nature biotechnology

Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments

Quin F Wills[1], Kenneth J Livak[2], Alex J Tipping[3], Tariq Enver[3], Andrew J Goldson[4], Darren W Sexton[5] & Chris Holmes[1,6–8]

Gene expression in multiple individual cells from a tissue or culture sample varies according to cell-cycle, genetic, epigenetic and stochastic differences between the cells. However, single-cell differences have been largely neglected in the analysis of the functional consequences of genetic variation. Here we measure the expression of 92 genes affected by Wnt signaling in 1,440 single cells from 15 individuals to associate single-nucleotide polymorphisms (SNPs) with gene-expression phenotypes, while accounting for stochastic and cell-cycle differences between cells. We provide evidence that many heritable variations in gene function—such as burst size, burst frequency, cell cycle–specific expression and expression correlation/noise between cells—are masked when expression is averaged over many cells. Our results demonstrate how single-cell analyses provide insights into the mechanistic and network effects of genetic variability, with improved statistical power to model these effects on gene expression.

Human, clinical, genome-wide association studies (GWAS) have been used to correlate genetic variants with disease and pharmacogenomic traits, typically in a hypothesis-free manner. To investigate the mechanisms underpinning these statistical associations, molecular phenotyping technologies have also been used to associate traits such as gene expression with genetic variants. Progress has been slow, with tissue specificity, the low resolution of DNA genotypes and the technical challenges of assaying molecular traits all thought to be important limiting factors[1].

The most commonly studied molecular trait is baseline, whole-tissue gene expression. In these studies, genetic variants that are associated with variation in gene expression are referred to as expression quantitative trait loci (eQTLs)[2]. A trait with 'smooth' variation between individuals due to many small contributing factors is treated as being quantitative (continuous) in epidemiological models. However, this is a simplistic treatment of gene expression as a trait, as its genetic perturbation need not result in the smooth variation of expression. This is most commonly seen when cancer somatic variations result in pronounced transcriptomic changes. Theoretical discussions around concepts such as self-organizing criticality[3] have proven popular as attempts at explaining this observed complexity of gene expression and its regulation. The same lack of 'smoothness' can be seen with the environmental perturbation of gene expression—as recently demonstrated with lipopolysaccharide-exposed immune cells[4]—and is likely to become an important theme in the pharmacogenomics of gene function.

Although a number of disease-associated genetic variants have been shown to be linked to eQTLs[5], in most cases no such correlation exists. We hypothesize that instead of influencing the average gene expression of a gene in a whole organism or a specific cell type or tissue, these variants might change the cell-to-cell variability, temporal dynamics or cell cycle dependence of gene expression at the single-cell level.

Here we explore whether studying individual cells can begin to provide greater mechanistic insights into how SNPs quantitatively affect gene function, as opposed to just assaying their effects on average tissue expression. We refer to these variants as single-cell quantitative trait loci (scQTLs).

To demonstrate the importance of cell-to-cell variability, we measured gene expression of selected genes in fresh, naive B lymphocytes from three individuals. Gene expression typically had much greater variability between cells within an individual than between individuals, and the distribution of gene expression values is very different between individuals for some genes. (**Fig. 1a**). The currently understood reasons for this large cell-to-cell noise are thermodynamic, regulatory and cellular (**Supplementary Fig. 1**).

As a basis for studying the association of single-cell phenotypes with genetic variants, we first sought to generate high-quality data in a large population of cells (1,440 cells). We measured gene expression using highly parallel qPCR validated with digital PCR, as single-cell RNA sequencing still faces notable technical challenges[6,7]. We focused on 92 genes affected by Wnt signaling, a major regulator of the cell cycle that has been highlighted as a key pathway in clinical GWAS and cancer epidemiology (**Supplementary Fig. 2**). Of the 92 genes studied, 46 are listed in the Catalog of Genome-Wide Association Studies (http://www.genome.gov/gwastudies). Wnt pathway genes

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Fluidigm Corporation, South San Francisco, California, USA. [3]Stem Cell Laboratory, UCL Cancer Institute, University College London, London, UK. [4]UEA Flow Cytometry Services, BioMedical Research Centre, School of Biological Sciences, University of East Anglia, Norwich, UK. [5]BioMedical Research Centre, Norwich Medical School, University of East Anglia, Norwich, UK. [6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [7]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [8]Medical Research Council Harwell, Harwell Science and Innovation Campus, UK. Correspondence should be addressed to Q.F.W. (wills@stats.ox.ac.uk).

