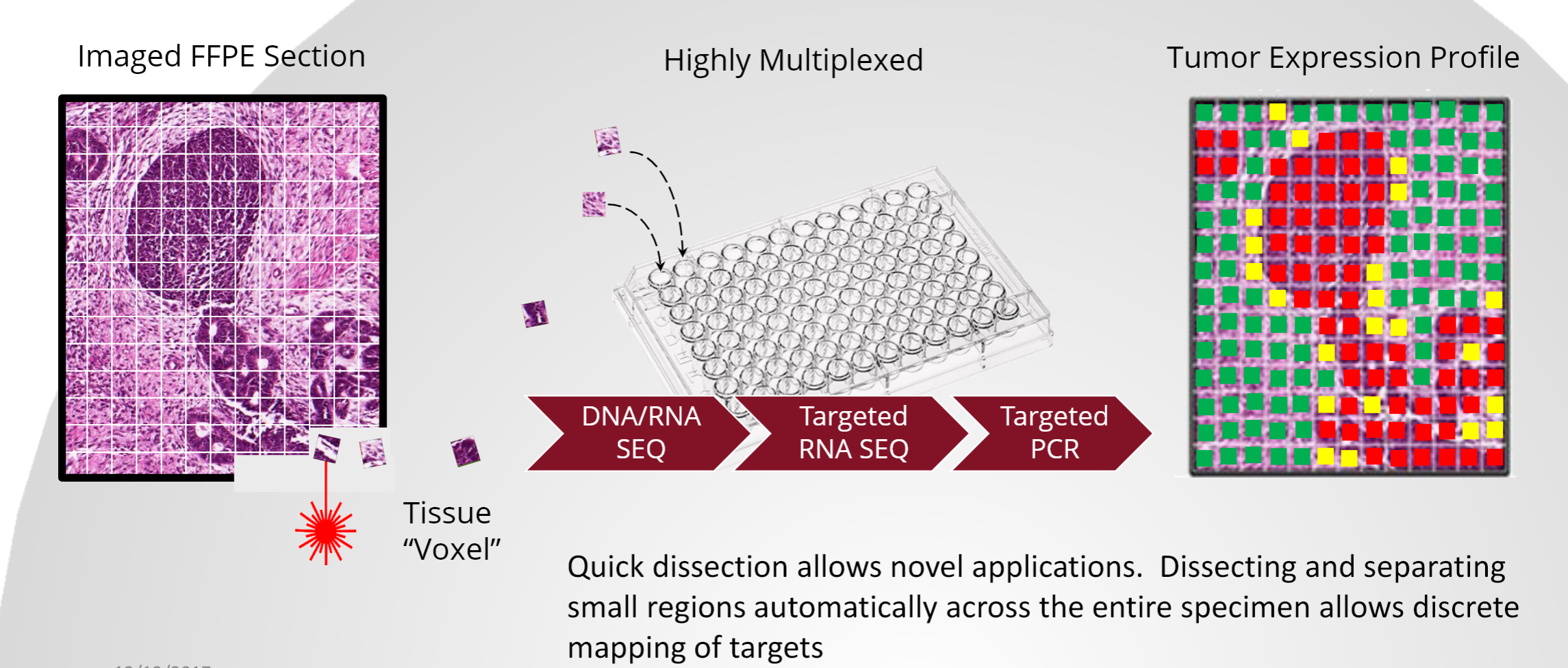
Tiling Training Set Setup

The following lists the key assumptions used in creating this sample dataset. These assumptions are representative of future findings, however, when we implement this type of analysis on a full scale, we will not know which genes are targets nor if they will be physically grouped or randomly distributed. Therefore, an ideal algorithm would be able to identify these relationships itself without knowing which genes to look for.



The data in the excel sheet is formatted such that individual tables for a single gene are physically measured microregions like the image above. The results for all genes are compiled into a large table on each Excel Tab. There are three groups of data: 1) healthy 2) Cancer, non-responder 3) Cancer, Responder. (responder indicating if the patient positively responded to a treatment). The Responder and Non-responder groups have 5 patients each (samples), and each patient sample has 10 tissues.

**Normal**

* Base levels of expression for each gene, with variation of ~3-6 Ct
* No “tumor” areas

**Responder**

* Sections have tumor areas
* Some genes co-localized and upregulated
  + CDKN2A, NOTCH, EGFR, IDH1, KRAS
* Genes co-localized, downregulated
  + SMARCB1, APC, GNA11, PIK3CA
* Genes upregulated surrounding tumor area
  + HRAS, SMO
* Genes randomly dispersed, upregulated
  + ABL1, JAK3

**Nonresponder**

* Sections have tumor areas
* Genes co-localized and upregulated
  + CDKN2A, EGFR, IDH1, KRAS
  + No upregulation in NOTCH
* Genes co-localized, downregulated
  + SMARCB1, APC, PIK3CA
  + No downregulation in GNA11

Goal is to detect difference in:

* tumor vs. normal
  + detect tumor areas on samples
* responder vs. non-responder
  + non-responders lack upregulated NOTCH in tumor area, downregulated GNA11 in tumor area, upregulated HRAS and SMO surrounding tumor area, and dispersed, upregulated ABL1 and JAK3