Pathology (Month 2014) **00(xx)**, pp. 1–11

Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes

James S. Wilmott[1,2], Matthew A. Field[3], Peter A. Johansson[4], Hojabr Kakavand[1,2], Ping Shang[1], Ricardo De Paoli-Iseppi[1], Ricardo E. Vilain[1,2], Gulietta M. Pupo[1,5], Varsha Tembe[1,5], Valerie Jakrot[1], Catherine A. Shang[6], Jonathan Cebon[7], Mark Shackleton[8], Anna Fitzgerald[6], John F. Thompson[1,2,9], Nicholas K. Hayward[4], Graham J. Mann[1,2,5] and Richard A. Scolyer[1,2,10]

[1]Melanoma Institute Australia, North Sydney, NSW, [2]Sydney Medical School, The University of Sydney, Camperdown, NSW, [3]Immunogenomics Laboratory, Australian National University, Canberra, ACT, [4]Oncogenomics Laboratory, QIMR Berghofer Medical Research Institute, Herston, Brisbane, Qld, [5]Centre for Cancer Research, The University of Sydney at Westmead Millennium Institute, Westmead, NSW, [6]Bioplatforms Australia, Macquarie University, North Ryde, NSW, [7]Ludwig Institute for Cancer Research, Olivia Newton-John Cancer and Wellness Centre, Austin Health, Heidelberg, Vic, [8]The Cancer Development and Treatment Laboratory, Peter MacCallum Cancer Centre and Sir Peter MacCallum Department of Oncology, The University of Melbourne, Vic, [9]Departments of Melanoma and Surgical Oncology, and [10]Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Camperdown, NSW, Australia; these authors contributed equally

Summary

Whole genome sequencing (WGS) of cancer patients' tumours offers the most comprehensive method of identifying both novel and known clinically-actionable genomic targets. However, the practicalities of performing WGS on clinical samples are poorly defined. This study was designed to test sample preparation, sequencing specifications and bioinformatic algorithms for their effect on accuracy and cost-efficiency in a large WGS analysis of human melanoma samples. WGS was performed on melanoma cell lines ($n = 15$) and melanoma fresh frozen tumours ($n = 222$). The appropriate level of coverage and the optimal mutation detection algorithm for the project pipeline were determined. An incremental increase in sequencing coverage from 36X to 132X in melanoma tissue samples and 30X to 103X for cell lines only resulted in a small increase (1–2%) in the number of mutations detected, and the quality scores of the additional mutations indicated a low probability that the mutations were real. The results suggest that 60X coverage for melanoma tissue and 40X for melanoma cell lines empower the detection of 98–99% of informative single nucleotide variants (SNVs), a sensitivity level at which clinical decision making or landscape research projects can be carried out with a high degree of confidence in the results. Likewise the bioinformatic mutation analysis methodology strongly influenced the number and quality of SNVs detected. Detecting mutations in the blood genomes separate to the tumour genomes generated 41% more SNVs than if the blood and melanoma tissue genomes were analysed simultaneously. Therefore, simultaneous analysis should be employed on matched melanoma tissue and blood genomes to reduce errors in mutation detection. This study provided valuable insights into the accuracy of SNV with WGS at various coverage levels in human clinical cancer specimens. Additionally, we investigated the accuracy of the publicly available mutation detection algorithms to detect cancer specific SNVs which will aid researchers and clinicians in study design and implementation of WGS for the identification of somatic mutations in other cancers.

Key words: Cancer, coverage, genomics, melanoma, methods, mutation, pathology, sequencing, treatment, whole genome.

Received 18 March, revised 22 June, accepted 28 June 2015

INTRODUCTION

Whole-genome sequencing (WGS) of fresh-frozen tumours has enabled the characterisation of the entire genomic profile of a patient's cancer, theoretically allowing the identification of virtually all possible genomic drivers, disease modifiers and risk factors, thereby facilitating the selection of the most appropriate treatment options for an individual patient's disease.[1,2] Unlike whole exome sequencing (WES), which only analyses the portion of DNA that is transcribed into proteins (~1.2% of the genome), WGS covers the entire genome. Until very recently, the use of WES or WGS in the clinical setting has been restricted by the prohibitive costs of sequencing and the additional necessary computing resources needed to perform downstream data analysis. However, with the advent of new technologies, the cost of WGS of a patient's genome has recently dropped below US$1000.[3] Therefore, we can expect its clinical use to rapidly expand due to greater accessibility to clinicians and affordability for patients. In fact many small scale trials have already implemented WGS in the clinic, resulting in altered clinical care based on the increase in knowledge of the cancer genome and the development of novel therapies that target the protein products of specifically mutated genes.[4–6]

The successful implementation of the technology in the clinic or in a research setting relies upon sound methodological

Print ISSN 0031-3025/Online ISSN 1465-3931 Copyright © 2015 Royal College of Pathologists of Australasia. All rights reserved.
DOI: 10.1097/PAT.0000000000000324

approaches and despite technological advances, some of the practicalities of performing WGS are not well outlined for clinical samples.[7] Methodological factors that need to be considered in performing WGS on clinical samples include: appropriate sample selection, storage, preservation and macrodissection for tumour enrichment, DNA preparation, sequencing specifications and accuracy of the variant detection algorithms. The importance of sample selection and DNA extraction can be overlooked, leading to unusable or inaccurate data.[8] Patient biopsies often contain an assortment of tumour and non-tumour cells such as immune, stromal and endothelial cells. The presence of the latter in substantial numbers reduces the ability to detect somatic mutations within tumours by genome sequencing and can lead to false negative mutation calls.[9–11] This can have serious consequences for patient care, as patients with false negative results could be incorrectly designated as ineligible for a personalised treatment that could have been effective.

Another major factor to consider when performing WGS is the number of replicate sequences (or reads) of the genome that is required to accurately identify all mutations present. Sequencing coverage (X) expresses the average number of times an average nucleotide base will have been read in a given sample.[12] However, coverage varies across the genome due to variability in sampling and ease of sequencing; it is desirable to maximise coverage to obtain data on a greater proportion of the genome. Selecting the appropriate coverage for a project is often a balancing act between sensitivity and costs, with increased levels of coverage producing more reliable and sensitive variant detection capabilities but at an increased cost.[7] For sequencing projects, this results in a choice between increased coverage per sample but lower overall sample numbers or higher sample numbers and reduced sensitivity to detect rare or low frequency events. Deciding upon the appropriate coverage levels for a WGS assay is a vital aspect of clinical and research WGS. Coverage levels of 10–30X to detect putative germline mutations (usually derived from blood DNA) have been suggested, while higher coverage levels (>40X) are considered necessary for mutation detection with tumour samples due to contamination by normal cells, tumour heterogeneity and amplification bias.[7,13]

The algorithms and processing tools that are used to convert raw sequencing data into annotated lists of mutations can often be the most expensive and infrastructure-intensive aspect of the WGS process. The sequencing platforms produce large text files of nucleotide sequences that need to be aligned to the reference genome using algorithms such as the Burrows–Wheeler aligner (BWA).[14] Variant detection algorithms such as SAMtools can then detect alterations between the genomes based on the probability of a variant occurring in that genomic region,[15] whilst taking into account sequencing error and predicted polymorphism rates. There are over 205 tools for WGS data analysis that differ in their statistical approach, number of variants identified and type of mutations detected by each algorithm.[16,17] However, for matched blood and melanoma tissue samples, these algorithms are founded on two basic principles of detecting somatic mutations in cancer: (1) subtraction method, whereby the mutations are compared to a synthetic reference genome separately for the matched blood and the tumour samples, then mutations present in the blood are removed from the tumour calls by subtraction or filtering; (2) simultaneous detection of the matched tumour and blood samples uses probability based statistics to filter mutations that are unique to the tumour sample.[18] The choice of the bioinformatics variant detection algorithm and methodology has an effect on the data produced from the raw WGS.[19]

In the present study, part of the Australian Melanoma Genome Project (AMGP) which will complete WGS of 400 human melanomas by the end of 2015, we performed a rigorous optimisation process for sample preparation, quality control, sequencing coverage level analysis and the mutation detection methodology via the WGS of 222 patients' matched melanoma and normal white blood cell genomes. The results of this study provided some valuable insights into the accuracy to detect single nucleotide variants (SNVs) at various coverage levels in different sample types. Additionally, we investigated the accuracy and specificity of the simultaneous versus the subtraction methodology using publicly available mutation detection algorithms.

