ARTICLE

Received 16 Jul 2014 | Accepted 10 Dec 2014 | Published 14 Jan 2015

DOI: 10.1038/ncomms7066

The statistical geometry of transcriptome divergence in cell-type evolution and cancer

Cong Liang^{1,2,†}, the FANTOM Consortium*, Alistair R.R. Forrest^{3,†} & Günter P. Wagner^{1,4,†}

In evolution, body plan complexity increases due to an increase in the number of individualized cell types. Yet, there is very little understanding of the mechanisms that produce this form of organismal complexity. One model for the origin of novel cell types is the sister cell-type model. According to this model, each cell type arises together with a sister cell type through specialization from an ancestral cell type. A key prediction of the sister cell-type model is that gene expression profiles of cell types exhibit tree structure. Here we present a statistical model for detecting tree structure in transcriptomic data and apply it to transcriptomes from ENCODE and FANTOM5. We show that transcriptomes of normal cells harbour substantial amounts of hierarchical structure. In contrast, cancer cell lines have less tree structure, suggesting that the emergence of cancer cells follows different principles from that of evolutionary cell-type origination.

¹Systems Biology Institute, Yale University, West Haven, Connecticut 06516, USA. ²Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. ³RIKEN Center for Life Science Technologies (CLST), Division of Genomic Technologies, Yokohama 230-0045, Japan. ⁴Department of Ecology and Evolutionary Biology, Yale University, Prospect Street 165, OML Room 327, New Haven, Connecticut 06520-8106, USA. † Present addresses: 300 Heffernan Drive, B31 Room 267, West Haven, Connecticut 06516, USA (C.L.); 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan (A.R.R.F.); 300 Heffernan Drive, B31 Room 271, West Haven, Connecticut 06516, USA (G.P.W.). Correspondence and requests for materials should be addressed to G.P.W. (email: gunter.wagner@yale.edu).

*A list of the consortium members and their affiliations is provided in Supplementary Table 4.

The number of recognizable cell types varies between metazoan lineages by at least two orders of magnitude, from five in the primitive metazoan *Trichoplax* to at least 500 in humans^{1–3}. The processes that create this diversity, however, are not well understood. The only published model for cell-type origination we are aware of is the so-called sister cell-type model⁴. The model suggests that novel cell types arise through sub-specialization of an ancestral cell type. As a consequence, new cell types arise in pairs, so-called sister cell types. Sister cell types are expected to have more similar gene expression profiles than each of them compared with other cell types, because they initially share most of the developmental pathway of their ancestral cell type. This model thus predicts that the transcriptomes of cell types in one species has substantial amounts of tree structure. An alternative model is that novel cell types arises through the recruitment of co-regulatory modules, recruited from unrelated cell types. Similarly, a new cell type could arise by ‘fusion’ of gene regulatory networks of two unrelated cell types. In either case, the resulting cell types would not be expected to harbour substantial amounts of tree structure. Here we present a statistical tool that allows us to assess the amount of tree structure in a set of cell-type transcriptomes and apply this tool to two data set of cell-type transcriptomes, ENCODE and FANTOM5. We find that the transcriptomes of normal cell types have substantial amounts of tree structure, consistent with the sister cell-type model. In contrast, cancer cells do not retain strong similarities with their cell types of origin, suggesting a different mode of transcriptome divergence in cancer progression than during evolutionary origin of novel cell types.

Results

Approach to data analysis. In our analysis, we quantify RNA expression profiles from ENCODE Illumina sequencing⁵ in terms

of *tpm* (transcripts per million transcripts) based on the frequency of RNA sequencing (RNA-seq) reads mapped to a genomic feature⁶. The FANTOM5 data⁷ is from CAGE (Cap Analysis Gene Expression) sequences and is quantified as tags per million, which is quantitatively equivalent to *tpm* based on Illumina RNA-seq. Thus, we call all RNA abundance measures *tpm*. We then discretize the *tpm* of each gene from ENCODE data using an operational threshold of *tpm* = 3, where genes with *tpm* > 3 are coded as expressed (1) and those with *tpm* ≤ 3 coded to be not expressed (0; ref. 8). This approach is justified by a statistical model of transcript abundance as well as by correlation with chromatin immunoprecipitation data of active chromatin marks⁹. For the FANTOM5 data, an expression threshold of ≥ 2 tags per million is used, and the results are robust with respect to different thresholds. The rational to focus on qualitative expression profiles is that the exact expression levels of the genes are not a cell-type-specific property but influenced by many environmental and experimental factors.

