Espaliers: A Visualization Method for Big Data

Max Robinson[1,*], Greg Eley[2,3], Joseph G. Vockley[3],
John E. Niederhuber[3], and Gustavo Glusman[1]
[1]Institute for Systems Biology, 401 Terry Ave N, Seattle, WA 98109
[2]Scimentis LLC, 1515 Evergreen Park, Statham, GA 30666-3629
[3]Inova Translational Medicine Institute, Inova Health System,
3300 Gallows Road, Falls Church, VA 22042
[*]Corresponding author
Oct 2, 2015

Abstract
As thousands of human genomes become available, there is pressing need for efficient and intuitive analysis and visualization methods. The genotypes observed in a set of genomes can be represented as a [genome x variant] matrix. Standard PCA-based visualizations of genotype matrices can reveal population structure, but give little insight into genetic admixture in individuals or the history of individual variants.

We present Espaliers, a novel visualization of non-negative matrices, including genotype matrices. Given an ordering of the genomes (columns), we compute a position for each variant (row) that reflects the information in the genotype matrix. An Espalier plots each variant by this position and its population frequency (row sum), which is related to the variant's age. The resulting Espalier plot resembles a parsimonious evolutionary tree connecting the genomes that is consistent with the input ordering of the genomes.

We compare Espaliers with PCA, provide examples of Espaliers for Big Data sets from genomics and transcriptomics, and discuss potential future directions.

Key Words: genotypes, visualization, PCA, non-negative matrices, linear dimensionality reduction

1. Introduction

Since the first human genome sequences were determined (Venter *et al.* 2001; Lander *et al.* 2001), DNA sequencing technology has become affordable and efficient enough to enable sequencing tens of thousands of whole genomes. The massive raw data from each sequenced DNA sample is mapped to a standard reference sequence and expressed as a few million "variants" from the reference. While other genotyping technologies evaluate a subset of known variant positions, whole genome sequencing (WGS) systematically collects nearly complete genotypes. As in other disciplines, studies of population structure are both enabled and challenged by the availability of such "Big Data"; new methods for analysis and especially for visualization of such large datasets are needed. Here we present Espaliers as a visualization metaphor for genotypes and other large datasets that share a common object-attribute matrix structure. We describe the construction of Espaliers and discuss Espalier Plot visualizations of large genotype and gene expression datasets to demonstrate the value of the Espalier metaphor.

The remaining sections are organized as follows. In Section 2 we describe object-attribute matrices that arise in many "Big Data" contexts, and introduce two examples. In Section 3 we present the Espalier metaphor and discuss Principal Component Analysis (PCA; Pearson 1901) as a potential means of constructing Espaliers. We then present Key Espalier Analysis (KEA), a true Espalier construction method. In Section 4 we compare principal components and KEA Espaliers derived from a set of 2,504 genotypes, and compare Espalier Plots and KEA biplots with the standard population structure visualization, the PCA biplot. In Section 5 we make similar comparisons for Espaliers constructed from a dataset of gene expression in 16 tissue samples. We summarize our conclusions in Section 6.

2. Object-Attribute Matrices

Many "Big Data" studies concern counts or other measured *amounts* a_{ij} of each attribute i of an object j. In text mining, for example, the data may be counts a_{ij} of word i in source text j. In gene expression studies, the amount a_{ij} of mRNA or protein i is measured in each tissue or blood sample j. In social networks, the data may be a binary indicator a_{ij} of a relationship of some particular type between "entities" i and j. Note that numerical categories are not amounts; two type 1 errors do not constitute a type 2 error. We make this distinction between quantitative "amount" data and numerical (but not quantitative) data because quantitative data lies in a metric space and therefore has a natural, Euclidean geometry, which numerical categories do not. Amount data are inherently non-negative and arise in so many contexts that analysis methods and visualizations well-suited to this class of data will have broad application.