Received 8 February; accepted 14 June; published online 21 July 2013; doi:10.1038/nbt.2642

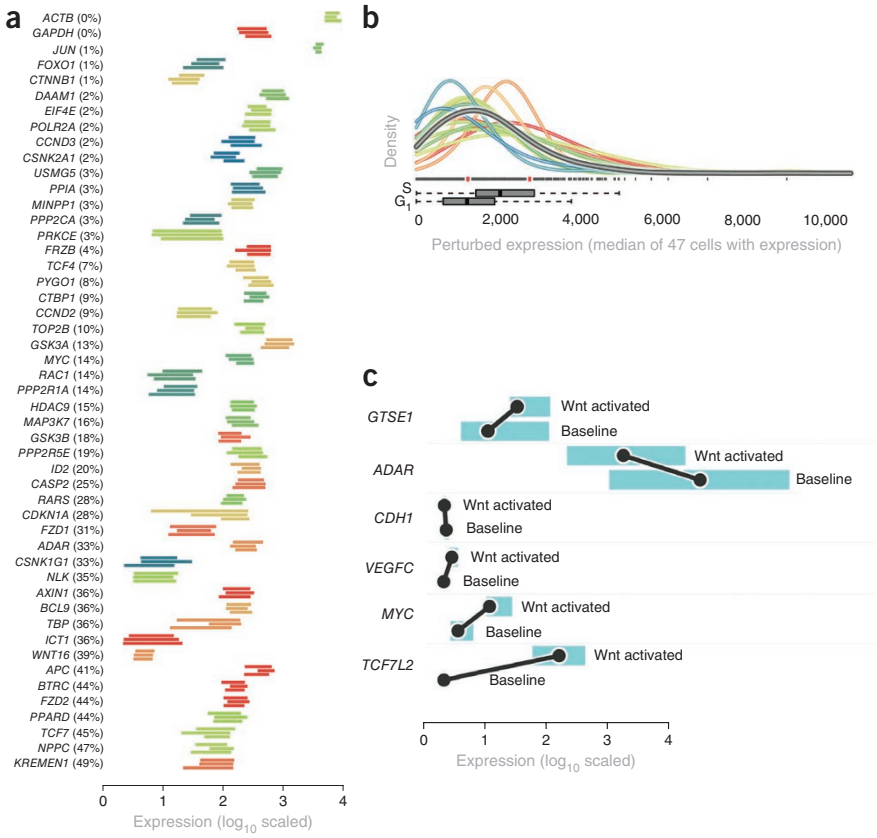

Figure 1 Single-cell gene expression distributions. (**a**) Wnt pathway genes from naive B lymphocytes in G_0 phase of their cell cycle were assayed in three human donors. Genes with expression in at least 50% of the cells are shown (the percentage of cells without detectable expression are shown in parentheses). The box plot provides the expression interquartile ranges, with the values for the three human donors per gene plotted together. Each of the genes show greater variability between cells than between individuals, with a level of 'noise' that is different for each gene. (**b**) The expression distributions of *PPP2R1A* are shown in 15 cell lymphoblast cell lines perturbed with a GSK3 inhibitor. Each cell line generated a median of 47 cells with *PPP2R1A* expression. Each curve represents a kernel density estimate of the distribution of the *PPP2R1A* gene expression of a single sample, with the gray curve providing the combined distribution of all samples. The primary data, shown as tick marks below the distributions, give the expression values for individual cells, and the two red points are the mean estimates for a mixture of two Poisson distributions. If the skewed distribution represents mostly promoter switching (**Supplementary Fig. 3**), the Poisson means and mixing proportion can be used as markers of gene burst size and frequency. The box plots provide the expression distributions for cells in G_1 versus early S phase. (**c**) Genes from 15 lymphoblastoid cell lines with statistically significant differential expression (when chemically perturbed) are compared. The black lines show change in median expression and the blue bars provide the inter-quartile expression ranges across cells. From these examples it can be seen that genes change not only whole-tissue expression but also their expression noise.

thus together form an ideal model for studying a clinically relevant system where we expect to be able to identify genetic drivers of expression behavior. In addition to Wnt system genes, classical reference genes such as *GAPDH* were assayed and found to demonstrate substantial variability in expression between cells. This makes it impossible to use such genes for traditional data normalization, and has been discussed elsewhere[8].

HapMap lymphoblastoid cell lines derived from 15 unrelated individuals of European descent[9] were perturbed for 24 h with 10 μM SB216763, a Wnt pathway agonist that inhibits GSK3 (ref. 10). Forty-eight cells in each of 30 samples (15 baseline and 15 perturbed cell lines) were assayed using a combination of flow cytometry and microfluidic gene expression chips[8] (Online Methods).

We observed variability both in the average gene expression values as well as distribution of expression values between the 15 individuals in both the unperturbed and inhibitor-treated cell populations (**Fig. 1b**,**c**). A number of parameters can be deduced from these data and the correlation of these to genetic variants tested.

