Next generation sequencing of pooled samples reveals new *SNRNP200* mutations associated with retinitis pigmentosa

Paola Benaglio[1], Terri L. McGee[2], Leonardo P. Capelli[1,3], Shyana Harper[2], Eliot L. Berson[2] and Carlo Rivolta[1]

[1]Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

[2]The Berman-Gund Laboratory for the Study of Retinal Degenerations, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, USA

[3]Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

Correspondence to:
Carlo Rivolta
Department of Medical Genetics
University of Lausanne
Rue du Bugnon 27
1005 Lausanne
Switzerland;
Phone: +41(21) 692-5451
FAX: +41(21) 692-5455
email: carlo.rivolta@unil.ch

ABSTRACT

The gene *SNRNP200* is composed of 45-exons and encodes a protein essential for pre-mRNA splicing, the 200 kDa helicase hBrr2. Despite the fact that complete lack of this protein is incompatible with cell survival, two independent heterozygous mutations in *SNRNP200* have recently been found to be associated with the retinal degenerative disease retinitis pigmentosa (RP) in two families from China. In this work the entire 35-Kb *SNRNP200* genomic region was analyzed in a cohort of 96 unrelated North American patients with autosomal dominant RP. To carry out this large-scale sequencing project, we performed ultra high-throughput sequencing of pooled, untagged PCR products and verified the presence of the detected DNA changes by Sanger sequencing of individual samples. One of the two previously known mutations (p.S1087L) was identified in 3 patients, and 4 new missense changes (p.R681C, p.R681H, p.V683L, p.Y689C) affecting highly conserved codons were identified in 6 unrelated individuals. We also critically evaluate this pooling approach, in particular with respect to the generation of false positive and negative results. We conclude that, although this method can be adopted for rapid discovery of new disease-associated variants, it still requires extensive validation to be used in routine DNA screening projects.

INTRODUCTION

Retinitis pigmentosa (RP) is a group of hereditary retinal diseases characterized by the progressive degeneration of rod and cone photoreceptors. The disorder typically begins with night blindness in adolescence and proceeds with gradual reduction of peripheral visual field with eventual development of tunnel vision and, in some cases, virtual total blindness. Early detection of this condition has been achieved by measuring retinal function by electroretinographic (ERG) testing (Berson, 1993), which represents the most reliable diagnostic tool for RP at all ages. RP is both clinically and genetically heterogeneous. This disorder affects almost 1 in 4000 people worldwide and can be inherited as an autosomal-dominant, autosomal-recessive or X-linked trait, and in rare cases also as a non-Mendelian trait (Hartong et al., 2006; Rivolta et al., 2002).

By linkage mapping and candidate gene screening, more than 60 genes have been associated so far with non-syndromic RP (RetNet database, http://www.sph.uth.tmc.edu/retnet/); however, mutations in these genes account for only about half of all reported cases (Hartong et al., 2006; Sullivan et al., 2006). Discovery of new causative genes by a candidate-functional approach is hampered by the labor intensive and costly methods of sequencing candidate genes in large numbers of patients. New and efficient methods of screening are therefore necessary to aid in the discovery of the remaining fraction of RP genes. In this context, the development of strategies based on "next-generation", or ultra high throughput DNA sequencing technologies is starting to provide new tools to analyze panels of different genes in several patients in a parallel fashion (Calvo et al.; Daiger et al., 2010).

Twenty causative genes for autosomal dominant forms of RP (adRP) have been identified so far, including several genes encoding pre-mRNA splicing factors: *PAP-1* (*RP9*) (Keen et al., 2002), *PRPF31* (*RP11*) (Vithana et al., 2001), *PRPF8* (*RP13*) (McKie et al., 2001), *PRPF3* (*RP18*) (Chakarova et al., 2002) and *SNRNP200* (*RP33*) (Li et al., 2010; Zhao et al., 2009).

Splicing is a ubiquitous process by which introns are removed from pre-mRNA to form mature mRNA. The enzymatic reactions take place in the spliceosome, a supermolecular complex containing five small nuclear ribonucleoproteins (snRNP) and ~200 other proteins (Jurica and Moore, 2003). The *SNRNP200* gene (or *ASCC3L1,* chromosome 2q11.2),

encoding for the 200-kDa helicase hBrr2, is essential for the unwinding of the U4/U6 snRNP duplex, which is a key step in the catalytic activation of the spliceosome (Laggerbauer et al., 1998; Raghunathan and Guthrie, 1998). This protein is homologous to Brr2 from yeast and belongs to the DExD/H box protein family. It consists of two consecutive Hel308-like modules, each composed of a DExD/H box domain with ATPase activity and a Sec63 domain (Lauber et al., 1996; Pena et al., 2009; Zhang et al., 2009). Recently, two different mutations of hBrr2 have been found in two Chinese families with adRP that showed linkage to the *RP33* locus (Li et al., 2010; Zhao et al., 2009; Zhao et al., 2006). These mutations, p.R1090L and p.S1087L, were identified following screening of candidate genes within the *RP33* linkage interval and are both located in the first Sec63 domain. It has been shown that the corresponding mutations in yeast affect the helicase activity of Brr2 (Zhao et al., 2009). No other hBrr2 mutations have been identified but the prevalence of mutations in this gene among patients with adRP is yet unknown. No genetic analyses have been performed so far on large cohorts of patients or in families that were not pre-selected for segregation of the diseases with the *SNRNP200* (*RP33*) genomic region.

