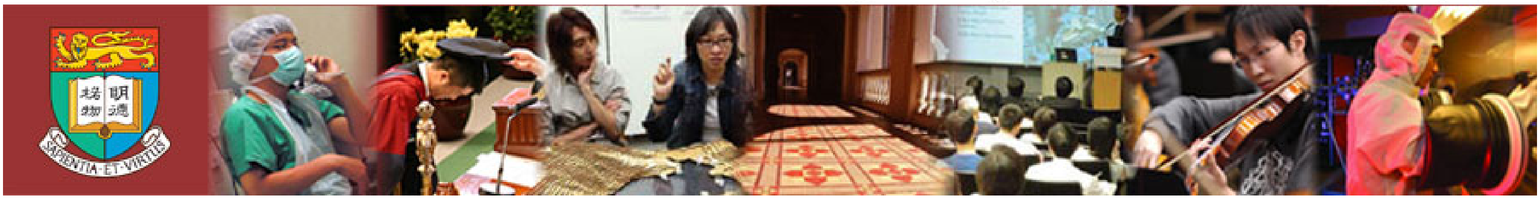

Targeted next-generation sequencing on Hirschsprung disease: a pilot studyexploitsDNA pooling

Hongsheng GUI[1,2], Jessie Yunjuan BAO[7,9], Clara Sze-Man TANG[1,2], Man-Ting SO[1], Diem-Ngoc NGO[6], Anh-Quynh TRAN[6], Duc-Hau BUI[6], Duy-Hien PHAM[6], Thanh-Liem NGUYEN[6], Amy TONG[7,8], Si LOK[7,8], Pak-Chung SHAM[2,3,4,5], Paul Kwong-Hang TAM[1,4], Stacey S CHERNY[2,3†], Maria-Mercè GARCIA-BARCELO[1,3,4†]

[1]Department of Surgery, [2]Department of Psychiatry, [3]Center for Genomic Sciences and [4]Centre for Reproduction, Development, and Growth of the Li KaShing Faculty of Medicine, The University of Hong Kong and, [5]State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong SAR, China; [6]Department of Human Genetics, National Hospital of Pediatrics, Hanoi, Vietnam; [7]Genome Research Centre, The University of Hong Kong; [8]Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China; [9]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China.

[†]Corresponding authors: Drs. Stacey CHERNY (The Hong Kong Jockey Club Building for Interdisciplinary Research,5 Sassoon Road, Pokfulam, Hong Kong; Phone (852) 28315073cherny@hku.hk) & Maria-Mercè GARCIA-BARCELO (The Hong Kong Jockey Club Building for Interdisciplinary Research,5 Sassoon Road, Pokfulam, Hong Kong; Phone: (852) 28315079,Fax:(852) 2819 9621; mmgarcia@hku.hk)

Running Head: Targeted resequencingand DNA pooling

SUMMARY

To adopt an efficient approach of identifying rare variants possibly related to Hirschsprung disease (HSCR), a pilot study was set up to evaluate the performance of a newly designed protocol for next generation targeted resquencing. In total 20 Chinese HSCR patients and 20 Chinese sexmatched individuals with no HSCR were included, for which coding sequences (CDS) of 62 genes known to be in signaling pathways relevant to enteric nervous system (ENS) development were selected for capture and sequencing. Blood DNAs from eight pools of five cases or controls were enriched by PCR-basedRaindance technology (RDT) and then sequenced on a 454 FLX platform. As technical validation, five patients from case Pool-3 were also independently enriched by RDT, indexed with barcode and sequenced with sufficient coverage. Assessment for CDS single nucleotide variants showed DNA pooling performed well (specificity/sensitivity at98.4%/83.7%) at the common variant level; but relatively worse (specificity/sensitivity at 65.5%/61.3%) at the rare variant level. Further Sanger sequencing only validated five out of 12 rare damaging variants likely involved in HSCR. Hence more improvement at variant detection and sequencing technology is needed to realize the potential of DNA pooling for large-scale resequencing projects.

