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

Kevin Litchfield[1], Max Levy[1], Giulia Orlando[1], Chey Loveday[1], Philip J. 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[15], Janet Shipley[16], Tom Grotmol[17], Fredrik Wiklund[2], Richard S Houlston[1], Clare Turnbull[1,18, 19]

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, SW3 6JB, 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.

26   11. Royal Marsden NHS Foundation Trust, London, SM2 5NG, UK

27   12. The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, 237

28   Fulham Road, London SW3 6JB, UK

29   13. Department of Applied Health Research, University College London, London, WC1E 6BT, UK

30   14. Academic Uro-oncology Unit, The Royal Marsden NHS Foundation Trust, Sutton, Surrey, SM2

31   5PT, UK

32   15. Academic Radiotherapy Unit, Institute of Cancer Research, Sutton, Surrey, SM2 5NG, UK

33   16. Division of Molecular Pathology, The Institute of Cancer Research, London, SM2 5NG, UK

34   17. Cancer Registry of Norway, Oslo, Norway.

35   18. William Harvey Research Institute, Queen Mary University, London, EC1M 6BQ , UK

36   19. Guys and St Thomas Foundation NHS Trust, Great Maze Pond, London, SE1 9RT, UK

37   * See supplementary note

38   Correspondence to: Clare Turnbull, Division of Genetics and Epidemiology, The Institute of Cancer

39   Research, London, SM2 5NG, UK; Tel: ++44 (0) 208 722 4485; E-mail: clare.turnbull@icr.ac.uk

40

42

43   **Genome-wide association studies (GWAS) have transformed our understanding of testicular germ**

44   **cell tumour (TGCT) susceptibility but much of the heritability remains unexplained. Here we report**

45   **a new GWAS, a meta-analysis with previous GWAS and a replication series, totalling 7,319 TGCT**

46   **cases and 23,082 controls. We identify 19 new TGCT risk loci, approximately doubling the number**

47   **of known TGCT risk loci to 44. By performing *in-situ* Hi-C in TGCT cells, we provide evidence for a**

48   **network of physical interactions between all 44 TGCT risk SNPs and candidate causal genes. Our**

49   **findings reveal widespread disruption of developmental transcriptional regulators as a basis of**

50   **TGCT susceptibility, consistent with failed primordial germ cell differentiation as an initiating step**

51   **in oncogenesis[1]. Defective microtubule assembly and dysregulation of KIT-MAPK signalling also**

52   **feature as recurrently disrupted pathways. Our findings support a polygenic model of risk and**

53   **provide insight into the biological basis of TGCT.**

54

55

56

57

58

59

60

61   Testicular germ cell tumour (TGCT) is the most common cancer in men aged 18-45, with over 52,000

62   new cases diagnosed annually worldwide[2]. The development of TGCT is strongly influenced by

63   inherited genetic factors, which contributes to nearly half of all disease risk[3] and is reflected in the 4-

64   to-8 fold increased risk shown in siblings of cases[4-7]. Our understanding of TGCT susceptibility has

65   been transformed by recent genome-wide association studies (GWAS), which have so far identified

66   25 independent risk loci for TGCT[8-18]. Although projections indicate that additional risk variants for

67   TGCT can be discovered by GWAS[19], studies to date have been based on comparatively small sample

68   sizes which have had limited power to detect common risk variants[20].

69   To gain a more comprehensive insight into TGCT aetiology we performed a new GWAS with

70   substantially increased power, followed by a meta-analysis with existing GWAS and replication

71   genotyping (totalling 7,319 cases/23,082 controls). Here we report both the discovery of 19 new

72   TGCT susceptibility loci and refined risk estimates for the previously reported loci. In addition, we

73   have investigated the gene regulatory mechanisms underlying the genetic associations observed at

74   all 44 TGCT GWAS risk loci by performing *in-situ* chromosome conformation capture in TGCT cells

75   (Hi-C) to characterize chromatin interactions between predisposition SNPs and target genes,

76   integrating these data with a range of publicly available TGCT functional genomics data.

77

78   We conducted a new GWAS using the Oncoarray platform (3,206 UK TGCT cases/7,422 UK controls),

79   followed by a meta-analysis combining the two largest published TGCT GWAS datasets[11,16] (986 UK

80   cases/4,946 UK controls, 1,327 Scandinavian cases/6,687 Scandinavian controls) (**Fig. 1**). To increase

81   genomic resolution, we imputed >10 million SNPs using the 1000 Genomes Project as a reference

82   panel. Quantile-Quantile (Q-Q) plots for SNPs with minor allele frequency (MAF) >5% post

83   imputation did not show evidence of substantive over-dispersion (λ_{1000}=1.03, **Supplementary Fig. 1**)**.**

84   We derived joint odds ratios (ORs) and 95% confidence intervals (CIs) under a fixed-effects model for

85   each SNP with MAF >0.01. Finally we sought validation of 37 SNPs associated at $P < 5.0 \times 10^{-6}$, which

86   did not map to known TGCT risk loci and displayed a consistent OR across all GWAS datasets, by

87   genotyping an additional 1,801 TGCT cases and 4,027 controls from the UK. After meta-analysis of

88   the three GWAS and replication series, we identified genome-wide significant associations (*i.e*. $P < 5$

89    $\times 10^{-8}$) at 19 new loci (**Table 1**). We found no evidence for significant interactions between risk loci.

90   To the extent that they have been deciphered, many GWAS risk loci map to non-coding regions of

91   the genome and influence gene regulation. Across the 44 independent TGCT risk loci (19 new and 25

92   previously reported), we confirmed a significant enrichment of enhancer/promoter associated

93   histone marks, including H3K4me1, H3K4me3 and H3K9ac, using available ChIP-Seq data from the

94   TGCT cell line NTERA2 ($P<5.0\times10^{-3}$) (**Supplementary Table 1**). Moreover this enrichment showed

95   tissue specificity when compared to 41 other cell lines from the ENCODE[21] project (**Supplementary**

96   **Fig. 2**). These observations support the assertion that the TGCT predisposition loci influence risk

97   through effects on *cis*-regulatory networks, and are involved in transcriptional initiation and

98   enhancement. Since genomic spatial proximity and chromatin looping interactions are fundamental

99   for regulation of gene expression we performed *in situ* capture Hi-C of promoters in NTERA2 cells to

100   link risk loci to candidate target genes. We also sought to gain insight into the possible biological

101   mechanisms for the associations by performing tissue-specific expression quantitative trait loci

102   (eQTL) analysis for all risk SNP and target gene pairs (**Supplementary Fig. 3, Supplementary Table 2**).

103   We analysed RNA-seq data from both normal testis (GTEx project[22]) and TGCT (TCGA),

104   acknowledging that the latter may be affected by the issue of tumour purity, in addition to

105   dysregulated gene expression that typifies cancer. Accepting this limitation and that further

106   validation may be required, eQTL analysis was conducted in both datasets based on the established

107   network of enhancer/ promoter variants, to maximise our ability to find statistically significant

108   associations after correcting for multiple testing. We additionally annotated risk loci with variants

109   predicted to disrupt binding motifs of germ cell specific transcription factors (TF) (see methods).

110   Finally, direct promoter variants and non-synonymous coding mutations for genes within the 44 risk

111   loci were denoted (**Table 2, Fig. 2**).

112

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

137   effect (GTEx data, P=4.9 x 10^{-7}) on *TEX14* (Testis-Expressed 14), which encodes an important

138   regulator of kinetochore-microtubule assembly in testicular germ cells[14,34,35]. At new risk locus

139   15q25.2 we identified a nominal eQTL association (rs2304416, TCGA data, P=3.2 x 10^{-2}) and

140   accompanying chromatin looping interaction with mitotic spindle assembly related gene *WDR73*[36]

141   (**Fig. 3b**). *WDR73* encodes a protein with a crucial role in the regulation of microtubule organization

142   during interphase[37] and biallelic mutations cause Galloway-Mowat Syndrome, a human syndrome of

143   nephrosis and neuronal dysmigration. Finally the functional analysis also highlighted microtubule

144   assembly related genes *PMF1*, *CENPE* and *PCNT* [38-41] as candidates at 1q22, 4q24 and 21q22.3

145   respectively.