We present here the results of the screening of the *SNRNP200* gene (45 exons, 44 introns) in 96 adRP families with unknown molecular genetic cause, mostly composed of Caucasian individuals. To reduce the time and costs required to screen such a large gene in several patients with classical techniques, we used ultra high throughput sequencing technology on pooled samples from multiple patients (Calvo et al.; Ingman and Gyllensten, 2009; Out et al., 2009). The potential advantages and the limitations of this method are evaluated.

MATERIALS AND METHODS

Patients and controls

This study was carried out in accordance with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Boards of the University of Lausanne and of Harvard Medical School and the Massachusetts Eye and Ear Infirmary, where the blood was collected and patients were followed. Written informed consent was obtained from patients who participated in this study before they donated 10-30 ml of their blood for research.

In addition to a regular ophthalmologic examination, our evaluation included ERG testing, performed as previously described (Berson et al., 1993). Patients were characterized as autosomal dominant if their families showed evidence of transmission of retinitis pigmentosa over two consecutive generations in at least one branch with or without evidence of reduced penetrance in other branches.

DNA from peripheral leukocytes was extracted from 191 unrelated patients with adRP. Ninety-six of these samples were used for screening with ultra high-throughput sequencing (UHTS), while the other 95 patients were analyzed only for those exons in which a mutation was confirmed after Sanger sequencing. Controls included 175 individuals with no history of retinal degeneration, and included 80 subjects with normal ERG. In instances where genomic DNA was insufficient for direct genetic screening, it was amplified by using a whole-genome amplification kit, following manufacturer's instructions (REPLI-g Mini Kit, Qiagen, Venlo, The Netherlands).

UHTS and sequence analysis

The general work flow followed in this work is schematically illustrated in Figure 1. Specifically, the *SNRNP200* gene was amplified in the initial set of 96 patients by 4 overlapping long-range (LR) PCRs of 9,009 bp, 10,474 bp, 12,145 bp and 5,594 bp in length, spanning in total an approximate 35-kb genomic region (chromosome 2, from position 96,300,768 to 96,335,728, of GenBank entry NC_000002.11). Primers were those used in the work by Hinds *et al.* (Hinds et al., 2005), adapted in some instances to the region of interest (Supp. Table 1). PCR reactions were performed in a 10 µl final volume, including LA PCR Buffer II (TaKaRa, Otsu, Shiga, Japan), 4 mM of $MgCl_2$, 1µM of each

primer, 0.4 mM of dNTPs and 1 U of TaKaRa LA Taq (TaKaRa), with slight modification in the cycling conditions suggested by the supplier. For the quantification of PCR products, we loaded 1 µl of each reaction (96 x 4) on E-Gel 48 2% agarose gels (Invitrogen, Carlsbad, CA) and analyzed them by densitometry. We pooled equimolar PCR products, according to the measured intensity of the bands. To avoid over-representation of the overlapping regions after shotgun library preparation and sequencing, we chose to pool only fragments from non-overlapping LR-PCRs (fragment #1 pooled with fragment #3 and #2 with #4). Sequencing was performed with 2 runs of Roche 454 GS FLX Titanium, according to manufacturer's protocols and by using a gasket that separated the 2 pools. All sequence analyses were carried out with the software package CLC Genomics Workbench (CLC bio, Aarhus, Denmark). Sequence reads were first trimmed and filtered according to their quality score and length (quality limit value set to 0.001, defined in the software manual; minimum length of a read set to 25 nt) and then assembled onto the reference sequence. We used default local-gapped alignment, allowing reads to align if they have at least 98% identity for more than 98% of their length.

For variant detection we applied the following restrictions on quality: within an 11-nt window, the average quality of the bases was set to 20 (PHRED score, corresponding to a base accuracy of 99%) and the maximum number of mismatches or indels accepted was 3 with respect to the reference sequence. We only considered calls having a minimal coverage of 1,000 reads, corresponding to at least 5 reads per allele per patient and to twice as much the threshold indicated previously for confident detection of variants (Ingman and Gyllensten, 2009). Finally, to be considered reliable DNA variants, all detected changes had to be present independently in the 2 technical replicates, represented by the 2 runs of sequencing, with at least a 0.5% frequency (corresponding roughly to 1 variant allele over 192 alleles).

Sanger sequencing and validation of mutations

To validate the changes detected by UHTS sequencing we individually analyzed the PCR products from each patient's DNA for 4 exons (16, 25, 37, and 38) containing putative mutations by the Sanger procedure (Sanger et al., 1977) on either long- or short-range PCR templates (Supp. Table 1 and Supp. Information). In addition, we sequenced exons 4 and 31 to further ascertain the precision of the variants called by the UHTS procedure. Sequencing

reactions were performed by mixing 5 μl of previously-purified PCR products (ExoSAP-IT, USB, Cleveland, OH), 0.75 μM of primers and 1 μl of BigDye Terminator v1.1 Cycle Sequencing kit (Applied Biosystems, Foster City, CA), and run on a ABI-3130XLS sequencer (Applied Biosystems).

To predict pathogenicity of amino acid substitutions we used both the PolyPhen (Ramensky et al., 2002) and MutPred (Li et al., 2009) software. Possible mutations affecting splicing were tested with the NNSPLICE 0.9 program (Reese et al., 1997). Protein sequences were aligned by using tools form the CLC Genomics Workbench.

RESULTS

Sequencing

We sequenced the *SNRNP200* gene with two runs of Roche 454 GS FLX Titanium, as a pool of individually obtained LR-PCRs in 96 unrelated patients with adRP (4,275 exons, 4,180 introns, or ~3.5 Mb in total). We obtained in total ~2.3 million raw sequences of 314 nt in length on average. Following trimming and quality filtering procedures, 87% of them aligned to the reference sequence. The average base coverage obtained was about 7,500 fold, corresponding to ~40 sequences per single allele per patient in the pool, assuming an even representation of each sample. Ninety-six percent of the reference sequence was covered by at least 1,000 reads.

