Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing

NATHAN R. CAMPBELL, STEPHANIE A. HARMON and SHAWN R. NARUM
Columbia River Inter-Tribal Fish Commission, 3059F National Fish Hatchery Road, Hagerman, ID 83332, USA

Abstract
Genotyping-in-Thousands by sequencing (GT-seq) is a method that uses next-generation sequencing of multiplexed PCR products to generate genotypes from relatively small panels (50–500) of targeted single-nucleotide polymorphisms (SNPs) for thousands of individuals in a single Illumina HiSeq lane. This method uses only unlabelled oligos and PCR master mix in two thermal cycling steps for amplification of targeted SNP loci. During this process, sequencing adapters and dual barcode sequence tags are incorporated into the amplicons enabling thousands of individuals to be pooled into a single sequencing library. Post sequencing, reads from individual samples are split into individual files using their unique combination of barcode sequences. Genotyping is performed with a simple perl script which counts amplicon-specific sequences for each allele, and allele ratios are used to determine the genotypes. We demonstrate this technique by genotyping 2068 individual steelhead trout (Oncorhynchus mykiss) samples with a set of 192 SNP markers in a single library sequenced in a single Illumina HiSeq lane. Genotype data were 99.9% concordant to previously collected TaqMan™ genotypes at the same 192 loci, but call rates were slightly lower with GT-seq (96.4%) relative to Taqman (99.0%). Of the 192 SNPs, 187 were genotyped in ≥90% of the individual samples and only 3 SNPs were genotyped in <70% of samples. This study demonstrates amplicon sequencing with GT-seq greatly reduces the cost of genotyping hundreds of targeted SNPs relative to existing methods by utilizing a simple library preparation method and massive efficiency of scale.

Keywords: amplicon sequencing, genotyping by sequencing, GT-seq, next-generation sequencing, SNP

Received 16 July 2014; revision received 24 November 2014; accepted 26 November 2014

Introduction
Single-nucleotide polymorphism (SNP) markers are used for a variety of research applications in the fields of ecological, conservation and agricultural genetics. This includes population genomics, genomewide association studies (GWAS), parentage analysis, and population identification (Narum et al. 2013). Relatively new genotyping techniques using next-generation sequencing (NGS) platforms now make it possible to both identify and genotype thousands to millions of SNPs directly from sequencing data. Methods such as restriction-site-associated DNA sequencing (RAD-seq; Baird et al. 2008), RNA sequencing and whole-genome shotgun can all be used for genotyping (Davey et al. 2011). The cost per genotype collected using these NGS methods is low due to the large number of SNP genotypes that can be generated. Despite the ever decreasing costs of NGS, these genotyping-by-sequencing (GBS) methods remain less cost effective than alternative 5’ exonuclease methods for applications that require only a few hundred SNP markers.

Historically, SNP markers were first identified through sequencing and then selected SNPs were subsequently converted into allele detection assays such as the 5’ exonuclease assay [examples include TaqMan™ (Life Technologies), KASPar (KBiosciences) and SNP type (Fluidigm)] for genotyping (Seeb et al. 2011). High-quality genotypes could then be produced for large numbers of individuals from DNA extracts of varying quality and quantity using such PCR-based genotyping methods (e.g. Campbell & Narum 2008). All 5’ exonuclease methods require a unique reaction for each combination of DNA sample and assay (locus-specific primers mixed with allele-specific fluorescent probes) to produce each SNP genotype. Therefore, the cost per SNP genotype is relatively high compared to high-density array genotyping (examples: Affymetrix SNP arrays, Illumina genotyping arrays). However, when the number of SNP loci to
be genotyped is low (50–1000) and the number of samples is high (hundreds to thousands), the 5’ exonuclease methods remain cost effective and widely used for SNP genotyping (e.g. Matala et al. 2014 with 9011 individual samples genotyped at 188 SNPs). Despite the utility of the 5’ exonuclease method, there is substantial potential to use GBS methods with NGS instruments to generate SNP genotypes at target amplicons for 50–1000 SNP markers.

The use of two or more primer pairs in a single PCR is known as multiplex PCR and has been used to improve efficiency in both microsatellite and SNP genotyping (Hayden et al. 2008). In the case of SNP genotyping, multiplex PCR is used to boost the number of template copies for subsequent genotyping assays (Schunter et al. 2014). However, these amplicons could be sequenced and genotypes may be determined directly from the sequencing data eliminating the need to perform 5’ exonuclease reactions. Sequencing platforms such as the Illumina HiSeq are capable of producing hundreds of millions of sequencing reads making genotyping with this approach highly cost effective given the coverage needed to accurately genotype SNP loci with GBS methods. Indeed, recent publications have demonstrated the feasibility of genotyping using NGS sequencing of multiplex PCR amplicons, although using only 9–11 loci per PCR (Wielstra et al. 2014; Ziebiński et al. 2014). The complexity of multiplex PCR has been an obstacle to further advances in amplicon sequencing with larger numbers of loci; however, highly multiplexed PCR panels such as the AmpliSeq™ Cancer Hotspot Panel v2 (207 amplicons) and the Illumina TruSeq Amplicon Cancer Panel (48 amplicons) are now commercially available (Sie et al. 2014; Tsongalis et al. 2014). Alternatively, sequence capture (also known as sequence hybridization) methods also offer the ability to enrich and sequence targeted loci for the purposes of genotyping (Rohland & Reich 2012). While both of these approaches have shown to be effective at small scales (up to 100 samples per library), neither approach has yet been demonstrated on the scale of hundreds of SNPs for thousands of individuals sequenced within a single Illumina HiSeq lane.

