 | 2 | 2 |
| MPI03 | 61.1 | 14.8 | 105 | 16 | 465,435 | 453,198 | 56 | 8 | 7 | 2 | 3 | 2 |
| MPI04 | 44.6 | 12.2 | 93 | 12 | 465,221 | 455,388 | 59 | 6 | 6 | 0 | 2 | 0 |
| MPI05 | 57 | 20.1 | 162 | 32 | 472,442 | 466,869 | 49 | 4 | 6 | 1 | 3 | 0 |
| MPI06 | 79.9 | 49 | 445 | 104 | 475,785 | 471,906 | 69 | 4 | 6 | 0 | 2 | 0 |
| MPI07 | 56.4 | 31.4 | 303 | 58 | 472,659 | 466,797 | 61 | 7 | 6 | 2 | 3 | 0 |
| MPI08 | 58.3 | 17.3 | 149 | 29 | 471,184 | 464,040 | 72 | 6 | 7 | 1 | 2 | 1 |
| MPI09 | 62 | 24.3 | 200 | 37 | 471,129 | 464,733 | 43 | 5 | 6 | 0 | 2 | 1 |
| MPI10 | 68.8 | 39.9 | 285 | 82 | 476,112 | 472,553 | 73 | 7 | 7 | 0 | 2 | 1 |
| MPI11 | 66.9 | 43.9 | 360 | 56 | 471,563 | 464,788 | 52 | 4 | 6 | 2 | 3 | 1 |
| MPI12 | 43 | 27 | 210 | 17 | 459,532 | 447,392 | 43 | 6 | 6 | 0 | 3 | 0 |
| MPI13 | 71.5 | 43.6 | 390 | 33 | 459,696 | 448,012 | 61 | 3 | 6 | 0 | 2 | 0 |
| MPI14 | 66.7 | 33.1 | 332 | 24 | 451,283 | 432,832 | 52 | 5 | 5 | 0 | 3 | 0 |
| MPI15 | 67.7 | 45.2 | 302 | 42 | 463,295 | 453,557 | 61 | 4 | 7 | 2 | 1 | 2 |
| MPI16 | 72.1 | 39.8 | 261 | 37 | 462,978 | 452,147 | 62 | 5 | 6 | 2 | 2 | 2 |
| MPI17 | 82.2 | 57.4 | 356 | 62 | 468,230 | 460,292 | 74 | 7 | 9 | 1 | 3 | 0 |
| MPI18 | 69.5 | 45.5 | 359 | 29 | 452,294 | 436,990 | 43 | 6 | 4 | 0 | 2 | 1 |
| MPI19 | 74.3 | 42.3 | 325 | 30 | 456,327 | 441,255 | 58 | 6 | 6 | 1 | 5 | 0 |
| MPI20 | 53.9 | 18.4 | 145 | 16 | 462,203 | 449,531 | 58 | 3 | 5 | 0 | 1 | 0 |
| MPI21 | 64.7 | 33.4 | 268 | 56 | 475,372 | 470,751 | 42 | 5 | 7 | 0 | 3 | 0 |
| MPI22 | 53.4 | 22.3 | 92 | 8 | 449,473 | 418,304 | 54 | 2 | 5 | 0 | 0 | 1 |
| MPI23 | 65.9 | 23.6 | 221 | 52 | 475,621 | 471,114 | 62 | 4 | 8 | 0 | 1 | 0 |
| MPI24 | 71.7 | 35.3 | 276 | 76 | 475,502 | 471,623 | 57 | 4 | 7 | 1 | 1 | 2 |

