doi:10.1038/nature15540

Genetic predisposition to neuroblastoma mediated by a *LMO1* super-enhancer polymorphism

Derek A. Oldridge[1,2]*, Andrew C. Wood[3]*, Nina Weichert-Leahey[4,5], Ian Crimmins[1], Robyn Sussman[1], Cynthia Winter[1], Lee D. McDaniel[1], Maura Diamond[1], Lori S. Hart[1], Shizhen Zhu[6], Adam D. Durbin[4,5], Brian J. Abraham[7], Lars Anders[7], Lifeng Tian[8], Shile Zhang[9], Jun S. Wei[9], Javed Khan[9], Kelli Bramlett[10], Nazneen Rahman[11], Mario Capasso[12,13], Achille Iolascon[12,13], Daniela S. Gerhard[14], Jaime M. Guidry Auvil[14], Richard A. Young[7], Hakon Hakonarson[8,15], Sharon J. Diskin[1,15,16], A. Thomas Look[4,5] & John M. Maris[1,15,16]

Neuroblastoma is a paediatric malignancy that typically arises in early childhood, and is derived from the developing sympathetic nervous system. Clinical phenotypes range from localized tumours with excellent outcomes to widely metastatic disease in which long-term survival is approximately 40% despite intensive therapy. A previous genome-wide association study identified common polymorphisms at the *LMO1* gene locus that are highly associated with neuroblastoma susceptibility and oncogenic addiction to LMO1 in the tumour cells[1]. Here we investigate the causal DNA variant at this locus and the mechanism by which it leads to neuroblastoma tumorigenesis. We first imputed all possible genotypes across the *LMO1* locus and then mapped highly associated single nucleotide polymorphism (SNPs) to areas of chromatin accessibility, evolutionary conservation and transcription factor binding sites. We show that SNP rs2168101 G>T is the most highly associated variant (combined $P = 7.47 \times 10^{-29}$, odds ratio 0.65, 95% confidence interval 0.60–0.70), and resides in a super-enhancer defined by extensive acetylation of histone H3 lysine 27 within the first intron of *LMO1*. The ancestral G allele that is associated with tumour formation resides in a conserved GATA transcription factor binding motif. We show that the newly evolved protective TATA allele is associated with decreased total *LMO1* expression ($P = 0.028$) in neuroblastoma primary tumours, and ablates GATA3 binding ($P < 0.0001$). We demonstrate allelic imbalance favouring the G-containing strand in tumours heterozygous for this SNP, as demonstrated both by RNA sequencing ($P < 0.0001$) and reporter assays ($P = 0.002$). These findings indicate that a recently evolved polymorphism within a super-enhancer element in the first intron of *LMO1* influences neuroblastoma susceptibility through differential GATA transcription factor binding and direct modulation of *LMO1* expression in *cis*, and this leads to an oncogenic dependency in tumour cells.

Genome-wide association study (GWAS) efforts frequently identify highly statistically significant genetic associations within non-coding regulatory regions of the genome, but the underlying causal DNA sequence variations have only been identified in a few instances. A neuroblastoma GWAS has identified several disease susceptibility loci[1–7], with the signal within the LIM domain only 1 (*LMO1*) locus at 11p15 (ref. 1), a transcriptional co-regulator containing two zinc finger LIM domains that nucleate and regulate transcription factor complexes, being most robust. The main members of the LMO gene family, *LMO1–4*, are all implicated in cancer including *LMO1* and *LMO2* translocations in T-cell leukaemia[8], and we previously provided the first evidence that *LMO1* was a bona fide neuroblastoma oncogene[1]. Here, we sought to identify the causal polymorphism(s) driving the *LMO1* genetic association with neuroblastoma susceptibility as a basis for understanding neuroblastoma initiation and addiction mechanisms.

We first performed fine mapping of associated germline SNPs and indels at the *LMO1* gene locus by imputation to the 1000 Genomes Project for our European-American neuroblastoma GWAS[6]. This identified 27 SNPs with minor allele frequency (MAF) >0.01 and an association $P < 1 \times 10^{-5}$ (Fig. 1a and Extended Data Table 1). We further prioritized associated variants by evolutionary conservation, and by their regulatory potential inferred through neuroblastoma-specific DNase I hypersensitivity mapping and chromatin immunoprecipitation sequencing (ChIP-seq) from the ENCODE Consortium (Fig. 1b). These data showed that the most significantly associated SNP at the *LMO1* locus (rs2168101, odds ratio = 0.67, $P = 4.14 \times 10^{-16}$) resides within a highly conserved and active enhancer region inferred by DNase I sensitivity and p300 binding in the SKNSH neuroblastoma cell line (Fig. 1b). Notably, we found no rare or common non-synonymous coding variants in *LMO1* in a combined cohort of 482 unique neuroblastoma cases with germline whole-genome ($n = 136$), whole-exome ($n = 222$) and/or targeted DNA sequencing ($n = 183$) (see Extended Data Table 2 and Supplementary Data).

Because rs2168101 genotypes were imputed in our analyses (Extended Data Fig. 1), we next directly genotyped this SNP in 146 out of 2,101 European-American cases, and measured an 86% imputation accuracy (Supplementary Table 1). We additionally directly genotyped rs2168101 in two independent cohorts from the UK and Italy, with both showing robust replication (Table 1). We did not observe replication in an independent African-American cohort. Notably, the protective T allele is common in Europeans (CEU HapMap: 28%) and East Asians (CHB+JPT HapMap: 32%), but is rare or absent in Africans, indicating recent expansion of the rs2168101 protective allele in non-African human populations. Meta-analysis demonstrated a combined association $P = 7.47 \times 10^{-29}$ across 8,553 controls and 3,254 cases (Table 1).

As causal SNPs driving GWAS associations may disrupt transcription factor binding at distal enhancers, we sought to identify candidate SNPs disrupting known JASPAR motifs[9], which revealed that lead

[1]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [2]Medical Scientist Training Program, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [3]Department of Molecular Medicine and Pathology, University of Auckland, Auckland, Auckland Region 1142, New Zealand. [4]Department of Pediatric Oncology, Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02215, USA. [5]Division of Pediatric Hematology/Oncology, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [6]Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minnesota 55905, USA. [7]Whitehead Institute for Biomedical Research and MIT, Boston, Massachusetts 02142, USA. [8]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [9]Pediatric Oncology Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. [10]Thermo Fisher Scientific, Austin, Texas 78744, USA. [11]The Institute of Cancer Research, London SM2 5NG, UK. [12]University of Naples Federico II, 80131 Naples, Italy. [13]CEINGE Biotecnologie Avanzate, 80131 Naples, Italy. [14]Office of Cancer Genomics, National Cancer Institute, Bethesda, Maryland 20892, USA. [15]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [16]Abramson Family Cancer Research Institute, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

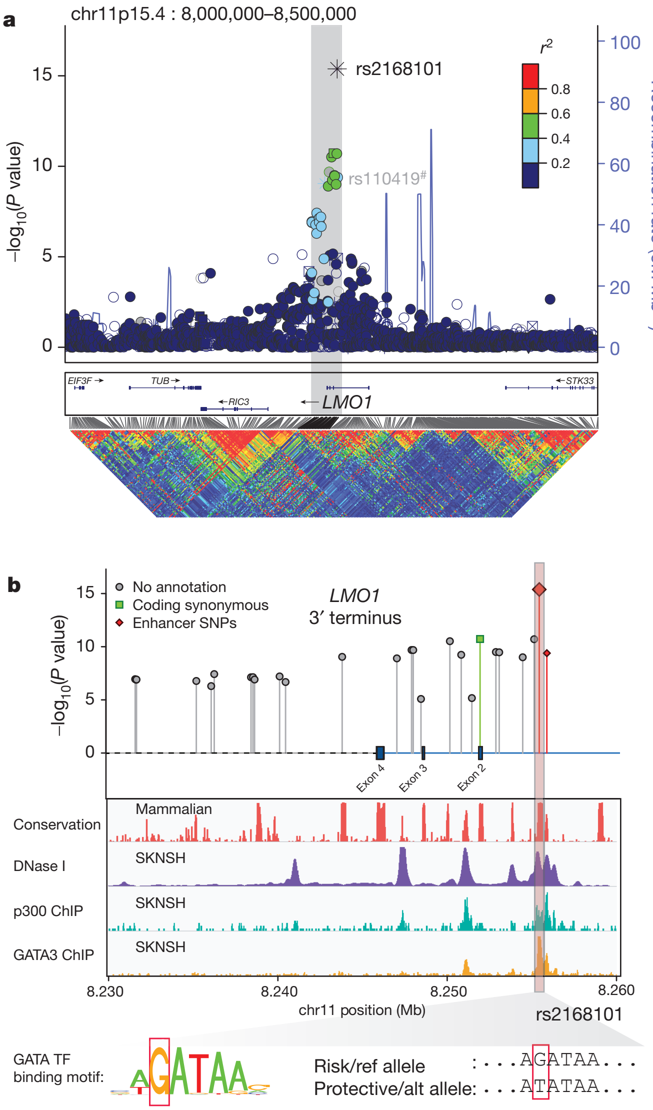


Figure 1 | Imputation-based GWAS and epigenomic profiling by ENCODE identifies rs2168101 as a candidate functional SNP at *LMO1*. **a**, Manhattan plot for neuroblastoma GWAS (cases = 2,101; controls = 4,202). The neuroblastoma-associated region falls within a 40-kilobase (kb) haplotype block (grey box) in Europeans, encompassing the *LMO1* 3′-terminus. rs2168101 is the most associated variant and is moderately correlated (maximum $r^2 = 0.52$) with other variants. The sentinel SNP reported previously, rs110419, is also highlighted (#). **b**, Associated variants ($P < 1 \times 10^{-5}$) are plotted with ENCODE tracks for neuroblastoma cell line SKNSH. Two SNPs, rs2168101 and rs7948497, were annotated 'enhancer SNPs' based on overlapping DNase peaks binding p300. The rs2168101 G>T SNP disrupts an evolutionarily conserved GATA transcription factor (TF) motif (5′-A[G/T]ATAA-3′). SKNSH has an rs2168101 = G/G genotype that preserves GATA binding, supported by ENCODE GATA3 ChIP-seq.

candidate SNP rs2168101 resides in a highly conserved GATA-binding motif (5′-A[G/T]ATAA-3′, mammalian phastCons score = 100%) (Fig. 1b). ENCODE transcription factor ChIP-seq confirmed GATA2 and GATA3 binding at the rs2168101 GATA motif in the neuroblastoma cell lines SKNSH and SHSY5Y, which are G/G homozygous, thereby preserving the GATA motif (Fig. 1b). No other associated variant showed this unique combination of evolutionary conservation, active enhancer localization, and disruption of a transcription factor binding motif, including the sentinel SNP rs110419 ($P = 1.17 \times 10^{-13}$) from our original report[1].