Here, we present a major advance in amplicon sequencing called Genotyping-in-Thousands by sequencing (GT-seq); a method of targeted SNP genotyping by multiplexed PCR amplicon sequencing. This method uses only unlabelled oligos and PCR master mix in two thermal cycling reactions for amplification of hundreds of targeted SNP loci, addition of sequencing adapters and dual barcode tags, enabling thousands of individual samples to be sequenced in a single Illumina HiSeq lane. Post sequencing, a simple bioinformatics pipeline is used to split sequencing reads by sample, count locus-specific allele sequences and generate genotypes. Tools to analyse library quality and identify the problems in the multiplex PCR are also presented. We demonstrate this technique by genotyping 2068 individual steelhead trout (Oncorhynchus mykiss) samples with a set of 192 SNP markers in a single library and compare genotype data to previously collected TaqMan™ genotypes at the same loci and samples. Our results demonstrate this technique is highly feasible as a low cost alternative to genotyping with 5’ exonuclease methods.

Methods

Primer design

As sequence regions with known polymorphisms are intended targets for amplicon sequencing, we utilized existing forward and reverse primer sequences from previous TaqMan™ SNP genotyping assays to target 192 specific loci (Matala et al. 2014). Existing primers were modified by adding Illumina sequencing primer tags, specifically the ‘small RNA sequencing primer’ and ‘read 2 sequencing primer’ for use in PCR1 (Fig. 1). These primers ranged from 50 to 74 bases in length and were ordered in 96-well plate format at a concentration of 100 µM from Integrated DNA technologies (IDT). These primers were mixed together in equal proportions and diluted to a stock concentration of 250 nM per primer. A set of 96 ‘i5’ tagging primers containing the Illumina dual-index F5 primer, unique 6 base barcode sequence and small RNA sequencing primer was designed for discriminating individual samples within each well of a 96-well plate (Fig. 1). These primers were diluted to 10 µM stock aliquots in 96-well PCR plates. Another set of 22 ‘i7’ tagging primers containing the Illumina i7 primer, unique 6 base barcode sequence and the read 2 primer were designed to assign sequence reads to up to 22 different 96-well plates (Fig. 1). The number of possible pairwise combinations for this set of i5 and i7 tags is 2112 (96 × 22), but higher multiplex levels are possible with additional tags. Specific primer sequences used in this study are available as supplemental data (Data S1, Supporting information).

GT-seq library preparation

DNA was extracted from 2068 individual O. mykiss fin tissue samples using Qiagen DNaseasy 96 Kits generating 22 extract plates with 94 samples per plate (two wells were empty per plate). Extracted DNA was not quantified but typically range from 5 to 25 ng/µL in concentration. The PCR cocktail for multiplex PCR amplification of target loci (PCR1, Fig. 1) was 2 µL of DNA extract, 3.5 µL of Qiagen Plus multiplex master mix and 1.5 µL
pooled primer mix (final volume = 7 μL; final primer concentration at each locus ≈54 nm). Thermal cycling was conducted in 96-well PCR plates with the following conditions for PCR1: 95 °C – 15 min; 5 cycles [95 °C – 30 s, 5% ramp down to 57 °C – 30 s, 72 °C – 2 min]; 10 cycles [95 °C – 30 s, 65 °C – 30 s, 72 °C – 30 s]; 4 °C hold. Following PCR1, the amplified samples were diluted 20-fold, and 3 μL of PCR product was transferred to a new 96-well PCR plate in preparation for PCR2 which adds indexes that effectively identify each sample by well and by plate (Fig. 1). For PCR2, 1 μL of 10 μM well-specific i5 tag primers were then added to the amplified samples using a multichannel pipette, and another 1 μL of plate-specific tagging primer was then added using a repeater pipette. Five microlitres of Qiagen Plus multiplex master mix (Q + MM) was then added to each well bringing the final reaction volume to 10 μL. Thermal cycling conditions for PCR2 were 95 °C – 15 min; 10 cycles [98 °C – 10 s; 65 °C – 30 s; 72 °C – 30 s]; 72 °C – 5 min; 4 °C hold. Following PCR2, each plate of samples was normalized using the SequalPrep™ Normalization Plate Kit (Applied Biosystems) according to the manufacturer’s instructions.
instructions. Following normalization, 10 μL of each sample per 96-well plate was pooled into a 1.5-mL Eppendorf tube for a total of 22 tubes.

A purification step was then performed on each of the 22 pooled aliquots by mixing 25 μL of Agencourt® AMPure® XP magnetic beads with 50 μL of pooled library in a fresh 0.2-mL PCR tube. Each tube was then placed on a magnetic rack, and the cleared supernatant was transferred to a fresh 0.2-mL PCR tube. Another 35 μL of magnetic beads were then mixed with the cleared supernatant, and each tube was placed in the magnetic rack. The supernatant was discarded, and the immobilized beads were washed twice with 70% ethanol. The purified library was then eluted with 15 μL of nuclease-free TE pH 8.0 and transferred to fresh 1.5-mL tubes before adding 1.5 μL of TE pH 8.0 with 1% Tween 20.

