SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
BIOPHYSICS
SYSTEMS BIOLOGY
BIOTECHNOLOGY

Received
16 November 2012

Accepted
3 December 2012

Published
3 January 2013

Correspondence and requests for materials should be addressed to C.T.Z. (ctzhang@tju.edu.cn)

A novel triangle mapping technique to study the *h*-index based citation distribution

Chun-Ting Zhang

Department of Physics, Tianjin University, Tianjin 300072, China.

The *h*-index has received wide attention in recent years. The area under the citation function is divided by the *h*-index into three parts, representing *h*-squared, excess and *h*-tail citations. The *h*-index by itself does not carry information for excess and *h*-tail citations, which can play an even more dominant role than *h*-index in determining the citation curve, and therefore it is necessary to examine the relations among them. A triangle mapping technique is proposed here to map the three percentages of these citations onto a point within a regular triangle. By viewing the distribution of mapping points, shapes of the citation functions can be studied in a perceivable form. As an example, the distribution of the mapping points for 100 most prolific economists is studied by this technique.

The *h*-index, proposed by Hirsch[1] for evaluating the academic impact of individual researchers, has received wide attention in recent years. The citations received by all papers of a given researcher can be characterized by a citation distribution function, where the y-axis corresponds to the citations received by a paper, whereas the x-axis represents the paper rank arranged in descending order of citations (Fig. 1). The distribution of citation *verse* paper rank is called the citation distribution function or curve, denoted by $C(x)$, in which the paper x receives $C(x)$ citations. The *h*-index was simply defined as $C(h)=h$[1]. The area under the citation distribution curve is divided by the *h*-index into two parts: those of the *h*-core[2] and the *h*-tail[3]. The former is further divided into another two parts: those of excess citations[4] and *h*-squared citations[1]. As a consequence, the total citations are divided into three different parts: *h*-squared, excess and *h*-tail citations (Fig. 1). Indeed, the *h*-index lacks information for the excess and the *h*-tail citations, keeping only the citations related to the *h*-index (h^2). Theoretically, only when h^2 is dominant among the three parts, the *h*-index can properly reflect the academic performance of the scientist under study, otherwise, the *h*-index leads to biased evaluation. The question that whether the *h*-index dominates the citations or not depends on the shape of citation distribution function.

As pointed out by Bornmann and co-workers[5], for an isohindex group (scientists having the same *h*-index), their associated citation distribution functions may display quite different shapes. Therefore, to study how to apply *h*-index fairly, it is necessary to study the shape of the citation distribution functions, and the current study aims to address this question by using a triangle mapping technique. One of the advantages of citation triangle method is that the comparison of different shapes of the citation distribution functions can be performed intuitively. By viewing the distribution of mapping points within the triangle, the shapes of the citation distribution functions can be studied with a perceivable manner. Based on this method, we are able to study the degree with which the *h*-index is applicable properly. It is hoped that the technique presented here is useful for using the *h*-index to evaluate academic performance in a more unbiased way.

Results

We here propose a novel triangle mapping technique to study the relations among *h*-squared, excess and *h*-tail citations. For a regular triangle, the sum of the distances from any interior point to the three sides is equal to a constant, the height of the triangle. Note that the sum of the percentages for h^2, e^2 and t^2 is also a constant, which equals to 1. Based on this characteristic, percentages for these 3 kinds of citations are mapped onto a point in a regular triangle (Fig. 2A). Refer to the Method section for details.

First of all, let us consider two concrete examples. According to the citation information provided by Dodson[6], $C_{total}=1700$, $h^2=625$ ($h=25$), $e^2=477$ ($e=21.84$), and we find $H=0.37$, $E=0.28$ and $T=0.35$. Therefore, the mapping point corresponding to Dodson is situated at the region No. 4, where the *h*-index is applicable (Fig. 2B).

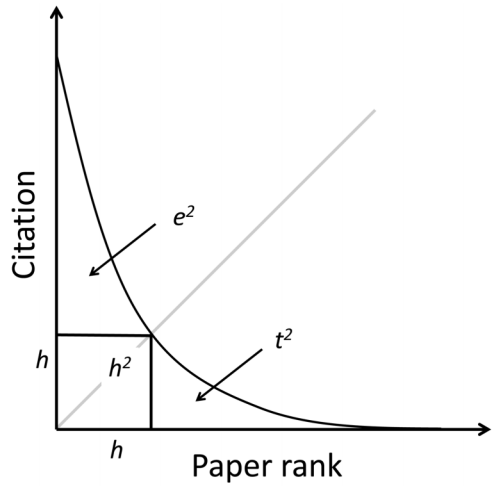

Figure 1 | The citation distribution curve. The y-axis corresponds to the citations received by a paper, whereas the x-axis represents the paper rank arranged in descending order of citations. The area under the citation distribution curve is divided by the *h*-index into three parts: h^2 , e^2(excess) and t^2(*h*-tail).