METHODS AND MATERIALS

Study overview

WGS was performed on genomic DNA extracted from matching blood and melanoma tissue samples on the Illumina HiSeq2000 platform (Illumina, USA). The study was designed to optimise the sample preparation, sequencing specifications and bioinformatic algorithms to decide upon the most accurate and cost-efficient methodology to be used in the AMGP. We began with a pilot study that performed WGS on a cell line and melanoma tissue samples to determine the appropriate level of coverage and optimise the mutation detection algorithms. Samples were sequenced in individual flow cell lanes and the data later combined to simulate incremental levels of coverage, 20–60X for blood genomes and 40–100X for tumour genomes. Variant detection was then performed at increasing coverage levels to determine the appropriate coverage for each DNA source. We then optimised the mutation detection algorithms to determine the methodology that produced the most reliable and accurate variant detection.

Specimen collection

The tissue and blood samples analysed in the current study were obtained from Australian melanoma biospecimen banks, which include the Melanoma Institute Australia (MIA) ($n = 198$), Queensland Institute of Medical Research (QIMR) ($n = 15$), Ludwig Institute for Cancer Research ($n = 5$), Peter MacCallum Cancer Centre/Victorian ($n = 4$) biospecimen banks. All tissues and bloods form part of a prospective collection of fresh-frozen samples accrued with written informed patient consent and institutional review board approval [MIA is covered by the Sydney South West Area Health Service institutional ethics review committee (Royal Prince Alfred Hospital zone), the Ludwig Institute for Cancer Research is covered by the Austin Hospital committee and the Peter MacCallum Cancer Center]. Clinical and follow-up details were collected on all patients as approved by the aforementioned research ethics committees.

Study population

Patients were selected for WGS based on the availability of fresh frozen melanoma tissue and blood that fulfilled the following clinical criteria:

1. Primary melanoma with a patient matched melanoma metastasis (any metastatic site).
2. Primary melanoma with greater than 3 years clinical follow-up and prior mRNA array data or known sentinel lymph node biopsy (SLNB) status.
3. Regional lymph node metastasis with greater than 3 years of clinical follow-up data and prior mRNA array data.[20]
4. Distant metastasis to the small intestine (limited to 8 samples) or brain (unlimited number of samples).
5. Melanomas of any stage of disease that arose from mucosal epithelium. Any sample with a primary melanoma that bordered cutaneous epithelium was excluded.
6. Melanomas of any stage of disease that arose from acral skin. Any sample with a primary melanoma that bordered non-acral skin was excluded.
7. Human melanoma cell lines with prior drug screen data were also included (limited to 15 samples).

Patients were excluded if they had received prior chemotherapy or radiotherapy to the biopsy site or if they had existing exome or WGS data available.

Tissue DNA extraction and quality assessment

Fresh surgical specimens were macro-dissected and tumour tissue was procured (with as little contaminating normal tissue as possible) and snap frozen in liquid nitrogen within 1 hour of surgery. For primary tumours, only macroscopically visible tumour was banked. The fresh-frozen tumour samples were sectioned on a cryostat (CM1520; Leica, Germany) and the slides were stained with haemotoxylin and eosin (H&E). Areas with high tumour content (>80% if possible) were marked and reviewed by pathologists (RV and RAS). The following features were evaluated for the selected area: percentage of tumour nuclei, percentage of non-tumour nuclei, percentage area displaying necrosis, degree of pigmentation (Fig. 1A–D; absent, mild, moderate, severe), predominant cell shape (Fig. 1E–G; epithelioid, spindle and mixed epithelioid and spindle), tumour cell size of the most cellular portion of the tumour, and the density of tumour-infiltrating lymphocytes (TILs) as previously described and depicted in Fig. 1H–K.[20–23] Cell size was measured as the longest dimension of the nuclei in ten representative cells using the measure tool in Leica's LAS software on photomicrographs taken using a 20× objective (Fig. 1L; DM2000; Leica).

The minimum tissue criterion needed for inclusion in the study was a macro or microdissectible tumour area containing greater than 80% tumour content and less than 30% necrosis as marked. Samples that needed tumour enrichment underwent macrodissection using a marked H&E slide as a reference to remove non-tumour or necrotic tissue under sterile conditions (Fig. 1M,N). Tumour DNA was then extracted using DNeasy Blood and Tissue Kits (69506; Qiagen, Germany), according to the manufacturer's instructions. Blood DNA was extracted from whole blood using the Flexigene DNA Kit (51206; Qiagen). All individual samples were quantified using the NanoDrop (ND1000; Thermo-Scientific, USA) and Qubit dsDNA HS Assay (Q32851; LifeTechnologies, USA) and DNA size and quality were tested using electrophoresis gels. Samples with a concentration of less than 50 ng/μL or absence of a high molecular weight band in electrophoresis gels were excluded from further analyses.

Cetyltrimethyl ammonium bromide (CTAB) DNA clean-up

Excessive melanin pigment was removed from tissue-derived DNA using a modified CTAB clean-up process.[24] Briefly, fresh CTAB was made with 50 mM Tris-HCL, 1% CTAB (#52365; Sigma, USA), 4 M urea (#U5378; Sigma), 1 mM EDTA and RNAse-free water. Per 100 μL of DNA elution buffer, 39 μL of 5 M NaCl was added, then 500 μL of CTAB-Urea solution added, mixed and left overnight at 4°C on a rotator. Samples were then spun at 15,000 g for 15 min at 4°C. Supernatant was discarded and the DNA pellet resuspended in 200 μL of buffer ATL, to which 200 μL of buffer AL and 200 μL of absolute ethanol were added and the subsequent solution pipetted onto a DNA extraction column. Columns were spun and washed in buffer AW1, AW2 and the DNA was eluted in 100 μL of buffer AE.