Ascertainment of DNA variants

Following the filtering and selection criteria described above, we identified 79 DNA variants, including 33 annotated SNPs and 18 changes associated with homopolymeric stretches (i.e. AAAA…, CCCC…, etc.). Since these latter changes represent a well known source of errors for Roche 454 technology (Huse et al., 2007), they were immediately discarded from further analyses along with the identified known SNPs. Of the remainder variants, 21 were located within noncoding regions, 3 were predicted to produce isocoding changes, and 4 involved nonsynonymous changes. Putative isocoding changes were tested *in silico* for possible interference with the canonical splicing process, and none of them were predicted to be pathogenic. More specifically, the c3315A>G (p.A1105=) variant was in fact predicted to create a new donor site, but its associated likelihood score was not particularly high (0.43 out of 1.00).

To confirm the presence of DNA variants and identify the actual carriers among the patients' DNA composing the pool, we sequenced all exons carrying nonsynonymous variants as well as p.A1105= by the Sanger method in individual DNA samples. Whenever a change could be confirmed, the screening of that particular exon (and of its intron vicinities) was extended to the genomes of additional unrelated 95 adRP patients (Table 1).

In exon 16, the non-synonymous DNA change p.Y689C was confirmed by Sanger sequencing to be present heterozygously in one patient. Two missense variations affecting

both codon 681 (c.2041C>T and c.2042G>A, or p.R681C and p.R681H, respectively) were also identified by Sanger sequencing in 2 patients from the first cohort. These variants were initially not detected by UHTS because they were present in the pool with frequency values that were below the 0.5% threshold and therefore they can be considered as false negatives of the first method of screening. Sequencing of the second cohort allowed the identification of p.R681C in 2 additional unrelated patients, as well as the detection of 2 new DNA changes, p.V683L and c.2160+42C>T in 2 patients. None of these variants was present in 350 control chromosomes.

In exon 25, the change p.S1087L, detected by UHTS with a frequency of 1.4%, was present in 2 patients from the first cohort and 1 patient of the second cohort. The isocoding change p.A1105= was also confirmed to be present in two patients, one in each cohort. Again, these DNA variants were absent in controls.

Sanger sequencing of the amplicons spanning exons 37 and 38 identified 2 false positives of the UHTS screening, p.F1717S and p.M1808V, both having a measured frequency corresponding exactly to the threshold value used in inclusion criteria. Alleles from SNPs rs772175 and rs78519182 were also confirmed to be present, with relatively high frequency.

Cosegregation analyses

The p.S1087L mutation, found in 3 unrelated patients from our cohorts, was previously reported to be present in a family with adRP by Zhao *et al.* (Zhao et al., 2009).

The 4 new missense changes detected in exon 16, p.R681C, p.R681H, p.V683L and p.Y689C involved highly conserved residues (Figure 2) and were all predicted to be deleterious by *in silico* analyses. Family members were available only from 2 probands carrying p.Y689C and p.R681C. In these pedigrees, both changes were present heterozygously in patients and absent in unaffected members, following the classical pattern of inheritance of alleles causing a dominant disease with complete penetrance (Figure 3).

The intronic change c.2160+42C>T, for which we also had other family members, did not co-segregate with the disease in the family and was therefore considered as non-pathogenic.

Evaluation of variant detection specificity

To test the performance of the pooling method adopted here, we re-analyzed the sequence obtained by UHTS for exons 4, 16, 25, 31, 37, and 38 in the initial set of 96 samples. Specifically, we ascertained the number of variants detected by using different frequency thresholds (0.1, 0.2, … 1.0%) and compared them with the results obtained by individual Sanger sequencing of such samples. The number of false positives increased as the threshold of detection decreased, following an exponential-like trend (Figure 4).

DISCUSSION

The *SNRNP200* gene, encoding the splicing factor hBrr2, has been recently discovered by linkage analysis as a new autosomal dominant retinitis pigmentosa gene in two families from China, among whom mutations p.S1087L and p.R1090L segregated with the disease (Li et al., 2010; Zhao et al., 2009). Based on the evidence that hBrr2 is part of the same snRNP that includes PRPF31, PRPF3, and PRPF8, also involved in adRP, prior to the publication of these studies we screened this candidate sequence in a large cohort of unrelated patients from North America. Because of the elevated number of exons to be analyzed and the high potential of UHTS, we adopted previously published protocols consisting of the parallel sequencing of pooled and untagged DNA samples and evaluated them as potential methods for research on adRP, i.e. a rare disease with elevated genetic heterogeneity.

Three out of 191 patients from our screening carried p.S1087L (c.3260C> T), located in the first Sec63 domain of the protein. These patients were of Mexican, French Canadian, and English/Irish descent, indicating either that this mutation represents a relative early event in human history or that nucleotide c.3260 is a mutational hotspot. Haplotype studies on other ethnical groups and extended cohorts of patients are needed to verify which one of these hypotheses is the correct one.

Importantly, we found 4 new missense variants in exon 16: p.R681C, p.R681H, p.V683L, and p.Y689C, which were present heterozygously in 6 patients and absent from 350 control chromosomes. Residues Arg681 and Tyr689 are phylogenetically very well conserved, and their replacement is predicted *in silico* to be damaging for the correct functioning of hBrr2. While both p.R681C and p.Y689C co-segregated with disease in the pedigrees analyzed, probands carrying p.R681H and p.V683L changes did not have other family members available for further genetic analyses. However, the p.R681H variation affects the very same conserved residue co-segregating with disease in the pedigree with p.R681C, strongly suggesting an association with RP. p.V683L was predicted to be possibly pathogenic by *in silico* analyses, it involved a conserved amino acid, and was absent in the controls. In the absence of additional data (e.g. cosegregation) it is difficult to speculate at the present time whether it represents a rare benign variant or a true RP mutation.

