Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor

Kevin Litchfield[1], Max Levy[1], Giulia Orlando[1], Chey Loveday[1], Philip Law[1], Gabriele Migliorini[1], Amy Holroyd[1], Peter Broderick[1], Robert Karlsson[2], Trine B Haugen[3], Wenche Kristiansen[3], Jérémie Nsengimana[4], Kerry Fenwick[5], Ioannis Assiotis[5], ZSofia Kote-Jarai[1], Alison M. Dunning[6], Kenneth Muir[8,9], Julian Peto[10], Rosalind Eeles[1,11], Douglas F Easton[6,7], Darshna Dudakia[1], Nick Orr[12], Nora Pashayan[13], UK Testicular Cancer Collaboration*, The PRACTICAL consortium*, D. Timothy Bishop[4], Alison Reid[14], Robert A Huddart[14], Janet Shipley[15], Tom Grotmol[16], Fredrik Wiklund[2], Richard S Houlston[1], Clare Turnbull[1,17]

1. Division of Genetics & Epidemiology, The Institute of Cancer Research, London, SM2 5NG, UK
2. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 171 77, Sweden
3. Faculty of Health Sciences, Oslo and Akershus University College of Applied Sciences, Oslo, Norway
4. Section of Epidemiology & Biostatistics, Leeds Institute of Cancer and Pathology, Leeds, LS9 7TF, UK
5. Tumour Profiling Unit, The Institute of Cancer Research, London, SM2 5NG, UK
6. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, CB1 8RN, UK
7. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK
8. Division of Health Sciences, Warwick Medical School, Warwick University, CV4 7AL, UK
9. Institute of Population Health, University of Manchester, M1 3BB, UK
10. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom.

11. Royal Marsden NHS Foundation Trust, London, SM2 5NG, UK
12. The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK
13. Department of Applied Health Research, University College London, London, WC1E 6BT, UK
14. Academic Radiotherapy Unit, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK
15. Division of Molecular Pathology, The Institute of Cancer Research, London, SM2 5NG, UK
16. Department of Research, Cancer Registry of Norway, Oslo, 0369, Norway
17. William Harvey Research Institute, Queen Mary University, London, EC1M 6BQ , UK

* See supplementary notes 1 and 2

Correspondence to: Clare Turnbull, Division of Genetics and Epidemiology, The Institute of Cancer Research, London, SM2 5NG, UK; Tel: ++44 (0) 208 722 4485; E-mail: clare.turnbull@icr.ac.uk

Key words: Testicular Cancer, Germ Cell Tumour, TGCT, GWAS, Oncoarray.

Genome-wide association studies (GWAS) have transformed our understanding of testicular germ cell tumour (TGCT) susceptibility but much of the heritability remains unexplained. Here we report a new GWAS, a meta-analysis with previous GWAS and a replication series, totalling 7,319 TGCT cases and 23,082 controls. We identify 19 new TGCT risk loci, approximately doubling the number of known TGCT risk loci to 44. By performing *in-situ* Hi-C in TGCT cells, we provide evidence for a network of physical interactions between all 44 TGCT risk SNPs and candidate causal genes. Our findings reveal widespread disruption of developmental transcriptional regulators as a basis of TGCT susceptibility, consistent with failed primordial germ cell differentiation as an initiating step in oncogenesis[1]. Defective microtubule assembly and dysregulation of KIT-MAPK signalling also feature as recurrently disrupted pathways. Our findings support a polygenic model of risk and provide insight into the biological basis of TGCT.

Testicular germ cell tumour (TGCT) is the most common cancer in men aged 18-45, with over 52,000 new cases diagnosed annually worldwide[2]. The development of TGCT is strongly influenced by inherited genetic factors, which contributes to nearly half of all disease risk[3] and is reflected in the 4-to-8 fold increased risk shown in siblings of cases[4-7]. Our understanding of TGCT susceptibility has been transformed by recent genome-wide association studies (GWAS), which have so far identified 25 independent risk loci for TGCT[8-18]. Although projections indicate that additional risk variants for TGCT can be discovered by GWAS[19], studies to date have been based on comparatively small sample sizes which have had limited power to detect common risk variants[20].

To gain a more comprehensive insight into TGCT aetiology we performed a new GWAS with substantially increased power, followed by a meta-analysis with existing GWAS and replication genotyping (totalling 7,319 cases/23,082 controls). Here we report both the discovery of 19 new TGCT susceptibility loci and refined risk estimates for the previously reported loci. In addition, we have investigated the gene regulatory mechanisms underlying the genetic associations observed at all 44 TGCT GWAS risk loci by performing *in-situ* chromosome conformation capture in TGCT cells (Hi-C) to characterize chromatin interactions between predisposition SNPs and target genes, integrating these data with a range of publicly available TGCT functional genomics data.

We conducted a new GWAS using the Oncoarray platform (3,206 UK TGCT cases/7,422 UK controls), followed by a meta-analysis combining the two largest published TGCT GWAS datasets[11,16] (986 UK cases/4,946 UK controls, 1,327 Scandinavian cases/6,687 Scandinavian controls) (**Fig. 1**). To increase genomic resolution, we imputed >10 million SNPs using the 1000 Genomes Project as a reference panel. Quantile-Quantile (Q-Q) plots for SNPs with minor allele frequency (MAF) >5% post imputation did not show evidence of substantive over-dispersion (λ_{1000}=1.03, **Supplementary Fig. 1**). We derived joint odds ratios (ORs) and 95% confidence intervals (CIs) under a fixed-effects model for each SNP with MAF >0.01. Finally we sought validation of 37 SNPs associated at $P < 5.0 \times 10^{-6}$, which

did not map to known TGCT risk loci and displayed a consistent OR across all GWAS datasets, by genotyping an additional 1,801 TGCT cases and 4,027 controls from the UK. After meta-analysis of the three GWAS and replication series, we identified genome-wide significant associations (*i.e*. $P < 5 \times 10^{-8}$) at 19 new loci (**Table 1**). We found no evidence for significant interactions between risk loci.

