

# Genetic diversity and multi-drug class resistance of a large HIV-1 sequence repository

**K. Van der Borgh<sup>1,2</sup>, H. van Vlijmen<sup>3</sup>, P. Lecocq<sup>1</sup>, M. Van Houtte<sup>1</sup>, L. Bacheler<sup>4</sup> and P. Sista<sup>4</sup>**

<sup>1</sup>Virco BVBA, Mechelen, <sup>2</sup>I-Biostat, Leuven, <sup>3</sup>Tibotec-Virco, Mechelen,

<sup>4</sup>VircoLab Inc., Chapel Hill, NC

# Introduction

**How to visualize a HIV-1 PR-RT database of half a million sequences and their resistance to 18 drugs?**

**1) Genetic diversity: Clustering**

- **Fast clustering.**
- **Retain sequences present in large enough clusters.**
- **To make a 2D plot: limit number of clusters to ~5,000.**

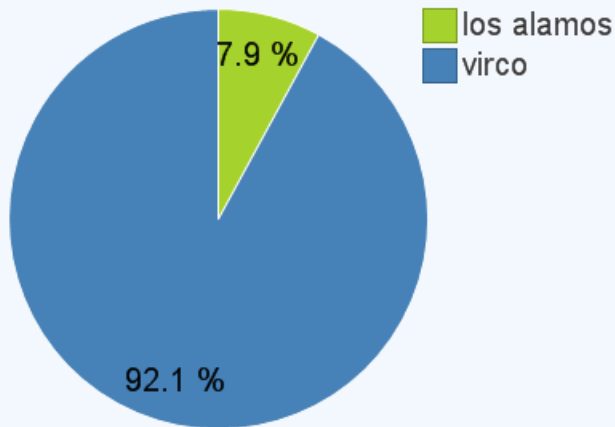
**2) Drug resistance: Factor analysis**

- **Analyze resistance correlation between drugs using full database with  $\ll$  18 factors.**
- **3 factors can be used to do resistance coloring of the clusters.**

# Repository

The HIV-1 PR-RT sequence repository in the analysis consisted of both the Virco and Los Alamos databases.

389,659 genotypes



PR-RT drug resistance phenotypes ( $IC_{50}$  Fold Change) were calculated for each sequence in the repository using the vircoTYPE HIV-1 prediction algorithm.

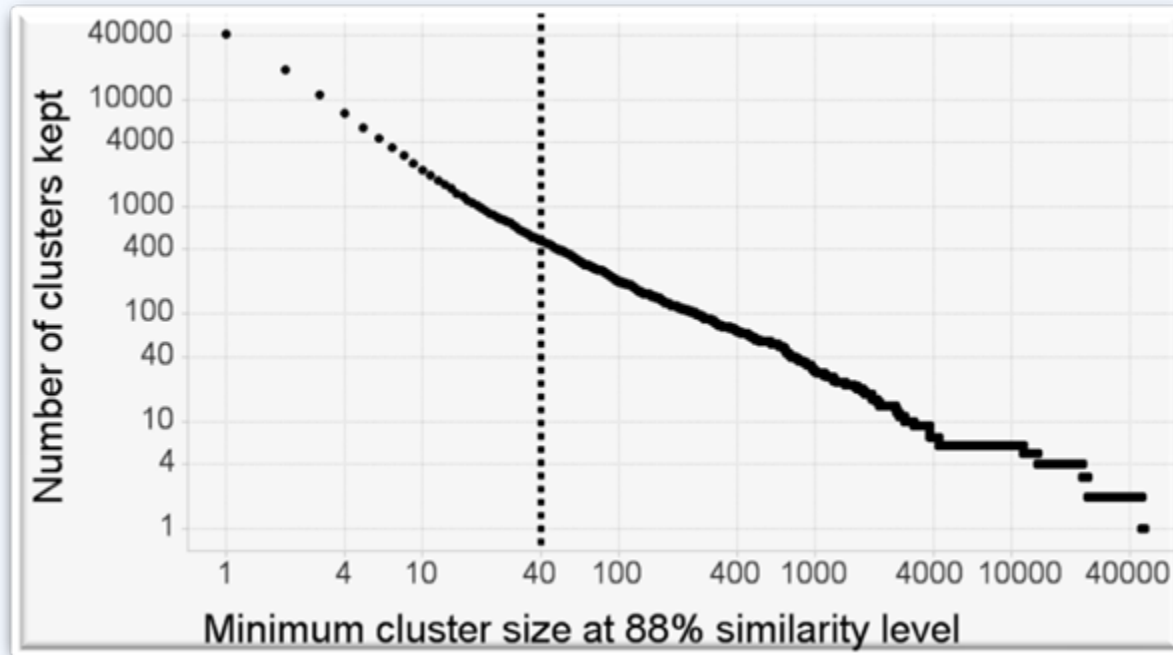
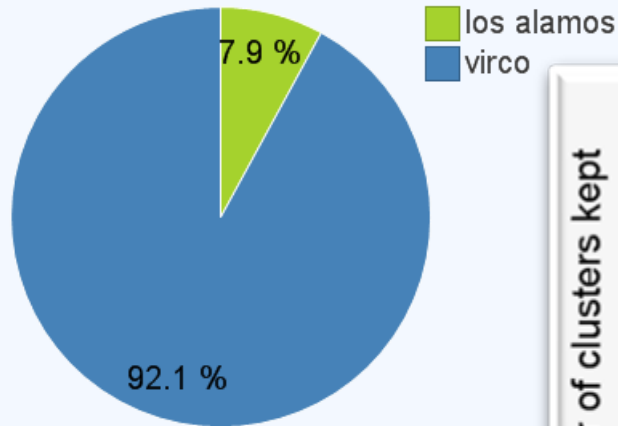
# Assign sequences to 5,743 clusters

389,659 genotypes



**6,659 clusters**  
at 90% similarity

*Fast incremental clustering\**



\*77 positions associated with resistance in PR (10-95) or RT (41-238) were considered, keep clusters with  $\geq 40$  sequences at 88% similarity level.

\*\*Keep clusters with BLAST similarity  $< 5E-37$  with at least two other clusters.