All the newly detected changes fall in the Brr2 protein region containing the first DExD-helicase domain, which has been demonstrated to be essential for the U4/U6 unwinding function *in vivo* and *in vitro* and for cell survival in yeast (Kim and Rossi, 1999; Raghunathan and Guthrie, 1998). The first of the two consecutive Hel308-like modules, consisting of a DExD/H domain and a Sec63 domain, shows the highest level of conservation among species, reflecting its importance at the functional level (Zhang et al., 2009). It is therefore remarkable that all adRP mutations in hBrr2 so far identified are located in this first Hel308-like module. We hypothesize that these new mutations, similar to the ones already described, would impair hBrr2 helicase/ATPase activity, leading to defects in spliceosome catalysis.

Because of the high genetic heterogeneity displayed by RP, a very effective strategy for the identification of new disease genes consists in the screening of candidate genes in large cohorts of patients (Dryja, 1997). UHTS technologies (reviewed in (Metzker, 2010)) allow obtaining unprecedented amounts of DNA sequencing data, which make them suitable for the screening of large genes. However, UHTS analysis of multiple samples is not a straightforward procedure, and unavoidably requires sample pooling to be economically sustainable. Current multiplexing procedures mostly rely in the addition of nucleotide barcodes to individual samples since the use of physical separators does not grant sufficient parallelization. (Craig et al., 2008; Lennon et al., 2010; Meyer et al., 2008). Detection of sequence variants in multiple samples can also be achieved through sequencing a pool of non-tagged DNA templates (for example PCR products covering the same gene) from different individuals and by ensuring an appropriate coverage in downstream UHTS procedures (Calvo et al.; Ingman and Gyllensten, 2009; Out et al., 2009). This approach bypasses the expensive and laborious procedure of barcoding multiple libraries, and can theoretically lead to identification of rare variants the frequency of which is as low as 0.5% with respect to the pool.

We followed this latter approach to analyze *SNRNP200* for mutations. In our screening we detected an unexpectedly high number of both false positive and false negative calls, which could be ascertained only by Sanger sequencing. False positive calling of mismatches is a necessary drawback, since in our study we were considering very low frequency variations that could also be caused by sequencing or alignment errors. We could reduce them by considering only the subset of variations detected in two independent sequencing runs, as demonstrated also by our simulation experiments using variable thresholds and as

indicated by the reduction of the number of DNA changes associated to homopolymeric stretches (from more than 100 to 18, data not shown). However, false positive calls could not be completely eliminated. Using a higher threshold of detection could correct the problem but would also hide potentially true signals (false negatives), especially for variants that could be penalized by uneven pooling of different PCRs products and/or unbalanced allelic amplification during pre-sequencing procedures (Benaglio and Rivolta, 2010). In our specific case, we failed for example to identify 2 true changes, p.R681C and p.R681H, since they were present at a frequency that was below the theoretical limit of 1 variant allele in 96 samples (0.4% and 0.1% respectively). While failure to detect the first change could be attributed to statistical fluctuation, the second false negative call is more likely to depend on the under representation of this allele in the pool, probably prior to sequencing. However, correcting the under detection of true positive calls through the mere operation of increasing the threshold would also result in an exponential increase of noise generated by false positive calls, making the fine tuning of this procedure a subtle and rather empirical process. A practicable possibility in this context could consist in pooling fewer samples and raise the threshold of detection proportionally. In our case, for example, pooling 48 samples instead of 96 would have allowed detecting a single allelic variation in 1% of the sequences (instead of 0.5%), allowing therefore to increase the detection rate while keeping the noise under control.

The DNA screening strategy used in this work has proven to be extremely advantageous, especially if it is compared to the alternative option of individually sequencing all *SNRNP200* exons in the several dozen patients and controls examined. Specifically, the triage operated by UHTS of pooled samples allowed reducing the number of exons to be analyzed by an order of magnitude (from 45 to 4). However, in contrast to classical exon-PCR analyses by Sanger sequencing or to UHTS of single samples, the results obtained have a stochastic component that depends heavily on the settings used.

In conclusion, we identified new mutations in *SNRNP200* and confirmed that adRP associated to hBrr2 impairment is not limited to the Chinese population. Furthermore, we also tested the use of next-generation sequencing technology on pooled and untagged samples, highlighting the advantages and the limitations of this methodology for DNA analyses involving multiple patients. Based on our work, we are persuaded that candidate gene screening for RP and other genetic diseases will greatly benefit from the high-throughput revolution in a very near future, but this would probably follow the development

of automated and inexpensive procedures for genetic barcoding or other solutions for sample multiplexing.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. M. Künzli, FGCZ, Zurich, Switzerland. Our work was supported by the Swiss National Science Foundation (grant # 320030-121929) and by the European Union (grant HEALTH-2007-201550).

This work was supported by a Center Grant from the Foundation Fighting Blindness, Columbia, MD. (ELB).

Leonardo P. Capelli was sponsored by the CAPES program (Process 3637/07-7).

REFERENCES

Benaglio P, Rivolta C. 2010. Ultra high throughput sequencing in human DNA variation detection: a comparative study on the NDUFA3-PRPF31 region. PLoS One 5.

Berson EL. 1993. Retinitis pigmentosa. The Friedenwald Lecture. Invest Ophthalmol Vis Sci 34:1659-1676.