To the extent that they have been deciphered, many GWAS risk loci map to non-coding regions of the genome and influence gene regulation. Across the 44 independent TGCT risk loci (19 new and 25 previously reported), we confirmed a significant enrichment of enhancer/promoter associated histone marks, including H3K4me1, H3K4me3 and H3K9ac, using available ChIP-Seq data from the TGCT cell line NTERA2 ($P<5.0\times10^{-3}$) (**Supplementary Table 1**). Moreover this enrichment showed tissue specificity when compared to 41 other cell lines from the ENCODE[21] project (**Supplementary Fig. 2**). These observations support the assertion that the TGCT predisposition loci influence risk through effects on *cis*-regulatory networks, and are involved in transcriptional initiation and enhancement. Since genomic spatial proximity and chromatin looping interactions are fundamental for regulation of gene expression we performed *in situ* capture Hi-C of promoters in NTERA2 cells to link risk loci to candidate target genes. We also sought to gain insight into the possible biological mechanisms for the associations by performing tissue-specific expression quantitative trait loci (eQTL) analysis for all risk SNP and target gene pairs (**Supplementary Fig. 3, Supplementary Table 2**). We analysed RNA-seq data from both normal testis (GTEx project[22]) and TGCT (TCGA), acknowledging that the latter may be affected by the issue of tumour purity, in addition to dysregulated gene expression that typifies cancer. Accepting this limitation and that further validation may be required, eQTL analysis was conducted in both datasets based on the established network of enhancer/ promoter variants, to maximise our ability to find statistically significant associations after correcting for multiple testing. We additionally annotated risk loci with variants predicted to disrupt binding motifs of germ cell specific transcription factors (TF) (see methods). Finally, direct promoter variants and non-synonymous coding mutations for genes within the 44 risk loci were denoted (**Table 2, Fig. 2**).

Although preliminary and requiring functional validation, three candidate disease mechanisms emerge from analysis across the 44 loci. Firstly, 10 of the risk loci contain candidate genes linked to developmental transcriptional regulation, as evidenced by Hi-C looping interactions (at 8p23.1, 20q13.2), eQTL effects (at 4q22.3, 8p23.1), promoter variants (at 8q13.3, 9p24.3, 12q15, 17q12, 19p12) and coding variants (at 2p13.3, 16q24.2) (**Table 2**). Notably the new TGCT risk locus at 8p23.1 features a looping chromatin interaction from risk SNP rs17153755 to the promoter of *GATA4*, which is supported by an overlapping predicted strong enhancer region and a nominal eQTL effect (TCGA data, P=3.1 x 10^{-2}) (**Fig. 3a**). The rs17153755 risk allele was associated with down-regulation of *GATA4* expression, consistent with the hypothesised role of *GATA4* as a tumor suppressor gene[23,24]. In addition the risk locus at 16q24.2 only contains a single gene *ZFPM1* (alias FOG, Friend of GATA1), which encodes an essential regulator of *GATA1*[25], in which we noted a predicted damaging [26] missense polymorphism (rs3751673, NP_722520.2:p.Arg22Gly). The GATA family of transcription factors are expressed throughout postnatal testicular development[27], and play a key role in ensuring correct tissue specification and differentiation[28]. We also observed promoter variants at 8q13.3 and 9p24.3, providing support respectively for the role of *PRDM14* and *DMRT1* in TGCT oncogenesis, both of which encode important transcriptional regulators of germ cell specification and sex determination[29-32]. Of final note the new locus at 20q13.2 was characterized by a predicted disrupted POU5F1 binding motif, together with a looping Hi-C contact from risk SNP rs12481572 to the promoter of *SALL4*, a gene associated with the maintenance of pluripotency in embryonic stem cells[33].

Secondly, candidate genes with roles related to microtubule/chromosomal assembly were implicated at five TGCT risk loci, supported by Hi-C looping interactions (at 1q22, 15q25.2), eQTL effects (at 15q25.2, 17q22), promoter variants (at 1q22, 4q24) and coding variants (at 21q22.3). Notably at locus 17q22 we observed a promoter variant (rs302875) which displays a strong eQTL

effect (GTEx data, P=4.9 x 10^{-7}) on *TEX14* (Testis-Expressed 14), which encodes an important regulator of kinetochore-microtubule assembly in testicular germ cells[14,34,35]. At new risk locus 15q25.2 we identified a nominal eQTL association (rs2304416, TCGA data, P=3.2 x 10^{-2}) and accompanying chromatin looping interaction with mitotic spindle assembly related gene *WDR73*[36] (**Fig. 3b**). *WDR73* encodes a protein with a crucial role in the regulation of microtubule organization during interphase[37] and biallelic mutations cause Galloway-Mowat Syndrome, a human syndrome of nephrosis and neuronal dysmigration. Finally the functional analysis also highlighted microtubule assembly related genes *PMF1*, *CENPE* and *PCNT* [38-41] as candidates at 1q22, 4q24 and 21q22.3 respectively.

