Journal of Bioinformatics and Computational Biology
Vol. 15, No. 4 (2017) 1750017 (14 pages)

DOI: 10.1142/S0219720017500172

Application of t-SNE to human genetic data

Wentian Li[*,§], Jane E. Cerise[*,¶], Yaning Yang[†,‖] and Henry Han[‡,**]

[*]*Robert S Boas Center for Genomics and Human Genetics*
Feinstein Institute for Medical Research
Northwell Health, Manhasset, NY 11030, USA

[†]*Department of Statistics and Finance*
University of Science and Technology of China
Hefei, Anhui, China

[‡]*Department of Computer and Information Sciences*
Fordham University, Lincoln Center
New York, NY, USA
[§]*wli@northwell.edu*
[¶]*jcerise@northwell.edu*
[‖]*ynyang@ustc.edu.cn*
[**]*xhan9@fordham.edu*

Received 30 March 2017
Accepted 30 March 2017
Published 18 July 2017

The t-distributed stochastic neighbor embedding t-SNE is a new dimension reduction and visualization technique for high-dimensional data. t-SNE is rarely applied to human genetic data, even though it is commonly used in other data-intensive biological fields, such as single-cell genomics. We explore the applicability of t-SNE to human genetic data and make these observations: (i) similar to previously used dimension reduction techniques such as principal component analysis (PCA), t-SNE is able to separate samples from different continents; (ii) unlike PCA, t-SNE is more robust with respect to the presence of outliers; (iii) t-SNE is able to display both continental and sub-continental patterns in a single plot. We conclude that the ability for t-SNE to reveal population stratification at different scales could be useful for human genetic association studies.

Keywords: t-SNE; PCA; SNP; dimension reduction.

1. Background

Genome-wide association study (GWAS)[1–6] is an approach to type common genetic variants in a general human population, and by comparing the allele frequency difference between a group of patients (cases) and a group of normal people (controls), discover statistical association signals. The genetic variant mostly encountered is the single nucleotide polymorphism (SNP).[7] These associated variants may

[§]Corresponding author.

reside in a protein coding region, indicating a possible change of transcription products which may play a role in the disease. It may also sit between genes which probably change a binding motif of a transcription factor which leads to a change of the transcription (expression) level.[8] A comprehensive database on statistically significant variants for many human diseases is the GWAS Catalog, first maintained by the NHRGI/NIH (National Human Genome Research Institute of National Institute of Health (USA)), then by the European Bioinformatics Institute (UK), which can be found at https://www.ebi.ac.uk/gwas/.[9,10]

A key step in GWAS analysis is to match the ethnicity background of cases and controls. Failure to do so would confound allele frequency difference due to disease-causing mutations and difference due to population genetic history. There have been many attempts to correct this "spurious association" due to "population stratification"[4,11]: "genomic control" uses non-functional variants to estimate the amount of population genetic history differences and the association signal is corrected accordingly;[12] "family-based association" uses untransmitted alleles as controls thus circumventing the population stratification issue completely when these type of data are available;[13–15] K-mean clustering to group sample genotype data towards the presumed K populations;[16] incorporating co-variance among samples to correct the association signal,[17–19] etc.

However, the most common practice in dealing with population stratification or other subtle/hidden structures in genetic data is to perform dimensional reduction techniques, such as MDS (multi-dimensional scaling), (principal component analysis PCA) and SVD (singular value decomposition). The reduced dimensions can be directly visualized and, in the case of PCA, can be used as covariates in the association analysis.[20–26] Within the European populations, it is well established that the first PC (PC1) aligns with the north–south direction (latitude), and the second PC (PC2) aligns with the east–west direction (longitude).[27]

Though the use of PCA is mostly satisfactory, the method is not without a problem. Most notably, PCA is highly affected by the presence of outliers. If most of the samples belong to one homogeneous population with a minority from another different population, the presence of genetically different minority samples can completely change the principal axes, thus changing the distribution of samples along the main PCs. This is because as a linear holistic dimension reduction method, PCA cannot capture local data characteristics well. Other problems include the determination of the number of SNPs to be included, the role common versus rare variants play in the result, and the number of PCs to be kept.

Here, we explore the application of a new dimensional reduction technique, t-SNE (t-distributed stochastic neighbor embedding),[28] to the genetic data. The t-SNE is one type SNE (stochastic neighbor embedding).[29] Similar to MDS, the aim of t-SNE is to preserve the pairwise distance in high-dimensional space to 2 or 3 lower dimensions. Unlike MDS or PCA, the preservation in t-SNE is nonlinear: t-SNE minimizes the Kullback–Leibler divergence between two distributions — one distribution that measures pairwise similarities of input samples in high-dimensional

space, another heavy-tailed Student's t-distribution that measures pairwise similarities of corresponding samples in the low-dimensional embedding space. t-SNE has demonstrated its built-in advantages in capturing local data characteristics and revealing subtle data structures in visualization, as shown in the original publication.[28] It is an embedding visual analytic algorithm that preserves the similarity and dissimilarity between data points in the low-dimensional embedding space.

t-SNE is a popular choice in the analysis of single-cell RNA-seq data (e.g. Refs. 30–32), but has not been applied extensively to genetic data. In the single previous publication where t-SNE was applied to genetic data,[33] the main conclusion is that if t-SNE is considered as a clustering technique, it performs better than PCA. We are particularly interested in t-SNE's claimed ability to "reveal structure at many different scales",[28] as major population stratification co-exists with other small-scaled shared evolutionary history among samples.

