Exome Capture for Variant Discovery and Analysis in Barley

Bayer, Micha; Morris, Jenny A.; Booth, Clare; Booth, Allan; Uzrek, Niki; Russell, Joanne R.

Published in:
Barley

DOI:
10.1007/978-1-4939-8944-7_18

Publication date:
2019

Document Version
Peer reviewed version

Link to publication in Discovery Research Portal

Citation for published version (APA):
Bayer, M., Morris, J. A., Booth, C., Booth, A., Uzrek, N., Russell, J. R., ... Hedley, P. E. (2019). Exome Capture for Variant Discovery and Analysis in Barley. In W. A. Harwood (Ed.), *Barley: Methods and Protocols* (Vol. 1900, pp. 283-310). (Methods in Molecular Biology). New York: Humana Press. https://doi.org/10.1007/978-1-4939-8944-7_18

Exome Capture for Variant Discovery and Analysis in Barley

Micha Bayer, Jenny Morris, Clare Booth, Allan Booth, Niki Uzrek, Joanne R. Russell, Robbie Waugh and Pete Hedley

The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK

SUGGESTED RUNNING HEAD: Barley Exome Capture

Summary

Exome capture is a reduced representation approach that selectively captures sequence from only the gene-bearing regions of a genome. It is based on probes targeted at these regions and, compared with whole genome shotgun sequencing, leads to a significant reduction in cost and data processing effort whilst still providing insights into the most relevant part of a genome. An exome capture array for barley was released in 2013 and this has opened the door to numerous studies that have put this technology to good use. In this chapter we detail the laboratory protocols required for enrichment and sequencing, and provide detailed step-by-step instructions for the bioinformatics analysis of the resulting data.

Key words:

Barley exome capture, library preparation, enrichment, data analysis, bioinformatics

1 Introduction

Next generation sequencing (NGS) has enabled unprecedented access to sequence variants in genomes, which can subsequently form the basis for identifying causal gene differences responsible for important traits of interest. Affordable access to NGS in recent years has ensured that even large-genome species such as barley, which has a 5.1 Gbp genome [1], can be mined efficiently for single nucleotide polymorphisms (SNPs), indels and deletions. Whole exome capture (EC) is a means of targeted re-sequencing only of gene regions, which account for less than 2% of the entire genome in barley [1]. Data generated can be used directly for population studies, QTL mapping and diversity analysis, or variants can be extracted to use on other lower-plexity platforms, including SNP chips, custom amplicon sequencing and KASP assays (LGC Genomics). The primary advantage of EC over other variant assays, which tend to be 'closed' platforms dependent upon previously characterised SNPs, is that it is not limited to known variants and is therefore an excellent means of novel unbiased SNP discovery.

A custom EC assay (SeqCap EZ Library; Roche Diagnostics) was established in barley through a consortium funded effort and covers 60 Mbp of exome, utilising long oligonucleotide probes designed to c. 40,000 genes annotated in the cultivar Morex 2012 reference sequence [1, 2]. The EC barley assay has been utilised in several projects [3-7] capturing a range of material, from cultivated to wild barleys, demonstrating the utility and robustness of the design. Recently, using a comprehensive set of EC data we have designed an Illumina SNP chip, representing 44,000 loci, for which there was an excellent assay translation rate.

Here we describe laboratory-based processing of barley DNA samples for EC and also provide a detailed bioinformatics workflow for data quality control and filtering to ensure that only the most robust and reliable variants are identified. The barley assay allows DNAs to be multiplexed prior to capture enrichment and sequencing, thereby increasing the efficiency and reducing costs of the processing.

2 Materials

2.1 Genomic DNA Extraction

1. DNeasy Plant Mini Kit (Part No: 69104; Qiagen).
2. Minigel electrophoresis system (eg. Part No: Sub-Cell®GT; Bio-Rad Laboratories).
3. SYBR®Safe DNA gel stain (Part No: S33102; Thermo Fisher Scientific).
4. 1 kb λ-DNA molecular ladder (Part No. G5711; Promega).
5. UV transilluminator (eg. Part No. UTX 20M; Uvitec)
6. General laboratory solutions: TBE, bromophenol blue.

2.2 Illumina Library Construction

1. Kapa Library Preparation Kit (Part No: 07137923001; Roche).
2. DNA vacuum concentrator (eg. Part No: 5305000100; Concentrator Plus, Eppendorf).
3. Ultra Sonicator (eg. Part No: M220; Covaris).
4. MicroTUBE AFA fiber screw-cap (6mm x 16mm, 50µl, Part No: 500096; Covaris).
5. Ethyl alcohol (Part No: E7023, Sigma).
6. DynaMag Magnet (Part No: 12321D; Life Technologies).
7. General laboratory equipment: heat block; water baths; microcentrifuge; spectrophotometer; vortex.
8. Bioanalyzer (Part No: 2100; Agilent Technologies); DNA 1000 Kit (Part No: 5067-1504; Agilent Technologies).
9. AMPure XP beads (Part No: A63880; Beckman Coulter).
10. General laboratory solutions: absolute ethanol; TE buffer (pH 8.0); elution buffer 10 mM Tris-HCl (pH 8.0).

2.3 Exome Capture

1. SeqCap EZ Developer Library, Barley Exome Design (Part No 120426_Barley_BEC_D04; Roche; (*see* **Note 1**).
2. SeqCap Adapter Kit (Part No: 07141530001; Roche).

3. SeqCap Hybridization and Wash Kit (Part No: 05634261001; Roche).
4. SeqCap EZ Accessory Kit (Part No: 07145594001; Roche).
5. SeqCap HE-Oligo Kit A (Part No: 06777287001; Roche).
6. SeqCap HE-Oligo Kit B (Part No: 06777317001; Roche).
7. SeqCap Pure Capture Bead Kit (Part No: 06977952001; Roche).

2.4 Data analysis

1. Linux server or compute cluster (CentOS 6.8 used here but most Linux distributions should work), with access through command line interface. We recommend one or more multiprocessor machines with at least 64 GB of RAM. (*See* **Note 25**).
2. List of software as shown in Table 1. All software is available freely for academic users. (*See* **Note 26**).

3 Methods

3.1 Genomic DNA Extraction

1. Use young seedling leaf tissue (approximately 3 to 5 cm) for gDNA extractions (*see* **Note 2**).
2. Isolate DNA using DNeasy plant mini-preparation kits following the manufacturer's instructions.
3. Visually assess the DNA by gel electrophoresis. Briefly a 1.5% agarose-gel is prepared with 1× Tris/Borate/EDTA (TBE) (pH 8.0) containing 10 µl SYBR®Safe DNA gel stain, placed in the electrophoresis chamber and covered with about 5 mm 1× TBE once it was cooled and solidified. For each sample, 5 µl of DNA is loaded in the wells together with 2.5 µl 1× bromophenol blue buffer. A 1kb λ-DNA molecular ladder is used as standard for determining the DNA concentration by visual comparison. Electrophoresis is conducted at a constant voltage of 100 V for 45 min and the results visualised and recorded using a UV transilluminator (*see* **Note 3**).

4. DNA is quantified using Pico-green (Fluroskan Ascent, labSystems) according to the manufacturer's instruction.

3.2 Illumina Library Construction

Illumina compatible whole genome shotgun libraries are made which include barcoding (indexing) to allow multiplexing prior to exome capture and downstream read identification (*see* **Note 4**). Here, the Kapa library construction is described as follows.