The second example is for the chemist Berni Alder, where $h^2=2500$ ($h=50$), $e^2=12996$ ($e=114$) and $C_{total}=18400$[4], and we find $H=0.14$, $E=0.71$ and $T=0.15$. Therefore, the mapping point corresponding to Alder is situated at the region No. 5, where the *e*-index is absolutely dominant (Fig. 2B). This example shows that Alder's *h*-index severely under-estimates his academic impact, and in this case, the *e*-index should be used together with the *h*-index for a fair evaluation[4].

In what follows, let us apply the citation triangle method to study the cases of citations of the 100 most prolific economists[7]. The data used to derive the corresponding *h*-index, *e*-index and *t*-index were kindly provided by Dr. Tol. As a consequence, we calculated the coordinates x and y of each mapping point corresponding to each economist. The distribution of the 100 points is showed in Fig. 3A. As we can see that only two points are situated at the region No. 3, i.e., an *h*-index dominant region. Meanwhile, only 11 points (11%) are situated above the horizontal line $H=1/3$ or $y=0$, where the *h*-index can be properly applicable. Accordingly, for the remaining cases (89%), where $H<1/3$ or $y<0$, the *h*-index should be used jointly with the *e*-index, even the *t*-index. The average *h*-index and *e*-index over the 100 points are 19 and 28.14, respectively, corresponding to the average H and E being 0.26 and 0.48 (average $x=-0.13$, $y=-0.07$). Overall, to have a fair and accurate evaluation, the *h*-index should be used together with the *e*-index even the *t*-index for most of the 100 most prolific economists.

The *h*-index captures only the information of the citation function partially. However, the above distribution of the 100 mapping points within the triangle provides more information about the shapes of the corresponding citation functions. For example, the mapping points within the small triangle No. 5 indicate that their citation distribution functions are peaked on the beginning part. On the contrary, the mapping points within the small triangle No. 9 indicate that their citation distribution functions are flat with a long tail. In both cases, the *h*-index seems not appropriate in capturing the main information of citation function. To complement the *h*-index, Bormann and co-workers[5] introduced three parameters: the h^2upper, h^2center and h^2lower, which correspond to E, H and T, respectively, in this paper. In other words, the triangle mapping technique provides an intuitive representation of the h^2upper, h^2center and h^2lower. Bornmann and co-workers[5] studied the shapes of the citation distribution functions of three scientists, A, B and C, belonging to an isohindex group with h = 14. For scientist A, E = 0.82, H = 0.15 and T = 0.03, corresponding to $x=-0.456$, $y=-0.183$. Its mapping point is situated at the small triangle No. 5, an *e*-index absolutely dominated regions. According to Bornmann et al[5] and Cole and Cole[8], Scientist A is called perfectionist-type scientist, who has rather few but very highly cited publications. For scientist B, E = 0.39, H = 0.48 and T = 0.13, corresponding to $x=-0.150$, $y=0.147$. Its mapping point is situated at the small triangle No. 2, a boundary region between *h*-index and *e*-index dominated regions. According to references[5,8], Scientist B is called a prolific-type scientist, who publishes a large number of high-impact papers. For scientist C, E = 0.10, H = 0.33 and T = 0.57, corresponding to $x=0.271$, $y=-0.003$. Its mapping point is situated at the small triangle No. 8, a *t*-index dominated region. Scientist C is called a mass producer[5,8], who publishes a larger number of papers that are lowly cited. It can be seen from the above analysis, the locations of the mapping points carry the information of the types of scientists. Therefore, the triangle mapping technique is particularly useful when the academic impact of a large number of scientists is studied. In that case, clustering analysis can be performed based on the mapping point locations, and therefore scientists can be classified according to their academic performance.

