SNP discovery performance of two second generation sequencing platforms in the NOD2 gene region

Espen Melum, Sandra May, Markus Schilhabel, Ingo Thomsen, Tom H. Karlsen, Philip Rosenstiel, Stefan Schreiber, Andre Franke

Human Mutation

SNP discovery performance of two second generation sequencing platforms in the *NOD2* gene region

Journal:	*Human Mutation*
Manuscript ID:	humu-2009-0537.R2
Wiley - Manuscript type:	Methods
Date Submitted by the Author:	16-Apr-2010
Complete List of Authors:	Melum, Espen; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology; Rikshospitalet, Oslo University Hospital, Medical Department May, Sandra; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology Schilhabel, Markus; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology Thomsen, Ingo; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology Karlsen, Tom; Rikshospitalet, Oslo University Hospital, Medical Department Rosenstiel, Philip; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology Schreiber, Stefan; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology; University Clinic S.-H. (Campus Kiel), 1st Medical Department Franke, Andre; Christian-Albrechts-University Kiel, Institute of Clinical Molecular Biology
Key Words:	Second generation sequencing, SOLiD, 454/FLX, NOD2, rare variants, benchmarking, coverage simulation, SNP discovery, mutation detection

SNP discovery performance of two second generation sequencing platforms in the *NOD2* gene region

Short title: NOD2 second generation sequencing

Espen Melum[1,2], Sandra May[1], Markus B. Schilhabel[1], Ingo Thomsen[1], Tom H. Karlsen[2], Philip Rosenstiel[1], Stefan Schreiber[1,3] and Andre Franke[1,¶]

[1] Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany

[2] Norwegian PSC Research Center, Clinic for specialized Medicine and Surgery, Oslo University Hospital Rikshospitalet, Oslo, Norway

[3] 1st Medical Department, University Clinic S.-H. (Campus Kiel), Kiel, Germany

¶ Corresponding author:

Andre Franke

Institute of Clinical Molecular Biology

Christian-Albrechts University Kiel

Schittenhelmstr. 12

D-24105 Kiel

Germany

e-mail: a.franke@mucosa.de

Tel.: +49 (0) 431-597 4138

Fax: +49 (0) 431-597 2196

Keywords: Second generation sequencing, SOLiD, 454/FLX, NOD2, rare variants, benchmarking, coverage simulation, SNP discovery, mutation detection

Abbreviations

SNP – Single Nucleotide Polymorphism

GWAS – Genome-Wide Association Studies

CD – Crohn's Disease

NOD2 – Nucleotide-binding Oligomerization Domain containing 2

PCR – Polymerase Chain Reaction

LR-PCR – Long-Range PCR

Abstract

A potentially important application of second generation sequencing technologies is to identify disease-associated variation. For comparison of the performance in SNP detection, the Crohn's disease (CD) associated *NOD2* gene was subjected to targeted resequencing using two different second generation sequencing technologies. Eleven CD patients were selected based on their haplotype background at the *NOD2* locus. The 40 kb large *NOD2* gene region was amplified using long-range PCR (LR-PCR), and sequenced with the Roche 454/FLX system, an Applied Biosystems SOLiD mate-pair library (2x25 bp), and a SOLiD fragment (50 bp) library. The entire *NOD2* region was also sequenced using conventional Sanger technology. Four-hundred and forty-two SNPs were discovered with the SOLiD mate-pair library, 454 with the fragment library and 441 with the 454/FLX. For the homozygous SNPs, 98% were confirmed by Sanger for the mate-pair library, 100% for the fragment library and 99% for the 454/FLX. Ninety-six percent of the heterozygous SNPs detected with the SOLiD mate-pair library, 91% with the fragment library and 96% with the 454/FLX were confirmed by Sanger. In a simulation, the SNP detection performance fell rapidly when the achieved coverage was below 40x. Due to uneven representation of the target region when using LR-PCR, oversequencing of other regions is necessary.

Introduction

During the last years, second generation sequencing technologies have made it possible to sequence genetic regions and complete genomes in a time-efficient manner with a low per-base cost (Schuster, 2008). Second generation sequencing has been applied to *de novo* sequencing of bacterial genomes (Margulies et al., 2005), resequencing of entire human genomes (Bentley et al., 2008; Wang et al., 2008; Wheeler et al., 2008), and exome sequencing (Choi et al., 2009; Ng et al., 2009). An equally important application of this technology is targeted resequencing of (entire) known susceptibility genes or loci of interest, as has been reported for Type I diabetes (Nejentsev et al., 2009). The number of well-established disease genes in complex diseases is currently growing at an unprecedented speed, after several hundred genome-wide association studies (GWAS) have been performed for different diseases and other human phenotypes (Hindorff et al., 2009). Crohn's disease (OMIM 266600, CD), besides ulcerative colitis, is one of the two major inflammatory bowel disease phenotypes (Loftus, 2004). In CD, the GWAS method has been especially successful, and to this end, more than 40 new disease genes have been identified (Franke et al., 2007; Parkes et al., 2007; Raelson et al., 2007; Barrett et al., 2008). CD associated variants in the nucleotide-binding oligomerization domain containing 2 (*NOD2)* gene (OMIM 605956) were well-known before GWAS (Hugot et al., 2001; Ogura et al., 2001), and have been widely replicated (Schreiber et al., 2005).

The SOLiD platform from Applied Biosystems (Foster City, CA, USA) follows the sequencing by ligation approach while the 454/FLX platform from Roche / 454 Life Sciences (Branford, CT, USA) uses the sequencing by synthesis principle (Shendure and Ji, 2008). The new sequencing methods contrast sequencing with end-termination chemistry which has been

the standard method for sequencing since the original publication by Sanger *et al.* in 1977 (Sanger et al., 1977). The method used in the 454/FLX system was first published in 2005, in a study reporting on shotgun *de novo* sequencing of *Mycoplasma genitalium* (Margulies et al., 2005), and has been further developed with increasing read lengths, throughput and accuracy (Droege and Hill, 2008). The SOLiD system was released commercially in 2007 by Applied Biosystems. Since then, the sequencing protocols and the analytical tools have been substantially improved and extended to approach new scientific questions. The 454/FLX and SOLiD system share several characteristics. With both systems short DNA templates are attached to a surface of beads, and on these beads an emulsion based polymerase chain reaction (PCR) takes place (Shendure et al., 2005). The beads, now containing multiple copies of the DNA template, are transferred to a glass plate or into micropores for SOLiD or 454/FLX, respectively. After this step there are substantial differences between the platforms (Metzker, 2010). For the down-stream analyses, the two most important features are the longer read lengths for the 454/FLX (Droege and Hill, 2008) and the inherent error-checking properties of the two-base pair encoding of the SOLiD system (McKernan et al., 2006). The SOLiD system also delivers significantly more sequence output at the cost of shorter reads compared to the 454/FLX system.

In this study, we wanted to explore the utility of these two technologies in targeted sequencing with a focus on single nucleotide polymorphism (SNP) discovery, the most abundant form of variation. For this, we aimed to resequence the entire coding and non-coding regions of the most widely replicated CD-susceptibility gene *NOD2* in 11 CD patients. To benchmark this data further, the identified variants were compared to Sanger sequencing data that were generated in parallel for the 40 kb large *NOD2* genomic region.

Several methods for enrichment of a target region for sequencing exist, including long-range PCR (LR-PCR), array-based enrichment (Summerer et al., 2009) and in-solution enrichment (Gnirke et al., 2009). The latter two exploit hybridization to complementary nucleotide stretches to capture the target regions of interest. For enrichment of the 40kb target region in this study, LR-PCR, which is applicable in any lab, was used. LR-PCR is not specifically optimized for any of the sequencing platforms and produces the same input material to both platforms.

The challenges in the present project will largely parallel the challenges that the genetic community faces when the new genes and regions identified through GWAS will be subjected to deep resequencing (Yeager et al., 2008).

Material and methods

Selection of samples

The precise haplotype structure of the *NOD2* locus has been described before (Croucher et al., 2003). Using the data generated by Croucher *et al.* and the same haplotype nomenclature, we selected 11 patients based on the haplotypes generated by the phasing software PHASE 2.1.1 (Stephens et al., 2001) (Supplementary Table S1). These patients were sequenced following the standard protocols for SOLiD mate-pair library, SOLiD fragment library and the 454/FLX platform.

All CD patients were recruited through the 1st Medical Department of General Internal Medicine of the University Clinic Schleswig-Holstein (UK-SH, Campus Kiel, Germany). Clinical, radiological and endoscopic (i.e. type and distribution of lesions) examinations were required to unequivocally confirm the diagnosis of CD, and histological findings had to be confirmative of, or compatible with, the diagnosis. All recruitment protocols were approved by the ethics committee of the UK-SH and participants gave their written, informed consent.

Genomic DNA was extracted from blood using the Invitek kit (Invitek, Berlin, Germany) and each DNA sample was evaluated by gel electrophoresis for the presence of high-molecular weight DNA and normalized to 50-60 ng/µl using Picogreen fluorescent dye (Invitrogen, Carlsbad, CA, USA).

Primer design and optimization for long-range PCR

LR-PCR primers (Table 1) were designed using the Primer3 software (available online at http://frodo.wi.mit.edu/) (Rozen and Skaletsky, 2000), and the HPLC-purified primers were

ordered from Metabion (Martinsried, Germany). Amplicon sizes varied from 3 kb to 11 kb, with an average size of 7 kb (Figure 1). Adjacent amplicons overlapped with >70 bases to allow for SNP detection underneath the primer binding sites. The 50 µl PCR reactions were performed using 50 ng of human genomic DNA, 1.5 µl (10 µM) of the forward and the reverse primer, 5 µl GeneAmp High Fidelity 10x PCR Buffer (Applied Biosystems), and 2.5 units of GeneAmp High Fidelity Enzyme Mix (Applied Biosystems). Primer optimization was performed for each amplicon on a Biometra Gradient cycler (Biometra, Goettingen, Germany) using the following gradient program: one cycle at 94°C for 2 min, followed by 35 cycles 94°C for 15 sec (initial denaturation), at gradient of 45-65°C for 30 sec (annealing), at 68°C for 9 min (extension), followed by a final extension step at 72°C for 7 min.

