Recitation 8

Single Cell RNA Sequencing & Genetics

Corban Swain

MIT - 6.802 / 6.874 / 20.390 / 20.490 / HST.506 - Spring 2020

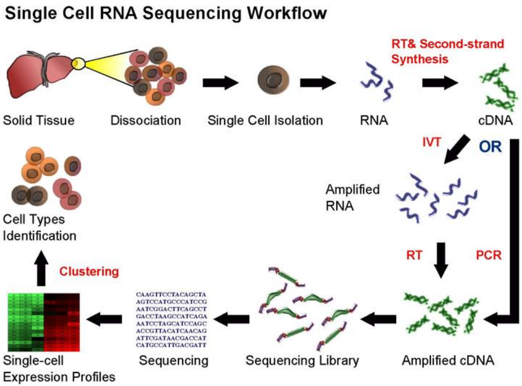
2020-04-09 / 2020-04-10

Outline

- Single Cell RNA Sequencing
- scRNA-Seq Innovations
- Genome Wide Association Studies
- Quantile-Quantile Plot Explanation
- Explanation of perplexity for tSNE

(Re-)Introduction to single-cell RNA Sequencing

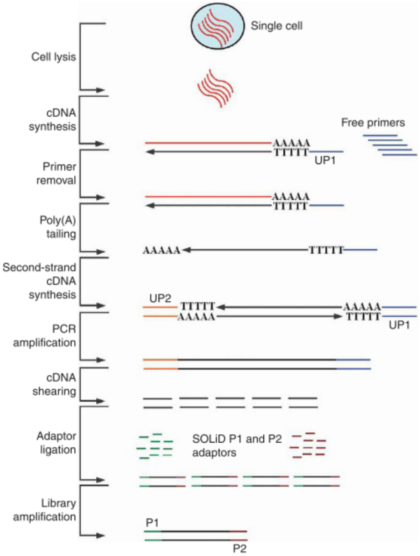
- Single cell RNA sequencing is a class of technologies that aim to isolate, sequence, map to the geneome, and quantify mRNA transcripts in a way that perseveres the distinction between individual cells, in contrast to bulk RNA seq.
- One of the key benefits of these technologies is that it enables us to capture the heterogeneity and stochastic nature of expression within populations of cells.



scRNA-Seq innovation: Template Switch Oligos I

- There are many technological innovations which have improved RNA sequencing from the original Tang, 2009 paper. Many of these were discussed in lecture. Let's breifly chat about one: SMART sequencing and Template Switch Oligos
- If we look at the following protocol for mRNA isolation and amplification we see that the reverse transcription step is performed by elongation of a ploy(T)-containing primer. After elongation for about 30 min by a reverse transcription enzyme the first-strand cDNA is tailed with a poly(A) sequence by a separate enzyme.

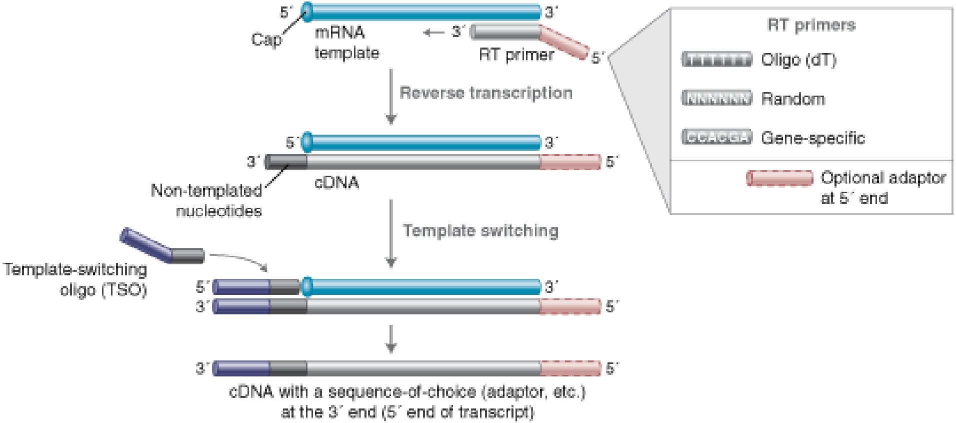
scRNA-Seq innovation: Template Switch Oligos II