Considering gene expression phenotypes in terms of single-cell distributions provides important information about gene regulation, as the noise from the regulation of transcription can be considered separately from the noise of RNA turnover (**Supplementary Fig. 1**). Thus constitutively expressed genes, for example, are expected to be less noisy, demonstrating mostly the thermodynamic noise of RNA turnover in the absence of variable regulation (**Supplementary Fig. 1**). Analyzing four, public, single-cell RNA sequencing data sets[6,7], we found that this is indeed the case (**Supplementary Notes**, section 1). A gene that is not constitutively expressed can be described in terms of how often it switches 'on' (burst frequency), the amount of RNA produced when 'on' (burst size) and the rate at which its RNA is degraded. A recent study[11] of 8,000 human loci found that almost all of them exhibited such 'bursty' expression, with certain loci modulating burst frequency and others modulating burst size.

Our data suggest that the genes in this study differ from each other in terms of burst size (**Fig. 2a**). Whereas an increase in both burst size and frequency elevates mean gene expression, an increase in only burst size raises the expression variance between cells relatively more than the expression mean. This is evidenced by an increase in the expression Fano factor (variance/mean). As genetic variants are expected to affect these dynamics, we propose that it is important for eQTL studies to begin considering expression dynamics in terms of parameters from models describing cell-to-cell variability. Gene expression between cells has been described as being log-Normal or Gamma distributed[12,13]. Although the log-Normal model maps to the standard Gaussian distribution, an advantage of the Gamma model is that its parameters relate directly to gene burst frequency and size (**Supplementary Notes**, section 2). We suggest that a more complete model should describe the Poisson-distributed thermodynamic contribution to a gene's noise and its mixing owing to gene bursting (**Supplementary Figs. 3** and **4**). Gene expression could, thus, be modeled as overdispersed Poisson noise (variance greater than would be expected with Poisson noise). For this, one suitable distribution is the negative binomial, which is equivalent to a Gamma-Poisson continuous mixture model where the Gamma distribution parameters of κ (shape) and ϕ (scale) increase with burst frequency and size, respectively. For this work, a discrete three-parameter Poisson