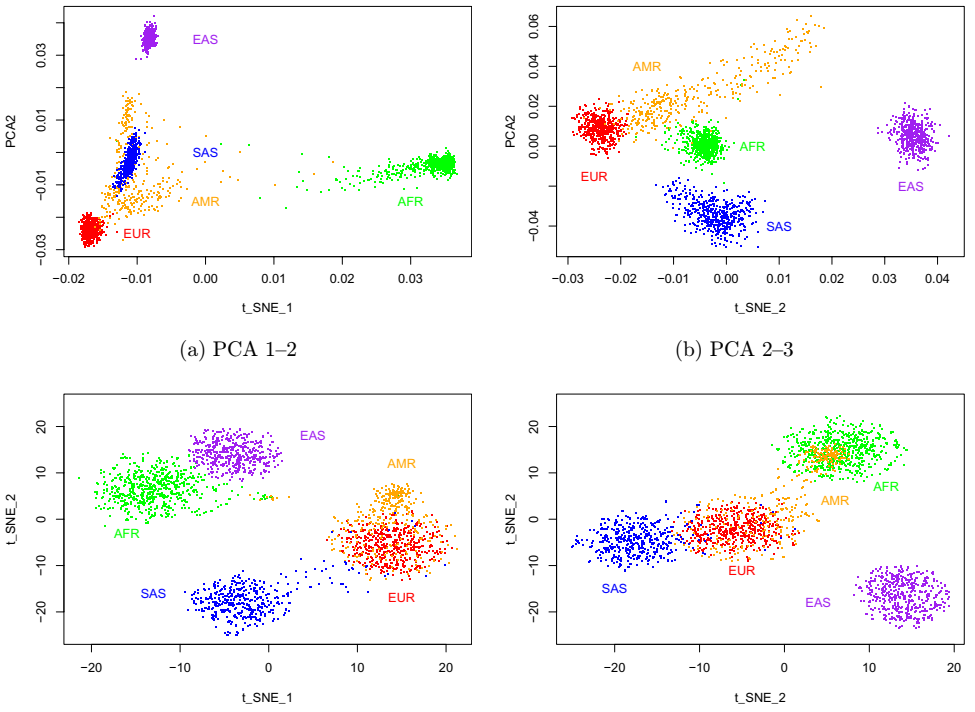
2. Results

2.1. *Continental separation*

We first examine the ability of t-SNE to separate continental populations (Africa, Asia, Europe, etc.). GWAS usually refers to genetic association studies using common variants. The new next-generation-sequencing (NGS) technique, though aiming at typing all variants, also produces common variants. We use one of the major public NGS data, the 1000 Genomes Project.[34–36] We extracted 3825 common variants which pass a quality control (QC) criteria, and are also present in the Illumina Global Screening Array chip. We use the KING program to calculate the inter-sample genetic distances. We override the default setting of KING to retain only three significant digits so that nine significant digits are kept, in order to improve our ability to distinguish subtle structures.

Figure 1 shows the results from PCA (top) and t-SNE (bottom) with the first and the second major dimensions (left column) and the second and third dimension. The number of samples in each continent (though Asia is split into east and south Asia) are more or less balanced: 661 Africans (AFR), 347 "Americans" (AMR), 504 East Asians (EAS), 489 South Asians, and 503 Europeans (EUR). Note that some subgroups living in continental America is not grouped with AMR: African–American in Southwest of US (ASW) and African–Caribbean in Barbados are in the AFR group, Utah CEPH families are in the EUR group, etc.

Although all methods are able to separate continental populations, PCA (1–2 dimensions) shows an overlap between South Asian and American, whereas t-SNE shows AMR has more overlap with Europeans. In the 2–3 dimensions, PCA shows some link between AMR and EUR, whereas t-SNE continue to show a strong connection between AMR and EUR. As some AMR samples, such as those from Colombia, Puerto Rico, and to some extent, Mexico (actually the samples are Mexican–American), are expected to contain European ancestor, the t-SNE result is



(a) PCA 1–2

(b) PCA 2–3

(c) t-SNE 1–2

(d) t-SNE 2–3

Fig. 1. A comparison of three different low-dimensional displays of 1000 Genomes Projects data with 3825 SNPs. The number of samples (dots) is 2504, with 661 AFR (African green), 347 AMR (American, orange), 504 EAS (East Asia, purple), 489 SAS (South Asia, blue), and 503 EUR (European, red). (a) PCA (PC1–PC2); (b) PCA (PC2–PC3); (c) t-SNE (1–2) and (d) t-SNE (2–3).

consistent with our external knowledge. As can be seen in the 3D version of t-SNE (Fig. 2), the overlap between AMR and AFR in Fig. 1(f) is an artifact, as AMR and AFR are actually separated.

