RESEARCH NOTE

REVISED Long read nanopore sequencing for detection of *HLA* and *CYP2D6* variants and haplotypes [version 2; referees: 2 approved]

Ron Ammar[1], Tara A. Paton[2], Dax Torti[1], Adam Shlien[3], Gary D. Bader[1,4,5]

[1]The Donnelly Centre, University of Toronto, Toronto, ON, M5S3E1, Canada
[2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, M5G0A4, Canada
[3]Department of Laboratory Medicine and Pathobiology, University of Toronto; Program in Genetics and Genome Biology & Department of Paediatric Laboratory Medicine The Hospital for Sick Children, Toronto, ON, M5G1X8, Canada
[4]Department of Computer Science, University of Toronto, Toronto, ON, M5S3G4, Canada
[5]Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S1A8, Canada

v2 **First published:** 21 Jan 2015, **4**:17 (doi: 10.12688/f1000research.6037.1)
Latest published: 20 May 2015, **4**:17 (doi: 10.12688/f1000research.6037.2)

Abstract

Haplotypes are often critical for the interpretation of genetic laboratory observations into medically actionable findings. Current massively parallel DNA sequencing technologies produce short sequence reads that are often unable to resolve haplotype information. Phasing short read data typically requires supplemental statistical phasing based on known haplotype structure in the population or parental genotypic data. Here we demonstrate that the MinION nanopore sequencer is capable of producing very long reads to resolve both variants and haplotypes of *HLA-A*, *HLA-B* and *CYP2D6* genes important in determining patient drug response in sample NA12878 of CEPH/UTAH pedigree 1463, without the need for statistical phasing. Long read data from a single 24-hour nanopore sequencing run was used to reconstruct haplotypes, which were confirmed by HapMap data and statistically phased Complete Genomics and Sequenom genotypes. Our results demonstrate that nanopore sequencing is an emerging standalone technology with potential utility in a clinical environment to aid in medical decision-making.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED **version 2** published 20 May 2015	✓ report	
version 1 published 21 Jan 2015	✓ report	✓ report

1 **Martin Kennedy**, University of Otago New Zealand

2 **Thomas Hoenen**, National Institute of Allergy and Infectious Diseases, National Institutes of Health USA

Discuss this article

Comments (0)

Corresponding authors: Ron Ammar (ron.ammar@mail.utoronto.ca), Gary D. Bader (gary.bader@utoronto.ca)

How to cite this article: Ammar R, Paton TA, Torti D *et al.* **Long read nanopore sequencing for detection of *HLA* and *CYP2D6* variants and haplotypes [version 2; referees: 2 approved]** *F1000Research* 2015, **4**:17 (doi: 10.12688/f1000research.6037.2)

Copyright: © 2015 Ammar R *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Grant information: This study was funded by a Large-scale Applied Project grant from Genome Canada and the Ontario Genomics Institute (grant ID OGI-068). We confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: R.A. is a member of the Oxford Nanopore Technologies Inc. MinION Access Programme and the MinION instrument and R7.3 flowcells were received free of charge.

First published: 21 Jan 2015, **4**:17 (doi: 10.12688/f1000research.6037.1)

REVISED **Amendments from Version 1**

Dear reviewers,

We would like to thank you for your thorough and useful review of our manuscript "*Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes*". Please find our point-by-point response in Supplementary file 2.

Both of you suggested we reproduce our results using the newest Oxford Nanopore Technology kits and pipeline. Therefore, we performed the sequencing experiment using the newest Oxford Nanopore Genomic DNA Sequencing protocol (SQK-MAP004). The library was sequenced on 3 fresh R7.3 flow cells, only one of which produced a significant number of reads (just over 1000 reads from the other two flow cells). As anticipated, the 2D consensus reads generated using the newer protocol were longer on average suggesting advances in basecalling accuracy since our initial experiments. Unfortunately, yield was significantly lower with a total of 885 aligned reads for all targeted loci (including our 3 genes-of-interest and the nanopore controls), which did not enable us to make accurate genotyping calls. These new read data have been deposited in the NCBI Sequence Read Archive along with our other data.

Thank you again for your feedback on our manuscript.

Sincerely,

The authors

See referee reports

Introduction

An important aspect of precision medicine is the study of how genes influence individual response to drug therapies, known as pharmacogenomics (PGx). PGx genotyping impacts the choice of drug dosing in many medical contexts. As an example, in acute lymphoblastic leukemia patients the metabolizer status of thiopurine methyltransferase (*TPMT*) must be considered when calculating the initial drug dose of mercaptopurine (6-MP) to ensure proper treatment and avoid fatal toxicity[1,2].

PGx data are typically collected by sequencing a small panel of known PGx genes via traditional Sanger sequencing, or targeted genotyping technologies[3]. Diagnostic labs are also exploring the use of whole genome or exome sequencing (WGS, WES) for PGx. However, existing methods have various limitations, which may lead to adverse drug responses. WGS and WES methods may fail to capture or provide adequate sequence coverage for certain PGx loci. Targeted genotyping approaches, such as Taqman (Life Technologies), Luminex (Luminex Corp.) or Sequenom (Agena Bioscience), can fail to detect novel loss-of-function mutations due to their selective interrogation of predefined genomic loci. In a diagnostic clinic, where results are often required within days of administering diagnostic tests, some existing technologies can delay the return of clinical results. Current massively parallel technologies have high capital costs (ranging from $100,000–$1,000,000) requiring clinical laboratories to purchase and maintain large instruments to perform in-house genotyping. Alternatively, a laboratory can send these PGx samples to a third party service for a fee, but may wait up to several months for a clinical report.

Clinical haplotypes, tightly-linked collections of inherited alleles, that are responsible for a plethora of medical phenotypes including patient drug response are important in many medical sequencing applications. Information about haplotypes, or genotype phase, can be inferred from parental genotypes or genetic pattern frequency in the human population, however, these predictions can be inaccurate if *de novo* or rare haplotypes are encountered in a patient[4,5]. Due to the chromosomal distance between alleles, current short read technologies in use for PGx are often unable to resolve haplotype information without supplemental statistical phasing or parental genotypic data.

Recently, Oxford Nanopore Technologies Inc. has developed the MinION, a real time nanopore-based DNA sequencing instrument which is compact, inexpensive and faster than most established DNA sequencing technologies. The time it takes from initiation of library preparation to basecalling the first sequence read is approximately 3 hours and the instrument is capable of detecting long sequence reads in excess of 50kb (according to the manufacturer, ONT). Nanopore-based technology promises major advances in DNA sequencing by offering an inexpensive (e.g. on the order of $1000) pocket-sized device for clinical diagnostics or field experiments.

Here we report nanopore-based sequencing of three clinically relevant PGx genes to identify medically actionable variants and haplotypes without statistical phasing.

Methods

PCR amplification

Primer sequences for *HLA-A*, *HLA-B* and *CYP2D6* are available in Table S1. For *CYP2D6*, we designed primers to specifically amplify *CYP2D6* while not amplifying the 94% identical *CYP2D7*. PCR primer specificity was verified using UCSC *in silico* PCR. We used the standard protocol (for fragments up to 8 kb) of the KAPA LongRange HotStart PCR system: 5× KAPALongRange Buffer (without Mg_2^+) 1×, $MgCl_2$ (25 mM) 1.75 mM, dNTPs (10 mM each dNTP) 0.3 mM, Fwd primer (10 µM) 0.5 µM, Rev primer (10 µM) 0.5 µM 50ng of genomic DNA (1ul of a 50ng/ul preparation), KAPA LongRange HotStart DNA Polymerase (2.5 U/µl) 1.25 U/50 µl, PCR grade water up to 50µl. For *HLA-A* and *HLA-B*, genomic sequence was downloaded from UCSC Browser with common SNPs masked. Primers were designed using Primer3 using parameters of 68°C for optimal annealing temp and 26bp minimum primer length[6]. PCR cycling conditions were: 94°C for 3 mins, followed by 35 cycles of 94°C for 20 sec, 68°C for 15 sec and 68°C for 5 mins followed by a final step of 72°C for 5 mins and hold at 10°C.

Oxford Nanopore genomic DNA library preparation

The DNA libraries were prepared using the Oxford Nanopore Genomic DNA Sequencing protocol (SQK-MAP003). 1.5µg of PCR product was used (instead of the suggested 1µg based on improved yield in earlier testing) with equimolar amounts of *CYP2D6*, *HLA-A* and *HLA-B* amplicons in solution. DNA was not fragmented because the PCR amplicons were already at the desired size for sequencing and downstream haplotyping (4–5Kbp; Table S1).

In accordance with the protocol, we end-repaired the DNA with the NEBNext end repair module (New England Biolabs, cat. no. E6050) and subsequently dA-tailed the sample using the NEBNext dA-tailing module (New England Biolabs cat. no. E6053), prior to ligation of nanopore-specific adapters. All purifications were accomplished with Agencourt AMPure XP beads (Beckman Coulter Inc., cat. no. A63880). Throughout the library preparation, care was taken not to vortex or vigorously pipette/mix the library to avoid shearing the DNA into smaller fragments.