Berson EL, Rosner B, Sandberg MA, Hayes KC, Nicholson BW, Weigel-DiFranco C, Willett W. 1993. A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa. Arch Ophthalmol 111:761-772.

Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, Wilson CJ, Altshuler D, Gabriel SB, Daly MJ, Thorburn DR, Mootha VK. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42:851-858.

Chakarova CF, Hims MM, Bolz H, Abu-Safieh L, Patel RJ, Papaioannou MG, Inglehearn CF, Keen TJ, Willis C, Moore AT, Rosenberg T, Webster AR, Bird AC, Gal A, Hunt D, Vithana EN, Bhattacharya SS. 2002. Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa. Hum Mol Genet 11:87-92.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5:887-893.

Daiger SP, Sullivan LS, Bowne SJ, Birch DG, Heckenlively JR, Pierce EA, Weinstock GM. 2010. Targeted high-throughput DNA sequencing for gene discovery in retinitis pigmentosa. Adv Exp Med Biol 664:325-331.

Dryja TP. 1997. Gene-based approach to human gene-phenotype correlations. Proc Natl Acad Sci U S A 94:12117-12121.

Hartong DT, Berson EL, Dryja TP. 2006. Retinitis pigmentosa. Lancet 368:1795-1809.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143.

Ingman M, Gyllensten U. 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. European Journal of Human Genetics 17:383-386.

Jurica MS, Moore MJ. 2003. Pre-mRNA splicing: awash in a sea of proteins. Mol Cell 12:5-14.

Keen TJ, Hims MM, McKie AB, Moore AT, Doran RM, Mackey DA, Mansfield DC, Mueller RF, Bhattacharya SS, Bird AC, Markham AF, Inglehearn CF. 2002. Mutations in a protein target of the Pim-1 kinase associated with the RP9 form of autosomal dominant retinitis pigmentosa. Eur J Hum Genet 10:245-249.

Kim DH, Rossi JJ. 1999. The first ATPase domain of the yeast 246-kDa protein is required for in vivo unwinding of the U4/U6 duplex. RNA 5:959-971.

Laggerbauer B, Achsel T, Luhrmann R. 1998. The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro. Proc Natl Acad Sci U S A 95:4188-4192.

Lauber J, Fabrizio P, Teigelkamp S, Lane WS, Hartmann E, Luhrmann R. 1996. The HeLa 200 kDa U5 snRNP-specific protein and its homologue in Saccharomyces

cerevisiae are members of the DEXH-box protein family of putative RNA helicases. EMBO J 15:4001-4015.

Lennon NJ, Lintner RE, Anderson S, Alvarez P, Barry A, Brockman W, Daza R, Erlich RL, Giannoukos G, Green L, Hollinger A, Hoover CA, Jaffe DB, Juhn F, McCarthy D, Perrin D, Ponchner K, Powers TL, Rizzolo K, Robbins D, Ryan E, Russ C, Sparrow T, Stalker J, Steelman S, Weiand M, Zimmer A, Henn MR, Nusbaum C, Nicol R. 2010. A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. Genome Biol 11:R15.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744-2750.

Li N, Mei H, MacDonald IM, Jiao X, Hejtmancik JF. 2010. Mutations in ASCC3L1 on 2q11.2 are associated with autosomal dominant retinitis pigmentosa in a Chinese family. Invest Ophthalmol Vis Sci 51:1036-1043.

McKie AB, McHale JC, Keen TJ, Tarttelin EE, Goliath R, van Lith-Verhoeven JJ, Greenberg J, Ramesar RS, Hoyng CB, Cremers FP, Mackey DA, Bhattacharya SS, Bird AC, Markham AF, Inglehearn CF. 2001. Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13). Hum Mol Genet 10:1555-1562.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.

Meyer M, Stenzel U, Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. Nat Protoc 3:267-278.

Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PE, Tops CM, Breuning MH, van Ommen GJ, den Dunnen JT, Devilee P, Hes FJ. 2009. Deep sequencing to reveal new variants in pooled DNA samples. Hum Mutat 30:1703-1712.

Pena V, Jovin SM, Fabrizio P, Orlowski J, Bujnicki JM, Luhrmann R, Wahl MC. 2009. Common design principles in the spliceosomal RNA helicase Brr2 and in the Hel308 DNA helicase. Mol Cell 35:454-466.

Raghunathan PL, Guthrie C. 1998. RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. Curr Biol 8:847-855.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894-3900.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. J Comput Biol 4:311-323.

Rivolta C, Sharon D, DeAngelis MM, Dryja TP. 2002. Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. Hum Mol Genet 11:1219-1227.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

Sullivan LS, Bowne SJ, Birch DG, Hughbanks-Wheaton D, Heckenlively JR, Lewis RA, Garcia CA, Ruiz RS, Blanton SH, Northrup H, Gire AI, Seaman R, Duzkale H, Spellicy CJ, Zhu J, Shankar SP, Daiger SP. 2006. Prevalence of disease-causing mutations in families with autosomal dominant retinitis pigmentosa: a screen of known genes in 200 families. Invest Ophthalmol Vis Sci 47:3052-3064.

Vithana EN, Abu-Safieh L, Allen MJ, Carey A, Papaioannou M, Chakarova C, Al-Maghtheh M, Ebenezer ND, Willis C, Moore AT, Bird AC, Hunt DM, Bhattacharya SS. 2001. A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies

autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). Mol Cell 8:375-381.

Zhang L, Xu T, Maeder C, Bud LO, Shanks J, Nix J, Guthrie C, Pleiss JA, Zhao R. 2009. Structural evidence for consecutive Hel308-like modules in the spliceosomal ATPase Brr2. Nat Struct Mol Biol 16:731-739.

