Common genetic variation drives molecular heterogeneity in human iPSCs

Helena Kilpinen1†, Angela Goncalves1‡, Andreas Lehå1‡, Vackar Afzal1, Kaur Alasoo1, Sofie Ashford4, Sendu Bala1, Dalila Bensaddek1, Francesco Paolo Casale1, Oliver J. Cullery1, Petr Danecek2, Adam Faulconbridge1, Peter W. Harrison1, Annie Kathuria2, Davis McCarthy1,6, Shane A. McCarthy1, Ruta Meleckyte2, Yasin Memari1,2, Nathalie Moens2, Filipa Soares7, Alice Mann2, Ian Streeter1, Chukwuma A. Agu1, Alex Alderton6, Rachel Nelson2, Sarah Harper2, Minal Patel1, Alastair White1, Sharad R. Patel2, Laura Clarke1, Reena Halai2, Christopher M. Kirton2, Anja Kolb–Kokocinski2, Philip Beales8, Ewan Birney1, Davide Danovi5, Angus I. Lammad3, Willem H. Ouwehand1,6,9, Ludovic Vallier2,7, Fiona M. Watt5‡, Richard Durbin2§, Oliver Stegle1§* & Daniel J. Gaffney1§*

Technology utilizing human induced pluripotent stem cells (iPS cells) has enormous potential to provide improved cellular models of human disease. However, variable genetic and phenotypic characterization of many existing iPS cell lines limits their potential use for research and therapy. Here we describe the systematic generation, genotyping and phenotyping of 711 iPS cell lines derived from 301 healthy individuals by the Human Induced Pluripotent Stem Cells Initiative. Our study outlines the major sources of genetic and phenotypic variation in iPS cells and establishes their suitability as models of complex human traits and cancer. Through genome-wide profiling we find that 5–46% of the variation in different iPS cell phenotypes, including differentiation capacity and cellular morphology, arises from differences between individuals. Additionally, we assess the phenotypic consequences of genomic copy-number alterations that are repeatedly observed in iPS cells. In addition, we present a comprehensive map of common regulatory variants affecting the transcriptome of human pluripotent cells.

iPS cells are powerful model systems for human disease1. A major open question is whether iPS cells can be used to study the functions of genetic variants associated with complex traits and normal human phenotypic variation. Previous work has suggested that individual iPS cell lines are highly heterogeneous2–5, although some of these differences may arise as a result of the genetic background of the donor6–7. High variability could make iPS cells unsuitable cellular models for genetic variants with small effects. In addition, existing iPS cell lines frequently have limited genetic and phenotypic data of variable quality available, or are derived from individuals with severe genetic disorders, thereby limiting their utility for studying other phenotypes.

The Human Induced Pluripotent Stem Cells Initiative (HipSci, http://www.hipscsi.org/) was established to generate a large, high-quality, open-access reference panel of human iPS cell lines. A major focus of the initiative is the systematic derivation of iPS cells from hundreds of healthy volunteers using a standardized and well-defined experimental pipeline. The cell lines are extensivly characterized and available to the wider research community along with the accompanying genetic and phenotypic data. Here we report initial results from the characterization of the first 711 iPS cell lines derived from 301 healthy individuals. We provide a high-resolution map of recurrent copy-number aberrations in iPS cells, identify putative candidate genes under selection in these regions and assess the functional consequences of these changes. We show that common genetic variants produce readily detectable effects in iPS cells and provide a comprehensive map of regulatory variation in human iPS cells. We also demonstrate that differences between donors have pervasive effects at all phenotypic levels in iPS cells, from the epigenome, transcriptome and proteome to cell differentiation and morphology.

Sample collection and iPS cell derivation

Samples were collected from healthy, unrelated research volunteers via the NIHR Cambridge BioResource (Methods). We established 711 iPS cell lines from 301 donors (more than 1 cell line for 82% of donors, more than 2 cell lines for 50%), which were profiled using an initial set of ‘tier 1’ assays (Fig. 1a). These assays included array-based genotyping and gene expression profiling of the iPS cells and their fibroblast progenitors, as well as an assessment of the pluriptocity and differentiation properties of the iPS cells. Using immunohistochemistry followed by quantitative image analysis (hereafter, cellometrics), we measured protein expression of pluriptocity markers in 307 cell lines, differentiated 372 cell lines into neuroectoderm, mesoderm and endoderm8, and measured the expression of three lineage-specific markers in each germ layer (Fig. 1a and Extended Data Fig. 1). We then selected 1–2 cell lines (hereafter ‘selected lines’) per donor to minimize the number of genetic abnormalities and performed further phenotyping (hereafter, ‘tier 2’ assays) using RNA-sequencing (RNA-seq), DNA-methylation arrays, quantitative proteomics and imaging of cell morphology in 239, 27, 16 and 24 cell lines, respectively (Supplementary Table 1).

1 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. 2Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. 3Centre for Gene Regulation & Expression, School of Life Sciences, University of Dundee, Dundee, DD1 5EH, UK. 4Department of Haematology, University of Cambridge, Cambridge CB2 0SZ, UK. 5Centre for Stem Cells & Regenerative Medicine, King’s College London, Tower Wing, Guy’s Hospital, Great Maze Pond, London SE1 9RT, UK. 6Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK. 7Centre for Stem Cells & Regenerative Medicine, King’s College London, Tower Wing, Guy’s Hospital, Great Maze Pond, London SE1 9RT, UK. 8UCL Great Ormond Street Institute of Child Health, University College London, London WC1N 1EH, UK. 9Wellcome Trust and MRC Cambridge Stem Cell Institute and Biomedical Research Centre, Anne McLaren Laboratory, Department of Surgery, University of Cambridge, Cambridge CB2 0SZ, UK. *These authors contributed equally to this work. §These authors jointly supervised this work.
Pluripotency and genetic stability

Using tier 1 expression data, 84% of the generated cell lines were classified as pluripotent by PluriTest\(^9\) (score of over 20) and 97% had a pluripotency score of over 10, which yields almost identical sensitivity and specificity in the PluriTest training set (Fig. 1b). Most iPS cell lines with a pluripotency score less than 20 (69%) had been cultured in feeder-free Essential 8 medium (odds ratio 5.4, \(P < 8 \times 10^{-13}\)), Fisher's exact test), which reflects the fact that PluriTest was primarily trained using stem cell lines that were grown in feeder-dependent conditions (Extended Data Fig. 2). Using the imaging data generated by cellomics, we quantified the fraction of cells expressing each pluripotency marker individually and estimated that, on average, between 18% and 62% of cells in the iPS cell lines co-expressed all pluripotency marker individually and estimated that, on average, between 18% and 62% of cells in the iPS cell lines co-expressed all pluripotency markers. Brachury is encoded by \(\text{CAP2}\), and nestin by \(\text{FATS}\), human dermal fibroblasts; \(\text{DEN}\), differentiated endoderm; \(\text{DM}\), differentiated mesoderm; \(\text{dmEC}\), differentiated neuroectoderm; \(\text{p}\), passage. Colours throughout indicate human iPS cell lines that were not selected (light blue) or selected (dark blue). b, PluriTest pluripotency versus novelty score. c, d, Percentage of cells expressing pluripotency and differentiation markers. Brachury is encoded by \(\text{T}\), and nestin by \(\text{NES}\). e, Cumulative distribution of the number of CNAs and the fraction of trisomies per chromosome (inset). f, Relationship between CNA counts and line passage number.

Figure 1 | iPS cell line generation and quality control. a, Schematic of iPC cell line generation and phenotyping. The x axis shows the median number of days, including freeze/thaw cycles (indicated by snowflakes), at each pipeline stage, with stage-specific success rates. hDF, human dermal fibroblasts; \(\text{DEN}\), differentiated endoderm; \(\text{DM}\), differentiated mesoderm; \(\text{dEC}\), differentiated neuroectoderm; \(\text{p}\), passage. Colours throughout indicate human iPS cell lines that were not selected (light blue) or selected (dark blue). b, PluriTest pluripotency versus novelty score. c, d, Percentage of cells expressing pluripotency and differentiation markers. Brachury is encoded by \(\text{T}\), and nestin by \(\text{NES}\). e, Cumulative distribution of the number of CNAs and the fraction of trisomies per chromosome (inset). f, Relationship between CNA counts and line passage number.

Figure 2 | Locations and consequences of recurrent CNAs. a, Genomic locations of CNAs. Colours denote the significance level of recurrence, NS, not significant. b, Genes that are differentially expressed between lines with copy numbers 2 and 3 for the recurrent chr17 CNA. The horizontal bar denotes the 1% FDR threshold (Benjamini–Hochberg). c, Top, genomic location versus number of lines with copy number 3 (grey) and with a CNA (black). Bottom, the \(\text{NAV}\) (average nuclei number reduction) gene score from ref. 22 and \(\log_2\) gene expression fold change between the iPS cell lines with copy numbers 2 and 3 (colour scale), in the region highlighted in red in the top panel. Highlighted genes are upregulated when copy number increases, known oncogenes or tumour-suppressor genes and/or genes with \(\text{NAV}\) score in the top 2%.

CNAs that are observed in pluripotent stem cells (PS cells) are known to recur at certain genomic locations\(^11–13\). We observed 35 regions where CNAs occurred significantly more often than expected under a uniform genomic distribution, including whole-chromosome duplication of the X chromosome (\(P = 1.5 \times 10^{-7}\)), 20 sub-chromosomal duplications, 11 deletions and three regions with both duplications and deletions (Fig. 2a and Supplementary Table 2). The three most frequent CNAs (X trisomy, chromosome 17 and 20) have been previously observed in PS cells\(^12,14,15\), but others have not been identified previously, to our knowledge.

Although recurrent CNAs could be due to mutational hot spots, we did not find a significant overlap between our recurrent CNA set and annotated fragile sites\(^16\) (17% overlap, \(P = 0.075\)). Recurrent CNAs could also arise if duplication or deletion of specific genes led to a selective advantage. To identify potential targets of selection, we defined peak regions of amplification (regions of maximum recurrence for example, Fig. 2c and Extended Data Fig. 4) within each CNA and identified expressed genes (read count > 0 in over 10% of cell lines). Fourteen candidate regions contained fewer than six expressed genes, including genes with established roles in cancer progression (DOCK1, FATS (also known as C10orf90), WWOX, STAG2 and XIAP)

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Characterization of multiple cell lines per donor enabled us to quantify the variance contributed by between-individual differences (hereafter, donor effects) and systematically compare this with variance from other factors, substantially extending previous analyses in smaller cohorts. We identified consistent donor effects for most measured iPSC cell phenotypes, ranging from DNA methylation, through mRNA and protein abundance to pluripotency, differentiation and cell morphology. After accounting for assay-specific batch factors (full list in Methods), donor effects explained 5.2–26.3% of the variance in the genome-wide assays (Fig. 3a, 21.4–45.8% in protein immunostaining (Fig. 3b) and 7.8–22.8% in cellular morphology (Fig. 3c). Collectively, these results indicate that differences between donor individuals affect most traits in iPSC cells.

