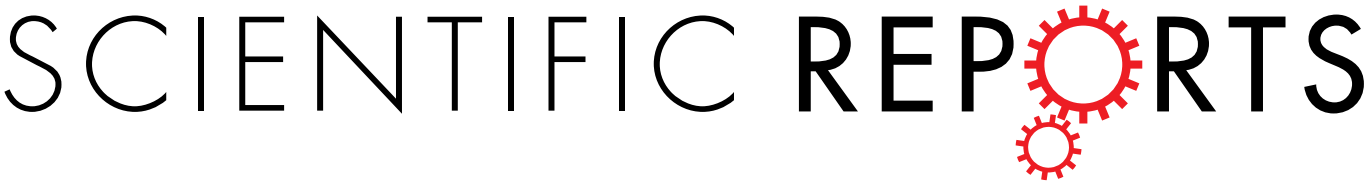

OPEN

Networks of plants: how to measure similarity in vegetable species

Received: 25 February 2016
Accepted: 06 May 2016
Published: 07 June 2016

Gianna Vivaldo[1], Elisa Masi[2], Camilla Pandolfi[2], Stefano Mancuso[2] & Guido Caldarelli[1,3,4]

Despite the common misconception of nearly static organisms, plants do interact continuously with the environment and with each other. It is fair to assume that during their evolution they developed particular features to overcome similar problems and to exploit possibilities from environment. In this paper we introduce various quantitative measures based on recent advancements in complex network theory that allow to measure the effective similarities of various species. By using this approach on the similarity in fruit-typology ecological traits we obtain a clear plant classification in a way similar to traditional taxonomic classification. This result is not trivial, since a similar analysis done on the basis of diaspore morphological properties do not provide any clear parameter to classify plants species. Complex network theory can then be used in order to determine which feature amongst many can be used to distinguish scope and possibly evolution of plants. Future uses of this approach range from functional classification to quantitative determination of plant communities in nature.

Plants are the building blocks of food production on Earth. Their role is crucial in the transformation and use of chemical energy to sustain the energy transfer in food webs and ultimately to feed any animal species. Despite their importance, they are seldom considered in ecological analysis of food webs and more generally they have attracted a relatively small interest for people studying complex networks. Actually, the study of vegetable world is revealing more and more evidence of the fact that different plants have many unexpected ties connecting them with each other. For instance, they are able to interact with the environment and to actively defend themselves from predators. On this respect, we note that the recipe that animal developed for the same purpose was to create an energetically expensive neural and locomotion system. This is mainly due to the fact that single individuals are indeed "in-dividual", that is they cannot be divided without killing them. Plants "individuals" instead can even propagate by their division and generally tolerate a loss of some of their parts. For this reason, in order to perform defensive tasks, they developed a series of features remarkably different than those of animal species. As a result, plants communicate, or "signal", with each other, using a complex internal analysis system to find nutrients, spread their species and even defend themselves against predators[1,2]. Plants have solved all these problems in different ways, shaping the plant growth, adapting to the different environmental constrains, using different kind of vectors in many phase of their life to overcome their immobility. One of the most critical stages in the life of any plant is the dispersal of seeds into a suitable habitat. To do this plants make effective use of many external agents such as wind, water, insects or higher animals. In order to track the many different series of strategies we need a measure to determine how much the same feature (i.e. fruit shape or diaspora mechanism) are different in two distinct species. A similar process is at the basis of taxonomic classification where plants are clustered according for example to properties (number of stamen) of plants, while cladistic classification is instead based on common ancestory.

In this paper we use network analysis[3–6] of some relational data about different plants with the aim of finding classes of "similar" plants. This analysis allows a clustering of species able to reveal and quantify similarity with respect to different species. In network theory, the various elements of an ensemble (i.e. plants in vegetal kingdom) are represented by vertices and they can be joined by considering common features they have. The number of common edges becomes then a quantitative proxy of relationships that are otherwise impossible to measure.

[1]IMT School for Advanced Studies, Piazza San Francesco 19, 55100 Lucca, Italy. [2]Università di Firenze, Dipartimento di Scienze Produzioni Agroalimentari e dell'Ambiente (DISPAA), Viale delle Idee, 30, 50019, Sesto Fiorentino Firenze, Italy. [3]London Institute for Mathematical Sciences, 35a South St. Mayfair, W1K 2XF London UK. [4]CNR-Istituto dei Sistemi Complessi (ISC), Roma, Italy. Correspondence and requests for materials should be addressed to G.C. (email: Guido.Caldarelli@imtlucca.it)

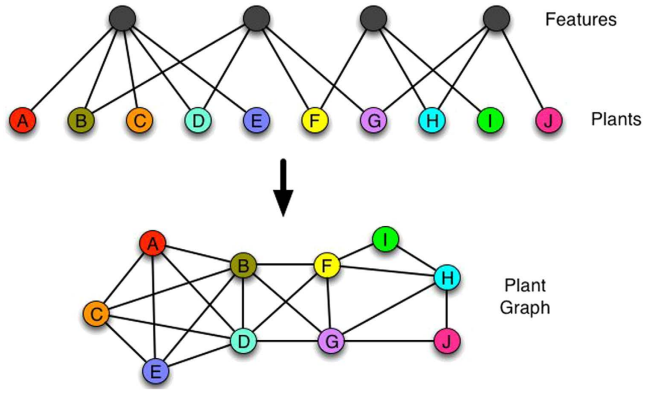

Figure 1. Bipartite network structure. From the original graph one can create a graph made by only one of the two sets.

In this respect this is similar to what happen in technological systems where the number of e-mails[7], likes on Facebook[8], or retweet between two persons[9], becomes a number assessing the strength of an acquaintance or even friendship. When passing to biology, network theory has been fruitfully used to determine structure and robustness of Food Webs[10], as well as the structure of protein interactions in the cell[11] with important applications to human diseases[12]. As previously mentioned, compared to other topics in biology, plants received a minor attention from networks scientists, despite some tentatives of comparing different ecosystems looking for steady (i.e. "universal") behaviours[13]. In order to adapt to the environment in which they live, plants have evolved an astonishing number of different mechanisms and structures to disperse their seeds. Typically, plants evolved in time to adapt to the environment in which they lived, so that only the mutations giving a comparative advantage with others were selected. Today, after 500 million years of plant evolution we are witnessing a huge differentiation in the features of plants as seed form and dispersal structure. Of the 250,000 today known flowering plants just a small fraction (5,000) has been classified in available databases on the basis of the variety of seed features.