scRNA-Seq innovation: Template Switch Oligos III

- This initial elongation can unfortunately lead to premature termination of the cDNA transcript and loss of the 5' end of the mRNA in the sequencing library.
- To address this issue Ramsköld *et al.* developed a method which uses a special reverse transcription enzyme which elongates the first strand cDNA using a poly(T) primer, same as the last protocol.
- However once the RT reaches the end of the mRNA transcript it auto-catalyzes the addition of a few non-templated cytosine nucleotides. These nucleotides serve as a site for a second, template switching ologonucleotide (TSO, which the the experimenter designs and adds to the mix) to bind.
- The RT continues elongation of the first-strand cDNA, extending it with the new TSO as a template to include primers onto the (and possibly a unique molecular identifier) onto the 5' end of the first-strand cDNA.
- In this way we ensure a higher probability of fully incorporating the 5' end of long mRNA transcripts into the sequencing library.
- *See lecture 15, slide 26*

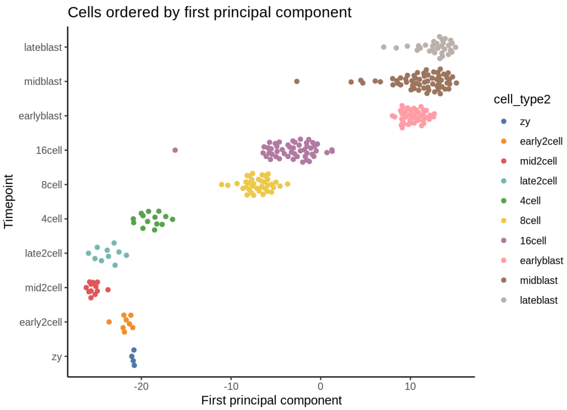
scRNA-Seq innovation: Template Switch Oligos IV



Trajectories & pseudotime analysis I

- In many biological contexts, we want to be able to study the transcriptional changes that occur over the course of the cell cycle, or over the course of differentiation.
- These process occur over time in each cell; however, traditional sequencing technologies involve destruction of the cell to isolate mRNAs and prepare the sequencing library.
- To get around this we can use the transcriptional profiles of single cells to represent "snapshots" at different points in a temporal process. There are a number of computational techniques that exist to map single cell transcriptional profiles into this "pseudotime".
- A very naive approach to do this ordering would be to order the data points along a principle component axis. More nuanced approaches were described in class, *Lecture 15, slides 98-106*
- Great References at https://scrnaseq-course.cog.sanger.ac.uk/website/index.html

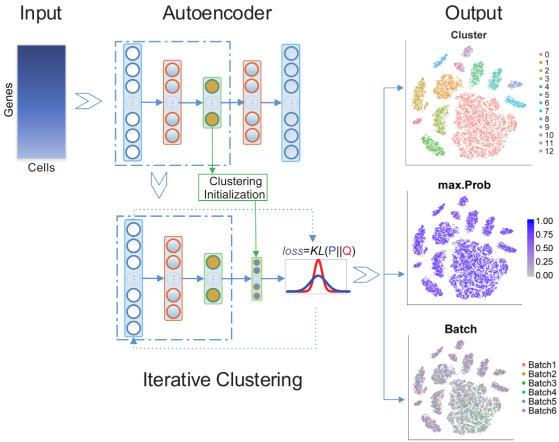
Trajectories & pseudotime analysis II



Lecture 15 Paper: "Deep learning enables accurate clustering and batch effect removal…" I

- **The Problem:** (1) Large scRNA-seq datasets containing thousands to millions of cells presents a problem for traditional clustering algorithms; they struggle to scale up to the cluster different cell populations in these datasets. (2) Systematic differences in gene expression from experiment to experiment (either across replicates or across different studies) can taint and confound the true biological differences leading to drawing inaccurate conclusions. This phenomenon is know as "batch effect".
- **The Goal:** Perform *both* clustering and batch correction of scRNA-seq datasets simultaneously since some cell populations are more prone to batch effect than others, ultimately improving the quality of both..
- **The Method:** Deep embedding algorithm for single-cell clustering (DESC). First train a stacked autoencoder to produce a low diemsional projection from the full gene space for each cell. Use the Louvian method to cluster the cells in the low dimensional space; this method both predicts the number of clusters and the centroids for each cluster in the low dimensional space. Then, iteratively improve the clustering by evaluating the similarity between each cell and it's assigned cluster center (q) and the distribution of the highest confidence cells in each cluster (p). The KL divergence between these two distributions is used to update both the cluster centers and the parameters of the autoencoder network by gradient descent over many epochs.