Following purification, each of the 22 plate libraries was quantified by qPCR using duplicate dilutions of 1:1000, 1:2000, 1:4000 and 1:8000. Five microlitres of each dilution was used as template in 15 μL reactions using 7.5 μL Power Sybr Green Master Mix (Life Technologies and 500 nM Illumina p5 and p7 primers. The qPCRs were performed in 384-well plates using a Life Technologies QuantStudio™ 6 Flex instrument (Life Technologies). Six dilutions of an Illumina library of known concentration were run in the same plate in triplicate as a standard. Following qPCR, the concentration of each plate library was calculated using their C\text{T} values in a linear regression of C\text{T} vs. LOG (conc.) generated using the standards. Each of the 22 plate libraries was then normalized to a concentration of 5 nM, and equal volumes of each were pooled to create the final GT-seq library for sequencing. The final library containing 2068 individuals was sequenced on an Illumina HiSeq 1500 instrument on one lane of a single-end read flow cell using 100 cycles with dual-index reads of six cycles each. A complete list of supplies and reagents is provided in Appendix I.

Data analysis

Sequencing data were concatenated into a single fastq file then split into plate-specific fastq files using the i7 index sequence. These files were then used to generate individual-specific fastq files using the i5 index sequence (Fig. 1). These two steps were performed using simple Linux shell scripts and are described in the bioinformatics pipeline (Data S2, Supporting information). The complete genotyping pipeline including custom perl scripts described in the text is also available online at https://github.com/GTseq/GTseq-Pipeline.

The efficiency of the library preparation for each plate of samples was tested by first using the HashSeqs.pl perl script (described in Miller et al. 2012; Hecht et al. 2013) to catalogue and count unique reads. Then, a custom perl script (GTseq_SeqTest.pl) was used to count the occurrence of each forward primer sequence, a locus/allele-specific internal probe sequence (in-silico probe) and when both occur in the same read (Data S2, Supporting information). The information from this analysis is used to identify primers that produce large amounts of ‘artefact’ reads within the sequencing data and examine read distribution among target loci. Primers producing large numbers of artefact reads in previous libraries were identified and redesigned using this approach. Of the initial 192 target amplicons, 8 primers were redesigned due to either large amounts of artefact sequences produced or because the SNP site was beyond the 100 base sequence read length. These previous libraries also helped to refine preparation techniques to improve read equality among target loci.

Individuals were genotyped with a custom perl script (GTseq_Genotyper.pl) that reads in locus information from a text file and uses allele-specific in-silico probe sequences to count the occurrence of each allele within individual fastq files (Fig. 1). Reads containing in-silico probe sequences are referred to throughout the manuscript as ‘on-target’ reads. The term ‘on-target percentage’ is also mentioned and is defined as the percentage of raw reads containing in-silico probe sequences. Once allele counts were completed, the ratio of allele 1 to allele 2 was used to determine a genotypes for each locus. Allele counts of zero were converted to a value of 0.1 for the purposes of generating a valid ratio. For each SNP, allele ratios >10.0 were called as homozygous for allele 1, ratios <0.1 were called as homozygous for allele 2, and ratios between 0.2 and 5.0 were called as heterozygous. Loci with total read counts <10 were given a genotype of ‘NA’ as 10× coverage is expected to yield accurate genotypes with GBS (Catchen et al. 2013). Likewise, allele ratios between 0.1 and 0.2 or 5.0 and 10.0 were classified as ‘NA’. The accuracy of genotyping using these cut-offs was also supported by empirical data (Figs 4 and 5). Output from the script was a list of loci, allele counts, allele ratios and genotypes. One genotype file was generated for each individual in the library which can be compiled in a number of ways for further analysis (Data S2, Supporting information).

For the sex determination marker, a separate perl script was used as this was a presence/absence type marker. This script (OmySEX_test.pl) determines a minimum threshold for the number of occurrences of the male-specific in-silico probe sequence within each of the individual fastq files based on the number of raw reads. Then, if the number of occurrences is greater than the threshold, the sex is reported as male otherwise the sex is reported as female. Samples containing <5000 raw reads were not given a genotypic sex and are reported as ‘NA’ (Data S2, Supporting information).
Analyses of read count variation among loci within individuals were performed by compiling read counts within each individual at each locus using another custom perl script (GTseq_GenoCompile_Counts.pl; Data S2, Supporting information). This script generates a single csv file containing individual raw sequence reads, on-target sequence reads, the percentage of on-target reads, the percentage of genotypes and read counts at each locus. The read count data were converted into percentages of total on-target reads for each individual at each locus. The average and standard deviation of these read percentages at each locus are presented in Fig. 3.