The PCR products were cleaned up using the Biorobot 8000 (Qiagen, Hilden, Germany) and the Qiaquick purification kit (Qiagen). The amplicons were subsequently quantified using Picogreen and then pooled at an equimolar ratio based on the following formula:

$$\text{amount of PCR product [ng]} = \frac{\text{length of the fragment [bp]} \times \text{amount of input DNA for library [ng]}}{\text{length of target [bp]}}$$

Second Generation Sequencing

Three types of libraries were prepared for each sample: 1) SOLiD mate-paired library, 2) SOLiD fragment library and 3) Roche/454 GS-FLX sequencing library.

SOLiD mate-pair run

A mate-paired library consists of two pieces of target DNA. These pieces of sequence orignates from the two ends of the same DNA fragment. SOLiD mate-pair libraries were constructed with an insert size of approximately 600 bp using 2 µg of DNA from the equimolar pooled amplicons. The DNA was sheared to generate fragments with a size of 500-650 bp, end-repaired and methylated. Methylation of EcoP15I sites in the target DNA prevented digestion by the methylation-sensitive EcoP15I enzyme. Then, EcoP15I CAP ligation added the EcoP15I CAP adaptors to the sheared and methylated DNA. After the circularization using DNA T4 ligase, EcoP15I cleaved 25-27 bp away from the unmethylated enzyme recognition sites in the CAP adaptors. Following the ECOP15I digestion another end repair of the DNA fragments was carried out and P1 and P2 adaptors were ligated to the ends of the fragments before PCR amplification. Clonal bead populations were prepared in microreactors by a water-in-oil emulsion PCR. The microreactors contained DNA template, PCR reagents, beads and primers. After the PCR reaction, the emulsion was broken with 2-butanol and enrichment of the beads containing a library template was performed by hybridization to polystyrene beads containing the complementary sequence of the P2 primer. A 3' modification of the template on the selected beads allowed for covalent binding to a glass slide. On the SOLiD Analyzer the templated beads were combined with the universal sequencing primer, ligase and a large pool of dibase probes. The dibase probes are fluorescently labeled with four dyes, each dye represents four of the sixteen possible dye nucleotide sequences. Complementary probes hybridized to the template sequence and were ligated during the sequencing process. After measurement of the fluorescence the dye was cleaved of. This process was repeated for five cycles. The synthesized strand was thereafter removed and a new primer was hybridized offset by one base and the ligation cycles were

repeated. This primer reset process was repeated for five cycles providing a dual measurement for each base. Each library was separately sequenced in one of the eight spots of an octant slide.

SOLiD fragment run

Short fragment DNA libraries were generated using 2 µg of DNA from the equimolar pooled amplicons and sequenced at Applied Biosystems in Beverly, MA (USA). LR-PCR fragments were sheared by sonication to a size of 150-200 bp. After end-repairing and P1 and P2 adaptor ligation to the resulting fragments, library-amplification was performed. The resulting library contained molecules representing the entire target sequence. Each molecule was clonally amplified on beads in an emulsion PCR. In the emulsion PCR clonal bead populations were generated in water-in-oil microreactors containing template, PCR reagents, primers and beads with sequence complementary to P1 adaptor on the surface. After the library molecules annealed to the beads, the sequence from the P1 adaptor was extended by a polymerase followed by template dissociation. The emulsion was thereafter broken with 2-butanol. Enrichment for beads containing a library template was performed by hybridization to polystyrene beads containing the complementary sequence of the P2 primer, and a 3'-end modification followed to allow attachment of the beads to the glass surface of the sequencing slide. The sequencing was based on the sequential ligation with dye labeled oligonucleotides as described for the SOLiD mate-paired libraries. For the 50bp fragment run each of the 5 primer cycles included 10 ligations of the dye labeled oligonucleotides. Each library was run on an individual spot on a quadrant slide.

Roche GS-FLX

FLX DNA libraries were generated at Roche Diagnostics in Penzberg (Germany). For each of the 11 samples consisting of equimolar pooled PCR products, 3 µg of DNA was used for the library preparation step, following the standard protocol for whole genome libraries. The pooled PCR fragments were nebulized to 400-800 nucleotide fragments. To get blunted ends, end-repair was performed. Then, short adaptors with 10 base identifier tags were ligated to both ends (adaptor A and B), which were used for purification, amplification and sequencing. The input to the emulsion PCR were quantitated library molecules flanked with proper adaptors for amplification and sequencing. According to the manufacturer the library fragments were mixed with beads coated by oligonucleotides complementary to the sequence of adaptor B. The amplification reaction was performed in a water-in-oil emulsion. The forward PCR primer was biotinylated for use later during the enrichment step. The reaction within the droplets of an emulsion resulted in beads carrying clonally amplified DNA fragments. After the emulsion PCR step, the emulsion was broken chemically and the beads were recovered and washed by filtration. During the emulsion PCR a certain fraction of beads that carry no amplified DNA (null beads) were generated. To reduce the percentage of null beads an enrichment step for template beads was performed. Streptavidin-coated magnetic enrichment beads were added to capture the successfully amplified fragments binding only to those which have incorporated the biotinylated primer. The template beads were finally annealed with sequencing primers and mixed with polymerase and cofactors. Then, they were loaded together with "enzyme beads" (sulfurylase and luciferase) and "packing beads" in a PicoTiterplate device and sequenced using a GS-FLX pyrosequencer (Roche/454 Life Science). During the sequencing process nucleotides were flowed sequentially across the

PicoTiterPlate device. A light signal was generated if the nucleotide was complementary to the template and thus incorporated. The signal intensity at each nucleotide flow indicated the number of nucleotides. If no nucleotide was incorporated, no light signal was generated and the complementary strand was not elongated. A two area gasket was used during the run for maximized output.

Sanger Sequencing

Ninety-eight PCR amplicons (size range 83-912 bp, average size 529 bp) were constructed for the same *NOD2* target region as described above (see Figure 1 for amplicon overview and Supplementary Table S2 for primer sequences). The corresponding PCRs were performed using AmpliTaq Gold Fast PCR Mastermix (Applied Biosystem), primers and 2.5 ng genomic DNA. For all fragments, cycling was performed with an initial denaturation step of 95°C for 10 min to activate the polymerase, followed by 35 cycles of 96°C for 6 sec, 62°C for 6 sec, 68°C for 8 sec, and a final step of 72°C for 10 sec. After PCR, the resulting amplicon was used as a target for sequencing with the BigDye Terminator v1.1 Cycle Sequencing Kit, and analyzed using an Applied Biosystems 3730*xl* Genetic Analyzer.

Matching against the reference sequence and SNP-calling

The reference sequence for the region was acquired from the human reference sequence hg18 assembly (NCBI's build 36) using the UCSC Genome Browser (http://genome.ucsc.edu/).

The reads from the SOLiD mate-pair runs were matched separately against the reference with the freely available Corona Lite version 4.2.2 software (Applied Biosystems) (http://solidsoftwaretools.com/gf/project/corona/) allowing two mismatches in color-space (2 per 25 bp) and counting valid adjacent errors as one. After matching the reads against the reference, they were paired in a test procedure allowing mate-pair distances from 0 to 60 kb. The actual ranges of mate-pair distances were determined manually for each individual sample using a histogram of the distances observed. The mate-pair distances observed for all samples were in the range of 260 – 690 bp. Pairing of the runs was performed again with these estimated mate-pair distances, and the results of this pairing formed the basis for SNP-calling.

The fragment runs were matched against the reference genome using Corona Lite version 4.2.2, allowing four mismatches in color-space (4 per 50 bp) and counting valid adjacent errors as one (i.e. allowing on average the same number of mismatches per base sequenced as in the mate-pair procedure). The matched fragment runs were used for SNP-calling.

SNP-calling for the SOLiD data was performed using with Corona Lite version 4.2.2 with default settings for both the mate-pair library and the fragment library.

The 454/FLX runs were mapped against the reference sequence and evaluated for SNPs with the Genome Sequencer FLX Software Package version 2.0.01.12 software from Roche. All of the matched sequences were evaluated for SNPs, and high confidence SNPs reported by the software were considered further. To avoid bias from areas with high coverage, additional

filtering was applied with the following criteria: base pair positions where between 30% and 70% of the reads showed alternative bases were called as heterozygous SNPs, while the base-pair positions where more than 70% of the reads differed from the reference were called as homozygous SNPs. As the objective of the present paper was to evaluate the performance in SNP detection, insertions and deletions were not evaluated.

All SNP positions were looked up in the NCBI's database of SNPs (dbSNP 130) (Sherry et al., 2001) and assigned the respective dbSNP ID ("rs-number") if the SNP had been annotated before. The novel SNPs that were confirmed by Sanger were submitted to dbSNP and their temporary accession numbers were used ("ss-numbers"). The remaining SNPs were assigned the names NOD2_01 to NOD2_12 for this publication. The exact genomic positions of all identified SNPs are shown in Supplementary Table S3.

Sanger SNP detection

Semi-automated SNP detection was carried out using the software novoSNP (http://www.molgen.ua.ac.be/bioinfo/novosnp/) (Weckx et al., 2005). All base pair positions where a SNP was reported by the second generation technologies were manually inspected.

Coverage

The achieved coverage was calculated from the number of reads covering a specific base pair position. For the SOLiD platform, the coverage was calculated based on the reads that mapped uniquely (non-unique matches were excluded) to the reference sequence, and paired

with a mate for the mate-pair library, while the coverage for the 454/FLX platform was calculated based on the reads that aligned to the reference sequence. These coverage figures represent the reads subsequently used for SNP-calling. Coverage depths for individual LR-PCR fragments were calculated using the R statistical package v2.91 (http://www.r-project.org/) and custom scripts. To compare the coverage between the different technologies, the coverage at all base-pair positions were averaged for all the samples and calculated in 100 bp bins.