Oxford Nanopore MinION sequencing and basecalling

The MinION flowcell (R7.3 flowcell chemistry) was run for 24 hours using the MinKNOW software (v47.3) producing 24,859 fast5 files, corresponding to individual reads from base detection events at specific nanopore channels. Online basecalling was performed using the Metrichor software (v2.23). The MinION outputs 3 reads for each dsDNA molecule that passes through a pore. The leading ssDNA is referred to as "1D template" and its complementary ssDNA strand is the "1D complement". When both 1D template and 1D complement reads are basecalled, a 2D consensus sequence is determined based on complementarity. All read information was extracted using the python HDF5 package h5py (http://github.com/h5py/h5py). We observed 19,655 1D template reads, 9,584 1D complement reads and 7,540 2D reads. The mean lengths were 2,693bp for 1D template, 2,706bp for 1D complement and 3,486bp for 2D consensus.

Read alignment

Existing massively parallel sequencing instruments, such as the Illumina HiSeq 2500, produce accurate short reads typically up to ~250bp in length. These sequencers can produce hundreds of millions of reads which need to be rapidly aligned to a reference genome. Current computational methods for accurate alignment of these reads, including BWA[7] and Bowtie2[8] are based on the Burrows-Wheeler Transform FM index, and they are designed to align short reads with minimal variation to a reference assembly. BWT-FM methods are insufficiently sensitive to align much longer reads with higher error rates[9]. These long reads, generated by single molecule sequencers such as the Oxford Nanopore MinION or the Pacific Biosciences RS II, have a significantly higher error rate, enriched for insertions or deletions (indels) rather than substitutions[9]. Mapping of these reads is best suited to aligners that were originally designed for whole genome alignments, such as LAST[10]. We chose to use BLASR, originally developed for the Pacific Biosciences system, to align our data, because it was designed to align long error-prone reads rather than genomes[9].

All reads were aligned to the human genome reference assembly GRCh37.p13 (hg19) using default BLASR parameters (gap open penalty = ten, gap extension penalty = zero, minimum seed length = 12). The use of other parameters, such as a gap-open penalty of zero (with default gap extension penalty = zero) did not alter the results, even though it may be expected to do so given the prevalence of indels expected in single molecule nanopore sequencing. The majority of successfully aligned long read fragments were obtained from 2D basecalls (Table 1), and these were of higher quality because they are consensus reads constructed from corresponding 1D template and complement. For the final alignment data, for each separate read event (1D template, 1D complement and 2D consensus), we selected the 2D read if it was available. Since the 1D reads typically had lower mapping accuracy and significantly shorter aligned fragments (Table 1), these were not included in our variant or haplotype calling analysis.

To repeat our findings, we performed the sequencing experiment using the newest Oxford Nanopore Genomic DNA Sequencing protocol (SQK-MAP004). The library was sequenced on 3 fresh R7.3 flow cells, only one of which produced a significant number of reads. The 2D consensus reads generated using the newer protocol were longer on average (3885.3bp compared to 2952.3bp, see Table 1) suggesting advances in basecalling accuracy since our initial experiments. However, yield was significantly lower with a total of 885 aligned reads for all targeted loci, which did not enable us to make accurate genotyping calls.

Finally, we performed two separate alignments depending on our desired sequencing application: a) gene targets with highly similar nearby genes; and b) highly polymorphic gene targets. The first analysis only selected the single best alignment for each long read. This was critical for the gene *CYP2D6* because the *CYP2D* locus on chromosome 22 harbors two paralogous pseudogenes *CYP2D7* and *CYP2D8P*[11]. In particular, *CYP2D6* and *CYP2D7* are highly similar (94% identity, BLAST E-value = 0.0) and are positioned in tandem on the chromosome. By allowing reads to only map to a single best hit, we were able to verify that the PCR selectively amplified *CYP2D6* and not nearby related genes (see coverage in Figure 1). Our second analysis was performed due to the high degree of polymorphism in the MHC locus on chromosome 6. As part of the MHC haplotype project[12], multiple reference contigs for this highly variable region are included in the GRCh37 reference assembly as indicated in the release notes (http://www.ncbi.nlm.nih.gov/genome/guide/human/release_notes.html). Since long reads from the NA12878 *HLA-A* and *HLA-B* genes mapped to different

Table 1. Basecall and read mapping statistics.

Read Type	Number of reads	Mean Length, unaligned (bp)	Number of reads aligned	% of reads aligned	Mean Length, aligned fragment (bp)	Mean substitution frequency	Mean deletion frequency	Mean insertion frequency	Mean mapping accuracy (for each read [match bp/total bp])
1D template	19655	2693.7	3793	19.3%	872.8	8.9%	13.9%	5.7%	71.5%
1D complement	9584	2705.7	2717	28.3%	292.7	7.3%	15.4%	4.1%	73.2%
2D consensus	7540	3486.3	4761	63.1%	2952.3	7.0%	13.3%	5.3%	74.3%

Figure 1. Integrate Genomics Viewer (IGV) diagram of MinION reads aligned to the *CYP2D* locus on chromosome 22 from 42,521,411 to 42,552,401. The majority of reads aligned across the entire length of *CYP2D6* as was expected by selective PCR amplification. Downstream, an insignificant number of read fragments aligned to *CYP2D7* and *CYP2D8* (*2D8* is located from 42,545,874 to 42,551,097; exon-intron diagram not shown in gene annotation track). Due to the extremely high coverage at *CYP2D6*, not all reads are shown in this pileup diagram.

chromosome 6 reference haplotype contigs, by allowing multiple alignments to the reference (up to 10) for each read, we could gather all reads for a single gene in a single pileup to any of the eight *HLA-A/B* loci to generate a consensus sequence. For this study, we used the reference chromosome 6 contig NC_000006.11 (not the MHC haplotype project contigs).

Variant detection and haplotype identification

Due to the long reads, high error rates and continuously evolving error profile of the MinION basecalls at this early stage of technology roll out, variant callers such as the Genome Analysis Toolkit's UnifiedGenotyper or HaplotypeCaller[13] were unable to identify variants or haplotypes in the MinION sequence data during our trials. Variant and haplotype level information, however, was readily accessible based on coverage of aligned reads, which we extracted using SAMTools via the Pysam wrapper (http://github.com/pysam-developers/pysam)[14].

Mean coverage was 1236.4× for *CYP2D6* (single best hit alignment), 785.5× for *HLA-A* (multi-hit alignment) and 1416.3× for *HLA-B* (multi-hit alignment).

Variants were detected using a naïve threshold requiring 1/3 of reads to contain the variant genotype at that position. While this was effective for substitution detection, we are likely to detect many false positive deletions due to the high deletion error rate (Table 1). Haplotype proportions were identified by interrogating clinical marker positions (Table S2) across all reads aligned to a particular gene to establish the proportion of reads corresponding to each haplotype. Pharmacogenomic haplotypes were verified by comparison to diagnostic data sets (see below) using the MedSavant software (www.medsavant.com; manuscript in preparation) with the pharmacogenomics app that we developed. The PGx app interprets human pharmacogenomic variants with medically actionable output based on published guidelines established by the Clinical Pharmacogenetics Implementation Consortium (CPIC) and the Pharmacogenomics Knowledgebase (www.pharmgkb.org).

Validation of genotypes by Complete Genomics, and clinical diagnostic Sequenom MassARRAY and qPCR

Complete Genomics WGS data for NA12878 were obtained from the public 69 genomes project (CG analysis pipeline version 2.0.0; http://www.completegenomics.com/public-data/69-Genomes/)[15].

10ng of genomic DNA from NA12878 was genotyped for 36 SNP, indel and copy number variants for *CYP2D6* using the iPLEX® ADME CYP2D6 Panel v1.0, developed by Assays by Agena (formerly Sequenom) on the MassARRAY4 System. Haplotype assignment and copy number determination was done using Typer software version 4.0 (Agena Biosciences).

In parallel, copy number estimation of *CYP2D6* was performed using the Taqman copy number assays Hs04502391_cn and Hs04083572_cn (Life Technologies) using the manufacturer's

recommended protocol (Figure S1). The assay was performed in quadruplicate on 10ng genomic DNA for each sample in a 96-well plate. The 10µL reaction mix consisted of 5µL 2× Taqman Genotyping Master Mix (Life Technologies), 0.5µL of 20X copy number assay (described above), 0.5µL TaqMan RNAse P Copy Number Reference Assay (Life Technologies cat. no. 4403326), 2µL water and 2µL of 5ng/µL genomic DNA. Cycling conditions for the reaction were 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. Samples were analyzed using the ViiA™ 7 Real-Time PCR System (Life Technologies) and analyzed using CopyCaller Software (Life Technologies). The HuRef sample (http://huref.jcvi.org) was used as a 2-copy calibrator sample.

Validation of haplotypes by statistical phasing and HapMap

Complete Genomics WGS and Sequenom MassARRAY genotypes were statistically phased using the BEAGLE software (v4.0) and the 1000 Genomes Project phase 3 reference panel (http://faculty.washington.edu/browning/beagle/beagle.html)[4]. For the *HLA-A/B* genes, phase information was obtained from the HapMap phase 2 data (http://hapmap.ncbi.nlm.nih.gov/downloads/index.html.en)[16]. *HLA-A/B* alleles were determined using the GATK HLACaller software package (http://gatkforums.broadinstitute.org/discussion/65/hla-caller).