Genotypes were compiled from all individuals in the library using a third custom perl script (GTseq_GenoCompile.pl; Data S2, Supporting information) to test concordance vs. genotypes collected at the same loci using TaqMan™ assays. TaqMan™ assay genotypes were generated using data collected using Fluidigm 96.96 micro-fluidic chips and EP-1 instrument following protocols described in Matala et al. (2014). Concordance was evaluated only in cases where both methods produced a genotype. In cases where one or both methods failed to produce a genotype, the comparison was counted as neither concordant nor discordant. A heat map of concordance data was generated using the R-script heat map 2 using a matrix file where concordant genotypes were given a value of 2, discordant genotypes were given a value of −1, missing GT-seq genotypes a value of 0 and missing TaqMan™ genotypes a value of 1. All computing steps were performed on a single node of a CentOS Linux cluster with 24 Intel Xeon 2.50 GHz processors and 32G RAM. Individual genotyping using the GTseq_Genotyper.pl script was the most time-consuming step taking approximately 10 h to generate genotypes running all 2068 individual samples simultaneously.

To illustrate the genotyping data generated by the bioinformatics pipeline, genotype plots were generated with concatenated genotype files from 100 individual samples (all 94 samples using index 19 and samples 1–6 using index 20). Allele 1 vs. allele 2 read counts for each genotypic classification at all tested loci was plotted on an XY coordinate graph and compared to a similar plot with fluorescence intensities collected for TaqMan™ assays interrogating the same SNPs.

Results
Sequencing from one HiSeq lane produced 157M total reads with 146M (93%) reads containing intact i5 and i7 barcode sequences. Read counts from each of the 22 sample plates ranged from 5.4M to 8.6M reads per plate and individual read counts averaged 68 671 (SD 24 866). The average percentage of target genotypes collected was 96.4% (SD 13.5%) among all samples. Of the 2068 samples tested, 1987 produced genotypes in at least 90% of the target loci. Of the remaining 81 samples, 40 samples produced genotypes for at least half the target loci while 41 samples were considered to have failed (<50% genotypes).

As a quantitative measure of the quality of the library, the number of reads containing in-silico probe sequences was divided by the total reads per individual to determine the on-target percentage. This evaluation is important for this type of genotyping as only on-target reads contribute to generating genotypes (Fig. 2A,B). Among individuals, the average on-target percentage was 47.2% (SD 16.1%). However, the on-target percentage also varied greatly among plates which had an impact on the percentage of genotypes generated within each plate (Fig. 2C). As each of the plates was treated identically for library preparation, the most likely cause of this variation was DNA quality among individual samples. We hypothesized that the fragmented ends of sheared DNA were forming incomplete primer sites in PCR1 which could be extended by the polymerase to generate artefact sequences (terminal tagging). To test this hypothesis, the DNA extracts from the worst performing plate of 94 samples (average on-target reads = 29%; SD 7.1%) were treated with exonuclease I and shrimp alkaline phosphatase (Exo-SAP) prior to GT-seq library preparation. This step removes overhangs and dephosphorylates the free ends of sheared DNA blocking extension from the ends of genomic DNA during PCR. Following sequencing, the average on-target percentage among the 94 individual samples in this plate improved to 67%; SD 6.4%. These results indicate that Exo-SAP treatment of DNA extracts prior to PCR can significantly improve data quality by increasing the percentage of on-target reads.

Read counts at each locus were converted to a percentage of each individual’s on-target reads to examine the variance among loci within the library. Ideally, each locus would amplify evenly resulting in approximately equal percentages of reads. However, in this case, the percentage of reads contributed by each locus was highly variable ranging from 0.02% to 3.5% (Fig. 3) with a mean of 0.52%. This range represents approximately 175-fold difference between the least and the most abundant loci. Not surprisingly, the least abundant loci also had the lowest percentage of genotypes among samples with a strong decline in genotyping success when on-target reads were below 0.10% (Fig. 3). Only 5 of the 192 loci had <0.10% on-target reads (not including the sex determination marker OmyY1_2SEXY), and thus, the remaining 187 SNPs were genotyped at ≥90% of the individual samples. Of the five SNPs with lower genotyping success, two were genotyped at >70%, two at >50%, and one was genotyped at 20% of samples.
Fig. 2 The percentages of genotypes generated vs. the number of individual raw reads (A) and individual on-target reads (B). The variation in the fraction of on-target reads among the 22 plates has a strong influence on genotyping success (C).
Using sequences unique to each locus and allele for quantification, this method provides more precise data than TaqMan™ with very low background signal (Fig. 4A,B). This characteristic allowed all samples to be genotyped at once using the same parameters for each locus. By contrast, TaqMan™ probes anneal to the alternate allele at varying rates requiring each locus to be genotyped independently using more complex genotyping software in addition to manual scoring. Genotypes and relative fluorescence intensities of TaqMan™ probes from the same loci (Fig. 4C) illustrate that background fluorescence signal was highly variable between these 5′ exonuclease assays. In spite of the differences in methodology and data, the genotypes produced with GT-seq were 99.3% concordant to those generated using TaqMan™ assays (Fig. 5A). Discrepant genotypes were observed most commonly in only a few loci and are largely the result of sequence variation within the in-silico probe sequence resulting in a miscount of one or both alleles. When these cases were identified, the in-silico probe sequence was simply altered to match existing sequence variants. Alterations were made to the in-silico probe sequences of 15 loci with <99.8% concordance to TaqMan™ genotypes (Fig. 5B). Modification of these in-silico probe sequences improved the overall concordance rate to 99.9% with individual locus concordance rates ranging from 97.2% to 100.0%.