Coverage simulation

To compare the performance at lower coverages than the maximum achieved, an *in silico* experiment with random and sequential removal of reads, aiming for an achieved coverage of 100x, 80x, 60x, 40x, 20x, 10x and 5x, was performed. The random removal of reads was performed irrespective of where in the sequence the read mapped. The benchmarking reference was defined as the genotype call made for the respective technology in the full data set (maximum achieved coverage). In samples where the full achieved coverage was below the aimed coverage in the simulation, the full achieved coverage was used for that simulation step.

Results

Sequence content and coverage

There were substantial differences in the coverage achieved between the different samples (Table 1). The SOLiD mate-pair and fragment runs had fixed read lengths of respectively 2x25 bp and 50 bp, while for the 454/FLX technology the read lengths differed with averages from 217 bp to 227 bp in the different samples. For the 454/FLX runs 90% of the reads were above 133 bp to 177 bp for the different samples. The overall GC content in the sequenced region was 48% (variation across the region can be seen from Figure 2). Forty-three percent of the sequence consisted of repetitive elements. The coverage varied corresponding to the LR-PCR amplicons for the SOLiD mate-pair and SOLiD fragment library, while for the 454/FLX this was not so pronounced (Figure 2). The per base Pearson's correlation coefficients between the different samples showed a good correlation within each technology, and were on average 0.81 (range 0.32-0.98) for SOLiD mate-pair, 0.88 (range 0.67-0.98) for SOLiD fragment and 0.78 (range 0.63-0.89) for the 454/FLX. Only four samples sequenced with the mate-pair library had more than 10 base pair positions not covered, that was for sample 6, 7, 8 and 9 where respectively 0.21%, 0.11%, 0.26% and 0.03% of the base pair positions were not covered. For the fragment library and the 454/FLX none of the samples had more than 10 base pair positions not covered.

SNP detection and Sanger confirmation

Summarizing all samples, there were a total of 442 SNPs discovered with the SOLiD mate-pair library, 454 with the SOLiD fragment library and 441 with the 454/FLX (Figure 3). The numbers of homozygous SNPs were 146, 141 and 147 for the mate-pair, fragment and

454/FLX respectively. Ninety-eight percent of the homozygous SNPs detected with the mate-pair library, 100% detected with fragment library and 99% detected with the 454/FLX were confirmed by Sanger sequencing. Out of the five homozygous SNPs (Mate-pair: 3 SNPs, Fragment: 0 SNPs, 454/FLX: 2 SNPs) not confirmed by Sanger sequencing, four were detected as heterozygous with Sanger, while one SNP identified with the mate-pair library was not detected at all. Calling of this SNP was based on 7 short sequence reads with distinct start points agreeing with the variant, while on average 70 sequence reads with distinct start points (counting positions on both of the mates and both strands) agreed with the variant for the Sanger confirmed mate-pair variants. Of the homozygous variants detected by Sanger sequencing 97%, 96% and 99% were detected with the SOLiD mate-pair, SOLiD fragment and 454/FLX, respectively.

For the heterozygous SNPs, the confirmation rates by Sanger were 96% for the SOLiD mate-pair library (10 SNPs not detected with Sanger and 1 SNP homozygous with Sanger), 91% for the SOLiD fragment library (28 SNPs not detected with Sanger and 1 SNP homozygous with Sanger) and 96% for the 454/FLX (10 SNPs not detected with Sanger and 1 SNP homozygous with Sanger). For the 10 heterozygous SNPs not detected with the mate-pair library an average of 17 short sequence reads with distinct start points agreed with the variant compared to 54 for the Sanger confirmed variants. The 28 heterozygous SNPs not detected with the fragment library had on average 30 short sequence reads with distinct start points agreeing with the variant compared to 89 for the Sanger confirmed variants (counting positions on both strands). The 10 heterozygous SNPs not detected with 454/FLX were on average seen in 41 reads compared to 40 reads for the Sanger confirmed variants. For the Sanger detected heterozygous SNPs, 99%, 98% and 98% were detected with the SOLiD

mate-pair, SOLiD fragment and 454/FLX, respectively. The NOD2_04 SNP (see Supplementary Table S3 for IDs and other characteristics of identified SNPs) was only seen with the SOLiD mate-pair library, and not confirmed by Sanger, while ten SNPs (NOD2_01, NOD2_02, NOD2_03, NOD2_05, NOD2_06, NOD2_07, NOD2_08, NOD2_09, NOD2_10 and NOD2_11) were only seen with the SOLiD fragment library and not confirmed by Sanger. The NOD2_12 SNP was only detected with the 454/FLX and not confirmed by Sanger sequencing, in dbSNP a deletion is reported at this position (rs5816717) while in the pre-release from the 1000genomes project (http://www.1000genomes.org) a SNP with the same call as we observed was reported.

Singleton and novel SNPs

In total 19 SNPs not reported in dbSNP b130 (Sherry et al., 2001) were identified. Seven of these were confirmed by Sanger sequencing, all were singleton heterozygous SNPs, only observed in one individual. The private SNP ss196000735 is a non-synonymous coding SNP in exon 4 (leading to an amino acid change from alanine to glycine) and was detected with all technologies. The rest of the Sanger confirmed private variants were non-coding SNPs. Thirty-six of the 67 detected and Sanger-validated SNPs had previously been genotyped by the HapMap consortium (http://www.hapmap.org), and were included in the HapMap data release 27 (genotype data for one of these 36 SNPs were not reported in the European population dataset). Of the reported SNPs, two SNPs were reported to be monomorphic (minor allele frequency of "0") while the rest had frequencies ranging from 3% to 48% in Utah residents with ancestry from northern and western Europe (see Supplementary Table S4). In the pre-release from the 1000genomes project (http://www.1000genomes.org), four of

the SNPs (ss196000734, NOD2_08, ss196000736 and NOD2_12) currently not listed in dbSNP b130 were identified.

Coverage simulation

The coverage achieved varied between the samples and technologies (see Table 1), and to explore the importance of coverage depth, a simulation of lower coverages was performed. This was done irrespective of the local coverage across the region. In this coverage simulation experiment, the SNP detection rate at a coverage of 20x dropped to 78%, 83% and 84% compared to full coverage for the mate-pair, fragment and 454/FLX respectively (Figure 4 and 5). The numbers of misclassified SNPs at coverages above 20x were low (Figure 4). During the simulations, SNPs that were not observed at full coverage started to appear for all technologies, this was most pronounced for the SOLiD fragment library, while only one such SNP was seen for the 454/FLX. All these SNPs are listed as false positives in Figure 4. As the present project included several techniques these SNPs were compared to the results from the other technologies. Six SNPs from the fragment library simulation, one SNP from the mate-pair library simulation and one SNP from the 454/FLX simulation turned out to represent true SNPs seen in the other technologies and confirmed by Sanger, while the rest of the SNPs appearing during the simulations were not seen in any of the other technologies applied, i.e. true false positives.

Discussion

Second generation sequencing has started to be the standard way of following up disease association findings from GWAS. These technologies have recently been used in several sequencing projects (Bentley et al., 2008; Wang et al., 2008; Wheeler et al., 2008; Ng et al., 2009; Choi et al., 2009). Comparsion of the different second generation technologies have previously been evaluated for mutational profiling in yeast (Smith et al., 2008) and a few human individuals (Harismendy et al., 2009). The present results report on a comparsion of resequencing performance of a target gene in individually selected patients. This sequencing scenario will be similar to what most researchers will face when they set out to resequence genetic regions discovered through GWAS.

For most of the samples included, we achieved saturating coverage in the target region, meaning that a further increase of the coverage beyond this point would not improve performance in terms of SNP detection. Importantly, the coverage figures used are the achieved coverage representing the reads that subsequently went into the SNP-calling pipelines of the SOLiD and 454/FLX systems, respectively. The coverage achieved is dependent upon the platform and the laboratory. Researchers planning a project need to take the platform and service provider efficacy into account when interpreting these numbers. Interestingly, from the coverage simulation, it seems that when using LR-PCR as the amplification method, a coverage depth of around 40x is necessary for all the technologies to achieve acceptable SNP detection rates. The uneven distribution of mappable reads implies that certain regions are oversequenced, a phenomenon which could preferably be reduced for cost reduction. With a more even representation of the different regions in the template, the average coverage needed for SNP detection may be lower. However, accurate equimolar

mixing of the templates cannot entirely overcome this problem. When aiming at a target coverage of for example 100x, a large overrepresentation of one amplicon will reduce the relative representation of other regions, and could therefore lead to an impaired total performance in the sample. The over-representation is especially prominent at the end of the fragments, a problem that could partly be reduced when using blocked primers (Harismendy and Frazer, 2009).

In the simulation experiment, some of the SNPs originally detected as heterozygous changed their status to homozygous, a process related to the stochastic nature of the random removal of reads. In studies aiming at the detection of SNPs for further follow-up, this is not of critical importance, as it is merely the SNP detection that is of importance, and not the true genotype. For studies aiming at identifying haplotypes, inheritance patterns or phenotype correlations, the coverage should be of such depth that the distinction between homozygous and heterozygous SNPs is unambiguous. In addition, we noted that some SNPs not detected at full coverage, started to appear at lower coverages and represented true SNPs confirmed by Sanger. More troublesome was the appearance of true false positives, and such SNPs were evident with the SOLiD mate-pair and fragment libraries.