2.1 Definition of an Object-Attribute Matrix

Triplets $<i, j, a_{ij}>$ are sufficient to define the set of edges for a weighted, directed graph, with the sets of attributes and objects defining the sets of source and target nodes, respectively, and the amounts providing the edge weights. Equivalently, a set of triplets define a sparse $m \times n$ adjacency matrix $A = \{a_{ij}\}$, where m is the number of distinct "attributes" and n the number of distinct "objects" mentioned in at least one edge. Entries a_{ij} in the (dense) adjacency matrix A without a corresponding triplet are not missing data, but mean the amount of attribute i of object j is zero. The terms "object" and "attribute" may have a context-specific meaning but have no mathematical meaning, and the sets of objects and attributes may overlap or be identical.

We will refer to any such non-negative adjacency matrix of amounts as an *object-attribute matrix*. While object-attribute matrices encompass diverse datasets, we note they specifically do *not* encompass graphs with negative edge weights, with multiple categories of edges, with multiple weights per edge, with node weights, or with multiple edges for the same object-attribute pair $<i, j>$. The object-attribute matrices we will use to illustrate the Espalier metaphor and Espalier Plots are described below.

2.2 Genotype Data: 1000 Genomes, phase 3

Large-scale genotyping studies determine the state of a large set of genetic markers over a population of DNA samples. Genotyping arrays interrogate hundreds of thousands of

nucleotides along the human genome that vary between individuals (single-nucleotide variants, SNVs); SNVs are selected for both assay reliability and relevance to ancestry, typically having two states (reference and variant) that are measured simultaneously for both the maternally-inherited and paternally-inherited chromosome. Each interrogated marker *i* is classified into one of three states: sample *j* contained the reference allele twice (homozygous reference), both the reference and variant alleles (heterozygous), or the variant allele twice (homozygous variant). Thus, a genotype array of this sort results in an object-attribute matrix $<i, j, a_{ij}>$, where $a_{ij} = 0, 1,$ or 2 is the amount of variant alleles observed at variant *j* in DNA sample *i*.

Phase 3 of the 1000 Genomes Project made public WGS data for 2,504 samples ascertained to belong to 26 populations. The variants in this data are from a combination of WGS, genotyping arrays, exome sequencing, and other methods, including imputation of some values, as described (1000 Genomes Project Consortium, 2012, 2013). As an object-attribute matrix, approximately 15% of all triplets are nonzero.

2.3 Gene Expression Data: BodyMap2

The Illumina BodyMap2 RNASeq dataset (NCBI GEO Accession No. GSE30611) contains counts of cDNA sequence reads from 16 human tissue samples (adipose, adrenal gland, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells) mapped to RNA transcripts; the identified transcripts were then processed and mapped to 29,663 genes as described (Glusman, G. and others 2013) . Triplets $<i, j, a_{ij}>$ from this dataset represent the count a_{ij} of reads, normalized by the relative length of the gene transcripts, from gene *i* detected in tissue sample *j*. As an object-attribute matrix, approximately 64% of all triplets are nonzero.

2.4 Visualizing Object-Attribute Matrices

While such data can be visualized as “hairball” weighted graph structures, such visualizations can be hard to comprehend, especially for large datasets. BioFabric (Longabaugh 2012) is a better visualization method for conveying the geometric structure of such datasets. An alternative approach is to visualize the object vectors (columns of the object-attribute matrix) in a Euclidean space with a dimensionality-reduction method such as PCA. While more general, non-linear dimensionality reduction methods have been developed, including kernel PCA (Schölkopf, Smola, and Müller 1998), Self-Organizing Maps (Yin 2007), and t-SNE (van der Maaten and Hinton 2008), PCA remains popular due to its simplicity and its implicit statistical interpretation. PCA is the first step in standard analyses of both genotype and expression data.

3. The Espalier Metaphor

Fruit trees or other garden plants are sometimes spread on a planar support, an espalier, for both aesthetic and practical reasons. The espalier transforms the 3D structure characteristic to the plant into a more convenient 2D plane, facilitating tasks such as grafting, pollination, and harvesting. Branches are a plant’s support structure, and positioning the branches on the espalier controls the positions of the leaves, flowers, and fruit.