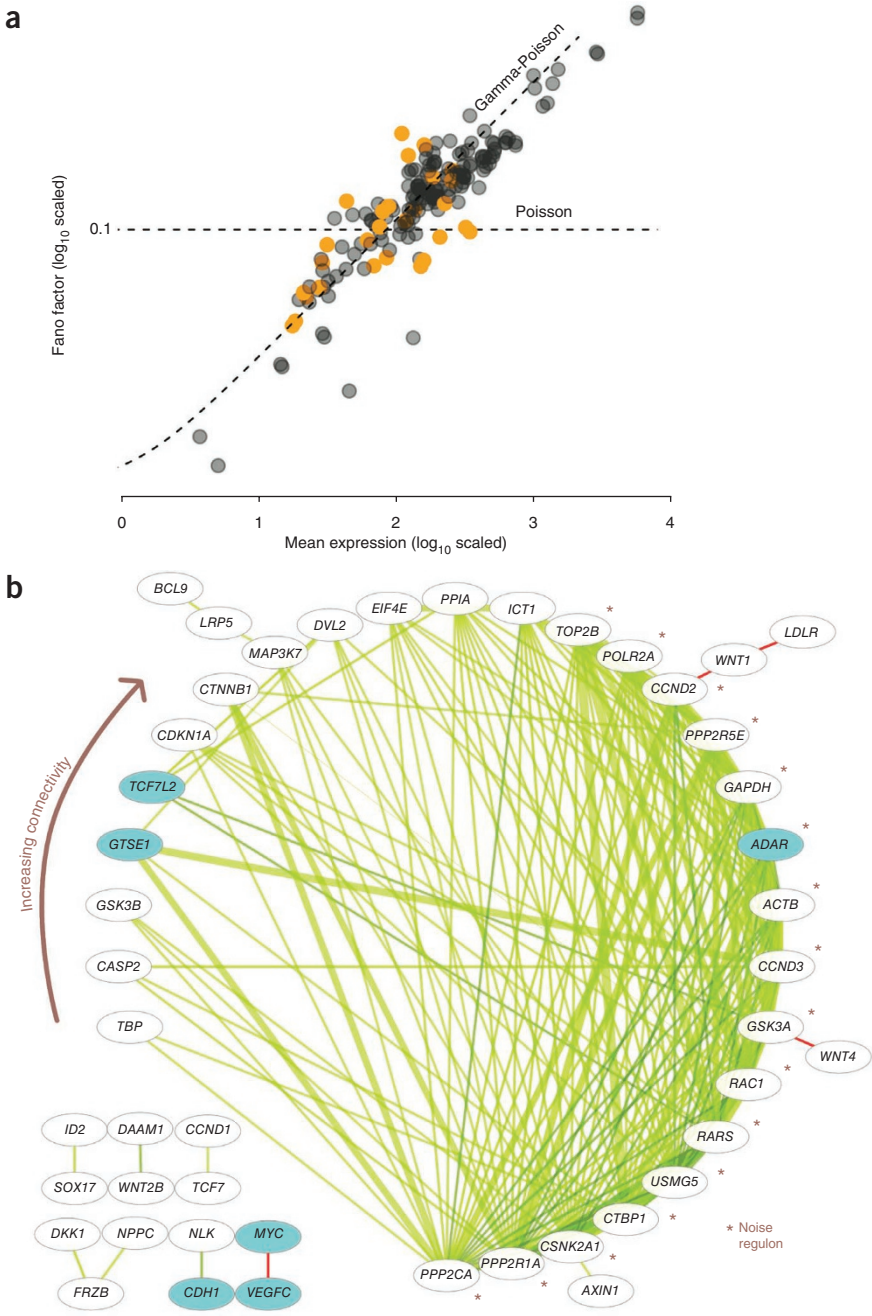

Figure 2 Properties of gene expression noise. (**a**) Fano factors (variance/mean) of the study's genes are plotted against their mean expression. Orange points are genes better fitted with Poisson than overdispersed Poisson distributions (Online Methods). The Fano factors of highly expressed genes are proportionally higher, in keeping with an overdispersed Gamma-Poisson model of gene expression (**Supplementary Fig. 4**). The Fano factor for a Gamma-Poisson model increases by $\phi + 1$, where ϕ is the scale parameter of the Gamma distribution, which increases with gene burst size. If genes differ by burst frequency, rather than burst size, the data would scatter along a line parallel to the Poisson line. (**b**) Baseline cell-to-cell expression correlations of $|\rho| > 0.5$ are shown for ~200 cells from sample GM10860. Genes are ordered clockwise according to increasing number of correlations (network connectivity). Red edges are negative correlations, whereas green edges are positive correlations ($\rho > 0.7$ are dark green). Correlations that increase with perturbation are plotted with bold lines. The right-hand side cluster is a hub of highly connected genes that tend to increase their correlations with perturbation. We refer to this as a 'noise regulon'. Notably, the regulon correlations vary without detectable change in mean expression of any genes except *ADAR*.

mixture was used as an approximation to the slower fitting, four-parameter, Beta-Poisson mixture (**Supplementary Fig. 4**). Such a Poisson mixture model not only describes the long-tailed behavior described by the negative binomial, but allows for the expected expression bimodality in genes with low burst frequency. These models and their rationales are further detailed in the Online Methods and **Supplementary Notes**, section 2.

Gene expression distributions can also be described in terms of heterogeneous cell subpopulations. We considered cells in different stages of the cell cycle, and their varying proportions between samples. Using flow cytometry, we excluded cells with increased DNA content, as would be expected in the S and G_2 cell cycle phases. We further subdivided cells into G_1 and early S-phase based on their expression of *GTSE1*, a cell division molecular switch that becomes highly expressed in the S and G_2 phases[14]. Almost two-thirds of the genes demonstrated altered expression between G_1 and early S-phase (*PPP2R1A* is shown in **Fig. 1b**), raising the question of how many associations are driven by differences in cell cycle subpopulation proportions between samples. Per cell culture, the proportion of cells without increased DNA content was found to be significantly anti-correlated with cell density after 48 h of growth (Spearman's ρ, $-\log_{10}P = 5.00$). We used this as a marker of cell line growth, which is a known confounder[15,16] and was hence adjusted for in the SNP associations. Growth was noted to be sample specific, batch specific, and correlated with the expression and noise of several genes (**Supplementary Fig. 5**). Counts of active Epstein-Barr virus replication—the agent used to immortalize the cells—were assessed but did not correlate with growth.