Whole genome sequencing

WGS was undertaken at three Australian sequencing facilities (Australian Genomic Research Facility, Ramaciotti Centre for Genomics, John Curtin School of Medical Research) managed by the Australian government infrastructure enabling body, Bioplatforms Australia, and also at Macrogen (South Korea). All facilities carried out the following protocols. Library construction was performed using TruSeq DNA Sample Preparation kits as per Illumina instructions. Sample DNA (1 μg) was fragmented into 300–400 base pair (bp) average insert size with 3′ or 5′ overhangs. End repair mix was then used to convert the fragmented DNA into blunt ends by removing the 3′ overhangs and the polymerase activity fills the 5′ overhang. The 3′ ends were then acetylated to add a single 'A' nucleotide to the 3′ to reduce chimera formation. Ligate adapters were then used to attach adapters to the DNA fragments so they could be loaded into a flow cell and purified to remove unligated adapters to generate a final product with an insert size of 300–400 bp. PCR was then used to selectively enrich DNA fragments with adapter molecules at both ends for sequencing. Post-amplification quality controls were performed using DNA High Sensitivity Labchips (Agilent Bioanalyzer; Agilent, USA). The libraries were then pooled and clustered using the iBOT and ready for sequencing. The 100 bp pair-end library was sequenced on a Hiseq2000 using a Truseq SBS V3-HS kit (Illumina). Each sample was analysed in the minimum number of lanes to generate the total data output (i.e., full lanes dedicated to one sample rather than multiplexing unless a half lane is needed to achieve the final desired output). FASTQ files were then sent to the bioinformatics pipeline for analysis.

Detection of somatic nucleotide variation

To detect somatic variants in each sample, a variant detection pipeline was implemented that was derived from an existing pipeline originally designed for mouse exome analysis.[25] The new pipeline consists of a MySQL tracking database Supplementary Fig. 1, http://links.lww.com/PAT/A38 in addition to a versioned code base containing custom Perl code, external annotation files, and external binaries. For each new patient, an automated script enters sample and sequencing metadata into the tracking database, information such as tumour and tissue type as well as the number of lanes for the each tumour and control sample. The script further creates all the necessary files for submission of jobs onto the 57,000-processor compute cluster raijin (http://nci.org.au/nci-systems/national-facility/peak-system/raijin/) hosted at the National Computational Infrastructure (NCI) at the Australian National University. The utilisation of the NCI was required as analysis of each pair of patient samples requires 1000 CPU hours and 1 terabyte of storage. Once entered into the system, each sample followed a pre-defined workflow (Fig. 2) ensuring reproducible results and further allowing users to make consistent comparisons across samples analysed at different stages of the project. This workflow consisted of three main phases: alignment, variant detection and annotation.

i) Alignment

Each individual paired lane was aligned to the human reference genome GRCh37 using the short read aligner BWA[14] with default parameters unless a non-standard base quality encoding was utilised. Repetitively aligned reads were filtered out from the SAM file and a sorted BAM file generated. Alignment statistics were generated from the BAM file using the SAMtools[15] flagstat command and only lanes with 90% or more of the reads aligning to the reference genome were further processed. When all the sequencing lanes for a patient had been aligned and had passed quality control steps, single merged BAM files were created for the control sample and each tumour sample using SAMtools merge. Duplicate reads were subsequently removed from each merged BAM file using the SAMtools rmdup command, a step that aims to avoid the detection of false positive variants arising due to PCR duplication. Coverage levels for each sample were computed and checked against minimum requirements of the project (40X for control / cell line tumour samples and 60X for tissue derived tumours).

ii) Variant detection

Variant detection utilised SAMtools and BCFtools to generate a VCF file for each chromosome. Variants were filtered first on their variant quality score (>40, higher score indicating reduced risk of false base call) and then their CLR score (>60), a score that measures the probability that the tumour and control had different genotypes. Variants from each chromosome were next merged and classified according to variant type (SNVs or indels) and further segregated into germline, LOH, or somatic categories with only somatic variants processed further.

iii) Annotation

All variants were overlapped with ENSEMBL's canonical transcripts and run through the variant effect predictor (VEP).[26] to determine whether the SNVs represented non-synonymous changes and to obtain additional annotations. Polyphen and SIFT scores were extracted in order to prioritise variants by predicting the functional impact of the missense mutations on the protein. Variants that overlapped with exon splice sites (defined as 10 bases either side of exon) were considered the most likely to be pathogenic. Finally, variants were overlapped with dbSNP[27] to determine whether the variant had been previously reported in the general population and to record allele frequencies of variants matching dbSNP.

To compare the somatic variants with previous cancer studies, we annotated genes in Vogelstein's influential paper[28] and provided a link to the COSMIC repository.[29] Variants found to have an exact coordinate match to existing COSMIC entries were further annotated with special attention paid to variants previously detected in melanoma studies.

The final step was the generation of separate variant reports for SNVs and small indels that integrated all the information calculated during analysis which was accessed from the tracking database. The reports contained sequence information for the variants in the tumour and control samples as well as extensive gene annotations and links to external resources.