**Sources of iPSC cell heterogeneity**

Characterization of multiple cell lines per donor enabled us to quantify the variance contributed by between-individual differences (hereafter, donor effects) and systematically compare this with variance from other factors, substantially extending previous analyses in smaller cohorts. We identified consistent donor effects for most measured iPSC cell phenotypes, ranging from DNA methylation, through mRNA and protein abundance to pluripotency, differentiation and cell morphology. After accounting for assay-specific batch factors (full list in Methods), donor effects explained 5.2–26.3% of the variance in the genome-wide assays (Fig. 3a), 21.4–45.8% in protein immunostaining (Fig. 3b) and 7.8–22.8% in cellular morphology (Fig. 3c). Collectively, these results indicate that differences between donor individuals affect most traits in iPSC cells.

**Identification of iPSC-cell-specific regulatory variants**

Using RNA-seq data from 166 unrelated donors (median sequencing depth 38 million reads), we next mapped eQTLs in a 1-Mb cis-window from the gene start. We identified 6,631 genes with an eQTL (FDR < 5%; hereafter, eGenes), 598 of which had a significant secondary eQTL (Supplementary Table 4). Power to discover eGenes in iPSC cells was comparable to that in somatic tissues given our sample size, and eQTLs in iPSC cells showed similar genomic properties to eQTLs in cell lines and tissues (Extended Data Fig. 7 and Supplementary Table 5).

As many eQTLs are shared among tissues, we sought to place eQTLs from iPSC cells in the broader context of somatic tissues.
We assessed eQTL replication from the iPS cells across 44 tissues from the Genotype–Tissue Expression (GTEx) project24 (lead eQTLs and all proxy variants, $r^2 > 0.8$, defining replication as $P < 0.01/45$, where 45 was the total number of tissues tested, see Methods), revealing 2,131 eQTLs that were specific to iPS cells (Fig. 4a). We also considered secondary eQTLs, identifying a similar proportion of genetic effects specific to iPS cells (both 32%). Most tissue-specific signals (72%) occurred in genes with at least one GTEx eQTL that was distinct from the lead iPS cell eQTL variant (not in high linkage disequilibrium, $r^2 < 0.8$), suggesting that iPS-cell-specific eQTLs are frequently driven by independent regulatory variants. Only 11% of the iPS-cell-specific eQTLs could be attributed to tissue-specific gene expression (Fig. 4b), despite higher numbers of genes that were expressed in iPS cells compared to somatic tissues (Extended Data Fig. 7). Similarly, most eQTLs specific to somatic tissues were also driven by alternative regulatory variants, with only testis showing a higher fraction (16%) of eQTLs attributable to tissue-specific gene expression (Fig. 4b). Using alternative methods for eQTL detection and assessing the extent to which eQTLs were shared between iPS cells and GTEx tissues, we confirmed that our conclusions were robust to methodological differences between GTEx and our study (Extended Data Fig. 8).

iPS cell eQTLs tag common disease variants

We next identified eQTLs of iPS cells that may be associated with disease. iPS cell eQTLs tagged 322 variants associated with 145 different traits in genome-wide association studies (GWAS), corresponding to a 1.4-fold global enrichment over control variants (Fisher’s exact test, $P = 1.4 \times 10^{-6}$) and trait-specific enrichments for seven traits (Supplementary Table 6), a comparable level of enrichment to eQTLs from most somatic tissues (Extended Data Fig. 9). We also observed that iPS cell eQTLs tagged a larger number of known cancer genes (COSMIC cancer census, 27 April 2016 (ref. 20)) than eQTLs from somatic tissues, with only eQTLs derived from cancer tissues tagging more (Extended Data Fig. 9).

We subsequently used statistical colocalization28 to identify loci where the same causal variant appeared to be driving both an iPS cell eQTL and an association with one of fourteen complex traits, identifying 233 loci where the posterior probability of a joint association exceeded 0.5 (Supplementary Table 6). Of these, 45 were iPS-cell-specific, including PTPN2, an iPS-cell-specific eQTL that strongly colocalized with risk variants for four autoimmune disorders (Fig. 5a). Previous eQTL studies in both immune cells29–31 and GTEx tissues have not identified a PTPN2 eQTL (Extended Data Fig. 9), suggesting that disease risk variants at PTPN2 may function in stem cells or early development.

Statistical colocalization analysis is limited to instances where full summary statistics are available for both traits. For other disease traits that are available in the GWAS catalogue, we searched for shared lead iPS-cell-specific eQTLs and GWAS variants. We found six instances

### Table 1: Fold enrichment of iPS cell eQTLs

| Tissue Type | Specific | Non-specific |
|-------------|----------|--------------|
| Endo        | 1.0      | 0.8          |
| Mesoderm    | 1.2      | 0.9          |
| Ectoderm    | 1.5      | 1.1          |
| LCL         | 2.0      | 1.5          |
| Testis      | 2.5      | 2.0          |

### Figure 4: Comparison of iPS cell and somatic tissue eQTLs.

**a.** Proportion of tissue-specific eQTLs in iPS cells and 44 GTEx tissues24. Ecto, ectoderm; endo, endoderm; meso, mesoderm; LCL, lymphoblastoid cell line. **b.** Most likely source of tissue-specific eQTLs in iPS cells (primary (1) and secondary (2) eQTLs; red, enriched for non-specific eQTLs; blue, enriched for iPS-cell-specific eQTLs; red, enriched for non-specific eQTLs). **c.** Heat map of the fold enrichment (FE) difference between iPS-cell-specific and non-specific eQTLs at chromatin states from the Roadmap Epigenomics Project26, shown for five aggregated clusters representing 127 cell types. iPS cell, PS cell-derived; SOM, somatic; TSS, transcription start site; ZNF, zinc-finger. **d.** Heat map of the fold enrichment (FE) difference between iPS-cell-specific and non-specific eQTLs at chromatin states from the Roadmap Epigenomics Project26, shown for five aggregated clusters representing 127 cell types. iPS cell, PS cell-derived; SOM, somatic; TSS, transcription start site; ZNF, zinc-finger. **d.** Heat map of the fold enrichment (FE) difference between iPS-cell-specific and non-specific eQTLs at chromatin states from the Roadmap Epigenomics Project26, shown for five aggregated clusters representing 127 cell types. iPS cell, PS cell-derived; SOM, somatic; TSS, transcription start site; ZNF, zinc-finger.
targets of selection. The majority of these recurrent loci are rare and were not reliably identified in previous studies with smaller sample sizes. Compared to previous work\textsuperscript{11,12}, we observed substantially lower levels of genetic aberrations. One possible explanation is that access to donor-matched reference samples helped us more accurately identify germline CNAs that would otherwise have inflated our estimates, while previous studies in ES cells were unable to perform similar comparisons.

Our study provides a high-resolution map of common regulatory variation in human PS cells. We show that variation in local gene regulation in IPS cells is similar to that in somatic tissues, with eQTLs driving cell-type-specific expression profiles through distal tissue-specific regulatory elements. We have identified eQTLs that function primarily in pluripotent cells, a subset of which tag loci associated with disease. These loci may drive disease susceptibility through molecular changes early in development or, more generally, in cells with ‘stem-like’ characteristics, which are not well captured by studies of differentiated primary tissues from adult individuals. A compelling example of this is the IPS-cell-specific eQTL regulating TERT expression. In human tissues, telomerase activity is mainly restricted to stem cells, with most somatic tissues silencing TERT expression. However, cancer cells bypass this tumour-suppressive mechanism by reactivating telomerase activity\textsuperscript{48}. This result highlights how IPS cells could be used to study the genetic effects of diseases that manifest in transient states during cellular growth and differentiation, including in cancer\textsuperscript{49}.

The analysis of the recurrence of CNAs and the eQTL map we present are based on a large sample, providing a high-confidence map of molecular associations in IPS cells. We have presented preliminary experimental characterization of some of the CNAs we detected, however our results are inconclusive. An important next step will be to perform more extensive functional characterization to understand how cellular phenotypes of IPS cells are influenced by CNAs and IPS cell eQTLs. We anticipate that the lines and data we have generated here will be a valuable starting point for future studies, to understand how germline and somatic genetic variation influences IPS cell growth and differentiation.

In summary, our study provides a detailed picture of the genetic and phenotypic variability in human PS cells, including major drivers of this variation. Data and cell lines from this study are being made available through HipSci (http://www.hipsci.org), the European Collection of Authenticated Cell Cultures (ECACC) and the European Bank of Induced Pluripotent Stem Cells (EBiSC). As the HipSci resource continues to expand in sample size and assays, it will enable the study of subtler genetic effects, under a wider range of conditions, in an increasing range of disease-relevant differentiated cell types.

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 11 May 2016; accepted 27 April 2017.

Published online 10 May 2017.

1. Sternebeck, J. L., Reinhardt, P. & Schöler, H. R. Investigating human disease using stem cell models. Nat. Rev. Genet. 15, 625–639 (2014).
2. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. Nature 467, 285–290 (2010).
3. Kim, K. et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat. Biotechnol. 29, 1117–1119 (2011).
4. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68–73 (2011).
5. Nazor, K. L. et al. Recurrent variations in DNA methylation in human induced pluripotent stem cells and their differentiated derivatives. Cell Stem Cell 10, 620–634 (2012).
6. Rouhani, F. et al. Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS Genet. 10, e1004452 (2014).
7. Burnes, C. K. et al. Genetic variation, not cell type of origin, underlies the majority of identifiable regulatory differences in iPSCs. PLoS Genet. 12, e1005793 (2016).

Discussion

Here we present a comprehensive analysis of genetic and phenotypic data from human IPS cell lines. Our study substantially expands on previous work\textsuperscript{46,47}, by demonstrating widespread functional consequences of genetic variation for many molecular and cellular phenotypes in human PS cell lines, including in the efficiency with which IPS cells differentiate\textsuperscript{40–42}. This is potentially a consequence of variation in core components of the regulatory networks that control cellular differentiation, as has been observed previously in haematopoietic cells and mouse and fly embryos\textsuperscript{43–45}, and in parallel large-scale projects in human IPS cells\textsuperscript{46,47}.

We have also created a high-resolution map of recurrent genetic abnormalities in human IPS cells and identified plausible candidate...
8. Vallier, L. et al. Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. Stem Cells 27, 2655–2666 (2009).

9. Müller, F. J. et al. A bioinformatic assay for pluripotency in human cells. Nat. Methods 8, 315–317 (2011).

10. Danecek, P. & McCarthy, S. A., HiSeq Consortium & Durbin, R. A method for checking genomic integrity in cultured cell lines from SNP genotyping data. PLoS ONE 13, e0195014 (2016).