Keywords: HSCR; targeted resequencing; CDS; rare variant; Raindance

INTRODUCTION

Targeted resequencing is a direct approach to pinpoint the causal or associated rare variants/mutations (RVs) in one or a few given genes (Rehm, 2013). The combination of PCR-based enrichment (for targeted sequence regions) and first-generation sequencing platforms (Sanger sequencing on ABI 3730 sequencer) is very efficient and is considered as the gold standard method for validating DNA mutations. However, due to the high cost, large-scale Sanger sequencing is not applicable to the screening of patients affected by multi-genic diseases where the number of candidate genes to be screened for RVs may easily be larger than 10.Next generation sequencing and microarray-based enrichment technologies (multiplex capture)offer not only high coverage of sequence reads, but also good capture specificity for targeted DNA regions withinaffordable research budgets(Summerer *et al.*, 2009). This combination, next generation targeted resequencing, has been adopted for clinical genetic testing ofhereditary cancers, cardiac diseases and sensory disorders where the genetic mechanisms are relatively well understood(Rehm, 2013). Meanwhile, it also enables population-based research that aims to further explore the roles of RVs in a batch of candidate genes identified by previous linkage or association studies.

Hirschsprung disease (HSCR, aganglionicmegacolon) is a complex genetic disorder of the enteric nervous system (ENS) characterized by the absence of enteric neurons along a variable length of the intestine. This is attributed to a failure in the migration of the enteric neuronprecursors, the neural crest cells (NCCs) (Torfs, 2004, Amiel *et al.*, 2008). The receptor tyrosine kinase gene (*RET*) is the major HSCR gene and its expression is crucial for the development of the enteric ganglia and hence the ENS (Amiel & Lyonnet, 2001). RVs in the coding sequence (CDS) of *RET* account for up to 50% of the familial cases and between 10%-20% of the sporadic cases. In addition to CDS mutations, *RET* common variants (single nucleotide polymorphisms -SNPs) are strongly associated with HSCR (Emison *et al.*, 2005). Yet, despite the importance of *RET*, additional genes (acting either in conjunction or independently from *RET*) are necessary to explain not only the disease incidence but also its complex pattern of inheritance. Other HSCR genes identified so far mainly code for protein members of inter-related signaling pathways involved in the ENS development: endothelin receptor B (*EDNRB*), the transcriptional regulator *SOX10*and the *NRG1/ERBB2*signaling pathway gene(Hofstra *et al.*, 1999, Heanue & Pachnis, 2007, Garcia-Barcelo *et al.*, 2009).CDS mutations in genes other than *RET* thus far only account for 7% of the cases, however the unexplained phenotypic variance in sporadic HSCR cases is still very high (>80%)(Gui *et al.*, 2013).This suggests that variations in other genes involved in ENS development are likely to contribute to the manifestation of the phenotype (Heanue & Pachnis, 2007). Nevertheless, not many of the known or candidate genes have been systematically screened in HSCR patients aside from *RET*. Hence it is promising to adopt multiplex capture and parallel sequencing to efficiently identify rare variants residing in genemembers of these signaling pathways relevant to processes that, when altered, give rise to diseases.

Targeted sequencing on complex diseases has previously shown that a population-based design involving > 1000 cases/controls is always required (Wu *et al.*, 2011). However, a few technical issues need to be handled before this large-scale sequencing project proceeds on HSCR. This includes selection of candidate genes, design of capture kits, and choice of multiplexing sequencing, among others (Mertes *et al.*, 2011). Therefore, we set up a pilot study in which a new strategy of NGS-based target resequencing was conducted on a relatively small sample of HSCR patients and normal controls. Evaluation of this strategy would provide technical guidance on future targeted resequencing studies on complex diseases including HSCR.

MATERIALS AND METHODS

Sample collection

HSCR diagnosis was based on histological examination of either biopsy or surgical resection material for absence of enteric plexuses. Patients were also fully characterized and a *pro-formaquestionare*was provided for each patient. In total 20 HSCR patients (16 males and fourfemales) were included in this pilot study. As controls, we included 20 Chinese individuals (16 males and four females) with no HSCR. Adult individuals gave written informed consent and for minors, written informed consent was obtained from their parents. The study was approved by the institutional review board of The University of Hong Kong together with the Hospital Authority (IRB:UW09-360).