146   Thirdly, the central role of KIT-MAPK signalling in TGCT oncogenesis was further supported at four

147   loci, by Hi-C looping interactions (at 11q14.1, 15q22.31), eQTL effects (at 6p21.31) and promoter

148   variants (at 6p21.31, 11q14.1, 15q22.31). Recent tumour sequencing studies have established that

149   *KIT* is the major somatic driver gene for TGCT[42] and a relationship between the previously identified

150   risk SNP rs995030 (12q21) and *KITLG* expression has been demonstrated through allele-specific p53

151   binding by Zeron-Medina et al[43]. Here we report a new locus at 15q22.31, containing a variant within

152   the promoter of *MAP2K1* (**Fig. 3c**), which raises the prospect of further elucidating mechanisms of

153   KIT-MAPK signalling in driving TGCTs. *MAP2K1* (alias *MEK1*) is downstream of c-Kit and MEK1

154   inhibition slows primordial germ cell growth in the presence of KIT ligand[44]. If *MAP2K1* is confirmed

155   as a causal gene at 15q22.31, the study of somatic *KIT* mutational status in patients carrying the risk

156   allele at 15q22.31 should be highly informative. In addition, within the 11q14.1 risk locus, we

157   identify a candidate promoter variant for *GAB2,* which encodes a docking protein for signal

158   transduction to MAPK and PI3K pathways which interacts directly with KIT[45]. Finally in our analysis

159   we identify both a candidate promoter variant and a nominal eQTL effect for *BAK1* (6p21.31)(TCGA

160   data, P=1.9 x 10^{-2}), which encodes a protein regulating apoptosis which binds with KIT [40]. While we

161   have sought to decipher the functional basis of risk loci based on the cumulative weight of evidence

162   across eQTL, Hi-C and ChIP-seq data, a limitation has been reliance on relatively small sample size for

163   eQTL analysis. Access to larger eQTL datasets in testicular tissue are likely in the future to address

164   this deficiency enabling a better definition of the causal basis of TGCT risk at each locus.

165

166   The 44 risk loci which have now been identified for TGCT collectively account for 34% of the (father-

167   to-son) familial risk and hence have potential clinical utility for personalized risk profiling. To assess

168   this potential, we constructed polygenic risk scores (PRS) for TGCT, considering the combined effect

169   of all risk SNPs modelled under a log-normal relative risk distribution. Using this approach the men in

170   the top 1% of genetic risk have a relative risk of 14 which translates to a 7% lifetime risk of TGCT

171   (**Supplementary Fig. 4**).

172

173   In summary, we have performed a new TGCT GWAS, identifying 19 new risk loci for TGCT,

174   approximately doubling the number of previously reported SNPs. Using capture Hi-C we have

175   generated a chromatin interaction map for TGCT, providing direct physical interactions between

176   non-coding risk SNPs and target gene promoters. Moreover integration of these data together with

177   ChIP-seq chromatin profiling and RNA-seq eQTL analysis, accepting certain caveats, has allowed us to

178   gain preliminary but unbiased tissue-specific insight into the biological basis of TGCT susceptibility.

179   This analysis suggests a model of TGCT susceptibility based on transcriptional dysregulation, which is

180   likely to contribute to the developmental arrest of primordial germ cells coupled with chromosomal

181   instability through defective microtubule function and accompanied upregulation of KIT-MAPK

182   signalling.

183

184

185 **METHODS**

186

187 **Sample description**

188 TGCT cases were from the UK (n=5,992) and Scandinavia (n=1,327). The UK cases were ascertained

189 from two studies (1) a UK study of familial testicular cancer and (2) a systematic collection of UK

190 collection of TGCT cases. Case recruitment was via the UK Testicular Cancer Collaboration, a group of

191 oncologists and surgeons treating TGCT in the UK (**Supplementary note**). The studies were co-

192 ordinated at the Institute of Cancer Research (ICR). Samples and information were obtained with full

193 informed consent and Medical Research and Ethics Committee approval (MREC02/06/66 and

194 06/MRE06/41). Additional (n=1,327) case samples of Scandinavian origin were used from a

195 previously published GWAS[16].

196 Control samples for the primary GWAS were all taken from within the UK. Specifically 2,976 cancer-

197 free, male controls were recruited through two studies within the PRACTICAL Consortium

198 (**Supplementary note**): (1) the UK Genetic Prostate Cancer Study (UKGPCS) (age <65), a study

199 conducted through the Royal Marsden NHS Foundation Trust and (2) SEARCH (Study of Epidemiology

200 & Risk Factors in Cancer), recruited via GP practices in East Anglia (2003-2009). 4,446 cancer-free

201 female controls from across the UK were recruited via the Breast Cancer Association Consortium

202 (BCAC). Controls from the UK previously published GWAS[11] were from two sources within the UK:

203 2,482 controls were from the 1958 Birth Cohort (1958BC), and 2,587 controls were identified

204 through the UK National Blood Service (NBS) and were genotyped as part of the Wellcome Trust

205 Case Control Consortium. Additional (n=6,687) control samples of Scandinavian origin were used in

206 the meta-analysis, and have been previously described[16]. Control samples for replication genotyping

207 (n=4,027) were taken from two studies, the national study of colorectal cancer genetics (NSCCG)[46]

208 and GEnetic Lung CAncer Predisposition Study (GELCAPS)[47]. NSCCG and GELCAP controls were

209 spouses of cancer patients with no personal history of cancer at time of ascertainment.

210

Primary GWAS

212 Genotyping was conducted using a custom Infinium OncoArray-500K BeadChip (Oncoarray) from

213 Illumina (Illumina, San Diego, CA, USA), comprising a 250K SNP genome-wide backbone and 250K

214 SNP custom content selected across multiple consortia within COGS (Collaborative Oncological

215 Gene-environment Study). Oncoarray genotyping was conducted in accordance with the

216 manufacturer's recommendations by the Edinburgh Clinical Research Facility, Wellcome Trust CRF,

217 Western General Hospital, Edinburgh EH4 2XU.

218

Published GWAS

220 The UK and Scandinavian GWAS have been previously reported[8,11,13]. Briefly the UK GWAS comprised

221 986 cases genotyped on the Illumina HumanCNV370-Duo bead array (Ilumina, San Diego, CA, USA)

222 and 4,946 controls genotyped on the Illumina Infinium 1.2M array. We analysed data on a common

223 set of 314,861 SNPs successfully genotyped by both arrays. The Scandinavian GWAS [16], comprised

224 1,326 cases and 6,687 controls genotyped using the Human OmniExpressExome-8v1 Illumina array.

225

Quality Control of GWAS

227 Oncoarray data was filtered as follows, we excluded individuals with low call rate (<95%), with

228 abnormal autosomal heterozygosity or with >10% non-European ancestry (based on multi-

229 dimensional scaling). We filtered out all SNPs with minor allele frequency <1%, a call rate of <95% in

230 cases or controls or with a minor allele frequency of 1–5% and a call rate of <99%, and SNPs

231 deviating from Hardy-Weinberg equilibrium (10^{-12} in controls and 10^{-5} in cases). The final number of

232 SNPs passing quality control filters was 371,504. Quality control (QC) procedures for the UK and

233 Scandinavian GWAS have been previously described[8,11,13,16].

234

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.

241

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.

246

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 an additive genetic 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**).

