Appendix A

Electronic files and supplementary information

A.1 Description of electronic files

I have placed these files in the accompanying CD, as they are large files that I do not think need to be printed. In this Appendix, I have written a short description detailing each of these, as well as the file name with which they can be located.

A.1.1 Pedigree structures for 24 melanoma-prone families sequenced as part of the discovery phase

File name: `Figure_A_1_1.pdf`

This PDF file contains pedigree structures for the 24 pedigrees that were sequenced as part of the discovery phase, one per page. Squares represent males, circles represent females, diamonds represent individuals of undisclosed sex. Individuals that had their whole exome sequenced are shown with a red outline. Types of cancer are indicated in the legend. Note that pedigrees have been adjusted to protect the identity of the families without loss of scientific integrity.

A.1.2 Detailed information about recurrently mutated genes

File name: `Table_A_1_2.xlsx`

This table contains the list of 344 genes that had mutations in two or more pedigrees. The number of families with mutations, the number of members in each family and the

coding length of the gene are indicated.

A.1.3 List of biological pathways ordered by *P*-value after hypergeometric tests on list of recurrently mutated genes

File name: `Table_A_1_3.xlsx`

This table contains the results of the analysis to identify overrepresented biological pathways in our set of 344 recurrently mutated genes. All pathways with a P-value<0.5 are in this table, pathways in green were the ones included in the list for targeted sequencing. Note that the acute myocardial infarction pathway from Biocarta has an adjusted P-value < 0.05 and was not included in the set for targeted sequencing; the reason is that the original analysis was performed with the R package HTSanalyzeR version 2.8.0, which had an error in the code for the hypergeometric function. This error was corrected for the next version (2.8.3), which is the one I used to generate the table presented here.

A.1.4 All genes included for capture in the replication phase

File name: `Table_A_1_4.xlsx`

All 701 genes in that were targeted for sequencing in an additional set of 94 melanoma cases. The Ensembl identifier, along with the coding length and the reason for inclusion are indicated. "Direct evidence" means the gene was mutated in two or more pedigrees, "ABC transporter", "Pantothenate and CoA biosynthesis pathway" and "Linkage to MAPK signalling for integrins" indicate that the gene was included because it is part of a pathway that was found to be overrepresented in the recurrently mutated genes. "Disruptive consequence" means that the gene was included because it had a premature stop codon, a frameshift variant or a splice acceptor or donor variant in any one of the familial melanoma pedigrees. Genes marked as "previous evidence for involvement in melanoma/cancer" were included because they have been found linked to processes relevant to melanoma.

A.1.5 List of ranked genes after prioritisation stage

File name: `Table_A_1_5.xlsx`

This table includes the genes captured in both the discovery and replication phases, with their respective values for the variables used in the prioritisation method. The

genes highlighted in red are regarded as uninformative, as they have either a single variant detected in the melanoma pedigrees and no variants in the control exomes, or were found to have a coding length of 0.

A.1.6 List of variants in the analysis of founder mutations and those in single pedigrees

File name: `Table_A_1_6.xlsx`

This table includes the list of variants predicted to affect protein function that were present in more than one pedigree for which co-segregation information was available. The position of the variants, the number of pedigrees with their respective number of members and the consequences of the variants are indicated.

A.1.7 Genes found with only one variant in multi-case pedigrees

File name: `Table_A_1_7.xlsx`

List of the genes that were found with variants segregating with melanoma in only one pedigree, and that were not present in the prioritisation stage.

A.1.8 Pedigrees sequenced as part of the integrative phase

File name: `Table_A_1_8.xlsx`

The number of cases in each pedigree is indicated, as well as the number of cases sequenced (exomes or whole genomes). Whole genomes were only sequenced as part of the QFMP cohort.

A.1.9 Genes with co-segregating variants from the 28 pedigrees for which we had sequence data for 3 or more members and their GO terms

File name: `Table_A_1_9.xlsx`

Genes that had variants co-segregating with melanoma from the 28 pedigrees for which we had sequence data for 3 or more members and their GO terms.

A.1.10 Wide variation in telomere measurements when samples have not been processed in the same manner

File name: `Figure_A_1_10.pdf`

This graph shows the telomere measurements for the 41 samples that belong to the discovery phase cohort and that were sequenced at the Sanger Institute alongside telomere length estimates for samples part of the UK10K, sequenced at the same institute (all shown in a white background), and samples processed at QFMP (in a blue background, sample origin is indicated at the bottom). Samples with *POT1* variants are indicated with red arrows, two whole genomes part of the QFMP cohort are indicated with green arrows. Samples within the QFMP cohort showed much more variability in telomere length measurement and thus could not be assessed with the bioinformatic method.

A.2 Supplementary tables, figures and notes

A.2.1 Removal of genes likely to be false positives after filtering

I manually scanned the list of candidates after filtering and decided to remove four that are likely to be false positives (titin [*TTN*], obscurin [*OBSCN*], dystrophin [*DMD*] and maestro heat-like repeat family member 2A [*MROH2A*]) due to their length and/or their ubiquity in other cancer screens (based on analyses performed by Vertebrate Resequencing Informatics at the Sanger). I also inspected the variants for their presence in repeat regions, and a further five genes were excluded given that the mutations we detected on these overlapped with the RepeatMasker track from the University of California, Santa Cruz (UCSC) Genome Browser [511]: (lysine-specific demethylase 6B [*KDM6B*], WD repeat domain 87 [*WDR87*], Zinc finger protein 589 [*ZNF589*], choline kinase alpha [*CHKA*] and abnormal spindle homolog, microcephaly associated [*ASPM*]). This left 344 genes for further consideration.