Recently, Baum introduced a new parameter, called Excess-Tail Ratio[9], denoted by$R$, where $R=E/T=e^2/t^2$. Baum found that for most cases he studied, $R<1$, even $R\ll1$.Only for few cases, $R>1$. The shapes of citation distribution functions for $R>1$ are peaked,

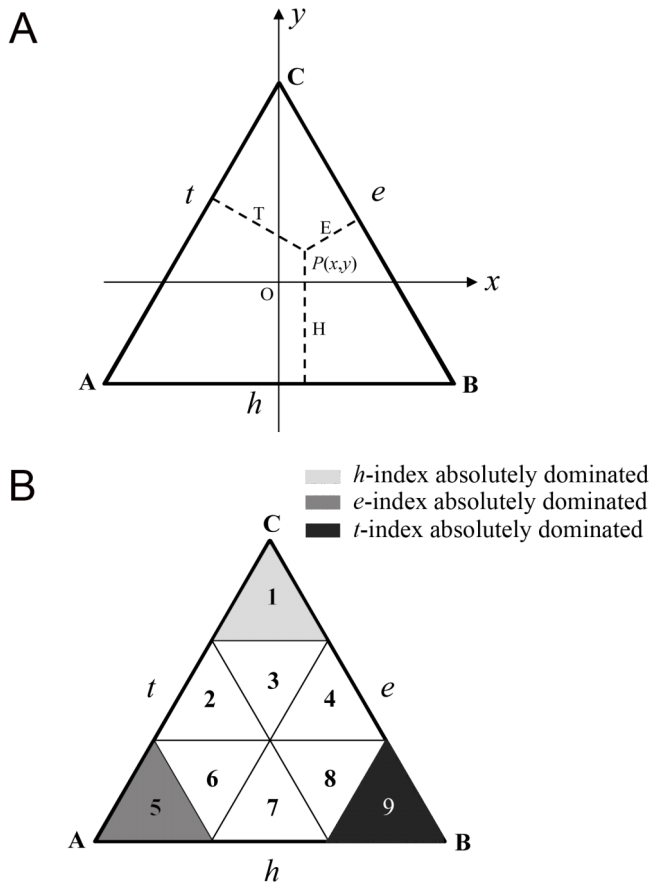

Figure 2 | The citation triangle method in studying *h*-index based citations. (A) A regular triangle ABC with its height being equal to 1 and center situated at O. A Descartes coordinate system $x-y$ is set up with its origin at O. The three sides of the triangle are denoted by h, e and t, and the distances of an interior point $P(x,y)$ to them are equal to H, E and T, respectively. Therefore, the point $P(x,y)$ is the mapping point for the three real numbers H, E and T. (B) The regular triangle is divided into nine smaller regular triangles. The intervals of H, E and T for each of the 9 smaller triangles are shown in Table 1.