Lecture 15 Paper: "Deep learning enables accurate clustering and batch effect removal. . ." II



Lecture 15 Paper: "Deep learning enables accurate clustering and batch effect removal. . . " III

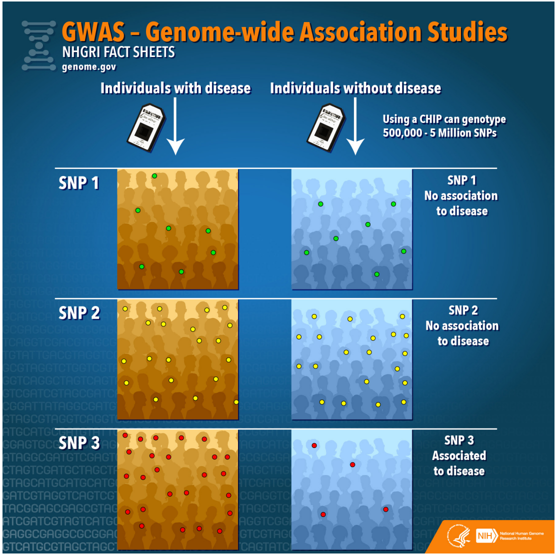
- **The Results:** The DESC method was able to successfully cluster datasets better than many existing methods for datasets containing $\approx 8,000$ and $\approx 24,000$ cells. Additionally, their method was able to successfully cluster by tissue type and remove batch effects from scRNA-seq experiments performed by multiple labs using the same platform.

Li, X., Lyu, Y., Park, J., Zhang, J., Stambolian, D., Susztak, K., … Li, M. (2019). Deep learning enables accurate clustering and batch effect removal in single-cell RNA-seq analysis. *BioRxiv*, 530378. https://doi.org/10.1101/530378

Genome Wide Association Studies I

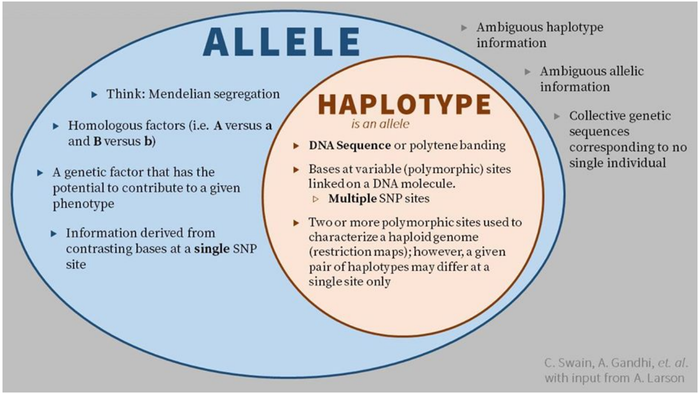
- The primary purpose of GWAS is to identify single nucleotide polymorphisims across the genome that are statistically over- or under-enriched in a specific population compared to a ”normal” population. Specific populations could include persons having a certain disease, people who respond adversely to a specific therapeutic, and other phenotypic classifications.
- The ”library” of SNPs to be tested is often on the order of 100k. Because of the large cohort sizes and the high coverage of genomic sites to GWAS is well suited to identify many SNPs related to ”polygenic” disorders, where each genomic site has a small, but significant, effect on phenotype perturbation.

Genome Wide Association Studies II


GWAS – Genome-wide Association Studies
NHGRI FACT SHEETS
genome.gov
Individuals with disease
Individuals without disease
Using a CHIP can genotype 500,000 - 5 Million SNPs
SNP 1
SNP 1 No association to disease
SNP 2
SNP 2 No association to disease
SNP 3
SNP 3 Associated to disease
NIH National Human Genome Research Institute

Haplotypes and Linkage Disequilibrium I

Evaluation of allelic groupings called “haplotypes” which arise because of the recombination-based mechanism of genomic variability. Such groupings lead to “linkage disequilibrium” (LD) meaning that a whole set of SNPs can tend to arise in specific patterns; this makes de-tangling biologically significant SNP sites from those that simply happen to be included in the haplotype difficult. *See Lecture 16, slide 39*