Cost of consumable laboratory supplies to genotype 192 SNPs for 2068 samples in a single GT-seq library was estimated to be $3.98 per sample (Table 1; all costs are reported in US dollars) which includes costs for DNA extraction ($0.51; Step 1), library preparation ($3.02; Steps 2–7) and sequencing ($0.44; Step 8). (A complete list of itemized supplies and costs is provided in Appendix I at ‘list’ prices). Cost of DNA extracts are shown for the Chelex 100 method (nondenaturing protocol) which has been shown to be effective in additional tests of our GT-seq protocol. While costs per sample for DNA extraction and library preparation are constant in the current protocol, costs for sequencing will vary depending on the number of individuals that are pooled in a single library (Fig. 6). At a minimum of one 96-well plate of samples in a single library, sequencing costs are $9.30 per sample and this amount decreases exponentially to $0.44 per sample for 2068 samples in library. Costs continue to decrease for larger numbers of samples in a library, but at diminishing returns of savings per sample (Fig. 6). This relationship of sequencing costs should allow flexibility in the number of samples that are pooled in a single library depending on the study. Total genotyping supplies cost of $3.98 for GT-seq is much lower in comparison to current 5′ exonuclease methods of genotyping on microfluidic chips at $16.50 per sample for 192 SNPs. In every case (Fig. 6), total genotyping costs are lower for GT-seq relative to 5′ exonuclease even if only 100 samples are included in a library (GT-seq = $1284, 5′ exonuclease = $1650) but savings can be extensive for large projects when thousands of samples can be pooled.

Costs for technician labour were not calculated directly since salaries, benefits, and overhead rates vary across laboratories, but we estimated that a single technician could prepare a 2000 sample library in approximately 4–6 weeks with the majority of the time being consumed by DNA extraction. For this study, the library preparation portion took only 3 days with two people working together. Sequencing and data analysis could take as little as an additional week depending on access to a HiSeq instrument and computing resources. Capital equipment costs for library preparation are expected to be minimal as the current protocol only requires
Fig. 4 These graphs plot allele 1 vs. allele 2 read counts for all target loci. The orange line indicates an even 1:1 allele 1 to allele 2 ratio ($y = x$). The other lines indicate the cut-off values used in the genotyping script. [below red = allele 1 homozygote; left of blue = allele 2 homozygote; between blue or red & black = no call; between black = heterozygote; Any data point below read depth of 10 = no call (Yellow Line)]; Genotypes are colour coded [red = allele 1 homozygote; blue = allele 2 homozygote; purple = heterozygote; yellow = no call]. Both A and B are the same plot zoomed to different scales to show an overview as well as highlight the accuracy at low read depths using GT-seq. Relative fluorescence intensities from each allele-specific probe from TaqMan™ assays for the same loci (C) are presented in a similar plot as a comparison of the two methods.
equipment that is standard in existing molecular laboratories (i.e. thermal cycler, single- and multichannel pipettes, small plate centrifuge). Sequencing costs shown in Table 1 and Fig. 6 only include the portion of a flow cell and sequencing reagents needed for a single SR 100 lane ($930 USD) divided by the number of samples in a library. However, sequencing costs will vary depending on the Illumina platform (HiSeq, MiSeq, NextSeq) and whether sequencing is performed internally or at a core facility. Overall, GT-seq is expected to be a highly cost efficient method for genotyping panels of 50–1000 SNP markers in several hundred or thousands of samples.

**Discussion**

This study introduces SNP genotyping by multiplex amplicon sequencing at large scale (GT-seq), and we demonstrate that this approach can provide high-quality genotype data for panels of at least 192 SNP loci for thousands of individual samples. This new method has several advantages over existing approaches and is highly cost efficient (Fig. 6). One advantage is that GT-seq libraries can be prepared easily using commercially available reagents and basic laboratory equipment (i.e. a thermal cycler). As GT-seq uses only unlabelled primers to generate libraries, this method also provides significant cost savings vs. custom 5' exonuclease assays that require expensive fluorescently labelled probes. Pairwise combinations of barcoded primers allow for thousands of individuals to be pooled together without the need to purchase thousands of unique adapters, and these samples can be sequenced on a single Illumina HiSeq SR 100 lane for high cost efficiency. Finally, as a relatively small number of known loci are targeted, sequencing alignment programmes are not necessary for genotyping. Rather, genotyping is performed with a simple bioinformatics pipeline using string matching to count allele-specific sequences for each SNP locus. Allele ratios were consistent among all loci with little or no background noise allowing the same parameters to be used for every locus. Overall, we demonstrate that GT-seq is a powerful approach for genotyping hundreds of SNPs for thousands of samples.

While our results demonstrate that GT-seq can be implemented immediately with existing protocols, we anticipate that further refinement of library preparation will yield higher throughput and offer more flexibility for various NGS platforms. The full potential of GT-seq has not yet been realized as a single Illumina HiSeq lane could have the capacity with approximately 200M reads to genotype many more samples or SNP loci if all sequencing reads represented ‘on-target’ amplicons with.