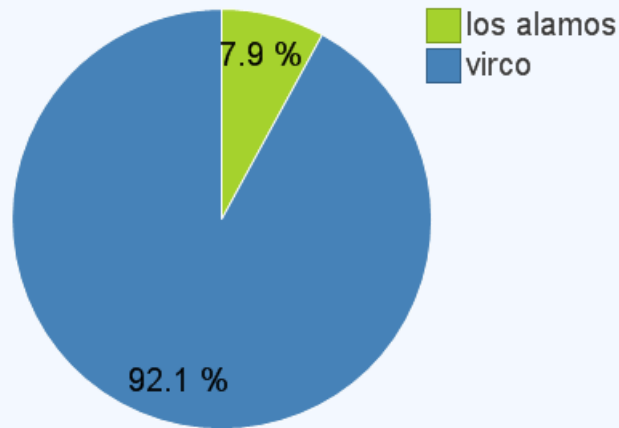
# Assign sequences to 5,743 clusters

389,659 genotypes



**6,659 clusters**  
at 90% similarity

*Fast incremental clustering\**



*BLAST\*\**

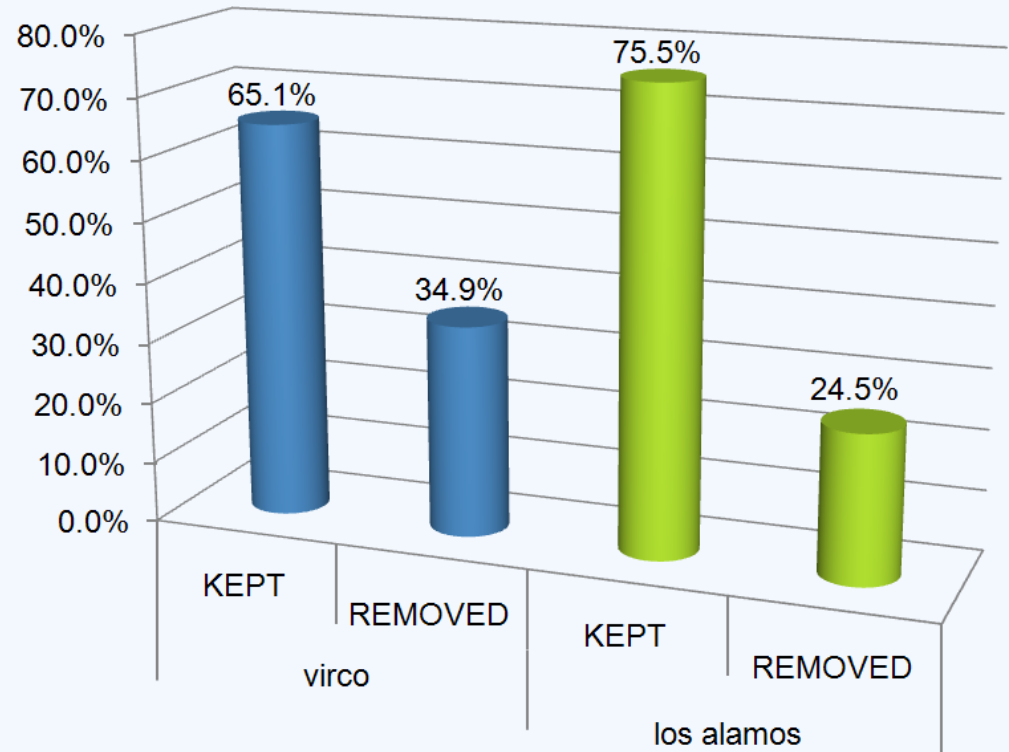
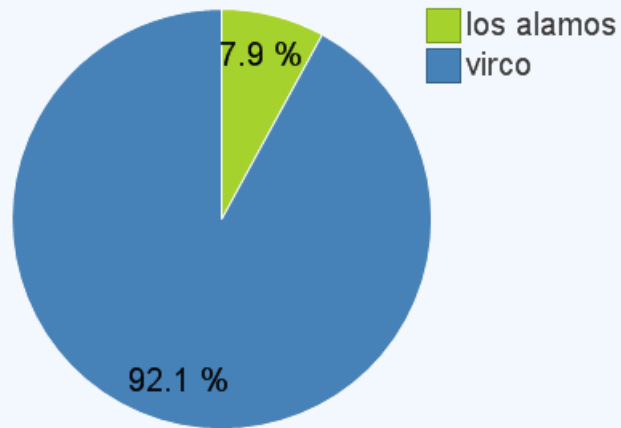
**5,743 clusters**

\*77 positions associated with resistance in PR (10-95) or RT (41-238) were considered, keep clusters with  $\geq 40$  sequences at 88% similarity level.

\*\*Keep clusters with BLAST similarity  $< 5E-37$  with at least two other clusters.

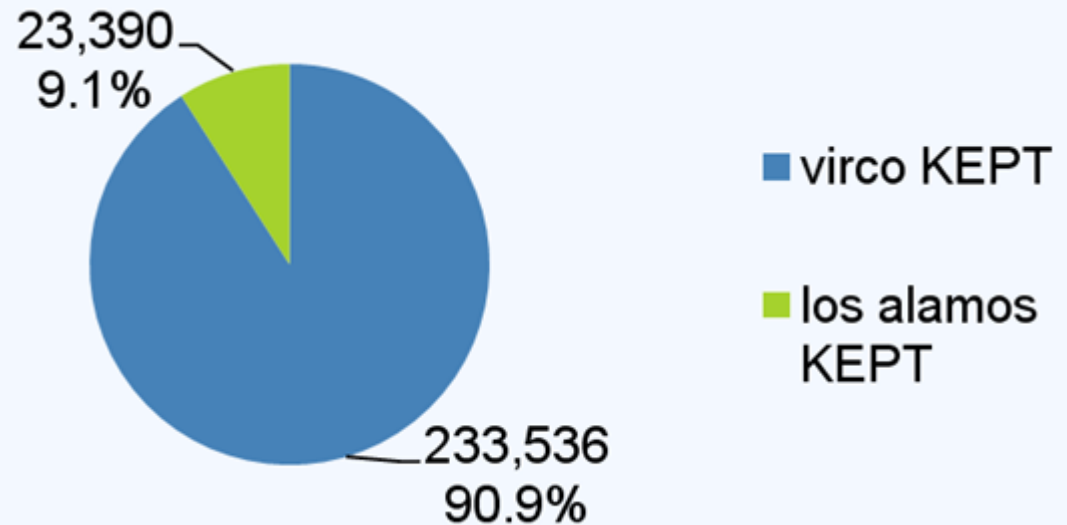
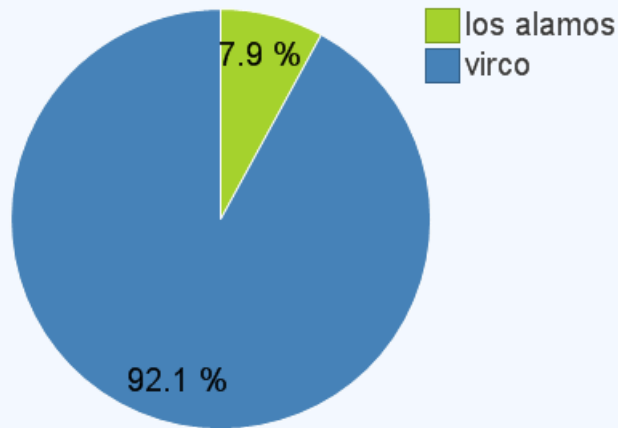
# Assign sequences to 5,743 clusters