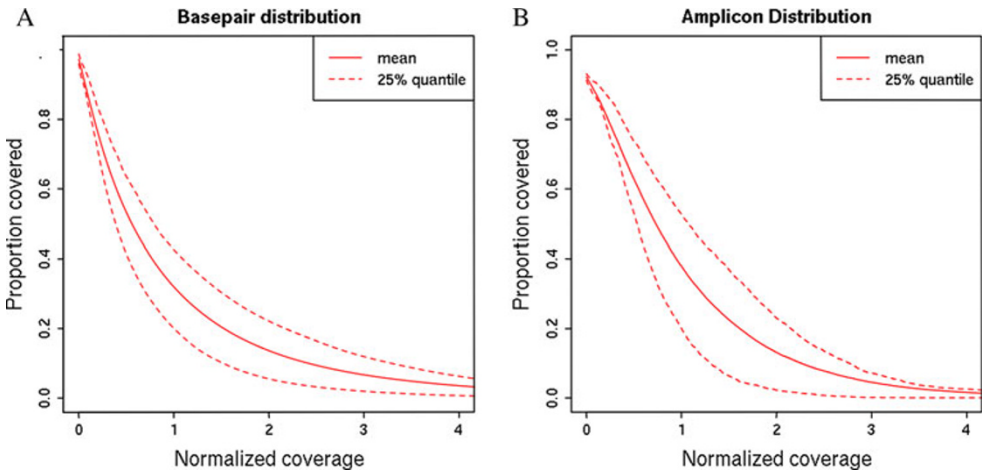


**Fig. 1** Normalised sequencing coverage distribution. **a** Normalised coverage cumulative distribution of the sequenced bases within all amplicons. **b** Normalised coverage cumulative distribution of all amplicons. Normalised coverage is the absolute coverage divided by the mean coverage. Amplicon coverage is the median coverage of all the bases within the amplicon. *Solid lines* represent the average among all the 24 samples and the *dashed lines* represent 25th and 75th percentiles

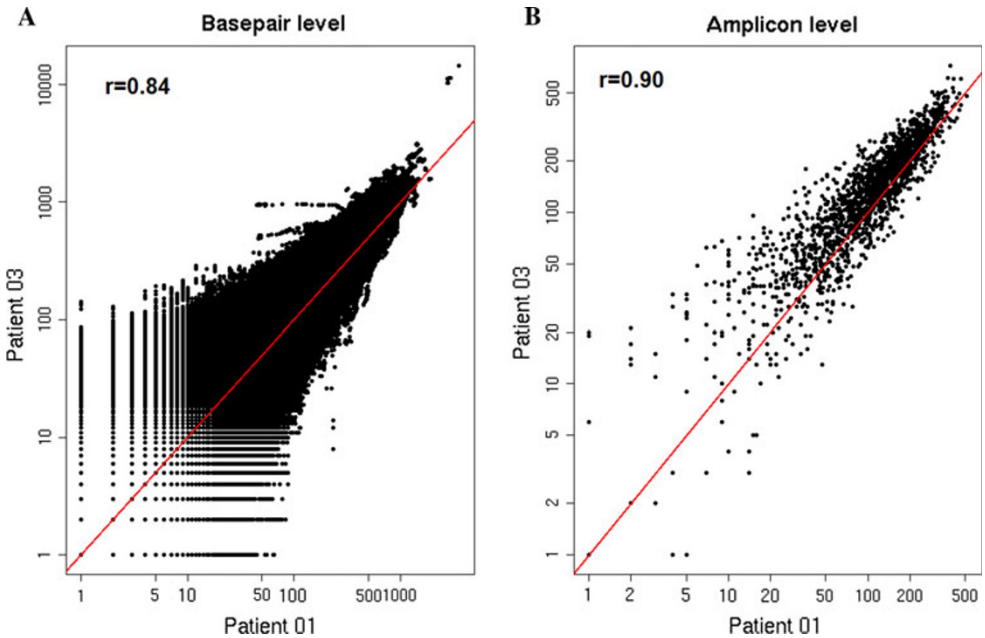


**Fig. 2** The correlation of sequencing coverage between 2 samples. **a** The sequence coverage of each base from sample 1 was plotted against sample 3. **b**. The sequence coverage of each amplicon from the same 2 samples was plotted. The sequence coverage of an amplicon is the median sequence coverage of all bases within the amplicon. The data shown here are representative for all 24 patients