These features can be represented by a graph of correlation, providing an effective taxonomy of vegetable species. The basic idea is to represent the information on plants, by means of a bipartite graph. A graph $G(N, E)$ is a mathematical object composed by N vertices and E edges. In a bipartite graph, vertices are divided in two sets, and the connections are made only from vertices of one set towards vertices of the other set. From one side we have the different plants, on the other side the various features. This information is transformed into two other graphs made by vertices of the same kind (see Fig. 1). In the first case we connect plants with plants on the basis of their common features. In the second case we connect features with features on the basis of how many plants have similar behaviour. Community detection[14] in such a graph is a powerful method to classify in a quantitative way the different vertices creating a taxonomic tree[15].

We present here the main results on the analysis conducted on the datasets considered; further detailed analysis is present in the Supplementary Information provided with this paper.

Results

The results presented here are computed on the dataset D^3 Dispersal and Diaspore Database[16] suitably represented as a network as shown in Fig. 1 and with the details presented in the section "Data".

Basic network analysis. Plants species networks G^P are defined by considering as vertices the plant species i and j in the database; two vertices are linked if they share at least one common property. The 2,662 plants species analysed are representative of 111 families, but the dataset is not homogeneous in terms of families percentages, being dominated by *Asteraceae* (12.81%), *Poaceae* (8.72%), *Cyperaceae* (5.63%), *Brassicaceae* (5.41%), *Rosaceae* (5.33%), and *Fabaceae* (4.58%). In the following we consider both properties related to diaspora morphology (G_1^P) as well as fruit typology (G_2^P). For the various networks, we considered size (number of edges), order (number of vertices), degree (average and its distribution), density (the ratio of actual vertices against the possible ones), clustering and finally (in the next section) the community structure.

Diaspora-based graph. A weight w_{ij} of each link e_{ij} can be defined by the total number of shared properties between plant i and plant j. The order of $G_1^P(N, E)$ is given by $N = 2{,}662$ vertices (plants species) and the size by $E = 1{,}176{,}968$ edges. The maximum and minimum number of properties shared by two plants are equal to 1 and 4, respectively. The 69.84% of plants share one property, only, and the proportion of edges with weight $w_{ij} = 1$ represents the 89.47% of E. On the contrary, just the 3.2% of the species share four properties, and $w_{ij} = 4$ links accounts for the 0.1% of the graph total number of edges E.

As regards the basic metrics, we can describe G_1^P as a weakly connected graph, whose density $\frac{2E}{N(N-1)}$ is equal to 0.332, the global weighted clustering coefficient is 0.84, and the nodes mean degree is $\overline{k} = \frac{1}{N}\sum_{i=1,N} k_i = \frac{2E}{N} = 884.27$. The network degree distribution $P(k)$, representing the fraction of vertices with degree $K > k$, is shown in Fig. 2 (panel A, black crosses). More in details, the log-line plot displays G_1^P degree complementary cumulative distribution function (CCDF). Analogously, panel B (black crosses) displays the

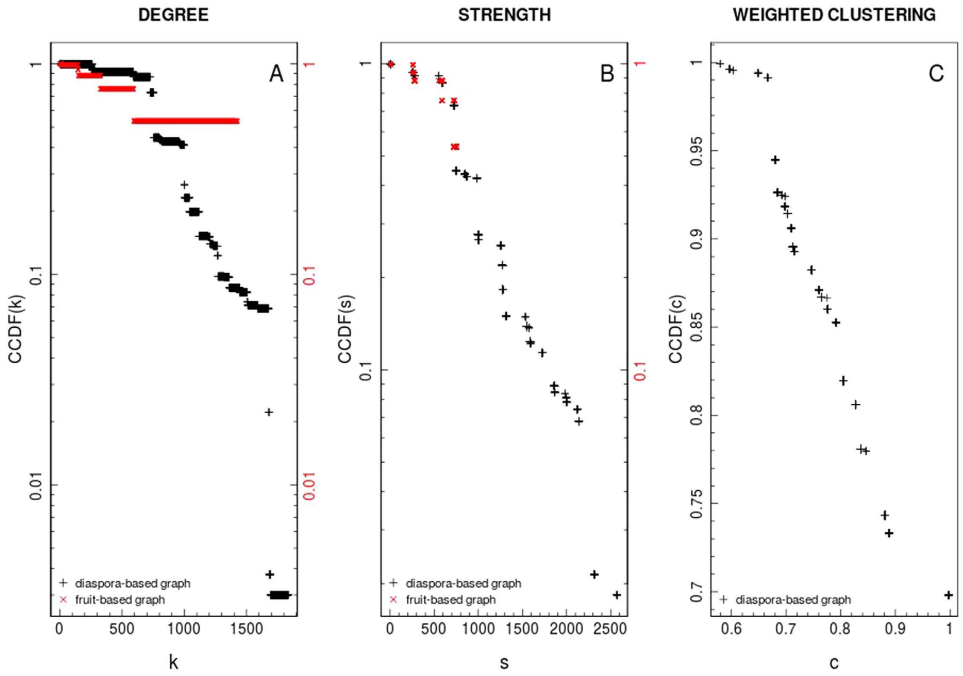

Figure 2. Basic network analysis. Complementary cumulative distribution functions (CCDF) of degree and strength are reported in log-line scale in panel A and B, respectively, for both diaspora-based network (black crosses) and fruit-based network (red crosses). Panel C, moreover, shows $G_1^P(N, E)$ weighted clustering coefficient distribution. More precisely, CCDF (on y-axis) is plotted versus the weighted clustering parameter (x-axis) on linear scale.

graph strength distribution, where the vertices strength s takes into account their connections total weight. Besides, panel C shows G_1^P local clustering coefficient, defined as the tendency among two vertices to be connected if they share a mutual neighbour. Taken as a whole, Fig. 2 suggests that plants network is not dominated by some central nodes with a huge amount of connections linking them to all the other minor vertices.