Thirdly, the central role of KIT-MAPK signalling in TGCT oncogenesis was further supported at four loci, by Hi-C looping interactions (at 11q14.1, 15q22.31), eQTL effects (at 6p21.31) and promoter variants (at 6p21.31, 11q14.1, 15q22.31). Recent tumour sequencing studies have established that *KIT* is the major somatic driver gene for TGCT[42] and a relationship between the previously identified risk SNP rs995030 (12q21) and *KITLG* expression has been demonstrated through allele-specific p53 binding by Zeron-Medina et al[43]. Here we report a new locus at 15q22.31, containing a variant within the promoter of *MAP2K1* (**Fig. 3c**), which raises the prospect of further elucidating mechanisms of KIT-MAPK signalling in driving TGCTs. *MAP2K1* (alias *MEK1*) is downstream of c-Kit and MEK1 inhibition slows primordial germ cell growth in the presence of KIT ligand[44]. If *MAP2K1* is confirmed as a causal gene at 15q22.31, the study of somatic *KIT* mutational status in patients carrying the risk allele at 15q22.31 should be highly informative. In addition, within the 11q14.1 risk locus, we identify a candidate promoter variant for *GAB2,* which encodes a docking protein for signal transduction to MAPK and PI3K pathways which interacts directly with KIT[45]. Finally in our analysis we identify both a candidate promoter variant and a nominal eQTL effect for *BAK1* (6p21.31)(TCGA data, P=1.9 x 10^{-2}), which encodes a protein regulating apoptosis which binds with KIT [40]. While we have sought to decipher the functional basis of risk loci based on the cumulative weight of evidence across eQTL, Hi-C and ChIP-seq data, a limitation has been reliance on relatively small sample size

for eQTL analysis. Access to larger eQTL datasets in testicular tissue are likely in the future to address this deficiency enabling a better definition of the causal basis of TGCT risk at each locus.

The 44 risk loci which have now been identified for TGCT collectively account for 34% of the (father-to-son) familial risk and hence have potential clinical utility for personalized risk profiling. To assess this potential, we constructed polygenic risk scores (PRS) for TGCT, considering the combined effect of all risk SNPs modelled under a log-normal relative risk distribution. Using this approach the men in the top 1% of genetic risk have a relative risk of 14 which translates to a 7% lifetime risk of TGCT (**Supplementary Fig. 4**).

In summary, we have performed a new TGCT GWAS, identifying 19 new risk loci for TGCT, approximately doubling the number of previously reported SNPs. Using capture Hi-C we have generated a chromatin interaction map for TGCT, providing direct physical interactions between non-coding risk SNPs and target gene promoters. Moreover integration of these data together with ChIP-seq chromatin profiling and RNA-seq eQTL analysis, accepting certain caveats, has allowed us to gain preliminary but unbiased tissue-specific insight into the biological basis of TGCT susceptibility. This analysis suggests a model of TGCT susceptibility based on transcriptional dysregulation, which is likely to contribute to the developmental arrest of primordial germ cells coupled with chromosomal instability through defective microtubule function and accompanied upregulation of KIT-MAPK signalling.

METHODS

Sample description

TGCT cases were from the UK (n=5,992) and Scandinavia (n=1,327). The UK cases were ascertained from two studies (1) a UK study of familial testicular cancer and (2) a systematic collection of UK collection of TGCT cases. Case recruitment was via the UK Testicular Cancer Collaboration, a group of oncologists and surgeons treating TGCT in the UK (**Supplementary note 1**). The studies were co-ordinated at the Institute of Cancer Research (ICR). Samples and information were obtained with full informed consent and Medical Research and Ethics Committee approval (MREC02/06/66 and 06/MRE06/41). Additional (n=1,327) case samples of Scandinavian origin were used from a previously published GWAS[16].

Control samples for the primary GWAS were all taken from within the UK. Specifically 2,976 cancer-free, male controls were recruited through two studies within the PRACTICAL Consortium (**Supplementary note 2**): (1) the UK Genetic Prostate Cancer Study (UKGPCS) (age <65), a study conducted through the Royal Marsden NHS Foundation Trust and (2) SEARCH (Study of Epidemiology & Risk Factors in Cancer), recruited via GP practices in East Anglia (2003-2009). 4,446 cancer-free female controls from across the UK were recruited via the Breast Cancer Association Consortium (BCAC). Controls from the UK previously published GWAS[11] were from two sources within the UK: 2,482 controls were from the 1958 Birth Cohort (1958BC), and 2,587 controls were identified through the UK National Blood Service (NBS) and were genotyped as part of the Wellcome Trust Case Control Consortium. Additional (n=6,687) control samples of Scandinavian origin were used in the meta-analysis, and have been previously described[16]. Control samples for replication genotyping (n=4,027) were taken from two studies, the national study of colorectal cancer genetics (NSCCG)[46] and GEnetic Lung CAncer Predisposition Study (GELCAPS)[47]. NSCCG and GELCAP controls were spouses of cancer patients with no personal history of cancer at time of ascertainment.

Primary GWAS

Genotyping was conducted using a custom Infinium OncoArray-500K BeadChip (Oncoarray) from Illumina (Illumina, San Diego, CA, USA), comprising a 250K SNP genome-wide backbone and 250K SNP custom content selected across multiple consortia within COGS (Collaborative Oncological Gene-environment Study). Oncoarray genotyping was conducted in accordance with the manufacturer's recommendations by the Edinburgh Clinical Research Facility, Wellcome Trust CRF, Western General Hospital, Edinburgh EH4 2XU.

Published GWAS

The UK and Scandinavian GWAS have been previously reported[8,11,13]. Briefly the UK GWAS comprised 986 cases genotyped on the Illumina HumanCNV370-Duo bead array (Ilumina, San Diego, CA, USA) and 4,946 controls genotyped on the Illumina Infinium 1.2M array. We analysed data on a common set of 314,861 SNPs successfully genotyped by both arrays. The Scandinavian GWAS [16], comprised 1,326 cases and 6,687 controls genotyped using the Human OmniExpressExome-8v1 Illumina array.