257 Statistical analysis of previously reported GWAS was performed as previously described[8,11,13,16,54].

258 Meta-analyses were performed using the fixed-effects inverse-variance method based on the β

259 estimates and standard errors from each study using META v1.6[55]. Cochran's Q-statistic to test for

260 heterogeneity and the I^2 statistic to quantify the proportion of the total variation due to

261 heterogeneity were calculated[56]. For each new locus we examined evidence of departure from a log-

262 additive (multiplicative) model, to assess any genotype specific effect. Using the Oncoarray data

263 individual genotype data ORs were calculated for heterozygote (OR_{het}) and homozygote (OR_{hom})

264 genotypes, which were compared to the per allele ORs. We tested for a difference in these 1d.f. and

265 2d.f. logistic regression models to assess for evidence of deviation ($P<0.05$) from a log-additive

266 model. Using Oncoarray data we examined for statistical interaction between any of the 44 TGCT

267 predisposition loci by evaluating the effect of adding an interaction term to the regression model,

268 adjusted for stage, using a likelihood ratio test (using a significance threshold of $P < 2.58 \times 10^{-5}$ to

269 account for 1,936 tests). Regional plots were generated using visPIG software[57] (**Supplementary Fig.**

270 **6**). Polygenic risk scores (PRS) were constructed using the methodology of Pharoah et al[58], based on

271 a log-normal distribution LN (μ, σ^2) with mean μ and variance σ^2 (*i.e.* relative risk is normally

272 distributed on a logarithmic scale). The 0.5% lifetime risk of TGCT risk was based on 2014 UK data[59],

273 multiplied by relative risk to give lifetime risk per percentile of the PRS. For calculation of the

274 proportion of TGCT genetic risk explained by the 44 loci, a father-to-son relative risk of four was

275 used.

276

277 **Chromatin mark enrichment analysis**

278 To examine enrichment in specific ChIP-seq tracks across risk loci we adapted the variant set

279 enrichment method of Cowper-Sal lari *et al*[60]. Briefly, for each risk locus, a region of strong LD was

280 defined (*i.e.* R^2 > 0.8 and D' > 0.8), and SNPs mapping to these regions were termed the associated

281 variant set (AVS). Histone ChIP-seq uniform peak data was obtained from ENCODE[21] for the NTERA2

282   cell line, and data was included for four histone marks. For each of these marks, the overlap of the

283   SNPs in the AVS and the binding sites was determined to produce a mapping tally. A null distribution

284   was produced by randomly selecting SNPs with the same LD structure as the risk associated SNPs,

285   and the null mapping tally calculated. This process was repeated 10,000 times, and approximate *P*-

286   values were calculated as the proportion of permutations where null mapping tally was greater or

287   equal to the AVS mapping tally. An enrichment score was calculated by normalizing the tallies to the

288   median of the null distribution. Thus the enrichment score is the number of standard deviations of

289   the AVS mapping tally from the mean of the null distribution tallies. Tissue specificity was assessed

290   by comparison of enrichment levels in NTERA2, compared to 41 other cell lines from ENCODE[21], with

291   analysis performed using the same method as above (**Supplementary Fig. 2**).

292

293   **Promoter Hi-C**

294   *In situ* Hi-C libraries were prepared as described by Rao et al.[61] with the following modifications: (i)

295   25 million cells were fixed and processed; (ii) HindIII enzyme (NEB, Ipswich, MA, USA) was used and

296   digestion was performed overnight; (iii) ligation was performed overnight at 16C; (iv) 3 µl of 15 µM

297   annealed PE adaptors were ligated incubating 3 µl of T4 DNA ligase (NEB, Ipswich, MA, USA) for 2h at

298   RT; (vi) 6 cycles of PCR were performed to amplify the libraries before capture. A Sure Select

299   (Agilent, Santa Clara, CA, USA) custom promoter kit was used to perform capture with the same

300   design as described by Misfud *et al.*[62]. For each capture reaction, 750 µg of Hi-C libraries were used.

301   Capture was performed following the manufacture protocol and employing a custom reagent kit

302   (Agilent, Santa Clara, CA, USA). Final PCR amplification was performed using 5 cycles to minimise PCR

303   duplicates. 2x100bp sequencing was performed using Illumina HiSeq2000 or 2500 technology

304   (Illumina, San Diego, CA, USA). The HiCUP pipeline[63] was used to process raw sequencing reads, map

305   di-tag positions against the reference human genome and remove duplicate reads. The protocol was

306   performed for two independent NTERA2 cell culture replicates, with cells obtained from the

307   laboratory of Prof. Janet Shipley (The Institute of Cancer Research, London) and their identity

308   independently confirmed through STR typing at an external laboratory (Public Health England,

309   Porton Down, UK). Cells were tested and found to be negative for mycoplasma contamination. Both

310   Hi-C libraries achieving the following quality control thresholds: >80% reads uniquely aligning, >80%

311   valid pair rate, >85% unique di-tag rate and >80% of interactions being *cis* (**Supplementary Table 4**).

312   Statistically significant interactions were called using the CHiCAGO pipeline[64], with both cell culture

313   replicates processed in parallel to obtain a unique list of reproducible NTERA2 contacts. Stability of

314   results across replicates was also verified by processing each sample individually and comparing the

315   significance scores of called interactions; strong correlation was observed between the replicates (r

316   = 0.8, $P < 5.0 \times 10^{-10}$, **Supplementary Fig. 7**). Interactions with a -log(weighted P-value) > 5 were

317   considered significant. To avoid short-range proximity bias interactions of <40kb were excluded. The

318   distribution of interaction distances closely matched the prior published dataset of Misfud *et al.*[62]

319   (**Supplementary Fig. 8**). A Hi-C track plotting read pair counts per HindIII fragment has been added

320   to region plot figures to demonstrate the underlying signal strength of significant Hi-C contacts.

321

322   **3C Validation**

323   3C was used to validate selected chromatin interactions detected by CHi-C (3p24.3, 4q24, 11q14.1,

324   15q22.31, 15q25.2, 16q12.1, and 16q23.1) (**Supplementary Fig. 9, Supplementary Table 5**). Three

325   cell culture replicates of *in situ* 3C libraries were prepared using NTERA2 cells. Cell pellets were

326   crosslinked, digested with HindIII, and ligated. Libraries were purified by phenol-chloroform

327   extraction.

328   For each loci one or more bacterial artificial chromosomes (BACs; Source BioScience, Nottingham,

329   UK) were used as an internal standard (**Supplementary Table 6**). Clones were streaked and grown

330   before extracting DNA using a QIAGEN Plasmid Maxi Kit (QIAGEN, Hilden, Germany) which was

331 purified by phenol-chloroform extraction. In loci covered by more than one clone, equimolar

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

333 equimolar solution of BACs.

334 Unidirectional primer pairs were designed to amplify ligation junctions of the bait and other

335 interacting HindIII fragment (promoter-element, P-E) and around the bait and a flanking control

336 HindIII fragment in between the promoter and distal element (promoter-control, P-C) using

337 Primer3[65] (**Supplementary Tables 7 and 8**). Regions were amplified using both P-E and P-C primer

338 pairs in BAC and NTERA2 libraries using a QIAGEN Multiplex PCR Kit (QIAGEN, Hilden, Germany). 5 ng

339 and 100 ng of BAC and NTERA2 library template DNA, respectively, were amplified using the

340 following procedure: initial 15 minute denaturation at 95°C followed by 38 cycles of 94°C for 0.5

341 minutes, annealing temperature specific to primer pair for 1.5 minutes seconds, 72°C extension for

342 1.5 minutes, followed by a final 10 minute extension at 72°C extension. 5 μl of each PCR reaction

343 was visualised on 2% agarose gels stained with ethidium bromide. ImageJ[66] was used to quantify

344 intensities of PCR products and normalise for differential primer efficiency by comparing to

345 equimolar BAC PCR products.

346 P-E fragments were Sanger sequenced in NTERA2 libraries to confirm fragments visualised on

347 agarose gels as expected (**Supplementary Fig. 10**).

348

349 **Chromatin state annotation**

350 We used ChromHMM[67] to infer chromatin states by integrating information on histone modifications

351 and DNaseI hypersensitivity data to identify combinatorial and spatial patterns of epigenetic marks.

352 Aligned next generation sequencing reads from ChIP-Seq and DNAse-Seq experiments on the

353 NTERA2 cells were downloaded from ENCODE[21]. Read-shift parameters for ChIP-Seq data were

354 calculated using PHANTOMPEAKQUALTOOLS. Genome-wide signal tracks were binarised (including

355 input controls for ChIP-Seq data) and a set of learned models were generated using ChromHMM

356   software[67]. The parameters of the highest scoring model were retained and model states were

357   iteratively reduced down from 30 to 5 states. A 27-state model found to be stable and was

358   subsequently used for segmenting the genome at 200bp resolution (**Supplementary Fig. 11**).