Gene selection

Candidate genes were selected according to the evidence oftheir involvement in ENS development (provided by either the phenotype presented by mouse models, *in vitro* data,or*in silico* data). All of the14 genes known to cause HSCR in human newborns and ENS failure in the embryonic mouse (**Table 1**) werecompulsorily included, while the other 48 genes (**Table S1**) were selected according to the following reasons: 1) key members in the same signaling pathway with know HSCR genes; 2) genes with suggestive association with HSCR in our previous GWAS study (Garcia-Barcelo *et al.*, 2009); 3) supported by murineor functional studies as outlinedin Table S1.

DNA pooling

To minimize the cost, a protocol of DNA pooling that was previously recommended in large-scale genotyping studies was adopted (Sham *et al.*, 2002). In detail, DNA wasextracted fromperipheral blood samples (20 cases and 20 controls) and was fragmented into 2~4Kb fragments by sonication. DNA concentrations were measured by picogreen reagent in triplicate and were normalized to 10ng/μl.These genomic DNA samples were then arranged into pools with five cases or controls each in order to maximize the usage of the multiplex design, given the length of gene locus and PCR amplification. In overall, eight DNA templates (pool of DNA from five individuals and PCR reagents) were prepared,

Enrichment and capture

RainDance Technology (RDT 1000; RainDance Technologies, Lexington, MA, USA) was chosen to conduct multiplex capture for targeted regions (all exons + 2kb of 5' and 3' flanking UTR) ofthe 62 candidate genes. Basically, RDT allows us to PCR-amplify up to 4000 selected 500bp length sequences per DNA template by providing a custom-made pooled primer-pair library (http://www.raindancetechnologies.com). In brief, aliquoted primers were loaded on the RDT 1000 enrichment chip along with a conventional PCR mix and 3-4μg of fractionated (to facilitate primer annealing) pooled-genomic DNA. During this process, each primer pair droplet merged with a droplet that contained genomic DNA, and PCR reagents. Approximately 1.5 million droplets were collected in a 0.2mL single tube and were directly PCR amplified. Amplification products were recovered by breaking the emulsion followed by purification using a MiniElute PCR Purification Kit (QIAGEN, QIAGEN Hong Kong Pte. Limited, Hong Kong).The whole strategy isshown in Figure

S1.

Sequencing

During preparation of the sequencing library, purified PCR products generated in the previous steps were ligated to two distinct oligonucleotide adapters to provide priming sites for subsequent amplification and sequencing (Shendure & Ji, 2008, Rothberg & Leamon, 2008).The templates were amplified and immobilized by compartmentalizing individual template molecules and 28-m DNA capture beads within droplets of an emulsion. PCR reactions conducted inside the droplets amplified the template molecules and complementary primers covalently attached to the DNA capture beads immobilizedthe template. Template-covered DNA capture beads were loaded into individual wells etched into the surface of a fiber-optic slide.The sequencing process is the same as that used in pyrosequencing (454 Genome Sequencer FLX, Roche Inc., Branford, USA). Raw sequence reads were handled by the GS Data Analysis Software package accompanied by the GSFLXsystem, using gsMapper for alignment (human reference genome 19) and Newbler for variant calling (Roche Inc.). Only those single nucleotide variants (SNVs) with relatively high quality annotations (total read depth > 40, alternative allele > 3, and phred score > 20) were included for downstream analysis.

Technical validation

To evaluate the proposed protocol at detecting true single nucleotide variants, one pool of samples (five HSCR patients in case pool-3) was individually sequenced with high coverage (>300×).This pool was selected since all those five patients were also genotyped by Affymetrix 500K in our previous genome-wide association study (Garcia-Barcelo *et al.*, 2009). The same RainDancetechnology was used to capture and enrich coding sequences of 62 candidate genes; however, DNA template was prepared using independent index primersfor each individual without pooling and was then sequenced on an IlluminaGAIIX platform (Illumina, San Diego,USA). Raw reads in fastq format were exported and aligned to human reference genome build 19 with BWA software (Li & Homer, 2010). The GATK pipeline was then used to preprocess each sequence (re-alignment, recalibration) and to call SNVs on multiple samples together (DePristo *et al.*, 2011).Only SNVs with total read depth > 8, allelic ratio for alternative allele > 0.25, and Phred score > 20) were reported.