In addition to considering expression distributions and cell cycle subpopulations, single-cell gene expression can be used to generate an expression network per sample. This is in contrast to most systems biology approaches in human genetic epidemiology that describe a network of all samples combined, rather than the more explicit comparison of network parameters between samples. Gene expression noise can thus be treated as a third form of perturbation, together with genetic and chemical perturbation. **Figure 2b** plots an example of cell-to-cell gene correlations in ~200 cells from one of the lymphoblastoid cell lines, for genes detected in at least 50% of the cells. This network provides an example of expression behavior detected only at the single-cell resolution: the correlated and anti-correlated expression between cells. One can define 'noise regulons', that is, groups of genes co-regulated within this expression heterogeneity, whose Spearman correlations alter

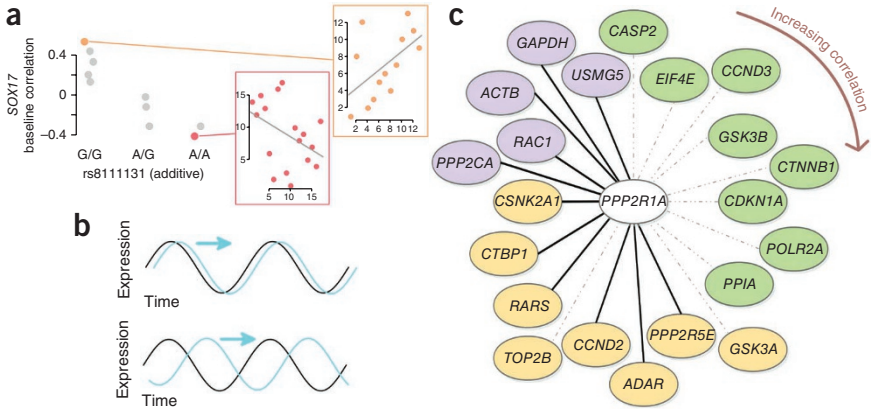

Figure 3 The heritability of single-cell expression. As shown in **Figure 1b**, the single-cell expression distribution of *PPP2R1A* is highly variable between individuals, suggesting heritable drivers. (**a**) *PPP2R1A* has a nominally significant additive SNP association with its baseline *SOX17* correlation. The left-hand side plot shows each individual's genotype versus their *PPP2R1A-SOX17* correlation. The right-hand side plots show the correlations for two selected points. In both plots, the *x*-axis is the ranked *PPP2R1A* expression and the *y*-axis is the ranked *SOX17* expression. (**b**) Single-cell correlations are a useful phenotype to study heritable gene relationships, but more complete interpretation requires temporal studies. Gene expression correlations have time lags and, as shown for two theoretical genes, this may be small (top plot) or large (bottom plot). The altered *SOX17* correlation in **a** may simply represent a change in the time lag with the putative altered *PPP2R1A* transcription rate by rs8111131. (**c**) The mean perturbed correlations of 21 genes in T allele homozygotes for *PPP2R1A* are ordered clockwise by increasing Spearman correlation. Those genes with $\rho > 0.5$ are in green, those with $\rho > 0.6$ are in orange, whereas those with $\rho > 0.7$ are in violet. All correlations decrease in individuals with a C allele of rs9304726, with those genes connected by dashed lines dropping below a mean correlation of 0.5.

with chemical perturbation, but without detectable change in mean (whole-tissue) expression.

For the SNP association testing we considered how such noisy correlations (co-expression of different genes between genetically homogeneous cells) of genes vary between individuals. We also considered how the network connectivity of each gene varies (the number of genes it correlates with).

Overall, for each gene the following phenotypes were measured and associated with SNPs within 50 kb of the gene: (i) whole-tissue expression, measured as the mean value over all cells; (ii) expression heterogeneity/noise, measured as the gene's Fano factor; (iii) burst size, inferred from a discrete Poisson mixture; (iv) burst frequency, inferred from a discrete Poisson mixture; (v) individual Spearman correlation strengths with the five most correlated genes; (vi) network connectivity, measured as number of correlations of $|\rho| > 0.5$; (vii) G_1 and S-phase expression based on detectable *GTSE1* expression and (viii) the number of cells for which expression could not be detected.

The motivation for the latter phenotype is that gene expression may be too low to be reliably detected, and thus zero-inflated, as shown in **Supplementary Figures 6** and **7**. All phenotypes were analyzed in the baseline state, perturbed state and as the log ratio of the perturbed to baseline states. The top 374 SNP associations (above $-\log_{10}P = 3$) are tabled with their corresponding *P* values in **Supplementary Table 1**. To control for multiple testing, we considered and permuted only the most statistically significant association per phenotype. For each permutation, the maximum $-\log_{10}$P was used to generate a null distribution of most significant *P* values, and so provide a family-wise error correction. As most of the genes are affected by a single pathway (that is, are not independent), a conservative global significance threshold of $-\log_{10}P = 4$ was used. We found 47 significant associations (0.9% of all association tests), which are detailed in **Supplementary Table 2**, where seven genes found in clinical GWAS studies that are without previously known eQTLs are highlighted. When considering only whole-tissue expression (and reducing the multiple testing threshold to $-\log_{10}P = 3$), this resulted in an eightfold reduction in the number of hits to 6; suggesting that a large portion of SNP effects typically go undetected. It is interesting to note from the list of hits the over-representation of Wnt receptor genes that have altered correlations with downstream genes, adding to the validity of correlations as a phenotype in genetic epidemiology, as one would expect altered correlations to reflect signaling pathway structure.