Zhao C, Bellur DL, Lu S, Zhao F, Grassi MA, Bowne SJ, Sullivan LS, Daiger SP, Chen LJ, Pang CP, Zhao K, Staley JP, Larsson C. 2009. Autosomal-dominant retinitis pigmentosa caused by a mutation in SNRNP200, a gene required for unwinding of U4/U6 snRNAs. Am J Hum Genet 85:617-627.

Zhao C, Lu S, Zhou X, Zhang X, Zhao K, Larsson C. 2006. A novel locus (RP33) for autosomal dominant retinitis pigmentosa mapping to chromosomal region 2cen-q12.1. Hum Genet 119:617-623.

Table 1. *SNRNP200* DNA variants in selected exons, detected by UHT and Sanger sequencing. With the exception of c.5317C>T, all changes were detected in heterozygous state.

DNA change*	Allele frequencies in the pool (%)	Coverage	Count of minor allele	Putative amino acid change	Detected with UHTS	Sanger in 1st cohort	Sanger in 2nd cohort	Controls	Pathogenicity
Exon 16									
c.2041C>T	99.6/0.4	10,627	44	p.R681C	No (false negative)	1	2	0	Likely
c.2042G>A	99.9/0.1	10,656	10	p.R681H	No (false negative)	1	0	0	Likely
c.2047G>T				p.V683L	No	0	1	0	Undetermined
c.2066A>G	99.3/0.7	9,677	69	p.Y689C	Yes	1	0	0	Likely
c.2160+42C>T				Intronic	Yes	0	1	0	No
Exon 25									
c.3260C>T	98.6/1.4	20,258	290	p.S1087L	Yes	2	1	0	Confirmed
c.3315A>G	99.1/0.9	17,203	153	p.A1105A	Yes	1	1	0	Undetermined
Exon 37									
c.5150T>C	99.5/0.5	12,064	65	p.F1717S	Yes (false positive)	0	ND	ND	No (not a real variant)
c.5317C>T (rs772175)	64.9/35.1	7,975	2802	p.L1773=	Yes	63 (alleles)	ND	ND	No
c.5324-31G>C (rs78519182)	98.0/2.0	7,553	149	Intronic	Yes	3	ND	ND	No
Exon 38									
c.5422A>G	99.5/0.5	9,409	51	p.M1808V	Yes (false positive)	0	ND	ND	No (not a real variant)

* with respect to Ensembl reference cDNA sequence ENST00000323853.

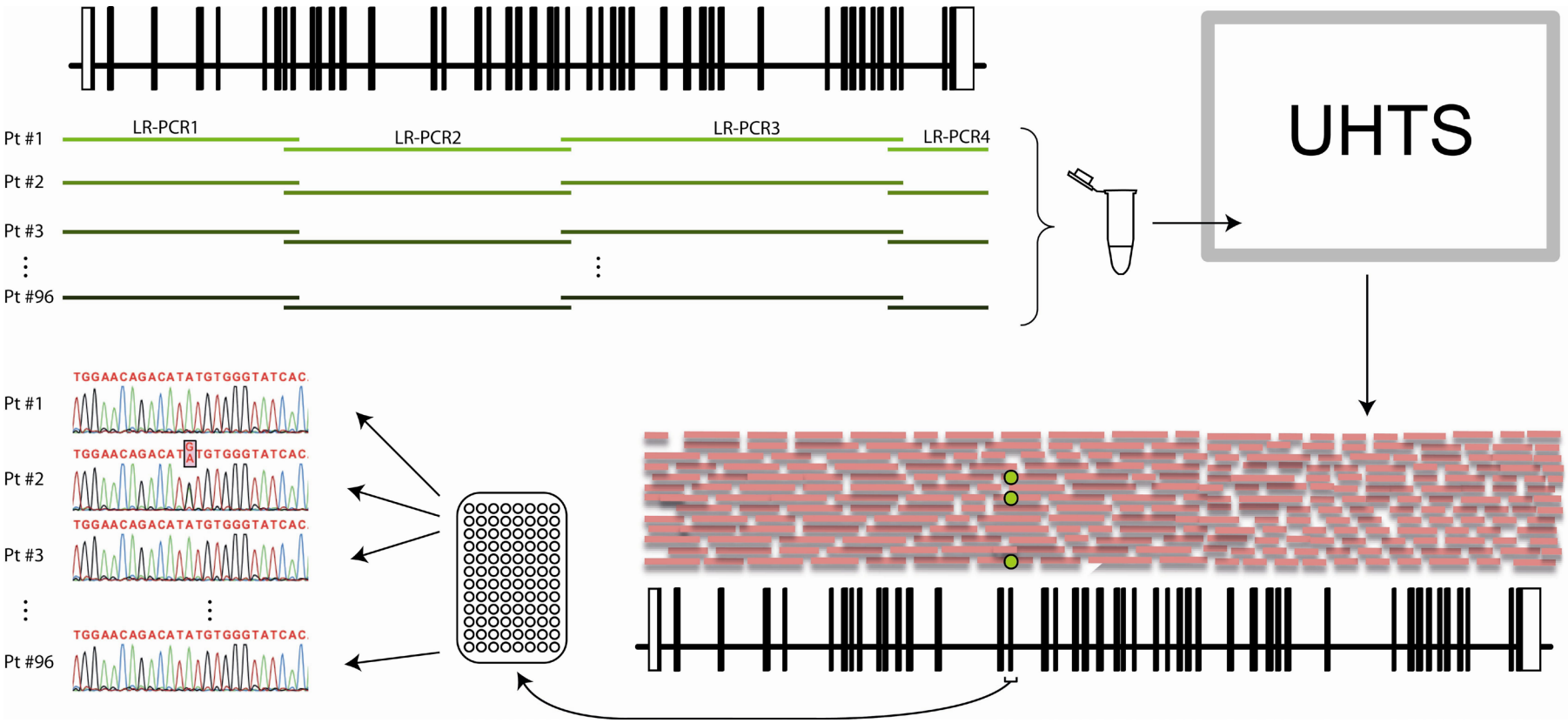
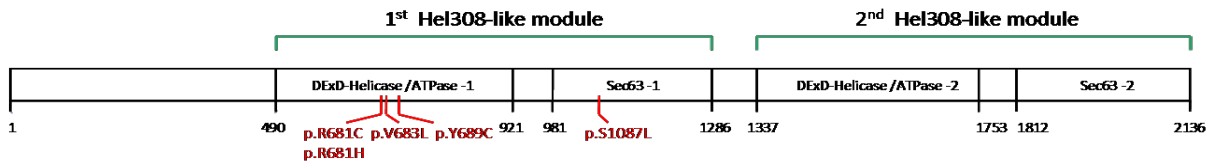