2.2. *Treatment of outliers*

We compare t-SNE and PCA in a more realistic setting where most of the samples belong to one ethnic group, whereas a few are either from distinct ethnic groups or a mixed race. For that purpose, we extract 99 Utah residents with North/West European ancestry (CEU), 91 England/Scotland samples (GBR), and five African–American in Southwest (ASW), five Mexican–American in Los Angeles (MXL), two Chinese in Beijing (CHB), two Chinese in South China (CHS).

PCA (Fig. 3(a)) moves ASW, CHB/CHS far away from the main cluster of roughly 200 Caucasians, whereas MXL samples, though still separated, are closer to the center. However, though t-SNE (Fig. 3(b)) shows ASW, CHB/CHS as separated groups, the MXL samples are much closer to other Caucasian samples. In order to see

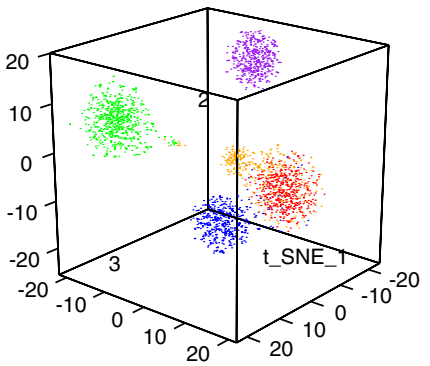
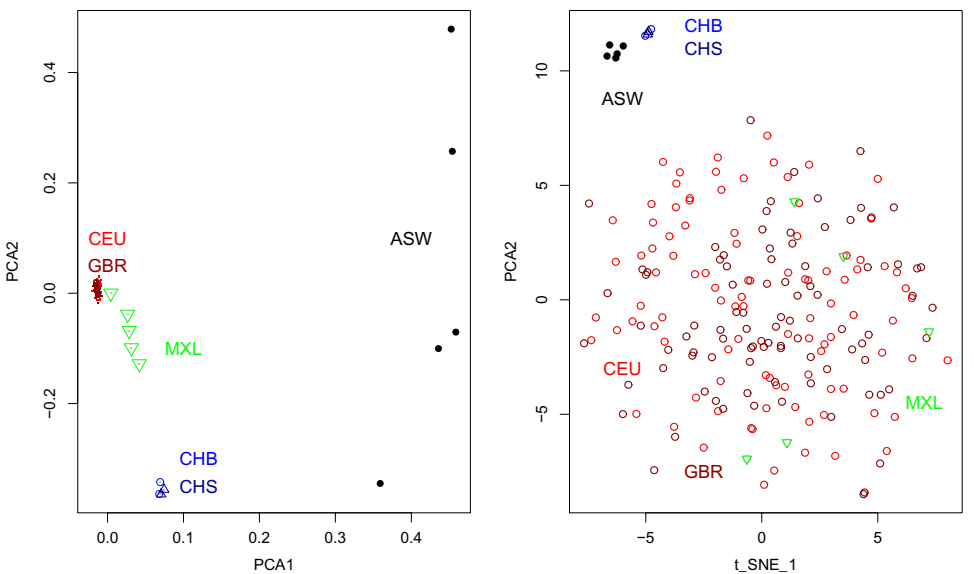


Fig. 2. Three-dimensional t-SNE which combines information from Figs. 1(c) and 1(d). Color scheme: green for AFR, orange for AMR, purple for EAS, blue for SAS, and red for EUR.

the structure within the Caucasian group through PCA, the usual procedure is to remove the "outliers" and re-run PCA again (e.g. recommended in smartpca). On the other hand, an advantage of t-SNE is to show both the "outliers" and the main cluster with detail simultaneously, capturing subtle local data structures and preserving global structures in the low-dimensional space.



(a) PCA (b) t-SNE

Fig. 3. A comparison of PCA and t-SNE on a mostly Caucasian dataset (99 Caucasians from Utah (CEU), USA and 91 Caucasians from UK (GBR)) with "outliers" included (five African–American in Southwest (ASW), five Mexican–American in Los Angeles (MXL), two Chinese in Beijing (CHB), two Chinese in South China (CHS)). (a) PCA (PC1–PC2) and (b) t-SNE.