To test for the possibility of multiple statistical signals or enhancers not marked by conservation or p300 at the *LMO1* locus, we repeated association testing conditional on imputed rs2168101 genotypes and

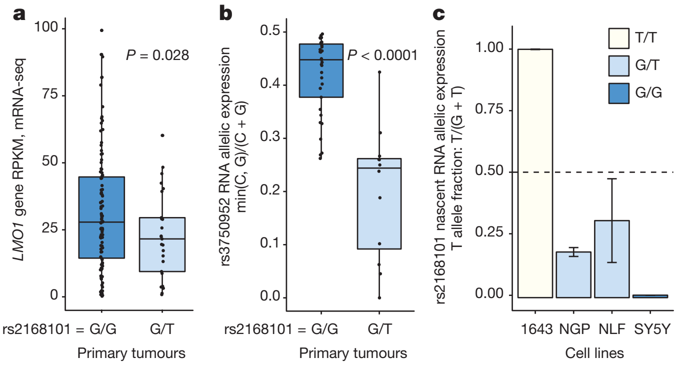


Figure 2 | RNA expression of *LMO1* associates with rs2168101 genotype consistent with regulation in *cis*. **a**, mRNA-seq across 127 primary tumours genotyped for rs2168101 (G/G = 102, G/T = 25, T = 0) revealed a significant decrease in *LMO1* gene expression between G/T and G/G tumours (*t*-test $P = 0.028$). RPKM, reads per kilobase per million reads. **b**, Using the synonymous exonic SNP, rs3750952, to measure allelic expression by mRNA-seq revealed significantly more allelic imbalance in 12 heterozygous neuroblastoma tumours (rs2168101 = G/T) than in 33 homozygous tumours (rs2168101 = G/G) (*t*-test $P = 5.3 \times 10^{-5}$). **c**, Allelic expression for rs2168101 from targeted nascent RNA-seq in four neuroblastoma cell lines. The two heterozygous cell lines (rs2168101 = G/T) exhibited significantly reduced T-allele expression compared to the G allele (*t*-test $P = 1.6 \times 10^{-4}$ and 1.5×10^{-2} for NGP and NLF, respectively; error bars denote 95% confidence intervals across $n = 3$ duplicate experiments).

observed no significant variants after multiple test correction (most significant variant: rs34544683, nominal $P = 9.0 \times 10^{-4}$, Bonferroni $P = 1$; Extended Data Fig. 2a). To test whether the rs2168101 signal can be equally captured by other variants, we also performed reciprocal association tests for rs2168101 conditioned on all 27 other SNPs within 1.5 megabases (Mb) of *LMO1* passing thresholds MAF > 0.01 and nominal $P < 1 \times 10^{-5}$. Notably, rs2168101 remained significant across all conditional tests (worst-case nominal $P = 2.6 \times 10^{-7}$, Bonferroni $P = 0.002$; Extended Data Fig. 2b). These results are consistent with a single underlying signal at the *LMO1* locus, and re-affirm that rs2168101 is the single best causal SNP candidate, because its association with neuroblastoma cannot be accounted for by other variants.

We next sought to determine whether rs2168101 genotypes were associated with *LMO1* expression by messenger RNA sequencing (mRNA-seq) of 127 primary high-risk neuroblastoma tumours. Genotyping rs2168101 yielded 102 G/G, 25 G/T and no T/T tumours (MAF = 9.8%). We observed significantly higher *LMO1* expression in G/G versus G/T genotype tumours (*t*-test $P = 0.028$; Fig. 2a). Notably, the absence of protective homozygous T/T genotypes in this high-risk neuroblastoma cohort is consistent with our previous observation that the risk alleles predispose to the high-risk phenotypic subset[1] (for clinical covariate associations, see Extended Data Table 3). Accordingly, the rs2168101 G/G genotype is highly associated with decreased neuroblastoma patient event-free ($P = 0.0004$) and overall ($P = 0.0004$) survival compared to G/T and T/T genotypes together in our European-American cohort (Extended Data Fig. 3). Two cell lines with homozygous T/T or T/− genotypes expressed *LMO1* at comparatively lower levels than cell lines containing the G allele (Extended Data Fig. 4a).

GATA transcription factors mediate chromatin looping and facilitate long-range enhancer–promoter interactions to regulate target gene expression[10]. We therefore sought to confirm allelic imbalance of *LMO1* transcripts (a hallmark of gene regulation in *cis*), which could result from differential GATA-binding caused by rs2168101. First, because the rs2168101 intronic SNP is not detectable by mRNA-seq, we identified the *LMO1* exonic synonymous SNP, rs3750952, which can measure allelic expression in the heterozygous state. We identified 45 tumours with

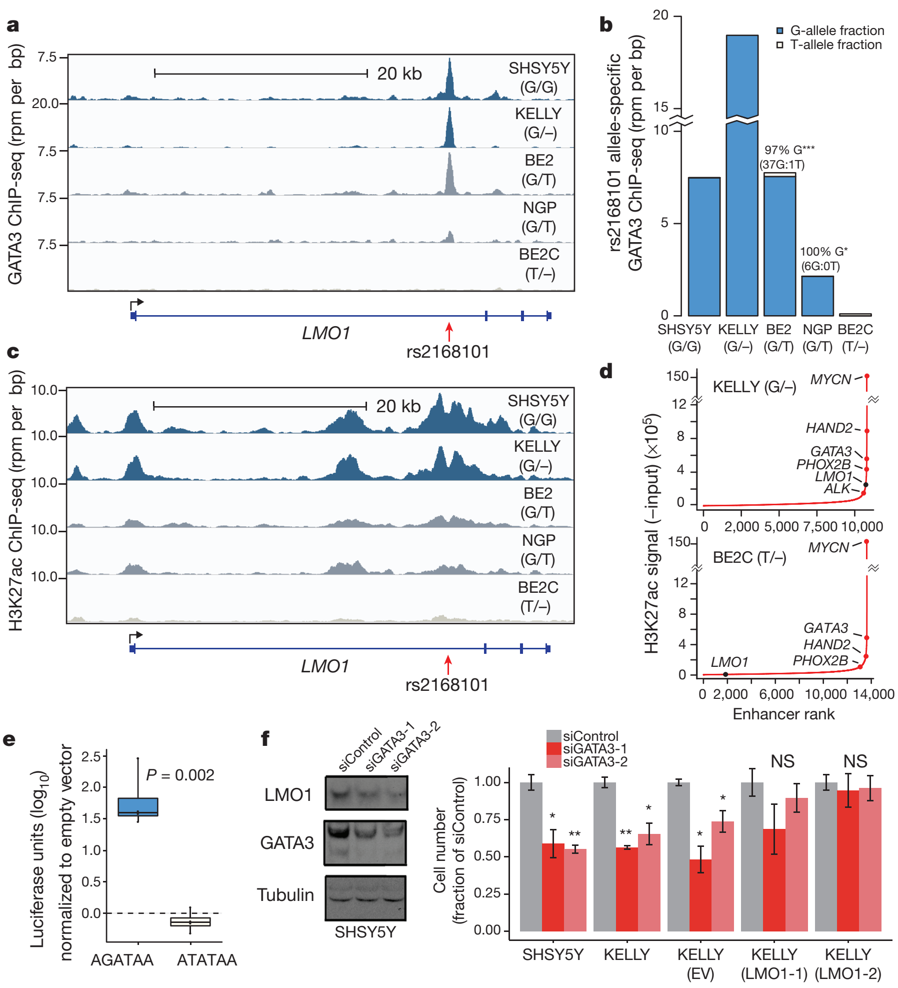


Figure 3 | The rs2168101 protective T allele disrupts GATA3 binding and negatively associates with the *LMO1* super-enhancer in neuroblastoma cells. a, Normalized GATA3 ChIP-seq signal at rs2168101 in five neuroblastoma cell lines (rs2168101 genotypes: SHSY5Y = G/G, KELLY = G/−, BE2 = G/T, NGP = G/T, BE2C = T/−). rpm, reads per million. **b**, Allele-specific binding of GATA3 at rs2168101. GATA3 binding highly favoured the risk G allele in heterozygous lines (BE2: 0.97 G-allele fraction from 38 reads, 95% confidence interval: 0.86–1.00, binomial $P = 2.8 \times 10^{-10}$; NGP: 1.00 G-allele fraction from 6 reads, 95% confidence interval: 0.54–1.00, binomial $P = 0.03$). **c**, Normalized ChIP-seq signal for H3K27ac at rs2168101. **d**, Ranked H3K27ac signal across all enhancers in MYCN-amplified KELLY and BE2C lines. Super-enhancers associate with key neuroblastoma genes, highlighted on the curve. There is an *LMO1*-associated super-enhancer in G-allele-containing lines SHSY5Y, KELLY and BE2, but not in BE2C, which lacks the G allele. **e**, Luciferase reporter assay for *LMO1* enhancer region. The risk G allele preserved enhancer activity (*t*-test $P = 0.002$ across $n = 4$ independent clones, each with $n = 5$ technical replicates), whereas the protective T allele was indistinguishable from empty vector. **f**, Left, protein blots for GATA3, LMO1 and tubulin in SHSY5Y cells treated with control (siControl) short interfering RNAs (siRNAs), or with siRNAs targeting GATA3 (siGATA3-1 and siGATA3-2), at 72 h post-treatment. Right, cell counts for cell lines SHSY5Y, KELLY, KELLY stably overexpressing control vector (EV) and KELLY with forced LMO1 overexpression (LMO1-1 and LMO1-2) treated with siRNAs at 72 h post-transfection (see Extended Data Fig. 7 for complete growth curves). Rescue of suppressed cell growth after GATA3 depletion by forced LMO1 expression was observed at 72 h. Error bars denote ± s.e.m. $^*P < 0.05$, $^{**}P < 0.001$ by *t*-test. $n = 3$ independent transfections, $n = 9$ technical replicates.

the necessary rs3750952 = C/G genotype, and then directly genotyped rs2168101 (G/G = 33, G/T = 12, T/T = 0) in this panel. By mRNA-seq, there was greater allelic imbalance in 12 tumours that were heterozygous for rs2168101 (G/T) than in 33 homozygous tumours (rs2168101 = G/G; *t*-test $P < 0.0001$; Fig. 2b). We next used targeted sequencing of nuclear-enriched nascent RNAs in four neuroblastoma cell lines (G/G = 1, G/T = 2, T/T = 1) to provide direct ascertainment of allele-specific expression at rs2168101. In both heterozygous lines, we observed allelic imbalance that significantly favoured the risk G allele over the protective T allele (Fig. 2c). Collectively, these results indicate that the intact GATA motif at rs2168101 results in significantly higher *LMO1* expression levels than the TATA coded by the alternative allele. Allelic imbalance of *LMO1* was not driven by somatic DNA alterations (for example, loss of heterozygosity) that could affect allelic dosage (Extended Data Fig. 4b).