3.2.1 Fragmentation of Genomic DNA

1. Dilute gDNA to a final concentration of 2 ng/µl in TE buffer (10 mM Tris-HCl (pH 8.0), 0.1 mM EDTA).
2. Transfer 53 µl of the input gDNA (~100 ng) to a microTUBE AFA fiber screw-cap.
3. Set Covaris M220 instrument to generate an average size range of 180-220 bp. The following settings have been successfully used: Peak Incident Power, 50 W; Duty Factor, 20%; Cycles per Burst, 200; Temperature, 20 °C; Duration, 280 s (*see* **Note 5**).
4. Transfer the fragmented gDNA to a 0.2 ml PCR tube.
5. Run 1 µl on a Bioanalyzer 2100 DNA High-Sensitivity chip. Successfully fragmented DNA is shown in Figure 1A, with an average fragment size between 180 and 250 bp.

3.2.2 End repair of Fragments

1. Prepare the End Repair Master Mix as follows: water, 8 µl; 10X KAPA End Repair Buffer, 7 µl; KAPA End Repair Enzyme, 5 µl.
2. Assemble each End Repair reaction as follows in well(s) of a 96 well plate: Fragmented gDNA, 50 µl; End Repair Master Mix, 20 µl.
3. Mix by pipetting, spin briefly and incubate at 20 °C for 30 min in a thermocycler. Allow the AMPure XP beads to warm to room temperature for at least 30 min, prior to next step.
4. Prepare fresh 80% ethanol using ethyl alcohol in a screw capped tube.
5. Proceed immediately to the next step once the End Repair reaction time is finished.

6. To each 70 µl End Repair reaction, add 120 µl Agencourt AMPure XP beads. Mix thoroughly by pipetting.
7. Incubate the plate at room temperature for 10 min to allow the DNA to bind to the beads.
8. Put the plate on a magnet to capture the beads. Incubate until the liquid is clear.
9. Carefully remove and discard the supernatant.
10. Keeping the plate on the magnet, add 200 µl of 80% ethanol. Incubate the plate on the magnet at RT for 30 s.
11. Carefully remove and discard the ethanol.
12. Repeat ethanol wash.
13. Seal the plate and spin briefly (up to 2000 rpm) then place again on magnet. Remove all residual ethanol without disturbing the beads.
14. Allow the beads to dry at room temperature (~ 3 min, *see* **Note 6**). Remove the plate from the magnet and proceed immediately.

3.2.3 A-Tailing of Fragments

1. Prepare the A-Tailing Master Mix as follows for each library preparation: water, 42 µl; 10X Kapa A-Tailing Buffer, 5 µl; Kapa A-Tailing Enzyme, 3 µl.
2. Thoroughly re-suspend the beads by pipetting then seal plate.
3. Incubate at 30 °C for 30 min in a thermocycler.
4. Prepare the Indexed Adapter required in the Adapter Ligation step as follows (*see* **Note 7**). Briefly spin required adaptors (from the SeqCap Adapter Kit A and/or B) to pellet contents. Add 50 µl cold, PCR-grade water (included in the SeqCap Adapter kit). Briefly vortex and spin down the re-suspended Index Adapter tubes then keep on ice. Following use, store at -20°C for future use.
5. Equilibrate the PEG/NaCl SPRI solution (Kapa kit) to room temperature, protected from the light.

6. After the 30 min at 30 °C, add 90 µl PEG/NaCl SPRI solution to each 50 µl A-tailing reaction with beads (*see* **Note 8**).
7. Mix thoroughly by pipetting and incubate at room temperature for 10 min to allow the DNA to bind to the beads.
8. Place the plate on a magnet to capture the beads. Incubate until the liquid is clear.
9. Carefully remove and discard the supernatant.
10. Keeping the plate on the magnet, add 200 µl 80% (v/v) ethanol.
11. Incubate at room temperature for 30 s.
12. Carefully remove and discard the ethanol.
13. Repeat ethanol wash.
14. Briefly spin the plate, place it on the magnet, and remove all residual ethanol without disturbing the beads.
15. Allow the beads to dry at room temperature (*see* **Note 9**). Once dried, remove the plate from the magnet and proceed immediately.

3.2.4 Adapter Ligation

1. Prepare the Ligation Master Mix as follows for each library preparation: Water, 32 µl; 5X Kapa Ligation Buffer, 10 µl; Kapa T4 DNA Ligase, 5 µl; Indexed Adapter, 3 µl.
2. Thoroughly re-suspend the beads by pipetting.
3. Seal the plate and incubate at 20 °C for 15 min in a thermocycler, then proceed immediately to the next step.
4. Add 50 µl PEG/NaCl SPRI solution to each 50 µl ligation reaction/beads.
5. Mix thoroughly by pipetting and incubate at room temperature for 10 min to allow the DNA to bind to the beads.
6. Place the plate on a magnet to capture the beads. Incubate until the liquid is clear.
7. Carefully remove and discard the supernatant.
8. Keeping the plate on the magnet, add 200 µl 80% (v/v) ethanol.

9. Incubate the plate at room temperature for 30 s.
10. Carefully remove and discard the ethanol.
11. Repeat ethanol wash.
12. Briefly spin the plate, place on the magnet, and remove all residual ethanol without disturbing the beads.
13. Allow the beads to dry at room temperature.
14. Re-suspend the beads in 100 µl of elution buffer (EB) and incubate 2 min at room temperature to allow the DNA to elute off the beads (*see* **Note 10**).
15. For a safe stopping point, store re-suspended beads at 4 °C for up to 24 h.

3.2.5 Size Selection

Libraries are size selected using beads to ensure an optimal size range (250 bp- 450 bp) is achieved for exome capture and sequencing.

1. Add 60 µl PEG/NaCl SPRI solution to the re-suspended beads.
2. Pipette to mix and incubate at room temperature for 10 min to allow library fragments > 450 bp to bind to the beads.
3. Place the plate on a magnet to capture the beads. Incubate until liquid is clear.
4. Carefully transfer 155 µl of the supernatant(s) containing library fragments <450 bp to a new plate.
5. Discard the old plate with the beads carrying library fragments > 450 bp.
6. Vortex the AMPure XP beads (previously equilibrated at room temperature) and add 20 µl to the plate containing 155 µl of the previous supernatant.
7. Thoroughly re-suspend the beads by pipetting and incubate at room temperature for 10 min to allow library fragments >250 bp to bind to the beads.
8. Place the plate on a magnet to capture the beads. Incubate until the liquid is clear.
9. Carefully remove and discard the supernatant.
10. Keeping the plate on the magnet, add 200 µl of 80% (v/v) ethanol. No need to mix.

11. Incubate the plate at RT for 30 s.
12. Carefully remove and discard the ethanol.
13. Repeat ethanol wash.
14. Briefly spin the plate, place it on the magnet, and remove all residual ethanol without disturbing the beads.
15. Allow the beads to dry at room temperature. Remove the plate from the magnet then proceed immediately to the next step.
16. Thoroughly re-suspend the beads in 25 µl of elution buffer (EB) buffer and incubate at room temperature for 2 min to allow the DNA to elute off the beads.
17. Place the plate on a magnet to capture the beads. Incubate until the liquid is clear.
18. Transfer 20 µl of the clear supernatant to a new plate (*see* **Note 11**). Keep aside a further 2 µl aliquot in a labelled 0.2 ml tube on ice. This will be run on Bioanalyzer 2100 (see Figure 1B) to check size selection.