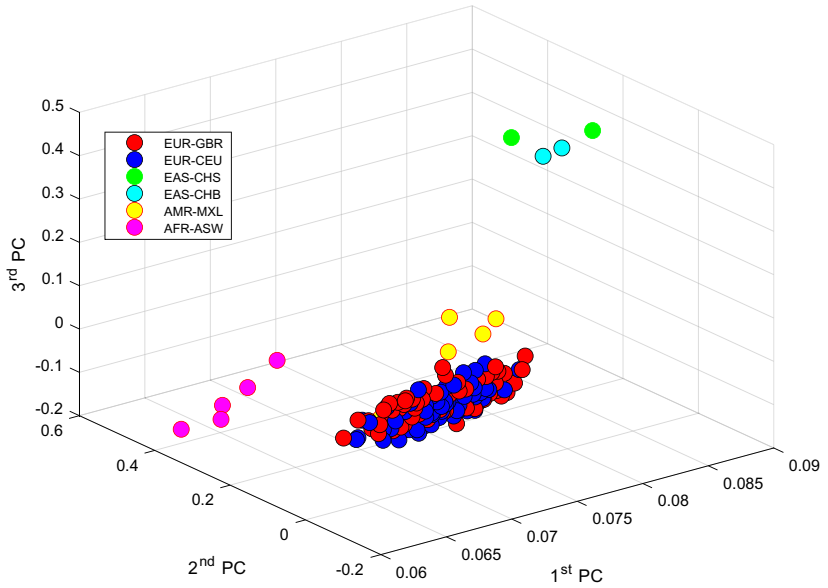


Fig. 4. Non-negative PCA (NPCA) of the data used in Fig. 3.

To see whether some PCA modifications may exhibit similar properties as t-SNE, we test two PCA variants: non-negative PCA (NPCA)[37,38] and derivative component analysis (DCA).[39] The purpose of imposing extra constraints in NPCA is that positive and negative terms in classic PCA may cancel each other, leading to a loss of local feature. The purpose of DCA is to use derivatives to capture latent patterns and to suppress noise level. Figures 4 and 5 show the 3D NPCA and DCA plots. Although NPCA and DCA are better than classic PCA in showing more spreading in CEU/GBR cluster, they are more consistent with classic PCA than with t-SNE.

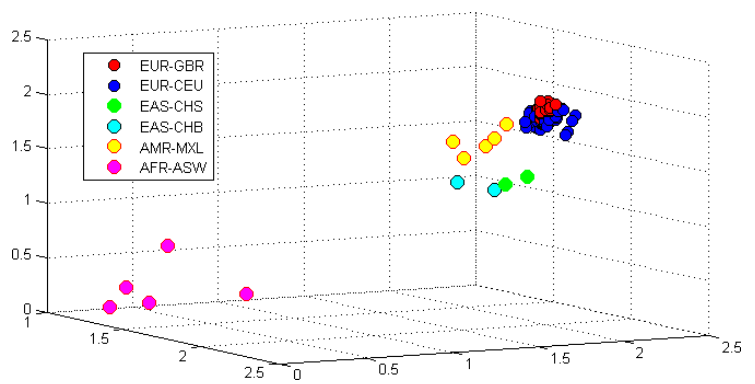


Fig. 5. Derivative component analysis (DCA) of the data used in Fig. 3.

2.3. *Sub-continental separation*

We examine whether t-SNE is able to display sub-continental patterns besides the continental separation. For that, we extract a denser set of SNPs from chromosome 1 in the 1000 Genomes Project with the criteria of alternative/minor allele frequency > 0.2, and spacing between neighboring SNPs $> 20,000$ bases. This leads to more than 9000 SNPs, roughly equivalent to a genome-wide $+100,000$ SNPs. The use of a single chromosome to represent the whole genome is justified, as PCAs based on any human chromosome are almost identical (Fig. S3 of Ref. 40). Only in the extreme case of using SNPs from a region with inversion, may the shape of PCA be different,[40,41] as the trimodal distribution of the samples reflects the three underlying configurations.[42]

Figure 6 shows the t-SNE with a finer group labeling. We paid particular attention to choose colors to represent a population's geographic information within the continent, which is shown in Fig. 7. Generally, we choose a dark color for the population towards the north, and a light color for those in the south. We also switch the t-SNE-1 and t-SNE-2 so that the north groups point to the up direction.

The lower-left cluster in Fig. 6 contains five groups in East Asian: they are, in the rough ordering from north to south, Japanese in Tokyo (JPT), Han Chinese in