ALLELE

- Think: Mendelian segregation
- Homologous factors (i.e. **A** versus **a** and **B** versus **b**)
- A genetic factor that has the potential to contribute to a given phenotype
- Information derived from contrasting bases at a **single** SNP site

HAPLOTYPE

is an allele

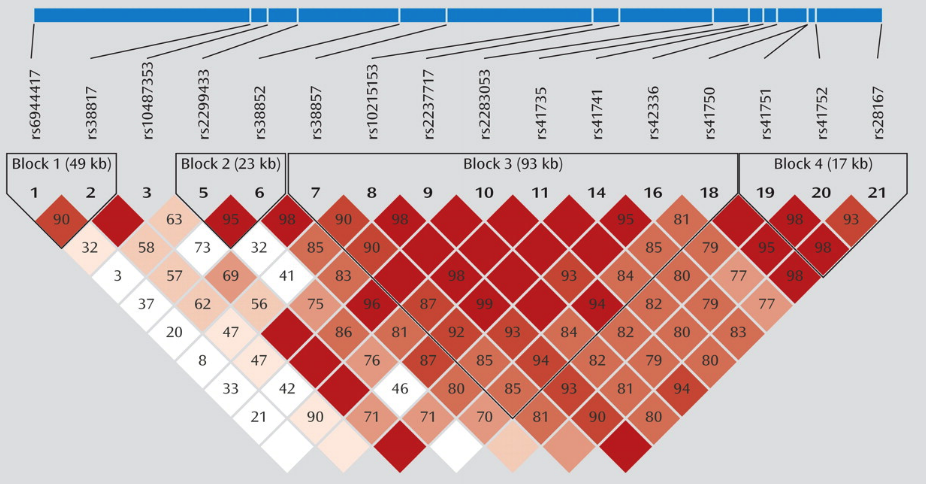
- **DNA Sequence** or polytene banding
- Bases at variable (polymorphic) sites linked on a DNA molecule.
  - **Multiple** SNP sites
- Two or more polymorphic sites used to characterize a haploid genome (restriction maps); however, a given pair of haplotypes may differ at a single site only

- Ambiguous haplotype information
- Ambiguous allelic information
- Collective genetic sequences corresponding to no single individual

C. Swain, A. Gandhi, *et. al.*
with input from A. Larson

Haplotypes and Linkage Disequilibrium II

Linkage disequilibrium Useful description of linkage diagrams at this link.



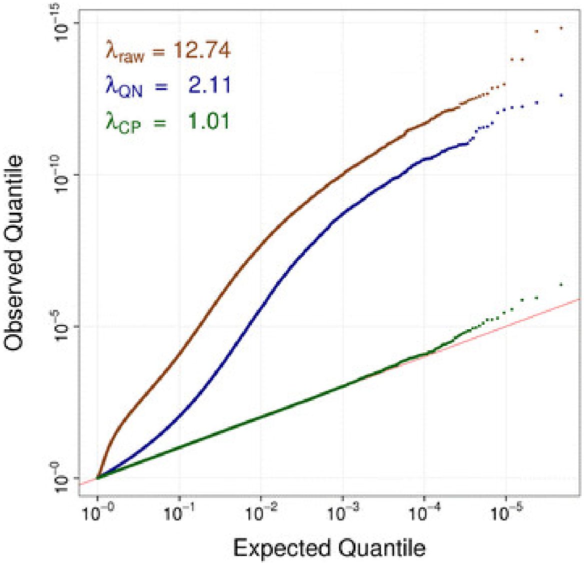
Applications of GWAS

- An important goal after performing GWAS is the identification of the causal link between a given allele and the the perturbation it is known to cause.
- Furthermore, in a disease context, we want to be able to use our knowledge of the causal mechanism to propose and design therapeutics to decrease of mitigate the risk presented by a certain genetic predisposition.
- Manolis went through en excellent example of this in lecture with respect to obesity and the *FTO* locus.