11. Laurent, L. C. et al. Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. Cell Stem Cell 8, 106–118 (2011).

12. The International Stem Cell Initiative. Screening ethnically diverse human embryonic stem cell lines identifies a chromosome 20 minimal amplicon conferring growth advantage. Nat. Biotechnol. 29, 1132–1144 (2011).

13. Grundberg, E. et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature 492, 438–442 (2012).

14. Mayshar, Y. et al. Identification and classification of chromosomal aberrations in human induced pluripotent stem cells. Cell Stem Cell 7, 521–531 (2010).

15. Taapken, S. M. et al. Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells. Nat. Biotechnol. 29, 313–314 (2011).

16. Osses, S. M. et al. Copy number variation and selection during reprogramming to pluripotency. Nature 471, 56–62 (2011).

17. Laurin, M. & Côté, J. F. Insights into the biological functions of Dock family guanine nucleotide exchange factors. Genes Dev. 28, 533–547 (2014).

18. Zhang, X. et al. FATS is a transcriptional target of p53 and associated with antitumor activity. Mol. Cancer 9, 244 (2010).

19. Lo, J. Y., Chou, Y. T., Lai, F. J. & Hsu, L. J. Regulation of cell signaling and apoptosis by tumor suppressor WWOX. Exp. Biol. Med. 240, 383–391 (2015).

20. Futreal, P. A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).

21. Dukett, C. S. et al. A conserved family of cellular genes related to the baculovirus ial gene and encoding apoptosis inhibitors. EMBO J. 15, 2685–2694 (1996).

22. Chia, N. Y. et al. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468, 316–320 (2010).

23. Belinsky, S. A. et al. PathCards: multi-source consolidation of human biological pathways. Database (Oxford) 2015, bav006 (2015).

24. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015).

25. Grundberg, E. et al. Mapping cis- and trans- regulatory effects across multiple tissues in twins. Nat. Genet. 44, 1084–1089 (2012).

26. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 187–193 (2015).

27. Xu, H. et al. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford) 2013, bato045 (2013).

28. Ghiambartolomei, C. et al. A Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383 (2014).

29. Dubois, P. C. et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 44, 295–302 (2010).

30. Stranger, B. E. et al. Population genomics of human gene expression. Nat. Genet. 39, 1217–1224 (2007).

31. Zeller, T. et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS One 9, e10693 (2010).

32. Purrington, K. S. et al. Genome-wide association study identifies 25 known copy number variation on gene expression in human induced pluripotent stem cells. Stem Cell Reports 6, 131–143 (2016).

33. Garratt, A. N. et al. Constraint and divergence of global gene expression in the mammalian embryo. eLife 4, e07918 (2015).

34. Wang, Z. et al. Expression quantitative trait loci associated with glioblastoma patient survival. Proc. Natl Acad. Sci. USA 109, 12538–12543 (2012).

35. Shu, N. et al. Constraint and divergence of global gene expression in the mammalian embryo. eLife 4, e07918 (2015).

36. Cavanagh, E. et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. Nature 541, 402–406 (2017).

37. Grundberg, E. et al. A genome-wide RNAi screen reveals determinants of human pluripotency. Stem Cell Reports 20, 518–532 (2017).

38. DeBoever, C. et al. Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. Cell Stem Cell 20, 533–546 (2017).

39. Kim, N. W. et al. Specific association of human telomerase activity with immortal cells and cancer. Science 266, 2011–2015 (1994).

40. Kelly, P. G. & Gilliland, D. G. Genetics of myeloid leukemias. Annu. Rev. Genomics Hum. Genet. 3, 179–198 (2002).

41. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

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

Acknowledgements This work was funded with a strategic award from the Wellcome Trust and UK Medical Research Council (WT098503). We thank the staff in the Cellular Genetics and Phenotyping and Sequencing core facilities at the Wellcome Trust Sanger Institute, Work at the Wellcome Trust Sanger Institute was further supported by Wellcome Trust grant WT098581. H.K. is supported by a MRC eMedLab Medical Bioinformatics career development award from the UK Medical Research Council (MR/L016311/1). F.M.W. acknowledges financial support from the Department of Health via the NIHR Biomedical Research Centre award to Guy’s and St Thomas’ National Health Service Foundation Trust in partnership with King’s College London and King’s College Hospital NHS Foundation Trust. We acknowledge the participation of all NIHR Cambridge BioResource volunteers, and thank the NIHR Cambridge BioResource centre staff for their contribution. We thank the National Institute for Health Research and NHS Blood and Transplant, The NIHR/Wellcome Trust Cambridge Clinical Research Facility supported the volunteer recruitment. We acknowledge Life Science Technologies Corporation as the provider of CytoTune. We thank F.-J. Müller for insights regarding the PluriTest method, and the GTEx consortium for making raw data and intermediate results available.

Author Contributions H.K., A.G., O.S. and D.J.G. wrote the paper with input from all authors. H.K., A.G., D.B., Y.M., I.S., P.D., D.M., A.M., D.D., A.I.L., O.S. and D.G. contributed to the Supplementary Information. H.K., A.G., A.L., F.P.C., P.D., O.M., K.A. and D.D. analysed the data, S.A. and W.H.O. managed and supervised collection of research volunteer samples. F.S., C.A.A., A.R., N.S.H., M.P., S.R.P., A.W. and C.M.K. generated iPSC cell lines, tier 1 assay data, cell growth data, RNA-seq and methylation data. V.A. and D.B. generated and processed the high-content cellular imaging data. S.A.M., S.B. and Y.M. carried out initial data quality control and bioinformatics processing/pipelines. A.F., P.W.H., I.S. and L.C. curated and managed data and the project website. R.H. and A.K.-K. coordinated the project. D.O., P.B., W.H.O., E.B., L.V., A.L., F.M.W., R.D., O.S. and D.G. performed tissue or disease specific research. H.K. and A.K.-K. contributed equally to this work; O.S. and D.J.G. contributed equally to this work.

Author Information 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. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.J.G. (dg13@sanger.ac.uk), O.S. (stegle@ebi.ac.uk), R.D. (rd@sanger.ac.uk) or F.M.W. (fiona.watt@kcl.ac.uk).

Reviewer Information Nature thanks E. Dermitzakis, S. Montgomery and the other anonymous reviewer(s) for their contribution to the peer review of this work.
METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

Generation of iPS cell lines. All samples for the HipSci resource were collected from consented research volunteers recruited from the NHIR Cambridge BioResource (http://www.cambridgebioresource.org.uk). Samples were collected initially under ethics for iPS cell derivation (REC 09/H0304/77, V2 04/01/2013), with later samples collected under a revised consent (REC 09/H0304/77, V3 15/03/2013).

Fibroblast isolation. Primary fibroblasts were derived from 2-mm skin punch biopsies from each donor. Biopsies were collected in fibroblast growth medium (advanced DMEM, 10% FBS, 1% l-Glutamine, 0.007% 2-mercaptoethanol and 1% penicillin and streptomycin (pen/strep)) in falcon tubes at room temperature. Biopsies were manually dissected using a microscope under a drop of fibroblast medium using sterile scalpels. The biopsy fragments were transferred onto a 60-mm Petri dish containing several drops of fibroblast growth medium. Sterile cover slips were placed onto the dissected pieces of tissue to hold them in place against the bottom of the plate. The explants were cultured for five days and the spent medium was removed and replaced with a few drops of medium (1 ml) to prevent dehydration. The explants were fed every five days with 1 ml fibroblast medium until fibroblast outgrowths appeared. The explants were screened for presence of mycoplasma using a standard PCR kit (EZ-PCR Kit, Gene flow (41106313-001)). On average outgrowths appeared within 14 days, with a small fraction of samples failing to produce outgrowths (12% of cases). Failures were due to contamination (0.5%) or lack of observed outgrowths after 30 days (11%). Approximately 30 days after dissection, when the fibroblasts had reached confluence, the culture was trypsinized and passaged into a 25 cm² tissue-culture dish in iPS cell medium consisting of advanced DMEM (Life technologies, A1517001) supplemented with E8 supplement (50 μg/ml 47,230; < 50 ng μl⁻¹ by the on-site sample-management team before submission for sequencing.

Genotyping (gtarray). Experimental processing of arrays. Samples were hybridized to the Illumina HumanCoreExome-12 BeadChip according to the manufacturer’s guidelines. In total, 4 μl (200 ng) of DNA is required for the pre-amplification reaction using a Tecan Freedom Evo. The process is automated, except for a manual agitation/centrifugation step midway through and at the end of the process. Post-amplification processes (fragmentation, precipitation, resuspension, hybridization to BeadChip and Illumina sXtain) were completed over three days as per Illumina protocol. Following the staining process, BeadChips were coated for protection and dried completely under vacuum before scanning on the Illumina iScan paired with Illumina Autoloader 2.x. Before downstream analysis, all samples were subjected to initial quality control to establish that the assay was successful. Sample call-rates below 92.5% were flagged before loading samples into Illumina’s GenomeStudio software. Using Illumina’s QC dashboard, sample performance was assessed by measuring dependent and non-dependent controls that are manufactured onto each BeadChip during production.

Genotype calling and imputation. After primary quality control, the genotyping (GT) module of the GenomeStudio software (Illumina) was used to call the genotypes. For each probe, the GT module estimates the log R ratio and B-allele frequency for each sample using a clustering model applied to the distribution of signal intensities. These statistics are used internally by GenomeStudio to assign signal intensities. These statistics are used internally by GenomeStudio to assign the sample genotypes for each marker. Variant coverage was further increased using statistical imputation and phasing. We constructed a reference panel of haplotypes from a combination of SNPs and small insertions and deletions (indels) in the UK10K cohorts and 1000 Genomes Phase 1 data22,23. Samples were independently imputed using IMPUTE2 version 2.3.1 (ref. 54) and subsequently phased using SHAPEIT version 2.r790 (ref. 55). This analysis was done in chunks of, on average, 5 Mb, with 300 kb buffer regions on each side. IMPUTE2 was used with its default MCMC options (-Ne 20000 -k 80) for autosomes and -Ne 15000 -k 100 for the X chromosome. SHAPEIT was run without MCMC iteration (-no-mcmc) so that each sample was phased independently using the reference panel as the haplotype scaffold, regardless of the phasing of the other samples. Single-sample VCFs were merged together and INFO scores were recalculated from genotype posterior probabilities. Variants with an INFO score less than 0.4 were excluded from further analysis. Cell lines originating from the same donor were checked for possible sample swaps using BCFtools (bcftools gcheck -G1). Swapped samples typically had a large number of discordant genotypes (>20%), whereas in samples from the same donor the number of discordant genotypes was low (<0.3%), even in the presence of large copy-number variation.