Examination of neuroblastoma transcriptome data for 127 primary tumours showed that *GATA2* and *GATA3* are overexpressed compared to other members of the GATA transcription factor family (Extended Data Fig. 5a), and that GATA3 is the most highly expressed. Additionally, protein immunoblotting showed that GATA3 is uniformly highly expressed in neuroblastoma cell lines, while LMO1 is highly expressed in the G/G (SKNSH and SHSY5Y), G/− (KELLY) and G/T (IMR32) cell lines, but only barely detectable in the BE2C cell line that lacks a G allele at the rs2168101 locus (Extended Data Fig. 5b). We therefore performed ChIP-seq using a GATA3 antibody in neuroblastoma cell lines, and observed robust GATA3 binding at rs2168101 in lines containing the G allele (SHSY5Y, KELLY, BE2 and NGP) but not in a line containing only a T allele (BE2C; Fig. 3a). We then specifically considered GATA3 ChIP-seq reads overlapping rs2168101, and we observed strong preferential binding to the G allele in the G/T

Table 1 | Replication and meta-analysis of rs2168101 association

SNP	Ref/alt (major/minor) allele	Cohort	MAF cases	MAF controls	Additive *P* value	Additive odds ratio	Het odds ratio (GT vs GG)	Hom odds ratio (TT vs GG)
rs2168101	G/T	European-American*	0.242 ($n = 2{,}101$)	0.313 ($n = 4{,}202$)	4.14×10^{-16}	0.67 (0.61–0.74)	0.69 (0.62–0.77)	0.52 (0.42–0.64)
		Italian	0.164 ($n = 420$)	0.250 ($n = 751$)	2.07×10^{-6}	0.61 (0.50–0.75)	0.57 (0.44–0.74)	0.40 (0.21–0.75)
		UK	0.190 ($n = 369$)	0.311 ($n = 1{,}109$)	5.86×10^{-10}	0.56 (0.47–0.68)	0.51 (0.39–0.66)	0.31 (0.18–0.53)
		African-American*	0.0865 ($n = 364$)	0.0891 ($n = 2{,}491$)	0.20	0.79 (0.56–1.13)	0.96 (0.71–1.30)	1.07 (0.38–3.04)
		Combined			7.47×10^{-29}	0.65 (0.60–0.70)	0.67 (0.61–0.73)	0.49 (0.41–0.59)

Alt, alternative; het, heterozygous; hom, homozygous; ref, reference. Odds ratio 95% confidence intervals are shown in parentheses.
*Imputed genotypes and correction for population stratification.

heterozygous cell lines BE2 (0.97 G-allele fraction from 38 reads, 95% confidence interval: 0.86–1.00, binomial test $P = 2.8 \times 10^{-10}$) and NGP (1.00 G-allele fraction from 6 reads, 95% confidence interval: 0.54–1.00, binomial test $P = 0.03$; Fig. 3b).

Acetylation of histone H3 at lysine 27 (H3K27ac) is a hallmark of active enhancers[11], and ChIP-seq analysis of SHSY5Y (G/G; not *MYCN* amplified), KELLY (G/−; *MYCN* amplified), BE2 (G/T; *MYCN* amplified) and NGP (G/T; *MYCN* amplified) neuroblastoma cells showed extensive H3K27 acetylation in the first intron of *LMO1* across rs2168101, which was not observed in BE2C (T/−; *MYCN* amplified; Fig. 3c). This region is classified as a super-enhancer in G-allele-containing lines SHSY5Y, KELLY and BE2 based on enhancer clustering and especially high H3K27ac signal (NGP was just below the threshold; see Methods), a pattern also observed for other known oncogenes and tumour suppressor genes in this disease[12] (Fig. 3d and Extended Data Fig. 6a). No super-enhancer was observed in BE2C or Jurkat T-ALL cells that also express *LMO1* (ref. 13), or in other non-neuroblastoma tissues from ENCODE (Fig. 3d and Extended Data Fig. 6b, c). These results are consistent with recent evidence that disease-associated SNPs frequently affect enhancers that are specific to disease-relevant cell lines and tumour histology, and control developmental stage and tissue-specific gene expression[12,14–18].

We next performed luciferase reporter assays to measure the effect of rs2168101 alleles on enhancer activity. HEK293T cells transfected with constructs containing the risk G allele demonstrated 30–300-fold higher normalized luminescence compared to the T allele (*t*-test $P = 0.002$, Fig. 3e), whereas luciferase activity of the T allele was not significantly different from empty vector, indicating that the intact GATA motif is required for robust enhancer activity. Finally, knockdown of GATA3 in SHSY5Y and KELLY cells resulted in both decreased LMO1 protein levels and suppression of cell growth that was rescued by LMO1 overexpression (Fig. 3f and Extended Data Fig. 7), indicating the central role of GATA3 in regulating LMO1 expression levels in neuroblastoma.

Taken together, these data demonstrate the underlying molecular mechanism for a highly robust genetic association to neuroblastoma, mediated by a single common causal SNP rs2168101 that disrupts a GATA transcription factor binding site within a tissue-specific super-enhancer element. The rarity or absence of the protective allele in African populations and its relative depletion in African-Americans may partially explain the more aggressive clinical course in African-American children[19]. Moreover, this work further confirms the utility of association studies to define clinically relevant oncogenic pathways. Finally, the dependence of neuroblastoma cells on super-enhancer-mediated *LMO1* expression provides another potential mechanism for the sensitivity of these tumours to inhibitors of the transcriptional machinery such as CDK7 and BET bromodomain proteins[14,16], demonstrating the potential of translating basic mechanistic insights of tumour initiation towards novel therapeutic strategies.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 February; accepted 2 September 2015.
Published online 11 November 2015.

1. Wang, K. *et al.* Integrative genomics identifies *LMO1* as a neuroblastoma oncogene. *Nature* **469,** 216–220 (2011).
2. Maris, J. M. *et al.* Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.* **358,** 2585–2593 (2008).
3. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459,** 987–991 (2009).
4. Capasso, M. *et al.* Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma. *Nature Genet.* **41,** 718–723 (2009).
5. Nguyễn lễ, B. *et al.* Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS Genet.* **7,** e1002026 (2011).
6. Diskin, S. J. *et al.* Common variation at 6q16 within HACE1 and LIN28B influences susceptibility to neuroblastoma. *Nature Genet.* **44,** 1126–1130 (2012).
7. Diskin, S. J. *et al.* Rare variants in TP53 and susceptibility to neuroblastoma. *J. Natl. Cancer Inst.* **106,** dju047 (2014).
8. Matthews, J. M., Lester, K., Joseph, S. & Curtis, D. J. LIM-domain-only proteins in cancer. *Nature Rev. Cancer* **13,** 111–122 (2013).
9. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** D142–D147 (2014).
10. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149,** 1233–1244 (2012).
11. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107,** 21931–21936 (2010).
12. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
13. Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22,** 209–221 (2012).
14. Chipumuro, E. *et al.* CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* **159,** 1126–1139 (2014).
15. Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110,** 17921–17926 (2013).
16. Puissant, A. *et al.* Targeting MYCN in neuroblastoma by BET bromodomain inhibition. *Cancer Discov.* **3,** 308–323 (2013).
17. Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res.* **73,** 4180–4184 (2013).
18. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
19. Henderson, T. O. *et al.* Racial and ethnic disparities in risk and survival in children with neuroblastoma: a Children's Oncology Group study. *J. Clin. Oncol.* **29,** 76–82 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported in part by NIH grants R01-CA124709 (J.M.M.), R01-CA180692 (J.M.M. and A.T.L.), R00-CA151869 (S.J.D.), RC1MD004418 to the TARGET consortium, 1K99CA178189 (S.Z.), T32-HG000046 (D.A.O.), R01-CA109901 (R.A.Y.), the Giulio D'Angio Endowed Chair (J.M.M.), the PressOn Foundation (J.M.M.), Andrew's Army Foundation (J.M.M.), the Abramson Family Cancer Research Institute (J.M.M.), the Brooke Mulford Foundation (J.M.M.), the University of Pennsylvania Genome Frontiers Institute, an Alex's Lemonade Stand Foundation Innovation Award (A.T.L.), young investigator awards from Alex's Lemonade Stand Foundation (S.Z., A.C.W.) and the CureSearch for Children's Cancer Foundation (S.Z.), grant from the German Cancer Aid 110801 (N.W.-L.), St Baldrick's Foundation Fellow award (A.C.W.), George L. Ohrstrom Jr foundation (A.C.W.), Wellcome Trust Senior Investigator Award Ref:100210/Z/12/Z (N.R.) and NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and the ICR (N.R.), Fondazione Italiana per la Lotta al Neuroblastoma (M.C.), Associazione Oncologia Pediatrica e Neuroblastoma (M.C.), and Associazione Italiana per la Ricerca sul Cancro (M.C.). We gratefully acknowledge the Children's Oncology Group (COG) for providing the specimens and clinical data from neuroblastoma patients and thank patients and families for participating in the COG, the UK-based Factors Associated with Childhood Tumors (FACT), and Italian cooperative group studies. We thank A. Renwick who performed the Taqman analyses and A. Zachariou for recruiting participants to the FACT study. We thank G. Blobel for scientific advice and discussion, and generously providing equipment and reagents for ChIP experiments, N. Saeki and H. Sasaki for providing the *LMO1* cDNA clone, and Y. Nakatan for providing the lentiviral vector pOZ-FHN.

Author Contributions J.M.M. and A.T.L. conceived the study, guided interpretation of results and guided preparation of the manuscript. D.A.O. and A.C.W. performed and/or oversaw most of the experiments, computational analyses and data interpretation. I.C., R.S., C.W., L.S.H., S.Z., N.W.-L., A.D.D., B.J.A., L.A., L.T., K.B. and R.A.Y. performed the genomic and epigenetic experiments and data analysis including DNA sequencing and ChIP sequencing. L.D.M., S.J.D. and H.H. performed the fine mapping and association testing. J.S.W. and J.K. performed the tumour RNA sequencing. N.R. and M.C. performed the validation genotyping and association testing. M.C. and A.I. replicated the SNP association in the Italian cohort. D.A.O. and A.C.W. drafted the manuscript, while A.T.L. and J.M.M. and other authors edited the manuscript.