overrepresented and underrepresented targets, the more sequencing will be required to cover all the target regions. Here, from the lowest to the highest 10% quantile, the depth per base pair differs up to 26-fold. 90% of the base pairs have at least 29% of the average depth (Fig. 1a). Two experimental procedures can generate this bias. On one hand, due to the nature of massive parallel sequencing, it is impossible to achieve an even distribution of sequencing reads across the entire region. On the other hand, uneven PCR amplification efficiency will result in different amplicons with various abundances. Since the latter is more relevant here, we also evaluated whether the sequencing depth is uniform across amplicons. The variation from the lowest to the highest 10% quantile does not change significantly from what is observed at the base pair level (25-fold). 90% of all amplicons are covered with a depth higher than 23% the average value, which is attained for ~50% of the amplicons (Fig. 1b). To assess the reproducibility of such amplification bias, we compared the normalised sequence coverage per amplicon between different samples. The bias between two samples was remarkably similar (see Fig. 2) and the average pair-wise correlation coefficient between individuals for base and amplicon coverage is 0.84 and 0.90, respectively, indicating the variability in amplification efficiency is mostly systematic instead of being stochastic. Potential causes of

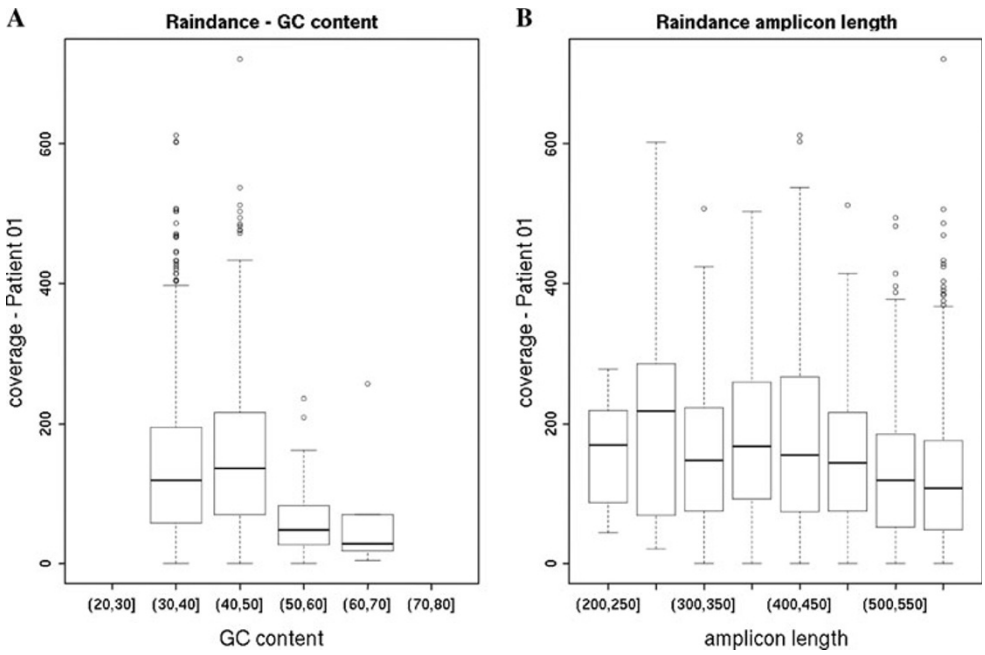


**Fig. 3** The dependence of sequencing coverage of amplicons on their GC content (**a**) or amplicon length (**b**). The amplicons were grouped depending on their GC content (**a**) or length (**b**), the distribution of sequencing coverage within each group is shown as *boxplot*

**Table 2** List of the seven potentially deleterious mutations

| Patient ID | Mutation | Gene | Amino acid change |
|---|---|---|---|
| MPI-2 | chrX 134907970T->C | SLC9A6 | p.L188P |
| MPI-3 | chrX 31132095-31132097delAAG | DMD | p.L1897del |
| MPI-8 | chrX 73660933G->A | SLC16A2 | p.R271H |
| MPI-10 | chrX 70256689A->G | MED12 | p.Y166C |
| MPI-11 | chrX 21920673A->G | SMS | p.Y328C |
| MPI-16 | chrX 153323964-153323965delAG | GDI1 | p.396frameshift |
| MPI-22 | chrX 153240569T->C | FLNA | p.Q1484R |

the variability could include GC content of the amplicon and amplicon length. However, these factors do not have a significant impact (see Fig. 3).