Gene expression arrays (gxarray). Experimental processing of arrays. 500 ng of total RNA for each sample was amplified and purified using the Illumina TotalPrep-96 RNA Amplification kit (Life Technologies), according to the manufacturer’s instructions. Biotin-labelled complementary RNA was then normalized to a concentration of 150 ng μl and 750 ng was hybridized to Illumina Human HT-12 v4 Expression BeadChips (Illumina) for 16h (overnight) at 58°C. Following hybridization, BeadChips were washed and stained with streptavidin-Cy3 (GE Healthcare). BeadChips were then scanned using the ReadArray reader and image data were processed using GenomeStudio software (Illumina).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
We first mapped the sequences allowing no mismatches (−n, seed disabled) and kept uniquely mapping probes with a minimum mapping quality (MAPQ) of 10 (−q 10). These sequences were then mapped again, this time allowing one mismatch (−n 1). Again, only uniquely mapping probes with MAPQ > 10 were retained, resulting in a total of n = 37,740 probes. We further removed all probes that overlapped with any variant with a minor allele frequency greater than 0.05 in the main imputed dataset (nmiss = 858). Remaining probes were annotated with Gencode version 19 gene annotations and only probes mapping uniquely to a single gene were selected (one-to-one probe number np1to1 = 25,604, representing 17,116 unique genes of which 14,569 are protein-coding.

Pre-processing and normalization of data. Gene expression profiles were measured with Illumina HumanHT-12 v4 Expression BeadChips. After limiting the dataset to iPS cell lines derived from fibroblast of healthy donors, we obtained data from 711 iPS cell lines and 301 somatic fibroblast lines. Probe intensity estimates were normalized separately for the two cell types using the variance-stabilizing transformation implemented in the R/Bioconductor vsn package [2,26]. After normalization, the datasets were limited to the final remapped set of probes (np1to1 = 25,604). We refer to this version of the gexarray data by vsn log(iPS cell/somatic).

Cellular differentiation assay (cellomics). Differentiation potential assays. Selected iPS cell lines were assessed for their pluripotency and differentiation properties by culturing the cells under conditions favouring the formation of the three embryonic germ layers, and subsequent immunostaining with markers specific for pluripotency and differentiation. Differentiation was performed as described previously [29]. In brief, iPS cells grown in feeder-dependent or feeder-free conditions were collected using either collagenase and dispase or EDTA, respectively. Colonies were collected, washed in medium and mechanically broken up before being re-plated onto 24-well MEF feeder plates or pre-coated gelatine/PBS plates. For pluripotency assays, feeder-dependent colonies were seeded on MEF feeder plates and feeder-free colonies onto Vitronectin plates. For the differentiation assay, colonies were grown on gelatine/PBS plates. Before differentiation into mesoderm, endoderm or neuroectoderm, cells were cultured overnight in pre-differentiation medium CDM-PVA supplemented with recombinant activin-A (10 ng ml⁻¹; CSCR, University of Cambridge) and zebrafish FGF2 (12 ng ml⁻¹; CSCR, University of Cambridge).

For differentiation into mesoderm following culture in pre-differentiation medium, spent medium was removed and replaced with fresh CDM-PVA medium containing bone morphogenetic protein 4 (BMP4, 40 ng ml⁻¹, R&D Systems), FGF2 (20 ng ml⁻¹; CSCR, University of Cambridge), recombinant activin-A (10 ng ml⁻¹; CSCR, University of Cambridge), BMP4 (10 ng ml⁻¹ R&D Systems), LY29004 (10 mM, Promega), CHIR99021 (5 mM, Selleckchem) and subsequently cultured for three days. Medium was changed daily.

For differentiation into endoderm, following culture in pre-differentiation medium further cells were cultured in differentiation medium for three days. In brief, on day one, medium was removed and replaced with fresh CDM-PVA medium supplemented with recombinant activin-A (100 ng ml⁻¹; CSCR, University of Cambridge), zebrafish FGF2 (80 ng ml⁻¹; CSCR, University of Cambridge), BMP4 (10 ng ml⁻¹ R&D Systems), LY29004 (10 mM) and CHIR90021 (3 mM). On day two, medium was removed and replaced with fresh CDM-PVA medium supplemented with recombinant activin-A (100 ng ml⁻¹), zebrafish FGF2 (80 ng ml⁻¹), BMP4 (10 ng ml⁻¹) and LY29004 (10 mM). On day three, medium was removed and replaced with RPMI medium supplemented with R27 (1 × Life Technologies), recombinant activin-A (100 ng ml⁻¹), zebrafish FGF2 (80 ng ml⁻¹) and Non-essential amino acids (1 × Life Technologies).

For differentiation into neuroectoderm, iPS cells were grown for 12 days in CDM-PVA medium supplemented with SB431542 (10 mM; Tocris Bioscience), FGF2 (12 ng ml⁻¹, CSCR University of Cambridge) and noggin (150 ng ml⁻¹, R&D Systems). Medium was changed daily.

Immunostaining for pluripotency and differentiation markers. For the detection of pluripotency and differentiation markers, cells grown in 24-well plates were fixed with cold ethanol for 20 min before being washed three times with 10% donkey serum and 0.1% Triton X-100 in PBS. Subsequently, cells were stained with primary antibodies overnight at 4 °C and finally incubated with fluorochrome-labelled secondary antibodies (Invitrogen). The primary antibodies used for detecting pluripotency markers were: anti-OCT4 (SC-5279, Santa-Cruz Biotech), anti-SOX2 (AF2018, R&D), anti-NANOG (AF1997, R&D). The primary antibodies used for detecting endoderm markers were: anti-SOX17 (AF1924, R&D), anti-CXC4R4 (MAB173-100, R&D) and anti-GATA4 (SC-25310, Santa-Cruz Biotech). The primary antibodies used for detecting mesoderm markers were: anti-brachyury (AF2085, R&D), anti-EOMES (ab23345, Abcam) and anti-MIXL1 (SC-98664, Santa-Cruz Biotech). The primary antibodies used for detecting neuroectoderm markers were: anti-nestin (ab22035, Abcam), anti-SOX1 (AF3369, R&D) and anti-SOX2 (AF2018, R&D). The secondary antibodies used were: donkey anti-goat AF488 (Invitrogen), donkey anti-mouse AF488 (Invitrogen), donkey anti-rabbit AF488 (Invitrogen). Additionally, DAPI staining was used to label the nucleus in order to facilitate cell segmentation.

Images were captured and quantified using a Cellomics Array Scan imaging system. In brief, images were taken in 24-well plastic plates. Individual plates were used to either measure pluripotency markers or markers to assess differentiation for one of the germ layers. Each plate contained cells from one or two cell lines, as well as technical replicates for each measurement. Three types of plate layouts were considered throughout the project: two-channel, three-channel and three-channel with single staining. For all layouts, the signal from the DAPI staining was read in the first channel. The first columns of each plate were used for marker staining; the second and third columns were used with the secondary antibody to measure background signal (Extended Data Fig. 1).

Processing of images on the Cellomics instrument. Individual wells in the plate were imaged consecutively, either until the whole plate was imaged or until 10,000 individual cells were detected. Cell detection was performed based on nucleus segmentation from DAPI staining. All considered markers, except for CXC4R4, are nuclear markers, so their signal intensities were measured in the segmented nuclear area. The cell-surface marker CXC4R4 was quantified in a circle around the segmented nucleus. For each cell and marker, we used the average intensity within the respective quantitative area as final readout. Each batch of lines for staining included the reference line (CTRLR214kfp-y). This reference line was used to determine parameter values for cell size (usually around 30–400) and an approximate intensity threshold for detecting responding cells.

To quantify cellomics phenotypes, we fit a Gamma mixture model to the cellomics raw intensities (Supplementary Information). In brief, this model was fit to primary wells as well as background wells (Extended Data Fig. 1), thereby estimating both the proportion of responding cells as well as the overall intensity (expression) of the corresponding cells.

For downstream analyses, technical replicates on each plate were aggregated using average values. Analogously to the processing steps for gene expression arrays, we regressed out batch (derived from the date of staining), medium type, gender, passage number, plating technician, fixation technician and the technician in charge of the staining. Analyses focused on quality control were based on the proportions of responding cells out of the total cell number (Fig. 1c, d). Variance component analyses (Fig. 3b) and marker correlations (Extended Data Fig. 1c) were based on quantitative expression estimates, averaged across individual markers for a given layer.

Proteomics. Sample preparation. Frozen iPS cell pellets were thawed and washed with PBS twice before lysis. The protein content of the cells was extracted by re-dissolving the pellets in 8 M TEAR, pH 8.5 and mixing at room temperature for 15 min. Next, the DNA content of the cells was sheared using ultrasonication. The protein amount was determined using a fluorescence-based assay (EZQ, Life Technologies) before double digestion using mass-spectrometry-grade lysyl endopeptidase (Wako) and trypsin (Pierce) in a substrate-to-enzyme ratio of 1:50 (w:w), at final urea concentrations of 2 M and 0.8 M, respectively. The digested proteins were desalted using sepk vacuum cartridges (Waters) and dried in vacuo. The desalted peptides were redisolved in 10 mM borate (pH 9.3):acetoniite at a ratio of 80:20 hydrophilic strong anion exchange (hSAX) fractionation using a 40 min gradient. A total of 16 fractions were collected, desalted and dried. The hSAX fractions were redisolved in 5% formic acid for label-free LC–MS analysis. In addition to individual samples, a composite reference sample (HPSI_composite_1503) was constructed by pooling together protein lysates from 43 iPS cell lines. For each line, 2 μg of protein was used. All samples in this reference were of fibroblast origin and reprogrammed with sendai virus.

For tandem mass tag (TMT)-based quantification, the dried peptides were re-dissolved in 100 mM TEAB (50 μl) and their concentration was measured using a fluorescent assay (CBQCA, Life Technologies). 100 μg of peptides from each cell line to be compared, in 100 μl of TEAB, were labeled with a different TMT tag (20 μg ml⁻¹ in 40 μl acetoniite) (Thermo Scientific), for 2 h at room temperature. Afterwards, the labeling reaction was enriched with 8 μl of 5% hydroxyamine (Pierce) for 30 min and the different cell lines/tags were mixed and dried in vacuo.

The TMT samples were fractionated using off-line high pH–reverse phase chromatography: samples were loaded onto a 4.6 × 250 mm Xbridge BEH130 C18 column with 3.5-μm particles (Waters). Using a Dionex biosRS system, the samples were separated using a 25-min multistep gradient of solvents A (10 mM formate at pH 9) and B (10 mM ammonium formate pH 9 in 80% acetonitiile), at a flow rate of 1 ml min⁻¹. Peptides were separated into 48 fractions, which were consolidated into 24 fractions. The fractions were subsequently dried and the peptides re-dissolved in 0.3% formic acid in MS/MS.