Fruit-based graph. We extended our analysis to the ecological properties of the fruit related to seed dispersal. Following the same approach, we created $G_2^P(N, E)$ as a projection of the bipartite graph where the plants are associated to fruit features. This creates a graph made up of $N = 2{,}662$ vertices (plants species) connected by $E = 1{,}265{,}831$ edges.

Also this graph is sparse with a density 0.357 and an average degree equal to $\overline{k} = 951.04$. The weight w_{ij} of each link e_{ij} is given by the total number of shared properties between plant i and plant j. The maximum number of properties shared by two plants is one, thus suggesting how fruit typology is a more strict parameter to classify plants behaviour related to diaspores, since plants cannot share more than a single trait. Moreover the properties are *mutually exclusive*, i.e. each species possesses just one of the eight properties analysed. That can be easily verified by building the bipartite projection of the fruit typology graph (not shown) made up by eight vertices, each one equal to a fruit typological property. The number of links of such a network is zero, meaning that two different properties do not share any species between them. Figure 2 (panel A, red crosses) shows the fruit-based graph degree CCDF by log-line scale, while G_2^P strength CCDF is displayed in panel B (red crosses). The weighted clustering coefficient distribution is not shown for that second graph since G_2^P is made up by fully connected isolated subgraphs, apart for a couple of nodes. Thus the local clustering coefficient is equal to 1 for all the vertices, while it is undefined for the two interconnected nodes (for a more deep description of the analysed network metrics, refer to Methods section).

Community detection analysis. *Diaspora-based graph.* We show the result of the community detection on the first graph G_1^P in Table 1. The communities detection results are obtained by using different algorithms: (i) fastgreedy (FG), (ii) walktrap (WT), (iii) Blondel's modularity optimisation algorithm (BL) and (iv) label propagation (LP) (see Methods). Each line corresponds to a different subgraph, i.e. a filtered-by-edges-weight version of G_1^P, with $w_{ij} \in [1, 2, 3, 4]$. Figure 3 shows the six communities detected by modularity algorithm (BL) in graph G_1^P. Colours refer both to cluster (panel A) and to families (panel B) membership. Looking to panel A, clusters 3 (cyan), 5 (red), and 6 (blue) are isolated components. The three bigger clusters and the corresponding families they embed are reported in Supplementary Information. Such communities are not homogeneous in terms of family composition (see panel B). Hereafter, the composition of every cluster is summarised, together with the morphological properties that the element families share each other. Notice that one property can be shared by more than a single species in the same cluster, since diaspore morphological features are not mutually exclusive.

FG	WT	BL	LP	Weight	E	N	Is.connected	Density
5	6	6	6	1	1176968	2662	FALSE	0.3323087
4	7	6	2	2	123939	803	TRUE	0.3849001
6	9	7	6	3	27009	343	FALSE	0.460488
4	4	4	4	4	1395	85	FALSE	0.3907563

Table 1. Plant species in diaspora-based plant graph are grouped on the basis of the common diaspore morphological properties. Four distinct communities detection algorithms were employed: FG = fast greedy, WT = walktrap algorithms, BL = Blondel modularity optimisation, LP = label propagation. Four filtered-by-edges-weight versions of the graph were analysed (one for each row). Graph edges weight integer values range from 1 to 4.

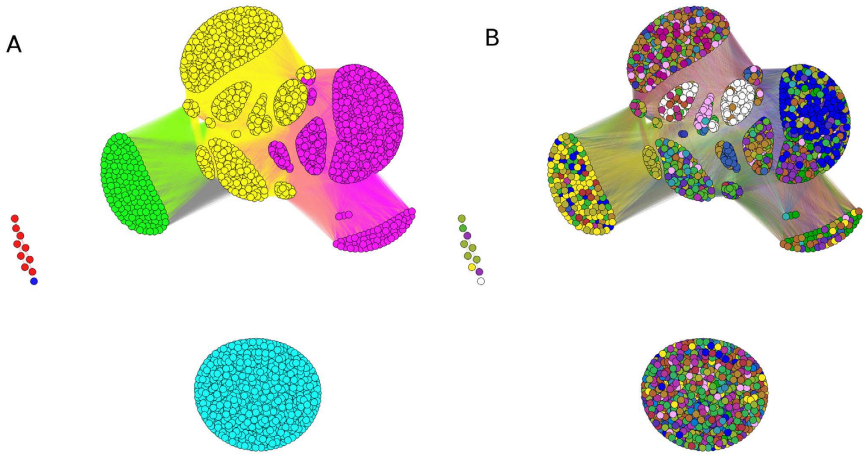

Figure 3. Communities detection based on diaspore morphology. The graphs refers to $G_1^P(N, E)$ communities detection by modularity method. Panel A shows the six communities which are detected: green, yellow, and fuchsia communities are highly connected components. On the contrary, red, blue and cyan clusters are isolated components. While cluster blue just embeds a single species (*X Calammophila baltica Brand*), cluster cyan is quite big, being composed by the 28.29% of total species present in the database D^3, for a total of 12 different families. Panel B shows the families belonging to each cluster. *Asteraceae* (blue, 12.81%), *Poaceae* (white, 8.72%), *Cyperaceae* (dark green, 5.63%), *Brassicaceae* (yellow, 5.41%), *Rosaceae* (cerise, 5.33%) are some of the most numerous. The heterogeneous distribution of families inside each clusters is evident.