Results

To evaluate the MinION for diagnostic PGx sequencing, we selectively amplified and sequenced the genes *CYP2D6*, *HLA-A*, and *HLA-B* from the CEPH/UTAH pedigree 1463 sample NA12878. *CYP2D6* is a pharmacogenetically vital cytochrome P450 gene because it encodes a protein responsible for metabolism of 20% of clinically used drugs[11]. The diagnostic relevance of *2D6* is derived from its significant polymorphism which contributes to dramatic inter-individual variability in enzyme activity[11]. Also important are the *HLA* genes which are clinically relevant for solid organ transplantation and accurate dosing of abacavir, allopurinol and carbamazepine, used to treat HIV/AIDS, hyperuricemia and seizure disorders, respectively[17–19]. The *HLA* genes are among the most polymorphic loci in the human genome, making their sequencing and confident typing difficult with current short read DNA sequencing methods. These three genes were also chosen for sequencing due to their length (4–5Kbp), which did not require long range PCR amplification methods.

PCR amplicons of these three genes from NA12878 (CEPH/Utah Pedigree 1463) were sequenced on the MinION instrument yielding 19655 read events. Each read event could be basecalled in multiple forms, as a template or complement strand (1D) or as a consensus of the two (2D), and we obtained 36779 1D and 2D reads in total. For the purpose of diagnostic evaluation, we chose to align only the consensus 2D reads due to their lower error rate and extended length (Table 1; see Methods). Reads were aligned to the human genome (GRCh37), with an abundance of aligned reads 4–5Kb in length representing full-length PCR amplicons. As well, smaller aligned read fragments were observed, some of which are speculated to be by-products of shearing during experimental DNA handling (Figure 2A, Table S1).

With depth of coverage of ~1000× for each of the genes, many chromosomal positions were called with 70–90% consensus, demonstrating that as coverage of loci increases on the MinION, confidence improves with regard to specific base calls (Figure 2B). While the MinION basecalls are emitted with a comparatively high error rate (Table 1), the majority of errors appear to be randomly distributed across the length of the reads, which is why increasing coverage can yield a consensus that matches variant calls from existing sequencing and genotyping platforms such as Illumina, Complete Genomics and Sequenom.

MinION-called variants and haplotypes were validated against statistically phased genotypes from multiple platforms including Complete Genomics and Sequenom MassARRAY (see Methods). Based on the statistically phased genotypes, we determined that NA12878 possesses both the *3 and *4 loss-of-function alleles for *CYP2D6*, and this *3/*4 diplotype is interpreted as reduced metabolism of drugs such as codeine (an opiate) and olanzapine (an atypical antipsychotic)[20–22].

CYP2D6 haplotype proportions in MinION data were identified by interrogating clinical marker positions across all aligned reads to establish the proportion of reads corresponding to each PGx haplotype (Table S2). Only reads spanning all clinical markers were included (n = 404), so that haplotypes could be measured by linkage of markers on a single DNA molecule. The MinION data confirmed the statistically-phased haplotypes by direct interrogation of markers from individual reads (Figure 2C). However, we also observed a prominent *2 haplotype, which we could not account for given that our Sequenom MassARRAY and qPCR results indicated that the *CYP2D6* locus was diploid (no copy number variation) and could only correspond to a *3/*4 diplotype. To determine whether the *2 haplotype could arise from mismatched *CYP2D7* DNA, which was not supposed to be PCR amplified (see Methods), we interrogated four positions with different bases between *CYP2D6* and *CYP2D7* reference sequences and found that all reads corresponded to *CYP2D6* (Supplementary File S1). Finally, we hypothesized that the *2 haplotype might arise due to duplexes forming between *3 and *4 oligonucleotides during PCR (effectively outcompeting the primer binding during the annealing step), but this was ruled out by identifying the *2 haplotype using only 1D reads. It is possible that this *2 haplotype arose either due to early cycle template switching during PCR or sample contamination[23]. Also, the relative proportion of haplotypes was potentially skewed by the PCR amplification step.

The *HLA-A* and *HLA-B* haplotypes were determined in the same way as the *CYP2D6* haplotypes, using predefined markers from the HapMap project. In *HLA-A* (only spanning reads, n = 203), the most abundant haplotype matched the transmitted haplotypes of the parents NA12891 and NA12892, which both transmitted an identical haplotype (Figure 2C). Accounting for the errors in MinION sequencing, when allowing for a single mismatch in the haplotype, ~85% of reads confirm the NA12878 diplotype. In *HLA-B* (only spanning reads, n = 202), the majority of reads corresponded to the transmitted and untransmitted haplotypes of the parent NA12891, with only 8.4% of reads corresponding to the

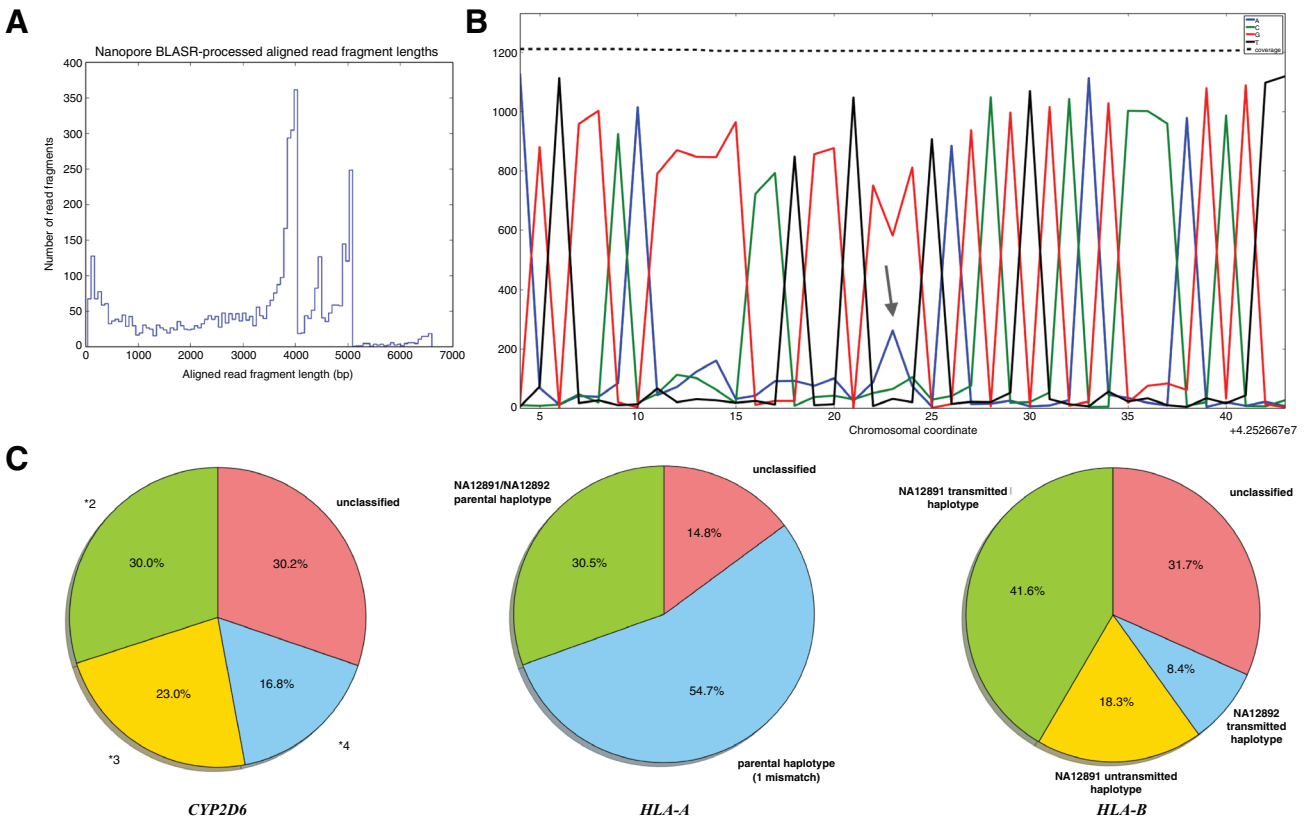


Figure 2. A. Length distribution of aligned reads. 4–5Kb reads represent full-length PCR amplicons. Slightly smaller fragments were likely byproducts of shearing during DNA handling in the experimental protocol. **B**. With depth of coverage of ~1000× for each of the genes, many chromosomal positions were called with 70–90% consensus. This is a short window of aligned reads for the *4 locus of CYP2D6 with over 1200× coverage. The heterozygous *4 allele rs1065852 is indicated with the arrow. **C**. Proportions of haplotypes of *CYP2D6, HLA-A* and *HLA-B* when directly measured from individual reads spanning all haplotype markers.

transmitted haplotype of parent NA12892 (Figure 2C). This could be a result of potential contamination suspected earlier in described above with *CYP2D6*. As suggested for CYP2D6, the relative proportion of HLA haplotypes was likely also affected by PCR bias during amplification. *HLA* alleles were called with 4-digit resolution using the GATK HLACaller, but due to the high error rates of nanopore reads, HLA alleles did not match with alleles called using HapMap data (Table S3)[24].