The number of detected homozygous and heterozygous SNPs and the sensitivity were comparable for the three technologies. There were only a few homozygous SNPs not confirmed by Sanger sequencing and only one of them was not detected at all, while the remaining SNPs were determined as heterozygous. For the heterozygous SNPs the number of false positive SNPs was considerably higher for the SOLiD fragment library compared to the

SOLiD mate-pair library and the 454/FLX. These SNPs turned out to be supported by only a small number of short sequence reads with distinct start points and probably represent sequencing errors towards the end of the reads. Importantly, these SNPs are easy to sort out manually before further experiments are initiated. Thus, identified SNPs supported by a low number of unique starting points should be flagged as likely problematic ones.

As expected, given the small number of study subjects, most of the SNPs detected were already annotated in the public databases, and were common in the European population. *NOD2* is one of the most studied genes in human complex diseases, and interestingly, we discovered as many as seven novel SNPs that were confirmed by Sanger sequencing. Six of these SNPs were intronic SNPs, while one exonic non-synonymous SNP (ss196000735) was seen in one patient with all technologies. Neither this SNP, nor the other two exonic SNPs that were not confirmed by Sanger (NOD2_04 and NOD2_09), were reported in an extensive investigation of the coding part of the *NOD2* gene, a study in which 612 inflammatory bowel disease cases and 103 controls were resequenced (Lesage et al., 2002). This finding demonstrates that deep resequencing of even well-known disease genes in carefully selected individuals can yield novel rare or likely private variants of potential mechanistic importance not covered by GWAS (Bodmer and Bonilla, 2008).

In conclusion, we have demonstrated that second generation sequencing technologies are capable of resequencing target genes with reproducible results confirmed by Sanger sequencing. The uneven representation of the target region that we observed when using LR-PCR renders it necessary to oversequence to get adequate coverage in all parts of the target

region. Array-based enrichment (Summerer et al., 2009) and in-solution enrichment (Gnirke et al., 2009) could possibly solve this, however, more benchmarking experiments as herein described need to be performed for these enrichment methods as well.

Funding

This work was supported by the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN), the popgen biobank and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° [HEALTH-F4-2008-201418] entitled READNA. The project received infrastructure support through the DFG excellence cluster “Inflammation at Interfaces”. E Melum and TH Karlsen were funded by the Norwegian PSC research center.

Acknowledgments

We thank all individuals with IBD, their families and physicians for their cooperation. We acknowledge the cooperation of the German Crohn and Colitis Foundation (Deutsche Morbus Crohn und Colitis Vereinigung e.V.), the BMBF competence network ”IBD” and the contributing gastroenterologists. We wish to thank Lena Bossen, Anita Dietsch, Catharina von der Lancken, Ina Elena Baumgartner, Magda Depta, Melanie Friskovec, Susan Ehlers, Rainer Vogler and Kristian Holm for expert technical help. This study received infrastructure support from the DFG cluster of excellence “Inflammation at Interfaces”. The Research Computing Services group at USIT, University of Oslo is acknowledged for making the Titan computing cluster available. We acknowledge Clarence Lee, Swati S. Ranade, and Fiona C. Hyland from Applied Biosystems for providing and supporting the SOLiD fragments run. We also thank Volker Strack and the support of the Operation groups of 454 Life Science for performing the the 454/FLX sequencing run.

Reference List

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955-962.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira CR, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi, Chang S, Neil CR, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott FW, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling NB, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris PD, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva RA, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna SJ, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695-701.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106:19096-19101.

Croucher PJP, Mascheretti S, Hampe J, Huse K, Frenzel H, Stoll M, Lu T, Nikolaus S, Yang SK, Krawczak M, Kim WH, Schreiber S. 2003. Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. Eur J Hum Genet 11:6-16.

Droege M, Hill B. 2008. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol 136:3-10.

Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Hasler R, Little RD, Huse K, Ruether A, Balschun T, Wittig M, ElSharawy A, Mayr G, Albrecht M, Prescott NJ, Onnie CM, Fournier H, Keith T, Radelof U, Platzer M, Mathew CG, Stoll M, Krawczak M, Nurnberg P, Schreiber S. 2007. Systematic association mapping identifies NELL1 as a novel IBD disease gene. PLoS ONE 2:e691.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182-189.

Harismendy O, Frazer K. 2009. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques 46:229-231.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10:R32.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362-9367.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599-603.

Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V, Finkel Y, Modigliani R, Gower-Rousseau C, Macry J, Merlin F, Chamaillard M, Jannot AS, Thomas G, Hugot JP. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet 70:845-857.

Loftus EV. 2004. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology 126:1504-1517.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, Mcdade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

McKernan K, Blanchard A, Kotler L, Costa G. 2006. Reagents, methods, and libraries for bead-based sequencing. Patent.

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

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. Science 324:387-389.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272-276.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603-606.

Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, Drummond H, Lees CW, Khawaja SA, Bagnall R, Burke DA, Todhunter CE, Ahmad T, Onnie CM, McArdle W, Strachan D, Bethel G, Bryan C, Lewis CM, Deloukas P, Forbes A, Sanderson J, Jewell DP, Satsangi J, Mansfield JC, Cardon L, Mathew CG. 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39:830-832.

Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van EP, Bradley WE, Croteau P, Nguyen-Huu Q, Segal J, Debrus S, Allard R, Rosenstiel P, Franke A, Jacobs G, Nikolaus S, Vidal JM, Szego P, Laplante N, Clark HF, Paulussen RJ, Hooper JW, Keith TP, Belouchi A, Schreiber S. 2007. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A 104:14747-14752.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-86.:365-386.

Sanger F, Nicklen S, Coulson AR. 1977. Dna Sequencing with Chain-Terminating Inhibitors. Proc Natl Acad Sci U S A 74:5463-5467.

Schreiber S, Rosenstiel P, Albrecht M, Hampe J, Krawczak M. 2005. Genetics of Crohn disease, an archetypal inflammatory barrier disease. Nat Rev Genet 6:376-388.

Schuster SC. 2008. Next-generation sequencing transforms today's biology. Nat Methods 5:16-18.

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

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728-1732.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308-311.

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, Stewart DA, Zhang L, Ranade SS, Warner JB, Lee CC, Coleman BE, Zhang Z, McLaughlin SF, Malek JA, Sorenson JM, Blanchard AP, Chapman J, Hillman D, Chen F, Rokhsar DS, McKernan KJ, Jeffries TW, Marth GT, Richardson PM. 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 18:1638-1642.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989.

Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stahler CF, Chee MS, Stahler PF, Beier M. 2009. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. Genome Res 19:1616-1621.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60-65.

Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De JP, Van BC, De RP. 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15:436-442.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies

M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-8U5.

Yeager M, Xiao NQ, Hayes RB, Bouffard P, Desany B, Burdett L, Orr N, Matthews C, Qi LQ, Crenshaw A, Markovic Z, Fredrikson KM, Jacobs KB, Amundadottir L, Jarvie TP, Hunter DJ, Hoover R, Thomas G, Harkins TT, Chanock SJ. 2008. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 124:161-170.

Figure 1 – NOD2 genetic structure, long-range and ordinary PCR amplicons

Panel A and **B** show the physical localization of the *NOD2* gene on chromosome 16q. **Panel C** shows the gene structure, with the open reading frame highlighted by thicker and dark blue boxes that correspond to exons. Positions are according to NCBI's build 36. The coverage of each of the LR-PCR products (**Panel D**) is shown along with the Sanger PCR products (**Panel E**). The lower **Panel F** shows the identified SNPs and their positions, with non-synonymous SNPs highlighted in red.

Figure 2 – Average coverage across all samples and genomic GC-content

The coverages for the SOLiD mate-pair, SOLiD fragment and 454/FLX are shown divided in 100 bp bins across the sequenced region. All the samples sequenced are averaged within each of the technologies. The coverage for the different technologies is given relative to the average coverage seen using that technology. The GC content in the region is shown in the gray shaded area and was calculated according to the same 100 bp bins that were used for the coverage calculation. Positions are according to NCBI's build 36.

Figure 3 – Number of SNPs detected for each individual sample

The upper **Panel A** shows the absolute number of homozygous SNPs detected for each individual sample, while the lower **Panel B** shows the corresponding number of heterozygous SNPs. All SNPs detected with the second generation technologies are included in the figure.

Figure 4 – Simulation of SNP detection at different coverage depths

The graphs show the number of SNPs discovered at the specified coverage summed up for all individuals and divided into homozygous and heterozygous SNPs (as defined from full coverage call). False positives are SNPs that appeared during the simulation and that were not detected at full coverage for that specific technology irrespective of what the other technologies showed. The different genotype call class constitutes SNPs that were detected at full coverage but changed their genotype call during the simulation. The cleaned numbers are the total, homozygous and heterozygous SNPs detected, with removal of the false positives and SNPs changing their genotype call. **Panel A** shows simulation results from the SOLiD mate-pair library, **Panel B** from the SOLiD fragment library and **Panel C** shows the results from the 454/FLX runs. If the original coverage was below 100x for the 454/FLX runs (see Table 1), the maximum coverage achieved was used for the simulations above that coverage.

Figure 5 – Comparison of simulation results

The graphs show a comparison of the simulation experiments for the SOLiD mate-pair, SOLiD fragment and 454/FLX library. The number of SNPs for each of the technologies corresponds to the number of SNPs that were detected and that were identical to what was seen at full coverage. The gray line represents the total number of SNPs that were detected with any of the second generation technologies and subsequently confirmed by Sanger sequencing.