Label-free analysis. RPLC was performed using a Dionex RSLCnano HPLC (Thermo Scientific). Peptides were injected onto a 75 μm × 2 cm PepMap-C18 pre-column and resolved on a 75 μm × 50 cm RP-C18 EASY-Spray temperature-controlled integrated column-emitter (Thermo Scientific) using a four-hour multistep gradient from 5% B to 35% B with a constant flow of 200 nl min⁻¹ as described.
Quantification and normalization of data. Mapped reads were quantified on the level of genes using HTSeq version 0.6.1p1 (ref. 65) and annotations from Gencode version 19 (ref. 57). We used the ‘union’ method of ‘htseq-count’ for unstranded quantification output (PXD003506). Pre-processing and normalization of data. Data for analysis were obtained from the ProteinGroups.txt output of MaxQuant. Contaminant and reverse hits (n = 3,419) were excluded from analysis. For each sample, the total protein abundance was calculated by summing-up protein intensity (Intensity) values across all proteins and protein groups. This value was then used to scale all quantification values (iBAQ) per sample. For a protein or a protein group to be considered, we required at least one unique peptide mapping to it. Overall, we quantified 10,097 protein groups (4,877 unique proteins) in at least one of the samples. Only unique protein entries quantified in at least half of the samples were used in the subsequent analyses (3,435 proteins). The mean pairwise correlation of samples was 0.87 for unique proteins (Spearman rank correlation). On the basis of the clustering of samples (principal component analysis and pairwise correlation of protein quantification; data not shown), one sample appeared as an outlier (HPS01713i-darw_1) and was excluded from further analyses.

DNA methylation (marray). Sample preparation and experimental processing of arrays. 500 ng of DNA was used for bisulphite conversion using the Zymo Research EZ-96 DNA Methylation kit. The bisulphite-converted DNA extracts were hybridized to Infinium 450 K ReadChip (Illumina). Owing to the differences in sample plates between the completed Zymo assay and the Illumina assay, pre-amplification of a number of samples was required to ensure consistent amplification efficiency across the entire batch.

RNA sequencing. Library preparation and sequencing. miRNA in total RNA was isolated and converted into non-stranded or stranded libraries. Non-stranded libraries were produced manually using reagents provided in the Illumina TrueSeq RNA Sample Preparation kit v2 in accordance with the manufacturer’s recommendations. The protocol was modified to produce size-selected libraries by modifying the fragmentation conditions and using a Caliper LabChip XT instrument. Stranded libraries were prepared using a NeoPrep Library Prep System and the reagents provided in the Illumina TrueSeq Stranded mRNA Library Preparation kit. The stranded library preparation workflow is similar to the non-stranded workflow, except that it involves additional ribosomal reduction chemistry to maximize the percentage of uniquely mapped reads. Following purification, the RNA was fragmented and synthesized into cDNA using reverse transcription. The products were then enriched with PCR (maximum of 10 cycles) to create the final cDNA library. Enriched libraries were subjected to 75-base paired-end sequencing using Illumina HiSeq 2000 v3 kits following the manufacturer’s instructions. Pre-processing of sequence data. Raw RNA-seq data were aligned using STAR version 2.4.0 (ref. 64) against the 1000 Genomes Phase 2 reference genome assembly that integrates the GRCh37 primary assembly with the human decay sequence 37d5. Exon–intron junctions derived from Gencode v19 transcript annotations68 were used to improve the alignments. The same approach was taken to re-align data for two tissues from the GTEx Project (Extended Data Fig. 8). Raw fastq files were obtained from dbGaP (http://www.ncbi.nlm.nih.gov/gap, accession phs004242.v6.p1) for adrenal gland (n = 126 samples) and oesophagus gastro–oesophageal junction (n = 127 samples; limited to 126 unique samples used in the GTeX V6p map). Quantification and normalization of data. Mapped reads were quantified on the level of genes using HTSeq version 0.6.1p1 (ref. 65) and annotations from Gencode version 19 (ref. 57). We used the ‘union’ method of ‘htseq-count’ for unstranded quantification output (PXD003506). To match the original GTEX V6p quantifications as closely as possible, we ran RNA-SeqC with the -strictMode flag and used custom exon annotations generated and used by GTEX (gencode.v19.genes.v6p_model.patched_contigs.gtf.gz) to obtain gene fragments per kilobase of transcript per million fragments (FPKMs). The same RNA-SeqC quantification pipeline was applied to the two re-mapped GTeX tissues.

High-content cellular imaging. Sample preparation and cellular imaging. Each line was cultured and passed as previously described48. In brief, 96-well plates were coated with three concentrations of fibronectin in alternating columns in a randomized fashion. Cell lines were also seeded in rows in a randomized fashion. Per well, 3,000 cells were plated and fixed after 24 h. EdU was incorporated 30 min before fixation. Plates were then fixed and stained with DAPI, cell mask and EdU staining. Images were acquired using the Operetta (Perkin Elmer) high-content device. Using the Harmony software, measurements were derived for each cell. Measurements included intensity features (DAPI, EdU), morphology features (cell area, cell roundness, cell width to length ratio, nucleus area, nucleus roundness, nucleus width to length ratio) and context features related to cell-adhesion properties (manually per clump). Processing, quantification and normalization of data were performed as previously described48. The fluorescence imaging data acquired by celiomics were used to quantify the fraction of cells expressing each protein marker independently. In the absence of co-staining information we used the marginal fractions of cells expressing each marker to calculate lower and upper bounds for the fractions of cells expressing all markers simultaneously as follows: let \( P(A_i) \) be the fraction of cells expressing protein marker \( A_i \) and \( P\left(\bigcup_{i=1}^{n} A_i\right) \) the fraction of cells expressing \( n \) markers simultaneously, then this value is bounded by:

\[
\max \left\{ \sum_{i=1}^{n} P(A_i) - n + 1, 0 \right\} \leq P\left(\bigcup_{i=1}^{n} A_i\right) \leq \min\left\{ P(A_1), \ldots, P(A_n) \right\}
\]

Variance component analysis. Feature (gene, protein or probe) intensity estimates for each of the assays were pre-processed and normalized as described in the

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
individual assay sections and subsequently transformed into a standard normal distribution across lines. For each feature in each assay, variance was partitioned using a linear mixed model (implemented in the lme4 R package) fitted with all metadata variables as random effects. Only lines with complete metadata information were included in each one of the analyses, these numbers are shown in parenthesis in Fig. 3a-c. The variance components were normalized to sum to one and subsequently averaged across the different sets of features considered (Fig. 3). The fraction of non-technical variance explained by each biological or experimental factor or random error explained by the model was divided by the total variance minus the variance explained by assay batches (see below for a definition of experimental and assay batch factors). Confidence intervals for the cellomics variance components were obtained with the ‘profile’ method implemented in the ‘confint’ function of the lme4 package. The following random effects were included for each assay as follows. (1) Metylation: donor, experimental factors (summed up to produce Fig. 3a; gender, passage interval at time of assay), assay batches (summed up to produce Fig. 3a; Illumina Sentrix array ID (slide number), Sentrix array (within Sentrix ID), bisulphite-conversion plate, year and month of assay, year and month of tier 1 assays). (2) Expression microarrays: donor, experimental factors (gender, passage interval at the time of assay, culture system at time of assay, trisomy status, recurrent CNA status), assay batches (array batch, year and month of assay, BeadChip ID, BeadChip array (within BeadChip ID), technician ID, assay performed before or after April 2014). (3) RNA-seq: donor, experimental factors (gender, passage interval at time of assay, trisomy status, recurrent CNA status, culture system at time of assay), assay batches (year and month of assay, year and month of tier 1 assays). (4) Proteomics (uniquely identified proteins): donor, experimental factors (gender), assay batches (year and month of QC assays, instrument, year and month of analysis). (5) Cellomics, donor, experimental factors (gender, culture system at time of assay, passage number of assayed cells), assay batches (date of staining, plating technician ID, fixation technician ID, staining technician ID, primary antibody lot and secondary antibody lot). (6) Cellular morphology assays (cell area, roundness, EDU, PC1 cellmorph): donor, experimental factors (gender, cell line, fibrinconcentration), assay batches (plate, row). CNA analysis. Pairwise fibroblast–iPS cell CNA detection. Copy-number differences between fibroblasts and iPS cell lines from the same donor were checked using a HMM algorithm implemented in BCFtools/cnv for this purpose18. In order to distinguish between normal variation and novel CNAs as well as to reduce the number of false calls, the program was run in the pairwise mode (bcftools cnv -c <donor> -s <derived> ) with default parameters. The CNA calls were filtered to exclude calls with a quality score smaller than 2, deletions with fewer than 10 markers, and duplications with fewer than 10 heterozygous markers. Three sets of CNA calls were generated: a more lenient set containing all calls >0.2Mb in length, a set with all calls >0.5Mb and a stricter subset of the previous with calls >1Mb.

Statistical significance of recurrent CNAs was estimated from the complementary cumulative distribution function of the binomial distribution, and the significance of sub-chromosomal events was estimated using a permutation test (Supplementary Information). Overlap with annotated regions. To assess the significance of the overlap between CNAs and annotated regions (namely chromatin fragile sites34 and recurrent somatic copy number altered regions in cancer35) we randomly generated a set of 2,000 matched control regions for each CNA. Control regions were generated so that they had the same size as the CNA, did not overlap with telomeres or centromers, and did not overlap the original CNA. Overlaps were determined between the CNA and the annotated regions and between the matched control regions and the annotated regions to calculate an empirical $P$ value.

Association between CNAs and gene expression. To determine the functional consequence of CNAs with regards to gene expression we first selected CNAs for which we had expression data on at least 5% of the cell lines. Then, we permutated the labels of the cell lines in the dataset and calculated the correlation between CNAs and gene expression for each of these CNAs we took the copy number at the region of peak coverage for each cell line (see CNA coverage plots in Fig. 2c and Extended Data Fig. 4). We then defined the set of expression array probes to test by choosing only expressed probes. Here a probe was defined as expressed if the number of RNA-seq fragment counts (normalized between samples for sequencing depth) mapping to the genomic regions targeted by the probe was greater than 0 in 10% or more of the lines. Finally, we used a linear mixed model to independently test for association between the copy number of each CNA and the intensity of each probe. We included culture condition, gender, and an interaction between copy number and culture condition as fixed effects; and used donor and assay batch as random effects. $q$ values were obtained for each CNA using the Benjamini–Hochberg test to adjust for multiple testing. The same approach was employed to test for association between X chromosome copy number and gene expression, but we limited the tests to female samples and to probes on the X chromosome.

