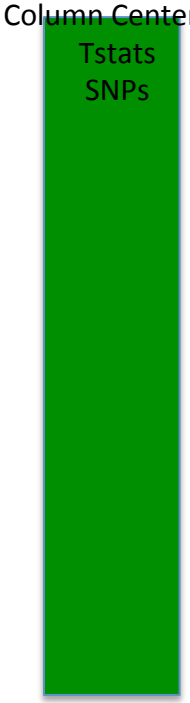
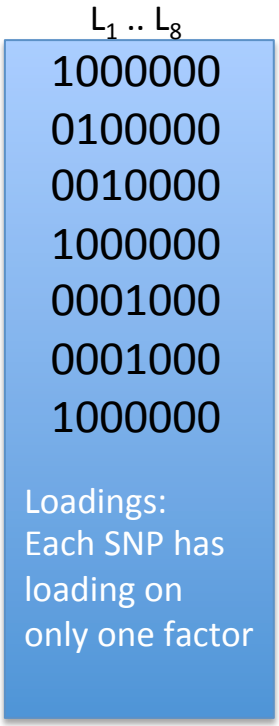
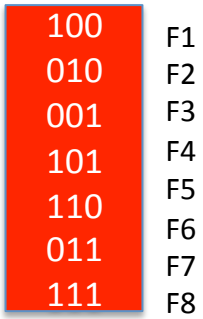


$N_{1000}xD_3$



NxK



F1
F2
F3
F4
F5
F6
F7
F8

KxD

- For each SNP, bj1 will receive only input from factors in which tissue 1 is active
- If the SNP has no loading on these factors, then the SNP will not have activity in this tissue
- For example, SNP 1 will have activity only in tissue 1 because it has loading on factor 1. SNP 5 will have activity in tissue 1 and 3 because it has loading on factor 4

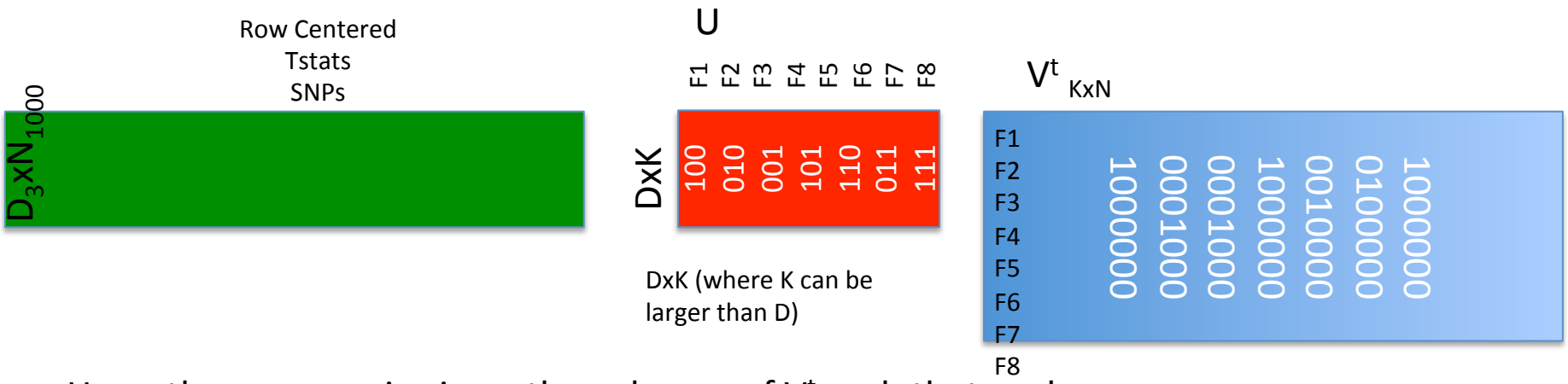
- Each SNP can be maximally loaded on 1 eigenconfig (Factor)
- This means that it will be active only in tissues that are active in this eigenconfig
- E.g., if SNP 1 has loading on factor 1, it is active only in tissue 1
- While loading on factor 4 corresponds to activity in tissue 1 and 3
- We are effectively putting a sparse prior on the **rows of L** (or the rows of U in SVD or on columns of V^t in tranposed case)