We see the garden espalier as an appropriate metaphor for organizing the objects and attributes of an object-attribute matrix. Mathematically, we define an Espalier as a *geometrically flat* axis (an Espalier Coordinate, EC) together with a set of $m+n$ positions $\{x_i\}$, one for each attribute i ($1 \leq i \leq m$) and one for each object j ($m < m+j \leq m+n$). We require an EC to be geometrically flat in the sense that positions along an EC are additive, like positions along a line, rather than sub-additive like the x- and y-coordinates of points on a circle, or multiplicative like logarithms of amounts (the sum of the logarithms of two amounts is the logarithm of the product of the amounts, not their sum). This requirement ensures that the induced geometry of a Cartesian product of a set of ECs will be a Euclidean geometry, just as the geometry of a Cartesian product of a set of measured amounts is Euclidean.

3.1 Principal Components are not Espaliers

PCA finds and orders axes within a high-dimensional Euclidean space in order to capture as much variance as possible in the first k dimensions. Standard PCA first normalizes each attribute to the corresponding z-score ($Z = \{ z_{ij} \} = (a_{ij}-\mu_i)/\sigma_i$). The eigenvalues and eigenvectors of the correlation matrix Z^TZ are found by singular value decomposition of Z (SVD; Beltrami 1873). The right eigenvectors are ranked in decreasing order of the corresponding eigenvalue, producing the principal components (PC_1, PC_2, …) of the object vectors; the left eigenvectors correspond to the variable (attribute) "loadings". The first k coordinates PC_1, …, PC_k computed by Standard PCA capture the most Pearson correlation possible for a linear reduction of the object vectors to a k-dimensional subspace.

The normalization to z-scores used in Standard PCA is motivated by statistical ideas about variance and Pearson correlations, not the geometric ideas of an Espalier. In particular, the variable loadings vary systematically with the total amount in each row, and the relationships between individual objects and attributes are neither emphasized nor visible in plots of principal components and the associated variable loadings.

3.2 Key Espalier Analysis: Constructing Espaliers by Design

The singular value decomposition used to compute principal components, however, has a natural geometric interpretation. Let u and v be a singular vector pair for an m x n nonnegative matrix M with singular value s; then M, s, u, and v satisfy the system of equations $u^TM = sv^T$, $Mv = su$. While PCA focuses on u and v as eigenvectors of the covariance (or in Standard PCA, correlation) matrix M^TM, geometrically the defining equations for singular vector pairs above identify u in R^m and v in R^n as *corresponding directions* invariant under multiplication by M. SVD identifies the full set of such correspondences intrinsic to the matrix M, along with the scaling constants s, and fully defines the bilinear, geometric transformation between R^m and R^n expressed in M as scaling along these corresponding directions.

Note however that u and v are direction vectors confined to the unit sphere in their respective spaces. These spheres are curved manifolds and are not additive, i.e. the sum of two left (or right) singular vectors for the same matrix M does not always lie on the unit sphere. However, we can project u and v into corresponding additive manifolds; specifically, geometrically flat manifolds tangent to the unit spheres at corresponding points. Letting u_0, v_0 be another left and right singular vector pair for M, note that u_0 and

u are orthogonal, and their component-wise ratio $u' = u/u_0$ is constrained to the (flat, additive) manifold tangent to the R^m unit sphere at u_0; by the same reasoning, $v' = v/v_0$ is constrained to the additive manifold tangent to the R^n unit sphere at v_0. This projection completes construction of a true Espalier, in which the objects and attributes have corresponding *positions* in a geometrically flat space, rather than corresponding *directions* in a geometrically curved space.