Gene set enrichment analysis. Pathway enrichment analysis of the genes regulated by chr17 was performed with GSEA20 on the full list of genes ordered by effects size of association between gene expression with copy number and with a custom set of pathways. The custom set of pathways considered comprised 1,156 super pathways from the PathCards database21 filtered to exclude pathways relating to infectious diseases and pharmacokinetics. Multiple testing corrections were performed as described in ref. 70.

eQTL mapping. Pre-processing of genotype data. Variants in the original VCF file were explained by forming a binary output file (‘chr_pose’ for example, ‘snp_236887241’ or ‘indel_2D_1_18847945’), and filtered for polymorphic and biallelic sites with VCFtools version 0.1.12b (ref. 71) (vcftools –gzvcf IN–mac 1–min-alleles 2- max-alleles 2–recode–recode-INFO-all–out OUT). The resulting VCF files were then converted to ‘.02’ format (vcftools–gzvcf IN–02-out OUT.0012), where 0, 1 and 2 represent the number of non-reference alleles, and further to HDF5 format using a converter function (-g012) from LIMIX (https://github.com/PMBio/limix). This resulted in a set of $n = 14,644,791$ variants. To obtain allele dosages, genotype likelihoods (GL) in the original VCF were converted to genotype probabilities (GP) with BCFtools (bctools –tag2Ing IN -O -o OUT–gg-tp-glo) and used to define allele dosage (REF, Reference; ALT, alternative) as follows. Dosage of alternative allele = GP(REF/ALT) + 2 × GP(ALT/ALT). Genotype dosage information was converted to HDF5 format using LIMIX converter (-g012 dosage). For eQTL mapping, we included autosomal variants with a minimum minor allele frequency of 1% in our samples and maximum 10% missing values across individuals. Variant sites were further required to have a minimum IMPUTE2 INFO score of 0.4 to assure good imputation quality. Missing genotypes were mean imputed and the dosage of the alternative allele used for mapping. Pre-processing of expression data. Scaled gene counts were filtered for missing values (maximum 90% missing values, that is, zero counts, allowed per gene). Zero values were offset by 1, after which the data was log10-transformed and quantile-normalized across individuals using the R limma normalizeQuantiles function22. We then ran PEER23 with the full pre-normalized dataset with the following parameters: $K = 30$; covariates = gender, iPS cell growth condition (feeder-dependent/E8), mean expression (addMean = True in PEER); maximum iterations = 10,000. Residuals for each gene were quantile normalized to a normal distribution before mapping. In total, we had 26,936 and 17,116 genes available for mapping with RNA-seq and ‘gexarray’, respectively.

eQTL mapping using linear mixed models. eQTLs were identified using a linear mixed model implemented in LIMIX24,25. eQTLs were mapped in cis, and included a window of 1 Mb around the start of the gene (as defined by Gencode version 19 annotations). We modelled the genotype as a fixed effect, with population structure included as random effect. Population structure was modelled with a kinship matrix, calculated as the dot product of the genotypes in trans for each cis window (realized relationship).

eQTL mapping was performed with the following datasets (Supplementary Table 4): (1) iPS cells, 166 donors (239 cell lines), RNA-seq data (hereafter the ‘main’ eQTL map); (2) iPS cells, 301 donors (711 cell lines), gexarray data. When multiple lines were available for a donor, the mean expression value of lines was used. In the array-based iPS cell dataset, which had the largest number of replicate lines available per donor (246 donors with multiple lines), we additionally mapped the eQTLs using two randomly drawn sets of individual lines per donor to assess the replicability of the iPS cell eQTLs (Extended Data Fig. 6f–h). In the main RNA-seq based map, we identified both primary and secondary eQTL effects. To identify secondary effects, we repeated the mapping with the genotypes of the lead eQTL variant included as a covariate in the model. Finally, an alternative version of the main eQTL map was generated using a pipeline matched with GTEX V6 eQTLs for quality control purposes (see Supplementary Information).

Multiple testing correction. For cis eQTLs (primary and secondary), to adjust for the number of permutations per sample line in each cis window 10,000 times, keeping everything else in the model constant. To derive an empirical $P$ value distribution, the test statistic of the most significant variant in each permutation round was stored. A region-wise adjusted $P$ value was derived from the proportion of permuted test statistics that were larger than the most significant observed test statistic in the region. These threshold $P$ values were further adjusted for genome-wide analysis using the Benjamini–Hochberg correction. A gene was considered an eGene if its genome-wide final Benjamini–Hochberg-adjusted $P$ value was less than 0.05, which is equivalent to FDR $= 0.05$. Identification of tissue-specific cis eQTLs. In this study, tissue-specific eQTLs were defined as eQTLs as any of the 44 tissues available by the GTeX Project24. Replication was tested at the level of individual eQTL variants between all pairs of tissues. For this analysis, we considered the full cis eQTL output of iPS cell eQTLs from HippSci and 44 tissues from GTEX (V6p results; 2,025 tissue pairs). The included tissues and cell lines are detailed in Supplementary Table 5a. For each discovery tissue, we tested for the replication of all lead eQTL effects (lead eQTL
Replication was defined as the query variant (original lead or proxy) having a nominal eQTL $P < 2.2 \times 10^{-9}$ for the same eGene (corresponding to $P \approx 0.01 / 45$, where 45 refers to the total number of tissues tested). A lenient threshold was chosen in order to rule out any evidence of replication. High-linkage-disequilibrium proxies for a lead were defined as having $r^2 > 0.8$ in the UK10K European reference panel and located in the same cis window. All available linkage-disequilibrium-proxies were tested for replication and the variant with the most significant eQTL $P$ value in the discovery tissue was stored as the proxy. If no linkage-disequilibrium-proxies could be tested, the cis variant with the most significant eQTL $P$ value in the discovery tissue overall was selected. Overall, the same lead variant was available to test in 95% of tissues across tissues (median of discovery tissues, with each discovery tissue represented as the median of tests across all replication tissues; 90% for iPS cell eQTLs). A high-linkage-disequilibrium proxy for the lead was tested 3.6% of the time (7.3% in iPS cells), while the best available cis variant was tested only 1.1% of the time (2.7% in iPS cells). Of the rare cases when the best available variant was tested, the selected variant was independent of the original eQTL effects ($r^2 < 0.1$) 0.2% of the time (1.2% in iPS cells), indicating that in the vast majority of cases, the same eQTL effect was tested for replication and the choice of variant is unlikely to have a marked effect on the results. SNPs in cis replication of the iPS cell eQTLs, including numbers of selected proxy variants, are provided in Supplementary Table 5b. If a gene was not tested for eQTLs in the query tissue (for example, because the gene was not expressed), replication information was defined as missing. A replication profile was derived for each eGene in the discovery tissue, indicating whether the lead eQTL effect replicated (yes/no) or could not be tested (NA). We then extracted eGenes, for which the lead eQTL effect did not show evidence of replication in any other tissue ($P > 2.2 \times 10^{-9}$) or could not be tested (hereafter referred to as tissue-specific eQTLs). Of note, in this analysis, for an eQTL effect to replicate, it has to affect the expression of the same gene in both tissues. If the same variant is an eQTL for two different genes in the two tissues, it is not considered replicating. We also investigated the impact of the specific replication threshold, considering a threshold of $P < 0.01$ and $P < 0.05$ (Extended Data Fig. 8c). This analysis showed that the ability to replicate an eQTL signal in a second tissue is primarily determined by the sample size of the replication tissue and the specific choice of threshold. This dependency needs to be taken into account when interpreting tissue-specific eQTL effects.

As an alternative strategy for comparing eQTLs among different tissues, we also calculated the $r_s$ statistic ($r_s = 1 - \pi_1 / \pi_0$; ref. 76) for all pairs of tissues (Extended Data Fig. 8b) using the qvalue package in R. The $r_s$ statistic provides a global measure of similarity between a pair of tissues by estimating the proportion of eQTL signals discovered in one tissue that shows evidence of replication in a second tissue. For this analysis, for each discovery tissue, we queried all significant variants per eGene in the full cis output of all other tissues and estimated $q_s$ with the ‘bootstrap’ method of the qvalue package.

Functional annotation of eQTLs. Matched sets of variants. For all functional enrichment analyses, 100 matched sets of variants (hereafter control variants) were used as null variants. These sets were generated with SNPsnap15 using unique lead eQTL variants from different datasets as the input. Variants were matched for minor allele frequency (MAF) and number of the Gene (MAF > 0.01 and number of the Gene > 5). The high-linkage-disequilibrium proxies (LD buddies, $r^2 > 0.5$) distance to the nearest gene, and gene density, allowing for maximum deviation of ±50% for each criterion. HLA SNPs (defined as falling between positions 25,000,000 and 35,000,000 on chromosome 6) were excluded from the analysis and all matched sets were non-overlapping with the input variants.

The 1000 Genomes Phase 3 European population was used as the genotype reference panel. Both target and control sets of variants were further expanded with their high-linkage-disequilibrium proxies ($r^2 > 0.8$) derived from the European UK10K reference panel. These expanded sets form the basis for all subsequent enrichment analyses.

Overlap with chromatin state annotations. Functional enrichment of eQTLs was assessed using chromatin state data from the Roadmap Epigenomics Project26. The data comprised 25 chromatin states derived from reference epigenomes from 127 cell types. We tested the overlap between target eQTL variants and each chromatin state and cell type separately. We also tested the overlap with 100 sets of control variants with the same annotations and derived an empirical $P$ value for enrichment for each annotation. This was defined as the number of control variant sets ($n$) that showed a higher overlap with the target annotation than the eQTL lead variants ($P = n / 100$). The empirical $P$ values were further adjusted for the number of tests (25 states × 127 cell types) within each eQTL set using the qvalue26 package in R. Annotations with a $P < 0.05$ were considered significantly enriched. We tested two sets of eQTLs for enrichment: iPS-cell-specific eQTLs ($n = 2,131$) and non-iPS-cell-specific eQTLs ($n = 4,500$).

For visualization purposes we aggregated the 127 cell types into five clusters, using k-means clustering. The means of clusters was chosen on the basis of the number of different sample types annotated by Roadmap (primary cell, primary culture, primary tissue, cell line, ES-cell-derived). A heatmap of the difference in fold enrichment between iPS-cell-specific and non-specific eQTLs ($DIF = FE_{specific} - FE_{non-specific}$) was generated with the heatmap package in R (Fig. 4d) to assess how well our definition of iPS-cell-specific enriches for functional elements active in stem cells.

Overlap with transcription factor binding sites. Functional enrichment of eQTLs was additionally assessed using ChiP–seq-based transcription factor binding sites from the ENCODE Project26. Specifically, we used a set of proximal and distal transcription factor binding sites, where proximal is defined as a 2,000 bp window centred on transcription start sites annotated with Gencode v19 and distal as everything outside this window. Binding sites were defined as ChiP–seq peaks for all available transcription factors and cell types that overlapped DNase peaks in the respective cell type. We limited our analysis to binding sites present in H1 HES cells. With chromatin state annotations, we overlapped target and control eQTL variants with binding sites for each individual factor (all sites where the given factor is bound or co-bound). An empirical $P$ value for the enrichment was derived on the basis of the control sets and adjusted for multiple testing using the Benjamini–Hochberg method. Factors with adjusted $P < 0.05$ were considered enriched. Factors with at least 10 observed overlaps are plotted in Fig. 4e.