After discretization, each sample is represented as a 0–1 vector of expressed and non-expressed genes (Fig. 1a). The similarity of gene expression profiles can be measured as Hamming distance H_{xy} , that is, the number of genes at which the expression state differs for cells *x* and *y*. The Hamming distances for all pairs of cell types determine a distance matrix for those samples (Fig. 1b). As mentioned above, the sister cell-type model predicts that gene expression profiles of the samples has substantial tree structure or, mathematically, the distance matrix of gene expression profile satisfies the four-point condition¹⁰ (see below). Here we present a statistical test for endogenous ‘treeness’ for gene expression data.

Tree structure of distance matrices has been investigated extensively in the field of statistical geometry^{10,11}. If a distance matrix is to be precisely described by a tree, it should satisfy the

Figure 1 | The logic of the treeness test. (a) Discretized expression data is represented in a 0–1 matrix where rows represent samples and columns represent genes. The entry is 0 if the gene is not expressed in a particular cell sample and 1 if it is expressed above threshold. (b) The data in the matrix can be represented as a Hamming distance matrix, where each entry is the number of genes where two cell types have different expression values. (c) Any collection of Hamming distances among four samples can be represented in a box graph, which consists of a internal box with dimensions *e* and *f*, and terminal branches with length *a*, *b*, *c* and *d*. (d) A tree is a special case of a box diagram where one and only one of the inner edges have zero length; *e* = 0 and *f* > 0. (e) Probability density of δ for random matrices. This distribution is relatively flat with a maximum at $\delta = 0.5$. The integral from $\delta = 0$ to δ_c is the type I error probability α for rejecting the null hypothesis, that is, the probability that a δ value of $\leq \delta_c$ is due to chance.

Figure 2 | Treeness tests on ENCODE data. (a) Distribution of α -values (significance values) of the treeness test for all tetrads from the 12 normal cell types in the ENCODE data. The high mode near $\alpha = 0$ indicates that there is a large number of tetrads where the null hypothesis is false, that is, where the tetrad has true tree structure. From this distribution, one can calculate that $100 \times (1 - \pi_0) = 67\%$ of the tetrads have true tree structure. (b) Tree relationship among normal cell types based on discretized transcriptomic data using maximum parsimony method²¹. Colours indicate different groups of cells. Magenta: fibroblast and myoblast cells; cyan: immune cells; orange: epithelial cells; black: others. (c) Neighbour net^{22,23} for the distance matrix for normal cells. Scale bar, 1,000. (d) Neighbour net of the distance matrix for neoplastic cells. Scale bar, 1,000. Note that the cancer cells have a greatly less resolved tree than normal cells, which is also reflected in the treeness statistic. Colour-marked samples are used for treeness test among cancer origin pairs.

four-point condition. That is, for any four samples (tetrad) x, y, u, v , with their pairwise distances H_{ij} , the two largest values of $H_{xy} + H_{uv}$, $H_{xu} + H_{yv}$ and $H_{xv} + H_{yu}$ should be the same (Fig. 1d). More generally, distance matrix of any tetrad can be represented in a box graph (Fig. 1c). Trees are limiting cases of box graphs, that is, a tree is a box graph in which one and only one of the inner edges is of zero length. Let the inner edges of the box graph be e and f , then the measure of treeness is the so-called delta statistic, $\delta = e/f$, $e \leq f$, which can be calculated directly from the distances¹¹. For example, the tetrad in fig. 1c, we have $H_{xy} + H_{uv} \leq H_{xu} + H_{yv} \leq H_{xv} + H_{yu}$, so δ is defined as

$$\delta = \frac{(H_{xv} + H_{yu}) - (H_{xu} + H_{yv})}{(H_{xv} + H_{yu}) - (H_{xy} + H_{uv})} \quad (1)$$