Quality Control of GWAS

Oncoarray data was filtered as follows, we excluded individuals with low call rate (<95%), with abnormal autosomal heterozygosity or with >10% non-European ancestry (based on multi-dimensional scaling). We filtered out all SNPs with minor allele frequency <1%, a call rate of <95% in cases or controls or with a minor allele frequency of 1–5% and a call rate of <99%, and SNPs deviating from Hardy-Weinberg equilibrium (10^{-12} in controls and 10^{-5} in cases). The final number of SNPs passing quality control filters was 371,504. Quality control (QC) procedures for the UK and Scandinavian GWAS have been previously described[8,11,13,16].

Imputation

Genome-wide imputation was performed for all GWAS datasets. The 1000 genomes phase 1 data (Sept-13 release) was used as a reference panel, with haplotypes pre-phased using SHAPEIT2[48]. Imputation was performed using IMPUTE2 software[49] and association between imputed genotype and TGCT was tested using SNPTEST [50], under a frequentist model of association. QC was performed on the imputed SNPs; excluding those with INFO score < 0.8 and MAF < 0.01.

Replication genotyping

Replication genotyping of the 37 SNPs was performed by allele-specific KASPar allele-specific SNV primers[51]. Genotyping was conducted by LGC Limited, Unit 1-2 Trident Industrial Estate, Pindar Road, Hoddesdon, UK.

Statistical Analysis

Study sample size was chosen in order to achieve >50% power to detect common variants, defined as MAF > 5%, OR > 1.3[20]. For Oncoarray data tests of association between imputed SNPs and TGCT was performed under a probabilistic dosage model in in SNPTESTv2.5[52], adjusting for principal components. Inflation in the test statistics was observed at only modest levels, λ_{1000}=1.03. The inflation factor λ was based on the 90% least-significant SNPs[53]. The adequacy of the case-control matching and possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of test statistics (**Supplementary Fig. 1**). Population ancestry structure for the UK and Scandinavian cohorts was assessed through visualisation of the first two principle components (**Supplementary Fig. 5**); stable ancestral clustering was observed (**Supplementary Table 3**).

Statistical analysis of previously reported GWAS was performed as previously described[8,11,13,16,54]. Meta-analyses were performed using the fixed-effects inverse-variance method based on the β estimates and standard errors from each study using META v1.6[55]. Cochran's Q-statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated[56]. For each new locus we examined evidence of departure from a log-additive (multiplicative) model, to assess any genotype specific effect. Using the Oncoarray data individual genotype data ORs were calculated for heterozygote (OR_{het}) and homozygote (OR_{hom}) genotypes, which were compared to the per allele ORs. We tested for a difference in these 1d.f. and 2d.f. logistic regression models to assess for evidence of deviation (P<0.05) from a log-additive model. Using Oncoarray data we examined for statistical interaction between any of the 44 TGCT predisposition loci by evaluating the effect of adding an interaction term to the regression model, adjusted for stage, using a likelihood ratio test (using a significance threshold of $P < 2.58 \times 10^{-5}$ to account for 1,936 tests). Regional plots were generated using visPIG software[57] (**Supplementary Fig. 6**). Polygenic risk scores (PRS) were constructed using the methodology of Pharoah et al[58], based on a log-normal distribution LN (μ, σ^2) with mean μ and variance σ^2 (*i.e.* relative risk is normally distributed on a logarithmic scale). The 0.5% lifetime risk of TGCT risk was based on 2014 UK data[59], multiplied by relative risk to give lifetime risk per percentile of the PRS. For calculation of the proportion of TGCT genetic risk explained by the 44 loci, a father-to-son relative risk of four was used.

Chromatin mark enrichment analysis

To examine enrichment in specific ChIP-seq tracks across risk loci we adapted the variant set enrichment method of Cowper-Sal lari *et al*[60]. Briefly, for each risk locus, a region of strong LD was defined (*i.e.* $R^2 > 0.8$ and D' > 0.8), and SNPs mapping to these regions were termed the associated variant set (AVS). Histone ChIP-seq uniform peak data was obtained from ENCODE[21] for the NTERA2

cell line, and data was included for four histone marks. For each of these marks, the overlap of the SNPs in the AVS and the binding sites was determined to produce a mapping tally. A null distribution was produced by randomly selecting SNPs with the same LD structure as the risk associated SNPs, and the null mapping tally calculated. This process was repeated 10,000 times, and approximate *P*-values were calculated as the proportion of permutations where null mapping tally was greater or equal to the AVS mapping tally. An enrichment score was calculated by normalizing the tallies to the median of the null distribution. Thus the enrichment score is the number of standard deviations of the AVS mapping tally from the mean of the null distribution tallies. Tissue specificity was assessed by comparison of enrichment levels in NTERA2, compared to 41 other cell lines from ENCODE[21], with analysis performed using the same method as above (**Supplementary Fig. 2**).