Author Information ChIP-seq data sets are available under Gene Expression Omnibus (GEO) super series GSE65664, and relevant accession numbers are shown in Supplementary Table 2. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.M. (maris@chop.edu).

METHODS

No statistical methods were used to predetermine sample size.

Genotype imputation and association testing. A primary European-American cohort of 2,101 cases and 4,202 matched controls were assayed with Illumina HumanHap550 v1, Illumina HumanHap550 v3, and Illumina Human610 SNP arrays as previously described[6]. Genotypes were phased using SHAPEIT v2.r790 and data from 1000 Genomes phase 1 version 3. Subsequently, imputation was performed using IMPUTE2 v2.3.1 for all SNPs and indel variants annotated in the 1000 Genomes phase 1 version 3. Testing for association with neuroblastoma under an additive genetic effect model was performed using the frequentist likelihood score method implemented in SNPTEST v2.4.1. Genotypes for a previously described African-American cohort of 365 cases and 2,491 controls[20] were imputed and tested for neuroblastoma association using the same analytic pipeline. Statistical adjustment for gender was performed in both cohorts. For population stratification adjustment, the first 20 multidimensional scaling (MDS) components were included as covariates in the European-American cohort, while a measure of African admixture as estimated by the ADMIXTURE software program was used in the African-American cohort. Manhattan plots of SNP position and statistical significance were generated using LocusZoom software. Linkage plots were generated by Haploview software based on HapMap CEU individuals (version 3, release 2) using default settings. All research subjects or their guardians provided informed consent for research, and all institutions involved in this research had regulatory approval for human subjects research.

Prioritization of candidate causal variants. All SNPs and indels reported in the 1000 Genomes phase 1 version 3 data were considered as candidate causal variants and were ranked based on a combination of (1) neuroblastoma association in the primary European-American cohort, (2) evolutionary conservation, (3) DNase I hypersensitivity, and (4) transcription factor binding motif matching. Neuroblastoma association in European-Americans was evaluated as described above. Conservation scores were computed as the average of the phastCons46way-Placental UCSC conservation track score for all bases from the −10 position to the +10 position surrounding each candidate variant. A DNase I hypersensitivity score was calculated by counting the number of sequencing tags from the −100 position to the +100 position around each candidate variant in ENCODE data for the neuroblastoma cell line, SK-N-SH. Position weight matrices representing transcription factor binding motifs were obtained from the JASPAR database, and candidate binding sites were identified by scanning the hg19 human reference genome using the MATCH-TM algorithm with a matrix similarity score (mSS) threshold of 0.90.

Neuroblastoma association replication and meta-analysis for rs2168101. We replicated the association of rs2168101 with neuroblastoma by direct genotyping of rs2168101 in independent Italian (cases = 420, controls = 751) and UK cohorts (cases = 369, controls = 1,109). Meta-analysis across the European-American, African-American, Italian and UK cohorts was performed using the inverse variance method provided in the METAL software program. Beta values (log-odds) and standard errors generated by SNPTEST, as described above, were used as input.

Survival analysis. We compared both overall survival and event-free survival over a 10-year follow-up period between G/G versus G/T and T/T rs2168101 genotypes in a case–case comparison between neuroblastoma patients from the European-American cohort. Because rs2168101 genotypes were imputed in this cohort, the most-probable genotype predicted by IMPUTE2 was used for each patient. In the event of insufficient follow-up, all data was right censored. Cox proportional hazard modelling was performed using 20 MDS components to account for population stratification, in addition to *MYCN* amplification status, as covariates. All statistical analysis and generation of Kaplan–Meier plots was performed in R using the CRAN repository package, 'survival'.

Total and allele-specific expression analysis. Total and allele-specific RNA expression analysis was performed based on poly-A-enriched RNA-sequencing data from 127 primary neuroblastoma tumours sequenced through the TARGET project. RNA-seq reads were aligned to the hg19 human reference genome using the STAR aligner (v2.4.0b). Aligned reads were assigned to RefSeq genes using HTSeq (v0.6.1) and normalized to RPKM for total gene expression measurements. DNA genotypes for rs2168101 were obtained either through matched whole-genome sequencing ($n = 69$) or targeted genotyping assays ($n = 58$ additional tumours). DNA genotypes for rs3750952 were obtained through either matched whole-genome or whole-exome sequencing.

Allele-specific RNA expression analysis was performed from a subset of 45 primary neuroblastoma tumours (out of 127) with the necessary synonymous exonic SNP genotypes (rs3750952 = C/G) to enable measurement of allelic expression by mRNA-seq. As a readout for allelic imbalance of rs3750952, we computed allelic fractions as min(C, G)/(C + G), since phasing between rs3750952 and rs2168101 alleles in each tumour was unknown. Statistical comparison between the two groups was performed by two-sided Welch's *t*-test, comparing 12 tumours heterozygous for rs2168101 (G/T) to the remaining 33 tumours that were homozygous for rs2168101 (G/G) as controls. DNA genotyping for rs2168101 was performed by whole-genome sequencing or a directed genotyping assay, whereas DNA genotyping for rs3750952 was determined from TARGET whole-exome or whole-genome sequencing. Where possible, integrity of sample matching was verified by measurement of genome-wide genotype concordance. All genotypes are reported with respect to the minus strand of the human reference genome, hg19.

To measure allele-specific expression directly at the intronic SNP we first purified the nuclear RNA fraction using the Cytoplasmic and Nuclear RNA purification Kit (Norgen Biotek, 21000) from four neuroblastoma cell lines (SNP rs2168101: SHSY5Y = G/G; NLF = G/T; NGP = G/G; NB1643 = T/T). Ion AmpliSeq Designer v3.4.3 (Life Technologies White Glove service) was used to design amplicons targeting the intronic SNP rs2168101 and three additional exonic SNPs in linkage disequilibrium. Custom AmpliSeq libraries were prepared in triplicate for each cell line, indexed, pooled and sequenced using an Ion 318 Chip on a Personal Genome Machine (Life Technologies). Reads were aligned to the hg19 reference genome and a synthetic genome showing the alternate allele at SNP rs2168101 at hg19 chr11:8255408 to account for any alignment bias. High-quality mapped reads containing the reference G allele or alternative T allele were counted and tested for significant deviation from 50:50 expression using a two-sided one-sample *t*-test (null hypothesis that allele fraction = 0.50) across three experimental replicates. Primer pair sequences: rs1042359 forward: 5′-GTGTGGGAGACAAAUTCTTCCUGA-3′, reverse: 5′-GCCGGGCGUTACTGAACUT-3′; rs3750952 forward: 5′-CGCAAGAUCAAGGACCGCTAUC-3′, reverse: 5′-GATGAGGTUGGCCTTGGTGUA-3′; rs2168101 forward: 5′-CCUTTCCUGAAGGAGCGCAAA-3′, reverse: 5′-CACTTTCCATUAAGGAGATAGCAUCCC-3′; rs204929 forward: 5′-CAAUCTAGGTUAAGAGCCGGACAAG-3′, reverse: 5′-GTGUCCAGCCGCAGCUA-3′.

Reporter assays. Primers were designed to clone a 553-bp genomic region (hg19, chr11:8255155–8255707) surrounding the candidate SNP rs2168101 at the GATA transcription factor binding site from neuroblastoma cell lines SKNSH (G/G) and matching site of BE2C (T/−). The cloned region did not contain other statistically significant SNPs at the *LMO1* locus. The primers were designed to introduce sequences for restriction sites 5′-XhoI and 3′-BglII, which are present in the MCS of pGL4.26[luc2/minP/Hygro] (Promega, E8441). XhoI/BglII restriction enzyme digested fragments were sequence verified, gel purified, ligated into pGL4.26[luc2/minP/Hygro], transformed into One Shot TOP 10 chemically competent cells (Life Technologies, C4040-10) and grown on LB plates containing 50 μg ml^{-1} ampicillin overnight at 37 °C. Colonies positive for the vector containing the insert were grown in 50 ml LB broth containing 50 μg ml^{-1} ampicillin and plasmids were purified using a Qiagen Plasmid Midi Prep Kit (Qiagen, 12143). Transfection into HEK293 cells which were approximately 50% confluent was accomplished using Fugene 6 Transfection reagent (Promega E2691) at a 3 μl:1 μg fugene:DNA ratio. Cells underwent selection in 150 μg ml^{-1} Hygromycin B (Mediatech, 30-240-CR) and individual colonies were picked and grown, and genotypes of constructs were confirmed by fragment size and Sanger sequencing. Subsequently, HEK293 + 553 bp insert cells and HEK294 + vector only cells were grown in 96-well optical plates. On day 2, the cells were transiently fugene transfected with the Renilla expression control vector pGL4.74[hRLuc/TK] (Promega, E6921) at a 1:500 dilution with respect to the luciferase vector. Luciferase assays were carried out 48 h after Renilla transfection using Dual Luciferase Reporter Assay System (Promega, E1910) with read-outs performed on a Dual Injector System for GloMax-Multi Detection System (Promega, E7081). Luciferase expression was normalized to Renilla expression. All reporter assays were performed in quintuplicate (five technical replicates each) across the experimental conditions: (1) HEK293T, (2) HEK293T with empty vector, (3)–(6) four independent clones of HEK293T with T allele construct, and (7)–(10) four independent clones of HEK293T with G allele construct. Results were averaged across technical replicates, normalized to empty vector, and reporter activities for T allele versus G allele clones (four biological replicates each) were analysed by two-sided Welch's *t*-test.

Construct risk allele (G):
GTAGGGGTTGGAGTTCAGCCTGTTTCCCCTCCAATGTTGTTCCCCC
CACATCCTGAGACTTAGGGGTGACCCTGGGTTGAGTGGACTGGTTTA
TTCTGCTGGGCCCAGCGCATGCATCTGAGTGTGTGCCCAGGCGTGCG
TGTCGGCGCAAACATCATCCATTGTGAAATATCAGTGTTTTCATGGGT
GAGTAGTAATTACTGGGTAATGCTTTAAAACCTTTCCTGAAGGAGCGC
AAAGCCATTTTTTTTCTAAAGTCAGGAGTACATTAAAAGGATTACCATG
TAGATTTGATTTTTA**GATA**ACACTAAAATGGATCCCAAATGGACTTCA
GCAAAGGGATGCTATCTCCTTAATGGAAAGTGCATGGCCCGAGGCTC
AGGTCCCAGAGCCAGGCTGGGGAAGGAGGGAGGGAAGAGGTGTCTG
CAGGGGGGCAGGCTGGCAGATTGGGTGGGGGCTAGGTGGGAATGGG
GAAGGCAGAGCAGGAGGGAGGGCCTGGACCCTGTGGGGAGCTTATC
CCTCCATCTGGGGAGCAGGAGACTACAGAGCCCCT.