δ -Values close to zero indicate that the distance data conforms to the constraints of a tree. In contrast, δ -values close to 1 are indicative of complete lack of tree structure. Below we will present a statistical model that predicts the probability that a δ -value of less than a certain value δ_c is caused by chance. Hence we will

calculate the probability that a tetrad from a random 0–1 matrix has a δ -value smaller than a given value, $\delta \leq \delta_c$. This probability can be interpreted as type-I error probability, α , for the null hypothesis that the distance matrix is random. Note that this approach specifically tests only one class of deviations from randomness, namely non-random with respect to ‘treeness’.

Significant tree structure requires very low δ -values. Preliminary analysis of transcriptomic data suggests that $\sim 40\%$ of mapped genes fulfill the operational criterion of being expressed in any one cell type. Hence, there is a high chance of random similarities among gene expression profiles. We calculated the probability density of the δ -statistic, assuming that the expected proportion of expressed genes in each cell type is r , where r is neither close to 0 nor to 1 (with N being the total number of genes considered). We further assume that the expression indicator of each gene in each cell type is distributed as identical, independent Bernoulli variables with ‘success’ probability r . For a genome with large-enough N and moderate expression rate r , we arrived at an

Figure 3 | Using the treeness test as a tool to identify distinct cell types. (a) Comparison of pairs of replicates between normal cell types. If the two cell samples are in fact two cell types, we expect the tetrad to have significant tree structure, meaning that replicates are more similar to each other than each is to a replicate from a different cell type. (b) To test whether cancer cells remain similar to their normal cell types from which they originate, we compare cancer cell transcriptomes with that of normal cell types from which the cancer originated, that is, each tetrad has two pairs, each consisting of a normal cell with their cognate cancer cell. (c) Results for ENCODE data. (d) Results for FANTOM5 data. Note that cancer cells are not more similar to their cells of origin than to unrelated cell types, indicating that cancer cells diverge more radically than novel cell types in evolution. Fibroblasts from different body parts have marginally significant tree structure, suggesting that they might be representing the same cell type in different locations of the body. Finally, replicates of the other normal cell types are highly significant, showing that their replicates are much more similar than they are to other cell types. In the box-and-whisker plot, the bottom and top of the box are the first and third quartiles. The ends of whiskers are the lowest datum still within 1.5 interquartile range of the lower quartile, and the highest datum within 1.5 interquartile range of the upper quartile. The red dotted line indicates $\alpha = 0.05$.

asymptotic null hypothesis distribution independent of N and r , as shown in Fig. 1e and below.

$$p(\delta) = \frac{3\sqrt{3}}{2\pi(\delta^2 - \delta + 1)} \quad (2)$$

The distribution for the δ statistic is quite close to uniform on the unit interval with a maximum at $\delta = 0.5$. We then calculated, by direct integration of equation (2), the probability for δ to be smaller than a given threshold δ_c . This probability can be interpreted as the type-I error probability for rejecting the null-hypothesis of random structure. This probability is $\alpha = \int_0^{\delta_c} p(\delta)d\delta$, with the asymptotic solution

$$\alpha = \frac{3}{\pi} \left(\arctan\left(\frac{2\delta_c - 1}{\sqrt{3}}\right) + \frac{\pi}{6} \right) \quad (3)$$

For a significance level of $\alpha = 0.05$, δ has to be smaller than 0.0587.

Transcriptomes of normal cells have extensive tree structure.