Like principal components, Espaliers can be ordered by their corresponding singular values, and the first k Espalier components provide a representation of the objects in a k-dimensional space that preserves the k most significant independent relationships between the objects and attributes. So long as the matrix is not too sparse, an appropriately modified version of the Sinkhorn and Knopp (1967) algorithm can find strictly positive vectors d_m of length m and d_n of length n so that the row and column vectors of the scaled matrix $W = diag(d_m)\, A\, diag(d_n)$ have constant length; i.e. the rows and columns lie on spheres in R^m and R^n, respectively. W is therefore a scaled version of A in which the differences between the rows and between the columns are entirely directional. This is the motivation for the normalization: to convert the length and direction information in the rows and columns of A into purely directional information in preparation for SVD, which finds a full set of corresponding directions between the space spanned by the row vectors and the space spanned by the column vectors of a matrix. The pairs of corresponding singular vectors for W therefore capture as much of the correlation structure of A as possible for any scaling of A.

In addition, normalizing the object-attribute matrix in this way renders the correlation between the EC positions of the objects and attributes independent of the *total amount* of data for each attribute and each object; it is therefore natural to visualize the objects and attributes on each Espalier coordinate against an appropriate measure of the amount of data for each attribute and object, the row and column sums of the object-attribute matrix. We call this visualization an Espalier Plot.

We call the Espalier construction method Key Espalier Analysis (KEA). In summary, A is scaled to W by finding the vectors d_m and d_n; the first k corresponding singular vectors of W are found using a sparse SVD algorithm such as the implicitly restarted Lanczos bidiagonalization algorithm (Baglama and Reichel, 2006); and each singular vector is projected to the geometrically flat tangent space by division by another singular vector pair. Due to the non-negativity of object-attribute matrices the singular vector pair corresponding to the largest singular value is typically strongly correlated with total amount, and we use this pair as the divisor. The resulting coordinates we call Espalier coordinates (ECs) in analogy to Principal Components (PCs). In the next two sections, we compare the results of PCA and KEA on two kinds of object-attribute matrices that arise in biomedical studies.

4. Espaliers for Genotype Data

The most common visualization of population structures is the PCA biplot (Fig.1, upper left). The DNA samples (colored dots) are positioned using their first few PCs; rarely, the variable loadings (black dots, one per variant) are also shown. Although the variable loadings and principal components of the DNA samples are mathematically related as left and right singular vectors, these relationships are generally not visually apparent in a

PCA biplot even when the variable loadings are shown, and provide little evolutionary insight.

The other plots in Figure 1 are Espalier Plots, showing the objects (DNA samples) and attributes (variants) of an object-attribute matrix on the horizontal axis against a measure of data quantity on the vertical axis. Here the measure of data quantity is the variant allele count across all samples, plotted on a logarithmic scale.

The Espalier Plot for PC1 (Fig.1, lower left) reveals that PCA systematically assigns a broader range of PC values to more common variants than to rarer variants, and PCA therefore organizes the DNA samples primarily by the frequencies of common variants. Since time is required for a new variant allele to reach high frequency in the population, rarer variants are expected to have arisen more recently (Kimura and Ohta, 1973), and the older, more common variants are believed to have arisen prior to the earliest separations of human populations, the so-called “Out of Africa” event(s). Differences in frequency of common variants between populations are believed to be due to the effect known as “genetic drift”: stochastic differences in the rate at which variant alleles are retained in independently-evolving populations.