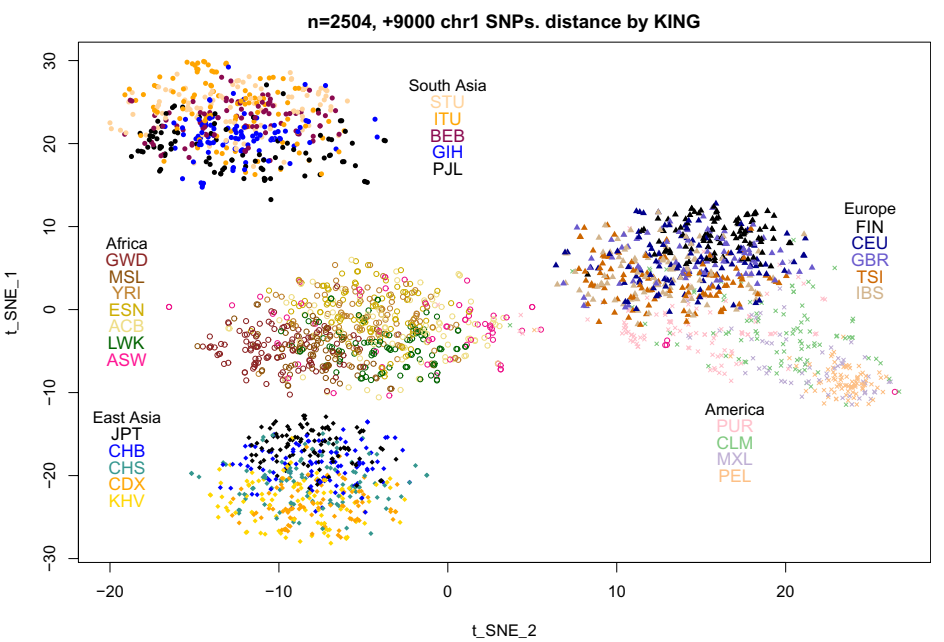


Fig. 6. t-SNE plot obtained from more than 9000 SNPs on chromosome 1 of the 1000 Genome Projects (alternative allele frequency > 0.2, spacing between neighboring SNPs longer than 20,000 bases. The explanation of group labels in Africa (ACB, ASW, ESN, GWD, LWK, MSL, YRI), America (CLM, MXL, PEL, PUR), East Asia (CDX, CHB, CHS, JPT, KHV), South Asia (BEB, GIH, ITU, PJL, STU) and Europe (CEU, FIN, GBR, IBS, TSI) are explained in the text.

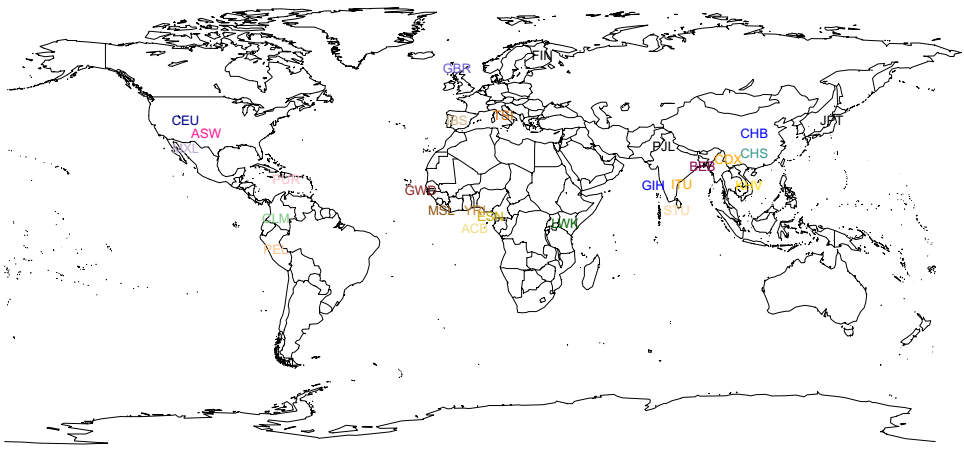


Fig. 7. A world map where the location of the 26 populations are marked. Note these specific choices: the Utah Caucasian CEU is not moved to Europe, and southwest African American ASW is not moved to Africa; the African Caribbean in Barbados ACB is moved to the west coast of Africa; the Mexican-American in California MXL is moved to Mexico.

Beijing (CHB), Han Chinese in South China (CHS), Chinese Dai in Xishuangbanna (CDX), and Kinh in Ho Chi Minh City, Vietnam (KHV). The change of color from light (south) to dark (north) can be clearly seen. However, the CHS points are more scattered.

The upper-left cluster in Fig. 6 is the South Asia (India subcontinent) populations, with this ordering in color lightness: Punjabi in Lahore Pakistan (PJL), Gujarati Indian in Houston US (GIH), Bengali in Bangladesh (BEB), Indian Telugu in UK (ITU), and Sri Lankan Tamil in UK (STU). There is no question that PJL in the north, and STU/ITU in the south should be the two ends of the ordering. The GIH to BEB is a direction from west to east. Nevertheless, the trend shading change from top to bottom is clear.

The African populations (middle-left cluster) are mostly limited to the western Africa: Gambian in Western Gambia (GWD), Mende in Sierra Leone (MSL), Yoruba in Ibadan, Nigeria (YRI) and Esan in Nigeria (ESN). The African Caribbean in Barbados (ACB) should also be of a western Africa origin. Only the Luhya in Webuye, Kenya (LWK) group is on the east coast. The origin of African–American in Southwest US (ASW) samples is not clear, so we use a distinctly different color. Interestingly, several ASW samples are closer to the America cluster, indicating a genetic admixture with native Americans. The west/east representatives of GWD and LWK do seem to anchor the two ends of the cluster in Fig. 6.

The European populations are color ordered by this sequence: Finnish (FIN), Utah American (CEU), British (GBR), TSI for Toscani in Italy, and IBS for Iberian in Spain. Again, we observe a better separation between the two extremes: FIN at the north, and TSI/IBS at the south, with CEU and GBR more scattered.