Figure 1. Schematic representation of the work flow used.

Four long-range PCRs (LR-PCR, in shades of green) targeting the *SNRNP200* gene (vertical boxes: exons; horizontal lines: introns) were performed on individual DNA samples from 96 patients, purified and pooled in equimolar quantities before UHTS sequencing. Following alignment to the reference sequence of the 2 million reads obtained (pink bars), DNA variants were identified in the pool (green circles). To verify the presence of the variants detected, as well as to ascertain which patients carried them, relevant regions of *SNRNP200* were re-amplified by regular PCR in all 96 patients and sequenced individually by the Sanger procedure.

A



B

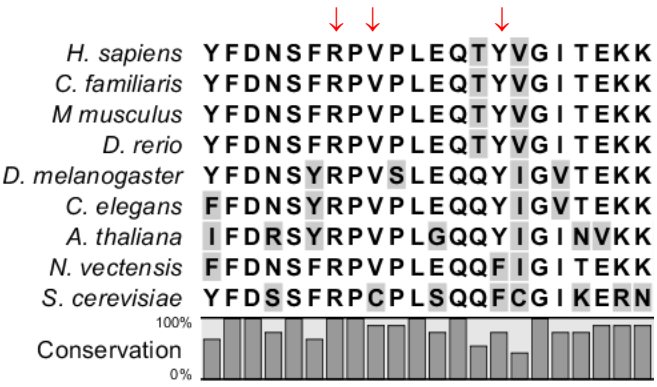


Figure 2. Structure and sequence of the hBrr2 protein

A) Functional domains of hBrr2. Location of the mutations found in this screening are indicated in red. B) Alignment of Brr2 protein sequences from human, dog, mouse, fish, fly, worm, plant, sea anemone and yeast. Non-conserved residues are shaded; arrows indicate the residues affected by DNA changes detected in exon 16.

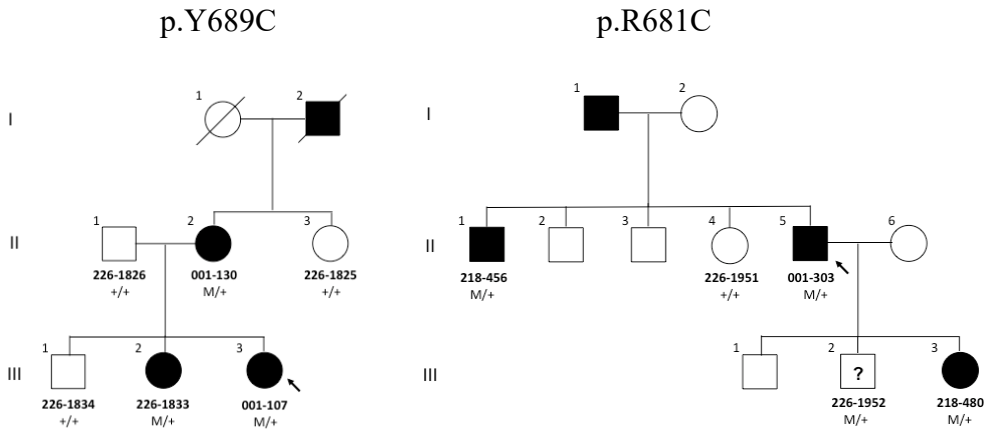


Figure 3. Pedigrees and allele segregation analysis of two families affected with adRP.

Pedigrees segregating the p.Y689C (family ID: 5632) and p.R681C (family ID: 0270) mutations (M) are shown. Black and white symbols represent clinically affected and unaffected members, respectively. The question mark indicates an individual for whom clinical examination was not possible. Arrows indicate probands analyzed in the UHTS screening.