We called sequence variations including single nucleotide variants (SNVs) and indels of 1–3 bp at positions with sufficient sequence coverage (≥2× coverage and Phred-like quality score ≥30; for details, see "Materials and methods"). In 24 patients, we have identified 310 SNVs and 27 small indels comprising 1–3 bp. Of these, 75 SNVs and 16 indels were not represented in dbSNP (version 130). The number of annotated and novel SNVs as well as indels found in each individual is listed in Table 1. In total,, we found 58 novel SNVs and 9 novel indels in exonic and 50-bp of flanking intronic regions. All 3 previously known mutations were detected (see "Materials and methods"). 7 other changes were also considered as potentially pathogenic, including 5 missense changes predicted by PolyPhen (Sunyaev et al. 2001) to be deleterious, a frameshift mutation and a deletion of one amino acid. All 7 could be confirmed by Sanger sequencing (see Table 2).

## Discussion

We have applied a novel droplet-based multiplex PCR method to enrich the coding regions of 86 known XLMR genes. Using emulsion PCR and microfluidics, the PCR reactions by individual primer pairs were compartmentalised, thereby avoiding interference between different primer pairs. This study demonstrated that the method has high specificity for selected regions and high uniformity between different amplicons. Combined with massive parallel sequencing, high sequence coverage of the target regions was achieved. This allowed us to identify novel and potentially disease-causing mutations in XLMR genes. Compared to previously described enrichment methods that are based on hybridisation, our approach is better suited for capturing short, closely neighboring exons. One obvious reason is that PCR amplification can be optimized for unique target regions whereas selection by hybridisation approaches will have to carry adjacent sequences together with real targets and thus reduce specificity. In the

approach used here, different amplicons are concatenated for shotgun sequencing. Most reads comprising the ends of different fragments cannot be mapped using current mapping strategies and thus although these reads are derived from the amplicons, they are not included to calculate the specificity of enrichment. Since we expect 14% percent of the reads to belong to this category (36 nt single end reads from amplicons of 500 bp), the 67.9%, i.e. the percentage of sequencing reads derived from the amplicons, represents the minimum specificity, although this is already well in excess of the ~50% reported by two recent studies employing hybridisation approaches (Gnirke et al. 2009; Ng et al. 2009). In addition to specificity, our approach has less allelic bias since most alleles can be equally well amplified in one PCR reaction. In contrast, selection by hybridisation will capture less of an allele which is significantly different from the reference sequence, e.g. large deletions overlapping with the probes. Upon mutation detection, this allelic bias can increase false negative rates,, especially in heterozygotes.

Another advantage of the enrichment protocol used here is that the representation of different amplicons appears to be more uniform than obtained with other, previously described methods, particularly those based on capture by circularisation, although recently, their performance has been improved (Turner et al. 2009). In this study, the sequencing depth was 29% of the mean for 90% of the targeted bases, and 80% of all bases were enriched to abundances within a 26-fold range. Since this bias is largely systematic instead of being stochastic, it can be further reduced by adjusting the relative proportion of different primer pairs in the PCR reaction.

High coverage, along with high uniformity and specificity of this enrichment technique render it very suitable for the reliable detection of sequence variants. We have correctly called variants at up to 99.5% of the uniquely mappable positions in the target regions. In addition to the 3 previously known mutations, we detected another seven putatively deleterious changes.