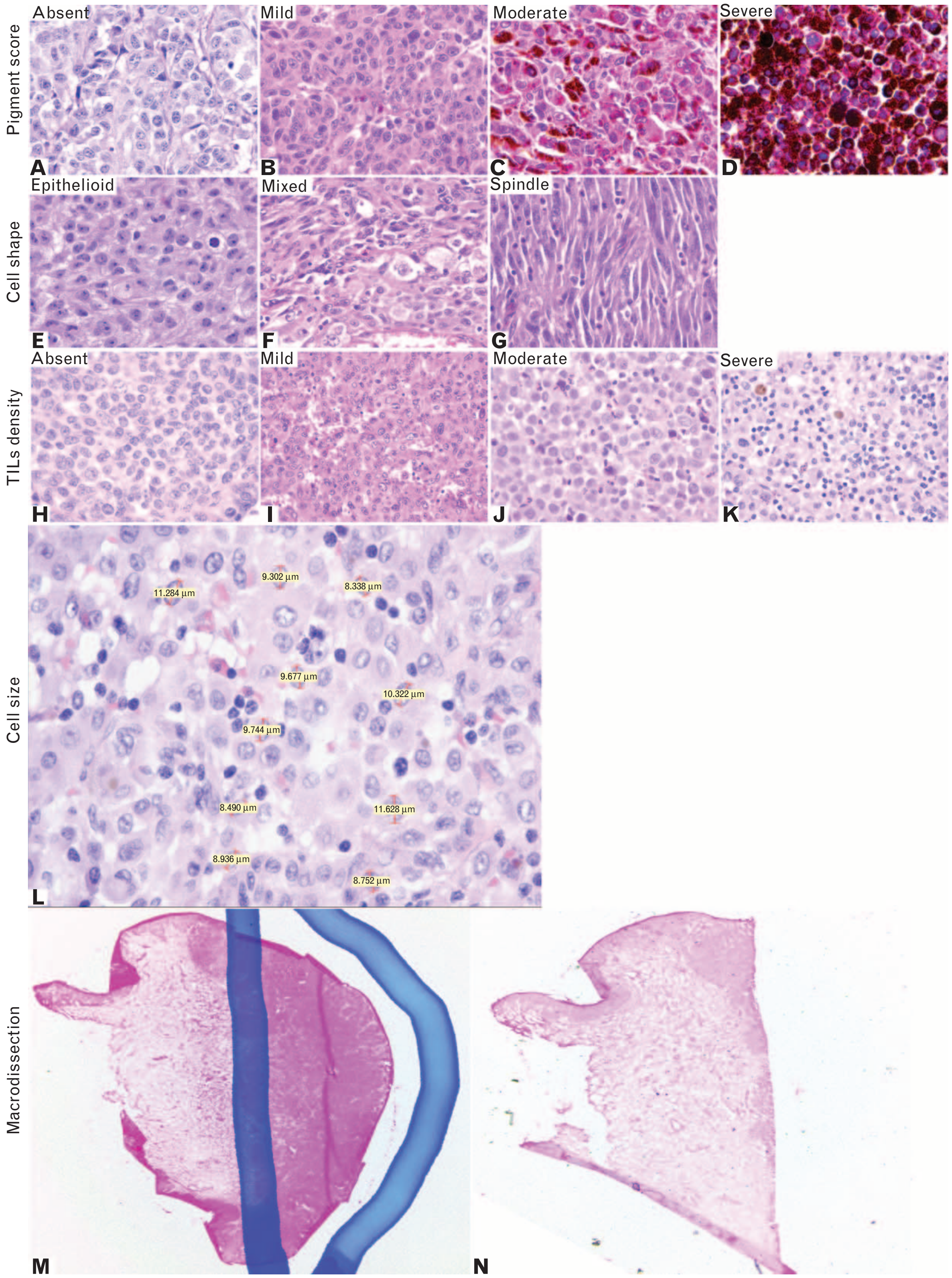

Fig. 1 Guide for pathological assessment of frozen tissue sections for WGS (H&E). (A–D) Increasing degrees of melanin pigmentation; (E–G) epithelioid, spindle, and mixed cell shape; (H–K) increasing density of tumour infiltrating lymphocytes; (L) an example of the measurements of cell size; (M) pre-macrodissection marked slide; and (N) post-macrodissection quality control slide.

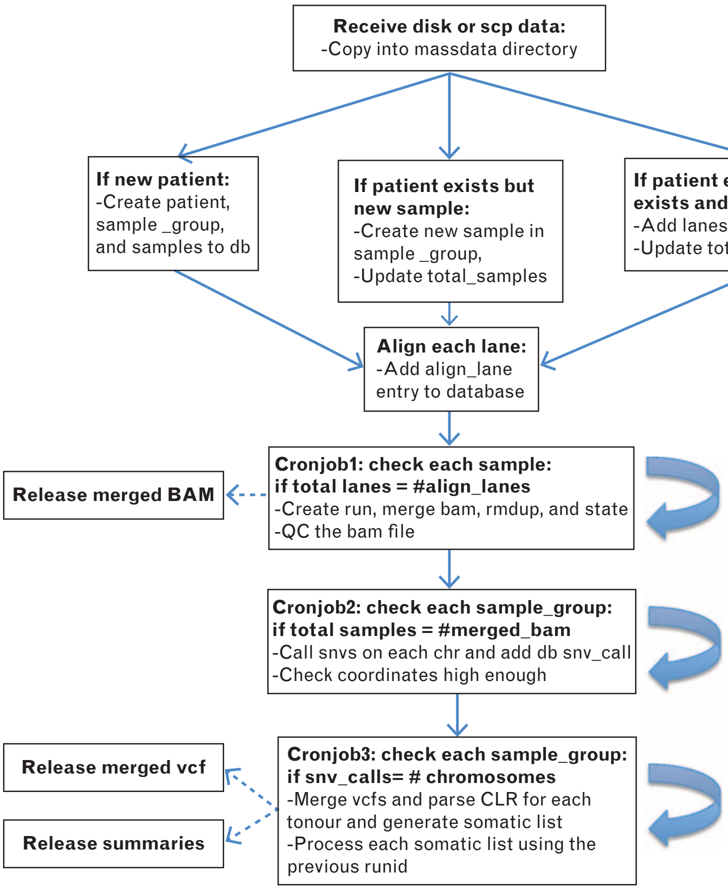
Receive disk or scp data:
-Copy into massdata directory
If new patient:
-Create patient, sample _group, and samples to db
If patient exists but new sample:
-Create new sample in sample _group,
-Update total_samples
If patient exists and sample exists and other lanes exist:
-Add lanes to sample
-Update total_lanes
Align each lane:
-Add align_lane entry to database
Cronjob1: check each sample:
if total lanes = #align_lanes
-Create run, merge bam, rmdup, and state
-QC the bam file
Release merged BAM
Cronjob2: check each sample_group:
if total samples = #merged_bam
-Call snvs on each chr and add db snv_call
-Check coordinates high enough
Cronjob3: check each sample_group:
if snv_calls= # chromosomes
-Merge vcfs and parse CLR for each tonour and generate somatic list
-Process each somatic list using the previous runid
Release merged vcf
Release summaries

Fig. 2 Workflow of patient data through the bioinformatics process.

RESULTS

Patients and melanoma samples

At the time of writing this report, 222 matched bloods and tissue samples from 204 patients had been contributed from the biospecimen banks of MIA ($n = 198$), QIMR ($n = 15$), Peter MacCallum Cancer Centre ($n = 5$) and the Ludwig Institute for Cancer Research ($n = 4$), all of which had undergone WGS. The melanoma samples comprised 18 primary melanomas with patient matched metastasis, 30 primary melanomas with long-term (>3 years) clinical follow-up data, 18 primary melanomas with known SLNB status, 62 regional lymph node metastases with long-term (>3 years) clinical follow-up data, 19 cerebral metastases, eight small intestine metastases, 15 melanoma cell lines with prior drug screen data, 25 (8 primary and 17 metastatic) acral melanomas and nine (6 primary and 3 metastatic) mucosal melanomas. Figure 3 depicts the selection and exclusion process of the samples that underwent WGS. In total, over 445 patients were identified as having fresh frozen melanoma tissue in the MIA biospecimen bank that fitted the criteria for the study. The major causes of exclusion from the study were lack of a blood or other germline sample ($n = 93$) and prior radiotherapy or chemotherapy ($n = 17$). Tissue and DNA quality controls excluded 83 samples due to low tumour content (<80%), high necrosis (>30%) and consequent poor quality DNA. A proportion of patients had undergone prior genome sequencing in other studies and therefore were excluded from this analysis ($n = 30$).

Seventy-nine primary melanomas were suitable for WGS, the subtypes of which included 26 nodular (33%), 14 superficial spreading (SS) (18%), 11 acral lentiginous (14%), nine desmoplastic (11%), five SS with a dominant dermal nodule (6%) and one lentigo maligna melanoma (1%). The subtype of 13 cases (17%) was unknown at time of analysis (Fig. 4A, B). Additionally, 142 metastatic melanoma samples were included in the WGS study, comprising 21 in-transit metastases, 72 regional lymph node metastases, 34 distant metastases and 15 cell lines (Fig. 4A, C). The subtype of the likely causal antecedent ('culprit') primary was determined using a published algorithm,[30] indicating that the metastatic samples were derived from 19 acral lentiginous, four desmoplastic, 40 SS, 40 NM and eight SS with a dominant dermal nodule. In eight cases the primary melanoma was occult. The primary melanoma subtype for 23 samples is still to be determined.