Variant assessment

All genetic variants were annotated by KGGSeq and then classified into noncoding or coding (defined by Refgene), common or rare (minor allele frequency < 0.01 in dbSNP137), and damaging or non-damaging by Polyphen2(Li *et al.*, 2012). Rare damaging variants detected in pool-3 patients were validated by Sanger sequencing. The sensitivity and specificity of detecting two types of variants (common and rare) by DNA pooling were calculated from the overlapped or non-overlapped counts between variants found in one sequenced pool and variants found in all individually sequenced patients.

Rare variant prioritization

In order to pinpoint the most deleterious mutations that separate cases from controls, a step-by-step filtering was conducted on all variants found in different pools according to MAFreported in public databases (1000 Genomes Project, dbSNP137 and NHLBI Exome sequencing project), their functional impact and variant/gene recurrence across pools. Burden analysis (binomial test for the variant count difference between cases and controls) was performed on variants or genes retained at each level(Zhao *et al.*, 2012). A final set of rare damaging variants from genes present in case pools only were sent for validation using Sanger sequencing.

RESULTS

Coverage (average read depth per base on targeted regions) and target specificity (unique reads on targets over total unique reads) were used to evaluate read-level performance. As shown in Table

2, the average read depth across the target regions for RDT was above 80×across eight pools, with range from 54.21× to 93.53×. In addition, the target specificity was 74.18% forRDT, while 74.12% of the RDT targeted bases were covered with depth >=40. In terms of variant calling, our protocol detected 943 raw SNVs (149 within CDS, 13.6%) per pool. Only those SNVs located in CDS regions were assessed. In overall, 89.5% CDS SNVs were also found in dbSNP137 (Table 2).

Variants detected in the pool-3 sample were compared between pooled sequencing and individual indexed sequencing on the same five cases. The validity of using individual sequence as technical validation was indicated by their read depth coverage (>300×), dbSNP137 coverage (~92%) and concordance with GWAS array (>95%) (Table 3). In general, the specificity and sensitivity for pooled sequencing were 92% (139 out of 151) and 79.4%(139 out of 175) respectively. When classified into common or rare variants according to dbSNP137 (MAF at 0.01), the specificity and sensitivity were98.4%/83.3.7%, or 65.5%/61.3% respectively (Table 4). This difference was verified by using small set common variants overlapped with the SNP array, and rare damaging variants validated by Sanger sequencing. By pooled sequencing, 11 out of 13 SNPs were found and all of them matched the minor/major allele by counting sequence reads covering reference or alternative alleles. On the other hand, six out of 10 rare damaging variants found by pooled sequencing were verified by Sanger, compared to seven out of eight variants for individual sequencing.

Among all eight pools, 366 coding sequence SNVsin20 different HSCR cases and 20 healthy controls were detected by our protocol. The distribution of variant count in cases and controls is shown in Table 5, as stratified by allele frequency and deleteriousness. No significant burden was found for mutation load at any level between cases and controls (p-value >0.05; Table 5). With regard to genes carrying rare damaging mutations, thesecould be classified into case-only, case-control and control-only by the presence of at least one mutation. None of these, however,were shown to have a significant burden in HSCR cases versus healthy controls(data not shown). Nevertheless, genes uniquely mutated in patients only may be considered as functional hits for further analysis. Therefore 12SNVscontained in these genes were selected for validation by Sanger sequencing. Finally,fiveout of these 12 variants were successfully validated (data not shown);at the phenotype level, only fourout of 20 patients (oneshort male, oneshort female and twolong female cases; 20%) carried at least one validated mutation with high functional impact.

DISCUSSION

In our current study,we designed a protocol combining DNA pooling, RDT multiplex capture and next generation sequencing together for targeted resequencing research. Its performance was carefully evaluated through the application on an HSCR case/control dataset. The idea of DNA pooling was widely applied for studying disease risk variants in large-scale population samples(Sham *et al.*, 2002). An efficient association can be done by comparing minor allele frequency between case pools and control pools, without knowing the exact individual genotype. This was supported by our pooled sequencing for common variants, for whichthe presence of a variant can be detected by counting reads harboring reference alleles and alternative alleles.However, this approach did not seem reasonable when examining rare variants, as revealed by the high false positive rate and high false negative rate found in our study.