389,659 genotypes → 5,743 clusters (256,926 genotypes)



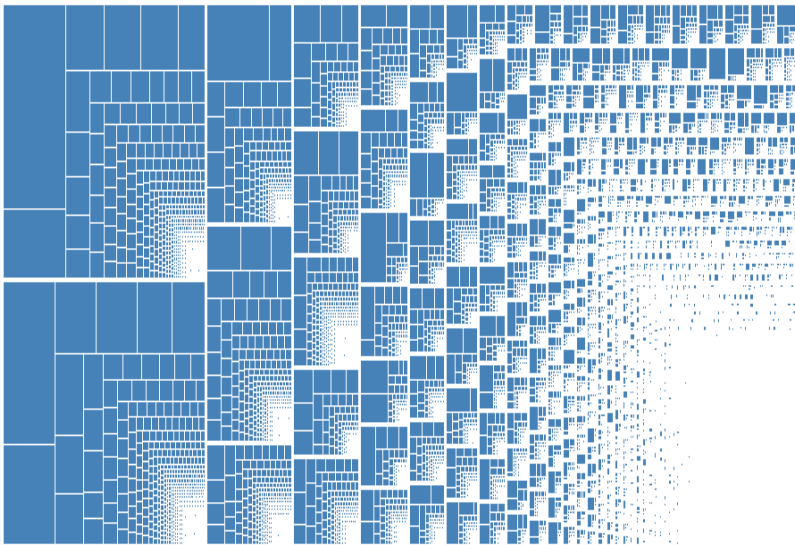
# Assign sequences to 5,743 clusters

389,659 genotypes → 5,743 clusters (256,926 genotypes)



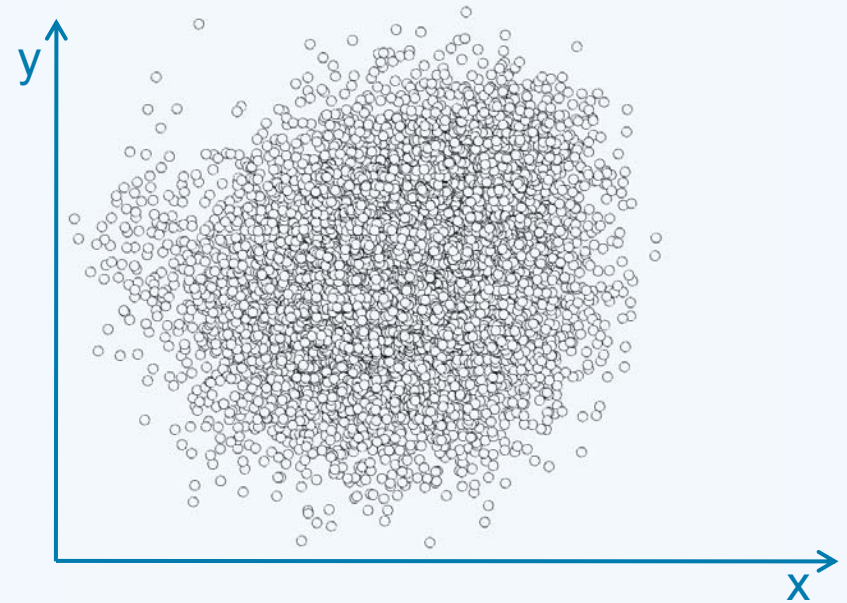
# Two displays of genetic diversity

## Treemap



- size of block= cluster size
- different similarity levels
- cluster coloring
- loss of sequence similarity between clusters

## Network



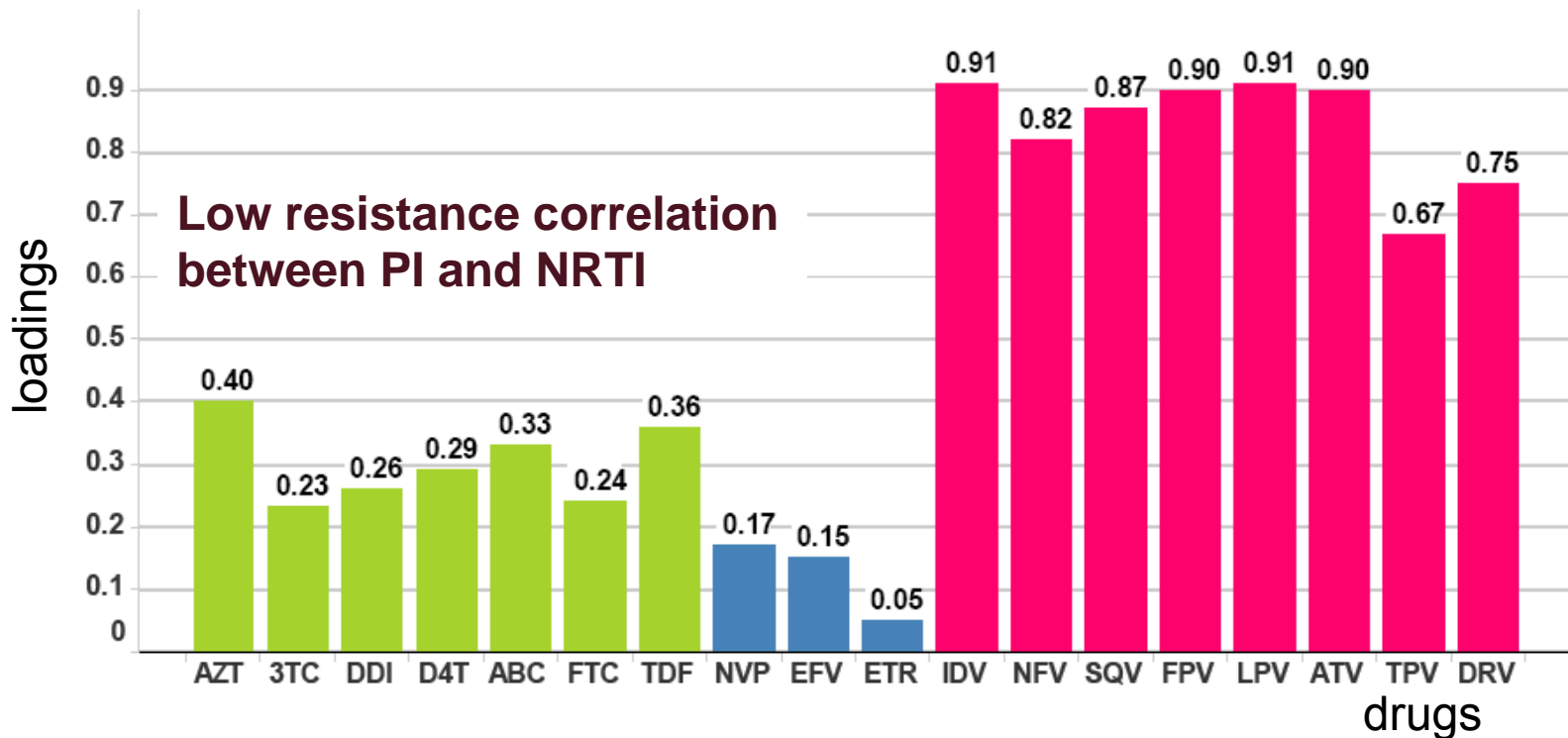
- similar clusters are grouped in xy-plane
- cluster coloring and z-axis can be used to add information
- two dimensions allow for crude mapping of sequence similarity only.