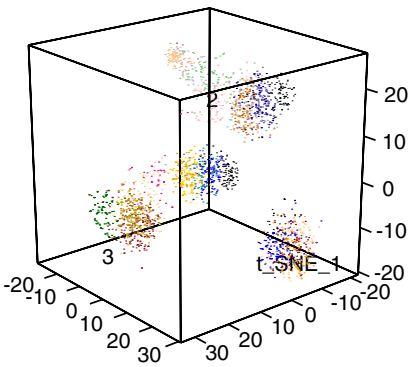


Fig. 8. Three-dimensional t-SNE which expands the plot in Fig. 4. The cloud of dots on left is the AFR group, that on the right is SAS. The middle cluster is EAS, and the top cluster is ERU with a "tail" towards AMR.

American samples are the only one which do not form its own cluster, due to the well-known admixture between indigenous America population and European colonists. Although we do not know the details of individual samples, most AMR samples do not seem to have a large Africa admixture (with a few exceptions). The Peruvian in Lima (PEL) samples are furthest away from the EUR cluster, which could be the center for a native America cluster. The Mexican–American in LA, US (MXL), Colombian in Medellin (CLM) and Puerto Rican (PUR), appear to form a gradient extending from the PEL samples to the EUR samples.

Finally, a 3D t-SNE plot is shown in Fig. 8. It illustrates both the continental separation in 4–5 large clusters, as well as sub continental group patterns with gradient of colors present with each cluster as discussed above. In comparison, the PCA plot (not shown) has sub-populations belonging to the same continent tightly squeezed, making the separation among them difficult.

3. Discussion

Application of PCA to genetic data to display sample population stratification is a common practice,[43] and the generic form of PCA plot is easily interpreted (e.g. Supplement Fig. S4 of Ref. 35). However, using t-SNE is new. We show in Fig. 1 that t-SNE can separate population from different continents as well as PCA.

Although some problems in the application of PCA to revealing population stratification remain in t-SNE, such as identification of the optimal number of SNPs to keep in the data in order to best show the population structure, whether rare variants should be used, etc. other problems are not an issue in t-SNE. For example, which high-order PCs should be kept to reveal the full population structure. In principle, one can use the proportion of variance explained to select the number of PCs. However, it is not clear what the threshold cutoff should be, and how to meaningfully interpret the contributions of a higher-order PC in visualization

compared to its lower-order ones. We are less concerned about this issue with t-SNE, as t-SNE conducts dimension reduction by optimizing Kullback–Leibler distance between the raw data distributions, and a low-dimensional distribution in at most three-dimensional space, instead of via a PC selection.

The main feature of t-SNE, the ability to exhibit structures at multiple scales, is responsible for two observations in this paper. One is that an outlier in PCA does not appear to be an outlier in t-SNE. In other words, the display of the pattern in t-SNE is more robust against a small number of outliers. This property might be compared to robust PCA,[44] nonlinear PCA,[45,46] local PCA[47] (note the meaning of "local" in this local PCA paper[40] refers to local genomic regions), or other robust methods in dimension reduction.[48] Another observation that both continental and sub-continental population structures can be viewed in a single t-SNE plot is also a consequence of this property of t-SNE. t-SNE is better than PCA at characterizing local structures, while equally well at preserving global structures.

The time computational complexity of standard PCA/SVD is $O(\min(Np^2 + p^3, pN^2 + N^3)$ where N is the number of samples and p the number of factors.[49] If $N < p$, the computational complexity is $O(N^3)$. On the other hand, the computational complexity of t-SNE is $O(N^2)$,[28] which has an advantage over PCA. Of course, in specialized applications such as sparse matrix, approximate results, or partial results, the computational complexity can be improved.[50–52]

In conclusion, though the current application of t-SNE in genomics is mostly limited to gene expression data such as those in single-cell RNA-seq, we found it quite useful in human genetic data, being able to display.

4. Data and Methods

1000 Genomes Project data: The 1000 Genomes Project is a major effort to sequence the whole genome of more than 1000 samples.[34–36] The phase 3 1000 Genomes Project data for 2504 individuals was downloaded from *http://www.internationalgenome.org/data*.

Genetic distance between samples: KING (Kinship-based INference for Genome-wide association studies), a computationally efficient program to calculate the person–person distance based on genetic data, was used: (*http://people.virginia.edu/˜wc9c/KING/*).[53]

t-SNE: The R (*http://www.r-project.org/*) implementation of the t-SNE, *Rtsne* version 0.11 (June 30, 2016) (*https://cran.r-project.org/web/packages/Rtsne/* or *https://github.com/jkrijthe/Rtsne*) was used. We use the default value of perplexity (30) which is not extreme enough to be causing visual problems (see, e.g. *http://distill.pub/2016/misread-tsne/*).

PCA: The smartpca in the EIGENSOFT package[20] (v6.1.4) *https://data.broadinstitute.org/alkesgroup/EIGENSOFT/* was used for PCA of genotype data.