Nanopore reads and alignments

Data File

http://dx.doi.org/10.6084/m9.figshare.1289717

Conclusions and discussion

Phasing of genotypes is critical to prevent misinterpretation of PGx variants. The importance of correct phasing of PGx genotypes is illustrated with the gene *TPMT*, which plays a critical role in the metabolism of thiopurine, a drug used to treat acute lymphoblastic leukemia. In a recent study[25], an individual was reported to have a *TPMT *3B/*3C* diplotype, based on observed heterozygous genotypes for the rs1142345 and rs1800460 variants, but this was a misinterpretation due to faulty haplotyping and **1/*3A* is the correct diplotype. The rs1800460 variant is present in both **3A* and **3B* haplotypes while the rs1142345 variant is present in both **3A* and **3C* haplotypes. As a result, it was possible for the individual to have a **1/*3A* diplotype or a **3B/*3C* diplotype. Clinically, a **1/*3A* diplotype corresponds to an intermediate metabolizer, requiring a 30–70% reduction in thiopurine dose, while a **3B/*3C* diplotype corresponds to a poor metabolizer with a 90% reduction in dose[2]. An individual who receives a standard dose and is a poor metabolizer can experience fatal toxicity, while a low dose for a normal metabolizer can lead to disease progression. Clinical trials have demonstrated the medical importance of TPMT haplotyping in treatment of myeloid leukemias and non-malignant immunologic disorders[2,26].

Long sequence reads aid haplotype identification by determining which genetic variants are in phase (i.e. on the same DNA strand). If the *TPMT* genotypes from the example above were sequenced using nanopore-based long read technology, the **1/*3A* diplotype would likely be called correctly (note that rs1142345 and rs1800460 are only 8,310bp apart).

While nanopore sequencing with the MinION is demonstrably error-prone in its current stage of development, we assert that this technology holds promise for clinical applications because accurate consensus sequences can be built with sufficient coverage given the high number of reads generated. As well, we have been able to successfully call haplotypes from long reads *de novo* in the absence of parental haplotypes or statistical phasing. The MinION device produced sufficiently long mappable reads to phase all variants in the loci examined. As error rates on the MinION decrease, we can expect to deconvolute these data into more accurate diplotypes with less noise and will be able to measure how much multi-sample multiplexing can be supported by a single run.

According to the CPIC, 63 genes and 132 drugs have guidelines for pharmacogenomic status (http://www.pharmgkb.org/cpic/pairs), and this list is constantly expanding. With increasing guidelines and demands for PGx in the clinic, affordable and rapid nanopore sequencing may hold great utility.

Data availability

figshare: Nanopore reads and alignments, doi: http://dx.doi.org/10.6084/m9.figshare.1289717[27]

Raw nanopore reads and alignment files are available at the NCBI Sequence Read Archive, accession SRP051851 (http://www.ncbi.nlm.nih.gov/sra/?term=SRP051851).

Author contributions

RA, AS and GDB conceived the study. TAP designed and performed the PCR amplification. RA and DT performed the library preparation. RA performed the sequencing and downstream analysis. RA, TAP and GDB drafted the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

R.A. is a member of the Oxford Nanopore Technologies Inc. MinION Access Programme and the MinION instrument and R7.3 flowcells were received free of charge.

Grant information

This study was funded by a Large-scale Applied Project grant from Genome Canada and the Ontario Genomics Institute (grant ID OGI-068).

We confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors thank Stephen Scherer and Peter Ray of the Hospital for Sick Children for contributing diagnostic validation data.

Supplementary material

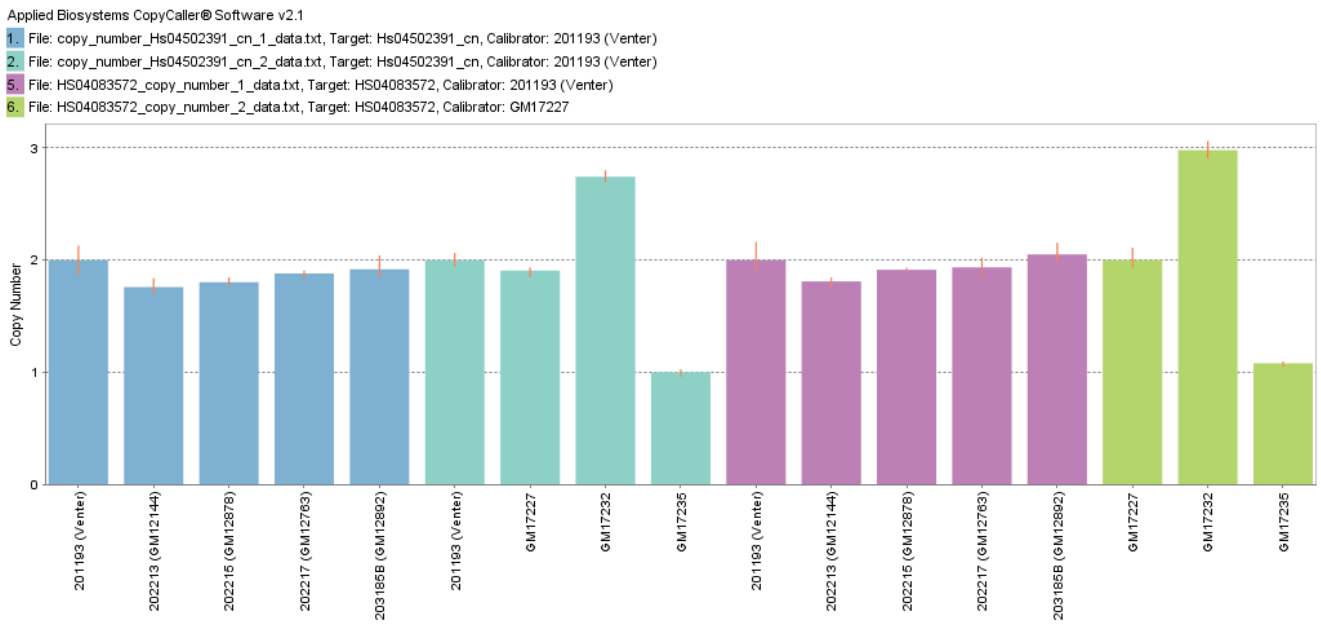


Figure S1 (Updated). CopyCaller software analysis of data from Taqman copy number assays Hs04502391_cn and Hs04083572_cn. *CYP2D6* from sample NA12878 was observed to be diploid. Samples known to be 1 copy or greater than 2 copies for CYP2D6 (GM17235 and GM17232 respectively[28]) were run for the two Taqman copy number assays as well.

Table S1. Primer sequences and amplicon lengths for *HLA-A*, *HLA-B* and *CYP2D6*.

Oxford Nanopore PGx primer list				
Gene	**Primer Name**	**Primer seq**	**Chromosomal Coordinates**	**Amplicon Size**
CYP2D6	CYP2D6-2F	TAGCTCCCTGACGCCATGATTTGTCTT	chr22:42,522,077-42,527,144	5,067 bp
CYP2D6	CYP2D6-2R	CCTGGTTATCCCAGAAGGCTTTGCAG		
HLA-A	HLAA-2F	AGAAGAGTCCAGGTGGACAGGTAAGGAGTG	chr6:29,909,854-29,913,805	3,951 bp
HLA-A	HLAA-2R	TTCTACTGAAGGGCCAAGGACAATGGAG		
HLA-B	HLAB-2F	TGGATTCAGCACCAAGATCACTAGAACCAG	chr6:31,321,279-31,325,303	4,024 bp
HLA-B	HLAB-2R	GTCTCTCCCTGGTTTCCACAGACAGATCCT		

Table S2. Haplotype translation table for *CYP2D6*.

Haplotype Id	CYP2D6	rs1065852	rs28371706	rs5030655	rs3892097	rs35742686	rs5030656	rs16947	rs28371725	rs1135840
PA165816576	*1	G	G	A	C	T	CTT	G	C	C
PA165816577	*2	G	G	A	C	T	CTT	A	C	G
PA165816578	*3	G	G	A	C	-	CTT	G	C	C
PA165816579	*4	A	G	A	T	T	CTT	G	C	G
PA165948092	*5	-	-	-	-	-	-	-	-	-
PA165816581	*6	G	G	-	C	T	CTT	G	C	C
PA165948317	*9	G	G	A	C	T	-	G	C	C
PA165816582	*10	A	G	A	C	T	CTT	G	C	G
PA165816583	*17	G	A	A	C	T	CTT	A	C	G
PA165816584	*41	G	G	A	C	T	CTT	A	T	G

Table S3. *HLA* alleles called with 4-digit resolution using the GATK HLACaller.

Locus	HapMap A1	HapMap A2	Nanopore A1	Nanopore A2
HLA-A	0101	1101	0132	0312
HLA-B	0801	5601	0765	5510

Supplementary File S1

BLAST alignment of *CYP2D6* (bottom sequence) and *CYP2D7* (top sequence). These genes have 94% identity (E-value = 0.0). To determine if our reads were generated from *CYP2D7* reads, we interrogated reads at the paralogous positions circled in red.