We also noted that clinical GWAS genes demonstrate greater G_1 and early S-phase inter-individual variability compared with other genes ($-\log_{10}P = 4.17$ and $-\log_{10}P = 5.27$, Mann-Whitney testing of baseline expression), which is more significant than the observed variability at the whole-tissue level ($-\log_{10}P = 3.13$). At the systems level, clinical GWAS genes appear, also, to have greater interindividual variability of their network connectivities ($-\log_{10}P = 2.37$ for perturbed expression (**Supplementary Fig. 8**), $-\log_{10}P = 2.24$ for baseline expression). Although we describe genes of only a single pathway, we speculate that such results could be used to identify genes that are key modulators of pathology risk and prognosis. Using cross-validation for model selection, growth was included as a variable in most of the significant associations, demonstrating a larger effect size than the co-associated SNP for 36% of scQTLs. This pervasive association with growth is not surprising considering the strong cell cycle activity of these genes. If considered individually, the mean R^2 values for genotype and growth associations in the 47 hits suggest a statistical power of 5.6% and 2.9%, respectively. When modeled together this raises the study power substantially to 36%, using almost half the sample size that would be required to achieve this power if growth was not considered.

In addition to the cell cycle phase, noise and growth associations, the results allowed us to propose mechanistic and systems-level hypotheses not possible with whole-tissue associations. Using *PPP2R1A* as an example, our results suggest the interindividual variability shown in **Figure 1b** to be a result of a gene burst size associations with SNPs rs8111131 and rs8108607. These are listed together with other associations in **Supplementary Table 1**. The latter SNP is immediately downstream of a binding site for transcriptional repressor CTCF[17], with a much smaller effect that would not be statistically significant if considered without the sample growth effect. Although not globally significant, rs8111131 also appears to drive a reversal of correlation with *SOX17* that may be a temporal effect ($-\log_{10}P = 2.7$, **Fig. 3a**,**b**). Both genes negatively regulate cell growth by inhibiting Wnt signaling. Added to this, the C allele of rs9304726 (in linkage disequilibrium with rs8108607) was found to almost halve *PPP2R1A*'s network connectivity from 21 to 11 (**Fig. 3c**). An interpretation of these results is that variation of *PPP2R1A*'s transcription properties might have broad systemic consequences, making it a promising candidate for

more detailed follow-up studies to investigate the molecular basis of the heritability of variations in Wnt signaling.

In conclusion, gene expression is not only different between individuals but also between cells. Genes display complex and heritable spatiotemporal expression variability, which we propose is largely masked without techniques that offer higher resolution. Using a clinically relevant model pathway, and the largest study known to us of individual cells to date (1,440 cells), we have provided evidence that this masking is likely to be important enough to require the inclusion of single-cell technologies as part of the standard genetic epidemiology toolbox.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

Many thanks to L. Toji at the Coriell Institute for her valuable input on the cell line growth and transformation characteristics. Also, thanks to the following people at Fluidigm: B. Jones for his overall support, G. Harris and D. Wang for their help with primer design, and the meticulous technical assistance of K. Datta and R. Mittal. C.H. and T.E. are funded by the Medical Research Council of the UK. T.E. is also funded by Leukaemia Lymphoma Research and EuroSyStem.

AUTHOR CONTRIBUTIONS

Q.F.W. and C.H. conceived and designed the study. A.J.T. and T.E. ran the initial flow cytometry characterization and cell culture optimization. A.J.G. and D.W.S. ran the main study's cell culture and flow cytometry, further optimizing the sample characterization. K.J.L. designed and optimized the single-cell RNA assays, and generated the gene expression chip data. Q.F.W. analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
2. Li, H. & Deng, H. Systems genetics, bioinformatics and eQTL mapping. *Genetica* **138**, 915–924 (2010).
3. Bak, P. *et al.* Self-organized criticality: an explanation of the 1/f noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
4. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
5. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
6. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
7. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
8. Livak, K.J. *et al.* Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* **59**, 71–79 (2013).
9. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
10. Coghlan, M.P. *et al.* Selective small molecule inhibitors of glycogen synthase kinase-3 modulate glycogen metabolism and gene transcription. *Chem. Biol.* **7**, 793–803 (2000).
11. Dar, R.D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. USA* **109**, 17454–17459 (2012).
12. Bengtsson, M., Stahlberg, A., Rorsman, P. & Kubista, M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res.* **15**, 1388–1392 (2005).
13. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule Sensitivity in single cells. *Science* **329**, 533–538 (2010).
14. Bublik, D.R.R., Scolz, M., Triolo, G., Monte, M. & Schneider, C. Human GTSE-1 regulates p21(CIP1/WAF1) stability conferring resistance to paclitaxel treatment. *J. Biol. Chem.* **285**, 5274–5281 (2010).
15. Choy, E. *et al.* Genetic analysis of human traits *in vitro:* drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* **4**, e1000287 (2008).
16. Im, H.K.K. *et al.* Mixed effects modeling of proliferation rates in cell-based models: consequence for pharmacogenomics and cancer. *PLoS Genet.* **8**, e1002525 (2012).
17. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19**, 24–32 (2009).