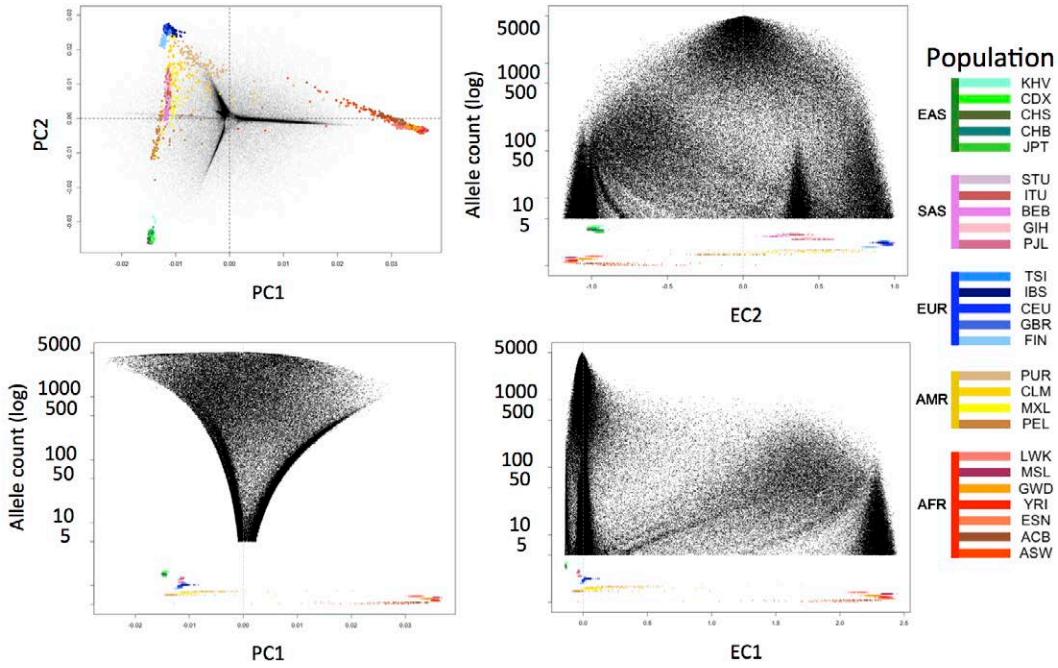

Figure 1: Espalier Plots. *Upper left*: Conventional PCA biplot showing DNA samples (colored dots) on the first two principal coordinates, augmented by corresponding variable loadings (black points). *Lower left*: Espalier plot of PC1 (horizontal axis) and variant allele count (vertical axis, log scale). *Lower right*: Espalier Plot showing the corresponding first Key Espalier Analysis (KEA) Espalier coordinate. PCA assigns large values to common variants, while KEA assigns large values to rare variants. *Upper right*: Espalier Plot of the second Espalier component. DNA samples are colored by 1000 Genomes population of origin (legend).

Conversely, the rarest variants are believed to have arisen by mutation after human populations separated geographically. O'Connor and others have recently shown (2015) that removing the common variants prior to performing PCA permits derivation of more accurate and detailed population structures. Espalier plots with ECs derived by KEA as the horizontal axis display vertically aligned subsets of variants, with a broader range of EC values for the rarer variants, and the resulting visualization approximates a phylogenetic tree. PCA and KEA therefore appear to be sensitive to the outcomes of different evolutionary forces: PCA to genetic drift, and KEA to patterns of mutation over time.

While PCA and KEA are both based on SVD of the same data, the normalization of the data differs and a comparison of PCA and KEA biplots (Fig. 2) shows that the methods lead to qualitatively different organizations of the data. Compare, for instance, the different positions of the genomes (objects) from African populations in biplots of PC2 and PC3 versus EC2 and EC3 (Fig. 2, upper plots). The two analyses differ dramatically in the organization of the variants (attributes): in the PCA biplots (Fig.2, left plots) the variants are organized in linear structures involving the origin, while in the KEA biplots (Fig.2, right plots) subsets of variants cluster around subsets of objects and along lines between objects. These biplots demonstrate that KEA exposes the richly interrelated structure of the genotype-variant matrix in a comprehensible, visibly geometric form.