BLAST ®

Basic Local Alignment Search Tool

NCBI/ BLAST/ blastn suite-2sequences/ **Formatting Results - 7WAFRJZK114**
Formatting options
Download
Blast report description

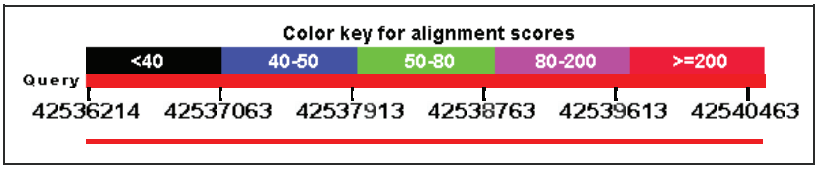
Blast 2 sequences

ref|NC_000022.10| (51304566 letters)

RID	7WAFRJZK114 (Expires on 12-04 04:07 am)		
Query ID	gi|224589814|ref|NC_000022.10|	**Subject ID**	gi|224589814|ref|NC_000022.10|
Description	Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly	**Description**	Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly See details
Molecule type	dna	**Molecule type**	dna
Query Length	51304566	**Subject Length**	51304566
		Program	BLASTN 2.2.30+

Graphic Summary

Distribution of 1 Blast Hits on the Query Sequence



Dot Matrix View

Descriptions

Sequences producing significant alignments:

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly	6582	6582	99%	0.0	94%	NC_000022.10

Alignments

Homo sapiens chromosome 22, GRCh37.p13 Primary Assembly

Sequence ID: **ref|NC_000022.10|** Length: 51304566 Number of Matches: 1
Range 1: 42522501 to 42526883

Score	Expect	Identities	Gaps	Strand	Frame
6582 bits(3564)	0.0()	4142/4408(94%)	91/4408(2%)	Plus/Plus	