It is true that barcoding individual samples in targeted sequencing would be a better approach for detecting rare variants (Cummings *et al.*, 2010). However, it will be much more laborious and expensive than DNA pooling in a large-scale population-based study. From the perspective of other technologies besides DNA pooling in our protocol, on-target reads showed capture specificity for RDT was relatively better than hybridization-based capture on the same 62 candidate genes (our own published data, the Centre for Genomic Sciences, the University of Hong Kong), which may be ascribed to its higher specificity for pseudo-genes (including other paralogous sequences) and total targeted regions < 1M base pairs(Mondal *et al.*, 2011). Given the advantage of multiplex capture technology been considered, the current drawback could be overcome by improving the performance of the NGS sequencer in terms of read accuracy together with cost-efficiency. It is sensible to envisage the potential of DNA pooling as sequencing technology advances.

Because of the intricate ways in which biological processes operate, rare (especially damaging) variants are likely to differ among individuals with the same disease, thus accounting for phenotypic variability. Of note is the fact that RVs identified in HSCR patients happen to belong to pathways involved in ENS development (Heanue & Pachnis, 2007, Heanue & Pachnis, 2006). However, we were not able to prioritize genes by binomial tests mainly because of limited power given the small sample size, low frequency of effective variants and relatively low specificity and sensitivity of the technologies (Kiezun *et al.*, 2012). A simplified estimation on variant level indicated that 600 case-control pairs is needed to detect RVs with odds ratio (OR) above 2 and minor allele frequency above 1% (supplementary Figure 2). To boost power of detecting more rare variants with moderate effect size (OR <2), gene-wise burden tests should be adopted (Bansal *et al.*, 2010).

While it is true that RVs need not always be in CDSs and that RVs may include short insertion/deletions (Indels) and longer structural variations (copy number variations, translocations), further investigation on other types of rare variants in candidate genes and whole genomic regions (exome sequencing or whole genome sequencing)are necessary to better explain our HSCR patients not carrying rare damaging SNVs. Full genomic profiles would also give an opportunity to later risk prediction that integrate both common risk variants (*RET* and *NRG1*) and rare risk mutations. In addition, replication is always needed since the involvement of a mutated gene/chromosomal region in the disease may only be consolidated by the identification of recurrent mutations or genes in an independent group of patients(Veltman & Brunner, 2012, Gilissen *et al.*, 2012, Vissers *et al.*, 2010).

ACKNOWLEDGEMENTS

This work was supported by research grants from the Hong Kong Research Grants Council HKU 778610M to PT, from The University of Hong Kong Seed Funding Programme 200911159190 to SSC and from the Innovation and Technology Support Programme ITS/245/09FP to SL. Support was also obtained from The University of Hong Kong Genomics Strategic Research Theme.

REFERENCES

Amiel, J. & Lyonnet, S. (2001) Hirschsprung disease, associated syndromes, and genetics: a review. *J Med Genet,* 38**,** 729-39.

Amiel, J., Sproat-Emison, E., Garcia-Barcelo, M., Lantieri, F., Burzynski, G., Borrego, S., Pelet, A., Arnold, S., Miao, X., Griseri, P., Brooks, A.S., Antinolo, G., De Pontual, L., Clement-Ziza, M., Munnich, A., Kashuk, C., West, K., Wong, K.K., Lyonnet, S., Chakravarti, A., Tam, P.K., Ceccherini, I., Hofstra, R.M., Fernandez, R. & Hirschsprung Disease, C. (2008) Hirschsprung disease, associated syndromes and genetics: a review. *J Med Genet,* 45**,** 1-14.

Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. (2010) Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet,* 11**,** 773-85.

Cummings, N., King, R., Rickers, A., Kaspi, A., Lunke, S., Haviv, I. & Jowett, J.B. (2010) Combining target enrichment with barcode multiplexing for high throughput SNP discovery. *BMC Genomics,* 11**,** 641.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., Mckenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat*

Genet, 43, 491-8.