3.2.6 Pre-Capture PCR Amplification

1. Prepare lyophilised Pre-LM-PCR Oligos for use the first time: briefly spin and add 550 µl PCR-grade water (from the kit) to the tube labelled 'Pre-LM-PCR Oligo 1 & 2 (LP1)'. Vortex briefly and spin.
2. Prepare the LM-PCR Master Mix on ice according to the following: Kapa HiFi HotStart Ready Mix, 25 µl; Pre LM-PCR Oligos 1 & 2 (5 µM), 5 µl.
3. Pipette 30 µl of the LM-PCR Master Mix into each 20 µl sample(s) on the plate (including a negative water control, *see* **Note 12**).
4. Mix well by pipetting five times. Do not vortex.
5. Amplify samples in thermocycler using the following Pre-Capture LM-PCR program: Step 1: 98 °C, 45 sec; Step 2: 98 °C, 15 sec; Step 3: 60 °C, 30 sec; Step 4: 72 °C, 30 sec; Step 5: repeat Steps 2-4 8 times (9 cycles in total); Step 6: 72 °C, 60 sec; Step 7: hold at 4 °C.

3.2.7 Pre-Capture Purification

1. Allow an aliquot of AMPure XP Beads to equilibrate to room temperature. Vortex the beads for 10 s.
2. Add 90 µl AMPure XP Beads to 50 µl amplified library and the negative water control.
3. Pipette to mix and incubate at room temperature for 10 min to allow the DNA to bind the beads.
4. Place the plate containing the bead-bound DNA on the magnet and allow the solution to clear.
5. Remove and discard the supernatant being careful not to disturb the beads.
6. Keeping on the magnet, add 200 µl of 80% (v/v) ethanol to each well and incubate at room temperature for 30 s.
7. Remove and discard the 80% ethanol.
8. Repeat ethanol wash.
9. Briefly spin the plate, place it on the magnet and remove all residual ethanol without disturbing the beads.
10. Allow the beads to dry at room temperature for 1 min.
11. Remove the plate from the magnet and add 52 µl PCR-grade water. Pipette up and down to mix to ensure that all of the beads are resuspended.
12. Incubate at RT for 2 min.
13. Place the plate back on the magnetic and allow the solution to clear.
14. Remove 50 µl supernatant that now contains the amplified sample library and transfer into a fresh plate. Seal the plate with lids and keep on ice. Transfer the water control sample to a labelled tube.
15. Measure A260/A280 on a NanoDrop (DNA setting) to determine the concentration. The sample library yield should be > 1 µg, with a A260/A280 of 1.7-2.0. The minimum required is 20 ng/µl.

16. Run 1 µl pre-capture LM-PCR product on a Bioanalyzer 2100 DNA High Sensitivity chip (Figure 1C, *see* **Note 13**). Also run the negative water control.
17. This is a safe stopping point. The library can be stored at -20 ^{o}C for up to 1 year.

3.2.8 Hybridising Sample and SeqCap EZ Probes

1. In case of multiplexing, mix together in a 1.5 ml microfuge tube, equal amounts (by mass) of each amplified DNA sample library to obtain a single pool with a combined mass of at least 1.25 µg ("Multiplex DNA Sample Library Pool").
2. Spin the lyophilized SeqCap HE Universal and required SeqCap HE Index oligo tubes briefly.
3. Add 120 µl PCR-grade water to the SeqCap HE Universal Oligo tube (1 mM final concentration). Vortex 5 s and spin.
4. Add 10 µl PCR-grade water to each required SeqCap HE Index Oligo tube (1 mM final concentration). Vortex 5 s and spin (*see* **Note 14**).
5. Mix together the HE oligos so that the resulting Multiplexing Hybridization Enhancing Oligo Pool contains, by mass, 50% SeqCap HE Universal Oligo 1 and 50% of a mixture of the appropriate SeqCap HE Index oligos (*see* **Note 15**).
6. Add 10 µl Developer Reagent to a new 1.5 ml tube.
7. Add 1 µg Multiplex DNA Sample Library to the 1.5 ml tube containing Developer Reagent.
8. Add 2 µl (2,000 pmol) of the specific Multiplex Hybridization Enhancing Oligo pool.
9. Close the tube and make 7 holes in the tube's cap with a syringe needle.
10. Dry the Multiplex DNA Sample Library Pool / Developer reagent / Multiplex Hybridization Enhancing Oligo Pool in a DNA vacuum concentrator at 60 °C for a minimum of 20 min.
11. Once the sample is dry, cover the holes with a sticker or piece of tape.
12. Add the following to each sample: 2X SC Hybridization Buffer, 7.5 µl; SC Hybridization Component A, 3 µl.
13. Vortex the sample for 20 s and centrifuge at maximum speed for 10 s.
14. Place each sample at 95 °C heat block for 10 min.

15. During the denaturation step, equilibrate the appropriate number of 4.5 µl SeqCap EZ capture probe pool aliquots (one per library or pooled library) to ice temperature.
16. After denaturation, centrifuge samples at maximum speed for 10 s.
17. Important: work quickly through the following steps.
18. Quickly transfer the library sample to the aliquot of SeqCap EZ probe pool (in a 0.2 ml tube).
19. Carefully pipette up and down 3-4 x to mix. Avoid introducing bubbles.
20. Incubate in a thermocycler at 47 °C for 16-20 h (*see* **Note 16**).

3.2.9 Washing and Recovering Captured DNA Sample

1. Remove the SeqCap capture beads from the Hyb + wash kit and an aliquot of Ampure XP beads, and equilibrate to room temperature.
2. Dilute 10x SC Wash Buffers (I, II, and III), 10x Stringent Wash Buffer and 2.5x Bead Wash Buffer with PCR-grade water to create 1x working solutions.
3. Place the working solutions at the appropriate temperatures (Stringent Wash Buffer & Wash Buffer I at 47 °C; Wash Buffer I and all others at room temperature).
4. Vortex the SeqCap beads for 15 s.
5. Aliquot 100 µl SeqCap beads for each capture into a single 1.5 ml microfuge tube (*see* **Note 17**).
6. Place the tube on a 1.5 ml tube magnet. When the liquid becomes clear (< 5 min), remove the supernatant, being careful to leave all of the beads in the tube.
7. Add 200 µl 1x Bead Wash Buffer.
8. Remove the tube from the magnet and vortex for 10 s.
9. Place the tube back into the magnet and, once clear, remove the supernatant.
10. Repeat wash steps.
11. Re-suspend the beads in 100 µl 1x Bead Wash Buffer.
12. Aliquot 100 µl re-suspended beads into a new 0.2 ml tube.
13. Hold the tube against a plate magnet to bind the beads, remove and discard the liquid.

14. Proceed to next step quickly as possible. Do not allow the Capture Beads to dry out.
15. Place the Capture Beads at 47 °C on a PCR block then quickly transfer the hybridization samples to the Capture Beads.
16. Mix by carefully pipetting 10 x. Avoid introducing bubbles and leaving liquid droplets up the side of the tube.
17. Leave the tube(s) containing the beads and the hybridized sample(s) at 47 °C for a minimum of 45 min (up to 3 h).
18. Vortex for 3 s at 15 min intervals, to keep the beads in suspension. Ensure samples are kept at 47 °C (*see* **Note 18**).
19. After 45 min maintain tubes on the 47 °C block, and add 100 µl of Wash Buffer I pre-heated to 47 °C.
20. Mix by vortexing 10 s. Hold the tube against the plate magnet and remove Wash Buffer I.
21. Return the tube to the 47 °C block and add 200 µl of 47 °C Stringent Wash Buffer to the beads.
22. Keep the tube on the block and pipette 10 x to mix. The temperature should not drop much below 47 °C.
23. Incubate at 47 °C for 5 min.
24. Hold the tube against the plate magnet and remove the Stringent Wash Buffer.
25. Return the tube to the 47 °C block and add 200 µl of 47 °C Stringent Wash Buffer to the beads.
26. Keep the tube on the block and pipette up and down 10 x to mix. The temperature should not drop much below 47 °C.
27. Incubate at 47 °C for 5 min, then transfer sample to a 1.5 ml tube.
28. Place the tube in the 1.5 ml magnet to bind the beads and remove the Stringent Wash Buffer.
29. Remove the tube from the magnet and, at room temperature, add 200 µl Wash Buffer I.