Table 1 – Long-range PCR primers and the coverage achieved for each of the products

Fragment	Primer pair	Annealing temperature	Amplicon length	Amplicon start – end (NCBI b36)	Average per base coverage for SOLiD mate-pair / SOLiD fragment / 454 FLX (% of base positions <20x)										
					Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Sample 11
1	GGAGTGGGCCTTGGAGTC / GTCCAGGACACTCTCGAAGC	60 °C	5963 bp	49285038 - 49291000	1621/6136/186 (0/0/0)	1724/6296/54 (0/0/2)	1238/4466/139 (0/1/0)	1998/7623/20 (0/0/41)	1579/8481/51 (0/1/2)	254/14688/130 (6/0/0)	362/15589/116 (4/0/0)	206/13244/113 (7/0/0)	979/14493/255 (0/0/0)	869/9572/75 (2/0/0)	359/17960/161 (5/0/0)
2	GCTTTTCAGGCACAGAGGAG / AAGTGTGGGGAATCTTGCAC	60 °C	9496 bp	49290923 - 49300418	824/3578/196 (0/0/0)	904/4728/76 (0/0/1)	819/2609/114 (0/0/0)	1404/4022/20 (0/0/55)	624/3959/57 (0/0/1)	46/12901/143 (10/0/0)	128/9852/126 (2/0/0)	34/8733/122 (25/0/0)	261/7602/259 (1/0/0)	145/6221/80 (3/0/0)	338/11339/188 (1/0/0)
3	CTGGCACTCTTTAGGGCTTG / GCTGCAACTGAATCCAGACA	60 °C	10,722 bp	49300077 - 49310798	2059/4961/209 (0/0/0)	2205/5165/64 (0/0/0)	1301/3370/138 (0/0/0)	2970/5549/23 (0/0/32)	1513/6016/59 (0/0/0)	297/15360/122 (3/0/0)	360/15632/126 (1/0/0)	230/12559/114 (4/0/0)	639/13541/269 (0/0/0)	317/9151/76 (1/0/0)	996/16578/170 (0/0/0)
4	GTTCTCCTAGCTGCCACACC / GTCCACACAACCGCTCCTAT	60 °C	4276 bp	49310609 - 49314884	1618/6776/220 (0/0/0)	1507/7434/65 (0/0/0)	1094/5051/162 (0/0/0)	1900/8959/25 (0/0/25)	1041/8398/62 (0/0/1)	293/14109/133 (3/0/0)	221/28576/141 (5/0/0)	152/11552/135 (8/0/0)	602/13348/353 (2/0/0)	448/10136/83 (2/0/0)	349/15215/169 (2/0/0)
5	CCTGGTGGGGAACAACATT / AGAGCAGGGCTTTCAAACAA	61 °C	8402 bp	49314713 - 49323114	481/2853/144 (0/0/0)	273/3326/47 (0/0/2)	444/2908/125 (0/0/0)	923/4421/19 (0/0/69)	475/5043/50 (0/0/2)	91/10803/102 (3/0/0)	110/14252/143 (3/0/0)	47/6594/75 (11/0/1)	74/9955/224 (22/0/0)	87/5645/59 (4/0/2)	234/10997/155 (2/0/0)
6	GCAGCGAATGCAGATATCAA / GCGAAAGGAGACTCAACACC	59 °C	3053 bp	49322284 - 49325336	1289/5820/276 (0/0/0)	1530/9077/81 (0/0/0)	1019/6399/212 (0/0/0)	1834/9719/32 (0/0/8)	978/13294/109 (1/0/0)	191/18700/173 (7/0/0)	192/16043/234 (8/0/0)	104/13703/147 (14/0/0)	572/30598/445 (4/0/0)	464/12931/133 (5/0/0)	306/19904/246 (6/0/0)

Primers used for generation of the Long-range PCR products and the coverage achieved for that product for the SOLiD mate-pair, SOLiD fragment library and the Roche 454/FLX. The numbers in parentheses list the percentages of base pair positions that had a coverage depth below 20x.

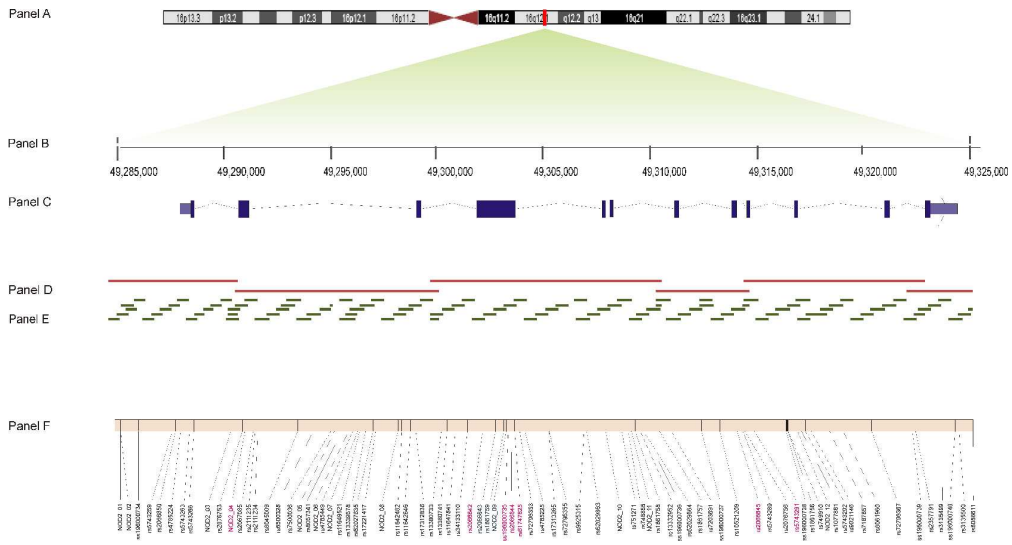

Figure 1 – NOD2 genetic structure, long-range and ordinary PCR amplicons
Panel A and B show the physical localization of the NOD2 gene on chromosome 16q. Panel C shows the gene structure, with the open reading frame highlighted by thicker and dark blue boxes that correspond to exons. Positions are according to NCBI's build 36. The coverage of each of the LR-PCR products (Panel D) is shown along with the Sanger PCR products (Panel E). The lower Panel F shows the identified SNPs and their positions, with non-synonymous SNPs highlighted in red.

283x151mm (600 x 600 DPI)

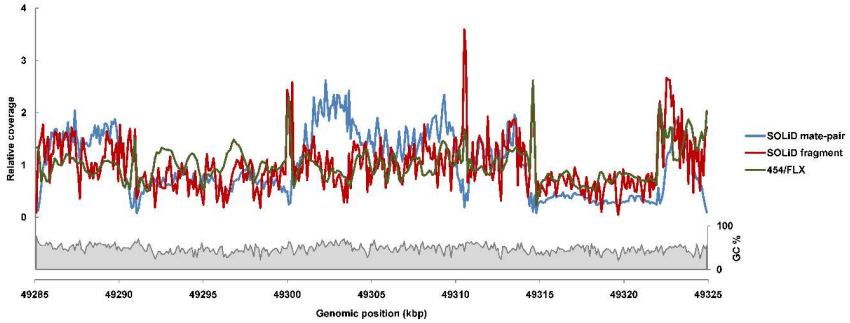

Figure 2 – Average coverage across all samples and genomic GC-content
The coverages for the SOLiD mate-pair, SOLiD fragment and 454/FLX are shown divided in 100 bp bins across the sequenced region. All the samples sequenced are averaged within each of the technologies. The coverage for the different technologies is given relative to the average coverage seen using that technology. The GC content in the region is shown in the gray shaded area and was calculated according to the same 100 bp bins that were used for the coverage calculation. Positions are according to NCBI's build 36.

266x200mm (600 x 600 DPI)

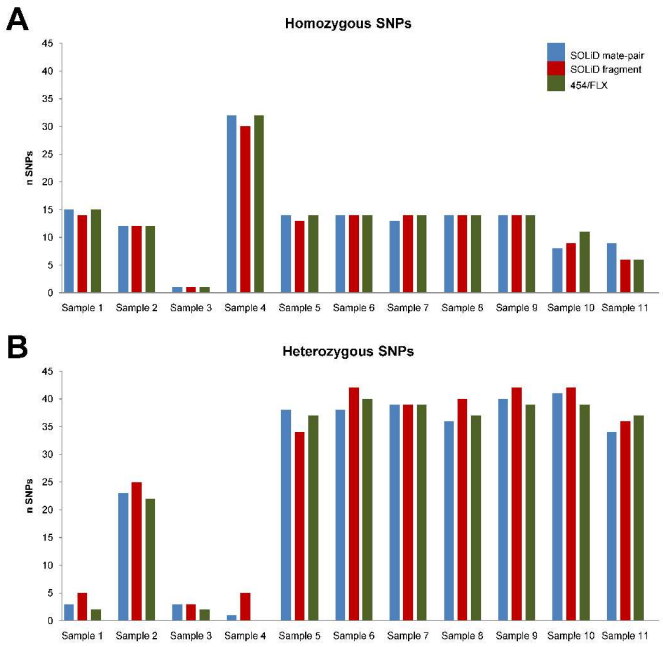

Figure 3 – Number of SNPs detected for each individual sample
The upper Panel A shows the absolute number of homozygous SNPs detected for each individual sample, while the lower Panel B shows the corresponding number of heterozygous SNPs. All SNPs detected with the second generation technologies are included in the figure.