Overlap with the GWAS catalogue. The NHGRI–EBI GWAS catalogue was downloaded on 18 April 2016 (released 10 April 2016). Entries with missing positional or $P$ value information were removed and the positions of the remaining entries were converted to hg19 with the UCSC LiftOver function in the rtracklayer R/Biocoductor package29 and the ‘hg38ToHg19’ chain file. This resulted in 24,861 catalogue entries, corresponding to 18,446 unique variants of which 6,681 were significantly associated with a trait ($P < 5 \times 10^{-8}$). However, many studies included in the complete catalogue are not reliable, so we parsed the sample size information in the catalogue and excluded studies that did not report effect sizes (odds ratio or regression coefficient), had a sample size below 1,000, or assayed fewer than 100,000 variants. Then, following the approach from ref. 79, we further filtered the set of remaining associations to retain only traits that had at least six significantly associated variants, and kept all associations with $P < 1 \times 10^{-4}$ for these traits. This approach yielded a filtered set of 9,562 associations for 6,059 variants with 358 diseases and traits. Each variant position was then parsed for trait overlap (all traits associated to the given variant). If a variant–trait association was reported multiple times (for example, by different studies), the most significant association was kept.

We first tested whether eQTLs in iPS cells and somatic tissues (lead variants and their high-linkage-disequilibrium proxies) showed global enrichment in this final set of disease-associated variants compared with matched sets of variants (Extended Data Fig. 9b). Fold enrichments were derived from a comparison to 100 sets of matched controls (mean of control overlaps per tissue). For iPS cells, we additionally tested enrichment for individual traits, deriving an empirical $P$ value. Traits with an adjusted $P < 0.05$ (Benjamini–Hochberg) and a minimum of five observed overlaps were considered enriched (Supplementary Table 6c).

Lastly, we parsed each disease variant position for overlap with iPS cell eQTLs, again considering lead variants and their high-linkage-disequilibrium proxies for each tissue. We report all disease variants that were tagged by iPS cell eQTLs (lead variants and their high-linkage-disequilibrium proxies) showed global enrichment in this analysis for each disease variant position for overlap with iPS cell eQTLs, again considering lead variants and their high-linkage-disequilibrium proxies for each tissue. We report all disease variants that were tagged by iPS cell eQTLs (lead variants and their high-linkage-disequilibrium proxies). We note that the selected proxy may differ across the replication tissues. To account for this, if the original lead variant would have been tested across all tissues. However, because of differences in genotyping methods and allele-frequency due to sample size, it was not always possible to query the exact same ePair in all tissues. To account for this, if the original lead variant was not available to test in the query tissue, a proxy variant was tested instead. We note that the selected proxy may differ across the replication tissues. The approach to define the proxy variant is summarized in the Supplementary Information.

Information.
association signals over a model with two distinct causal variants. We excluded all colocalization results from the MHC region (GRCh37: 6:28477897–33448354), because these could exhibit an elevated false positives rate due to the complicated linkage-disequilibrium patterns in this region. We kept only results where the minimal GWAS $P$ value was $<5 \times 10^{-8}$.

**Overlap with the COSMIC genes.** To assess the overlap of iPS cell eQTLs with known cancer genes in the context of somatic and cancerous tissues, we calculated the cumulative number of cancer genes (COSMIC cancer census 27 April 2016, $p_{min} = 0.1$ (ref. 20)) regulated by eQTLs in iPS cells, somatic tissues (GTEx V6p), and three different cancers (ER positive and negative breast cancer, colorectal cancer) 52,53 (Extended Data Fig. 9a).

**Splicing of TERT.** Alternative splicing of the TERT gene was analysed using Leafcutter 92, which focuses on introns and quantifies both known and novel alternative splicing events by quantifying reads mapping to exon–exon junctions. Annotation results were derived from Gencode v19. Introns supported by fewer than 30 reads (n30; default) across all samples were removed. We obtained quantifications for eight intron clusters within TERT. After removing individual introns with a mean intron usage of zero, we had a total of 22 introns to test. We used the intron exclusion proportions to assess genotype-dependent effect of rs10080900 on TERT splicing (linear model between genotype and excision proportion, Bonferroni correction of $P$ values for the total number of introns tested). One intron showed evidence of a splicing-QTL effect ($P < 0.05$, Bonferroni adjusted; Extended Data Fig. 10).

**Code availability.** Scripts that were used to process the raw data and for implementing the statistical analyses presented are available from https://github.com/hipscica Nature2017.

**Data availability.** The assay data used in this publication are listed in the BioStudies archive (https://www.ebi.ac.uk/biostudies/studies) with accession identifier S-BXS55. All data can be accessed via the HipSci data portal (http://www.hipscici.org), which references to the EMBL-EBI archives that are used to store the HipSci data. Managed access data from all assays are accessible in the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) under the study EGAS00001001465. Open access genotyping array data and RNA-seq data are available in the European Nucleotide Archive (ENA) under the studies PRJEB11752 and PRJEB7388. Open access gene expression array data are available in the ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-4057. The mass spectrometry proteomics data have been deposited in ProteomeXchange (http://proteomcentral.proteomexchange.org/cgi/GetDataset) via the PRIDE repository with the dataset identifiers PXD003903 and PXD005506. Image data for this study is available in the Image Data Resource at http://dx.doi.org/10.17867/10000107. Data types from specialized assays for which none of the existing archives are appropriate are available from the HipSci FTP site (ftp://ftp.hipscici.ebi.ac.uk/vol1/ftp). Intermediate result files for this study, such as processed gene expression levels, can be found at: ftp://ftp.hipscici.ebi.ac.uk/vol1/ftp/data. For full details see Supplementary Information. Details of the data generated during the project, including archive accession identifiers for obtaining the data, are described in the Supplementary Information. The HipSci website (http://www.hipscici.org) also has full details of all publicly available data and instructions for researchers to apply for access to data in the European Genome-phenome Archive (EGA).

51. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676 (2006).
52. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature 526, 82–90 (2015).
53. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015).
54. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
55. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. Nat. Methods 9, 179–181 (2011).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transformer. Bioinformatics 25, 1754–1760 (2009).
57. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).
58. Huber, V., von Heydebreck, A., Siltmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 (Suppl 1), S96–S104 (2002).
59. Valleri, L. et al. Early cell fate decisions of human embryonic stem cells and mouse embryonic stem cells are controlled by the same signalling pathways. PLoS ONE 4, e6062 (2009).
60. Ly, T. et al. A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. eLife 3, e01630 (2014).
61. Bensaddek, D. et al. Micro-proteomics with iterative data analysis: proteome analysis in C. elegans at the single worm stage. Proteomics 16, 381–392 (2016).
62. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372 (2008).
63. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Illumina Methylation microarrays. Bioinformatics 30, 1363–1369 (2014).
64. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
65. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
66. Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).
67. Deluca, D. S. et al. RNA-Seq QC metrics for quality control and process optimization. Bioinformatics 28, 1530–1532 (2012).
68. Leha, A. et al. A high-content platform to characterise human induced pluripotent stem cell lines. Methods 96, 85–96 (2016).
69. Zick, T. I. et al. Instructions for researchers to apply for access to data in the European Genome-phenome Archive (EGA).
Extended Data Figure 1 | Overview of the cellomics assay. a, Example plate layout for the cellular differentiation assay. Images are shown for the pluripotency markers (OCT4, SOX2 and NANOG) as they are measured in the Cellomics imaging device. Each line is measured in two rows of the same plate as technical replicates. The secondary antibody used for each marker is shown in parenthesis (aG, anti-goat antibody; aM, anti-mouse antibody). Each plate also has measurements for staining with the secondary antibody only, which serves as a means to assess background fluorescence. The red channel shows the signal from the DAPI staining, the green channel the marker signal. As expected, there is only a small signal from the green channel in the wells stained for the secondary antibody only. Image acquisition stops as soon as 10,000 cells have been detected. b, Detailed variance components of the cellomics markers (Methods). Substantial proportions of the marker variance could be attributed to batch factors, including staining, technician effects and antibody lots. These effects mean that the fraction of cells expressing particular markers need to be interpreted with caution (Fig. 1c, d). c, Pairwise Pearson’s correlation (r) between quantitative expression scores derived from immunostaining for pluripotency and differentiation and the PluriTest score (P values from a Student’s t-test).
Extended Data Figure 2 | Pluritest scores in the two culture conditions.
a–c, Comparison of PluriTest novelty score versus pluripotency score for the 711 lines generated. Lines grown on feeder-free conditions (E8 medium) scored systematically lower than feeder-dependent lines \((P = 1.62 \times 10^{-43}; \text{t-test}, \text{for pluripotency score})\). We note that, while we cannot rule out that feeder-free lines are less pluripotent, feeder-free conditions are not well represented in the PluriTest training dataset, which may explain this result (of the 204 ES cell/iPS cell lines in the PluriTest paper that have medium metadata available, none were cultured in E8 and only 37 were cultured in a variety of other feeder-free formulations such as MTSER).
d, Despite lower pluripotency scores, lines grown on feeder-free conditions have higher fractions of cells expressing canonical protein markers of pluripotency.
Extended Data Figure 3 | Extended CNA analysis. Relationship between the number of CNAs, using three CNA minimum length thresholds for calling CNAs: 200 kb, 500 kb and 1,000 kb and other experimental factors. Values on the x axis have been 'jittered' (that is, small random 'noise' has been added to the true values) to enhance the visualization. Data points underlying the box plots are shown as semi-transparent blue dots.