Quantile-Quantile Plots Explained I

- Let's say that we obtain *p*-values from from 10^5 SNPs computed using the χ^2 distribution between two populations.
- If our quality control is performed well, and our SNP distribution is indeed normal (recall that the χ^2 distribution represents the squared distance between normal random variables) then we would expect our *p*-values to be uniformly distributed.
- So if we rank-order our observed *p*-values from smallest to largest and plot it against the expected uniform distribution of 10^5 p values evenly spaced from $\frac{1}{10^5}$ to 1, we would expect to observe a straight line.
- To evaluate this we can look at the goodnes of fit graphically or compute the genomic inflation factor λ which effectively compares the median test statistic to the expected median test statistic.
- All of this is to ensure that our population data is, generally, not biased or prone to producing many false positives.
- We do expect, however, for the SNPs biologically associated with a pheotype of interest to fall off above this line of expectation and be more significant than would be expected if all data conformed to the null hypothesis.
- *See lecture 16, slide 18*

Quantile-Quantile Plots Explained II



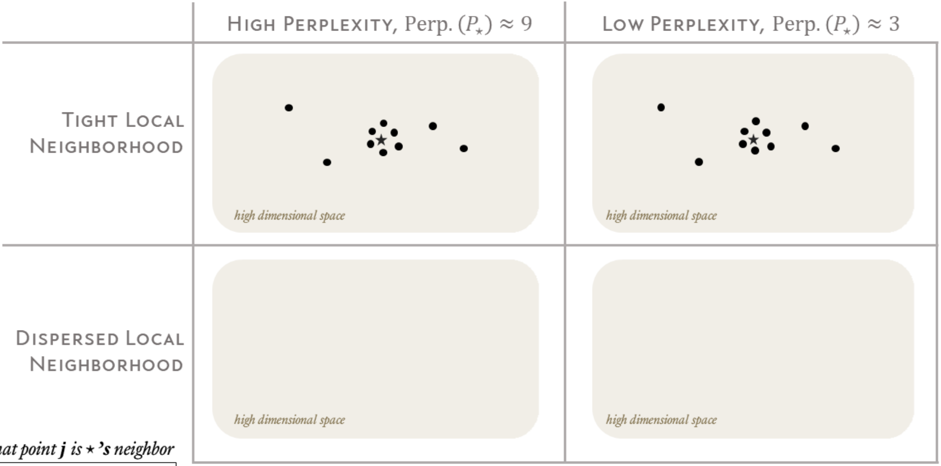
*t*SNE Perplexity Graphical Explanation I

	HIGH PERPLEXITY, Perp. $(P_\star) \approx 9$	LOW PERPLEXITY, Perp. $(P_\star) \approx 3$
TIGHT LOCAL NEIGHBORHOOD	*high dimensional space*	*high dimensional space*
DISPERSED LOCAL NEIGHBORHOOD	*high dimensional space*	*high dimensional space*

*probability that point **j** is $\star$'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

*t*SNE Perplexity Graphical Explanation II



High Perplexity, Perp. $(P_\star) \approx 9$

Low Perplexity, Perp. $(P_\star) \approx 3$

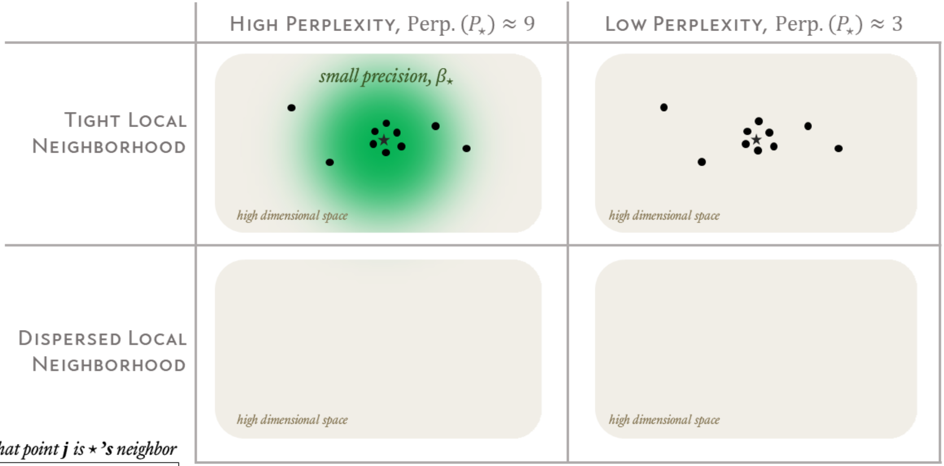
Tight Local Neighborhood

Dispersed Local Neighborhood

*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star} \exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