266x200mm (600 x 600 DPI)

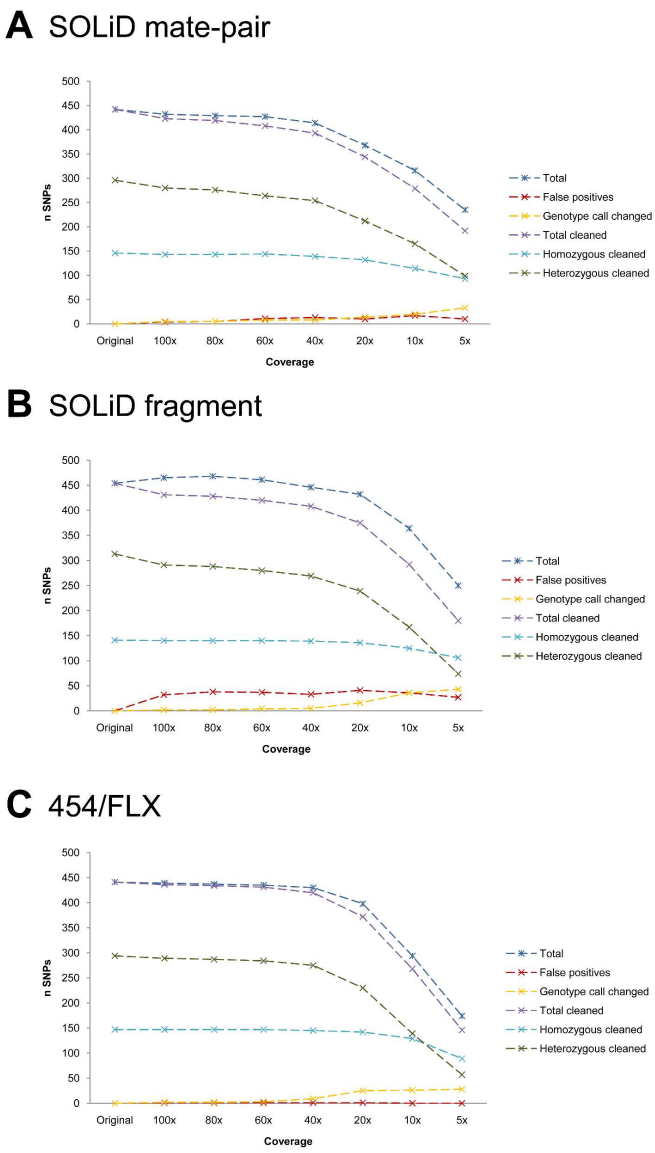

Figure 4 – Simulation of SNP detection at different coverage depths

The graphs show the number of SNPs discovered at the specified coverage summed up for all individuals and divided into homozygous and heterozygous SNPs (as defined from full coverage call). False positives are SNPs that appeared during the simulation and that were not detected at full coverage for that specific technology irrespective of what the other technologies showed. The different genotype call class constitutes SNPs that were detected at full coverage but changed their genotype call during the simulation. The cleaned numbers are the total, homozygous and heterozygous SNPs detected, with removal of the false positives and SNPs changing their genotype call. Panel A shows simulation results from the SOLiD mate-pair library, Panel B from the SOLiD fragment library and Panel C shows the results from the 454/FLX runs. If the original coverage was below 100x for the 454/FLX runs (see Table 1), the maximum coverage achieved was used for the simulations above that coverage.

165x220mm (600 x 600 DPI)

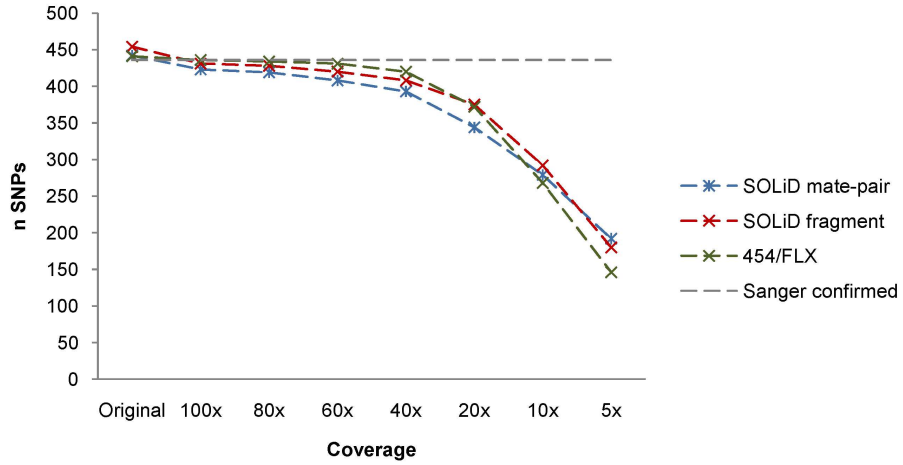

Figure 5 – Comparison of simulation results
The graphs show a comparison of the simulation experiments for the SOLiD mate-pair, SOLiD fragment and 454/FLX library. The number of SNPs for each of the technologies corresponds to the number of SNPs that were detected and that were identical to what was seen at full coverage. The gray line represents the total number of SNPs that were detected with any of the second generation technologies and subsequently confirmed by Sanger sequencing.

120x64mm (600 x 600 DPI)

Supplementary Material

SNP discovery performance of two second generation sequencing platforms in the NOD2 gene region

Authors: Espen Melum, Sandra May, Markus B. Schilhabel, Ingo Thomsen, Tom H. Karlsen, Philip Rosenstiel, Stefan Schreiber and Andre Franke

Supplementary Table S1 – Haplotype background of the included patients

Sample	Haplotype 1 (f_{co}; f_{ca}; p-value)	Haplotype 2 (f_{co}; f_{ca}; p-value)
1	H09 (0.02; 0.02; ns)	H09 (0.02; 0.02; ns)
2	H07 (0.03; 0.01; p=0.0319)	H01 (0.41; 0.29; p<0.0001)
3	H02 (0.23; 0.17; p=0.0133)	H02 (0.23; 0.17; p=0.0133)
4	H04 (0.04; 0.11; p<0.0001)	H04 (0.04; 0.11; p<0.0001)
5	H08 (0.00; 0.04; p<0.0001)	H01 (0.41; 0.29; p<0.0001)
6	H06 (0.02; 0.02; ns)	H01 (0.41; 0.29; p<0.0001)
7	H06 (0.02; 0.02; ns)	H01 (0.41; 0.29; p<0.0001)
8	H08 (0.00; 0.04; p<0.0001)	H01 (0.41; 0.29; p<0.0001)
9	H08 (0.00; 0.04; p<0.0001)	H01 (0.41; 0.29; p<0.0001)
10	H10 (0.02; 0.01; ns)	H01 (0.41; 0.29; p<0.0001)
11	H09 (0.02; 0.02; ns)	H01 (0.41; 0.29; p<0.0001)

The table lists the haplotype background of the patients included in this study as determined by PHASE. The nomenclature used is the same as Croucher *et al.* reported. The number in parentheses represents the frequency of the haplotype in healthy controls (f_{co}), in patients with Crohn's disease (f_{co}) and the significance level (**p-value**) for the association from that publication.

Croucher PJP, Mascheretti S, Hampe J, Huse K, Frenzel H, Stoll M, Lu T, Nikolaus S, Yang SK, Krawczak M, Kim WH, Schreiber S. 2003. Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. Eur J Hum Genet 11:6-16.