Genomic DNA quantification and quality controls

During the DNA extraction process it became apparent that the eluted DNA remained pigmented when extracted from pigmented melanoma tissue. This was a concern as melanin has been shown to inhibit PCR reactions and can interfere with downstream library preparation.[13] Additionally, the DNA from pigmented samples demonstrated a significant over-estimation in DNA concentration when measured with the spectrometry-based NanoDrop compared with the fluorometric based Qubit assay ($p = 0.008$; Supplementary Fig. 2A, http://links.lww.com/PAT/A38): pigmented samples often measured many fold higher using the NanoDrop compared to the Qubit assay due to the absorbance attributable to melanin. The pigmented DNA

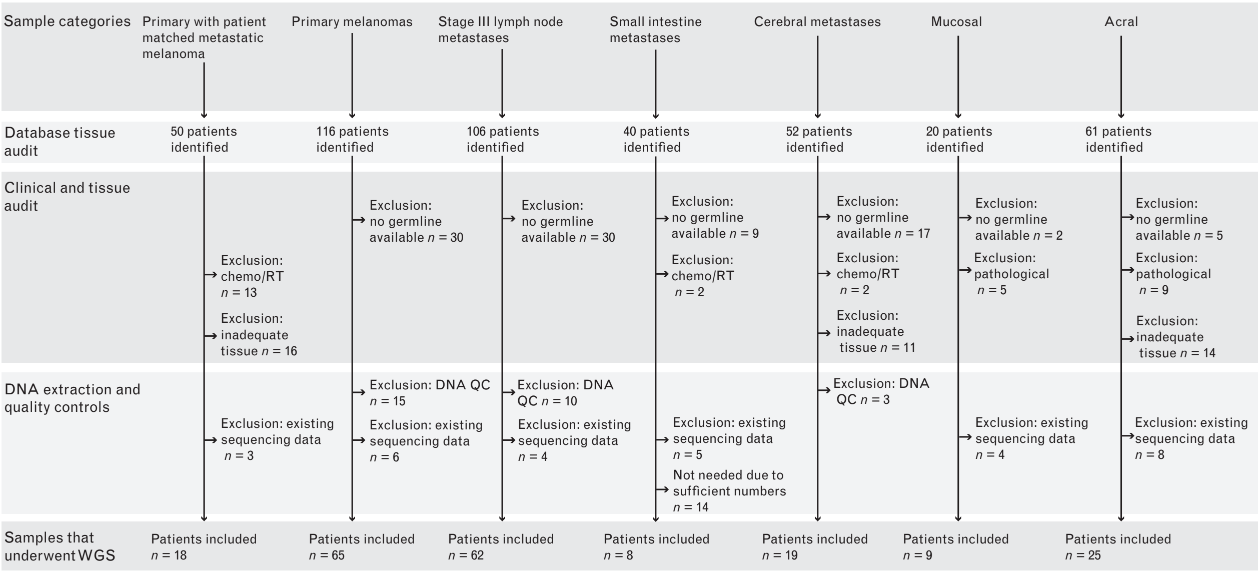
Sample categories
Primary with patient matched metastatic melanoma
Primary melanomas
Stage III lymph node metastases
Small intestine metastases
Cerebral metastases
Mucosal
Acral
Database tissue audit
50 patients identified
116 patients identified
106 patients identified
40 patients identified
52 patients identified
20 patients identified
61 patients identified
Clinical and tissue audit
Exclusion: chemo/RT n = 13
Exclusion: inadequate tissue n = 16
Exclusion: no germline available n = 30
Exclusion: no germline available n = 30
Exclusion: no germline available n = 9
Exclusion: chemo/RT n = 2
Exclusion: no germline available n = 17
Exclusion: chemo/RT n = 2
Exclusion: inadequate tissue n = 11
Exclusion: no germline available n = 2
Exclusion: pathological n = 5
Exclusion: no germline available n = 5
Exclusion: pathological n = 9
Exclusion: inadequate tissue n = 14
DNA extraction and quality controls
Exclusion: existing sequencing data n = 3
Exclusion: DNA QC n = 15
Exclusion: existing sequencing data n = 6
Exclusion: DNA QC n = 10
Exclusion: existing sequencing data n = 4
Exclusion: existing sequencing data n = 5
Not needed due to sufficient numbers n = 14
Exclusion: DNA QC n = 3
Exclusion: existing sequencing data n = 4
Exclusion: existing sequencing data n = 8
Samples that underwent WGS
Patients included n = 18
Patients included n = 65
Patients included n = 62
Patients included n = 8
Patients included n = 19
Patients included n = 9
Patients included n = 25

Fig. 3 Patient inclusion and exclusion process for the patient cohorts. A database audit of the biospecimen banks yielded candidate samples. Patient samples were excluded on the presence of appropriate germline samples, prior radio or chemotherapy, poor quality tumour (<80% tumour content or >30% necrosis), poor quality or quantity of genomic DNA (lack of high molecular weight band in electrophoresis gel, concentration <50 ng/μL and <1 μg total) or existence of prior exome or whole genome sequencing data.

also appeared heavily smeared on electrophoresis gels (Supplementary Fig. 2B, http://links.lww.com/PAT/A38).

Therefore any pigmented tissue samples that produced pigmented DNA, smearing on electrophoresis gels or had abnormally high NanoDrop readings underwent a cleanup process using a modified CTAB method.[24] We measured the DNA concentration before and after the clean-up process on both assays. The concentrations of the CTAB clean-up DNA was reduced in the NanoDrop readings but the Qubit readings were often stable or increased due to differences in elution volumes, suggesting minimal loss of double stranded DNA and removal of melanin ($p = 0.476$). Subsequent electrophoresis gels of the cleaned-up DNA show less DNA fragmentation following CTAB treatment (Supplementary Fig. 2B, http://links.lww.com/PAT/A38) and the copy number plots from the WGS data shows an intact genome (Supplementary Fig. 2C, http://links.lww.com/PAT/A38).

Optimal coverage level determination

In order to determine optimal coverage levels, two melanoma tissue samples and one melanoma cell line were sequenced to depths of at least 100X. Sequence data were then subsampled to test tumour coverage levels in one sequencing lane increments, ranging from 30–132X against a control cover level of three lanes of sequencing. For each coverage level, somatic variants were identified, average quality scores calculated and variants compared to dbSNP v135.[31]

For the melanoma tissue samples (SCC09 and SMU11), Table 1 shows <1% additional variants were identified at 124–132X compared to 61–64X coverage levels. Furthermore, somatic variants identified at 61–64X have mean SNV qualities of 212 and 216 while the variants added at 124–132X are of lower quality with means of 90 and 65 (Fig. 5A–D). However, 0.7–2.4% of SNVs were added at coverage levels less than 60X (Table 1). Consequently, for melanoma tissue samples we concluded that coverage above 60X only results in a minor increase in the total number of additional variants, the majority of which have quality scores that are indicative of false positive variants. The coverage analysis was then performed on the genome derived from melanoma cell line A15. This showed that increasing coverage from 30X to 40X resulted in a 1% increase in the total number of SNVs that were detected at 103X (Fig. 5E), with the gain in the number of SNVs levelling off below 1% beyond 43X. Less than 1% more variants were identified at 103X compared to 43X and these variants had a mean quality of 70 compared to 209 mean quality for variants at 43X (Fig. 5F). Therefore, we selected a minimum coverage of 40X for all cell line derived samples. Likewise, coverage below 40X for blood derived DNA resulted in a reduction of 1.5–5% of total SNVs compared to coverage greater than 40X (Table 1 and Fig. 5G–I). Therefore coverage of at least 40X was recommended for blood derived genomes Fig. 4D.

Variant detection optimisation: independent versus paired variant detection

Two options for somatic variant detection were initially considered. The first option was to call variants independently on the tumour and normal samples using SAMtools/BCFtools followed by extraction of tumour unique somatic variants by detecting variants present in the tumour sample and absent in the normal sample. The second option was to call variants simultaneously using SAMtool's paired mode followed by an additional filtering step on the CLR score. Several samples were analysed to determine which method generated more reliable tumour specific variant calls.