Promoter Hi-C

In situ Hi-C libraries were prepared as described by Rao et al.[61] with the following modifications: (i) 25 million cells were fixed and processed; (ii) HindIII enzyme (NEB, Ipswich, MA, USA) was used and digestion was performed overnight; (iii) ligation was performed overnight at 16C; (iv) 3 µl of 15 µM annealed PE adaptors were ligated incubating 3 µl of T4 DNA ligase (NEB, Ipswich, MA, USA) for 2h at RT; (vi) 6 cycles of PCR were performed to amplify the libraries before capture. A Sure Select (Agilent, Santa Clara, CA, USA) custom promoter kit was used to perform capture with the same design as described by Misfud *et al*.[62]. For each capture reaction, 750 µg of Hi-C libraries were used. Capture was performed following the manufacture protocol and employing a custom reagent kit (Agilent, Santa Clara, CA, USA). Final PCR amplification was performed using 5 cycles to minimise PCR duplicates. 2x100bp sequencing was performed using Illumina HiSeq2000 or 2500 technology (Illumina, San Diego, CA, USA). The HiCUP pipeline[63] was used to process raw sequencing reads, map di-tag positions against the reference human genome and remove duplicate reads. The protocol was performed for two independent NTERA2 biological replicates, with cells obtained from the

laboratory of Prof. Janet Shipley (The Institute of Cancer Research, London) and their identity independently confirmed through STR typing at an external laboratory (Public Health England, Porton Down, UK). Cells were tested and found to be negative for mycoplasma contamination. Both Hi-C libraries achieving the following quality control thresholds: >80% reads uniquely aligning, >80% valid pair rate, >85% unique di-tag rate and >80% of interactions being *cis* (**Supplementary Table 4**). Statistically significant interactions were called using the CHiCAGO pipeline[64], with both biological replicates processed in parallel to obtain a unique list of reproducible NTERA2 contacts. Stability of results across replicates was also verified by processing each sample individually and comparing the significance scores of called interactions; strong correlation was observed between the replicates (r = 0.8, $P < 5.0 \times 10^{-10}$, **Supplementary Fig. 7**). Interactions with a -log(weighted *P*-value) > 5 were considered significant. To avoid short-range proximity bias interactions of <40kb were excluded. The distribution of interaction distances closely matched the prior published dataset of Misfud *et al*.[62] (**Supplementary Fig. 8**). A Hi-C track plotting read pair counts per HindIII fragment has been added to region plot figures to demonstrate the underlying signal strength of significant Hi-C contacts.

3C Validation

3C was used to validate selected chromatin interactions detected by CHi-C (3p24.3, 4q24, 11q14.1, 15q22.31, 15q25.2, 16q12.1, and 16q23.1) (**Supplementary Fig. 9, Supplementary Table 5**). Three replicates of *in situ* 3C libraries were prepared using NTERA2 cells. Cell pellets were crosslinked, digested with HindIII, and ligated. Libraries were purified by phenol-chloroform extraction.

For each loci one or more bacterial artificial chromosomes (BACs; Source BioScience, Nottingham, UK) were used as an internal standard (**Supplementary Table 6**). Clones were streaked and grown before extracting DNA using a QIAGEN Plasmid Maxi Kit (QIAGEN, Hilden, Germany) which was purified by phenol-chloroform extraction. In loci covered by more than one clone, equimolar

solutions of clones were prepared. Randomly ligated 3C libraries were generated for each BAC or equimolar solution of BACs.

Unidirectional primer pairs were designed to amplify ligation junctions of the bait and other interacting HindIII fragment (promoter-element, P-E) and around the bait and a flanking control HindIII fragment in between the promoter and distal element (promoter-control, P-C) using Primer3[65] (**Supplementary Tables 7 and 8**). Regions were amplified using both P-E and P-C primer pairs in BAC and NTERA2 libraries using a QIAGEN Multiplex PCR Kit (QIAGEN, Hilden, Germany). 5 ng and 100 ng of BAC and NTERA2 library template DNA, respectively, were amplified using the following procedure: initial 15 minute denaturation at 95°C followed by 38 cycles of 94°C for 0.5 minutes, annealing temperature specific to primer pair for 1.5 minutes seconds, 72°C extension for 1.5 minutes, followed by a final 10 minute extension at 72°C extension. 5 µl of each PCR reaction was visualised on 2% agarose gels stained with ethidium bromide. ImageJ[66] was used to quantify intensities of PCR products and normalise for differential primer efficiency by comparing to equimolar BAC PCR products.

P-E fragments were Sanger sequenced in NTERA2 libraries to confirm fragments visualised on agarose gels as expected (**Supplementary Fig. 10**).

Chromatin state annotation

We used ChromHMM[67] to infer chromatin states by integrating information on histone modifications and DNaseI hypersensitivity data to identify combinatorial and spatial patterns of epigenetic marks. Aligned next generation sequencing reads from ChIP-Seq and DNAse-Seq experiments on the NTERA2 cells were downloaded from ENCODE[21]. Read-shift parameters for ChIP-Seq data were calculated using PHANTOMPEAKQUALTOOLS. Genome-wide signal tracks were binarized (including input controls for ChIP-Seq data) and a set of learned models were generated using ChromHMM software[67]. The parameters of the highest scoring model were retained and model states were

iteratively reduced down from 30 to 5 states. A 27-state model found to be stable and was subsequently used for segmenting the genome at 200bp resolution (**Supplementary Fig. 11**).

Expression quantitative trait locus analysis

We investigated for evidence of association between the SNPs at each locus and tissue specific changes in gene expression using two publically available resources: (i) RNAseq and Affymetrix 6.0 SNP data for 150 TGCT patients from The Cancer Genome Atlas and (ii) normal testicular tissue data from GTEx from 157 samples[22]. Associations between normalized RNA counts per-gene and genotype were quantified using R package 'Matrix eQTL'. Box plots of all eQTL associations are presented in **Supplementary Fig. 3** and the tissue in which the association was observed (TGCT or normal testis), along with any other tissues resulting in a positive association, are denoted in **Supplementary Table 2**. To reduce multiple testing, association tests were only performed between SNP and gene pairs where either: (i) a direct promoter variant was observed (as per column six of **Table 2**) or (ii) a Hi-C contact to a gene promoter was observed (as per column nine of **Table 2**), together with functionally active chromatin (as per column seven of **Table 2**). The SNP used for testing at each locus was selected based on the closest available proxy (highest R^2) to the functional variant (*i.e.* the promoter or Hi-C contact variant), rather than using the sentinel SNP with the strongest TGCT association. Finally, as a comparison all possible gene/variant eQTL combinations were also tested at each locus (ignoring the functional Hi-C/promoter/CHiP-seq data), to provide a reference overview of all possible eQTL associations at each locus (**Supplementary Table 9**).