Construct protective allele (T):
GTAGGGGTTGGAGTTCAGCCTGTTTCCCCTCCAATGTTGTTCCCCC
CACATCCTGAGACTTAGGGGTGACCCTGGGTTGAGTGGACTGGTTTA
TTCTGCTGGGCCCAGCGCATGCATCTGAGTGTGTGCCCAGGCGTGCG
TGTCGGCGCAAACATCATCCATTGTGAAATATCAGTGTTTTCATGGGT
GAGTAGTAATTACTGGGTAATGCTTTAAAACCTTTCCTGAAGGAGCGC
AAAGCCATTTTTTTCTAAAGTCAGGAGTACATTAAAAGGATTACCATG
TAGATTTGATTTTTA**TATA**ACACTAAAATGGATCCCAAATGGACTTCAG
CAAAGGGATGCTATCTCCTTAATGGAAAGTGCATGGCCCGAGGCTCA
GGTCCCAGAGCCAGGCTGGGGAAGGAGGGAGGGAAGAGGTGTCTGC
AGGGGGGCAGGCTGGCAGATTGGGTGGGGGGCTAGGTGGGAATGGG
GAAGGCAGAGCAGGAGGGAGGGCCTGGACCCTGTGGGGAGCTTATC
CCTCCATCTGGGGAGCAGGAGACTACAGAGCCCCT.

Cell culture and protein lysates. Jurkat T-ALL and neuroblastoma cell lines were sourced from the American Type Tissue Culture Collection, and kept in growth medium of RPMI+10% heat-inactivated FCS with 1% penicillin–streptomycin, as previously described[13]. Cells were lysed for protein, with subsequent protein quantified by spectrophotometry, as previously described[21]. Protein was resolved on 8–14% Tris-Bis gels, transferred to PVDF membranes, blocked and subjected to primary and secondary antibodies, as previously described[21]. Primary antibodies were anti-GATA3 (Pierce Biotechnology, 1:1,000), anti-LMO1 (Bethyl Laboratories, 1:1,000) and α-tubulin (Cell Signaling Technologies, 1:1,000). Blots were developed with secondary horseradish peroxidase (HRP)-conjugated antibodies (Cell Signaling Technologies, 1:5,000) and Protein-plus Dura ECL Reagent (Thermo-Fisher Scientific). All cell lines are genotyped semiannually to assure identity and also tested routinely for mycoplasma contamination.

Genome-wide occupancy analysis. ChIP coupled with massively parallel DNA sequencing (ChIP-seq) was performed as previously described[22,23]. The following antibodies were used for ChIP: anti-H3K27ac (Abcam, ab4729) and anti-GATA3 (Santa Cruz, sc-22206X). For each ChIP, 10 μg of antibody was added to 3 ml of sonicated nuclear extract. Illumina sequencing, library construction and ChIP-seq analysis methods were previously described[23].

ChIP-seq processing. Reads were aligned to build hg19 of the human genome using bowtie with parameters -k 2 -m 2 -e 70 -best and -l set to the read length[24]. For visualization in the UCSC genome browser in Fig. 3a, c and Extended Data Fig. 6 (ref. 25), WIG files were created from aligned ChIP-seq read positions using MACS 1.4.2 with parameters -w -S -space = 50 -nomodel -shiftsize = 200 to artificially extend reads to be 200 bp and to calculate their density in 50-bp bins[26]. Read counts in 50-bp bins were then normalized to the millions of mapped reads, giving reads per million values.

ChIP-seq allele specificity analysis. To determine preferential ChIP-seq coverage of one allele, which implies preferential binding of protein to one allele versus another, we counted the reads at rs2168101 using samtools mpileup[27]. By using the aligned reads described above, this gave us a count of reads with a given base at this position. The fraction of reads with the risk allele versus the protective allele is reported in Fig. 3b. Statistical tests for preferential allelic binding were performed by two-sided binomial test.

Enriched regions. Regions enriched in ChIP-Seq signal were identified twice using MACS with corresponding control and parameters -keep-dup = all and -p 1e-9 or -keep-dup = 1 and -p 1e-9. Super-enhancers in SHSY5Y and KELLY were identified using ROSE (https://bitbucket.org/young_computation/rose)[18,28] with modifications based on ref. 14. In brief, peaks of H3K27ac were identified using MACS as described above and their union was used as constituent enhancers. These peaks were stitched computationally if they were within 12,500 bp of each other, although peaks fully contained within ±2,000 bp from a RefSeq promoter were excluded from stitching. These stitched enhancers were ranked by their H3K27ac signal (length × density) with input signal in the corresponding region subtracted. Super-enhancers were separated from typical enhancers by geometrically determining the point at which the line $y = x$ is tangent to the curve of stitched enhancer rank versus stitched enhancer signal. Those stitched enhancers above this point are considered super-enhancers.

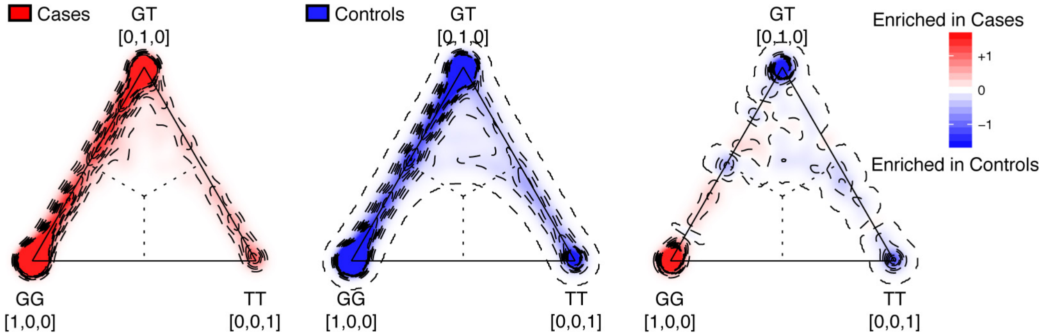
To account for the known focal amplification of the *MYCN* locus in KELLY, BE2, BE2C and NGP neuroblastoma cells, which contain enhancers, we modified our pipeline slightly. Because MACS is insensitive for the identification of peaks in focally amplified DNA, we identified peaks of H3K27ac versus input using MACS2 callpeak (https://pypi.python.org/pypi/MACS2) with parameters -broad -keep-dup = 1 -p 1e-9 and -broad -keep-dup = all -p 1e-9. The union of these MACS2 calls was used as constituent enhancers for ROSE with the remaining parameters as described above. For Fig. 3d, most of the curve represents the analysis performed using MACS-identified constituents; the rank and signal of the MYCN-associated enhancer comes from this MACS2-identified set of constituents to remain consistent with the conclusions and methods as previously described[14]. The curve output from the MACS-identified enhancers was vertically compressed and a point representing the signal of the MYCN-associated super-enhancer from the MACS2-identified enhancers was added in Illustrator. Super-enhancers were assigned to the single expressed RefSeq transcript whose transcription start site was nearest the centre of the stitched region. Expressed genes were in the top 2/3 of RefSeq transcripts ranked by their promoter (transcription start site ±500 bp) H3K27ac signal determined by bamToGFF (https://github.com/BradnerLab/pipeline) with parameters -e 200 -m 1 -r -d.

Clone cell generation. *LMO1* cDNA was amplified from pcDNA3-LMO1 and subcloned into the XhoI and NotI site of the lentiviral vector pOZ-FHN. Lentivirus expressing FH-LMO1 was propagated in HEK293T cells by cotransfection with psPAX2 and pMD2.G plasmids (adgene) using FUGENE 6 (Roche) by standard methodologies[29]. Viral supernatant was recovered and KELLY cells were infected with lentivirus expressing FH-LMO1 or empty vector alone, as previously described[13]. Cells were sorted for expression of the IL2R, and positive expression was used to establish single cell clones. Expression of FH-LMO1 was assessed by western blotting as above to confirm overexpression.