359

360   **Expression quantitative trait locus analysis**

361   We investigated for evidence of association between the SNPs at each locus and tissue specific

362   changes in gene expression using two publically available resources: (i) RNAseq and Affymetrix 6.0

363   SNP data for 150 TGCT patients from The Cancer Genome Atlas and (ii) normal testicular tissue data

364   from GTEx from 157 samples[22]. Associations between normalized RNA counts per-gene and

365   genotype were quantified using R package 'Matrix eQTL'. Box plots of all eQTL associations are

366   presented in **Supplementary Fig. 3** and the tissue in which the association was observed (TGCT or

367   normal testis), along with any other tissues resulting in a positive association, are denoted in

368   **Supplementary Table 2**. To reduce multiple testing, association tests were only performed between

369   SNP and gene pairs where either: (i) a direct promoter variant was observed (as per column six of

370   **Table 2**) or (ii) a Hi-C contact to a gene promoter was observed (as per column nine of **Table 2**),

371   together with functionally active chromatin (as per column seven of **Table 2**). The SNP used for

372   testing at each locus was selected based on the closest available proxy (highest R^2) to the functional

373   variant (*i.e.* the promoter or Hi-C contact variant), rather than using the sentinel SNP with the

374   strongest TGCT association. Finally, as a comparison all possible gene/variant eQTL combinations

375   were also tested at each locus (ignoring the functional Hi-C/promoter/CHiP-seq data), to provide a

376   reference overview of all possible eQTL associations at each locus (**Supplementary Table 9**).

377

378   **Transcription factor binding motif analysis**

379   The impact of variants on regulatory motifs was assessed for a set of transcription factors (TF)

380   associated with germ cell development. A germ cell specific TF set was utilized, rather than all TF

381   globally, to provide increased specificity. An OMIM[68] search-term-driven method was used to define

382   the germ cell development TF set, using the following search terms: "germ cell" AND "development"

383   AND "transcription factor" (n=46). The TF list was then intersected with predicted TF binding motifs

384   based on a library of position weight matrices computed by Kheradpour and Kellis (2014)[69 70]. The

385   intersected dataset contained motif position data for 10 TFs: DMRT1, GATA, KLF4, LHX8, NANOG,

386   POU5F1, PRDM1, SOX2, SOX9, and CTCF. To validate the specificity of these motifs for TGCT we

387   conducted variant set enrichment analysis, using the same method as detailed above (based on

388   Cowper-Sal lari *et al*[60]), which confirmed enrichment for disruption of these 10 motifs in the 44 TGCT

389   risk loci compared to the null distribution (**Supplementary Table 10**).

390

391   **Integration of functional data**

392   For the integrated functional annotation of risk loci LD blocks were defined as all SNPs in $R^2 > 0.8$

393   with the sentinel SNP. Risk loci were then annotated with six types of functional data: (i) presence of

394   a Hi-C contact linking to a gene promoter, (ii) presence of an expression quantitative trait locus, (iii)

395   presence of a ChIP-seq peak, (iv) presence of a disrupted transcription factor binding motif, (v)

396   presence of a variant within a gene promoter boundary, with boundaries defined using the Ensembl

397   regulatory build[71], (vi) presence of a non-synonymous coding change. Candidate causal genes were

398   then assigned to TGCT risk loci using the target genes implicated in annotation tracks (i), (ii), (v) and

399   (vi). Where the data supported multiple gene candidates, the gene with the highest number of

400   individual functional data points was assigned to be the candidate. Where multiple genes have the

401   same number of data points all genes are listed. Direct non-synonymous coding variants were

402   allocated additional weighting. Competing mechanisms for the same gene (e.g. both coding and

403   promoter variants) were allowed.

404

445

446   **AUTHOR CONTRIBUTIONS**

447   C.T., K.L., and R.S.H designed the study. Case samples were recruited by A.R., R.H. and through

448   UKTCC. R.E., A.D, K.M, J.P., Z.K-J, N.P. and D.E supplied Oncoarray control data. N.O. administrated

449   genotyping of Oncoarray case samples. D.D. coordinated all case sample administration and

450   tracking. K.L., M.L., A.H. and P.B. prepared samples for genotyping experiments. K.L., M.L., G.O., C.L.,

451   K.F. and I.A. conducted all Promotor HiC and 3C laboratory experiments. Bioinformatics and

452   statistical analyses were designed by C.T., R.S.H and K.L.. K.L., G.M., C.L. and M.L. conducted all

453   Promotor HiC and 3C data analysis. K.L. and P.L. conducted transcription factor enrichment analysis.

454   K. L., C.L. and M.L. performed all other bioinformatics and statistical analyses. R.K., T. H., W. K., T.G.

455   and F.W. provided Scandinavian GWAS data. K. L. drafted the manuscript with assistance from C.T.,

456   R.S.H., M.L., J.S., J.N. and T.B. All authors reviewed and contributed to the manuscript.

457

458   **DATA AVAILABILITY**

459   Case Oncoarray GWAS data and the Hi-C dataset utilized in this paper have both been deposited in

460   the European Genome–phenome Archive (EGA), which is hosted by the European Bioinformatics

461   Institute (EBI), under the accession codes EGAS00001001836 and EGAS00001001930 respectively.

462

463   **COMPETING FINANCIAL INTERESTS**

464   The authors declare no competing financial interests.

465

466   **FIGURES AND TABLE LEGENDS**

467   **Figure 1** - **Study design.**

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

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

483

484 **Table 1 - Summary of genotyping results for all genome-wide TGCT risk SNPs (n=44). New loci (n=19) discovered through this study are marked in bold.**