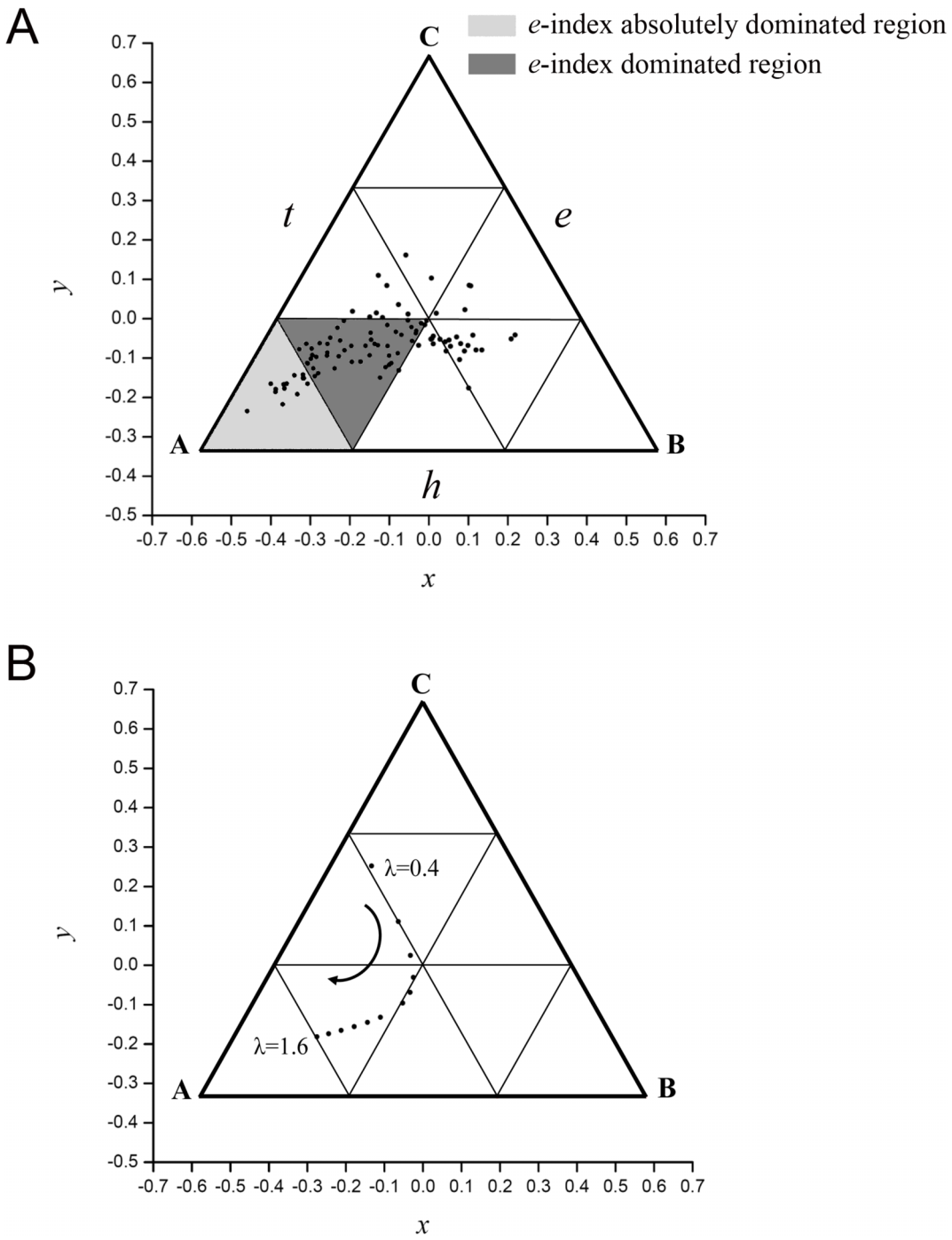

Figure 3 | Distributions of mapping points in the citation triangle. (A) The distribution of the 100 mapping points for each of the 100 most prolific economists. Note that only 11 points (11%) are situated at the regions where the *h*-index can be applicable ($H > 1/3$), indicating that the *h*-index should be used jointly with the *e*-index, even the *t*-index, for the remaining 89 economists. (B) An example to demonstrate that the power parameter λ is one of the key factors, which determines the position of the mapping point. Given $C_1 = 512$ and $N = 100$, starting from the region No. 3 (the *h*-index dominated region) with $\lambda = 0.4$, the mapping point moves to the region No. 6 (the *e*-index dominated region) with $\lambda = 1.6$. Interestingly, the track of the mapping points forms a clockwise rotating curve.

whereas for $R < 1$ the shapes of the citation functions are flat with a long tail. Therefore, the Excess-Tail ratio is an appropriate parameter to capture the overall shapes of the citation functions. According to eq. (12), $R > 1$or $R < 1$ corresponds to $x < 0$, or $x > 0$, respectively.

Discussion

In what follows, we want to explore the key factors that determine the shape of the citation distribution function. As previously, we assume a simple mathematical model for the citation distribution curve $C(x)$[4]

$$C(x) = \frac{C_1}{x^{\lambda}},\ C_1 = C(1) > 0,\ x \geq 1,\ \lambda > 0. \quad (1)$$

The total citations received by N papers, C_{total}, is

$$C_{total} = \int_1^N C(x)dx = \frac{C_1}{1-\lambda}\left(N^{1-\lambda} - 1\right). \quad (2)$$

Table 1 | Intervals of H, E and T for each of the 9 regions (small triangles) within the citation triangle

No.	H	E	T	Feature remark
1	$H>2/3$	$E+T<1/3$	$E+T<1/3$	The *h*-index absolutely dominated region
2	$H>1/3$	$E>1/3$	$T<1/3$	Boundary between *h*-index and *e*-index dominated regions
3	$H>1/3$	$E<1/3$	$T<1/3$	The *h*-index dominated region
4	$H>1/3$	$E<1/3$	$T>1/3$	Boundary between *h*-index and *t*-index dominated regions
5	$H+T<1/3$	$E>2/3$	$H+T<1/3$	The *e*-index absolutely dominated region
6	$H<1/3$	$E>1/3$	$T<1/3$	The *e*-index dominated region
7	$H<1/3$	$E>1/3$	$T>1/3$	Boundary between *e*-index and *t*-index dominated regions
8	$H<1/3$	$E<1/3$	$T>1/3$	The *t*-index dominated region
9	$H+E<1/3$	$H+E<1/3$	$T>2/3$	The *t*-index absolutely dominated region