Transcription factor binding motif analysis

The impact of variants on regulatory motifs was assessed for a set of transcription factors (TF) associated with germ cell development. A germ cell specific TF set was utilized, rather than all TF

globally, to provide increased specificity. An OMIM[68] search-term-driven method was used to define the germ cell development TF set, using the following search terms: “germ cell” AND “development” AND “transcription factor” (n=46). The TF list was then intersected with predicted TF binding motifs based on a library of position weight matrices computed by Kheradpour and Kellis (2014)[69 70]. The intersected dataset contained motif position data for 10 TFs: DMRT1, GATA, KLF4, LHX8, NANOG, POU5F1, PRDM1, SOX2, SOX9, and CTCF. To validate the specificity of these motifs for TGCT we conducted variant set enrichment analysis, using the same method as detailed above (based on Cowper-Sal lari *et al*[60]), which confirmed enrichment for disruption of these 10 motifs in the 44 TGCT risk loci compared to the null distribution (**Supplementary Table 10**).

Integration of functional data

For the integrated functional annotation of risk loci LD blocks were defined as all SNPs in $R^2 > 0.8$ with the sentinel SNP. Risk loci were then annotated with six types of functional data: (i) presence of a Hi-C contact linking to a gene promoter, (ii) presence of an expression quantitative trait locus, (iii) presence of a ChIP-seq peak, (iv) presence of a disrupted transcription factor binding motif, (v) presence of a variant within a gene promoter boundary, with boundaries defined using the Ensembl regulatory build[71], (vi) presence of a non-synonymous coding change. Candidate causal genes were then assigned to TGCT risk loci using the target genes implicated in annotation tracks (i), (ii), (v) and (vi). Where the data supported multiple gene candidates, the gene with the highest number of individual functional data points was assigned to be the candidate. Where multiple genes have the same number of data points all genes are listed. Competing mechanisms for the same gene (e.g. both coding and promoter variants) were allowed.

ACKNOWLEDGEMENTS

We thank the subjects with TGCT and the clinicians involved in their care for participation in this study. We thank the patients and all clinicians forming part of the UK Testicular Cancer Collaboration (UKTCC) for their participation in this study. A full list of UKTCC members is included in **Supplementary note 1**. We acknowledge National Health Service funding to the National Institute for Health Research Biomedical Research Centre. We thank the UK Genetics of Prostate Cancer Study (UKGPCS) study teams for the recruitment of the UKGPCS controls. Genotyping of the OncoArray was funded by the US National Institutes of Health (NIH) [U19 CA 148537 for ELucidating Loci Involved in Prostate cancer SuscEptibility (ELLIPSE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract number HHSN268201200008I]. Additional analytic support was provided by NIH NCI U01 CA188392 (PI: Schumacher). The PRACTICAL consortium was supported by Cancer Research UK Grants C5047/A7357, C1287/A10118, C1287/A16563, C5047/A3354, C5047/A10692, C16913/A6135, European Commission's Seventh Framework Programme grant agreement n° 223175 (HEALTH-F2-2009-223175), and The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148537-01 (the GAME-ON initiative). A full list of PRACTICAL consortium members is included in **Supplementary note 2.** We would also like to thank the following for funding support: The Institute of Cancer Research and The Everyman Campaign, The Prostate Cancer Research Foundation, Prostate Research Campaign UK (now Prostate Action), The Orchid Cancer Appeal, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. We are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust. This study would not have been possible without the contributions of the following: M. K. Bolla (BCAC), Q. Wang (BCAC), K. Michailido (BCAC), J. Dennis (BCAC), P. Hall (COGS); D.F. Easton (BCAC), A. Berchuck (OCAC), R. Eeles (PRACTICAL), G. Chenevix-Trench (CIMBA), J. Dennis, P. Pharoah, A. Dunning, K. Muir, J. Peto, A. Lee, and E. Dicks. We also thank the following for their contributions to this project: Jacques Simard, Peter Kraft, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory; and Kimberly F. Doheny and the staff of the Center for Inherited Disease Research (CIDR) genotyping facility. The results published here are in

part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This study makes use of data generated by the Wellcome Trust Case Control Consortium 2 (WTCCC2). A full list of the investigators who contributed to the generation of the data is available from the WTCCC website. We acknowledge the contribution of Elizabeth Rapley and Mike Stratton to the generation of previously published UK GWAS case data. We acknowledge funding from the Swedish Cancer Society (CAN2011/484 and CAN2012/823), the Norwegian Cancer Society (grants number 418975 – 71081 – PR-2006-0387 and PK01-2007-0375) and the Nordic Cancer Union (grant number S-12/07). This study was supported by the Movember foundation and the Institute of Cancer Research. K. Litchfield is supported by a PhD fellowship from Cancer Research UK. R.S.H. and P.B. are supported by Cancer Research UK (C1298/A8362 Bobby Moore Fund for Cancer Research UK). We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out.