					Oncoarray		Discovery Meta		Replication		Combined-Meta	
SNP[1]	Chr.	Pos. (b37)	Alleles[2]	RAF[3]	OR[4] (95% CI)	Ptrend[5]	OR (95% CI)	Ptrend	OR (95% CI)	Ptrend	P meta[6]	(I²)[7]
rs4240895	**1**	**9713386**	**C/T**	**0.39**	**1.13 (1.07-1.19)**	**7.8X10-05**	**1.14 (1.09-1.19)**	**1.7X10-07**	**1.24 (1.16-1.32)**	**1.7X10-07**	**6.2X10-13**	**47**
rs2072499	1	156169610	A/G	0.36	1.14 (1.08-1.20)	2.2X10-05	1.18 (1.13-1.23)	1.9X10-10	-	-	1.9X10-10	45
rs3790672	1	165873392	T/C	0.29	1.16 (1.10-1.23)	3.7X10-06	1.20 (1.14-1.25)	5.3X10-11	-	-	5.3X10-11	17
rs7581030	**2**	**71572455**	**C/T**	**0.24**	**1.15 (1.08-1.22)**	**5.9X10-05**	**1.17 (1.12-1.23)**	**8.3X10-09**	**1.17 (1.08-1.26)**	**6.2X10-04**	**2.1X10-11**	**0**
rs10510452	3	16625048	A/G	0.70	1.15 (1.08-1.21)	2.4X10-05	1.18 (1.13-1.23)	9.5X10-10	-	-	9.5X10-10	43
rs11705932	3	141818850	C/T	0.80	1.18 (1.11-1.26)	1.1X10-05	1.17 (1.11-1.23)	4.5X10-07	-	-	4.5X10-07	39
rs1510272	3	156300724	C/T	0.75	1.19 (1.12-1.26)	4.8X10-07	1.23 (1.17-1.28)	1.7X10-12	-	-	1.7X10-12	33
rs6821144	**4**	**76520651**	**G/A**	**0.89**	**1.18 (1.08-1.27)**	**9.7X10-04**	**1.22 (1.14-1.30)**	**1.9X10-06**	**1.35 (1.23-1.47)**	**9.8X10-07**	**2.3X10-11**	**18**
rs17021463	4	95224812	T/G	0.43	1.14 (1.09-1.20)	6.4X10-06	1.14 (1.10-1.19)	3.3X10-08	-	-	3.3X10-08	0
rs2720460	4	104054686	A/G	0.63	1.24 (1.18-1.31)	2.2X10-12	1.26 (1.21-1.31)	6.6X10-20	-	-	6.6X10-20	10
rs4862848	**4**	**188921440**	**A/G**	**0.35**	**1.18 (1.10-1.25)**	**1.2X10-05**	**1.21 (1.16-1.27)**	**3.5X10-12**	**1.10 (1.02-1.18)**	**2.4X10-02**	**1.9X10-12**	**61**
rs2736100	5	1286516	C/A	0.51	1.25 (1.19-1.30)	2.5X10-13	1.28 (1.24-1.33)	9.0X10-25	-	-	9.0X10-25	27
rs3805663	5	134342720	C/A	0.58	1.09 (1.03-1.15)	6.0X10-03	1.12 (1.07-1.17)	1.2X10-05	-	-	1.2X10-05	20
rs4624820	5	141681788	G/A	0.56	1.46 (1.40-1.52)	3.8X10-36	1.47 (1.43-1.52)	2.5X10-57	-	-	2.5X10-57	0
rs210138	6	33542538	A/G	0.21	1.42 (1.35-1.49)	1.0X10-21	1.48 (1.42-1.54)	3.5X10-37	-	-	3.5X10-37	70
rs11155671	**6**	**149972132**	**G/A**	**0.66**	**1.16 (1.09-1.22)**	**6.1X10-06**	**1.15 (1.10-1.20)**	**1.9X10-07**	**1.14 (1.05-1.22)**	**3.3X10-03**	**2.3X10-09**	**0**
rs12699477	7	1968953	T/C	0.39	1.22 (1.16-1.28)	1.1X10-10	1.20 (1.15-1.25)	5.9X10-13	-	-	5.9X10-13	0
rs17689040	**7**	**40920313**	**C/G**	**0.43**	**1.15 (1.10-1.21)**	**2.1X10-06**	**1.15 (1.10-1.19)**	**5.6X10-08**	**1.17 (1.09-1.25)**	**1.2X10-04**	**3.1X10-11**	**0**
rs17153755	**8**	**11611500**	**C/G**	**0.65**	**1.19 (1.13-1.25)**	**1.6X10-08**	**1.16 (1.11-1.21)**	**1.5X10-08**	**1.05 (0.97-1.14)**	**2.8X10-01**	**4.4X10-08**	**68**
rs7010162	8	70976505	C/T	0.62	1.13 (1.07-1.19)	7.6X10-05	1.15 (1.10-1.20)	4.5X10-08	-	-	4.5X10-08	0
rs7040024	9	845516	A/C	0.77	1.48 (1.41-1.55)	2.3X10-27	1.53 (1.47-1.59)	2.2X10-45	-	-	2.2X10-45	42
rs7107174	11	77996403	C/A	0.22	1.19 (1.11-1.27)	2.6X10-05	1.17 (1.10-1.24)	5.3X10-06	-	-	5.3X10-06	0
rs648090	**11**	**125071163**	**A/G**	**0.71**	**1.18 (1.12-1.25)**	**4.7X10-07**	**1.15 (1.10-1.20)**	**7.6X10-08**	**1.24 (1.15-1.33)**	**2.4X10-06**	**2.2X10-12**	**24**
rs2900333	12	14653867	C/T	0.63	1.17 (1.11-1.23)	3.7X10-07	1.20 (1.15-1.25)	9.3X10-13	-	-	9.3X10-13	17
rs4931000	**12**	**32141495**	**A/G**	**0.22**	**1.16 (1.09-1.23)**	**3.4X10-05**	**1.16 (1.11-1.22)**	**1.2X10-07**	**1.23 (1.13-1.32)**	**2.3X10-05**	**1.9X10-11**	**0**
rs7315956	**12**	**70563865**	**A/G**	**0.34**	**1.13 (1.06-1.19)**	**2.0X10-04**	**1.13 (1.08-1.18)**	**6.5X10-07**	**1.16 (1.08-1.25)**	**2.9X10-04**	**8.7X10-10**	**0**
rs3782181	12	88953561	C/A	0.81	2.06 (1.99-2.14)	1.4X10-80	2.07 (2.01-2.12)	3.3X10-129	-	-	3.3X10-129	40
rs1009647	**14**	**55880047**	**G/A**	**0.73**	**1.13 (1.06-1.19)**	**4.1X10-04**	**1.15 (1.10-1.21)**	**5.9X10-07**	**1.12 (1.03-1.21)**	**1.7X10-02**	**3.4X10-08**	**0**
rs11071896	**15**	**66821250**	**A/G**	**0.26**	**1.17 (1.10-1.24)**	**4.9X10-06**	**1.19 (1.14-1.25)**	**9.2X10-10**	**1.18 (1.09-1.27)**	**2.1X10-04**	**8.4X10-13**	**0**
rs56046484	**15**	**85605427**	**G/T**	**0.80**	**1.17 (1.09-1.24)**	**4.5X10-05**	**1.17 (1.11-1.23)**	**2.7X10-07**	**1.11 (1.01-1.21)**	**3.8X10-02**	**4.6X10-08**	**0**
rs4561483	16	11920037	A/G	0.35	1.11 (1.05-1.17)	1.1X10-03	1.13 (1.08-1.18)	5.5X10-07	-	-	5.5X10-07	3
rs7404843	**16**	**15530708**	**T/G**	**0.11**	**1.21 (1.12-1.30)**	**4.7X10-05**	**1.28 (1.21-1.35)**	**1.1X10-11**	**1.17 (1.05-1.29)**	**9.8X10-03**	**7.9X10-13**	**42**
rs8046148	16	50142944	A/G	0.79	1.10 (1.03-1.17)	7.8X10-03	1.16 (1.10-1.21)	4.5X10-07	-	-	4.5X10-07	58
rs4888262	16	74670458	C/T	0.50	1.17 (1.11-1.23)	1.2X10-07	1.18 (1.13-1.23)	6.9X10-12	-	-	6.9X10-12	0
rs55637647	16	88549264	C/G	0.38	1.15 (1.09-1.22)	7.2X10-06	1.17 (1.12-1.22)	2.9X10-09	-	-	2.9X10-09	0

rs7501939	17	36101156	T/C	0.61	1.22 (1.16-1.28)	7.8X10-11	1.25 (1.21-1.30)	2.8X10-20	-	-	2.8X10-20	55
rs9905704	17	56632543	G/T	0.68	1.27 (1.20-1.33)	2.4X10-13	1.27 (1.22-1.32)	3.4X10-20	-	-	3.4X10-20	0
rs9966612	**18**	**649311**	**A/G**	**0.32**	**1.16 (1.08-1.23)**	**9.8X10-05**	**1.17 (1.12-1.23)**	**2.1X10-08**	**1.13 (1.05-1.22)**	**5.1X10-03**	**4.4X10-10**	**0**
rs2195987	19	24149545	C/T	0.83	1.18 (1.08-1.28)	1.3X10-03	1.22 (1.15-1.28)	9.9X10-09	-	-	9.9X10-09	0
rs2241024	**19**	**28257393**	**G/A**	**0.80**	**1.24 (1.17-1.31)**	**4.7X10-09**	**1.23 (1.16-1.29)**	**1.4X10-10**	**1.32 (1.22-1.42)**	**6.3X10-08**	**1.0X10-16**	**29**
rs4599029	**19**	**54284689**	**G/T**	**0.74**	**1.18 (1.11-1.25)**	**1.1X10-06**	**1.16 (1.11-1.22)**	**7.6X10-08**	**1.10 (1.01-1.19)**	**3.8X10-02**	**1.4X10-08**	**19**
rs12481572	**20**	**50708054**	**A/T**	**0.20**	**1.20 (1.13-1.28)**	**9.5X10-07**	**1.20 (1.14-1.27)**	**1.4X10-08**	**1.23 (1.13-1.33)**	**3.7X10-05**	**2.5X10-12**	**0**
rs2839186	21	47690068	C/T	0.48	1.17 (1.11-1.23)	1.6X10-07	1.18 (1.14-1.23)	7.1X10-12	-	-	7.1X10-12	0
rs739525	**22**	**21332441**	**T/C**	**0.53**	**1.13 (1.06-1.19)**	**1.8X10-04**	**1.14 (1.09-1.19)**	**2.0X10-07**	**1.10 (1.02-1.18)**	**2.2X10-02**	**1.9X10-08**	**0**

485

486 [1] dbSNP rs number

487 [2] Alleles

488 [3] Risk Allele Frequency

489 [4] OR: per allele odds ratio

490 [5] P_{trend}: *P*-value for trend, via logistic regression

491 [6] P_{meta}: *P*-value for fixed effects meta-analysis

492 [7] I^2 heterogeneity index (0-100)

493

494 **Table 2 – Summary of functional annotation of all 44 TGCT risk loci**. Novel risk loci are highlighted in bold.