Factors:
Each Tissue Can be active in more than one Config

SFA

- Factor Represents the 'Eigenconfig' - i.e., the D Dimensional vector of directions active in a particular eigenconfig
- E.g., Factor 2 might consist of expression in the direction of tissue 1, Factor 4 might consist of expression in the direciton of tissues 1 and 2, etc.
- We have hierarchically reduced K by estimating the proportional membership of each SNP in a config type

Or, Equivalent to



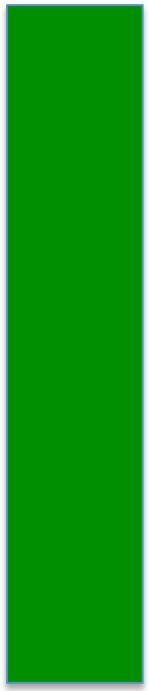
Here, the sparse prior is on the columns of V^t such that each SNP can be maximally a member of one Factor, but there are more factors than Tissues

We learn that in order to translate SFA to SVD, the matrices must be transposed

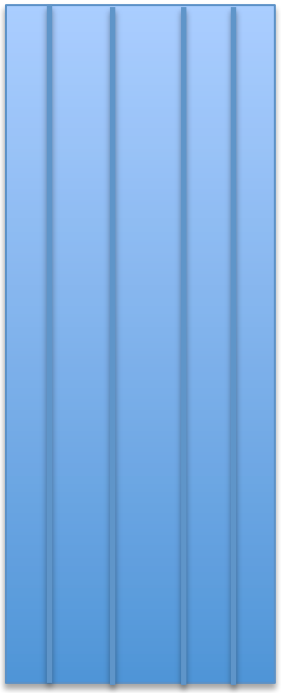
The sparse prior goes from the rows of L to the columns of V^t

Understanding …

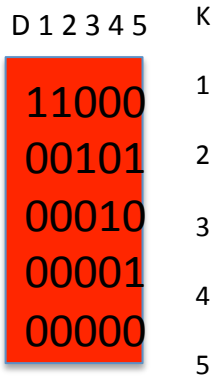
- Alternatively, we could consider the pxr X matrix as a representation of L F, where L corresponds to the 'sparse' loadings of every SNP on at most one 'eigenconfiguration' (of which there are k) and each eigenconfiguration corresponds to the direction of expression across a particular set of tissues. This is not precisely analogous to the Engelhardt *etal* paper, because in that framework, the goal was to reduce n individuals into a smaller subset of d populations, where here we actually expand the d tissues into at most $k = 2^d$ configurations. However, we reduce the SNPs into 'clusters', such that each effect size clusters into one characteristic configuration pattern, in which the SNP can be active in a particular election of tissues. This is naturally appropriate for the ash model, in which we attempt to gleam shared information from all the β_j, who we assume to share characteristic expression profiles.
- We are effectively putting a sparse prior on the rows of U.