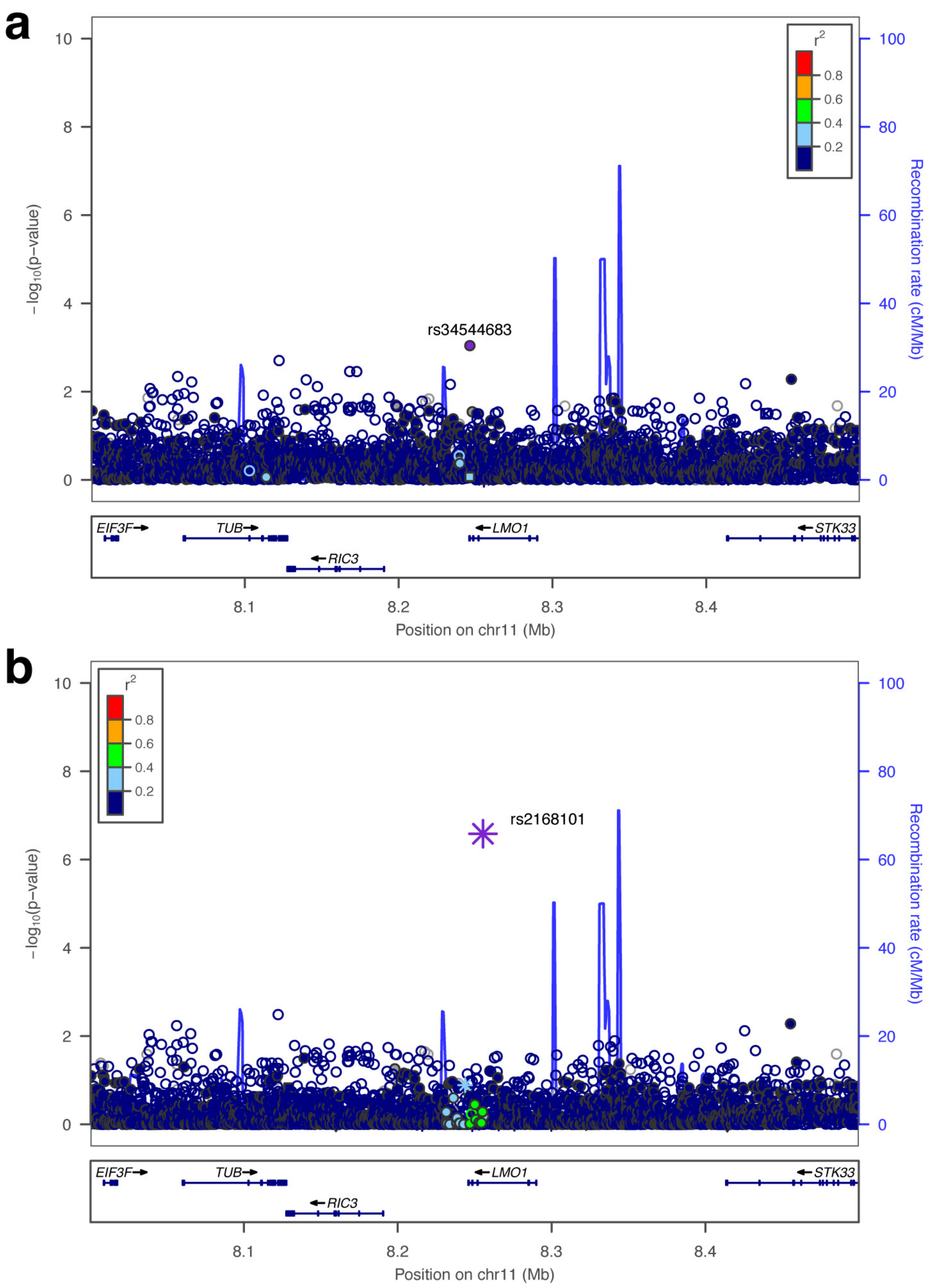
siRNA and growth assays. SHSY5Y, KELLY and KELLY clone cells were reverse transfected with 100 nM concentrations of either non-targeted (control siRNA-1) or GATA3-targeted siRNA-1 or -2 (Ambion) for 6 h with lipofectamine 2000 (1:1,000) in Optimem I before being replated into growth assays in normal RPMI growth media. Cells (2×10^5) were replated in triplicate for counting at 24, 48 and 72 h post-transfection by manual hemocytometry. Cells (5×10^5) were replated for protein lysates at the same time points. All experiments were repeated in triplicate, with a technical replicate number of 9 for all cell growth assays as described[30]. Statistical tests were performed by two-sided Welch's *t*-test.

Data access. GWAS and sequencing data used for this analysis are available in dbGaP under accession phs000124 and phs000467. The tumour genomics data are also available through the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) data matrix portal (http://target.nci.nih.gov/dataMatrix/TARGET_DataMatrix.html). Data generated through the ENCODE project including DNase I hypersensitivity sequencing and ChIP-sequencing data were obtained from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/. Aligned sequencing read (bam) files were used as provided from the FTP site. The mammalian evolutionary conservation track representing 46 mammalian species (phastCons46wayPlacental) was obtained from the UCSC Table Browser http://genome.ucsc.edu/cgi-bin/hgTables?command=start. JASPAR-annotated transcription factor binding site position frequency matrices were obtained from http://jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/pfm/nonredundant/pfm_all.txt. New ChIP-seq data sets generated in this study are available under super series GSE65664.

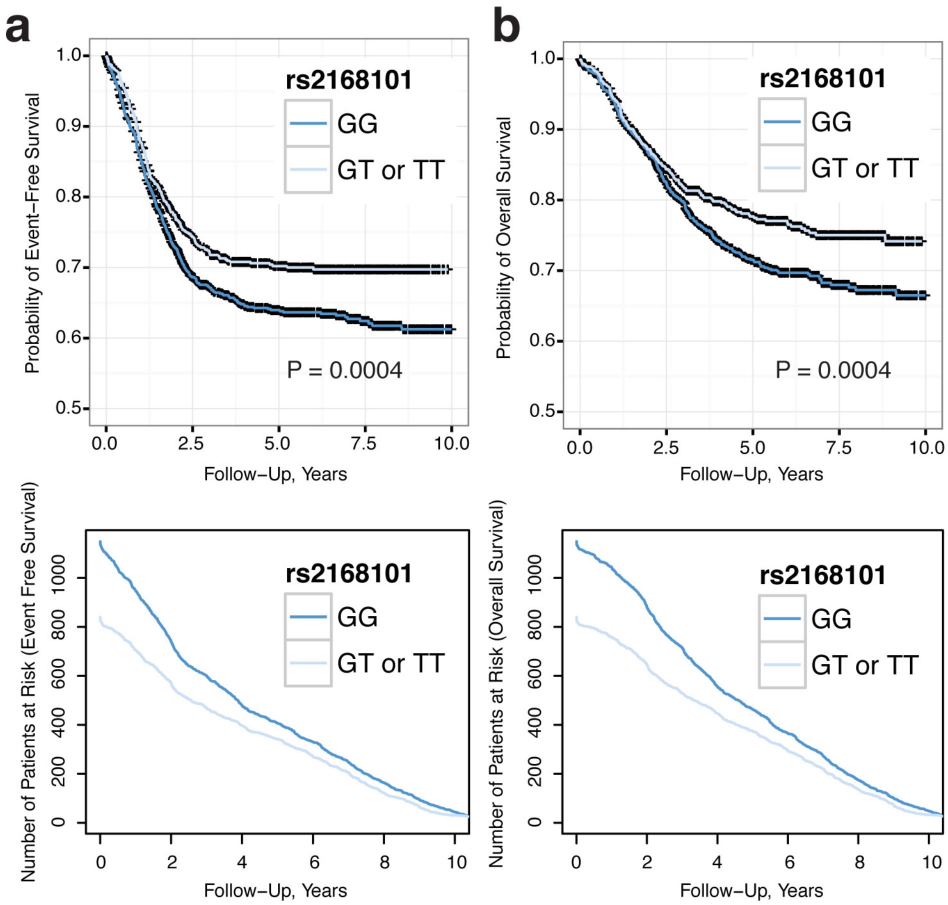
20. Latorre, V. *et al.* Replication of neuroblastoma SNP association at the *BARD1* locus in African-Americans. *Cancer Epidemiol. Biomarkers Prev.* **21,** 658–663 (2012).
21. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346,** 1373–1377 (2014).
22. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols* **1,** 729–748 (2006).
23. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533 (2008).
24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
25. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
28. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153,** 320–334 (2013).
29. Nakatani, Y. & Ogryzko, V. Immunoaffinity purification of mammalian protein complexes. *Methods Enzymol.* **370,** 430–444 (2003).
30. Durbin, A. D. *et al.* JNK1 determines the oncogenic or tumor-suppressive activity of the integrin-linked kinase in human rhabdomyosarcoma. *J. Clin. Invest.* **119,** 1558–1570 (2009).



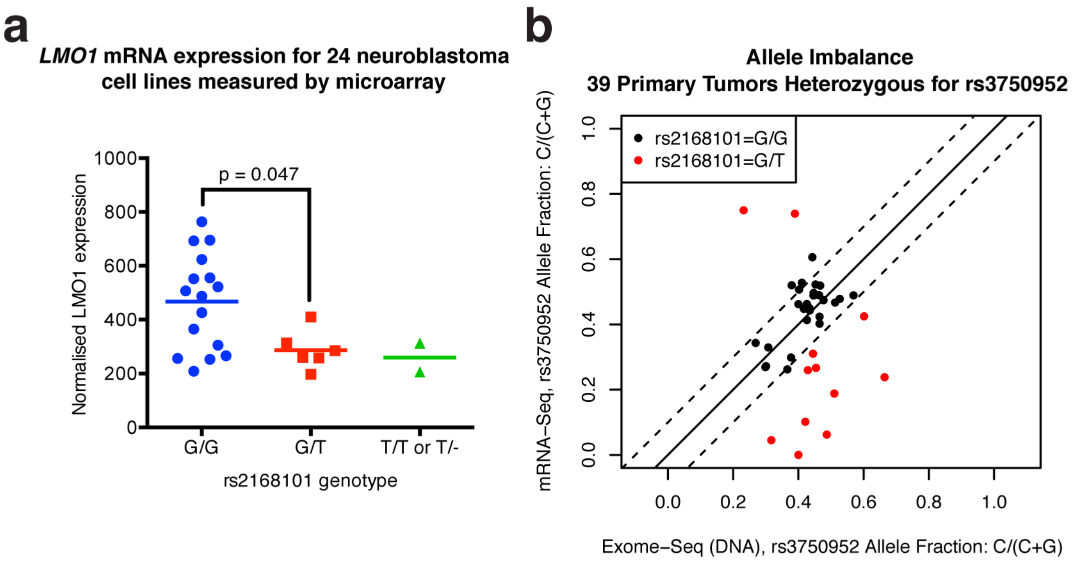
Extended Data Figure 1 | The imputed SNP, rs2168101, is associated with neuroblastoma, and the risk 'G' allele is enriched in neuroblastoma cases. Ternary density plots of genotype probability vectors [P(G/G), P(G/T), P(T/T)] output from IMPUTE2 for rs2168101 in the European-American cohort. Vertices represent 'perfect' confidence calls in which P(genotype) = 1; dotted lines represent decision boundaries for genotype calling based on most probable genotype. All plots were normalized by the total number of individuals studied and subjected to 2D Gaussian kernel smoothing. Left, 2,101 cases (red); centre, 4,202 controls (blue); right, difference between cases and controls highlights enrichment of G/G genotype (homozygous risk) in cases and of G/T and T/T genotypes in controls. Validation efforts using PCR-based genotyping in 146 out of 2,101 European-American cases confirmed an 86% concordance with imputation based on most probable genotypes (Supplementary Table 1).