Features:

```
Query  42536214  TTGGAACTACCACATTGCTTTATTGTACATTAGAGCCTCTGGCTAGGGAGCAGGCTGGGG  42536273
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522501  TTGGAACTACCACATTGCTTTATTGTACATTAGAGCCTCTGGCTAGGGAGCAGGCTGGGG  42522560

Query  42536274  ACTAGGTACCCCATTCTAGCGGGGCACAGCACAAAGCTCGTAGGGGGATGGGGTCACCAG  42536333
                 |||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  42522561  ACTAGGTACCCCATTCTAGCGGGGCACAGCACAAAGCTCATAGGGGGATGGGGTCACCAG  42522620

Query  42536334  -AAAGCTGACGACACGAGAGTGGCTGGGCCCGGGGCTGTCCGGCGGCCACGGAGAAGCTGA  42536392
                  |||||| | |||||| | ||||||||||||||||||||| | || |||| ||||||||||
Sbjct  42522621  GAAAGC-AAAGACACCATGGTGGCTGGGCCCGGGGCTGTCCAGTGGGCACCGAGAAGCTGA  42522679

Query  42536393  AGTGCTGCAGCAGGGAGGTGAAGAAGAGGAAGAGCTCCATGCGGGCCAGGGGCTCCCCGA  42536452
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522680  AGTGCTGCAGCAGGGAGGTGAAGAAGAGGAAGAGCTCCATGCGGGCCAGGGGCTCCCCGA  42522739

Query  42536453  GGCATGCACGGCGGCCTGTGGGGAGGGGAGGGGCGTCAGTGAGCCTGGCTCCTGGGTGAT  42536512
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522740  GGCATGCACGGCGGCCTGTGGGGAGGGGAGGGGCGTCAGTGAGCCTGGCTCCTGGGTGAT  42522799

Query  42536513  ACCCCTGCAAGACTCCACGGAAGGGGACAGGGAGCCGGGCTCCCCACAGGCACCTGCTGA  42536572
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522800  ACCCCTGCAAGACTCCACGGAAGGGGACAGGGAGCCGGGCTCCCCACAGGCACCTGCTGA  42522859

Query  42536573  GAAAGGCAGGAAGGCCTCCGGCTTCACAAAGTGGCCCTGGGCATCCAGGAAGTGTTCGGG  42536632
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522860  GAAAGGCAGGAAGGCCTCCGGCTTCACAAAGTGGCCCTGGGCATCCAGGAAGTGTTCGGG  42522919

Query  42536633  GTGGAAGCGGAAGGGCTTCTTCCAGACGGCCTCATCCTTCAGCACCGATGACAGGTTGGT  42536692
                 |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  42522920  GTGGAAGCGGAAGGGCTTCTCCCAGACGGCCTCATCCTTCAGCACCGATGACAGGTTGGT  42522979

Query  42536693  GATGAGTGTCGTTCCCTGGGCAGGAGATGCAGGGTGAGAGTGGGGACTGGACTCTAGGAT  42536752
                 |||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  42522980  GATGAGTGTCGTTCCCTGGGCAGAAGATGCAGGGTGAGAGTGGGGACTGGACTCTAGGAT  42523039

Query  42536753  GCTGGGACCCCTGCCACCAAACACACGGGGGACACACACTGCCTGGCACACAGCTGGACT  42536812
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42523040  GCTGGGACCCCTGCCACCAAACACACGGGGGACACACACTGCCTGGCACACAGCTGGACT  42523099

Query  42536813  CTGTCAACTAGTCCTGCGCCCGAGAAGCTCCACAGTACCCTCTCCGACCCCACAGCAGGG  42536872
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  42523100  CTGTCAACTAGTCCTGCGCCCGAGAAGCTCCACAGTACCCTCTCCGACCCCACAGCAGGG  42523159

Query  42536873  CGCAGTCACACCTCTCAGAGGCACCCACACTGCCCCCTCTCCCTGCAGGCGCTGGGTCCT  42536932
                 |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||
Sbjct  42523160  CGCAGTCACACCTCTCAGAGGCACCCACACTGCCCCCTCTCCCTGCAGGTGTTGGGTCCT  42523219

Query  42536933  CCAACATTCTGGCAGGTCCTGATTTGTCTTCCCCACTAGACGGGGGCTCTGGATGGACAG  42536992
                 |||||||||||||||||||| |||||| ||||||||||||||||||||||||||||||||
Sbjct  42523220  CCAACATTCTGGCAGGTCCTGGTTTGTC-TCCCCACTAGACGGGGGCTCTGGATGGACAG  42523278

Query  42536993  GCCAGCCCTGCCTATACTCTGGACCCCCCATCCAAGCGGGGACAGTCAGTGTGGTGGCAT  42537052
                 |||||||||||||||||||| ||||| ||||||||||||||||||||||||||||||||||
Sbjct  42523279  GCCAGCCCTGCCTATACTCTGGACCCCCCACCCAAGTGGGGACAGTCAGTGTGGTGGCAT  42523338
```

```
Query  42537053  TGAGGACTAGGTGGCCAGGGTTCCTAGAGTGGGCCCACCTGGCAGTAGCCATGCTGGGGC  42537112
Sbjct  42523339  TGAGGACTAGGTGGCCAGGGTTCCTAGAGTGGGCCCACCTGGCAGTAGCCATGCTGGGGC  42523398
Query  42537113  TACCACC-GGGGCTGATGCTGAGCTGGGGTGAGGAGGGCGCCAGGCCTACCTTAGGGATG  42537171
Sbjct  42523399  TATCACCAGGGGCTGGTGCTGAGCTGGGGTGAGGAGGGCGCCAGGCCTACCTTAGGGATG  42523458
Query  42537172  CGGAAGCCCTGTACTTCGATGTCATGGGATGTCATATGGGTCACACTCAGGGGGATGATG  42537231
Sbjct  42523459  CGGAAGCCCTGTACTTCGATGTCACGGGATGTCATATGGGTCACACCCAGGGGGACGATG  42523518
Query  42537232  TCCCCAAAGCGCTGCACCTCGTGAATCACGGCAGTGGTGTAGGGCATGTGAGCCTGGTCA  42537291
Sbjct  42523519  TCCCCAAAGCGCTGCACCTCATGAATCACGGCAGTGGTGTAGGGCATGTGAGCCTGGTCA  42523578
Query  42537292  CCCATCTCTGGTCGCCGCACCTGCCCTATCACGTCGTCGATCTCCTGTTGGACACGGACT  42537351
Sbjct  42523579  CCCATCTCTGGTCGCCGCACCTGCCCTATCACGTCGTCGATCTCCTGTTGGACACGGCCT  42523638
Query  42537352  GGACAGACATGCGTCCCCACAATGGGTCAGCACCCAGGGGA-CA--C--T--C--TCCTT  42537402
Sbjct  42523639  GGACAGACATGCGTCCCCACAATGGGTCAGCACCCAGGGGGTCCGGCCCTGACACTCCTT  42523698
Query  42537403  C--G--TCCTGTGTTGGAGGAAGTTAGGCTTACAGGAGCCTGGCCACGCCTGTGCTGGAA  42537458
Sbjct  42523699  CTTGCCTCCTATGTTGGAGGAGGTCAGGCTTACAGGATCCTGGTCAAGCCTGTGCTTGGA  42523758
Query  42537459  GCCCCGGGTGTCCCAGCTAAGCCCAGGGGCCCCCAGCTGTACCCTTCCTCCCTCAGTCCC  42537518
Sbjct  42523759  GCCCCGGGTGTCCCAGCAAAGTTCATGGGCCCCCGCCTGTACCCTTCCTCCCTCGGCCCC  42523818
Query  42537519  TGCCTTGGGCCCCAGCTGGGCTCACGCTGCACATCCAGGTGTAGGATCATGAGCAGGAGG  42537578
Sbjct  42523819  TGCACTGTTTCCCAGATGGGCTCACGCTGCACATCCGGATGTAGGATCATGAGCAGGAGG  42523878
Query  42537579  CCCCAGGCCAGCGTGGTCAAGGTGGTCACCATCCCGGCAAGGAACAGGTTACCCACCACT  42537638
Sbjct  42523879  CCCCAGGCCAGCGTGGTCGAGGTGGTCACCATCCCGGCAGAGAACAGGTCAGCCACCACT  42523938
Query  42537639  ATGCGCAGGTTCTCATCATTGAAGCTGCTCTCAGGGCTCCCCTTGGCCTGAGCAGGGCCG  42537698
Sbjct  42523939  ATGCACAGGTTCTCATCATTGAAGCTGCTCTCAGGGTTCCCCTTGGCCTGAGCAGGGCCG  42523998
Query  42537699  AGAGGATACTCAGGGGATAGAACGGGGTAGCCCCCAAATGACCTCCAATTCTGCACCTGT  42537758
Sbjct  42523999  AGAGCATACTC--GGGACAGAACGGGGTAGCCCCCAAATGACCTCCAATTCTGCACCTGT  42524056
Query  42537759  CAGCCCAGATGCGGCTCGCCGGGTGATGCACTGGTCCAACCTTTTGCCCAGCCTCCCCTC  42537818
Sbjct  42524057  CAGCCCAGATGCGGCTCGCCGGGTGATGCACTGGTCCAACCTTTTGCCCAGCCTCCCCTC  42524116
Query  42537819  ATTCCTCCTGGGACGTTCAACCCACCACCCTTGCCCCCCACCGTGGCAGCCACTCTCACC  42537878
Sbjct  42524117  ATTCCTCCTGGGACGCTCAACCCACCACCCTTGCCCCCCACCGTGGCAGCCACTCTCACC  42524176
Query  42537879  TTCTCCTTCTTTGCCAGGAAGGCCTCAGTCAGGTCTCGGGGTGGCTGGGCTGGGTCCCAG  42537938
Sbjct  42524177  TTCTCCATCTCTGCCAGGAAGGCCTCAGTCAGGTCTCGGGGGGGCTGGGCTGGGTCCCAG  42524236
Query  42537939  GTCATCCTGTGCTCAGTTAGCAGCTCATCCAGCTGGGTCAGGAAAGCCTTTTGGAAGCGT  42537998
Sbjct  42524237  GTCATCCTGTGCTCAGTTAGCAGCTCATCCAGCTGGGTCAGGAAAGCCTTTTGGAAGCGT  42524296
Query  42537999  AGGACCTTGCCAGCCAGCGCTGGGATGTGCAGGAGGACGGGGACAGCATTCAGCACCTAC  42538058
Sbjct  42524297  AGGACCTTGCCAGCCAGCGCTGGGATATGCAGGAGGACGGGGACAGCATTCAGCACCTAC  42524356
Query  42538059  ACCAGACAGAACGGGGTCTCAATCCCTCCTGTGCTCTGCGTTCATCTGGACCAGTCTCAG  42538118
Sbjct  42524357  ACCAGACAGAACGGGGTCTCAATCCCTCCTGTGCTCTGCGTTCACCTGGACAAGTCTCAG  42524416
Query  42538119  GCCCCAGCCATCTCCAGGAAGACCCAGGGCCTGCCTGTCCTTACCACTGACCTCACCAAG  42538178
Sbjct  42524417  GCCCCAGCCATCTCCAGGTAGACCCAGGGCCTGCCTGTCCTTACCACTGACCTCACCAAG  42524476
Query  42538179  TCCCTCCCCAAGTGCCAGCCTCCACCCTCTCTCTCCTTGCCCAGAGGAGAGACCTAAAAT  42538238
Sbjct  42524477  TCCCTCCCCAAGTGCCAGCCTCCACC--CTCTCTCCTTGCCCAGAGGAGAAACCTAAAAT  42524534
Query  42538239  CGAAATCTCCAACGTGGACGGG-GGTACAGAGTCCTTGGCCTCTCCTGGTGCCCCCTGAC  42538297