Supplementary Table S2 – Primers used for Sanger sequencing

	Forward primer	Reverse primer	Genetic region
Fragment 1	GGAGTGGGCCTTGGAGTC	TAGGAGAGCAAACAGCTGGG	49285038-49285533
Fragment 2	AAATCCATCCCTTCCCAGAC	ATGCTTCCAGCCTCAGATCC	49285463-49285937
Fragment 3	TCCCGATCACTCAGCCAT	ATCCTGCTTCCCAGCTTTGT	49285838-49286329
Fragment 4	CCCTGCTTCTGAGTTTGCTC	TCTCGTGTCTCATCCAGTGC	49285656-49286184
Fragment 5	GAAAGGAATGGACAGCAATGA	CTATCGACACGGCCTTTGAA	49286240-49286729
Fragment 6	ATGGCCCAGAGGTGGTTAAA	TTCATTTCCTCCAGGACACC	49286655-49287136
Fragment 7	CCAGTGTTAGGAACCAGCCA	TGTCCTGAGTCTCTGCTGGG	49287064-49287537
Fragment 8	AGTGTTAGGGAGGGAGAATGC	CCTTAACATCCTTCCTCTTACTCC	49287466-49287924
Fragment 9	TGACAGATTTCGCCTGAAGA	GCCTACGTGTGAGCCATACA	49287857-49288320
Fragment 10	TCCTGGTGTCCACAGAAGC	ATGTCCGAGGAGGACACTTG	49288261-49288721
Fragment 11	CAGCCTAATGGGCTTTGATG	CCACTGGAAGCCAGGATCTA	49288639-49289110
Fragment 12	TGCTTAGATGTGGGCACAAG	CAGGGACATTCCCAGACACT	49289039-49289507
Fragment 13	ACTTGGCATCTGCTCCCTT	ATTGCTCTAAAGGTCCTGTCC	49289468-49289908
Fragment 14	AGAACGCTCGGAAACAACTT	TCAGGGAATTCATTCTCTTCG	49289838-49290309
Fragment 15	ATCCAGCCAGCATGAAAGAG	CCCAGAGAAGGCAAGAAACA	49290252-49290733
Fragment 16	GCTGGTCTAGCGCTCCTG	GAGGTGGAAGCCCTCGTAGT	49290551-49291054
Fragment 17	ACCAAGCATGGATATGGGAG	GTCCAGGACACTCTCGAAGC	49290641-49291000
Fragment 18	CCAAGCATGGATATGGGAGT	TTCTGACAGGCCCAAGTACC	49290642-49291131
Fragment 19	GTCGAGCTGCTGGTCTCAG	TTTACCTGCCTGGCTACCTTT	49290950-49291506
Fragment 20	CTCATGAAGTCAGCCTGTGG	AAGGGAGACTGAGGCTGGA	49291437-49291971
Fragment 21	AACCCTGTGTGATCTCAGGC	TGGTCTCAACTTCCCGATTT	49291917-49292481
Fragment 22	TCCTGAGAGGCAGATGGAAC	GAAACTAAGGTCAGTCCAATAACGA	49292420-49292978
Fragment 23	AGCAGAAGTGGTGCAGTCCT	CTTCAGCCCAGAAGTTCGAG	49292903-49293459
Fragment 24	AGGGGTTTGGAGAAGGTAGC	CCCAACCTTCAATGTTTGCT	49293084-49293724
Fragment 25	TCACTGTGTCACCCAGGCT	TTGGGAAGACATGTTGGTTG	49293384-49293950
Fragment 26	ATTTGGTGGCCTTCCCTTAG	GGAATCAGGCAATTCAGAGC	49293870-49294451
Fragment 27	TCAGTTGGGCATAGCAGTTG	CCTGACCTTGTGATCCACCT	49294316-49294942
Fragment 28	GCCAAGAATAGTGGCATGTG	TACAAAGCAGCCCACTTCCT	49294905-49295367
Fragment 29	AGGAAGTGGGCTGCTTTGTA	GAGAGGATAATGGAATAATGCAATC	49295348-49295430
Fragment 30	GGATTGATAGTTTGGCTGGG	CCTGTAAACTAATCAAGAACATTTCCC	49295827-49296405
Fragment 31	GGCCATCATTTTCCTTCAGA	GCAAAAAGACACAGAGATGGAA	49295863-49296524
Fragment 32	CCACCACCACACCCTTTAGT	CAGTCAAACCTACCCCTTGC	49296093-49296620
Fragment 33	CCAAGAGGTACTTTCAATCTGGA	GGTAGACAAGTTTCACCCGC	49296340-49296895
Fragment 34	TGGGCTCATGATACTCTTTGG	TCACCTTCTAGCACACGCTG	49296807-49297377
Fragment 35	TACACACACTCCAGTCAGGG	CTGCTTCTGTTCCTGCCC	49297306-49297866
Fragment 36	GATTAGACTCTGGCTGTGGCA	CCTAAGGTCCGGAGTTCCA	49297811-49298346
Fragment 37	GGTTTCACCATGTTGTCCAG	CTTGCACATCCAACTGCCTA	49298301-49298848
Fragment 38	TTGGGACATGCTTTGAAGGT	GCATCTCTCTTCCATCCAGAA	49298608-49299181
Fragment 39	CGTTCAATTCACCATCCAAC	CCCTCAATATATGGTAGCCCAA	49299085-49299635
Fragment 40	AGCTTTCATTATCTCCCACACA	GCCAATGAGCTTTCCCTACA	49299580-49300128
Fragment 41	CAAGCCTGGCACTCTTTAGG	CCTGGCCTGGTCTTACAGTG	49300072-49300398
Fragment 42	CTGGCACTCTTTAGGGCTTG	TATTGCTGAGCTTTCTGCCC	49300077-49300634
Fragment 43	GTCTAAGTTCAGGGATCCTGGT	AAAGGAACTCTGAGATGTGATAAAG	49300584-49301143
Fragment 44	GCTAGGTCCACCCAGATCAA	GGAGGAGTTCTAGATGACGGG	49301069-49301617
Fragment 45	GCCACTATGGCAGGACAGAT	CCAGACCTCCAGGACATTCT	49301537-49302103
Fragment 46	GCTCAGTCTCGCTTCCTCAG	ATTCTTCAGCAGGTTGCCCT	49302017-49302577
Fragment 47	CTGTCCAGACCCTGCTCTTC	AAATGTCATCAGGGCTGACC	49302528-49303064
Fragment 48	TGCTGCTACGTGTTCTCAGC	GGATGAGCCAGATGAACCC	49303004-49303553
Fragment 49	AGCACTTCCACTCCATCCC	GCCAGGGTGGTTGTAGAGAT	49303473-49304047
Fragment 50	ATTGACTGGCCTATGTGCTG	CCCAAGATGTGGAAAGAGGC	49303972-49304516
Fragment 51	TTTGCCCTTTGGAGAATCTG	TTTAAGTGTGTCATCTGTTTCCTC	49304460-49305000
Fragment 52	TGATTATACCCATCATGGTCTTG	TGCACATTCTTATTTCTATCAGGG	49304954-49305490
Fragment 53	TGAGTGGGAATTTGGAAACC	CTACGGCCTTGGCAGTTT	49305429-49305977
Fragment 54	GCAATTACTGTCCTTGAGCGA	GACTGCAGGTTGCAGGTCAT	49305909-49306471
Fragment 55	TTTTACTGTGTACCGGGTGCT	TGGAGCTGGCAATCAAGTATC	49306759-49307339
Fragment 56	GTACAGTGGCACAATCTCGG	CTGTCTGCCACTCCTCTTGG	49306880-49307437
Fragment 57	GGGAAGCTTTCAGATGATGTT	ATTCATCCCTGAAGTGCCAG	49307368-49307917
Fragment 58	GTGCTCCATCTATGCAGGGT	AAACCTGGGCTCACCTCAAT	49307860-49308400
Fragment 59	CTCCATGGCTAAGCTCCTTG	GGGCAATTTAAAGGCATCAG	49308339-49308905
Fragment 60	GGTTTGCAGTCCATCTGGTT	GCAGGTGTGTAGGTATTGGGA	49308808-49309391
Fragment 61	CAGTGGATGCCTGAAACCAT	TGCTATTCTGATCACGTGCC	49309318-49309850
Fragment 62	TTTGAGCAACAATGAGGTCG	TACCCATCATCTTTCAGCCA	49309699-49310226
Fragment 63	CCACATGAAGGGCTGCAC	CCCTAGGCTTCCTGGGATAA	49310158-49310736
Fragment 64	GTTCTCCTAGCTGCCACACC	GCTGCAACTGAATCCAGACA	49310609-49310798
Fragment 65	GTTCTCCTAGCTGCCACACC	TAGTTTACTCAGCGGCTCCC	49310609-49311187
Fragment 66	CTGCTGTGCAAACAGCGT	CTGAGGATCAGAGGGCTTCC	49311109-49311680
Fragment 67	GAGGACAGTGCCCACGAG	GAAATGGGCTGGAACACCTA	49311634-49312180
Fragment 68	CCTGGGACTGTAGGAAGAGAGA	GGTGCCCTAATAAATATGTGTTGAA	49312125-49312705
Fragment 69	AGGTCACATCTGCCACATCA	GCAAGGTGGCTCACTTGTGT	49312646-49313198
Fragment 70	CTGCTTGAGGTCACATCTGC	GGCTCACTTGTGTAATCCCAA	49312639-49313191

	Forward primer	Reverse primer	Genetic region
Fragment 71	CAGGTGATCCTCCCTCCTTT	GGTAACTGCCACCCACAAGT	49313138-49313689
Fragment 72	TGTTATGGGCAGGTACTGGAG	CAGCTCCTCCCTCTTCACCT	49313638-49314178
Fragment 73	TGGCTCAGGTAAGCTTCAGA	TGTCAGAGCAGGGAAAGACC	49314110-49314686
Fragment 74	CAGGAGAGCACCACGAATTT	GTCCACACAACCGCTCCTAT	49314619-49314884
Fragment 75	AACATTGGCAGTGTGGGTG	GCAGCAGCAGCAGTAATAGTAG	49314726-49315302
Fragment 76	TGCCTGGCATATTTGTAAGTG	CTGATGCTCTCATAACAAGGTTC	49315221-49315771
Fragment 77	TTTGCCACTTGGAATTTCTT	CCTACACTGCCTACATCCAGC	49315720-49316269
Fragment 78	GAGGTTTAAACAAGGAAAGGGTT	GGATTCTCAATTCCAGGGAC	49316206-49316758
Fragment 79	TGAGAGCCCTTGGAGTGAAT	TGAGCTCCAATTTCCCTCAT	49316705-49317268
Fragment 80	GGGAGGTGGTCCCAGTAAAT	CTAGGCAGATGCCACTTTCC	49317154-49317736
Fragment 81	AGAGTGGTGAAGTGAGATGGG	TTCCTAGGACCACATTTGCC	49317183-49317755
Fragment 82	GAAATGGGCTGGGAAGTTG	CCAGTCTAGGGTGCAATGGT	49317617-49318528
Fragment 83	CCAGGCTTAGTTTCTGCCAA	TTTGCTCCAACCTACAACTCA	49318229-49318784
Fragment 84	GATGAGTGCCTCAGTTGCCT	ATCCCTGCGAAACCATTACA	49318667-49319253
Fragment 85	GCTGCAGGTGTTGCAGTTT	TGCAGTGGTGGTTCAAGAAG	49319235-49319715
Fragment 86	CCTGCCAGTACACTCCTCATC	GGAGACAGGATAGGGCCAG	49319659-49320226
Fragment 87	GAGCAACCTGGGAAACACAT	GAAGACCCACCACTCTCTGC	49320169-49320710
Fragment 88	GGGACACAGCCAAACCATA	ACAATGGTGAGCTTCAAGGA	49320644-49321215
Fragment 89	CTGGCTAACTCCTGCAGTCTC	CACTGTGCAACCTGCTCCTA	49321141-49321720
Fragment 90	ATTGGATCAGACCTTCTGCG	ATGAGTTGCCAAAGGGTCC	49321657-49322214
Fragment 91	TCTTCTCAGATGGGCAAACC	ACTTTGCCCAGAACCTGATG	49322131-49322693
Fragment 92	ATGATTGTGATAGCTCTGTGGC	AGAGCAGGGCTTTCAAACAA	49322643-49323114
Fragment 93	GGGTCTGATGTTATTTGCTGC	TTCAGCCTTAGGCAGGACAG	49322795-49323353
Fragment 94	TTGTGAGCAGGCTGTGAGTT	GGGTGCTGTAAGTGAAATATCAAA	49323263-49323825
Fragment 95	TCCTTCCTCTGAGGCTGAAA	ACAAAGGCTTCAACATACATCA	49323759-49324314
Fragment 96	AAATTAGGCCGTTCCTTCAAA	TTCCTCCGCTTCAAGTAGAAA	49324245-49324808
Fragment 97	CTGGGCAACAAGAGTGGAAC	CCATAGACACACTCCCATGC	49324697-49325246
Fragment 98	TGGTCACCCAGCAAGGTAAA	GCGAAAGGAGACTCAACACC	49325143-49325336

The forward and reverse primers used for Sanger sequencing of the genetic region. The base pair positions that the fragments cover are listed according to NCBI's build 36.