Detection of variants simultaneously yielded an average of 40% less somatic variants with virtually all variant calls being a subset of variants detected using the independent detection method (Fig. 6). Further examination of the 40% of variants detected solely with the independent method revealed that they generally fell into the following categories: (1) the tumour sample contained a variant while the normal sample had insufficient coverage to make a variant call either way, or (2) the tumour sample variant was slightly above the variant cut-off score while the normal sample was slightly below the variant cut-off (Supplementary Fig. 3, http://links.lww.com/

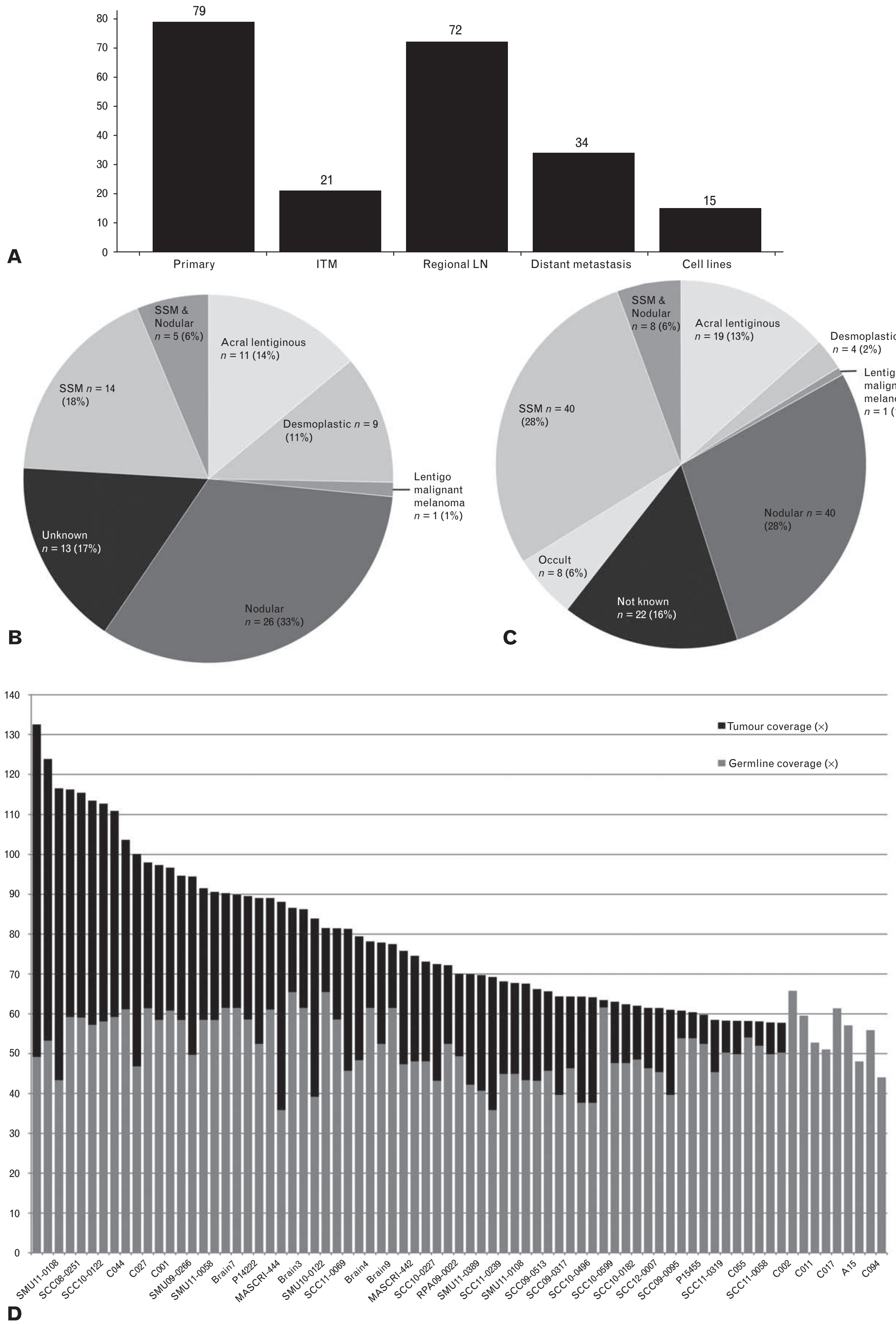

Fig. 4 Summary of melanoma sample type, melanoma subtype of the primary melanoma and coverage levels of the samples within the study. (A) Histogram of the samples included in the WGS study and their various tissue types; (B) melanoma subtype of the primary melanoma samples; (C) melanoma subtype of the culprit primary melanoma of all the metastatic melanoma samples; and (D) histogram depicting the coverage levels of the blood (mean 51X) and tumour (mean 79X) genomes of the samples process at time of publication.

Table 1 Depth of coverage and single nucleotide variant (SNV) detection

Sample	Coverage (Lanes)	Total somatic SNVs	Number SNVs added at this coverage level	Mean SNV quality of added SNVs	Percentage of total SNVs detected
A15 (tumour)	30X (3)	98297	N/A	N/A	98.4
A15 (tumour)	43X (4)	99267	970	76	99.4
A15 (tumour)	56X (5)	99543	276	67	99.7
A15 (tumour)	70X (6)	99738	158	65	99.9
A15 (tumour)	80X (7)	99740	2	62	99.9
A15 (tumour)	93X (8)	99868	128	66	100.0
A15 (tumour)	103X (9)	99884	16	64	
SCC09 (tumour)	38X (3)	157338	N/A	N/A	97.6
SCC09 (tumour)	51X (4)	159352	2014	85	98.8
SCC09 (tumour)	64X (5)	160255	903	87	99.4
SCC09 (tumour)	79X (6)	160699	444	90	99.7
SCC09 (tumour)	92X (7)	160955	256	90	99.8
SCC09 (tumour)	105X (8)	161098	143	97	99.9
SCC09 (tumour)	119X (9)	161157	59	97	100.0
SCC09 (tumour)	132X (10)	161214	57	97	
SMU11 (tumour)	36X (3)	322406	N/A	N/A	98.3
SMU11 (tumour)	47X (4)	325846	3440	77	99.3
SMU11 (tumour)	61X (5)	327149	1303	69	99.7
SMU11 (tumour)	74X (6)	327660	511	66	99.9
SMU11 (tumour)	86X (7)	327891	231	63	99.9
SMU11 (tumour)	99X (8)	328051	160	61	100.0
SMU11 (tumour)	111X (9)	328089	38	61	
SMU11 (tumour)	124X (10)	328142	53	60	
A15 (Lymphoblast cell line)	33X (3)	93664	N/A	N/A	93.9
A15 (Lymphoblast cell line)	46X (4)	98212	4548	178	98.5
A15 (Lymphoblast cell line)	55X (5)	99738	1526	136	
SCC09 (blood)	36X (3)	152682	N/A	N/A	95.0
SCC09 (blood)	49X (4)	160699	8017	183	
SMU11 (blood)	39X (3)	322659	N/A	N/A	98.5
SMU11 (blood)	53X (4)	327660	5001	159	

PAT/A38). As variants in these categories are likely to either be false positives or inconclusive we concluded that the simultaneous detection method yielded more reliable somatic variant calls and was utilised in our analysis.

DISCUSSION

The accuracy and sensitivity of WGS data depends on appropriate sample preparation, sequencing specifications and the selection of an appropriate mutation detection algorithm. The current study determined the optional WGS coverage for human melanoma tissue and blood genomes and optimised freely available mutation detection algorithms to accurately call mutations unique to the tumour genome, whilst limiting false positive calls.