# 4 factors explain phenotype variability

- **Factor analysis on the 389,659 phenotypes x 18 drugs was used to analyze the resistance correlation between these PR-RT drugs.**
- **Using 4 factors, each of the factors could be interpreted as being a drug class phenotype:  
PI (factor 1), NRTI1 (factor 2), NNRTI (factor 3) and NRTI2 (factor 4).**
- **Using 4 factors, 82.7% of the phenotype variability with 18 drugs was explained.**

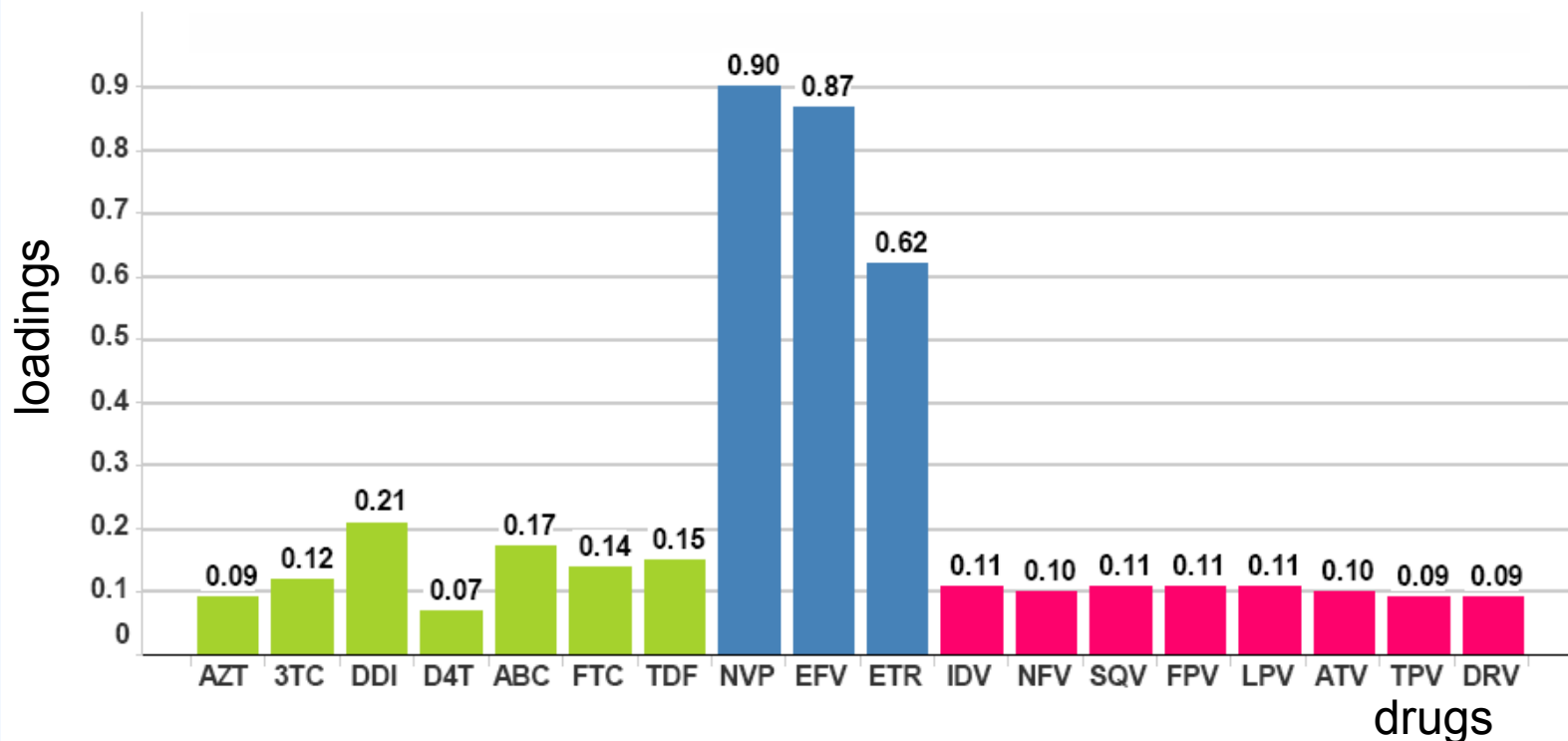
# Factor 1: Protease Inhibitors

PI factor: *Less cross-resistance between TPV/DRV and the other PI.*