Sbjct  42524535  CGAAATCTCTGACGTGGATAGGAGGTACAGAGTCCTTGGCCTCTCCTGGTGCCCCCTGAC  42524594
Query  42538298  CCGGGCACACCTCTCCCACGACCATGTCTGAGATGTCCCCTCCTCCTCCAGGCCCTTCTT  42538357
Sbjct  42524595  CCGGGCACACCTCTCCCACGACCATGTCTGAGATGTCCCCTCCTCCTCCAGGCCCTTCTT  42524654
Query  42538358  ACAGTGGGGTCTCCTGGAATGTCCTTTCCCAAACCCATCTACGCAAATCCTGCCCTTCGG  42538417
Sbjct  42524655  ACAGTGGGGTCTCCTGGAATGTCCTTTCCCAAACCCATCTATGCAAATCCTGCTCTTCCG  42524714
Query  42538418  AGGCCCCAGTCCAGCCCCGGCACCTCTCAGGAGCTCGCCCTGCAAAGAC-CCT---T---  42538470
Sbjct  42524715  AGGCCCCAGTCCAGCCCCGGCACCTCTCGGGAGCTCGCCCTGCAGAGACTCCTCGGTCTC  42524774
Query  42538471  --GCTCCGCACCTCGCGCAGGAAGCCCGACTCCTCCTTCAGTCCCTCCTGAGCTAGGTCC  42538528
Sbjct  42524775  TCGCTCCGCACCTCGCGCAGAAAGCCCGACTCCTCCTTCAGTCCCTCCTGAGCTAGGTCC  42524834
```

```
Query  42538529  AGCAGCCTGAGGAAGCGAGGGTCGTCGTACTCGAAGCGGCGCCCGCAGGTGAGGGAGGCG  42538588
Sbjct  42524835  AGCAGCCTGAGGAAGCGAGGGTCGTCGTACTCGAAGCGGCGCCCGCAGGTGAGGGAGGCG  42524894
Query  42538589  ATCACGTTGCTCACGGCTTTGTCCAAGAGACCGTTGGGGCGAAAGGGGCGTCCTGGGGGT  42538648
Sbjct  42524895  ATCACGTTGCTCACGGCTTTGTCCAAGAGACCGTTGGGGCGAAAGGGGCGTCCTGGGGGT  42524954
Query  42538649  GGGAGATGCGGGTAAGGGGTCGCCTTCTCCGTCCCCCGCCTTCCCAGTTCCCGCTGTGTG  42538708
Sbjct  42524955  GGGAGATGCGGGTAAGGGGTCGCCTTCCCCGTCCCCCGCCTTCCCAGTTCCCGCTTTGTG  42525014
Query  42538709  CCCTTCTGCCCATCACCCACCGGCTTGGTCGGCGAAGGCGGCACAAAGGCAGGCGGCCTC  42538768
Sbjct  42525015  CCCTTCTGCCCATCACCCACCGGAGTGGTTGGCGAAGGCGGCACAAAGGCAGGCGGCCTC  42525074
Query  42538769  CTCGGTCACCCACTGCTCCAGCGACTTCTTGCCCAGGCCCAAGTTGCGCAAGGTGGACAC  42538828
Sbjct  42525075  CTCGGTCACCCACTGCTCCAGCGACTTCTTGCCCAGGCCCAAGTTGCGCAAGGTGGAGAC  42525134
Query  42538829  GGAGAAGCGCCTCTGCTCGCGCCACGCGGGCCCATAGCGCGACAGGAACACCCCTGGGGG  42538888
Sbjct  42525135  GGAGAAGCGCCTCTGCTCGCGCCACGCGGGCCCATAGCGCGCCAGGAACACCCCTGGGGG  42525194
Query  42538889  CGGGACGGACACGTGGGCGTTGCCATGAAGGCCTTGGCCCCACCCTCCGCCACCCACTCC  42538948
Sbjct  42525195  TGGGACGGGCACGTGCGCGTGGCCATGAAGGCATTAGCCCCACCATCCACCACCCACTCC  42525254
Query  42538949  AACCCTGGCGCTCCACAAGGTCTCCCGCAGTCCCTAGCCCGGTCCAGCTGGGCACAGGGC  42539008
Sbjct  42525255  AACCCTAT-GCTCCCCCTGGTCTCCCGCAGTCCCTGGCTCTGTCCAGCTGGTCACAGGGC  42525313
Query  42539009  CCACTCTTTGCTCACCCACATTGCTCCCCTGCCTGGGGCGGGGTTTGGCCCCACCTCGTC  42539068
Sbjct  42525314  CCACTCTTTGTGCATCCACCTTGCTCCCTTGGCTGGGGCAGGGCTTTGCCCCACCTCGTC  42525373
Query  42539069  TCTGCCCACCCTGACCACCTTTCCACTCAAGGAAGATCCCGCCCGTCCCG-----CCCAC  42539123
Sbjct  42525374  TCTGCCCACCCTGACCGCCTTTGCACTCAGGGAAGACCCCGCGGGCCCCGCGCCACCCAC  42525433
Query  42539124  ACTGAGCCCGCAGCATAGGCGCGGTCCCCGCCACCGCCACTTCGACGC---AT-C-AGCC  42539178
Sbjct  42525434  ACTGAGCTTACAGCACAGGTGCGGTCCCCGCC-CC-CCACTTCGACACCGGATTCCAGCT  42525491
Query  42539179  ------T-CGCC--C---ACC----GGGCTTCTGGCGGGTCTGGGCAGTAGCCCCGCCCC  42539222
Sbjct  42525492  GGGAAATGCGCCAGCCTCACCCATTGGGCTCCTGCCAGGTCTCGGCAGTGGCCCCGCCCA  42525551
Query  42539223  CTC--C-CA-GCCCA-CA---G-----A--CTCGCACCTCCCCCGTGCAGGTGGTTTCCT  42539267
Sbjct  42525552  CTCGTCACAAGCCCCGCCCTCGTCCCCATGCTCACACCTCCCTAGTGCAGGTGGTTTCTT  42525611
Query  42539268  GGCCCACTGTCCTCAGCCCACTCGCTGGCCTTTATCTCTGTTTCACGTCCAGGACCCCAC  42539327
Sbjct  42525612  GGCCCGCTGTC-----CCCACTCGCTGGC--------CTGTTTCATGTCCACGACCCCGC  42525658
Query  42539328  GCCCTGTCGGCGCTGCTTGGGCTACGGTCACTGTCCACCCGGGGCCCACGGAAACGCGGT  42539387
Sbjct  42525659  GCCCTCTCTGCCCAGCTCGGACTACGGTCATCACCCACCCGGGTCCCACGGAAAT-CTGT  42525717
Query  42539388  CTCTGTCCCCCACCGCCGCTTGCCTTGGGAACGCGGCCCGAAGCCCAGGACCTGGTAGAT  42539447
Sbjct  42525718  CTCTGT-CCCCACCGCTGCTTGCCTTGGGAACGCGGCCCGAAACCCAGGATCTGGGTGAT  42525776
Query  42539448  GGCCGCAGGCGGGCGGTCGGCCGTGTCCTCGCCGCGGGTCACCATCGCCTCGCGCACGGC  42539507
Sbjct  42525777  GGCCACAGGCGGGCGGTCGGCGGTGTCCTCGCCGTGGGTCACCAGCGCCTCGCGCACGGC  42525836
Query  42539508  CGCCAGCCCATTGAGCACGACCACCGGCGTCCAGGCCAGCTGCAGGCTGAACACGTCCCC  42539567
Sbjct  42525837  CGCCAGCCCATTGAGCACGACCACCGGCGTCCAGGCCAGCTGCAGGCTGAACACGTCCCC  42525896
Query  42539568  GAAGCGGCGCCGCAACTGCAGAGGGAGGGTCAGGGCCTCTTGTCAAGCCAGGATCCCCCC  42539627
Sbjct  42525897  GAAGCGGCGCCGCAACTGCAGAGGGAGGGTCAGGGCCTCTTGTCAAGCCAGGATCCCCCC  42525956
Query  42539628  AGACTACAGGTCCTAGTCCTATTTGAACCTTGGACGACCCCCGGGGCTACCAGGAGTGAG  42539687
Sbjct  42525957  AGACTACAGGTCCTAGTCCTATTTGAACCTTGGACGACCCCCGGGGCTACCAGGAGTGAG  42526016
Query  42539688  CAGGTGGAAGGAGGAGACCCAGCCTCCTGATCCTggggcgggggtgggggTCACACCTTC  42539747
Sbjct  42526017  CAGGTGGAAGGAGGAGACCCAGCCTCCTGATCCTGGGGCGGGGGTGGGGGTCACACCTTC  42526076
Query  42539748  TGTGATGGAGGAACTCAGTTTGGATGCGTCACCCAGGTATGACCTTGCAAGAGTCACCAA  42539807
Sbjct  42526077  TGTGATGGAGGAACTCAGTTTGGATGCGTCACCCAGGTATGACCTTGCAAGAGTCACCAA  42526136
Query  42539808  AATTGCCGAGAGGCCCCAGTTAGCATCCCATTCCCAGATGATGGTCCATGCCGGTGAGCA  42539867
Sbjct  42526137  AATTGCCGAGAGGCCCCAGTTAGCATCCCATTCCCAGATGATGGTCCATGCCGGTGAGCA  42526196
Query  42539868  GTGAGGCCCGAGGACCCACAGTGCAAAAGGTTTGAACCGGGTCACTGCACCCCCTTCATC  42539927
Sbjct  42526197  GTGAGGCCCGAGGACCCACAGTGCAAAAGGTTTGAACCGGGTCACTGCACCCCCTTCATC  42526256
Query  42539928  CTCGATTTCGTGATTTAAACGGCACTCAGGACTAACTCATCTTCCATTCCCAAGGCCTTT  42539987
Sbjct  42526257  CTCGATTTCGTGATTTAAACGGCACTCAGGACTAACTCATCTTCCATTCCCAAGGCCTTT  42526316
Query  42539988  CCTTCTGGTGTCAGCAGAAGGGACTTTGTACTCCATAACATATGTTGCCCAATGGGCTTG  42540047
```

```
Sbjct  42526317  CCTTCTGGTGTCAGCAGAAGGGACTTTGTACTCCATAACATATGTTGCCCAATGGGCTTG  42526376
Query  42540048  CATGCCCACTGCCAAGTCCAGCTCCACCTCCAGGCCCTTGCCCTACTCTTCCTTGGCCTT  42540107
Sbjct  42526377  CATGCCCACTGCCAAGTCCAGCTCCACCTCCAGGCCCTTGCCCTACTCTTCCTTGGCCTT  42526436
Query  42540108  TGGAAAATCCAGTCCTTCATGCCATGTATAAATGTCCTTCCCCAGGACGTCCCCCAAACC  42540167
Sbjct  42526437  TGGAAAATCCAGTCCTTCATGCCATGTATAAATGCCCTTCTCCAGGAAGTCCCCCAAACC  42526496
Query  42540168  TGCTTCCCCTTCTCAGCCTGGCTTCTGATCCAGCCTGTGGTTTAACCCACCACCCATGTT  42540227
Sbjct  42526497  TGCTTCCCCTTCTCAGCCTGGCTTCTGGTCCAGCCTGTGGTTTCACCCACCACCCATGTT  42526556
Query  42540228  TGCTGGTGGTGGGGCATCCTCAGGACCTCTGCCGCCCTCCAGGACCTCCTCCCTCACCTG  42540287
Sbjct  42526557  TGCTGGTGGTGGGGCATCCTCAGGACCTCTGCCGCCCTCCAGGACCTCCTCCCTCACCTG  42526616
Query  42540288  GTCGAAGCAGTATGGTGTGTTCTGGAAGTCCACATGCAGCAAGGTTGCCCAGCCCGGGCA  42540347
Sbjct  42526617  GTCGAAGCAGTATGGTGTGTTCTGGAAGTCCACATGCAGC-AGGTTGCCCAGCCCGGGCA  42526675
Query  42540348  GTGGCAGGGGACCTGGCGGGTAGCGTGCAGCCCAGCGTTGGTGCCGGTGCATCAGGTCCA  42540407
Sbjct  42526676  GTGGCAGGGGGCCTGGTGGGTAGCGTGCAGCCCAGCGTTGGCGCCGGTGCATCAGGTCCA  42526735
Query  42540408  CCAGGAGCAGGAAGATGGCCACTATCATGGCCAGGGGCACCAGTGCTTCTAGCCCCATGG  42540467
Sbjct  42526736  CCAGGAGCAGGAAGATGGCCACTATCACGGCCAGGGGCACCAGTGCTTCTAGCCCCATAC  42526795
Query  42540468  CTGCCTCACTACCAACTGGGCTCCTCTGGACACACCTGGCACCCCCACCCCACCAGGCAC  42540527
Sbjct  42526796  CTGCCTCACTACCAAATGGGCTCCTCTGGACACACCTGGCACCCCCACCCCACCAGGCAC  42526855
Query  42540528  AGAGGACCAGGCAGGACACTCTCAGCAC  42540555
Sbjct  42526856  AGAGGACCAGGCAGGACACTCTCAGCAC  42526883
```