- cluster 1: 884 species (33.21% of database D^3 total species); prevailing families: *Poaceae*, *Fabaceae*, *Rosaceae*, *Plantaginaceae*, *Polygonaceae*. 709 species have nutrient diaspores, followed by 447 showing flat/wings diaspore morphology; 204 times is encountered the elongated feature;
- cluster 2: 858 species (32.23%) dominant families: *Asteraceae*, *Cyperaceae*, *Ranunculaceae*, *Rosaceae*, *Apiaceae*, *Amaranthaceae*, *Salicaceae*, *Caprifoliaceae*, *Potamogetonaceae*. The vast majority of the species (782) show elongated diaspore trait; other common observed properties are: hooked (220), ballo/aerenchym (224), and flat/wings (140);
- cluster 3: 753 species (28.29%), sharing property *no specialization*. Notwithstanding its big dimensions, that cluster is a completely isolated component robust to changes in clustering algorithms. The leading families belonging to this cluster are summarized in Supplementary Information. They all share the same no specialization property concerning diaspore morphology. That category refers to species whose diaspores can have either a structured surface and no further appendages or specializations (e.g. many *Caryophyllaceae*), or a smooth surface and no further appendages or specializations (e.g. many *Brassicaceae*). *Caryophyllaceae* and *Brassicaceae* are two of the most numerous families with 86 and 43 species each respectively, besides *Orchidaceae* (61) and *Orobanchaceae* (48). Many species found in this cluster are characterized by very small, dust-like seeds, whose dispersal is easily achieved through the wind movements, even without specialized structures;
- cluster 4: 157 species (5.9%); prevailing families: *Brassicaceae*, *Juncaceae*, *Plantaginaceae*, *Asteraceae*, *Lamiaceae*. All these species share mucilaginous diaspore property;
- cluster 5: 9 plants species belonging to *Hydrocharitaceae*, *Brassicaceae*, *Polygonaceae*, and *Araceae* families. They all show other specialization concerning diaspore morphology. More in detail, 7 out of 9 are aquatic plants (5 species of *Hydrocharitaceae* and 2 of *Araceae* family); 1 species belongs to *Brassicaceae* and 1 to *Polygonaceae*. The 5 species of *Hydrocharitaceae* are strictly related: like other *Hydrocharitaceae*, they are aquatic plants that release their diaspore in water and that, conversely to other plants of the same family, have seeds with very low nutrients content; more, they do not set seeds regularly, preferring asexual reproduction; in both cases (sexual or asexual reproduction) water movements allow the dispersal; the 2 other aquatic

Cluster	Species	%	Families	%
1	884	33.21%	73	65.76%
2	858	32.23%	44	39.64%
3	753	28.29%	57	51.35%
4	157	5.9%	12	10.81%
5	9	0.34%	4	3.6%
6	1	0.04%	1	0.9%

Table 2. $G_{I,P}(N, E)$ clusters composition on the basis of diaspore morphological properties. The total number of species corresponds to the order $N = 2{,}662$ of the graph. The total number of families is equal to 111. Species and family percentage are referred to that values.

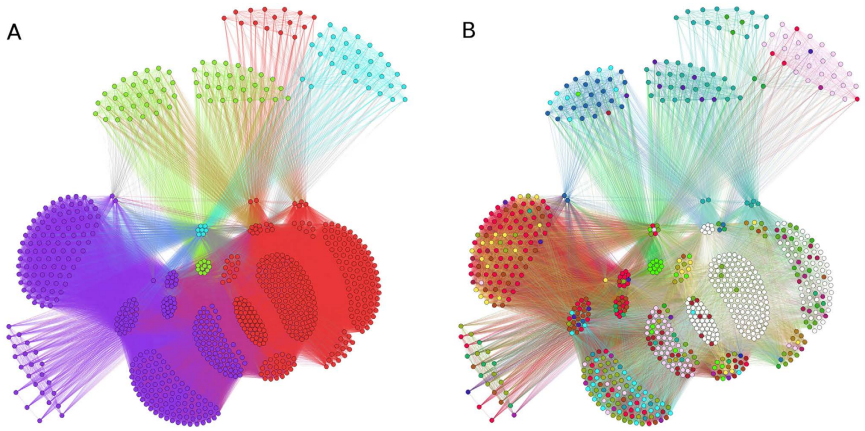

Figure 4. Communities detection on a filtered version of $G_1^P(N, E)$ graph. In that case, edges with weight $w_{ij} = 1$ are removed from the original graph. Four clusters are detected (panel A). Clearly each cluster is highly heterogeneous in terms of families composition, but more correspondences are found, and some families begin to dominate some cluster (especially red and cyan clusters of left panel). Prevailing families are visible in panel B: *Poaceae* (white), *Cyperaceae* (red), *Rosaceae* (cerise).

(*Araceae*) also prefer asexual reproduction; having no or little roots, the whole plants can float and disperse; the species belonging to the family of *Brassicaceae* has dehishent fruits; finally, the species of *Polygonaceae* rarely produces viable seeds and reproduction is normally asexual (by bulbils);

- cluster 6: 1 isolated plant, *X Calammophila baltica Brand (Poaceae)* which doesn't show any of the used morphological properties with the other species.

The total number of species which are part of each cluster, and the corresponding total number of families to which they belong are shown in Table 2. Notice the persistent heterogeneity of each cluster. The percentage reported in the third column of Table 2 refers to the relative number of species inside each cluster with respect to the total number of species present in D^3 database (2,662). Analogously, the relative number of families inside each cluster (last column, Table 2) is referred to the total amount of families inside the dataset, i.e. 111. Each plant belongs to a single cluster, while different families can characterize different clusters. The results are generally robust to changes in the detection algorithm, and to sizes of the filters employed over edges weights. In general we note that the network G_1^P is made up of a small number of clusters. Some of them behave like weakly-connected components that can be split into a different number of sub-clusters, depending on the applied methodology. For this reason we also made the same analysis on filtered versions of G_1^P to better focus on the largest components.

Communities after pruning of Diaspora-based graph. The same modularity analysis was performed on three filtered-by-edges-weight versions of the seeds features graph. Figure 4 shows the four communities detected by BL algorithm, after filtering by edges weight $w_{ij} > 1$, thus retaining plants connected by more that a single property. In that way only $N = 803$ vertices/plants species organized into 46 families and 123,939 links survive the pruning. Colors here keep the same meaning of Fig. 3, so that each color in the right panel corresponds to one of the 46 families present in the filtered dataset.

Again, detected communities are not homogeneous in terms of family composition. Anyway, more correspondences can be observed between the two panels of Fig. 4. Cluster 1 (red) and clusters 3 (cyan), for example, are less heterogeneous, being composed by *Poaceae* and *Rosaceae* families, respectively (white and cerise dots in the right panel). Table 3 reports species and families amount and the corresponding percentage present in each cluster. A brief description of the four clusters identified by BL method is the following.