![Fig. 5 (A) Concordance heat map of genotypes generated using GT-seq vs. TaqMan assays using unaltered in-silico probe sequences. (B) Concordance heat map of genotypes after alteration of in-silico probe sequences to match identified sequence variants showing improved genotype concordance. Each row is an individual sample, and each column is a target locus. (green = matching genotype; yellow = TaqMan™ method failed to produce genotype; orange = GT-seq method failed to produce genotype; red = each method produced a different genotype).](image)

| Steps for GT-seq                                                                 | Cost per 96-well plate | Cost per Individual sample |
|---------------------------------------------------------------------------------|------------------------|---------------------------|
| (1) DNA extraction                                                             | $48.18                 | $0.51                     |
| (2) PCR 1: target loci amplification                                           | $61.19                 | $0.64                     |
| (3) Dilution of amplicons                                                      | $19.38                 | $0.20                     |
| (4) PCR 2: dual barcoding                                                      | $96.42                 | $1.00                     |
| (5) Sample normalization & pooling samples                                      | $87.31                 | $0.91                     |
| (6) Bead size selection                                                        | $18.75                 | $0.20                     |
| (7) qPCR & normalize each pooled 96-well plate                                  | $7.73                  | $0.08                     |
| (8) * Illumina HiSeq (SR100 reads)                                             | $42.26                 | $0.44                     |
| Total                                                                           | $381.41                | $3.98                     |

*Steps 7 and 8 assume 22 96-well plates for a single library as described in text.*
an even distribution among individuals and loci. The current capacity falls short of this potential due to a significant proportion of artefact sequences (29.5–87.5%) within the data as well as a large amount of variation in read distribution among target loci. To a lesser extent, variation in read counts among individual samples also has an effect on genotyping capacity as each sample will require a minimum threshold to produce an adequate percentage of target genotypes. It is possible that an additional step to normalize DNA quantity of each individual before library preparation may provide more even amplification across samples (e.g. Rohland & Reich 2012), but this step may not provide much additional advantages over our current protocol that normalizes each sample after amplification (Step 5; Fig. 1). However, achieving consistently high percentages of on-target reads while improving evenness among target loci will deliver the most benefit as this technique continues to be refined and advances to the multiplex PCR step should yield significant improvement in these areas. For example, although we tested several multiplex PCR master mixes and PCR conditions before conducting this large-scale test, exhaustive testing has not yet been performed (e.g. adjusting primer concentrations, DNA template concentrations or testing PCR additives such as Qiagen’s ‘Q-solution’). It was also demonstrated that Exo-SAP treatment can improve on-target read percentage by more than double for lower quality DNA samples and could be incorporated into the protocol for minimal additional costs ($0.05–$0.10 per sample). Improvements to on-target read percentage at this level would also provide more tangible solutions for applying GT-seq with smaller ‘benchtop’ sequencers that currently yield approximately 22–70M reads (e.g. Illumina MiSeq, Ion Torrent; Glenn 2011) and would allow for approximately 500–1600 samples to be genotyped for 200 SNPs in a single run.

Finally, there may be solutions to reduce artefact reads from being generated for specific loci during PCR that would reduce the number of target loci that are excluded from GT-seq panels. In our case, we initially tested 192 target amplicons on a small number of samples (n = 96) to determine which of these loci could be successfully genotyped with GT-seq. Of the 192 target amplicons, 35 loci were initially excluded for either low or discordant genotyping or high percentages of artefact (off-target) reads. Subsequently, several libraries were made and evaluated using a truncated panel of 157 loci which produced higher quality genotypes. However, after analysis of sequencing data from loci producing large numbers of sequence artefacts, we were able to identify two primers that were generating large numbers of artefacts.
GT-SEQ: SNP GENOTYPING BY AMPLICON SEQUENCING

of heterodimers within the multiplex PCR and redesign them. We also redesigned primers from four loci which generated amplicons in which the SNP site of interest occurred after the 100 bases of sequence collected. These were modified simply by swapping the sequencing primer sites such that the amplicon would then be sequenced from the reverse direction. This modification does not affect the genotyping pipeline as the genotyping script searches for both the forward and reverse complement of the in-silico probe sequence for each locus. The 29 remaining excluded loci were unchanged except that some of their in-silico probe sequences were modified. These simple changes allowed the conversion of all 192 loci for genotyping by GT-seq and will be essential initial steps to the development of new GT-seq panels. We have therefore included a troubleshooting guide as part of the GT-seq online bioinformatics package stored at https://github.com/GTseq/GTseq-Pipeline/ which outlines the initial steps necessary for testing new panels of primers.

In summary, we present GT-seq as a low cost solution for studies that require genotyping thousands of individuals at a few hundred SNPs. The novel method presented here includes library preparation and a bioinformatics pipeline for generating genotypes. Libraries are prepared using a simple cost-effective two-step PCR process requiring no other enzymatic reactions. The genotypes produced are high quality and match those collected using TaqMan™ assays at a rate of 99.9%. The bioinformatics tools also provide the opportunity for quantitative evaluation of multiplex PCR conditions from sequencing data allowing for testing of a variety of reagents and library preparation methods.