ONLINE METHODS

Culture and perturbation of lymphoblastoid cell lines. Fifteen lymphoblastoid cell lines from unrelated HapMap individuals of European descent were supplied by the Coriell Institute. Further details on these samples are provided in the **Supplementary Data**. An initial 20 samples were selected based on age (samples from older subjects tend to grow more poorly) and no known poor growth characteristics. Four samples were removed based on slow growth characteristics and unsuitable IgM immunophenotypes (see "Flow cytometry and sample stratification"). Slow growth was found to broadly correlate with poor cell viability based on ATP content (CellTiterGlo, http://www.promega.com). All samples were seeded at 4×10^5cells/ml in standard media (RPMI 1640 containing L-glutamine (http://www.lifetechnologies.com/, 21875), 15% fetal calf serum (http://www.gehealthcare.com, A15-104), and penicillin/streptomycin (100 units per ml/100 mg per ml final concentration (http://www.lifetechnologies.com/, 15140-122)). In order to avoid batch to batch variations for cell growth, the standard media for all cell cultures were obtained from single batches of each of the cell culture constituents. Cells were initially passaged in T-25 flasks with all perturbations occurring in 24-well plates. Passage numbers were the same for all cells lines used and never exceeded 6. Treatment with 22.5 mg/ml acycloguanosine (Acyclovir) to suppress Epstein-Barr-Virus (EBV) activity was not found to have any observable effect on growth or gene expression and was, thus, omitted (data not shown). Seeded cells were grown for an initial 24 h, then perturbed with 10 μM SB216763 (http://www.sigmaaldrich.com/) or left unperturbed (baseline) for a further 24 h, before sorting.

Flow cytometry and sample stratification. Protein expression flow cytometry markers suggested widespread heterogeneity within and between samples; however, the implication of this is not clear. Although it is feasible that EBV transformation is not monoclonal, some of the markers appeared to drift (vary in proportion) within samples over time. Nevertheless, the following three markers were selected to minimize unwanted heterogeneity:

1. *DNA content (phase of cell cycle).* G_1 cells were selected using nuclear Hoechst staining. This also helped protect against 'doublets' (deposits of two cells instead of one).
2. *IgM expression.* In keeping with the nondifferentiated nature of the lymphoblasts, most samples cultured predominantly IgM^- cells[18]. These were selected for, to rule out any EBV-related heterogeneity that may be occurring with the inclusion of IgM^+ cells.
3. *CD27 expression.* CD27 expression is a marker for memory and plasma B cells[19], and, so, would not be expected in naive and undifferentiated cells. $CD27^+$ cells were excluded.

A BD FACS Aria II (Becton Dickinson) flow cytometer was used to perform single-cell sorting following the manufacturer's aseptic sort protocol. Cells were counted and viability assessed using a hemocytometer and trypan blue dye exclusion before staining. Nuclear DNA was stained using Hoechst 33342 (2 μg/ml) in buffer (pH 7.2) containing HBSS, 20 mM HEPES (http://www.invitrogen.com/), 5.55 mM glucose, 10% fetal calf serum, 50 μM Verapamil for 90 min at 37 °C, with gentle vortexing every 15 min. Cells were subsequently stained with PE-Cy7 CD27 (http://www.ebioscience.com) and Biotin IgM (http://www.bdbiosciences.com) antibodies for 20 min and Streptavidin APC-eFluor 780 (http://www.ebioscience.com) secondary antibody staining for a further 15 min. Antibody concentrations used were those recommended by the manufacturer and all antibody staining was done on ice in the Hoechst buffer specified above. Hoechst 33342 staining was detected using 375 nm laser illumination and 450/40 nm band pass–filtered detection; PE-Cy7 CD27 was detected using 488 nm laser excitation and 780/60 nm band pass–filter detection; and IgM APC-eFluor 780 was detected using 633 nm laser excitation and 780/60 nm band pass filter detection. Individual cells were sorted using the following gating criteria: debris discrimination using forward and orthogonal 488 nm laser scatter (cells selected), doublet discrimination using orthogonal pulse height and width (individual cells selected), and the above listed markers. In order to obtain maximum purity, cells were sorted twice using the defined gating strategy. Initially, sorted cells were collected as a pooled sample and subsequently re-sorted for single-cell deposition directly into pre-aliquoted lysis solution (see the section on "highly parallel qPCR").