SNP	Cyto-band	bp (b37)	Genes in LD Block	Functional Evidence							Candidate causal Gene(s)	Functional Pathway
				Coding Variant	Promoter Variant	Functional Chromatin (ChIP-seq peaks)	TF binding motif disruption	Hi-C Contact(s)	eQTL	Functional Study		
rs4240895	**1p36.22**	**9,713,386**	***PIK3CD C1orf200***				**KLF4**					
rs2072499	1q22	156,169,610	*KIAA0446 SLC25A44*		*PMF1*	H3k4me1, H3k4me3, H3k9ac	PRDM1, CTCF	*BGLAP, CCT3, PAQR6, PMF1, SEMA4A, UBQLN4*	*CCT3*[1]		*PMF1*	**Microtubule/ chromosomal assembly**
											CCT3	
rs3790672	1q24.1	165,873,392	*UCK2*				GATA, NANOG, LHX8,					

							POU5F1, SOX9, PRDM1, CTCF					
rs7581030	**2p13.3**	**71,572,455**	*ZNF638*	*ZNF638*		**H3k4me3, H3k9ac**	**NANOG, POU5F1**	*PAIP2B*			*ZNF638*	**Transctiptional Regulation**
rs10510452	3p24.3	16,625,048	*DAZL*				GATA, NANOG, POU5F1	*OXNAD1*			*OXNAD1*	
rs11705932	3q23	141,818,850	*TFDP2*									
rs1510272	3q25.31	156,300,724					GATA, POU5F1, CTCF					
rs6821144	**4q21.1**	**76,520,651**	*CDKL2 G3BP2*			**H3k4me3, H3k9ac**	**SOX2**	*G3BP2*			*G3BP2*	
rs17021463	4q22.3	95,224,812	*SMARCAD1 HPGDS*	*SMARCAD1*	*SMARCAD1*	H3k4me3, H3k9ac	GATA, KLF4, NANOG, POU5F1, PRDM1	*ATOH1*	*SMARCAD1²* *ATOH1²*		*SMARCAD1*	**Transctiptional Regulation**
rs2720460	4q24	104,054,686	*CENPE*		*CENPE*	H3k4me3, H3k9ac	GATA, NANOG, LHX8, POU5F1, DMRT1	*MANBA, NFKB1, SLC39A8, TACR3*	*MANBA²*		*CENPE* / *MANBA*	**Microtubule/ chromosomal assembly**
rs4862848	**4q35.2**	**188,921,440**	*ZFP42*			**H3k4me3, H3k9ac**						
rs2736100	5p15.33	1,286,516	*TERT*									
rs3805663	5q31.1	134,366,200	*CATSPER3 PITX1 AK026965*			H3k4me3, H3k9ac						
rs4624820	5q31.3	141,681,788	*SPRY4*									
rs210138	6p21.31	33,542,538	*BAK1 AY383626 C6orf227*		*BAK1*	H3k4me3, H3k9ac		*GRM4*	*BAK1²*		*BAK1*	**KIT-MAPK**

rs11155671	**6q25.1**	**149,972,132**	***KATNA1 LATS1***				**GATA, KLF4, NANOG, SOX2, POU5F1, DMRT1, SOX9, PRDM1, CTCF**					
rs12699477	7p22.3	1,968,953	*MAD1L1*			H3k4me1, H3k9ac	NANOG, CTCF					
rs17689040	**7p14.1**	**40,920,313**										
rs17153755	**8p23.1**	**11,611,500**	***GATA4 c8orf***			**H3k4me3, H3k9ac**	**CTCF**	***GATA4***	***GATA4²***		***GATA4***	**Transctiptional Regulation**
rs7010162	8q13.3	70,976,505	*PRDM14*		*PRDM14*		GATA, PRDM1				*PRDM14*	**Transctiptional Regulation**
rs7040024	9p24.3	845,516	*DMRT1*		*DMRT1*	H3k4me3, H3k9ac	GATA, KLF4, CTCF				*DMRT1*	**Transctiptional Regulation**
rs7107174	11q14.1	77,997,936	*GAB2*	*USP35*	*GAB2*	H3k4me3, H3k9ac	GATA, KLF4, NANOG, LHX8, SOX2, POU5F1, DMRT1, SOX9, PRDM1, CTCF	*ALG8, GAB2, NARS2*			*GAB2* / *USP35*	**KIT-MAPK**
rs648090	**11q24.2**	**125,071,163**	***PKNOX2***			**H3k4me1**	**CTCF**					
rs2900333	12p13.1	14,653,867	*ATF7IP PLBD1*	*ATF7IP*	*ATF7IP*		GATA, NANOG, SOX2, POU5F1, CTCF				*ATF7IP*	**Telomerase Function**
rs4931000	**12p11.21**	**32,141,495**	***C12orf35***	***KIAA1551***			**NANOG, CTCF**				***KIAA1551***	
rs7315956	**12q15**	**70,563,865**	***CNOT2 KCNMB4***		***CNOT2***		**GATA, NANOG, LHX8, SOX2**				***CNOT2***	**Transctiptional Regulation**

rs ID	cytoband	position	genes			histone marks	transcription factors	gene list	gene	KITLG	final gene	pathway
rs3782181	12q21.32	88,953,561	*KITLG*			H3k4me3, H3k9ac	GATA, NANOG, SOX2, POU5F1, DMRT1, PRDM1, CTCF			*KITLG*	*KITLG*	**KIT-MAPK**
rs1009647	**14q22.3**	**55,880,047**	***ATG14 DLGAP5 FBXO34 LGALS3 TBPL2***		***ATG14***				*ATG14*[1]		***ATG14***	**Autophagy**
rs11071896	15q22.31	66,821,250	***MAP2K1 TIPIN***	***ZWILCH***	***MAP2K1***	**H3k4me1, H3k4me3, H3k9ac**	**GATA, NANOG, SOX2, DMRT1, SOX9, PRDM1, CTCF**	***DENND4A, IGDCC4, LCTL, MAP2K1, RAB11A, RPL4, SCARNA14, SNAPC5***	*SNAPC5*[2]		*MAP2K1* / *SNAPC5* / *ZWILCH*	**KIT-MAPK**
rs56046484	**15q25.2**	**85,605,427**	***PDE8A SLC28A1***				**GATA, NANOG, LHX8, CTCF**	***SLC28A1, WDR73***	*WDR73*[2]		***WDR73***	**Microtubule/ chromosomal assembly**
rs4561483	16p13.13	11,920,037	*BCAR4 CATX-11 RSL1D1*				KLF4, NANOG, CTCF	*LITAF*	*GSPT1*[3]		*GSPT1*	**Cell cycle**
rs7404843	**16p13.11**	**15,530,708**	***MPV17L***				**GATA, POU5F1, CTCF**					
rs8046148	16q12.1	50,142,944	*HEATR3 AF086132*		*HEATR3*	H3k4me1, H3k4me3, H3k9ac	GATA, SOX2, CTCF	*HEATR3*	*HEATR3*[1]		*HEATR3*	
rs4888262	16q23.1	74,670,458	*RFWD3*		*RFWD3*	H3k4me1, H3k4me3, H3k9ac	GATA, KLF4, NANOG, PRDM1	*CLEC18C, LDHD, RFWD3, WDR59*	*RFWD3*[1]		*RFWD3*	**Apoptosis/p53 pathway**
rs55637647	16q24.2	88,549,264	*ZFPM1*	*ZFPM1*		H3k4me1, H3k9ac	KLF4				*ZFPM1*	**Transctiptional Regulation**