XLMR is a very heterogeneous disorder, and although more than 80 XLMR genes are already known, they account only for a portion of all disease causing mutations in MR families with unambiguous X-linkage. In a meta-analysis of all mutations found so far in the large cohort of families collected by the European Mental Retardation Consortium which does not include patients with Fragile-X syndrome, mutations in other known XLMR genes were found to be responsible for at least 42% and likely to account for even half of the XLMR families (de Brouwer et al. 2007). Since our study includes additional XLMR genes that were identified (Tarpey et al. 2009) after the afore-mentioned analysis, we expected to find mutations in an even higher proportion of families. However, this was not the case: in the present study we detected putatitve and possible mutations in no more than 7 out of 21 families (33%). Several factors may account for this discrepancy. First, our cohort was small ($n = 21$), and the statistical power is therefore low. Second, our families were part of the EuroMRX cohort, i.e. they were mutation-negative "leftovers" after the screening of a variable number of XLMR genes by conventional Sanger sequencing. Third, the coverage of the method employed here was not complete: on average, 88.5% of all targeted exonic and 50 bp flanking intronic sequences (515,598 bp) were covered at a sufficient depth. With our alignment settings, 131 amplicons cannot be unambiguously mapped and thus can not be analyzed. In addition, 34 amplicons were not successfully amplified in at least one of the 24 patients (see Supplementary Table). These regions may harbour additional point mutations. Most amplification failures are probably due to technical reason. To rule out larger deletions encompassing two or several exons, we have screened the data for missing adjacent amplicons in one or a few patients. No such clustering of missing amplicons was observed in any of the patients. Finally, due to the limitations of 36 nt single end sequencing, we were only able to detect indels of 1–3 bp, and larger indels might well account for some of the missing mutations.

It is, however, probable that in most mutation-negative XLMR families, many of the missing causative mutations are either hidden in the non-coding regions of known XLMR genes, or the mutations are within X-chromosomal genes which have hitherto not been implicated in mental retardation, including sequences for non-coding RNA (Mercer et al. 2008). In the near future, large-scale exon enrichment and parallel sequencing will make it possible to screen the coding regions of all X-chromosomal genes. This will not only lead to the identification of additional XLMR genes, but will also provide the basis for comprehensive diagnostic tests for XLMR families.

In conclusion, the results of the present study demonstrate the power of droplet-based multiplex PCR and next-generation sequencing of the amplification products for parallel mutation screening in a large number of genes. This paves the way for a dramatic improvement in diagnosing genetically heterogeneous disorders and to elucidate novel disease genes.

## Materials and methods

### Patients

Twenty-four families with X-linked mental retardation (XLMR) were selected from the cohort of families collected by the European Mental Retardation consortium

(EUROMRX; www.euromrx.com). In all families, the affected members suffered from non-syndromic mental retardation, and the pedigree structure was strongly suggestive of an underlying X-linked gene defect. No causative gene defects were known for 21 of these 24 families. In 11 families, linkage analysis had been performed previously, which provided additional evidence of an X-linked disorder and indicated linkage intervals that were supposed to harbour the mutated genes.

Three families (15, 18 and 24) were included in which the underlying gene defects were already known: an insertion of one nucleotide in *JARID1C* (c.202_203insC) in family 15 (Jensen et al. 2005), a base pair exchange in *IL1RAPL1* (c.1460G>A) in family 18 (Tabolacci et al. 2006), and a base pair exchange in *SLC6A8* (c.1661C>T) in family 24 (Rosenberg et al. 2004).

## Primer design

The primer library was designed using the manufacturer's design parameters (RainDance Technologies) and the Primer3 algorithm (http://frodo.wi.mit.edu/primer3/). All SNPs from dbSNP build 129 were filtered from the primer selection region. Repeat masking was not performed on the input regions to the primer design pipeline. The primer design pipeline performed an exhaustive primer selection across all of the regions submitted. The primer library design pipeline defined 2,075 unique amplicons successfully designing primers for 2,048 amplicons (98.7% success rate). 27 amplicons could not be designed based on the Primer3 parameter setting (amplicon GC content and primer Tm). From the 2,048 amplicons 1,912 were selected for the final design.

## Droplet based multiplex PCR

Genomic DNA samples were first fragmented to 2–4 kb using a nebulization kit (Invitrogen, K7025-05) following the manufacturer's recommended protocol. To prepare the input DNA template mixture for targeted amplification, 1.5 μg of the purified Genomic DNA fragments were added to 4.7 μl 10× High-Fidelity Buffer (Invitrogen, 11304-029), 1.26 μl of $MgSO_4$ (Invitrogen, 11304-029), 1.71 μl 10 mM dNTP (New England Biolabs, NO447S/L), 3.6 μl Betaine (Sigma, B2629-50G), 3.6 μl of RDT Droplet Stabilizer (RainDance Technologies, 30-00826), 1.8 μl dimethyl sulfoxide (Sigma, D8418-50 ml) and 0.72 μl 5 units/μl of Platinum High-Fidelity Taq (Invitrogen, 11304-029). The samples were brought to a final volume of 25 μl with Nuclease Free Water.