# Factor 3: NNRTIs

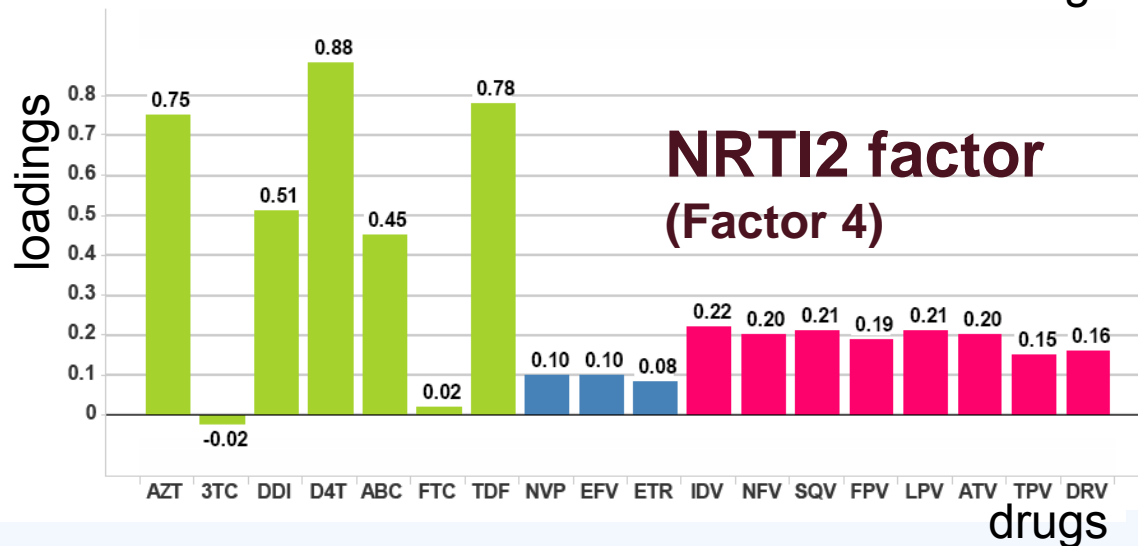
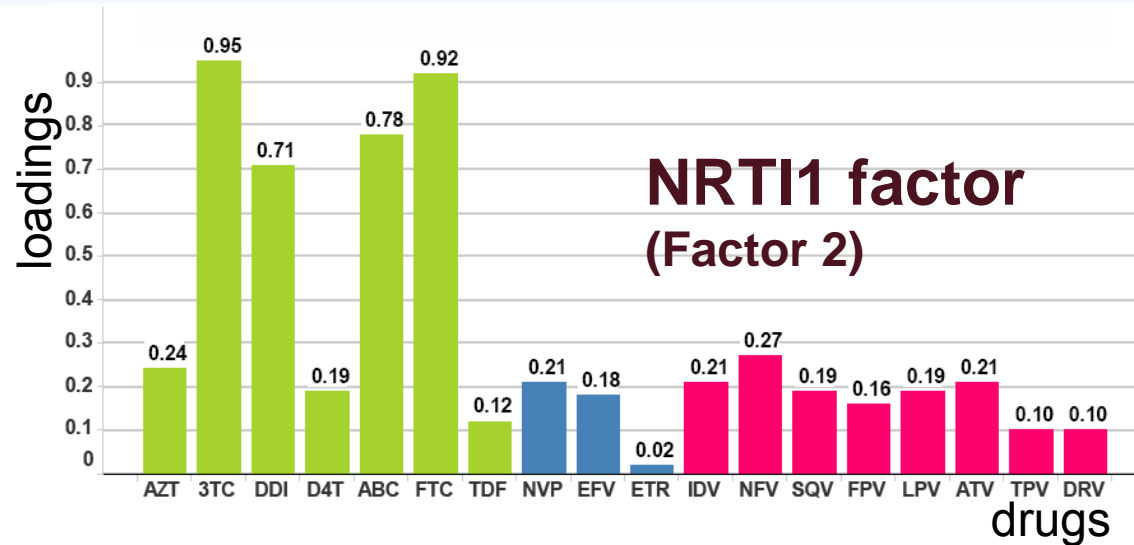
**NNRTI factor:** *Less cross-resistance between ETR and the other NNRTI*



# Factors 2 and 4: NRTIs

## NRTI load on two factors.

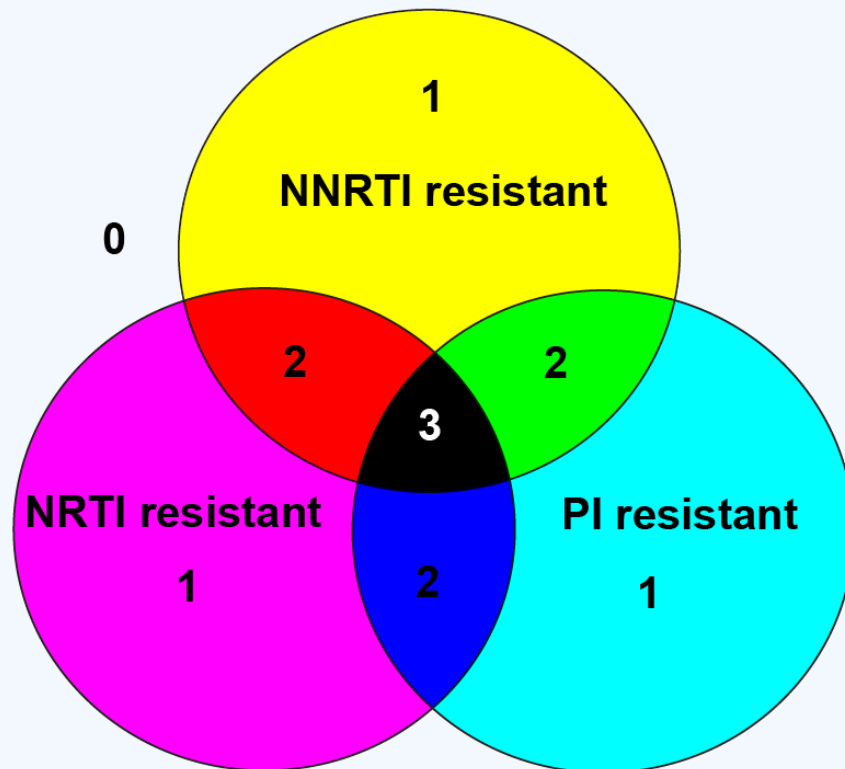
- Genetic profiles causing 3TC/FTC resistance differ from those causing AZT/D4T/TDF resistance.
- ABC/DDI can be used as an alternative or third NRTI in the treatment regimen, having intermediate loadings on both factors.