Other programs: PLINK *http://pngu.mgh.harvard.edu/~purcell/plink/*[54] was used for genotype file management and conversion. The 3D plot was generated by the *rgl* R package (*https://cran.r-project.org/web/packages/rgl/index.html*, 0.97.0). R (*http://www.r-project.org/*) statistical package was used for other general analysis and graphics.

Acknowledgments

We would like to thank Andrew Shih for discussions. WL and JEC acknowledge the support from the Robert S. Boas Center for Genomics and Human Genetics, YY acknowledges the support from NSFC (11671375), and HH acknowledges the partial grant support from Fordam's IPGF.

References

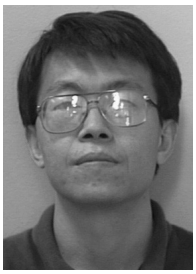
1. Risch N, Merikangas K, The future of genetic studies of complex human diseases, *Science* **273**:1516–1517, 1996.
2. Hirschhorn JN, Daly MJ, Genome-wide association studies for common diseases and complex traits, *Nat Rev Genet* **6**:95–108, 2005.
3. Blading DJ, A tutorial on statistical methods for population association studies, *Nat Rev Genet* **7**:781–791, 2006.
4. Li W, Three lectures on casecontrol genetic association analysis, *Brief Bioinf* **9**:1–13, 2008.
5. Manolio TA, Genomewide association studies and assessment of the risk of disease, *New Eng J Med* **363**:166–176, 2010.
6. Li W, Genome-wide association studies, in *Encyclopedia of System Biology*, Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds.), Springer, Berlin, 2013.
7. Li W, Genetic marker, in *Encyclopedia of System Biology*, Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, Springer, Berlin, 2013, pp. 821–824.
8. Edwards SL, Beesley J, French JD, Dunning AM, Beyond GWASs: Illuminating the dark road from association to function, *Am J Hum Genet* **93**:779–797, 2013.
9. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucl Acids Res* **42**(D1):D1001–D1006, 2014.
10. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), *Nucl Acids Res* **45**(D1):D896–D901, 2016.
11. Cardon LR, Palmer LJ, Population stratification and spurious allelic association, *Lancet* **361**:598–604, 2003.
12. Devlin B, Roeder, Genomic control for association studies, *Biometrics* **55**:997–1004, 1999.
13. Terwilliger JD, Ott J, A haplotype-based haplotype relative risk'approach to detecting allelic associations, *Hum Heredity* **42**:337–346, 1992.
14. Spielman RS, Ewens WJ, The TDT and other family-based tests for linkage disequilibrium and association, *Am J Hum Genet* **59**:983–989, 1996.
15. Laird NM, Horvath S, Xu X, Implementing a unified approach to family-based tests of association, *Genet Epi* **19**:S36–S42, 2000.

16. Pritchard JK, Stephens M, Donnelly P, Inference of population structure using multilocus genotype data, *Genetics* **155**:945–959, 2000.
17. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nat Genet* **38**:203–208, 2006.
18. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer MB, Sabatti C, Eskin E, Variance component model to account for sample structure in genome-wide association studies, *Nat Genet* **42**:348–354, 2010.
19. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL, Advantages and pitfalls in the application of mixed-model association methods, *Nat Genet* **46**:100–106, 2014.
20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D, Principal components analysis corrects for stratification in genome-wide association studies, *Nat Genet* **38**:904–909, 2006.
21. Patterson N, Price AL, Reich D, Population structure and eigenanalysis, *PLoS Genet* **2**:e190, 2006.
22. Reich D, Price AL, Patterson N, Principal component analysis of genetic data, *Nat Genet* **40**:491–492, 2008.
23. Novembre J, Stephens M, Interpreting principal component analyses of spatial population genetic variation, *Nat Genet* **40**:646–649, 2008.
24. Tzeng J, Lu HS, Li WH, Multidimensional scaling for large genomic data sets, *BMC Bioinfo* **9**:179, 2008.
25. Zhu C, Yu J, Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types, *Genetics* **182**:875–888, 2009.
26. Galinsky KL, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL, Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and east Asia, *Am J Hum Genet* **98**:456–472, 2016.
27. Lao O *et al.*, Correlation between genetic and geographic structure in Europe, *Curr Biol* **18**:1241–1248, 2008.
28. van der Maaten LJP, Hinton GE, Visualizing high-dimensional data using t-SNE, *J Machine Learning Res* **9**:2579–2605, 2008.
29. Hinton G, Roweis S, Stochastic neighbor embedding, in *Advances in Neural Information Processing Systems*, Vol. 15, MIT Press, Cambridge, 2002.
30. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Peer D, viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia, *Nat Biotech* **31**:545–552, 2013.
31. Wagner A, Regev A, Yosef N, Revealing the vectors of cellular identity with single-cell genomics, *Nat Biotech* **34**:1145–1160, 2016.
32. Poirion OB, Zhu X, Ching T, Garmire L, Single-cell transcriptomics bioinformatics and computational challenges, *Front Genet* **7**:163, 2016.
33. Platzer A, Visualization of SNPs with t-SNE, *PLoS ONE* **8**:e56883, 2013.
34. The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, *Nature* **467**:1061–1073, 2010.
35. The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, *Nature* **491**:56–65, 2012.
36. The 1000 Genomes Project Consortium, A global reference for human genetic variation, *Nature* **526**:68–74, 2015.
37. Plumbley M, Oja E, A 'nonnegative PCA' algorithm for independent component analysis, *IEEE Trans Neural Net* **15**:66–76, 2004.