Acknowledgements

Thanks to Jeff Stephenson for the compiled costs of laboratory supplies provided in Appendix I and Amanda Matala for sequencing and data processing. Funding for this project was provided by the Bonneville Power Administration under grant 2008-907-00.

References

Baird NA, Etter PD, Atwood TS, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One, 3, e3576.
Campbell NR, Narum SR (2008) Quantitative PCR assessment of microsatellite and SNP genotyping with variable quality DNA extracts. Conservation Genetics, 10, 779–784.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Molecular Ecology, 22, 3124–3140.
Davey JW, Hohenlohe PA, Etter PD, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews. Genetics, 12, 499–510.
Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources, 11, 759–769.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. BMC Genomics, 9, 80–92.
Hecht BC, Campbell NR, Holecek DE, Narum SR (2013) Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. Molecular Ecology, 22, 3061–3076.
Matala AP, Ackerman MW, Campbell MR, Narum SR (2014) Relative contributions of neutral and non-neutral genetic differentiation to inform conservation of steelhead trout across highly variable landscapes. Evolutionary Applications, 7, 682–701.
Miller MR, Brunelli JP, Wheeler PA et al. (2012) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Molecular Ecology, 21, 237–249.
Narum SR, Baerlekde CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology, 22, 2841–2847.
Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research, 22, 939–946.
Schunter C, Garza JC, Macpherson E, Pascual M (2014) SNP development from RNA-seq data in a nonmodel fish: how many individuals are needed for accurate allele frequency prediction? Molecular Ecology Resources, 14, 157–165.
Seeb JE, Carvalho G, Hauser L et al. (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources, 11(Suppl 1), 1–8.
Sie D, Snijders PJF, Meijer GA et al. (2014) Performance of amplicon-based next generation DNA sequencing for diagnostic gene mutation profiling in oncopathology. Cellular Oncology, 31, 353–361.
Tsongalis CJ, Peterson JD, de Abreu FB et al. (2014) Routine use of the Ion Torrent AmpliSeq™ Cancer Hotspot Panel for identification of clinically actionable somatic mutations. Clinical Chemistry and Laboratory Medicine, 52, 707–714.
Wielstra B, Duijts E, Lagler P et al. (2014) Parallel tagged amplicon sequencing of transcriptome-based genetic markers for Triticus newts with the Ion Torrent next-generation sequencing platform. Molecular Ecology Resources, 14, 1080–1089.
Zielinski P, Stuglik MT, Dudek K, Konczal M, Babik W (2014) Development, validation and high-throughput analysis of sequence markers in nonmodel species. Molecular Ecology Resources, 14, 352–360.

Nathan Campbell designed the library preparation technique, wrote the genotyping scripts, performed data analyses, and wrote the paper. Stephanie Harmon prepared the sequencing libraries, helped to refine library preparation conditions, and performed data analysis. Shawn Narum supervised the development of GT-seq, assisted with writing the paper, and performed cost analysis.

Data accessibility

The complete GT-seq pipeline and necessary scripts are available both in Data S2 (Supporting information) and online at https://github.com/GTseq/GTseq-Pipeline. Individual fastq files and input files necessary for generating genotypes using them are available in the dryad database (doi: 10.5061/dryad.ff4 fs).
### Appendix 1 List of supplies reagents and costs for GT-seq.