SNP	Cytoband	Position	Gene	Gene	Gene	Histone	TF			Gene	Pathway
rs7501939	17q12	36,101,156	HNF1B		HNF1B	H3k4me3, H3k9ac	GATA, NANOG, SOX2, POU5F1, SOX9			HNF1B	**Transctiptional Regulation**
rs9905704	17q22	56,632,543	TEX14		TEX14		GATA, SOX2, DMRT1, SOX9		TEX14[1]	TEX14	**Microtubule/ chromosomal assembly**
rs9966612	**18p11.32**	**649,311**	**CLUL1 ENOSF1 TYMS**					**THOC1**		**THOC1**	**Apoptosis/p53 pathway**
rs2195987	19p12	24,149,545	AK125686		ZNF254	H3k4me3, H3k9ac	SOX2, POU5F1			ZNF254	**Transctiptional Regulation**
rs2241024	**19q11**	**28,257,393**									
rs4599029	**19q13.42**	**54,284,689**	**NLRP12**				**GATA, KLF4, NANOG**				
rs12481572	**20q13.2**	**50,708,054**					**POU5F1**	**NFATC2, SALL4**		**SALL4**	**Transctiptional Regulation**
rs2839186	21q22.3	47,690,068	MCM3APAS MCM3AP	C21orf58, PCNT		H3k4me1, H3k4me3, H3k9ac	KLF4, NANOG, DMRT1, CTCF			PCNT	**Microtubule/ chromosomal assembly**
rs739525	**22q11.21**	**21,332,441**	**AIFM3**		**AIFM3**	**H3k4me1, H3k4me3, H3k9ac**	**NANOG**			**AIFM3**	**Apoptosis/p53 pathway**

[1] Signficant vs threshold corrected for 96 multiple tests

[2] Nominally significant at P<0.05, see supplementary table 9 for exact P-values

[3] eQTL identified in previous study

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).

Running title: Identification of 19 new risk loci for TGCT

544 21. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome.
545 *Nature* **489**, 57-74 (2012).
546 22. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5
547 (2013).
548 23. Agnihotri, S. *et al.* A GATA4-regulated tumor suppressor network represses formation of
549 malignant human astrocytomas. *J Exp Med* **208**, 689-702 (2011).
550 24. Hellebrekers, D.M. *et al.* GATA4 and GATA5 are potential tumor suppressors and biomarkers
551 in colorectal cancer. *Clin Cancer Res* **15**, 3990-7 (2009).
552 25. Tsang, A.P. *et al.* FOG, a multitype zinc finger protein, acts as a cofactor for transcription
553 factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* **90**, 109-19 (1997).
554 26. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense
555 mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
556 27. Ketola, I. *et al.* Developmental expression and spermatogenic stage specificity of
557 transcription factors GATA-1 and GATA-4 and their cofactors FOG-1 and FOG-2 in the mouse
558 testis. *Eur J Endocrinol* **147**, 397-406 (2002).
559 28. Zheng, R. & Blobel, G.A. GATA Transcription Factors and Cancer. *Genes Cancer* **1**, 1178-88
560 (2010).
561 29. Kurimoto, K., Yamaji, M., Seki, Y. & Saitou, M. Specification of the germ cell lineage in mice: a
562 process orchestrated by the PR-domain proteins, Blimp1 and Prdm14. *Cell Cycle* **7**, 3514-8
563 (2008).
564 30. Ohinata, Y. *et al.* A signaling principle for the specification of the germ cell lineage in mice.
565 *Cell* **137**, 571-84 (2009).
566 31. Yamaji, M. *et al.* Critical function of Prdm14 for the establishment of the germ cell lineage in
567 mice. *Nat Genet* **40**, 1016-22 (2008).
568 32. Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J. & Sinclair, A.H. Conservation of a sex-
569 determining gene. *Nature* **402**, 601-2 (1999).
570 33. Rao, S. *et al.* Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. *Mol Cell*
571 *Biol* **30**, 5364-80 (2010).
572 34. Greenbaum, M.P. *et al.* TEX14 is essential for intercellular bridges and fertility in male mice.
573 *Proc Natl Acad Sci U S A* **103**, 4982-7 (2006).
574 35. Mondal, G., Ohashi, A., Yang, L., Rowley, M. & Couch, F.J. Tex14, a Plk1-regulated protein, is
575 required for kinetochore-microtubule attachment and regulation of the spindle assembly
576 checkpoint. *Mol Cell* **45**, 680-95 (2012).
577 36. Jinks, R.N. *et al.* Recessive nephrocerebellar syndrome on the Galloway-Mowat syndrome
578 spectrum is caused by homozygous protein-truncating mutations of WDR73. *Brain* **138**,
579 2173-90 (2015).
580 37. Colin, E. *et al.* Loss-of-function mutations in WDR73 are responsible for microcephaly and
581 steroid-resistant nephrotic syndrome: Galloway-Mowat syndrome. *Am J Hum Genet* **95**, 637-
582 48 (2014).
583 38. Petrovic, A. *et al.* The MIS12 complex is a protein interaction hub for outer kinetochore
584 assembly. *J Cell Biol* **190**, 835-52 (2010).
585 39. Rao, C.V., Yamada, H.Y., Yao, Y. & Dai, W. Enhanced genomic instabilities caused by
586 deregulated microtubule dynamics and chromosome segregation: a perspective from
587 genetic studies in mice. *Carcinogenesis* **30**, 1469-74 (2009).
588 40. Barisic, M. *et al.* Mitosis. Microtubule detyrosination guides chromosomes during mitosis.
589 *Science* **348**, 799-803 (2015).
590 41. Ma, W. & Viveiros, M.M. Depletion of pericentrin in mouse oocytes disrupts microtubule
591 organizing center function and meiotic spindle organization. *Mol Reprod Dev* **81**, 1019-29
592 (2014).

Running title: Identification of 19 new risk loci for TGCT

593   42.   Litchfield, K. *et al.* Whole-exome sequencing reveals the mutational spectrum of testicular
594   germ cell tumours. *Nat Commun* **6**, 5973 (2015).
595   43.   Zeron-Medina, J. *et al.* A polymorphic p53 response element in KIT ligand influences cancer
596   risk and has undergone natural selection. *Cell* **155**, 410-22 (2013).
597   44.   De Miguel, M.P., Cheng, L., Holland, E.C., Federspiel, M.J. & Donovan, P.J. Dissection of the c-
598   Kit signaling pathway in mouse primordial germ cells by retroviral-mediated gene transfer.
599   *Proc Natl Acad Sci U S A* **99**, 10458-63 (2002).
600   45.   Yu, M. *et al.* The scaffolding adapter Gab2, via Shp-2, regulates kit-evoked mast cell
601   proliferation by activating the Rac/JNK pathway. *J Biol Chem* **281**, 28615-26 (2006).
602   46.   Penegar, S. *et al.* National study of colorectal cancer genetics. *Br J Cancer* **97**, 1305-9 (2007).
603   47.   Eisen, T., Matakidou, A., Houlston, R. & Consortium, G. Identification of low penetrance
604   alleles for lung cancer: the GEnetic Lung CAncer Predisposition Study (GELCAPS). *BMC*
605   *Cancer* **8**, 244 (2008).
606   48.   Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands
607   of genomes. *Nat Methods* **9**, 179-81 (2012).
608   49.   Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate
609   genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*
610   **44**, 955-9 (2012).
611   50.   Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev*
612   *Genet* **11**, 499-511 (2010).
613   51.   Cuppen, E. Genotyping by Allele-Specific Amplification (KASPar). *CSH Protoc* **2007**, pdb
614   prot4841 (2007).
615   52.   Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for
616   genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
617   53.   Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-
618   scale, case-control association study. *Nat Genet* **37**, 1243-6 (2005).
619   54.   Litchfield, K. *et al.* Multi-stage genome wide association study identifies new susceptibility
620   locus for testicular germ cell tumour on chromosome 3q25. *Hum Mol Genet* (2014).
621   55.   Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking
622   quantity. *Nat Genet* **42**, 436-40 (2010).
623   56.   Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**,
624   1539-58 (2002).
625   57.   Scales, M., Jager, R., Migliorini, G., Houlston, R.S. & Henrion, M.Y. visPIG--a web tool for
626   producing multi-region, multi-track, multi-scale plots of genetic data. *PLoS One* **9**, e107497
627   (2014).
628   58.   Pharoah, P.D.P. *et al.* Polygenic susceptibility to breast cancer and implications for
629   prevention. *Nature Genetics* **31**, 33-36 (2002).
630   59.   Cancer Research UK 2014, Testicular Cancer Life Time Risk, Cancer Research UK.
631   60.   Cowper-Sal lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of
632   chromatin for FOXA1 and alter gene expression. *Nat Genet* **44**, 1191-8 (2012).
633   61.   Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of
634   chromatin looping. *Cell* **159**, 1665-80 (2014).
635   62.   Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution
636   capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
637   63.   Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310
638   (2015).
639   64.   Jonathan Cairns, P.F.-P., Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol,
640   Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser,