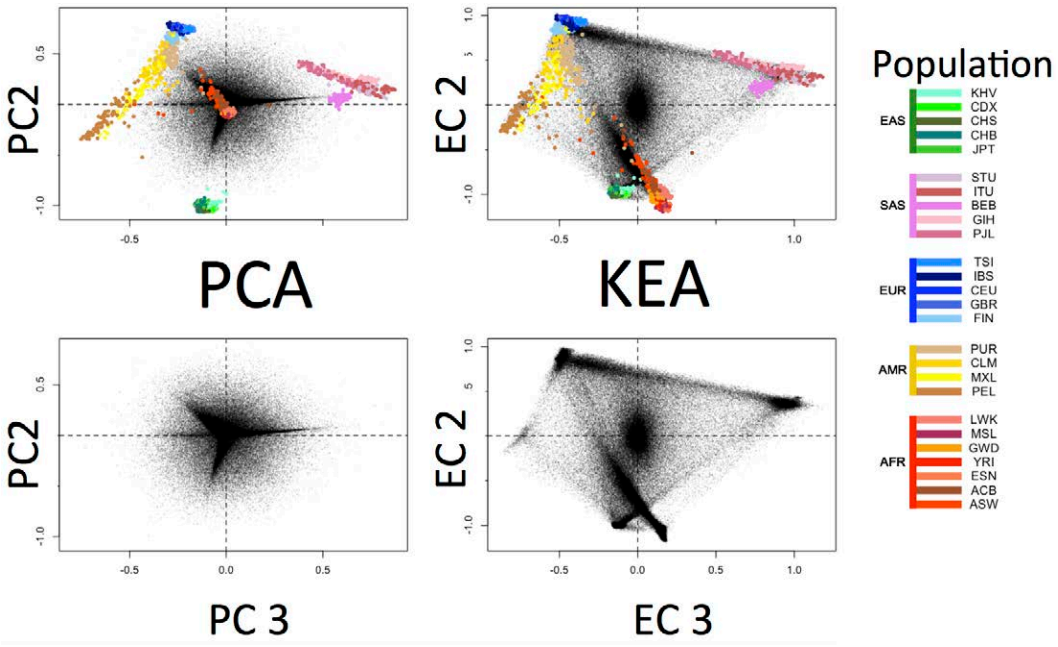

Figure 2: PCA and KEA biplots for genotype data. Same 1000 Genomes object-attribute matrix as shown in Fig. 1, but shown as biplots on coordinates determined by PCA or KEA. Genomes (objects) are shown as colored dots; variants (attributes) are shown as translucent black points. *Upper left*: PCs 3 (x-axis) and 2 (y-axis) determined by PCA; attributes correspond to variable loadings rescaled to the same range as the object principal components. *Upper right*: EC3 (x-axis) vs. EC2 (y-axis) biplot. KEA computes EC coordinates for both objects and attributes, scaled to have the same range. *Lower figures*: Same as upper figures, but showing only the attributes.

5. Espaliers for Expression Data

When applied to the Illumina BodyMap2 gene expression dataset (Fig.3), KEA again finds correspondences between objects (tissue types) and attributes (genes) that are independent of total amounts (gene expression level summed across all tissues; Fig.3, lower right). In Espalier Plots (Fig.3, lower right), genes expressed in a single tissue align directly above the tissue. In KEA biplots (Fig.3, upper right), note that tissues with known functional similarities, such as heart and skeletal muscle, kidney and liver, and tissues of the immune system such as lymph nodes and white blood cells, are organized along lines marked by dense, linear subsets of genes; upon examination, these gene subsets are observed to have higher expression in the aligned tissues relative to other tissues. Comparison of tissue and gene positioning is therefore informative about both tissue function and gene function.