Doray, B., Salomon, R., Amiel, J., Pelet, A., Touraine, R., Billaud, M., Attie, T., Bachy, B., Munnich, A. & Lyonnet, S. (1998) Mutation of the RET ligand, neurturin, supports multigenic inheritance in Hirschsprung disease. *HumMolGenet,* 7, 1449-52.

Emison, E.S., Mccallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., Cutler, D.J., Green, E.D. & Chakravarti, A. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature,* 434, 857-63.

Garcia-Barcelo, M.M., Tang, C.S., Ngan, E.S., Lui, V.C., Chen, Y., So, M.T., Leon, T.Y., Miao, X.P., Shum, C.K., Liu, F.Q., Yeung, M.Y., Yuan, Z.W., Guo, W.H., Liu, L., Sun, X.B., Huang, L.M., Tou, J.F., Song, Y.Q., Chan, D., Cheung, K.M., Wong, K.K., Cherny, S.S., Sham, P.C. & Tam, P.K. (2009) Genome-wide association study identifies NRG1 as a susceptibility locus for Hirschsprung's disease. *Proc Natl Acad Sci U S A,* 106, 2694-9.

Gilissen, C., Hoischen, A., Brunner, H.G. & Veltman, J.A. (2012) Disease gene identification strategies for exome sequencing. *Eur J Hum Genet,* 20, 490-7.

Griseri, P., Vos, Y., Giorda, R., Gimelli, S., Beri, S., Santamaria, G., Mognato, G., Hofstra, R.M., Gimelli, G. & Ceccherini, I. (2009) Complex pathogenesis of Hirschsprung's disease in a patient with hydrocephalus, vesico-ureteral reflux and a balanced translocation t(3;17)(p12;q11). *EurJ HumGenet: EJHG,* 17, 483-90.

Gui, H., Tang, W.K., So, M.T., Proitsi, P., Sham, P.C., Tam, P.K., Sau-Wai Ngan, E., Cherny, S.S. & Garcia-Barcelo, M.M. (2013) RET and NRG1 interplay in Hirschsprung disease. *HumGenet,* 132, 591-600.

Heanue, T.A. & Pachnis, V. (2006) Expression profiling the developing mammalian enteric nervous system identifies marker and candidate Hirschsprung disease genes. *Proc Natl Acad Sci U S A,* 103, 6919-24.

Heanue, T.A. & Pachnis, V. (2007) Enteric nervous system development and Hirschsprung's disease: advances in genetic and stem cell studies. *Nat Rev Neurosci,* 8, 466-79.

Hofstra, R.M., Valdenaire, O., Arch, E., Osinga, J., Kroes, H., Loffler, B.M., Hamosh, A., Meijers, C. & Buys, C.H. (1999) A loss-of-function mutation in the endothelin-converting enzyme 1 (ECE-1) associated with Hirschsprung disease, cardiac defects, and autonomic dysfunction. *Am J Hum Genet,* 64, 304-8.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., Mclaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., Hultman, C.M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y.Y., Price, A.L., De Bakker, P.I., Purcell, S.M. & Sunyaev, S.R. (2012) Exome sequencing and the genetic basis of complex traits. *Nat Genet,* 44**,** 623-30.

Li, H. & Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *BriefBioinform,* 11**,** 473-83.

Li, M.X., Gui, H.S., Kwan, J.S., Bao, S.Y. & Sham, P.C. (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res,* 40**,** e53.

Mertes, F., Elsharawy, A., Sauer, S., Van Helvoort, J.M., Van Der Zaag, P.J., Franke, A., Nilsson, M., Lehrach, H. & Brookes, A.J. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics,* 10**,** 374-86.

Mondal, K., Shetty, A.C., Patel, V., Cutler, D.J. & Zwick, M.E. (2011) Targeted sequencing of the human X chromosome exome. *Genomics,* 98**,** 260-5.

Rehm, H.L. (2013) Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet,* 14**,** 295-300.

Rothberg, J.M. & Leamon, J.H. (2008) The development and impact of 454 sequencing. *Nat Biotechnol,* 26**,** 1117-24.