We applied our statistical test to all normal human cell types from the ENCODE RNA-seq data. We obtained read files from 24 samples representing 12 non-cancerous cell types, including fibroblasts, embryonic stem cells, immune cells and others (Supplementary Table 1). We mapped the data and quantified gene expression levels, and converted the data into operationally defined gene expression indicators as expressed and non-expressed. A Hamming distance matrix was calculated and the 24 samples were arranged into 7,740 tetrads, with replicates from the same cell type deliberately avoided. Based on the null distribution (equation (3)) each of the tetrads was assigned an α -value for the treeness test. The distribution of all 7,740 α -values was analysed according to the methods of Benjamini and Hochberg (Fig. 2a)^{12,13}. With this method, it is possible to estimate the number of tetrads that are truly conforming to the null hypothesis, even though one cannot identify them individually. From this α -value distribution, we calculated the fraction of true null hypotheses $\pi_0 = 0.33$. The value $\pi_0 = 0.33$ implies that $1 - \pi_0 = 0.67$ is the fraction of tetrads that actually have tree structure. To estimate the confidence interval for this estimate, we performed a jack-knife procedure yielding a 95%

confidence interval of (0.57, 0.77). In other words, $67 \pm 10\%$ of the tetrads reflect non-random tree structure, suggesting strong evidence for tree structure among the majority of cell types.

We next analysed the more extensive data of 127 normal human cell types from the FANTOM5 consortium. We proceeded in the same way as described above for the ENCODE data and calculated the α -value distribution for transcriptomes of normal cells. We again found evidence for extensive tree structure. Specifically, the data suggests that $50 \pm 5\%$ of the tetrads have significant tree structure. Finally, we analysed 35 normal mouse cell types from FANTOM5 and found $74 \pm 5\%$ of the tetrads has tree structure. The exact fraction of tetrads with tree structure varies between data sets, most probably reflecting differences in celltype sampling. Overall, these results support the notion that the transcriptomes of normal human and mouse cells have substantial tree structure, consistent with the sister cell-type model⁴ (Supplementary Fig. 1a,b).

Next, we took a look at those tetrads with significant low α -values. One thousand and six hundred out of 7,740 tetrads from the ENCODE normal human cells have α -values < 0.05 . In each tetrad with a tree structure, the four cell types are grouped into two pairs, as shown in Fig. 1d (x, y) and (u, v). Among these tetrads, some cell-type pairs are observed much more frequently than random expectation (Supplementary Fig. 1c), suggesting that these cell types are more closely related to each other and two randomly sampled cell types. For instance, mesenchymal cells are more frequently paired with other mesenchymal cells in tree-like tetrads than expected by chance.

As the α -value distribution indicates extensive tree structure for those normal cell types, we built a phylogenetic tree of those samples using parsimony on the expression profile of protein-coding genes (for ENCODE data, see Fig. 2b,c). It is clear that fibroblast and myoblast cells (magenta) aggregated together, immune cells (cyan) also aggregated together and epidermal cells (orange) from ectoderm form another cluster. Similar results are achieved with only transcription factor-coding genes. The rationale for analysing transcription factor gene expression separately is that the total similarity of total transcriptomes can be influenced by similarity of function rather than phylogenetic relationships. Similarly, immune cells are more frequently paired with other immune cells and so on, implying that the tree

structure detected by our treeness test are biologically meaningful (for FANTOM5 data, see Supplementary Figs 2–3).

Some samples probably represent the same cell type. Cells analysed in this study have been designated as ‘different cell types’ based on their sampling location. For instance, there are a number of fibroblasts that have been sampled from different locations in the body, but it is not clear that they represent truly different cell types or just the same cell type in different parts of the body. We suggest that the ‘treeness test’ presented here can also be used as a first pass criterion to determine whether different samples could be from different cell types.

In a comparison of two cell types, each represented by two replicates, one would expect that the replicates are more similar to each other than each replicate is to that of another cell type. Hence, tetrads consisting of two pairs of replicates from different cell types are expected to have significant tree structure (Fig. 3a). However, replicates that in fact are all from the same cell type will follow the null model and fail to reject the null hypothesis. This is in fact the case (Fig. 3c). In the comparison among skin and lung fibroblasts from human normal ENCODE cell samples the average α -values are marginal (median α -value = 0.042), while for the comparison among all normal cell-type pairs we have lower α -values (median α -value = 0.0091). This statistic shows that the fibroblasts from the skin and the lung are only marginally more different than two replicates of fibroblasts from the same location. FANTOM5 data show the same pattern (Fig. 3d), with fibroblast median α -value = 0.049 and normal cells median α -value = 0.017. This suggests that fibroblasts from different parts of the body could be the same cell type, even though their gene expression patterns might be slightly different due to modulation by local factors in the skin and the lung for instance.