Supplementary File S2

Detailed responses to reviewers' comments.

Click here to access the data

References

1. Evans WE, Hon YY, Bomgaars L, *et al.*: **Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine.** *J Clin Oncol.* 2001; **19**(8): 2293–2301.
PubMed Abstract
2. Relling MV, Gardner EE, Sandborn WJ, *et al.*: **Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing.** *Clin Pharmacol Ther.* 2011; **89**(3): 387–391.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Mizzi C, Peters B, Mitropoulou C, *et al.*: **Personalized pharmacogenomics profiling using whole-genome sequencing.** *Pharmacogenomics.* 2014; **15**(9): 1223–1234.
PubMed Abstract | Publisher Full Text
4. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet.* 2007; **81**(5): 1084–1097.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Browning SR, Browning BL: **Haplotype phasing: existing methods and new developments.** *Nat Rev Genet.* 2011; **12**(10): 703–714.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Untergasser A, Cutcutache I, Koressaar T, *et al.*: **Primer3--new capabilities and interfaces.** *Nucleic Acids Res.* 2012; **40**(15): e115.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2009; **25**(14): 1754–1760.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Chaisson MJ, Tesler G: **Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory.** *BMC Bioinformatics.* 2012; **13**: 238.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Kiełbasa SM, Wan R, Sato K, *et al.*: **Adaptive seeds tame genomic sequence comparison.** *Genome Res.* 2011; **21**(3): 487–493.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Zanger UM, Schwab M: **Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation.** *Pharmacol Ther.* 2013; **138**(1): 103–41.
PubMed Abstract | Publisher Full Text
12. Horton R, Gibson R, Coggill P, *et al.*: **Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project.** *Immunogenetics.* 2008; **60**(1): 1–18.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Van der Auwera GA, Carneiro MO, Hartl C, *et al.*: **From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline.** *Curr Protoc Bioinformatics.* 2013; **11**(1110): 11.10.1–11.10.33.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Drmanac R, Sparks AB, Callow MJ, *et al.*: **Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.** *Science.* 2010; **327**(5961): 78–81.
PubMed Abstract | Publisher Full Text
16. Frazer KA, Ballinger DG, Cox DR, *et al.*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature.* 2007; **449**(7164): 851–861.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Leckband SG, Kelsoe JR, Dunnenberger HM, *et al.*: **Clinical Pharmacogenetics Implementation Consortium guidelines for HLA-B genotype and carbamazepine dosing.** *Clin Pharmacol Ther.* 2013; **94**(3): 324–8.
PubMed Abstract | Free Full Text
18. Martin MA, Hoffman JM, Freimuth RR, *et al.*: **Clinical Pharmacogenetics Implementation Consortium Guidelines for HLA-B Genotype and Abacavir Dosing: 2014 update.** *Clin Pharmacol Ther.* 2014; **95**(5): 499–500.
PubMed Abstract | Free Full Text
19. Hershfield MS, Callaghan JT, Tassaneeyakul W, *et al.*: **Clinical Pharmacogenetics Implementation Consortium guidelines for human leukocyte antigen-B genotype and allopurinol dosing.** *Clin Pharmacol Ther.* 2013; **93**(2): 153–8.
PubMed Abstract | Free Full Text
20. Zhou SF: **Polymorphism of human cytochrome P450 2D6 and its clinical significance: part II.** *Clin Pharmacokinet.* 2009; **48**(12): 761–804.
PubMed Abstract | Publisher Full Text
21. Crews KR, Gaedigk A, Dunnenberger HM, *et al.*: **Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update.** *Clin Pharmacol Ther.* 2014; **95**(4): 376–82.
PubMed Abstract | Free Full Text
22. Hicks JK, Swen JJ, Thorn CF, *et al.*: **Clinical Pharmacogenetics Implementation Consortium guideline for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants.** *Clin Pharmacol Ther.* 2013; **93**(5): 402–8.
PubMed Abstract | Free Full Text
23. Odelberg SJ, Weiss RB, Hata A, *et al.*: **Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I.** *Nucleic Acids Res.* 1995; **23**(11): 2049–2057.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Listgarten J, Brumme Z, Kadie C, *et al.*: **Statistical resolution of ambiguous HLA typing data.** *PLoS Comput Biol.* 2008; **4**(2): e1000016.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Brownstein CA, Margulies DM, Manzi SF: **Misinterpretation of *TPMT* by a DTC genetic testing company.** *Clin Pharmacol Ther.* 2014; **95**(6): 598–600.
PubMed Abstract | Publisher Full Text
26. Stanulla M, Schaeffeler E, Flohr T, *et al.*: **Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia.** *JAMA.* 2005; **293**(12): 1485–1489.
PubMed Abstract | Publisher Full Text
27. Ammar R, Paton TA, Torti D, *et al.*: **Nanopore reads and alignments.** *figshare.* 2015.
Data Source
28. Pratt VM, Zehnbauer B, Wilson JA, *et al.*: **Characterization of 107 genomic DNA reference materials for *CYP2D6, CYP2C19, CYP2C9, VKORC1*, and *UGT1A1*: a GeT-RM and Association for Molecular Pathology collaborative project.** *J Mol Diagn.* 2010; **12**(6): 835–846.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 21 May 2015

doi:10.5256/f1000research.7015.r8715

Martin Kennedy
Gene Structure and Function Laboratory, Department of Pathology, University of Otago, Christchurch, New Zealand

Thanks for attending to these points. I am comfortable with the amendments that have been made.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 30 March 2015

doi:10.5256/f1000research.6463.r7919

Thomas Hoenen
Laboratory of Virology, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA

This manuscript describes the analysis of three PCR-fragments generated from a human reference sample using a MinION sequencing device. Despite the preliminary nature of the data, and the fact that the paper is based on a single run of a single sample, it will nevertheless be of significant interest, and publication as a research note is in my view justified. In fact, given the broad interest in this emerging new technology, and its applicability to all life sciences (not just human (pharmaco-)genetics), the readership might be much wider than anticipated by the authors. This constitutes at the same time the biggest weakness of the paper, since it is in parts rather inaccessible for readers from fields other than human genetics (e.g. virology or microbiology), for whom this paper might nevertheless be highly relevant.

In order to address the concern that the manuscript is based on a single MinION run, and given the nature of the MAP program, it should be no problem for the authors to repeat the experiment with another sample, and while doing so specifically address the concerns of Dr. Kennedy. In addition, the authors should strive to improve accessibility to a wider audience wherever possible. Finally, it would be helpful to include additional experimental details that are currently missing.

Specific comments:

- PCR cycling conditions should be provided.
- The authors refer to the SQK-MAP003 sequencing protocol. As far as I am aware, Oxford Nanotechnologies does not make the detailed protocol available to people outside the MAP program, although there have been indications that a non-technical version of the protocol will be made available for publication purposes. The authors should reference such a protocol (including a link) as soon as possible, and approach ONT about making it available to the general public, if this hasn't happened already.
- Did the authors perform a PCR purification prior to the library preparation? If so, what was the volume/ratio of Agencourt beads to sample?
- How did the authors extract reads from the fast5 files? Did they use poretools, or another tool (which should be referenced)?
- In general, providing more details about the exact bioinformatics workflow would be helpful.
- What was the rationale for the cut-off of 1/3 of reads for variant calling?
- It seems odd that the length of the aligned fragments from the 1D reads is so much shorter than the read length. In our hands (using a similar approach on ~2 kB PCR products amplified from virus genomes, albeit with a later chemistry/protocol version (SQK-MAP004) and using LAST for the alignment) we get much longer average alignments (92% for 2D reads, 82% for template reads, and 85% for complement reads, vs. 85%, 23% and 11% reported by the authors). This could either indicate significant advances in base-calling accuracy since the authors performed their experiments, or that their alignment is suboptimal. It might be very interesting to see whether the authors can get longer alignments using LAST or other alignment softwares in their workflow.
- It would be very helpful if the authors could repeat the experiment using the newest chemistry/protocol/software versions, which have changed considerably over the last months. At the same time this would allow them to address many of the concerns of Dr. Kennedy.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I am a participant in the MinION access program. Other than that, I have no connections to the authors, and I do not know them. I do not have any competing interests.

Referee Report 06 February 2015

doi:10.5256/f1000research.6463.r7404

Martin Kennedy
Gene Structure and Function Laboratory, Department of Pathology, University of Otago, Christchurch, New Zealand

General comments:

This research note describes preliminary results from the application of a new nanopore sequencing device (the MinIon), under development by Oxford Nanopore Technologies, to the analysis of three amplicons of pharmacogenetic interest.

The paper is suitable for a research note. Although the findings are very preliminary, this is a new technology and there is a lot of interest in understanding its current capabilities, longer-term potential and

limitations.
Positive aspects of the report are description of long-read nanopore sequencing on long amplicons (4-5kb), and description of data handling and bioinformatics approaches this team is using, which may aid others working with the MinIon device. Negative aspects of the report are that all the data are from only one reference sample, and one run on the device; rather than being utterly convincing, the data suggests haplotyping on unknown samples may be possible only once error rates on the MinIon reduce; and there is an unresolved question about the CYP2D6 CNV or haplotype analyses, potentially due to sample contamination.

Despite these issues, the report is still of merit and will be of interest to many in the field.

Specific comments:

Methods
Long PCR has been widely used for specific amplification of CYP2D6, with reaction conditions and primer sequences well established. It is not clear why the authors designed their own primers for this task, or how these novel primers were validated. This should be spelt out more clearly.

It would have been useful to have non-diploid control samples for the CYP2D6 CNV assay – for example a haploid (CYP2D6*5) case or multicopy case, to provide confidence that the assay was working as expected. This is relevant because of the question raised by the *2 haplotype which shows up in the MinIon analyis of NA12878.

Results
Page 6, first para: This description of the possible origins of the mystery CYP2D6*2 haplotype needs some editing for improved clarity. Not clear what is meant by "*3 and *4 duplexes forming". Also not clear what would cause PCR biases alluded to in the last sentence of this paragraph (and of the following paragraph).

Typos/suggested edits:

Abstract
MinIOn > MinIon
Should refer to NA12878 as "reference sample" rather than just "sample"
Suggest sentence be modified thus for clarity: "…statistically phased genotype data from Complete Genomics and Sequenom."
Suggest delete "Standalone" in penultimate line.

Introduction
Suggest this sentence be changed: However, existing methods have various limitations, which may lead to adverse drug responses. > However, existing methods have various limitations, which may lead to failure to detect variants of pharmacogenetic significance.

Methods
Page 3, 2nd para: indels expects > indels expected
Page 4 para 2 – clarify "The HuRef sample…"
Figure 1 title: Integrate > Integrative

Results

Page 5, para 4. First sentence should read thus, for improved clarity: “CYP2D6 haplotype proportions in MinIon data were identified by interrogating clinical marker positions…”

Supp File S1
It would be helpful to indicate which of the sequences is CYP2D6 and which is CYP2D7.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: I am also a participant in the MinIon Access Programme, and my laboratory is working with the device. I met Dr Ammar at a meeting organised by ONT, and we have a shared interest in use of this technology for pharmacogenetic analysis. However, I have not been involved in the work described here.