30. Mix by vortexing for 2 min.

31. Place back on the magnet to bind the beads and remove Wash Buffer I.

32. Remove from the magnet and add 200 µl of Wash Buffer II.

33. Mix by vortexing for 1 min.

34. Place back on the magnet to bind the beads and remove Wash Buffer II.

35. Remove from the magnet and add 200 µl of Wash Buffer III.

36. Mix by vortexing for 30 s.

37. Place back on the magnet to bind the beads and remove Wash Buffer III.

38. Remove from the magnet and add 50 µl PCR-grade water to the bead-bound captured sample. Vortex briefly to mix, then sit on ice.

3.2.10 Post-Capture PCR

1. Prepare lyophilised Post-LM-PCR Oligos 1&2 (LP2) for use the first time: add 480 µl PCR-grade water to the tube.

2. Vortex briefly to re-suspend the oligos and spin down. They should be stored at -20 °C after use.

3. Prepare the Post-Capture LM-PCR Master Mix in a 1.5 ml tube as follows: Kapa HiFi HotStart Ready Mix (KH), 50 µl; Post-LM-PCR Oligos 1&2 (LP2, 5 µM) 10 µl.

4. Vortex the bead-bound captured DNA from the previous stage to ensure homogenous mixture of beads.

5. Aliquot the following mix into 2 wells of a PCR plate (per capture). Mix well by pipetting up and down (do not vortex). Bead-bound Captured gDNA, 20 µl; LM-PCR Master Mix, 30µl.

6. Amplify samples in thermocycler using the following Post-Capture LM-PCR program: Step 1: 98 °C, 45 s; Step 2: 98 °C, 15 s; Step 3: 60 °C, 30 s; Step 4: 72 °C, 30 s; Step 5: repeat Steps 2-4 13 times (for a total of 14 cycles); Step 6: 72 °C, 60 s; Step 7: hold at 4 °C.

7. Allow your aliquot of AMPure XP Beads (or the equivalent ones from the SeqCap Pure Capture Bead Kit) to warm to RT while the PCR is running.

3.2.11 Post-Capture Purification

1. Equilibrate AMPure XP Beads to room temperature.
2. Pool the corresponding amplified captured Multiplex DNA Sample Library into a 1.5 ml microcentrifuge tube (*see* **Note 19**).
3. Vortex the AmPure XP Beads for 10 s before use to ensure a homogenous mixture.
4. Add 180 µl AMPure XP Beads to the 100 µl pooled amplified captured Multiplex DNA Sample Library/beads.
5. Vortex briefly and incubate at room temperature for 10 min to allow the DNA to bind the beads.
6. Place the tube on a 1.5 ml magnet and allow the solution to clear.
7. Once clear, remove and discard the supernatant being careful not to disturb the beads.
8. Add 200 µl of 80% (v/v) ethanol, leaving the tube in the magnet. No mixing required.
9. Incubate at room temperature for 30 s.
10. Remove and discard the ethanol.
11. Repeat the ethanol wash.
12. Briefly spin the tube, place it on the magnet, and remove all residual ethanol without disturbing the beads.
13. Allow the beads to dry on the magnet at room temperature with the tube lid open (*see* **Note 20**).
14. Remove the tube from the magnet and resuspend the DNA using 52 µl EB buffer. Pipette up and down 10 x to mix.
15. Incubate at room temperature for 2 min.
16. Place the tube back on the magnetic and allow the solution to clear.
17. Transfer 50 µl supernatant (containing the amplified sample library) to a new 1.5 ml microfuge tube and place on ice.

18. Measure A260/A280 on a NanoDrop spectrophotometer to determine the concentration and quality (*see* **Note 21**). Use EB as a blank.
19. Run 1 µl post-capture LM-PCR product on a Bioanalyzer 2100 High Sensitivity DNA chip (*see* **Note 22**).
20. A successfully constructed library should look like Figure 1d, with an average fragment size 150-500 bp. Store the library at -20 °C, ready for quantification (*see* **Note 23**) and sequencing.

3.3 Illumina sequencing

Standard paired-end Illumina sequencing is performed on the EC-enriched library pool. For barley EC, we recommend 2x 75 bp read length sequencing with a minimum predicted on-target coverage of 50 x (*see* **Note 24**).

4 Data analysis

4.1 Conventions

Command line statements are displayed in grey boxes, as shown in this example:

```
myCommand parameter1 parameter2
```

Longer commands use the newline escape character (“\”) to indicate that text on subsequent lines is meant to be on the same line but has been wrapped for readability:

```
myCommand \
parameter1 \
parameter2 \
parameter3 \
parameter4 \
parameter5 \
```

4.2 Quality control

A commonly used tool for the purpose of quality control of raw Illumina data is FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Command line invocation is by means of the wrapper script provided and involves just a single parameter, the name of the FASTQ formatted read file (uncompressed or GZIP compressed):

```
fastqc myData.fastq
```

(See **Note 27** for more detail).

4.3 Read mapping

The exome capture reads are mapped to the reference sequence using a short read mapper. We recommend the BWA-MEM algorithm for this (http://bio-bwa.sourceforge.net/). (*See* **Note 28** for more detail).

First, the reference sequence has to be indexed using the "bwa index" command:

```
bwa index myReferenceSequence.fasta
```

This builds a number of index files that allow the mapping algorithm to quickly and efficiently match each read to the appropriate part of the genome.

The read mapping algorithm itself is invoked as follows:

```
bwa mem myReferenceSequence.fasta readsR1.fastq readsR2.fastq \
-R "@RG\tID:mySampleName\tSM:mySampleName" > alignment.sam
```

The standard output stream of BWA is raw SAM format and in this example is redirected to a file called alignment.sam using the ">" character. Use of the –R flag and its argument will result in read group information to be added to the SAM output (see http://www.htslib.org/ and http://samtools.github.io/hts-specs/SAMtags.pdf for more information on the SAM format and the

RG tag). This allows downstream software to associate reads with sample name and other information. (*See* **Note 29**).

The initial SAM output is then converted to compressed, indexed BAM format using the samtools view command [8] while simultaneously removing unmapped reads to save disk space (-F 4 flag). BAM is the *de facto* standard for storing short read alignments and most downstream analysis software requires its input data to be in this format.

An additional filtering step uses the bamtools toolkit (https://github.com/pezmaster31/bamtools) to remove reads that contain excessive numbers of mismatches, based on their alignment score (AS) flag in the SAM/BAM output. Mismatch cut-offs in read mapping are essential for the accuracy of downstream analysis as read mismapping caused by overly relaxed mismatch parameters can lead to dramatically increased false positive rates in variant calling [9]. The minimum alignment score cut-off is calculated as (read length in bp) – (maximum allowed number of mismatches * default mismatch penalty). The default mismatch penalty for BWA-MEM is 5, and the cutoff of 80 in the example below is based on a read length of 100 and a maximum of 4 mismatches per read (4% mismatch rate).

The initial BAM output is then piped into the “samtools sort” utility which sorts reads by both contig and start position, a requirement for further downstream analysis:

```
samtools view -F 4 -b -h alignment.sam |  \
bamtools-2.2.3 filter -tag "AS:>=80" | \
samtools sort -o alignment.sorted.bam -
```

In this example, the resulting BAM file would be called “alignment.sorted.bam”. The trailing dash (“-”) indicates that the input for the samtools sort command is its standard input stream, i.e. the output of bamtools. Please refer to the bamtools manual page for details on the other command line options used above (http://www.htslib.org/doc/samtools.html).

4.4 BAM file preprocessing and GVCF file production

The GATK Best Practices workflow [10] involves several further preprocessing steps before the alignment data can be used for the final variant calling stage.