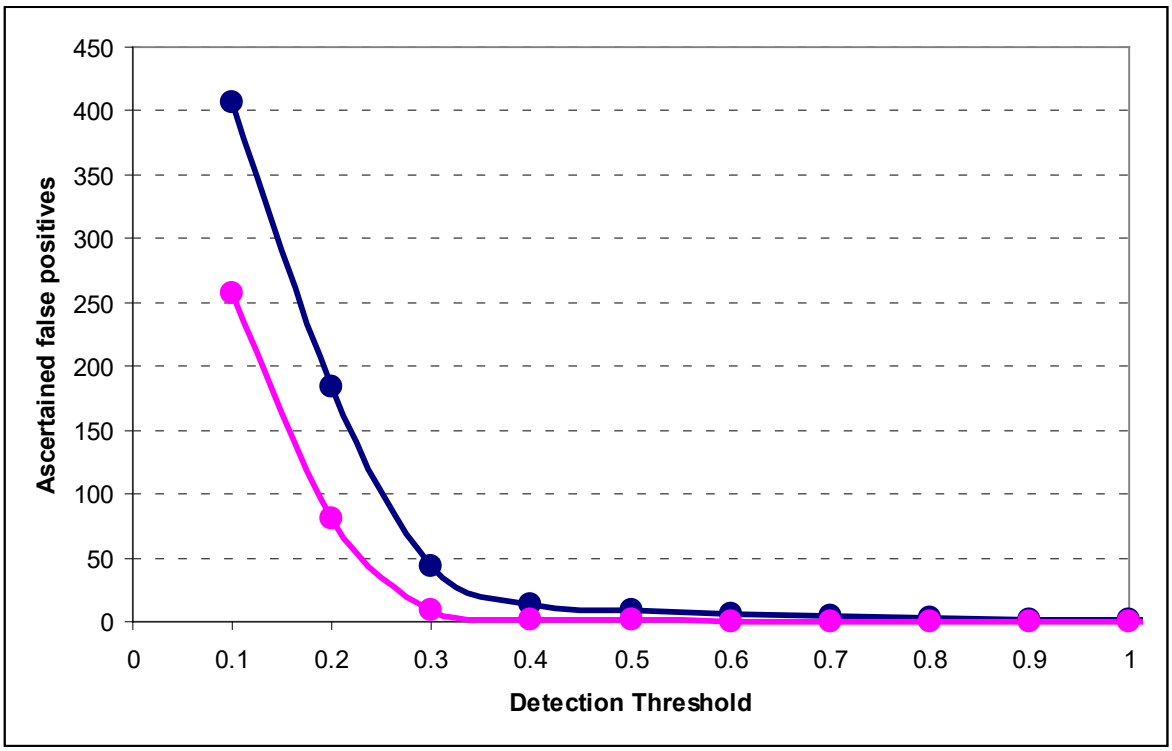


Figure 4. False positives identified as a function of different frequency thresholds.

False positives were ascertained after Sanger sequencing of exonic DNA, covering in total ~3% of the entire *SNRNP200* region sequenced by UHTS. Blue line: variants detected in any of the technical replicates. Pink line: variants detected in both technical replicates.

SUPPLEMENTARY INFORMATION

Supplementary Table 1.

Sequences of primers used for long-range (LR), regular (SR) and sequencing (Seq) PCRs.

Name	Sequence (5'-3')
LR_1.F	ACAGAGAAGACTTTGTAGCTGGGGAAGA
LR_1.R	CAGCCCTGAAATAAACTATATATGAAACAAGG
LR_2.F	TGGTTTGGTCATGAGACCAGTGACCTG
LR_2.R	ACACAAAACACAGTCATTAAAGGCAGACACTG
LR_3.F	GCTGCTCTTCATCTTTACCTCTAAGAA
LR_3.R	TAAACATGACAGTATCTGGTTTCTGCTATCAA
LR_4.F	CTGGTAGCTGGCTTGGTCAGGTGTCAACTCAC
LR_4.R	TGATGGGGAGGTGGCCTTCTGGAAGTATCAG
SR_ex16.F	GTTTTAGAAGGGCCTTTGGG
SR_ex16.R	TTTTAATTTCTGTCAATCTTCCCC
SR_ex25.F*	ACCGTGTGTAGAGTGGCTCA
SR_ex25.R*	TTCCCATCAGACCCTTGG
SR_ex37-38.F	GCGTATTGTCCACCAGTGATG
SR_ex37-38.R	TCCTCGATGCTGATGCACTT
Seq_ex4.F	TCCTTTAGTTGTGGCATCAGC
Seq_ex25.F	GCCGCAGCACTCTTCTAATTGT
Seq_ex31.R	TTTGGAATAGGGCAGCAGGTAG
Seq_ex37.F	TTAGGTCTCACACAGGGACCATG

* From Li *et al.* (Li et al., 2010)

Long-Range PCRs

Reaction mix

1x Buffer LA PCR™ Buffer II
4 mM MgCl2
1uM each primer
0.4 mM each dNTP
1 unit of LA Taq™ (TaKaRa)
10 ul final volume

Cycling conditions

LR 1
95˚C 5'(95˚C 30", 67˚C 1',68˚C 14')x14, (95˚C 30", 62˚C 1',68˚C 14')x16, 72˚C 10'

LR 2, 3, and 4
94˚C 1'(98˚C 5",68˚C 15')x30, 72˚C 10'

Regular PCRs

Exon 16

1x PCR Buffer
2 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
0.5 unit of HotStarTaq DNA Polymerase (Qiagen)
20 ul final volume

95˚C 15'(95˚C 30", 56˚C 30", 72˚C 1')x35, 72˚C 10'

Exon 25

1x PCR Buffer
1.5 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
1 unit of HotStarTaq DNA Polymerase
25 ul final volume

98˚C 8'(94˚C 30", 56˚C 30", 72˚C 1') x5, (94˚C 30", 54˚C 30", 72˚C 1')x5, (94˚C 30", 52˚C 30", 72˚C 1')x15, (94˚C 30", 50˚C 30", 72˚C 1')x15 72˚C 5'

Exons 37-38

1x PCR Buffer
0.5 mM MgCl2
0.1 uM each primer
0.2 mM each dNTP
0.5 unit of HotStarTaq DNA Polymerase
20 ul final volume

95˚C 15'(95˚C 30", 60˚C 30", 72˚C 1') x35, 72˚C 10'