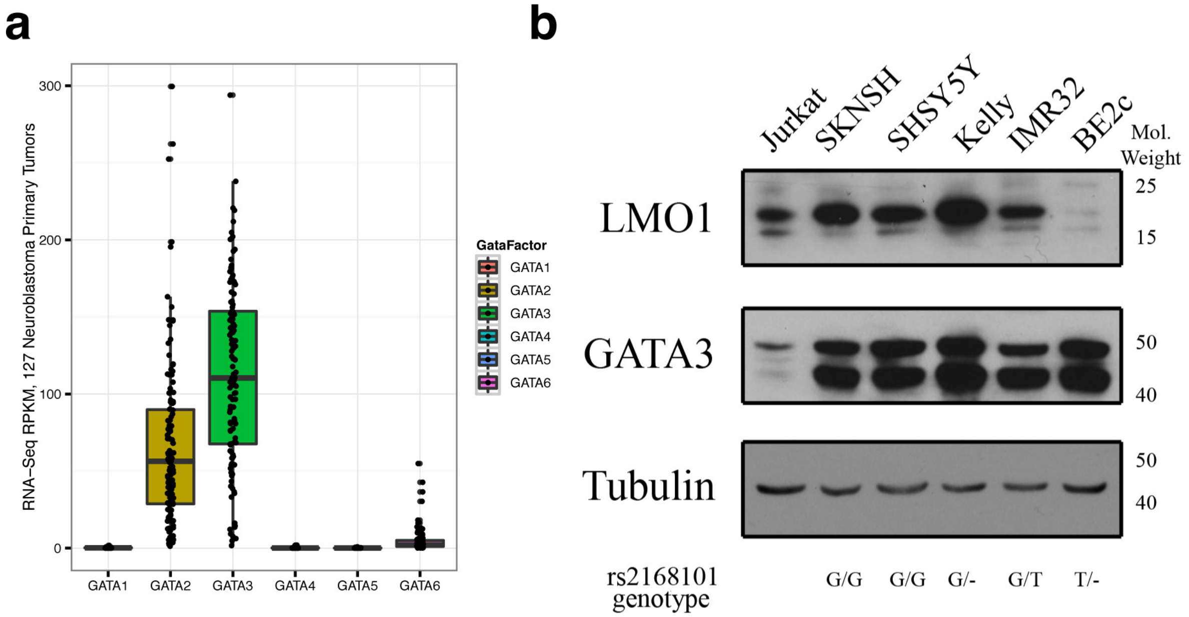
Extended Data Figure 2 | Conditional analysis reveals a single neuroblastoma association signal at the *LMO1* locus and that rs2168101 is the most associated variant. a, Imputation-based neuroblastoma association study conditional on rs2168101. No variants remain significant after conditioning on rs2168101 (most significant variant: rs34544683, nominal $P = 9.0 \times 10^{-4}$, Bonferroni $P = 1$). **b**, Reciprocal analysis conditioned on each of 27 SNPs with a nominal $P < 1 \times 10^{-5}$. For rs2168101, the maximum (least significant) P value across all non-rs2168101 conditional tests is shown, in order to illustrate the extent to which the signal at rs2168101 can be accounted for by other variants (a similar maximum P value statistic is plotted for other variants). Notably, rs2168101 remained significant (worst-case nominal $P = 2.6 \times 10^{-7}$, Bonferroni $P = 0.002$) across all tests. These results are consistent with a single underlying signal at the *LMO1* locus, and re-affirm that rs2168101 is the single best causal SNP candidate because its association with neuroblastoma cannot be accounted for by other single variants.



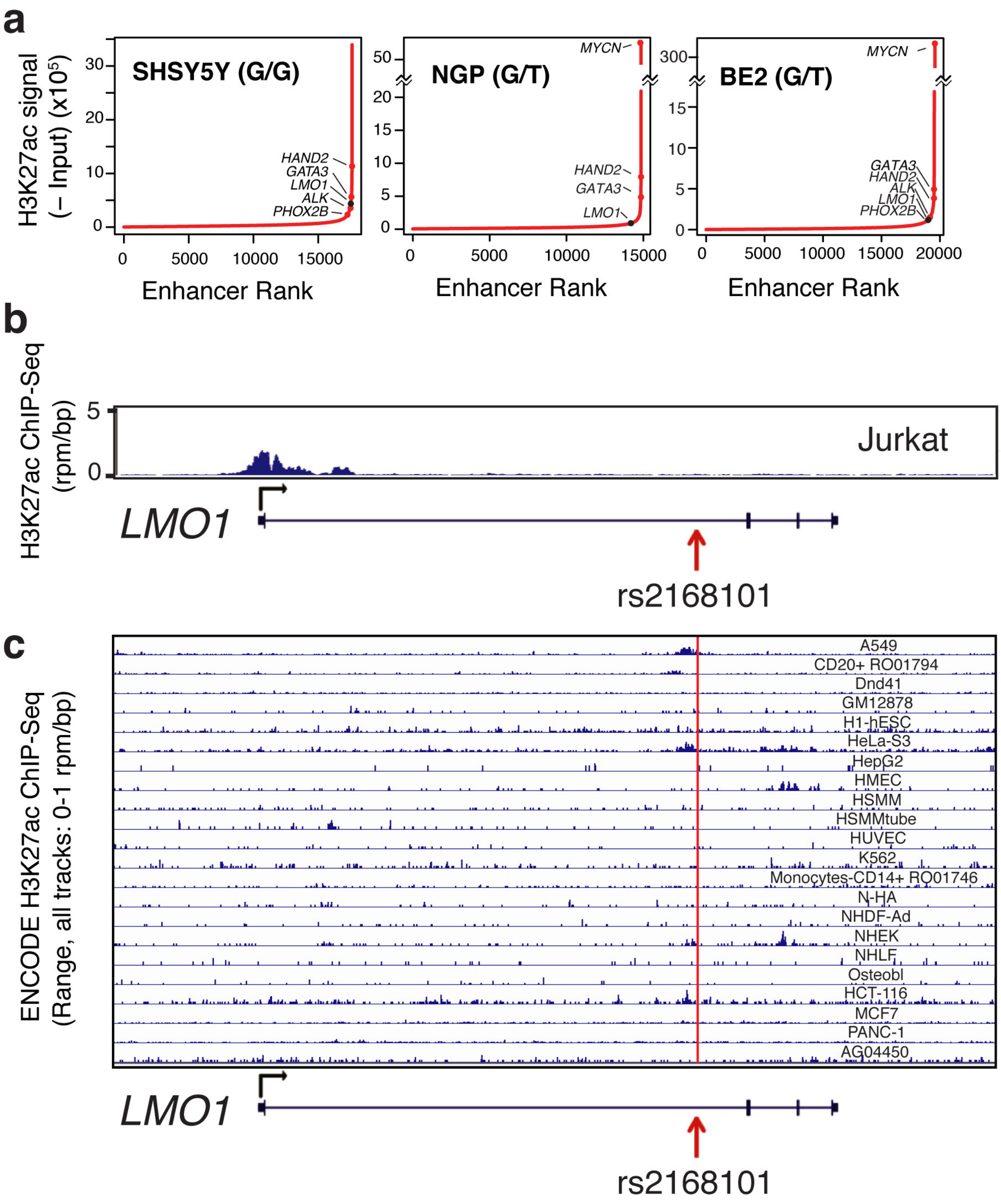
Extended Data Figure 3 | The risk G allele of rs2168101 is associated with decreased event-free and overall survival in the European-American discovery cohort. Because genotypes for rs2168101 are imputed within the European-American discovery cohort, the most likely genotype for each neuroblastoma case was called based on the maximum of P(G/G), P(G/T) and P(T/T) from IMPUTE2. *P* values reflect Cox proportional hazards regressions adjusted for *MYCN* amplification status and the first 20 MDS components to adjust for population stratification. **a**, Kaplan–Meier plot for event-free survival. Neuroblastoma cases with rs2168101 = G/G versus rs2168101 = G/T or T/T showed significantly worse event-free survival ($P = 0.0004$). **b**, Kaplan–Meier plot for overall survival. Neuroblastoma cases with rs2168101 = G/G versus rs2168101 = G/T or T/T showed significantly worse overall survival ($P = 0.0004$). Censored data points are shown as black crosses. Number of at risk patients at every time point for both event-free survival and overall survival are plotted below each respective Kaplan–Meier plot.



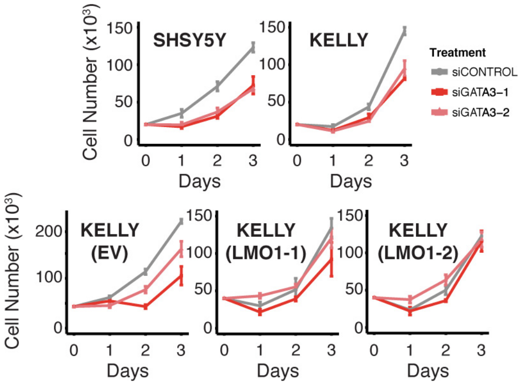
Extended Data Figure 4 | rs2168101 genotype is associated with total and allele-specific *LMO1* expression in neuroblastoma cell lines and primary tumours, and allele-specific expression differences are not driven by somatic DNA copy number alterations. a, Neuroblastoma cell line *LMO1* mRNA expression as quantified by Affymetrix U95Av2 oligonucleotide arrays and normalized as described[11] was significantly higher in cell lines harbouring homozygous risk alleles (G/G) compared to heterozygous alleles (G/T) ($P = 0.047$, Mann–Whitney two-tailed). **b**, Allele-specific expression measured by RNA-seq from primary neuroblastoma tumours. Since rs2168101 is an intronic SNP that is spliced out in mRNA, the synonymous exonic SNP rs3750952 was used as a surrogate for measuring allele-specific expression in 39 primary tumours which are heterozygous for rs3750952 (C/G genotype). The DNA allelic fraction for rs3750952 determined by whole-exome sequencing is plotted on the *x* axis, whereas the RNA allele fraction for rs3750952 determined by mRNA-seq is plotted on the *y* axis. The solid line indicates where DNA and RNA allele fractions are equal and dotted lines indicate the boundary where DNA and RNA allele fractions are within 10% of each other. Tumours that are heterozygous for rs2168101 (G/T genotype, red dots) exhibit greater RNA allelic imbalance ($P = 5.3 \times 10^{-5}$) than homozygous controls (rs2168101 = G/G genotype, black dots). By contrast, DNA allelic imbalance is no different between G/T versus G/G tumours ($P = 0.79$), indicating that a *cis*-acting regulatory mechanism, rather than somatic DNA alterations, drives *LMO1* allelic expression differences.



Extended Data Figure 5 | Expression of LMO1 and GATA-family transcription factors in neuroblastoma primary tumours and cell lines. **a**, RPKM expression measurements from mRNA-seq are summarized via boxplots for 127 primary neuroblastoma tumours for paralogues *GATA1* through *GATA6*. Both *GATA2* (median RPKM: 56) and *GATA3* (median RPKM: 110) are more highly expressed by 1–4 orders of magnitude on average compared to other members of the GATA family in neuroblastoma. **b**, Neuroblastoma cell lines were lysed for protein and resolved by SDS–PAGE as previously described[21]. Jurkat T-ALL cells are shown as a positive control for LMO1 and GATA3 expression. Data are representative of at least three independent blots. The rs2168101 genotype is shown below individual cell lines.