The first of these is the removal of duplicate reads. These represent non-independent observations which may skew the downstream analysis and should therefore either be removed or flagged up as duplicates. In order to reduce disk storage requirements, removal is preferable, and we use the "samtools rmdup" command for this purpose [8]:

```
samtools rmdup alignment.sorted.bam alignment.rmduped.bam
```

The output file from this command is alignment.rmduped.bam.

The second BAM file preprocessing step consists of the local realignment of reads around indels. This adjusts the placement of reads that have been aligned sub-optimally around indels, and thereby removes base mismatches that could be misinterpreted as variants in the downstream analysis. Both of the steps above are designed to keep the false positive SNP rate to a minimum [11]. This stage consists of two separate steps, both of which involve GATK tools. In the first of these, the BAM file is scanned and a list of target sites is identified for realignment:

```
java -jar GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R myReferenceSequence.fasta \
-I alignment.rmduped.bam \
-o target_intervals.list
```

In the second step, the actual realignment itself is carried out and a new BAM file is produced:

```
java -jar GenomeAnalysisTK.jar \
-T IndelRealigner \
-R myReferenceSequence.fasta \
-I alignment.rmduped.bam \
-targetIntervals target_intervals.list \
-o alignment.realigned.bam
```

We then need to index the newly created realigned BAM file:

```
samtools index alignment.realigned.bam
```

The realigned BAM file is then run through the actual variant caller component of the GATK, the HaplotypeCaller, to produce an initial VCF output file:

```
java -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R myReferenceSequence.fasta \
-I alignment.realigned.bam \
-o initialVariants.vcf \
-dontUseSoftClippedBases
```

This has to be filtered using the vcffilter tool (https://github.com/vcflib/vcflib#vcffilter) to produce a second VCF file containing only high quality variants with a variant quality score of >=20. This score is phred-based, and a value of 20 equates to a likelihood of 1% of a variant having been called in error:

```
vcffilter -f "QUAL > 20" initialVariants.vcf \
> initialVariants.filteredQ20.vcf
```

This filtered VCF file is a requirement for the next step in the pipeline – the base quality score recalibration (BQSR). (*See* **Note 30**).

The first step of the BQSR procedure consists of computing a recalibration table which is then used as input for the second stage:

```
java -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R myReferenceSequence.fasta \
-I alignment.realigned.bam \
-knownSites initialVariants.filteredQ20.vcf \
-o recalibrationTable.txt
```

In the second step, the recalibration table is applied to the input BAM file and a new, recalibrated BAM file is produced:

```
java -jar GenomeAnalysisTK.jar \
-T PrintReads \
-R myReferenceSequence.fasta \
-I alignment.realigned.bam \
-BQSR recalibrationTable.txt \
-o alignment.recalibrated.bam
```

During this final run of the HaplotypeCaller a GVCF file is produced, which is the endpoint of the single sample processing stage:

```
java -jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R myReferenceSequence.fasta \
-I alignment.recalibrated.bam \
-o finalVariants.g.vcf \
-ERC GVCF \
--variant_index_type LINEAR \

--variant_index_parameter 128000 \

-dontUseSoftClippedBases
```

The output from this step is a GVCF format file. (*See* **Note 31**).

4.5 Variant calling

GATK provides a cohort protocol for the purpose of comparing multiple samples to one another during the final variant and genotype calling stage. This is implemented as a tool that combines multiple GVCF files into one or more cohort GVCF files which are then processed by the joint genotyper tool. Below is a hypothetical example where three samples are combined into one cohort GVCF file:

```
java -Xmx50g -jar GenomeAnalysisTK.jar \

-T CombineGVCFs \

-R myReferenceSequence.fasta \

-o cohort.1.g.vcf \

--disable_auto_index_creation_and_locking_when_reading_rods \

--variant sample1.final.variants.g.vcf \

--variant sample2.final.variants.g.vcf \

--variant sample3.final.variants.g.vcf
```

(*See* **Note 32**).

In the final stage, we use the joint genotyper (GATK's "GenotypeGVCFs" command) to call variants and genotypes for all the samples in our cohort files (here, two cohorts are shown as an illustration):

```
java –Xmx100g -jar GenomeAnalysisTK.jar \

-T GenotypeGVCFs \

-R myReferenceSequence.fasta\
-o jointGenotyperSNPs.vcf \

-nt 32 \

-V cohort.1.g.vcf \

-V cohort.2.g.vcf
```

The final output from this step in our example would be a file named “jointGenotyperSNPs.vcf”. (*See* **Note 33**).

4.6 Filtering

Filtering of raw variant calls is recommended in order to remove potential false positives. (*See* **Note 34**). Types of filters that should be routinely applied include variant likelihood (“QUAL” in VCF files, see the VCF file format specification at http://samtools.github.io/hts-specs/VCFv4.2.pdf) and a depth filter that removes sites where mismapping of reads to a secondary location leads to excess coverage and false positive variants [12].

Using the vcffilter library as above, we can combine multiple filters into a single statement:

```
vcffilter \

-f "QUAL > 30" \

-f "DP > 100"\

jointGenotyperSNPs.vcf \

> jointGenotyperSNPs_filtered.vcf
```

We recommend a QUAL filter of > 30 as shown. (*See* **Note 35**).

Depth filter cut-offs need to be appropriate for the read depth expected. Our recommendation would be to remove sites with coverage greater than 1.5x mean read depth. The latter can be calculated as follows:

(# reads mapped * read length) / size of exome capture space in bp

Example:

(40,000,000 reads mapped * 100bp) / 60,000,000 bp exome = 66.7x mean coverage

In this case an appropriate read depth cut-off would be 66.7 * 1.5 = 100. If this is a per-sample value then the cut-off for the final joint genotyper SNPs (which will be based on multiple samples) needs to be multiplied by the number of samples to provide the correct joint coverage (i.e. if the final variant calls were from 40 samples our coverage cut-off should be 40 x 100 = 4,000).

4.7 Visualization

Data visualization is of paramount importance in quality control of NGS data processing. It can be used to spot excessive numbers of read errors, abnormal variant distribution patterns or general problems with the underlying read mapping such as excess numbers of read mismatches or lack of coverage.

We use the Tablet assembly viewer for visualization of BAM files, variants and exonic regions [13, 14]. (*See* **Note 36**). The partial Tablet screenshot in Figure 2 shows a typical example of mapped exome capture reads from a single sample, in a region with multiple exons. Forward and reverse reads in read pairs are shown in green and blue, respectively. A BED file with exon coordinates has been imported as a feature track, and these are shown above the main canvas as light green bars. Likewise, a VCF file with initial variant positions has been imported too, and these are shown below the exon annotation as individual dark blue markers. On the main canvas itself, variants are visible as thin white vertical lines.

4.8 On-target rate computation

Exome capture is an enrichment approach, which means variable outcomes may be achieved in terms of the degree of enrichment. It is therefore of interest to quantify the success of the capture process itself by computing the proportion of bases that map to the enrichment target regions (here exons). A popular tool for this is CalculateHsMetrics from the Picard suite of utilities (http://broadinstitute.github.io/picard/). We can use this to compute for each sample's BAM file the percentage of bases mapped in the exonic target regions.