Sham, P., Bader, J.S., Craig, I., O'donovan, M. & Owen, M. (2002) DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet,* 3**,** 862-71.

Shendure, J. & Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

Summerer, D., Wu, H., Haase, B., Cheng, Y., Schracke, N., Stahler, C.F., Chee, M.S., Stahler, P.F. & Beier, M. (2009) Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res,* 19**,** 1616-21.

Torfs, C. (2004) An epidemiological study of Hirschsprung' disease in a multiracial California population. In: *Third International Meeting: Hirschsprung's disease and related neurocristophaties*) *Third International Meeting: Hirschsprung's disease and related neurocristophaties.* Rvian, France.

Turner, K.N., Schachner, M. & Anderson, R.B. (2009) Cell adhesion molecule L1 affects the rate of differentiation of enteric neurons in the developing gut. *DevDyn,* 238**,** 708-15.

Veltman, J.A. & Brunner, H.G. (2012) De novo mutations in human genetic disease. *Nature reviews. Genetics,* 13**,** 565-75.

Vissers, L.E., De Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., De Vries, P., Van Lier, B., Arts, P., Wieskamp, N., Del Rosario, M., Van Bon, B.W., Hoischen, A., De Vries, B.B., Brunner, H.G. & Veltman, J.A. (2010) A de novo paradigm for mental retardation. *Nat Genet,* 42**,** 1109-12.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet,* 89**,** 82-93.

Zhao, Q., Peng, L., Huang, W., Li, Q., Pei, Y., Yuan, P., Zheng, L., Zhang, Y., Deng, J., Zhong, C., Hu, B., Ding, H., Fang, W., Li, R., Liao, Q., Lin, C., Deng, W., Yan, H., Hou, J., Wu, Q., Xu, T., Liu, J., Hu, L., Peng, T., Chen, S., Lai, K.N., Yuen, M.F., Wang, Y., Maini, M.K., Li, C., Li, M., Wang, J., Zhang, X., Sham, P.C., Wang, J., Gao, Z.L. & Wang, Y. (2012) Rare inborn errors associated with chronic hepatitis B virus infection. *Hepatology,* 56**,** 1661-70.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Table S1 Genes encoding members of pathways involved in the ENS development

Figure S1 Target sequence enrichment by Raindance technology on Hirschsprung disease

Figure S2 Statistical power calculations for rare variants

TABLES

Table 1: Genes reported to cause HSCR when mutated

Genes	Type of gene	Chr	Length[1]	Exons	Transcript[1]	Human[+]	Mouse[∫]	Frequency[*]
RET	Cell surface receptor	10q11	53,283	21	5,611	HSCR	Absence EN	50% F/7-35%S
GDNF	Neurotrophic factor	5p13	27,004	3	3,752	HSCR	Absence EN	Very rare
GFRA1	Cell surface receptor	10q25	210,024	10	1,539	HSCR	Absence EN	1 patient
NRTN	Neuronal growth factor	19p13	4,518	2	1,103	HSCR	Abnormal ENS	Very rare
PHOX2B	Transcription factor	4p13	4,888	3	3,029	Haddad Syndrome[0]	Abnormal ENS	-
NKX2.1	Transcription factor	14q13	3,815	3	2,165	HSCR	ND	Very rare
SOX10	Transcription factor	22q13	15,123	4	2,861	WS4[0]	Absence EN	-
NRG1	Signaling protein	8p12	1,124,806	17	3,000	HSCR	Abnormal NCC migration	-
EDNRB	Transmembrane receptor	13q22	80,049	8	4,310	HSCR/WS4[0]	Absence EN	3-7%
EDN3	Vasoactive peptide	20q13	25,566	6	2,397	WS4[0]	Absence EN	<5%

ECE-1	Enzyme	1p36	126,921	19	2,764	HSCR	Absence EN	Very rare
KIAA1279	Binding protein	10q21	28,246	7	2,524	GSM syndrome[0]	ND	-
ZFXH1B	Binding protein	2q22	132,334	10	5,558	MW syndrome[0]	Abnormal NCC migration	-
L1CAM	Cell adhension molecules	Xq28	24,630	29	5,113	Hydrocephalus/HSCR[#]	Delayed NCC differentiation[#]	-
TOTAL				142	45,726			