Cluster	Species	%	Families	%
1	352	43.84%	31	27.9%
2	345	42.96%	27	24.32%
3	37	4.61%	7	6.3%
4	69	8.59%	7	6.3%

Table 3. Families and species composition for each cluster detected by BL method on a filtered version of G_1^P graph ($w_{ij} > 1$). After filtering just $N = 803$ vertices survive, corresponding each one to a different plant species. The total number of families is equal to 41. Families percentage is referred to the total amount of families into the dataset (111).

- cluster 1: 352 species (43.84% of database D^3 total species); *Poaceae* with 228 species are clearly the prevailing family: see white nodes in panel B of Fig. 4. They are followed by *Juncaceae* (14 plants), *Fabaceae*, *Santalaceae*, *Caprifoliaceae*, *Pinaceae*. All these species share that common properties: nutrients (315), flat/wings (312), elongated (240). They do not show (almost most of them) ballo/aerenchyms and mucilaginous surfaces;
- cluster 2: 345 species (42.96%); dominant families: *Cyperaceae* (89), *Rosaceae* (48), *Ranunculaceae* (42), *Asteraceae* (29). *Cyperaceae* are visible as red dots in Fig. 4 (panel B) in the position corresponding to violet cluster of panel A. That cluster embeds species joined by elongated (317) and hooked (211) diaspores shape. Ballo/aerenchyms and flat/wings are shared by 175 and 112 species, respectively. Just 4 species shows mucilaginous surfaces;
- cluster 3: 37 species (8.95%); *Rosaceae* family dominates with 23 species, visible as cerise vertices in Fig. 4 (panel B) in the position corresponding to cyan cluster in panel A. Almost all of them share clearly two properties: nutrients and ballo/aerenchyms surfaces;
- cluster 4: 69 species (4.61%), dominated by those belonging to *Potamogetonaceae* (20), *Plantaginaceae* (19), and *Amaranthaceae* (12) families. All the species have mucilaginous surfaces, some of them show flat diaspores (39), in particular species belonging to *Plantaginaceae* and *Juncaceae* families; other individuals show elongated diaspore (41), especially *Amaranthaceae*, *Asteraceae*, *Potamogetonaceae*.

Notice that after pruning G_1^P, the species dataset reduces to 803 species/vertices and it is made up especially of *Poaceae* (28.39%), *Cyperaceae* (11.96%), and *Rosaceae* (9.09%). Different clusters are dominated by different families: *Poaceae* (cluster 1), *Cyperaceae* (cluster 2), and *Rosaceae* (dominant family in cluster 3, and second dominant family in cluster 2).

In any case, some general conclusions can be drawn after pruning G_1^P. *Poaceae* family dominates cluster 1 with 228 species. This is a robust result, since before filtering out plants sharing a single property, *Poaceae* were rather well grouped into a single cluster. *Cyperaceae* family is present in cluster 4 with 89 species. Before pruning, that family was already one of the most copious in cluster 2 with 134 species, after *Asteraceae*. On the contrary, *Asteraceae*, which previously were copious (dominant family with 279 species in cluster 2, i.e. the magenta cluster in Fig. 3 (panel A)), now are quite disappearing, and just a thirty of them survives. The same happens for *Caryophyllaceae*, which go from a hundred of species to no one taxa surviving the pruning. *Rosaceae* family is present in cluster 2 with 48 species, and in cluster 3 with 23 species. Two single species belong to cluster 1, i.e. *Aremonia agrimonoides (L.) DC.* and *Potentilla alba L.* In the previous clustering related to the original graph G_1^P, *Rosaceae* were already split into two different clusters (cluster 1 with 66 species, and cluster 2 with 54 species).

The same approach was followed for the other two subgraphs corresponding to G_1^P filtered version by $w_{ij} > 2$ and $w_{ij} > 3$ (not shown). The species sharing 3 or 4 morphological properties were retained as vertices in the network. In this case, the number of analysed species drastically reduced to the 13% and 3.2% of the D^3 total amount of species. Thus, communities detection on such a highly reduced dataset had to be intended as a merely quantitative investigation. The most relevant insight confirmed previous result: *Poaceae* family survived severe filtering, and they gathered in two different ways. Some *Poaceae* species were grouped on the basis of three morphological properties, mainly: nutrient, elongated, and flat diaspore type. Some other species, usually found in the same community embedding *Rosaceae* species, also showed mucilaginous diaspore surfaces.

We can conclude that the high family heterogeneity in each cluster survives the edges-weight based filtering: diaspore morphology seems not to be a good classifier, and further analyses on different datasets are required.

Fruit-based graph. Communities detection results are summarized in Table 4 while a graphical view is provided in Fig. 5 where eight giant components are revealed. The detected clusters are clearly separated one from each other, and the vertices (plants species) are fully-connected inside each community. In other terms, the plants belonging to a cluster all share a single precise property. As for previous cases, no particular homogeneity in terms of family composition is observed (more information in the Supplementary Information).

Graph of properties, G^F from diaspore morphology. Similarly to what has been done so far we also considered the second projection giving the graph of features shown in Fig. 6. Such graph G^F is composed by $N = 8$ vertices and $E = 15$ edges. Two nodes are completely isolated, and they correspond to properties other specialization and no specialization, in agreement with the previous findings (see Fig. 3 (panel A), clusters 3 (cyan) and 5 (red)), looking like isolated components of the graph, that is to say, that those species showing properties that do not share any other property with the other species. The dispersal of plants characterized by such properties, also not sharing any other properties with other species, may be not crucially linked to seed or fruit morphology (and

Cluster	Species	%	Families	%
1	1426	53.57%	47	42.3%
2	593	22.28%	42	37.83%
3	326	12.25%	24	21.62%
4	149	5.6%	30	27.02%
5	143	5.37%	11	9.9%
6	13	0.49%	3	2.7%
7	10	0.38%	5	4.5%
8	2	0.08%	1	0.9%

Table 4. $G_2^P(N, E)$ clusters composition on the basis of fruit typology categorical traits. Species percentage is referred to the relative amount of species inside each cluster with respect to the total number of species present in the database (2,662). Families percentage is referred to the total number of families (111) present in the dataset. The majority of species belong to the first three clusters, which are also the most heterogeneous in terms of families composition.