PCR droplets were generated on the RDT1000 instrument (RainDance Technologies, 20-01000). To process a single sample the user placed onto the RDT1000 a single tube containing 25 μl of Genomic DNA Template Mix, the primer droplet library (RainDance Technologies) and a disposable microfluidic chip (RainDance Technologies). The primer droplet library consists of a collection of individual primer droplets where each primer droplet contains matched pairs of forward and reverse primer (1.1 μM per primer) for each amplicon that is in the primer library. The RDT1000 generated each PCR droplet by pairing a single gDNA template droplet with a single primer droplet. The paired droplets flow past an electrode embedded in the chip and are instantly merged together. All of the resulting PCR droplets were automatically dispensed as an emulsion into a PCR tube and transferred to a standard thermal cycler for PCR amplification. Each single sample generated more than 1,000,000 single plex PCR droplets.

Samples were cycled in a Bio-Rad PTC-225 thermocycler as follows: initial denaturation at 94°C for 2 min; 55 cycles at 94°C for 15 s, 58°C for 15 s, 68°C for 30 s; final extension at 68°C for 10 min and hold at 4°C.

After PCR amplification the emulsion of PCR droplets was broken to release each individual amplicon from the PCR droplets. For each sample an equal volume of RDT 1000 Droplet Destabilizer (RainDance Technologies, 40-00830) was added to the emulsion of PCR droplets, the sample was vortexed for 15 s and then spun in a microcentrifuge at 12,000*g* for 10 min. The oil from below the aqueous phase was carefully removed from the sample. The remaining sample was then purified using a MinElute column (Qiagen, 28004) following the manufacturer's recommended protocol. The purified amplicon DNA was then tested on an Agilent Bioanalyzer to confirm that the amplicon profile matches the expected histogram profile.

## RainDance amplicon concatenation and shearing

The purified amplicons were chloroform extracted and ethanol precipitated to remove Taq-polymerase that may have remained bound to the ends of the amplicons. The ends of the amplicons were blunt ended and phosphorylated by adding all of the purified DNA to 2.5 μl 10× Blunting Buffer (NEB, E1201), 2.5 μl 1 mM dNTP Mix (NEB, E1201L) and sterile water to a total reaction volume of 25 μl. The reaction was incubated at 22°C for 15 min and immediately heated to 70°C for 5 min followed by placement on ice for 10 min. The amplicons were then concatenated using the NEB Quick Ligation kit according to the manufacturer's protocol. The chloroform extraction was repeated as before and the resulting pellet was resuspended in 100 μl of low TE. The sample was then fragmented as described in the standard Illumina workflow.

### Solexa sequencing

DNA fragments were then repaired to generate blunt ends by T4 polymerase and Klenow DNA polymerase, and phosphorylated with T4 polynucleotide kinase. After adding a single 'A' base to the 3′ end of the DNA fragments using Klenow exo (3′ to 5′ exo minus), we ligated Solexa adaptors (with a 'T' overhang) with the DNA fragments using DNA ligase. Ligated products (size range 150–200 bp) were gel purified on 2% agarose, followed by 18 cycles of PCR-amplification. We measured the DNA concentration with a Nanodrop 7500 spectrophotometer, and a 1 μl aliquot was diluted to 10nM. Adaptor-ligated DNA was hybridized to the surface of flow cells, and DNA clusters were generated using the Illumina/Solexa cluster station, followed by 36 cycles of sequencing on the Illumina/Solexa 1G analyzer, in accordance with the manufacturer's protocols.

### Data analysis

Sequence reads were compiled using a manufacturer-provided computational pipeline consisting of the open source Firecrest and Bustard applications (Illumina). All the reads with low quality were removed and the remaining reads were aligned onto human genome reference (NCBI36.1) with SOAP2.20 (Li et al. 2008) allowing up to 2 mismatches and without gap. The uniquely mappable reads in the alignment results were used to calculate the sequence depth of target regions.