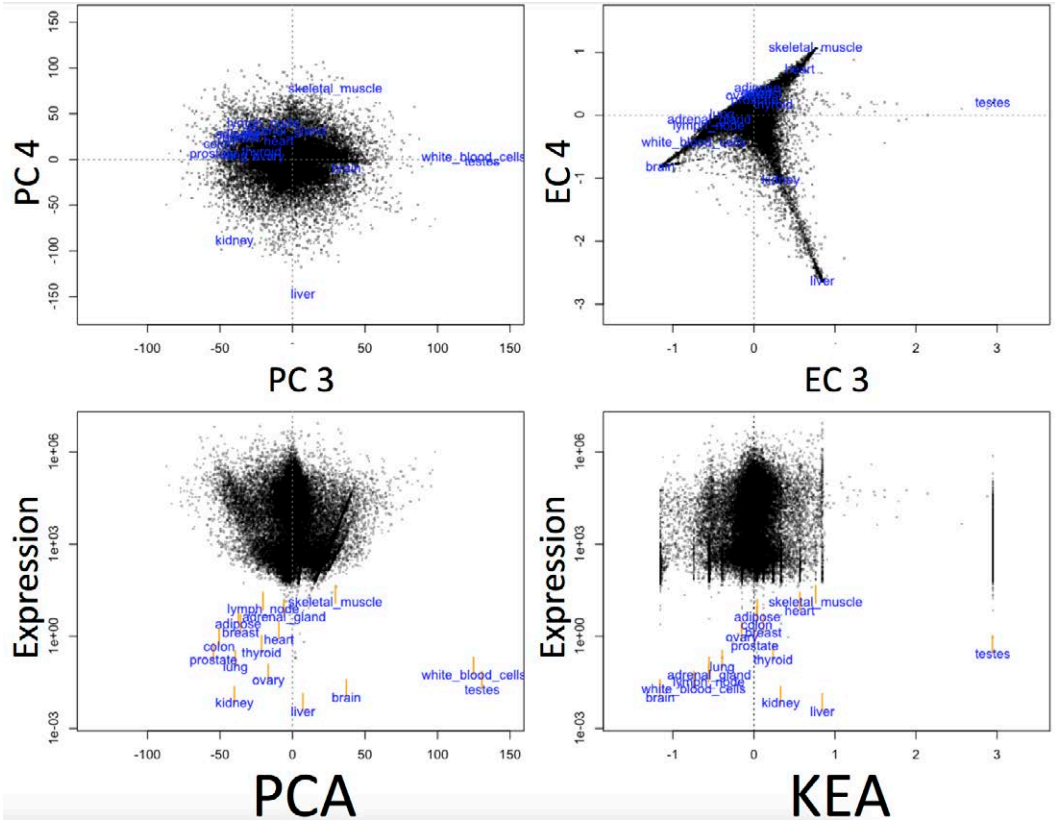

Figure 3: PCA and EC biplots for expression data. The BodyMap2 dataset visualized using PCA or KEA. Tissue samples are labeled in blue; genes are shown as translucent black points. *Upper left*: PC3 (x-axis) and PC4 (y-axis) determined by PCA; attributes correspond to variable loadings rescaled to the same range as the object principal components. *Lower left*: Espalier plot of gene expression level vs. PC3. *Upper right*: EC3 (x-axis) vs. EC4 (y-axis) biplot. KEA computes EC coordinates for both objects and attributes, scaled to have the same range. *Lower right*: Espalier plot of gene expression level vs. EC3.

In contrast, organization of tissues in the PCA biplot (Fig.3, upper left) is more dispersed and does not include linear arrangements of tissues, nor does the organization of the genes guide interpretation of tissue positions. In the PCA-derived Espalier Plot (Fig.3, lower left), some lines of genes are discernable, however these structures are not

vertically aligned, nor is their association with a specific tissue visible. Thus while PCA is guaranteed to maximize the variance in the data represented in a given number of dimensions, the Espalier coordinates provide a geometric correspondence between objects and attributes that guides insight and interpretation of the relationships between gene expression and tissue function in both Espalier Plots and biplots.

6. Conclusions

We have introduced the garden espalier as a useful metaphor for visualization of relationships between objects and the attributes on which they have been measured. This metaphor is applicable to a wide variety of "amount" data that is being collected and analyzed in massive quantities in this age of "Big Data". We have also shown how to realize this metaphor mathematically via Key Espalier Analysis, which is based on a novel, geometrically motivated normalization, followed by singular value decomposition and a novel projection into the geometrically flat tangent manifold.