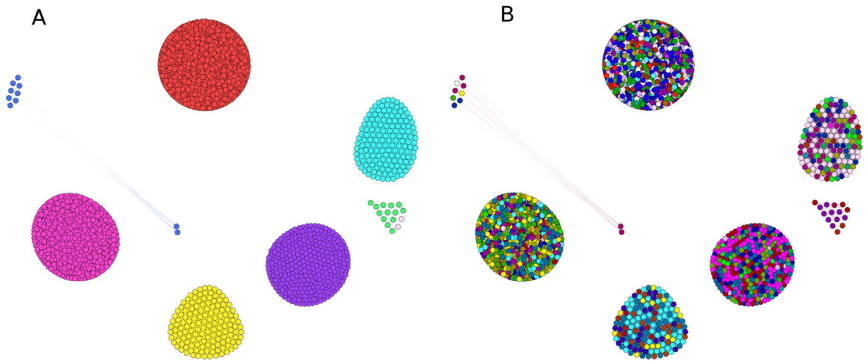

Figure 5. Fruit typology graph communities. $G_P^2(N, E)$ communities detection by modularity method (BL). Only edges with weight $w_{ij} = 1$ are present. Eight isolated communities are detected (panel A), and the corresponding families composition is displayed (panel B). Clearly each cluster is highly heterogeneous in terms of families composition, but not in terms of shared properties between the species belonging to each cluster. A single fruit topological property, in fact, is associated to each cluster and species. Main families are visible: *Poaceae* (white), *Asteraceae* (blue), *Cyperaceae* (red), *Rosaceae* (cerise), *Fabaceae* (cyan), *Caryophyllaceae* (fuchsia).

typology). Edges thickness is proportional to the number of common plants sharing the two properties connected by that link. In that sense, the elongated and flat appendages properties are common to a huge number of species. More in detail, the properties flat-elongated, flat-nutrient, hooked-elongated, elongated-nutrient share several species between them, respectively 323, 277, 272 and 219, and they have to be considered aggregative properties over the set of morphological seeds properties.

Discussion

Plants diaspore morphological features have been analysed in order to classify the various species. Data have been extracted from the D^3 Dispersal and Diaspore Database[16], developed as a partial solution to the gap about dispersal-related traits of plant species. In this paper we applied various quantitative measures, based on Complex Network Theory[17–19], in order to measure effective similarities between various species[20].

In particular we applied different communities detection algorithms

- to inspect plants species with the final goal to underline salient structures characterising our data;
- to identify the degree of similarity among the different species;
- to organise data in smaller structures and to gain insight into general hypothesis and properties of the whole dataset.

At a first glance, diaspores morphology did not turn out to be a good classification parameter for species. Indeed, different species share more than one common property, and each community shows a huge heterogeneity in terms of family composition. An explanation of this fact is that during their evolution plants were subjected to a strong selective pressure in order to colonise suitable habitats, mostly throughout the dispersal of seeds. To solve this problem, plants converged in the production of secondary structures such as plumes, samaras, hooks, wings, aerenchimas and mucilagines. Such convergent evolution determines that very similar solutions are found in species belonging to distant families. This is in accord to our results, where very different and genetically

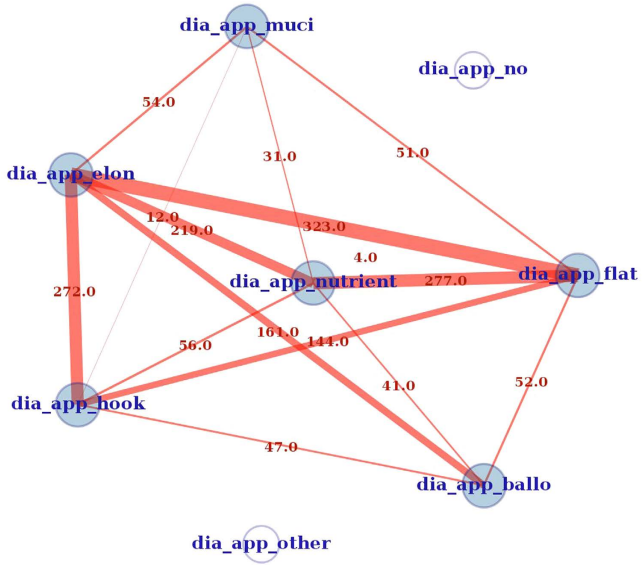

Figure 6. Morphological properties network. The following figure shows the second bipartite projection obtained by connecting features between them ($G^F(N, E)$) on the basis of how many plants share a given property. $N = 8$ vertices (blue circles) and $E = 15$ edges (red lines). Two nodes are isolated: they correspond to properties *other specializations* and *no specialization*. Thicker edges correspond to edges with higher weights (whose values are reported near each node), while nodes dimensions are proportional to the node degree (larger for higher degrees). It follows that properties *flat-elongated*, *flat-nutrient*, *hooked-elongated*, *elongated-nutrient* share more species between them, respectively 323, 277, 272, and 219.

unrelated plants cluster in stable groups. We observed the same behaviour also after a severe filtering that was applied on plants graph. Complex networks analysis main results in terms of basic quantities have been confirmed after pruning by edges weight, that is by removing species which shared a small number of properties.

On the other hand, species can be classified by their fruit topology, which prove to be a good categorical trait. A first explanation is that probably the selection did not push enough plants to provide convergent solutions for the environment where they lived. In the same spirit we intend in the future to do further analysis on the other features provided by D^3 Dispersal and Diaspore Database, such as diaspore typology, exposure of diaspores, heterodiaspory to improve the present findings. In conclusion, complex networks analysis seems to be an advantageous tool to investigate plants relationships related to morphological features. We believe that a similar approach may be applied with success to the study of many other fields of plant science, such as plant ecology, phytosociology and plant communication.

Materials and Methods

Data. Data are collected in the D^3 Dispersal and Diaspore Database[16] available at website http://www.seed-dispersal.info/. D^3 database is developed as a partial solution to the lack of knowledge about dispersal/related traits of plant species, with the aim to simplify traditional ecological and evolutionary analysis. Currently the database provides several information related to seed dispersal of plant species, such as empirical studies, functional and heritable traits, dispersal units image analysis and ranking indices (i.e. parameters which quantify the adaptation of a species to certain seed dispersal mode, in relation to a larger species set). More than 5,000 plant species are reported. Available raw data are mainly provided by DIASPORUS[21], BIOPOP[22] and LEDA[23] databases of plants traits. Here we focused on the well documented 2,662 Central European taxa, by exploiting the detailed ecomorphological categorizations of the diaspore and fruit, as well as information on prevailing dispersal modes. For every species we took into account diaspore morphology and fruit topology.