This tool requires a targets file which contains start and end locations for any targets used in the capture itself. The current barley exome capture array [2] was designed from the Morex v.3 assembly from 2012 [1] and although the design file itself is readily available from the manufacturer, Nimblegen (https://sftp.rch.cm/diagnostics/sequencing/nimblegen_annotations/ez_barley_exome/barley_exome.zip), it cannot be used for on-target computation with other assemblies/reference sequences. To derive the locations of the exon targets on other reference sequences, the target sequences have to first be mapped onto the new reference using the BLASTN command line tool run [15, 16]:

```
blastn \
-query barley_mapping_sequence.fa \
-db myReferenceSequence.fasta \
-max_target_seqs 1 \
-max_hsps_per_subject 1 \
-evalue 1e-10 \
-perc_identity 90 \
-out BLAST_output.txt
```

The BLAST output with the positions of the exome capture targets then needs to be converted to BED file format (https://genome.ucsc.edu/FAQ/FAQformat.html#format1). We can use Linux's built-in awk command line tool for this:

```
awk '{ print $2 "\t" $9 "\t" $10 "\t" $1}' BLAST_output.txt \
> exon_coords.bed
```

A sequence dictionary for the reference sequence then has to be created using Picard's CreateSequenceDictionary tool:

```
java -jar picard.jar CreateSequenceDictionary \
R=myReferenceSequence.fasta \
O=myReferenceSequence.dict
```

The BED file with the exon coordinates then needs to be converted to the intervals file format expected as input by the Picard CalculateHsMetrics tool (http://broadinstitute.github.io/picard/):

```
java -jar picard.jar BedToIntervalList \
INPUT=exon_coords.bed \
SEQUENCE_DICTIONARY=myReferenceSequence.dict \
OUTPUT=exon_coords.intervals
```

Once these steps are complete, we can run the actual CalculateHsMetrics tool:

```
java -jar picard.jar CalculateHsMetrics \
BAIT_INTERVALS=exon_coords.bed \
TARGET_INTERVALS=exon_coords.bed \
INPUT=alignment.recalibrated.bam \
OUTPUT=picard_Hs_metrics.txt \
VALIDATION_STRINGENCY=LENIENT
```

The output file picard_Hs_metrics.txt contains the value we are interested in – the proportion of bases among all reads that are on target. This is labeled “PCT_USABLE_BASES_ON_TARGET”. See http://broadinstitute.github.io/picard/picard-metric-definitions.html#HsMetrics for a full explanation of all the metrics contained in the output file.

5 Notes

1. Nimblegen offer a custom design service for exome capture and smaller targeted genome regions. The Barley Exome Design [2] uses a custom SeqCap EZ Developer Library which was established as an international collaborative effort to reduce design costs. Other cheaper options are available, including those produced by Agilent Technologies (SureSelect) and Mycroarray (MYbaits).
2. Use of young (< 2 weeks old) barley seedling material for gDNA extraction is important. This produces the best possible DNA quality in combination with Qiagen DNeasy Extraction Kits. Older material can suffer from lower yields and polysaccharide contamination, resulting in DNA less amenable to downstream processing.
3. Visualising DNA on a gel should not be used to quantify the DNA for library preparation but is a good 'first step' to determining whether the DNA is intact and of good quality. If a discrete band is not observed this generally means that the DNA has been degraded and unsuitable for downstream analysis.
4. Any Illumina-compatible whole genome shotgun library protocols can be used so long as standard Illumina adapters and indices are utilised. This is to ensure efficient blocking of adapter sequences is achieved downstream during the exome capture process. The numbers of samples which can be multiplexed prior to EC is dependent upon the final Illumina sequencing platform and read length selected, which ultimately determines the predicted coverage. We recommend pooling no greater than 12 barley lines prior to EC per single NextSeq 500 or HiSeq 2500 lane.
5. The exact time for fragmentation may need optimising for different species, we have increased it to 280s for barley. During the sonication process, briefly spin down the tube every 30 s, as well as at the end.
6. Over-drying the beads may result in dramatic yield loss. You should notice a 'crack' in the beads once dried and they will no longer be shiny.

7. Use Illumina's Experiment Manager software to check compatibility of index adaptors.
8. Pipette PEG/NaCl SPRI solution carefully as it is quite viscous.
9. Over-drying the beads may result in dramatic yield loss.
10. This is a safe stopping point. Store the re-suspended beads at 4 °C for up to 24 hours.
11. Do not transfer any beads with the supernatant.
12. It is important to set up a water control here and process in the same way as the samples including bead stages. Therefore, pipette 20 µl PCR- grade water into an empty well alongside your samples.
13. A successfully constructed library should have an average fragment size between 250-500 bp. Sometimes dual peaks are observed, this can be due to slight overloading of the Bioanalyzer (as shown in Figure 1C), or daisy-chaining due to excess PCR cycling.
14. Once resuspended the Hybridization Enhancing (HE) oligos can be aliquoted into smaller volumes to minimize the number of freeze/thaw cycles. Store them at -20 °C.
15. The total combined mass of the Multiplex Hybridization Enhancing Oligo Pool should be 2,000 pmol, which is the amount required for a single Sequence Capture experiment. Example: If a multiplex DNA Sample Library Pool contains four DNA sample libraries prepared with SeqCap Adapters Indexes 2, 4, 6, and 8, respectively, then the Multiplex Hybridization Enhancing Oligo Pool would contain the following: SeqCap HE Universal Oligo, 1,000 pmol (1 µl of 1,000 µM); SeqCap HE Index 2 Oligo, 250 pmol (0.25 µl of 1,000 µM); SeqCap HE Index 4 Oligo, 250 pmol (0.25 µl of 1,000 µM); SeqCap HE Index 6 Oligo, 250 pmol (0.25 µl of 1,000 µM); SeqCap HE Index 8 Oligo, 250 pmol (0.25 µl of 1,000 µM); Total 2,000 pmol (2 µl of 1,000 µM).
16. The heated lid of the thermocycler should be turned on and set to maintain 57 °C. Samples can be left for up to 72 h if needed.
17. Low-binding microfuge tubes are recommended. Enough beads for six captures can be prepared in a single tube.

18. Work quickly keeping samples at 47 ^{o}C to ensure only 'on target' fragments get through the process.
19. Some beads might stick a little to the sides of wells, however just take as much as possible. Turn the tube to spread wet beads around and help drying. High volume of beads for this step, so will take 5 min or longer. Over drying of the beads can result in yield loss.
20. The sample library yield should be >500 ng, with an A260/A280 between 1.7-2.0.
21. It is probable that a small aliquot of the samples will require diluting 1:10 to be within range of the Bioanalyzer chip.
22. It is best to quantify captured libraries using qPCR (Kapa Library Quantification Kit) due to presence of 'daisy chain' concatamers, which do not affect sequencing but can lead to unreliable Qubit quantification.
23. We recommend sequencing on NextSeq or HiSeq equipment. Please calculate read coverage based upon individual service provider's specifications for output.
24. Parallel processing of samples. The workflow described here is highly compute-intensive and hence time-consuming, but it was the most accurate option available at the time of writing. Run times for a single barley exome capture sample vary between 1 and 2 CPU days with our setup, depending on coverage. This means that large datasets with multiple samples have to be processed in parallel if the data analysis is to complete in a reasonable time frame. The tools described above are a mixture of multi- and single-threaded programs, but single-threaded programs are rate-limiting as they produce computational bottlenecks. Therefore, running a single sample with multiple threads is a poor use of resources, and a better approach is to process each sample as a single thread, with many of these jobs running in parallel. There are a number of options available to achieve this, including individual multiprocessor servers, compute clusters and cloud computing. Our setup for processing datasets with hundreds of exome capture samples is a small in-house compute cluster with 144 compute cores and up to 256 GB RAM per node. This has generally been adequate for

dealing with data on this scale, albeit with careful management of resources and occasional bottlenecks.