# Resistance colour coding

- 0 Zero class resistant**
- 1 One class resistant**
- 2 Two class resistant**
- 3 Three class resistant**

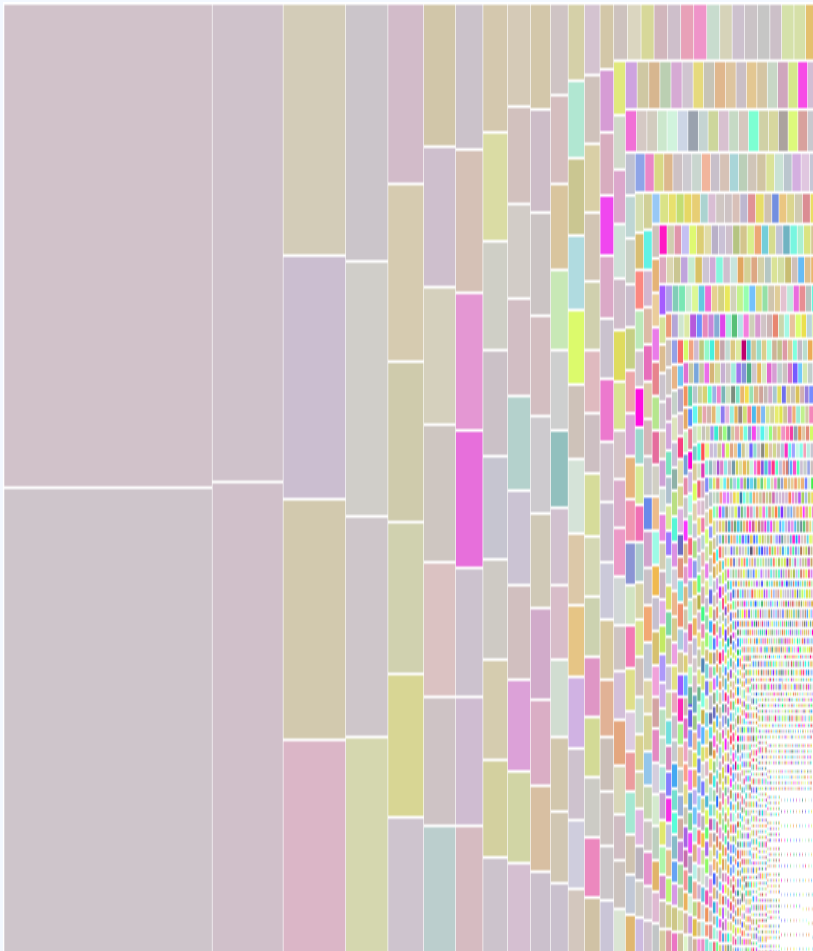
# Resistance colour coding



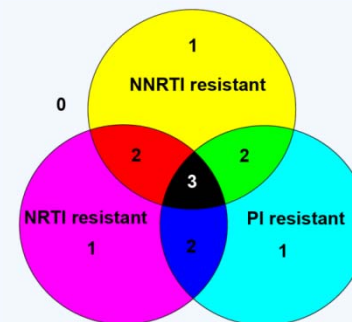
- 0** Zero class resistant
- 1** One class resistant
- 2** Two class resistant
- 3** Three class resistant