Diaspore Morphology. Morphology was treated technically as a set of binary traits. During the first test, eight features were taken into account for the categorization of diaspore morphology (see Supplementary information for more details): (1) nutrients: (2) elongated body; (3) hooked body; (4) flat/wings; (5) ballo/aerenchym; (6) mucilaginous; (7) none of the above: diaspores without any of the above mentioned specializations; (8) vegetative specialization. Such categorization scheme was inspired by the LEDA approach[23]. However, diaspore morphology represents an original dataset, which was derived either from visual inspection of the diaspores and respective images, or from an intensive and web research.

Fruit Typology. Fruit typology is a categorical trait which describes those ecological characteristics of the fruit which are related to seed dispersal. In the following analysis five categorization of ecological fruit types were taken into account. Fruit typology was categorized by visual inspection of fruits or respective images in addition to an intensive literature and web research[24]. Schematically (more detail on the Supplementary Information) they are divided into (1) indehiscent fruit: the pericarp is not opening during ripening; the above is further divided in (1a) non-fleshy;

(1b) fleshy fruit; (1c) pepo. (2) dehiscent fruit: the pericarp opens during ripening; further divided in (2a) fruit with upright aperture; (2b) fruit with lateral aperture. (3) explosive release. (4) *Gymnospermae* type. (5) not applicable: reserved for the following species (5a) sterile hybrid (e.g. *Betula x aurata*); (5b) for vegetative diaspore types.

Building the graph: projection in the space of plants/features. From the data written in the form of a bipartite graph (where every species N is connected to its features) we obtain two different projection graphs with the procedure shown in Fig. 1. Once a bipartite graph is built, it can also be described by a matrix $A(p, f)$ whose element a_{ij} is 1 if plant p has the feature f. The most immediate way to measure correlation between species is counting how many seeds features a couple of species share and similarly how many plants share the same couple of seeds features. In formulas, this corresponds to consider the matrix of species $P(p, p) = AA^T$ and the matrix of seeds features $F(f, f) = A^TA$. In detail we focused on the graph having as nodes the different plants, i.e. on the *Plants graph* $G^P(N, E)$ where edges weights were proportional to the number of common features shared by a couple of plants (this could be diaspora-based or fruit-based). Second, in order to catch the predominant properties in terms of seeds dispersal, we analysed the second bipartite projection, i.e. the *Features graph*, $G^F(N, E)$, whose nodes represented the different diaspore morphological traits taken into account. In that case edges weights were proportional to the number of plants sharing the same feature. Both a network metrics analysis, and a basic cluster analysis were performed to obtain an alternative classification of plants.

Basic network analysis. As regards network analysis, we computed some global and local basic metrics, described hereafter.

- *Graph density* is defined as the ratio between the numbers of existing edges and the possible number of edges, in a N-size network it is given by $\left(\frac{2E}{N(N-1)}\right)$.
- *Network clustering coefficient* is the overall measure of clustering in a undirected graph in terms of probability that the adjacent vertices of a vertex are connected. More intuitively, global clustering coefficient is simply the ratio of the triangles and the connected triples in the graph. The corresponding local metric is the *local clustering coefficient*, which is the tendency among two vertices to be connected if they share a mutual neighbour. In this analysis we used a local vertex-level quantity[5] defined in Eq. (1):

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{jh} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}, \tag{1}$$

- The normalization factor $\frac{1}{S_i(k_i - 1)}$ accounts for the weight of each edge times the maximum possible number of triplets in which it may participate, and it ensures that $0 \leq c_i^w \leq 1$. That metric combines the topological information with the weight distribution of the network, and it is a measure of the local cohesiveness grounding on the importance of the clustered structure on the basis of the amount of interaction intensity actually found on the local triplets[5].

- *Network strength* (s) is obtained by summing up the edge weights of the adjacent edges for each vertex[5]. That metric is a more significant measure of the network properties in terms of the actual weights, and it is obtained by extending the definition of *vertex degree* $k_i = \sum_j a_{ij}$, with a_{ij} elements of the network adjacent matrix A. In formulas $s_i = \sum_{j=1}^{N} a_{ij} w_{if}$.

Grouping plants from graph: communities detection analysis. Communities detection aims essentially at determine a finite set of categories (clusters or communities) able to describe a data set, according to similarities among its objects[25]. More in general, hierarchy is a central organising principle of complex networks, able to offer insight into many complex network phenomena[26].

In the present work we adopted the following method:

- Fast greedy (FG) hierarchical agglomeration algorithm[27] is a faster version of the preceding greedy optimisation of modularity[15]. FG gives identical results in terms of found communities. However, by exploiting some shortcuts in the optimisation problem and using more sophisticated data structures, it runs far more quickly, in time $O(md \log n)$, where d is the depth of the "dendrogram" describing the network community structure.
- Walktrap community finding algorithm (WT) finds densely connected subgraphs from a undirected locally dense graph *via* random walks. The basic idea is that short random walks tend to stay in the same community[28]. Starting from this point, WT is a measure of similarities between vertices based on random walks, which captures well the community structure in a network, working at various scales. Computation is efficient and the method can be used in an agglomerative algorithm to compute efficiently the community structure of a network.
- Louvain or Blondel method (BL)[29] to uncover modular communities in large networks requiring a coarse-grained description. *Louvain* method (*BL*) is an heuristic approach based on the optimisation of the modularity parameter (Q) to infer hierarchical organization. Modularity (Eq. (2)) measures the strength of a network division into modules[15,30], as it follows:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w) = \sum_{i=1}^{c} (e_{ii} - a_i^2), \tag{2}$$

- where, e_{ii} is the fraction of edges which connect vertices both lying in the same community i, and a_i is the fraction of ends of edges that connect vertices in community i, in formulas: $e_{ii} = \frac{1}{2m}\sum_{vw}[A_{vw}\delta(c_v, c_w)]$, and $a_i = \frac{k_i}{2m}$; **A** is the adjacent matrix for the network; c the number of communities; $k_i = \sum_w A_{\{i,w\}}$ the degree of the vertex-i, n and $m = \frac{1}{2}\sum_{\{vw\}} A_{\{aw\}}$ the number of graph vertices and edges, respectively. Delta function, $\delta(i, j)$, is 1 if $i = j$, and 0 otherwise.