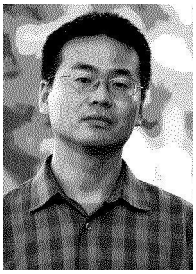
38. Han X, Nonnegative principal component analysis for cancer molecular pattern discovery, *IEEE/ACM Trans Comp Biol Bioinf* **7**:537–549, 2010.
39. Han H, Derivative component analysis for mass spectral serum proteomic profiles, *BMC Med Genomics* **7**(suppl 1):S5, 2014.
40. Li H, Ralph P, Local PCA shows how the effect of population structure differs along the genome, *bioRxiv*, https://doi.org/10.1101/070615.
41. Ma J, Amos CI, Investigation of inversion polymorphisms in the human genome using principal components analysis, *PLoS ONE* **7**:e40224, 2012.
42. Salm M, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, Shoulders CC, The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism, *Genome Res* **22**:1144–1153, 2012.
43. Lu D, Xu S., Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia, *Front Genet* **4**:127, 2013.
44. Candés EJ, Li X, Ma Y, Wright J, Robust principal component analysis, *J ACM* **58**:11, 2009.
45. Linting M, Meulman JJ, Groenen PJF, van der Kooij A, Nonlinear principal components analysis: Introduction and application, *Psych Methods* **12**:336–358, 2007.
46. Mori Y, Kuroda M, Makino N, *Nonlinear Principal Component Analysis and Its Applications*, Springer, Berlin, 2016.
47. Kambhatla N, Leen TK, Dimension reduction by local principal component analysis, *Neural Comp* **9**:1493–1516, 1997.
48. Farcomeno A, Greco L, *Robust Methods for Data Reduction*, Chapman and Hall/CRC, London, 2015.
49. Zou H, Hastie T, Tibshirani R, Sparse principal component analysis, *J Comp Graph Stat* **15**:262–286, 2006.
50. Rokhlin V, Szlam A, Tygert M, A randomized algorithm for principal component analysis, *SIAM J Matrix Anal Appl* **31**:1100–1124, 2009.
51. Halko N, Martinsson PG, Shkolnisky Y, Tyger M, An algorithm for the principal component analysis of large data sets, *SIAM J Sci Comp* **33**:2580–2594, 2011.
52. Abraham G, Inouye M, Fast principal component analysis of large-scale genome-wide data, *PLoS ONE* **9**:e93766, 2014.
53. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM, Robust relationship inference in genome-wide association studies, *Bioinformatics* **26**:2867–2873, 2010.
54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC, PLINK: A toolset for whole-genome association and population-based linkage analysis, *Am J Human Genet* **81**:559–575, 2007.

Wentian Li received Ph.D in 1989 in physics and complex systems from the Columbia University in New York City, and held positions at the University of Illinois, Santa Fe Institute, Rockefeller University, Cold Spring Harbor Laboratory, Columbia Medical Center, and New York State Psychiatric Institute. He started the current position at Feinstein Institute for Medical Research in 2001. He serves as an editorial board member of *Journal of Theoretical Biology*, co-editor-in-chief of *Computational Biology and Chemistry*, and a past member of editorial board at *Bioinformatics*.

Jane Cerise received her Ph.D. in Physics from the University of Houston University Park, Houston, Texas in 1999, and a M.S. in Statistics from the State University of New York, Stony Brook, New York in 2010. She was a Postdoctoral Fellow (2011–2014) and an Associate Research Scientist (2014–2016) at Columbia University in New York City. She is currently a Research Scientist at the Feinstein Institute for Medical Research.



Yaning Yang received his B.S. degree in Mathematics from the University of Science and Technology of China and his Ph.D. degree in Statistics from the Rutgers University in 2000. He was at the Rockefeller University as Postdoctoral Fellow and Research Assistant (2000–2004) and is now holding faculty position at Department of Statistics and Finance in the University of Science and Technology of China.



Henry Han received his Ph.D. in Computational Sciences from the University of Iowa in 2004. He is an Associate Professor in the Computer and Information Science Department at Fordham University in New York, the Associate Chair of the Department at Lincoln Center campus, and co-founder and director of Fordhams master program in cybersecurity (2013–2015). He is also an affiliated faculty member in Quantitative Proteomics Center at Columbia University. He published nearly 50 papers in leading journals and conferences in bioinformatics, big data, data mining, machine learning, and data analytics fields.