The geometric nature of Espaliers is valuable for both visualization and computational analysis of objects and attributes represented on Espalier coordinates, and the ability of Key Espalier Analysis to expose the inherent geometric structures of object-attribute matrices makes the Espalier metaphor an important new tool in the analysis of Big Data across many applications. As shown here, KEA maps objects and attributes in a manner that allows the relationships between them to be visible as simple geometric structures. KEA is particularly useful for identifying specific attributes as markers for the objects with which they are closely associated. This identification of markers is particularly visible in Espalier Plots, in which marker attributes are vertically aligned above the associated object(s). In KEA biplots, the geometric nature of the bilinear transformation an object-attribute matrix represents results in linear structures that are likely to be interpretable in the context from which the matrix arose, as discussed for the gene expression dataset. In future work we intend to demonstrate the value of this "marker" relationship, and to provide intuitive interpretations for the linear structures visible in KEA biplots of genotype data.

Acknowledgements

This work was supported by the Inova Translational Medicine Institute, and by NIH grants P50 GM076547 and U54 EB020406.

References

1000 Genomes Project Consortium (2012), "An integrated map of genetic variation from 1,092 human genomes", *Nature* 491(7422):56–65, DOI: 10.1038/nature11632.

1000 Genomes Project Consortium (2013), Phase 3 processing description available at "ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/README_phase3_callset_20150220".

Baglama, J. and Reichel, L. (2006), "Restarted Block Lanczos Bidiagonalization Methods", *Numerical Algorithms* 43:251-272.

Beltrami, E. (1873), “Sulle funzioni bilineari”, Giornale di Matematische ad Uso degli Studenti Delle Universita 11:98-106.

Glusman, G., Caballero, J., Robinson, M., Kutlu, B., and Hood, L. (2013). “Optimal scaling of digital transcriptomes”, *PLoS One* 8(11):e77885. DOI: 10.1371/journal.pone.0077885; erratum in: *PLoS One* 9(3):e93244 (2014), DOI: 10.1371/annotation/8b05a9ab-c8ad-4276-a851-1e265055fb65.

Kimura, M., and Ohta, T. (1973), “The age of a neutral mutant persisting in a finite population”, *Genetics* 75:199-212.

Lander, E. S. and the International Human Genome Sequencing Consortium (2001), “Initial sequencing and analysis of the human genome”, *Nature* 409:860-921.

Longabaugh WJ. (2012) “Combing the hairball with BioFabric: a new approach for visualization of large networks”, *BMC Bioinformatics* 13:275. doi: 10.1186/1471-2105-13-275.

O’Connor, T.D., Fu, W., NHLBI GO Exome Sequencing Project; ESP Population Genetics and Statistical Analysis Working Group, Turner, E., Mychaleckyi, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal, S.M., Smith, J.D., Reider, M.J., Bamshad, M.J., Nickerson, D.A., and Akey, J.M. (2015), “Rare variation facilitates inferences of fine-scale population structure in humans”, *Molecular Biology and Evolution*, 32(3):653-660. DOI 10.1093/molbev/msu326.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space", *Philosophical Magazine* 2(11):559–572. DOI:10.1080/14786440109462720.

Schölkopf. B., Smola, A., and Müller, K.-R. (1998), “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”, *Neural Computation* 10(5):1299-1319, doi:10.1162/089976698300017467.

Sinkhorn, R., and Knopp, P. (1967), "Concerning nonnegative matrices and doubly stochastic matrices", *Pacific J. Math.* 21:343–348.

van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing High-Dimensional Data Using t-SNE", Journal of Machine Learning Research 9: 2579–2605.

Venter, J. C., Adams, M. D., Myers, E. W. and others (2001), “The Sequence of the Human Genome”, *Science* 291(5507):1304-1351.

Yin, H. (2007), “Learning Nonlinear Principal Manifolds by Self-Organising Maps”, 58:68-95 in A.N. Gorban, B. Kégl, D.C. Wunsch, and A. Zinovyev (Eds.), “Principal Manifolds for Data Visualization and Dimension Reduction”, *Lecture Notes in Computer Science and Engineering*, Berlin, Germany: Springer.