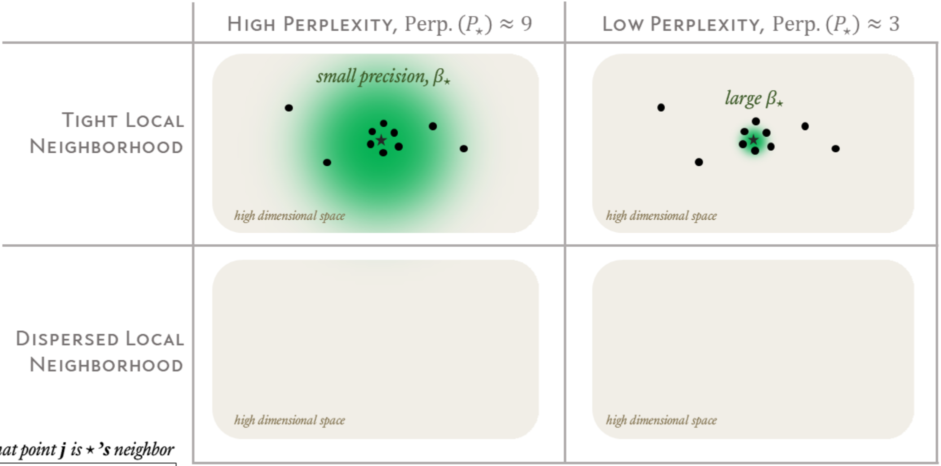
*t*SNE Perplexity Graphical Explanation III


HIGH PERPLEXITY, Perp. $(P_\star) \approx 9$
LOW PERPLEXITY, Perp. $(P_\star) \approx 3$
TIGHT LOCAL NEIGHBORHOOD
DISPERSED LOCAL NEIGHBORHOOD
small precision, $\beta_\star$
high dimensional space
high dimensional space
high dimensional space
high dimensional space

*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star} \exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

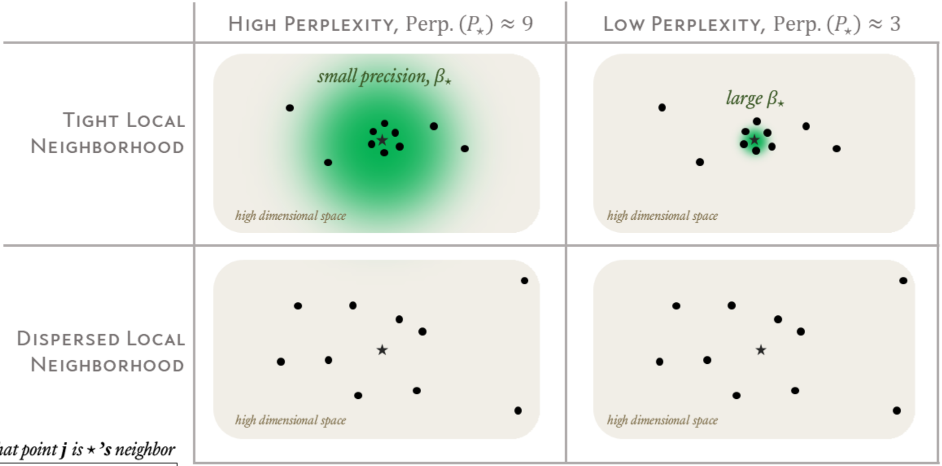
*t*SNE Perplexity Graphical Explanation IV


HIGH PERPLEXITY, Perp. $(P_\star) \approx 9$
LOW PERPLEXITY, Perp. $(P_\star) \approx 3$
TIGHT LOCAL NEIGHBORHOOD
small precision, $\beta_\star$
high dimensional space
large $\beta_\star$
high dimensional space
DISPERSED LOCAL NEIGHBORHOOD
high dimensional space
high dimensional space

probability that point ***j*** *is* $\star$ *'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star} \exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