25. Most projects involving exome capture are aimed at small variant discovery. However, a more recent trend is to also use exome capture data for the discovery of structural variants (SV) such as larger (kbp-scale) deletions or insertions. A potential caveat here is that SV discovery with exome capture sequencing is more challenging than with whole genome data, due to the uneven and sparse read coverage. A number of tools have been published in recent years that have been specifically designed for this purpose: ExomeDepth [17], Splitread [18], CoNIFER [19], EXCAVATOR [20] and Scalpel [21, 22]. As is often the case, different tools can generate very different result sets [21], and care must be taken in the interpretation of results. Two recent reviews have also summarised the tools available for this purpose [23, 24].

26. FASTQC comes with a both command line interface and a graphical user interface, with the latter implemented as cross-platform, standalone desktop client software. It provides easy-to-interpret summary plots of metrics that quantify important quality traits such as base qualities, duplication levels, contaminant sequences and others. This allows analysists to gauge whether sequence quality is adequate before proceeding to the analysis stage. Further details on the use of this tool and the interpretation of its output are provided on the FASTQC manual page at http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/Help/.

27. Throughout our protocol, we are following the recommendations of the Genome Analysis Toolkit (GATK) Best Practices [10], which for the purpose of read mapping recommends the use of raw, untrimmed data, rather than carrying out quality-trimming and adapter removal first. This enables accurate removal of read duplicates, which relies on matching start and end coordinates of untrimmed reads. Parts of reads that are mismatched with the reference sequence are soft-clipped in BWA-MEM, i.e. the read is included in the BAM file in its

entirety, but the CIGAR string entry [8] specifies which part of the read has actually been aligned and is thus suitable for downstream analysis.

28. It is computationally preferable to write the SAM output to disk, although in theory pipes could be used to stream the output directly into the downstream tools. However, the *bwa mem* command can be parallelised with the –t option which means that reads are getting mapped in parallel threads. The downstream tools are single-threaded and thus represent a computational bottleneck if everything is done as a single pipe operation, and even with the overhead of disk access, writing the SAM file to disk is still the quickest option.
29. The base quality score recalibration (BQSR) step 's purpose is to remove the bias found in raw base quality scores, which tends to be associated with a base's position in the read and also the identity of neighbouring bases [11]. It produces a final BAM file where base qualities have been adjusted as appropriate, and this is the file that is used as input for the second (and here final) run of the variant caller module, the HaplotypeCaller. This procedure requires a truth set of known variants which can be used to guide the recalibration. In the absence of a publicly available benchmark dataset (e.g. those available for human genomics), users are encouraged to produce their own calibration datasets by means of a bootstrapping approach that produces a high quality call set. The recommendation is to aim for convergence between expected and observed base qualities, which can require several iterations of the BQSR, but we found that a single iteration provides an acceptable degree of convergence.
30. This is a proprietary GATK file format which represents a variant of the standard VCF format which contains additional information designed to allow easy comparison of multiple samples in cohorts (http://gatkforums.broadinstitute.org/gatk/discussion/4017/what-is-a-gvcf-and-how-is-it-different-from-a-regular-vcf). Note the .g.vcf file extension, which is required by the GATK components.

31. In practice, small numbers of samples (below ~ 50, based on our experience) do not require this additional step, and instead their GVCF files can be fed into the joint genotyper directly. We also found that for larger numbers of samples, a cohort size of around 20 represents a good compromise between memory consumption and speed.
32. The GenotypeGVCFs component is compute-intensive and can be multi-threaded using the –nt option as shown. It is also memory-intensive, and in our example we have increased the default amount of allocated memory to 100 GB using the –Xmx option. The exact amount of memory required varies with the number of cohorts and their size, but as a guide we recommend between 0.5-1GB of RAM per sample for a cohort size of 20.
33. The type of filters and their parameters depend to a large extent on the intended purpose of the dataset. A trade-off exists between completeness of the callset and reliability of the constituent variants. Aggressive filtering generally means fewer false positives but more false negatives, whereas little filtering leads to an inverse outcome. If the purpose of the callset is SNP discovery for e.g. a genotyping panel, where reliability of SNPs is of paramount importance, then an aggressive filtering approach should be taken. Conversely, if the callset is intended for the identification of a single SNP linked to a given phenotype, then a conservative filtering approach should be taken that removes very few SNPs, in order not to remove the SNP of interest along with false positives.
34. The variant quality is a phred-like quality score (https://en.wikipedia.org/wiki/Phred_quality_score) that quantifies the likelihood of a variant being genuine rather than artefactual. A score of 30 signifies a 0.1% chance of the variant having been called in error, and we deem this a sufficiently low probability.
35. Tablet is standalone desktop software that can be installed easily on Windows, Mac and Linux platforms using ready-made installers bundled with their own version of Java (see https://ics.hutton.ac.uk/tablet/download-tablet/).

6 References

1. Mayer K.F.X., Waugh R., Langridge P., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome*. Nature* **491**(7426), 711-+
2. Mascher M., Richmond T.A., Gerhardt D.J., et al. (2013) Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond*. Plant Journal* **76**(3), 494-505
3. Pankin A., Campoli C., Dong X., et al. (2014) Mapping-by-Sequencing Identifies HvPHYTOCHROME C as a Candidate Gene for the early maturity 5 Locus Modulating the Circadian Clock and Photoperiodic Flowering in Barley*. Genetics*,
4. Wendler N., Mascher M., Nöh C., et al. (2014) Unlocking the secondary gene-pool of barley with next-generation sequencing*. Plant Biotechnology Journal* **12**(8), 1122-1131
5. Nice L.M., Steffenson B.J., Brown-Guedira G.L., et al. (2016) Development and Genetic Characterization of an Advanced Backcross-Nested Association Mapping (AB-NAM) Population of Wild × Cultivated Barley*. Genetics* **203**(3), 1453-1467
6. Hisano H., Sakamoto K., Takagi H., et al. (2017) Exome QTL-seq maps monogenic locus and QTLs in barley*. BMC Genomics* **18**(1), 125
7. Russell J., Mascher M., Dawson I.K., et al. (2016) Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation*. Nat Genet* **48**(9), 1024-1030
8. Li H., Handsaker B., Wysoker A., et al. (2009) The Sequence Alignment/Map format and SAMtools*. Bioinformatics* **25**(16), 2078-2079
9. Ribeiro A., Golicz A., Hackett C., et al. (2015) An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome*. BMC Bioinformatics* **16**(1), 382
10. Van der Auwera G.A., Carneiro M.O., Hartl C., et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*. Curr Protoc Bioinformatics* **43**, 11.10.1-33
11. DePristo M.A., Banks E., Poplin R., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data*. Nature Genetics* **43**(5), 491-+
12. Ribeiro A., Golicz A., Hackett C.A., et al. (2015) An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome*. BMC Bioinformatics* **16**(1), 1-16
13. Milne I., Stephen G., Bayer M., et al. (2013) Using Tablet for visual exploration of second-generation sequencing data*. Briefings in Bioinformatics* **14**(2), 193-202
14. Milne I., Bayer M., Cardle L., et al. (2010) Tablet--next generation sequence assembly visualization*. Bioinformatics* **26**(3), 401-2
15. Altschul S.F., Gish W., Miller W., et al. (1990) Basic local alignment search tool*. J Mol Biol* **215**(3), 403-10
16. Camacho C., Coulouris G., Avagyan V., et al. (2009) BLAST+: architecture and applications*. BMC Bioinformatics* **10**, 421
17. Plagnol V., Curtis J., Epstein M., et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling*. Bioinformatics* **28**(21), 2747-2754
18. Karakoc E., Alkan C., O'Roak B.J., et al. (2012) Detection of structural variants and indels within exome data*. Nat Meth* **9**(2), 176-178
19. Krumm N., Sudmant P.H., Ko A., et al. (2012) Copy number variation detection and genotyping from exome sequence data*. Genome Research* **22**(8), 1525-1532
20. Magi A., Tattini L., Cifola I., et al. (2013) EXCAVATOR: detecting copy number variants from whole-exome sequencing data*. Genome Biology* **14**(10), R120
21. Narzisi G., O'Rawe J.A., Iossifov I., et al. (2014) Accurate de novo and transmitted indel detection in exome-capture data using microassembly*. Nat Meth* **11**(10), 1033-1036