1: in base pairs**.** **EN:** entericneurons.**ENS:** enteric nervous system. **NCC:** neural crest cell. **+:**when not specified, isolated (non-syndromic) HSCR. **∫:** Mouse Genome Informatics search tool; phenotype refers only to gut/ENS system. *****: % of isolated HSCR patients with coding region mutations in these genes, data from (Heanue & Pachnis, 2007, Amiel & Lyonnet, 2001, Doray *et al.*, 1998). **-**: no patients with isolated HSCR reported. #: Referred from Human disease model (Griseri *et al.*, 2009) or mouse model (Turner *et al.*, 2009). **F=** familial. **S=** sporadic. **[0]:** HSCR is part of the syndrome.**WS4:** Waardenburg-Shah type 4 syndrome. **GSM:** Goldberg-ShprintzenMegacolon. **MW:**Mowat-Wilson. **ND:** no abnormal gut phenotype described.

Table 2 Quality summary of sequencing and SNV calling per sample pool

Sample Group[*]	Unique Reads on Targets[#]	Target specificity (%) [$]	Mean Coverage	Target Regions with depth ≥ 40 (%)	Raw SNVs	CDS CNVs	dbSNP137 coverage (%)
HS1	235,854	68.58	71	67.17	1015	131	88.55
HS2	320,415	74.63	99	81.93	1407	181	82.32
HS3	276,321	76.00	90	78.75	1200	151	90.07
HS4	263,895	74.12	85	72.35	1121	153	92.16
HS5	276,517	72.59	90	72.87	1131	153	90.2
HS6	296,667	69.76	91	72.13	1140	162	85.19
HS7	229,107	77.81	61	54.21	811	109	95.41
HS8	271,618	79.93	76	93.53	910	131	92.37
Mean	271,299	74.18	83	74.12	1092	149	89.53

*, Each sample pool (HS1-4 for cases only, HS5-8 for controls only) contains fiveindependent samples.$, target specificity was measured as unique reads on targets over total unique reads

Table 3 Quality summary for individual sequencing of Pool-3 samples

SampleID	Targeted Read Count	Average Read depth	RawSNVs*	CDS SNVs*	dbSNP137coverage (%)	GWAS concordance (%)#
HK109C	10,455,485	304.1×	876	93	92.39	95.40
HK113C	12,892,678	379.3×	967	96	95.83	96.60
HK119C	14,576,191	430.6×	958	92	93.41	97.70
HK128C	13,142,489	386.7×	970	93	96.77	95.40
HK138C	12,671,332	372.1×	941	112	94.64	93.00
Mean	13,653,900	374.6×	942	97	94.61	95.60

*, there were 1605 raw SNVs (including 175 CDS SNVs) that are unique from all fiveindividuals. #, 87 SNVs (covered by both GWAS arrays and targeted resequencing) were used to estimate the genotype concordance.Genotype concordance was calculated as the total number of matching genotypes out of all valid comparisons for which the same variants (position and reference/alternative allele) were genotyped bybothGWAS arrays and resequencing.

Table 4 Comparison of CDS SNVs detected by individual sequencing and pooled sequencing on Pool-3 patients

Technology	All	Common	SNP array	Rare	Rare damaging (Sanger)
Individual sequencing	175	144	13	31	8 (7)
Pooled sequencing	151	122	11	29	10 (6)
Overlapped	139	120	11$	19	6 (6)

[$] individual genotype was not available, concordant variants were roughly estimated by comparing the consistency of major allele and minor allele.

Table 5 Summary of variants from all eight pools forHSCR cases and healthy controls

Variant set	Total	Case	Control	*P*-value (burden test)
All variant	366	284	246	NA
Rare variants (RVs)	166	99	74	0.265 [1]
Rare damaging variants (RDVs)	35	16	19	0.131 [2]
RDVs present in Case-unique genes (Sanger validated)	12 (5)	12 (5)	NA	NA

[1] burden test for overall count of rare variants among all variant; [2] burden test for overall count of RDVs among RVs; NA means not available