Extended Data Figure 6 | The *LMO1* super-enhancer is observed in neuroblastoma cell lines containing the G allele of rs2168101 and is highly tissue-specific. a, H3K27ac signal across all enhancers in SHSY5Y (*MYCN* not amplified; rs2168101 = G/G), BE2 (*MYCN*-amplified; rs2168101 = G/T) and NGP (*MYCN*-amplified; rs2168101 = G/T) is shown. Enhancers are ranked by their signal of H3K27ac minus input signal and are geometrically divided into two populations (see Methods). Super-enhancers are those at the high end of the population and are associated with key genes in neuroblastoma, highlighted on the curve. *LMO1*-associated super-enhancers were identified in BE2, KELLY and SHSY5Y cells, which all contain the G allele of rs2168101, but not in BE2C cells in which the G allele is absent. **b**, H3K27ac ChIP-seq in the Jurkat cell line. **c**, All ENCODE non-neuroblastoma cell lines with H3K27ac ChIP-seq profiling. All non-neuroblastoma cell lines considered showed little to no evidence for an active enhancer element within the first intron of the *LMO1* gene locus, consistent with a tissue and disease-specific enhancer overlying the neuroblastoma causal SNP rs2168101.



Extended Data Figure 7 | Depletion of GATA3 results in suppression of cell growth that is rescued by forced LMO1 expression in neuroblastoma. Neuroblastoma cells SHSY5Y, KELLY, KELLY overexpressing control vector (EV) and KELLY with forced LMO1 overexpression (LMO1-1 and LMO1-2) were treated with non-targeted (siControl) or GATA3-targeting (siGATA3-1, siGATA3-2) siRNAs and cells were counted at 24, 48 and 72 h after transfection. Rescue of suppressed cell growth after GATA3 depletion by forced LMO1 expression in LMO1-1 and LMO1-2 after 72 h is shown on the bottom. Growth curves over the time of 72 h are shown (to accompany Fig. 3f). Error bars denote ±s.e.m., $n = 9$ technical replicates.

Extended Data Table 1 | Germline variants from 1000 Genomes Project with $P < 1 \times 10^{-5}$ association with neuroblastoma susceptibility from imputation-based SNPTEST analysis of European-American cohort

Variant ID (rsID)	Chromosome	Position (hg19)	Alleles (Ref/Alt)*	Alt Allele Frequency Cases	Alt Allele Frequency Controls	P-Value†	Odds Ratio†
rs191871553	11	8222464	C/T	0.035 (n=2101)	0.054 (n=4202)	7.49E-06	0.64 (0.53-0.78)
rs11041809	11	8231605	A/G	0.498 (n=2101)	0.440 (n=4202)	1.13E-07	0.80 (0.74-0.87)
rs11041811	11	8231665	C/T	0.492 (n=2101)	0.434 (n=4202)	1.28E-07	0.80 (0.74-0.87)
rs11041812	11	8231684	C/T	0.492 (n=2101)	0.433 (n=4202)	1.22E-07	0.80 (0.74-0.87)
rs11041813	11	8235207	T/C	0.478 (n=2101)	0.420 (n=4202)	1.67E-07	0.81 (0.75-0.87)
rs10839999	11	8236083	G/A	0.480 (n=2101)	0.423 (n=4202)	5.06E-07	0.81 (0.75-0.88)
rs10769885	11	8236262	C/A	0.513 (n=2101)	0.453 (n=4202)	3.77E-08	0.80 (0.74-0.87)
rs4758049	11	8238428	A/C	0.511 (n=2101)	0.452 (n=4202)	7.48E-08	0.81 (0.74-0.87)
rs4758050	11	8238545	G/C	0.511 (n=2101)	0.452 (n=4202)	7.34E-08	0.81 (0.74-0.87)
rs4758051	11	8238639	G/A	0.510 (n=2101)	0.452 (n=4202)	1.22E-07	0.81 (0.75-0.87)
rs10840000	11	8240113	G/C	0.509 (n=2101)	0.450 (n=4202)	6.22E-08	0.80 (0.74-0.87)
rs7933766	11	8240464	G/A	0.511 (n=2101)	0.453 (n=4202)	2.09E-07	0.81 (0.75-0.88)
rs11041816	11	8243798	A/G	0.397 (n=2101)	0.456 (n=4202)	8.99E-10	0.77 (0.71-0.84)
rs4315061	11	8247020	T/C	0.425 (n=2101)	0.490 (n=4202)	1.25E-09	0.78 (0.72-0.84)
rs72474792	11	8247885	TATAAAA/T	0.524 (n=2101)	0.456 (n=4202)	2.04E-10	0.77 (0.71-0.84)
rs12797723	11	8247984	C/T	0.443 (n=2101)	0.514 (n=4202)	2.05E-10	0.77 (0.71-0.84)
rs2290451	11	8248440	C/G	0.295 (n=2101)	0.255 (n=4202)	8.20E-06	1.23 (1.12-1.34)
rs7952320	11	8250143	G/C	0.408 (n=2101)	0.480 (n=4202)	3.03E-11	1.31 (1.21-1.42)
rs4758317	11	8250811	C/A	0.514 (n=2101)	0.447 (n=4202)	5.76E-10	0.78 (0.72-0.84)
rs11041820	11	8251438	G/A	0.294 (n=2101)	0.253 (n=4202)	6.77E-06	1.23 (1.12-1.34)
rs3750952	11	8251921	G/C	0.408 (n=2101)	0.481 (n=4202)	1.89E-11	0.76 (0.70-0.83)
rs110419	11	8252853	A/G	0.441 (n=2101)	0.511 (n=4202)	3.16E-10	0.78 (0.72-0.84)
rs110420	11	8253049	T/C	0.441 (n=2101)	0.511 (n=4202)	3.36E-10	0.78 (0.72-0.84)
rs204928	11	8254433	A/G	0.444 (n=2101)	0.512 (n=4202)	9.85E-10	0.78 (0.72-0.85)
rs204926	11	8255106	G/A	0.440 (n=2101)	0.510 (n=4202)	1.97E-11	0.76 (0.70-0.82)
rs2168101	11	8255408	C/A	0.242 (n=2101)	0.313 (n=4202)	4.14E-16	0.67 (0.61-0.74)
rs7948497	11	8255855	C/G	0.479 (n=2101)	0.419 (n=4202)	4.05E-10	1.30 (1.20-1.41)

*Forward strand hg19, imputed genotypes from IMPUTE2, frequencies as reported by SNPTEST.

†SNPTEST, frequentist score test with additive model, adjusted for gender and top 20 MDS components.

Extended Data Table 2 | Clinical characteristics for patients in referenced sequencing data sets

Characteristic	Whole Exome Sequencing (Blood and Tumor) N=222[*]	Whole Genome Sequencing (Blood and Tumor) N = 136[*]	*LMO1* Targeted Sequencing (Blood) N = 183[*]	Whole Transcriptome mRNA Sequencing (Tumor) N = 127
Age				
< 18 mos	0 (0%)	32 (24%)	82 (45%)	8 (6%)
>= 18 mos	219 (100%)	103 (76%)	101 (55%)	119 (94%)
Not Available	3	1	0	0
INSS Stage[†]				
Stage 1	0 (0%)	0 (0%)	39 (21%)	0 (0%)
Stage 2A	0 (0%)	0 (0%)	13 (7%)	0 (0%)
Stage 2B	0 (0%)	1 (1%)	18 (10%)	0 (0%)
Stage 3	0 (0%)	6 (4%)	27 (15%)	6 (5%)
Stage 4	219 (100%)	105 (78%)	78 (43%)	121 (95%)
Stage 4S	0 (0%)	23 (17%)	8 (4%)	0 (0%)
Not Available	3	1	0	0
MYCN				
Not Amplified	143 (67%)	102 (76%)	151 (83%)	95 (75%)
Amplified	71 (33%)	32 (24%)	30 (17%)	31 (25%)
Not Available	8	2	2	1
Histology				
Favorable	4 (2%)	29 (23%)	95 (54%)	9 (8%)
Unfavorable	187 (98%)	96 (77%)	82 (46%)	107 (92%)
Not Available	31	11	6	11
DNA Index				
Hyperdiploid	117 (54%)	81 (61%)	121 (67%)	67 (53%)
Diploid	98 (46%)	52 (39%)	59 (33%)	59 (47%)
Not Available	7	3	3	1
Risk				
Low	0 (0%)	15 (11%)	64 (35%)	0 (0%)
Intermediate	0 (0%)	14 (10%)	49 (27%)	6 (5%)
High	219 (100%)	106 (79%)	69 (38%)	121 (95%)
Not Available	3	1	1	0

[*] There is an overlap of 59 neuroblastoma patients with both whole-exome and whole-genome sequencing. Patients with targeted sequencing are all unique and do not overlap with whole-exome or whole-genome cases, yielding 482 unique patients with exonic DNA sequencing of *LMO1*.

[†] International Neuroblastoma Staging System (INSS).

Extended Data Table 3 | Association of rs2168101 with clinical/biological co-variates

Clinical/Biological Co-variate	rs2168101 genotypes*			Association result	
	GG	GT	TT	P-Value†	Odds Ratio†
Stage‡ 4	530 (62%)	280 (33%)	49 (6%)	0.01198	0.81 (0.69-0.95)
Not Stage 4	611 (56%)	400 (37%)	74 (7%)		
MYCN Amplified	183 (55%)	115 (34%)	36 (11%)	0.00297	1.39 (1.12-1.73)
MYCN Non-Amplified	881 (59%)	525 (35%)	83 (6%)		
High-Risk	523 (63%)	263 (32%)	47 (6%)	0.00174	0.76 (0.65-0.90)
Not High-Risk	594 (56%)	398 (37%)	73 (7%)		
Unfavorable Histology	454 (61%)	237 (32%)	48 (6%)	0.14479	0.88 (0.73-1.05)
Favorable Histology	527 (57%)	336 (36%)	62 (7%)		
DNA Index Hyperdiploid	685 (59%)	412 (35%)	71 (6%)	0.32009	0.91 (0.76-1.09)
DNA Index Diploid	324 (57%)	198 (35%)	43 (8%)		
Age >= 18 mos	621 (61%)	346 (34%)	55 (5%)	0.01448	0.82 (0.69-0.96)
Age < 18 mos	529 (57%)	338 (36%)	68 (7%)		

*Reverse strand hg19, imputed genotypes from IMPUTE2, genotype frequencies as reported by SNPTEST.
†SNPTEST, frequentist score test with additive model, adjusted for gender and top 20 MDS components.
‡International Neuroblastoma Staging System.