Transcriptomes of cancer cells have minimal tree structure.

Interestingly, for the cancerous cell types from ENCODE (Supplementary Table 2), we arrived at a much lower fraction of tetrads with true tree structure, $1 - \pi_0 = 0.23$, implying that only 23% of tetrads among neo-plastic cells have true tree structure (Fig. 2d and Supplementary Fig. 1d). The failure to find, among neoplastic cells, similar degrees of tree structure as with normal cells could be due to differences in the cell-type sampling. To further test whether this difference is biologically meaningful, we analysed tetrads with two normal cells (mammary epithelial cell and monocyte) and their derived cancer cells (mammary gland adenocarcinoma and leukemia cell) (Fig. 3b). If the cancer cells and their cells of origin are much more related to each other, the tetrad would have significant tree structure. However, if the origin of neoplastic cells is due to the recruitment of unrelated transcriptional modules, the tetrad would not have significant tree structure. This is actually the case as shown in Fig. 3c (median α -value = 0.40 over tetrads with different replicates of these four cell samples). A very similar pattern is found in the FANTOM5 data, where the cancer cells compared with the cell types of origin have a median α -value of 0.32. This indicates that the relationships between normal cell types are lost on neoplastic transformation either through general downregulation of cell state-specific gene products or through the breakdown of correlated expression profiles caused by heterogeneous and independent genomic applications and deletions seen in different cancer genomes.

Discussion

Detecting significant tree structure in the transcriptomic data from a set of normal cell types has two broad implications. For one, it supports the sister cell-type model of cell origination in

evolution⁴, and the second is that the analysis of cell-type trees can yield mechanistic insights. Here we briefly discuss these two points in turn.

Although our results are consistent with the sister cell-type model, we note that the precise nature of this inference is a ‘failed attempt to falsify the sister cell-type model,’ in the sense of Karl Popper’s philosophy of science, rather than a conclusive proof, if conclusive proof is possible at all in the empirical sciences. If our results would have failed to find the tree structure, the sister cell-type model would have lost its standing as a probable model of cell-type evolution. We also note that the sister cell-type model predicts that closely related cell types have similar gene expression patterns, because they share the developmental trajectory that they inherited from their ancestral cell type. Hence, the sister cell-type model also predicts a degree of congruence between the similarity of transcriptomes (interpreted here as a phylogenetic signal) and the ontogenetic cell lineage relationships⁴. In our analysis, this is reflected in the transcriptomic similarity between macrophages and dendritic cells, which are both derived from monocytes, as well as the separation between lymphoid and myeloid cells, which represent separate ontogenetic cell lineages (Supplementary Fig. 4)¹⁴ but the congruence is limited probably because ontogenetic trajectories are to some degree flexible.

The alternative to the sister cell-type model is analogous to hybridization of species. It would be a process where the new cell types arise by combining gene regulatory modules from very different cell types to create a new cell type. One consequence of this model is that new cell types do not arise as closely related and similar pairs. Naturally, under the ‘hybridization model’ of cell-type origination, the similarity structure among cell-type transcriptomes would be non-tree-like. In this context, it is interesting that we found substantially less tree structure among cancer cells, even when compared with their normal cell types of origin for each cancer (Fig. 3). This finding suggests that tumorigenesis is associated with transcriptional dysregulation, rather than an orderly hierarchical differentiation process.