a. Number of CNAs per line versus passage number. P values are shown from a generalized linear mixed model (Poisson regression) with donor random effect. b. Box plot of the number of autosomal CNAs per line versus growth medium. P values are shown from a Poisson regression for culture condition. c, d. Number of autosomal CNAs per line versus PluriTest pluripotency and novelty scores. P values are for a linear mixed model of the number of autosomal CNAs per line with a random effect for donor. e, f. Number of CNA counts per donor versus gender and donor age. CNA counts refer to the total number of unique CNAs across all lines derived from the same donor. CNAs that are shared between lines of the same donor (overlap by at least one base) are counted only once. P values are shown for a Poisson regression for either gender or age.
Extended Data Figure 4 | Location and consequence of the recurrent CNA on chr20. Related to Fig. 2. Top, genomic location versus number of lines with copy number 3 (grey) and with a CNA (black). Bottom, the NAV gene score from ref. 22 and log2 gene expression fold change between the iPS cell lines with copy number 2 and 3a (colour scale), in the region highlighted in red in the top panel. Highlighted genes are upregulated when copy number increases, known oncogenes/tumour-suppressor genes and/or genes with a NAV score in the top 2%.
Extended Data Figure 5 | See next page for caption.
Extended Data Figure 5 | Functional assessment of CNAs using growth assays. **a**–c, Cell growth rate (**a**), proliferation (**b**) and apoptosis (**c**) in cell lines with copy number 2 (wild type, blue dots) or copy number 3 (mutant, red dots) in a recurrently duplicated region in iPS cells on chromosome 1, 17 or 20. Plot titles show the donor name and the genomic coordinates of the CNA. **a**, Cell counts taken on successive days in culture, for pairs of lines (one mutant, one wild type) grown on the same 24-well plates are shown. Asterisks denote significance levels for statistical interactions between day and copy number in a linear mixed model, using fixed effects to fit day and copy number, and random effects to account for culture plate effects. EIF4A3 denotes whether a copy number variant overlaps one of the suspected candidate genes on chromosome 17. *P* < 0.05; **P** < 0.01; ***P** < 0.001. **b**, Protein expression level measured using TMT-based quantification using the Q-exactive plus (labelled QE Plus) orbitrap and a fusion (labelled Fusion) orbitrap mass spectrometry platforms. **c**, Estimated fraction of fluorescing nuclei following an EdU assay in mutant and wild-type lines, following exposure to mitomycin (Treated) or a control sample (Untreated). **d**, Estimated fraction of fluorescing nuclei following a terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL) assay in mutant and wild-type lines, following exposure to mitomycin (Treated) or a control sample (Untreated). Solid trend lines are least squares regression fits. *P* values in **b** and **c** denote the significance of statistical interactions between copy number and mitomycin treatment condition (treated or untreated).
Extended Data Figure 6 | Effect of passage number on tier 1 and tier 2 data and overview of iPS cell cis eQTLs mapped with tier 1 gene expression array data. a, b, Passage number versus PluriTest pluripotency and novelty scores shows no significant association between passage number and pluripotency. Trend lines are fit using a linear regression of the PluriTest scores and the passage number (score $P = 0.66$, novelty $P = 0.21$). Association was also not significant when including gender and medium as fixed effects and donor as random effects (score $P = 0.3$, novelty $P = 0.14$). c, Passage number versus log10 RNA-seq expression of pluripotency factors NANOG and OCT4 shows no significant association between passage number and pluripotency. Trend lines are fit using linear regression of log10 expression and passage number ($\text{NANOG, } P = 0.5$; $\text{POUSF1, } P = 0.15$). Association was also not significant when considering the two genes together and when including gender and medium as fixed effects and batch variables and donor as random effects (passage, $P = 0.28$; passage–gene interaction, $P = 0.96$). d, e, Variance component analysis for tier 2 assays, showing that for the majority of genes gender and passage explained little of the total variance. f, g, Comparison of lead variant effect sizes ($\beta^2$) in gexarray-based eQTL maps. The eQTL maps were derived by using mean expression levels per donor (‘main’ map) and with two sets of individual lines (one per donor), drawn randomly (‘replicate’ maps). The effect sizes for all tested genes are shown in black, with FDR < 5% eGenes from the main map indicated in blue. Effect sizes are compared between the two replicate maps (f, Spearman’s rank correlation $\rho = 0.47$ genome-wide, $\rho = 0.80$ eGenes only, both $P < 2.2 \times 10^{-16}$) and between the main map and one replicate map (g, Spearman’s rank correlation $\rho = 0.57$ genome-wide, $\rho = 0.88$ eGenes only, both $P < 2.2 \times 10^{-16}$). The effect sizes obtained using the mean expression values per donor are higher than when using individual lines. h, Pairwise correlation between gene expression levels in iPS cells measured with RNA-seq and gexarray. The Spearman rank correlation coefficients of either gene (pink) or gexarray probe (blue) region based read counts are shown, showing a higher correlation for probe-based counts.
Extended Data Figure 7 | Properties of iPS cell cis eQTLs in comparison to somatic eQTLs. 

a, b, The power to detect eQTLs is plotted, comparing 44 somatic tissues from GTEx24 (V6p) and the HipSci RNA-seq-based eQTL map (purple triangle), considering either the absolute (a) or relative (b) number of eQTLs identified (eGenes, FDR < 5%). The major determinant of eQTL detection power is sample size.

c, Cumulative fraction of RNA-seq reads relative to the number of protein-coding genes expressed. The mean read count derived from 20 iPS cell lines (10 donors, two lines each) is plotted, five fibroblast lines, and two ES cell (ESC) lines. In iPS cells, half of the reads are explained by the expression of 1,071 genes, whereas 75% and 90% of the reads are explained by the expression of 3,159 and 5,814 genes, respectively (total protein-coding genes with non-zero counts n = 17,332).

d, Distribution of iPS cell eQTLs around the annotated gene start position. The $-\log_{10}$ (eQTL P value) is plotted against the distance (in bp) from the gene start for lead eQTL variants genome-wide, highlighting significant eQTLs (FDR < 5%) in orange.

e, Comparison of the magnitude of eQTL effect size (absolute beta; left) and minor allele frequency (MAF; right) between iPS-cell-specific (n = 2,131; labelled as S) and non-specific eQTLs (n = 4,500; labelled as NS), demonstrating that overall, iPS-cell-specific eQTLs have smaller effects on the transcriptome than eQTLs shared among multiple tissues ($P = 9.97 \times 10^{-161}$, Wilcoxon rank-sum test) and have a lower minor allele frequency ($P = 1.08 \times 10^{-35}$, Wilcoxon rank-sum test).
Extended Data Figure 8 | Comparison of eQTL mapping pipelines between HipSci and GTEx (V6p). a, Proportion of tissue-specific eQTLs as a function of the discovery sample size. For iPSC cells, the two sets of tissue-specific eQTLs obtained with the two different mapping pipelines (Methods) are shown, namely the standard HipSci pipeline (iPSC; purple triangle) and the alternative ‘GTEx-like’ pipeline (iPSC2; purple triangle). Points other than iPSC cells are from the GTEx Consortium (44 somatic tissues and cell lines)24. b, Heat map of pairwise $\pi_1$ values ($\pi_1 = 1 - \pi_0$) between iPSC cells and GTEx tissues, with rows representing the discovery tissue and columns the replication tissue. Clustering of tissues is based on euclidean distance (R hclust, method = average). c, Effect of eQTL replication threshold on the definition of tissue-specific effects. The replication profile of iPSC cell eQTLs across GTEx tissues relative to discovery sample size in each replication tissue is shown. The proportion of lead eQTLs from iPSC cells that replicate in each tissue is plotted, with replication defined using two different replication thresholds (TH1: nominal eQTL $P < 0.01 / n_{tissue}$; TH5: $P < 0.05 / n_{tissue}$; plotted as dots and triangles, respectively). d, Enrichment of alternative iPSC cell eQTLs (GTEx-like) at proximal and distal (defined as less than or greater than 2 kb from the transcription start site) transcription factor binding sites of promoters in H1 hES cells from the ENCODE Project50. Fold enrichments per factor are shown for iPSC-cell-specific and non-specific eQTLs (minimum 10 observed overlaps) (Methods). Pluripotency-associated factors are indicated with an asterisk. The profile of enrichments is comparable to that obtained with the standard HipSci pipeline (Fig. 4d).
Extended Data Figure 9  |  iPS cell eQTLs and disease. a, Cumulative number of cancer genes (COSMIC cancer census 27 April 2016; \(n_{\text{genes}} = 571\); ref 20) regulated by eQTLs in iPS cells, somatic tissues (GTEx V6p), and three different cancers (ER positive and negative breast cancer, colorectal cancer) 32,33. b, Enrichment of iPS cells and somatic eQTLs (lead variants and their high-linkage-disequilibrium proxies) at disease-associated variants in the NHGRI–EBI GWAS catalogue (10 April 2016). The fold enrichment of eQTLs over 100 random sets of matched variants for each tissue relative to eQTL discovery sample size is shown. The tissues showing the highest fold enrichment are liver and brain (cerebellar hemisphere; BrainCH). c, Somatic eQTL signal for the PTPN2 (protein tyrosine phosphatase, non-receptor type 2) locus on chromosome 18. This locus contains a colocalizing association signal for PTPN2 gene expression in iPS cells and five immunological disease phenotypes (Fig. 5a). d, Somatic eQTL signal for the TERT (telomerase reverse transcriptase) locus on chromosome 5 (Fig. 5b). In both c and d, the lead eQTL variant locations are indicated with red and orange vertical lines for iPS cells and somatic tissues, respectively. The focal gene regions are indicated in solid grey and gene start positions of other protein-coding genes on the same strand with vertical grey lines.
Extended Data Figure 10 | Tissue expression and alternative splicing results for the TERT locus. a, b, Normalized RNA-seq per-base coverage across the TERT locus stratified by rs10069690 genotype. The full locus (a) or zoomed view of the region (b) around the lead eQTL and cancer risk variant rs10069690 are shown. rs10069690 is indicated with a dotted line on each plot. Grey regions indicate annotated exons from Ensembl version 75. Coverage was computed from indexed BAM files using the coverageBed function from the bedtools (version 2.25.0)\(^93\).

Raw coverage was divided by total library size in millions (total number of mapped reads) per sample to obtain normalized coverage, which was then averaged over samples with the same rs10069690 genotype to obtain mean normalized coverage for each genotype group. c, Profile of TERT expression in iPS cells and across somatic tissues from GTEx. The gene FPKM values obtained with RNA-SeQC (GTEx V6p) are shown.

d, Splicing-QTL of TERT. We quantified TERT intron retention rates using Leafcutter\(^92\) and identified one alternative splicing event associated with rs10069690, the lead iPS cell eQTL variant for TERT (Fig. 5b). The TERT intron 4 retention ratio (PSI, per cent spliced in) is shown in iPS cell lines of all individual donors stratified by their genotype at rs10069690. This variant affects the splicing of the intron where it is located, with the minor allele (T) increasing the fraction of TERT transcripts in which intron 4 is retained ($P = 1.7 \times 10^{-9}$, Bonferroni-adjusted linear regression).
Corrigendum: Common genetic variation drives molecular heterogeneity in human iPSCs

Helena Kilpinen*, Angela Goncalves*, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, Dalila Bensaddek, Francesco Paolo Casale, Oliver J. Culley, Petr Danecek, Adam Faulconbridge, Peter W. Harrison, Annie Kathuria, Davis McCarthy, Shane A. McCarthy, Ruta Meleckyte, Yasin Memari, Nathalie Moens, Filipa Soares, Alice Mann, Ian Streeter, Chukwuma A. Agu, Alex Alderton, Rachel Nelson, Sarah Harper, Minal Patel, Alistair White, Sharad R. Patel, Laura Clarke, Reena Halai, Christopher M. Kirton, Anja Kolb–Kokocinski, Philip Beales, Ewan Birney, Davide Danovi, Angus I. Lamond, Willem H. Ouwehand, Ludovic Vallier, Fiona M. Watt§, Richard Durbin§, Oliver Stegle*§ & Daniel J. Gaffney*§

*These authors contributed equally to this work.
§These authors jointly supervised this work.

In this Article, the authors Fiona M. Watt and Richard Durbin should also have been included as 'jointly supervising' authors, and authors Oliver Stegle and Daniel J. Gaffney should also have been noted as 'equally contributing' authors. In addition, the Author Contributions section should have included the sentence: 'H.K. and A.G. contributed equally to this work; O.S. and D.J.G. contributed equally to this work', as further clarification. The original Article has been corrected online.