- Label propagation (LP) community detection method is a fast, nearly linear time algorithm for detecting community structure in networks[14]. Vertices are initialised with a unique label and, at every step, each node adopts the label that most of its neighbours currently have, that is by a process similar to an 'updating by majority voting' in the neighbourhood of the vertex. Moreover, *LP* uses the network structure alone to run, without requiring neither optimisation of a predefined objective function nor *a-priori* information about the communities, thus overcoming the usual big limitation of having communities which are implicitly defined by the specific algorithm adopted, without an explicit definition. In this iterative process densely connected groups of nodes form a consensus on a unique label to form communities.

References

1. Brenner, E. D. *et al.* Plant neurobiology: an integrated view of plant signaling. *Trends Plant Sci.* **11,** 413–419 (2006).
2. Baluška, F. & Mancuso, S. Plant neurobiology as a paradigm shift not only in the plant sciences. *Plant Signal. Behav*. **2,** 205–207 (2007).
3. Caldarelli, G. *Scale-Free Networks: complex webs in nature and technology* (Oxford University Press, 2007).
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. U. Complex networks: Structure and dynamics. *Phys. Rep*. **424,** 175–308 (2006).
5. Barrat, A., Barthelemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *P. Natl. Acad. Sci. USA* **101,** 3747–3752 (2004).
6. Dehmer, M. & Emmert-Streib, F. *Quantitative Graph Theory: Mathematical Foundations and Applications* (CRC Press, 2014).
7. Caldarelli, G., Coccetti, F. & De Los Rios, P. Preferential exchange: strengthening connections in complex networks. *Phys. Rev. E* **70,** 027102 (2004).
8. Zollo, F. *et al.* Emotional dynamics in the age of misinformation. *PLoS ONE* **10,** e0138740 (2015).
9. Eom, Y.-H., Puliga, M., Smailović, J., Mozetič, I. & Caldarelli, G. Twitter-based analysis of the dynamics of collective attention to political parties. *PLoS ONE* **10,** e0131184 (2015).
10. Dunne, J. A., Williams, R. J. & Martinez, N. D. Food-web structure and network theory: the role of connectance and size. *P. Natl. Acad. Sci. USA* **99,** 12917–12922 (2002).
11. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122,** 957–968 (2005).
12. Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity. *P. Natl. Acad. Sci. USA* **105,** 9880–9885 (2008).
13. Caretta-Cartozo, C., Garlaschelli, D., Ricotta, C., Barthélemy, M. & Caldarelli, G. Quantifying the universal taxonomic diversity in real species assemblage. *J. Phys. A-Math. Gen*. **41,** 224012 (2008).
14. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **76,** 036106 (2007).
15. Newman, M. E. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69,** 026113 (2004).
16. Hintze, C. *et al.* D^3: the Dispersal and Diaspore Database – Baseline data and statistics on seed dispersal. *Perspect. Plant Ecol.* **15,** 180–192 (2013).
17. Ma, J., Shi, Y., Wang, Z. & Yue, J. On wiener polarity index of bicyclic networks. *Sci. Rep.* **6,** doi: 10.1038/srep19066 (2016).
18. Li, X., Li, Y., Shi, Y. & Gutman, I. Note on the homo-lumo index of graphs. *MATCH Commun. Math. Comput. Chem*. **70,** 85–96 (2013).
19. Cao, S., Dehmer, M. & Shi, Y. Extremality of degree-based graph entropies. *Inform. Sciences* **278,** 22–33 (2014).
20. Emmert-Streib, F., Dehmer, M. & Shi, Y. Fifty years of graph matching, *network alignment and comparison*. Information Sciences, Vol. 346–347, 180–197 (2016)
21. Bonn, S., Poschlod, P. & Tackenberg, O. "Diasporus"–a database for diaspore dispersal concept and application in case studies for risk assessment. *Z. Ökol. Nat.schutz* **9,** 85–97 (2000).
22. Poschlod, P., Kleyer, M., Jackel, A. K., Dannemann, A. & Tackenberg, O. Biopop a database of plant traits and internet application for nature conservation. *Folia Geobot*. **38,** 263–271 (2003).
23. Kleyer, M. *et al.* The leda traitbase: a database of life-history traits of the northwest european flora. *J. Ecol.* **96,** 1266–1274 (2008).
24. Bojnansky, V. & Fargašová, A. *Atlas of seeds and fruits of Central and East-European flora: the Carpathian Mountains region* (Springer Science & Business Media, 2007).
25. Campello, R. A fuzzy extension of the Rand Index and other related indexes for clustering and classification assessment. *Pattern Recogn. Lett*. **28,** 833–841 (2007).
26. Clauset, A., Moore, C. & Newman, M. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (2008).
27. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70,** 066111 (2004).
28. Pons, P. & Latapy, M. Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, 284–293 (Springer, 2005).
29. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory E*. **2008,** P10008 (2008).
30. Newman, M. E. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69,** 066133 (2004).

Acknowledgements

The authors acknowledge support from EU FET Open Project PLEASED nr. 296582. GV and GC also acknowledge EU FET Integrated Project MULTIPLEX nr. 317532. SM and GC are particularly indebted with C. Tomei for many interesting discussions in his Lab.

Author Contributions

G.V., E.M., C.P., S.M. and G.C. contributed equally to the analysis of the dataset and to the interpretation of the results of this analysis, both from the point of view of Network Theory as well as in terms of biological

implications. G.V., E.M. and C.P. made the data analysis. All the authors contributed equally to the writing and reviewing of the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

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

How to cite this article: Vivaldo, G. *et al.* Networks of plants: how to measure similarity in vegetable species. *Sci. Rep.* **6**, 27077; doi: 10.1038/srep27077 (2016).