For mutation calling, to avoid PCR artefacts during Solexa sequencing library construction, we removed all but one read with the same mapping coordinate and strandness. Sequence calls were then performed only on positions $\geq 2\times$ coverage and Phred-like quality score $\geq 30$. To call hemizygous single nucleotide variants (SNV), the minimum percentage of reads representing the specific allele was required to be 70%. The SNVs were compared with dbSNP (version 130) to identify the known SNPs. The unknown SNVs were evaluated by PolyPhen for functional effects (Sunyaev et al. 2001).

The sequencing reads that cannot be mapped using SOAP2.20 were aligned onto human genome reference (NCBI 36.1) using SOAP1.11, allowing a gap length of up to 3 bp and without any mismatch. The uniquely mappable reads were then used to construct the non-redundant alignment as above. To call hemizygous indels, the minimum percentage of reads representing the specific allele was required to be 70%. The results were compared with dbSNP (version130) to identify the known indels.

In order to assess the uniformity of enrichment, we calculated the depth of sequencing coverage for all the targeted positions except, the first and last 36 nt of each amplicon as well as the positions without full mappability. The reason for the removal of the terminal 36 nt on both sides of one amplicon were that they were often found in chimeric sequencing reads that comprised the ends of two concatenated amplicons. Since the chimeric reads could not be mapped using our alignment setting, the sequence depths obtained for these positions were underestimated. We also calculated the median depth of coverage for each amplicon according to the base depth. The reproducibility of the enrichment uniformity was evaluated by calculating the average correlation coefficients of sequencing depth across all pairwise comparisons.

The sequencing data have been deposited in NCBI Short Reads Archive (SRA) with accession number SRA010105.

**Acknowledgment** This work was supported by a grant from the Max-Planck Innovation Funds (to HHR).

## References

Albert TJ, Molla MN et al (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4(11):903–905

Bau S, Schracke N et al (2009) Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal Bioanal Chem 393(1):171–175

de Brouwer AP, Yntema HG et al (2007) Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium. Hum Mutat 28(2):207–208

Edwards MC, Gibbs RA (1994) Multiplex PCR: advantages, development, and applications. PCR Methods Appl 3(4):S65–S75

Gnirke A, Melnikov A et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27(2):182–189

Hodges E, Xuan Z et al (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39(12):1522–1527

Jensen LR, Amende M et al (2005) Mutations in the JARID1C gene, which is involved in transcriptional regulation and chromatin remodeling, cause X-linked mental retardation. Am J Hum Genet 76(2):227–236

Li R, Li Y et al (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24(5):713–714

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

Mercer TR, Dinger ME et al (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105(2):716–721

Meuzelaar LS, Lancaster O et al (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat Methods 4(10):835–837

Ng SB, Turner EH et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461(7261):272–276

Okou DT, Steinberg KM et al (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4(11):907–909

Olson M (2007) Enrichment of super-sized resequencing targets from the human genome. Nat Methods 4(11):891–892

Porreca GJ, Zhang K et al (2007) Multiplex amplification of large sets of human exons. Nat Methods 4(11):931–936

Ropers HH (2007) New perspectives for the elucidation of genetic disorders. Am J Hum Genet 81(2):199–207

Ropers HH (2008) Genetics of intellectual disability. Curr Opin Genet Dev 18(3):241–250

Rosenberg EH, Almeida LS et al (2004) High prevalence of SLC6A8 deficiency in X-linked mental retardation. Am J Hum Genet 75(1):97–105

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145

Sunyaev S, Ramensky V et al (2001) Prediction of deleterious human alleles. Hum Mol Genet 10(6):591–597

Tabolacci E, Pomponi MG et al (2006) A truncating mutation in the IL1RAPL1 gene is responsible for X-linked mental retardation in the MRX21 family. Am J Med Genet A 140(5):482–487

Tarpey PS, Smith R et al (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 41(5):535–543

Tewhey R, Warner JB et al (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27(11): 998–999

Turner EH, Lee C et al (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6(5):315–316

Varley KE, Mitra RD (2008) Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res 18(11):1844–1850