# Resistance coloring of the two displays

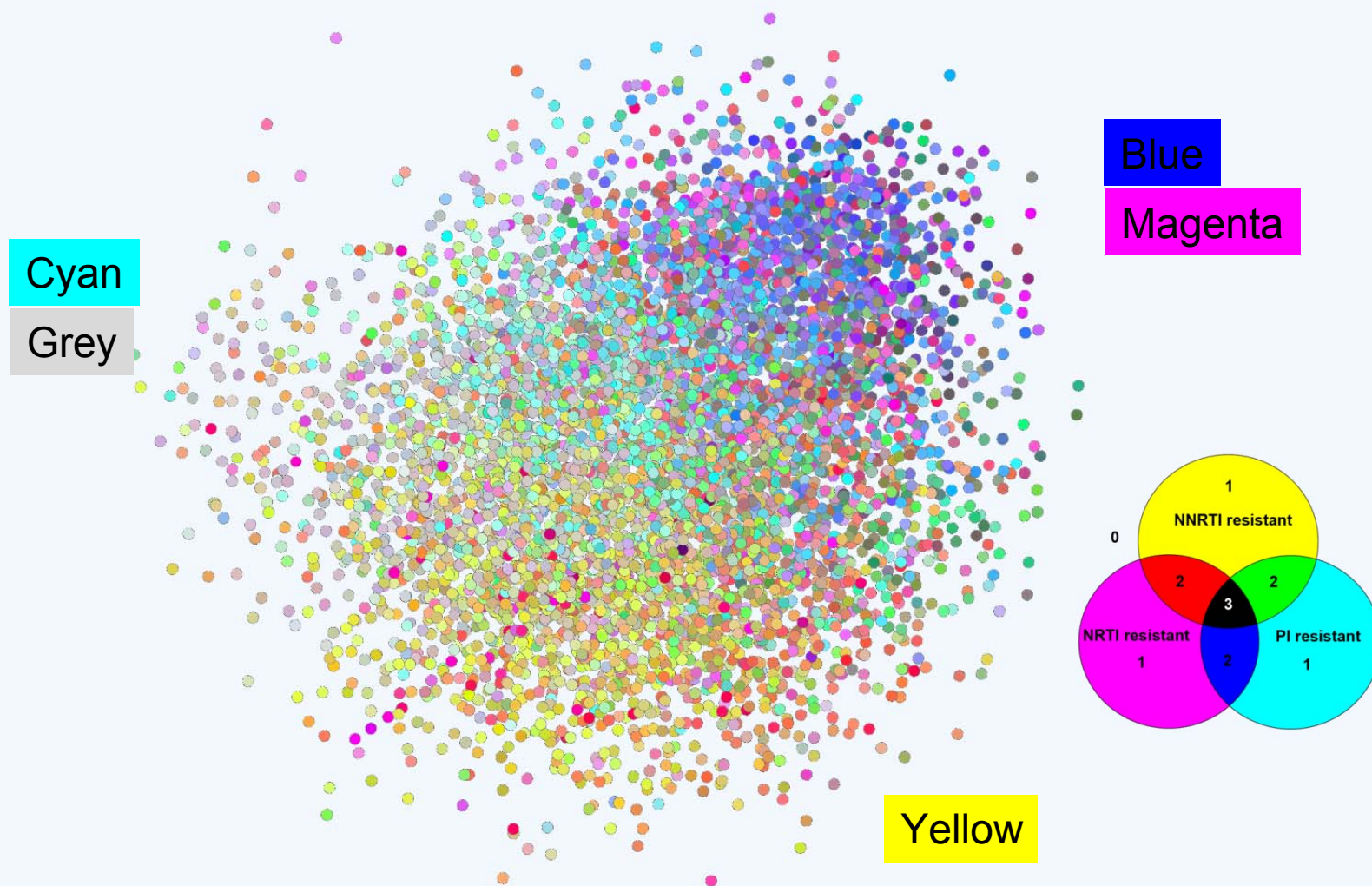
## Treemap



## Network

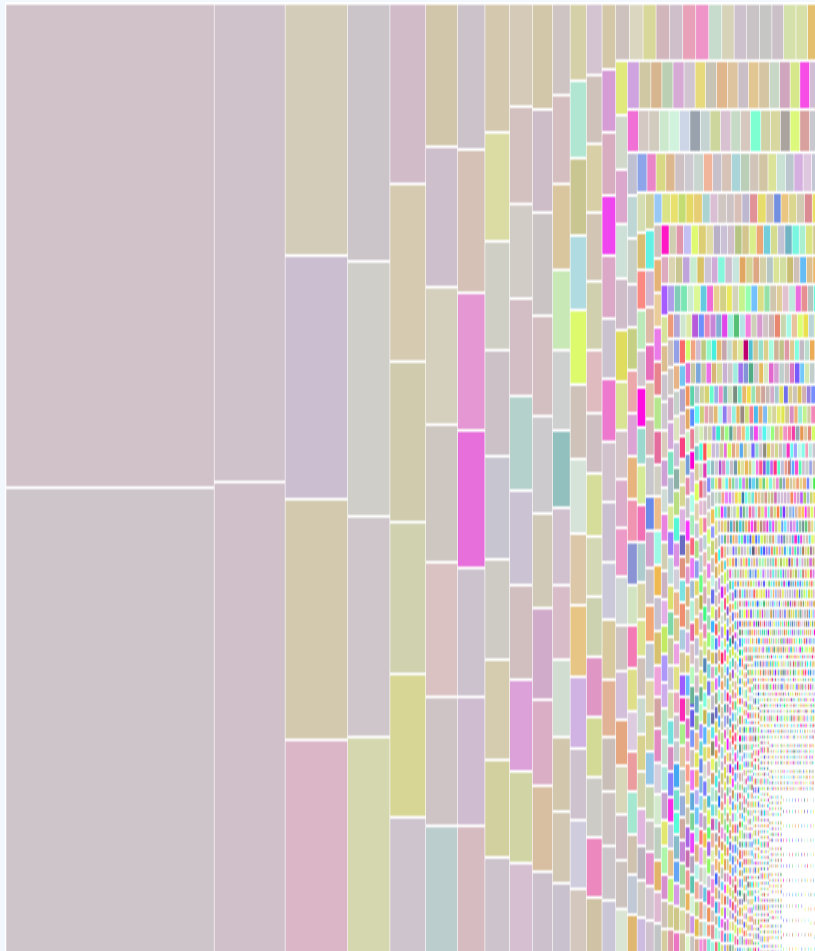


# Network: Blue-Grey-Yellow areas

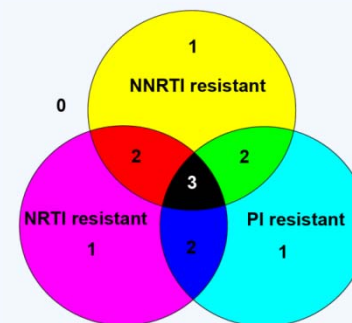
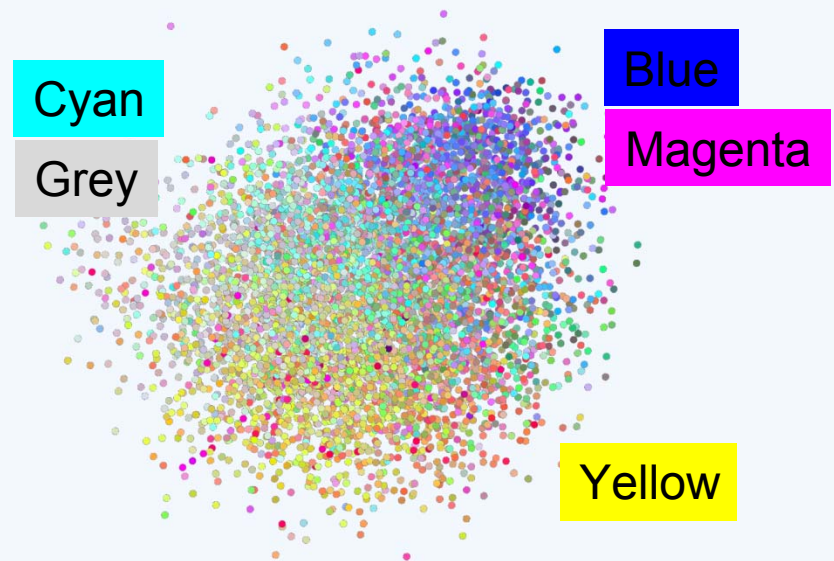