Running title: Identification of 19 new risk loci for TGCT

641   Mikhail Spivakov. CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C
642   data. *BioRxiv* (2016).

643 65.   Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115
644   (2012).

645 66.   Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image
646   analysis. *Nat Methods* **9**, 671-5 (2012).

647 67.   Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and
648   characterization. *Nat Methods* **9**, 215-6 (2012).

649 68.   Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian
650   Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.
651   *Nucleic Acids Res* **33**, D514-7 (2005).

652 69.   Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation,
653   and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*
654   **40**, D930-4 (2012).

655 70.   Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in
656   ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).

657 71.   Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T. & Flicek, P.R. The ensembl regulatory
658   build. *Genome Biol* **16**, 56 (2015).

659

660

Running title: Identification of 19 new risk loci for TGCT
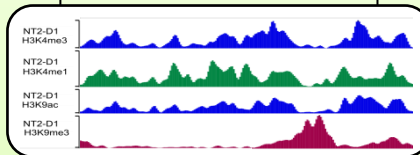
GWAS & replication

Oncoarray
3,206 cases
7,422 controls

Published GWAS
2,312 cases
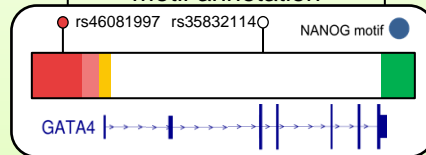11,633 controls

Meta-GWAS discovery
5,518 cases / 19,055 controls

Replication genotyping
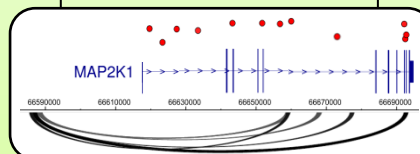1,801 cases / 4,027 controls
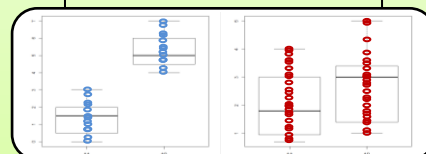
Functional annotation

ChIP-seq
chromatin profiling

NT2-D1 H3K4me3
NT2-D1 H3K4me1
NT2-D1 H3K9ac
NT2-D1 H3K9me3

Promoter/coding
variant & TF binding
motif annotation

rs46081997   rs35832114   NANOG motif

GATA4

HiC promoter capture

MAP2K1

66590000 66610000 66630000 66650000 66670000 66690000
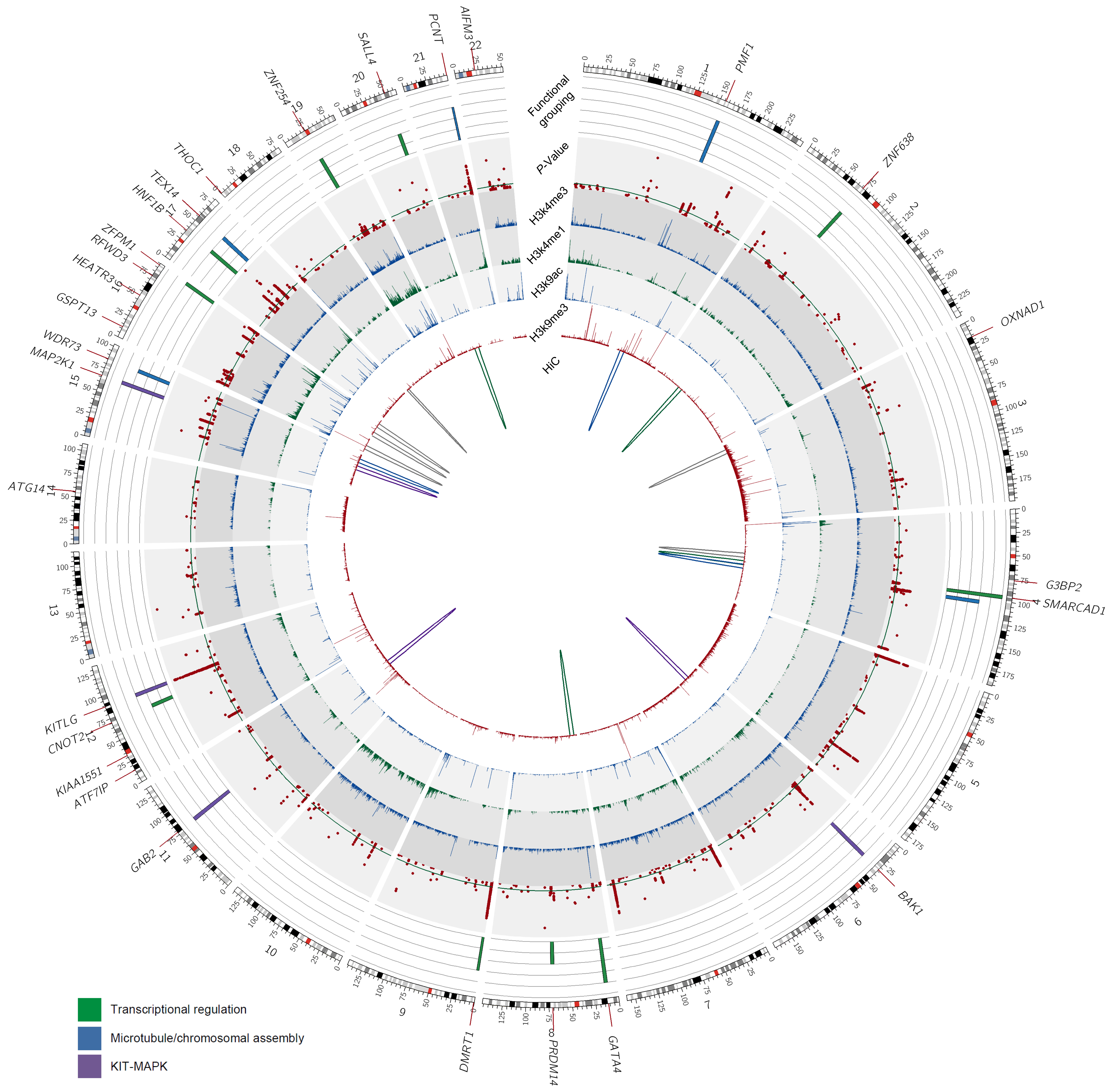
eQTL analysis

Functional grouping

P-Value

H3k4me3

H3k4me1

H3k9ac

H3k9me3

HiC

Transcriptional regulation

Microtubule/chromosomal assembly

KIT-MAPK

Figure 3A.

Figure 3B.

Figure 3C.

■ Active promoter	■ Insulator
■ Weak promoter	■ Transcriptional transition/elongation
■ Inactive/poised promoter	■ Weakly transcribed
■ Strong enhancer	■ Polycomb-repressed
■ Weak/poised enhancer	■ Heterochromatin; low signal, repetitive or CNV
● TF Binding Motif	▲ Sentinel SNP