| Item | Cost | Unit | Qty | 96-well Plate | Sample |
|------|------|------|-----|---------------|--------|
| (1) DNA extraction | | | | | |
| Extraction-Chelex 100 | 48.18 | Plate | 1.0 | $48.18 | 0.5125 |
| Subtotal | $48.18 | $0.51 |
| (2) PCR1: Pre-amp reaction | | | | | |
| Low profile PCR plate | 4.91 | Plate | 1.0 | 4.91 | 0.0511 |
| 10 µL PCR tips | 10.89 | Tray | 1.0 | 10.89 | 0.1134 |
| Repeater tips | 1.17 | Tip | 1.0 | 1.17 | 0.0122 |
| Qiagen Plus master mix | 0.09 | µL | 150.0 | 30.90 | 0.3219 |
| PCR1 primers | 0.09 | µL | 0.0 | 0.00 | 0.0000 |
| Nuclease-free water | 0.03 | ml | 0.0 | 0.00 | 0.0000 |
| Subtotal | $61.19 | $0.64 |
| (3) Dilution of amplicons | | | | | |
| Low profile PCR plate | 4.91 | Plate | 1.0 | 4.91 | 0.0511 |
| Heat seals | 0.94 | Seal | 1.0 | 0.94 | 0.0098 |
| Reagent res 100 mL | 0.67 | Box-96 | 1.0 | 0.67 | 0.0069 |
| 10 µL PCR tips | 10.89 | Tray | 1.0 | 10.89 | 0.1134 |
| 300 µL PCR tips | 10.75 | Tray | 0.1 | 1.08 | 0.0112 |
| Nuclease-free water | 0.03 | ml | 30.0 | 0.90 | 0.0094 |
| Subtotal | $19.38 | $0.20 |
| (4) PCR2: Dual barcoding | | | | | |
| Low profile PCR plate | 4.91 | Plate | 1.0 | 4.91 | 0.0511 |
| Heat seals | 0.94 | Seal | 1.0 | 0.94 | 0.0098 |
| 10 µL PCR tips | 10.89 | Tray | 2.0 | 21.77 | 0.2268 |
| Repeater tips | 1.17 | Tip | 1.0 | 1.17 | 0.0122 |
| Qiagen Plus master mix | 0.09 | µL | 535.0 | 47.23 | 0.4920 |
| GT I5 dual-index tags | 0.09 | µL | 100.0 | 9.47 | 0.0987 |
| GT I7 Index primer | 0.10 | µL | 107.0 | 10.93 | 0.1138 |
| Subtotal | $96.42 | $1.00 |
| (5) Sample normalization & pooling samples | | | | | |
| SequalPrep normalization plate | 51.40 | Plate | 1.0 | 51.40 | 0.5354 |
| 10 µL PCR tips | 10.89 | Tray | 1.0 | 10.89 | 0.1134 |
| Repeater tips | 1.17 | Tip | 2.0 | 2.35 | 0.0244 |
| Heat seals | 2.19 | Plate/cover | 3.0 | 6.57 | 0.0684 |
| 100 µL filter tips | 16.12 | Tray | 0.5 | 8.06 | 0.0839 |
| 10 µL PCR tips | 10.89 | Tray | 0.1 | 1.09 | 0.0113 |
| Reagent res 25 mL dvd | 0.75 | Box-96 | 3.0 | 2.25 | 0.0235 |
| Low profile PCR plate | 4.91 | Plate | 1.0 | 4.91 | 0.0511 |
| Subtotal | $87.51 | $0.91 |
| (6) Bead size selection | | | | | |
| 10 µL PCR tips | 10.89 | Tray | 1.0 | 10.89 | 0.1134 |
| Reagent res 25 mL | 0.75 | Box-96 | 1.0 | 0.75 | 0.0078 |
| Microcentrifuge tube | 0.08 | Tube | 1.0 | 0.08 | 0.0009 |
| 100 µL filter tips | 16.12 | Tray | 0.1 | 1.61 | 0.0168 |
| Microcentrifuge tube | 0.08 | Tube | 1.0 | 0.08 | 0.0009 |
| Ampure beads | 0.06 | µL | 60.0 | 3.60 | 0.0375 |
| Ethanol | 11.37 | Pint | 0.1 | 1.14 | 0.0118 |
| 15-mL conical centrifuge tube | 0.5082 | Tube | 1.0 | 0.51 | 0.0053 |
| TE (1×) | 0.03 | ml | 1.0 | 0.03 | 0.0003 |
| EB Tween | 0.04 | ml | 1.5 | 0.06 | 0.0006 |
| Subtotal | $18.75 | $0.20 |
| (7) qPCR & normalize plate libraries (22 × 96-well plates for a single library) | | | | | |
| Low profile PCR plate | 4.91 | Plate | 0.045 | 0.22 | 0.0023 |
| ABI optical cover | 2.19 | Plate/cover | 0.045 | 0.10 | 0.0010 |
| ABI optical plate 384 | 6.14 | Plate | 0.045 | 0.28 | 0.0029 |
| Sybr green MM | 0.07 | µL | 72.727 | 5.24 | 0.0545 |

© 2014 The Authors. *Molecular Ecology Resources* Published by John Wiley & Sons Ltd.
## Appendix I (Continued)

| Item                        | Cost | Unit | Qty | 96-well Plate | Sample |
|-----------------------------|------|------|-----|---------------|--------|
| 10 μL PCR tips              | 10.89| Tray | 0.005 | 0.05          | 0.0005 |
| Repeater tips               | 1.17 | Tip  | 0.045 | 0.05          | 0.0006 |
| Strip tube                  | 1.10 | Strips | 0.136 | 0.15          | 0.0016 |
| TE Tween                    | 0.04 | ml   | 9.091 | 0.36          | 0.0038 |
| Illumina qPCR primers       | 0.03 | μL   | 14.545 | 0.43        | 0.0045 |
| Epp 100 μL Filters          | 16.12| Tray | 0.045 | 0.73          | 0.0076 |
| TE Tween                    | 0.04 | ml   | 0.545 | 0.02          | 0.0002 |
| Microcentrifuge tube        | 0.08 | Tube | 1.091 | 0.09          | 0.0009 |
| Subtotal                    | $7.73|      |      | $0.08         |        |

(8) Illumina HiSeq SR100 lane (22 × 96-well plates for a single library as described in text)

| Item                        | Cost | Unit | Qty | 96-well Plate | Sample |
|-----------------------------|------|------|-----|---------------|--------|
| Illumina Cluster Kit single read | 462.25 | SR lane | 1.0 | 21.01        | 0.22    |
| Illumina reagents           | 467.52| SR lane | 1.0 | 21.25        | 0.22    |
| Subtotal                    | $42.26|      |      | $0.44         |        |

Total $381.41 $3.98

Vendors are only listed when a specific item is recommended for the GT-seq protocol.

All prices are ‘list price’ and include no discounts.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Data S1.** GTseq primer sequences.

**Data S2.** Pipeline Scripts.