Table A.1: Tools and parameters used for read alignment and variant calling in the discovery and replication phases

Step	Reference dataset	Tool	Version	Parameters
Discovery phase				
Read alignment to reference genome	GRCh37	BWA [377]	0.5.8c, 0.5.9	`-q 15 -t 6`
Alignment improvement				
Duplicate marking	-	Picard MarkDuplicates [378]	1.47	-
Indel realignment	dbSNP 129	GATK IndelRealigner [379]	1.1-5	`-LOD 0.4 -model KNOWNS_ONLY -entropy 0.15`
Quality score recalibration		GATK TableRecalibration [379]	1.1-5	-
Variant calling	-	SAMtools mpileup [380]	0.1.17	`-DRS -d 10000 -C50 -m2 -F0.0005 -aug -P ILLUMINA`
Replication phase				
Read alignment to reference genome	GRCh37d5	BWA [377]	0.5.9	`-q 15 -t 6`
Alignment improvement				
Duplicate marking	-	Picard MarkDuplicates [378]	1.5-9	-
Indel realignment	1000 Genomes Phase 2 Indels [15]	GATK IndelRealigner GATK CountCovariates [379]	1.5-9	`-LOD 0.4 -model KNOWNS_ONLY`
Quality score recalibration	-	GATK TableRecalibration [379]	1.5-9	-
Variant calling	-	SAMtools mpileup / Bcftools [380]	SAMTools: 0.1.18 Bcftools: 0.1.17-dev	SAMtools: `-EDS -d 10000 -C50 -m2 -F0.0005` Bcftools: `view -p 0.99 -vcgN`

Appendix B

Articles published during my PhD

During the course of my PhD, I was part of several publications, that were either directly associated with the work described in this dissertation or represented other work I carried out. In this Appendix, I list these publications and provide a short explanation of their main findings and my contribution to each of them. The physical articles can be found at the end of the dissertation. An asterisk denotes that those authors contributed equally.

B.1 Articles directly associated with this dissertation

- Robles-Espinoza CD*, Harland M*, Ramsay AJ*, Aoude LG*, Quesada V, Ding Z, Pooley KA, Pritchard AL, Tiffen JC, Petljak M, Palmer JM, Symmons J, Johansson P, Stark MS, Gartside MG, Snowden H, Montgomery GW, Martin NG, Liu JZ, Choi J, Makowski M, Brown KM, Dunning AM, Keane TM, López-Otín C, Gruis NA, Hayward NK, Bishop DT, Newton-Bishop JA and Adams DJ (2014). *POT1* loss-of-function variants predispose to familial melanoma. *Nature Genetics* **45**(5): 478-81. doi: 10.1038/ng.2947.

This is the main publication explaining the results from my dissertation. This letter explains the finding of familial melanoma pedigrees with rare variants in *POT1* and their consequences in carriers. I performed most of the bioinformatic analyses, including data processing and filtering, analysis and comparison with control exomes, computational assessment of variant conservation and pathogenicity, and analysis of telomere length data.

- Aoude LG*, Pritchard AL*, Robles-Espinoza CD*, Wadt K*, Harland M*, Choi J, Gartside M, Quesada V, Johansson P, Palmer JM, Ramsay AJ, Zhang X, Jones K, Symmons J, Holland EA, Schmid H, Bonazzi V, Woods S, Dutton-Regester K, Stark MS, Snowden H, van Doorn R, Montgomery GW, Martin NG, Keane TM, López-Otín C, Gerdes AM, Olsson H, Ingvar C, Borg A, Gruis NA, Trent JM, Jonsson G, Bishop DT, Mann GJ, Newton-Bishop JA, Brown KM, Adams DJ and Hayward NK (2014). Nonsense mutations in the shelterin complex genes *ACD* and *TERF2IP* in familial melanoma. *Journal of the National Cancer Institute* (Accepted for publication).

This publication explains an extended search for variants in members of the shelterin complex in 510 melanoma-prone pedigrees from Australia, UK, The Netherlands, Denmark and Sweden. We found additional nonsense variants that co-segregate with the disease in *ACD* and *TERF2IP*, providing further support for telomere dysregulation as an important contributor to this phenotype. I did the data analysis for samples from the UK and The Netherlands, as well as comparisons with control exomes.

- Robles-Espinoza CD, del Castillo Velasco-Herrera M, Hayward NK and Adams DJ (2014). Telomere-regulating genes and the telomere interactome in familial cancers. *Molecular Cancer Research* (Accepted for publication).

This review explores the role that telomere-regulating proteins, including members of the shelterin complex, have in cancer predisposition. The structures and functions of telomerase, shelterin and the telomere interactome are discussed. I wrote most of this review.

- Rashid M, Robles-Espinoza CD, Rust AG and Adams DJ (2013). Cake: A bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics* **29**(17): 2208-10. doi: 10.1093/bioinformatics/btt371.

This application note provides a tool for somatic variant discovery from NGS data in cancer genomes. It combines other software tools previously developed and provides a strategy to optimise the sensitivity and accuracy of candidate variant calls when compared to any of these tools alone. I wrote some of the functions in this piece of software, which were adapted mainly from the analyses I did during the discovery phase of this study.

B.2 Other articles

- Robles-Espinoza CD and Adams DJ (2013). Cross-species analysis of mouse and human cancer genomes. *Cold Spring Harbor Protocol*s **2014**(4).

  doi: 10.1101/pdb.top078824

- van der Weyden L, Rust AG, McIntyre RE, Robles-Espinoza CD, del Castillo Velasco-Herrera M, Strogantsev R, Ferguson-Smith AC, McCarthy S, Keane TM, Arends MJ and Adams DJ (2013). *Jdp2* downregulates *Trp53* transcription to promote leukaemogenesis in the context of *Trp53* heterozygosity. *Oncogene* **32**(3): 397-402. doi: 10.1038/onc.2012.56