Based on eq. (1), it was shown that[4,10]

$$h^{\lambda+1}=C_1. \tag{3}$$

However, we should have $h<N$, which leads to

$$\lambda>\lambda_0=\frac{\ln C_1}{\ln N}-1. \tag{4}$$

Meanwhile, we have[4]

$$e^2=\frac{1}{\lambda-1}\left(C_1-\lambda C_1^{\frac{2}{\lambda+1}}\right). \tag{5}$$

Using eqs. (1)–(5), we find

$$H=\frac{h^2}{C_{total}}=\frac{(1-\lambda)\times C_1^{\frac{1-\lambda}{1+\lambda}}}{N^{1-\lambda}-1},\ \lambda>\lambda_0,\ \lambda\neq 1. \tag{6}$$

$$E=\frac{e^2}{C_{total}}=\frac{\lambda\times C_1^{\frac{1-\lambda}{1+\lambda}}-1}{N^{1-\lambda}-1},\ \lambda>\lambda_0,\ \lambda\neq 1. \tag{7}$$

Therefore, the condition under which the *h*-index can be dominant should satisfy $H>1/3$, or

$$\frac{(1-\lambda)\times C_1^{\frac{1-\lambda}{1+\lambda}}}{N^{1-\lambda}-1}>\frac{1}{3}. \tag{8}$$

To have an intuitive picture, we consider some numerical examples as follows. Taking $C_1=512$, $N=100$ and letting $\lambda=0.4$, 0.5, 0.6, 0.7, 0.8, 0.9, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, respectively, we calculate the values of H and E for each case. Using eq. (12), we find 12 mapping points in the triangle, as shown in Fig. 3B. It is interesting to see that with the increase of the λ value, the track of the mapping points forms a clockwise rotating curve. This example shows that the power parameter λ is one of the key factors to determine the shape of the citation function. Given C_1 and N, there is a threshold of λ, when λ is less than this threshold, the *h*-index can no longer be properly applicable. In fact, $\lambda\to\infty$, $h\to 1$.

The main contribution of this paper is to propose the citation triangle method, by which the shapes of citation distribution functions can be studied in a perceivable form. Based on the distribution of mapping points, applicability and limitation of the *h*-index can be studied. Generally, the *h*-index is not properly applicable in the *e*-index or *t*-index dominated regions. In those cases, the *h*-index should be jointly applied together with the *e*-index or *t*-index. The proposed mapping technique provides a platform to study the academic impact of a group of scientists, because some mathematical methods, such as clustering analysis, can be used to study the distribution of mapping points, and the academic impact of these scientists can then be classified and compared.

Methods

The *h*-index was proposed by Hirsch in 2005[1]. The set of h papers of a scientist was called the *h*-core[2], in which at least h citations were received by each of the h papers. The *e*-index was proposed by Zhang[4], which was defined as the square root of excess citations over those used for calculating the *h*-index. Therefore, the total citations received by the papers in the *h*-core are equal to h^2+e^2. The *h*-index divides the total citations of a scientist into two parts: the first part is of the *h*-core, whereas the second one is of the *h*-tail[3]. For convenience, we define the square root of citations received by all papers in the *h*-tail as the *t*-index. Therefore, the number of total citations received by all papers of a scientist, C_{total}, is composed of three parts: h^2, e^2 and t^2, i.e.,

$$C_{total}=h^2+e^2+t^2, \tag{9}$$

where h, e and t are the *h*-, *e*- and *t*-index, respectively. Letting

$$H=h^2/C_{total},\ E=e^2/C_{total},\ T=t^2/C_{total}, \tag{10}$$

we have

$$H+E+T=1. \tag{11}$$

For any regular triangle, the sum of the distances from any interior point to the three sides is equal to the height of the triangle. Consider a regular triangle ABC with its height equal to 1 (Fig. 2A). Let the center of the triangle be denoted by O, and an x − y coordinate system is set up as shown in Fig. 2A. Based on eq. (11) and the feature of the regular triangle, the set of three real numbers H, E and T is mapped onto a point P(x, y) within the triangle, as shown in Fig. 2A. Simple calculation shows that

$$\begin{cases} x=(T-E)/\sqrt{3}=(1-H-2E)/\sqrt{3}, \\ y=H-1/3. \end{cases} \tag{12}$$