AUTHOR CONTRIBUTIONS

C.T., K.L., and R.S.H designed the study. Case samples were recruited by A.R., R.H. and through UKTCC. R.E., A.D, K.M, J.P., Z.K-J, N.P. and D.E supplied Oncoarray control data. N.O. administrated genotyping of Oncoarray case samples. D.D. coordinated all case sample administration and tracking. K.L., M.L., A.H. and P.B. prepared samples for genotyping experiments. K.L., M.L., G.O., C.L., K.F. and I.A. conducted all Promotor HiC and 3C laboratory experiments. Bioinformatics and statistical analyses were designed by C.T., R.S.H and K.L.. K.L., G.M., C.L. and M.L. conducted all Promotor HiC and 3C data analysis. K.L. and P.L. conducted transcription factor enrichment analysis. K. L., C.L. and M.L. performed all other bioinformatics and statistical analyses. R.K., T. H., W. K., T.G. and F.W. provided Scandinavian GWAS data. K. L. drafted the manuscript with assistance from C.T., R.S.H., M.L., J.S., J.N. and T.B. All authors reviewed and contributed to the manuscript.

DATA AVAILABILITY

Case Oncoarray GWAS data and the Hi-C dataset utilized in this paper have both been deposited in the European Genome–phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under the accession codes EGAS00001001836 and EGAS00001001930 respectively.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURES AND TABLE LEGENDS

Figure 1 - **Study design.**

Figure 2 - **Circos plot of integrated functional analysis for all 44 TGCT risk loci.** Inner-most ring represents the presence of a Hi-C contact in the NTERA2 cell line, the next four rings are narrow-peak histone ChIP-seq tracks for NTERA2, the sixth ring represents -log *P* values of TGCT risk association from the Oncoarray GWAS data with green line denoting genome-wide significance and the seventh ring (outer-most) is the functional annotation and classification of candidate causal genes.

Figure 3A-C – Regional plots of three new TGCT loci at A) 8p23.1, B) 15q25.2 and C) 15q22.31. Shown by triangles are the −log10 association P values of genotyped SNPs, based on Oncoarray data. Shown by circles are imputed SNPs at each locus. The intensity of red shading indicates the strength of LD with the sentinel SNP (labelled). Also shown are the SNP build 37 coordinates in mega-bases, recombination rates in centi-morgans (in light blue) and the genes in the region. Below the gene transcripts are Hi-C next generation sequencing read pair counts (gaps represent bait locations) and significant Hi-C interactions. Below the axis is a zoomed-in section displaying the surrounding genes for each SNP, the predicted chromHMM states along with an arc depiction of the same Hi-C contact(s).

Table 1 – Summary of genotyping results for all genome-wide TGCT risk SNPs (n=44).

Table 2 – Summary of functional annotation.

REFERENCES

1. Manku, G. *et al.* Changes in the expression profiles of claudins during gonocyte differentiation and in seminomas. *Andrology* **4**, 95-110 (2016).
2. Le Cornet, C. *et al.* Testicular cancer incidence to rise by 25% by 2025 in Europe? Model-based predictions in 40 countries using population-based registry data. *Eur J Cancer* **50**, 831-9 (2014).
3. Litchfield, K. *et al.* Quantifying the heritability of testicular germ cell tumour using both population-based and genomic approaches. *Sci Rep* **5**, 13889 (2015).
4. Swerdlow, A.J., De Stavola, B.L., Swanwick, M.A. & Maconochie, N.E. Risks of breast and testicular cancers in young adult twins in England and Wales: evidence on prenatal and genetic aetiology. *Lancet* **350**, 1723-8 (1997).
5. McGlynn, K.A., Devesa, S.S., Graubard, B.I. & Castle, P.E. Increasing incidence of testicular germ cell tumors among black men in the United States. *J Clin Oncol* **23**, 5757-61 (2005).
6. Hemminki, K. & Li, X. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. *British Journal of Cancer* **90**, 1765-1770 (2004).
7. Kharazmi, E. *et al.* Cancer Risk in Relatives of Testicular Cancer Patients by Histology Type and Age at Diagnosis: A Joint Study from Five Nordic Countries. *Eur Urol* **68**, 283-9 (2015).
8. Rapley, E.A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nat Genet* **41**, 807-10 (2009).
9. Turnbull, C. & Rahman, N. Genome-wide association studies provide new insights into the genetic basis of testicular germ-cell tumour. *Int J Androl* **34**, e86-96; discussion e96-7 (2011).
10. Kanetsky, P.A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat Genet* **41**, 811-5 (2009).
11. Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat Genet* **42**, 604-7 (2010).
12. Kanetsky, P.A. *et al.* A second independent locus within DMRT1 is associated with testicular germ cell tumor susceptibility. *Hum Mol Genet* **20**, 3109-17 (2011).
13. Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. *Nat Genet* **45**, 686-9 (2013).
14. Bojesen, S.E. *et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* **45**, 371-84, 384e1-2 (2013).
15. Chung, C.C. *et al.* Meta-analysis identifies four new loci associated with testicular germ cell tumor. *Nat Genet* **45**, 680-5 (2013).
16. Kristiansen, W. *et al.* Two new loci and gene sets related to sex determination and cancer progression are associated with susceptibility to testicular germ cell tumor. *Hum Mol Genet* (2015).
17. Litchfield, K. *et al.* Multi-stage genome-wide association study identifies new susceptibility locus for testicular germ cell tumour on chromosome 3q25. *Hum Mol Genet* **24**, 1169-76 (2015).
18. Litchfield, K. *et al.* Identification of four new susceptibility loci for testicular germ cell tumour. *Nat Commun* **6**, 8690 (2015).
19. Litchfield, K., Shipley, J. & Turnbull, C. Common variants identified in genome-wide association studies of testicular germ cell tumour: an update, biological insights and clinical application. *Andrology* **3**, 34-46 (2015).
20. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies (vol 38, pg 209, 2006). *Nature Genetics* **38**, 390-390 (2006).

21. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
22. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
23. Agnihotri, S. *et al.* A GATA4-regulated tumor suppressor network represses formation of malignant human astrocytomas. *J Exp Med* **208**, 689-702 (2011).
24. Hellebrekers, D.M. *et al.* GATA4 and GATA5 are potential tumor suppressors and biomarkers in colorectal cancer. *Clin Cancer Res* **15**, 3990-7 (2009).
25. Tsang, A.P. *et al.* FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* **90**, 109-19 (1997).
26. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
27. Ketola, I. *et al.* Developmental expression and spermatogenic stage specificity of transcription factors GATA-1 and GATA-4 and their cofactors FOG-1 and FOG-2 in the mouse testis. *Eur J Endocrinol* **147**, 397-406 (2002).
28. Zheng, R. & Blobel, G.A. GATA Transcription Factors and Cancer. *Genes Cancer* **1**, 1178-88 (2010).
29. Kurimoto, K., Yamaji, M., Seki, Y. & Saitou, M. Specification of the germ cell lineage in mice: a process orchestrated by the PR-domain proteins, Blimp1 and Prdm14. *Cell Cycle* **7**, 3514-8 (2008).
30. Ohinata, Y. *et al.* A signaling principle for the specification of the germ cell lineage in mice. *Cell* **137**, 571-84 (2009).
31. Yamaji, M. *et al.* Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat Genet* **40**, 1016-22 (2008).
32. Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J. & Sinclair, A.H. Conservation of a sex-determining gene. *Nature* **402**, 601-2 (1999).
33. Rao, S. *et al.* Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. *Mol Cell Biol* **30**, 5364-80 (2010).
34. Greenbaum, M.P. *et al.* TEX14 is essential for intercellular bridges and fertility in male mice. *Proc Natl Acad Sci U S A* **103**, 4982-7 (2006).
35. Mondal, G., Ohashi, A., Yang, L., Rowley, M. & Couch, F.J. Tex14, a Plk1-regulated protein, is required for kinetochore-microtubule attachment and regulation of the spindle assembly checkpoint. *Mol Cell* **45**, 680-95 (2012).
36. Jinks, R.N. *et al.* Recessive nephrocerebellar syndrome on the Galloway-Mowat syndrome spectrum is caused by homozygous protein-truncating mutations of WDR73. *Brain* **138**, 2173-90 (2015).
37. Colin, E. *et al.* Loss-of-function mutations in WDR73 are responsible for microcephaly and steroid-resistant nephrotic syndrome: Galloway-Mowat syndrome. *Am J Hum Genet* **95**, 637-48 (2014).
38. Petrovic, A. *et al.* The MIS12 complex is a protein interaction hub for outer kinetochore assembly. *J Cell Biol* **190**, 835-52 (2010).
39. Rao, C.V., Yamada, H.Y., Yao, Y. & Dai, W. Enhanced genomic instabilities caused by deregulated microtubule dynamics and chromosome segregation: a perspective from genetic studies in mice. *Carcinogenesis* **30**, 1469-74 (2009).
40. Barisic, M. *et al.* Mitosis. Microtubule detyrosination guides chromosomes during mitosis. *Science* **348**, 799-803 (2015).
41. Ma, W. & Viveiros, M.M. Depletion of pericentrin in mouse oocytes disrupts microtubule organizing center function and meiotic spindle organization. *Mol Reprod Dev* **81**, 1019-29 (2014).

42. Litchfield, K. *et al.* Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours. *Nat Commun* **6**, 5973 (2015).
43. Zeron-Medina, J. *et al.* A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. *Cell* **155**, 410-22 (2013).
44. De Miguel, M.P., Cheng, L., Holland, E.C., Federspiel, M.J. & Donovan, P.J. Dissection of the c-Kit signaling pathway in mouse primordial germ cells by retroviral-mediated gene transfer. *Proc Natl Acad Sci U S A* **99**, 10458-63 (2002).
45. Yu, M. *et al.* The scaffolding adapter Gab2, via Shp-2, regulates kit-evoked mast cell proliferation by activating the Rac/JNK pathway. *J Biol Chem* **281**, 28615-26 (2006).
46. Penegar, S. *et al.* National study of colorectal cancer genetics. *Br J Cancer* **97**, 1305-9 (2007).
47. Eisen, T., Matakidou, A., Houlston, R. & Consortium, G. Identification of low penetrance alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). *BMC Cancer* **8**, 244 (2008).
48. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
49. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
50. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
51. Cuppen, E. Genotyping by Allele-Specific Amplification (KASPar). *CSH Protoc* **2007**, pdb prot4841 (2007).
52. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
53. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243-6 (2005).
54. Litchfield, K. *et al.* Multi-stage genome wide association study identifies new susceptibility locus for testicular germ cell tumour on chromosome 3q25. *Hum Mol Genet* (2014).
55. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-40 (2010).
56. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
57. Scales, M., Jager, R., Migliorini, G., Houlston, R.S. & Henrion, M.Y. visPIG--a web tool for producing multi-region, multi-track, multi-scale plots of genetic data. *PLoS One* **9**, e107497 (2014).
58. Pharoah, P.D.P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* **31**, 33-36 (2002).
59. CRUK. (2014).
60. Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* **44**, 1191-8 (2012).
61. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
62. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
63. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310 (2015).
64. Jonathan Cairns, P.F.-P., Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser,

Mikhail Spivakov. CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C data. *BioRxiv* (2016).
65. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115 (2012).
66. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-5 (2012).
67. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
68. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-7 (2005).
69. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
70. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).
71. Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. & Flicek, P.R. The ensembl regulatory build. *Genome Biol* **16**, 56 (2015).