X_{nxd}



U_{nxd}



V^t_{dxd}

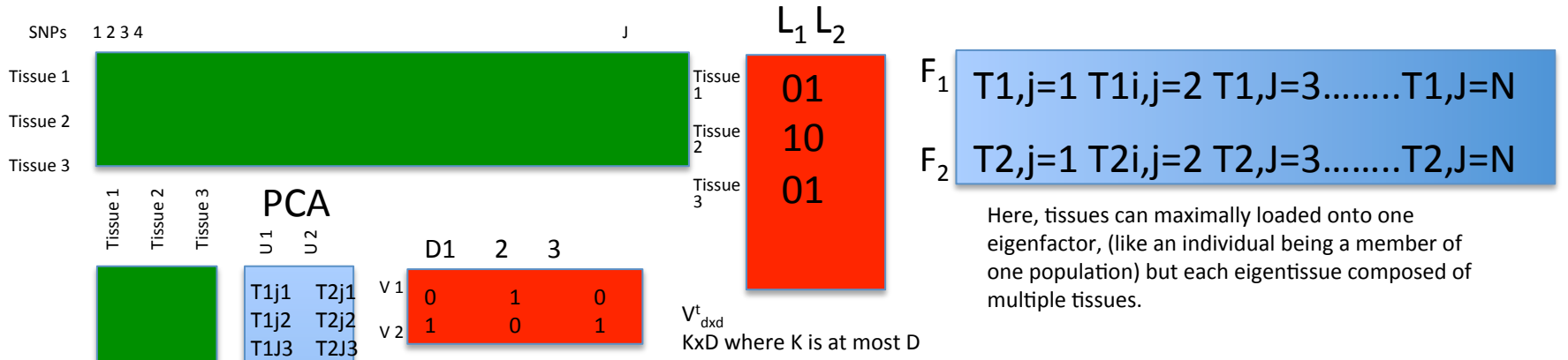
KxD where K is at most D PCA

Here, tissues can be a member of at most one eigentissue, but each eigen tissue might be composed of several tissues (i.e., Eigentissue 1 is composed of tissues 1 and 2, while eigentissue 5 is false)

- Columns of U are linear combinations of tissues at a given SNP U_1 represents direction of expression across the genome for eigentissue1
- Previously 'avg' given gene expression across tissues, *Engelhardt et al*: the expression of a given gene in eigentissue1
- The difference is that we might reduce the number of tissues (because K is at most D), where our goal is to reduce the patterns of expression, but we actually expand D
- If we put sparse prior on the rows of U, it means, that a SNP can be active in only one 'eigenconfig'
- Rows of V^t are linear combinations of rows of X, each element of $V^{t[1,]}$ represents the genome-wide expression of eigenconfig 1 for each tissue
- Will only receive weight from SNPs active in eigenconfig 1. Previously, 'average expression of all SNPs in tissue 1'
- Now expression of SNPs active in eigenconfig1
- Sparse prior on columns of V^t means that each each tissue can be at most composed of 1 eigentissue

What I Had Before ….

Imagine D = 3 Tissues, P = 10 Gene-SNP Pair, 2 Factors



- Here, set possible loadings to be "sparse" (loaded on one factor each) means that each tissue can be loaded on one eigentissue each, which doesn't make sense because our aim is not to reduce the number of tissues but the number of ways in which a gene-SNP pair can act (sparse always on rows of L)
- Equivalent to Sparse prior on columns of V^t
- In this situation, each tissue could be maximally a member of one 'eigentissue'
- Here, we can see that the Vector of T statistics across all P genes for eigentissue 1 represents the first vector, while the vector of t statistics across all P genes represents the direction of the second vector.
- We would model the posterior on each multivariate vector $beta_j$ as a mixture of these characteristic directions.
- Thus each multivariate component will correspond to the importance of the direction of Factor 1…K in tissue D. Each tissue will contain only the expression of SNPs active in its respective factor loading, meaning those SNPs active in characteristic eigentissue K, analogous to containing only the genotype of your populations.
- Analogous to population framework, in which tissues are individuals and each individual is a member of only one population. Then his or her genotype vector is derived from only the factor according to his identify.
- **In PCA**
- Columns of U are linear combinations of tissues at a given SNP U_1 represents direction of expression across the genome for eigentissue1
- Previously 'avg' given gene expression across tissues, *Engelhardt et al*: the expression of a given gene in eigentissue1
- The difference is that we might reduce the number of tissues (because K is at most D), where our goal is to reduce the patterns of expression, but we actually expand D
- Rows of V^t are linear combinations of rows of X, each element of $V^{t[1,]}$ represents the genome-wide expression of eigenconfig 1 for each tissue
- Previously, 'average expression of all SNP. Now expression of SNPs active in eigenconfig1
- Sparse prior on columns of V^t means that each each tissue can be at most composed of 1 eigentissue