Our study found that an increase in sequencing coverage from 60X to 120X in melanoma samples resulted in a minimal increase in the detection of SNVs, the quality of which indicated they could be false positive calls or in a minor subpopulation of tumour cells. Likewise, melanoma cell line genomes reach saturation for SNV discovery at around 40X, most likely due to their greater purity and increased homogeneity. The study suggests that these coverage levels empower the detection of 99% of informative SNVs, which would represent a sensitivity level at which clinical decision making or landscape research projects can be carried out with a high degree of confidence. However, lower coverage WGS (5–15X) has been used to detect high frequency variants in genome wide association studies and by The Cancer Genome Atlas,[32] but the sensitivity to detect rare or low frequency events is diminished. Nevertheless, variant detection algorithms and sequencing technologies are constantly evolving, with improvements being made to the accuracy, speed and ability to accommodate lower levels of coverage.[33,34] Therefore, the appropriate coverage levels for a specific project may vary depending on the tissue type, sequencing technologies, detection algorithms and the type of questions the data needs to address.

Likewise the detection methodology, subtraction versus the simultaneous variant detection, had significant effect on the number and quality of SNVs detected, finding that the subtraction methodology with SAMtools resulted in a 41% increase in the number SNVs detected over the simultaneous detection method. These results are similar to those of other studies that found the additional SNV calls are the result of library preparation bias, sequencing errors and mutation detection artifact due to read depth imbalances that represent false positive calls.[25] Therefore, the simultaneous detection approach should be employed on matched tumour and control genomes to reduce false positive variant calls.

The current study highlights some basic steps that are required to achieve high quality input DNA and accurate sequencing data from clinical melanoma specimens. The key to achieving this is the accurate evaluation and selection of areas of tumour for macrodissection by a skilled pathologist. With careful macrodissection of the pathologist-selected tumour region and subsequent quality controls, factors that contribute to erroneous results such as low tumour content, excessive necrosis or misclassified samples can be better avoided. If the tissue analysed is suboptimal, this can have long-term implications for further research as the genome sequencing results are made publicly available for use by the wider cancer research community. Therefore the process of sample selection and preparation is critical in WGS as subsequent results depend upon the accuracy of these steps.

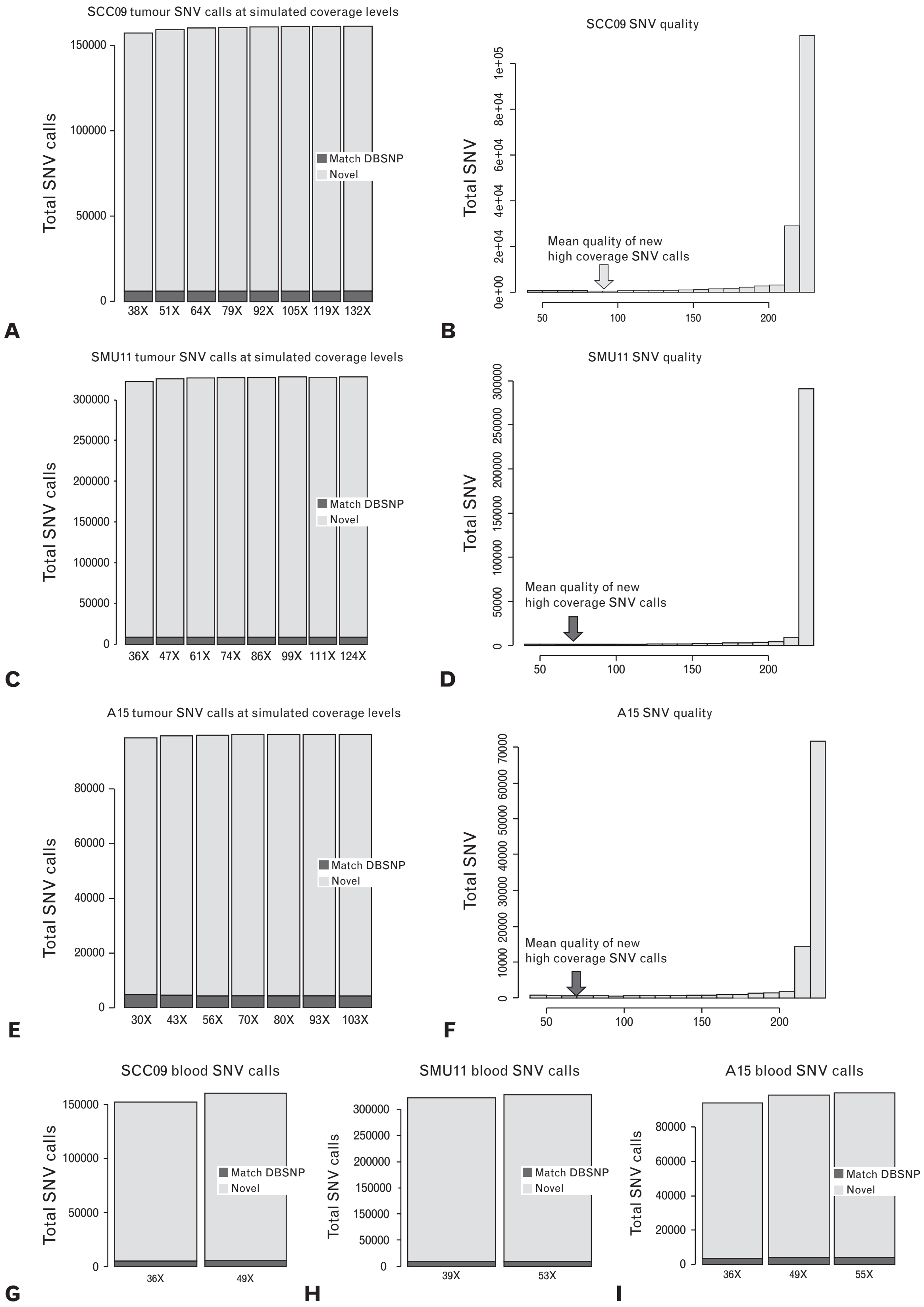

Fig. 5 Number and quality of SNV detection at simulated coverage levels. (A,B) SCC09 tumour genome: increasing coverage from 38 to 132X only generated 2.5% additional SNVs, with a total of 161,214 SNVs identified at 132X at an average quality of 212 (arrow, mean quality of SNVs detected greater than 32X). (C,D) SMU11 tumour genome: increasing coverage from 36 to 124X only generated 1.8% additional SNVs, with a total of 328,142 SNVs identified at 124X at an average quality of 216 (arrow, mean quality of SNVs detected greater than 36X). (E,F) A15 tumour genome: increasing coverage from 30 to 103X only generated 1.6% additional SNVs, with a total of 99,543 SNVs at an average quality of 209 (arrow, mean quality of SNVs detected greater than 30X). (G) SCC09 blood genome: increasing the coverage from 36 to 49X added 8017 (5%) additional SNVs with a mean quality of 212 for all SNVs identified. (H) SMU11 blood genome: increasing the coverage from 39 to 53X added 5001 (1.5%) additional SNVs with a mean quality of 216 for all SNVs identified. (I) A15 lymphoblast cell line: increasing the coverage from 33 to 46X added 4548 (5%) and from 46 to 55X added another 1526 (1.5%) SNVs with a mean quality of 209 for all SNVs.