22. Fang H., Bergmann E.A., Arora K., et al. (2016) Indel variant analysis of short-read sequencing data with Scalpel*. Nat. Protocols* **11**(12), 2529-2548

23. Tattini L., D'Aurizio R., and Magi A. (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data*. Frontiers in Bioengineering and Biotechnology* **3**(92),

24. Guan P. and Sung W.-K. (2016) Structural variation detection using next-generation sequencing data: A comparative technical review*. Methods* **102**, 36-49

Tables and figure legends

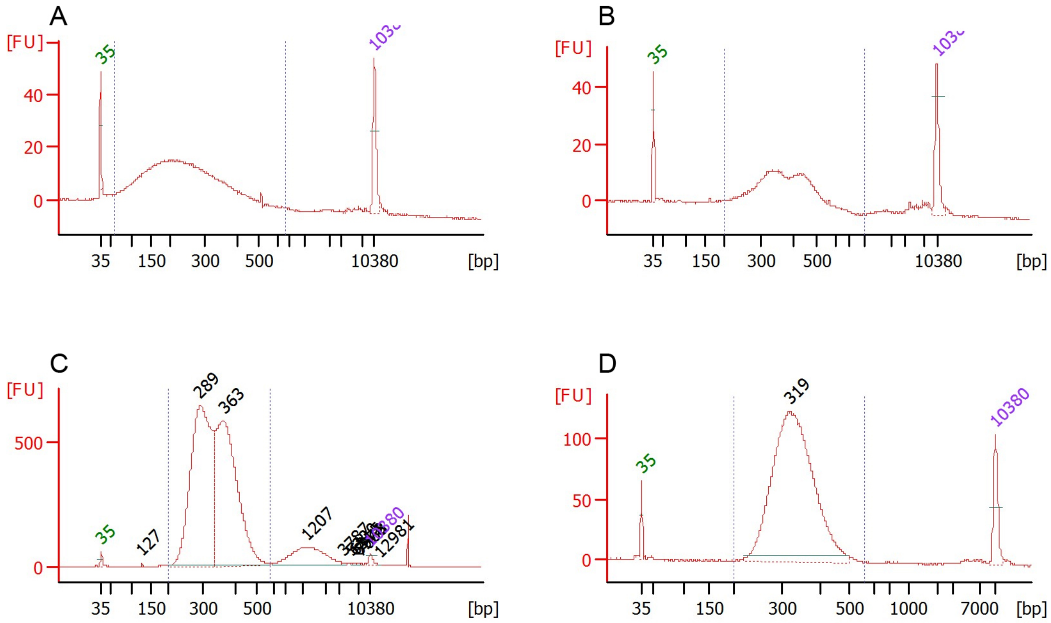
Table 1. List of software and versions. The list of software shown specifies versions used in the pipeline implemented by the authors. Scientific software is subject to ongoing development and later or earlier versions of the software used here may not behave as expected, or may even lack some of the parameters used here. While most software developers attempt to preserve backward compatibility during the development process, there is no guarantee that a different version will provide the same outcome as the versions used here. With the exception of GATK, which is cross-platform, all the tools used here are designed to run on the Linux operating system. All tools are executed through command line statements.

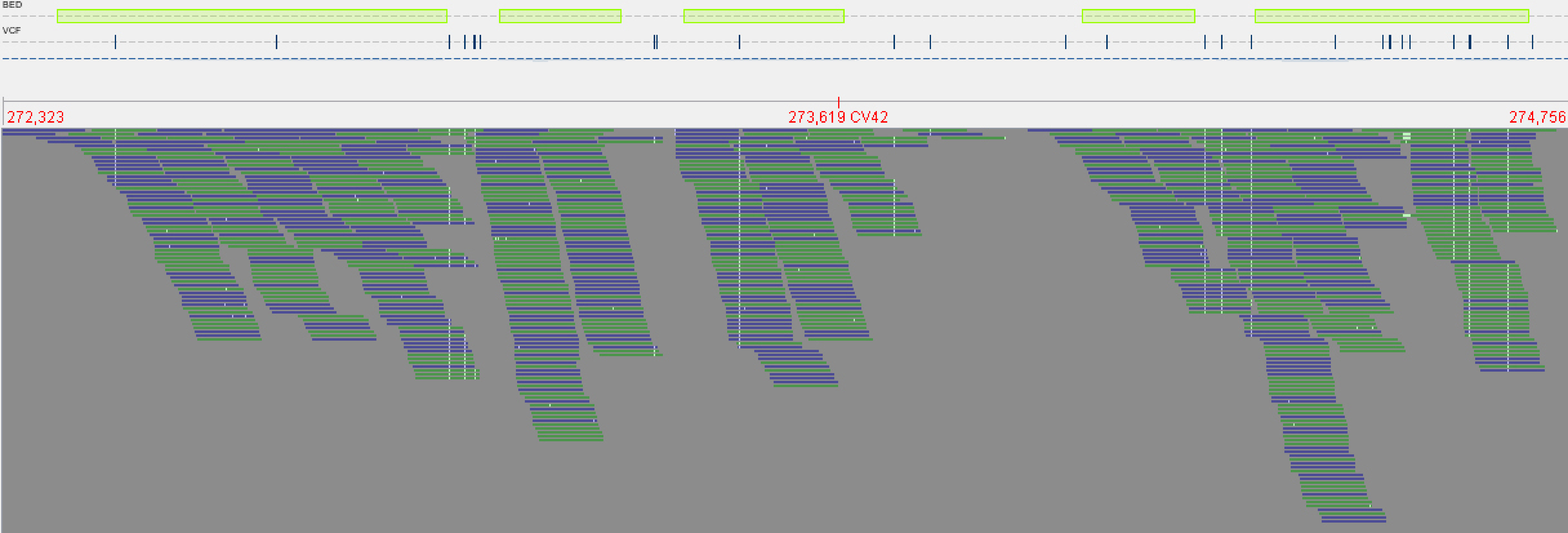
Tool	Version	URL
GATK	3.4.0	https://software.broadinstitute.org/gatk/
bamtools	2.2.3	https://github.com/pezmaster31/bamtools
bwa mem	0.7.10	http://bio-bwa.sourceforge.net/
samtools	1.3.1	http://samtools.sourceforge.net/
vcflib	20140627	https://github.com/vcflib/vcflib
Picard toolkit	1.138	http://broadinstitute.github.io/picard/

Figure 1. Bioanalyzer traces showing examples of successful sample preparation through the exome capture process: A. Post-fragmentation gDNA; B. Post size-selection library; C. Pre-capture library; D. Final captured library.

Figure 2. Partial screenshot of the Tablet assembly viewer software, showing on the main canvas mapped exome capture reads from a single sample (green = forward, blue = reverse) and SNPs (thin white vertical lines). The annotation tracks above the main canvas show exon annotation (top track,

light green horizontal bars) and SNP information imported from the VCF file (second track from top, thin blue vertical marks).


A
[FU]
40
20
0
35
103
35
150
300
500
10380
[bp]
B
[FU]
40
20
0
35
103
35
150
300
500
10380
[bp]
C
[FU]
500
0
35
127
289
363
1207
10380
12981
35
150
300
500
10380
[bp]
D
[FU]
100
50
0
35
319
10380
35
150
300
500
1000
7000
[bp]


BED
VCF
272,323
273,619 CV42
274,756