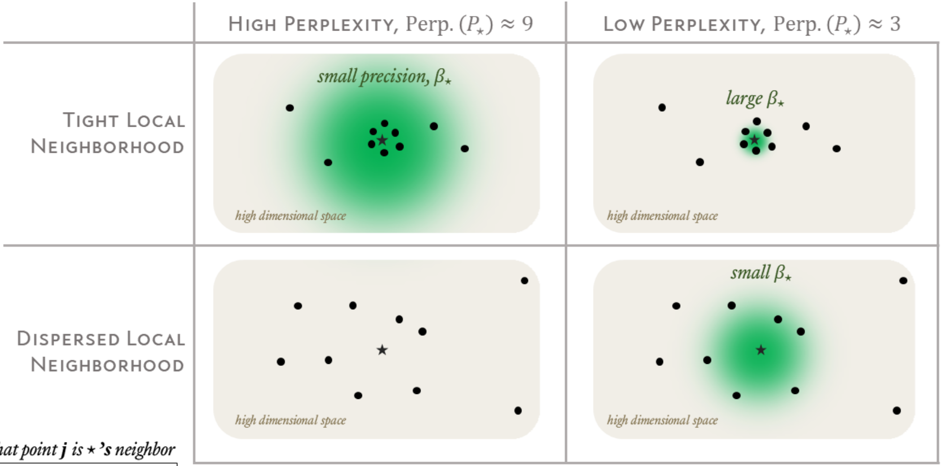
tSNE Perplexity Graphical Explanation V


High Perplexity, Perp. $(P_\star) \approx 9$
Low Perplexity, Perp. $(P_\star) \approx 3$
Tight Local Neighborhood
Dispersed Local Neighborhood
small precision, $\beta_\star$
large $\beta_\star$
high dimensional space
high dimensional space
high dimensional space
high dimensional space

*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

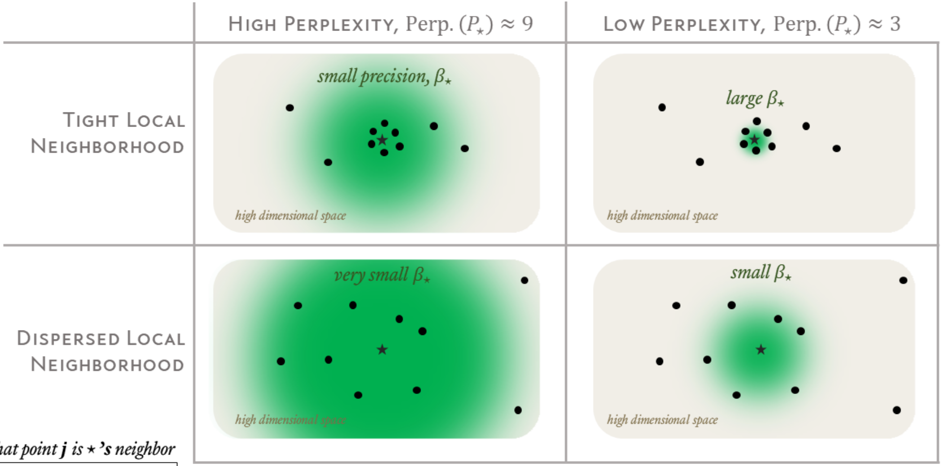
*t*SNE Perplexity Graphical Explanation VI



*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

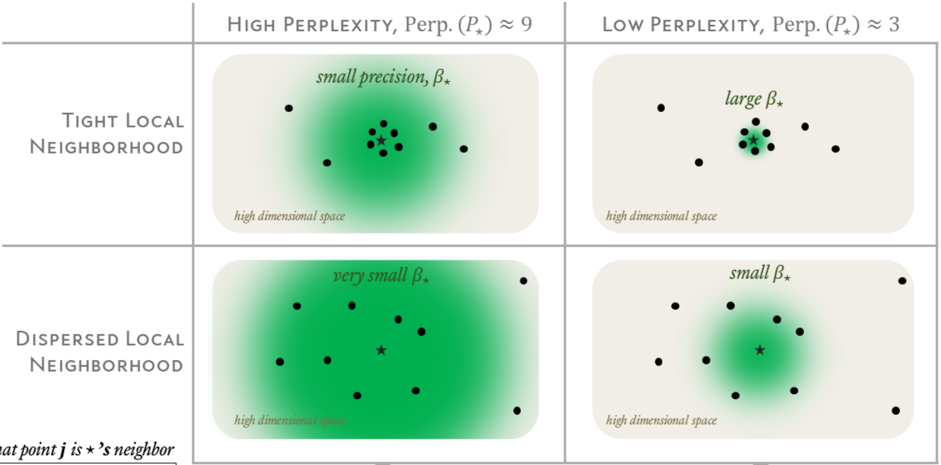
tSNE Perplexity Graphical Explanation VII