Supplementary Table S3 – SNPs identified with the second generation sequencing technologies compared with the results from Sanger sequencing

			Sample 1				Sample 2				Sample 3				Sample 4				Sample 5				Sample 6				Sample 7				Sample 8				Sample 9				Sample 10				Sample 11			
SNP	BP	Ref	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S
NOD2_01	49,285,092	C		S																																								
NOD2_02	49,285,093	G		S				S								S																												
ss196000734	49,285,914	C																																					Y	Y	Y	Y				
rs5743259	49,287,589	G																																					K	K	K	K				
rs2066850	49,287,730	G	A		A	A	R		R	R					A		A	A	R		R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs4785224	49,287,947	G					R	R	R	R									R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs5743263	49,288,366	T									Y	Y	Y	Y																														
rs5743266	49,288,597	G	A	A	A	A	R	R	R	R					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
NOD2_03	49,290,339	G														S								S								S											
rs2076753	49,290,875	G	T	T	T	T	K	K	K	K					T	T	T	T	K	K	K	K		K	K	K	K	K	K	K		K	K	K	K	K	K	K	K	T	K	K	K
NOD2_04	49,290,920	G	S				S				S				S				S																S							C			
rs2067085	49,291,360	C					S	S	S	S									S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
rs2111235	49,291,470	A	G	G	G	G	G	G	G	G					G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
rs2111234	49,291,534	G	A	A	A	A	A	A	A	A					A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs8045009	49,293,442	C																	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
rs6500328	49,294,157	A																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs7500036	49,294,999	C	T	T	T	T	T	T	T	T					T	T	T	T	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	T	Y	Y	Y
NOD2_05	49,295,099	A																						R				R				R				R							
rs8057341	49,295,481	A	G	G	G	G	G	G	G	G					G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
NOD2_06	49,295,958	C		Y				Y												Y																								
rs4785449	49,296,048	T	C	C	C	C	C	C	C	C					C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
NOD2_07	49,296,287	G						R				R				R																												
rs11649521	49,296,331	G	A	A	A	A	A	A	A	A					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs13339578	49,296,606	G	A	A	A	A	A	A	A	A					A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
rs62027635	49,296,951	G	A	A	A	A	A	A	A	A					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs17221417	49,297,083	C	G	G	G	G	G	G	G	G					G	G	G	G	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
NOD2_08	49,298,110	C																						Y																				
rs11642482	49,298,265	C																	S		S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
rs11642646	49,298,687	C																	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
rs17312836	49,298,963	A																	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
rs13380733	49,300,182	G	A	A	A	A	A	A	A	A					A	A		A	R	R		R	R	R	R	R	R	R	R	R	R	R		R	R	R	R	R	R		R	R	R	R
rs13380741	49,300,439	G	A	A	A	A	A	A	A	A					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs11647841	49,300,832	G																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
rs34133110	49,301,420	C																	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
rs2066842	49,302,125	C					Y	Y	Y	Y					T		T	T	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		
rs2066843	49,302,700	C					Y	Y	Y	Y					T		T	T	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		
rs1861759	49,303,084	T																	K		K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
NOD2_09	49,303,168	G						S								S																												
ss196000735	49,303,302	C																									S	S	S	S														
rs2066844	49,303,427	C														T	T	T																										
rs61747625	49,303,587	C																									Y	Y	Y	Y														
rs72796353	49,303,795	A																					M	M	M	M									M	M	M	M						
rs4785225	49,304,047	C					S	S	S	S					G	G	G	G	G		G	G	G	G	G	G		G	G	G	G	G	G	G	G	G		G	G	G	S	S	S	S
rs17313265	49,305,205	C					Y	Y	Y	Y					T	T	T	T	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y				
rs72796355	49,305,240	C					S	S	S	S					G	G	G	G	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S				
rs9925315	49,306,875	A					W	W	W	W					T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	W	W	W	W

				Sample 1				Sample 2				Sample 3				Sample 4				Sample 5				Sample 6				Sample 7				Sample 8				Sample 9				Sample 10				Sample 11			
	SNP	BP	Ref	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S	M	F	R	S
1	rs62029863	49,306,909	G					K	K	K	K					T	T	T	T	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K								
2	NOD2_10	49,307,824	C																														M														
3	rs751271	49,308,676	T					K	K	K	K					G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	K	K	K	K
4	rs748855	49,308,899	A																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
5	NOD2_11	49,309,063	G														S																														
6	rs1861758	49,309,288	G																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
7	rs13332952	49,309,473	G					R	R	R	R					A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	R	R	R	R	R	R	R	R
	ss196000736	49,309,610	C									Y	Y	Y	Y																																
8	rs62029864	49,309,665	G					R	R	R	R					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R								
9	rs1861757	49,310,316	T																	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
10	rs7203691	49,310,925	G																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
11	ss196000737	49,312,786	T	K	K	K	K																																						
12	rs10521209	49,313,210	T																	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
	rs2066845	49,314,041	G																	S	S	S	S									S	S	S		S	S	S							
13	rs5743289	49,314,275	C					Y	Y	Y	Y					T	T	T	T																										
14	rs2076756	49,314,382	A					R	R	R	R					G	G	G	G	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R						
	rs5743291	49,314,777	G																																					R	R	A	R				
15	ss196000738	49,316,297	G					R	R	R	R																																				
16	rs1861756	49,316,339	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
17	rs749910	49,316,350	G					R	R	R	R					A	A	A	A	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R						
	NOD2_12	49,316,769	C																			M								M								M						M	
18	rs1077861	49,317,048	A					W	W	W	W					T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	W	W	W	W	W	W	W	W
	rs5743292	49,317,232	T													A	A	A	A																												
19	rs9921146	49,317,481	A					M	M	M	M					C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	M	M	M	M	M	M	M	M
20	rs7187857	49,318,339	A					R	R	R	R					G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	R	R	R	R	R	R	R	R
21	rs8061960	49,318,935	C																	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
22	rs72796367	49,320,272	T																					Y	Y	Y	Y									Y	Y	Y	Y								
23	ss196000739	49,322,271	G	R	R	R	R																																						
24	rs2357791	49,322,337	G																	R		R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R		R	R	R	R	R	R	R	
25	rs3135499	49,323,628	A																	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
	ss196000740	49,324,255	G																					R	R	R	R																				
26	rs3135500	49,324,387	G																	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
	rs8056611	49,325,148	A																	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	R	R	R	G	G	R	R	R

Overview of all SNPs identified divided for the different technologies. The SNPs listed in the columns named **M**, **F**, **R** and **S** are from the samples sequenced with the SOLiD mate-pair library, SOLiD fragment library, Roche 454/FLX and Sanger, respectively. SNPs that had already been reported in the NCBI's database of SNPs (dbSNP 130) are listed with NCBI's dbSNP IDs (the rs-numbers) and the Sanger confirmed novel variants are listed with their temporary accession numbers (the ss-numbers). The rest of the SNPs are labeled NOD2_01 to NOD2_12. The coding non-synonymous SNPs are highlighted by red SNP-IDs.

Supplementary Table S4 – Allele frequencies of SNPs reported in the HapMap populations

		Allele frequencies			
SNP	**BP**	**CEU**	**CHB**	**JPT**	**YRI**
rs5743259	49,287,589	0.00	0.00	0.00	0.00
rs4785224	49,287,947	0.37	0.07	0.04	0.30
rs5743263	49,288,366	0.03	0.00	0.00	0.00
rs5743266	49,288,597	0.36	0.00	0.00	0.13
rs2076753	49,290,875	0.34	0.00	0.01	0.02
rs2067085	49,291,360	0.39	0.06	0.02	0.25
rs2111235	49,291,470	0.75	0.35	0.28	0.67
rs2111234	49,291,534	0.73	0.29	0.27	0.63
rs6500328	49,294,157	0.38	0.30	0.24	0.50
rs8057341	49,295,481	0.72	0.24	0.22	0.83
rs11649521	49,296,331	0.36	0.00	0.00	0.00
rs13339578	49,296,606	0.74	0.30	0.26	0.56
rs17221417	49,297,083	0.35	0.00	0.00	0.01
rs11642646	49,298,687	0.38	0.26	0.19	0.16
rs17312836	49,298,963	0.38	0.26	0.19	0.15
rs13380733	49,300,182	n/a	0.00	0.00	0.00
rs13380741	49,300,439	0.36	0.00	0.00	0.06
rs11647841	49,300,832	0.39	0.23	0.20	0.17
rs2066842	49,302,125	0.34	0.00	0.01	0.01
rs2066843	49,302,700	0.33	0.00	0.00	0.01
rs1861759	49,303,084	0.37	0.14	0.15	0.16
rs4785225	49,304,047	0.71	0.29	0.22	0.55
rs17313265	49,305,205	0.34	0.02	0.04	0.04
rs751271	49,308,676	0.71	0.24	0.23	0.54
rs748855	49,308,899	0.37	0.21	0.18	0.16
rs1861758	49,309,288	0.37	0.21	0.17	0.21
rs1861757	49,310,316	0.39	0.18	0.16	0.07
rs10521209	49,313,210	0.39	0.21	0.17	0.17
rs5743289	49,314,275	0.23	0.00	0.00	0.00
rs2076756	49,314,382	0.32	0.00	0.00	0.01
rs5743291	49,314,777	0.10	0.00	0.00	0.00
rs1861756	49,316,339	1.00	1.00	1.00	1.00
rs1077861	49,317,048	0.72	0.21	0.17	0.18
rs3135499	49,323,628	0.40	0.25	0.23	0.57
rs3135500	49,324,387	0.39	0.24	0.22	0.57
rs8056611	49,325,148	0.48	0.32	0.30	0.77

The frequencies of the alternate allele for the detected SNPs are shown for the four HapMap populations (CEU: CEPH (Utah residents with ancestry from northern and western Europe), CHB: Han Chinese in Beijing, China, JPT: Japanese in Tokyo, Japan, YRI: Yoruba in Ibadan, Nigeria). The frequencies are based on frequencies for SNPs reported in the HapMap data release 27.