The second consequence of our result is that the phylogenetic analysis of cell-type transcriptomes can allow biologically meaningful inferences, rather than just arbitrary classifications. Cell-type trees are hypotheses of the phylogenetic relationships among the cell types, similar to character trees as suggested by Oakley *et al.*^{15,16} and Geeta¹⁷ for plant parts. Consequently, phylogenetic reconstruction of ancestral transcriptomes, using standard phylogenetic methods, can reveal the history of gene recruitment events associated with the origin of a novel cell type in evolution and can lead to testable hypotheses about the molecular mechanisms of cell-type origination.

Methods

Data processing. RNA-seq data used in this analysis is from ENCODE⁵ (available at <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshLongRnaSeq/>). Samples are selected by ‘extraction method = longPolyA’ and ‘localization = cell’. Raw sequencing data are downloaded and aligned with Tophat¹⁸ (version 2.0.6) to UCSC hg19. HTSeq¹⁹ (version 0.6.1p1) is used to count reads with Ensembl²⁰ gene assembly *Homo sapiens* GRCh37.73. The tpm value for each gene is calculated based on read counts. A threshold of tpm = 3 is used to discretize the tpm values. CAGE data used in this analysis is generated by FANTOM5 (ref. 7; available at <http://fantom.gsc.riken.jp/5/data/>). Only differentiated and not externally treated cell samples are selected for the treeness test. Expression table with tpm values for refSeq genes is downloaded and a threshold of tpm = 2 is used to get the discretized expression profile.

Derivation of null model δ -value distribution. A full derivation of the null model δ distribution is available in Supplementary Methods (Supplementary Figs 5 and 6, and Supplementary Table 3).

References

1. Bonner, J. T. *The Evolution of Complexity* (Princeton University Press, 1988).
2. Valentine, J. W., Collins, A. G. & Meyer, C. P. Morphological complexity increase in metazoans. *Paleobiology* **20**, 131–142 (1994).
3. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
4. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
7. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
8. Wagner, G. P., Kin, K. & Lynch, V. J. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci.* **132**, 159–164 (2013).
9. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7**, 497 (2011).
10. Gusfield, D. *Algorithms on Strings, Trees and Sequences* (Cambridge University Press, 1997).
11. Holland, B. R., Huber, K. T., Dress, A. & Moulton, V. delta plots: a tool for analyzing phylogenetic distance data. *Mol. Biol. Evol.* **19**, 2051–2059 (2002).
12. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. (Stat. Method)* **57**, 289–300 (1995).
13. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *PNAS* **100**, 9440–9445 (2003).
14. Chaplin, D. D. Overview of the immune response. *J. Allergy Clin. Immunol.* **125**, S3–23 (2010).
15. Oakley, T. H. The eye as a replicating and diverging, modular developmental unit. *Trends Ecol. Evol.* **18**, 623–627 (2003).
16. Oakley, T. H., Plachetzki, D. C. & Rivera, A. S. Furcation, field-splitting, and the evolutionary origins of novelty in arthropod photoreceptors. *Arthropod. Struct. Dev.* **36**, 386–400 (2007).
17. Geeta, R. Structure trees and speies trees: what they say about morphological development and evolution. *Evol. Dev.* **5**, 609–621 (2003).
18. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
19. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* doi:10.1093/bioinformatics/btu638 (2014).
20. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
21. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
22. Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).
23. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).

Acknowledgements

FANTOM5 was made possible by the following grants: Research Grant for RIKEN Omics Science Center from MEXT to Y.H.; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y.H.; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Y.H. C.L. is gratefully acknowledging the receipt of a graduate fellowship from the China Scholarship Council, and G.P.W. acknowledges support from the Yale University Science Development Fund, and the John Templeton Foundation grant #54860 (the opinions expressed in this article are not those of the JTF). We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the data-set and thank GeNAS for data production.

Author contributions

C.L. derived the statistical model, analysed the data and edited the manuscript. A.R.R.F. provided data and contributed to the manuscript. G.P.W. conceived the project and wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

How to cite this article: Liang, C. *et al.* The statistical geometry of transcriptome divergence in cell-type evolution and cancer. *Nat. Commun.* 6:6066 doi: 10.1038/ncomms7066 (2015).