*probability that point **j** is $\star$'s neighbor*

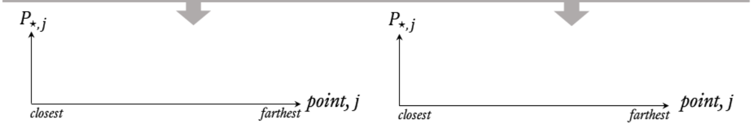
$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star} \exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

*t*SNE Perplexity Graphical Explanation VIII

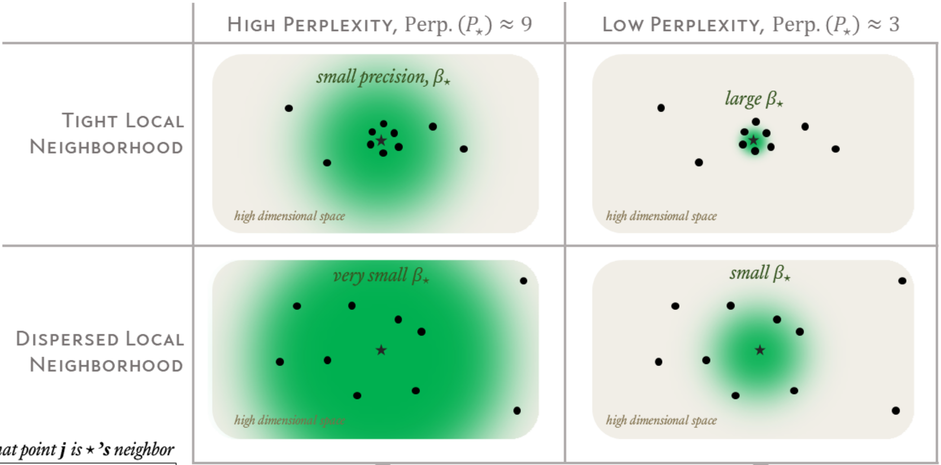


*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$

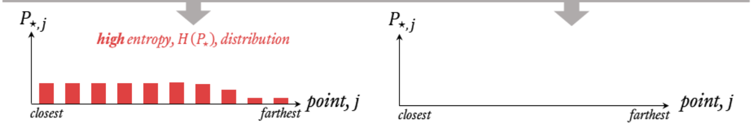


tSNE Perplexity Graphical Explanation IX

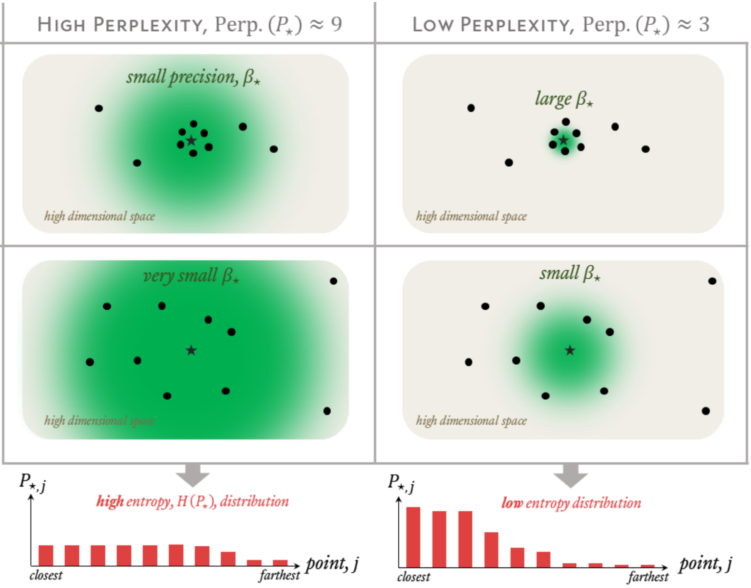


*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\dfrac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\dfrac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$



*t*SNE Perplexity Graphical Explanation X



*probability that point **j** is ⋆'s neighbor*

$$P_{\star,j} = \frac{\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_j\|^2}{2/\beta_\star}\right\}}{\sum_{k\neq\star}\exp\left\{-\frac{\|\mathbf{x}_\star - \mathbf{x}_k\|^2}{2/\beta_\star}\right\}}$$