Details of flow cytometry and EBV assays related to cell growth are provided in the **Supplementary Notes**, section 3. Details of the fresh naive B lymphocyte isolation and phenotyping are also provided in the **Supplementary Notes**, section 4, and **Supplementary Figure 10**.

Cell culture reproducibility. The hit genes listed in the **Supplementary Notes**, section 8 were validated by comparing their expression in six of the cell lines (GM12239, GM11881, GM12752, GM06991, GM07029, GM07019) with a duplicated cell culture batch. Cell culture duplicate QQ plots of Cq values from the combined samples are shown in **Supplementary Figure 9** for each gene. The adjacent bar plots show the interquartile ranges in the six baseline (B) and six perturbed (P) samples. Duplicate samples from the two cell culture batches are plotted in the same color. Most genes were found to be have highly reproducible expression, expect for the three marked with borders. These were genes expressed at low levels. As can be seen with some of the other genes, poor reproducibility occurs with Cq values >18 (marked in gray on the QQ plots). This is due to library generation and qPCR effects of very low starting RNA; these are discussed in the **Supplementary Notes**, section 5.

Tests of SNP association. Generation, QC and normalization of the expression data are described in the **Supplementary Notes**, sections 5–7, and **Supplementary Figure 11**. All phenotypes per gene were associated with the publicly available HapMap SNP genotypes (http://hapmap.ncbi.nlm.nih.gov) located 50 kb either side of the gene. Associations with less than 10 genotype-phenotype pairwise complete values were omitted from further analysis. As genes were selected for absence of nearby CNVs, these were not considered. Additive, dominant and recessive genotype effects were tested against the described phenotypes together with the growth effect described in the main text. Using leave-one-out cross-validation for each genotype, the model with the lowest predictive error was selected (genotype-only model versus growth-only model versus genotype-plus-growth model). In addition to ordinary least-squares (OLS) regression, robust Theil-Sen estimation and Kendall's τ were used to improve on the potential type II error rate with associations departing from parametric assumptions. The Theil-Sen estimate of association is an unbiased nonparametric linear regression approach[20], being distribution free while still retaining a high precision. As a measure of association it is simply the median of all pairwise slopes. Under parametric assumptions the Theil-Sen estimate demonstrates a 91% Pitman efficiency, and has been shown to be more efficient that OLS regression when data are not normal and skewed[21]. This efficiency and robustness makes the Theil-Sen approach an attractive option. As recommended by Sen[22], Kendall's τ was used to determine significance, whereas the Theil-Sen estimate was used for model selection. Theil-Sen multiple regression was by using the linear combination of genotype and growth that minimized the variability of Theil-Sen pairwise slopes. The exact *P* value from Kendall's τ tested the null hypothesis that $\tau = 0$, whereas OLS regression tested the null hypothesis that the slope coefficient $\beta = 0$. If the Theil-Sen estimate was more statistically significant than the OLS estimate, it was the estimate taken forward for multiple testing correction. To control for multiple testing, only the most statistically significant association per phenotype was considered. If the $-\log_{10}P$ proved greater than 3, the phenotype was permuted and retested against all genotypes 10^4 times. For each permutation, the maximum $-\log_{10}P$ was used to generate a null distribution of most significant *P* values, and so provide a family-wise error correction. As most of the genes are affected by a single pathway (that is, are not independent), a conservative corrected significance threshold of $-\log_{10}P = 4$ was used.

18. Hardy, R.R. & Hayakawa, K. B cell development pathways. *Annu. Rev. Immunol.* **19**, 595–621 (2001).
19. Wu, B., Piatkevich, K.D., Lionnet, T., Singer, R.H. & Verkhusha, V.V. Modern fluorescent proteins and imaging technologies to study gene expression, nuclear localization, and dynamics. *Curr. Opin. Cell Biol.* **23**, 310–317 (2011).
20. Siegel, A.F. Robust regression using repeated medians. *Biometrika* **69**, 242–244 (1982).
21. Johnstone, I.M. & Velleman, P.F. The resistant line and related regression methods. *J. Am. Stat. Assoc.* **80**, 1041–1054 (1985).
22. Sen, P.K. Estimates of the regression coefficient based on Kendall's Tau. *J. Am. Stat. Assoc.* **63**, 1379–1389 (1968).