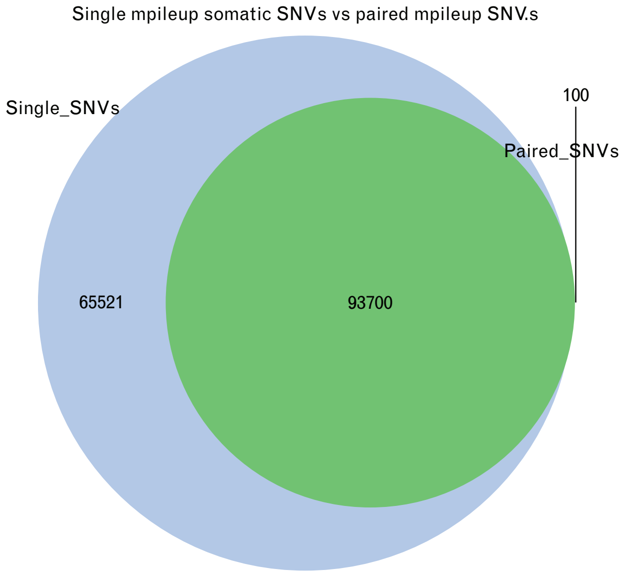

Fig. 6 Venn diagram depicting the overlap between the subtraction and simultaneous detection methods using SAMtools mpileup. 70% more candidate SNVs were identified with the subtraction method (SNV single) and virtually no new variants added with paired method (paired SNVs).

Melanoma tissue preparation for WGS is especially problematic for sequencing due to naturally occurring melanin which interferes with spectrometer-based DNA quantification and PCR reactions.[24] The presence of melanin in the library preparation can inhibit library preparation and may even induce sequencing artifact. We sought to remove pigmentation prior to library preparation using the CTAB clean-up method. Results show this process removes pigment and improves the quality of the genomic material on electrophoresis gels, which is then suitable for WGS. The results show that cleaning the DNA with CTAB results in minimal loss of genomic material as measured via Qubit and the copy number plots of the WGS of these samples appears to be consistent with non-pigmented tumour genomes. Therefore, a fluorometric based method should be used to accurately quantify double-stranded DNA concentrations in melanoma-derived DNA and a modified CTAB clean-up protocol can be used to produce suitable genomic material for WGS from heavily pigmented samples.

The Australian Melanoma Genome Project has performed WGS on 222 human melanoma samples and gained valuable insight into WGS and analysis of clinical melanoma samples. Data generated by the project will be made publicly available in order to make a substantial impact on future melanoma diagnosis and treatment. The current study outlines practical guidelines for sample preparation, quality control, sequencing methodology and mutation detection algorithms that will aid researchers and clinicians in sequencing patient-derived melanoma samples on a large scale.

Conflicts of interest and sources of funding: This work was supported by Melanoma Institute Australia, the Australian Government through Bioplatforms Australia, the NSW Ministry of Health, and a Cancer Council NSW project grant (RG12/13), as well as by Program grants (402761, 633004) of NHMRC (to JT, GM, RS, NH), Translational Research Program Grants (05TPG1-01, 10TPG1-02) of Cancer Institute NSW (to GM, JT, RS), Australian Research Council discovery grant (DP1301004 to GM), a NHMRC Research Fellowship (to NH) and a Cancer Institute NSW Clinical Research Fellowship and NHMRC Practitioner Fellowships (to RS).

Address for correspondence: Prof Richard A. Scolyer, Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital, Missenden Road, Camperdown, NSW 2050, Australia. E-mail: richard.scolyer@sswahs.nsw.gov.au

References

1. Scolyer RA, Thompson JF. Biospecimen banking: The pathway to personalized medicine for patients with cancer. *J Surg Oncol* 2013; 107: 681–2.
2. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time. *Genet Med* 2011; 13: 499–504.
3. Hayden EC. Is the $1,000 genome for real? *Nature News* 15 Jan 2014. http://www.nature.com/news/is-the-1-000-genome-for-real-1.14530.
4. Van Allen EM, Wagle N, Stojanov P, *et al*. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014; 20: 682–8.
5. Roychowdhury S, Iyer MK, Robinson DR, *et al*. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* 2011; 3: 111–21.
6. Biesecker LG, Green RC. Diagnostic clinical genome and exome sequencing. *N Engl J Med* 2014; 370: 2418–25.
7. Rehm HL, Bale SJ, Bayrak-Toydemir P, *et al*. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013; 15: 733–47.
8. Hirsch FR, Wynes MW, Gandara DR, Bunn PA. The tissue is the issue: personalized medicine for non-small cell lung cancer. *Clin Cancer Res* 2010; 16: 4909–11.
9. Long GV, Wilmott JS, Capper D, *et al*. Immunohistochemistry is highly sensitive and specific for the detection of V600E BRAF mutation in melanoma. *Am J Surg Pathol* 2013; 37: 61–5.
10. Song S, Nones K, Miller D, *et al*. Qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One* 2012; 7: e45835.
11. Van Loo P, Nordgard SH, Lingjaerde OC, *et al*. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010; 107: 16910–5.
12. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988; 2: 231–9.
13. Bentley DR, Balasubramanian S, Swerdlow HP, *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; 456: 53–9.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; 25: 1754–60.
15. Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; 25: 2078–9.
16. Ruffalo M, LaFramboise T, Koyuturk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 2011; 27: 2790–6.
17. Pabinger S, Dander A, Fischer M, *et al*. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2014; 15: 256–78.
18. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* 2014; 15: 244.
19. O'Rawe J, Jiang T, Sun G, *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013; 3: 28.
20. Mann GJ, Pupo GM, Campain AE, *et al*. BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J Invest Dermatol* 2013; 133: 509–17.
21. Viros A, Fridlyand J, Bauer J, *et al*. Improving melanoma classification by integrating genetic and morphologic features. *PLoS Med* 2008; 5: e120.
22. Wilmott JS, Long GV, Howle JR, *et al*. Selective BRAF inhibitors induce marked T-cell infiltration into human metastatic melanoma. *Clin Cancer Res* 2012; 18: 1386–94.
23. Long GV, Wilmott JS, Haydu LE, *et al*. Effects of BRAF inhibitors on human melanoma tissue before treatment, early during treatment, and on progression. *Pigment Cell Melanoma Res* 2013; 26: 499–508.
24. Stefania Lagonigro M, Cecco LD, Carninci P, *et al*. CTAB–Urea method purifies RNA from melanin for cDNA microarray analysis. *Pigment Cell Res* 2004; 17: 312–5.
25. Andrews TD, Whittle B, Field MA, *et al*. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biol* 2012; 2: 120061.

26. McLaren W, Pritchard B, Rios D, *et al*. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; 26: 2069–70.
27. Sherry ST, Ward MH, Kholodov M, *et al*. DbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; 29: 308–11.
28. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med* 2004; 10: 789–99.
29. Forbes SA, Bhamra G, Bamford S, *et al*. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* 2008; Chapter 10: Unit 10.11.
30. Murali R, Brown PT, Kefford RF, *et al*. Number of primary melanomas is an independent predictor of survival in patients with metastatic melanoma. *Cancer* 2012; 118: 4519–29.
31. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (Oct 13, 2011 edn, Vol. dbSNP 135: dbSNPv135). http://www.ncbi.nlm.nih.gov/SNP/.
32. Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* 2015; 161: 1681–96.
33. Bizon C, Spiegel M, Chasse SA, *et al*. Variant calling in low-coverage whole genome sequencing of a Native American population sample. *BMC Genomics* 2014; 15: 85.
34. Li Y, Sidore C, Kang HM, Boehnke MA, Becasis GR. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res* 2011; 21: 940–51.