# Resistance coloring of the two displays

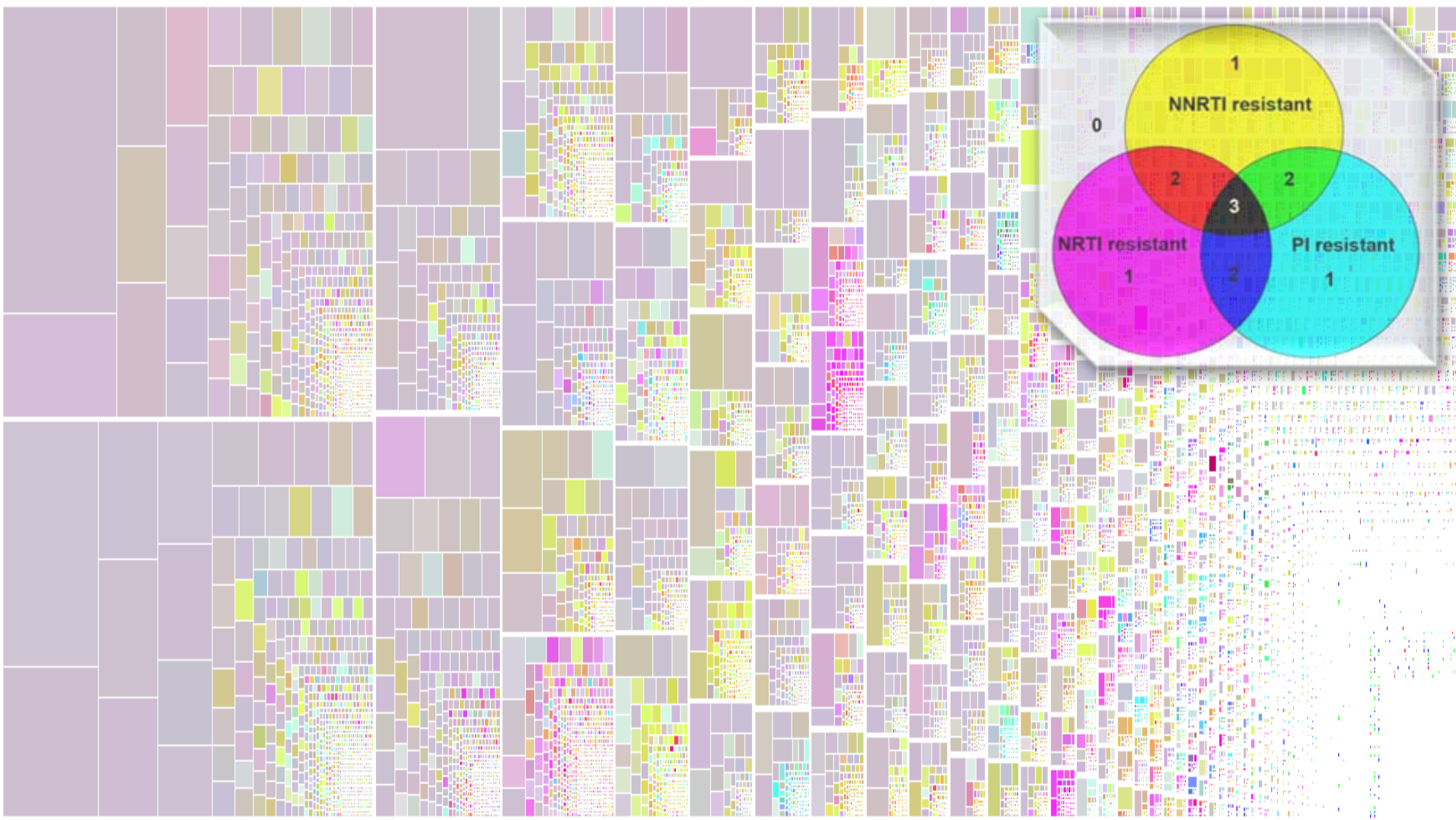
## Treemap



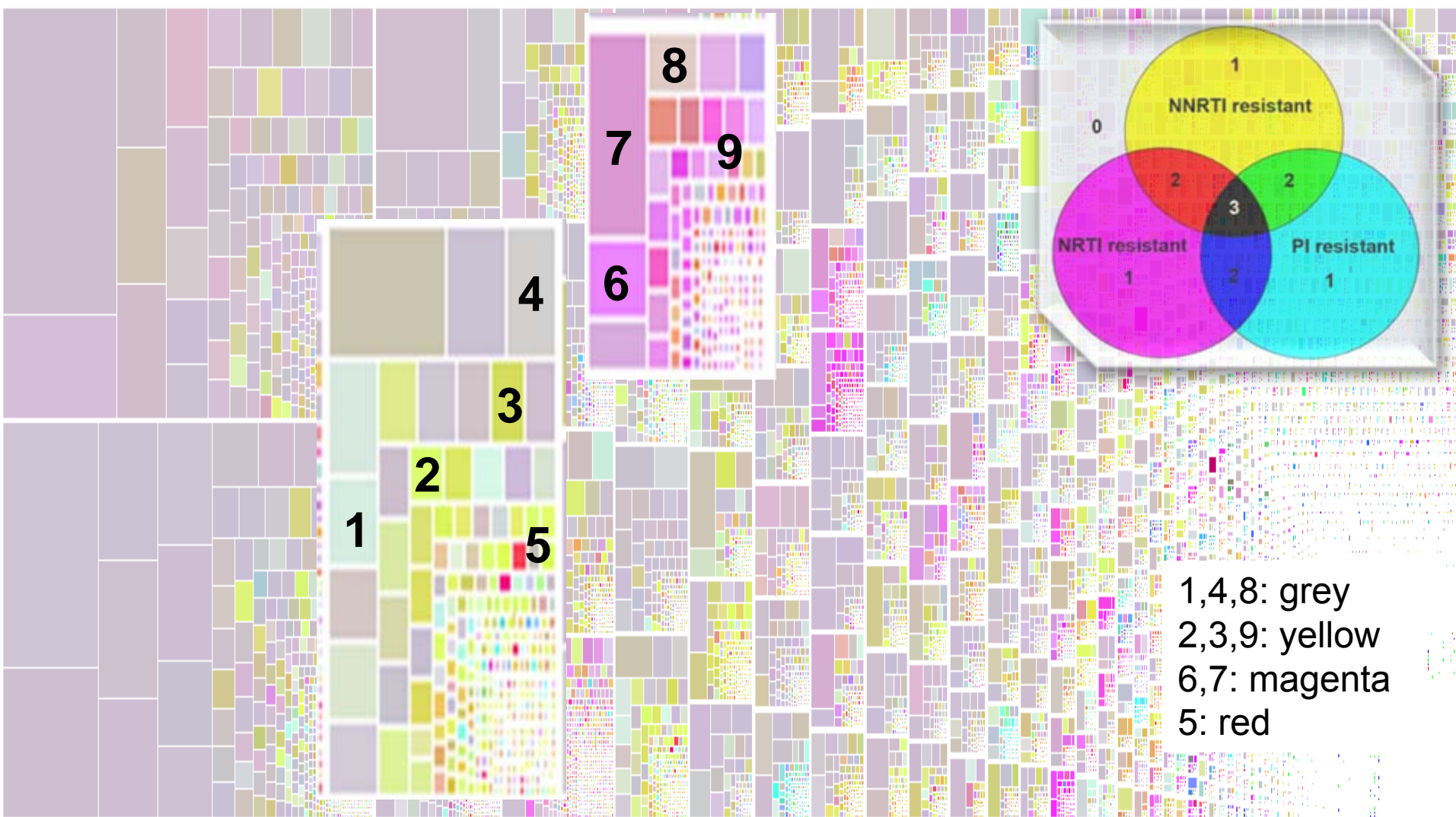
## Network



# Treemap: showing 44,014 clusters



# Treemap: showing 44,014 clusters



# Conclusions

- **4 drug class level phenotypes (factors) explain the phenotype variability: one PI factor, one NNRTI factor and two NRTI factors.**
- **Genetic diversity of ~5,000 clusters can be displayed in a 2D network. From a high level view, genetically similar clusters have the same resistance profile.**
- **Genetic diversity of ~50,000 clusters can be visualized using a treemap. From a low level view, sequence similarity does not necessarily lead to the same resistance profile.**