The triangle can be divided into 9 smaller triangles (regions) as shown in Fig. 2B. We denote them by No. 1 through No. 9, respectively. Each region is characterized by a special interval of the three real numbers H, E and T, respectively. For example, at the region No.1, $H>2/3$ and $E+T<1/3$, indicating that h^2 is absolutely dominant at this region as compared with e^2 and t^2. Similarly, at the region No. 5, $E>2/3$ and $H+T<1/3$, indicating that e^2 is absolutely dominant as compared with h^2 and t^2. At the region No. 9, $T>2/3$ and $H+E<1/3$, indicating that t^2 is absolutely dominant as compared with h^2 and e^2. Furthermore, at the region No. 3, $H>1/3$, $E<1/3$, $T<1/3$, so, it is called an *h*-index dominant region; at the region No. 6, $E>1/3$, $H<1/3$, $T<1/3$, so, it is called an *e*-index dominant region; and at the region No. 8, $T>1/3$, $H<1/3$, $E<1/3$, so, it is called a *t*-index dominant region. Finally, the region No. 2 is the boundary region between the *h*-index and *e*-index dominant regions, the region No. 4 is the boundary region between the *h*-index and *t*-index dominant regions, and the region No. 7 is the boundary region between the *e*-index and *t*-index dominant regions. The above description has symmetry of a regular triangle. The total description is summarized in Table 1.

The three real numbers H, E and T are the percentages of citations associated with the *h*-, *e*- and *t*-index, respectively. In general, H should be greater than 1/3 (or $y>0$), where the *h*-index is properly applicable, otherwise, if $H<1/3$ (or $y<0$), the *h*-index under-evaluates the academic impact of the researcher concerned. Therefore, the four regions No.1, No.2, No.3 and No.4 are the regions where the *h*-index can be properly applied ($H>1/3$). The regions No.2, No.5, No.6 and No.7 are the regions where the *e*-index can be properly applied ($E>1/3$), whereas those of No.4, No.7, No.8 and No.9 are the regions where the *t*-index can be properly applied ($T>1/3$). In summary, the *h*-index can only be properly applied in the regions No.1, No.2, No.3 and No.4 ($H>1/3$ or $y>0$); and the *h*-index should be jointly applied together with the *e*-index or *t*-index in the remaining regions No. 5 through No.9 ($H<1/3$ or $y<0$).

1. Hirsch, J. E. An index to quantify an individual's scientific research output. *P Natl Acad Sci USA* **102**, 16569–16572 (2005).
2. Rousseau, R. New developments related to the Hirsch index. *Science Focus* **1**, 23–25 (2006).
3. Ye, F. Y. & Rousseau, R. Probing the h-core: an investigation of the tail-core ratio for rank distributions. *Scientometrics* **84**, 431–439 (2010).
4. Zhang, C. T. The e-Index, Complementing the h-Index for Excess Citations. *Plos One* **4**, e5429 (2009).
5. Bornmann, L., Mutz, R. & Daniel, H. D. The h index research output measurement: Two approaches to enhance its accuracy. *J Informetr* **4**, 407–414 (2010).

6. Dodson, M. V. Citation analysis: Maintenance of h-index and use of e-index. *Biochem Bioph Res Co* **387**, 625–626 (2009).
7. Tol, R. S. J. The h-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics* **80**, 317–324 (2009).
8. Cole, S. & Cole, J. R. Scientific Output and Recognition - Study in Operation of Reward System in Science. *Am Sociol Rev* **32**, 377–390 (1967).
9. Baum, J. The Excess-Tail Ratio: Correcting Journal Impact Factors for Citation Quality. *SSRN*, http://ssm.com/abstract=2038102 (2012).
10. Egghe, L. & Rousseau, R. An informetric model for the Hirsch-index. *Scientometrics* **69**, 121–129 (2006).

Acknowledgements

I thank Dr. F. Gao and Dr. K. Song and for helps in preparing Figures 2–3. I am grateful to Dr. Richard Tol for kindly providing the data used to calculate the *e*-index and *t*-index for each of the 100 most prolific economists.

Author contributions

CTZ designed the study, performed most of the experiments, analyzed data and wrote the manuscript. All authors reviewed the manuscript.

Additional information

Competing financial interests: The author declares no competing financial interests.

How to cite this article: Zhang, C.T. A novel triangle mapping technique to study the *h*-index based citation distribution. *Sci. Rep.* **3**, 1023; DOI:10.1038/srep01023 (2013).